
Stata 学术论文专题

视频教程

论文集

连玉君
中山大学 岭南学院
arlionn@163.com
<http://goo.gl/tRXba>

目 录

Chang and Wong (2009, JCF).....	1
Chang, E. C., S. M. L. Wong, 2009, Governance with multiple objectives: Evidence from top executive turnover in China, <i>Journal of Corporate Finance</i> , 15 (2): 230-244.	
Cleary (1999, JF).....	16
Cleary, S., 1999, The Relationship between Firm Investment and Financial Status, <i>Journal of Finance</i> , 54 (2): 673-692.	
Faulkender and Wang (2006, JF)	37
Faulkender, M., R. Wang, 2006, Corporate Financial Policy and the Value of Cash, <i>Journal of Finance</i> , 61 (4): 1957-1990.	
Fazzari et al. (1988, BPEA)	71
Fazzari, S., R. Hubbard, B. Petersen, A. Blinder, J. Poterba, 1988, Financing Constraints and Corporate Investment, <i>Brookings Papers on Economic Activity</i> , 1988 (1): 141-206.	
Flannery and Ragan (2006, JFE)	138
Flannery, M. J., K. P. Rangan, 2006, Partial adjustment toward target capital structures, <i>Journal of Financial Economics</i> , 79 (3): 469-506.	
Hansen (1999, JE)	176
Hansen, B., 1999, Threshold Effects in Non-dynamic Panels: Estimation, Testing, and Inference, <i>Journal of Econometrics</i> , 93 (2): 345-368.	
Kumbhakar and Christopher (2009, JPA)	200
Kumbhakar, S., F. Christopher, 2009, The effects of bargaining on market outcomes: Evidence from buyer and seller specific estimates, <i>Journal of Productivity Analysis</i> , 31 (1): 1-14.	

Lian et al. (2011, FBRC, PSM)	214
Lian, Y., Z. Su, Y. Gu, 2011, Evaluating the effects of equity incentives using PSM: Evidence from China, <i>Frontiers of Business Research in China</i> , 5 (2): 266-290.	
Love and Zicchino (2006, QREF, PVAR)	239
Love, I., L. Zicchino, 2006, Financial development and dynamic investment behavior: Evidence from panel VAR, <i>Quarterly Review of Economics and Finance</i> , 46 (2): 190-210.	
Opler et al. (1999, JFE)	260
Opler, T., L. Pinkowitz, R. Stulz, R. Williamson, 1999, The Determinants and Implications of Corporate Cash Holdings, <i>Journal of Financial Economics</i> , 52(1): 3-46.	
Wang (2003, JBES, Het-SFA)	304
Wang, H., 2003, A Stochastic Frontier Analysis of Financing Constraints on Investment, <i>Journal of Business and Economic Statistics</i> , 21 (3): 406-419.	
连玉君和苏治 (2009, 管理评论)	318
连玉君, 苏治, 2009, 融资约束、不确定性与上市公司投资效率, 管理评论, 1: 19-26.	
卢洪友、连玉君和卢盛峰 (2011, 经济研究)	326
卢洪友, 连玉君, 卢盛峰, 2011, 中国医疗服务市场中的信息不对称程度测算, 经济研究, (4): 94-106.	
叶德珠、连玉君和黄有光 (2012, 经济研究)	339
叶德珠, 连玉君, 黄有光, 李东辉, 2012, 消费文化、认知偏差与消费行为偏差, 经济研究, (2): 80-92.	
连玉君和钟经樊 (2007, 南方经济)	353
连玉君, 钟经樊, 2007, 中国上市公司资本结构动态调整机制研究, 南方经济, (1): 23-38.	
连玉君 (2011, 面板数据模型, 书稿)	369
连玉君 (2010, 一份不太长的 Stata 简介)	451



Governance with multiple objectives: Evidence from top executive turnover in China

Eric C. Chang^{a,*}, Sonia M.L. Wong^b

^a Faculty of Business and Economics, The University of Hong Kong, Pokfulam Road, Hong Kong

^b Department of Finance and Insurance, Lingnan University, Tuen Mun, Hong Kong

ARTICLE INFO

Article history:

Received 25 October 2006

Received in revised form 3 October 2008

Accepted 7 October 2008

Available online 12 November 2008

JEL classifications:

P31

P34

G34

Keywords:

Managerial turnovers

Multiple firm objectives

Firm performance

State ownership

ABSTRACT

We examine the relationship between Chief Executive Officer (CEO) turnover and the performance of listed Chinese firms and obtain two results. First, we find a negative relationship between the level of pre-turnover profitability and CEO turnover when firms are incurring financial losses, but no such relationship when they are making profits. Second, there is an improvement in post-turnover profitability in loss-making firms, but no such improvement in profit-making firms. These results indicate the existence of a time-varying objective function, whereby shareholders have a greater incentive to discipline their CEOs on the basis of financial performance when their firms are incurring financial losses rather than profits.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Despite the massive waves of privatization in recent decades, many firms around the globe today, particularly in vital industries such as telecommunications, energy, public utilities, and banking, still remain at least partially state-owned.¹ Notwithstanding its importance, empirical evidence on the monitoring of managers in state-owned firms remains scarce. The existing studies on managerial turnover focus primarily on firms that are controlled by private owners. A substantial body of literature shows that forced managerial turnover is preceded by a large and significant decline in financial performance and is then followed by improved performance, which reflects the effectiveness of the various corporate control mechanisms at work in these firms (e.g., Kaplan, 1994; Denis and Denis, 1995; Denis et al., 1997; Kang and Shivdasani, 1995; Huson et al., 2001; Volpin, 2002; McNeil et al., 2004; Huson et al., 2004).²

This study examines the relationship between managerial turnover and firm performance in China's listed firms in which the majority of controlling shareholders are state-owned entities. Unlike the shareholders of typical listed firms, state shareholders are not real owners, but rather bureaucrats who run the firms on behalf of the government. As agents of the government, their decisions are subject to the influence and control of the government, which tends to use a firm's resources to promote social and political objectives (Shleifer and Vishny, 1994, 1997; Dixit, 1997). Similar to managers in the traditional agency model, state shareholders can possess multiple personal interests, such as the accumulation of personal wealth, job security, and others (Alchian, 1965; Shleifer and Vishny, 1997). A salient characteristic of state shareholders is therefore the existence of multiple

* Corresponding author. Tel.: +852 28578347.

E-mail address: ecchang@business.hku.hk (E.C. Chang).

¹ Bauer (2005) documents that 49.2% of fixed access lines were still operated by either fully or partially state-owned telecommunication operators at the end of 2004. Based on a study of the 10 largest banks in 92 countries, La Porta et al. (2002) document that 42% of their assets are controlled by state-owned banks.

² Conflicting evidence is provided by Dalton and Kesner (1983), Friedman and Singh (1989), and Davidson et al. (1990).

objectives. We explore how the existence of these multiple objectives on the part of state shareholders affects the relationship between managerial turnover and firm financial performance.

According to Jensen (2001), it is impossible to maximize more than one objective at any given time if there are tradeoffs among various objectives. Therefore, shareholders must place different levels of importance on these objectives and integrate them into a single objective function. We expect the level of importance that the state shareholders of listed Chinese firms attach to the objective of improving firm financial performance to be a function of actual firm performance and that they will attach greater importance to this objective when their firms are experiencing financial losses than they will in times of profit. We reason that, in general, state shareholders do not possess a strong incentive to maximize financial performance because the pursuit of political and/or personal objectives often lessens ex post firm profit (Dixit, 1997; Bai et al., 2000; Chang and Wong, 2004; Bai et al., 2006).³ However, controlling state shareholders are likely to face greater government pressure to improve financial performance when their firms are experiencing financial losses. This is because the government may eventually have to bail out loss-making firms through the provision of fiscal subsidies and/or low-cost loans (Qian and Roland, 1998). Therefore, the shareholders of loss-making firms, who are obliged to subscribe to the government's objective function, will have little choice in such circumstances but to attach greater importance to the objective of improving financial performance. The external pressure from government to improve financial performance also limits the latitude and resources available to state shareholders to serve their own personal objectives, such as on-the-job consumption and the accumulation of personal wealth, thus further increasing their incentive to dismiss relatively poorly performing managers. As a result, we expect the shareholders of firms in a loss-making state to have a greater incentive than their counterparts in profit-making state to monitor managers on the basis of firm performance.

In addition to being a determinant of managerial turnover, we expect the existence of multiple objectives in shareholders' objective functions to also affect post-turnover performance changes. This is because the different levels of importance to firm performance in shareholders' objective functions will lead to different incentives to fire the incumbent managers and to hire and monitor the new managers, which will in turn affect post-turnover performance changes.⁴ When shareholders attach a higher level of importance to firm performance, they have a greater incentive to remove the poorly performing managers and to identify managers with the ability to improve that performance. They also have a greater incentive to monitor new managers on the basis of it. The result is a greater likelihood of post-turnover operating performance improvements for loss-making firms. When shareholders attach a low level of importance to firm performance, in contrast, that performance may not be the major cause of managerial turnover in the first place. Instead, it may be due to such non-performance reasons as organizational politics or personal considerations. The existing literature suggests that factors other than performance (e.g., social and political factors) also play an important role in determining managerial turnover in private firms (Fredrickson et al., 1988; Gibelman and Gelman, 2002; Shen and Cannella, 2002). Such turnover is even more likely to occur in state-controlled firms, given the existence of multiple objectives and the weak profit motive on the part of state owners. As performance is not the major cause of managerial turnover, new managers are more likely to be selected and monitored on the basis of political connections and personal favoritism than on their ability to improve firm performance. This, in turn, means that post-turnover performance improvements are less likely in profit-making firms.

Based on a sample of the Chief Executive Officer (CEO) turnovers experienced by listed Chinese firms between 1995 and 2001, we provide two pieces of evidence on the turnover-performance relationship. First, there is a significant negative relationship between pre-turnover profitability and CEO turnover when firms are experiencing financial losses, but no such relationship when they are making profits. Second, there is significant improvement in the post-turnover profitability of loss-making firms, but not in that of profit-making firms. Overall, our results indicate that there are differences in the turnover-performance relationship between loss- and profit-making firms.

Our study provides a useful addition to the existing literature on the monitoring activities of state-owned firms. Groves et al. (1995) examine the relationship between labor productivity and managerial turnover for a sample of state-owned firms in China. They offer evidence that the managerial turnover is not associated with ex ante labor productivity but is followed by a significant increase in productivity. Kole and Mulherin (1997) study the managerial turnover and performance of 17 U.S. firms controlled by the federal government during and after World War II. They find that the turnover and performance did not differ significantly from those of private-sector firms. Based on a sample of Czech firms, Claessens and Djankov (1999) find that managers appointed by state asset management agencies are associated with weaker performance improvements than are managers appointed by private owners. Firth et al. (2006) find that managerial turnover in listed Chinese firms is inversely related to a firm's profitability. However, they do not find evidence of post-turnover performance improvement. Also based on a sample of listed Chinese firms, Kato and Long (2006) show that CEO turnover is significantly and negatively related to a firm's financial performance.

All of the aforementioned studies on the relationship between managerial turnover and firm performance in state-owned firms assume that shareholders have the same incentive structure to discipline managers under all circumstances. In this study, we

³ Bai et al. (2000) further demonstrate that, in addition to having a weak profit motive, state shareholders provide their managers with weak-profit incentives. Consistent with this theory, the compensation schemes of CEOs in China's listed firms are characterized by weak profit incentives, whereby the main component is a low and undifferentiated civil-service-ranked salary. Stock-based incentives are also weak, given that the average shareholding of managers in listed firms, as of the end of 1999, was only 0.006%, and stock options were non-existent until the early 2000s (Chang and Wong, 2004).

⁴ The improved management hypothesis and the scapegoat hypothesis on post-turnover performance changes both suggest that shareholders have an invariant incentive structure to hire and monitor the new managers. The improved management hypothesis suggests that managers differ in quality and that shareholders always have the incentive to identify and hire a new, superior manager who is capable of improving a firm's performance (Denis and Denis, 1995; Huson et al., 2004). The scapegoat hypothesis holds that the quality of managerial abilities does not vary significantly across different individuals, and therefore a newly hired manager is unable to alter a firm's fundamentals and improve its performance (Khanna and Poulsen, 1995; Huson et al., 2004).

assume, as well as provide evidence to support the assumption, that shareholders have a time-varying objective function that depends on firm performance. By estimating the performance–turnover sensitivities for profit- and loss-making firms separately, we show that the negative relationship between pre-turnover firm performance and turnover found by Firth et al. (2006) and Kato and Long (2006) exists only in loss-making firms, not in profit-making firms. Furthermore, we find a significant improvement in post-turnover profitability in loss-making firms.⁵

China is becoming increasingly integrated into the global financial market, with a growing number of Chinese firms seeking listing on overseas exchanges and an increasing number of international institutional investors being attracted to China's domestic market. Despite recent drastic attempts to reduce the percentage of state shareholding, China's government is still unlikely to fully privatize all of its listed firms in the foreseeable future (Wong, 2006). In light of this, the question of whether and how the state shareholders of listed firms in China have an incentive to exercise effective corporate control in order to strive for the maximization of their shareholders' wealth should be of great interest to international investors.

The remainder of this paper is structured as follows. Section 2 provides a brief discussion on the corporate governance and incentive structure of the state shareholders of China's listed firms. Section 3 discusses the data and research methods, and Section 4 presents the empirical results and robustness checks. Finally, Section 5 presents the study's conclusions.

2. Corporate governance and state shareholder incentive in China's listed firms

The majority of China's listed firms are controlled by state shareholders who retain their dominant control through the ownership of about two-thirds of total equity in the form of non-tradable state-owned shares (Sun and Tong, 2003). State-owned shares are officially classified into state shares and legal person shares.⁶ Earlier studies have taken the view that state shares are owned by government agencies and legal person shares are owned by state-owned commercial firms (for example, Xu and Wang, 1999; Sun and Tong, 2003; Firth et al., 2006). Recent studies, however, indicate that the official classification scheme does not yield valid measures of ownership identity. Delios et al. (2006), for example, find that the first-level owners of state and legal person shares can be either government agencies or state-owned commercial firms and that 39.7% of legal person shares are actually held by state asset investment bureaus. Liu and Sun (2005) find that the official classification of shares fails to identify the ultimate owners of listed firms and that legal person shares can ultimately be owned by either private or state entities.

China's Company Law of 1992 requires listed Chinese firms to adopt a formal governance structure whereby CEOs are monitored by boards of directors. In spite of this regulation, however, local governments maintain control over partially privatized listed firms not only through the formal voting power vested in the controlling shareholdings, but also through the formal authority to approve appointments and dismiss key personnel as recommended by the boards of directors (Wong et al., 2004). As a result, the decisions to appoint and remove CEOs in China's listed firms are in the hands of state controlling shareholders, which are ultimately controlled by the governments through their administrative controls over the state shareholders and their authority to approve CEO appointments.

Two different sets of assumptions have been made in studies analyzing the objective functions of state shareholders. The first assumption is that state shareholders are good stewards of government who seek to serve its interests—the main one being to promote social and political goals—such as by correcting market failures and providing additional employment opportunities and social security to the public (Shleifer and Vishny, 1994, 1997; Dixit, 1997). The second assumption is that these shareholders are motivated by self-interest and will, therefore, use state-owned firm resources to promote their own personal interests (Shleifer and Vishny, 1994, 1997; Jones, 1985; Krueger 1990). Both views suggest that state shareholders tend to have a weak incentive to maximize firm profits because they enjoy control rights but not cash flow right. Furthermore, as the pursuit of most political and personal objectives often detracts from a firm's financial performance, these shareholders face a tradeoff between the pursuit of political and personal objectives on the one hand and the delivery of higher ex post firm financial performance on the other (Dixit, 1997; Bai et al., 2000, 2006; Chang and Wong, 2004). Shareholders with a greater incentive to pursue political and personal objectives will attach less weight to financial performance in their objective function, and thus will have less incentive to monitor managers on the basis of it.

However, state shareholders will attach greater weight to financial performance when their firms are experiencing financial losses. Note that, although the government uses firms to serve its political objectives at the expense of their financial performance, it also has incentives to minimize the amount of the financial losses that firms incur. This is because the occurrence of consistent and significant financial losses will eventually backfire and ultimately place a burden on government budgets and on state-owned banks when the firms have to be bailed out with either government subsidies or bank loans (Qian and Roland, 1998). Loss-making firms thus face tremendous government pressure to improve their operational efficiency, and hence financial performance.

⁵ In addition to these published papers, there are three unpublished manuscripts that also examine the relationship between managerial turnover and firm performance in listed Chinese firms. Chen and Wang (2004) examine the relative effectiveness of government agencies and state-owned firms in monitoring CEOs, whereas Chen et al. (2006) analyze how the effects of the delegation of control rights affect performance–turnover links. Cheng et al. (2007) examine the responsiveness of turnover decisions to different accounting performances. All of these working papers estimate performance–turnover links for whole sample firms regardless of their financial performance.

⁶ According to official classification, state shares are created as a consequence of a government agency contributing its assets to the formation of a shareholding firm. Legal person shares, in contrast, represent the contribution by government-invested state-owned enterprises (SOEs) of their legally owned assets to the formation of a shareholding firm. To maintain the dominance of state ownership, these two types of shares were formerly not allowed to be traded on China's two stock exchanges. However, reforms were introduced in August 2005 to allow these shares to be traded on the exchanges after the state shareholders pay adequate compensation to and obtain consent from individual private investors.

Table 1

Annual CEO turnover rate and performance in China's listed firms: 1995–2001

	1995	1996	1997	1998	1999	2000	2001	1995–2001
Number of listed firms	307	510	715	821	918	1054	1136	5461
Total number of CEO turnovers	47	81	136	210	273	332	314	1393
Annual turnover rate	15.31%	15.88%	19.02%	25.58%	29.74%	31.50%	27.64%	25.51%
Number of CEO turnovers after consolidation	44	80	130	196	254	303	284	1291
Annual turnover rate after consolidation	14.33%	15.69%	18.18%	23.87%	27.67%	28.75%	25.00%	23.64%

This table reports CEO turnovers in China's listed firms from 1995 to 2001. The number of listed firms includes all non-financial firms listed on the A-share markets of the Shanghai and Shenzhen Stock Exchanges. The total number of CEO turnovers refers to the number of CEO turnovers, including multiple turnovers during a single year. The number of CEO turnovers after consolidation represents the number of CEO turnovers after multiple CEO turnovers for a given firm in a given fiscal year is consolidated into one observation.

In addition to concerns about their careers, state shareholders also have a stronger personal motive to improve firm performance during periods of financial loss than during those of financial gain. There exists an obvious incompatibility between delivering a higher ex post financial performance and the current and future resources available for state shareholders for the pursuit of their private interests. For firms that are already in the loss-making state, if there is no improvement in a firm's operation efficiency, the need to deliver higher financial performance would also significantly reduce the amount of resources available for state shareholders to serve their own interests. The state shareholders of loss-making firms, who are under government pressure to improve financial performance, therefore also have the personal incentive to replace poorly performing CEOs as a way of restoring the resource base for their pursuit of those interests. In this situation, the ability to improve performance will be an important consideration in the selection and appointment of replacement CEOs, who will also be more closely monitored with a focus on performance improvement. As a result, we hypothesize that managerial turnover in loss-making firms is more likely to be followed by significant improvements in performance.

In contrast, when firms are profitable, neither the government nor state shareholders seeks to maximize firm performance, and thus managerial turnover is more likely to have non-performance-related causes. As a result, the state shareholders of profit-making firms are less likely to have the incentive to seek out CEOs with the ability to improve firm performance and less likely to monitor new CEOs on the basis of that performance. As suggested by Zhang (2006), and Zhang (1998), the state shareholders in China's listed firms have little incentive to select CEOs who will ensure that firms are efficiently and profitably operated. Zhang (1998) further suggest that state shareholders too often base their selections of CEOs on personal connections (*guanxi*) because they want to appoint friendly CEO to facilitate their tunneling of firm resources. This shareholder incentive problem suggests that managerial turnover in profit-making firms is less likely to be followed by significant performance improvements.

3. Data, sample selection, and research methods

3.1. Data sources and classification of managerial turnovers

Our study is based on all of the non-financial firms listed on the Shanghai and Shenzhen Stock Exchanges from 1995 to 2001. We obtain our data on CEO turnover from the China Corporate Governance Research Database (CCGRD) developed by the GTA Information Technology Co. Persons holding the formal title of either General Manager or Chief Executive are identified as CEOs. Table 1 documents the extent of CEO turnover for all of the listed firms. Of the 1136 non-financial firms listed on the exchanges at

Table 2

Stated reasons for CEO turnover in China's listed enterprises

	Full sample		Consolidated sample	
	Number	Percentage of sample	Number	Percentage of sample
1. Change of job	442	31.73%	406	31.45%
2. Retirement	35	2.51%	33	2.56%
3. Contract expiration	255	18.31%	247	19.13%
4. Change in controlling shareholders	78	5.60%	78	6.04%
5. Resignation	175	12.56%	154	11.92%
6. Dismissal	64	4.59%	53	4.11%
7. Health	44	3.16%	41	3.18%
8. Personal reasons	4	0.29%	4	0.31%
9. Corporate governance reform	207	14.86%	198	15.34%
10. Legal disputes	11	0.79%	10	0.77%
11. No reason given	64	4.59%	56	4.34%
12. Completion of acting duties	14	1.01%	11	0.85%
Total number of observations	1393	100.00%	1291	100.00%

This table reports the frequencies of the stated reasons for CEO turnovers in China's listed firms between 1995 and 2001. The full sample refers to the total number of CEO turnovers, including multiple turnovers during a single year. The consolidated sample is obtained by consolidating multiple changes in a year into one single observation.

Table 3

Destinations of departing CEOs

Destination	No. of observations	Percentage of sample
1. Information unavailable	265	28.46%
2. New position ranked lower than CEO position	199	21.37%
3. CEO position taken up at an unlisted, smaller firm	11	1.18%
4. Arrested or under investigation	27	2.90%
5. Important government position taken up	26	2.79%
6. Remaining as board chairman or vice chairman	158	16.97%
7. Promoted to board chairman or vice chairman	173	18.58%
8. CEO position taken up at another listed firm or parent firm	64	6.87%
9. Health problems	5	0.54%
10. Going abroad to study	3	0.32%
Total	931	100.00%

This table reports the destinations of departing CEOs for which the stated reasons for turnovers fall under the categories of change of job, contract expiration and resignation, dismissal, personal reasons, completion of acting duties, and no reason given. This information is obtained from the China Economic News Database and the China's Listed Firms Database provided by Infobank, the annual reports of China's listed firms, the China's Listed Firms Database provided at <http://stock.sina.com.cn/>, and internet materials available at <http://www.baidu.com>.

the end of 2001, 755 had undergone at least one turnover between 1995 and 2001; the total number of turnovers was 1393. There was a significant increase in the annual turnover rate during the period, rising from 15.31% in 1995 to 27.64% in 2001. The average annual turnover rate is 25.51%, which is significantly higher than the rates documented by Denis and Denis (1995) and Huson et al. (2004) for U.S. firms (12.7% and 9.3%, respectively) and Kang and Shivdasani (1995) for Japanese firms (12.88%). In line with previous studies, we consolidate multiple turnovers for a given enterprise in a given fiscal year. Thus, if a firm underwent two or more turnovers in the same year, only one is recorded. This reduces the number of turnovers from 1393 to 1291 and the average annual turnover rate from 25.51% to 23.64% in our consolidated sample.

The CCGRD provides information on the reasons stated for a turnover (if any): (1) change of job, (2) retirement, (3) contract expiration, (4) change in controlling shareholders, (5) resignation, (6) dismissal, (7) health, (8) personal reasons, (9) corporate governance reform, (10) legal disputes, (11) no reason given, and (12) completion of acting duties. Table 2 summarizes the distribution of turnovers across different stated reasons for the full and consolidated samples. In the full sample, change of job is the most commonly stated reason, accounting for 31.73% of the turnovers. The second most commonly stated reason is contract expiration, which accounts for 18.31% of the turnover, and the third is corporate governance reform (14.86%).⁷ Only 4.59% the turnovers fall in the dismissal category. Our consolidated sample shows a similar distribution of turnovers across different stated reasons.

To assess the effectiveness of the corporate control exercised by shareholders, we distinguish between forced and non-forced turnovers because only the former reflect the disciplinary efforts of shareholders. As many researchers (e.g., Denis and Denis, 1995; Kang and Shivdasani, 1995; Huson et al., 2004) have recognized, it is difficult to distinguish between forced and non-forced turnovers based on publicly available information because very few press reports indicate clearly whether a turnover was voluntary or forced. We face similar identification problems. For example, a turnover for which the stated reason is a job change can either be forced or non-forced depending on the new job that the departing manager takes up. The turnover is likely to be non-forced if the new job is a better one, but forced if the new job is less desirable than the old one.

We adopt the following procedures and assumptions to identify forced turnovers. We first exclude from the forced turnover sample the 360 turnovers for which the stated reasons are retirement, health (including death), corporate governance reform, and a change in controlling shareholders.⁸ Additionally, we exclude those cases that involve legal disputes because these turnovers are not directly initiated by state shareholders as a result of their normal monitoring activities. For the remaining turnovers, we trace the destinations of the departing managers to ascertain their nature. We exclude those turnovers in which the departing managers subsequently take up a position that is better than their previous managerial position. Our search for the destinations of departing CEOs is based on five data sources: the annual reports of the firms, Infobank's China Economic News Database, Infobank's China's Listed Firms Database, China's Listed Firms Database, which is available at <http://www.sina.com.cn>, and internet materials available at <http://www.baidu.com>. The results are reported in Table 3.

Of the 931 turnovers in our original sample, we excluded 456 cases that we considered to be voluntary departures. These include 26 turnovers in which the CEOs left their firms to take up important government positions such as city governors and provincial leaders; 158 cases in which the departing CEOs retained their positions as board chairman or vice-chairman; 173 turnovers in which the CEOs were promoted to the position of board chairman or vice-chairman; 64 cases in which the departing CEO took up a new managerial position at another listed firm or at the listed firms' parent group; five cases in which they left

⁷ This refers to two types of turnovers that are unique to China's listed firms. The first type of turnover involves the division of the combined position of chairperson of the board of directors and CEO into two separate positions (i.e., the CEO resigns from his managerial position, but retains the chairperson position) with the stated objective of improving corporate governance. The second type of turnover refers to those that result from regulations imposed by the China Securities Regulatory Commission in 1999 that require CEOs who also hold senior managerial positions in the parent firms to retire from either position to minimize the conflict of interest between holding firms and minority shareholders.

⁸ We exclude those turnovers for which the stated reason was corporate governance reform because the departing CEOs either retained their positions as chairpersons of the board of directors or their key managerial positions in the parent firms.

Table 4
Summary statistics of variables

Variables	Number	Mean	Standard Deviation	Minimum	Maximum
<i>Panel A: Control variables</i>					
List	3916	3.763	2.211	1.000	10.000
Age	3916	48.242	7.188	26.000	72.000
Tenure	3916	2.435	1.506	0.000	11.500
Duality	3916	0.283	0.451	0.000	1.000
Leverage	3916	0.000	0.166	−0.457	0.615
Size	3916	20.815	0.868	18.543	24.784
State	3916	0.140	0.347	0.000	1.000
<i>Panel B: Performance variables</i>					
ROA	3916	0.040	0.056	−0.288	0.219
IROA	3916	−0.003	0.054	−0.325	0.209
MROA	3916	0.047	0.051	−0.271	0.240
MIROA	3916	0.001	0.050	−0.325	0.191

This table reports the number of observations, the mean, standard deviation, minimum, and maximum values for the variables used in our models. List is the number of years that a firm has been listed. Age is the age of a CEO. *Tenure* is the number of years a CEO has been in his or her current position. *Duality* is a dummy variable that equals 1 if a CEO is also a board chairperson and 0 otherwise. *Leverage* is the industry-adjusted capital structure of a listed firm, measured as the ratio of the book value of total debt over the book value of total assets less the median ratio in its industry. *Size* is the size of a listed firm, measured as the natural logarithm of the book value of total assets. *State* takes the value of 1 if a firm has a higher percentage of state shares relative to legal person shares, and 0 otherwise. *ROA* is the unadjusted return on assets, measured as the ratio of pretax operating income to the beginning period book value of total assets. *IROA* is the industry-adjusted return on assets, measured as ROA less the median value of ROA for all firms in the same industry. MROA is the moving average of ROA over a CEO's tenure. MIROA is the moving average of IROA over a CEO's tenure.

because of health problems; 27 cases in which they were arrested or placed under legal investigation⁹; and three cases in which they were reported as having gone abroad for further education.

We treat the remaining 475 turnovers as forced. These include 199 cases in which the departing CEOs took up new positions that were less prestigious than their former positions, 11 cases in which they took up managerial positions at unlisted and/or smaller-sized firms, and 265 cases in which we were unable to trace the destinations of the departing CEOs. We treat the turnovers for which no such information is available as forced because our data sources provide comprehensive information on the business activities of the major firms in China. It is highly unlikely that there would be no information available if a departing CEO were to take up a position better than his/her previous role. For this sample of forced turnovers, we exclude 77 cases in which the CEO's tenure was less than one year because CEOs with such a short tenure are unlikely to have left on account of poor performance. We add to our turnover sample two cases for which the stated reason was retirement, but the age of the departing CEO was less than 55. Our final sample contains 400 cases of forced turnovers, which represents 30.98% of all turnovers. This proportion is higher than the rates reported by Denis and Denis (1995) and Huson et al. (2004) for U.S. firms (13.3% and 18% respectively) and by Kang and Shivdasani (1995) for Japanese firms (24.14%).

3.2. Regression models for determinants of turnovers

We employ a logit regression model to examine the sensitivity of turnover to firm performance.

$$\text{Probability(Forced CEO turnover)} = f(\text{Performance, Control Variables}) \quad (1)$$

The dependent variable is a dummy variable that equals 1 if there was a forced turnover during the period in question. *Performance* denotes four performance measures. The first is the unadjusted return on assets (ROA), measured as the ratio of year-end pretax operating income to the beginning period book value of total assets. The second is the industry-adjusted return on assets (IROA), measured as ROA less the median value of ROA for all firms in the same industry.¹⁰ These two variables measure the recent accounting performance of a listed firm. In addition, we also use the three-year moving average of ROA over a CEO's tenure (MROA) and the three-year moving average of IROA over a CEO's tenure (MIROA) as two other measures of a CEO's performance. As previously discussed, the CEO turnover decisions in China's listed firms are made by bureaucrats who are ultimately concerned with their job security and employment prospects. As bureaucrats are expected to fulfill their assigned responsibilities by following a set of fixed rules and procedures, their decisions tend to be slow and pressure-driven to maintain prudence and conformity (Merton, 1940; Fligstein, 1987). Average performance is a lagging indicator of a manager's overall performance. Li and Zhou (2005) provide evidence that China's central government tends to evaluate provincial leaders by the average economic performance over their tenure rather than annual performance. We also expect that the shareholders of listed Chinese firms also rely more on average performance than on annual performance when evaluating their CEOs.

⁹ These 27 cases are not included the category of legal disputes in CSMAR's database because the stated reasons for these turnovers are related to legal disputes.

¹⁰ We use the industry classification system of the Chinese Securities and Regulatory Commission, which classifies all listed firms into 13 industries.

We follow [Huson et al. \(2001\)](#) in using current year performance if a turnover occurred in the last six months of the year and previous year performance if a turnover occurred in the first six months of the year. The use of a half-year lag allows us to partially deal with the issue of endogeneity. Furthermore, we use half-year rather than full-year lag performance measures because the average tenure of CEOs in China's listed firms is only 2.435 years (see [Table 4](#)).

We focus on accounting performance alone, rather than on stock price performance, for two reasons. First, stock prices are not good indicators of the performance of CEOs because of the prevalence of noise trading in China's emerging stock market. During the period of our investigation (1995–2001), the turnover velocity of stocks, defined as the total transaction volume divided by the total number of tradable shares, was about 500% ([Wong, 2006](#)). [Morck et al. \(2000\)](#) also find that 80% of the stocks listed on China's two exchanges move in the same direction, which suggests that the country's stock prices tend to capitalize on market-level information rather than on firm-specific information. Second, state-owned shares in China are not tradable on the stock exchanges, which suggest that state shareholders are less inclined to discipline their CEOs on the basis of stock prices.

We introduce a set of control variables to eliminate possible confounding effects. First, we control for the departing CEOs' age (Age), as earlier studies have found that managerial turnover is positively related to age (e.g., [Kang and Shivdasani, 1995](#)). Second, prior studies have shown that managerial turnover is negatively related to both a manager's number of years of service (e.g., [Kang and Shivdasani, 1995](#)) and whether he or she also holds the position of board chairperson ([Dalton et al., 1998](#)). Therefore, we control for the number of years that a CEO has served in a listed firm (Tenure) and the existence of a duality structure (Duality). Third, we control for three firm characteristics: capital structure, size, and the ownership nature of the largest shareholders. We control for capital structure (Leverage) and firm size (Size) because debtors play a role in disciplining managers ([Jensen, 1986](#)), and managers are more entrenched in larger firms ([Dalton and Kesner, 1983](#)).¹¹ In addition, [Sun and Tong \(2003\)](#) and [Wang et al. \(2004\)](#) show that the performance of a listed Chinese firm is worse if it is controlled by holders of state shares rather than holders of legal person shares. Although recent studies show that the official classifications of shares fail to identify the true identity of state shareholders ([Liu and Sun 2005](#); [Delios et al., 2006](#)), we nevertheless create a dummy variable (State) to indicate whether a listed firm is controlled by state shareholders or legal person shareholders. State takes the value of 1 if a firm has a greater percentage of state shares relative to legal person shares, and 0 otherwise. Finally, a series of dummy variables indicating the year of the turnover (Year_dum) and a variable indicating the number of years a listed firm has been listed (List) are also used to control for time-specific factors.

We estimate the coefficients for the profit- and loss-making samples separately. A listed firm is classified as profit-making if its current pre-tax operating income is non-negative and loss-making if its current pre-tax operating income is negative.

3.3. Post-turnover performance changes and mean reversion

In our analysis of the post-turnover performance changes, we follow [Huson et al. \(2004\)](#) in using control-group, adjusted-performance measures to isolate the component of a performance change that is attributable to the mean reversion of accounting performance. CEOs may attempt to manage reported earnings. Outgoing CEOs may have the incentive to increase reported earnings to save their jobs, and incoming CEOs may have the incentive to reduce reported earnings immediately upon taking office with the aim of blaming the poor performance on their predecessors. To alleviate the possible biases that are caused by earnings management, we follow [Denis and Denis \(1995\)](#) in using firm performance in both year 0 and year -1 as the benchmarks for evaluating post-turnover performance changes. Depending on the different benchmarks (year 0 or year -1), we create two separate control groups in which the control firms are matched on the basis of firm performance in the corresponding year. We first match each firm that has had a CEO turnover to a firm in the same industry whose firm performance in the corresponding year was within +/- 20% of the sample firm's performance but with no turnover occurring in the event year and in the three years preceding the turnover. If multiple firms satisfy these criteria, then we include the firm whose asset size is closest to that of the turnover firm as the control firm. Of our forced turnover sample, only 285 turnovers have complete financial data for the seven years surrounding the turnovers. For the control group in which year 0 (year -1) was used as the benchmark, we find only 158 (137) firms that are matched on the basis of both performance and industry. We then match performance within the filter bound, regardless of industry, for the remaining turnovers and obtain another 112 (138) performance-matched control firms. In our analysis of the changes in performance following managerial turnovers, we exclude 15 (10) turnover firms from the sample because they had no matching control firms.

4. Empirical results

4.1. Sample selection and descriptive statistics

There are a total of 5461 firm-year observations from 1995 to 2001 after excluding those firm-year observations that involve firms in the finance industry and firms listed only in the B-share market.¹² To focus on the monitoring activities of state shareholders, we exclude 615 firm-year observations for which the listed firms have private shareholders as the ultimate

¹¹ The data on capital structure and size are also obtained from the CSMAR Financial Databases. Leverage is measured by the industry-adjusted ratio of debt over total assets, and Size is measured by the logarithm of sales.

¹² The B-share market was formerly open to foreign investors, but not to domestic investors, although individual domestic investors have been allowed to invest in B-shares since February 2001.

controlling shareholders, because their objective function is likely to be different from that of state shareholders.¹³ We also exclude from this data 48 firm-year observations that involve firms with negative equity and 442 observations that involve firms which listed for less than six months. Our ROA and *Leverage* data have some extreme values. To minimize the possibility of biases in our results due to outliers, we winsorize these two variables at the 1st and 99th percentiles. After further eliminating observations with missing values in the variables included in our regression analysis, our final sample consists of 3916 firm-year observations.

Table 4 shows the summary statistics of the variables included in our model.

Our sample firms have been listed, on average, for 3.763 years. The average age and length of tenure of the managers are 48.242 and 2.435 years, respectively. Duality is not a common feature of the listed firms' corporate governance structure, with only 28.3% of CEOs also serving as board chairpersons. The average ROA for all of the listed firms is 4.0%, and the average MROA is 4.7%, which suggests that, on average, our sample firms experienced a performance decline during the study period.

Table 5 shows the average turnover rates and the turnover rates by quartiles of performance for all of the sample firms, the profit-making firms, and the loss-making firms. For our four performance measures, the average turnover rates of the loss-making firms are approximately 3.9% higher than those of the profit-making firms. Additionally, there is a greater dispersion of turnover rates between the best-performing and worst-performing firms if those firms are loss-making rather than profit-making. For example, in panel D, in which the firms are sorted by MROA, the best performing–worst performing turnover differential is 7.9% for loss-making firms but only 2.1% for profit-making firms. This seems to suggest that the turnover in loss-making firms is more sensitive to financial performance than it is in profit-making firms.¹⁴

4.2. Regression results on the determinants of turnovers

Two estimation issues are worth noting before we move to a discussion of our results. First, the *t*-statistics for MROA (MIROA) are potentially overstated, as there is a lack of independence across observations for a given CEO. We therefore estimate the model using the Huber/White/sandwich robust method with adjustment for within-cluster correlations for each CEO (Wooldridge, 2002).¹⁵ Second, we conduct a Pearson correlation test and find that all of the correlations among the variables included in our models are lower than 0.5. To further ensure that multicollinearity is not a problem, we calculate the variance inflation factors (VIF) for each independent variable. These VIFs never exceed 2, which suggest that our models are not prone to serious multicollinearity problems.

Tables 6 and 7 report our estimates of the performance–turnover sensitivities for the profit- and loss-making firms, respectively.¹⁶ For the profit-making samples, the coefficients for our four performance measures are statistically insignificant. For the loss-making firms, the coefficients for ROA and IROA are significantly negative at 10% and 5%, and those for MROA and MRIOA are significantly negative at 1%. These results are consistent with our hypothesis that state shareholders have a greater incentive to discipline their CEOs on the basis of financial performance when their firms are undergoing financial losses rather than making profits. Among the loss-making firms, the relationships between performance and turnover are statistically significant at the 10% and 5% level for the annual performance models and at the 1% level for average the performance models. This suggests that turnovers in Chinese listed firms are more sensitive to average performance than to annual performance. It is interesting to note that the coefficients for the control variables of *Duality* and *Tenure* are significantly negative at the 1% level in the sample of profit-making firms. For loss-making firms, the coefficients for *Duality* become statistically insignificant, and those for *Tenure* are statistically significant at 5% only in the two regressions in which annual performance measures are used. *Duality* and *Tenure* are measures of a CEO's formal and informal power in a listed firm. The results suggest that CEO turnover in profit- versus loss-making firms is associated with different political dynamics in which a CEO's power has different effects on the possibility of turnover.

4.3. Additional tests and robustness checks

We offer evidence of a non-linear relationship between performance and turnover and explain such relationship by the existence of a time-varying objective function whereby state shareholders attach greater importance to the objective of improving firm performance when their firms are incurring financial losses rather than making profits. Our hypothesis has additional implications for the sensitivities of performance to turnover in profit- and loss-making firms. When firms are making profits, state shareholders tend to place less emphasis on the objective of improving firm performance and are left with more room to serve different political and social objectives. The performance–turnover relationship in the profit-making firms is then likely to be

¹³ We obtained our data on private owners for the 1998–2001 period from the Ultimate Ownership of China's Listed Firms Dataset provided by Sinofin. For observations before 1998, we collected the information from the annual reports of the firms, Infobank's China Economic News Database, Infobank's China's Listed Firms Database, the China's Listed Firms Database available at <http://stock.sina.com.cn>, and internet materials available at <http://www.baidu.com>.

¹⁴ It should be noted that the performance spreads between the best-performing and worst-performing firms for the loss-making samples are greater than the corresponding spreads for the profit-making samples. The most obvious case is in Panel B in which performance is measured by IROA: the performance spread is 9.2% for profit-making firms, but 13.5% for loss-making firms. The best performing–worst performing turnover rate differential therefore has to be interpreted by taking the differences in performance spread into consideration. Nevertheless, the ratio of the best performing–worst performance turnover rate differential to performance spread is relatively larger in the loss-making sample than it is in the profit-making sample. Furthermore, in Panel C, in which the performance spreads of both the profit- and loss-making samples are comparable, there is still a substantial best performing–worst performing turnover rate differential between the profit- and loss-making firms.

¹⁵ Consistent results can be obtained if we make an adjustment for the within-cluster correlation for each firm.

¹⁶ For brevity, the year dummy coefficients are not reported.

Table 5

Turnovers rates and performance of China's listed firms

		(1) = lowest performance	(2)	(3)	(4)	(5) = highest performance	Overall turnover rate	Performance spread (5)–(1)	Turnover rate spread (5)–(1)	T test of Turnover rate spread
Panel A: Observations are sorted according to ROA										
All firms	median ROA	–0.010	0.020	0.042	0.063	0.100		0.110		
(obs.: 3916)	turnover rate	0.114	0.078	0.054	0.059	0.068	0.074		0.046	3.184***
Profit-making firms	median ROA	0.011	0.031	0.049	0.070	0.105		0.094		
(obs.:3340)	turnover rate	0.082	0.082	0.052	0.061	0.064	0.069		0.018	1.259
Loss-making firms	median ROA	–0.128	–0.065	–0.027	–0.009	–0.003		0.125		
(obs.: 576)	turnover rate	0.158	0.078	0.129	0.122	0.052	0.108		0.106	2.662***
Panel B: Observations are sorted according to IROA										
All firms	median IROA	–0.056	–0.022	0.000	0.020	0.055		0.111		
(obs.: 3916)	turnover rate	0.110	0.082	0.059	0.059	0.063	0.074		0.047	3.357***
Profit-making firms	median IROA	–0.031	–0.009	0.006	0.025	0.060		0.092		
(obs.:3340)	turnover rate	0.096	0.072	0.052	0.057	0.066	0.069		0.030	2.009**
Loss-making firms	median IROA	–0.171	–0.102	–0.073	–0.054	–0.036		0.135		
(obs.: 576)	turnover rate	0.165	0.096	0.122	0.087	0.069	0.108		0.096	2.293**
Panel C: Observations are sorted according to MROA										
All firms	median MROA	–0.006	0.026	0.047	0.069	0.107		0.113		
(obs.: 3916)	turnover rate	0.110	0.087	0.055	0.054	0.066	0.074		0.044	3.046***
Profit-making firms	median MROA	0.016	0.038	0.054	0.075	0.113		0.097		
(obs.:3340)	turnover rate	0.099	0.064	0.051	0.058	0.070	0.069		0.028	1.869*
Loss-making firms	median MROA	–0.088	–0.040	–0.013	–0.001	0.016		0.103		
(obs.: 576)	turnover rate	0.148	0.157	0.114	0.070	0.051	0.108		0.097	2.482**
Panel D: Observations are sorted according to MIROA										
All firms	median MIROA	–0.052	–0.018	0.002	0.022	0.059		0.111		
(obs.: 3916)	turnover rate	0.111	0.079	0.057	0.054	0.070	0.074		0.041	2.829***
Profit-making firms	median MIROA	–0.030	–0.006	0.009	0.027	0.063		0.093		
(obs.:3340)	turnover rate	0.096	0.067	0.051	0.054	0.075	0.069		0.021	1.371
Loss-making firms	median MIROA	–0.135	–0.083	–0.057	–0.042	–0.024		0.111		
(obs.: 576)	turnover rate	0.157	0.165	0.078	0.061	0.078	0.108		0.079	1.873*

This table reports the average fraction of CEOs involuntarily replaced by quartiles of performance for all sample firms, profit-making firms, and loss-making firms. The observations are sorted into five classes according to their performance (1 = low, 5 = high). *ROA* is the unadjusted return on assets, measured as the ratio of pretax operating income to the book value of total assets at the beginning of the period. *IROA* is the industry-adjusted return on assets, measured as *ROA* less the median value of *ROA* for all firms in the same industry. *MROA* is the moving average of *ROA* over a CEO's tenure. *MIROA* is the moving average of *IROA* over a CEO's tenure. The performance spread is the difference in median performance between the best-performing firms and the worst-performing firms. Turnover rate spread is the best-performing and worst-performing turnover rate differential, and it is tested using a two-tailed *t*-test. *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

Table 6

Logit regression estimation of turnover–performance links in China's profit-making firms

	(1)	(2)	(3)	(4)
List	0.108 (3.298)***	0.107 (3.301)***	0.108 (3.200)***	0.109 (3.317)***
Age	0.046 (4.551)***	0.046 (4.491)***	0.046 (4.561)***	0.046 (4.546)***
Tenure	−0.266 (4.744)***	−0.267 (4.767)***	−0.262 (4.687)***	−0.263 (4.691)***
Duality	−1.183 (5.377)***	−1.177 (5.353)***	−1.183 (5.368)***	−1.182 (5.365)***
Leverage	−0.140 (0.281)	−0.229 (0.456)	−0.137 (0.267)	−0.123 (0.238)
Size	−0.295 (3.148)***	−0.293 (3.128)***	−0.292 (3.057)***	−0.294 (3.092)***
State	0.274 (1.384)	0.266 (1.344)	0.280 (1.419)	0.278 (1.412)
ROA	−1.898 (0.874)			
IROA		−3.032 (1.364)		
MROA			−1.473 (0.678)	
MIROA				−1.344 (0.627)
Constant	0.899 (0.446)	0.813 (0.403)	0.816 (0.403)	0.787 (0.386)
Observations	3340	3340	3340	3340
Pseudo R-squared	0.067	0.068	0.067	0.067

This table reports the logit regression estimation of the probabilities of forced CEO turnovers in China's profit-making firms. The sample period is from 1995 to 2001. *List* is the number of years that a firm has been listed. *Age* is the age of a CEO. *Tenure* is the number of years a CEO has been in his or her current position. *Duality* is a dummy variable that equals 1 if a CEO is also a board chairperson and 0 otherwise. *Leverage* is the industry-adjusted capital structure of a listed firm, measured as the ratio of the book value of total debt over the book value of total assets less the median ratio in this industry. *Size* is the size of a listed firm, measured as the natural logarithm of the book value of total assets. *State* takes the value of 1 if a firm has a higher percentage of state shares relative to legal person shares, and 0 otherwise. *ROA* is the unadjusted return on assets, measured as the ratio of pretax operating income to the beginning period book value of total assets. *IROA* is the industry-adjusted return on assets, measured as ROA less the median value of ROA for all firms in the same industry. *MROA* is the moving average of ROA over a CEO's tenure, and *MIROA* is the moving average of *IROA* over a CEO's tenure. *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

sensitive to the level of importance that different state shareholders attach to those social and political objectives. Specifically, those shareholders who attach greater importance to social and political objective have less incentive to discipline relatively poorly performing CEOs than do shareholders who attach relatively less importance to them. As a result, the former type of shareholder exhibits a lower level of performance–turnover sensitivity than does the latter. When firms incur financial losses, state shareholders tend to place more emphasis on improving firm performance. When that objective becomes overriding and dominates other social and political objectives, different state shareholders exhibit similar performance–turnover sensitivities, even though they have different performance–turnover sensitivities when the firms are making profits.

To provide additional evidence to support our hypothesis, we examine the performance–turnover sensitivities for different state shareholders who have different incentives to serve political and social objectives when their firms are making profits. First, we expect the shareholders of firms owned by the central government to place greater emphasis on social and political objectives than those of firms owned by local governments.¹⁷ The central government in China tends to retain controls over firms in industries that are considered to be strategically and politically important (such as utilities, telecommunications, and energy) (Nee et al., 2007). Given that prices in these industries tend to be regulated, it is likely that the state shareholders of these firms would attach greater importance to social and political objectives than to improving firm performance. Second, fiscal decentralization in China has provided local governments with a greater incentive to improve economic performance. Feltenstein and Iwata (2005), for example, offer evidence that this fiscal decentralization has induced local governments to place greater emphasis on economic

¹⁷ Prior studies (Chen and Wang, 2004; Firth et al., 2006; Chen et al., 2006) have examined whether there are any differences in monitoring activities between firms that are ultimately owned by government agencies (GAs) and state-owned commercial firms (SCFs). The evidence is far from conclusive. Firth et al. (2006) and Chen and Wang (2004) offer evidence that SCFs are more effective than GAs with respect to the monitoring of managers, whereas Chen et al. (2006) show that the latter are more effective than the former. These conflicting results may result from the differences between, as well as the difficulty involved in, the identifications of the ultimate ownership of the listed Chinese firms. Firth et al. (2006) use the official share classification as a proxy for ownership identity, whereas Chen et al. (2006) and Chen and Wang (2004) focus on the identity of the first-level and ultimate owners, respectively. Classifications based on first-level owners fail to identify the true ultimate owners of listed firms. Although classifications based on ultimate owners can distinguish clearly whether the ultimate owners are private or state entities, it is very difficult to determine whether a listed firm is ultimately owned by a GA or SCF, because a SCF is by definition ultimately owned by the government. Strictly speaking, no listed firm is ultimately owned by a state-owned firm. In fact, listed firms are inconsistent in their own reporting of their ultimate owners, as revealed by the database on the Ultimate Ownership of China's Listed Companies provided by SinoFin. Some listed firms sometimes reported that they were ultimately owned by commercial firms and sometimes that they were ultimately owned by governments, even though they had undergone no changes in ownership.

Table 7

Logit regression estimation of turnover–performance links in China's loss-making firms

	(1)	(2)	(3)	(4)
List	0.138 (1.748)*	0.143 (1.797)*	0.107 (1.327)	0.115 (1.426)
Age	0.043 (2.305)**	0.042 (2.269)**	0.041 (2.184)**	0.040 (2.173)**
Tenure	−0.193 (2.003)**	−0.191 (1.973)**	−0.155 (1.622)	−0.155 (1.602)
Duality	0.100 (0.301)	0.124 (0.371)	0.109 (0.325)	0.151 (0.449)
Leverage	−0.401 (0.572)	−0.438 (0.631)	−0.692 (0.999)	−0.715 (1.034)
Size	−0.034 (0.204)	−0.038 (0.232)	0.044 (0.263)	0.029 (0.175)
State	−0.016 (0.042)	−0.023 (0.059)	−0.043 (0.112)	−0.047 (0.121)
ROA	−4.710 (1.895)*			
IROA		−4.866 (1.999)**		
MROA			−8.800 (3.304)***	
MIROA				−8.668 (3.294)***
Constant	−4.512 (1.211)	−4.591 (1.249)	−5.880 (1.583)	−5.984 (1.636)
Observations	576	576	576	576
Pseudo R-squared	0.044	0.045	0.060	0.060

This table reports the logit regression estimation of the probabilities of forced CEO turnovers in China's loss-making firms. The sample period is from 1995 to 2001. *List* is the number of years that a firm has been listed. *Age* is the age of a CEO. *Tenure* is the number of years a CEO has been in his or her current position. *Duality* is a dummy variable that equals 1 if a CEO is also a board chairperson and 0 otherwise. *Leverage* is the industry-adjusted capital structure of a listed firm, measured as the ratio of the book value of total debt over the book value of total assets less the median ratio in this industry. *Size* is the size of a listed firm, measured as the natural logarithm of the book value of total assets. *State* takes the value of 1 if a firm has a higher percentage of state shares relative to legal person shares, and 0 otherwise. *ROA* is the unadjusted return on assets, measured as the ratio of pretax operating income to the beginning period book value of total assets. *IROA* is the industry-adjusted return on assets, measured as ROA less the median value of ROA for all firms in the same industry. *MROA* is the moving average of ROA over a CEO's tenure, and *MIROA* is the moving average of IROA over a CEO's tenure. *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

performance relative to political objectives (controlling inflation) in their objective functions. We therefore expect that the listed firms owned by local governments will tend to display a stronger incentive to discipline poorly performing CEOs than will firms owned by the central government.

In addition to different levels of government, we also expect that firms in regions with different fiscal conditions will have different incentives to serve social and political objectives. Bai et al. (2000) show theoretically that local governments in regions with poor fiscal conditions have a greater incentive to use SOEs to serve their social and political objectives. Chen et al. (2004) also show that listed Chinese firms appoint more politically connected CEOs if they are located in regions with larger fiscal deficits. We therefore expect the state shareholders of listed firms in regions with larger fiscal deficits to attach more importance to social and political objectives than the shareholders of firms in regions with better fiscal conditions, and thus display weaker performance–turnover sensitivities.

We obtain our ultimate ownership data for the 1998–2001 period from the ultimate ownership of Chinese listed firm databases provided by Sinofin and the WIND Information Corporation.¹⁸ We double-check these data and also hand-collect additional data for the 1995–1997 period from the same data sources that we used to collect our data on the destinations of departing CEOs. We trace the ultimate owners of the listed firms and find that 11.9% are privately-owned, 14.5% are centrally-owned, and 66.8% are locally owned. There are 353 observations (6.8%) for which we cannot ascertain the ownership nature. We create a dummy variable (LOCAL) that equals 1 if a listed firm is owned by a local government and 0 if it is owned by the central government, and we collect the data on the fiscal conditions of 34 province-level administrative units in China from the Statistical Yearbook of China. We create another dummy variable (DEFICIT) that equals 1 if the region's budgetary deficit is larger than the median level of the country. We include these two dummy variables (LOCAL and DEFICIT) and their interaction terms with performance measures to capture the effects of government ownership and fiscal conditions on the sensitivity of turnover to performance. As turnover in China's listed firms is more sensitive to average performance measures than to annual performance measures, Table 8 reports only the results obtained using MROA and MIROA as the performance measures.

The results are consistent with our expectations. For the profit-making sample, the coefficients for LOCAL are significantly positive, which suggest that firms owned by local governments tend to have a higher incidence of CEO turnover. The coefficients for the interaction terms between LOCAL and the two performance measures are significantly negative, which indicates the existence

¹⁸ As suggested by Kato and Long (2006), the database provided by Sinofin can only distinguish between state and private ultimate owners. The database provided by WIND information offers more detailed classifications that allow us to distinguish between firms controlled by different levels of government.

Table 8

Logit regression estimation of turnover–performance links in China's listed firms with interactions

	Profit-making firms			Loss-making firms	
	(1)	(2)		(3)	(4)
List	0.096 (2.816)***	0.099 (2.955)***	List	0.078 (0.957)	0.084 (1.030)
Age	0.050 (4.714)***	0.050 (4.710)***	Age	0.048 (2.462)**	0.048 (2.462)**
Tenure	−0.255 (4.290)***	−0.255 (4.278)***	Tenure	−0.143 (1.434)	−0.143 (1.420)
Duality	−1.174 (5.136)***	−1.181 (5.180)***	Duality	0.033 (0.095)	0.075 (0.220)
Leverage	−0.117 (0.217)	−0.069 (0.126)	Leverage	−1.070 (1.456)	−1.064 (1.449)
Size	−0.333 (3.258)***	−0.339 (3.318)***	Size	0.066 (0.355)	0.050 (0.275)
LOCAL	0.905 (2.513)**	0.532 (2.476)**	LOCAL	0.103 (0.237)	0.165 (0.276)
DEFICIT	−0.123 (0.484)	0.088 (0.551)	DEFICIT	−0.259 (0.634)	−0.274 (0.480)
MROA	2.999 (0.679)		MROA	−7.778 (1.476)	
LOCAL*MROA	−8.308 (1.799)*		LOCAL*MROA	−0.365 (0.068)	
DEFICIT*MROA	4.888 (1.356)		DEFICIT*MROA	−2.434 (0.492)	
MIROA		3.106 (0.737)	MIROA		−8.604 (1.739)*
LOCAL*MIROA		−8.652 (1.881)*	LOCAL*MIROA		0.403 (0.076)
DEFICIT*MIROA		6.172 (1.691)*	DEFICIT*MIROA		−1.572 (0.317)
Constant	0.818 (0.371)	1.031 (0.469)	Constant	−6.439 (1.623)	−6.590 (1.695)*
Observations	3127	3127	Observations	497	497
Pseudo R-squared	0.074	0.074	Pseudo R-squared	0.068	0.068

This table reports the logit regression estimation of the probabilities of forced CEO turnovers in China's profit- and loss-making firms with interactions. The sample period is from 1995 to 2001. *List* is the number of years that a firm has been listed. *Age* is the age of a CEO. *Tenure* is the number of years a CEO has been in his or her current position. *Duality* is a dummy variable that equals 1 if a CEO is also a board chairperson and 0 otherwise. *Leverage* is the industry-adjusted capital structure of a listed firm, measured as the ratio of the book value of total debt over the book value of total assets less the median ratio in this industry. *Size* is the size of a listed firm, measured as the natural logarithm of the book value of total assets. *LOCAL* is a dummy variable that equals 1 if a listed firm is owned by a local government and 0 if it is owned by the central government. *DEFICIT* is a dummy variable that equals 1 if the region's budgetary deficit is larger than the median level of the country. *MROA* is the moving average of ROA (Return on Assets) over a CEO's tenure, and *MIROA* is the moving average of IROA (Industry-adjusted ROA) over a CEO's tenure. *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

of a more negative relationship between firm performance and turnover for firms owned by local governments. The coefficients for *DEFICIT* are statistically insignificant, but the coefficients for the interaction terms between *DEFICIT* and the two performance measures are positive and significant when *MIROA* is used as the performance measure, which suggests that firms in regions with poorer fiscal conditions tend to exhibit weaker performance–turnover sensitivities. For the loss-making sample, the coefficients for *LOCAL* (*DEFICIT*) and its interaction terms with the two performance measures are statistically insignificant. Overall, the results are consistent with our hypothesis about the existence of a time-varying objective function whereby state shareholders attach greater importance to improving firm performance when their firms are incurring financial losses rather than making profits.¹⁹

Some earlier studies have suggested that net income is an important decision factor in motivating the actions of boards of directors (e.g., Jensen and Murphy, 1990; Kaplan, 1994). Therefore, we replicate our regression equations using net income rather than pretax operating income as the performance measure. These results are consistent with those based on pretax operating income performance.

¹⁹ As Powers (2005) discusses, interpreting the interaction terms in logit models can be problematic because of model non-linearity. We follow McNeil et al. (2004) in using the delta method to check the statistical significance of the predicted probability of turnovers and the sensitivities with respect to a change in performance. By assuming that all other variables are equal to the median values of each sample, we calculate the predicted probabilities and derivatives at the 25th, 50th, and 75th percentiles of *MROA* and *MIROA* for firms owned by the central government and local governments (firms in regions with poor or good fiscal conditions), respectively. We test the differences in turnover rates and the predicted performance–turnover sensitivities between firms owned by the central government and local governments (firms in regions with poor or good fiscal conditions). For the profit-making sample, the differences in both the turnover rate and the predicted performance–turnover sensitivity between firms owned by the central government and local governments (firms in regions with poor or good fiscal conditions) are statistically significant, though the significance levels when *MROA* is used as the performance measure are marginal only. For the loss-making sample, neither the difference in the turnover rates nor the predicted performance–turnover sensitivities between firms owned by the central government and local governments (firms in regions with poor or good fiscal conditions) is statistically significant.

Table 9

Changes in post-turnover performance in China's listed firms

	Panel A: Median changes in ROA		P value of difference in ROA change between (1) and (2)	Panel B: Median changes in IROA		P value of difference in IROA change between (3) and (4)	Panel C: Median changes in CROA		P value of difference in CROA change between (5) and (6)	Panel D: Median changes in CIROA		P value of difference in CIROA change between (7) and (8)
	Profit-making firms (1)	Loss-making Firms (2)		Profit-making firms (3)	Loss-making Firms (4)		Profit-making firms (5)	Loss-making Firms (6)		Profit-making firms (7)	Loss-making Firms (8)	
Median performance at $t = -1$	0.047	-0.022		0.000	-0.066		0.001	0.001		0.001	0.000	
+1 to -1	-0.014***	0.004	0.004	-0.003*	0.011**	0.005	0.001	0.042**	0.030	0.001	0.039***	0.025
+2 to -1	-0.022***	0.012**	0.000	-0.006**	0.026**	0.000	-0.001**	0.031**	0.021	-0.001*	0.033*	0.012
+3 to -1	-0.028***	0.024**	0.000	-0.008**	0.038***	0.000	-0.001*	0.006	0.380	0.001	0.008	0.475
Median performance at $t = 0$	0.037	-0.050		-0.001	-0.092		-0.001	-0.001		0.004	0.001	
+1 to 0	-0.005***	0.033***	0.000	-0.002***	0.039***	0.000	-0.012*	0.010*	0.013	-0.013**	0.010*	0.010
+2 to 0	-0.013***	0.040***	0.000	-0.005***	0.050***	0.000	-0.003	0.012	0.283	-0.004	0.007	0.240
+3 to 0	-0.019***	0.052***	0.000	-0.007***	0.064***	0.000	0.002	0.014	0.782	-0.004	0.014	0.747

This table presents the changes in the post-turnover performance of China's listed firms. The sample period is from 1995 to 2001. Panel A reports the median change in unadjusted return on assets (ROA), measured by the ratio of pretax operating income to total assets. Panel B shows the median change in industry-adjusted return on assets (IROA), measured by the ratio of pretax operating income to total assets minus the median of the corresponding ratio in the industry. Panel C reports the median changes in control-group-adjusted return on assets (control-group-adjusted ROA), measured by the ratio of pretax operating income to total assets minus the median of the corresponding ratio in the control group. Panel D shows the medium changes of control-group and industry-adjusted return on assets (control-group-adjusted ROA), measured by the industry-adjusted return on assets minus the median of the corresponding ratio in the control group. Significance of median changes is tested using the Wilcoxon signed rank test.

We also check the sensitivity of the results to our classification of turnovers. We use either ages 60 or 65 as the benchmark for the classification of forced retirement and also include turnovers that are associated with legal disputes as forced turnovers. Consistent results are obtained with these alternative classification schemes.

The involvement of a board chairperson in the management of a listed firm may affect managerial turnover. The more active involvement of a chairperson indicates the existence of more intensive monitoring, which increases the probability of turnover. However, if it is the chairperson rather than the CEO who actually maintains the daily decision-making responsibilities, then the CEO is less likely to be held responsible for poor firm performance. For the observations from 1998 to 2001 for which data on the frequency of board meetings are available from Sinofin, we use that frequency to capture the involvement of a chairperson and include the variable as an additional control variable. For the observations from 1999 to 2001 for which data on a chairperson's compensation package is available from CSMAR, we also attempt to capture that person's involvement in the management of a listed firm by using the information on whether he or she receives a salary or only an honorarium. We create a dummy variable that equals 1 if a chairperson receives a salary and 0 if he or she receives only an honorarium and use it as an alternative control variable. Our results show that a greater frequency of board meetings is positively and significantly related to forced turnovers. However, the existence of a chairperson who receives a salary is negatively related to the turnover rate, which suggests that a CEO faces a lower probability of turnover if it is the chairperson, rather than the CEO, who maintains daily decision-making responsibilities. This relationship, however, is statistically insignificant. All of the other results pertaining to the performance–turnover relationship are retained, which suggests that our evidence on the different performance–turnover links in profit- and loss-making firms are robust to the varying degree of chairperson involvement.

4.4. Changes in performance surrounding turnover

Table 9 presents the median post-turnover changes in ROA and IROA and control-group adjusted ROA (CROA) and IROA (CIROA) for all of the sample firms. CROA is measured by ROA minus the median of the corresponding ratio in the control group. CIROA is measured by the industry-adjusted ROA minus the median of the corresponding ratio in the control group.²⁰

Panel A reports the post-turnover changes in ROA. For the profit-making firms, these changes are significantly negative at the 1% level in all of the years using either year 0 or year -1 as the reference year, which indicates that profitability declined in the post-turnover period for these firms. However, the post-turnover changes for the loss-making firms are significantly positive, except for year +1, when year -1 is used as the reference year, which indicates that ROA improved.

Panel B reports the changes in IROA. After adjusting for industry performance, the post-turnover performance changes of the profit-making firms become statistically less significant, and the extent of the performance declines becomes smaller than the corresponding declines measured by ROA. The positive performance changes for the loss-making firms, in contrast, remain statistically significant, and the size of the improvements is larger than the corresponding improvements measured by ROA.

²⁰ Consistent results can be obtained if we focus on the mean post-turnover changes.

Panels C reports the median changes in CROA after adjusting for the performance of the control firms. For years +2 and +3 when year –1 is used as the reference and for year +1 when year 0 is the benchmark, the post-turnover performance declines in the profit-making firms remains statistically significant. This result suggests that the profit-making firms experienced performance declines relative to the control firms. The median changes in CROA for the loss-making firms, in contrast, remain positive. However, the median changes in CROA in year +3 when year –1 is used as the reference year and years +2 and +3 when year 0 is used as the benchmark become statistically insignificant. This indicates the presence of mean reversions in which the loss-making control firms also experienced performance improvements. Nevertheless, the significant positive changes in the other years suggest that the performance improvements in the loss-making firms cannot be entirely attributed to the mean reversion of the time series.

The median changes in CROA after adjusting for both industry and control group performance are shown in Panel D. For the profit-making firms, the changes in CROA are significantly negative for year +2 when year –1 is used as the benchmark and for year +1 when year 0 is used as the reference year. The changes in CROA remain statistically insignificant in the other years. This result suggests that the turnover in profit-making firms is not followed by significant performance improvements after controlling for industry and control group performance. The changes in CROA remain positive and are statistically significant in year +1 and year +2 when year –1 is used as the benchmark and in year +1 when year 0 is used as the reference year. Overall, these results indicate that there was a significant improvement in post-turnover profitability among the loss-making sample firms, but not among the profit-making sample firms. The results are consistent with the different incentive structures of shareholders and CEOs discussed in Section 2.

5. Conclusion

This study examines the relationship between CEO turnover and performance in listed Chinese firms in which the majority of the controlling shareholders are state shareholders with multiple objectives. We provide evidence of the existence of different turnover–performance sensitivities in profit- and loss-making firms. We also demonstrate that there is a noticeable improvement in the post-turnover profitability of loss-making firms, but not in that of profit-making firms. These results are consistent with our hypothesis that the shareholders of listed firms tend to attach greater importance to firm performance, and thus have a greater incentive to discipline managers on the basis of that performance, when their firms are incurring losses rather than making profits.

Our study provides a useful addition to the existing literature on monitoring activities in state-owned firms. Our findings are also relevant to not-for-profit and collectively owned organizations, which are similarly characterized by the absence of dominant private owners and the presence of multiple objectives (Dixit, 1997; Brickley and Van Horn, 2002; Eldenburg and Krishnan, 2003). The existing studies assume a time-invariant objective function for these organizations and offer mixed evidence on the performance–turnover relationship. Based on a sample of not-for-profit hospitals, Brickley and Van Horn (2002) find that the turnover of CEOs is significantly related to financial performance. Eldenburg and Krishnan (2003), in contrast, find that CEO turnover in government-owned not-for-profit hospitals is unrelated to financial performance. In view of this mixed evidence, our study suggests that exploring the trade-offs among different objectives and ascertaining how the relative level of importance attached to firm performance affects the performance–turnover relationship would offer greater insight into how managers in not-for-profit and collectively-owned organizations are monitored.

Our findings may also have some implications for the performance–turnover relationship in private firms because value-maximizing private shareholders also very often have multiple and conflicting objectives. Jensen (2001) suggests that a firm cannot effectively maximize its value if it ignores the interests of its stakeholders, who include, not only financial claimants, but also employees, customers, and communities. Furthermore, firm decisions very often have multiple effects on different dimensions (such as profits, market share, and financial risks), and these could have conflicting implications for firm financial performance both at a given time and across time. Given that financial losses usually trigger larger declines in stock prices (Francis et al., 2005) and that the costs of financial distress can be substantial (Andrade and Kaplan, 1998), private shareholders, although they tend to have a stronger profit motive than do state shareholders, may also place more emphasis on firm performance and have a greater incentive to discipline managers when their firms experience poor performance than when they experience good performance. The existing studies, however, estimate the performance–turnover relationship for all sample firms regardless of financial performance (e.g., Huson et al., 2001; Volpin, 2002; McNeil et al., 2004), even though some of them indicate that the turnover rate for CEOs in financially distressed firms is significantly higher than that in non-distressed private firms (e.g., Gilson, 1989; Hotchkiss, 1995). Whether private firms also have different turnover–performance sensitivities when they experience different levels of financial performance is another issue worthy of future investigation.

Acknowledgements

This paper benefited immensely from insightful suggestions and comments from our editor, David J. Denis and an anonymous referee. We acknowledge the financial support from the Hong Kong Research Grants Council (RGC) Competitive Earmarked Research Grant Awards 2004–2005 (LU7236/04H). Zhang Xuan and Yang Yong provided excellent research assistance.

References

- Alchian, A., 1965. Some economics of property rights. *Il Politico* 30, 816–829.
- Andrade, G., Kaplan, S., 1998. How costly is financial (not economic) distress? Evidence from highly leveraged transactions that became distressed. *Journal of Finance* 53, 1443–1493.
- Bai, C.E., Li, D., Tao, Z., Wang, Y., 2000. A multitask theory of state enterprise reform. *Journal of Comparative Economics* 28, 716–738.

- Bai, C.E., Lu, J.Y., Tao, Z.G., 2006. The multitask theory of state enterprise reform: empirical evidence from China. *American Economic Review* 96, 353–357.
- Bauer, J.M., 2005. Regulation and state ownership: conflicts and complementarities in EU telecommunications. *Annals of Public and Cooperative Economics* 76 (2), 151–177.
- Brickley, J., Van Horn, R.L., 2002. Managerial incentives in nonprofit organizations: evidence from hospitals. *Journal of Law and Economics* 95 (1), 227–250.
- Chang, E., Wong, S., 2004. Political control and performance in China's listed firms. *Journal of Comparative Economics* 20, 1–20.
- Chen, K.C.W., Wang, J.W., 2004. A comparison of shareholder identity and governance mechanisms in the monitoring of listed companies in China. Working paper, Hong Kong University of Science and Technology.
- Chen, D.H., Fan, P.H.J., Wong, T.J., 2004. Politically-connected CEOs, corporate governance and post-IPO performance of China's partially privatized firms. Working paper, Chinese University of Hong Kong.
- Chen, C., Li, Z., Su, X., Tsui, J., 2006. CEO turnover and accounting measures: effect of delegation of control rights. Working paper, City University of Hong Kong.
- Cheng, P., Li, J.L., Tong, W.H.S., 2007. What triggers top management turnovers in China? Working paper, Hong Kong Polytechnic University.
- Claessens, S., Djankov, S., 1999. Enterprise performance and management turnover in Czech Republic. *European Economic Review* 43 (4–6), 1115–1124.
- Dalton, D., Kesner, I., 1983. Inside/outside succession and organizational size: the pragmatics of executive replacement. *Academy of Management Journal* 26, 736–742.
- Dalton, D., Daily, C., Ellestrand, A., Johnson, J., 1998. Meta-analytic reviews of board composition, leadership structure, and financial performance. *Strategic Management Journal* 19, 269–290.
- Davidson, W., Worrel, D., Cheng, L., 1990. Key executive succession and stockholder wealth: the influence of successor's origin, position, and age. *Journal of Management* 16 (3), 647–664.
- Delios, A., Zhi, Wu, J., Zhou, N., 2006. A new perspective on ownership identities in China's listed companies. *Management and Organization Review* 2 (3), 319–343.
- Denis, D., Denis, D., 1995. Performance changes following top management dismissals. *Journal of Finance* 50, 1029–1057.
- Denis, D., Denis, D., Sarin, A., 1997. Ownership structure and top executive turnover. *Journal of Financial Economics* 40, 193–221.
- Dixit, A., 1997. Power of incentives in private versus public organizations. *AEA Papers and Proceedings* 87 (2), 378–382.
- Eldenburg, L., Krishnan, R., 2003. Public versus private governance: a study of incentives and operational performance. *Journal of Accounting and Economics* 35 (3), 377–404.
- Feltenstein, A., Iwata, S., 2005. Decentralization and macroeconomic performance in China: regional autonomy has its costs. *Journal of Development Economics* 76, 481–501.
- Firth, M., Fung, M.Y., Rui, M., 2006. Firm performance, governance structure, and top management turnover in a transitional economy. *Journal of Management Studies* 43 (6), 1289–1330.
- Fligstein, N., 1987. The intraorganizational power struggle: rise of finance personnel to top leadership in large corporations, 1919–1979. *American Sociological Review* 52, 44–58.
- Francis, J., Schipper, K., Vincent, L., 2005. Earnings and dividend informativeness: when cash flow rights are separated from voting rights. *Journal of Accounting and Economics* 39, 329–360.
- Fredrickson, J.W., Hambrick, D.C., Baumrin, S., 1988. A model of CEO dismissal. *Academy of Management Review* 13, 255–270.
- Friedman, S., Singh, H., 1989. CEO succession and stockholder reaction: the influence of organizational context and event context. *Academy of Management Journal* 32, 718–744.
- Gibelman, M., Gelman, S.R., 2002. On the departure of a chief executive officer: scenarios and implications. *Administration in Social Work* 26, 63–82.
- Gilson, S.C., 1989. Management turnover and financial distress. *Journal of Financial Economics* 25, 241–262.
- Groves, T., Hong, Y., McMillan, J., Naughton, B., 1995. China's evolving managerial labor market. *Journal of Political Economy* 103, 873–892.
- Hotchkiss, E.S., 1995. The post-emergence performance of firms emerging from Chapter 11. *Journal of Finance* 50, 3–21.
- Huson, M., Parrino, R., Starks, L., 2001. Internal monitoring mechanisms and CEO turnover: a long-term perspective. *Journal of Finance* 56, 2265–2298.
- Huson, M., Malatesta, P., Parrino, R., 2004. Managerial succession and firm performance. *Journal of Financial Economics* 74, 237–275.
- Jensen, M., 1986. Agency costs of free cash flow, corporate finance and takeovers. *American Economic Review* 76, 323–339.
- Jensen, M., 2001. Value maximization, stakeholder theory, and the corporate objective function. *European Financial Management Review* 7 (3), 297–317.
- Jensen, M., Murphy, K., 1990. Performance pay and top management incentives. *Journal of Political Economy* 98, 225–264.
- Jones, L.P., 1985. Public enterprise for whom? Perverse distributional consequences of public operational decisions. *Economic Development and Cultural Change* 33 (2), 333–348.
- Kang, J., Shivdasani, A., 1995. Firm performance, corporate governance, and top executive turnover in Japan. *Journal of Financial Economics* 38, 29–58.
- Kaplan, S., 1994. Top executive rewards and firm performance: a comparison of Japan and the U.S. *Journal of Political Economy* 102, 510–546.
- Kato, T., Long, C., 2006. CEO turnover, firm performance and enterprise reform in China: evidence from micro data. *Journal of Comparative Economics* 34, 796–817.
- Khanna, N., Poulsen, A., 1995. Managers of financially distressed firms: villains or scapegoats? *Journal of Finance* 50 (3), 919–940.
- Kole, S., Mulherin, J., 1997. The government as a shareholder: a case from the United States. *Journal of Law and Economics* 40, 1–22.
- Krueger, A.O., 1990. Government failures in development. *Journal of Economic Perspectives* 4 (3), 9–23.
- La Porta, R., López de Silanes, F., Shleifer, A., 2002. Government ownership of banks. *Journal of Finance* 57, 265–301.
- Li, H., Zhou, L., 2005. Political turnover and economic performance: the disciplinary role of personnel control in China. *Journal of Public Economics* 89 (9–10), 1743–1762.
- Liu, G.S., Sun, P., 2005. The class of shareholdings and its impacts on corporate performance: a case of state shareholding composition in Chinese public corporations. *Corporate Governance—An International Review* 13 (1), 46–59.
- Merton, R., 1940. Bureaucratic structure and personality. *Social Forces* 17, 560–568.
- McNeil, C., Niehaus, G., Powers, E., 2004. Management turnover in subsidiaries of conglomerates versus stand-alone firms. *Journal of Financial Economics* 72, 63–69.
- Morck, R., Yueng, B., Yu, W., 2000. The information content of the stock market: why do emerging markets have synchronous stock price movement? *Journal of Financial Economics* 58, 215–260.
- Nee, V., Oppen, S., Wong, S.M.L., 2007. Developmental state and corporate governance in China. *Management and Organization Review* 3 (1), 19–53.
- Powers, E.A., 2005. Interpreting logit regressions with interaction terms: an application to management turnover literature. *Journal of Corporate Finance* 11, 504–522.
- Qian, Y., Roland, G., 1998. Federalism and the soft budget constraint. *The American Economic Review* 88, 1143–1162.
- Shen, W., Cannella, A.A., 2002. Political dynamics within top management and their impacts on CEO dismissal followed by inside succession. *Academy of Management Journal* 45, 1195–1208.
- Shleifer, A., Vishny, R., 1994. Politicians and firms. *Quarterly Journal of Economics* 109, 995–1025.
- Shleifer, A., Vishny, R., 1997. A survey of corporate governance. *Journal of Finance* 52, 737–783.
- Sun, Q., Tong, W., 2003. China share issue privatization: the extent of its success. *Journal of Financial Economics* 70, 183–222.
- Volpin, P., 2002. Governance with poor investor protection: evidence from top executive turnover in Italy. *Journal of Financial Economics* 64 (1), 61–90.
- Wang, X., Xu, L., Zhu, T., 2004. State-owned enterprises going public: the case of China. *Economics of Transition* 12 (3), 467–487.
- Wong, M.L.S., 2006. China's stock market: a marriage of capitalism and socialism. *Cato Journal*, 26, 1–35.
- Wong, M.L.S., Oppen, S., Hu, R.Y., 2004. Shareholding structure, depoliticization and enterprise performance: lessons from China's listed companies. *Economics of Transition* 12 (1), 29–66.
- Wooldridge, J.M., 2002. *Economic Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.
- Xu, X., Wang, Y., 1999. Ownership structure and corporate governance in Chinese stock companies. *China Economic Review*, 10, 75–98.
- Zhang, L.Y., 2006. Market socialism revisited: the case of Chinese state-owned enterprises. *Issues & Studies* 42 (3), 1–46.
- Zhang, W.Y., 1998. China's SOEs reform: A corporate governance perspective. Working paper, Institute of Business Research, Beijing University.



The Relationship between Firm Investment and Financial Status

Sean Cleary

The Journal of Finance, Vol. 54, No. 2 (Apr., 1999), 673-692.

Stable URL:

<http://links.jstor.org/sici?sici=0022-1082%28199904%2954%3A2%3C673%3ATRBFA%3E2.0.CO%3B2-L>

The Journal of Finance is currently published by American Finance Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/afina.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

<http://www.jstor.org/>
Thu Feb 16 20:48:30 2006

The Relationship between Firm Investment and Financial Status

SEAN CLEARY*

ABSTRACT

Firm investment decisions are shown to be directly related to financial factors. Investment decisions of firms with high creditworthiness (according to traditional financial ratios) are extremely sensitive to the availability of internal funds; less creditworthy firms are much less sensitive to internal fund availability. This large sample evidence is based on an objective sorting mechanism and supports the results of Kaplan and Zingales (1997), who also find that investment outlays of the least constrained firms are the most sensitive to internal cash flow.

A FIRM'S FINANCIAL STATUS IS IRRELEVANT for real investment decisions in a world of perfect and complete capital markets, as has been demonstrated by Modigliani and Miller (1958). However, financial structure may be relevant to the investment decisions of companies facing uncertain prospects that operate in imperfect or incomplete capital markets where the cost of external capital exceeds that of internal funds. For example, Greenwald, Stiglitz, and Weiss (1984), Myers and Majluf (1984), and Myers (1984) provide a foundation for these market imperfections by appealing to asymmetric information problems in capital markets. Alternatively, Bernanke and Gertler (1989, 1990) and Gertler (1992) demonstrate that agency costs can also cause a premium on external finance that increases as borrower net worth decreases. The investment decisions of firms operating in such environments are sensitive to the availability of internal funds because they possess a cost advantage over external funds.

Fazzari, Hubbard, and Petersen (1988) and a number of subsequent empirical studies provide strong support for the existence of this financing hierarchy, which is most prevalent among firms that have been identified as facing a high level of financial constraints.¹ These studies categorize firms according to characteristics (such as dividend payout, size, age, group membership, or debt ratings) that are designed to measure the level of financial

* Saint Mary's University, Halifax. I am grateful to Laurence Booth, Glenn Hubbard, Donald Brean, Paul Halpern, Varouj Aivazian, Raymond Kan, Tom McCurdy, Steve Hadjiyannakis, and participants at the 1996 Northern Finance Association meetings for their valuable comments. The article was improved substantially by incorporating comments from the editor and an anonymous referee. All errors are the responsibility of the author.

¹ Other examples include studies by Hoshi, Kashyap, and Scharfstein (1991), Oliner and Rudebusch (1992), Whited (1992), Schaller (1993), and Gilchrest and Himmelberg (1995). Refer to Hubbard (1998) for an extensive summary of this literature.

constraints faced by firms. The results suggest that investment decisions of firms that are more financially constrained are more sensitive to firm liquidity than those of less constrained firms.

Debate over this matter has been fueled by the recent work of Kaplan and Zingales (1997) who challenge the generality of the conclusions summarized above. Kaplan and Zingales (hereafter KZ) classify firms according to their degree of financial constraint, based on quantitative and qualitative information obtained from company annual reports. Contrary to previous evidence, they find that investment decisions of the least financially constrained firms are the most sensitive to the availability of cash flow.

This study follows the approach of Kaplan and Zingales by classifying firms according to financial variables that are related to financial constraints. Firm financial status is determined using multiple discriminant analysis, similar to Altman's Z factor for predicting bankruptcy. This multi-variate classification scheme effectively captures desired cross-sectional properties of firms. It also allows reclassification of firm financial status every period, and group composition is allowed to vary over time to reflect changing levels of financial constraints at the level of the firm. This differs from previous studies that do not allow group composition to vary, implicitly assuming that financial obstacles faced by firms do not change over time.

BS A major focus of this literature is the comparison of investment-liquidity sensitivities across different groups of firms. I employ a bootstrap methodology to determine significance levels of observed differences in coefficient estimates. This represents an improvement over previous studies whose conclusions are based primarily on the observed differences in magnitude and level of significance of the liquidity variable coefficient estimates.

Investment decisions of all firms are found to be very sensitive to firm liquidity, which is consistent with previous evidence. Similar to the KZ results, firms that are more creditworthy exhibit greater investment-liquidity sensitivity than those classified as less creditworthy. This provides strong support for the KZ conclusions using an objective classification scheme and a large, diversified sample of 1,317 U.S. firms.

The remainder of the paper is organized as follows. The next section reviews existing literature and discusses the motivation for the present study. Section II provides details of the data and methodology utilized, and Section III examines the regression results. Conclusions are offered in the final section.

I. Background

A. Evidence of Financing Hierarchies

An important empirical study of firm investment decisions in the presence of financial constraints was conducted by Fazzari, Hubbard, and Petersen (1988) (hereafter FHP88). They use *Value Line* data for 422 large U.S. man-

ufacturing firms over the 1970 to 1984 time period to analyze differences in investment behavior by firms classified according to earnings retention.² FHP88 argue that firms with higher retention ratios face higher informational asymmetry problems and are more likely to be liquidity constrained.

FHP88 run the following regression for several models of investment:

$$(I/K)_{it} = f(X/K)_{it} + g(\underline{CF/K})_{it} + u_{it}, \quad (1)$$

where I_{it} represents investment in plant and equipment for firm i during period t ; K is the beginning-of-period book value for net property, plant, and equipment; $g(CF/K)$ is a function of current cash flow which measures firm liquidity; $f(X/K)$ is a function of variables related to investment opportunities; and u_{it} is an error term. Their analysis focuses on the q theory of investment, which suggests that $f(X/K)$ is represented by a firm's Tobin's q value. The investment of firms that exhaust all their internal finance is found to be much more sensitive to fluctuations in cash flow than that of mature, high dividend firms. FHP88 attribute these results to a financing hierarchy in which internal funds have a cost advantage over new equity and debt.

Subsequent studies have confirmed the central FHP88 result by dividing samples according to other a priori measures of financial constraint. For example, Hoshi et al. (1991) conclude that the investment outlays of 24 Japanese manufacturing firms that are not members of a keiretsu are much more sensitive to firm liquidity than that of 121 firms that are members of a keiretsu and are presumed to be less financially constrained. Oliner and Rudebusch (1992) examine 99 NYSE-listed firms and 21 over-the-counter firms during the 1977 to 1983 period. They find that investment is most closely related to cash flow for firms that are young, whose stocks are traded over-the-counter, and that exhibit insider trading behavior consistent with privately held information. Schaller (1993) studies 212 Canadian firms over the 1973 to 1986 period and concludes that investment for young, independent, manufacturing firms with dispersed ownership concentration is the most sensitive to cash flow.

Whited (1992) and Bond and Meghir (1994) employ an Euler equation approach to directly test the first-order condition of an intertemporal maximization problem, which does not require the measurement of Tobin's q . The strategy is implemented by imposing an exogenous constraint on external finance and testing whether that constraint is binding for a particular group of firms. Whited uses a sample of 325 U.S. manufacturing firms for the 1972 to 1986 period, and Bond and Meghir use an unbalanced panel of

² In particular, FHP88 classify firms into the following three groups based on their dividend behavior over the 1970 to 1984 period: (1) those that have a ratio of dividends to income of less than 0.10 for at least 10 years; (2) those that have a dividend-income ratio between 0.10 and 0.20 for at least 10 years; and (3) all other firms.

626 U.K. manufacturing companies for the 1974 to 1986 period. Both of these studies find the exogenous finance constraint to be particularly binding for the constrained groups of firms, which supports the basic FHP88 result. All of these results support FHP88's informational asymmetry argument.

A related study by Mayer (1990) examines the sources of industry finance of eight developed countries from 1970 to 1985 and reveals a number of stylized facts regarding global corporate financing behavior which also support the existence of financing hierarchies. He finds that: (i) retentions are the dominant source of financing in all countries; (ii) the average firm in any of these countries does not raise substantial amounts of financing from security markets in the form of short-term securities, bonds, or equities; and, (iii) the majority of external financing comes from bank loans in all countries.

B. *Conflicting View*

Kaplan and Zingales (1997) challenge the generality of the conclusions described above. They use a combination of qualitative and quantitative information extracted from company annual reports to rank firms in terms of their apparent degree of financial constraint. A firm is classified as financially constrained if the cost or availability of external funds precludes the company from making an investment it would have chosen to make had internal funds been available. Their classification scheme uses data from letters to shareholders, management discussions of operations and liquidity (when available), financial statements, notes to those statements for each firm-year, and financial ratios obtained from the COMPUSTAT database.³

The KZ sample consists of the 49 low-dividend paying firms identified by FHP88 as having extremely high investment-cash flow sensitivity. Contrary to FHP88's prediction that this entire group would face severe financial constraints, KZ find that "in only 15 percent of firm-years is there some question as to a firm's ability to access internal or external funds to increase investment. In fact, almost 40 percent of the sample firms, including Hewlett-Packard (cited above), could have increased investment in every year of the sample period" (p. 171). Contrary to previous research, KZ find that the least financially constrained firms exhibit the greatest investment-cash flow sensitivity. They suggest these controversial results "capture general features of the relationship between corporate investment and cash flow" (p. 204), and are not specific to the sample or techniques utilized.

³ KZ determine firm financial constraint status every year; however, they classify firms into one of three groups for the entire period for regression purposes. Firms are categorized as not financially constrained in a particular year if they "initiated or increased cash dividends, repurchased stock or explicitly indicated in its annual report that the firm had more liquidity than it would need for investment in the foreseeable future." Firms were "more likely" to be classified as not constrained if they had a large cash position (relative to investment), or if the firm's lenders did not restrict the firm from making large dividend payments (relative to investment). This classification scheme suggests unconstrained firms tend to include financially healthy companies with low debt and high cash.

C. Motivation

KZ's finding that investment outlays of the least financially constrained firms are the most sensitive to cash flow contradicts a large body of empirical results, which implies the importance of examining the generality of their conclusions. The results are puzzling because they suggest that managers choose to rely primarily on internal cash flow for investment, despite the availability of additional low cost external funds. An important implication is that policies designed to make credit more available during recessions may not lead to an increase in investment by firms with high investment-cash flow sensitivities, which has been a policy implication of the existing literature.

The classification of firm financial constraint status according to traditional financial ratios has intuitive appeal because it represents a direct measure of the premium paid for bank loans by firms. The importance of this type of measure is highlighted by Mayer's (1990) evidence that bank loans are the primary source of external finance for firms in developed countries. However, a major limitation of the KZ study is the fact that their sample consists of only 49 manufacturing firms that could be considered fairly high quality firms, or they would not have been included in the *Value Line* database. They further subdivide this sample into groups of 22, 19, and 8, which leaves very few firms in the groups for comparison purposes. The use of such a small homogeneous sample implies the behavior of a very few firms could be driving their results, and it may be ambitious to make general conclusions based on these observations. Further, KZ are criticized by Fazzari, Hubbard, and Petersen (1996) and Schiantarelli (1995) because their sorting criteria are somewhat subjective and rely on possibly self-serving managerial statements.

II. Research Design

A. Sample Characteristics

The sample consists of 1,317 U.S. firms that have complete financial information available for the 1987 to 1994 period on the SEC Worldscope Disclosure data set.⁴ Because the majority of firms have a December fiscal year-end, firms are included only if their last available financial statements were reported for fiscal year-ends occurring between July 1994 and June 1995.

- Banks, insurance companies, other financial companies, and utility companies were deleted from the sample. Details of the calculation of the financial variables are included in the Appendix. Included firms were required to have positive values for sales, total assets, net fixed assets, and market-to-book ratio.

⁴ The requirement of complete information availability over the entire sample period is imposed to allow comparison of results with previous studies. The rationale underlying the use of this criterion is to focus attention on firms that have wealth to distribute.

A number of observations are “winsorized” (if the value of the variable exceeded cutoff values) according to the following rules: (i) assign a value of 100 percent (–100 percent) if growth in sales is greater (less) than 100 percent (–100 percent); (ii) assign a value of 2 (–2) if investment/net fixed assets is greater (less) than 2 (–2); (iii) assign a value of 5 (–5) if cash flow/net fixed assets is greater (less) than 5 (–5); (iv) assign a value of 10 if market-to-book is greater than 10; (v) assign a value of 10 if current ratio is greater than 10; (vi) assign a value of 100 percent (–100 percent) if net income margin is greater (less) than 100 percent (–100 percent); and (vii) assign a value of 100 (–0.1) if fixed charge coverage is greater (less) than 100 (0). This approach reduces the impact of extreme observations and allows the use of a larger number of observations than would be possible if these *extreme* observations were deleted (1,317 versus 1,080 firms).⁵

The sample includes 709 NYSE listed companies, 416 Nasdaq companies, and 192 companies listed on the AMEX or other U.S. exchanges. It is diversified across industries as measured by primary SIC code: 843 manufacturing firms (SIC codes 2000–3999); 99 agricultural, mining, forestry, fishing and construction firms (SIC codes 1–1999); 201 retail and wholesale trade firms (SIC codes 5000–5999); and 174 service firms (SIC codes 7000–8999). Summary statistics for the entire sample are included in Panel A of Table I.

B. Classification Methodology

Firms are classified into groups according to a beginning-of-period financial constraint index (Z_{FC}). Firm classification is allowed to change every period to reflect the fact that financial status changes continuously.⁶ The index is determined using multiple discriminant analysis, similar to Altman’s Z factor for predicting bankruptcy.⁷ An advantage of this approach is that it considers an entire profile of characteristics shared by a particular firm and transforms them into a univariate statistic.

The first step in discriminant analysis is to establish two or more mutually exclusive groups according to some explicit group classification. For example, Altman’s two groups consist of firms that went bankrupt and those that did not. It is difficult, if not impossible, to categorize explicitly which firms are financially constrained without making reference to a number of variables. However, it is still possible to establish two mutually exclusive groups by making use of the knowledge that firms do not like to cut dividends and are hesitant to increase them unless they can be maintained. This suggests dividing our sample into three categories: group 1 firms increase dividends and are likely not financially constrained; group 2 firms cut dividends and are likely financially constrained; and group 3 firms do not change

⁵ I thank an anonymous referee for this suggestion.

⁶ This point is acknowledged by Fazzari et al. (1996) who suggest that assuming firms are in one group for the entire period is an empirical convenience. Schiantarelli (1995) discusses the importance of accounting for this matter in detail.

⁷ Refer to Altman (1968) or Altman, Haldeman, and Narayanan (1977).

Table I
Sample Summary Statistics (1988–1994)

Panel A reports financial variable means for the sample of 1,317 firms. All financial variables are for the beginning of the fiscal year, except for cash flow and investment which represent firm cash flow and capital expenditures during period t . K is the firm's beginning-of-period net fixed assets value. The discriminant score (Z) is calculated using discriminant analysis according to equation (2). A full description of the variables is included in the Appendix. Dividend Group 1 includes firms whose dividend per share (DPS) increased in year t , Dividend Group 2 includes firms whose DPS decreased in year t , and Dividend Group 3 includes firms that had no change in DPS in year t . Panel B shows the number (percentage) of firms falling into the these three dividend categories over the sample period.

Panel A: Selected Financial Ratio Means (1988–1994)				
	Total Sample	Dividend Group 1 (increased DPS)	Dividend Group 2 (decreased DPS)	Dividend Group 3 (no change in DPS)
Net fixed assets (<i>K</i>)	\$650m	\$1076m	\$913m	\$360m
Current ratio	2.57	2.40	2.36	2.71
Debt ratio	0.22	0.20	0.26	0.23
Fixed charge coverage	12.1	16.8	7.4	9.9
Net income margin (%)	3.0	6.8	1.0	1.0
Market-to-book ratio	2.18	2.64	1.62	1.97
Sales growth (%)	10.1	11.4	1.6	10.3
Slack/ <i>K</i>	1.71	1.42	1.45	1.92
Cash flow/ <i>K</i>	0.47	0.58	0.27	0.42
Investment/ <i>K</i>	0.26	0.26	0.19	0.24
Discriminant score (<i>Z</i>)	−0.31	0.17	−0.87	−0.61

Panel B: Number of Firms per Dividend Group								
Dividend Group	Total Sample	1988	1989	1990	1991	1992	1993	1994
1 (increased DPS)	3241 (35.1%)	547 (41.5%)	543 (41.2%)	478 (36.4%)	408 (31.0%)	411 (31.2%)	420 (31.9%)	434 (33.0%)
2 (decreased DPS)	634 (6.9%)	53 (4.0%)	68 (5.2%)	94 (7.2%)	127 (9.6%)	110 (8.4%)	91 (6.9%)	91 (6.9%)
3 (no change in DPS)	5344 (58.0%)	717 (54.5%)	706 (53.6%)	745 (56.6%)	782 (59.4%)	796 (60.4%)	806 (61.2%)	792 (60.1%)

dividend payments. Group 3 firms are not utilized for purposes of the discriminant analysis; however, they are assigned Z_{FC} scores and are used in the subsequent regression analyses.⁸

Panel A of Table I reports summary statistics for the 1988 to 1994 period which confirm that firms reducing dividends appear to be more financially constrained according to traditional financial ratios. Firms that cut dividends exhibit lower current ratios, higher debt ratios, lower fixed charge coverage, lower net income margins, lower market-to-book ratios, and lower sales growth, and have lower slack/net fixed assets values than firms that increased dividends.⁹ Table I also shows the standard ratio performance for firms that did not increase or decrease dividend payments was between the other two groups.

Panel B of Table I indicates that the number of firms increasing (or decreasing) dividends changes through the years in response to changing economic conditions. The largest number of firms increasing dividends (547) occurred in the prerecessionary year of 1988; the largest number of firms cutting dividends (127) occurred in the recessionary year of 1991. This evidence supports the notion that firms face changing levels of financial constraints every year. Because the purpose of classifying firms is to examine the behavior of groups that face different levels of financial barriers, it is logical to allow group composition to change over time. Schiantarelli (1995) argues that studies which assign a firm to one group for the entire period are “neglecting the information that the financial constraints may be binding for the same firm in some years but not in others. It would be more advisable in these cases to allow firms to transit between different financial states” (p. 21).

Discriminant analysis uses a number of variables that are likely to influence characterization of a firm in one of the two mutually exclusive groups of interest. The present study uses the following beginning-of-period variables that are chosen to proxy for firm liquidity, leverage, profitability, and growth: current ratio, debt ratio, fixed charge coverage (FCCov), net income margin (NI%), sales growth, and slack/net fixed assets (SLACK/K).¹⁰ The hypothesis is that these variables will enable us to predict if firms will in-

⁸ This group of firms represents 58 percent of the sample (5,341 out of 9,219 firm-year observations) and can be categorized by reference to their Z_{FC} value (discussed below) as those that “fit the profile” of constrained or unconstrained firms. This enables the use of an increased sample size and requires less reliance on firm dividend policy for the purpose of a priori classification.

⁹ Slack is calculated as: cash + short term investments + (0.50 * inventory) + (0.70 * accounts receivable) – short term loans. It is included as a proxy for cash + unused line of credit, which is a measure of liquidity used by Kaplan and Zingales (1997). The calculation is based on traditional credit line arrangements that enable firms to establish operating loans up to 50 percent of inventory and 70–75 percent of good accounts receivable. Net fixed assets is the net property, plant, and equipment figure obtained from the firm’s balance sheet, and is used for scaling purposes.

¹⁰ Alternative specifications, including the one used in Altman (1968), are also employed. They produce similar results but have a slightly lower success rate in predicting which firms will cut or increase dividends.

crease or decrease dividend payments in the subsequent period. Coefficient values are estimated that best distinguish each independent variable between the two groups according to the following Z_{FC} value:

$$Z_{FC} = \beta_1 \text{Current} + \beta_2 \text{FCCov} + \beta_3 \text{SLACK/K} + \beta_4 \text{NI\%} + \beta_5 \text{Sales Growth} + \beta_6 \text{Debt}. \quad (2)$$

Univariate significance levels indicate that net income margin, sales growth, debt ratio, and fixed charge coverage are all significant at the 1 percent significance level. Table II displays correlations among these variables, as well as those used in the subsequent regression analysis. The largest correlations between Z_{FC} and the independent variables are 0.80 with NI% and 0.55 with sales growth. These observations suggest that firms tend to increase dividends during periods of stable and increasing profits. Current ratio and SLACK/K both exhibit small, negative correlations with Z_{FC} , which accounts for their insignificance in classifying firms. This is somewhat surprising because one would expect dividend increases to be closely tied to a firm's liquidity status as measured by these variables.

Overall, the variables do a good job of successfully predicting which firms will cut or increase their dividends, with group 1 and group 2 firms being properly classified 74 percent of the time. Despite the practical importance of being able to accurately predict dividend changes, it is not the primary concern of this paper.¹¹ The focus here is to classify firms according to their financial status, and the summary statistics for the predicted group classification of firms presented in Table III indicate success in achieving this objective. In particular, firms that are classified as group 1 (likely to increase dividends) appear more solid in terms of the reported financial variables.

Firms are classified every year according to their Z_{FC} value to reflect the fact that their financial constraint status is changing continuously. The top one-third of the firms each year are categorized as not financially constrained (NFC), the next one-third as partially financially constrained (PFC), and the bottom one-third as financially constrained (FC). Summary statistics for these groups presented in Table III indicate the classification scheme has successfully captured the desired cross-sectional properties. The financial ratios are superior for the NFC group, inferior for the FC group, with the PFC group lying somewhere in between.¹² The importance of classifying firm financial status every year is highlighted by the observed turnover rates for the NFC, PFC, and FC groups which average 40.9, 52.3, and 37.3 percent per year. Further, 75 percent (or 986) of the total 1,317 firms are clas-

¹¹ In fact, if the purpose was to predict changes in dividend behavior, it would be incorrect to use *in-sample* observations for the discriminant analysis.

¹² This trend persists for similarly formed subgroups within dividend payout categories, exchange groups, and industry classifications, although the results are not reported here.

Table II
Correlations among Variables

All financial variables are for the beginning of the fiscal year, except cash flow and investment, which represent firm cash flow and capital expenditures during period t . Cash flow, investment, and slack are all scaled by net fixed assets at the beginning of fiscal year t . The discriminant score (Z) is calculated using discriminant analysis according to equation (2). A full description of the variables is included in the Appendix.

	Cash Flow/ Fixed Assets	Current Ratio	Debt Ratio	Fixed Charge Coverage	Investment/ Fixed assets	Market-to- Book Ratio	Net Income Margin (%)	Sales Growth (%)	Slack/ Fixed Assets	Discriminant Score (Z)
Cash flow/Fixed assets	1.00									
Current ratio	0.11*	1.00								
Debt ratio	-0.18*	-0.33*	1.00							
Fixed charge coverage	0.21*	0.19*	-0.43*	1.00						
Investment/Fixed assets	0.37*	0.17*	-0.23*	0.18*	1.00					
Market-to-book ratio	0.21*	0.02	-0.12*	0.21*	0.24*	1.00				
Net income margin (%)	0.34*	0.08*	-0.14*	0.24*	0.13*	0.10*	1.00			
Sales growth (%)	0.19*	0.02	-0.01	0.11*	0.24*	0.20*	0.21*	1.00		
Slack/Fixed assets	0.38*	0.47*	-0.33*	0.13*	0.40*	0.08*	0.02	0.05*	1.00	
Discriminant score (Z)	0.32*	-0.07*	-0.29*	0.32*	0.18*	0.19*	0.80*	0.55*	-0.08*	1.00

*Significant at the 1 percent level.

Table III
Selected Financial Ratio Means for Financially
Constrained Groups (1988–1994)

All financial variables are for the beginning of the fiscal year, except cash flow and investment, which represent firm cash flow and capital expenditures during period t . K is the firm's beginning-of-period net fixed assets value. The discriminant score (Z) is calculated using discriminant analysis according to equation (2). A full description of the variables is included in the Appendix. Predicted Group 1 includes firms that are classified as likely to increase dividends in year t according to discriminant analysis, Predicted Group 2 includes firms that are classified as likely to decrease dividends per share (DPS) in year t . The FC, PFC, and NFC groups are formed by sorting all firms according to their discriminant scores. Every year, the firms with the lowest discriminant scores (the bottom one-third) are categorized as financially constrained (FC); the next one-third are categorized as partially financially constrained (PFC); and the top one-third are categorized as not financially constrained (NFC).

	Predicted Group 1 (likely to increase DPS)	Predicted Group 2 (likely to decrease DPS)	FC firms (financially constrained)	PFC firms (partially financially constrained)	NFC firms (not financially constrained)
Net fixed assets (K)	\$803m	\$591m	\$507m	\$787m	\$656m
Current ratio	2.37	2.54	2.74	2.37	2.62
Debt ratio	0.18	0.28	0.31	0.22	0.14
Fixed charge coverage	18.3	4.8	3.0	8.8	24.6
Net income margin (%)	7.2	-1.2	-4.8	4.2	9.6
Market-to-book ratio	2.58	1.50	1.65	1.91	2.99
Sales growth (%)	15.1	-0.6	-2.3	9.0	23.5
Slack/ K	1.30	1.30	1.93	1.46	1.75
Cash flow/ K	0.52	0.24	0.23	0.42	0.75
Investment/ K	0.27	0.19	0.21	0.24	0.33
Discriminant score (Z)	0.51	-1.45	-1.77	-0.21	1.05

sified as NFC in at least one year, with figures of 83 and 74 percent for the PFC and FC groups. This indicates that individual firm financial status does change significantly from one year to the next. In fact, only 17 firms are classified as PFC for all seven years, and only 49 and 80 are classified as NFC and FC for the entire period.

C. Regression Estimation

The following variation of the FHP88 regression equation is estimated using fixed firm and year effects:

$$I/K_{it} = \beta_{M/B} (M/B)_{it} + \beta_{CF/K} (CF/K)_{it} + u_{it}. \quad (3)$$

I represents investment in plant and equipment during period t ; K is the beginning-of-period book value for net property, plant, and equipment; CF represents current period cash flow to the firm as measured by net income plus depreciation plus the change in deferred taxes; and M/B represents the firm's common equity market-to-book ratio based on the previous year's actual market value at year-end. Fixed effects estimation maintains separate

intercepts for each firm and for each year in order to account for unobserved relationships between investment and the independent variables, and to capture business-cycle influences.¹³

The use of market-to-book ratio to proxy for growth opportunities follows the approach of KZ. This differs from FHP88 who calculate Tobin's q based on replacement costs and the average market value over the last quarter of the previous year; however, Perfect and Wiles (1994) indicate that improvements obtained from the more involved computation of Tobin's q are limited. Further, KZ point out that using year-end market values can only be regarded as a methodological improvement because "the FHP88 measure will not distinguish between a firm whose stock price declines from 20 to 10 and a firm whose stock price increases from 10 to 20 at the end of the previous year" (p. 179). Current period cash flow (CF), scaled by K , is used to measure the liquidity variable. This follows the specification of most previous studies including FHP88 and KZ, and facilitates comparison of results with previous evidence.

D. Determination of Significance Levels

A major focus of this literature is the comparison of investment-liquidity sensitivities across different groups of firms. However, traditional tests designed to detect differences in coefficients are not appropriate because the error terms likely violate the required assumptions.¹⁴ As a result, conclusions regarding the existence of differences in investment-liquidity sensitivity across groups have been largely based on observing differences in magnitude and level of significance of the coefficient on the liquidity variable in regression estimates. This paper uses simulation evidence to determine the significance of observed differences in coefficient estimates.¹⁵

A bootstrapping procedure is used to calculate empirical p -values that estimate the likelihood of obtaining the observed differences in coefficient estimates if the true coefficients are, in fact, equal. Observations are pooled

¹³ Regression estimates are obtained using OLS and using fixed firm and year effects. The fixed effects estimates are obtained using two standard approaches that transform the actual observations before running regressions using the transformed values. The first approach involves subtracting firm means and year means from the actual observations; the second approach transforms the actual observations by taking first differences and using time dummy variables. The reported results are the *demeaned* or *within* fixed firm and year estimates, which coincide with estimates presented by FHP88 and KZ. The coefficients estimated using OLS and *first differences* are not reported here, however they are consistent with the reported estimates in terms of magnitude and observed patterns across groups. Hsiao (1986), Griliches and Hausman (1986), and Schaller (1993) suggest that obtaining consistent estimates from alternative panel data estimation techniques provides evidence of no serious errors in variables problems.

¹⁴ Traditional tests are generally designed for testing changes in parameters across time series data, where it may sometimes be reasonable to assume no heteroscedasticity in the resulting residuals. Panel data, with emphasis on cross-sectional data, likely violate the required assumptions. For example, the Chow test requires that the disturbance variance be the same for both regressions, while the standard Wald test requires independence of the error terms. These conditions are unlikely to be satisfied by panel data residuals.

¹⁵ I thank Raymond Kan for this suggestion.

from the two groups whose coefficient estimates are to be compared. Using n_1 and n_2 to denote the number of annual observations available from each group, we end up with a total of $n_1 + n_2$ observations every year. Each simulation randomly selects n_1 and n_2 observations each year from the pooled distribution and assigns them to group 1 and group 2, respectively. Coefficient estimates are then determined for each group using these observations, and this procedure is repeated 5000 times. The empirical p -value is the percentage of simulations where the difference between coefficient estimates (d_i) exceeds the actual observed difference in coefficient estimates (d_{Sample}). This p -value tests against the one-tailed alternative hypothesis that the coefficient of one group is greater than that of the other group ($H_1: d > 0$). For example, a p -value of 0.01 indicates that only 50 out of 5000 simulated outcomes exceeded the sample result, which implies the sample difference is significant, and supports the notion that $d > 0$.

III. Discussion of Results

A. Results

Regression estimates for the entire sample are presented in Table IV and indicate that firms' investment decisions are sensitive to investment opportunities as proxied by market-to-book, but are even more sensitive to liquidity variables. This is consistent with evidence from previous studies. Regression results for the FC, PFC, and NFC groups are also presented in Table IV. They indicate that liquidity and market-to-book are significant determinants of investment (at the 1 percent significance level) for all three groups. The adjusted R^2 values range from 7.78 percent to 18.24 percent, which is consistent with previous studies. The coefficients for market-to-book ratios are not significantly different across the three groups.

Coefficients for liquidity variables are all positive and significant, which suggests firm investment decisions are sensitive to the availability of internal funds. More important, the investment outlays of the NFC firms are significantly more sensitive to liquidity than that of PFC and FC firms, and PFC firms are more liquidity sensitive than FC firms. The estimated cash flow coefficients for the NFC, PFC, and FC groups are 0.153, 0.090, and 0.064. The observed differences between the NFC coefficient estimates and those for the other two groups are significant at the 1 percent significance level, and the difference between the PFC estimate and the FC estimate is significant at the 8.30 percent level. These results provide strong support for the KZ conclusions, using a much larger, broader sample and an objective classification scheme.

It is important to ensure that this is a general result across different categories of firms. In order to obtain more homogeneous groups and reduce the potential impact of dividend policy, the entire sample is divided into dividend payout groups, similar to the original FHP88 approach. In particular, firm-year observations are delegated to three groups: (i) those with zero dividend payout (Pay0); (ii) those with greater than zero but less than

Table IV
Regression Results for the Total Sample (1317 firms)

Reported coefficients are the *within* fixed firm and year estimates over the 1988–1994 sample period (*t*-statistics are in parentheses). Capital expenditures divided by net fixed assets is the dependent variable. The firm's market-to-book ratio and cash flow/net fixed assets are the independent variables. The FC, PFC, and NFC groups are formed by sorting all firms according to their discriminant scores. Every year, the firms with the lowest discriminant scores (the bottom one-third) are categorized as financially constrained (FC); the next one-third are categorized as partially financially constrained (PFC); and the top one-third are categorized as not financially constrained (NFC). The empirical *p*-values are determined using the simulation procedure described in Section II. They are estimated based on the null hypothesis that the coefficients are equal for the two groups under consideration. The alternative hypothesis is that the coefficient for the first group is greater than that of the second group. For example, the *p*-value of 0.9168 in the market-to-book column for NFC versus PFC suggests that the market-to-book coefficient for the NFC group is greater than that for the PFC group at the 91.68 percent significance level. The 0.0046 *p*-value in the next column suggests that the coefficient estimate for Cash Flow/Net Fixed Assets is greater for the NFC group than for the PFC group (at the 0.46 percent level of significance).

	Market-to-Book	Cash Flow/Net Fixed Assets	Adjusted <i>R</i> ²	Number of Observations
Regression estimates				
Total sample	0.024 (12.3)	0.096 (29.7)	11.76%	9219
FC firms (financially constrained)	0.020 (5.8)	0.064 (14.0)	7.78%	3073
PFC firms (partially financially constrained)	0.028 (7.7)	0.090 (14.1)	9.28%	3073
NFC firms (not financially constrained)	0.018 (5.8)	0.153 (23.5)	18.24%	3073
Empirical <i>p</i> -values				
PFC versus FC	0.1344	0.0830		
NFC versus FC	0.5890	0.0000		
NFC versus PFC	0.9168	0.0046		

30 percent payout ($\text{Pay} < 30$); and (iii) those with 30 to 100 percent payout ($\text{Pay} > 30$).¹⁶ These dividend payout groups are then sub-divided according to discriminant scores every year as above to determine the FC, PFC, and NFC groups within each dividend payout category. Table V presents regression results for these subgroups that confirm the general conclusions above—namely, that investments of the NFC firms are the most sensitive to liquidity, followed by the PFC firms, and finally by the FC firms. This result is strongest for the zero payout group, which is similar to the group analyzed by KZ, lending additional support to their conclusions. I also examine the generality of these results by dividing the sample into groups based on exchange listing and industry classification. These groups are then subdivided according to discriminant scores as above to determine the FC, PFC, and NFC groups within each category. Regression results for these subgroups, which are not reported here, confirm the general results above.¹⁷

An additional test is performed to examine the robustness of results to the influence of firm leverage. The importance of controlling for firm leverage is demonstrated by Lang, Ofek, and Stulz (1996), who find that future growth and investment are negatively related to leverage, particularly for firms with low Tobin's q values and high debt ratios. This implies the significance of examining whether the pattern of investment-liquidity sensitivities detected in this study could be attributed to a systematic tendency of the classification scheme to assign firms to a group whose investment decisions are more sensitive to firm leverage than those of other groups. This hypothesis is tested by running regressions that include debt to total assets as an independent variable in the regression specification, in addition to market-to-book and CF/K. The results are not reported here; however, the coefficient on the debt to total assets variable is found to be negative and significant for all three groups, which confirms the results of Lang et al. Despite the relevance of firm leverage, the cash flow coefficients remain virtually identical for all of the groups, which is the primary concern of the present study.¹⁸ This evidence suggests that the observed pattern of investment liquidity-sensitivities is not attributable to a *leverage effect*.

¹⁶ This approach differs slightly from the FHP88 classification scheme, which divides firms based on payout ratios over the entire sample period and does not allow group composition to vary through time. The FHP88 approach is also used, with no resulting change in conclusions.

¹⁷ The result that the least constrained firms are most sensitive to liquidity is robust to a number of alternative sorting arrangements whose results are not reported, including: (i) whether the sample is divided into two or three groups; (ii) groups formed using absolute discriminant score cutoff points for the entire period to create the NFC, PFC, and FC groups, rather than dividing the sample into thirds each year; (iii) groups formed based on the dividend groups (as defined in Table I); and (iv) groups formed on predicted dividend groups (as defined in Table III).

¹⁸ In particular, the coefficients on CF/K for the FC, PFC, and NFC groups changed from 0.064, 0.090, and 0.153 to 0.064, 0.086, and 0.149, and the adjusted R^2 values increased from 7.78, 9.28, and 18.24 percent to 10.17, 11.43, and 19.63 percent.

Table V
Regression Results for Dividend Payout Groups

Reported coefficients are the *within* fixed firm and year estimates over the 1988–1994 sample period (*t*-statistics are in parentheses). Capital expenditures divided by net fixed assets is the dependent variable. The firm’s market-to-book ratio and cash flow/net fixed assets are the independent variables. Pay0 represents the group formed using firm year observations where the firm’s dividend payout was zero; Pay<30 represents payouts greater than zero but less than 30 percent; and Pay>30 represents payouts of 30 to 100 percent. The FC, PFC, and NFC groups are formed by sorting firms within a given payout group according to their discriminant scores. Every year, the firms in the group with the lowest discriminant scores (the bottom one-third) are categorized as financially constrained (FC); the next one-third are categorized as partially financially constrained (PFC); and the top one-third are categorized as not financially constrained (NFC). The number of observations for the PFC group may be larger than the other two because the leftover firms are assigned to the PFC group when the total number of firms in a payout group during a given year is not a multiple of three. The empirical *p*-values are determined using the simulation procedure described in Section II. They are estimated based on the null hypothesis that the coefficients are equal for the two groups under consideration. The alternative hypothesis is that the coefficient for the first group is greater than that of the second group. For example, the *p*-value of 0.6114 in the market-to-book column for NFC versus PFC in the Pay0 group suggests that the market-to-book coefficient for the NFC group is greater than that for the PFC group at the 61.14 percent significance level. The 0.0002 *p*-value in the next column suggests that the coefficient estimate for Cash Flow/Net Fixed Assets is greater for the NFC group than for the PFC group in the Pay0 group (at the 0.02 percent level of significance).

	Market-to-Book	Cash Flow/ Net Fixed Assets	Adjusted R^2	Number of Observations
Panel A: Pay0 Group				
Regression estimates				
FC firms	0.021 (4.8)	✓ 0.057 (9.9)	8.23%	1520
PFC firms	0.028 (5.3)	0.080 (11.5)	11.03%	1529
NFC firms	0.024 (4.7)	<u>0.159 (18.7)</u>	22.47%	1520
Empirical p -values				
PFC versus FC	0.2694	0.1378	Bootstrap	
NFC versus FC	0.3834	<u>0.0000</u>		
NFC versus PFC	0.6114	0.0002		
Panel B: Pay<30 Group				
Regression estimates				
FC firms	0.035 (3.7)	✓ 0.054 (3.6)	4.06%	709
PFC firms	0.017 (2.4)	0.133 (6.1)	6.66%	712
NFC firms	0.005 (0.9)	<u>0.147 (8.4)</u>	10.67%	709
Empirical p -values				
PFC versus FC	0.9008	0.1826		
NFC versus FC	0.9878	<u>0.0788</u>		
NFC versus PFC	0.8584	0.3974		

Table V—Continued

	Market-to-Book	Cash Flow/ Net Fixed Assets	Adjusted R^2	Number of Observations
Panel C: Pay>30 Group				
Regression estimates				
FC firms	0.018 (2.6)	✓0.051 (2.6)	1.80%	839
PFC firms	0.010 (2.0)	0.105 (6.3)	5.91%	842
NFC firms	0.023 (4.5)	<u>0.119 (10.0)</u>	13.40%	839
Empirical p -values				
PFC versus FC	0.8008	0.1640		
NFC versus FC	0.3008	<u>0.1738</u>		
NFC versus PFC	0.1026	0.4382		

B. Interpretation

The high investment liquidity sensitivity of the unconstrained firms appears puzzling at first glance. However, it is consistent with Mayer's (1990) empirical evidence that internal financing is the dominant source of financing for all firms, which implies that investment decisions of the majority of firms are sensitive to current liquidity. It also concurs with the results of Lamont (1997) who documents a large decrease in the capital expenditures of non-oil subsidiaries of oil conglomerates in reaction to the 1986 drop in oil prices. Lamont concludes that large reductions in cash flow and collateral value lead to decreased investment, independent of changes in available investment opportunities.

This behavior supports the free cash flow argument presented by Jensen (1986) that firms increase investment in response to the availability of cash flows. Jensen argues that "managers have incentives to cause firms to grow beyond optimal size" since "growth increases managers' power by increasing the resources under their control" (p. 323). It is also consistent with the conclusion of Bernanke and Gertler (1990) that "both the quantity of investment spending and its expected return will be sensitive to the *creditworthiness* of borrowers (as reflected in their net worth positions)" (p. 89). Alternatively, KZ suggest that "managerial risk aversion" may contribute to the correlation between investment and liquidity. Given the size and changing group composition of the approach used in this study, the observed sensitivities are not likely to be driven by overly risk-averse managers in a particular group, and this may in fact, be a general behavioral characteristic of most firm managers.

IV. Conclusions

The sensitivity of firm investment decisions to liquidity status is examined using data for 1,317 U.S. firms over the 1988 to 1994 period. Following the basic approach of Kaplan and Zingales (1997), firms are classified according to financial statement variables that are related to their ability to raise external finance. An objective multivariate classification index, similar to Altman's Z factor, is used to determine firm financial status and this status is allowed to vary from one period to the next. The approach captures desired cross-sectional properties of a large number of firms and successfully classifies firms that increase or decrease dividends 74 percent of the time. Additionally, a bootstrap methodology is used to determine significance levels of observed differences in coefficient estimates across different firm categories.

Large sample evidence demonstrates that the investment decisions of firms with high creditworthiness are significantly more sensitive to the availability of internal funds than are firms that are less creditworthy. This strongly supports the small-sample evidence of Kaplan and Zingales (1997), who also find that the least constrained firms are the most sensitive to cash flow availability, contrary to the conclusions of several previous studies.

Appendix

The financial variables utilized are calculated as follows:

$$(1) \text{ Current ratio} = \frac{\text{current assets}}{\text{current liabilities}};$$

$$(2) \text{ Debt ratio} = \frac{\text{current portion of long-term debt} + \text{long-term debt}}{\text{total assets}};$$

$$(3) \text{ Fixed charge coverage ratio}$$

$$= \frac{\text{earnings before interest and taxes}}{\text{interest expense} + \text{preferred dividend payments} \times \left(\frac{1}{1 - \text{tax rate}} \right)};$$

$$(4) \text{ Net income} = \text{net income before extraordinary items} \pm \text{extraordinary items and discontinued operations};$$

$$(5) \text{ Net income margin} = \frac{\text{net income}}{\text{net sales}};$$

$$(6) \text{ Cashflow} = \text{net income} + \text{depreciation and/or amortization expense} + \text{change in deferred taxes};$$

$$(7) \text{ Investment} = \text{net capital expenditures}$$

- (8) Net sales growth = $\frac{\text{net sales}_t - \text{net sales}_{t-1}}{\text{net sales}_{t-1}}$;
- (9) Dividend payout = $\frac{\text{total dividends paid}}{\text{net income}}$;
- (10) Slack = cash + short term investments + (0.50 × inventory) + (0.70 × accounts receivable) – short term loans;
- (11) Net fixed assets = net property, plant and equipment;
- (12) Market-to-book = $\frac{\text{market value of common equity}}{\text{book value of common equity}}$.

REFERENCES

- Altman, Edward I., 1968, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* 23, 589–609.
- Altman, Edward I., Robert G. Haldeman, and P. Narayanan, 1977, Zeta analysis: A new model to identify bankruptcy risk of corporations, *Journal of Banking and Finance* 1, 29–54.
- Bernanke, Ben, and Mark Gertler, 1989, Agency costs, net worth, and business fluctuations, *American Economic Review* 79, 14–31.
- Bernanke, Ben, and Mark Gertler, 1990, Financial fragility and economic performance, *Quarterly Journal of Economics* 105, 97–114.
- Bond, Stephen, and Costas Meghir, 1994, Dynamic investment models and the firm's financial policy, *Review of Economic Studies* 61, 197–222.
- Fazzari, Steven, R. Glenn Hubbard, and Bruce Petersen, 1988, Financing constraints and corporate investment, *Brookings Papers on Economic Activity* 19, 141–195.
- Fazzari, Steven, R. Glenn Hubbard, and Bruce Petersen, 1996, Financing constraints and corporate investment: Response to Kaplan and Zingales, NBER Working Paper No. 5462.
- Gertler, Mark, 1992, Financial capacity and output fluctuation in an economy with multi-period financial relationship, *Review of Economic Studies* 59, 455–472.
- Gilchrist, Simon, and Charles Himmelberg, 1995, Evidence for the role of cash flow in investment, *Journal of Monetary Economics* 36, 541–572.
- Greenwald, Bruce, Joseph Stiglitz, and Andrew Weiss, 1984, Information imperfections and macroeconomic fluctuations, *American Economic Review* 74, 194–199.
- Griliches, Zvi, and Jerry A. Hausman, 1986, Errors in variables in panel data, *Journal of Econometrics* 31, 93–118.
- Hoshi, Takeo, Anil K. Kashyap, and David Scharfstein, 1991, Corporate structure liquidity and investment: Evidence from Japanese panel data, *Quarterly Journal of Economics* 106, 33–60.
- Hsiao, Cheng, 1986, *Analysis of Panel Data* (Cambridge University Press, Cambridge, U.K.).
- Hubbard, R. Glenn, 1998, Capital market imperfections and investment, *Journal of Economic Literature* 36, 193–225.
- Jensen, Michael C., 1986, Agency costs of free cash flow, corporate finance, and takeovers, *American Economic Review* 76, 323–329.
- Kaplan, Steven N., and Luigi Zingales, 1997, Do financing constraints explain why investment is correlated with cash flow?, *Quarterly Journal of Economics* 112, 169–215.
- Lamont, Owen, 1997, Cash flow and investment: Evidence from internal capital markets, *Journal of Finance* 52, 83–109.
- Lang, Larry, Eli Ofek, and René M. Stulz, 1996, Leverage, investment and firm growth, *Journal of Financial Economics* 40, 3–29.

- Mayer, Colin, 1990, Financial systems, corporate finance, and economic development; in R. Glenn Hubbard, ed.: *Asymmetric Information, Corporate Finance and Investment* (The University of Chicago Press, Chicago).
- Modigliani, Franco, and Merton H. Miller, 1958, The cost of capital, corporation finance, and the theory of investment, *American Economic Review* 48, 261–297.
- Myers, Stewart C., 1984, The capital structure puzzle, *Journal of Finance* 39, 575–592.
- Myers, Stewart C., and Nicholas Majluf, 1984, Corporate financing and investment decisions when firms have information that investors do not have, *Journal of Financial Economics* 13, 187–221.
- Oliner, Stephen D., and Glenn D. Rudebusch, 1992, Sources of the financing hierarchy for business investment, *Review of Economics and Statistics* 74, 643–654.
- Perfect, Steven, and Kenneth Wiles, 1994, Alternative constructions of Tobin's q : An empirical comparison, *Journal of Empirical Finance* 1, 313–341.
- Schaller, Huntley, 1993, Asymmetric information, liquidity constraints, and Canadian investment, *Canadian Journal of Economics* 26, 552–574.
- Schiantarelli, Fabio, 1995, Financial constraints and investment: A critical review of methodological issues and international evidence, Working paper, Boston College.
- Whited, Toni, 1992, Debt, liquidity constraints, and corporate investment: Evidence from panel data, *Journal of Finance* 47, 1425–1460.

Corporate Financial Policy and the Value of Cash

MICHAEL FAULKENDER and RONG WANG*

ABSTRACT

We examine the cross-sectional variation in the marginal value of corporate cash holdings that arises from differences in corporate financial policy. We begin by providing semi-quantitative predictions for the value of an extra dollar of cash depending upon the likely use of that dollar, and derive a set of intuitive hypotheses to test empirically. By examining the variation in excess stock returns over the fiscal year, we find that the marginal value of cash declines with larger cash holdings, higher leverage, better access to capital markets, and as firms choose greater cash distribution via dividends rather than repurchases.

WHAT VALUE DO SHAREHOLDERS PLACE ON THE CASH THAT FIRMS HOLD, and how does that value differ across firms? While an extensive literature attempts to estimate the value of adding debt to a firm's capital structure, the search for estimates of the value of additional cash has not received nearly as much attention. This is a non-trivial oversight considering that corporate liquidity enables firms to make investments without having to access external capital markets, and to thereby avoid both transaction costs on either debt or equity issuance and information asymmetry costs that are often associated with equity issuances. Moreover, corporate liquidity reduces the likelihood of incurring financial distress costs if the firm's operations do not generate sufficient cash flow to service obligatory debt payments. Corporate liquidity comes at a cost, however, since interest earned on corporate cash reserves is often taxed at a higher rate than interest earned by individuals. Furthermore, cash may provide funds for managers to invest in projects that offer non-pecuniary benefits but destroy shareholder value (Jensen and Meckling (1976)). Given the extent to which the literature examines the effect of these same frictions on capital structure, it is surprising that the value implications of holding cash in the presence of these frictions have not been similarly explored.¹

*Faulkender and Wang are at the Olin School of Business, Washington University in St. Louis. We thank Luca Benzoni, Murillo Campello, Gerald Garvey, Robert Goldstein, Todd Milbourn, Mitchell Petersen, Robert Stambaugh (the editor), Rene Stulz, Rohan Williamson, an anonymous referee, the associate editor, and seminar participants at Washington University in St. Louis and the 2004 Western Finance Association Annual Conference for helpful comments.

¹ There has been some work that estimates the value implications of excess cash flow. For instance, Hanson (1992) and Smith and Kim (1994) both find that bidding firms with high excess free cash flow exhibit low excess stock returns around merger announcements. Their estimated coefficients can be interpreted as the value destruction associated with high levels of excess free cash flow.

Recent empirical studies of corporate cash holdings (e.g., Opler et al. (1999), Harford (1999)) examine the cross-sectional variation in the level of cash holdings related to the above theoretical benefits and costs.² Consistent with the hypothesized effects, they find that firms with stronger growth opportunities, riskier cash flows, and more limited access to capital markets hold higher cash balances. Now that we understand the characteristics that determine how much cash firms hold, we move to the question of what value the market places on the cash holdings of firms and how that value varies cross-sectionally.

In generating empirical predictions, we argue that the value (to the equity holder) of one additional dollar of cash reserves should vary considerably depending upon whether that dollar is more likely to go to: (1) increasing distributions to equity via dividend payments or share repurchases, (2) decreasing the amount of cash that needs to be raised in the capital markets, depending upon the firm's capital market accessibility, or (3) servicing debt or other liabilities of the firm.

For firms whose cash reserves appear to greatly exceed their needs in the foreseeable future, an additional dollar of cash reserves is more likely to be distributed to equity holders through dividends and/or stock repurchases. However, because of the "dividend tax," only the fraction $(1 - \tau_d)$ ends up in the hands of shareholders.³ As such, the marginal value of cash is reduced to $(1 - \tau_d)$, which can be significantly below \$1. Additionally, if firms use their cash to pay down debt or other liabilities, a small increase in cash reserves partially goes to increasing debt value, not solely to increasing equity value. Thus, the equity market will place a lower value on an additional dollar of cash for high leverage firms relative to the marginal value of cash for a firm with little debt. In contrast, for those firms that need to raise cash from external markets because they have value-enhancing investment opportunities but their internal funds are low, the marginal value of cash should be higher than \$1, with the exact amount depending upon the transactions costs (direct or otherwise) that are incurred by accessing the capital markets. Therefore, the marginal value of cash should decline as cash holdings increase because as the cash position of the firm improves, firms become more likely to distribute funds and less likely to raise cash.

We also argue that for firms that face greater financing constraints, especially those with valuable investment opportunities, the marginal value of cash should be higher than for firms that can easily raise additional capital. While financial constraints are often associated with information asymmetries between firms and capital providers, they can be thought of as tantamount to higher transactions costs in accessing external capital. In such a context, an additional dollar of internal funds enables a constrained firm to avoid

² Other related papers include Kim, Mauer, and Sherman (1998), Pinkowitz and Williamson (2001), Billett and Garfinkle (2004), Faulkender (2004), Ozkan and Ozkan (2002), Mikkelsen and Parth (2003), Hartzell, Titman, and Twite (2005), and Dittmar, Mahrt-Smith, and Servaes (2003).

³ During the sample period, the appropriate tax rate τ_d varied considerably depending upon whether the cash distribution was done through a dividend payment or through a stock repurchase. We discuss this point in detail below.

these higher costs of raising funds, thereby, rendering additional internal funds relatively more valuable.

Below, we use these arguments to formalize hypotheses about how the marginal value of cash should vary across firm characteristics. We then test these hypotheses empirically and find broad support for them. Indeed, our main empirical results include:

- (1) The average marginal value of cash across all firms is \$0.94.
- (2) As firms' cash levels and leverage increase, their marginal value of cash decreases significantly.
- (3) For those firms that distribute cash, the marginal value of cash is \$0.13 higher if they do so by stock repurchase rather than by dividend payments. This number is consistent with a dividend tax rate that is 13% higher than the capital gains tax rate on repurchases for the marginal shareholder.⁴
- (4) The average marginal value of cash for those firms that are likely to have more difficulty accessing capital is significantly higher than for those firms that are less likely to be constrained.
- (5) The difference in the marginal value of cash between constrained firms and unconstrained firms is especially large among those firms that appear to have valuable investment opportunities but low levels of internal funds.

In a similar paper, Pinkowitz and Williamson (2004) also examine the marginal value of cash, focusing largely on the cross-sectional variation related to the firm's investment opportunity set.⁵ Using the methodology of Fama and French (1998), they find that shareholders of a firm with better growth options and more volatile investment opportunities place higher values on the firm's cash than a firm with fewer, more stable growth opportunities. In contrast, we focus on how the value of cash varies with firm financial characteristics and we use a methodology that examines the variation in excess equity returns rather than in the level of the market-to-book ratio. Because we normalize all independent variables by the firm's equity value at the end of the previous fiscal year, we can interpret our estimated coefficients as the change in equity value associated with a \$1 change in the corresponding independent variable. Using this methodological approach, we report estimated coefficients that appear to be both quantitatively and qualitatively consistent with all of our hypotheses.

In Section I, we argue that there are essentially three different cash regimes and that the marginal value of cash depends upon the likelihood with which a firm will find itself in each of these different regimes. We then generate a set of hypotheses for how the marginal value of cash should be affected by

⁴ We would expect that this value differential would shrink in the future following the recent reduction in the dividend tax rate for individuals. We do not yet have sufficient data to verify that this has indeed occurred.

⁵ In another related paper, Pinkowitz, Stulz, and Williamson (2006) extend the examination to cross-country differences in the marginal value of cash.

changes in the level of corporate liquidity, the amount of debt in the firm's capital structure, and the accessibility of external capital. In Section II, we discuss the empirical methodology that we utilize to test these hypotheses and provide further explanations for why we prefer this approach in estimating the value associated with a particular firm characteristic. The data sources and summary statistics are provided in Section III.

Section IV contains the results of testing our empirical hypotheses. We begin with our baseline specification, which estimates the effects of leverage and the level of cash on the marginal value of cash for the firms in our sample. Because our approach uses excess returns, estimating the marginal value of cash requires estimating the *unexpected* change in the firm's cash position over the corresponding return period. We therefore conduct numerous robustness tests in which we estimate the expected change in cash and then use the difference between the realized change and the expected change in our analysis. As the results demonstrate, our findings are quite stable, both statistically and economically, across these different measures. We then move on to utilize multiple definitions of being financially constrained to examine how capital market accessibility impacts the value that the market places on additional corporate cash. Finally, we examine subsamples of firm-years that are most likely to fall into our three cash regimes and further demonstrate that the estimated marginal values of cash have differences that are consistent with our hypotheses. Section V concludes.

I. Three Cash Regimes

As mentioned in the Introduction, the value (to the equity holder) of one additional dollar of cash reserves should vary considerably depending upon the cash regime to which a firm is likely to belong. The identification of the three different regimes here is similar to that of Hennessy and Whited (2005), who investigate optimal dynamic capital structure.⁶ This identification is important because it not only allows us to make qualitative predictions for how the marginal value of cash should vary cross-sectionally, but it also allows us to provide semi-quantitative estimates for the marginal value of cash cross-sectionally.

A. Regime I: Distributing Cash

Consider a firm that is currently carrying excess cash in that the sum of current cash on hand and expected short-term earnings is more than sufficient to fund both the short-term liabilities of the firm and any possible investments in new value-enhancing projects that may arise. If there were no costs to holding cash, then it would be optimal for the firm to retain large cash reserves

⁶ Hennessy and Whited (2005) argue that an additional dollar of debt is more valuable if it goes to reducing costly external equity issuance rather than increasing cash distributions. However, since they only investigate one-period risk-free debt, they do not have a situation similar to our second regime discussed below.

rather than distribute the excess cash in order to guarantee that the firm will not need to incur the transaction costs associated with raising cash. However, taxes and agency costs generate costs to holding excess cash. First, because the corporate tax rate is generally higher than the personal tax rate paid on interest income, investors are better off if they rather than the firm hold excess cash. Second, agency costs due to the “free cash flow” problem (Jensen (1986)) are more likely for firms with excess cash reserves. Hence, it will be optimal for firms with excess cash to distribute funds to shareholders via dividends or share repurchases, and shareholders will not place a high value on a marginal dollar of cash for these firms.

Specifically, we argue that the marginal value of cash for firms that are likely to distribute large sums of cash is less than \$1. Defining τ_d as the tax rate on dividends, only $(1 - \tau_d)$ of every dollar distributed by the firm in the form of dividends finds its way into the hands of shareholders. Moreover, the fact that the corporate tax rate τ_c on earned interest is typically greater than the tax rate τ_i on earned interest for individuals implies that the marginal value of any excess cash that is not immediately distributed is significantly lower than $(1 - \tau_d)$. Indeed, consider the extreme example of a firm with no debt whose only asset is cash placed into a risk-free security. Without taxes and payouts, the cash holdings grow according to

$$dC_t = rC_t dt. \quad (1)$$

However, if we assume that earnings are taxed at τ_c and that the cash payout is a fraction β of the after-tax earnings, then the cash holdings grow according to

$$dC_t = rC_t(1 - \tau_c)(1 - \beta)dt, \quad (2)$$

implying that

$$C_t = C_0 e^{rt(1-\tau_c)(1-\beta)}. \quad (3)$$

Moreover, the distribution to shareholders over the interval dt would be

$$dX = rC_t(1 - \tau_c)\beta dt, \quad (4)$$

which would be taxed at the dividend rate τ_d . Note that this after-tax cash flow is risk free so the appropriate discount rate for this stream of cash flows is the personal after-tax risk-free rate $r(1 - \tau_i)$. Hence, the value of this equity claim is

$$\begin{aligned} E(C_0) &= \int_0^\infty dX(1 - \tau_d)e^{-rt(1-\tau_i)} \\ &= C_0(1 - \tau_d)(1 - \tau_c)r\beta \int_0^\infty e^{rt(1-\tau_c)(1-\beta)} e^{-rt(1-\tau_i)} dt \\ &= C_0(1 - \tau_d) \frac{(1 - \tau_c)\beta}{(1 - \tau_i) - (1 - \tau_c)(1 - \beta)} \end{aligned} \quad (5)$$

and the marginal value of cash for this firm is

$$\frac{\partial E}{\partial C} = (1 - \tau_d) \frac{(1 - \tau_c)\beta}{(1 - \tau_i) - (1 - \tau_c)(1 - \beta)}. \quad (6)$$

Note that in the special case in which $\tau_c = \tau_i$, that is, interest earned by the corporation is not taxed more heavily than interest earned by individuals, equation (6) reduces to

$$\frac{\partial E}{\partial C} = (1 - \tau_d). \quad (7)$$

However, even small differences between τ_c and τ_i can have large effects on the marginal value of cash for levels of β that are observed in the data. For example, if we consider the base case $\tau_d = 0.25$, $\tau_c = 0.35$, $\tau_i = 0.30$, and $\beta = 0.25$, we find that the marginal value of cash for this “cash cow” is

$$\frac{\partial E}{\partial C} = 0.57, \quad (8)$$

which is significantly lower than $(1 - \tau_d) = 0.75$ due to the dividend tax alone. This result is reminiscent of the insights of Berk and Stanton (2004) who demonstrate that the closed-end fund discount can be explained by a small cost given that the payout ratio is small. This result suggests that the marginal value of cash for firms with excess cash (i.e., bad investment opportunities and high cash levels) are predicted to be well below \$1. The presence of agency costs due to free cash flow problems would only reduce this estimate further.

B. Regime II: Servicing Debt or Other Liabilities

For highly leveraged firms, contingent claims analysis (e.g., Black and Scholes (1973), Merton (1973)) predicts that almost all firm value is in the hands of the debt holders. As such, a small increase in cash reserves goes largely to increasing debt value, not equity value, implying in turn that the equity market will place a low value on an additional dollar of cash for these firms. Furthermore, this “option theory” predicts that the marginal value of cash to equity holders should increase as leverage declines, since the probability of avoiding bankruptcy, and therefore the probability of the extra dollar finding its way into the pocket of equity holders, increases.

C. Regime III: Raising Cash

We argue that the marginal value of cash for firms that are likely to raise cash in the near future should be higher than \$1, and that the amount varies depending upon the ease with which the firm can access the capital markets.

Consider two firms that need to raise capital immediately because they have a value-enhancing project and their current cash holdings are low. Assume that these firms are identical except that Firm A has one additional dollar of cash

reserves. Hence, Firm B needs to raise one more dollar of cash than Firm A to fund the investment. In the presence of a proportional transactions cost (direct or otherwise) f that is incurred by accessing the capital markets, raising this additional dollar will cost Firm B an additional $(\frac{1}{1-f})$.

For firms that raise cash optimally, the marginal value of cash should reach an upper bound of $\frac{1}{1-f}$. The argument is straightforward: If the market currently values an additional dollar of cash at higher than $\frac{1}{1-f}$, then the firm can increase its equity value by raising additional cash now. Hence, under the objective of shareholder-maximizing behavior, firms should raise their cash levels so that the marginal value of cash never exceeds $\frac{1}{1-f}$. Assuming transactions costs are not too high, this argument suggests that the marginal value of cash will be slightly greater than \$1 for “unconstrained” firms that are at the margin of raising cash. As firms face financing constraints, which can be thought of as larger transactions costs f (whether direct or indirect), they are expected to have even higher marginal values of cash, all else equal.

D. Empirical Predictions

Now that we have explained why the marginal value of cash should vary considerably depending upon which regime the firm is likely to face, we seek to link firm financial characteristics to these regimes, by specifying a set of hypotheses that we empirically test below.

Hypothesis 1: The marginal value of cash is decreasing in the level of the firm’s cash position.

A firm with a low level of cash reserves is more likely than firms with high cash balances to be in the third cash regime, that is, needing to access the external capital markets to fund its short-term liabilities and investments. Due to the transactions costs (direct and indirect) incurred by accessing the capital markets, the value of an additional dollar of cash for such a firm is greater than one. Holding profitability constant, as cash holdings increase, the firm is less likely to access capital markets in the near future and is instead more likely to return cash to shareholders. Thus, greater cash levels reduce the probability of the firm being in regime three and instead, the first cash regime becomes more likely, in which case the value of an additional dollar of cash could be significantly lower than one, due to higher corporate tax rates relative to investor tax rates and the free cash flow problem. Therefore, for firms that are not near bankruptcy, the marginal value of cash should be a decreasing function of the cash level, as the likelihood of being in the high marginal cash value regime diminishes and the likelihood of being in the lower marginal value of cash regime increases.

Hypothesis 2: An extra dollar of cash holdings is less valuable for shareholders in highly levered firms than in firms with low leverage.

This hypothesis is common in most capital structure models. As firms generate more cash flow or accumulate higher cash balances, if the debt is risky, the increase in firm value is shared by the debt and equity holders. For firms with low leverage, and therefore less risky debt, an increase in the firm's cash position has very little impact on the probability of the debt holders being paid in full. As leverage increases, all else equal, more of the firm value generated by additional cash benefits the debt holders. This effect can be motivated by interpreting an equity security as a call option on the firm's value and thinking of the debt holders as being short a put option on the value of the firm. As the strike price increases, that is, as the firm takes on more debt, holding constant the value of the firm, the delta of the option decreases. So, while an increase in cash increases the value of the underlying firm, thereby increasing the value of both the debt and the equity, more of the value associated with the increase in cash will accrue to the equity holders as the firm has less leverage.⁷

Hypothesis 3: An extra dollar of cash holdings is more valuable for shareholders in financially constrained firms.

Returning to our discussion of firms that access capital markets, a firm that faces financial constraints can be thought of as facing a higher cost f when raising external funds. As a result, the marginal value of cash may be higher for these firms since internal funds enable the firm to avoid incurring this higher cost. Additionally, if the firm has investment opportunities, the higher the cost of raising external funds, the more likely it is that these value-enhancing projects will be forgone if internal funds are insufficient. Fazzari, Hubbard, and Petersen (1988) document the presence of an investment cash flow sensitivity that is consistent with financial constraints deterring firms from being able to make investments when internal funds are insufficient to fund them. If capital market access were perfect, then regardless of the firm's liquidity, it would always be able to fund positive net present value (NPV) projects. As access to capital becomes more difficult, forgoing positive NPV projects is more likely, absent internal funds. Therefore, for constrained firms, higher cash holdings increase the likelihood of taking positive NPV projects that would otherwise be forgone, whereas liquidity provides no such benefit for unconstrained firms. This effect should be most prevalent for firms that are more likely to have investment opportunities but little internal cash with which to fund those investments.

There are numerous other studies that present evidence consistent with our third hypothesis. Korajczyk and Levy (2003) find that target leverage is counter cyclical for relatively unconstrained firms, but pro-cyclical for relatively constrained firms; unconstrained firms time their security issuance to coincide

⁷ In the presence of agency costs, in which case there are conflicts between the interests of shareholders and the interests of debt holders, we would similarly expect to find the value of cash holdings to equity holders decreasing with leverage. When the firm has a large amount of debt, positive NPV projects could predominately benefit debt holders, leading to a debt overhang (or Myers's (1977) underinvestment) problem. Highly levered firms are more likely to have debt overhang problems and pass up good projects.

with periods of favorable macroeconomic conditions, while constrained firms do not. Almeida, Campello, and Weisbach (2004) find that financially constrained firms systematically save cash out of cash flow while unconstrained firms do not. Acharya, Almeida, and Campello (2004) build upon those results by separating out constrained firms based upon the correlation between cash flow and investment opportunities. They show that financially constrained firms whose investment opportunities arise when operating cash flows are relatively low save cash rather than pay down debt. On the other hand, unconstrained firms and constrained firms with a high correlation between the presence of investment opportunities and high cash flows pay down debt rather than save cash. These previous empirical results support the hypothesis that the accessibility of capital affects the capital structure and liquidity choices of firms, which should be accompanied by differences in the value of cash across firms with differential access.⁸

II. Empirical Methodology

In this paper, the basic questions we investigate are what value do shareholders place on an extra dollar of cash held by firms, and what financial characteristics affect that value? If shareholders believe that difficulty in accessing capital markets may sometimes lead firms to forgo value-creating investments, then a dollar of cash may be worth more than a dollar. Alternatively, if shareholders believe that extra cash only serves to increase (taxable) distributions, or only generates free cash flow problems, then the marginal value of cash may be significantly less than \$1.

To test these hypotheses, we develop a methodology that generates estimates of the additional value the market incorporates into equity values that result from changes in the cash position of firms over the fiscal year. Following Grinblatt and Moskowitz (2004) and Daniel and Titman (1997), our dependent variable is a stock's excess return over the fiscal year, which is defined to be stock i 's return during fiscal year t less the return of stock i 's benchmark portfolio during fiscal year t . The benchmark portfolios, defined below, are designed to offset the expected return component of stock i due to its size and market-to-book ratio at the beginning of the fiscal year. We regress that excess return on changes in firm characteristics, focusing on the estimated coefficient that corresponds to the variable measuring the ratio of the unexpected change in cash

⁸ An exception is DeAngelo, DeAngelo, and Wruck (2002), who suggest that financial constraints may actually be beneficial if the constrained firm is likely to waste additional cash on negative NPV projects. So, if financially unconstrained firms are more likely to be run by managers that only invest in positive NPV projects, whereas constrained firms are associated with relatively worse managers, then additional cash would actually be valued higher by shareholders of unconstrained firms. If this effect dominates, then we would expect the opposite of our hypothesis, namely, that cash is more valuable for unconstrained firms. However, we do not believe that our measures of financial constraints identify firms that invest in value-destroying projects, so our hypothesis is unchanged.

to the firm's lagged equity value.⁹ Since both the dependent and independent variables are standardized by the lagged market value of equity, the coefficient measures the dollar change in shareholder value resulting from a one dollar change in the amount of cash held by the firm.

We argue that our methodology for estimating the value associated with a firm characteristic is an improvement over the Fama and French (1998) methodology, which focuses on the cross-sectional variation in the market-to-book ratio, for two important reasons. First, we incorporate time-varying risk factors into our estimation. Part of the time-series variability in the market-to-book ratio used in Fama and French (1998) should come from differences over time in the compensation for risk, and therefore the market value of the firm. Their methodology controls for firm-specific characteristics that affect expected cash flows, but does not include measures that capture differences in sensitivities to risk factors, and therefore differences in discount rates. We address this by using a stock's benchmark return to control for the time-series variation in risk factors and the cross-sectional variation in exposures to those factors.¹⁰ Second, with regard to the dependent variable, unlike the ratio of market-to-book, equity returns are easy to measure and interpret. Fama and French (1998) note that they would "prefer to measure assets at replacement cost, but we do not have the necessary data." As a result, part of the variability in market-to-book may result from the cross-sectional differences in accounting for the book value of assets relative to their true replacement cost. If accounting methods across firms are correlated with liquidity, this correlation might bias the estimates of the marginal value of cash.¹¹

We recognize that stock returns should be affected both by common risk factors and by changes in firm-specific characteristics. Since firm-specific risk factors are very noisy and can be diversified away, most papers in the asset pricing literature only look at portfolio returns. However, since the emphasis of this paper is how changes in cash holdings affect shareholder's wealth, we need to examine individual stocks instead of portfolios. While we are interested in the change in equity value associated with changes in the cash holdings of firms, it is important to control for other factors that may be correlated with changes in cash that may also affect firm value. Therefore, we regress the excess equity return over the fiscal year on not only the change in cash holdings, but also on changes in a firm's profitability, financing policy, and investment policy. We initially assume that firms have the same sensitivity to these

⁹ Initially, we examine the realized change in cash over the fiscal year, essentially assuming that the market's expectation of the level of cash at the end of the fiscal year is the cash level at the end of the previous fiscal year. In robustness checks that follow the initial specification, we use three alternative measures of the expected change in cash and then use the realized change in cash net of the estimated expected change in cash.

¹⁰ As a robustness check, we also use stock returns in excess of the risk-free rate as our dependent variable and include the Fama and French three factors in the regression. Our results are robust to such a specification.

¹¹ In unreported regressions, we standardized by the lagged book value of assets rather than the lagged market value of equity and find strikingly different results, consistent with our criticism of standardizing by book values. Results are available upon request.

firm-specific factors. We then test our hypotheses by including interaction terms and by examining differences in coefficients across subsamples. Throughout the analysis, our focus is on the value of the unexpected change in cash, captured by its coefficient and the coefficients corresponding to interactions with other financial variables.

To arrive at our estimate of the excess return, we use the 25 Fama and French portfolios formed on size and book-to-market as our benchmark portfolios. A portfolio return is a value-weighted return based on market capitalization within each of the 25 portfolios. For each year, we group every firm into one of 25 size and BE/ME portfolios based on the intersection between the size and book-to-market independent sorts. Fama and French (1993) conclude that size and the book-to-market of equity proxy for sensitivity to common risk factors in stock returns, which implies that stocks in different size and book-to-market portfolios may have different expected returns. Therefore, stock i 's benchmark return at year t is the return of the portfolio to which stock i belongs at the beginning of fiscal year t . To form a size- and BE/ME-excess return for any stock, we simply subtract the return of the portfolio to which it belongs from the realized return of the stock.¹²

Our baseline regression model is:

$$\begin{aligned} \underline{r_{i,t} - R_{i,t}^B} = & \gamma_0 + \gamma_1 \frac{\Delta C_{i,t}}{M_{i,t-1}} + \gamma_2 \frac{\Delta E_{i,t}}{M_{i,t-1}} + \gamma_3 \frac{\Delta NA_{i,t}}{M_{i,t-1}} + \gamma_4 \frac{\Delta RD_{i,t}}{M_{i,t-1}} \\ & + \gamma_5 \frac{\Delta I_{i,t}}{M_{i,t-1}} + \gamma_6 \frac{\Delta D_{i,t}}{M_{i,t-1}} + \gamma_7 \frac{C_{i,t-1}}{M_{i,t-1}} + \gamma_8 L_{i,t} + \gamma_9 \frac{NF_{i,t}}{M_{i,t-1}} \\ & + \gamma_{10} \frac{C_{i,t-1}}{M_{i,t-1}} * \frac{\Delta C_{i,t}}{M_{i,t-1}} + \gamma_{11} \underline{L_{i,t}} * \frac{\Delta C_{i,t}}{M_{i,t-1}} + \epsilon_{i,t}, \end{aligned} \quad (9)$$

where the term ΔX indicates unexpected changes in the variable X . As previously stated, we initially use the realized change, assuming that the expected change is zero, and then conduct a number of robustness tests with varying estimates of the unexpected change in cash.

The dependent variable in our regression is the excess stock return, $r_{i,t} - R_{i,t}^B$, where $r_{i,t}$ is the stock return for firm i during fiscal year t and $R_{i,t}^B$ is stock i 's benchmark return at year t . The independent variables are firm-specific factors that control for sources of value other than cash that may be correlated with cash holdings. The financing variables that we are interested in include the cash holdings of firm i at time t ($C_{i,t}$), interest expense ($I_{i,t}$), total dividends ($D_{i,t}$), market leverage at the end of fiscal year t ($L_{i,t}$), and the firm's

¹² While the Fama and French 25 portfolios are formed at the end of each June, the fiscal year-end of a firm could be any month during the year. Therefore, a firm could change the portfolio to which it belongs during the year. Consider a firm whose fiscal year ends in December in year $t - 1$. From January to June of year t , it belongs to the portfolio according to the size and BE/ME breakpoints of year $t - 1$ and from July to December of year t , it belongs to the portfolio according to the size and BE/ME breakpoints of year t . Since we have value-weighted monthly returns of the portfolios, we calculate the benchmark return by annualizing the monthly returns from the portfolio it belongs to each month.

net financing during the fiscal year t ($NF_{i,t}$). We also control for changes in the firm's profitability using earnings before interest and extraordinary items ($E_{i,t}$) and changes in the firm's investment policy by controlling for total assets net of cash ($NA_{i,t}$) and R&D expenditures ($RD_{i,t}$). To avoid having the largest firms dominate the results, we deflate the firm-specific factors (except leverage) by the 1-year lagged market value of equity ($M_{i,t-1}$). Since the stock return is the spread of $(M_{i,t} - M_{i,t-1})$ divided by $M_{i,t-1}$, this standardization enables us to interpret the estimated coefficients as the dollar change in value for a one-dollar change in the corresponding independent variable.

Additionally, we add interaction terms to test the hypotheses stated in the previous section. We use $\frac{C_{i,t-1}}{M_{i,t-1}} * \frac{\Delta C_{i,t}}{M_{i,t-1}}$ in order to estimate the effect of changes in the value of cash for different levels of cash holdings. Following the first hypothesis, we expect the coefficient γ_{10} to be negative, indicating that the marginal value of cash is decreasing in the amount of cash the firm has. We also include $L_{i,t} * \frac{\Delta C_{i,t}}{M_{i,t-1}}$ in the regression to capture the effect of leverage on the marginal value of cash holdings. Based upon our second hypothesis, we expect γ_{11} to be negative, indicating that as firms have more leverage, less of the value created by the presence of extra cash accrues to shareholders. In these regressions, we also include the lagged cash position and the level of leverage to ensure that our estimated coefficients on the interaction terms are due to the interaction, and not due to the cash position or leverage individually.

The methodology we use is essentially a long-term event study. Generally, the focus of event studies is to estimate the effect of a firm event on the return of its common stock. In standard event study methodology, the net present value of the event is estimated by looking at the abnormal return experienced around the time of the event. The expected return is estimated using a performance model whose parameters are estimated outside the event window. In this paper, we focus on how the change of cash holdings affects stock returns, controlling for other relevant changes in the firm's financial status. The event in which we are interested is the unexpected change of cash holdings, and the event window is defined to be the fiscal year. Since there is not an estimation window, we instead estimate the expected return by using the benchmark returns of the 25 size and book-to-market portfolios. By subtracting the benchmark return from the stock return, we control for the expected return of the stock. The unexpected changes in the firm-specific factors should therefore explain the abnormal returns, similar to an event study.

III. Data and Summary Statistics

The data for this paper come from the 2001 COMPUSTAT tapes (numbers in parentheses are COMPUSTAT data item numbers) and 2001 CRSP tapes over the 1971 to 2001 period. We exclude all financial firms and utility firms (SIC codes between 6,000 and 6,999, and between 4,900 and 4,999, respectively). Our measure of stock returns includes distributions during the fiscal year. The breakpoints for the 25 portfolios formed on size and BE/ME and the

portfolio monthly returns are from Kenneth R. French's web page.¹³ All returns correspond to the 12-month period representing the fiscal year of the firm.

All data are converted to real values in 2001 dollars using the consumer price index (CPI). The market value of equity is defined as the number of shares (54) multiplied by the stock's closing price at the fiscal year-end (199). Cash holdings equals cash plus marketable securities (1). Net assets is total assets (6) minus cash holdings. Following Fama and French (1998) and Pinkowitz and Williamson (2004), earnings are calculated as earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits (18+15+50+51). Total dividends are measured as common dividends paid (21). Leverage is defined as the market debt ratio, calculated as total debt (9+34) over the sum of total debt and the market value of equity. Net financing is total equity issuance (108) minus repurchases (115) plus debt issuance (111) minus debt redemption (114). We also use R&D expenditures (46), which equals zero if missing, and interest expense (15).

□ We trim our firm-specific factors and dependent variable at the 1% tails measured using the full sample, to reduce the impact of outliers. Since we require 1 year of changes for some variables, our usable sample starts in 1972. We eliminate firm-years for which net assets are negative, the market value of equity is negative, or dividends are negative. Our final sample consists of 82,187 firm-years. Summary statistics for the sample can be found in Table I.

Recall that all independent variables, excluding leverage (L_t), are deflated by the lagged market value of equity, thereby allowing us to interpret our results as the dollar increase in value associated with a one-dollar change in the explanatory variable. We see that the median firm has a -8.45% 1-year excess (abnormal) stock return while the mean is slightly negative at -0.50% , consistent with the distribution of abnormal stock returns being right-skewed.¹⁴ The mean and median changes in cash holdings are close to zero, suggesting that the distribution of the change in cash holdings is relatively symmetric. However, the median cash holdings level is equivalent to 9.45% of market equity value at the beginning of the fiscal year, while the mean is much higher at 17.26% , suggesting that cash holdings are right-skewed. These two numbers are slightly higher than the cash ratios in Opler et al. (1999). Our statistics are not directly comparable to summary statistics in most other cash papers in the literature, however, because most papers use net assets or book assets to scale independent variables, whereas we use the lagged market value of equity, consistent with both the discussion of our hypotheses and the normalization of our variables. Note that the median leverage ratio of 22.65% and mean of 27.78% are consistent with Opler et al. (1999).

Table I also shows that on average, profitability has been increasing over time as the changes in earnings are positive both at the mean and the median,

¹³ See <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french>. We thank him for graciously providing the data.

¹⁴ Recall that the observations are trimmed at the 1% tails, which explains the non-zero mean.

Table I
Summary Statistics for the 1972–2001 Sample

This table provides summary statistics for the variables in our sample of firm-years from U.S.-based publicly traded firms over the period 1972 to 2001. $r_{i,t} - R_{i,t}^B$ is the excess stock return, where $r_{i,t}$ is the annual stock return of firm i at time t (fiscal year-end) and $R_{i,t}^B$ is stock i 's benchmark portfolio return at time t . All variables except L_t and excess stock return are deflated by the lagged market value of equity (M_{t-1}). C_t is cash plus marketable securities, E_t is earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits, and NA_t is total assets minus cash holdings. I_t is interest expense, total dividends (D_t) are measured as common dividend paid, L_t is market leverage, and NF_t is the total equity issuance minus repurchases plus debt issuance minus debt redemption. ΔX_t is compact notation for the 1-year change, $X_t - X_{t-1}$. The subscript $t - 1$ means the value of the variable is at the beginning of fiscal year t or at the end of fiscal year $t - 1$.

Variable	Mean	1 st Quartile	Median	3 rd Quartile	SD
$r_{i,t} - R_{i,t}$	-0.0050	-0.3403	-0.0845	0.2014	0.5592
ΔC_t	0.0036	-0.0382	-0.0005	0.0348	0.1514
C_{t-1}	0.1726	0.0346	0.0945	0.2155	0.2248
ΔE_t	0.0105	-0.0382	0.0063	0.0461	0.2137
ΔNA_t	0.0190	-0.0871	0.0292	0.1599	0.5464
ΔRD_t	0.0009	0.0000	0.0000	0.0009	0.0196
ΔI_t	0.0008	-0.0040	0.0000	0.0070	0.0349
ΔD_t	-0.0003	0.0000	0.0000	0.0004	0.0100
L_t	0.2778	0.0616	0.2265	0.4445	0.2416
NF_t	0.0518	-0.0291	0.0015	0.0866	0.2604

consistent with findings in Pinkowitz and Williamson (2004). Firms' research and development expenditures have also increased on average over time. In contrast, interest expense and dividend payments appear to be quite stable.

In order to test our third hypothesis, we must analyze separately those firms that face greater financing constraints than others. There is a great deal of debate in the literature on how to measure financial constraints. Following Almeida et al. (2004), we use four alternative schemes to partition our sample.¹⁵

1. *Payout ratio*: The payout ratio is measured as total dividends (total common dividends plus repurchases) over earnings. For each year from 1972 to 2001, we sort firms according to their annual payout ratios and assign to the financially constrained (unconstrained) group those firms whose payout ratios are less (greater) than or equal to the payout ratio of the firm at the 30th (70th) percentile of the annual payout ratio distribution.¹⁶ Firms with high payout ratios are more likely to have ample internal funds to cover their debt obligations and to finance their investments, and should

¹⁵ Almeida et al. (2004) actually use five alternative schemes. Since they do not find that the Kaplan–Zingales (1997) index is effective, we do not use it.

¹⁶ In this way we make sure that all firms with the same payout ratio are in the same group, which generates an unequal number of observations being assigned to each of our groups.

therefore receive lower benefits from cash holdings than firms with low payout ratios. Additionally, Fazzari et al. (1988) document that financially constrained firms have significantly lower payout ratios.

2. *Firm size:* Larger firms are thought to be better known and have better access to capital markets than smaller firms, and should therefore face fewer constraints when raising capital to fund its investments. We use sales (12) as our measure of firm size.¹⁷ For each year from 1972 to 2001, we rank all firms by their sales at the end of the previous fiscal year and assign to the financially constrained (unconstrained) group those firms whose sales are less (greater) than or equal to the sales in the bottom (top) three deciles of the annual size distribution.

3. *Long-term bond rating:* Firms that have access to public debt markets are able to raise funds from a source of capital that those without a rating may not be able to access. The former firms are usually better known, and should face less difficulty in raising funds for their investment opportunities. COMPUSTAT provides data on firms' bond ratings starting in 1985. We assign to the financially unconstrained group those firm-years in which the firm has a bond rating when it reports positive debt and to the constrained group those firm-years in which the firm does not have a bond rating but reports positive amounts of debt.¹⁸ Faulkender and Petersen (2006) find that firms with a public debt rating (either a long-term bond rating or commercial paper rating) have significantly higher leverage ratios than firms without a debt rating, and the difference cannot be explained by firm characteristics previously found to determine observed capital structure. This finding is consistent with rated firms having better access to debt capital. They should therefore be not as reliant on internal funds as those firms without a debt rating, reducing their marginal value of cash.

4. *Commercial paper rating:* Firms with a commercial paper rating are an even more exclusive set and are considered among the safest group of publicly traded firms. We use the same categorization approach as above except that we look at the commercial paper rating instead of the long-term bond rating. The percentage of firm-years classified as having a commercial paper rating is 9.0% relative to 21.7% of firm-years classified as having a public bond rating.

IV. Empirical Results

This section contains results of regressions that test our empirical predictions. We begin in Section IV.A by testing the first two hypotheses, looking at the full sample over the entire period. We then demonstrate in Section IV.B the robustness of these results using three alternative measures of the unexpected

¹⁷ The results are robust to the use of total assets instead of sales.

¹⁸ Whited (1992), Kashyap, Lamont, and Stein (1994), and Gilchrist and Himmelberg (1995) similarly categorize constrained and unconstrained firms.

change in cash over the fiscal year. In Section IV.C, we examine the effect of capital market accessibility (our third hypothesis) by using our various measures of financial constraints to subdivide the sample, testing the differences in the marginal value of cash across the subsamples. In additional robustness checks in Section IV.D, we examine three subsets of firms that are most likely to fall into the three cash regimes discussed above, based upon their cash position, cash generation, and investment opportunities. We also revisit the effects of capital market accessibility by examining the subset of firms that are most likely to want to raise capital, and we look for differences in the marginal value of cash based upon our four measures of financial constraints.

A. Findings for Cash Level and Leverage

One of our primary objectives is to measure the marginal value of cash for the average firm. The results obtained from the estimation of our regression model (equation (1)) are presented in Table II. The initial coefficient estimate corresponding to the change in cash holdings suggests that an extra dollar of cash is only valued by shareholders at \$0.75. Our results change dramatically, though, when we allow the change in cash to interact with the level of cash ($C_{t-1} * \Delta C_t$) and with leverage ($L_t * \Delta C_t$), as seen in column 2 of Table II. These results indicate that the marginal value of cash is sensitive to both the amount of cash the firm already has on hand and to the percentage of the firm's capital structure that consists of debt. Recall that these are the variables that we added to test our first two empirical predictions. Having added these variables, the estimated marginal value of cash for a firm with zero cash and no leverage is \$1.47.

As hypothesized, as firms' cash positions improve, the value of an additional dollar of cash decreases. The estimated coefficient corresponding to the interaction of the level of cash holdings with the change in cash is negative and statistically significant at better than 1%.¹⁹ Economically, the estimate suggests that for two otherwise identical firms, a firm with cash holdings of 5% of equity has a marginal value of cash that is nearly 7.4 cents higher than a firm with cash holdings equal to 15% of its equity. In other words, for a firm with no leverage and cash holdings equal to 5% of their equity market capitalization, the value of an additional dollar of cash is \$1.43 (= \$1.466 + (-0.738 * 5%)), relative to \$1.36 for an otherwise equivalent firm with cash holdings equivalent to 15% of the value of their equity. This finding is consistent with our first hypothesis that firms with little or no cash on hand are likely to raise costly external funds and therefore would receive the highest benefits from having additional internal funds.

The results are also consistent with our second hypothesis that the marginal value of cash is decreasing in the amount of leverage. The significantly negative coefficient on $L_t * \Delta C_t$ suggests that an extra dollar of cash in an all-equity firm is worth 14.3 cents more to shareholders than an extra dollar in a firm with a

¹⁹ All reported regressions use White (1980) heteroscedastic-consistent errors, corrected for correlation across observations of a given firm.

Table II
Regression Results for the Whole Sample

This table presents the results of regressing the excess stock return $r_{i,t} - R_{i,t}^B$ on changes in firm characteristics over the fiscal year. All variables except L_t and excess stock return are deflated by the lagged market value of equity (M_{t-1}). C_t is cash plus marketable securities, E_t is earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits, and NA_t is total assets minus cash holdings. I_t is interest expense, total dividends (D_t) are measured as common dividends paid, L_t is market leverage, and NF_t is the total equity issuance minus repurchases plus debt issuance minus debt redemption. We also use $R\&D$ expenditures (RD_t), which is set to zero if missing and Re is the percentage of distributions to shareholders that occur in the form of repurchases (repurchase/(repurchase+dividend)). ΔX_t is compact notation for the 1-year change, $X_t - X_{t-1}$. The subscript $t - 1$ means the value of the variable is at the end of fiscal year $t - 1$. The third regression is only on the subset of firms with positive earnings and positive payout in the corresponding fiscal year. White heteroscedastic-consistent standard errors, corrected for correlation across observations of a given firm, are in parentheses (White (1980)).

Independent Variables	I	II	III
ΔC_t	0.751*** (0.020)	1.466*** (0.038)	1.030*** (0.054)
ΔE_t	0.529*** (0.013)	0.524*** (0.013)	0.806*** (0.027)
ΔNA_t	0.151*** (0.007)	0.161*** (0.007)	0.155*** (0.010)
ΔRD_t	1.350*** (0.139)	1.302*** (0.138)	1.423*** (0.234)
ΔI_t	-1.516*** (0.085)	-1.448*** (0.084)	-1.922*** (0.116)
ΔD_t	2.534*** (0.188)	2.504*** (0.186)	2.851*** (0.229)
C_{t-1}	0.337*** (0.012)	0.263*** (0.013)	0.212*** (0.015)
L_t	-0.475*** (0.009)	-0.477*** (0.009)	-0.417*** (0.011)
NF_t	0.087*** (0.013)	0.059*** (0.012)	0.017 (0.017)
$C_{t-1} * \Delta C_t$		-0.738*** (0.055)	-0.433*** (0.069)
$L_t * \Delta C_t$		-1.433*** (0.074)	-1.086*** (0.104)
Re_t			-0.014** (0.006)
$Re_t * \Delta C_t$			0.130* (0.070)
Intercept	0.057*** (0.003)	0.061*** (0.003)	0.082*** (0.004)
Observations	82,187	82,187	46,444
Adj R^2	0.19	0.20	0.20

*Corresponds to significant at 10%; ** significant at 5%; and *** significant at 1%.

10% leverage ratio. This finding is consistent with debt holders receiving some of the benefit associated with an increase in the amount of cash the firm holds. As cash increases, the likelihood of the firm defaulting on the debt decreases, meaning that some of the value associated with the firm having additional cash accrues to debt holders. The value of cash to shareholders is higher when the firm has very little debt since the change in the likelihood of default is lower than when there is a large amount of leverage.

Having included the effects of both leverage and the level of cash holdings and discussed the related findings, we can now estimate the marginal value of cash for the mean firm in our sample. Since most firms have some cash and some leverage, the marginal value of cash estimate is a function of the estimated coefficient on the change in cash and the interactions with the level of cash holdings and with leverage. So, using the estimates from the second column, an extra dollar of cash holdings increases shareholder wealth by \$1.466 if the firm has zero cash and no leverage at the beginning of the fiscal year. However, the mean firm has cash holdings equivalent to 17.26% of the market capitalization of equity at the beginning of the fiscal year, and the mean leverage ratio is 27.78%. Therefore, the marginal value of cash to shareholders in the mean firm is $\$0.94 (= \$1.466 + (-\$0.738 * 0.1726) + (-\$1.433 * 0.2778))$. This finding suggests that less than the full value of the extra dollar of cash is incorporated into stock prices, consistent with shareholders valuing cash held by the firm at its after-shareholder-tax value.²⁰

We also examine how the value of cash differs depending upon how cash is distributed to shareholders. As discussed above, for those firms that pay out cash to shareholders, we expect the value of an additional dollar of cash to be valued at less than a dollar since shareholders will have to pay taxes on that dollar when it is distributed. However, the tax rate applied to that dollar depends partially upon how it is paid. Throughout our sample period, the tax rate on dividends was higher than the tax rate on capital gains, the rate that would normally apply in the case of a share repurchase. Therefore, we expect that the marginal value of cash would be higher over our sample period for those firms that predominately return cash to shareholders in the form of repurchases rather than dividends.²¹

The results are consistent with our conjecture. In the third column of Table II, we examine just those firm-years with positive earnings in which cash is distributed to equity holders. When we include in our specification a

²⁰ We add year dummies to our regression model to verify that the results are robust to year effects that may be correlated with changes in firm characteristics. The results are not significantly different and are available upon request.

²¹ This analysis assumes that the market believes that future distributions will follow approximately the same mix of dividends and repurchases that the firm has used in the past year. Stephens and Weisbach (1998) document that firms take up to 3 years to complete a repurchase program and numerous studies show that dividend payments are rather sticky (for a comprehensive review, see Allen and Michaely (2002)). Together, these findings suggest that there should be autocorrelation in distribution methods. We therefore argue that this is a reasonable assumption for the market to make.

variable interacting the change in cash with the percentage of distributions that occur in the form of repurchases, we find a significantly higher value on cash for firms that predominately repurchase. Statistically, the coefficient is significant at better than 10%. The economic magnitude suggests that the equity market values an additional dollar of cash for a firm that carries out 100% of its equity payments in the form of repurchases 13.0 cents higher than an otherwise equivalent firm that pays out 100% of its equity payments in the form of dividends. These results are consistent with our hypothesis that the differential tax schedules on dividends and capital gains faced by shareholders play an important role in the value they place on the cash that firms hold.

Our estimated value difference naturally leads to the question of why dividend-paying firms would forgo the additional value that would be created by altering their payout structure. This question has puzzled researchers for years, leading to extensive examinations of payout policy. The findings of this literature suggest that firms recognize the value to be gained by moving toward repurchases. For instance, Grullon and Michaely (2002) find that beginning in 1999, “industrial firms spent more money on share repurchases than on dividend payments,” (p. 1649) and that payout initiations are much more likely to be in the form of repurchases. While firms have not uniformly adopted repurchases as their only form of cash distribution to shareholders, the empirical evidence does suggest that firms are slowly moving more toward share repurchases, consistent with the additional value creation we document.

B. Alternative Measures of the Expected Change in Cash

So far, we use the entire change in cash in our econometric specifications. Since we examine changes in market values, the expected change in cash should be incorporated into the market equity value of the firm at the beginning of the fiscal year and the change in value should correspond to just the portion of the change in cash that is unexpected. Thus, the results that we present so far assume that the expected level of cash at the end of the fiscal year is equal to the value of cash at the end of the previous fiscal year. We now conduct robustness checks that use three alternative measures of the expected change in cash over the fiscal year, and we use the difference between the realized change and the expected change in our analysis.

The first measure of the expected change in cash uses the average change in cash in the benchmark portfolio during the corresponding fiscal year. If most firms in the same size and book-to-market portfolio increase their cash positions during the fiscal year, then the benchmark return should already reflect the effect of the average increase in cash, and the excess return should be the response to the change not already reflected in the benchmark return. As an example, assume that two firms both increase their cash position by 2% of the market value of the firm's equity. If most firms in the same size and book-to-market portfolio as the first firm also have a similar increase in their cash position, then the market's response to the cash increase would be incorporated into the average return of the firms in the benchmark portfolio, and the excess

return should be close to zero. On the other hand, if the second firm belongs to a size and book-to-market portfolio in which most firms decrease their cash position, then the corresponding benchmark return would include this average decrease in cash. Assuming that increases in cash increase the market value of the equity, we would expect the excess return of the second firm in this example to be higher than that for the first firm, all else equal, since the second firm increases its cash relative to the expected drop in cash whereas the first firm increases its cash position, as it was expected to do. Results using this alternative measure of the change in cash are presented in the first column of Table III.

For the other two measures of the unexpected change, we use two models from Almeida et al. (2004) to estimate the expected change of cash holdings, controlling for industry fixed effects.²² In both cases, changes in cash are regressed on factors that represent sources and uses of cash. We use the realizations of these factors at the end of the previous fiscal year to estimate the change that is expected to occur in the current fiscal year and then subtract those estimates from the realized change to obtain the unexpected piece. Since our objective is to estimate the market's expectation of the change of cash holdings that occurs during the current fiscal year, we restrict ourselves to information the market had at the beginning of the fiscal year. The first of these specifications is

$$\Delta CashHoldings_{i,t} = \alpha_0 + \alpha_1 CashFlow_{i,t-1} + \alpha_2 Q_{i,t-1} + \alpha_3 Size_{i,t-1} + \epsilon_{i,t}, \quad (10)$$

where *Size* is measured as the natural log of book assets. The second equation from Almeida et al. (2004) adds capital expenditures, acquisitions, the change in net working capital, and the change in short-term debt as additional explanatory variables, all lagged and deflated by the lagged market value of assets. The results using these estimates of the unexpected change in cash appear in columns 2 and 3, respectively, of Table III.²³

The results in Table III are nearly identical to those using the realized change in cash found in column 2 of Table II, discussed above. The effects of leverage and cash levels are extremely similar in magnitude and have the same strong statistical significance. More debt in the firm's capital structure and higher cash levels both correspond to significantly lower marginal values of cash to the shareholders. After incorporating the leverage ratio and cash position of the average firm in the sample, we estimate the marginal value of cash to be \$0.95, \$0.93, and \$0.95, respectively. Recall that we estimate a value of \$0.94 when we use the entire change in cash rather than estimates of the unexpected change in cash. Even though these alternative methods should generate better estimates of the unexpected change in cash, they are still highly correlated with the overall

²² We deviate slightly from their model by normalizing the change in cash holdings and cash flow by the lagged market value of assets rather than current book value of assets, consistent with all of the other normalizations in this paper.

²³ The estimates are robust to estimating the expected change in cash industry-by-industry.

Table III
Regressions with Alternative Definitions of the Expected Change
in Cash Holdings

This table presents the results of regressing the excess stock return $r_{i,t} - R_{i,t}^B$ on changes in firm characteristics over the fiscal year. All variables except L_t and excess stock return are deflated by the lagged market value of equity (M_{t-1}). ΔC_t is compact notation for the realized 1-year change in cash relative to the expected change in cash for that specification (details provided in Section IV.B). C_t is cash plus marketable securities, E_t is earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits, and NA_t is total assets minus cash holdings. I_t is interest expense, total dividends (D_t) are measured as common dividends paid, L_t is market leverage, and NF_t is the total equity issuance minus repurchases plus debt issuance minus debt redemption. We also use $R\&D$ expenditures (RD_t), which is set to zero if missing. ΔX_t is compact notation for the 1-year change, $X_t - X_{t-1}$. The subscript $t - 1$ means the value of the variable is at the end of fiscal year $t - 1$. White heteroscedastic-consistent standard errors, corrected for correlation across observations of a given firm, are in parentheses (White (1980)).

Independent Variables	Portf. Ave.	ACW (1)	ACW (2)
ΔC_t	1.463*** (0.038)	1.511*** (0.038)	1.520*** (0.043)
ΔE_t	0.523*** (0.013)	0.498*** (0.013)	0.483*** (0.014)
ΔNA_t	0.161*** (0.007)	0.175*** (0.007)	0.189*** (0.007)
ΔRD_t	1.309*** (0.138)	1.408*** (0.14)	1.363*** (0.152)
ΔI_t	-1.478*** (0.085)	-1.406*** (0.085)	-1.520*** (0.097)
ΔD_t	2.589*** (0.184)	2.715*** (0.186)	2.862*** (0.223)
C_{t-1}	0.250*** (0.013)	0.238*** (0.013)	0.252*** (0.014)
L_t	-0.497*** (0.009)	-0.489*** (0.009)	-0.496*** (0.01)
NF_t	0.066*** (0.012)	0.039*** (0.012)	0.033** (0.014)
$C_{t-1} * \Delta' C_t$	-0.789*** (0.055)	-0.739*** (0.061)	-0.839*** (0.07)
$L_t * \Delta' C_t$	-1.361*** (0.074)	-1.619*** (0.075)	-1.530*** (0.085)
Intercept	0.073*** (0.003)	0.073*** (0.003)	0.080*** (0.004)
Observations	82,187	81,979	67,859
Adj R^2	0.20	0.20	0.21

*Corresponds to significant at 10%; ** significant at 5%; and *** significant at 1%.

change in cash. So, while these alternative measures likely have less noise, results using the absolute change in cash appear to be relatively unbiased. The stability of the estimated coefficients demonstrates the robustness of our findings regarding the value of additional corporate cash and how that value

is influenced by the level of cash and the portion of the firm's capital structure that consists of debt.

C. Financial Constraints Results

Moving to the empirical implication that shareholders place a higher marginal value on the cash of constrained firms than on the cash of unconstrained firms, we split the sample using the four criteria outlined above. Table IV presents summary statistics for the constrained and unconstrained groups under the four different financial constraints criteria. The letter (C) stands for constrained groups and (U) for unconstrained groups. The first row for each variable reports the mean value for the corresponding variable, with medians in brackets. There is a positive but imperfect association among the groups generated by the four criteria. Under all four criteria, the median change in cash holdings is negative for constrained firms, whereas the median change is zero or positive for unconstrained firms under three of the four criteria. This suggests that firms that have greater difficulty accessing capital (constrained firms) are more likely to draw down their cash holdings relative to unconstrained firms. Additionally, constrained firms have higher cash holdings than unconstrained firms under all four criteria, which is consistent with the findings in Almeida et al. (2004). The intuition is that constrained firms are more reliant on internal funds and therefore hold higher levels of cash than do firms that can easily access more cash when they need it. It is worth noting that with the exception of the use of firm size as our measure of a constrained firm, the number of financially constrained firms is much higher than the number of unconstrained firms.

Using all four criteria, the results from splitting the firm-years into constrained and unconstrained subsamples are strongly consistent with our hypothesis. We find that the marginal value of cash is significantly higher for constrained firms than for unconstrained firms. As displayed in Table VI, the estimated marginal value of cash for constrained firms, controlling for the interaction with the level of cash and with leverage, is significantly higher than the estimate for unconstrained firms, both statistically and economically.²⁴ The difference between the coefficients for the two different subsamples is significant at better than 5% under all four criteria. Firms that can easily raise funds when they need cash should not carry a lot of cash, and the market does not place a high value on such cash because of the costs associated with holding cash (such as tax effects and agency costs). However, the market places a rather high value on liquidity for those firms that may face problems raising external capital when they need to raise additional cash.

The coefficients estimated on the interaction terms with cash holdings and leverage retain the statistical significance found in Table II in almost all of

²⁴ We only show the results in which the dependent variable is left-censored at the 1% tail and right-censored at the 99% tail. The main results do not change if the dependent variable is left-censored at 5% and right-censored at 95%. The results are available upon request.

Table IV
Summary Statistics for Constrained and Unconstrained Groups

This table presents summary statistics for key variables across groups of financially constrained and unconstrained firms (see text for definitions) from 1972 to 2001. The first number corresponds to the mean and the medians are in brackets. We use letter (C) for constrained firms and (U) for unconstrained firms. $r_{i,t} - R_{i,t}^B$ is the excess stock return, where $r_{i,t}$ is the annual stock return of firm i at time t (fiscal year-end) and $R_{i,t}^B$ is stock i 's benchmark portfolio return at time t . All variables except L_t and excess stock return are deflated by the lagged market value of equity (M_{t-1}). C_t is cash plus marketable securities, E_t is earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits, and NA_t is total assets minus cash holdings. I_t is interest expense, total dividends (D_t) are measured as common dividends paid, L_t is market leverage, and NF_t is the total equity issuance minus repurchases plus debt issuance minus debt redemption. ΔX_t is compact notation for the 1-year change, $X_t - X_{t-1}$. The subscript $t - 1$ means the value of the variable is at the beginning of fiscal year t or at the end of fiscal year $t - 1$.

Financial Criteria	Payout Ratio		Firm Size		Bond Ratings		Comm. Paper Ratings	
	(C)	(U)	(C)	(U)	(C)	(U)	(C)	(U)
dC_t	0.0034 [-0.0015]	-0.0078 [-0.0027]	-0.0051 [-0.0050]	0.0063 [0.0004]	-0.0005 [-0.0010]	0.0085 [0.0004]	0.0015 [-0.0007]	0.0015 [0.0000]
C_{t-1}	0.1884 [0.0993]	0.1686 [0.0958]	0.2152 [0.1203]	0.1344 [0.0718]	0.1559 [0.0791]	0.1222 [0.0519]	0.1569 [0.0789]	0.0626 [0.0334]
dE_t	0.0249 [0.0092]	-0.0315 [-0.0057]	0.0196 [0.0049]	0.0032 [0.0057]	0.0095 [0.0044]	0.0036 [0.0052]	0.0091 [0.0050]	-0.0008 [0.0031]
dNA_t	-0.0116 [0.0271]	-0.0193 [0.0076]	0.0278 [0.0178]	0.0203 [0.0301]	0.0370 [0.0321]	0.0776 [0.0379]	0.0456 [0.0351]	0.0492 [0.0243]
dRD_t	0.0007 [0.0000]	0.0007 [0.0000]	0.0014 [0.0000]	0.0005 [0.0000]	0.0009 [0.0000]	0.0005 [0.0000]	0.0009 [0.0000]	0.0004 [0.0000]
dI_t	-0.0005 [0.0000]	0.0007 [0.0000]	0.0012 [0.0000]	0.0006 [0.0002]	0.0004 [0.0001]	0.0036 [0.0007]	0.0011 [0.0002]	0.0009 [0.0001]
dD_t	-0.0013 [0.0000]	-0.0005 [0.0000]	-0.0001 [0.0000]	-0.0003 [0.0000]	-0.0007 [0.0000]	-0.0005 [0.0000]	-0.0007 [0.0000]	0.0001 [0.0004]
L_t	0.2873 [0.2176]	0.2369 [0.1874]	0.2149 [0.1268]	0.3079 [0.2694]	0.2537 [0.1913]	0.3466 [0.3046]	0.2771 [0.2210]	0.2447 [0.2177]
NF_t	0.0744 [0.0073]	0.0124 [-0.0046]	0.0762 [0.0028]	0.0321 [0.0012]	0.0549 [0.0019]	0.0619 [0.0043]	0.0604 [0.0028]	0.0154 [-0.0023]
Observations	32,822	24,436	25,152	24,022	34,691	9,608	40,300	3,999

✓

Table V
Regressions for Constrained and Unconstrained Groups

This table presents regression results across groups of financially constrained and unconstrained firms (see text for definitions) from 1972 to 2001. We use letter (C) for constrained firms and (U) for unconstrained firms. The dependent variable in all regressions is $r_{i,t} - R_{i,t}^B$, the excess stock return of firm i during fiscal year t and $R_{i,t}^B$ is stock i 's benchmark portfolio return during fiscal year t . All variables except L_t and excess stock return are deflated by the lagged market value of equity (M_{t-1}). C_t is cash plus marketable securities, E_t is earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits, and NA_t is total assets minus cash holdings. I_t is interest expense, total dividends (D_t) are measured as common dividends paid, L_t is market leverage, and NF_t is the total equity issuance minus repurchases plus debt issuance minus debt redemption. We also use $R\&D$ expenditures (RD_t), which is set to zero if missing. ΔX_t is compact notation for the 1-year change, $X_t - X_{t-1}$. The subscript $t - 1$ means the value of the variable is at the end of fiscal year $t - 1$. White heteroscedastic consistent standard errors, corrected for correlation across observations of a given firm, are in parentheses (White, (1980)).

Independent Variables	Payout Ratio		Firm Size		Bond Ratings		Comm. Paper Ratings	
	(C)	(U)	(C)	(U)	(C)	(U)	(C)	(U)
ΔC_t	1.674*** (0.054)	1.066*** (0.072)	1.621*** (0.059)	1.123*** (0.088)	1.706*** (0.060)	1.339*** (0.161)	1.685*** (0.056)	1.707*** (0.217)
p -value ($C - U \neq 0$)	0.00		0.00		0.03		0.00	
ΔE_t	0.480*** (0.016)	0.459*** (0.029)	0.490*** (0.023)	0.505*** (0.028)	0.497*** (0.018)	0.537*** (0.043)	0.500*** (0.017)	0.522*** (0.075)
ΔNA_t	0.160*** (0.009)	0.123*** (0.014)	0.180*** (0.013)	0.118*** (0.011)	0.190*** (0.011)	0.095*** (0.018)	0.177*** (0.01)	0.122*** (0.026)
ΔRD_t	1.399*** (0.171)	0.474 (0.314)	1.655*** (0.202)	-0.105 (0.287)	0.933*** (0.189)	0.301 (0.538)	0.946*** (0.182)	-0.802 (0.771)
ΔI_t	-1.249*** (0.118)	-1.084*** (0.182)	-1.010*** (0.175)	-1.710*** (0.147)	-1.640*** (0.152)	-1.525*** (0.251)	-1.604*** (0.134)	-3.220*** (0.58)
ΔD_t	2.115*** (0.377)	2.703*** (0.266)	3.725*** (0.405)	1.750*** (0.317)	1.661*** (0.311)	0.665 (0.599)	1.459*** (0.286)	0.569 (1.009)
C_{t-1}	0.321*** (0.021)	0.206*** (0.021)	0.272*** (0.023)	0.326*** (0.024)	0.309*** (0.023)	0.420*** (0.049)	0.327*** (0.022)	0.312*** (0.082)
p -value ($C - U \neq 0$)	0.02		0.00		0.01		0.00	
L_t	-0.555*** (0.013)	-0.317*** (0.016)	-0.547*** (0.017)	-0.446*** (0.015)	-0.657*** (0.014)	-0.638*** (0.026)	-0.642*** (0.013)	-0.405*** (0.035)
NF_t	0.083*** (0.018)	-0.027 (0.026)	0.132*** (0.024)	-0.006 (0.021)	0.064*** (0.021)	0.059* (0.033)	0.049* (0.019)	0.104 (0.065)
$C_{t-1} * \Delta C_t$	-0.904*** (0.083)	-0.600*** (0.098)	-1.014*** (0.111)	-0.219*** (0.111)	-0.860*** (0.093)	-0.145 (0.184)	-0.756*** (0.085)	0.271 (0.287)
$L_t * \Delta C_t$	-1.594*** (0.104)	-0.836*** (0.14)	-1.451*** (0.119)	-1.229*** (0.118)	-1.646*** (0.125)	-1.716*** (0.319)	-1.724*** (0.115)	-1.089* (0.581)
Intercept	0.057*** (0.006)	-0.014*** (0.005)	0.022*** (0.007)	0.080*** (0.005)	0.078*** (0.005)	0.134*** (0.009)	0.084*** (0.005)	0.067*** (0.009)
Observations	32,822	24,436	25,152	24,022	34,691	9,608	40,300	3,989
Adj R^2	0.22	0.16	0.20	0.19	0.22	0.22	0.22	0.11

*Corresponds to significant at 10%; ** significant at 5%; and *** significant at 1%.

the specifications. However, the coefficients differ from each other across the subsamples. Under all four criteria, the coefficient corresponding to the interaction of the change in cash holdings with the level of cash is significantly more negative for constrained firms than for unconstrained firms. Constrained firms have a higher marginal value of cash when their current cash position is extremely small, consistent with internal cash being most valuable for firms that are likely to want to access external capital and that would face higher transactions costs when doing so. However, if the firm has a large cash balance, it is likely to distribute cash and the fact that it is constrained will not impact the value of cash. This effect manifests itself in the form of the marginal value of cash declining faster for constrained firms as cash holdings increase.

The differences in coefficients for the variable interacting the change in cash and leverage are not as stable across subsamples, but the coefficients are statistically different from zero in all of the specifications. This finding is consistent with debt holders receiving some of the additional value derived from higher cash holdings for the firm, regardless of whether the firm is financially constrained or unconstrained.

As before, since most firms have some cash on hand and debt in their capital structure, testing our third empirical implication requires incorporating the three coefficient estimates that include the change in cash as part of the corresponding variable's measure. Using the summary statistics from Table IV, we estimate that shareholders of a mean firm that is classified as financially constrained under the payout ratio criterion place a value of \$1.04 ($= \$1.674 + (-\$0.904 * 0.188) + (-\$1.594 * 0.288)$) on an extra dollar of cash, while shareholders of the mean financially unconstrained firm only place a value of \$0.77 ($= \$1.066 + (-\$0.600 * 0.169) + (-\$0.836 * 0.2369)$) on an extra dollar of cash. These numbers support our third empirical implication in which we hypothesize that shareholders value the marginal dollar of cash for a constrained firm more highly than they do the marginal value of cash for an unconstrained firm. The estimated marginal values of cash under the other three constraint classifications are \$1.09 versus \$0.72 using size as the constrained criterion, \$1.15 versus \$0.73 using access to public debt markets, and \$1.09 versus \$0.46 using access to the commercial paper market. Under all four criteria, cash is more highly valued for constrained firms than it is for unconstrained firms. The differences in the marginal value of cash are all significant at better than 1%. Economically, the estimates range from \$0.27 to \$0.63, demonstrating how costly the market perceives difficulty in accessing capital markets to be, and the extent to which firms are rewarded with higher valuations for holding cash that helps them mitigate potential underinvestment.²⁵

²⁵ While this examination is motivated theoretically by transaction costs, we are not suggesting that the differences in our estimates for constrained and unconstrained firms arise solely from direct transaction costs. Certainly, a portion of the magnitude may arise from the value effects of differences in the information the market possesses across our constraint classifications, consistent with the financial intermediation literature, as well as from potential differences in moral hazard across the subsets.

Looking at the differences in the other coefficients for the constrained versus unconstrained firms, another interesting result emerges. The change in net assets has a higher coefficient for the constrained firms relative to the unconstrained firms, a difference that is statistically significant at better than 5% under all four specifications.²⁶ While not the focus of our analysis, this finding suggests that the market responds more positively to new investments made by constrained firms. Together, these findings are consistent with the market responding more favorably when constrained firms are able to fund investment (represented by the net assets result) and when they are able to generate cash to fund future investments (represented by the findings for cash).

As with the results that examine the first two hypotheses, we now seek to verify that our results regarding the effects of financial constraints are robust to other estimates of the unexpected change in cash holdings during the fiscal year. We therefore once again divide the firms based upon our four constraints criteria and use the change in cash net of the average change in cash in the benchmark portfolio that year in our regression specification.²⁷ The results of these tests are located in Table VI.

When we replicate the examination of the marginal value of cash for constrained versus unconstrained firms using this alternative measure of the unanticipated change in cash, the results once again suggest that the marginal value of cash is significantly higher for constrained firms relative to unconstrained firms. The coefficients of both the unexpected change in cash variable and the interaction term corresponding to the product of the unexpected change in cash and the level of cash are statistically different from each other under all four constraints categorizations. Evaluated at the means for the two subsamples, the marginal value of cash is \$1.06 for the constrained group versus \$0.77 for the unconstrained group using the payout ratio as our classification criterion. Under the other three criteria, the values are \$1.10 versus \$0.73, \$1.19 versus \$0.77, and \$1.13 versus \$0.39, respectively. Once again, these differences are significantly different from each other at better than 1% under all four classification criteria. These results further confirm our third hypothesis that the market places a significantly higher value on an additional dollar of cash for those firms that are likely to face difficulty in accessing external capital markets.

D. Subsample Tests

To verify the robustness of our cross-sectional results, we focus on three subsamples of firms that are most likely to correspond to our three cash regimes and that should therefore differ in the value shareholders place on additional

²⁶ The statistical significance of this difference is also found in the robustness checks we run using alternative measures of the expected change in cash (Table VI).

²⁷ We also reestimate the results using the two estimates of the expected change in cash holdings following Almeida et al. (2004); the results are economically and statistically similar. The results are available upon request.

Table VI
Robustness Checks for Constrained and Unconstrained Firms

This table presents regression results across groups of financially constrained and unconstrained firms (see text for definitions) from 1972 to 2001. We use letter (C) for constrained firms and (U) for unconstrained firms. The dependent variable in all regressions is $r_{i,t} - R_{i,t}^B$, the excess stock return, where $r_{i,t}$ is the annual stock return of firm i during fiscal year t and $R_{i,t}^B$ is stock i 's benchmark portfolio return during fiscal year t . All variables except L_t and excess stock return are deflated by the lagged market value of equity (M_{t-1}). ΔC_t is compact notation for the 1-year change in cash relative to the average 1-year change in cash for the average firm in the corresponding benchmark portfolio. C_t is cash plus marketable securities, E_t is earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits, and NA_t is total assets minus cash holdings. I_t is interest expense, total dividends (D_t) are measured as common dividends paid, L_t is market leverage, and NP_t is the total equity issuance minus debt repurchases plus debt issuance minus debt redemption. We also use RD_t expenditures (RD_t), which is set to zero if missing. ΔX_t is compact notation for the 1-year change, $X_t - X_{t-1}$. The subscript $t - 1$ means the value of the variable is at the end of fiscal year $t - 1$. White heteroscedastic-consistent standard errors, corrected for correlation across observations of a given firm, are in parentheses (White (1980)).

Independent Variables	Payout Ratio		Firm Size		Bond Ratings		Comm. Paper Ratings	
	(C)	(U)	(C)	(U)	(C)	(U)	(C)	(U)
ΔC_t	1.677*** (0.054)	1.064*** (0.072)	1.627*** (0.059)	1.110*** (0.088)	1.715*** (0.059)	1.338*** (0.16)	1.698*** (0.056)	0.638*** (0.216)
p -value ($C - U \neq 0$)	0.00		0.00		0.02		0.00	
ΔE_t	0.476*** (0.016)	0.458*** (0.029)	0.482*** (0.023)	0.509*** (0.028)	0.494*** (0.018)	0.531*** (0.043)	0.496*** (0.017)	0.528*** (0.068)
ΔNA_t	0.161*** (0.009)	0.122*** (0.014)	0.181*** (0.013)	0.119*** (0.011)	0.192*** (0.011)	0.095*** (0.018)	0.179*** (0.01)	0.119*** (0.028)
ΔRD_t	1.396*** (0.171)	0.525* (0.313)	1.652*** (0.203)	-0.139 (0.284)	0.941*** (0.189)	0.277 (0.534)	0.950*** (0.182)	-0.77 (0.786)
ΔI_t	-1.291*** (0.119)	-1.105*** (0.185)	-1.030*** (0.177)	-1.771*** (0.148)	-1.649*** (0.153)	-1.463*** (0.252)	-1.598*** (0.135)	-3.190*** (0.599)
ΔD_t	2.171*** (0.373)	2.714*** (0.266)	3.789*** (0.41)	1.852*** (0.309)	1.672*** (0.314)	0.838 (0.616)	1.522*** (0.289)	0.359 (1.002)
C_{t-1}	0.299*** (0.021)	0.191*** (0.021)	0.247*** (0.022)	0.328*** (0.025)	0.292*** (0.023)	0.411*** (0.05)	0.308*** (0.022)	0.317*** (0.082)
L_t	-0.582*** (0.014)	-0.331*** (0.016)	-0.572*** (0.017)	-0.460*** (0.015)	-0.688*** (0.014)	-0.661*** (0.026)	-0.671*** (0.013)	-0.432*** (0.034)
NP_t	0.088*** (0.018)	-0.019 (0.026)	0.140*** (0.024)	-0.004 (0.021)	0.071*** (0.021)	0.061 (0.034)	0.054*** (0.019)	0.113* (0.064)
$C_{t-1} * \Delta C_t$	-0.370*** (0.083)	-0.658*** (0.098)	-1.090*** (0.093)	-0.217*** (0.109)	-0.961*** (0.093)	-0.281 (0.18)	-0.868*** (0.085)	0.254 (0.285)
p -value ($C - U \neq 0$)	0.01		0.00		0.00		0.00	
$L_t * \Delta C_t$	-1.517*** (0.105)	-0.766*** (0.139)	-1.365*** (0.12)	-1.155*** (0.18)	-1.481*** (0.124)	-1.547*** (0.317)	-1.572*** (0.113)	-1.083* (0.583)
Intercept	0.076*** (0.006)	-0.007 (0.005)	0.042*** (0.007)	0.086*** (0.005)	0.092*** (0.005)	0.144*** (0.009)	0.098*** (0.005)	0.073*** (0.009)
Observations	32,509	23,182	24,702	23,505	34,367	9,314	39,882	3,799
Adj R^2	0.22	0.16	0.2	0.19	0.23	0.22	0.22	0.11

*Corresponds to significant at 10%; ** significant at 5%; and *** significant at 1%.

cash in the firm. Specifically, we divide the firms into quartiles based first upon a measure of interest coverage and separately upon their average industry market-to-book ratio, defined by their two-digit SIC code.²⁸ We define interest coverage as the sum of the beginning cash position of the firm and its earnings in that fiscal year divided by the interest expense over the fiscal year. A high interest coverage firm has less of its cash and cash flow obligated to debt and therefore has relatively more funds available for investment or distribution. We interpret the industry market-to-book ratio as a measure of the firm's investment opportunities, so the highest quarter of market-to-book firm-years should have a higher value placed on an extra dollar of internal cash. We use an industry-level value because we want a measure that has not been affected by the firm's financing constraints and the corresponding likelihood that the firm will be able to capitalize on these investment opportunities.²⁹ Such effects would likely be incorporated by the market into the firm's value.

First, we focus on the firms that have the lowest coverage and lowest industry market-to-book ratio. These are firms that have a relatively high portion of their cash and cash flow obligated to interest payments and that do not have good investment opportunities. We would therefore expect these firms to have low marginal cash values since they have relatively few investment opportunities and a significant part of any additional cash the firm attains is likely to go to the debt holders. The second group of firm-years has both low coverage and high market-to-book ratios. Such firms are also low on cash holdings but are likely to have investment opportunities, making them likely to have to access capital markets in order to take advantage of their available investments. Internal cash is therefore expected to be highly valued by these firms because it reduces the amount of costly external finance the firm would need to raise, thereby making it more likely that the investments are made. Finally, we look at firms that are in the top quarter of coverage ratios and the lowest quarter of market-to-book ratios. Such firms are expected to have low marginal values of cash since they hold larger balances, generate a great deal of cash, and do not have numerous investment opportunities, that is, they are likely to be distributing cash. In addition to tax effects, these cash cow firms are also likely to suffer from agency costs described by Jensen (1986).

When we individually examine these three subsets of firm-years, the results of which are reported in Table VII, we find further empirical support for our hypotheses.³⁰ When we evaluate the first subset of firm-years for the low coverage and low market-to-book ratio firm-years, an additional dollar of cash is

²⁸ We also estimate these values using different cutoffs, such as thirds (9 groups) instead of quarters (16 groups). As expected, the magnitude of the differences in the estimates of the marginal value of cash across the subsamples increases as we increase the number of groups. However, since the number of observations in each group decreases, the statistical difference between the estimates does not change a great deal.

²⁹ We also conduct robustness tests in which we use the individual firm's market-to-book ratio and the results are similar, statistically and economically.

³⁰ Note that in these regressions, we no longer interact the change in cash holdings with either the leverage ratio or the cash level, similar to the results presented in the first column of

Table VII
Results for Three Different Coverage and M/B Groups

This table presents the results of regressing the excess stock return $r_{i,t} - R_{i,t}^B$ on changes in firm characteristics over the fiscal year. All variables except L_t and excess stock return are deflated by the lagged market value of equity (M_{t-1}). C_t is cash plus marketable securities, E_t is earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits, and NA_t is total assets minus cash holdings. I_t is interest expense, total dividends (D_t) are measured as common dividends paid, L_t is market leverage, and NF_t is the total equity issuance minus repurchases plus debt issuance minus debt redemption. We also use R&D expenditures (RD_t), which is set to zero if missing. ΔX_t is compact notation for the 1-year change, $X_t - X_{t-1}$. The subscript $t - 1$ means the value of the variable is at the end of fiscal year $t - 1$. Regression I is on the subset of firms in the bottom quarter of interest coverage and the bottom quarter of the industry market-to-book ratio. Interest coverage is defined to be (cash+earnings)/interest expense. Regression II is on firms in the bottom quarter of interest coverage and the top quarter of the industry market-to-book ratio. Regression III is on firms in the top quarter of interest coverage and the bottom quarter of the industry market-to-book ratio. White heteroscedastic-consistent standard errors, corrected for correlation across observations of a given firm, are in parentheses (White (1980)).

Independent Variables	I	II	III
ΔC_t	0.448*** (0.06)	1.159*** (0.116)	0.528*** (0.07)
ΔE_t	0.275*** (0.023)	0.320*** (0.046)	0.551*** (0.085)
ΔNA_t	0.091*** (0.012)	0.090*** (0.03)	0.250*** (0.047)
ΔRD_t	0.43 (0.482)	0.929** (0.431)	0.225 (1.056)
ΔI_t	-0.756*** (0.146)	-0.383 (0.372)	-1.624** (0.639)
ΔD_t	0.997** (0.412)	2.309*** (0.893)	5.436*** (0.674)
C_{t-1}	0.431*** (0.048)	0.831*** (0.094)	0.104*** (0.031)
L_t	-0.439*** (0.036)	-0.614*** (0.047)	-0.439*** (0.075)
NF_t	0.022 (0.025)	0.169*** (0.051)	-0.029 (0.099)
Intercept	-0.024 (0.02)	0.020 (0.022)	-0.029*** (0.010)
Observations	6,000	3,234	3,981
Adj R^2	0.15	0.19	0.14

*Corresponds to significant at 10%; ** significant at 5%; and *** significant at 1%.

only valued at 45 cents. This finding is consistent with a firm having few investment opportunities and with a large portion of an additional dollar of

Table II. Since the cash level and interest expense are both used in classifying into which group the firm-year belongs, the cash level and leverage variables will have less variation within most of subsamples. Therefore, having controlled for their effects in the group categorization process, they have a much smaller impact in the subsample regressions, so we omit them.

Table VIII
Constrained and Unconstrained Firms with High M/B and Low Coverage

This table presents regression results across groups of financially constrained and unconstrained firms with low coverage ratios and high industry market-to-book ratios (see text for definitions) from 1972 to 2001. We use letter (C) for constrained firms and (U) for unconstrained firms. These are results of regressing the excess stock return $r_{i,t} - R_{i,t}^B$ on changes in firm characteristics over the fiscal year. All variables except L_t and excess stock return are deflated by the lagged market value of equity (M_{t-1}). C_t is cash plus marketable securities, E_t is earnings before extraordinary items plus interest, deferred tax credits, and investment tax credits, and NA_t is total assets minus cash holdings. L_t is interest expense, total dividends (D_t) are measured as common dividends paid, L_t is market leverage, and NF_t is the total equity issuance minus repurchases plus debt issuance minus debt redemption. We also use $R\&D$ expenditures (RD_t), which is set to zero if missing. ΔX_t is compact notation for the 1-year change, $X_t - X_{t-1}$. The subscript $t - 1$ means the value of the variable is at the end of fiscal year $t - 1$. White heteroscedastic-consistent standard errors, corrected for correlation across observations of a given firm, are in parentheses (White (1980)).

Independent Variables	Payout Ratio		Firm Size		Bond Ratings		Comm. Paper Ratings	
	(C)	(U)	(C)	(U)	(C)	(U)	(C)	(U)
ΔC_t	1.159*** (0.161)	0.510** (0.245)	1.166*** (0.229)	0.489*** (0.174)	1.506*** (0.174)	0.186 (0.212)	1.343*** (0.153)	1.934*** (0.722)
$p\text{-value } (C - U \neq 0)$	0.03		0.02		0.00		0.41	
ΔE_t	0.277*** (0.064)	0.355*** (0.095)	0.275*** (0.104)	0.464*** (0.070)	0.282*** (0.064)	0.637*** (0.137)	0.286*** (0.061)	0.555** (0.219)
ΔNA_t	0.011 (0.049)	0.092* (0.055)	-0.036 (0.068)	0.063 (0.039)	0.090** (0.044)	0.036 (0.063)	0.089** (0.040)	-0.231** (0.100)
ΔRD_t	0.953 (0.597)	0.349 (0.960)	1.380 (0.843)	0.466 (0.681)	0.647 (0.546)	0.714 (0.796)	0.704 (0.526)	3.021* (1.752)
ΔI_t	-0.386 (0.670)	-0.380 (0.610)	0.957 (0.882)	-1.361*** (0.440)	-1.062* (0.593)	-1.165 (0.798)	-1.196** (0.566)	0.887 (1.605)
ΔD_t	1.721 (1.647)	3.247** (1.314)	7.371*** (2.589)	1.959* (1.140)	0.960 (1.874)	2.427* (1.428)	0.939 (1.546)	5.295 (3.488)
C_{t-1}	0.830*** (0.170)	0.728*** (0.14)	0.686*** (0.245)	0.586*** (0.120)	0.987*** (0.178)	0.475** (0.224)	0.902*** (0.154)	1.064* (0.635)
L_t	-0.504*** (0.065)	-0.649*** (0.101)	-0.566*** (0.092)	-0.567*** (0.103)	-0.745*** (0.132)	-0.754*** (0.172)	-0.708*** (0.059)	-0.367 (0.231)
NF_t	0.341*** (0.079)	0.100 (0.087)	0.395*** (0.117)	0.064 (0.076)	0.233*** (0.080)	0.172 (0.122)	0.269*** (0.076)	0.270* (0.142)
Intercept	-0.109*** (0.031)	0.061 (0.044)	-0.036 (0.043)	0.129*** (0.048)	-0.022 (0.029)	0.234*** (0.063)	-0.020 (0.028)	0.037 (0.076)
Observations	1,404	821	861	843	1,678	399	1,827	126
Adj R^2	0.20	0.16	0.21	0.17	0.25	0.24	0.25	0.18

*Corresponds to significant at 10%; ** significant at 5%; and *** significant at 1%.

cash being claimed by the debt holders, thereby reducing the value of additional cash to the equity holders. For the second set of firms, those with low interest coverage and many investment opportunities, we find a marginal value of cash of \$1.16. Consistent with our expectation, shareholders in firms with investment opportunities and low internal funds place a relatively higher value on the additional cash these firms attain. Finally, for the “cash cow” firms, we find a marginal value of cash of only 53 cents. Combining these results, consistent with the results for the full sample, we find that the market places a higher value on the cash of those firms that are more likely to reinvest their cash into the firm and a lower value on the cash of those firms that are likely to distribute the cash to debt or equity holders.

To further estimate the importance of capital market access, we focus on the subset of firms that we estimate to have high marginal values of cash, namely, firms with investment opportunities but low coverage. Using this group of firms, we once again estimate the differences in the marginal value of cash for those that are categorized as constrained relative to those that are unconstrained. Since constrained firms are the firms that are most likely to want to access the capital markets in the near future, differences in the market value of cash relative to capital accessibility should be the greatest for this subset of firms.

Consistent with our earlier results, there are significant differences in the market value of cash based upon how difficult it is expected to be for this subset of firms to raise additional capital. As indicated by the results in Table VIII, the marginal value of cash is higher for firms that are categorized as constrained under all four measures. Statistically, the differences are significant at better than 5% in three of the four specifications.³¹ Looking at the three criteria under which the difference is statistically significant, the economic interpretation of the difference in the marginal value of cash between constrained and unconstrained firms in this subsample ranges from 65 cents to \$1.32. These differences are much greater than the differences we estimate for the entire sample, consistent with our hypothesis that these firms are the most likely to be affected by their ability to access external capital. The results confirm that access to capital markets is an extremely important factor in the value the market places on an additional dollar of cash held by firms.

V. Conclusion

We use a revised event study methodology that examines market returns over firm fiscal years to test empirical predictions about the cross-sectional variation in the market value of cash. We find results consistent with all of our hypotheses. Specifically, we estimate that for the mean firm-year in the sample, the marginal value of cash is \$0.94. Additional cash is most highly valued by shareholders of firms with low levels of cash holdings, low leverage,

³¹ Few firm-years in the sample (126) with access to the commercial paper market also have high market-to-book ratios and low coverage. This may explain the insignificant difference that obtains when we use the presence of a commercial paper rating as our measure of financial constraint.

and constraints in accessing financial markets. The marginal value of cash for the mean constrained firm-year ranges from 28 to 63 cents higher than the mean unconstrained firm, depending upon the constraints criterion. The results are even stronger when we focus on the subset of firms that are likely to need to raise external capital in the near future.

Our results suggest that the market perceives the presence of market frictions that make raising outside capital costly. The market rewards firms that retain liquidity with higher valuations, consistent with such firms being able to create more value than an otherwise equivalent firm with less internal cash. However, the results also suggest that the value of additional cash diminishes in the level of cash, implying that there may be an upper bound on the amount of cash for which the firm is rewarded for holding. This finding is consistent with both tax effects and agency costs.

Unlike examinations that focus on the cross-sectional variation in cash holdings, we focus on the value associated with those cross-sectional differences. Our methodology enables us to estimate the value of liquidity more precisely than can be done by studies of differences in levels of cash across firms or across time. As a result, we can estimate the magnitude of the value loss associated with the market frictions we examine and the extent to which liquidity can overcome these losses.

Considering the extent to which the market-to-book ratio is used to estimate value creation, and the potential biases that we argue may be associated with it, a methodology such as ours that analyzes changes in value may have numerous other applications. As long as there is sufficient time-series variation in the underlying firm characteristics, estimating market reactions to such changes should provide more precise estimates of the value the equity holders place on such characteristics of interest.

REFERENCES

- Acharya, Viral V., Heitor Almeida, and Murillo Campello, 2004, Is cash negative debt? A hedging perspective on corporate financial policies, London Business School IFA Working Paper Series.
- Allen, Franklin, and Roni Michaely, 2002, Payout policy, in George Constantinides, Milton Harris, and Rene Stulz, eds.: *Handbook of Economics* (North-Holland).
- Almeida, Heitor, Murillo Campello, and Michael S. Weisbach, 2004, The cash flow sensitivity of cash, *Journal of Finance* 59, 1777–1804.
- Berk, Jonathan, and Richard Stanton, 2004, A rational model of the closed-end fund discount, Working paper, U.C. Berkeley.
- Billett, Matthew T., and Jon A. Garfinkle, 2004, Financial flexibility and the cost of external finance for U.S. bank holding companies, *Journal of Money, Credit and Banking* 36, 827–852.
- Black, Fisher, and Myron Scholes, 1973, The pricing of options and corporate liabilities, *Journal of Political Economy* 81, 637–654.
- Daniel, Kent, and Sheridan Titman, 1997, Evidence on the characteristics of cross-sectional variation in common stock returns, *Journal of Finance* 52, 1–34.
- DeAngelo, Harry, Linda DeAngelo, and Karen H. Wruck, 2002, Asset liquidity, debt covenants, and managerial discretion in financial distress: The collapse of L.A. Gear, *Journal of Financial Economics* 64, 3–34.

- Dittmar, Amy, Jan Mahrt-Smith, and Henri Servaes, 2003, International corporate governance and corporate cash holdings, *Journal of Financial and Quantitative Analysis* 38, 111–133.
- Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.
- Fama, Eugene F., and Kenneth R. French, 1998, Taxes, financing decisions, and firm value, *Journal of Finance* 53, 819–843.
- Faulkender, Michael, 2004, Cash holdings among small businesses, Working paper, Washington University.
- Faulkender, Michael, and Mitchell A. Petersen, 2006, Does the source of capital affect capital structure? *Review of Financial Studies* 19, 45–79.
- Fazzari, Steven M., R. Glenn Hubbard, and Bruce Petersen, 1988, Financing constraints and corporate investment, *Brooking Papers on Economic Activity* 1, 141–195.
- Gilchrist, Simon, and Charles Himmelberg, 1995, Evidence on the role of cash flow for investment, *Journal of Monetary Economics* 36, 541–572.
- Grinblatt, Mark, and Tobias J. Moskowitz, 2004, Predicting stock price movements from past returns: The role of consistency and tax-loss selling, *Journal of Financial Economics* 71, 541–579.
- Grullon, Gustavo, and Roni Michaely, 2002, Dividends, share repurchases and the substitution hypothesis, *Journal of Finance* 57, 1649–1684.
- Hanson, Robert C., 1992, Tender offers and free cash flow: An empirical analysis, *The Financial Review* 27, 185–209.
- Harford, Jarrard, 1999, Corporate cash reserves and acquisitions, *Journal of Finance* 54, 1969–1997.
- Hartzell, Jay C., Sheridan Titman, and Garry Twite, 2005, Why do firms hold so much cash? A tax-based explanation, Working paper, University of Texas at Austin.
- Hennessy, Christopher A., and Toni M. Whited, 2005, Debt dynamics, *Journal of Finance* 60, 1129–1165.
- Jensen, Michael, 1986, Agency cost of free cash flow, corporate finance and takeovers, *American Economic Review* 76, 323–329.
- Jensen, Michael, and William H. Meckling, 1976, Theory of the firm: Managerial behavior, agency costs and ownership structure, *Journal of Financial Economics* 3, 305–360.
- Kaplan, Steven, and Luigi Zingales, 1997, Do financing constraints explain why investment is correlated with cash flow? *Quarterly Journal of Economics* 112, 169–215.
- Kashyap, Anil K., Owen A. Lamont, and Jeremy C Stein, 1994, Credit conditions and the cyclical behavior of inventories, *Quarterly Journal of Economics* 109, 565–592.
- Kim, Chang-Soo, David C. Mauer, and Ann E. Sherman, 1998, The determinants of corporate liquidity: Theory and evidence, *Journal of Financial and Quantitative Analysis* 33, 335–359.
- Korajczyk, Robert A., and Amnon Levy, 2003, Capital structure choice: Macroeconomic conditions and financial constraints, *Journal of Financial Economics* 68, 75–109.
- Merton, Robert C., 1973, Theory of rational option pricing, *Bell Journal of Economics and Management Science* 4, 141–183.
- Mikkelsen, Wayne, and Megan Partch, 2003, Do persistent large cash reserves hinder performance, *Journal of Financial and Quantitative Analysis* 38, 275–294.
- Myers, Stewart, 1977, Determinants of corporate borrowing, *Journal of Financial Economics* 5, 147–175.
- Opler, Tim, Lee Pinkowitz, René Stulz, and Rohan Williamson, 1999, The determinants and implications of cash holdings, *Journal of Financial Economics* 52, 3–46.
- Ozkan, Aydin, and Neslihan Ozkan, 2002, Corporate cash holdings: An empirical investigation of UK companies, Working paper, University of York.
- Pinkowitz, Lee, Rene Stulz, and Rohan Williamson, 2006, Does the contribution of corporate cash holdings and dividends to firm value depend on governance? A cross-country analysis, *Journal of Finance*, forthcoming.
- Pinkowitz, Lee, and Rohan Williamson, 2001, Bank power and cash holdings: Evidence from Japan, *Review of Financial Studies* 14, 1059–1082.

- Pinkowitz, Lee, and Rohan Williamson, 2004, What is a dollar worth? The market value of cash holdings, Working paper, Georgetown University.
- Smith, Richard L., and Joo-Hyun Kim, 1994, The combined effects of free cash flow and financial slack on bidder and target stock returns, *Journal of Business* 67, 281–310.
- Stephens, Clifford P., and Michael S. Weisbach, 1998, Actual share reacquisitions in open-market repurchase programs, *Journal of Finance* 53, 313–333.
- White, Halbert, 1980, A heteroscedasticity-consistent covariance matrix estimator and a direct test of heteroscedasticity, *Econometrica* 48, 817–838.
- Whited, Toni, 1992, Debt, liquidity constraints, and corporate investment: Evidence from panel data, *Journal of Finance* 47, 1425–1460.



Financing Constraints and Corporate Investment

Steven M. Fazzari; R. Glenn Hubbard; Bruce C. Petersen; Alan S. Blinder; James M. Poterba

Brookings Papers on Economic Activity, Vol. 1988, No. 1 (1988), 141-206.

Stable URL:

<http://links.jstor.org/sici?sici=0007-2303%281988%291988%3A1%3C141%3AFCACI%3E2.0.CO%3B2-O>

Brookings Papers on Economic Activity is currently published by The Brookings Institution.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/brookings.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

<http://www.jstor.org/>
Wed Feb 22 21:15:10 2006

STEVEN M. FAZZARI

Washington University

R. GLENN HUBBARD

Columbia University

BRUCE C. PETERSEN

Federal Reserve Bank of Chicago

Financing Constraints and Corporate Investment

EMPIRICAL models of business investment rely generally on the assumption of a “representative firm” that responds to prices set in centralized securities markets. Indeed, if all firms have equal access to capital markets, firms’ responses to changes in the cost of capital or tax-based investment incentives differ only because of differences in investment demand. A firm’s financial structure is irrelevant to investment because external funds provide a perfect substitute for internal capital. In general, with perfect capital markets, a firm’s investment decisions are independent of its financial condition.

An alternative research agenda, however, has been based on the view that internal and external capital are not perfect substitutes. According to this view, investment may depend on financial factors, such as the availability of internal finance, access to new debt or equity finance, or the functioning of particular credit markets. For example, a firm’s internal cash flow may affect investment spending because of a “financ-

We are grateful to members of the Brookings Panel for helpful comments and suggestions and to Charles Himmelberg and Jaewoon Koo for excellent research assistance. Financial support from the Federal Reserve Bank of Chicago is acknowledged. Hubbard acknowledges support from a John M. Olin Fellowship at the National Bureau of Economic Research. The views expressed here are not necessarily those of the Federal Reserve Bank of Chicago.

ing hierarchy'' in which internal funds have a cost advantage over new debt or equity finance. Under these circumstances, firms' investment and financing decisions are interdependent.

In this article, we link conventional models of investment to the recent literature on capital market imperfections and disparities in the access of individual firms to capital markets. Conventional representative firm models in which financial structure is irrelevant to the investment decision may well apply to mature companies with well-known prospects. For other firms, however, financial factors appear to matter in the sense that external capital is not a perfect substitute for internal funds, particularly in the short run. To provide a foundation for such an "imperfection," we appeal to problems in capital markets, especially asymmetric information, that make it very costly, even impossible, for providers of external finance to evaluate the quality of firms' investment opportunities. As a result, the cost of new debt and equity may differ substantially from the opportunity cost of internal finance generated through cash flow and retained earnings.

We begin by reviewing the role of financial factors in investment studies. We then document differences in financing patterns by size of firms and consider a variety of explanations why internal and external finance are not perfect substitutes. We use manufacturing firm data to analyze differences in investment in firms classified according to their earnings retention practices. If the cost disadvantage of external finance is small, retention practices should reveal little or nothing about investment: firms will simply use external funds to smooth investment when internal finance fluctuates, regardless of their dividend policy. If the cost disadvantage is significant, firms that retain and invest most of their income may have no low-cost source of investment finance, and their investment should be driven by fluctuations in cash flow.

We present tests of this hypothesis for the q , neoclassical, and accelerator models of investment. In each case, the investment of firms that exhaust all their internal finance is more sensitive to fluctuations in cash flow than that of mature, high-dividend firms. We also find a difference across firms in the sensitivity of investment to balance sheet variables that measure liquidity. Financial effects on investment are greatest at times when capital market information problems are likely to be most severe for high-retention firms, a finding that reinforces our thesis that financing constraints in capital markets affect investment.

We test the robustness of these results to a wide variety of changes in estimation techniques and specifications.

We conclude by discussing the implications of our findings. For firms that face financing constraints, investment may be sensitive to the *average* tax burden as well as to marginal tax rates. Our results may also shed light on problems in industrial organization, such as financial motivations for conglomerate mergers. Finally, while capital market information problems arise at the level of the firm, financial constraints have a clear macroeconomic dimension because fluctuations in firms' cash flow and liquidity are correlated with movements of the aggregate economy over the business cycle. To the extent that a significant subset of firms faces financing constraints, their behavior may help explain aggregate movements of investment, and conclusions from models that maintain the representative firm assumption must be reexamined.

Finance and the Study of Investment

Early investment research, especially the work of John Meyer and Edwin Kuh, emphasized the importance of financial considerations in business investment.¹ Indeed, financial effects on many aspects of real economic activity received broad attention during the early postwar period.² Most research since the middle 1960s, however, has isolated real firm decisions from purely financial factors. Franco Modigliani and Merton Miller provided the theoretical basis for that approach by

1. John R. Meyer and Edwin Kuh, *The Investment Decision: An Empirical Study* (Harvard University Press, 1957). Other contributions associated with the "Charles River School" of investment include James S. Duesenberry, *Business Cycles and Economic Growth* (McGraw-Hill, 1958); Kuh and Meyer, "Investment, Liquidity, and Monetary Policy," in Commission on Money and Credit, *Impacts of Monetary Policy* (Prentice Hall, 1963), pp. 339-474; and Meyer and Robert R. Glauber, *Investment Decisions, Economic Forecasting, and Public Policy* (Division of Research, Graduate School of Business Administration, Harvard University, 1964).

2. The influence of financial factors in real activity is provided by the "debt deflation" school associated with Irving Fisher, Hyman Minsky, and Charles Kindleberger. See Irving Fisher, "The Debt-Deflation Theory of Great Depressions," *Econometrica*, vol. 1 (October 1933), pp. 337-57; Hyman P. Minsky, *John Maynard Keynes* (Columbia University Press, 1975); Charles Kindleberger, *Manias, Panics and Crashes* (Basic Books, 1978). For the role of firm financial capacity in the credit intermediation process, see John G. Gurley and E. S. Shaw, "Financial Aspects of Economic Development," *American Economic Review*, vol. 45 (September 1955), pp. 515-38.

demonstrating the irrelevance of financial structure and financial policy for real investment under certain conditions.³ Their key insight was that a firm's financial structure will not affect its market value in perfect capital markets. Thus, if the Modigliani-Miller assumptions are satisfied, real firm decisions, motivated by the maximization of shareholders' claims, are independent of financial factors such as internal liquidity, debt leverage, or dividend payments.

Applied to capital investment, this general finding provided a foundation for the neoclassical theory of investment developed by Dale Jorgenson and others, in which the firm's intertemporal optimization problem could be solved without reference to financial factors.⁴ Firms were assumed to face a cost of capital, set in centralized securities markets, that did not depend on the firm's particular financial structure. Since the development of the neoclassical theory, much empirical work, with both aggregate and firm-level data, has been devoted to tests of the relative success of various investment demand models, often without reference to the possible influence of financial factors.

Using data on 15 large manufacturing firms, Jorgenson and Calvin Siebert found the neoclassical model superior to internal funds theories of investment. Apart from their results, they preferred the neoclassical theory because it was consistent with the Modigliani-Miller finding that firm financial policy is irrelevant for investment. However, with a larger sample of 184 firms, J. W. Elliott reversed the Jorgenson-Siebert rankings, assigning the best ranking to the liquidity model⁵

Subsequent comparative studies of investment demand models using aggregate time series data ranked alternative specifications based on statistical prediction error or goodness of fit. As an econometric issue, it is not obvious why these criteria are appropriate for comparative

3. Franco Modigliani and Merton H. Miller, "The Cost of Capital, Corporation Finance and the Theory of Investment," *American Economic Review*, vol. 48 (June 1958), pp. 261–97; Merton H. Miller and Franco Modigliani, "Dividend Policy, Growth, and the Valuation of Shares," *Journal of Business*, vol. 34 (October 1961), pp. 411–33.

4. The neoclassical model is outlined in Robert E. Hall and Dale W. Jorgenson, "Tax Policy and Investment Behavior," *American Economic Review*, vol. 57 (June 1967), pp. 391–414.

5. Dale W. Jorgenson and Calvin D. Siebert, "A Comparison of Alternative Theories of Corporate Investment Behavior," *American Economic Review*, vol. 58 (September 1968), pp. 681–712; J. W. Elliott, "Theories of Corporate Investment Behavior Revisited," *American Economic Review*, vol. 63 (March 1973), pp. 195–207.

analysis. Moreover, with formal nonnested specification tests of investment models estimated from quarterly time series data, and accounting for first-order serial correlation of the residuals, Ben Bernanke, Henning Bohn, and Peter Reiss find that all of the standard models are rejected by at least one other model.⁶

Apart from econometric issues, the assumption of representative firms is common to all this research—that is, the same empirical model applies to all firms regardless of the specification. Therefore, tests could not ascertain whether the observed empirical sensitivity of investment to financial variables differed in different kinds of firms.⁷ Thus, the representative firm paradigm limited the explanations that could be provided for financial effects.

The empirical work in this article relates the traditional study of financial effects on investment to recent literature on capital market imperfections by studying investment behavior in groups of firms with different financial characteristics.⁸ This change in empirical technique may help explain some aspects of the empirical paradoxes evident in past investment studies. If only certain classes of firms face capital market imperfections and corresponding financial constraints, the finding of Elliott, for example, that financial effects for a comparatively

6. For comparative studies, see Charles W. Bischoff, "Business Investment in the 1970s: A Comparison of Models," *BPEA*, 1:1971, pp. 13–58; Richard W. Kopcke, "The Behavior of Investment Spending during the Recession and Recovery, 1973–76," *New England Economic Review* (November–December 1977), pp. 5–41; Peter K. Clark, "Investment in the 1970s: Theory, Performance, and Prediction," *BPEA*, 1:1979, pp. 73–113; Ben Bernanke, Henning Bohn, and Peter C. Reiss, "Alternative Non-Nested Specification Tests of Time-Series Investment Models," *Journal of Econometrics*, vol. 37 (March, 1988).

7. Robert Eisner's extensive study of firm-level data provides an exception to the typical assumption of representative firms. Eisner found that the timing of investment in small firms is more sensitive to profits than it is in large firms. Robert Eisner, *Factors in Business Investment* (Ballinger Press, 1978).

8. Much recent work has studied general financial effects on real economic activity. See Mark Gertler, "Financial Structure and Aggregate Economic Activity: An Overview," *Journal of Money, Credit and Banking*, forthcoming; Alan S. Blinder, "Credit Rationing and Effective Supply Failures," *Economic Journal*, vol. 97 (June 1987), pp. 327–52; Charles W. Calomiris and R. Glenn Hubbard, "Price Flexibility, Credit Availability, and Economic Fluctuations: Evidence from the United States, 1879–1914" (Northwestern University, 1987); Ben S. Bernanke, "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review*, vol. 73 (June 1983), pp. 257–76; Charles W. Calomiris, R. Glenn Hubbard, and James H. Stock, "The Farm Debt Crisis and Public Policy," *BPEA*, 2:1986, pp. 441–79.

broad sample of firms are significant need not conflict with the findings of Jorgenson and Siebert that a model emphasizing only real factors explains investment better for a group of well-known, mature firms. Both empirical approaches are appropriate in certain contexts. The problem, common to both, is the use of the representative firm assumption to explain investment for all firms. Therefore, the issue need not be posed as whether firm financial conditions “matter” for investment in some aggregate sense, or whether models of investment that emphasize financial variables fit the aggregate data better or forecast better than models that include only real variables.

Sources and Cost of Finance

As information on their sources and uses of funds shows, the financing practices of U.S. firms vary widely.⁹ Table 1 summarizes the financing practices of manufacturing firms during 1970–84, the same period covered by the sample of manufacturing firms analyzed later. We report the percentage of total finance coming from short-term bank debt, long-term bank debt, other long-term debt, and retained earnings for six firm size classifications. We also report the average retention ratio. The data exclude new equity issues, which are small in the aggregate. Financing obtained by small firms constitutes a nontrivial portion of the aggregate. Firms with under \$10 million in assets accounted for 14 percent of the total finance raised over the period; firms with under \$100 million in assets, for 26 percent of the total.

Internal finance in the form of retained earnings generates the majority of net funds for firms in all size categories.¹⁰ The importance of internal

9. Early case studies suggested that small firms have more limited access to external finance than do large firms. See J. Keith Butters and John Lintner, *Effect of Federal Taxes on Growing Enterprises* (Division of Research, Graduate School of Business Administration, Harvard University, 1945); Meyer and Kuh, *The Investment Decision*; Gordon Donaldson, *Corporate Debt Capacity: A Study of Corporate Debt Policy and the Determination of Corporate Debt Capacity* (Division of Research, Graduate School of Business Administration, Harvard University, 1961).

10. This pattern has been true historically as well. U.S. manufacturing firms have relied heavily on internal finance for growth and development since at least the end of the nineteenth century. See, for example, the discussions by Lawrence H. Seltzer, *A Financial History of the American Automobile Industry* (Houghton Mifflin, 1928); and Meyer and Kuh, *The Investment Decision*.

Table 1. Sources of Funds, by Asset Class, U.S. Manufacturing Firms, 1970–84

Firm size	Source of funds (percent of total) ^a			Retained earnings	Percentage of long-term debt from banks	Average retention ratio
	Short-term bank debt	Long-term bank debt	Other long-term debt			
All firms	0.6	8.4	19.9	71.1	29.6	0.60
<i>Asset class</i>						
Under \$10 million	5.1	12.8	6.2	75.9	67.3	0.79
\$10–50 million	5.9	17.4	6.9	69.8	71.6	0.76
\$50–100 million	3.1	12.9	5.3	78.7	71.0	0.68
\$100–250 million	–0.2	13.3	12.0	74.9	52.4	0.63
\$250 million–\$1 billion	–2.3	10.6	15.4	76.3	40.8	0.56
Over \$1 billion	–0.6	4.8	27.9	67.9	14.7	0.52

Source: Authors' calculations based on data taken from U.S. Department of Commerce, Bureau of the Census, *Quarterly Financial Reports of Manufacturing, Mining, and Trade Corporations*, various issues. The data underlying the calculations are expressed in 1982 dollars.

a. Funds raised from new equity issues are excluded from the calculations.

finance would be even greater if we were able to include information on depreciation allowances, a source of internal funds roughly equal to retained earnings. Furthermore, the proportion of earnings retained by firms differs substantially by size classes. The average retention ratio is almost 80 percent for the smallest firms in table 1; it drops monotonically as firm size increases, to a low of approximately 50 percent for firms with assets of more than \$1 billion.¹¹

Differences in debt finance across size groupings are also important. Firms in the smallest classes accounted for the majority of net new short-term bank debt. Firms with assets of less than \$250 million got most of their debt finance from *banks*—lending institutions specializing in monitoring borrowers through customer relationships—while firms with

11. It is not likely that differences in retention rates by size grouping are traceable solely to the relative tax price of dividends in determining payout for small corporations with concentrated ownership. For example, Dun and Bradstreet surveyed 365 (“small,” “medium-sized,” and “large”) manufacturing concerns in 1937 to determine the sources of increased net worth from 1920 to 1928, a period in which the relative price of dividends and retentions (capital gains) to shareholders was virtually unity. Of small firms, 94 percent obtained more than 90 percent of their finance from retention, compared with 70 percent for large firms. Sixteen percent of large firms obtained at least half of their finance from new share issues over the period, compared with only 1 percent of the small firms. The survey results are reviewed in detail in Willard L. Thorp and Edwin B. George, “An Appraisal of the Undistributed Profits Tax,” *Dun’s Review* (September 1937), pp. 5–36.

assets of more than \$1 billion financed more than 85 percent of their new debt through *nonbank* sources.

Independent evidence by Philip Srinivasan indicates that manufacturing corporations with assets of less than \$100 million raised only 2 percent of their total finance from net new share issues from 1960 to 1980.¹² Srinivasan also finds that internal finance is more volatile over the business cycle in small and medium-sized corporations than in large corporations. Moreover, during downturns, large firms have greater relative access to short-term and long-term debt markets. Hence, if internal and external sources of funds are not perfect substitutes, business recessions and changes in corporate tax policy that affect internal finance will likely have a greater effect on the growth rates and investment behavior of small, immature enterprises.

THE COST OF INTERNAL VERSUS EXTERNAL FINANCE

To provide a microfoundation for links between a firm's financial structure and its real investment spending, one must identify reasons why internal and external finance are not perfect substitutes in practice. In fact, explanations why internal finance may be less costly than new share issues and debt finance abound. Among the most prominent are transaction costs, tax advantages, agency problems, costs of financial distress, and asymmetric information. We emphasize asymmetric information between managers and potential new investors or creditors.

New Share Issues. New share issues of seasoned equity in the United States are typically carried out by underwriters who purchase a block of new shares and resell it. Relative to gross proceeds, the cost of a new share issue, including underwriting discounts, registration fees and taxes, and selling and administrative expenses, can vary substantially by size of offering. Costs for small offerings can be high.¹³ In addition, both direct and indirect costs of offerings are higher for initial public offerings than for seasoned offerings.

12. Philip Vijay Srinivasan, "Credit Rationing and Corporate Investment" (Ph.D. dissertation, Harvard University, October 1986).

13. Transaction costs were recognized as a substantial impediment to the ability of small and medium-sized firms to raise equity capital in the 1930s. See U.S. Securities and Exchange Commission, "Cost of Flotation for Registered Securities, 1938-1939" (Washington, D.C.: Research and Statistics Section, Trading and Exchange Division, Securities and Exchange Commission, March 1941).

The design of the corporate tax system in the United States and in other countries has historically imparted a cost advantage to internal equity finance over external equity finance. In the United States for many years, the effective tax rate on capital gains has been much lower than the tax rate on dividends. Recent studies show that this differential gives a cost advantage to internal finance; while no tax savings accrue from the issue of new shares, tax savings do arise when earnings are retained rather than paid out, because a dividend tax is replaced with a lower tax on capital gains.

Mervyn King and Alan Auerbach calculate shadow prices for the cost of internal finance (r) and the cost of new share issues (s).¹⁴ They establish that $r = \rho/(1-\tau)(1-c)$ and $s = \rho/(1-\tau)(1-\theta)$, where ρ is the after-tax rate of return required by the capital market, τ is the corporate tax rate, and c and θ are the tax rates on capital gains and dividends, respectively. The tax cost of new share issues can be expressed as $(s-r)/r = (\theta-c)/(1-\theta)$. Alternatively, within a q framework, the threshold marginal q value a project must attain to be undertaken depends on how it is financed. Shareholders benefit from externally financed projects only if their marginal q exceeds unity. On the other hand, projects financed with retentions need only attain a q of $(1-\theta)/(1-c) < 1$.

Asymmetric information can generate potentially significant cost disadvantages of external finance for some kinds of firms. The theoretical arguments that support this view draw heavily on the "lemons" problem first considered by George Akerlof.¹⁵ The core of the argument is that

Clifford Smith finds that total costs as a percentage of proceeds in a sample of underwritten issues from 1971 to 1975 vary from 14 percent for issues under \$1 million to 4 percent for issues over \$100 million. Similar estimates of the cost differential by size of issue have been made in other studies. Clifford W. Smith, Jr., "Alternative Methods for Raising Capital: Rights versus Underwritten Offerings," *Journal of Financial Economics*, vol. 5 (December 1977), table 1, p. 277.

14. Mervyn A. King, *Public Policy and the Corporation* (London: Chapman and Hall, 1977); Alan J. Auerbach, "Wealth Maximization and the Cost of Capital," *Quarterly Journal of Economics*, vol. 93 (August 1979), pp. 433–46. See also David F. Bradford, "The Incidence and Allocation Effects of a Tax on Corporate Distributions," *Journal of Public Economics*, vol. 15 (February 1981), pp. 1–22; and the review of alternative approaches in James M. Poterba and Lawrence H. Summers, "The Economic Effects of Dividend Taxation," in Edward I. Altman and Marti G. Subrahmanyam, eds., *Recent Advances in Corporate Finance* (Homewood, Illinois: Richard D. Irwin, 1985), pp. 227–84.

15. George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, vol. 84 (August 1970), pp. 488–500.

some sellers with inside information about the quality of an asset or a security will be unwilling to accept the terms offered by a less-informed buyer. This may cause the market to break down, or at least force the sale of an asset at a price lower than it would command if all buyers and sellers had full information.

These ideas are applied to the problem of equity finance by Stewart Myers and Nicholas Majluf and by Bruce Greenwald, Joseph Stiglitz, and Andrew Weiss. In these 'pecking order' or 'financing hierarchy' theories, the firm's managers are assumed to have full information about the value of the firm's existing assets and the returns from new investment projects.¹⁶ Thus, to the extent that managers control sufficient internal funds to finance all profitable investment projects, investment demand models based on a representative firm in a perfect capital market apply. Suppose, however, that a firm exhausts all its internal funds and requires external finance to undertake a desirable project. In the Myers and Majluf model, external investors cannot distinguish the quality of firms; they value them all at the population average. Consequently, new shareholders implicitly demand a premium to purchase the shares of relatively good firms to offset the losses that will arise from funding lemons. The premium can raise the cost of new equity finance faced by managers of relatively high-quality firms above the opportunity cost of internal finance faced by existing shareholders. ✓

The intuition behind the lemons premium can be described in terms of the q model of investment. Following Myers and Majluf, we can say that an investment that requires new share issues will be undertaken only if it increases the wealth of existing shareholders. For good firms, the true gross returns from assets in place are denoted by Y and the returns from a new project by Y' . Myers and Majluf show that new shares will be issued only if

$$Y'/I \geq Y/V,$$

16. Stewart C. Myers and Nicholas S. Majluf, "Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have," *Journal of Financial Economics*, vol. 13 (June 1984), pp. 187–221; Bruce Greenwald, Joseph E. Stiglitz, and Andrew Weiss, "Information Imperfections in the Capital Market and Macroeconomic Fluctuations," *American Economic Review*, vol. 74 (May 1984, *Papers and Proceedings*, 1983), pp. 194–99. The pecking-order view is described in Stewart C. Myers, "The Capital Structure Puzzle," *Journal of Finance*, vol. 39 (July 1984), pp. 575–92.

where I is the cost of the new investment and V is the market value assigned to both good firms and lemons. This condition is equivalent to requiring that the marginal q on the new project at least equal the ratio of the firm's true average q —call it q^* —to the average q assigned to all firms by the market (\bar{q}). With full information, $q^*/\bar{q} = 1$, and the threshold q value for issuing new shares would be unity, as in conventional models. When good firms initially cannot be distinguished from lemons, however, q^*/\bar{q} will exceed unity for good firms. This ratio indicates how much dilution occurs when such firms issue new shares. The quantity $(q^*/\bar{q}) - 1$ is the lemons premium that we denote by Ω .

Debt Finance. Standard treatments of the effects of leverage on the firm's cost of funds posit an increasing marginal cost of new debt due to costs of financial distress and agency costs. Financial distress costs arise when a firm has difficulties meeting its principal and interest obligations—the extreme case being bankruptcy. Agency costs arise from the limited-liability feature of debt contracts that creates incentives for firm managers to act counter to the interests of creditors under some circumstances.

Debt finance, particularly long-term debt, creates agency problems. The greater the debt-equity ratio, the more the incentives of managers who act in the interest of equity owners diverge from the interests of creditors. Managers may forgo some investment opportunities with positive net present values and accept others with negative present values. They also have incentives to issue new debt that raises the riskiness and lowers the value of existing debt. Because creditors understand the conflicts of interest that exist between themselves and equity holders, they demand covenants that restrict the behavior of managers, particularly with respect to new debt issues.¹⁷ As a result, covenants typically stipulate target debt-equity ratios. While they may provide a second-best solution to the contracting problem given the potential for opportunism, they are not costless, and their restrictions on financial flexibility limit management's choices of investment opportunities, as well as the ability to finance investment opportunities when internal funds are low. If covenants impose working capital requirements, for example, the supply of internal funds available to finance investment may be reduced. Hence, shocks to working capital, such as

17. See the description of covenants in Clifford W. Smith, Jr., and Jerold B. Warner, "On Financial Contracting: An Analysis of Bond Covenants," *Journal of Financial Economics*, vol. 7 (June 1979), pp. 117–61.

a debt deflation or a decline in internal finance, will make debt finance more expensive at the margin, probably at a time when the need for new debt is most acute.

Asymmetric information in markets for debt can cause distortions similar to those discussed previously for new share issues. Asymmetric information may increase the cost of new debt, or even result in credit rationing. Dwight Jaffee and Thomas Russell show that the market interest rate must rise, and loan size may be limited, when lenders cannot distinguish borrower quality.¹⁸ Stiglitz and Weiss demonstrate that “equilibrium credit rationing” can arise from adverse selection. Again, the lemons argument is critical. Lenders cannot price discriminate between good borrowers and bad in loan contracts because of asymmetric information. Thus, when interest rates rise, relatively good borrowers drop out of the market, increasing the probability of default and possibly reducing the lenders’ expected profit. In equilibrium, lenders may set an interest rate that leaves an excess demand for loans in the market. Some borrowers receive loans while other observationally equivalent borrowers are rationed.¹⁹

Calomiris and Hubbard add heterogeneous debt markets and agents that are restricted from borrowing in some markets to the Stiglitz-Weiss structure.²⁰ Two credit markets, a “full-information” market (bond or commercial paper, for example) and a bank loan market, coexist. The banks specialize in financing projects of borrowers for which information problems are more severe, in the sense that costs of obtaining borrower information are high and lenders can reduce average information costs by maintaining long-term relationships. The central proposition in this work is that, depending on per capita levels of internal net worth, the

18. Dwight M. Jaffee and Thomas Russell, “Imperfect Information, Uncertainty and Credit Rationing,” *Quarterly Journal of Economics*, vol. 90 (November 1976), pp. 651–66.

19. Joseph E. Stiglitz and Andrew Weiss, “Credit Rationing in Markets with Imperfect Information,” *American Economic Review*, vol. 71 (June 1981), pp. 393–410.

20. Charles W. Calomiris and R. Glenn Hubbard, “Firm Heterogeneity, Internal Finance, and Credit Rationing,” Working Paper 2497 (National Bureau of Economic Research, January 1988). In addition, the importance of borrower net worth for obtaining external finance is stressed by Hayne E. Leland and David H. Pyle, “Informational Asymmetries, Financial Structure, and Financial Intermediation,” *Journal of Finance*, vol. 32 (May 1977), pp. 371–87; Myers and Majluf, “Corporate Financing Decisions”; Ben S. Bernanke and Mark Gertler, “Financial Fragility and Economic Performance,” Working Paper 2318 (NBER, July 1987).

allocation of new funds to classes of borrowers could either follow the full-information credit allocation or ration funds away from some classes of borrowers who would receive credit in the absence of asymmetric information. A “financial collapse” may occur, in which some or all classes of asymmetric-information borrowers are denied loans.

Finally, while it is generally true that higher leverage entails a higher shadow price of funds, only the largest and most mature firms are likely to face a smoothly increasing loan interest rate. Several features of heterogeneity are important here. Small and medium-sized firms are less likely to have access to impersonal centralized debt markets. Indeed, outside the Fortune 500 companies, the overwhelming majority of bond finance has been obtained historically through private placements, usually with life insurance companies or pension funds. Two features of private placements are significant. First, they are more restrictive than typical bond arrangements, requiring minimum levels of working capital and stockholders’ equity and often limiting dividend payments and capital spending. Second, during periods of tight credit, small and medium-sized borrowers are often denied loans in favor of better-quality borrowers, who could also obtain funds from centralized securities markets. Similarly, bank loans and lines of credit, the typical source of finance for smaller industrial firms, restrict operating flexibility and require particular levels for certain financial operating ratios.²¹ With constant investment opportunities, it is precisely in times of a decline in

21. With respect to private placements, see the extensive discussion in Eli Shapiro and Charles Wolf, who note that from 1953 to 1970, Fortune 500 companies obtained an average of 37 percent of their bond finance through private placements, compared with an average of 75 percent for other manufacturing firms. Eli Shapiro and Charles R. Wolf, *The Role of Private Placements in Corporate Finance* (Division of Research, Graduate School of Business Administration, Harvard University, 1972), p. 150.

With respect to bank finance, see the analysis of data for manufacturing firms from the *Quarterly Financial Reports* of the U.S. Bureau of the Census in Srinivasan, “Credit Rationing,” chap. 3. Although small businesses can borrow from commercial banks, the banks cannot (absent secured mortgages) furnish long-term funds as a substitute for equity or bonds; maturities of from three to five years are typically the longest available. The Small Business Administration, which can guarantee loans of longer maturities, is not active in industrial finance; see Barry P. Bosworth, Andrew S. Carron, and Elizabeth H. Rhyne, *The Economics of Federal Credit Programs* (Brookings, 1987).

For the use of financial ratios as a predictor of bankruptcy, see Edward I. Altman, “Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy,” *Journal of Finance*, vol. 23 (September 1968), pp. 589–609.

internal finance that such firms cannot obtain debt finance on the margin for capital spending projects.

As we noted before, covenants in debt contracts protect the interests of bondholders from opportunistic behavior on the part of shareholders. To the extent that difficulties in contracting in debt markets are related only to agency problems and not to asymmetric information, equity markets could provide the marginal source of external finance for firms. However, firms facing asymmetric information problems in credit markets will also probably need to pay a premium to obtain new equity. Therefore, equity finance will not, in general, solve asymmetric information problems associated with debt.

“Financing Hierarchies” and Investment

The preceding discussion of the cost premium that some firms must pay for external finance can be integrated into a model of firm financial and investment decisions developed in the public finance literature (see Appendix A and the references therein). In the standard model, the value of a firm, V , is the present value of the posttax dividend stream adjusted for the amount of new share issues, V^N , that current equity holders would have to purchase to maintain their proportional claim on the firm. Formally, the value of the firm is

$$(1) \quad V_t = \sum_{i=0}^{\infty} \left(1 + \frac{\rho}{1-c}\right)^{-(i-1)} \left[\left(\frac{1-\theta}{1-c}\right) D_{t+i} - V_{t+i}^N \right],$$

where ρ is the required return on equity, D_t represents the dividend payment in period t , θ is the tax rate on dividends, and c is the tax rate on capital gains. Managers maximize the value of existing shareholders' stock subject to a set of constraints on the distribution of earnings (see Appendix A). The solution for the case of $\theta > c$ is well known; it is never optimal to issue new shares and pay dividends at the same time. Here, whenever internal finance exceeds desired investment, q is $(1 - \theta)/(1 - c) < 1$ in equilibrium, as discussed previously. A value-maximizing firm will issue new shares only after it exhausts internal finance and $q > 1$. Thus, the breakeven q a project must attain depends on how it will be financed.

The same kind of logic applies to firms facing asymmetric information, but the cost differential between internal and external finance may be

much larger. The above expression for V_t can be modified to include a lemons premium demanded by potential new equity investors when asymmetric information problems exist. We reduce V_t in equation 1 by an amount Ω_t per dollar of new equity issued, or

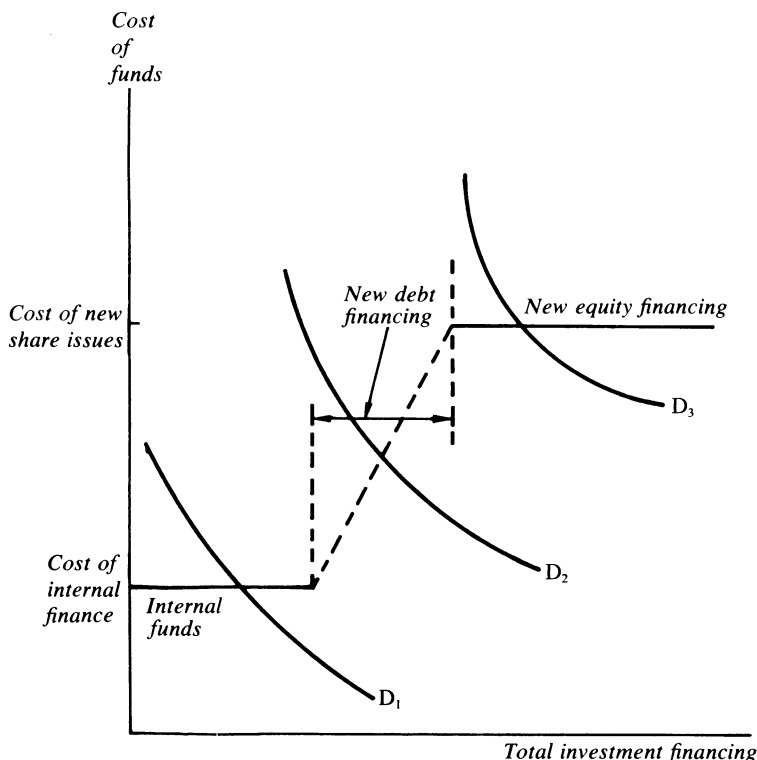
$$(2) \quad V_t = \sum_{i=0}^{\infty} \left(1 + \frac{\rho}{1-c}\right)^{-(i-1)} \left[\left(\frac{1-\theta}{1-c}\right) D_{t+i} - \left(1 + \Omega_{t+i}\right) V_{t+i}^N \right],$$

where Ω reflects the additional value that new investors demand from good firms to compensate them for the losses they incur from inadvertently funding lemons. With this modification to the model, the breakeven q value for investment projects financed by new share issues becomes $1 + \Omega$.

This financing hierarchy is depicted in figure 1. The solid lines in the figure represent a simple case of a discontinuous differential in the costs of internal and external equity finance.²² When investment demand is low, as with the D_1 schedule, capital spending can be financed from internally generated funds, at the expense of extra dividends. At very high levels of investment demand, as with the D_3 schedule, firms will issue new shares. The higher the value of Ω , the greater the likelihood that internal finance will constrain a firm's investment, as illustrated by the D_2 schedule. Of course, the lemons premium can vary both across firms and over time for the same firm. If information problems become less severe, the top horizontal schedule in figure 1 will shift downward toward unity.

Debt finance can also be incorporated. To the extent that debt can be secured, or obtained from lenders, such as commercial banks, that specialize in monitoring the borrower, information problems in debt markets will be less severe than those in external equity markets, but the marginal cost of debt will increase with leverage, as discussed

22. Some recent studies have tested for implied cost differences between internal and external equity finance. See Robert L. McDonald and Naomi Soderstrom, "Dividend and Share Changes: Is There a Financing Hierarchy?" Working Paper 2029 (NBER, September 1986); Avner Kalay and A. Shimrat, "On the Payment of Equity-Financed Dividends" (New York University, December 1985); Kalay and Shimrat, "Firm Value and Seasoned Equity Issues: Price Pressure, Wealth Distribution, or Negative Information," Working Paper 894/86 (New York University, March 1986). Also see Paul Asquith and David W. Mullins, Jr., "Equity Issues and Offering Dilution," *Journal of Financial Economics*, vol. 15 (January–February 1986), pp. 61–89; and Ronald Masulis and A. N. Korwar, "Seasoned Equity Offerings: An Empirical Investigation," *Journal of Financial Economics*, vol. 15 (January–February 1986), pp. 91–118.

Figure 1. Investment and Financing Decisions

previously. This modified hierarchy is illustrated by the dotted line in figure 1 that connects the two horizontal segments in the middle range of the figure. Hence, intermediate levels of investment demand, as illustrated by the D_2 schedule, will be financed by a mix of internal funds and debt.

This financing hierarchy has a number of implications for q values and investment behavior. First, all other things equal, observed q values will differ in firms with different information characteristics. For firms facing asymmetric information, the observed q value will be the value assigned by the imperfectly informed market. The model also predicts that q must be substantially higher to induce a new share issue for limited-information (high Ω) firms than for full-information (low Ω) firms.

The true marginal q is unobservable; we can, however, observe the

average q assigned by the market and its relationship to new share issues. Observed q can move independently from the true valuation for limited-information firms. For example, the market may reappraise the underlying probability that a firm is a lemon. If the asymmetric information problem is important empirically, observed q values should be high relative to historical values before new share issues for limited-information firms.

Finally, internal finance constrains spending for firms that do not pay dividends and face an investment demand schedule like D_2 in figure 1. When q is sufficiently high, new shares are issued, and movements in q lead to movements in investment. Otherwise, investment will be driven by changes in internal finance. In the limiting case, with a vertical debt supply schedule, variations in the length of the retention segment in figure 1 should cause corresponding variations in investment for firms that pay no dividends. More generally, the slope of the debt supply schedule will determine the extent to which firms can offset reductions in internal finance with greater leverage. Therefore, the larger the lemons premium, Ω , the greater the chances that a firm will have an investment demand curve like D_2 , where investment opportunities, as measured by a project's marginal q , can vary, while investment responses are affected by the availability of internal finance. Such a pattern resembles the predictions of sales-accelerator models of investment; we discuss this point in more detail later.

In summary, if the cost of capital differs by source of funds, the availability of finance will likely have an effect on the investment practices of some firms. In financing hierarchy models like the one summarized in figure 1, the availability of internal funds allows firms to undertake desirable investment projects without resorting to high-cost external finance. In addition, to the extent that a firm seeks debt finance at the margin, greater internal cash flow enhances its balance sheet and net worth positions, lowering the cost of new debt.

Differences in Firm Financing Practices

To examine the empirical importance of these ideas for explaining investment, we use a large panel of Value Line data for manufacturing firms. The details of the sample structure and definitions of the empirical variables are discussed in Appendix B. The firms in this data base are

typically large, and their stock is publicly traded. Evidence that some of these firms face financing constraints should indicate that the phenomenon is widespread.

Our approach is to study differences in financing and investment in groups of firms with different characteristics. Observed retention practices provide a useful a priori criterion for identifying firms that are likely to face relatively high costs of external finance. If the cost disadvantage of external finance is large, it should have the greatest effect on firms that retain most of their income. If the cost disadvantage is slight, then retention practices should reveal little about financing practices, q values, or investment behavior.²³

Our classification scheme divides firms into three groups. Class 1 firms have a ratio of dividends to income less than 0.1 for at least 10 years. Class 2 firms have a dividend-income ratio less than 0.2, but more than 0.1, for at least 10 years. Class 3 includes all other firms.

We considered further divisions of the high-payout firms in class 3, but we did not find substantial differences between firms that paid out 20–40 percent of their income on average as dividends and firms that paid out more than 40 percent. Because of possible outliers of the dividend-income ratio, due to abnormally low income in a particular year, this approach is more robust than classifying firms according to their average retention ratio.

One reason why firms might pay low dividends is that they require investment finance that exceeds their internal cash flow and retain all of the low-cost internal funds they can generate. A second is that they have little or no income to distribute. We are interested in the first group and, for this reason, have included only those firms in the sample that had positive real sales growth from 1969 through 1984. To avoid any biases across retention classes, this restriction was applied to all firms in the sample, not just the low-dividend class. The results that follow were not changed substantially by including firms with negative sales growth in the sample.

23. Our scheme for grouping firms according to differences in dividend behavior is similar to tests for the presence of liquidity constraints on consumption, in which households are grouped into high-wealth and low-wealth categories. See for example Fumio Hayashi, "The Effect of Liquidity Constraints on Consumption: A Cross-Sectional Analysis," *Quarterly Journal of Economics*, vol. 100 (February 1985), pp. 183–206; and Stephen P. Zeldes, "Consumption and Liquidity Constraints: An Empirical Investigation," Working Paper 24-85 (Rodney L. White Center for Research, Wharton School, University of Pennsylvania, November 1985).

Table 2. Summary Statistics: Sample of Manufacturing Firms, 1970–84

Statistic	Category of firm		
	Class 1 ^a	Class 2 ^b	Class 3 ^c
<u>Number of firms</u>	49	39	334
<u>Average retention ratio</u>	0.94	0.83	0.58
Percent of years with positive dividends	33	83	98
Average real sales growth (percent per year)	13.7	8.7	4.6
Average investment-capital ratio	0.26	0.18	0.12
Average cash flow-capital ratio	<u>0.30</u>	0.26	0.21
Average correlations of cash flow with investment (deviations from trend) ^d	0.92	0.82	0.20
Average of firm standard deviations of investment-capital ratios	0.17	0.09	0.06
Average of firm standard deviations of cash flow-capital ratios	0.20	0.09	0.06
Capital stock (millions of 1982 dollars)			
Average capital stock, 1970	100.6	289.7	1,270.0
Median capital stock, 1970	27.1	54.2	401.6
Average capital stock, 1984	320.0	653.4	2,190.6
Median capital stock, 1984	94.9	192.5	480.8

Source: Authors' calculations based on samples selected from the Value Line data base. See Appendix B.

a. Firms with dividend-income ratios of less than 0.1 for at least 10 years.

b. Firms with dividend-income ratios greater than 0.1 but less than 0.2 for at least 10 years.

c. Firms with dividend-income ratios greater than 0.2.

d. Estimated from time series constructed by aggregating the sample data within each category.

Several summary statistics for the firms in each class are presented in table 2. Our class 1 firms, those that we hypothesize will more likely face binding financial constraints, retained an average of 94 percent of their income and paid a dividend in only 33 percent of the years. Many of these firms paid no dividends for the first 7 to 10 years and a small dividend in the remaining years. In fact, 20 firms never paid a dividend.

Class 1 firms experienced much more rapid growth in the fixed capital stock than the mature firms in class 3. Mean values of the capital stock are, of course, influenced by extreme values. The growth pattern for median values is similarly striking. While class 1 firms are smaller than firms in class 3, they are still large relative to U.S. manufacturing corporations in general; 85 percent of manufacturing corporations had

smaller capital stocks in 1970 than the average class 1 firm.²⁴ Firms in class 1 have a high mean investment-to-capital ratio, and they exhaust nearly all of their cash flow on investment spending. Firms in class 3 spend a much lower proportion of their cash flow on investment. Both cash flow and investment are more volatile in class 1, as the standard deviation statistics in table 2 indicate. Table 2 also shows a striking difference in the correlation of deviations from exponential trends of cash flow and investment between classes 1 and 2 and class 3. These statistics are estimated from time series constructed by aggregating the sample data within each class. The correlations suggest the greater sensitivity of investment to cash flow in classes 1 and 2 that we find in the regression equations that follow.

The data in table 3 present information on new share issues, debt finance, and q values for firms in the various classes.²⁵ Other things being equal, one would expect firms in class 1 to rely more heavily on new share issues than firms in the remaining classes. The typical firm in class 1 has an investment demand schedule like D_2 or D_3 in figure 1. The typical firm in class 3 has a demand schedule like D_1 and should not simultaneously pay dividends and issue new shares, given the historical differences in dividend and capital gains tax rates. Consistent with their rapid growth, firms in class 1 issue new shares more frequently—approximately one year in every four—than the firms in the other classes. Firms in the first class also raise a greater proportion of total finance from new shares. Even for class 1 firms, however, new share issues provide a much smaller proportion of total funds than internal cash flows.

24. We estimated a probit model for the probability that a firm is included in class 1—as a function of size (capital stock in 1977), average real sales growth over the sample period, the average value of q , the average value of the ratio of outstanding debt to the market value of debt and equity, and the standard deviation of earnings (measured relative to the capital stock). The results are consistent with what one would expect based on the summary statistics reported in table 2. While firms in class 1 are smaller on average than firms in class 3, size as such does not appear to be the dominant factor explaining why firms fall into the high-retention class 1. The size variable in the probit equation has a negative estimated coefficient, but it is not as statistically or economically significant as the estimated coefficients for most of the other variables.

25. Some firms reported infrequent, but very small new share issues that were probably associated with executive stock option plans. In the calculations presented in table 3, we excluded such small issues by requiring that funds raised from new common stock exceed 10 percent of the firm's cash flow in the same year; stock splits are also excluded.

Table 3. New Share Issues, Tobin's q , and Debt Statistics for Manufacturing Firms, 1970–84

<i>Item</i>	<i>Class 1</i>	<i>Class 2</i>	<i>Class 3</i>
Average percentage of years with new share issues	28	19	10
Average value of share issues as a percentage of cash flow	23	13	8
Average annual q values ^a	3.8 (0.4)	2.4 (0.2)	1.6 (0.1)
Median q values	1.6	1.4	1.0
Average difference in q values between periods of new share issues and periods of no new share issues ^a	1.6 (0.8)	0.9 (0.4)	0.2 (0.1)
Average ratio of debt to capital stock	0.57	0.52	0.33
Average ratio of interest payments to sum of interest payments plus cash flows	0.27	0.21	0.17
Correlation of the earnings-to-capital ratio and the change in total debt-to-capital ratio (averaged over firms)	0.23	0.15	0.09

Source: Same as table 2.

a. The standard error of the mean appears in parentheses.

The last three lines of table 3 provide information on debt use. Although one would expect the firms in class 3 to have higher debt capacities, the debt-to-capital and interest expense ratios are higher for classes 1 and 2. These results are consistent with a financing hierarchy and support the idea that constrained firms borrow up to their debt capacity.²⁶ Nor is there any indication in the data that debt issues smooth fluctuations in cash flow. For 43 of the 49 class 1 firms, the correlation of the earnings-to-capital ratio with the change in the total debt-to-capital ratio is positive. As shown in table 3, the average correlation of earnings

26. The pattern of debt leverage across classes also holds for debt-equity ratios measured as the book value of debt divided by the book value of common equity. For the empirical effect of debt service on investment, see Allen Sinai and Otto Eckstein, "Tax Policy and Business Fixed Investment Revisited," *Journal of Economic Behavior and Organization*, vol. 4 (June–September 1983), pp. 131–62; Steven M. Fazzari and Michael J. Athey, "Asymmetric Information, Financing Constraints, and Investment," *Review of Economics and Statistics*, vol. 69 (August 1987), pp. 481–87.

with the change in debt is positive for all classes, but it is largest for class 1. This result also holds up in regressions that control for investment opportunities through q . The change in either long-term or total debt is positively related to cash flow when it is regressed on q and cash flow (all variables were deflated by the capital stock). Therefore, changes in debt appear to reinforce rather than offset fluctuations in cash flow, especially for class 1 firms, for which the positive estimated sensitivity of changes in debt to cash flow fluctuations was the largest.²⁷

Table 3 also reports Tobin's q measures for all three classes of firms.²⁸ The average q value for the first two classes is significantly greater than the averages for the third. The asymptotic t -statistic for the null hypothesis that the first class mean equals the third class mean is 5.8. This result also holds for every year in the sample individually. Similar patterns hold for median q values.

One might interpret the high q values observed in class 1 as the result of high expected growth rates. As table 1 shows, firms in this class did indeed grow quickly over our sample period. Their high q values, however, beg the question of why they did not invest even more. As an alternative to financing constraints, high adjustment costs could slow convergence of q to a full-information equilibrium. Then, one would expect no systematic relation between q and new share issues. Firms would invest at an optimal pace to push q uniformly toward equilibrium, and new shares would be issued as necessary to finance capital spending.

The statistics in table 3, however, strongly contradict this view. We calculate the differences in q values in years with and without new share

27. A more detailed examination shows that for 21 of the class 1 firms, cash flow declined 25 percent or more on one or two occasions. In almost all cases, cash flow growth returned to normal in the next period, and the cash flow shock appeared to be temporary. In 26 cases, the debt-to-capital ratio either fell or remained unchanged in the next period; the ratio increased in only 5 instances. The evidence also indicates that debt is not on average an important source of bridge finance between new equity issues for these firms. If new debt were issued in the interim between new stock offerings, and the proceeds from the new equity were used to pay off debt, one would expect a negative correlation between new share issues and the change in debt. For the firms in our sample, however, the correlations between new equity finance and the changes in both total and long-term debt were essentially zero in all classes.

28. For measures of tax-adjusted Q (see the definition in Appendix B), the patterns were even more pronounced. The unadjusted q values reported were calculated with the book value of debt. The results were almost identical with various estimates of the market value of debt (see Appendix B).

issues on a firm-by-firm basis and then average these differences.²⁹ As noted in the table, for the three classes of firms, this procedure yields differences of, respectively, 1.6, 0.9, and 0.2. As discussed earlier, these results are consistent with a financing hierarchy.

Financial Constraints in Empirical Models of Investment

The theories discussed here imply that the supply of investment finance is not perfectly elastic for firms that face asymmetric information problems in capital markets. This result is independent of how one models the demand side of the investment decision. Indeed, the investment demand curve presented in figure 1 could be based on a q model of investment or a neoclassical model. Regardless of the true economic process at the foundation of investment demand, the supply of low-cost finance, and therefore the level of internal cash flow, enters the reduced-form investment equation of firms for which internal and external finance are not perfect substitutes.

In view of the longstanding debates in the literature over the appropriate specification of the model's demand side, we examine three broad empirical specifications that encompass the most common approaches: models based on q that emphasize market valuations of the firm's assets as the determinant of investment, sales accelerator models in which fluctuations in sales or output motivate changes in capital spending, and neoclassical models that combine measures of output and the cost of capital to explain investment demand. The most extensive tests of alternative specifications and estimation techniques are presented for the q model. These tests lead to similar conclusions for the other models.

29. The average differences reported in table 3 are computed as follows. We first compute the average difference on a firm-by-firm basis for all firms that issued shares, as defined above, in at least one of our sample years. These statistics are then averaged across firms in each class to obtain the results in table 2. Thus, differences in average q levels between firms that issue shares and firms that do not would not affect the reported statistics. Similar results can be obtained by regressing q on year dummies and a dummy variable for the new share issues.

An alternative explanation of the high q values in the firms in class 1 is the relative importance of "intangibles" for such firms. It is difficult, however, to link that explanation to the large differences in q values between periods in which new shares are issued and periods in which they are not.

The general form of the reduced-form investment equations that we examine is

$$(3) \quad (I/K)_{it} = f(X/K)_{it} + g(CF/K)_{it} + u_{it},$$

where I_{it} represents investment in plant and equipment for firm i during period t ; X represents a vector of variables, possibly including lagged values, that have been emphasized as determinants of investment from a variety of theoretical perspectives; and u is an error term. The function g depends on the firm's internal cash flow (CF); it represents the potential sensitivity of investment to fluctuations in available internal finance—after investment opportunities are controlled for through the variables in X .³⁰ We analyze other measures of internal liquidity later. All variables are divided by the beginning-of-period capital stock K .

As we stressed in our review of the implications of information problems in capital markets, empirical analysis must allow for systematic differences in the effect of potential finance constraints across firms. Our classification scheme based on retention practices identifies firms that are most likely to face capital market imperfections and the corresponding finance constraints. The evidence on firm financial behavior and q values across our retention classes, presented in the previous section, supports this view. If information problems in capital markets lead to financing constraints on investment, they should be most evident for the classes of firms that retain most of their income. If internal and external finance are nearly perfect substitutes, however, then retention practices should reveal little about investment by the firm. Firms would simply use external finance to smooth investment when internal finance fluctuates.

This test does not simply restate the accounting identity that sources equal uses of funds. Investment spending must be financed somehow, and cash flow provides a source of finance. Under an assumption of perfect capital markets, however, there is no reason to expect internal finance fluctuations to have different effects in firms with different

30. Other empirical studies that consider the effect of internal funds on investment include Kuh and Meyer, "Investment, Liquidity, and Monetary Policy"; Robert M. Coen, "The Effect of Cash Flow on the Speed of Adjustment," in Gary Fromm, ed., *Tax Incentives and Capital Spending* (Brookings, 1971), pp. 131–94; Eisner, *Factors in Business Investment*; Steven M. Fazzari and Tracy L. Mott, "The Investment Theories of Kalecki and Keynes: An Empirical Study of Firm Data, 1970–1982," *Journal of Post Keynesian Economics*, vol. 9 (Winter 1986–87), pp. 171–87.

retention behavior. Internal funds constitute only one possible source of investment finance, and their availability should not constrain investment unless the firm must pay a premium for new debt or equity finance.

INTERNAL FUNDS IN A Q MODEL OF INVESTMENT

We begin our empirical investigation of financing constraints and investment within the q -theory framework.³¹ The intuition of the model is that, absent considerations of taxes or capital market imperfections, a value-maximizing firm will invest as long as the shadow value of an additional unit of capital, marginal q , exceeds unity. In equilibrium, the value of an extra unit of capital is just its replacement cost, so that marginal q is unity. The conceptual advantage of this framework in modeling the effects of internal finance on investment is that q controls for the market's evaluation of the firm's investment opportunities.³²

We employ an empirical specification derived from an adjustment cost technology, and follow Lawrence Summers in specifying a cost of adjustment per unit of investment relative to capital. In the absence of financing constraints, Fumio Hayashi and Summers have linked the shadow price to the market value of existing capital (that is, average q). In that approach, under quadratic adjustment costs, investment is determined according to

$$(4) \quad (I/K)_{it} = \mu_i + \mu_1 Q_{it} + u_{it},$$

where μ_i is the normal value of (I/K) for the i th firm and u_{it} is an error term.³³ The term Q represents the value of q at the beginning of the

31. See the original discussions in William C. Brainard and James Tobin, "Pitfalls in Financial Model Building," *American Economic Review*, vol. 58 (May 1968, *Papers and Proceedings*, 1967), pp. 99–122; and James Tobin, "A General Equilibrium Approach to Monetary Theory," *Journal of Money, Credit and Banking*, vol. 1 (February 1969), pp. 15–29.

32. Andrew Abel and Olivier Blanchard found important roles for profits and output in aggregate investment equations relying on q , suggesting problems of aggregation or that alternative sources of finance are not perfect substitutes. Andrew B. Abel and Olivier J. Blanchard, "The Present Value of Profits and Cyclical Movements in Investment," *Econometrica*, vol. 54 (March 1986), pp. 249–73.

33. The quadratic adjustment cost framework that motivates a linear relationship between the investment-capital ratio and q , and the adjustments of q for corporate and personal taxation were developed by Andrew Abel, Lawrence Summers, and Fumio Hayashi. Assume that adjustment costs, A , follow: $A_{it} = (2\mu_1)^{-1} [(I/K)_{it} - \mu_i - u_{it}]^2 K_{it}$, if $[(I/K)_{it} - \mu_i] \geq 0$; and $A_{it} = 0$, otherwise. We also assume that shocks occur during the

period and is defined as the sum of the value of equity and debt less the value of inventories divided by the replacement cost of the capital stock, adjusted for corporate and personal tax considerations (see Appendix B for details). Estimates based on unadjusted q are very similar.

Table 4 presents estimates of the Q investment model, including cash flow, for each of the three retention classes. The equations were estimated with fixed firm and year effects.³⁴ Results are reported over three time periods, 1970–75, 1970–79, and 1970–84. There are two reasons to expect that the sensitivity of investment to cash flow in class 1 will be most pronounced in the shorter periods. First, most class 1 firms (26 out of 49) began paying dividends in the last two years of the sample, and were no longer exhausting all their internal funds. Second, as firms mature and more observations of project realizations and balance sheets are collected, asymmetric information problems should become less severe.

The structure of the Value Line data permits an interesting test of this possibility. A firm is not added to the data base until it is “of interest to subscribers and the financial community.” Once a firm is added, however, observations on items from its income statements and balance sheets are collected for at least 10 years prior to the date it is added to the Value Line data base. Most class 1 firms were not recognized until near the end of the sample period even though our data for these firms extend back to 1969.³⁵ Therefore the strongest case for asymmetric

period t so that the Q observed by the firm in formulating the capital spending decision is uncorrelated with the unanticipated components of the shocks. Andrew B. Abel, *Investment and the Value of Capital* (Garland Publishing Company, 1979); Lawrence H. Summers, “Taxation and Corporate Investment: A q -Theory Approach,” *BPEA*, 1:1981, pp. 67–127; Fumio Hayashi, “Tobin’s Marginal q and Average q : A Neoclassical Interpretation,” *Econometrica*, vol. 50 (January 1982), pp. 213–24.

34. Fixed time effects are included to capture aggregate business-cycle influences. Fixed firm effects account for unobserved time-invariant links between investment and the explanatory variables. That is, the “within” effect of Q or cash flow on investment is captured by our estimates. Problems of high values of average Q stemming from monopoly rents not captured in our formulation will be eliminated by using fixed-effects methods as long as the markup of price over marginal cost is constant over the period. See Eric B. Lindenberg and Stephen A. Ross, “Tobin’s q Ratio and Industrial Organization,” *Journal of Business*, vol. 54 (January 1981), pp. 1–32; Michael A. Salinger, “Tobin’s q , Unionization, and the Concentration-Profits Relationship,” *Rand Journal of Economics*, vol. 15 (Summer 1984), pp. 159–70.

35. Only 10 of the 49 firms were in the data base as of 1973. By 1980, 29 firms, over half the sample, were yet to be added. We thank Maria Latorraca of Value Line for providing information about the procedure used to add firms to the sample.

Table 4. Effects of Q and Cash Flow on Investment, Various Periods, 1970–84^a

Independent variable and summary statistic	Class 1	Class 2	Class 3
<i>1970–75</i>			
Q_{it}	–0.0010 (0.0004)	0.0072 (0.0017)	0.0014 (0.0004)
$(CF/K)_{it}$	<u>0.670</u> (0.044)	0.349 (0.075)	<u>0.254</u> (0.022)
\bar{R}^2	0.55	0.19	0.13
<i>1970–79</i>			
Q_{it}	0.0002 (0.0004)	0.0060 (0.0011)	0.0020 (0.0003)
$(CF/K)_{it}$	<u>0.540</u> (0.036)	0.313 (0.054)	<u>0.185</u> (0.013)
\bar{R}^2	0.47	0.20	0.14
<i>1970–84</i>			
Q_{it}	0.0008 (0.0004)	0.0046 (0.0009)	0.0020 (0.0003)
$(CF/K)_{it}$	<u>0.461</u> (0.027)	0.363 (0.039)	<u>0.230</u> (0.010)
\bar{R}^2	0.46	0.28	0.19

Source: Authors' estimates of equation 3 based on a sample of firm data from Value Line data base. See text and Appendix B.

a. The dependent variable is the investment-capital ratio $(I/K)_{it}$, where I is investment in plant and equipment and K is beginning-of-period capital stock. Independent variables are defined as follows: Q is the sum of the value of equity and debt less the value of inventories, divided by the replacement cost of the capital stock adjusted for corporate and personal taxes (see Appendix B); $(CF/K)_{it}$ is the cash flow–capital ratio. The equations were estimated using fixed firm and year effects (not reported). Standard errors appear in parentheses.

information between firms and outside investors can be made for the shorter time periods, 1970–79 and particularly 1970–75.

The results in table 4 show large estimated cash flow coefficients for firms in class 1. As expected, the cash flow coefficient is largest (0.670) in the earliest period, when most of these firms had yet to be recognized by Value Line. The coefficient is the smallest (0.461) for 1970–84. Furthermore, as the sample period is extended one year at a time from 1970–75 to 1970–84, the estimated cash flow coefficients for these firms decline monotonically.³⁶ The cash flow coefficients in classes 2 and 3 are

36. The coefficients for the periods 1970–75 through 1970–84 are: 0.670, 0.571, 0.566, 0.554, 0.540, 0.520, 0.510, 0.494, 0.481, and 0.461. The corresponding coefficients for firms in the third class are: 0.254, 0.176, 0.160, 0.173, 0.185, 0.204, 0.217, 0.221, 0.230, and 0.230. The coefficients of firms in class 2 always fall in the middle.

positive and approximately stable over time. That the cash flow coefficient is different from zero even for the mature firms in class 3 is not surprising given the limitations of the Q model.³⁷

It is the difference in the estimated coefficients across classes that we stress. These differences range from 0.416 for 1970–75 to 0.231 for 1970–84, the smallest difference for any period in our sample. These differences are always statistically significant at very high confidence levels. The t -statistic under the null hypothesis that the class 1 cash flow coefficient equals the class 3 coefficient is 12.1 for the 1970–84 sample period, in which the difference is the smallest. That the difference between the classes narrows as the time period is extended is expected; asymmetric information should not cause permanent differences in investment behavior for ultimately successful firms like the ones in our sample. On the other hand, that the differences remain substantial for so long indicates that the phenomenon is quite persistent.³⁸

The model explains a greater proportion of the variance of I/K in class 1 as a result of the inclusion of cash flow. In class 1, 46 percent to 55 percent of the variance in I/K is explained, depending on the time period analyzed, primarily due to the variation in cash flow alone. The first column of table 5 presents the Q model estimated without cash flow. Adding cash flow increases the R^2 by 0.23 for class 1, 0.11 for class 2, and only 0.08 for class 3, confirming the greater statistical importance of cash flow for firms in the first class.

Furthermore, the economic significance of these results is reinforced by the high variability of cash flow in the first class. Investment is two to three times more sensitive to cash flow fluctuations in this class than it is in the third, while the underlying variations in cash flow for the first

37. To the extent that firms are experiencing tax losses or are unable to take full advantage of investment incentives, our tax-adjusted Q is mismeasured. A positive coefficient on cash flow in the estimated investment equation could reflect to some extent this mismeasurement of Q . Moreover, because of the greater volatility of earnings in firms in class 1, such firms may be more likely to experience the problem. We did not adjust the tax measures for each firm, but reestimating the models reported in table 4 using unadjusted q produced virtually identical estimated effects of internal cash flow on investment.

38. Because the firms in the first two classes are smaller on average than those in the third, one might expect that these results reflect differences due to size rather than retention practices. But the third class contains many small firms as well. When the sample is split into thirds by firm size, as measured by average capital stock, small firms have relatively low cash flow coefficients.

Table 5. Effects of Q and Cash Flow on Investment: Consideration of Measurement Error, 1970–84^a

Independent variable and summary statistic	Ordinary least squares ^b	Ordinary least squares ^b with (CF/K)	Instrumental variable ^{b,c}	First difference ^d	Second difference ^e
Class 1					
Q_{it}	0.0045 (0.0004)	0.0008 (0.0004)	0.0065 (0.0009)	−0.0021 (0.0006)	−0.0040 (0.0010)
$(CF/K)_{it}$. . .	0.464 (0.027)	0.455 (0.029)	0.496 (0.034)	0.457 (0.040)
\bar{R}^2	0.23	0.46	0.53	0.25	0.22
Class 2					
Q_{it}	0.0073 (0.0009)	0.0046 (0.0009)	0.0035 (0.0011)	0.0106 (0.0015)	0.0090 (0.0019)
$(CF/K)_{it}$. . .	0.363 (0.039)	0.418 (0.038)	0.268 (0.046)	0.364 (0.054)
\bar{R}^2	0.17	0.28	0.28	0.14	0.13
Class 3					
Q_{it}	0.0044 (0.0002)	0.0020 (0.0003)	0.0024 (0.0004)	0.0032 (0.0004)	0.0036 (0.0005)
$(CF/K)_{it}$. . .	0.230 (0.010)	0.238 (0.010)	0.223 (0.013)	0.228 (0.014)
\bar{R}^2	0.11	0.19	0.19	0.08	0.07

Source: Same as table 4.
a. Dependent variable is the investment-capital ratio $(I/K)_{it}$. All variables are as defined in table 4, note a. Standard errors appear in parentheses.
b. Estimated using fixed firm and year effects.
c. The instrumental variable procedure uses lagged Q as an instrument for Q .
d. All variables expressed as first differences.
e. All variables expressed as second differences.

class are more than three times larger than those in the third class, measured by the standard deviation of CF/K reported in table 2.

ALTERNATIVE ESTIMATION METHODS AND SPECIFICATIONS FOR THE Q MODEL

In this section, we examine the robustness of the results presented to this point for the Q model with respect to changes in estimation technique and specification. There are at least two problems in measuring Q that might affect the econometric results for cash flow. First, to the extent the stock market is excessively volatile, Q may not reflect market fundamentals. Second, the replacement capital stock in Q may be measured with error. The results of tests to deal with these problems are

ME reported in table 5. First, using lagged Q as an instrument for Q , we obtained similar coefficients on the Q and cash flow terms.³⁹ Second, we estimated the model using first differences and second differences (as opposed to the conventional fixed-effects, within-group estimator) to address measurement-error problems; coefficient estimates on cash flow are similar in all cases.⁴⁰

Across all the tests reported in table 5, differences between the class 1 and class 3 cash flow coefficients range between 0.217 and 0.273. This range is consistent with the difference of 0.231 estimated with the basic Q model over 1970–84. If these tests are run on earlier time periods, the estimated difference in the cash flow effects across classes rises, but the differences remain remarkably consistent across different estimation techniques for a given period. The differences between classes 1 and 3 for 1970–79, for example, range between 0.331 and 0.355.

Table 6 reports estimates of alternative specifications to analyze further the robustness of the difference in cash flow effects in different retention classes. Results are reported both for 1970–79 and for 1970–84. Some rejections of the strongest versions of the q theory result from a significant effect of lagged Q in explaining investment. The second model presented in table 6 includes lagged Q . In the third class, lagged Q does have a statistically significant estimated coefficient, and the coefficient on the current Q variable becomes positive in the first class when lagged Q is included. The pattern of cash flow coefficients across classes for both time periods, which is the result of primary interest here, is virtually identical when lagged Q is included in the equation.⁴¹

We also report the effect of including additional lags of cash flow in table 6. Lagged values of cash flow may have explanatory power for

39. This finding also addressed the concern of Fumio Hayashi and Tohru Inoue that disturbances in the cost of adjustment function are incorporated into the beginning-of-period Q , making Q endogenous. Fumio Hayashi and Tohru Inoue, "Implementing the Q Theory of Investment in Micro Data: Japanese Manufacturing, 1977–1985" (Osaka University, June 1987).

40. Zvi Griliches and Jerry Hausman argue that measurement error will lead to different biases across potential estimators that are similar in that they control for firm-specific effects, but differ in their signal-to-noise ratios, making it possible to place bounds on the importance of measurement error. Zvi Griliches and Jerry A. Hausman, "Errors in Variables in Panel Data," *Journal of Econometrics*, vol. 31 (February 1986), pp. 93–118.

41. We also considered the possibility that the adjustment cost function was nonlinear by adding Q^2 to the equations. This change did not materially affect the cash flow coefficient pattern.

Table 6. Effects of Q and Cash Flow on Investment: Alternative Specifications, Various Periods, 1970–84^a

Independent variable and summary statistic	Class 1		Class 2		Class 3	
	1970–79	1970–84	1970–79	1970–84	1970–79	1970–84
<i>Model with additional cash flow lags</i>						
Q_{it}	–0.0002 (0.0004)	0.0007 (0.0004)	0.0059 (0.0011)	0.0044 (0.0009)	0.0011 (0.0003)	0.0011 (0.0003)
$(CF/K)_{it}$	<u>0.508</u> (0.035)	0.400 (0.029)	0.245 (0.059)	0.304 (0.045)	<u>0.146</u> (0.015)	0.168 (0.012)
$(CF/K)_{i,t-1}$	0.216 (0.045)	0.167 (0.039)	0.100 (0.062)	0.095 (0.053)	0.092 (0.021)	0.116 (0.018)
$(CF/K)_{i,t-2}$	0.179 (0.043)	0.115 (0.037)	0.132 (0.063)	0.073 (0.052)	0.116 (0.020)	0.074 (0.017)
\bar{R}^2	0.54	0.49	0.23	0.30	0.16	0.21
<i>Model including lagged Q</i>						
Q_{it}	0.0037 * (0.0015)	0.0033 * (0.0013)	0.0064 (0.0016) *	0.0052 (0.0014) *	0.0014 * (0.0004)	0.0015 * (0.0004)
$Q_{i,t-1}$	0.0011 * (0.0006)	0.0015 * (0.0006)	0.0004 (0.0015)	–0.0002 (0.0013)	0.0011 * (0.0004)	0.0008 * (0.0003)
$(CF/K)_{it}$	<u>0.528</u> (0.041)	0.426 (0.030)	0.287 (0.059)	0.345 (0.041)	<u>0.183</u> (0.014)	0.225 (0.010)
\bar{R}^2	0.58	0.53	0.22	0.29	0.14	0.20

Source: Same as table 4.

a. Dependent variable is the investment-capital ratio $(I/K)_{it}$. All variables are as defined in table 4, note a. Equations are estimated with fixed firm and year effects (not reported). Standard errors appear in parentheses.

investment in a time-to-build context, for example. Collinearity among the cash flow variables reduces the current cash flow coefficient in all classes when additional lags are included, but the pattern across classes remains clear. Indeed, the differences between the current cash flow coefficients in the classes 1 and 3 are almost identical to the differences between the current cash flow coefficients in table 4. The differences in the sums of the cash flow coefficients between the first and third classes rise substantially when more lags are added.⁴² Also the current cash flow coefficient relative to the lagged coefficients is much larger for class 1 than for class 3. To the extent that the difference in the cash flow effects across classes reflects the impact of financial constraints on investment,

42. The t -statistic for the null hypothesis that the sum of the cash flow coefficients is equal across the first and third classes is 10.6. When a third lag of cash flow was included in the equation, its coefficient was not significantly different from zero at the 10 percent level in any of the classes.

one would expect the difference to be most evident in the coefficient on current cash flow, especially because these data are annual.⁴³ The effects of the lagged coefficients may well reflect shortcomings in the empirical performance of Q . That the estimated coefficient on Q for the mature firms in class 3 is only half as large when longer lags on cash flow are included supports this interpretation.

L.Q. A different interpretation of the effect of cash flow on investment is that movements in cash flow reflect productivity shocks not captured in the beginning-of-period Q (that is, cash flow may be correlated with the disturbance in the adjustment cost function). To explain our results, one would have to account for the different effect of productivity shocks in firms grouped only by their retention behavior. From a broader perspective, it is also possible that current cash flow contains “news” about investment opportunities not captured in the beginning-of-period Q . To address these points, we reestimated the basic Q model in two ways, first treating CF/K as endogenous and using instrumental variables techniques and then adding Q dated at the end of the current period—that is, incorporating all news arriving in the current period—to the ordinary least squares model. With both alterations, the differences in the estimated cash flow coefficients across classes remained.⁴⁴

✓ In summary, the results presented here suggest important effects of fluctuations in the availability of internal finance on investment. Internal funds help explain investment in all classes, even for firms that have much more cash flow than investment. Most likely, that finding indicates the pitfalls in using average Q in empirical studies. For our purposes, however, the fundamental finding is the substantial difference across classes in the effect of cash flow on investment. Several possible issues involving measurement error have been addressed by instrumenting Q and estimating the basic model with first and second differences. We have also considered several alternative specifications, including lagged Q , additional lags of cash flow, and treating cash flow as endogenous.

43. Abel and Blanchard consider three quarterly lags of profits in a q model estimated from aggregate data for the manufacturing sector. This time period falls within our contemporaneous annual observation. Abel and Blanchard found only the coefficient on the first lag of profits to be statistically significantly different from zero. Abel and Blanchard, “The Present Value of Profits and Investment.”

44. The difference in the effect of cash flow across classes generally widened when current cash flow was instrumented with lagged variables. This result also suggests that the possible dependence of current cash flow on current investment is not responsible for the observed pattern of cash flow coefficients.

In all these models, the estimated difference in cash flow effects in the different retention classes is always statistically significant at very high confidence levels. Furthermore, the estimated differentials are larger over shorter periods when the firms in class 1 are less mature and probably face more severe asymmetric information problems. The results over shorter periods are also remarkably consistent across the various models and estimation techniques. For example, the differential for 1970–79 between the estimated cash flow coefficients for classes 1 and 3 was between 0.33 and 0.38 over all the tests reported in tables 4 through 6. The range for 1970–75 was 0.36 to 0.42. These results are consistent with the cost differential between internal and external finance predicted by the models described earlier and with the differences in the q values we found across classes. The economic importance of these findings is magnified by the fact that cash flow is highly variable for the rapidly growing firms in the first class, while mature firms in the third class experience much less variation in cash flow.

Because the firms we examine, even the rapidly growing firms in class 1, are large manufacturing corporations by economywide standards, the significance of internal finance for capital spending may well be greater for smaller companies, which may have more difficult, or no, access to centralized securities markets.

SALES ACCELERATOR INVESTMENT DEMAND MODELS

From a theoretical standpoint, the Q investment demand model has many attractive features. In practice, however, other approaches have performed better empirically. Some of the most successful empirical investment models are based on the traditional acceleration principle, which links the demand for capital goods to the level or change in a firm's output or sales.⁴⁵ Below we test whether the pattern of cash flow effects across retention classes holds up in models that include sales. Certainly one possible explanation for the effect of the cash flow variables in all the retention classes is that internal finance is correlated with sales.

45. Traditional accelerator models are based on the change in sales rather than its level. For a given number of lags, this approach imposes one restriction on the estimated coefficients. In a recent paper, Abel and Blanchard present an accelerator model that includes delivery and installation lags. In this more general approach, estimating the model with levels of sales is appropriate. Andrew Abel and Olivier Blanchard, "Investment and Sales: Some Empirical Evidence," Working Paper 2050 (NBER, October 1986).

Table 7 presents estimated equations for the three retention classes that include cash flow and current and lagged values of sales. Two equations are reported, one that includes only sales variables augmented by cash flow and one that adds Q . Most of the sales terms are statistically significant individually, and they are highly significant jointly. Moreover, some of the cash flow effects in the Q model can indeed be explained by the correlation of cash flow and sales; the cash flow coefficients decline in all three classes when the sales variables are added to the equation. This may indicate discrepancies between average and marginal Q or accelerator effects. The pattern of the cash flow coefficients across classes, however, remains about the same as in the models without sales. The results in table 7 are for the full 1970–84 period; greater estimated differences in the cash flow coefficients arise for shorter sample periods. The different effects of cash flow between the classes 1 and 3 for all sample periods are similar to the results obtained from the Q model without sales. These results show that including sales variables does not change the primary result presented above.

The results for the equation that includes Q also provide an interesting perspective on a point often raised in the investment literature. It is typical to find significant effects of both sales and profits or cash flow in an investment equation. In that case, however, the question remains whether the cash flow variable should be interpreted as a signal of the profitability of investment not captured in the simple accelerator formulation, or whether the significance of cash flow arises because it represents an additional supply of low-cost investment finance for firms that must pay a premium for external funds.

以下
的双重
作用

Including Q in the estimated equation helps to resolve this question. Because Q is based on asset prices determined in forward-looking markets, it should capture the prospective profitability of investment better than lags of past profits. The results show that including Q reduces the cash flow effect somewhat in classes 2 and 3, but cash flow still has a strong effect in all the dividend-payout classes. To the extent that Q captures the effect of future profitability on the demand for investment, this result supports the financing constraint interpretation. Again, that the cash flow effect remains significant in the class of high-payout firms suggests caution in this regard. The *difference* in cash flow effects across classes remains the strongest evidence supporting the finance constraint view.

Table 7. Effects of Sales and Cash Flow on Investment, 1970–84^a

Independent variable and summary statistic	Class 1	Class 2	Class 3
<i>Model with sales-capital ratio</i>			
$(CF/K)_{it}$	0.277 (0.033)	0.256 (0.047)	0.120 (0.013)
$(S/K)_{it}$	0.041 (0.007)	0.045 (0.009)	0.027 (0.002)
$(S/K)_{i,t-1}$	-0.015 (0.011)	-0.016 (0.011)	-0.001 (0.003)
$(S/K)_{i,t-2}$	0.031 (0.012)	0.015 (0.011)	0.008 (0.003)
$(S/K)_{i,t-3}$	-0.036 (0.009)	-0.020 (0.008)	-0.010 (0.003)
\bar{R}^2	0.54	0.30	0.23
<i>Model with sales-capital ratio and Q</i>			
Q_{it}	-0.0004 (0.0004)	0.0049 (0.0009)	0.0019 (0.0003)
$(CF/K)_{it}$	0.286 (0.035)	0.178 (0.047)	0.086 (0.013)
$(S/K)_{it}$	0.042 (0.007)	0.047 (0.009)	0.029 (0.002)
$(S/K)_{i,t-1}$	-0.013 (0.011)	-0.021 (0.011)	-0.003 (0.003)
$(S/K)_{i,t-2}$	0.029 (0.012)	0.015 (0.011)	0.008 (0.003)
$(S/K)_{i,t-3}$	-0.036 (0.009)	-0.012 (0.008)	-0.009 (0.003)
\bar{R}^2	0.54	0.34	0.24

Source: Authors' calculations based on a sample of firm data from Value Line data base. See text description and Appendix B.

a. The dependent variable is the investment-capital ratio $(I/K)_{it}$ defined as in table 4, note a. Q and $(CF/K)_{it}$ are also as defined in table 4, note a. $(S/K)_{it}$ is the ratio of sales, S , to the beginning-of-period capital stock. All equations were estimated with fixed time and firm effects (not reported). Standard errors appear in parentheses.

INTERNAL FINANCE IN THE NEOCLASSICAL INVESTMENT MODEL

A common criticism of the sales accelerator model is that it does not incorporate the relative price of capital or capital services in the empirical specification. This issue is addressed by the neoclassical investment

model pioneered by Jorgenson.⁴⁶ In its most general form, the neoclassical model is derived from the solution to a dynamic factor demand problem that determines the firm's optimal level of capital services through time. The change in the demand for capital services along with the depreciation of existing capital determines investment.

With perfectly competitive input and output markets, the firm's optimal demand for capital services depends ultimately on the price of output and the relative prices of various inputs, including the cost of capital. To simplify the empirical specification, however, Jorgenson used a transformation of the reduced form of the optimal demand for capital based on a Cobb-Douglas production function. The transformation allows the demand for capital to be expressed as a function of the relative cost of capital services alone; the effect of other factor prices is captured by including the level of output or sales in the model. In this case, the neoclassical model with partial-adjustment assumptions takes a form similar to the accelerator model, except that the sales or output term is modified by a cost of capital measure. If firms have Cobb-Douglas production functions, the desired capital stock is proportional to the ratio of sales to the tax-adjusted relative price of capital.⁴⁷ This variable is denoted by J in table 8.

The first equation in table 8 includes the cost of capital and cash flow variables. Again, the pattern of coefficients across the retention classes shows that cash flow has a substantially higher effect for firms that pay low dividends than for mature, high-payout firms. The neoclassical model is subject to the same criticism that is raised against the accelerator model: the equation is specified with backward-looking variables. However, adding tax-adjusted Q to the equation, as we do in the second equation reported in table 8, does not change the results substantially.

Though not reported here, we have also estimated the investment equations outlined before with instrumental variables for Q , cash flow, and sales to attempt to correct for "news" in cash flow and measurement error problems. The results depend on the specific instruments used,

46. For a survey of much of the relevant literature, see Dale W. Jorgenson, "Econometric Studies of Investment Behavior: A Survey," *Journal of Economic Literature*, vol. 9 (December 1971), pp. 1111-47; and Clark, "Investment in the 1970s."

47. The general form of the tax adjustments to cost of capital we use in the empirical work presented here is based on the original development by Hall and Jorgenson, "Tax Policy." The cost of capital definition is presented in Appendix B.

Table 8. Effects of Cost of Capital and Cash Flow on Investment, 1970–84^a

<i>Independent variable and summary statistic</i>	<i>Class 1</i>	<i>Class 2</i>	<i>Class 3</i>
<i>Model with adjusted sales–cost of capital ratio</i>			
$(CF/K)_{it}$	0.337 (0.029)	0.331 (0.043)	0.199 (0.011)
$(J/K)_{it}$	0.273 (0.043)	0.177 (0.039)	0.081 (0.009)
$(J/K)_{i,t-1}$	–0.100 (0.072)	–0.070 (0.055)	–0.023 (0.012)
$(J/K)_{i,t-2}$	0.152 (0.079)	0.046 (0.057)	0.025 (0.013)
$(J/K)_{i,t-3}$	–0.123 (0.060)	–0.069 (0.044)	0.002 (0.010)
\bar{R}^2	0.52	0.28	0.20
<i>Model with adjusted sales–cost of capital ratio and Q</i>			
Q_{it}	0.0005 (0.0004)	0.0050 (0.0009)	0.0020 (0.0003)
$(CF/K)_{it}$	0.319 (0.033)	0.248 (0.044)	0.163 (0.011)
$(J/K)_{it}$	0.275 (0.043)	0.190 (0.038)	0.086 (0.009)
$(J/K)_{i,t-1}$	–0.114 (0.073)	–0.090 (0.053)	–0.030 (0.012)
$(J/K)_{i,t-2}$	0.158 (0.079)	0.051 (0.055)	0.026 (0.012)
$(J/K)_{i,t-3}$	–0.125 (0.060)	–0.037 (0.043)	0.003 (0.010)
\bar{R}^2	0.53	0.32	0.21

Source: Same as table 7.

a. The equations are as specified in table 7 except that the sales term used in table 7 is modified by a cost of capital measure (see text). The variable, defined J , enters the equations above as a ratio to the capital stock at the beginning of the period, K . All equations were estimated with fixed time and firm effects (not reported). Standard errors appear in parentheses.

but several general features of the estimates are clear. First, the pattern of declining cash flow coefficients as one moves to the higher payout classes remains. The differential between classes 1 and 3 is generally at least as large as in the reported results. Second, the cash flow effects in class 3 remain as large as or larger than in the OLS/fixed-effect equations. Therefore, no simple correction for measurement error resolves the puzzle of why cash flow has a persistent effect for mature firms in each of the alternative specifications of investment demand we examined.

Regardless of the conclusion reached about the source of cash flow effects in mature firms, however, the difference in the cash flow effects reported here establishes that firm heterogeneity is an important aspect of the link between finance and real investment.

INVESTMENT EQUATIONS AT THE INDUSTRY LEVEL

Another dimension of firm heterogeneity that may be important for investment behavior is differences across *industry* categories. Table 9 provides estimates of the basic Q model augmented with cash flow by retention class for several two-digit Standard Industrial Classification (SIC) code manufacturing industry categories. The results reported are robust to the alternative investment demand specifications reviewed before. The number of observations in classes 1 and 2 is small in the separate industry categories. We have reported estimates for these two classes combined for individual two-digit industries that have at least five firms in the combined class. For comparison, we also report the estimated coefficients for the model from a sample that combines the remaining two-digit industries.

In six out of the seven cases, the cash flow effect is larger for the high-retention classes than for the more mature firms in class 3. That the effect of cash flow on investment is greatest for low-payout firms, with industry effects held constant, casts further doubt on a productivity shock interpretation of the differential effect. Because of the small samples, the differentials vary substantially. The one case (chemicals, industry 28) in which the cash flow coefficient for the third class is higher than that for the first two classes has only two firms from the first class, the lowest number for any industry group. These results indicate that greater sensitivity of investment to cash flow in high-retention firms is not a phenomenon restricted to particular industries. The high-technology computer firms in industry 36 have a high differential, for example, but the differential in the food-processing firms in industry 20 is even greater.

BALANCE SHEETS, INTERNAL FINANCE, AND INVESTMENT

The results presented to this point have examined how changes in the *flow* of internal funds affect investment spending in different kinds of

Table 9. Effects of Q and Cash Flow on Investment, Various Industries, 1970–84^a

Industry	Standard Industrial Classification code	Classes 1 and 2			Class 3		
		Q	CF/K	\bar{R}^2	Q	CF/K	\bar{R}^2
Food	20	-0.003 (0.008)	<u>0.613</u> (0.135)	0.19	0.007 (0.002)	<u>0.247</u> (0.054)	0.14
Chemicals	28	0.006 (0.001)	✓ 0.190 (0.068)	0.36	-0.001 (0.001)	✓ 0.413 (0.036)	0.28
Machinery, except electrical	35	0.000 (0.001)	0.545 (0.041)	0.59	0.014 (0.002)	0.280 (0.039)	0.42
Electrical and electronic machinery	36	0.002 (0.001)	[0.293 (0.045)	0.21	0.000 (0.001)	[0.207 (0.022)	0.27
Transportation	37	0.008 (0.002)	0.401 (0.053)	0.62	0.019 (0.003)	0.161 (0.054)	0.27
Measuring instruments	38	0.006 (0.002)	0.457 (0.108)	0.29	0.003 (0.001)	0.349 (0.047)	0.47
All others		0.011 (0.003)	0.394 (0.056)	0.34	0.003 (0.001)	0.191 (0.017)	0.14

Source: Same as table 4.

a. For each industry the equations are exactly the same as the equations in table 4, except that the firms in classes 1 and 2 are aggregated. All equations were estimated with fixed firm and year effects (not reported). Standard errors are in parentheses.

firms. Of course, *stock* measures of a firm's internal liquidity might also have an effect on investment for firms that face high costs of external funds due to information problems in capital markets. Cash and marketable securities provide a low-cost source of investment finance for firms that must pay a premium for external funds. To the extent that such firms have accumulated liquid resources, they have a financial cushion that may reduce the sensitivity of their investment to cash flow fluctuations. Therefore, one might expect to observe a positive effect of stock measures of liquidity for the high-retention firms, whose investment is especially sensitive to fluctuations in cash flow.

The motivation for this test is analogous to considerations of precautionary saving. If managers know that they will have to pay a premium for external funds, they should accumulate a stock of liquid assets when cash flow is high. That stock of liquid assets will help smooth investment over downturns and spare firms the need to obtain potentially costly capital from external sources. It might also provide the necessary collateral to obtain new debt as suggested by some of the models considered earlier. Finally, as discussed, debt finance may entail cove-

Table 10. Effect of Balance Sheet Variables on Investment, 1970–84^a

<i>Independent variable and summary statistic</i>	<i>Class 1</i>	<i>Class 2</i>	<i>Class 3</i>
<i>Model including cash and equivalents variable</i>			
Q_{it}	0.0001 (0.0004)	0.0045 (0.0009)	0.0019 (0.0003)
$(CF/K)_{it}$	0.372 (0.027)	0.348 (0.039)	0.224 (0.011)
$(CASH/K)_{it}$	0.112 (0.011)	0.052 (0.020)	0.010 (0.007)
\bar{R}^2	0.53	0.30	0.19
<i>Model including working capital</i>			
Q_{it}	0.0003 (0.0004)	0.0043 (0.0009)	0.0021 (0.0003)
$(CF/K)_{it}$	0.365 (0.030)	0.351 (0.039)	0.230 (0.010)
$(WCMI/K)_{it}$	0.077 (0.011)	0.021 (0.015)	-0.011 (0.006)
\bar{R}^2	0.51	0.29	0.19
<i>Model including current and lagged values of cash and sales</i>			
Q_{it}	-0.0005 (0.0004)	0.0042 (0.0009)	0.0012 (0.0003)
$(CASH/K)_{it}$	0.099 (0.011)	0.058 (0.020)	0.000 (0.008)
$(CF/K)_{it}$	0.163 (0.036)	0.119 (0.054)	-0.005 (0.016)
$(CF/K)_{i,t-1}$	0.168 (0.044)	0.089 (0.061)	0.153 (0.022)
$(CF/K)_{i,t-2}$	0.071 (0.047)	0.002 (0.059)	0.091 (0.020)
$(S/K)_{it}$	0.044 (0.007)	0.053 (0.009)	0.038 (0.003)
$(S/K)_{i,t-1}$	-0.035 (0.012)	-0.032 (0.012)	-0.017 (0.004)
$(S/K)_{i,t-2}$	0.026 (0.014)	0.018 (0.012)	0.001 (0.004)
$(S/K)_{i,t-3}$	-0.020 (0.010)	-0.015 (0.009)	-0.005 (0.003)
\bar{R}^2	0.60	0.35	0.26

Source: Same as table 7.

a. The dependent variable is the investment-capital ratio $(I/K)_{it}$, where I is investment in plant and equipment and K is beginning-of-period capital stock. Q_{it} is the sum of the value of equity and debt less the value of inventories, divided by the replacement cost of the capital stock adjusted for corporate and personal taxes (see Appendix B); $(CF/K)_{it}$ is the cash flow-capital ratio; $(S/K)_{it}$ is the ratio of sales to capital; $CASH$ is cash on hand plus liquid securities; and $WCMI$ is working capital less the book value of inventories. Standard errors appear in parentheses.

nants and restrictions that constrain firms' ability to use stocks of liquidity. Thus, when financially constrained firms experience increased liquidity, they may be able to finance increased investment.

On the other hand, mature firms that pay a substantial portion of their income as dividends are unlikely to derive any particular benefit for investment from higher stocks of liquid assets. If retained earnings fall below the level necessary to finance desired investment in these firms, they could reduce dividends, or, if managers perceive dividend cuts as negative signals to the market, they could likely obtain relatively low-cost funds from external capital markets. Therefore, one would expect little estimated significance for stock measures of liquidity in the investment of the high-dividend firms in our third class.

Table 10 reports the results of including stock liquidity measures in an augmented Q investment equation similar to the equations presented earlier. We used two alternative liquidity stock variables—cash and equivalents (defined as cash on hand plus securities readily convertible into cash), *CASH*, and working capital less the book value of inventories, *WCMI*, where working capital is defined as current assets minus current liabilities. Both variables were measured at the beginning of the period and were deflated by the firm's capital stock. The results clearly support the view that changes in balance sheet positions and liquidity have a significant effect on investment for the low-payout firms. On the other hand, the estimated coefficients on the liquidity variables are not statistically different from zero for the mature firms. The results for the firms in class 2 fall in the middle. These results are also remarkably robust in equations that include sales accelerator variables (not reported here). As discussed, the cash flow coefficients drop for all the classes when lags of sales are included. The coefficients on the stock liquidity variables, however, are virtually identical in models that include sales. Similar results were obtained when we included current assets alone or working capital alone without subtracting inventories.

It is not especially surprising that the results across classes are so strong for the liquidity variables from the balance sheet. Cash flow is closely correlated with profits, and to the extent that there are problems with the Q model or other investment demand specifications, one would expect cash flow to enter an investment equation positively, even for mature, high-dividend firms that are unlikely to face important cost disadvantages of external funds. On the other hand, stock measures of

liquidity are less likely to indicate much about profitability of new investment. The evidence supports the hypothesis that these variables have no important effect for firms like the ones in our class 3 sample. For firms in classes 1 and 2, however, the results using balance sheet variables present strong evidence of the imperfect substitutability of internal and external finance at the margin.

We have examined the robustness of these results to alternative specifications. Because we have only current cash flow in the reported regressions, the estimated liquidity effects may be proxies for longer lags of cash flow, or they could capture accelerator effects of sales. To test this possibility, we included current and three lagged values of cash flow and sales in the model. The results are reported in the last half of table 10. The effects of these additional variables were statistically significant, but the pattern of estimated coefficients for the cash-and-equivalents and working-capital-less-inventories variables are virtually identical to the patterns found in the models without sales or lags of cash flow.

INTERNAL FINANCE AND INVESTMENT IN HIGH-PAYOUT FIRMS

In some specifications of the investment models presented here, the estimated coefficient on cash flow is both statistically significant and economically important for the high-payout firms in class 3. This finding was quite robust. We tried further splits of those firms based on the level of payout rates over the sample. We also divided those firms into groups based on dividend growth, rather than levels, to test the hypothesis that investment of firms that increase their dividends would be less sensitive to cash flow than firms that paid stable or falling dividends. The estimated cash flow coefficients for these subgroups were roughly the same as the estimated coefficients from the full class 3 sample.

Because class 3 firms pay substantial dividends, such findings may seem inconsistent with our emphasis on the imperfect substitutability of internal and external finance. That is, if external funds are more costly than internal finance, why would these firms not cut dividends rather than investment when cash flow falls? One explanation is that agency costs of internal finance (that is, potential “managerial waste” on less productive investments) account for this link between cash flow and

investment in mature firms.⁴⁸ While these agency problems may be important, they do not seem to explain the entire cash flow effect for class 3 firms. The class 3 cash flow effect is small when sales variables are included, suggesting that the apparent correlation between cash flow and investment in mature firms may be due to the omission of output terms important in reconciling the difference between marginal and average Q . Nor is there any measured effect of beginning-of-period stocks of liquidity on investment in these firms.

Furthermore, evidence of “sticky” dividends suggests that, in the presence of even small cost differentials between internal and external finance, investment may be sensitive to internal finance for mature firms with substantial payout.⁴⁹ If these firms are reluctant to cut dividends when cash flow falls, maybe for signaling reasons, they may reduce investment somewhat rather than seek more costly external finance. This kind of behavior would, of course, magnify the importance of financial constraints for macroeconomic fluctuations in investment, a possibility that should be considered in more depth in future research.

Conclusions and Applications

Our results show that financial factors affect investment. Our approach emphasizes that the link between financing constraints and investment varies by type of firm. Recent literature on asymmetric information and capital market imperfections demonstrates that a firm’s opportunity cost of internal funds can be substantially lower than its cost of external finance. Under these circumstances, the investment of firms that exhaust nearly all of their low-cost internal funds should be more sensitive to fluctuations in their cash flow than that of firms that pay high dividends. Also, liquidity should have a greater effect on investment for low-dividend firms than for high-dividend firms.

48. For a discussion of the agency costs associated with “free cash flows” in the petroleum industry, see Michael C. Jensen, “Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers,” *American Economic Review*, vol. 76 (May 1986, *Papers and Proceedings*, 1985), pp. 323–29.

49. See, for example, the review of studies presented in James M. Poterba, “Tax Policy and Corporate Saving,” *BPEA*, 2:1987, pp. 455–503.

To test these hypotheses, we estimated investment functions across groups of firms classified by their dividend behavior. Financial effects were generally important for investment in all firms. But the results consistently indicated a substantially greater sensitivity of investment to cash flow and liquidity in firms that retain nearly all of their income. This statistically and economically significant difference was robust to a wide variety of model specifications and estimation techniques. It was largest for sample periods in which the low-dividend firms were the youngest and had yet to be recognized by major financial data services. These empirically important differences across firms are consistent with financial constraints arising from capital market imperfections. The results also cast doubt on the longstanding interpretation of empirical financial effects on investment as proxies for misspecified “real” influences.

If capital market imperfections lead to binding financial constraints on investment, several important implications arise for the study of macroeconomic investment fluctuations and the impact of public policy on capital spending. We consider these points briefly, as well as some suggested directions for future research, in the remainder of the paper.

INTERNAL FINANCE, INVESTMENT, AND ECONOMIC FLUCTUATIONS

Financial constraints in capital markets can magnify the macroeconomic effect of shocks to cash flow or liquidity that reduce some firms' access to low-cost finance and worsen their balance sheet positions. To examine this issue more closely, we consider the extent to which internal finance effects on investment can account for the variability of aggregate investment. Since 1970, the standard deviation of the ratio of nonresidential gross investment to the replacement value of the stock of plant and equipment has been 0.87 percent (with a mean value of 12.46 percent). How much of this variance can be explained by our estimated effect of changes in cash flow in investment?

From the investment model estimated from the full sample with Q , current cash flow, and lags of sales, the cash flow coefficients for the dividend classes 1 through 3 are 0.309, 0.167, and 0.085, respectively. We make the conservative assumptions that the effect of cash flow for the mature, high-payout firms in class 3 is not related to finance

constraints, and that the portion of the class 1 and 2 coefficients equal to the class 3 coefficient should be attributed to effects other than finance constraints. Then the net cash flow effects for classes 1 and 2 are 0.224 and 0.082, respectively. The predicted changes in the investment-capital ratio resulting from a one standard deviation change in the cash flow to capital ratios are 4.48 and 0.74 percentage points for classes 1 and 2, respectively.

The aggregate investment-capital ratio can be expressed as a weighted average of the ratio for each class, with weights equal to the proportion of the aggregate capital stock in each class. To predict the effect of cash flow changes for firms like those in classes 1 and 2 for aggregate investment fluctuations, therefore, one needs to estimate the proportion of the aggregate capital stock in similar firms. We begin very conservatively by assuming that the aggregate proportions are the same as our Value Line sample proportions. Then, one standard deviation changes in the class 1 and 2 cash-flow-to-capital ratios explain about 13 percent of the standard deviation in the aggregate investment-capital ratio.

This result, however, almost certainly understates the true effect because large, mature firms constitute a greater proportion of our Value Line sample than they do of the aggregate economy. Indeed, data for our sample period from the *Quarterly Financial Reports* of the U.S. Department of Commerce indicate that approximately 20 percent of aggregate assets are held by firms with total assets less than \$100 million. The median capital stock figure for our Value Line firms in class 1, certainly less than their total assets, was \$95 million in 1984. Class 2 firms had a median capital stock of \$193 million. These statistics imply that the *aggregate* importance of firms as small as or smaller than our class 1 and class 2 firms is much greater than our sample proportions would indicate, and the 13 percent figure derived above may well be a loose lower bound. The aggregate retention data also suggest that low-dividend firms are much more numerous and account for a much greater fraction of investment and capital in the economy as a whole than in our Value Line sample. Firms with assets less than \$100 million retained about 77 percent of their income. Therefore, the part of a representative aggregate shock to investment that could be explained by the kind of financial effects estimated here could be substantial, and financing constraints could account for a large proportion of the aggregate variability of investment. While only suggestive, such calculations provide

further impetus to research that links aggregate economic fluctuations to problems in financial markets.

FINANCE CONSTRAINTS, INVESTMENT, AND TAX POLICY

Most studies of the effects of tax policy on investment assume that firms respond to prices set in centralized securities markets, such as market interest rates on Tobin's q , and that the availability of finance does not limit investment. The implications for tax policy are clear: what matters for investment is the marginal tax rate on returns from a new project, not the firm's average tax burden on returns from its investments in place. As we have emphasized, however, for firms that face imperfect markets for external finance, it is not sufficient to focus solely on the cost of funds determined in centralized securities markets. For these firms, the amount of earnings devoted to taxes, and therefore the *average* tax rate on returns from existing projects, matters for investment, possibly along with incentive effects of marginal tax rates. Thus, the cash flow effects of changes in the investment tax credit or depreciation allowances may be more important for many firms than the associated cost of capital effects of such policies.⁵⁰

That average tax rates matter for some firms does not, however, necessarily imply a policy opportunity. To the extent that policymakers can distinguish project types no better than private financiers, the lemons problem remains. An additional concern relates to agency issues. Policies that increase internal finance might encourage managers concerned, for example, with corporate size as well as the value of shareholders' claims to overinvest.⁵¹ Nevertheless, understanding the impact of public poli-

50. These issues are considered in greater detail in Steven Fazzari, R. Glenn Hubbard, and Bruce Petersen, "Investment, Financing Decisions, and Tax Policy," *American Economic Review*, vol. 78 (May 1988, *Papers and Proceedings*, 1987), pp. 200–05.

51. At first glance, our finding that internal finance influences investment spending in addition to q , especially in firms with low payout, could be consistent with a managerial waste hypothesis: available internal finance is invested in projects at levels not justified by market signals alone. Our results show, however, that it is rapidly growing firms with high q values, *not* large, mature or declining firms, that have low average payout and the greatest sensitivity of investment to the supply of internal funds. Therefore, tax changes that increase internal cash flow and liquidity could lead to higher levels of productive investment in some firms.

cies on investment through their effect on internal finance can be important. As an example, asymmetric information problems reduce the likelihood that households can “pierce the corporate veil.” Redistributions of funds away from firms, either to shareholders or to taxpayers, may change both the level of investment and its allocation to the extent that firms face information-related finance constraints.

FURTHER EXTENSIONS AND LINKS TO OTHER CURRENT RESEARCH

The link between the financial influences on investment and information imperfections in capital markets suggests that research on “information capital” accumulation through financial intermediation is important for understanding the investment process. One channel through which information capital can be accumulated is financial institutions that specialize in long-term borrower relationships and in the evaluation of balance sheet positions. These institutions can figure prominently in the finance of smaller firms lacking cost-effective access to commercial paper, bond, and equity markets. Also, venture capitalists can be viewed as specialists in the accumulation of information on balance sheet positions and investment prospects in growing enterprises. The existence of a lemons premium in equity issues does not, however, imply that large arbitrage profits exist, where any cash-rich firm or individual could buy a constrained firm. Rather, “profits” arise from the costly activity of investigating and overcoming information asymmetries.

The existence of finance constraints has implications for research in industrial organization. Kenneth Judd and Petersen argue, for example, that large differentials in the cost of internal and external finance can rationalize predatory and limit-pricing strategies. In addition, interesting evidence provided by David Ravenscraft and F. M. Scherer supports the view that many mergers appear to match different corporations that face different costs of capital on the margin.⁵² Such combinations would

52. Kenneth L. Judd and Bruce C. Petersen, “Dynamic Limit Pricing and Internal Finance,” *Journal of Economic Theory*, vol. 39 (April 1986), pp. 268–99; David J. Ravenscraft and F. M. Scherer, *Mergers, Sell-Offs, and Economic Efficiency* (Brookings, 1987).

permit reallocations of capital that bypass capital markets. This possibility suggests other research questions, some which have been addressed by Ravenscraft and Scherer. Do mergers of companies in related activities perform better than purely conglomerate mergers, and if so, are the reasons information-related? How do young firms that are independent perform relative to those acquired by cash-rich mature companies? Similarly, how do start-up ventures of cash-rich companies perform relative to independent start-up ventures?

Our empirical results on firm investment suggest that models should address links between net worth and credit allocation and the possibility of precautionary retentions by many firms. Theoretical research is proceeding along these lines.⁵³

Future research should consider the role of internal finance in investment decisions in other countries, examining differences in tax policies, the structure of capital markets, and organization of firms. A particularly interesting topic would be the analysis of differences in the sensitivity of investment to internal finance according to the extent to which lenders participate in corporate decisionmaking. Research in these areas is just beginning, but the importance of internal finance for investment has been confirmed using firm data for Japan and for the United Kingdom.⁵⁴

These results are also relevant to debates over the source of aggregate fluctuations. The importance of firm heterogeneity in capital markets suggests that representative agent, real business-cycle models, in which financial factors are irrelevant and productivity shocks drive macroeconomic movements, are not likely to be adequate descriptions of cyclical fluctuations. On a formal level, models should consider channels through which exogenous shocks are magnified by information imperfections in capital markets.

53. Roger E. A. Farmer, "A New Theory of Aggregate Supply," *American Economic Review*, vol. 74 (December 1984), pp. 920–30; Bernanke and Gertler, "Financial Fragility"; Calomiris and Hubbard, "Firm Heterogeneity"; Bruce Greenwald and Joseph E. Stiglitz, "Information, Finance Constraints, and Business Fluctuations" (Princeton University, 1986).

54. Takeo Hoshi, Anil K. Kashyap, and David Scharfstein, "Corporate Structure and Investment: Evidence from Japanese Panel Data" (MIT, May 1988); Richard Blundell, Stephen Bond, Michael Devereux, and Fabio Schiantarelli, "Does Q Matter for Investment? Some Evidence from a Panel of U.K. Companies," Working Paper 8712 (London: Institute for Fiscal Studies, December 1987).

APPENDIX A

Dividends, Investment, and Q under Alternative Financing Regimes

WE BEGIN with a simple model of equity finance, dividends, and investment.⁵⁵ In tax-based models, there are differences in the costs of internal and external finance because of the differential taxation of capital gains and dividends at the personal level. In any period t , an existing shareholder's after-tax return R_t is the sum of a dividend return (taxed at rate θ) and a capital gain (taxed at an accrual-equivalent rate c), so that

$$(A.1) \quad R_t = \frac{(1 - \theta)D_t + (1 - c)(V_{t+1} - V_t)}{V_t},$$

where D_t represents the dividend payment by the firm, V_t is the value of the firm's equity, and V_{t+1} is the value in period $t + 1$ of the shares outstanding in period t . In period $t + 1$, the total value of the firm is

$$(A.2) \quad V_{t+1} = {}_tV_{t+1} + V_t^N,$$

where V_t^N represents new share issues.

In equilibrium, owners of equity earn their required return ρ , so that

$$(A.3) \quad \rho V_t = (1 - \theta)D_t - (1 - c)V_t^N + (1 - c)V_{t+1} - (1 - c)V_t,$$

and the value of the firm is given by

$$(A.4) \quad V_t = \sum_{i=0}^{\infty} \left(1 + \frac{\rho}{1 - c}\right)^{-(i+1)} \left[\left(\frac{1 - \theta}{1 - c}\right) D_{t+i} - V_{t+i}^N \right].$$

That is, the total value of the firm is the present value of the posttax dividend stream adjusted for the present value of new share issues that would have to be bought by current equity holders to maintain their proportional claim on the firm.

55. See the discussions in Alan J. Auerbach, "Taxes, Firm Financial Policy and the Cost of Capital: An Empirical Analysis," *Journal of Public Economics*, vol. 23 (February–March 1984), pp. 27–57; and Poterba and Summers, "The Economic Effects of Dividend Taxation."

To take into account the lemons premium associated with new equity issues, as we discussed in the text, we reduce V in equation A.4 by an amount Ω per dollar of new equity issued. That is,

$$(A.5) \quad V_t = \sum_{i=0}^{\infty} \left(1 + \frac{\rho}{1-c}\right)^{-(i-1)} \left[\left(\frac{1-\theta}{1-c}\right) D_{t+i} - (1 + \Omega_{t+i}) V_{t+i}^N \right].$$

The firm maximizes its market value subject to a set of four constraints.

—*Capital accumulation*: $K_t = (1 - \delta)K_{t-1} + I_t$, where K_t is the capital stock at the end of period t , I represents investment, and δ represents a constant rate of depreciation.

—*Sources equal uses of funds*: $(1 - \tau)\pi(K_t) + V_t^N = D_t + I_t$, where $\pi(K)$ represents pretax profits and τ is the corporate income tax rate.

—*Dividends*: $D_t \geq 0$.

—*New share issues*: $V_t^N \geq V^N$; that is, new share issues are assumed to be bounded from below by some minimum (negative) level, V^N .

In summary, the firm chooses I , K , V^N , and D so as to maximize V subject to the constraints described above. That is,

$$(A.6) \quad \max \sum_{i=0}^{\infty} \left(1 + \frac{\rho}{1-c}\right)^{-(i-1)} \left\{ \left[\left(\frac{1-\theta}{1-c}\right) D_t - (1 + \Omega_t) V_t^N \right] \right. \\ \left. - \lambda_t \left[K_t - (1 - \delta)K_{t-1} - I_t \right] \right\} \\ - \alpha_t \left[(1 - \tau)\pi(K_t) + V_t^N - D_t - I_t \right] \\ - \beta_t (V_t^N - V^N) - \gamma_t D_t,$$

where λ , α , β , γ are the Lagrange multipliers associated with the constraints.

The solution for the case where internal finance exceeds investment is familiar. In that case, if the dividend tax rate exceeds the capital gains tax rate ($\theta > c$), it is never optimal to issue new shares and pay dividends at the same time. Abstracting from corporate tax considerations, the equilibrium value of an additional unit of capital—marginal q —is equal to $(1 - \theta)/(1 - c)$. This is the q value at which shareholders are indifferent between a dollar of retentions reinvested in the firm and taxed at rate c , and a dollar of dividends taxed at rate θ .

New shares are issued only when internal finance is exhausted and the marginal q on additional projects exceeds $1 + \Omega$. The range of q values over which firms neither pay dividends nor issue new shares can be derived as follows. When firms are not paying dividends and internal finance is exhausted, we know that $\beta_t = 0$ and

$$(A.7) \quad \alpha_t = -1 - \Omega_t.$$

Given the lemons discount, firms will choose to issue shares only when

$$(A.8) \quad \lambda_t \geq 1 + \Omega_t,$$

so that the supply-of-funds schedule facing the firm has a discontinuity at the point where retentions are exhausted.

APPENDIX B

Data Base and Variables

OUR DATA SAMPLE was the annual Value Line data base, updated in April 1986. The data cover manufacturing firms (two-digit SIC codes between 20 and 39, inclusive). Firms were included in the sample only if they had observations for each year from 1969 through 1984. The 1969 data were used only for constructing lags. We used earlier data, when available, to construct longer lags for some of the tests described in tables 4 and 5. We chose 1969 as the starting point because inventory data necessary to construct the Q variable were available only from 1969 onward. We excluded 1985 because the number of firms with observations in 1985 dropped substantially.

Firms that had mergers valued at more than 10 percent of their capital stock were excluded from the sample because large mergers could lead to inconsistencies when constructing the ratios used in the regressions. Merger data were taken from the COMPUSTAT data base. The merger deletions occurred almost exclusively among mature firms, and they did not materially affect the reported results. Several observations were deleted because of missing data for individual variables necessary for the regressions. Three firms were deleted because of major inconsistencies between their capital stock and investment data. Two firms were moved from the first to the second class, and one firm from the first to

the third class, because of substantial and frequent share repurchases that functioned like dividends. Share repurchases in the remainder of the first class firms were zero or negligible. Further details concerning the data are available from the authors.

Market value of equity (V). The value of common stock at the beginning of the year is the average price over the last fiscal quarter of the previous year times the number of shares outstanding at the end of the previous fiscal year. For the preferred stock, we compute the market value by dividing preferred dividends by the preferred stock yield from Standard and Poor's.

Value of debt (B). The results in the text are based on the book value of short-term and long-term debt. We also considered the effect of estimating the market value of long-term debt as follows. Value Line data provide the interest paid on long-term debt. The ratio of this variable to the book value of long-term debt gives an estimate of the debt's average coupon rate (r_c). To avoid the effect of outliers, this ratio was limited at a 10 percent premium over the Baa corporate bond rate. Following Michael Salinger and Lawrence Summers, we assumed all long-term debt carries a Baa rating.⁵⁶ Then the market value of long-term debt can be estimated by $[(1 + r_{Baa})/(1 + r_c)]^M$ times the book value, where r_{Baa} is the market rate on Baa debt and M is the average time to maturity of the existing debt. We made this adjustment for M values of 5, 10, 15, and 20 years, reflecting the fact that the maturity of outstanding debt across our retention classes is likely different. None of these calculations, however, changed the pattern of the reported Q values or regression results for any of the M values, relative to the results with book values presented in the text.

We also considered the possibility that the debt of firms in the first class was more risky than Baa debt, in which case the adjustments described above would overstate the value of debt in class 1 and could bias the q measurements upward. We assumed that any difference between r_c and r_{Baa} was a risk premium, and computed q with the debt discounted accordingly. Again, this modification produced virtually no difference in the statistics relative to the book-value calculations.

56. Michael A. Salinger and Lawrence H. Summers, "Tax Reform and Corporate Investment: A Microeconomic Simulation Study," in Martin Feldstein, ed., *Behavioral Simulation Methods in Tax Policy Analysis* (University of Chicago Press, 1983), pp. 247-81.

Replacement value of the capital stock (K). K_t represents the capital stock at the beginning of period t . The replacement value of property, plant, and equipment is estimated from book values using a method similar to that of Salinger and Summers. We set the initial value of K to the value of net plant (adjusted to market value with aggregate data) for the first year the firm appears on the Value Line data base. The capital stock is then defined iteratively as

$$K_t = [I_t + (P_t/P_{t-1}) K_{t-1}] (1 - 1/LIFE),$$

where P_t is the implicit price deflator for fixed nonresidential investment, I_t is the firm's capital spending, and $LIFE$ is the average service life implicit in the firm's book depreciation costs. The final term is based on the assumption that economic depreciation is single-declining balance. Our results did not change substantially when we assumed double-declining balance economic depreciation. For mature firms, the starting point for this procedure generally stretched back to the late 1950s. For newer firms, the initial book value of their capital stock probably is a good estimate of its replacement cost. Thus, the capital stock estimates should exhibit little inflationary bias for our sample that begins in 1969.

Tax parameters for Q . As in Salinger and Summers, we assume that tax policy parameters remain constant, and that the sum of the required rates of return on investment and expected inflation is equivalent to the nominal Baa bond rate plus 0.06. That is, we let

$$X_t = \tau z \left[\frac{1 - \theta}{1 - c} \right] K_t,$$

where τ represents the corporate income tax rate, π represents inflation, and K_t is the nominal replacement value of the capital stock and

$$z = \left[\frac{\delta}{\delta + \frac{\rho + \pi}{1 - c}} \right].$$

Tax depreciation is assumed to be double-declining balance at rate $\delta = 2/LIFE$. The average effective tax rate on dividends (θ) and capital gains (c) are taken from James Poterba.⁵⁷ The corporate tax rate τ was set at the statutory maximum marginal rate.

57. Poterba, "Tax Policy."

Market value of inventories (N). Because inventories are included in the market valuation of the firm, but not in the replacement cost of the fixed capital stock, we subtract N from the market value of the firm. There was no substantial difference in the results when N was instead added to the replacement cost of the firm's capital stock. Inventories for each firm are converted from book value to market value using the procedure outlined in Salinger and Summers and Value Line data concerning whether the firm uses LIFO and FIFO methods of inventory accounting.

Investment tax credit (k). Information on legislated values of the investment tax credit was taken from the Washington University Macro Model. Information on the mix between equipment and structure was taken from aggregate data.

Cash flow (CF). Cash flow, as defined by Value Line, equals income after interest and taxes, plus all noncash deductions from income (principally depreciation allowances and amortization). Dividends were not subtracted from cash flow.

Q definitions. Using these components, we have constructed three Q measures:

Tobin's $q = (V + B - N)/K$;

Tax-adjusted $Q = (1 - \tau)^{-1} \left[\frac{V + B - X - N}{K} - (1 - k - \tau z) \right]$; and
(no dividends paid)

Tax-adjusted $Q = (1 - \tau)^{-1} \left[\left(\frac{1 - c}{1 - \theta} \right) \left(\frac{V - X}{K} \right) + \frac{B - N}{K} - (1 - k - \tau z) \right]$
(dividends paid)

Cost of capital (r). The cost of capital is given by

$$r = \left(\frac{p_k}{p} \right) \left(\frac{1 - k - \tau z}{1 - \rho} \right) \left[(1 - L) \left(\frac{1 - \theta}{1 - c} \right) i + (1 - \tau) i L - \pi^e + \delta \right],$$

where

p_k = implicit price deflator for capital goods

p = implicit price deflator for nonfarm business output

τ = corporate income tax rate

k = investment tax credit rate

z = present value of one dollar of depreciation allowances

θ = marginal effective personal tax rate on dividend income

c = marginal effective personal tax rate on capital gains

L = average proportion of marginal investment financed with debt

i = average nominal Baa corporate bond rate

π^e = expected inflation rate

δ = economic depreciation rate.

Comments and Discussion

Alan S. Blinder: A few years ago, in revising my graduate course reading list, I looked for some modern literature on liquidity constraints and investment analogous to the burgeoning literature on liquidity effects on consumption. There was none. Now there is, thanks to the sterling efforts of Steven Fazzari, Glenn Hubbard, and Bruce Petersen. So, lest what I have to say sound critical, I want to state clearly that the potential effects of cash flow on investment was a research question crying out to be asked theoretically and then answered empirically. The authors, in this paper and its predecessor, attempt to do both. For that, they deserve credit, maybe even cash.

Empirically, there are striking parallels between consumption and investment. As we all know, consumption seems to respond strongly to current income and weakly, if at all, to interest rates. The stylized facts from business investment equations are much the same: a strong response to sales or output and a weak response to the cost of capital. These four econometric findings pose challenges to economic theory.

I start with income sensitivities, since they are most germane to the authors' work. Milton Friedman and Franco Modigliani suggested decades ago that if consumption decisions arise from intertemporal optimization, then current income should have little effect on current consumption. Yet the observed effect is strong. Modern consumption theorists append rational expectations to the Friedman-Modigliani framework and offer two explanations: the theory is right, but current income is an excellent predictor of future income; the theory is wrong, perhaps because of liquidity constraints.

In the case of investment, the empirical puzzle runs deeper and the explanations run shallower. Basic neoclassical theory denies any role to current output; only relative factor prices should drive investment. As economists realized in the 1950s, but forced themselves to forget in the

1960s and 1970s, liquidity constraints offer one possible explanation: short-run fluctuations of GNP have large effects on cash flow, which is a cheaper source of finance than external funds. The authors resurrect this 1950s view, but rationalize it not by transactions costs—though they do mention them—but rather by 1980s-style theorizing based on informational asymmetries. I like this line of theorizing, though I think there is a tendency to carry it too far. For example, were most capital markets closed to Steven Jobs in 1975 because of the lemons problem, or was it because the risk was so great? Similarly, did General Motors finance its recent multibillion dollar investment campaign so easily because information was symmetric or because its pockets were so deep? We should insist on evidence that informational problems are more important in practice than simpler explanations like transactions costs.

Now, what of interest rates? It is by now widely agreed that saving is not sensitive to rates of return. The standard explanation is that income effects cancel substitution effects. This explanation, of course, will not do for investment because profit maximization precludes income effects. Yet the stylized fact is much the same: you have to torture the data pretty ruthlessly before they confess to an interest elasticity of investment. Why? One possibility is that business managers do not maximize profits. I return to that heresy at the end of my comments.

The authors' explanation is, once again, the financing hierarchy. If the marginal cost of funds looks like a staircase with narrow treads and big risers (see the authors' figure 1), then many firms will find their optimum on the risers rather than on the treads. For such firms, a vertical upward or downward shift of the whole staircase (a change in the cost of capital) will have no effect on investment, but a widening of the relevant tread (a change in credit availability) will change investment. Obviously, the story is more important empirically when the risers are tall than when they are short (again, see figure 1). In the authors' theory, the heights depend on the severity of informational asymmetries. In a more naive theory, they depend on transactions costs.

Although the model favored by the authors is consistent with the stylized facts, it is not the only possible explanation. Matthew Shapiro offered a different explanation for these same facts two years ago at a meeting of this panel.¹ His was that frequent, large shocks to productivity

1. Matthew D. Shapiro, "Investment, Output, and the Cost of Capital," *BPEA*, 1:1986, pp. 111–52.

simultaneously raise output, investment, and interest rates. As I recall, Shapiro was all but hooted out of the room—Washington being too far from the Great Lakes to make his story believable, especially in a crowd more favorably disposed toward liquidity constraints. But we should still insist on empirical evidence.

Fazzari, Hubbard, and Petersen provide some. Their basic empirical idea is a good one. To see whether investment spending is sensitive to cash flow, they try to identify, on a priori grounds, the firms most likely to encounter liquidity constraints. They suggest dividend behavior as the telltale sign: firms with very low dividend payout rates are arguably more likely to be liquidity constrained than firms with more normal payout rates. I understand the argument. But it makes me a bit uneasy because it is so puzzling that firms pay any dividends at all. It takes exceedingly clever theoretical arguments to rationalize this apparently irrational behavior.

I also have an econometric source of unease. Dividend payout rates are endogenous and, in particular, are probably sensitive to unobserved investment prospects. The authors' basic regression is:

$$(1) \quad I/K = aQ + b(\text{Cash flow}/K) + u.$$

Firms that draw large positive u 's will probably choose low payouts and hence wind up in classes 1 and 2 while firms with large negative u 's will wind up in class 3. That starts to sound like truncating on the error term. I'm only a good enough econometrician to worry about that problem, not to figure out whether including fixed effects, as the authors do, takes care of it.

It seems to me that there are other ways to divide the sample—old versus young firms or small versus large ones—that are freer of this problem and relate better to the information-based theories to which the authors appeal. Of course, these attributes are correlated with dividend policy; so perhaps the results would look much the same. However, dividend policy, age, and size are not perfectly correlated; so alternatives are perhaps worth exploring.

Dividing the sample in different ways has one further virtue. As I have noted, the financing staircase can arise from several sources. The lemons explanation that the authors favor suggests that young versus old might be the key distinction. Theories based on deep versus shallow pockets or on fixed flotation costs suggest that small versus large may be the key distinction.

✓ The results the authors obtain are stunningly strong and important. In regressions like equation 1, estimates of b are large and significant, even though very small firms and start-ups are not in the Value Line sample. In fact, the results are too strong and too robust. Cash flow seems to affect investment strongly even in class 3 firms, which have an average 1984 capital stock of \$2 billion and an average payout rate of about 40 percent. Look, for example, at table 4, which uses the authors' favorite theory, the Q theory. The equation for the full period says that, at the margin, a one dollar increase in cash flow raises investment spending by 23 cents. That's a lot. Can we really believe that lending to one of these billion-dollar firms is like buying a used car from a stranger? I know I'd rather buy a used bond from Chrysler than a used Chrysler from Bond.

Here is a second problem. It seems to me that the staircase theory argues not only that cash flow should be more important in classes 1 and 2, which the authors always find, but also that cost of capital effects should be less important. In table 4, this is not true: Q matters most in class 2. And in table 8, the Jorgenson term matters most in class 1 and least in class 3.

Finally, let me say something about the most boring issue in macro-economics: stock versus flows. It seems to me that liquidity constraints should pertain to stocks, not to flows. I can understand why a firm with limited access to external capital might find its holdings of physical capital constrained by internal funds. But I have a hard time understanding how a low current cash flow could constrain the net acquisition of capital by a firm with a large accumulated stock of cash. Yet table 10 shows that cash flows matter more than cash stocks and that adding stocks does not reduce the coefficients of cash flow very much.

I can think of two possible explanations. The first is that the equation is misspecified: it should relate the desired capital stock to cash stocks and append an adjustment mechanism through which current cash flows influence the adjustment of actual to desired capital. The second is that the constraining variable for current investment is actually opening cash stock plus current cash flow, and cash flows are bigger and more variable than opening stocks, so they dominate econometrically. I have no idea if either of these explanations holds water.

One last remark. At the end of their paper, the authors dismiss the "managerial waste" hypothesis: that managers invest internal funds even if the investments are not profitable. I would not dismiss it so

lightly. Perhaps managers of large firms treat internal funds as costless and hate to go to the market. How else are we to explain the influence of cash flow on the investment of billion-dollar firms? Maybe managers are like mountain climbers: they invest the money “because it is there.” That, I suppose, is what Carl Icahn and Boone Pickens believe. They are certainly rich. Maybe they are also smart.

James M. Poterba: Empirical comparisons between the simple accelerator, neoclassical accelerator, Q theory, and cash flow models of aggregate U.S. investment have usually favored the simple accelerator specification. Nevertheless, textbook and classroom expositions of business investment tend to rely on either the neoclassical accelerator or Q model, since they can be grounded more formally in economic theory. This provocative and important paper seeks to change the way we think about the investment function in two ways. First, it marshals a convincing theoretical case based on credit market imperfections for the proposition that cash flow may significantly affect investment outlays. Economic theory suggests many reasons why firms may be cash constrained when making investment outlays. Second, after removing the central obstacle to the respectability of the cash flow model, the paper shows that cash flow variables substantially improve the explanatory power of investment equations estimated using individual firm data. The paper breaks new ground in explicitly modeling firm heterogeneity with respect to investment rules and in demonstrating that cash flow plays a more important role in investment decisions of small firms that retain most of their earnings.

There is more to compliment than to quarrel with in this paper. My comments will reflect this, focusing on three questions that arise in evaluating the paper. First, is the link between cash flow and investment operative primarily for low-dividend firms, or is it likely to be significant for mature firms as well? Second, do the paper’s empirical results significantly sharpen our knowledge of how cash flow affects investment? Third, how well do the present results, for a sample of manufacturing firms, extrapolate to the economy at large? I shall consider these questions in turn.

The authors are undoubtedly correct in arguing that some small, low-dividend firms face cash flow constraints when undertaking new investments. Even for mature dividend-paying firms, however, I suspect (and

the paper's empirical results confirm) the potential importance of cash flow. Several strands of prior evidence buttress the view that cash flow may be more influential for large firms than the authors claim. First, mature dividend-paying firms cannot costlessly reduce their dividends. Share prices fall when firms cut their dividends: the most recent study shows a 2 percent decline in prices when a firm reduces its dividend, and an 8 percent decline if a firm completely omits a dividend.¹ For a firm with a dividend yield of 4 percent a year, omitting the dividend for a one-year period will reduce share values by twice as much as the increment to investment funds. This suggests significant costs to dividend cuts, but it may also place an upper bound on the potential cost of external funds for mature firms. Anecdotal evidence also suggests the difficulty of dividend reduction. In 1968 when General Utilities tried to omit its dividends to finance investment projects, shareholders protested violently and eventually the management agreed to continue the dividend and resort to external finance.²

Second, previous empirical studies of rates of return are consistent with the view that internal finance is perceived as less costly than external funds.³ Ex post profit rates are higher for firms that use external finance, particularly external equity, than for firms that rely on internal finance. These results are difficult to interpret because they may demonstrate only that firms with good earnings prospects can convince investors of their favorable future returns, but they are nevertheless consistent with this paper's results. They are not restricted to small firms, although it might be interesting to reexamine the earlier tests using the type of firm stratification rule developed in the present paper.

Third, the asymmetric information problems that are invoked to explain credit market failures for small firms appear to affect both large and small firms. The voluminous literature on the valuation consequences of changes in capital structure, finding positive returns to transactions that add debt or replace equity with debt, and negative

1. Kenneth M. Eades, Patrick J. Hess, and E. Han Kim, "Market Rationality and Dividend Announcements," *Journal of Financial Economics*, vol. 14 (December 1985), pp. 581-604.

2. "A Case for Dropping Dividends," *Fortune*, June 15, 1968, p. 181.

3. References to this literature, and some constructive empirical evidence, may be found in Alan J. Auerbach, "Taxes, Firm Financial Policy and the Cost of Capital: An Empirical Analysis," *Journal of Public Economics*, vol. 23 (February-March 1984), pp. 27-57.

returns for equity issues, shows that firms are affected across size categories. Of course, it may be that *if* significant capital structure changes were observed for smaller firms, the valuation effects would be even larger than those for mature firms. This evidence nevertheless suggests the potential importance of imperfect information even for large firms.

While the a priori case for believing this paper's central theme is strong, that does not simplify the task of determining how much investment results from shocks to corporate cash flow. That must be answered on the basis of the empirical results, where some caution is required. The authors report investment equations for three groups of firms stratified on the basis of dividend payout and show that the link between investment and cash flow is substantially stronger for low-payout than for high-payout firms, even after controlling for Tobin's q . The key question is whether shocks to cash flow are transmitted to investment outlays, or whether other uses of funds, such as repurchasing shares or buying back or issuing debt, serve as shock absorbers when earnings fluctuate.

Earlier studies of investment and cash flow were dismissed partly because shocks to cash flow signal two things: an increase in current liquidity and a potential improvement in future profitability. The present paper is much more careful about this problem than previous investment studies. By controlling for the beginning-of-year value of Tobin's q , the investment equations reduce the informational content of current cash flow. They do not eliminate it, however, and this clouds the results. There are many reasons for suspecting that measured Q is not a sufficient statistic for future cash flows. These range from difficulties in measuring the replacement cost of the firm's assets, to concern over whether average Q is a good proxy for marginal Q , to questions about the informational content of stock prices themselves. If for any of these reasons the measured Q variable provides an error-ridden indicator of the firm's true prospects, then econometric results may find that current cash flow affects investment only because this variable, just like measured Q , is correlated with the "true" marginal Q variable that firms consider in making investment decisions. The pattern of results across different classes of firms could be explained on this view because Q is measured with more error for smaller firms, which tend to be lower-dividend firms. The authors recognize these potential difficulties, and

現金流
的雙重功能

allude to instrumental variable results where the current cash flow variable is treated as endogenous. These results are unfortunately not reported, even though they are easier to interpret than the ordinary least squares estimates. Similarly, the authors mention but do not report equations including Tobin's q from the end of the current period as well as the end of the previous period. The coefficients on cash flow in these equations are somewhat cleaner than those from the models with only lagged Q , since they avoid biases that result when cash flow incorporates later information than the Q variable.

One particular source of error that illustrates these problems concerns tax losses. Although the paper uses microeconomic data, the authors assume that all firms face identical tax parameters. In practice, some firms have tax loss carryforwards that prevent them from taking advantage of the investment tax credit and depreciation allowances that are available to the "representative firm." For tax loss firms, the assumption that they can claim full tax benefits induces a measurement error in Q . Moreover, since a firm's current cash flow is almost certainly correlated with its tax status, the measurement error is correlated with the cash flow variable. A standard errors-in-variables argument could therefore account for the cash flow coefficients. Instrumental variables estimates using the lagged value of Q , or equations that ignore the tax factors completely, may fail to remedy these problems. Further work, with more explicit modeling of the measurement error dynamics, would help, since definitive support for credit market effects must resolve these issues.

The final question I consider involves the authors' efforts to generalize their results. Within the sample, approximately 1 percent of total investment was undertaken by firms in class 1, and another 2.3 percent by firms in class 2. This understates the importance of cash flow factors as sources of investment fluctuation, however, since the authors correctly observe that cash flow is more variable for their class 1 and 2 firms than for the mature class 3 corporations. The paper's extrapolations are probably too sweeping, however. The paper notes that over 20 percent of assets in manufacturing are held by firms that are as small as, or smaller than, the firms in class 1. The trouble with inferring that they all face tight borrowing constraints is that firms with traded equity (a precondition for being in the sample) may be a selected group that has both substantial investment needs and weak access to bank credit. It

may therefore be difficult to extrapolate the results to the rest of the manufacturing sector.

It is even more difficult to generalize to nonmanufacturing firms, which held over 70 percent of corporate plant and equipment at the end of 1986. Some assets, such as cars, cash registers, and computers, can serve as collateral for bank loans. Firms that invest heavily in such standardized assets probably face much easier hurdles on external finance than do more specialized manufacturing firms that purchase unique assets. Firms outside manufacturing are also likely to experience more stable cash flow: a 1 percent change in GNP translates into more than a 2.2 percent change in manufacturing output. This suggests that the cash flow considerations that are highlighted here may be less central in other parts of the economy. Conclusions about the importance of cash flow factors in these sectors must therefore await evidence on the behavior of nonmanufacturing firms.

In testimony to the important and provocative nature of this paper, studies generalizing the present methodology to other samples of firms, in other industries and other countries, have already begun to appear. There is little doubt that future research on corporate investment and capital markets more generally will have to reckon with the authors' revivification of the cash flow model of capital spending.

General Discussion

Some participants discussed the reliability of the authors' empirical results. Elaborating on Alan Blinder's comments, Christopher Sims suggested that the authors should have grouped the firms according to some essentially exogenous characteristic such as size or age rather than by their dividend-income ratio. It is not sufficient to argue that all class 1 firms are small or young, because a considerable percentage of the small and young firms might be in classes 2 and 3. Even in that case a simultaneity bias will remain. James Tobin noted that the firm jointly determines investment, dividend payments, and other ways of allocating its cash flow. Therefore, he suggested that the authors model investment and dividends as depending on the same set of explanatory variables.

Sims went on to describe two other potential pitfalls of the authors' econometric method. First, cash flow may be a key source of information

to the firm about future profitability. Hence investment should be correlated with cash flow even with perfect capital markets. The present results may simply indicate that the information content of cash flow is greater for class 1 firms, which are almost all small and young. Second, even in the absence of a correlation between investment opportunities and cash flow in the entire population of firms, it is possible that the authors' method of classification will group together firms that, by chance, have cash flow roughly equal to their investment needs.

William Brainard concurred with Sims's argument, observing that the typical class 1 firm is likely to have a low variance of its dividend payout ratio as well as a low average. Since dividends themselves tend to be infrequently changed, most of the variation in a firm's payout ratio is likely to reflect variations in the denominator, its earnings. High earnings variability presumably reduces the information content of current earnings for the profitability of investment. Hence the firms excluded from class 1 would be expected to have a lower correlation of cash flow and investment, even with perfect capital markets. Joseph Stiglitz suggested a more powerful method to test for the importance of the cash flow constraint. If the cash flow constraint is actually binding, then one should find a clustering of investment levels around the constraint. On the other hand, if investment is far away from the constraint, then it is likely that a significant coefficient on cash flow is spurious.

Robert Hall was generally skeptical about the progress of empirical work on investment. He noted that most investment equations, including his own earlier work with Jorgenson and the present equations of the authors, suffer from an identification problem. Because the right-hand-side variables are invariably endogenous, there is no way to determine what is driving what.

Discussion turned to Blinder's question of why cash flow rather than the stock of cash is the relevant variable for investment equations. It is difficult to argue that a firm with low cash flow is constrained if it holds substantial liquid assets. James Poterba noted that a firm that builds up large stocks of cash for future investments is considered a cash cow: a prime target for takeovers. A firm may therefore soak up excess cash flow by investing incrementally rather than acquiring stocks of cash. This would tend to make investment more highly correlated with cash flow than with stocks of cash. Stiglitz suggested that the liquidity of a firm includes its lines of credit as well as its stock of cash. This is an

alternative explanation of why the stock of cash has little explanatory power in cross-sectional investment equations even if finance constraints are important. Stiglitz also noted that for a variety of reasons firms may want to maintain a certain ratio of capital to cash on their balance sheet. Thus the stock of cash may actually increase with investment, contrary to what would be expected in a liquidity-constrained world.

Stiglitz noted that imperfect information is a key reason for constraints on external financing, for both large and small firms. Therefore he was not surprised by the economically significant cash flow coefficients even for the larger class 3 firms. Ben Bernanke drew parallels between the authors' work and earlier work of Feldstein and Horioka, who found that for smaller countries investment often equals savings. Thus small countries, as well as small corporations, apparently face external finance constraints.

Partial adjustment toward target capital structures[☆]

Mark J. Flannery^{a,*}, Kasturi P. Rangan^b

^a*Graduate School of Business, University of Florida, Gainesville, FL 32611-7168, USA*

^b*Weatherhead School of Management, Case Western Reserve University, Cleveland, OH 44106, USA*

Received 12 May 2004; received in revised form 21 December 2004; accepted 16 March 2005

Available online 10 October 2005

Abstract

The empirical literature provides conflicting assessments about how firms choose their capital structures. Distinguishing among the three main hypotheses (“tradeoff”, pecking order, and market timing) requires that we know whether firms have long-run leverage targets and (if so) how quickly they adjust toward them. Yet many previous researchers have applied empirical specifications that fail to recognize the potential for incomplete adjustment. A more general, partial-adjustment model of firm leverage indicates that firms *do* have target capital structures. The typical firm closes about one-third of the gap between its actual and its target debt ratios each year.

© 2005 Elsevier B.V. All rights reserved.

JEL classification: G 32

Keywords: Leverage; Tradeoff theory; Target; Speed of adjustment

1. Introduction

Since Modigliani and Miller’s irrelevance proposition in 1958 (Modigliani and Miller, 1958), researchers have investigated firms’ decisions about how to finance their operations.

[☆]We would like to thank, without implicating, Jay Ritter, Arturo Bris, Ralf Elsas, Vidhan Goyal, Rongbing Huang, Mike Lemmon, Peter MacKay, Sam Thomas, Ivo Welch, Jeff Wurgler, and seminar participants at Arizona State University, the Atlanta Finance Forum, Case Western Reserve University, the Federal Deposit Insurance Corporation, George Mason University, New York University, Southern Methodist University, the University of Texas, and Washington University for comments on previous drafts of this paper. Murray Frank (the referee) provided advice that substantially improved the paper. George Pennacchi and Ajai Singh provided helpful advice about a related paper.

*Corresponding author. Tel.: 352 392 3184; fax: 352 392 0301.

E-mail address: flannery@ufl.edu (M.J. Flannery).

Initially, they asked whether the irrelevance proposition is consistent with the available data, or, whether instead capital market imperfections make firm value depend on capital structure. In the latter case, it was argued, firms would select target debt-equity ratios, trading off their costs and benefits of leverage. Survey evidence by [Graham and Harvey \(2001\)](#) shows that indeed, 81% of firms consider a target debt ratio or range when making their debt decisions. However, alternative theories remain plausible. [Myers \(1984\)](#) contrasts this tradeoff theory of capital structure with an updated version of [Donaldson's \(1961\)](#) pecking order theory, according to which information asymmetries lead managers to perceive that the market generally underprices their shares. Accordingly, investments are financed first with internally generated funds, the firm issues safe debt if internal funds prove insufficient, and equity is used only as a last resort. In a pecking order world, observed leverage reflects primarily a firm's historical profitability and investment opportunities. Firms have no strong preference about their leverage ratios and, a fortiori, no strong inclination to reverse leverage changes caused by financing needs or earnings growth.

Two additional theories of capital structure also reject the notion of timely convergence toward a target leverage ratio. First, [Baker and Wurgler \(2002\)](#) argue that a firm's observed capital structure reflects its cumulative ability to sell overpriced equity shares: that is, share prices fluctuate around their "true" values, and managers tend to issue shares when the firm's market-to-book ratio is high. Unlike the pecking order hypothesis, this market timing hypothesis asserts that managers routinely exploit information asymmetries to benefit current shareholders; like the pecking order hypothesis, there is no reversion to a target capital ratio if market timing is the dominant influence on firm leverage. Second, [Welch \(2004\)](#) argues that managerial inertia permits stock price changes to have a prominent effect on market-valued debt ratios: "... over reasonably long time frames, the stock price effects are considerably more important in explaining debt-equity ratios than previously identified proxies" (p. 107).

The pecking order, market timing, and inertia theories of capital structure imply that managers perceive no great leverage effect on firm value and therefore make no effort to reverse changes in leverage. In contrast, the tradeoff theory maintains that market imperfections generate a link between leverage and firm value, and firms take positive steps to offset deviations from their optimal debt ratios. The speed with which firms reverse deviations from their target debt ratios depends on the cost of adjusting leverage. With zero adjustment costs, the tradeoff theory implies that firms should never deviate from their optimal leverage. At the other extreme, if transaction costs are infinite we should observe no movements toward a target. [Baker and Wurgler \(2002\)](#) emphasize the connection between adjustment costs and observed capital structure:

The basic question is whether market timing has a short-run or a long-run impact. One expects at least a mechanical, short-run impact. However, *if firms subsequently rebalance away from the influence of market timing financing decisions, as normative capital structure theory recommends*, then market timing would have no persistent impact on capital structure. (page 2, emphasis added)



Estimating the effect of capital adjustment costs is thus a key first step in testing competing theories of capital structure.

The empirical model in this paper accounts for the potentially dynamic nature of a firm's capital structure. The model is general enough that we can test whether there is indeed a

leverage target and if so, what is the (adjustment) speed with which a firm moves toward its target. Our evidence indicates that firms do target a long run capital structure, and that the typical firm converges toward its long-run target at a rate of more than 30% per year. This adjustment speed is roughly three times faster than many existing estimates in the literature, and affords targeting behavior an empirically important effect on firms' observed capital structures. When we add market timing or pecking order variables to our base specification, we do find some support for these theories. However, more than half of the observed changes in capital structures can be attributed to targeting behavior while market timing and pecking order considerations explain less than 10% each. Unlike Welch (2004), we find that stock price changes have only transitory effects on capital structure.

Our findings are not consistent with many recent empirical papers on capital structure (e.g., Shyam-Sunder and Myers, 1999; Baker and Wurgler, 2002; Fama and French, 2002; Huang and Ritter, 2005; Welch, 2004). However, the literature also offers some precedents for our rapid estimated adjustment speeds (Marcus, 1983; Jalilvand and Harris, 1984; Roberts, 2002). We argue that some previous empirical models impose unwarranted, yet testable, assumptions about the adjustment speed and/or the dynamic properties of target leverage. These assumptions materially influence the estimation results and consequently the conclusions drawn. Part of our paper's contribution is to identify why previous research produces such disparate estimated adjustment speeds.

The paper is organized as follows. Section 2 derives our preferred regression specification for testing the tradeoff theory in a partial adjustment framework. Section 3 describes the Compustat—CRSP data we use to estimate our regression models. Section 4 presents our basic results. After showing that our regressions are robust to various estimation methods, we establish the statistical and economic significance of a target debt ratio and relate our results to previous discussions of the tradeoff theory. Section 5 explicitly compares our model to the pecking order, market timing, and inertia models. Section 6 presents a series of robustness tests and the final section concludes. An appendix discusses the econometric issues associated with estimating the dynamic panel regression that constitutes our base specification.

2. Regression model specification

A regression specification used to test for tradeoff leverage behavior must permit each firm's target debt ratio to vary over time, and must recognize that deviations from target leverage are not necessarily offset quickly. Both of these requirements are satisfied in a model with partial (incomplete) adjustment toward a target leverage ratio that depends on firm characteristics.

2.1. Target leverage

Our primary leverage measure is a firm's market debt ratio,¹

$$MDR_{i,t} = \frac{D_{i,t}}{\underline{D_{i,t}} + \underline{S_{i,t}P_{i,t}}}, \quad (1)$$

¹Finance theory tends to downplay the importance of book ratios, with previous research largely analyzing market-valued debt ratios (including Hovakimian et al., 2001; Hovakimian, 2003; Fama and French, 2002; Welch,

where $D_{i,t}$ denotes the book value of firm i 's interest-bearing debt (the sum of Compustat items 9 plus 34) at time t , $S_{i,t}$ equals the number of common shares outstanding (Compustat item 199) at time t , and $P_{i,t}$ denotes the price per share (Compustat item 25) at time t .

We model the possibility that target leverage might differ across firms or over time by specifying a target capital ratio of the form

$$MDR_{i,t+1}^* = \beta X_{i,t}, \quad (2)$$

where $MDR_{i,t+1}^*$ is firm i 's desired debt ratio at $t+1$, $X_{i,t}$ is a vector of firm characteristics related to the costs and benefits of operating with various leverage ratios, and β is a coefficient vector. Under the tradeoff hypothesis, $\beta \neq 0$, and the variation in $MDR_{i,t+1}^*$ should be nontrivial.

2.2. Adjustment to target leverage

In a frictionless world, firms would always maintain their target leverage. However, adjustment costs may prevent immediate adjustment to a firm's target, as the firm trades off its adjustment costs against the costs of operating with suboptimal leverage. We estimate a model that permits incomplete (partial) adjustment of the firm's initial capital ratio toward its target within each time period. The data can then indicate a typical adjustment speed.

A standard partial adjustment model is given by

$$MDR_{i,t+1} - MDR_{i,t} = \lambda(MDR_{i,t+1}^* - MDR_{i,t}) + \tilde{\delta}_{i,t+1}. \quad (3)$$

Each year, the typical firm closes a proportion λ of the gap between its actual and its desired leverage levels. Substituting (2) into (3) and rearranging gives the estimable model

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + (1 - \lambda)MDR_{i,t} + \tilde{\delta}_{i,t+1}. \quad (4)$$

Eq. (4) says that managers take 'action' or 'steps' to close the gap between where they are ($MDR_{i,t}$) and where they wish to be ($\beta X_{i,t}$). The specification further implies that

- (1) The firm's actual debt ratio eventually converges to its target debt ratio, $\beta X_{i,t}$.
- (2) The long-run impact of $X_{i,t}$ on the capital ratio is given by its estimated coefficient, divided by λ .
- (3) All firms have the same adjustment speed (λ).²

The smooth partial adjustment in Eq. (4) may only approximate an individual firm's actual adjustments. A reasonable alternative model would permit small deviations from

(footnote continued)

2004; Leary and Roberts, 2005). When authors analyze both market and book leverage ratios, the results are generally comparable. We report similar results below in Table 5. Table 11 presents evidence that our conclusions are robust across a range of reasonable definitions for "leverage."

²We experiment with modeling λ as a function of firm-specific variables (Y), that is,

$$MDR_{i,t+1} = (\lambda(Y)\beta)X_{i,t} + (1 - \lambda(Y))MDR_{i,t} + \delta_{i,t+1}. \quad (5)$$

Although we find some evidence that firm characteristics affect adjustment speeds (the coefficients on Y are statistically significant), we do not report this evidence here because the mean adjustment speeds ($\bar{\lambda}(Y)$) and the coefficients on $X_{i,t}$ are very similar to the results of estimating (4). Roberts (2002) analyzes this issue further.

the target to persist because adjustment costs outweigh the gains from removing small deviations between actual and target leverage (Fischer et al. (1989); Mauer and Triantis (1994); Titman and Tsyplakov (2004); Leary and Roberts (2005); Ju et al. (2002)). Indeed, Figs. 1 and 2 below indicate that the mean change in book leverage substantially exceeds the median, a phenomenon also observed by Frank and Goyal (2003, p. 228), Leary and Roberts (2005); and Halov and Heider (2004, Table 1).

We investigate the impact of infrequent adjustments on the parameters estimated by our smooth adjustment specification (4) by simulating 20 sets of panel data, each with 100,000 data points. The data are generated by assuming that while each firm's target changes stochastically every year, the actual debt ratio is adjusted only periodically. For the randomly chosen periods in which debt is adjusted, the simulated firm adjusts completely to its target ratio. When we estimate a partial adjustment model on these generated data sets, we find that the estimated adjustment speed exceeds the true proportion of adjusting firms by less than 2%. (That is, if an average of 30% of sample firms move to their target each year, the estimated adjustment speed is less than 0.306). The average bias is statistically significant, but economically unimportant. We therefore interpret the

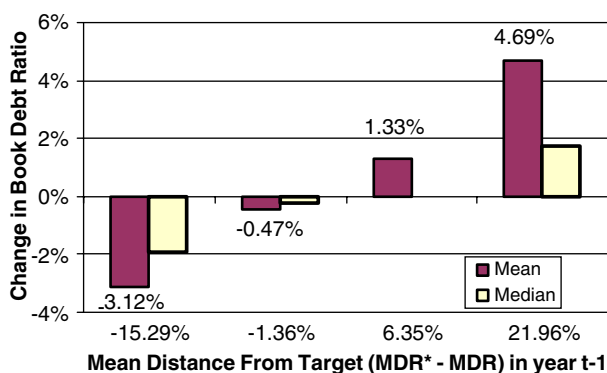


Fig. 1. Subsequent year's change in book debt ratio.

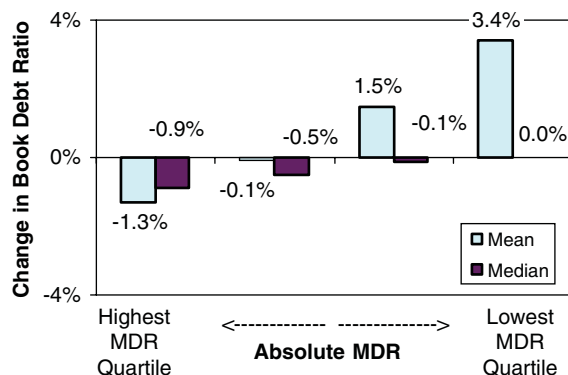


Fig. 2. Mean reversion in leverage.

Table 1

Summary statistics

Sample includes all Industrial Compustat firms with complete data for two or more adjacent years during 1965 to 2001. Total: 12,919 firms; 111,106 firm years. All variables are winsorized at the 1st and 99th percentiles to avoid the influence of extreme observations.

	Number of observations	Mean	Median	Std. Dev.	Min.	Max.
<i>MDR</i>	111,106	0.2783	0.2252	0.2439	0.0000	0.9174
<i>SPE</i>	111,106	0.0038	0.0000	0.0917	-0.3383	0.4334
<i>BDR</i>	111,106	0.2485	0.2296	0.1925	0.0000	0.8635
<i>EBIT_TA</i>	111,106	0.0517	0.0935	0.2142	-1.6371	0.4096
<i>MB</i>	111,106	1.6153	1.0415	1.7888	0.2690	13.6372
<i>DEP_TA</i>	111,106	0.0451	0.0381	0.0327	0.0000	0.2338
<i>LnTA</i>	111,106	18.2400	18.1126	2.0184	12.7227	23.3787
<i>FA_TA</i>	111,106	0.3220	0.2754	0.2200	0.0005	0.9220
<i>R&D_DUM</i>	111,106	0.4654	0.0000	0.4988	0.0000	1.0000
<i>R&D_TA</i>	111,106	0.0337	0.0000	0.0830	0.0000	0.8290
<i>Rated</i>	111,106	0.1035	0.0000	0.3046	0.0000	1.0000
<i>IND_Median</i>	111,106	0.2240	0.2145	0.1339	0.0000	0.8164
<i>MB_EFWA</i>	81,343	1.7552	1.2828	1.3918	0.2690	9.9976
<i>L3MDR</i>	98,709	0.2698	0.2289	0.2189	0.0000	0.9174
<i>FINDEF</i>	111,106	0.0818	0.0074	0.2228	-1.8180	2.4829
<i>MDR₁</i>	110,659	0.2110	0.1737	0.1848	0.0000	0.9918
<i>MDR₂</i>	111,093	0.4121	0.3979	0.2445	-0.0016	1.0000
<i>MDR₃</i>	108,995	0.2064	0.1459	0.2090	0.0000	1.6114

MDR: market debt ratio = book value of (short-term plus long-term) debt (Compustat items [9] + [34])/market value of assets (Compustat items [9] + [34] + [199]*[25])

SPE_t: the surprise impact of share price change on a firm's *MDR* during (*t*, *t* + 1).

$$SPE_t = \left(\frac{Debt_t}{(Debt_t + MarketEquity_t(1 + \hat{R}_{t,t+1}))} \right) - MDR_t,$$

where $\tilde{R}_{i,t+1}$ is the realized return in the i th firm's stock between t and $t+1$.

BDR: book debt ratio: (long-term [9] + short-term [34] debt)/total assets [6].

EBIT_TA: profitability: earnings before interest and taxes (Compustat items [18] + [15] + [16]), as a proportion of total assets (Compustat item [6]).

Market Equity: market value of outstanding common stock (Compustat items [199 × 25]).

MB: market to book ratio of assets: book liabilities plus market value of equity (Compustat items [9] + [34] + [10] + [199] × [25]) divided by book value of total assets (Compustat item [6]).

DEP_TA: depreciation (Compustat item [14]) as a proportion of total assets (Compustat item [6]).

lnTA: log of asset size, measured in 1983 dollars (Compustat item (6) × 1,000,000, deflated by the consumer price index.

FA_TA: fixed asset proportion: property, plant, and equipment (Compustat item [14])/total assets (Compustat Item [6]).

R&D_DUM: dummy variable equal to one if firm did not report R&D expenses.

R&D_TA: R&D expenses (Compustat item (46)) as a proportion of total assets (Compustat item [6]).

Rated: dummy variable equal to one (zero) if the firm has a public debt rating in Compustat (Item [280]).

Ind_Median: median industry **MDR** (excluding the instant firm) calculated for each year based on the industry groupings in Fama and French (2002).

MB_EFWA: “external finance weighted average” of a firm's past market-book ratios (as defined in Baker and Wurgler, 2002, p. 12).

L3MDR: trailing three-year average of the firm's own **MDR**.

FINDEF: ‘financial deficit’ variable constructed as per, used to test the pecking order hypothesis. As defined in Frank and Goyal (2003) (see Table 2),

FINDEF = dividend payments + investments + change in working capital – internal cashflow.

$$MDR_1 = \frac{Long\ term[9] + Short\ Term\ Debt[34]}{Total\ assets[6] - Book\ Equity[216] + Market\ Equity[199*25]},$$

$$MDR_2 = \frac{Total\ Liabilities[181]}{Total\ Liabilities[181] + Market\ Equity[199*25]},$$

$$MDR_3 = \frac{Long\ term\ Debt[9]}{Total\ assets[6] - Current\ Liabilities[181] - Book\ Equity[216] + Market\ Equity[199*25]}.$$

adjustment speed (λ) as the average speed for a “typical” firm. Table 8 (below) provides further evidence that the partial adjustment specification (4) fits the data well.

3. Data

We construct our sample from all firms included in the Compustat Industrial Annual tapes between the years 1965 and 2001. Following previous research, we exclude financial firms (SIC 6000–6999) and regulated utilities (SIC 4900–4999), whose capital decisions may reflect special factors. Because our regression specification includes lagged variables, we must also exclude any firm with fewer than two consecutive years of data. These exclusions leave us with complete information for 111,106 firm-year observations, which consist of 12,919 firms with an average of 9.6 years each.³ Some prior studies exclude smaller firms from the analysis, because their adjustment costs may be unusually large or their leverage determinants might be significantly different. We include all firms in our estimations, but Table 9 reports estimates of the main regression model for various firm size classes. We define annual observations on the basis of fiscal (as opposed to calendar) years because sample firms use a variety of fiscal yearends. Table 1 defines the variables used in our study and reports their summary statistics. All of these variables are winsorized at the 1st and 99th percentiles to avoid the influence of extreme observations. Most of our variables are expressed as ratios; where this is not the case (e.g. *LnTA*), we deflate the nominal magnitudes by the consumer price index to express nominal values in 1983 dollars.

To model a target debt ratio, we use a set of firm characteristics ($X_{i,t}$) that appear regularly in the literature (Rajan and Zingales, 1995; Hovakimian, 2003; Hovakimian et al., 2001; Fama and French, 2002). Their expected effects on the target debt ratio are as follows:

EBIT_TA: A firm with higher earnings per asset dollar could prefer to operate with either lower or higher leverage. Lower leverage might occur as higher retained earnings mechanically reduce leverage, or if the firm limits leverage to protect the “franchise” producing these high earnings. Higher leverage might reflect the firm’s ability to meet debt payments out of its relatively high cash flow.

MB: Market to book ratio of assets. A higher **MB** is generally taken as a sign of more attractive future growth options, which a firm tends to protect by limiting its leverage.

DEP_TA: Depreciation as a proportion of total assets. Firms with more depreciation expenses have less need for the interest deductions provided by debt financing.

LnTA: Log of (real) total assets. Larger firms tend to operate with more leverage, perhaps because they are more transparent, have lower asset volatility, or have better access to public debt markets.

FA_TA: Fixed asset proportion. Firms operating with greater tangible assets have a higher debt capacity.

R&D_TA: Research and development expenses as a proportion of total assets. Firms with more intangible assets in the form of R&D expenses will prefer to have more equity.

³The minimum number of years per firm is two, the maximum is 37, and the median is six. In the parlance of panel data analysis, this constitutes a “large N , small T ” data set.

R&D_DUM: A dummy variable equal to one for firms with missing R&D expenses. About 55% of our sample firm-years do not report R&D expenses. For these firms, we set R&D expense to zero and set **R&D_DUM** equal to one.

Ind_median: The firm's lagged industry median debt ratio (using Fama and French, 1997 industry definitions), to control for industry characteristics not captured by other explanatory variables. (See also Hovakimian et al., 2001; Roberts, 2002).

In addition to these “usual” determinants of target leverage, we include firm-specific unobserved effects (μ_i) to capture the impact of intertemporally constant, but unmeasured, effects on each firm's target leverage. We find that these unobserved effects explain a large proportion of the cross-sectional variation in target debt ratios, without displacing the other firm characteristics in $X_{i,t}$. At the same time, however, firm fixed effects complicate the estimation problem by making the regression (4) a dynamic panel model (Bond, 2002). We discuss some of these econometric issues in the next section, and provide further details in the appendix.

4. Partial adjustment and the tradeoff theory

4.1. Appropriate estimation techniques

The first column in Table 2 presents Fama and MacBeth (1973) (FM) estimates of (4).⁴ Most of the lagged variables representing the target debt ratio carry significant coefficients with appropriate signs. (Only **MB** and **LnTA** have insignificant coefficients.) The coefficient on lagged **MDR** implies that firms close 13.3% ($= 1 - 0.867$) of the gap between current and desired leverage within one year. At this rate, it takes approximately five years to close half the gap between a typical firm's current and desired leverage ratios. This slow adjustment is consistent with the hypothesis that other considerations—e.g., pecking order or market timing – outweigh the cost of deviating from optimal leverage. With such a low estimated adjustment speed, convergence toward a long-run target seems unlikely to explain much of the variation in firms' debt ratios.

While the FM estimates have some attractive features, they fail to recognize the data's panel characteristics. A panel regression with unobserved (fixed) effects is more appropriate if firms have relatively stable, unobserved variables affecting their leverage targets. Column (2) of Table 2 reports a fixed effects panel regression, whose estimated coefficients on the determinants of target leverage generally resemble their FM counterparts, except for **LnTA**. The statistical significance of most variables is greater, and the fixed effects on target **MDR** are well justified: an F-test for the joint significance of the unobserved effects in column (2) rejects the hypothesis that these terms are equal across all firms ($F(12918, 98178) = 2.24$; $pr = 0.000$). A prominent difference between columns (1) and (2) are the estimated coefficients on lagged **MDR**, which indicate a substantially faster adjustment speed (38%) in the panel model. This estimated adjustment speed implies that the typical firm closes half of a leverage gap in about 18 months.

⁴Fama and French (2002) recommend FM estimators to avoid understating coefficient standard errors. OLS yields similar coefficient estimates for similar specifications, as shown in column (2) of Table 3 or column (2) of Table A.1 in the appendix.

Table 2

Alternate estimation methods for specification (4)

Regression results for the model

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + (\hat{\lambda} - \lambda)MDR_{i,t} + \delta_{i,t+1}, \quad (4)$$

where MDR is the market debt ratio. The (lagged) “ X ” variables determine a firm’s long-run target debt ratio, and include:

EBIT_TA: earnings before interest and taxes as a proportion of total assets;

MB: the market-to-book ratio of firm assets;

DEP_TA: depreciation expense as a proportion of total assets;

LnTA: natural log of total assets;

FA_TA: fixed assets as a proportion of total assets;

☐ **R&D_DUM**: dummy variable indicating that the firm did not report R&D expenses;

R&D_TA: R&D expenses as a proportion of total assets;

Ind_Median: median debt ratio of firm i ’s Fama and French (2002) industry classification at time t ; and

Rated: dummy variable equal to one if the firm has a public debt rating in Compustat, zero otherwise.

Models (2) and (4)–(7) include firm fixed effects and models (4)–(7) include year dummies. T -statistics are shown in parentheses. Reported R^2 numbers for models including fixed effects are “within” R^2 statistics.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	FM	FE panel	FM Demeaned	FE Panel (with year dummy)	IV panel	IV panel, Middle 50th percentile	“Base” specification
MDR_{i,t}	0.867 (67.01)	0.620 (218.03)	0.639 (53.63)	0.620 (225.14)	0.656 (172.42)	0.636 (67.19)	0.656 (171.58)
EBIT_TA	−0.035 (−3.97)	−0.037 (−11.80)	−0.051 (−4.70)	−0.039 (−12.96)	−0.030 (−9.64)	−0.039 (−7.64)	−0.030 (−9.66)
MB	−0.001 (−1.53)	0.000 (−0.34)	−0.001 (−1.45)	−0.001 (−3.38)	0.000 (−0.68)	0.002 (2.71)	0.000 (−0.81)
DEP_TA	−0.225 (−7.59)	−0.280 (−13.46)	−0.338 (−7.67)	−0.224 (−10.97)	−0.226 (−11.07)	−0.209 (−6.61)	−0.226 (−11.06)
LnTA	0.000 (−0.43)	0.026 (38.22)	0.025 (14.42)	0.027 (37.52)	0.026 (34.56)	0.034 (30.28)	0.025 (34.00)
FA_TA	0.022 (2.68)	0.058 (12.82)	0.066 (10.33)	0.059 (13.42)	0.053 (11.85)	0.058 (8.33)	0.053 (11.93)
R&D_DUM	0.005 (3.97)	−0.006 (−3.87)	0.000 (0.23)	0.000 (−0.14)	0.000 (−0.01)	0.001 (0.35)	0.000 (0.02)
R&D_TA	−0.081 (−3.56)	−0.038 (−3.80)	−0.072 (−3.10)	−0.036 (−3.63)	−0.025 (−2.55)	−0.074 (−4.09)	−0.025 (−2.57)
Ind_Median	0.063 (5.71)	0.054 (9.89)	0.087 (4.43)	0.050 (6.34)	0.034 (4.29)	0.028 (2.42)	0.034 (4.30)
Rated							0.003 (1.71)
Fixed effects?	No	Yes	No	Yes	Yes	Yes	Yes
N	111,106	111,106	111,106	111,106	111,106	55,526	111,106
R^2	0.756	0.426	0.462	0.467	0.466	0.330	0.466

The more rapid adjustment speed in column (2) might reflect either the addition of firm fixed effects to the target specification, or the panel regression constraint that the slope coefficients remain constant over time. To distinguish between these two possibilities, the regression in column (3) applies the FM method to de-meaned data. That is, each variable is expressed as a deviation from that firm’s mean value. Most of the FM estimates in

column (3) are very close to the panel results in column (2). We conclude that firm-specific unobserved effects substantially influence estimated adjustment speeds, apparently because they substantially sharpen estimates of the target debt ratio.⁵ We return to this issue in Section 4.3.

Column (4) estimates a revised panel model, which includes a separate dummy variable for each year in the sample (except 1966, to avoid a dummy variable trap). The resulting (within) adjusted- R^2 statistic rises slightly from column (2) and the other coefficients remain essentially the same. We include year dummy variables in our subsequent panel regressions to absorb any unmodeled time-varying influences on capital structure. We also estimate this specification with a correction for first-order serial correlation within each panel (not reported). The estimated AR(1) coefficient is sufficiently small (-0.03) that we proceed under the assumption that serial correlation is not a significant effect in our study.

Consistently estimating the adjustment speed in a dynamic panel requires careful attention to the serial correlation properties of the dependent variable and the regression's residuals (Baltagi, 2001, Chapter 8; Wooldridge, 2002). Column (5) addresses the correlation between a panel's lagged dependent variable and the error term, which can bias the estimated adjustment speed. We substitute a fitted value for the lagged dependent variable, using the lagged book value of leverage and X_t as instruments (Greene, 2003).⁶ The estimated MDR_t coefficient rises slightly (from 0.620 to 0.656) but the other coefficient estimates remain close to the estimates in column (4). The implied adjustment speed of 34.4% indicates that the typical firm completes more than half of its required leverage adjustment in less than two years—far faster than estimated by many previous authors. Such a rapid adjustment toward a firm-specific capital ratio suggests that pecking order or market timing does not dominate most firms' debt ratio decisions. We return to this issue in Table 4 below.

Column (6) of Table 2 addresses the possibility that the rapid adjustment speed in column (5) reflects the bounded nature of $MDR_{i,t}$ between zero and unity. A firm with a very high leverage thus has nowhere to go but down, and vice versa. Column (6) reports the results of estimating our instrumental variables specification for only the middle 50% of observed $MDR_{i,t}$ values. The 25th and 75th percentile cutoffs for $MDR_{i,t}$ vary across years, but average 6.5% and 41.6%, respectively. All of the coefficient estimates in column (6), including the adjustment speed, are very similar to the results using the entire sample. We are therefore confident that “hard-wired” mean reversion in the dependent variable is not the cause of our high estimated adjustment speeds.

The last column of Table 2 presents our “base” specification that is used going forward. This specification includes an instrumental variable correction for MDR_t . Explanatory variables include firm and time fixed effects, plus an additional explanatory variable in the “ X ” matrix:

Rated equals unity when a firm has a public debt rating, and zero otherwise.⁷

⁵When we replace the firm fixed effects in column (2) with a set of 46 industry dummy variables (constructed as in Fama and French, 1997), the estimated coefficients closely resemble those in column (1), which also excludes firm fixed effects.

⁶More recent estimation techniques like that of Arellano and Bond (1991) improve upon this approach under some circumstances, but not for our sample. See the appendix for details.

⁷Because Compustat does not report this variable before 1981, we cannot compute Fama–MacBeth estimates comparable to the other specifications in Table 2 if **Rated** is included.

Faulkender and Petersen (2005) control for sample selectivity in their paper because *Rated* may be endogenous. We simply include *Rated* as an additional dependent variable, for two reasons. First, the impact of bond ratings is not our central concern. Second, the other results are completely insensitive to the inclusion or exclusion of *Rated* from the set of variables determining a firm's target debt ratio. This dummy variable carries a marginally significant positive coefficient (as in Faulkender and Petersen, 2005), but its introduction has no meaningful effect on the other coefficient estimates.

Our base specification in column (7) indicates that the typical firm's target debt ratio varies quite a lot. The cross-sectional mean target debt ratio starts at 32.1% in 1966, rises to a maximum of 64.0% in 1974, and ends the period at 27.0% in 2001. Over the entire sample, the estimated target has an average of 30.7% and a standard deviation of 25.1%. (In comparison, the actual *MDR*'s mean and standard deviation are 27.8% and 24.4%, respectively.) Firm characteristics, fixed effects, and time all contribute to the variation in target debt ratios. The set of nine *X* variables explain 16.0% of the total sample standard deviation of *MDR*, the unobserved (fixed) effects explain 25.2%, and the year dummies explain 9.0%. Within each year, the nine *X* variables alone explain between 12.5% and 17.6% of the annual, cross-sectional variations in target debt ratios, with an average (across all years) of 15.03%. In short, our computed leverage targets vary substantially across firms and across time.

4.2. Convergence toward the target

If we estimate meaningful leverage targets, we should find that firms adjust toward these targets over time. Fig. 1 illustrates managers' financing decisions conditional on the firm's deviation from its computed (estimated) target leverage. For each year between 1966 and 2000, we sort firms into quartiles on the basis of their deviations from target leverage ($MDR^* - MDR$). The horizontal axis in Fig. 1 indicates that the firms in Quartile 1 appear to be substantially overleveraged, by an average (median) of 15.29% (13.59%) of assets. Conversely, our model indicates that the firms in Quartile 4 are underleveraged by a mean (median) of 21.96% (19.70%). The vertical axis in Fig. 1 describes the subsequent year's change in book debt ratios (*BDR*), which should reflect the firm's explicit efforts to move toward its target. (In contrast, *MDR* confounds the effects of managerial actions and changes in the firm's stock price.) The evidence in Fig. 1 is consistent with convergence. The mean (median) overleveraged firm in Quartile 1 reduces its book leverage the following year by 3.12% (1.94%). Conversely, the underleveraged firms in Quartile 4 raise their *BDR* by a mean (median) of 4.69% (1.75%) during the subsequent year. Firms in the middle two quartiles also move toward their target debt ratios, but with much smaller adjustments.

While the results in Fig. 1 are consistent with targeting behavior, they might reflect merely a tendency of firms with relatively high or low debt ratios to move back toward the mean, as indicated by Leary and Roberts' (2005) hazard function estimates. Indeed, Fig. 2 illustrates this tendency in the data. The horizontal axis describes four quartiles formed on the basis of the prior year's absolute *MDR*. As in Fig. 1, the vertical axis of Fig. 2 plots the subsequent year's mean and median changes in book debt ratio (*BDR*). Independent of their position relative to their target, highly levered firms tend to reduce their book

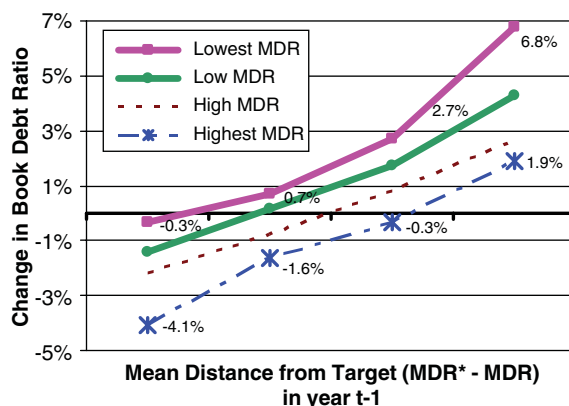


Fig. 3. Subsequent year's change in book debt ratio.

leverage the following year. Conversely, firms in the lowest *MDR* quartile tend to increase their *BDR* during the subsequent year.⁸

How much of the targeting behavior in Fig. 1 reflects this general tendency for extremely levered firms to revert toward the mean? We evaluate this question using a two-way sort of the data. First, we form four quartiles based on absolute leverage (MDR_{t-1}) as in Fig. 2. Within each leverage quartile, we construct quartiles based on the firm's deviation from its target leverage. Each line in Fig. 3 then plots the change in *BDR* against the prior year's deviation from target (according to our model) for a set of firms with similar absolute *MDR* values. As in Fig. 1, the tradeoff theory implies that firms to the left (right) on the horizontal axis in Fig. 3 are overleveraged (underleveraged) and should be acting to reduce (increase) leverage in the subsequent year. This is exactly what we find. Regardless of their absolute leverage, the most overleveraged firms reduce their *BDR*. Firms with high absolute leverage move toward their target more quickly than those with low absolute leverage, suggesting that deviations from target are more costly for more highly leveraged firms. At the other extreme, the mean underleveraged firm raises *BDR* regardless of its absolute leverage level. Among these underleveraged firms, those with the lowest absolute leverage act most aggressively to increase *BDR*.

4.3. Previous estimates of optimal capital structure

The rapid adjustment speed estimated in Table 2 (34.4% per year) constitutes the most notable feature of our empirical results. Although some prior research has supported such rapid adjustment, the conventional wisdom holds that a firm's annual adjustment speed lies in the neighborhood of 8% to 15%, which Fama and French (2002) consider insufficient for the tradeoff theory to explain the range of observed variation in firms' leverage data. Previous research uses a variety of regression specifications to study the determinants of a firm's capital structure. Why should our specification be preferred? We

⁸Fig. 2 is not inconsistent with targeting behavior, since firms with high (low) leverage are more likely to be above (below) their target leverage.

contend that most previous studies impose unwarranted, but testable, assumptions on the data, which have led to incorrect or misleading inferences.

Table 3 reports a set of capital structure models estimated over the same panel data set. Column (1) presents a typical simple, cross-sectional specification used by many prior studies to infer the determinants of a firm's optimal leverage (e.g., Hovakimian et al., (2001); Fama and French (2002); Korajczyk and Levy (2003); Kayhan and Titman (2005)).⁹ The estimated coefficients resemble those of the earlier studies. In particular, higher earnings, *MB* ratio, R&D expenditures, and depreciation expenses lower target leverage, while asset size and fixed assets raise it.

The specification in column (1) constrains the coefficient on lagged *MDR* to be zero. In other words, a firm's observed capital ratio is also its desired (target) ratio. Column (2) indicates that this hypothesis is strongly rejected by the data. When we add the lagged dependent variable to the specification in column (1), it carries a very highly significant coefficient (0.864). The simple cross-sectional regression in column (1) thus appears to omit an important variable. We also know from Table 2 that column (2)'s exclusion of firm fixed effects is unwarranted.

Column (3) specifies partial adjustment toward a target capital ratio that includes firm fixed effects.¹⁰ Several of the estimated coefficients differ substantially from those for the simple linear model in column (1). For example, column (1) indicates that the coefficient on *EBIT_TA* is -0.282 , more than three times the estimated long-run magnitude in column (3) ($(-0.03/0.343) = -0.087$). The long-run coefficients on *MB*, *R&D_Dum*, *R&D_TA*, and *Rated* are also substantially lower in column (3), while the long-run impact of firm size (*LnTA*) increases by a factor of six. It appears, therefore, that the estimated determinants of target leverage are materially affected by the omitted variables in column (1).¹¹

Regressions like that in column (1) are sometimes used to generate leverage target proxies for use in partial adjustment models. Two-stage estimates based on such a proxy have largely formed the conventional wisdom that firms adjust slowly toward any leverage target. Column (4) of Table 3 presents an estimated partial adjustment model based on a target debt ratio (*TDR*_{OLS}), which is computed from column (1). The estimated adjustment speed (9.1%) resembles other estimates derived from target proxies containing no fixed effects. This alone does not indicate a problem with the two-stage estimation. However, the coefficient on *TDR*_{OLS} is much smaller than theory would predict. The longrun elasticity of observed *MDR* with respect to its target should be unity. Here, it is only 0.56 ($= 0.051/0.091$), which differs from unity at a very high confidence level

⁹Hovakimian et al. (2001) and Korajczyk and Levy (2003) estimate a target leverage using simple OLS, then use deviations from their computed targets to help predict whether a firm subsequently issues debt or equity.

¹⁰This regression omits one determinant of the target debt ratio from our base specification, namely, the lagged industry median *MDR*. We omit this variable in Table 3 in order to provide a cleaner test of the two-stage approach to estimating a partial adjustment coefficient.

¹¹Hovakimian et al. (2004) estimate a cross-sectional regression like our column (1) for a set of firms that have recently issued large amounts of both debt and equity, and find that the estimated coefficients differ substantially from those estimated for the rest of the Compustat universe. The authors argue that the security issuers have moved close to their optimal leverage ratios, while the other firms are scattered more widely relative to their target capital ratios.

Table 3

The importance of recognizing partial adjustment

Regression results for the models

$$(1) MDR_{i,t+1} = \beta X_{i,t} + \delta_{i,t+1},$$

$$(2) MDR_{i,t+1} = (\lambda \beta) X_{i,t} + (1-\lambda) MDR_{i,t} + \delta_{i,t+1},$$

$$(3) MDR_{i,t+1} = (\lambda \beta) X_{i,t} + (1-\lambda) MDR_{i,t} + v_i + \delta_{i,t+1},$$

$$(4) MDR_{i,t+1} = \lambda (TDR_{OLS,i}) + (1-\lambda_1) MDR_{i,t} + \delta_{i,t+1},$$

$$(5) MDR_{i,t+1} = \beta_1 L3MDR_{i,t} + (1-\lambda_1) MDR_{i,t} + \delta_{i,t+1},$$

where *MDR* is the market debt ratio. The (lagged) “*X*” variables determine a firm’s long-run target debt ratio, and include:

EBIT_TA: earnings before interest and taxes as a proportion of total assets;

MB: the market-to-book ratio of firm assets;

DEP_TA: depreciation expense as a proportion of total assets;

LnTA: natural log of total assets;

FA_TA: fixed assets as a proportion of total assets;

R&D_DUM: dummy variable indicating that the firm did not report R&D expenses;

R&D_TA: R&D expenses as a proportion of total assets; and

Rated: dummy variable equal to one if the firm has a public debt rating in Compustat, and zero otherwise.

L3MDR is the average of the three-years’ lagging *MDR* values, and *TDR_{OLS}* is the fitted value from model (1). All regressions include (unreported) year dummies. *T*-statistics are shown in parentheses. Reported *R*² numbers for models including fixed effects are “within” *R*² statistics.

	(1)	(2)	(3)	(4)	(5)
<i>MDR_{i,t}</i>		0.864 (469.79)	0.657 (174.09)	0.909 (382.34)	0.935 (194.48)
<i>EBIT_TA</i>	-0.282 (-75.25)	-0.026 (-11.62)	-0.030 (-9.73)		
<i>MB</i>	-0.041 (-103.01)	-0.002 (-6.84)	0.000 (-1.07)		
<i>DEP_TA</i>	-0.566 (-24.87)	-0.216 (-16.38)	-0.225 (-11.03)		
<i>LnTA</i>	0.011 (26.39)	-0.001 (-3.11)	0.025 (33.85)		
<i>FA_TA</i>	0.128 (36.32)	0.027 (13.16)	0.054 (12.19)		
<i>R&D_DUM</i>	0.035 (24.33)	0.007 (8.73)	0.000 (0.03)		
<i>R&D_TA</i>	-0.518 (-50.79)	-0.116 (-19.51)	-0.026 (-2.64)		
<i>Rated</i>	0.066 (26.45)	0.005 (3.68)	0.003 (1.75)		
<i>TDR_{OLS}</i>				0.051 (12.31)	
<i>L3MDR</i>					-0.023 (-4.85)
Fixed effects?	No	No	Yes	No	No
<i>N</i>	111,106	111,106	111,106	111,106	81,343
<i>R</i> ²	0.262	0.753	0.466	0.750	0.767

($t = 11.48$).¹² This provides strong evidence against the two-step estimation procedure used previously in the literature.

Target leverage has also been proxied by the trailing average of a firm's actual leverage.¹³ In column (5), we test whether **L3MDR**, a three-year trailing average **MDR**, provides an adequate proxy for target capital. The estimated adjustment speed is now only 6.5%, and the impact of a one-unit increase in **L3MDR** actually reduces a firm's longrun debt ratio. Thus, it appears that **L3MDR** measures target leverage quite poorly.

We now investigate more formally how target measurement noise affects a partial adjustment model. Could measurement error alone account for the difference between the adjustment speeds in columns (2) and (3) of Table 3? Substitute a noisy proxy for **MDR*** into Eq. (3) to obtain

$$\underline{\mathbf{MDR}_{i,t+1}} = (1 - \lambda)\mathbf{MDR}_{i,t} + \lambda(\underline{\mathbf{MDR}_{i,t+1}^*} + \tilde{\xi}_{i,t}) + \tilde{\delta}_{i,t+1}, \quad (3a)$$

where $\tilde{\xi}_{i,t}$ is a standard normal variate with zero mean. Adding noise to an explanatory variable usually biases the associated coefficient toward zero, which implies an estimated coefficient on $\mathbf{MDR}_{i,t}$ biased toward unity. To quantify this effect, we assume that $\mathbf{MDR}^* = \mathbf{TDR}_{\text{Panel}}$, a target series constructed from the estimated coefficients in column (3) of Table 3. We then increase the standard deviation of $\tilde{\xi}_{i,t}$ from 0.0 to 0.5 across the columns of Table 4.¹⁴ The results indicate that a noisier target measure lowers both the estimated adjustment speed (from 0.345 to 0.104) and the long-run effect on **MDR** of a change in the target's value (from 1.00 to 0.31).

Which column of Table 4 is most relevant for our data set? Over the entire sample period, the observed **MDR** series has a standard deviation of 24.4%. A good proxy should be similarly distributed and $\mathbf{TDR}_{\text{Panel}}$ has a standard deviation of 25.1%. In contrast, the **TDR**_{OLS} series' standard deviation is much smaller (12.5%). The standard deviation of the difference between these two target proxies ($\mathbf{TDR}_{\text{Panel}} - \mathbf{TDR}_{\text{OLS}}$) is about 22%, which lies between the assumed noise volatilities in columns (4) and (5) of Table 4. We therefore see that a noise volatility of 20% to 25% roughly halves the estimated adjustment speed; from 34.5% ($\approx 1 - 0.656$) to something in the neighborhood of 17%.¹⁵

We conclude from Tables 3 and 4 that partial adjustment and firm fixed effects should be included in a model of firm capital structure choice. A few previous studies include such

¹²The importance of fixed effects in computing leverage targets can be assessed by reestimating the regression in column (1) with firm fixed effects. When the resulting fitted values are used as leverage targets, the adjustment speed in column (4) rises to 40% and its long-run impact on **MDR** is 1.05.

¹³Marsh (1982); Jalilvand and Harris (1984); and Shyam-Sunder and Myers (1999) have used various forms of this proxy, with mixed results. In Section 5.3 below, we point out that Welch's (2004) main regression specification can be viewed as a single lag of the dependent variable as a leverage target.

¹⁴The $\mathbf{TDR}_{\text{Panel}}$ series is itself a noisy estimate of the true target, so $\tilde{\xi}_{i,t}$ measures additional "noise," not the total estimation error.

¹⁵Further qualitative evidence that measurement error depresses estimated adjustment speeds comes from comparing various versions of Kayhan and Titman (2005). Although their focus is quite different from ours, they estimate a regression of the form

$$\mathbf{MDR}_t - \mathbf{MDR}_{t-k} = \alpha X + \beta(\mathbf{MDR}_{t-k} - \mathbf{TDR}_{t-k}) + \varepsilon_t, \quad (6)$$

where \mathbf{TDR}_{t-k} is the target debt ratio computed from a cross-sectional OLS regression using firm characteristics from year $(t-k-1)$. When $k = 5$ (as it does in Table 6 of their November 2003 manuscript), $\hat{\beta} = 0.45$, which implies an annual leverage adjustment rate of 7.7%. In the corresponding table of their May 2004 revision, they set $k = 10$ and the estimated annual adjustment speed falls to 4.6%. Although this change has no effect on their variables of interest, it does illustrate the effect of noisy targets on the estimated adjustment speed.

Table 4

Effect of Adding Noise to the Target Debt Ratio

Regression results for the model

$$MDR_{i,t+1} = (1 - \lambda)MDR_{i,t} + \lambda(MDR_{i,t+1}^* + \tilde{\xi}_{i,t}) + \tilde{\delta}_{i,t+1}. \quad (3a)$$

where MDR is the market debt ratio, MDR^* is the estimated target debt ratio from model (3) of Table 3, and $\tilde{\xi}_{i,t}$ is a white noise term with zero mean. Models (1) through (6) are estimated at various levels of standard deviation for $\tilde{\xi}_{i,t}$. T -statistics are shown in parentheses.

	(1)	(2)	(3)	(4)	(5)	(6)
Standard deviation of $\tilde{\xi}_{i,t}$	0%	5%	10%	20%	25%	50%
MDR_t	0.656 (230.82)	0.680 (246.07)	0.731 (278.24)	0.818 (345.06)	0.844 (370.89)	0.896 (432.96)
$MDR_{i,t+1}^* + \tilde{\xi}_{i,t}$	0.345 (145.05)	0.313 (138.98)	0.248 (123.97)	0.130 (90.45)	0.100 (78.10)	0.033 (46.19)
N	111,106	111,106	111,106	111,106	111,106	111,106
R^2	0.811	0.806	0.795	0.774	0.768	0.755
Long-run effect of target	1.000	0.979***	0.920***	0.729***	0.632***	0.312***

***Significantly different from one at the 1% level.

features in their regression models and produce rapid estimated adjustment speeds. For example, Marcus (1983) estimates a panel model with firm fixed effects for large U.S. banks over the period 1965–1977. His estimated adjustment speed for market leverage is 20–24% per year for the full sample. For the 1965–1971 subperiod, he estimates annual adjustment speeds as high as 32.5%. Roberts (2002) estimates even higher adjustment speeds in his Kalman filter model of partial adjustment. Because the standard Kalman filter specifies that all variables have zero means, he de-means each data series, which “implicitly...accounts for firm-specific effects in the intercepts” (p. 13). Using quarterly data over the period 1980–1998, he estimates a separate model similar to Eq. (4) for the firms in each of 53 industries. His results imply annual adjustment speeds (λ) ranging from a low of 18% to a high of more than 100%.¹⁶

In a multifaceted paper, Leary and Roberts (2005) use a quarterly Compustat data set (1984–2001) to estimate hazard functions for substantial net debt or equity adjustments.¹⁷ While their primary concern is to infer the form of capital adjustment costs (e.g., fixed vs. proportional vs. convex), but they indirectly address the question of adjustment speeds. They find that a typical firm changes the book value of its debt (equity) by more than 5% of book assets *about once per year*, and conclude that “Firms do indeed respond to equity issuances and equity price shocks by appropriately rebalancing their leverage over the next one to four years” (p. 32). If we define “appropriately rebalancing” as closing 90% of the initial leverage gap, “one to four years” corresponds to an adjustment speed in Eq. (4) that exceeds 40%.

¹⁶Across all industries, the mean annual adjustment speed is approximately 43%. His equal-weighted average of 53 industries’ adjustment speeds cannot be compared directly to our estimated λ . We weight each sample firm equally, which gives different industries different weights in our estimate of λ . Still, it is comforting to learn that applying Kalman filters to a comparable data set yields roughly similar results.

¹⁷They describe the estimated model as “similar in spirit to a nonlinear dynamic panel regression *with firm-specific random effects*” (p. 17, emphasis added).

[Alti \(2004\)](#) estimates relatively rapid adjustment speeds without fixed effects in his target variable specification. He compares the capital ratios chosen by 1,363 firms that go public in “hot” vs. “cold” issue markets. He finds that firms making their IPOs in a hot (cold) market tend to have unusually low (high) leverage immediately after going public. He then estimates a model similar to Eq. (4), but without fixed effects, for one and two years after the IPO date. The estimated adjustment speed is surprisingly rapid for a model without fixed effects: 30% per year. We conjecture that this rapid adjustment reflects his sample: rapidly growing, young firms tend to seek additional financing soon after their IPO and hence confront unusually low costs of adjusting leverage.

In conclusion, our model specification is theoretically preferable to many earlier models, and the specification differences account for our main results: firms adjust rapidly toward time-varying target leverage ratios, which depend on plausible firm features.

5. Other capital structure theories

The pecking order model has strong intuitive appeal and a long pedigree. Various forms of the model have been recently studied, with mixed empirical results ([Shyam-Sunder and Myers, 1999](#); [Frank and Goyal, 2003](#); [Lemmon and Zender, 2004](#); [Halov and Heider, 2004](#)). The literature also includes two more recent explanations of capital structure, namely, [Baker and Wurgler’s \(2002\)](#) market timing theory (also studied in [Huang and Ritter, 2005](#)) and [Welch’s \(2004\)](#) mechanical stock price explanation. We now compare our results to these three alternatives.

5.1. The pecking order and market timing theories: regression evidence

Previous authors estimate models in which the financing deficit (pecking order) or the weighted sum of past market-to-book ratios (market timing) compete with variables associated with the tradeoff theory of capital structure. The idea is that the variables associated with the “true” theory are more important than their competitors. [Frank and Goyal \(2003\)](#) explain:

The pecking order theory implies that the financing deficit ought to wipe out the effects of other variables. If the financing deficit is simply one factor among many that firms tradeoff, then what is left is a generalized version of the tradeoff theory. (p. 219)

[Baker and Wurgler \(2002\)](#) make similar statements, particularly concerning their Table III.

Pecking order behavior implies that a firm’s financing deficit explains contemporaneous changes in its book debt ratio. We test this proposition by evaluating a book-value analog to our main specification in Eq. (4):¹⁸

$$\Delta BDR_{i,t+1} = (\lambda\beta)X_{i,t} - \lambda BDR_{i,t} + \gamma_2 \underline{FINDEF}_{i,t+1} + \varepsilon_{i,t+1}, \quad (7)$$

where **BDR** is the ratio of (long-term plus short-term debt) to total assets, and **FINDEF** is the firm’s financing deficit. In their Table 2, [Frank and Goyal \(2003\)](#) define **FINDEF** as:

¹⁸The bias associated with including the lagged dependent variable in a panel regression applies to (7) and (8) as well as it does to (4). In these cases, we instrument for lagged **BDR** using the firm’s lagged **MDR**.

(dividend payments + investments + change in working capital – internal cashflow)/(total assets).

Baker and Wurgler (2002) assert that managers issue relatively overvalued securities, which can be either debt (when the firm's q -ratio is low) or equity (when q is high). They construct a backward-looking “external finance weighted average” book-market ratio, which they find to be correlated with a firm's book debt ratio over periods as long as ten years. We test their hypothesis here by estimating the following regression for the level of BDR:

$$BDR_{i,t+1} = (\lambda\beta)X_{i,t} - (1 - \lambda)BDR_{i,t} + \gamma_1 MB_EFWA_{i,t} + \varepsilon_{i,t+1}, \quad (8)$$

where MB_EFWA is the firm's external finance weighted average book-market ratio defined on 12 of Baker and Wurgler (2002).

In both regressions (7) and (8), the question is whether MB_EFWA or $FINDEF$ affects the estimated coefficients on $X_{i,t}$ or the lagged dependent variable. To set a baseline for the BDR regressions, the first column in Table 5 reports the results of explaining BDR with only our standard $X_{i,t}$ (including firm fixed effects), year dummies, and a lagged dependent variable. The partial adjustment model fits BDR well: the lagged dependent variable's coefficient implies a rapid adjustment (36.1% annually) and the variables meant to capture target leverage carry significant coefficients with the appropriate signs. Adding Baker and Wurgler's (2002) external finance weighted market-book ratio as an explanatory variable in column (2) lowers the estimated adjustment speed only slightly (from 36.1% to 34.2%). The other coefficients remain basically unchanged, except that the simple MB coefficient loses significance. The Baker and Wurgler (2002) variable is at best marginally significant (p -value = 0.093, two-tailed test), perhaps because the market-to-book effect is split between the two variables (Hovakimian, 2003).

The specification in column (3) explains the change in BDR with changes in our standard explanatory variables and the firm's financing deficit. The $FINDEF$ coefficient is significantly positive (t -statistic = 9.96), but does not substantially alter the other variables' signs and significance levels. Thus, the pecking order forces appear to be part of a “generalized version of the tradeoff theory” (Frank and Goyal, 2003, p. 219), rather than a unique determinant of financial leverage.

We next include both MB_EFWA and $FINDEF$ in the same regression. This is probably an inappropriate specification because MB_EFWA is meant to explain the level of BDR while $FINDEF$ affects the change in BDR . Nevertheless, the estimated coefficients in column (4) resemble their values in columns (2) and (3), and the tradeoff variables retain their usual values.

Although both the pecking order and market timing hypotheses apply to book-valued debt ratios, columns (5)–(7) of Table 5 incorporate MB_EFWA and $FINDEF$ into regressions explaining the market debt ratio. The market-timing variable (MB_EFWA) in column (5) now carries a significant coefficient with the proper sign. As in column (2), the simple MB variable again loses significance. The pecking order conclusions in column (6) replicate those for ΔBDR in column (3): $FINDEF$ carries a significantly positive coefficient but does not displace the tradeoff-related variables from the regression. Once again, including both of the competing theories' variables in column (7) yields significant coefficients without substantially changing the estimated coefficients on the tradeoff or partial adjustment variables.

Table 5

Pecking order and market timing explanations of book debt ratio

In columns (1) (2), and (4), we estimate a regression that explains a firm's book debt ratio:

$$BDR_{i,t+1} = (\lambda\beta)X_{i,t} + (1 - \lambda)BDR_{i,t} + \gamma_1 Z_{i,t} + e_{i,t+1}.$$

In column (4), we estimate a regression that explains a firm's change in book debt ratio:

$$\Delta BDR_{i,t+1} = (\lambda\beta)X_{i,t} - \lambda BDR_{i,t} + \gamma_2 Z_{i,t} + e_{i,t+1},$$

In columns (5) and (7) we estimate our main regression specification that explains a firm's market debt ratio:

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} - (1 - \lambda)MDR_{i,t} + \gamma_1 Z_{i,t} + \delta_{i,t+1}. \tag{4}$$

In column (6) we estimate a regression that explains a firm's change in market debt ratio:

$$\Delta MDR_{i,t+1} = (\lambda\beta)X_{i,t} - \lambda MDR_{i,t} + \gamma_2 Z_{i,t} + \delta_{i,t+1}.$$

The (lagged) “X” variables determine a firm's long-run target debt ratio, and include:

EBIT_TA: earnings before interest and taxes as a proportion of total assets;

MB: the market-to-book ratio of firm assets;

DEP_TA: depreciation expense as a proportion of total assets;

LnTA: natural log of total assets;

FA_TA: fixed assets as a proportion of total assets;

R&D_DUM: dummy variable indicating that the firm did not report R&D expenses;

R&D_TA: R&D expenses as a proportion of total assets;

Ind_Median: median debt ratio of firm *i*'s Fama and French (2002) industry classification at time *t*; and

Rated: dummy variable equal to one if the firm has a public debt rating in Compustat, and zero otherwise.

Z_t are additional explanatory variables and include:

FINDEF = a measure of the firm's financial deficit, defined as dividend payments + investments + change in working capital – internal cashflow; or

MB_EFWA = an “external finance weighted average” market-book ratio defined by Baker and Wurgler (2002, p. 12).

All models include (unreported) year dummies. *T*-statistics are shown in parentheses. Reported *R*² numbers for models including fixed effects are “within” *R*² statistics.

Panel A: Estimation results

Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<i>BDR</i>	<i>BDR</i>	<i>BDR</i>	Δ <i>BDR</i>	<i>BDR</i>	<i>MDR</i>	Δ <i>MDR</i>	<i>MDR</i>
<i>BDR_{i,t}</i>	0.639 (167.21)	0.658 (166.42)	-0.375 (-90.32)	0.645 (150.22)			
<i>MDR_{i,t}</i>					0.670 (163.84)	-0.341 (-86.53)	0.645 (146.10)
<i>EBIT_TA</i>	-0.024 (-9.48)	-0.026 (-8.57)	-0.018 (-6.61)	-0.019 (-5.89)	-0.030 (-7.70)	-0.021 (-6.63)	-0.020 (-4.69)
<i>MB</i>	-0.001 (-2.88)	0.000 (-0.07)	-0.001 (-4.90)	-0.001 (-2.06)	0.001 (1.53)	-0.001 (-2.20)	-0.001 (-1.54)
<i>DEP_TA</i>	-0.194 (-11.32)	-0.191 (-10.01)	-0.144 (-7.67)	-0.152 (-7.28)	-0.275 (-11.35)	-0.155 (-7.33)	-0.205 (-7.79)
<i>LnTA</i>	0.008 (13.42)	0.009 (13.36)	0.008 (12.22)	0.008 (11.13)	0.026 (31.27)	0.026 (33.12)	0.026 (28.30)
<i>FA_TA</i>	0.042 (11.17)	0.041 (10.15)	0.041 (9.92)	0.038 (8.59)	0.052 (10.36)	0.057 (12.24)	0.056 (10.14)
<i>R&D_DUM</i>	-0.001 (-0.79)	-0.001 (-1.01)	-0.001 (-0.75)	-0.001 (-0.77)	-0.001 (-0.33)	0.001 (0.48)	0.001 (0.40)
<i>R&D_TA</i>	-0.026 (-3.18)	-0.021 (-1.85)	-0.019 (-2.18)	-0.013 (-1.08)	-0.033 (-2.31)	-0.010 (-1.08)	-0.015 (-1.04)
<i>Rated</i>	0.008 (5.28)	0.006 (4.24)	0.008 (4.97)	0.007 (4.25)	0.001 (0.53)	0.003 (1.77)	0.003 (1.35)
<i>IND_Median</i>	0.032 (4.93)	0.030 (4.66)	0.032 (4.47)	0.031 (4.39)	0.031 (3.77)	0.030 (3.68)	0.037 (4.09)
<i>MB_EFWA</i>		-0.001 (-1.68)		-0.001 (-1.19)	-0.003 (-4.54)		-0.004 (-4.99)
<i>FINDEF</i>			0.021 (9.96)	0.028 (9.82)		0.023 (9.71)	0.048 (13.31)
Fixed effects?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>N</i>	111,106	94,235	98,709	83,790	94,235	98,709	83,790
<i>R</i> ²	0.384	0.420	0.198	0.398	0.475	0.240	0.443

Table 5 (continued)

Panel B: Relative economic significance of capital structure theories									
	Column (2) of Panel A		Column (3) of Panel A		Column (5) of Panel A		Column (6) of Panel A		
	Impact on <i>BDR</i>		Impact on ΔBDR		Impact on <i>MDR</i>		Impact on ΔMDR		
	Absolute	% of <i>BDR</i> 's Std. Dev.	Absolute	% of ΔBDR 's Std. Dev.	Absolute	% of <i>MDR</i> 's Std. Dev.	Absolute	% of ΔMDR 's Std. Dev.	
Tradeoff Theory	0.0617	32.94%	0.0711	65.49%	0.0676	27.82%	0.0748	60.14%	
Market Timing Theory	-0.0014	-0.73%	NA	NA	-0.007	-2.88%	NA	NA	
Pecking Order Theory	NA	NA	0.0047	4.31%	NA	NA	0.0022	1.77%	

Panel B of Table 5 assesses the economic significance of the tradeoff, market timing, and pecking order models by comparing their ability to explain variations in **BDR** and **MDR**. The first row in Panel B applies the coefficients from column (2) of Panel A. We calculate that changing the target leverage (**BDR***) by one standard deviation changes the (short-run) fitted **BDR** by 0.0617, which is about one-third of its standard deviation (0.1925). In contrast, changing **MB_EFWA** by one standard deviation lowers **BDR** by 0.0014 (less than 1% of its standard deviation). Using coefficient estimates from column (3) of Panel A, a one-standard deviation change in **BDR*** changes ΔBDR by 0.0711, while a similar change in **FINDEF** affects ΔBDR by only 0.0047. It therefore appears that variation in target leverage is much more important than mispricing or a financing deficit in explaining book debt ratios. Similar assessments of the three capital structure theories emerge from the last two columns in Panel B, which are based on the regressions in columns (5) and (6) of Panel A. Once again, the target leverage ratio explains far more of the variation in market debt ratios than either the market timing or the pecking order variables.

Although the pecking order and market timing theories each add some information to the regressions, we conclude that neither can replace our model of partial adjustment toward a target debt ratio. Targeting behavior consistent with the tradeoff theory seems to explain the bulk of the observed capital structure.

5.2. Targeting leverage vs. stockpiling debt capacity

Lemmon and Zender (2004) (LZ) suggest a modified version of the pecking order theory, in which each firm has a “debt capacity,” or maximum attainable leverage. Empirically, their debt capacity depends on a set of firm characteristics that overlap somewhat with our determinants of target leverage. As we do in Fig. 1, LZ classify a firm as “underleveraged” or “highly leveraged” relative to its estimated debt capacity. They argue that overleveraged firms have no choice but to reduce leverage because they cannot borrow further in the market. However, LZ observe that the tradeoff and pecking order theories have opposite implications for underleveraged firms (our Quartile 4 in Fig. 1). The tradeoff theory predicts that underleveraged firms should move toward their (higher) leverage target by issuing debt (if **FINDEF** > 0) or by retiring equity (if **FINDEF** < 0). In contrast, the LZ pecking order theory would have these firms “stockpile” debt capacity if **FINDEF** < 0 by using their excess cash exclusively to retire outstanding debt.

Table 6 examines the impact of **FINDEF** on leverage for the most overleveraged (underleveraged) firms defined in our Table 1. We explicitly discuss only the left-hand side of Table 6, which reports mean values. The Table’s right-hand side reports medians, which tell a similar story. Consider first the “Most overleveraged” firms. Row (1) indicates that excess leverage does not vary greatly with the subsequent period’s financing deficit. As predicted by LZ, the subsequent year’s change in **BDR** (row (2)) is negative for all three groups. Inconsistent with the pecking order theory, however, **BDR** declines by a similar proportion across a wide range of **FINDEF** values. LZ note that this “evidence” against the pecking order theory may arise simply because the capital market is unwilling to lend more to such highly levered firms.

In contrast, the underleveraged firms in rows (4)–(6) can issue additional debt freely. Under LZ’s form of the pecking order hypothesis, variations in **FINDEF** should therefore induce different amounts of new debt and hence substantially different increases in leverage. Yet we find that all three **FINDEF** groups in row (5) raised their **BDR** by

Table 6

Underleveraged firms' efforts to converge on targets

The table presents a two-way sort of our sample observations. Observations are first sorted into quartiles based on the distance from target leverage (MDR^*), where target leverage is estimated from model (7) of Table 2. Within each of the quartiles, firms are further sorted into terciles based on the financial deficit ($FINDEF$) values. $FINDEF$ is defined as dividend payments + investments + change in working capital—internal cashflow.

	Mean values			Median values		
	Quartile #1 of 4: Most overleveraged					
	<i>FINDEF</i> (<i>t</i> , <i>t</i> + 1)			<i>FINDEF</i> (<i>t</i> , <i>t</i> + 1)		
	Low (−6.51%)	Medium (1.17%)	High (19.78%)	Low (−3.97%)	Medium (0.58%)	High (13.27%)
(1) Distance from target @ <i>t</i> : (<i>MDR</i> *− <i>MDR</i>)	−15.90%	−14.74%	−15.40%	−14.00%	−12.88%	−13.57%
(2) $\Delta BDR_{i,t+1}$	−4.03%	−2.71%	−3.53%	−2.99%	−1.63%	−1.91%
(3) $\Delta MDR_{i,t+1}$	−7.31%	−5.43%	−4.47%	−5.84%	−3.66%	−2.73%
	Quartile #4 of 4: Most underleveraged					
	<i>FINDEF</i> (<i>t</i> , <i>t</i> + 1)			<i>FINDEF</i> (<i>t</i> , <i>t</i> + 1)		
	Low (−6.32%)	Medium (1.16%)	High (20.57%)	Low (−3.59%)	Medium (0.33%)	High (14.73%)
	(4) Distance from target @ <i>t</i> : (<i>MDR</i> *− <i>MDR</i>)	20.84%	21.31%	23.57%	18.89%	19.00%
(5) $\Delta BDR_{i,t+1}$	3.94%	4.87%	5.64%	1.09%	1.26%	3.22%
(6) $\Delta MDR_{i,t+1}$	6.51%	7.31%	10.15%	3.28%	3.31%	7.76%

approximately the same amount. Most importantly, while the “Low *FINDEF*” firms in Quartile 4 have the opportunity to use their net cash inflow (*FINDEF* = −6.32%) to retire some debt, they did not do so. Instead, these firms choose to increase leverage — a decision that is consistent with the tradeoff theory but not the pecking order view that underleveraged firms should stockpile debt capacity.

5.3. The “stock price mechanics” explanation

Welch (2004) finds no evidence of targeting, but concludes that managers passively tolerate almost any change in *MDR* caused by share price fluctuations. These conclusions are based on the regression

$$MDR_{i,t+1} = a_0 + a_1 MDR_{i,t} + a_2 IDR_{i,t+1} + \mu_{i,t+1}, \quad (9)$$

where $IDR_{i,t+1}$, the implied debt ratio equals $D_{i,t}/(D_{i,t} + S_{i,t}P_{i,t}(1 + \tilde{R}_{i,t+1}))$. Note that *IDR* is the mechanical effect of stock price changes on leverage, assuming that managers change neither D_t nor S_t during the following period. The variable $\tilde{R}_{i,t+1}$ is the realized appreciation in firm i 's share price during the period between t and $t+1$, and a_0 , a_1 , and a_2 are parameters to be estimated. (The intercept, a_0 , is sometimes omitted.) Welch's (2004) main conclusions result from Eq. (9)'s implicit and inappropriate constraints on the firm's adjustment process. To understand these constraints, we adjust our base specification in Eq. (4) to allow for the possibility that managers partly counteract the effects of share price changes on leverage. We begin by augmenting our basic partial adjustment model in Eq. (3) as follows:¹⁹

$$MDR_{i,t+1} - MDR_{i,t} = \lambda_1(MDR_{i,t}^* - MDR_{i,t}) + (1 - \lambda_2)(Share\ price\ effect) + \varepsilon_{i,t+1}, \quad (10)$$

where λ_2 is the adjustment speed to share price effects. Eq. (10) says that the observed change in the debt ratio equals the sum of the partial movement to the target leverage ($\lambda_1(MDR_{i,t}^* - MDR_{i,t})$) and the portion of the share price effect that is not offset within the year $((1 - \lambda_2)(Share\ price\ effect))$. We measure the “share price effect” as the change in *MDR* due exclusively to share price changes, that is,

$$SPE_{t+1} = \left(\frac{D_t}{D_t + S_t P_t (1 + \tilde{R}_{t,t+1})} \right) - MDR_t. \quad (11)$$

Substituting (11) into (10) gives²⁰

$$IDR_{i,t+1} = SPE_{i,t+1} + MDR_{i,t}. \quad (12)$$

The specification in Eq. (12) has two important features. First, finding $\lambda_2 = 0$ does not mean that managers never adjust *MDR* in response to share price changes. The residual effect of a price change during the period $(t, t+1)$ is impounded in the next period's lagged *MDR* and hence is offset at an annual rate λ_1 in the years following the initial price shock. Second, both *SPE* and the dependent variable contain the realized value of $\tilde{R}_{i,t+1}$, which

¹⁹Marcus (1983) also includes a separate share price effect in his model.

²⁰Replacing our base specification (5) with (7) has no important effect on any of the conclusions reported in this paper.

biases the coefficient on *SPE* toward unity. We therefore estimate (10) using an instrumental variable for *SPE*.²¹

The definition of *SPE* (11) implies that

$$MDR_{i,t+1} - MDR_{i,t} = \lambda_1(MDR_{i,t}^* - MDR_{i,t}) + (1 - \lambda_2)SPE_{i,t+1} + \varepsilon_{i,t+1}. \quad (11a)$$

Substituting (11a) into (12) gives Welch's specification in our notation:

$$MDR_{i,t+1} = \alpha_0 + a_1MDR_{i,t} + a_2SPE_{i,t+1} + \alpha_2MDR_{i,t} + \mu_{i,t+1}. \quad (13)$$

Juxtaposing (13) and our augmented specification (12) indicates that Welch's (2004) specification can be interpreted as:

- Using a multiple of $(a_0 + a_1 MDR_{i,t})$ as the firm's target leverage. This means that a share price shock becomes fully acceptable to the firm after one year, when the shock has passed into its lagged *MDR* value.
- Using *MDR*_{*i,t*} as a starting point for adjustments, as we do. This implies that one should interpret $a_2 = (1 - \lambda_1)$.
- Constraining *MDR* and *SPE* to have the same coefficient (a_2). That is, managers react equally quickly to anticipated and surprise leverage deviations from target.

Table 7 reports the results of estimating several versions of (12) and (13). Column (1) replicates Welch's (2004) specification for our data. Using the Fama and MacBeth (1973) methodology favored by Welch (2004), we find similar results: the lagged debt ratio yields an estimated coefficient close to zero, and the implied debt ratio's coefficient is near unity. These estimates seem to imply that firms permanently accept the impact of share price changes on their leverage. A stock price change is not offset at all before the end of period *t*, and at the start of period (*t* + 1) the leverage shock becomes fully incorporated into the firm's target. Column (2) replaces the assumption that firms target their previous period's debt ratio with a model of the firm's target debt ratio as $\beta X_{i,t}$ (*without firm fixed effects*). The estimated coefficient on *IDR* indicates that firms offset only 9.1% ($= 1 - \$0.909$) of the market's impact on firm leverage.

From (11a), we know that *IDR* has two components, *SPE* and *MDR*_{*i,t*}. Which part adjusts? We separate the two components of *IDR* in column (3). The estimated coefficient on *SPE* indicates that firms do not respond initially to share price surprises ($\lambda_2 = -0.029$).²² Typically for a simple OLS regression, however, the estimated adjustment speed for recognized deviations from target is 11.1% ($\lambda_1 = 1 - 0.889$)/year.

The first three columns in Table 7 exclude firm fixed effects, which we show above are important to the estimated adjustment speed. Column (4) therefore reports our base specification, augmented by a share price effect. Comparing these estimates to the last column of Table 2 indicates that the addition of *SPE* has little effect on our base

²¹We thank Yakov Amihud for pointing out this econometric issue. We regress *SPE* on the regression's predetermined variables and the realized return to the average firm in the same industry. In unreported results, we find that failing to instrument for *SPE* yields an estimated value of 0.06 for λ_2 , compared to the 0.029 we report in the last column of Table 7. Adding *SPE* to our base specification has only a small effect on the estimates of λ_1 and the β coefficients.

²²Intuitively, it may not be surprising that *SPE* carries a coefficient near unity. If some firms are above their targets and others are below, then any *SPE* moves some firms closer to their targets and other firms further away. This effect may average out to zero across all firms in the sample.

Table 7

Variations on the regression specification in Welch (2004)

Regression results for the models

$$\text{Column (1): } MDR_{i,t+1} = a_0 + a_1 MDR_{i,t} + a_2 IDR_{i,t+1} + \mu_{i,t+1} \quad (12)$$

$$\text{Column (2): } MDR_{i,t+1} = \alpha_0 + (1-\lambda_1) IDR_{i,t+1} + (\lambda_1 \beta) X_{i,t} + \mu_{i,t+1}$$

$$\text{Column (3): } MDR_{i,t+1} = \alpha_0 + (1-\lambda_1) MDR_{i,t} + (1-\lambda_2) SPE_{i,t+1} + (\lambda_1 \beta) X_{i,t} + \mu_{i,t+1}$$

$$\text{Column (4): } MDR_{i,t+1} = \alpha_0 + a_1 MDR_{i,t} + a_2 MDR_{i,t} + \alpha_2 SPE_{i,t+1} + \mu_{i,t+1}, \quad (13)$$

where MDR is the market debt ratio, IDR is the market debt ratio at time t augmented by the firm's realized return in $(t, t+1)$, $SPE = IDR - MDR_{i,t}$ measures the impact of share price changes on MDR during $(t, t+1)$, and the (lagged) “ X ” variables determine a firm's long-run target debt ratio, and include:

EBIT_TA: earnings before interest and taxes as a proportion of total assets;

MB: the market-to-book ratio of firm assets;

DEP_TA: depreciation expense as a proportion of total assets;

LnTA: natural log of total assets;

FA_TA: fixed assets as a proportion of total assets;

R&D_DUM: dummy variable indicating that the firm did not report R&D expenses;

R&D_TA: R&D expenses as a proportion of total assets; and

Ind_Median: median debt ratio of firm i 's Fama and French (2002) industry classification at time t .

T -statistics are shown in parentheses. Models (1)–(3) are undertaken using the Fama and MacBeth methodology. Reported R^2 numbers for models including fixed effects are “within” R^2 statistics.

	(1)	(2)	(3)	(4)
$MDR_{i,t}$	−0.075 (−6.48)		0.889 (201.57)	0.658 (226.06)
$IDR_{i,t+1}$	0.990 (71.93)	0.909 (175.79)		
$SPE_{i,t}$			1.029 (30.98)	0.971 (87.18)
$EBIT_TA$		0.026 (4.78)	0.022 (4.22)	−0.017 (−6.99)
MB		−0.003 (−5.10)	−0.004 (−6.63)	−0.005 (−18.52)
DEP_TA		−0.146 (−8.29)	−0.148 (−9.14)	−0.142 (−9.12)
$LnTA$		0.000 (−0.08)	0.001 (1.50)	0.002 (2.67)
FA_TA		0.019 (4.44)	0.022 (6.03)	0.045 (13.22)
$R\&D_DUM$		0.005 (6.80)	0.005 (6.90)	−0.001 (−0.54)
$R\&D_TA$		0.011 (0.71)	0.014 (0.96)	−0.023 (−3.04)
Ind_Median				0.078 (12.95)
$Rated$				0.007 (4.88)
Fixed effects?	No	No	No	Yes
N	111,106	111,106	111,106	111,106
R^2	0.860	0.862	0.860	0.691

specification's coefficient estimates. The coefficient on $MDR_{i,t}$ (0.658) implies a rapid adjustment speed to known deviations from target, while managers effectively ignore stock price changes in the year they occur (coefficient on SPE is 0.971). However, the net SPE passes into $MDR_{i,t}$ in subsequent periods and is offset at the rate of approximately 34% a year. We conclude, therefore, that our regressions provide little support for Welch's (2004) hypothesis that managers passively accept the impact of share price changes on their firm's leverage.

6. Robustness

Our conclusions about target debt ratios and the speed with which firms adjust toward their targets are robust to changes in the estimation horizon, the sample series, the time period, and the definition of leverage.

6.1. Stability over estimation horizons

The standard partial adjustment model only approximates firm behavior, and theory mandates no specific time interval between empirical observations. We therefore estimated Eq. (4) for time intervals between one and five years to assess whether the estimated adjustment speeds behave consistently with our assumption that a fixed proportion of the remaining leverage gap is closed each period.²³ Given Table 2's one-year adjustment speed of 0.344, a geometric decline would make the two-year coefficient 0.570 ($= 1 - (1 - 0.344)^2$). If the partial adjustment model fits the data well, then the estimated adjustment for a two-year period should be about 57%. The actual estimate in the first column of Table 8 indicates that 59.2% ($= 1 - 0.408$) of the initial gap would be closed by the end of year two. The longer horizons are also consistent with a continuous rate of adjustment. Following a smooth annual adjustment path of 34.4%, the respective three-, four-, and five-year coefficients would be 71.8%, 81.5%, and 87.9%, vs. the freely-estimated values of 73.4%, 82.9%, and 89.4%. Such a close correspondence between the theoretical and empirical estimates of adjustment provides further support for the hypothesis that our continuous partial adjustment specification appropriately captures variation in the data.

6.2. Stability across firm size

Thus far, we present regression results only for the entire sample of firms. Some previous writers omit relatively small firms, presumably because they might encounter prohibitively large transaction costs when making small leverage adjustments. On the other hand, some small firms grow quickly, which may reduce their costs of adjusting leverage (Alti, 2004). To assess the stability of regression (4), we reestimate our regressions for size-based subsamples. For each year, we group firms according to CRSP's NYSE size deciles for equity market value. Decile 1 contains the largest firms, and Decile 10 the smallest.²⁴ The results in Table 9 show that the partial adjustment model (4) fits all firm sizes quite well. Most of

²³The standard errors in Table 8 are not adjusted for the effect of overlapping observations, but our purpose here is to examine the coefficients' (unbiased) estimated values, not their statistical significance.

²⁴The uneven distribution of firms across the deciles reflects the fact that NYSE size deciles are being used to categorize firms drawn from the NYSE-AMEX-Nasdaq universe.

Table 8

Estimates over differing forecast horizons

We estimate variants of a regression based on

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + (1 - \lambda)MDR_{i,t} + \delta_{i,t+1} \quad (4)$$

using different intervals (“ k ” = 2, 3, 4, or 5 years). MDR is the market debt ratio. The (lagged) “ X ” variables determine a firm’s long-run target debt ratio, and include:

EBIT_TA: earnings before interest and taxes as a proportion of total assets;

MB: the market-to-book ratio of firm assets;

DEP_TA: depreciation expense as a proportion of total assets;

LnTA: natural log of total assets;

FA_TA: fixed assets as a proportion of total assets;

R&D_DUM: dummy variable indicating that the firm did not report R&D expenses;

R&D_TA: R&D expenses as a proportion of total assets;

Ind_Median: median debt ratio of firm i ’s Fama and French (2002) industry classification at time t ; and

Rated: Dummy variable equal to one (zero) if the firm has a public debt rating in Compustat.

All regressions include (unreported) year dummies. T -statistics are shown in parentheses. Reported R^2 numbers for models including fixed effects are “within” R^2 statistics.

	$k = 2$ years	$k = 3$ years	$k = 4$ years	$k = 5$ years
$MDR_{i,t}$	0.408 (93.17)	0.266 (55.62)	0.171 (32.94)	0.106 (19.29)
$EBIT_TA$	−0.176 (−46.78)	−0.224 (−51.97)	−0.245 (−50.95)	−0.252 (−47.64)
MB	−0.012 (−26.21)	−0.016 (−31.45)	−0.019 (−32.48)	−0.021 (−32.28)
DEP_TA	−0.543 (−20.49)	−0.419 (−13.97)	−0.362 (−10.99)	−0.380 (−10.54)
$LNTA$	0.051 (55.39)	0.059 (55.41)	0.060 (51.12)	0.059 (46.52)
FA_TA	0.146 (26.28)	0.180 (28.72)	0.187 (27.08)	0.190 (25.48)
$R\&D_DUM$	−0.001 (−0.53)	−0.004 (−1.87)	−0.005 (−2.07)	−0.005 (−2.12)
$R\&D_TA$	−0.146 (−11.49)	−0.219 (−14.74)	−0.258 (−15.20)	−0.274 (−14.36)
$Rated$	0.014 (6.59)	0.025 (10.49)	0.031 (12.81)	0.034 (13.31)
Ind_Median	−0.021 (−2.16)	−0.016 (−1.46)	−0.031 (−2.69)	−0.028 (−2.40)
Fixed effects?	Yes	Yes	Yes	Yes
N	97,590	85,958	75,886	67,052
R^2	0.278	0.208	0.176	0.166

the X variables’ coefficient estimates do not vary greatly across the four subsamples.²⁵ Each subsample exhibits a reasonable adjustment speed, which exceeds most past estimates in the literature. Note that the largest firms adjust least rapidly (27.3%). Perhaps larger

²⁵Note, however, that the uniformly positive coefficients on **MB** in Table 9 contrast with the more common negative coefficients estimated for the full sample. We have no good explanation for this result.

Table 9

Stability across firm sizes

Each year, we divided the universe of firms into ten groups based on CRSP's equity value size deciles for NYSE-traded firms in that year. Decile 1 contains the largest firms. For each size grouping, we then estimate the regression

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + (1 - \lambda)MDR_{i,t} + \delta_{i,t+1}, \quad (4)$$

where MDR is the market debt ratio. The (lagged) “ X ” variables determine a firm's long-run target debt ratio, and include:

EBIT_TA: earnings before interest and taxes as a proportion of total assets;

MB: the market-to-book ratio of firm assets;

DEP_TA: depreciation expense as a proportion of total assets;

LnTA: natural log of total assets;

FA_TA: fixed assets as a proportion of total assets;

R&D_DUM: dummy variable indicating that the firm did not report R&D expenses;

R&D_TA: R&D expenses as a proportion of total assets;

Ind_Median: median debt ratio of firm i 's Fama and French (2002) industry classification at time t ; and

Rated: Dummy variable equal to one if the firm has a public debt rating in Compustat, zero otherwise.

All regressions include (unreported) year dummies. T -statistics are shown in parentheses. Reported R^2 numbers for models including fixed effects are “within” R^2 statistics.

	Firms in Deciles 1 & 2	Firms in Deciles 3 & 4	Firms in Deciles 5, 6, & 7	Firms in Deciles 8, 9, & 10
$MDR_{i,t}$	0.727 (56.33)	0.574 (39.51)	0.486 (47.97)	0.608 (123.47)
EBIT_TA	0.009 (0.53)	0.059 (3.53)	0.013 (1.20)	-0.028 (-7.54)
MB	0.001 (0.63)	0.010 (9.25)	0.009 (10.15)	0.002 (4.72)
DEP_TA	-0.354 (-5.71)	-0.224 (-3.05)	-0.421 (-8.04)	-0.138 (-5.50)
LnTA	0.017 (8.61)	0.063 (22.93)	0.073 (38.04)	0.042 (38.91)
FA_TA	0.019 (1.76)	0.059 (4.56)	0.082 (7.92)	0.063 (11.06)
R&D_DUM	0.002 (0.52)	-0.008 (-2.13)	-0.004 (-1.34)	0.003 (1.63)
R&D_TA	-0.049 (-1.12)	0.089 (1.85)	0.094 (3.03)	-0.010 (-0.83)
Rated	0.001 (0.22)	-0.004 (-0.97)	0.004 (1.45)	0.021 (5.41)
Ind_Median	-0.072 (-4.73)	-0.036 (-1.99)	-0.031 (-2.06)	0.037 (3.32)
Fixed effects?	Yes	Yes	Yes	Yes
N	9,313	9,178	18,050	74,565
R^2	0.590	0.499	0.476	0.434

firms rely more on public debt, which is more expensive to adjust than the private (“bank”) debt used by smaller firms. Because public debt has few covenants, the external pressure for a large firm to reverse a leverage increase may be less intense than for a small firm whose bank lenders can enforce relatively tight covenants. Alternatively, if larger firms have less volatile cash flows, they may bear lower costs when they are away from their target leverage.

6.3. Stability over time

Table 10 reports estimation results for three equal time periods, 1966–1977, 1978–1989, and 1990–2001. We find that estimated adjustment speeds are quite similar across time periods, although the estimates in Table 10 all exceed the 0.344 value from Table 2. The other coefficient signs and significance are also generally consistent across periods, with two exceptions: the impact of depreciation expense (*DEP_TA*) on *MDR*^{*} declines from –0.500 in the first time period to –0.096 in the last, and the sign on *Ind_Median* reverses between the first and second periods.

6.4. Alternative “leverage” definitions

Previous research defines leverage in a variety of ways. Table 11 shows that our conclusions about targets and adjustment speeds do not depend on our definition of leverage. We reestimate Eq. (4) using three alternative (new) definitions of the market debt ratio:

$$MDR_1 = \left[\frac{\text{Long Term Debt} + \text{Short Term Debt}}{\text{Total assets} - \text{Book Equity} + \text{Market Equity}} \right],$$

$$MDR_2 = \left[\frac{\text{Total Liabilities}}{\text{Total Liabilities} + \text{Market Equity}} \right],$$

$$MDR_3 = \left[\frac{\text{Long Term Debt}}{\text{Total assets} - \text{Current Liabilities} - \text{Book Equity} + \text{Market Equity}} \right].$$

The estimated adjustment speeds are all rapid (36.6–40.5% annually) and the determinants of target leverage generally carry significant coefficients with appropriate signs.

7. Summary and conclusions

We find strong evidence that nonfinancial firms identified and pursued target capital ratios during the 1966–2001 period. The evidence is equally strong across size classes and time periods, and indicates that a partial adjustment model with firm fixed effects fits the data very well. As earlier research finds, target debt ratios depend on well-accepted firm characteristics. Firms that are under- or overleveraged by this measure soon adjust their debt ratios to offset the observed gap. Targeting behavior is evident in both market-valued and book-valued leverage measures.

Unlike some recent studies, we estimate that firms return relatively quickly to their target leverage ratios when they are shocked away from their targets. The mean sample firm acts to close its (market) leverage gap at the rate of more than 30% per year. While one might dispute whether a 30% annual adjustment speed is “slow” or “rapid,” it is

Table 10
Stability over Time

For three equal-sized time periods we estimate the regression

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + (1 - \lambda)MDR_{i,t} + \delta_{i,t+1}, \quad (4)$$

where *MDR* is the market debt ratio. The (lagged) “*X*” variables determine a firm’s long-run target debt ratio, and include:

- EBIT_TA*: earnings before interest and taxes as a proportion of total assets;
- MB*: the market-to-book ratio of firm assets;
- DEP_TA*: depreciation expense as a proportion of total assets;
- LnTA*: natural log of total assets;
- FA_TA*: fixed assets as a proportion of total assets;
- R&D_DUM*: dummy variable indicating that the firm did not report R&D expenses;
- R&D_TA*: R&D expenses as a proportion of total assets;
- Ind_Median*: median debt ratio of firm *i*’s Fama and French (2002) industry classification at time *t*; and
- Rated*: dummy variable equal to one if the firm has a public debt rating in Compustat, and zero otherwise.

All models include (unreported) year dummies. *T*-statistics are shown in parentheses. Reported *R*² numbers for models including fixed effects are “within” *R*² statistics. Note that the *Rated* coefficients are not estimated for the first sample period (1966–1977), because Compustat does not report that variable before 1981.

	1966–1977	1978–1989	1990–2001
<i>MDR_{i,t}</i>	0.566 (53.80)	0.509 (68.72)	0.516 (75.59)
<i>EBIT_TA</i>	−0.025 (−1.77)	−0.052 (−8.77)	−0.025 (−6.01)
<i>MB</i>	−0.004 (−3.13)	−0.001 (−1.71)	0.000 (0.21)
<i>DEP_TA</i>	−0.500 (−6.79)	−0.114 (−2.99)	−0.096 (−3.14)
<i>LnTA</i>	0.062 (22.12)	0.048 (27.38)	0.040 (29.82)
<i>FA_TA</i>	0.066 (5.09)	0.058 (6.73)	0.060 (7.87)
<i>R&D_DUM</i>	−0.003 (−1.07)	−0.003 (−0.99)	0.003 (0.80)
<i>R&D_TA</i>	−0.129 (−2.32)	−0.014 (−0.57)	−0.002 (−0.18)
<i>Rated</i>		−0.003 (−0.90)	0.009 (3.01)
<i>IND_Median</i>	−0.058 (−3.27)	0.088 (5.88)	0.079 (5.51)
Fixed effects?	Yes	Yes	Yes
<i>N</i>	21,032	38,779	51,295
<i>R</i> ²	0.551	0.338	0.336

surely not zero. Indicators of the pecking order and market timing (à la Baker and Wurgler, 2002) theories carry statistically significant coefficients, but their economic effects are swamped by movements toward a firm-specific leverage target. Share price fluctuations have a short-term impact on market debt ratios, but efforts to reach the target leverage ratio offset these transitory effects within a few years.

Table 11

Alternative definitions of leverage

Estimates of the basic regression specification

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + (1 - \lambda)MDR_{i,t} + \delta_{i,t+1}, \quad (4)$$

for three alternative definitions of MDR:

$$MDR_1 = \left[\frac{\text{Long Term Debt} + \text{Short Term Debt}}{\text{Total assets} - \text{Book Equity} + \text{Market Equity}} \right],$$

$$MDR_2 = \left[\frac{\text{Total Liabilities}}{\text{Total Liabilities} + \text{Market Equity}} \right],$$

$$MDR_3 = \left[\frac{\text{Long term Debt}}{\text{Total assets} - \text{Current Liabilities} - \text{Book Equity} + \text{Market Equity}} \right],$$

where *MDR* is the market debt ratio. The (lagged) “*X*” variables determine a firm’s long-run target debt ratio, and include

EBIT_TA: earnings before interest and taxes as a proportion of total assets;

MB: the market-to-book ratio of firm assets;

DEP_TA: depreciation expense as a proportion of total assets;

LnTA: natural log of total assets;

FA_TA: fixed assets as a proportion of total assets;

R&D_DUM: dummy variable indicating that the firm did not report R&D expenses;

R&D_TA: R&D expenses as a proportion of total assets;

Ind_Median: median debt ratio of firm *i*’s Fama and French (2002) industry classification at time *t*;

Rated: dummy variable equal to one (zero) if the firm has a public debt rating in Compustat.

All regressions include (unreported) year dummies. *T*-statistics are shown in parentheses. Reported *R*² numbers for models including fixed effects are “within” *R*² statistics.

Dependent variable:	<i>MDR</i> ₁	<i>MDR</i> ₂	<i>MDR</i> ₃
<i>MDR</i> _{<i>i,t</i>}	0.634 (184.09)	0.614 (132.79)	0.595 (153.82)
<i>EBIT_TA</i>	−0.023 (−9.73)	−0.055 (−17.75)	−0.039 (−13.89)
<i>MB</i>	0.000 (0.67)	−0.001 (−2.47)	−0.001 (−2.44)
<i>DEP_TA</i>	−0.192 (−12.33)	−0.183 (−9.33)	−0.191 (−9.99)
<i>LnTA</i>	0.020 (35.58)	0.028 (39.45)	0.024 (33.66)
<i>FA_TA</i>	0.047 (13.74)	0.047 (11.03)	0.055 (13.10)
<i>R&D_DUM</i>	0.000 (0.23)	−0.001 (−0.52)	−0.001 (−0.94)
<i>R&D_TA</i>	−0.017 (−2.26)	−0.070 (−7.47)	−0.035 (−3.93)
<i>Rated</i>	0.002 (1.77)	0.000 (−0.02)	0.005 (2.82)
<i>IND_Median</i>	0.051 (6.01)	0.033 (4.67)	0.044 (4.92)
Fixed effects?	Yes	Yes	Yes
<i>N</i>	110,253	111,088	108,476

Appendix A. Estimating our dynamic panel model

Consider a (simplified) dynamic panel data model of the form:

$$\mathbf{MDR}_{i,t+1} = \alpha \mathbf{MDR}_{i,t} + (\mu + \varepsilon_{i,t+1}), \quad (\text{A.1})$$

where i indexes firms and t indicates the time period. The error term in (A.1) has two components, μ_i , an unobserved, time-invariant, firm-specific effect, and $\varepsilon_{i,t+1}$, the usual residual. Because the residual component of $\mathbf{MDR}_{i,t}$ is correlated with the unobserved effect in the error term, an OLS-estimated coefficient on $\mathbf{MDR}_{i,t}$ will be biased upwards (Anderson and Hsiao, 1981; Baltagi, 2001; Bond, 2002).

A common way to estimate panel data models with unobserved effects is to perform a “within” transform of (A.1) and then estimate using OLS. A within transformation expresses all variables as deviations from their firm-specific time-series means. This eliminates μ_i from the regression as it is time invariant and thus provides consistent estimates. However, in the presence of a lagged dependent variable, the within transform introduces correlation of the transformed lagged dependent variable with the transformed error term by construction (Wooldridge, 2002; Baltagi, 2001; Hsiao, 2003; etc.). To see this, note that the within transforms of the lagged dependent variable and error terms are

$$\left(\mathbf{MDR}_{i,t} - \sum_{t=1}^{T_i} \mathbf{MDR}_{i,t} / T_i \right) \text{ and } \left(\varepsilon_{i,t+1} - \sum_{t=1}^{T_i} \varepsilon_{i,t+1} / T_i \right),$$

respectively, where T_i is the number of available observations for firm i . Since $\mathbf{MDR}_{i,2}$ is correlated with $\varepsilon_{i,2}$, $\mathbf{MDR}_{i,3}$ is correlated with $\varepsilon_{i,3}$, and so on, the transformed variables are correlated with the transformed error term. As a result, the coefficient on the lagged dependent variable, α , is biased downwards by a factor of (approximately) $1/T$ (Wooldridge, 2002). In panel data sets with large T , the bias becomes insignificant, but in panel data sets such as ours, with large N and small T , the bias can be substantial and needs to be addressed to obtain consistent estimates.

The first two columns of Table A.1 report regressions that contain biased estimates of the lagged dependent variable’s coefficient. The OLS estimate (0.86) in column (1) is biased upwards, while the panel estimate (0.62) in column (2) is biased downwards. As Bond (2002) points out, the true lagged dependent variable coefficient must therefore lie in the interval (0.62, 0.86).

Greene (2003) observes that we can obtain unbiased estimates of the levels regression (A.1) via two-stage least squares if an instrument can be found that is correlated with the lagged dependent variable but not the error term. This approach forms the basis for nearly all our results in the paper. The third column of Table A.1 reports a levels panel regression in which the lagged book debt ratio (**BDR**) instruments for the lagged dependent variable. The estimated coefficient on the lagged dependent variable (0.656) lies between the OLS and “within” estimates, as predicted by Bond (2002).

The estimates in column (3) rely on the book debt ratio being a reasonable instrument for the market debt ratio. As always, however, finding reliable instruments can be difficult and a number of techniques have been developed in the literature to estimate unbiased coefficients for a model such as (A.1). (Baltagi, 2001 and Arellano and Honore, 2001 provide surveys.) These alternative techniques generally first-difference the model (A.1) to eliminate fixed effects and use lagged dependent variables to instrument for the lagged

Table A.1

Alternative methods for estimating dynamic panel regressions

In columns (1)–(3) we estimate our main regression specification that explains a firm's market debt ratio:

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + (1 - \lambda)MDR_{i,t} + \delta_{i,t+1}, \quad (4)$$

In columns (4)–(6) we estimate our main regression specification in first differences:

$$\Delta MDR_{i,t+1} = (\lambda\beta)\Delta X_{i,t} - \lambda\Delta MDR_{i,t} + (\delta_{i,t+1} - \delta_{i,t})$$

The (lagged) “*X*” variables determine a firm's long-run target debt ratio, and include**EBIT_TA**: earnings before interest and taxes as a proportion of total assets;**MB**: the market-to-book ratio of firm assets;**DEP_TA**: depreciation expense as a proportion of total assets;**LnTA**: natural log of total assets;**FA_TA**: fixed assets as a proportion of total assets;**R&D_DUM**: dummy variable indicating that the firm did not report R&D expenses;**R&D_TA**: R&D expenses as a proportion of total assets;**Ind_Median**: median debt ratio of firm *i*'s Fama and French (2002) industry classification at time *t*; and**Rated**: dummy variable equal to one if the firm has a public debt rating in Compustat, and zero otherwise.All models include (unreported) year dummies. *T*-statistics are shown in parentheses. Reported *R*² numbers for models including fixed effects are “within” *R*² statistics. For the first-differenced models, *R*² cannot be computed.

	(1)	(2)	(3)	(4)	(5)	(6)
	Simple OLS	Simple FE	Base Specification, column (7) of Table 2	First-Difference Estimates (Anderson and Hsiao, 1981)		Arellano and Bond (1991)
				<i>MDR</i> _{<i>t,t-1</i>} Instruments for $\Delta MDR_{i,t}$	<i>BDR</i> _{<i>t,t-1</i>} Instruments for $\Delta MDR_{i,t}$	
<i>MDR</i> _{<i>t</i>}	0.860 (457.28)	0.620 (224.50)	0.656 (171.58)	0.467 (31.78)	0.522 (20.69)	0.476 (74.01)
<i>EBIT_TA</i>	-0.025 (-11.22)	-0.039 (-12.93)	-0.030 (-9.66)	-0.217 (-72.93)	-0.221 (-65.67)	-0.231 (-67.75)
<i>MB</i>	-0.001 (-3.86)	-0.001 (-3.52)	0.000 (-0.81)	-0.020 (-50.40)	-0.020 (-49.23)	-0.021 (-43.76)
<i>DEP_TA</i>	-0.206 (-15.62)	-0.224 (-10.96)	-0.226 (-11.06)	-0.269 (-10.49)	-0.309 (-10.27)	-0.294 (-10.39)
<i>LnTA</i>	-0.001 (-3.73)	0.027 (36.69)	0.025 (34.00)	0.125 (72.42)	0.128 (63.59)	0.131 (85.96)
<i>FA_TA</i>	0.024 (11.41)	0.060 (13.52)	0.053 (11.93)	0.237 (37.81)	0.238 (37.01)	0.246 (34.38)
<i>R&D_DUM</i>	0.006 (7.21)	0.000 (-0.10)	0.000 (0.02)	-0.001 (-0.37)	-0.001 (-0.23)	-0.002 (-0.77)
<i>R&D_TA</i>	-0.104 (-17.27)	-0.036 (-3.64)	-0.025 (-2.57)	-0.073 (-6.85)	-0.070 (-6.34)	-0.082 (-6.27)
<i>Ind_Median</i>	0.047 (12.20)	0.049 (6.32)	0.034 (4.30)	-0.137 (-10.79)	-0.163 (-10.12)	-0.102 (-8.57)
<i>Rated</i>	0.005 (3.47)	0.005 (2.56)	0.003 (1.71)	0.025 (8.71)	0.024 (8.13)	0.021 (6.82)
Fixed effects?	No	Yes	Yes	No ^a	No ^a	No ^a
<i>N</i>	111,106	111,106	111,106	94,591	94,591	82,395
<i>R</i> ²	0.753	0.467	0.466	—	—	—
<i>H</i> ₀ : No serial correlation				***		***

^aFixed effects are implicitly incorporated through the first-differencing procedure.

***Significantly different from one at the 1% level.

first-difference. For example, [Anderson and Hsiao \(1981\)](#) convert (A.1) into

$$\Delta MDR_{i,t+1} = \alpha \Delta MDR_{i,t} + \delta_{i,t+1} \quad (\text{A.2})$$

and then use the dependent variable's second lag ($MDR_{i,t-1}$) to instrument for the (first-differenced) lagged dependent variable ($\Delta MDR_{i,t}$). [Arellano and Bond \(1991\)](#) show that the instrument space can include further lags of the dependent variable under some circumstances.²⁶

These first-difference methodologies rely on two key assumptions to produce unbiased and consistent estimates. First, the error term $\varepsilon_{i,t+1}$ in (A.1) should be serially uncorrelated, because first-order serial correlation would make the lagged dependent variable correlated with the (differenced) regression residual. In other words, lags of the dependent variable fail the exogeneity test if the residual is serially correlated. Second, the dependent variable should not have (near) unit root properties. If the dependent variable series has high persistence then the first-difference will be close to zero and hence the instruments used in the procedure will be weak.

Columns (4)–(6) in [Table A.1](#) present results based on several first-differencing methods of estimating a dynamic panel model. These can be viewed as potential substitutes for the estimation method we use in the text for Eq. (4). Column (4) presents results based on [Anderson and Hsiao \(1981\)](#). We first-difference to remove the unobserved firm effects and then use the second lag of the dependent variable as an instrument. This procedure produces a coefficient on $MDR_{i,t}$ of 0.467, which is inconsistent with [Bond's \(2002\)](#) assertion that the true coefficient lies between 0.62 and 0.86. We also reject the hypothesis of zero serial correlation of ε in (A.1), casting serious doubt on the validity of the instrument used. In column (5) we modify the Anderson-Hsiao method by using the second lag of BDR as an instrument in place of the second lag of MDR . This instrument should be immune to the serial correlation problem identified in column (4). We find that the estimate (0.522) is higher than column (4)'s, but still outside the interval defined by the OLS and “within” estimates. The problem appears to be that of weak instruments given persistence in the dependent variable. Indeed, the correlation in the lagged levels of our dependent variable is 0.98.

Given the serial correlation in the error terms and the high persistence in our dependent variable, [Arellano and Bond's \(1991\)](#) GMM procedure is unlikely to yield consistent results. To verify, we run the GMM estimation (using Stata's XTABOND procedure) and present the results in column (6). The coefficient on $MDR_{i,t}$ (0.476) implies very rapid adjustment, but both the tests of serial correlation and an (unreported) Sargan test of overidentifying restrictions are rejected at the 1% level. [Arellano and Bond's \(1991\)](#) technique thus cannot be applied to our data.

References

- Ahn, S., Schmidt, P., 1995. Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68, 5–28.
- Alti, Aydoğan, 2004. How persistent is the impact of market timing on capital structure? University of Texas Working paper.

²⁶Subsequent authors refine [Arellano and Bond's \(1991\)](#) General Methods of Moments procedures (e.g., [Arellano and Bover, 1995](#); [Ahn and Schmidt, 1995](#)).

- Anderson, T., Hsiao, C., 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76, 598–606.
- Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte-Carlo evidence and an application to employment equations. *Review of Economic Studies* 38, 277–297.
- Arellano, M., Bover, O., 1995. Another look at instrumental-variable estimation of error-components models. *Journal of Econometrics* 68, 29–52.
- Arellano, M., Honore, B., 2001. Panel data models: some recent developments. In: Heckman, J.J., Learner, E.E., (Eds.), *Handbook of Econometrics*, vol 5, North-Holland, Amsterdam.
- Baker, M., Wurgler, J., 2002. Market timing and capital structure. *The Journal of Finance* 57, 1–32.
- Baltagi, B., 2001. *Econometric Analysis of Panel Data*, 2nd edn. John Wiley, New York.
- Bond, S., 2002. Dynamic panel data models: a guide to micro data methods and practice. Working paper.
- Donaldson, G., 1961. Corporate debt capacity: a study of corporate debt policy and the determination of corporate debt capacity. Harvard Business School, Division of Research, Harvard University.
- Fama, E., French, K., 1997. Industry costs of equity. *Journal of Financial Economics* 43, 153–193.
- Fama, E., French, K., 2002. Testing trade-off and pecking order predictions about dividends and debt. *Review of Financial Studies* 15, 1–34.
- Fama, E., MacBeth, J., 1973. Risk, return, and equilibrium: empirical tests. *Journal of Political Economy* 81, 607–636.
- Faulkender, M., Petersen, M., 2005. Does the source of capital affect capital structure? *Review of Financial Studies*, forthcoming.
- Fischer, E., Heinkel, R., Zechner, J., 1989. Dynamic capital structure choice: theory and tests. *Journal of Finance* 44, 19–40.
- Frank, M., Goyal, V., 2003. Testing the pecking order theory of capital structure. *Journal of Financial Economics* 67, 217–248.
- Graham, J., Harvey, C., 2001. The theory and practice of corporate finance: evidence from the field. *Journal of Financial Economics* 60, 187–243.
- Greene, W., 2003. *Econometric Analysis*, 5th edn. Upper Saddle River, Prentice Hall.
- Halov, N., Heider, F., 2004. Capital structure, risk and asymmetric information. NYU Working paper.
- Hovakimian, A., 2003. Are observed capital structures determined by equity market timing? Baruch College Working paper.
- Hovakimian, A., Opler, T., Titman, S., 2001. The debt-equity choice: an analysis of issuing firms. *Journal of Financial and Quantitative Analysis* 36, 1–24.
- Hovakimian, A., Hovakimian, G., Tehranian, H., 2004. Determinants of target capital structure: the case of dual debt and equity issues. *Journal of Financial Economics* 71, 517–540.
- Hsiao, C., 2003. *Analysis of Panel Data*, 2nd edn. Cambridge University Press, Cambridge.
- Huang, R., Ritter, J., 2005. Testing the market timing theory of capital structure. University of Florida Working paper.
- Jalilvand, A., Harris, R., 1984. Corporate behaviour in adjusting to capital structure and dividend targets: an econometric study. *Journal of Finance* 39, 127–145.
- Ju, N., Parino, R., Potoshman, A., Weisbach, M., 2002. Horses and rabbits optimal dynamic capital structure from shareholder and manager perspectives. Working paper.
- Kayhan, A., Titman, S., 2005. Firms' histories and their capital structure. University of Texas Working paper.
- Korajczyk, R., Levy, A., 2003. Capital structure choice: macroeconomic conditions and financial constraints. *Journal of Financial Economics* 68, 75–109.
- Leary, M., Roberts, M., 2005. Do firms rebalance their capital structure? *Journal of Finance* forthcoming.
- Lemmon, M., Zender, J., 2004. Debt capacity and tests of capital structure theories. University of Utah and University of Colorado Working paper.
- Marcus, A., 1983. The bank capital decision: a time series-cross section analysis. *Journal of Finance* 38, 1217–1232.
- Marsh, P., 1982. The choice between equity and debt: an empirical study. *Journal of Finance* 37, 121–144.
- Mauer, D., Triantis, A., 1994. Interactions of corporate financing and investment decisions: a dynamic framework. *Journal of Finance* 49, 1253–1277.
- Modigliani, F., Miller, M., 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48, 655–669.
- Myers, S., 1984. The capital structure puzzle. *The Journal of Finance* 39, 575–592.
- Rajan, R., Zingales, L., 1995. What do we know about capital structure: some evidence from international data. *Journal of Finance* 50, 1421–1460.

- Roberts, M., 2002. The dynamics of capital structure: an empirical analysis of a partially observable system. Duke Working paper.
- Shyam-Sunder, L., Myers, S., 1999. Testing static tradeoff against pecking order models of capital structure. *Journal of Financial Economics* 51, 219–243.
- Titman, S., Tsyplakov, S., 2004. A dynamic model of optimal capital structure. University of Texas Working paper.
- Welch, I., 2004. Capital structure and stock returns. *Journal of Political Economy* 112, 106–131.
- Wooldridge, J., 2002. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge.



ELSEVIER

Journal of Econometrics 93 (1999) 345–368

JOURNAL OF
Econometrics

www.elsevier.nl/locate/econbase

Threshold effects in non-dynamic panels: Estimation, testing, and inference

Bruce E. Hansen¹

*Department of Economics, University of Wisconsin, Social Science Building, 1180 Observatory Drive,
Madison, WI 53706, USA*

Received 1 June 1997; received in revised form 1 March 1999; accepted 1 April 1999

Abstract

Threshold regression methods are developed for non-dynamic panels with individual-specific fixed effects. Least squares estimation of the threshold and regression slopes is proposed using fixed-effects transformations. A non-standard asymptotic theory of inference is developed which allows construction of confidence intervals and testing of hypotheses. The methods are applied to a 15-year sample of 565 US firms to test whether financial constraints affect investment decisions. © 1999 Elsevier Science S.A. All rights reserved.

JEL classification: C33; C12; C13

Keywords: Threshold regression; Panel data; Liquidity constraints; Investment; Non-standard distribution

1. Introduction

Are regression functions identical across all observations in a sample, or do they fall into discrete classes? This question may be addressed using threshold

¹ Homepage: <http://www.ssc.wisc.edu/~bhansen/>.

E-mail address: bhansen@ssc.wisc.edu (B.E. Hansen)

regression techniques. Threshold regression models specify that individual observations can be divided into classes based on the value of an observed variable. Despite their intuitive appeal, econometric techniques have not been well developed for threshold regression.

This paper introduces econometric techniques appropriate for threshold regression with panel data. Least squares estimation methods are described. An asymptotic distribution theory is derived which is used to construct confidence intervals for the parameters. A bootstrap method to assess the statistical significance of the threshold effect is also described. The methods are similar to those developed in earlier work by the author (Hansen, 1996, 1999).

The methods are used to investigate whether financial constraints affect the investment practices of firms. The classical theory of the firm suggests that financing should have no allocative effects (e.g., the Modigliani-Miller theorem). Investment decisions should only be based on the marginal Q of a specific project, since banks will be willing to extend finance. In the context of imperfect information, external financing may be limited, and debt-constrained firms may need to finance investment out of cash flow. If this is the case, investment will be correlated with cash flow for constrained firms. This observation led Fazzari et al. (1988) to divide a sample of US firms into classes based on their degree of financial constraints and estimate the differing effects of cash flow on investment among these classes. Their analysis suffered from two problems. First, they used an endogenous variable (dividend to income ratio) rather than an exogenous variable to form their sample splits. Second, they used an ad hoc method to select their sample splits. We repeat their analysis on an analogous data set using appropriate econometric techniques and find qualitatively similar results.

Other authors have investigated the implications of non-linear q models of investment. Abel and Eberly (1994) propose a model which implies that the response of investment to q may be non-linear in q . Abel and Eberly (1996) use panel data to estimate a similar model, and find evidence for non-linearities in the investment function. Barnett and Sakellaris (1998) find similar results using a threshold regression approach. Hu and Schiantarelli (1998) use a switching regression framework to study the same problem. Our paper extends and reinforces this growing literature.

The next section introduces the model and notation. Section 3 discusses estimation by fixed effects. Section 4 outlines our asymptotic theory of inference. A distribution theory is developed for the threshold estimate and the slope coefficients. Section 5 reports the empirical application to firms' investment decisions. Section 6 concludes. Proofs of the asymptotic theory are provided in the appendix. GAUSS programs and data which replicate the empirical work are available from the author's homepage.

2. Model

The observed data are from a balanced² panel $\{y_{it}, q_{it}, x_{it}; 1 \leq i \leq n, 1 \leq t \leq T\}$. The subscript i indexes the individual and the subscript t indexes time. The dependent variable y_{it} is scalar, the threshold variable q_{it} is scalar, and the regressor x_{it} is a k vector. The structural equation of interest is

$$y_{it} = \mu_i + \beta'_1 x_{it} I(q_{it} \leq \gamma) + \beta'_2 x_{it} I(q_{it} > \gamma) + e_{it}. \quad (1)$$

where $I(\cdot)$ is the indicator function. An alternative intuitive way of writing (1) is

$$y_{it} = \begin{cases} \mu_i + \beta'_1 x_{it} + e_{it}, & q_{it} \leq \gamma, \\ \mu_i + \beta'_2 x_{it} + e_{it}, & q_{it} > \gamma. \end{cases}$$

Another compact representation of (1) is to set

$$x_{it}(\gamma) = \begin{pmatrix} x_{it} I(q_{it} \leq \gamma) \\ x_{it} I(q_{it} > \gamma) \end{pmatrix}$$

and $\beta = (\beta'_1 \ \beta'_2)'$ so that (1) equals

$$y_{it} = \mu_i + \beta' x_{it}(\gamma) + e_{it}. \quad (2)$$

The observations are divided into two ‘regimes’ depending on whether the threshold variable q_{it} is smaller or larger than the threshold γ . The regimes are distinguished by differing regression slopes, β_1 and β_2 . For the identification of β_1 and β_2 , it is required that the elements of x_{it} are not time invariant. We also assume that the threshold variable q_{it} is not time invariant. The error e_{it} is assumed to be independent and identically distributed (iid) with mean zero and finite variance σ^2 . The iid assumption excludes lagged dependent variables from x_{it} . It is unclear how to extend the results to allow for dynamic models and/or heteroskedastic errors. The analysis is asymptotic with fixed T as $n \rightarrow \infty$.

3. Estimation

3.1. Least squares estimation

One traditional method to eliminate the individual effect μ_i is to remove individual-specific means. While straightforward in linear models, the non-linear specification (1) calls for a more careful treatment. Note that taking

² It is unknown if the results extend to unbalanced panels.

averages of (1) over the time index t produces

$$\bar{y}_i = \mu_i + \beta' \bar{x}_i(\gamma) + \bar{e}_i \quad (3)$$

where $\bar{y}_i = T^{-1} \sum_{t=1}^T y_{it}$, $\bar{e}_i = T^{-1} \sum_{t=1}^T e_{it}$, and

$$\begin{aligned} \bar{x}_i(\gamma) &= \frac{1}{T} \sum_{t=1}^T x_{it}(\gamma) \\ &= \begin{pmatrix} \frac{1}{T} \sum_{t=1}^T x_{it} I(q_{it} \leq \gamma) \\ \frac{1}{T} \sum_{t=1}^T x_{it} I(q_{it} > \gamma) \end{pmatrix}. \end{aligned}$$

Taking the difference between (2) and (3) yields

$$y_{it}^* = \beta' x_{it}^*(\gamma) + e_{it}^* \quad (4)$$

where

$$y_{it}^* = y_{it} - \bar{y}_i,$$

$$x_{it}^*(\gamma) = x_{it}(\gamma) - \bar{x}_i(\gamma),$$

and

$$e_{it}^* = e_{it} - \bar{e}_i.$$

Let

$$y_i^* = \begin{bmatrix} y_{i2}^* \\ \vdots \\ y_{iT}^* \end{bmatrix}, \quad x_i^*(\gamma) = \begin{bmatrix} x_{i2}^*(\gamma) \\ \vdots \\ x_{iT}^*(\gamma)' \end{bmatrix}, \quad e_i^* = \begin{bmatrix} e_{i2}^* \\ \vdots \\ e_{iT}^* \end{bmatrix}$$

denote the stacked data and errors for an individual, with one time period deleted. Then let Y^* , $X^*(\gamma)$ and e^* denote the data stacked over all individuals, for example

$$X^*(\gamma) = \begin{bmatrix} x_1^*(\gamma) \\ \vdots \\ x_i^*(\gamma) \\ \vdots \\ x_n^*(\gamma) \end{bmatrix}.$$

Using this notation, (4) is equivalent to

$$Y^* = X^*(\gamma)\beta + e^*. \quad (5)$$

For any given γ , the slope coefficient β can be estimated by ordinary least squares (OLS). That is,

$$\hat{\beta}(\gamma) = (X^*(\gamma)'X^*(\gamma))^{-1}X^*(\gamma)'Y^*. \quad (6)$$

The vector of regression residuals is

$$\hat{e}^*(\gamma) = Y^* - X^*(\gamma)\hat{\beta}(\gamma)$$

and the sum of squared errors is

$$\begin{aligned} S_1(\gamma) &= \hat{e}^*(\gamma)'\hat{e}^*(\gamma) \\ &= Y^{*'}(I - X^*(\gamma)(X^*(\gamma)'X^*(\gamma))^{-1}X^*(\gamma))Y^*. \end{aligned} \quad (7)$$

Chan (1993) and Hansen (1999) recommend estimation of γ by least squares. This is easiest to achieve by minimization of the concentrated sum of squared errors (7). Hence the least squares estimators of γ is

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} S_1(\gamma). \quad (8)$$

It is undesirable for a threshold $\hat{\gamma}$ to be selected which sorts too few observations into one or the other regime. This possibility can be excluded by restricting the search in (8) to values of γ such that a minimal percentage of the observations (say, 1% or 5%) lie in each regime.

Once $\hat{\gamma}$ is obtained, the slope coefficient estimate is $\hat{\beta} = \hat{\beta}(\hat{\gamma})$. The residual vector is $\hat{e}^* = \hat{e}^*(\hat{\gamma})$ and residual variance

$$\hat{\sigma}^2 = \frac{1}{n(T-1)}\hat{e}^{*'}\hat{e}^* = \frac{1}{n(T-1)}S_1(\hat{\gamma}). \quad (9)$$

3.2. Computation issues

The computation of the least squares estimate of the threshold γ involves the minimization problem (8). Since the sum of squared error function $S_1(\gamma)$ depends on γ only through the indicator functions $I(q_{it} \leq \gamma)$, the sum of squared error function is a step function with at most nT steps, with the steps occurring at distinct values of the observed threshold variable q_{it} . Thus the minimization problem (8) can be reduced to searching over values of γ equalling the (at most nT) distinct values of q_{it} in the sample.

To implement the minimization, the following approach may be taken. Sort the distinct values of the observations on the threshold variable q_{it} . Eliminate the smallest and largest $\eta\%$ for some $\eta > 0$. The remaining N values constitute the values of γ which can be searched for $\hat{\gamma}$. For each of these N values, regressions (6) are estimated yielding the sum of squared errors (7). The smallest value of the latter yields the estimate $\hat{\gamma}$.

In practice, N may be a very large number, and the optimization search describe above may be numerically intensive. A simplifying shortcut which yields nearly identical results is to restrict the search to a smaller set of values of γ . Instead of searching over all values of q_{it} (between the $\eta\%$ and $(1 - \eta)\%$ quantile) the search may be limited to specific quantiles, perhaps integer valued. This greatly reduces the number of regressions performed in the search. The estimates from such an approximation are likely to be sufficiently precise for most applications of interest. For the empirical work reported in Section 4, we used the grid $\{1.00\%, 1.25\%, 1.50\%, 1.75\%, 2\%, \dots, 99.0\%\}$ which contains 393 quantiles.

4. Inference

4.1. Testing for a threshold

It is important to determine whether the threshold effect is statistically significant. The hypothesis of no threshold effect in (1) can be represented by the linear constraint

$$H_0: \beta_1 = \beta_2.$$

Under H_0 the threshold γ is not identified, so classical tests have non-standard distributions. This is typically called the ‘Davies’ Problem’ (see Davies, 1977, 1987) and has been recently investigated by Andrews and Ploberger (1994) and Hansen (1996). The fixed-effects equations (4) fall in the class of models considered by Hansen (1996) who suggested a bootstrap to simulate the asymptotic distribution of the likelihood ratio test.

Under the null hypothesis of no threshold, the model is

$$y_{it} = \mu_i + \beta_1' x_{it} + e_{it}. \quad (10)$$

After the fixed-effect transformation is made, we have

$$y_{it}^* = \beta_1' x_{it}^* + e_{it}^*. \quad (11)$$

The regression parameter β_1 is estimated by OLS, yielding estimate $\tilde{\beta}_1$, residuals \tilde{e}_{it}^* and sum of squared errors $S_0 = \tilde{e}^{*'} \tilde{e}^*$. The likelihood ratio test of H_0 is based on

$$F_1 = (S_0 - S_1(\hat{\gamma}))/\hat{\sigma}^2. \quad (12)$$

The asymptotic distribution of F_1 is non-standard, and strictly dominates the χ_k^2 distribution. Unfortunately, it appears to depend in general upon moments of the sample and thus critical values cannot be tabulated. Hansen (1996) shows

that a bootstrap procedure attains the first-order asymptotic distribution, so p-values constructed from the bootstrap are asymptotically valid.³ Given the panel nature of the data we recommend the following implementation of the bootstrap. Treat the regressors x_{it} and threshold variable q_{it} as given, holding their values fixed in repeated bootstrap samples. Take the regression residuals \hat{e}_{it}^* , and group them by individual: $\hat{e}_i^* = (\hat{e}_{i1}^*, \hat{e}_{i2}^*, \dots, \hat{e}_{iT}^*)$. Treat the sample $\{\hat{e}_1^*, \hat{e}_2^*, \dots, \hat{e}_n^*\}$ as the empirical distribution to be used for bootstrapping. Draw (with replacement) a sample of size n from the empirical distribution and use these errors to create a bootstrap sample under H_0 . (Notice that the test statistic F_1 does not depend on the parameter β_1 under H_0 , so any value of β_1 may be used.) Using the bootstrap sample, estimate the model under the null (11) and alternative (4) and calculate the bootstrap value of the likelihood ratio statistic F_1 (12). Repeat this procedure a large number of times and calculate the percentage of draws for which the simulated statistic exceeds the actual. This is the bootstrap estimate of the asymptotic p-value for F_1 under H_0 . The null of no threshold effect is rejected if the p-value is smaller than the desired critical value.

4.2. Asymptotic distribution of threshold estimate

When there is a threshold effect ($\beta_1 \neq \beta_2$) Chan (1993) and Hansen (1999) have shown that $\hat{\gamma}$ is consistent for γ_0 (the true value of γ) and that the asymptotic distribution is highly non-standard. Hansen (1999) argues that the best way to form confidence intervals for γ is to form the ‘no-rejection region’ using the likelihood ratio statistic for tests on γ . To test the hypothesis $H_0: \gamma = \gamma_0$, the likelihood ratio test is to reject for large values of $LR_1(\gamma_0)$ where

$$LR_1(\gamma) = (S_1(\gamma) - S_1(\hat{\gamma})) / \hat{\sigma}^2. \quad (13)$$

Note that the statistic (13) is testing a different hypothesis from the statistic (12) introduced in the previous section. $LR_1(\gamma_0)$ is testing $H_0: \gamma = \gamma_0$ while F_1 is testing $H_0: \beta_1 = \beta_2$.

Theorem 1. Under Assumptions 1–8 given in the Appendix, and $H_0: \gamma = \gamma_0$,

$$LR_1(\gamma) \rightarrow_d \xi$$

as $n \rightarrow \infty$, where ξ is a random variable with distribution function

$$P(\xi \leq x) = (1 - \exp(-x/2))^2. \quad (14)$$

³ Since the asymptotic distribution is non-pivotal, bootstrap size will not have an accelerated rate of convergence relative to conventional asymptotic approximations. A referee suggested that pre-pivoting as in Beran (1987) may improve the convergence rate. This is an interesting suggestion and would be a constructive subject for future research.

Theorem 1 shows that the asymptotic distribution of the likelihood ratio statistic is non-standard yet free of nuisance parameters. The technical assumptions include the rather unusual condition that $(\beta_2 - \beta_1) \rightarrow 0$ as $n \rightarrow \infty$, and is borrowed from the changepoint literature. The condition means that the difference in the slopes between the two regimes is ‘small’ relative to sample size. Its practical relevance is that the asymptotic approximation implied by Theorem 1 is likely to hold better for cases where $\beta_2 - \beta_1$ is small than for cases where $\beta_2 - \beta_1$ is large. If the threshold effect is large, however, the threshold will be quite precisely estimated.

Since the asymptotic distribution in Theorem 1 is pivotal, it may be used to form valid asymptotic confidence intervals. Furthermore, the distribution function (14) has the inverse

$$c(\alpha) = -2 \log(1 - \sqrt{1 - \alpha}), \quad (15)$$

from which it is easy to calculate critical values. For example, the 10% critical value is 6.53, the 5% is 7.35 and the 1% is 10.59. A test of $H_0: \gamma = \gamma_0$ rejects at the asymptotic level α if $LR_1(\gamma_0)$ exceeds $c(\alpha)$.

To form an asymptotic confidence interval for γ , the ‘no-rejection region’ of confidence level $1 - \alpha$ is the set of values of γ such that $LR_1(\gamma) \leq c(\alpha)$, where $LR_1(\gamma)$ is defined in (13) and $c(\alpha)$ is defined in (15). This is easiest to find by plotting $LR_1(\gamma)$ against γ and drawing a flat line at $c(\alpha)$.

One of the convenient features of this confidence region is that it is a natural by-product of model estimation. In order to find the LS estimate $\hat{\gamma}$, the sequence of sum of squared errors $S_1(\gamma)$ were calculated. The likelihood ratio sequence $LR_1(\gamma)$ is a simple re-normalization of these numbers, and require no further computation.

4.3. Asymptotic distribution of slope coefficients

The estimator $\hat{\beta} = \hat{\beta}(\hat{\gamma})$ depends on the threshold estimate $\hat{\gamma}$, which appears to complicate inference on β . Chan (1993) and Hansen (1999) show that the dependence on the threshold estimate is not of first-order asymptotic importance, so inference on β can proceed as if the threshold estimate $\hat{\gamma}$ were the true value. Hence $\hat{\beta}$ is asymptotically normal with a covariance matrix which can be estimated by

$$\hat{V} = \left(\sum_{i=1}^n \sum_{t=1}^T x_{it}^*(\hat{\gamma}) x_{it}^*(\hat{\gamma})' \right)^{-1} \hat{\sigma}^2.$$

While we need the assumption that the errors are iid for the purposes of constructing confidence intervals for γ , it would seem appropriate to relax this assumption when constructing confidence intervals for the slope coefficients. If

the errors are allowed to be conditionally heteroskedastic, the natural covariance matrix estimator for $\hat{\beta}$ is

$$\hat{V}_h = \left(\sum_{i=1}^n \sum_{t=1}^T x_{it}^*(\hat{\gamma}) x_{it}^*(\hat{\gamma})' \right)^{-1} \left(\sum_{i=1}^n \sum_{t=1}^T x_{it}^*(\hat{\gamma}) x_{it}^*(\hat{\gamma})' (\hat{e}_{it}^*)^2 \right) \\ \times \left(\sum_{i=1}^n \sum_{t=1}^T x_{it}^*(\hat{\gamma}) x_{it}^*(\hat{\gamma})' \right)^{-1}.$$

5. Multiple thresholds

Model (1) has a single threshold. In some applications there may be multiple thresholds. For example, the double threshold model takes the form

$$y_{it} = \mu_i + \beta'_1 x_{it} I(q_{it} \leq \gamma_1) + \beta'_2 x_{it} I(\gamma_1 < q_{it} \leq \gamma_2) + \beta'_3 x_{it} I(\gamma_2 < q_{it}) + e_{it} \quad (16)$$

where the thresholds are ordered so that $\gamma_1 < \gamma_2$. We will focus on this double-threshold model since the methods extend in a straightforward manner to higher-order threshold models. We discuss three relevant statistical issues: (1) Estimation; (2) Testing for the presence of a double threshold; (3) Construction of confidence intervals for the threshold parameters γ_1 and γ_2 .

5.1. Estimation

For given (γ_1, γ_2) , (16) is linear in the slopes $(\beta_1, \beta_2, \beta_3)$ so OLS estimation is appropriate. Thus for given (γ_1, γ_2) the concentrated sum of squared errors $S(\gamma_1, \gamma_2)$ is straightforward to calculate (as in the single threshold model). The joint LS estimates of (γ_1, γ_2) are by definition the values which jointly minimize $S(\gamma_1, \gamma_2)$. While these estimates might seem desirable, they may be quite cumbersome to implement in practice. A grid search over (γ_1, γ_2) requires approximately $N^2 = (nT)^2$ regressions which may be prohibitively expensive.

A remarkable insight allows us to escape this computational burden. It has been found (Chong, 1994; Bai, 1997; Bai and Perron, 1998) in the multiple changepoint model that sequential estimation is consistent. The same logic appears to apply to the multiple threshold model. The method works as follows. In the first stage, let $S_1(\gamma)$ be the single threshold sum of squared errors as defined in (7) and let $\hat{\gamma}_1$ be the threshold estimate which minimizes $S_1(\gamma)$. The analysis of Chong and Bai suggests that $\hat{\gamma}_1$ will be consistent⁴ for either γ_1 or γ_2 (depending on which effect is ‘stronger’).

⁴ The reason why $\hat{\gamma}_1$ is consistent is because the single-threshold sum of squared errors function $S_1(\gamma)$ asymptotically converges to a limit function which has two local minima at γ_1 and γ_2 .

Fixing the first-stage estimate $\hat{\gamma}_1$, the second-stage criterion is

$$S_2^r(\gamma_2) = \begin{cases} S(\hat{\gamma}_1, \gamma_2) & \text{if } \hat{\gamma}_1 < \gamma_2 \\ S(\gamma_2, \hat{\gamma}_1) & \text{if } \gamma_2 < \hat{\gamma}_1 \end{cases} \tag{17}$$

and the second-stage threshold estimate is

$$\hat{\gamma}_2^r = \underset{\gamma_2}{\operatorname{argmin}} S_2^r(\gamma_2). \tag{18}$$

Since it is undesirable to have a small number of observations in any given ‘regime’, we can restrict the search in (18) so that a minimum number of observations fall in each of the three regimes.

Bai (1997) has shown that $\hat{\gamma}_2^r$ is asymptotically efficient, but $\hat{\gamma}_1$ is not. This is because the estimate $\hat{\gamma}_1$ was obtained from a sum of squared errors function which was contaminated by the presence of a neglected regime. The asymptotic efficiency of $\hat{\gamma}_2^r$ suggests that $\hat{\gamma}_1$ can be improved by a third-stage estimation. Bai (1997) suggests the following *refinement* estimator. Fixing the second-stage estimate $\hat{\gamma}_2^r$, define the refinement criterion

$$S_1^r(\gamma_1) = \begin{cases} S(\gamma_1, \hat{\gamma}_2^r) & \text{if } \gamma_1 < \hat{\gamma}_2^r, \\ S(\hat{\gamma}_2^r, \gamma_1) & \text{if } \hat{\gamma}_2^r < \gamma_1, \end{cases} \tag{19}$$

and the refinement estimate

$$\hat{\gamma}_1^r = \underset{\gamma_1}{\operatorname{argmin}} S_1^r(\gamma_1). \tag{20}$$

Bai (1997) shows that the refinement estimator $\hat{\gamma}_1^r$ is asymptotically efficient in changepoint estimation, and we expect similar results to hold in threshold regression.

5.2. Determining number of thresholds

In the context of model (16), there are either no thresholds, one threshold, or two thresholds. In Section 3.1 we introduced F_1 as a test of no thresholds against one threshold, and suggested a bootstrap to approximate the asymptotic p-value. If F_1 rejects the null of no threshold, in the context of model (16) we need a further test to discriminate between one and two thresholds.

The minimizing sum of squared errors from the second-stage threshold estimate is $S_2^r(\hat{\gamma}_2^r)$ with variance estimate $\hat{\sigma}^2 = S_2^r(\hat{\gamma}_2^r)/n(T - 1)$. Thus an approximate likelihood ratio test of one versus two thresholds can be based on the statistic

$$F_2 = \frac{S_1(\hat{\gamma}_1) - S_2^r(\hat{\gamma}_2^r)}{\hat{\sigma}^2}.$$

The hypothesis of one threshold is rejected in favor of two thresholds if F_2 is large.

Since the null asymptotic distribution of the likelihood ratio test is non-pivotal⁵ we suggest using a bootstrap procedure to approximate the sampling distribution. To generate the bootstrap samples, hold the regressors x_{it} and threshold variable q_{it} fixed in repeated bootstrap samples. The bootstrap errors will be drawn from the residuals calculated under the alternative hypothesis, so should be the residuals from LS estimation of model (16). Group the regression residuals \hat{e}_{it}^* by individual: $\hat{e}_i^* = (\hat{e}_{i1}^*, \hat{e}_{i2}^*, \dots, \hat{e}_{iT}^*)$, and treat the sample $\{\hat{e}_1^*, \hat{e}_2^*, \dots, \hat{e}_n^*\}$ as an empirical distribution. Draw (with replacement) error samples from the empirical distribution. Let $e_i^\#$ denote a generic $T \times 1$ draw. The dependent variable y_{it} should be generated under the null hypothesis of a single threshold (1), so use the equation

$$y_{it}^\# = \hat{\beta}_1' x_{it} I(q_{it} \leq \hat{\gamma}) + \hat{\beta}_2' x_{it} I(q_{it} > \hat{\gamma}) + e_i^\#, \quad (21)$$

which depends on the parameter values $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\gamma}$, the least-squares estimates from the single threshold model. From the bootstrap sample, the test statistic F_2 may be calculated, and this procedure repeated multiple times to calculate the bootstrap p-value.

From Eq. (21) it is clear that unlike the null sampling distribution of F_1 , which asymptotically did not depend on γ , β_1 or β_2 , the null sampling distribution of F_2 depends asymptotically on both γ and the regression parameters β_1 and β_2 , though it only depends on the latter through $\beta_1 - \beta_2$. This leads us to expect that the bootstrap may not produce as accurate critical values for F_2 as for F_1 , and neither is expected to be second-order accurate.

5.3. Confidence region construction

We finally consider the construction of confidence intervals for the two threshold parameters $\{\gamma_1, \gamma_2\}$. Bai (1997) showed (for the analogous case of change-point models) that the refinement estimators of Section 5.1 have the same asymptotic distributions as the threshold estimate in a single threshold model. This suggests that we can construct confidence intervals in the same way as in Section 4.2.

Let

$$LR_2^r(\gamma) = \frac{S_2^r(\gamma) - S_2^r(\hat{\gamma}_2)}{\hat{\sigma}^2}$$

⁵ It is important to remember that this differs from the changepoint case (see Chong, 1994; Bai, 1997; Bai and Perron, 1998), where the asymptotic distribution of F_2 is known and pivotal.

and

$$LR_1^r(\gamma) = \frac{S_1^r(\gamma) - S_1^r(\hat{\gamma}_1^r)}{\hat{\sigma}^2},$$

where $S_2^r(\gamma)$ and $S_1^r(\gamma)$ are defined in (17) and (19), respectively. Our asymptotic $(1 - \alpha)\%$ confidence intervals for γ_2 and γ_1 are the set of values of γ such that $LR_2^r(\gamma) \leq c(\alpha)$ and $LR_1^r(\gamma) \leq c(\alpha)$, respectively.

6. Investment and financing constraints

Classical models of the firm assume the existence of perfect financial markets on which firms can borrow the needed resources for investment projects. Alternative models of financing place restrictions on the extent of external financing. An important empirical question is whether or not there exist firms which behave as though they are subject to such constraints.

A well-cited paper which explored the empirical implications of financing constraints is Fazzari et al. (1988), henceforth FHP. These authors argue that the presence of financing constraints implies that a firm's cash flow will be positively related to its investment rate only when the firm faces constraints on external financing. If a firm is free to borrow on external financial markets, cash flow will be irrelevant for investment. This distinction motivated FHP to test for financing constraints by estimating separate investment regressions for 'constrained' and 'unconstrained' firms to see if there are differing effects of contemporaneous cash flow. To distinguish constrained and unconstrained firms, they used the dividend to income ratio, as their theory suggests that a financially constrained firm will choose to retain earnings rather than pay dividends. Hence the firms which have low levels of dividend payments are the financially constrained firms.

FHP divide their sample into three classes, depending on whether the dividend to income ratio was less than 0.1 for 10 yr in the sample, between 0.1 and 0.2 for over 10 yr, and all other firms. Thus they are estimating a double-threshold regression on panel data, where q_{it} is the largest dividend-income ratio over the 10-yr period, and the thresholds γ are set at 0.1 and 0.2.

There are two obvious problems with the FHP regression. First, it treats the dividend-income ratio as exogenous, while their theory explicitly treats dividend payments as decision variables. The use of an endogenous threshold variable may bias their results. Second, they select their threshold levels arbitrarily, rather than estimating these parameters from the sample. In this section we explore whether our methods allow for a re-appraisal of FHP's analysis.

The original data used by FHP is no longer available. We use a similar dataset, extracted from the dataset used by Hall and Hall (1993), which is an unbalanced panel of US firms originally taken from Compustat. Our methods

are designed for balanced panels, so we took the subset of 565 firms which are observed for the years 1973–1987.

The threshold variable should be an exogenous indicator of a firm's access to external financing. A natural candidate is the existing debt level. It seems reasonable to believe that banks will be reluctant to lend money to debt-heavy firms. This choice is similar to that of Hu and Schiantarelli (1998) who estimate a switching regression using the debt–asset ratio as one variable in their switching equation.

To fix notation, let I_{it} be the ratio of investment to capital; Q_{it} be the ratio of total market value to assets; CF_{it} be the ratio of cash flow to assets; and D_{it} be the ratio of long-term debt to assets, where stock variables are defined at the end of year. Summary statistics of the four variables are given in Table 1.

We use the multiple threshold regression model

$$\begin{aligned}
 I_{it} = & \mu_i + \theta_1 Q_{it-1} + \theta_2 Q_{it-1}^2 + \theta_3 Q_{it-1}^3 + \theta_4 D_{it-1} \\
 & + \theta_5 Q_{it-1} D_{it-1} + \beta_1 CF_{it-1} I(D_{it-1} \leq \gamma_1) \\
 & + \beta_2 CF_{it-1} I(\gamma_1 < D_{it-1} \leq \gamma_2) + \beta_3 CF_{it-1} I(\gamma_2 < D_{it-1}) + e_{it}, \quad (22)
 \end{aligned}$$

where (22) represents a double threshold model for illustration. Model (22) falls in the class of models (1) setting $q_{it} = D_{it-1}$ and $x_{it} = CF_{it-1}$. There are also the additional regressors (Q_{it-1} , Q_{it-1}^2 , Q_{it-1}^3 , D_{it-1} , $Q_{it-1} D_{it-1}$). The latter can be viewed as a special case of (1) by constraining the slope coefficients on these variables to be the same in the two regimes, which has no effect on the distribution theory. The reason model (22) has only the slope coefficient on cash flow switch between regimes is to focus attention on this key variable of interest. The non-linear terms in the regression (namely, Q_{it-1}^2 , Q_{it-1}^3 , and $Q_{it-1} D_{it-1}$) were included to reduce the possibility of spurious correlations due to omitted variables bias. The choice of the particular non-linear terms was data-based, as the variables D_{it-1}^2 and D_{it-1}^3 were insignificant and omitted to reduce computation costs.

To determine the number of thresholds, model (22) was estimated by least squares, allowing for (sequentially) zero, one, two, and three thresholds. The test statistics F_1 , F_2 and F_3 , along with their bootstrap⁶ p-values, are shown in Table 2. We find that the test for a single threshold F_1 is highly significant with a bootstrap p-value of 0.003, and the test for a double threshold F_2 is also strongly significant, with a bootstrap p-value of 0.017. On the other hand, the test for a third threshold F_3 is not close to being statistically significant, with a bootstrap p-value of 0.723. We conclude that there is strong evidence that there are two thresholds in the regression relationship. For the remainder of the analysis we work with this double threshold model.

⁶ 300 bootstrap replications were used for each of the three bootstrap tests.

Table 1
Summary statistics

	Minimum	25% quantile	Median	75% quantile	Maximum
I_{it}	0.001	0.049	0.076	0.113	1.66
Q_{it-1}	0.021	0.371	0.675	1.31	111.8
CF_{it-1}	− 0.94	0.12	0.22	0.32	8.71
D_{it-1}	0.000	0.089	0.206	0.320	4.67

Table 2
Tests for threshold effects

<i>Test for single threshold</i>				
F_1				32.6
P-value				0.003
(10%, 5%, 1% critical values)				(12.4, 14.8, 26.2)
<i>Test for double threshold</i>				
F_2				25.8
P-value				0.017
(10%, 5%, 1% critical values)				(12.3, 14.9, 42.9)
<i>Test for triple threshold</i>				
F_3				4.2
P-value				0.723
(10%, 5%, 1% critical values)				(10.9, 13.3, 22.9)

The point estimates of the two thresholds and their asymptotic 95% and 99% confidence intervals are reported in Table 3. The estimates are 0.016 and 0.536, which are very small (and very large) values in the empirical distribution of the debt/assets threshold variable. Thus the three classes of firms indicated by the point estimates are those with ‘very low debt’, ‘very high debt’ and ‘other’. The asymptotic confidence intervals for the threshold are very tight, indicating little uncertainty about the nature of this division. More information can be learned about the threshold estimates from plots of the concentrated likelihood ratio function $LR_1(\gamma)$, $LR_2^t(\gamma)$ and $LR_1^t(\gamma)$ in Figs. 1–3 (corresponding to the first-stage estimate $\hat{\gamma}_1$ and the refinement estimators $\hat{\gamma}_2^t$ and $\hat{\gamma}_1^t$). The point estimates are the value of γ at which the likelihood ratio hits the zero axis, which is in the far left part of the graph. The 95% confidence intervals for γ_2 and γ_1 can be found from $LR_2^t(\gamma)$ and $LR_1^t(\gamma)$ by the values of γ for which the likelihood ratio lies beneath the dotted line.

It is interesting to examine the unrefined first-step likelihood ratio function $LR_1(\gamma)$, which is computed when estimating a single threshold model. The

Table 3
Threshold estimates

	Estimate	95% confidence interval	99% confidence interval
$\hat{\gamma}_1^r$	0.0157	[0.0139, 0.0181]	[0.0120, 0.0239]
$\hat{\gamma}_2^r$	0.5362	[0.5305, 0.5629]	[0.5190, 0.5693]

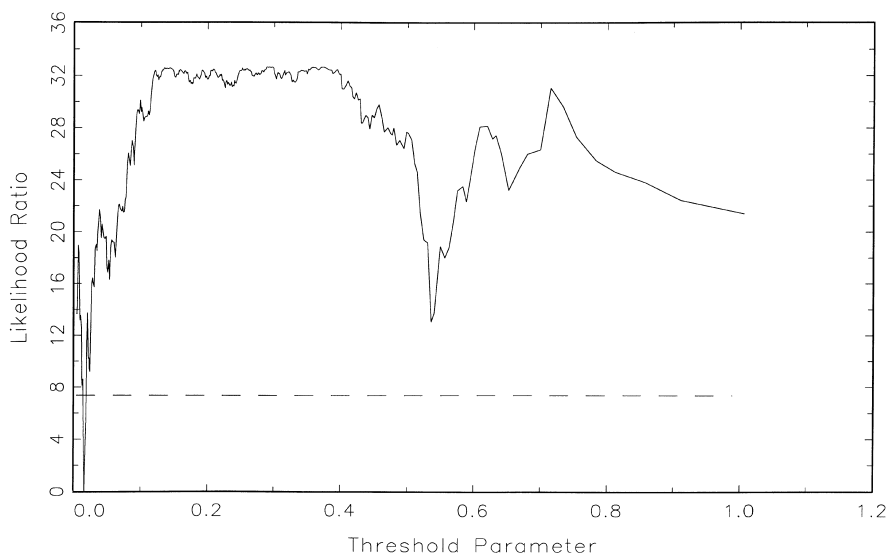


Fig. 1. Confidence interval construction in single threshold model.

first-step threshold estimate is the point where the $LR_1(\gamma)$ equals zero, which occurs at $\hat{\gamma}_1 = 0.0157$. There is a second major dip in the likelihood ratio around the second-step estimate 0.53. Thus the single threshold likelihood conveys information that suggests that there is a second threshold in the regression.

Table 4 reports the percentage of firms which fall into the three regimes each year. We see that the percentage of firms in the ‘very low debt’ category ranges from 10% to 16% of the sample over the years. The ‘very high debt’ firms range from 4% to 16% of the sample in a given year. It is interesting to note that the last two years of the sample (1986 and 1987) saw a large increase in the number of firms with very high debt ratios.

The regression slope estimates, conventional OLS standard errors, and White-corrected standard errors are displayed in Table 5. We see that Q_{it-1} and its powers are statistically significant, indicating a positive (and very slightly

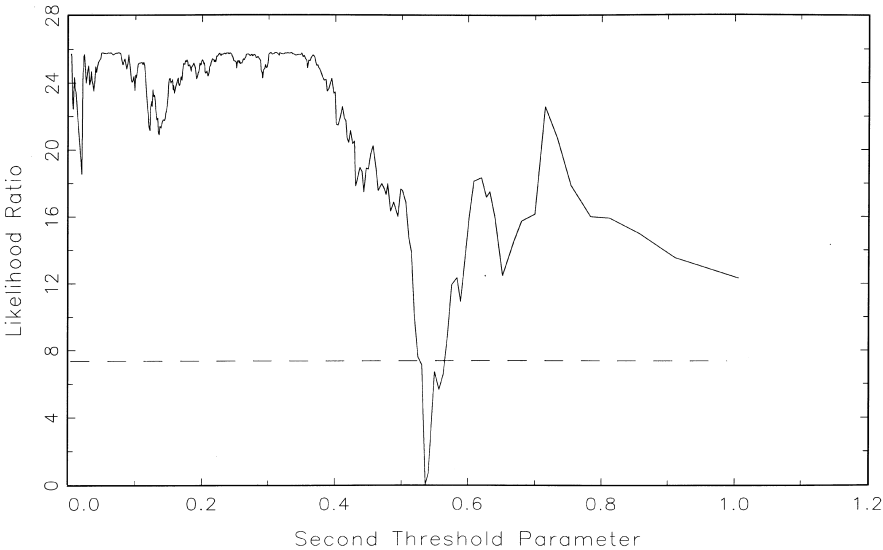


Fig. 2. Confidence interval construction in double threshold model.

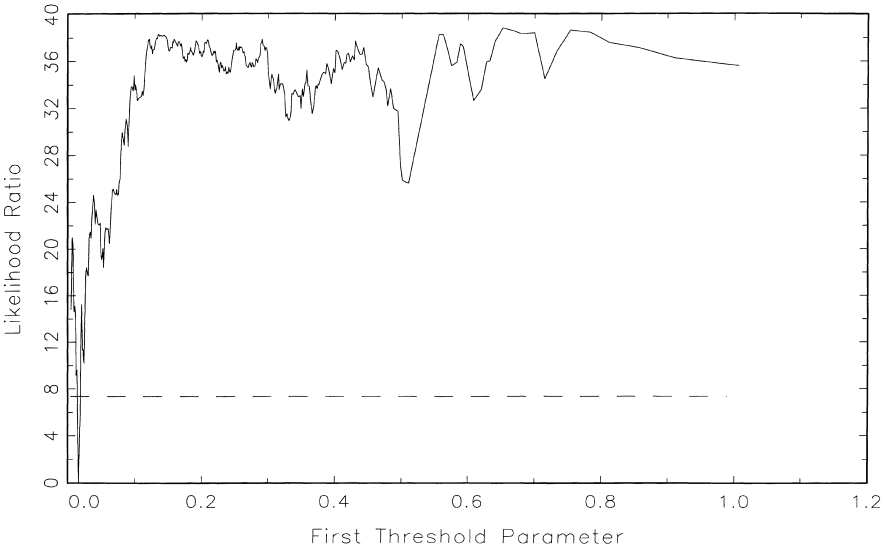


Fig. 3. Confidence interval construction in double threshold model.

non-linear) relationship between q and investment. The debt level D_{it-1} has a negative and significant effect on investment, and there is no apparent interaction effect between q and the debt level.

Table 4
Percentage of firms in each regime by year

Firm class	Year													
	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
$D_{it-1} \leq 0.0157$	16	14	14	15	15	13	13	11	10	10	10	10	10	11
$0.0157 < D_{it-1} \leq 0.5362$	78	79	78	81	81	84	82	85	86	85	84	82	77	73
$0.5362 < D_{it-1}$	6	7	8	5	4	4	5	4	4	5	6	8	13	16

Table 5
Regression estimates: double threshold model

Regressor	Coefficient estimate	OLS SE	White SE
Q_{it-1}	0.010	0.001	0.002
$Q_{it-1}^2/10^3$	-0.198	0.026	0.064
$Q_{it-1}^3/10^6$	1.047	0.199	0.448
D_{it-1}	-0.016	0.005	0.009
$Q_{it-1}D_{it-1}$	0.001	0.001	0.002
$CF_{it-1}I(D_{it-1} \leq 0.0157)$	0.063	0.006	0.014
$CF_{it-1}I(0.0157 < D_{it-1} \leq 0.5362)$	0.098	0.006	0.010
$CF_{it-1}I(0.5362 < D_{it-1})$	0.039	0.012	0.031

The coefficients of primary interest are those on cash flow. The point estimates suggest that investment is positively related to cash flow, with ‘very low debt’ firms having a lower coefficient (about one-third smaller in magnitude) than the typical firm. What is quite unexpected is that the firms with the highest debt levels have the smallest coefficient of 0.039. The White standard error on this last coefficient, however, is quite high, indicating that there is still considerable uncertainty in the estimate.

The conventional OLS standard errors and the White-corrected standard errors are considerably different, with the White-corrected ones roughly twice as big. This is evidence in favor of heteroskedasticity, which violates one of the maintained assumptions of our asymptotic analysis. Based on the theory (Hansen, 1999) for least squares threshold regression (the model without fixed effects), we would expect the threshold estimates to be consistent and the distribution theory of Theorem 1 to be correct up to a scale effect, so that asymptotic confidence intervals would still take the form given in Table 3, but would require a different critical value.

7. Conclusion

This paper has developed new empirical methods for panel data. We have defined a threshold regression model with individual-specific effects, and shown that the model is rather straightforward to estimate using a fixed-effects transformation. The asymptotic theory is non-standard, but confidence intervals for the threshold can be constructed by inverting the likelihood ratio statistic, and this construction is a natural by-product of the estimation method.

The methods are applied to the investment decisions of a panel of 565 US firms for the period 1973–1987. We find overwhelming evidence of a double threshold effect which separates the firms based on their debt to asset ratio. The estimates are somewhat consistent with the theory of financing constraints. The notable difference between our work and that of Fazzari et al. (1988) is that we are also able to quantify the extent of financing constraints in the economy rather than assuming the degree of such constraints.

Several extensions of our methods would be desirable, including allowing for heteroskedasticity, lagged dependent variables, endogenous variables, and random effects. It would also be interesting to compare our results with alternative approximations based on smooth transition threshold models, which replace the indicator functions by smooth distribution functions. These would be useful subjects for future research.

Acknowledgements

This research was supported by a grant from the National Science Foundation and a Sloan Foundation Research Fellowship. Special thanks go to Fabio Schiantarelli for many helpful discussions and motivation, to Richard Blundell for an insightful discussion, two referees for correcting some of my errors, and to Maria Laura Parisi for research assistance.

Appendix. Mathematical proofs

We need the following technical assumptions. Let γ_0 denote the true value of γ . Let $\theta = \beta_2 - \beta_1$ and $C = n^\alpha \theta$, where $\alpha \in (0, 1/2)$. Let $f_i(\gamma)$ denote the density function of q_{it} , set $z_{it} = C'x_{it}$,

$$D(\gamma) = \sum_{t=1}^T E(z_{it}^2 | q_{it} = \gamma) f_i(\gamma),$$

and $D = D(\gamma_0)$. Let $f_{k|l}(\gamma_1 | \gamma_2)$ denote the conditional density of q_{ik} given q_{il} .

Assumptions

1. For each t , (q_{it}, x_{it}, e_{it}) are independent and identically distributed (iid) across i .
2. For each i , e_{it} is iid over t , is independent of $\{(x_{ij}, q_{ij})_{j=1}^T\}$, and $E(e_{it}) = 0$.
3. For each $j = 1, \dots, k$, $P(x_{i1}^j = x_{i2}^j = \dots = x_{iT}^j) < 1$, where x_{it}^j is the j th element of x_{it} .
4. $E|x_{it}|^4 < \infty$ and $E|e_{it}|^4 < \infty$.
5. For some fixed $C < \infty$ and $0 < \alpha < 1/2$, $\theta = n^{-\alpha}C$.
6. $D(\gamma)$ is continuous at $\gamma = \gamma_0$.
7. $0 < D < \infty$.
8. For $k > t$, $f_{k|t}(\gamma_0 | \gamma_0) < \infty$.

Assumptions 1–4 are standard for fixed effect panel models with strictly exogenous regressors. Assumption 5 is more unusual, setting $\theta = n^{-\alpha}C \rightarrow 0$ as $n \rightarrow \infty$. The renormalization is to force θ to be ‘small’, reducing the information in the sample concerning the threshold and hence slowing down the rate of convergence of the threshold estimate. This assumption need not be viewed as very restrictive since the rate at which θ decreases to zero can be set quite low. It does suggest, however, that the asymptotic approximation is more likely to provide good approximations when θ is small relative to the case where θ is large.

Assumption 6 excludes threshold effects which occur simultaneously in the marginal distribution of the regressors and in the regression function. Assumption 7 excludes continuous threshold models (see Chan and Tsay, 1998), and requires that the threshold variable q_{it} be continuously distributed with positive support at the threshold γ_0 . Assumption 8 excludes the possibility that $q_{it} = \gamma_0$ for $t = 1, \dots, T$.

The proof of the theorem is based on the following lemma. Let ‘ \Rightarrow ’ denote weak convergence with respect to the uniform metric, and let $\lambda_n = n^{1-2\alpha}$.

Lemma A.1. As $n \rightarrow \infty$, uniformly over $v \in [-\bar{v}, \bar{v}]$,

$$S_1(\gamma_0) - S_1(\gamma_0 + v/\lambda_n) \Rightarrow q(v),$$

where

$$q(v) = -D_T|v| + 2\sqrt{\sigma^2 D_T}W(v)$$

$D_T = D(1 - 1/T)$, and $W(v)$ is a double-sided standard Brownian motion on $(-\infty, \infty)$.

Proof. Let

$$\nabla z_{it}(\gamma) = z_{it}I(q_{it} \leq \gamma) - z_{it}I(q_{it} \leq \gamma_0).$$

The regression equation (1) holds when $\gamma = \gamma_0$, the true value. For values of $\gamma \neq \gamma_0$, note that (1) can be re-written as

$$\begin{aligned}
 y_{it} &= \mu_i + \beta'_1 x_{it} I(q_{it} \leq \gamma_0) + \beta'_2 x_{it} I(q_{it} > \gamma_0) + e_{it} \\
 &= \mu_i + \beta'_1 x_{it} I(q_{it} \leq \gamma) + \beta'_2 x_{it} I(q_{it} > \gamma) \\
 &\quad - \beta'_1 x_{it} [I(q_{it} \leq \gamma) - I(q_{it} \leq \gamma_0)] - \beta'_2 x_{it} [I(q_{it} > \gamma) - I(q_{it} > \gamma_0)] + e_{it} \\
 &= \mu_i + \beta' x_{it}(\gamma) + (\beta_2 - \beta_1)' x_{it} [I(q_{it} \leq \gamma) - I(q_{it} \leq \gamma_0)] + e_{it} \\
 &= \mu_i + \beta' x_{it}(\gamma) + n^{-\alpha} \nabla z_{it}(\gamma) + e_{it}.
 \end{aligned} \tag{A.1}$$

Eq. (A.1) makes explicit the regression error for $\gamma \neq \gamma_0$. \square

The fixed effect transformation is linear, so can be applied to (A.1) to yield

$$y_{it}^* = \beta' x_{it}^*(\gamma) + n^{-\alpha} \nabla z_{it}^*(\gamma) + e_{it}^*$$

which is the correct representation of (4) for $\gamma \neq \gamma_0$. Hansen (1999) shows that the asymptotic distribution of $\hat{\gamma}$ is not affected by the estimation of β , and this holds in our environment as well. We can thus simplify matters by assuming that β is known and only γ is estimated, so that the regression residual (for fixed γ) is

$$\hat{e}_{it}(\gamma) = n^{-\alpha} \nabla z_{it}^*(\gamma) + e_{it}^*. \tag{A.2}$$

Using (A.2),

$$\begin{aligned}
 S(\gamma_0) - S(\gamma) &= \sum_{i=1}^n \sum_{t=1}^T \hat{e}_{it}(\gamma_0)^2 - \sum_{i=1}^n \sum_{t=1}^T \hat{e}_{it}(\gamma)^2 \\
 &= \sum_{i=1}^n \sum_{t=1}^T e_{it}^{*2} - \sum_{i=1}^n \sum_{t=1}^T (n^{-\alpha} \nabla z_{it}^*(\gamma) + e_{it}^*)^2 \\
 &= -n^{-2\alpha} \sum_{i=1}^n \sum_{t=1}^T \nabla z_{it}^*(\gamma)^2 - 2n^{-\alpha} \sum_{i=1}^n \sum_{t=1}^T \nabla z_{it}^*(\gamma) e_{it}^*.
 \end{aligned} \tag{A.3}$$

We now show that as $n \rightarrow \infty$, uniformly over $v \in [-\bar{v}, \bar{v}]$,

$$n^{-2\alpha} \sum_{i=1}^n \sum_{t=1}^T \nabla z_{it}^*(\gamma_0 + v/\lambda_n)^2 \Rightarrow D_T |v|. \tag{A.4}$$

We prove (A.4) for the case $v \in [0, \bar{v}]$. We will show that for $\gamma = \gamma_0 + v/\lambda_n$,

$$E \left(n^{-2\alpha} \sum_{i=1}^n \sum_{t=1}^T \nabla z_{it}^*(\gamma)^2 \right) = \lambda_n \sum_{t=1}^T E(\nabla z_{it}^*(\gamma))^2 \rightarrow D_T |v|. \tag{A.5}$$

Arguments similar to those in the proof of Lemma A.10 of Hansen (1999) show that (A.5) implies (A.4) under the assumptions. Expansion of the quadratic yields

$$\sum_{t=1}^T E(\nabla z_{it}^*(\gamma))^2 = \sum_{t=1}^T E(\nabla z_{it}(\gamma))^2 - \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^T E(\nabla z_{it}(\gamma) \nabla z_{ik}(\gamma)). \quad (\text{A.6})$$

Consider the first sum on the right-hand-side of (A.6). Observe that since $\gamma = \gamma_0 + v/\lambda_n \rightarrow \gamma_0$,

$$\begin{aligned} \lambda_n P(\gamma_0 < q_{it} \leq \gamma) &= v \frac{P(q_{it} \leq \gamma) - P(q_{it} \leq \gamma_0)}{\gamma - \gamma_0} \\ &\rightarrow v f_t(\gamma_0) \end{aligned} \quad (\text{A.7})$$

as $n \rightarrow \infty$. Thus

$$\begin{aligned} \lambda_n \sum_{t=1}^T E(\nabla z_{it}(\gamma))^2 &= \lambda_n \sum_{t=1}^T E(z_{it}^2 I(\gamma_0 < q_{it} \leq \gamma)) \\ &= \sum_{t=1}^T E(z_{it}^2 | \gamma_0 < q_{it} \leq \gamma) \lambda_n P(\gamma_0 < q_{it} \leq \gamma) \\ &\rightarrow \sum_{t=1}^T v E(z_{it}^2 | q_{it} = \gamma_0) f_t(\gamma_0) = vD. \end{aligned} \quad (\text{A.8})$$

Next consider the double-sum on the right-hand-side of (A.6). By Assumption 8, for $k > t$,

$$\begin{aligned} &P(\gamma_0 < q_{ik} \leq \gamma | \gamma_0 < q_{it} \leq \gamma) \\ &= v \lambda_n^{-1} \frac{[P(q_{ik} \leq \gamma | \gamma_0 < q_{it} \leq \gamma) - P(q_{ik} \leq \gamma_0 | \gamma_0 < q_{it} \leq \gamma)]}{\gamma - \gamma_0} \\ &= v \lambda_n^{-1} (f_{k|t}(\gamma_0 | \gamma_0) + o(1)) \\ &\rightarrow 0, \end{aligned}$$

and combined with (A.7), for $k > t$,

$$\begin{aligned} &\lambda_n P(\gamma_0 < q_{it} \leq \gamma, \gamma_0 < q_{ik} \leq \gamma) \\ &= \lambda_n P(\gamma_0 < q_{ik} \leq \gamma) P(\gamma_0 < q_{it} \leq \gamma | \gamma_0 < q_{ik} \leq \gamma) \rightarrow 0. \end{aligned} \quad (\text{A.9})$$

Eq. (A.9) also holds for $k < t$ by symmetry. Hence

$$\begin{aligned} &\lambda_n \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^T E(\nabla z_{it}(\gamma) \nabla z_{ik}(\gamma)) \\ &= \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^T \lambda_n E(z_{it} z_{ik} I(\gamma_0 < q_{it} \leq \gamma) I(\gamma_0 < q_{ik} \leq \gamma)) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^T E(z_{it}^2 | \gamma_0 < q_{it} \leq \gamma, \gamma_0 < q_{ik} \leq \gamma) \\
&\quad \cdot \lambda_n P(\gamma_0 < q_{it} \leq \gamma, \gamma_0 < q_{ik} \leq \gamma) \\
&= \frac{1}{T} \sum_{t=1}^T E(z_{it}^2 | \gamma_0 < q_{it} \leq \gamma) \lambda_n P(\gamma_0 < q_{it} \leq \gamma) + o(1) \\
&\rightarrow \frac{1}{T} \sum_{t=1}^T E(z_{it}^2 | q_{it} = \gamma_0) v f_t(\gamma_0) = \frac{1}{T} v D.
\end{aligned} \tag{A.10}$$

by (A.9). Eqs. (A.6), (A.8) and (A.10) imply (A.5) and hence (A.4). Next, we wish to show that uniformly over $v \in [-\bar{v}, \bar{v}]$,

$$n^{-\alpha} \sum_{i=1}^n \sum_{t=1}^T \nabla z_{it}^*(\gamma_0 + v/\lambda_n) e_{it}^* \Rightarrow \sqrt{\sigma^2 D_T} W(v). \tag{A.11}$$

By the properties of least squares projection, $\sum_{t=1}^T \nabla z_{it}^*(\gamma) e_{it}^* = \sum_{t=1}^T \nabla z_{it}^*(\gamma) e_{it}$, and since the e_{it} are iid,

$$\begin{aligned}
E \left(n^{-\alpha} \sum_{i=1}^n \sum_{t=1}^T \nabla z_{it}^*(\gamma) e_{it}^* \right)^2 &= \lambda_n E \left(\sum_{t=1}^T \nabla z_{it}^*(\gamma) e_{it} \right)^2 \\
&= \lambda_n \sum_{t=1}^T E (\nabla z_{it}^*(\gamma))^2 \sigma^2 \\
&\rightarrow D_T |v| \sigma^2
\end{aligned} \tag{A.12}$$

by (A.5). This establishes that the finite dimensional distributions of the stochastic process are those of the stated double-sided Brownian motion. By arguments identical to those in the proof of Lemma A.11 of Hansen (1999), (A.12) and Assumption 1 are sufficient to establish (A.11). Eqs. (A.4) and (A.11) combine with (A.3) to yield the stated result.

Proof of Theorem 1. Since $\hat{\gamma}$ minimizes $S(\gamma)$,

$$\begin{aligned}
LR_1(\gamma_0) &= \max_{\gamma} \left(\frac{S_1(\gamma_0) - S_1(\gamma)}{\hat{\sigma}^2} \right) \\
&= \max_v \left(\frac{S_1(\gamma_0) - S_1(\gamma_0 + v/\lambda_n)}{\hat{\sigma}^2} \right),
\end{aligned} \tag{A.13}$$

where the final equality makes the change-of-variables $\gamma = \gamma_0 + v/\lambda_n$, Hansen (1999) shows that under Assumption 1,

$$\hat{v} \equiv \lambda_n(\hat{\gamma} - \gamma_0) = O_p(1). \tag{A.14}$$

We will not repeat the proof of (A.14) here. The stochastic boundedness of (A.14) shows that for any $\eta > 0$, there is some $\bar{v} < \infty$ such that

$$P(|\hat{v}| \leq \bar{v}) \geq 1 - \eta. \quad (\text{A.15})$$

Let

$$\begin{aligned} \widetilde{LR} &= \max_{|v| \leq \bar{v}} \left(\frac{S_1(\gamma_0) - S_1(\gamma_0 + v/\lambda_n)}{\hat{\sigma}^2} \right) \\ &\Rightarrow \sigma^{-2} \max_{|v| \leq \bar{v}} q(v) \end{aligned}$$

where the stated weak convergence follows from Lemma A.1 and the continuous mapping theorem. Eq. (A.15) shows that

$$P(LR_1(\gamma_0) = \widetilde{LR}) \geq 1 - \eta.$$

Since η is arbitrary we conclude that

$$LR_1(\gamma_0) \Rightarrow \sigma^{-2} \max_{-\infty < v < \infty} q(v) = \zeta,$$

say. Hansen (1999, Proof of Theorem 2) shows that the distribution function of ζ is $P(\zeta \leq x) = (1 - \exp(-x/2))^2$. \square

References

- Abel, A.B., Eberly, J.C., 1994. A unified model of investment under uncertainty. *American Economic Review* 84, 1369–1384.
- Abel, A.B., Eberly, J.C., 1996. Investment and q with fixed costs: An empirical analysis. Working paper, University of Pennsylvania.
- Andrews, D.W.K., Ploberger, W., 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62, 1383–1414.
- Bai, J., 1997. Estimating multiple breaks one at a time. *Econometric Theory* 13, 315–352.
- Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66, 47–78.
- Barnett, S.A., Sakellaris, P., 1998. Non-linear response of firm investment to q : Testing a model of convex and non-convex adjustment costs. *Journal of Monetary Economics* 42, 261–288.
- Beran, R., 1987. Pivoting to reduce the level error of confidence sets. *Biometrika* 74, 457–468.
- Chan, K.S., 1993. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics* 21, 520–533.
- Chan, K.S., Tsay, R.S., 1998. Limiting properties of the least squares estimator of a continuous threshold autoregressive model. *Biometrika* 85, 413–426.
- Chong, T. T-L., 1994. Consistency of change-point estimators when the number of change-points in structural change models is underspecified. Working paper, Chinese University of Hong Kong.
- Davies, R.B., 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–254.
- Davies, R.B., 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74, 33–43.

- Fazzari, S.M., Glenn Hubbard, R., Petersen, B.C., 1988. Financing constraints and corporate investment. *Brookings Papers on Economic Activity*. pp. 141–195.
- Hall, B.H., Hall, R.E., 1993. The value and performance of U.S. corporations. *Brookings Papers on Economic Activity*. pp. 1–34.
- Hansen, B.E., 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64, 413–430.
- Hansen, B.E., 1999. Sample splitting and threshold estimation. *Econometrica*, forthcoming.
- Hu, X., Schiantarelli, F., 1998. Investment and capital market imperfections: A switching regression approach using U.S. firm panel data. *Review of Econometrics and Statistics* 80, 466–479.

The effects of match uncertainty and bargaining on labor market outcomes: evidence from firm and worker specific estimates

Subal C. Kumbhakar · Christopher F. Parmeter

Published online: 31 October 2008
© Springer Science+Business Media, LLC 2008

Abstract In this paper we examine wage dispersion in labor markets across currently employed workers. We argue that differences in the potential productivity of a match (typically assumed to be known in the previous literature) generates a surplus between the minimum wage the worker is willing to accept and the maximum wage the firm is willing to offer for the job. Existence of this surplus leads to wage dispersion due to negotiating over the amounts extracted by each agent. Our objective is to estimate the surplus extracted by each firm-worker pair and the effect of the net extracted surplus on the wage, for each firm-worker pair using the two-tier stochastic frontier model. An empirical application finds that, on average, firms paid workers less than their expected productivity. More specifically, at the mean, the net effect of productivity uncertainty leads to equilibrium wages which are 3.33% below the expected productivity of matches.

Keywords Expected productivity · Random matching · Two-tier frontier

JEL Classifications C2 · J3 · J15 · J41

The matching process that brings together workers and employers fails to weed out all bad matches
(Pries 2004, p. 194).

1 Introduction

Labor markets are designed to pair workers and firms efficiently to ensure a productive outcome. However, there are many obstacles that stand in the way of a perfectly efficient market sorting process. A major impediment to an efficient labor market is heterogeneity across both the worker and the firm concerning the productivity of the job which has implications for the wages they are willing to accept/pay. The effects of an inefficient market sorting process can be seen on the wage outcomes of worker-firm pairings. Some pairs will be characterized by identical workers with differing wages, while other pairs will be composed of identical firms, paying different wages for the same job. Thus, the impact of matching on labor market outcomes is key in understanding why observed wages are dispersed. The ability to pull out the effect that match quality and bargaining have on wage outcomes due to agent-specific heterogeneity is the central focus of this paper.

We argue that labor markets do not work perfectly. What this tells us is that all job formations are not ideal; ideal in the sense that the value of the job is the same to both worker and firm. We use a standard search/matching/bargaining model that accounts for wage dispersion generated by bargaining over an expected surplus given that the quality of any job match is only known imperfectly to each agent. While a litany of theoretical models have attempted to discern the nature and causes of wage dispersion, to our knowledge there does not exist a

S. C. Kumbhakar (✉)
Department of Economics, Binghamton University, Binghamton,
NY 13902-6000, USA
e-mail: kkar@binghamton.edu

C. F. Parmeter
Department of Agricultural and Applied Economics,
Virginia Polytechnic Institute and State University, Blacksburg,
VA 24061-0401, USA
e-mail: parms@vt.edu

microeconomic model that is designed to pull out the effects of this manifest match uncertainty.

Our attempt to capture wage dispersion generated by unknown match quality comes directly from the standard framework of search/matching/bargaining models that are commonly accepted as a genuine precipitator of wage dispersion. However, these models generate an equilibrium wage distribution for the labor market as a whole, while we focus on wage dispersion once person specific (and firm specific) effects have been accounted for. We modify the intuition of these models slightly to account for observed individual heterogeneity and provide a new twist to the exact nature of wage dispersion. While intimately linked to bargaining, our idea goes beyond simple negotiation of wages to a more important issue, viz., one of worker productivity.

Many authors have assumed that once a match is made the productivity of the job at hand is perfectly revealed (e.g., Flinn and Heckman 1982b; Hosios 1990; Mortensen and Pissarides 1994). Realistically the case is, however, more likely that there is an *expected* productivity of the match, known to both parties based on observables, but the *actual* productive value of the match is unknown.¹ Here we consider a static setting and are only concerned with the effect that this unknown productive value plays on current wage formation and dispersion. Intuitively, firms must caution against initially overpaying workers whom they will lock into long term contracts and cannot terminate simply because they overestimated their productivity, i.e., not all matches are ‘good’. This leads to firms attempting to pay workers less than their perceived productive value until the true value of the worker is revealed. Alternatively, workers attempt to receive a wage higher than their productive value as a means to guard against possible failure within the job and having to resort to a lower paying job in the future.

Unlike the standard models that formulate the wage being paid as the sum of the reservation wage and the part of the surplus extracted by the worker, it is more meaningful to consider the fact that the wage paid is related to the expected productivity of the match with fluctuations around the mean. This view is motivated by the insights of match uncertainty dating back to the work of Johnson (1978), Jovanovic (1979a, b), Viscusi (1979, 1980a, b, c, 1983), Wilde (1980), and Flinn (1986).² While the recent resurgence has been in the spirit of understanding macroeconomic unemployment fluctuations through dynamic

modelling procedures, our focus is more on the static microeconomic implications of productivity uncertainty. Specifically, our model is an attempt to uncover the effects of match quality on wage dispersion across worker and firm types.³

The objective of this paper is twofold: (i) we provide an intuitive explanation about the creation of productive surplus in the labor market in a reduced form setting which provides the impetus for the use of the two-tier estimation procedure, and (ii) we show how to obtain firm- and worker-specific extracted surpluses. Both points differ from the seminal two-tier papers of Polachek and Yoon (1987, 1996). First, Polachek and Yoon (1987) formulated the two-tier model to capture ignorance while searching for a job, we take this further to incorporate bargaining between worker and firm. Second, Polachek and Yoon (1987, 1996) only estimated the expected impacts of ignorance on behalf of the firm and the worker, we use the intuition of Jondrow et al. (1982) to derive observation specific expected values of the impact of bargaining.

The remainder of the paper is organized as follows: Sect. 2 provides an explanation for the existence of productivity uncertainty stemming from match inefficiency that has arisen in the labor economics literature. It continues with an investigation of a traditional textbook bargaining model and tries to imbed productivity uncertainty within it, generating a two-tier frontier model. Section 3 briefly reviews the structure and intuition of the two-tier frontier estimation procedure, showcasing the appeal of this estimation technique for the problem at hand. Section 4 presents the derivations of the conditional distributions that allow us to estimate agent-specific extracted surplus. Section 5 presents an empirical application that demonstrates the use of our decomposition, while our final thoughts and further avenues for extensions are contained in Sect. 6.

2 An overview of wage dispersion and match productivity

2.1 Wage dispersion—a short history

Given the extent of the labor market search literature, both theoretical and empirical, it is neither necessary nor desirable to discuss here all the existing avenues that have attempted to explain wage dispersion. A short list includes surveys by Rothschild (1973) (for markets in general), Lippman and McCall (1976) (for labor markets explicitly) and recently Eckstein and Van den Berg (forthcoming).

¹ It might be revealed at a later date or not revealed at all.

² Recently ideas paralleling these works can be found in Pries (2004) and Nagypál (2004). A different, but relevant idea on match uncertainty, miss-matching, can be found in Marimon and Zilibotti (1999).

³ Given that we have a supply side dataset we leave the effects of firm type on uncertainty for future research.

The penultimate paper that was the motivation behind endogenous wage dispersion was Diamond (1971) with major theoretical advances put forth in Butters (1977), Burdett and Judd (1983), Albrecht and Axell (1984), Burdett and Vishwanath (1988), Mortensen and Pissarides (1994), and Burdett and Mortensen (1998). Excellent empirical inspections are found in Eckstein and Wolpin (1990), Bowlus et al. (1995), Abowd et al. (1998), and Dey and Flinn (2005); as well as contributions with both a theoretical and empirical flavor, Flinn and Heckman (1982a, b), Bontemps et al. (1999a, b), Postel-Vinay and Robin (2002, 2003), and Flinn (2006).⁴

One of the main insights of equilibrium wage dispersion is that there is not one key element that generates it, whether it be differences in productivity across firms, allowing on-the-job search, heterogeneity in reservation wages, search frictions, or some as of yet undiscovered reason. Even today models capable of resolving the ‘Diamond Paradox’ are still being developed, see Gaumont et al. (2006) and Shapiro (2006). From an empirical standpoint there have been many studies that attempt to quantify the magnitude of dispersion, the effects of minimum wages on wage dispersion, structural estimation of the search theoretic models listed previously, as well as many applications of standard regression techniques to uncover the sources of equilibrium wage dispersion.

2.2 Uncertainty of match productivity

After Diamond (1971) laid out a theoretical model of price adjustment that did not generate equilibrium price dispersion, labor economists and search theorists alike tried to come up with strategies that allowed for the existence of wage dispersion in a cross section of homogeneous workers. It was Rothschild (1973) who noted that any model of search that cannot generate an endogenous non-degenerate equilibrium wage distribution was unsatisfactory.

Here we argue that when the match value is imperfectly known, perhaps a better way to model the wage setting mechanism is through the expected value of the match with fluctuations around the mean predicated upon workers and firms attempting to shield themselves from “bad” matches, i.e., the larger the surplus the more likely a match is bad for one party or another. Note that from the perspective of wages, matches can be considered bad for one party or the other, but not both. By relaxing the assumption of perfect knowledge of the match value we can, in a reduced form

setting, attribute equilibrium wage dispersion to negotiation over the initial, unknown “true” value of the match and determine which types of workers and/or firms are benefitting from this uncertainty.

Firms may try to mitigate the occurrence of ‘bad’ matches by posting specific skill requirements *a priori* (Acemoglu 1999; Mortensen and Pissarides 1994). However, this will not eliminate all lesser skilled workers from applying. Alternatively, workers may try to only apply for jobs in a certain region or of a specific type, in the hope of aligning themselves in such a way to reduce the probability of engaging in untenable employment, but this does not guarantee that a perfect match is conceived. Thus, while both workers and firms can attempt to insulate themselves there is no reason to expect the matching process to work perfectly and consequently existing market inefficiencies can result in inappropriate firm-worker pairings.

Our motivation of match uncertainty is through the plausible assumption of firm and worker heterogeneity. We assume that for a given set of characteristics for the worker, there is a distribution of productive ability. Similarly, for a given set of firm characteristics, there is a distribution of productive outcomes. Both of these distributions can be seen as proxies for limit wages. More productive firms will have the ability to pay more while more productive workers will place a higher value on their time, i.e., have higher reservation wages. Additionally, if we assume that the matching process is not perfect (Pries 2004), then we have the foundation for productivity uncertainty.

To explain this further, assume that r represents the distribution of reservation wages associated with worker heterogeneity for a given set of characteristics and p distinguishes the distribution of maximum wage offers for a given set of characteristics based on firm heterogeneity. A match function takes a draw from each distribution and one of two outcomes occurs: either $r > p$ and no match is consummated, or $p \geq r$ creating a surplus which the agents negotiate over. By repeating this process for any combination of worker and firm attributes one can obtain upper and lower bounds on the potential wage outcome given the observed characteristics of the agents. The difference in the upper and lower bound of the potential wage is necessary for surplus generation.

A key difference with our definition of wage dispersion compared to the previous literature is that we have wage dispersion for any combination of characteristics, while many of the previous empirical papers arbitrarily divide the labor market into distinct segments or search for a seemingly homogeneous sample to work with. In other words we are looking at vertical wage dispersion (based on specific agent characteristics) versus horizontal wage dispersion (based on the assumption that everyone is identical).

⁴ For a current synopsis of the state of the literature see the special issue of the *European Economic Review* (2006) in honor of Dale Mortensen.

2.3 Capturing productivity uncertainty in a reduced form setting

One particular model⁵ of interest that exemplifies the analysis of the impact of surplus extraction on wage variation is a matching/bargaining model with a distribution of productivity across job matchings. This distribution of productivities implies that there is also a distribution of surpluses that arises from these, as of yet, unknown productivities. Which agent (worker or firm) extracts more of the surplus has been shown to depend upon their bargaining power and information (see Osbourne and Rubinstein 1990, Chap. 5). Following Pissarides (2000, Chap. 1) the optimal wage rate is,

$$wage = \underline{wage} + \eta(\overline{wage} - \underline{wage}) \quad (1)$$

where \underline{wage} represents the worker's reservation wage, \overline{wage} represents the firm's maximum wage offer ($\overline{wage} \geq \underline{wage}$), and η ($0 \leq \eta \leq 1$), is the bargaining power of workers.⁶ In (1), $\eta(\overline{wage} - \underline{wage})$ represents the share of the surplus created from the formation of the job match that the worker receives when the job is filled.⁷ The reservation wage of a worker is unobserved, and due to bargaining it is unlikely that the observed wage is equal to the reservation wage. From an econometric standpoint (1) is not operational because the reservation wage and the maximum wage offer are unobserved. Another shortcoming of (1) is that it only provides insight on the impact of bargaining from a worker's standpoint, and, given productivity uncertainty it does not help in unveiling what happens when negotiations take place over an unknown surplus. Thus, even with an estimate of $\eta(\overline{wage} - \underline{wage})$ nothing could be said about what is happening on either the firm's side of the market, or of the productive value of the match.

To make (1) operational we transform the model so that it not only captures the impact of worker's bargaining, but the impact of firm's bargaining as well. For this, we first denote the expected productivity of the match conditional on a vector of characteristics x by $\mu(x) = E(\theta | x)$, where θ is the actual, but unknown, productivity of the match. We condition on x as it is intuitive, both from a worker's as well as a firm's perspective that observable characteristics may influence productivity and are certainly used in hiring

decisions by firms. By construction, $\underline{wage} \leq \mu(x) \leq \overline{wage}$ for those matches where a job is consummated. Consequently, $(\overline{wage} - \mu(x))$ is the firm's expected surplus from the match and $(\mu(x) - \underline{wage})$ is the worker's expected surplus from the match. We then use these notions of surplus to rewrite (1) as

$$\begin{aligned} wage &= \mu(x) - \mu(x) + \underline{wage} \\ &\quad + \eta(\overline{wage} - \underline{wage} + \mu(x) - \mu(x)) \\ &= \mu(x) + (\underline{wage} - \mu(x)) + \eta(\overline{wage} - \mu(x)) \\ &\quad - \eta(\underline{wage} - \mu(x)) \\ &= \mu(x) + \eta(\overline{wage} - \mu(x)) - (1 - \eta)(\mu(x) - \underline{wage}). \end{aligned} \quad (2)$$

In this framework the worker can raise his/her wage by extracting a share of the firm's surplus, $\eta(\overline{wage} - \mu(x)) \geq 0$, while the firm can lower the wage paid by extracting a share of the worker's surplus, $(1 - \eta)(\mu(x) - \underline{wage}) \geq 0$. The size of the extracted surplus by the worker depends upon the bargaining power of the worker, η , and the firm's expected surplus, $(\overline{wage} - \mu(x))$. Similarly, the level of the surplus extracted by the firm depends upon the firm's bargaining power, $(1 - \eta)$, and the worker's expected surplus, $(\mu(x) - \underline{wage})$.⁸

A more intuitive way to understand Eq. 2 is to note that at the time of the match neither the worker nor the firm knows the productive value of the match. What each knows is the expected productivity of the match given observable characteristics. Thus firms have an incentive to offer lower wages to protect themselves against bad hiring policies,⁹ while workers have an incentive to negotiate for higher wages to avoid entering into inefficient contracts. Thus workers try to extract some (all) of the surplus the firm is obtaining by hiring the worker, while the firm is trying to extract some (all) of the surplus the worker is acquiring by accepting the job. So, heuristically, our model is drawing upon previous studies that also look at the productive value of the match but they treat it as known and/or arising from a distribution (see Flinn and Heckman (1982b), Acemoglu and Shimer (2000), Postel-Vinay and Robin (2002, 2003), (Flinn 2006) for more on bargaining and productivity of matches).

Simply put, the idea of unknown productivity is very similar to the market for new Ph.D.s. When universities make hiring decisions, the productive value of the match is not known until a later date (usually when the candidate goes up for tenure) so the negotiated salary depends upon the job candidates skills and the characteristics of the university

⁵ We thank an anonymous referee for suggesting the following framework to us.

⁶ Pissarides (2000) used p instead of \overline{wage} and rU instead of \underline{wage} . Also, we used η to represent relative bargaining power of workers, instead of β as in Pissarides. Furthermore, in our modeling framework η can be observation specific.

⁷ The actual wage also represents a weighted average of the maximum offer and the reservation wage.

⁸ Using these notions we can define the expected productivity, $\mu(x)$ formally as the conditional expectation of $wage$ given x when either there is no surplus to extract or surplus extracted by workers and firms are equal.

⁹ See Shapiro (2006) for a similar idea along these lines.

hiring the candidate, both of which influence the productivity, but are not a perfect indicator of it. Given that the wage contract signed is for a predetermined number of years, it is in the interest of the candidate to get as high an amount as possible while the university should offer a lower wage not knowing if the candidate will be a good researcher. Thus, the university can extract the candidate's surplus until a later date when the contract is renegotiated based on a clearer picture of the productivity of the match. However, the candidate can extract surplus from the university in a similar manner to guard against being exploited until a later date when the contract is renegotiated. Our framework takes into account the influence of each party on the extent of wage dispersion around the expected productivity.

The wage equation in (2) has three distinct components. The first term, $\mu(x)$, represents the expected (productive value) wage of the worker given his/her characteristics x and is labelled as the benchmark wage (market value of the match). The second component shows the surplus extracted by the worker, while the third term (without the minus sign) is the surplus extracted by the firm. Since both workers and firms bargain and the effect of workers bargaining (surplus extraction) is to increase wages while the opposite happens due to firms bargaining (surplus extraction), what is relevant from the practical point of view is the net surplus, $NS = \eta(\overline{wage} - \mu(x)) - (1 - \eta)(\mu(x) - \underline{wage})$, which indicates the overall effect of bargaining on wage. Thus the observed wage can be more or less than the benchmark wage, $\mu(x)$ depending upon the sign of NS . Individually, neither of these components has a meaningful interpretation unless the other component is zero.

One model that may be seen as an overarching model for those described above is that of Postel-Vinay and Robin (2002, 2003). They setup a model that allows the equilibrium wage to be a random variable composed of a worker effect, a firm effect, and a market friction effect. Here we could say that the worker effect is positive, the firm effect is negative, and the market friction effect is two-sided. Thus, while our idea of a three component effect falls in line with their research, our model does not fit in with the structural nature of Postel-Vinay and Robin. What is interesting though is the fact that we can control for wage dispersion due to differences in worker characteristics while at the same time allowing for wage dispersion propagated by the three random factors. In their model only the three components lead to wage dispersion as there was no human capital accumulation thus accounting for age, education, tenure, and experience were not relevant to their discussion.¹⁰

¹⁰ Although they did mention that the next logical step in their modelling framework would be to introduce human capital accumulation.

One may even go as far as describing our method as a reduced form cousin to the modelling framework proposed by Flinn (1986), except that we explicitly allow productivity to depend on observable characteristics with fluctuations propagated by negotiations over the unknown surplus that exists due to the inherent uncertainty of the job outcome. Many of the same issues with identification that arose in his paper apply here as well. We have a reduced form equation that cannot nonparametrically identify the dispersion parameters of interest without distributional assumptions, a common theme in stochastic frontier models as well. And, while our model is incapable of providing estimates of the structural parameters associated with the Pissarides model from which it is derived, it does encapsulate information about an important aspect of the wage formation process, viz., uncertainty. This model is also one of the first to investigate the affect of productivity uncertainty in a microeconomic setting on equilibrium wage dispersion.¹¹ Are these dispersion effects large or small, do they represent a large share of the variation in wages or a small share relative to unobserved worker heterogeneity? While these issues have been around for quite some time we feel that our analysis is important in that it provides estimates of the impact of match quality and uncertainty on equilibrium wage dispersion.

Given that we are working in a reduced form setting a pertinent question becomes "Is our framework consistent with optimizing behavior of economic agents?" If we treat match productivity as uncertain and predicate wages upon productivity, making them uncertain as well, then both workers and firms attempt to optimize some expected criterion, rather than a deterministic one. Thus workers and firms bargain over the unknown surplus that is created from the match. Previous studies have used alternative rules, such as treating the surplus as known and splitting it in half (Flinn and Heckman 1982b), treating the surplus as known but bargaining over the relative amounts (Burdett and Mortensen 1998; Dey and Flinn 2005; and Flinn 2006), or treating surplus as known, but using other firms to enter in a Bertrand type game that will extract most or all of the surplus after the job has been accepted (Postel-Vinay and Robin 2002, 2003). Our approach is consistent with the last two of the three cases prevalent in the literature today, except that the surplus is *unknown*. To our knowledge no other study has treated the match surplus as unknown.

At this point it is worth mentioning that match uncertainty and negotiating power is dependent upon market structure. Some labor markets work better than others to ensure highly productive matches, for example the market for nuclear physicists is expected to work better than the

¹¹ See Shi (2006) for a recent theoretical insight into the effects of productivity on wages.

market for high school teachers. There are fewer candidates for nuclear physicists and the costs of obtaining the training to become one is more tedious and costly than that of obtaining a degree that allows one to teach in a high school. Also, because teachers have summers off, there are more candidates interested in being a teacher than a physicist. Thus, an extension of the model developed here, would be to test whether match certainty and surplus extraction differs based on metrics of market structure. One interesting example would be the labor markets of professional sports.¹²

3 A two-tier stochastic frontier model

3.1 Linking the bargaining model to a two-tier stochastic frontier model

The salient feature of the model in the preceding section is that the outcome variable has a lower and an upper bound. Polachek and Yoon (1987, 1996) (PY hereafter) used this notion and developed the two-tier frontier model in which these bounds are taken into account when estimating the model. In this section we put the labor market match quality example, discussed in the preceding section, in general terms and write the regression equation for the i th observation ($i = 1, \dots, n$) in the format of a two-tier stochastic frontier model, viz.,

$$y_i = x_i' \delta + \varepsilon_i, \quad (3)$$

where y_i is the outcome variable, x_i is a vector of covariates, δ is the corresponding parameter vector, and $\varepsilon_i = v_i - u_i + w_i$ represents the composite error term encapsulating the difference between the observed outcome variable and $\mu(x) = x' \delta$.¹³ In a buyer and seller framework $\mu(x)$ is the market value of the good. The lower-boundary (frontier) of price (y) is the minimum that the seller is willing to accept and is given by $\mu(x) - u$, $u \geq 0$. Similarly, the upper boundary (frontier) indicates the maximum that the buyer is willing to pay and is given by $\mu(x) + w$, $w \geq 0$. Because of the natural upper and lower boundaries of the outcome variable the frontier terminology is aptly used by PY. Furthermore, the frontiers are also likely to be affected by the presence of the noise term, v , that can take both positive and negative values and hence capture effects of random shocks.

To link the bargaining model in (2) to the two-tier model in (3) we rewrite the regression counterpart of (2) as

$$wage = \mu(x) - u + w + v \quad (4)$$

where $u = (1 - \eta)(\mu(x) - \overline{wage}) \geq 0$, $w = \eta(\overline{wage} - \mu(x)) \geq 0$, and v is the classical error term. As mentioned before the worker can raise his/her wage by extracting a share of the firm's surplus, denoted by w . Similarly, the firm can lower the wage paid by extracting a share of the worker's surplus, denoted by u . The size of these extracted surpluses depends on the bargaining power parameter, η , the firm's expected surplus, $(\overline{wage} - \mu(x))$, and the worker's expected surplus, $(\mu(x) - \overline{wage})$.

If one believes the argument put forth in Postel-Vinay and Robin (2002, 2003), then wages should represent the maximum possible productivity of a worker within a firm. On the other hand, if the scenario is closer to the textbook model of Pissarides (2000), then workers are paid their reservation wage plus something extra, representing the surplus extracted from the match. Both these models fit into the single-tier stochastic frontier models. So there are models which impose limits (at least implicitly) on how high or low wages can be. However, neither of these approaches is complete as each ignores one of the extremities relating to observed wages. Furthermore, stochastic frontier approaches which impose these limits in estimation are not explored in any of these models (see Kumbhakar and Lovell (2000) for a variety of such models).

A model closer to an actual frontier estimator is that of Flinn and Heckman (1982b). In their model they used the lowest wage in the sample (first order statistic) as an estimate of the (common) reservation wage for the population of interest. This paper was one of the first to consider how truncation at the reservation wage impacted the econometric analysis. While this paper is similar in spirit to our ideas, there are some dissimilarities, notably, the fact that we are focusing on each worker having a (possibly) different reservation wage, an upper bound on wages due to firms limiting their wage offers, and the introduction of worker/firm specific characteristics in each surplus term.

Thus the two-tier stochastic frontier technique seems an adequate avenue to go down when exploring bargaining within a matching framework. Although it is possible to estimate u and w (details are given in the following section), without further assumptions one cannot recover the relative bargaining parameter (η which can be worker-specific), worker's surplus ($\mu(x) - \overline{wage}$) and firm's surplus ($\overline{wage} - \mu(x)$) from the estimates of u and w . Thus our focus is not on the bargaining power *per se* but the surpluses extracted by the worker and firm for each observed

¹² We thank an anonymous referee for bringing this link with the model to our attention.

¹³ Although in (3) we are assuming $\mu(x) = x' \delta$ thereby making the assumption that $\mu(x)$ is linear in parameters, the linearity assumption is not necessary for the frontier model to work. One can, in principle, assume any functional form on $\mu(x)$.

match. In fact, estimates of these extracted surpluses are more useful than bargaining power because the end result of the bargaining process is to alter the wage in favor of a particular agent. In fact, the extent of productivity uncertainty on wages can be found from estimates on $w - u$.¹⁴ If this turns out to be positive then workers hold an advantage due to productivity uncertainty (by increasing their wages),¹⁵ while the opposite is true if this measure is negative.

3.2 Distributional assumptions and likelihood function

The δ parameters in (3) can be obtained using standard regression techniques. For example, the OLS procedure will give unbiased estimates of the slope coefficients. Since u and w are one-sided, $E(\varepsilon)$ may not be zero, even if $E(v) = 0$. Consequently, the OLS estimate of the intercept will be biased.¹⁶ Thus, if the objective is to estimate the δ parameters then the OLS estimator of the slope coefficients will be unbiased (unless one thinks along the lines of tenure and wage dispersion being correlated, in which case an instrument will be needed) and consistent. However, we are interested in not only estimating the δ parameters but also the surplus extraction components, i.e., to disentangle the one-sided error terms from the composed error term ε . For this reason, we estimate the model using the maximum likelihood (ML) method based on the following distributional assumptions of the error components, viz., u , v , and w . We assume that: (i) $v_i \sim i.i.d. N(0, \sigma_v^2)$, (ii) $u_i \sim i.i.d. Exp(\sigma_u, \sigma_u^2)$,¹⁷ (iii) $w_i \sim i.i.d. Exp(\sigma_w, \sigma_w^2)$, and (iv) the error components are distributed independently of each other and from the regressors, x . The use of an exponential distribution is commonplace in standard single-tier stochastic frontier studies when ML is used. It should be noted that the two-tier frontier distribution is nonparametrically underidentified. Thus these distributional assumptions are necessary to conduct an empirical analysis.

Based on the above distributional assumptions, it is straightforward (but tedious) to derive the probability density function (pdf) of ε_i , $f(\varepsilon_i)$ which is¹⁸

$$f(\varepsilon_i) = \frac{\exp\{\alpha_i\}}{\sigma_u + \sigma_w} \Phi(\beta_i) + \frac{\exp\{a_i\}}{\sigma_u + \sigma_w} \int_{-b_i}^{\infty} \phi(z) dz \\ = \frac{\exp\{\alpha_i\}}{\sigma_u + \sigma_w} \Phi(\beta_i) + \frac{\exp\{a_i\}}{\sigma_u + \sigma_w} \Phi(b_i) \quad (5)$$

where $a_i = \frac{\sigma_v^2}{2\sigma_w^2} - \frac{\varepsilon_i}{\sigma_w}$; $b_i = \frac{\varepsilon_i}{\sigma_v} - \frac{\sigma_v}{\sigma_w}$; $\alpha_i = \frac{\varepsilon_i}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2}$; $\beta_i = -(\frac{\varepsilon_i}{\sigma_v} + \frac{\sigma_v}{\sigma_u})$.

The log likelihood function for a sample of n observations is

$$\ln L(x; \theta) = -n \ln(\sigma_u + \sigma_w) + \sum_{i=1}^n \ln[e^{\alpha_i} \Phi(\beta_i) + e^{a_i} \Phi(b_i)] \quad (6)$$

where $\theta = \{\delta, \sigma_v, \sigma_u, \sigma_w\}$. The ML estimates of all the parameters can be obtained by maximizing the above log likelihood function. It should be noted that identification of all three standard deviations is achieved due to the fact that σ_u and σ_w appear in the likelihood equation separately, i.e., σ_u appears in α_i and β_i while σ_w appears in a_i and b_i .

The reason for the assumption of the exponential distributions for surpluses extracted by the firm and the worker is that the likelihood function can be expressed in a closed form and identification of the variance parameters is trivial. Moreover, while matching is a random process, we assume that markets work well enough that the generated surplus for any match is low. Thus, high values of extracted surplus, while probable in our model, occur with low probability. However, this does not preclude the use of other distributions for the one-sided error terms, such as log normal, gamma, half normal, truncated normal, etc.¹⁹

4 Measuring observation-specific extracted surplus

The main objective of estimating a two-tier stochastic frontier function is to obtain observation-specific estimates of extracted surplus by the worker and the firm, i.e., u_i and w_i from the composed error term ε_i , an estimate of which is obtained from the residuals of the wage equation, $y_i - x_i' \hat{\delta}$. In the standard single-tier frontier model, decomposition of the residual into inefficiency and noise components was accomplished by Jondrow et al. (1982). Here, we extend their technique to obtain observation-specific estimates of u

¹⁴ Recall that this is our measure of net surplus introduced prior.

¹⁵ Possibly from having relatively more bargaining power than firms.

¹⁶ Note that although $E(u)$ and $E(w)$ are non-zero, $E(w - u)$ might be zero. If this happens then the OLS estimator of the intercept will also be unbiased. This, however, does not mean that surplus does not exist in the market.

¹⁷ Here $Exp(\sigma_z, \sigma_z^2)$ denotes a random variable z that is exponentially distributed with mean σ_z and variance σ_z^2 .

¹⁸ The full derivations of all results are contained in the Appendix.

¹⁹ In fact, further research into the distributional assumptions of the two-tiered method, aside from making the technique more general, may also provide greater insight into wage variations once the error decomposition has taken place. See Tsionas (2008) for estimation of the two-tier model using Gamma distributions instead of exponentials. Also, the effect of distributional assumptions on the ranking of firms in efficiency studies has been found to have minor differences in the rankings of producers (see Kumbhakar and Lovell 2000, p. 90).

and w . For this, we need to derive the conditional distributions of u_i and w_i , viz., $f(u_i|\varepsilon_i)$ and $f(w_i|\varepsilon_i)$. These are

$$f(u_i|\varepsilon_i) = \frac{\lambda \exp\{-\lambda u_i\} \Phi(u_i/\sigma_v + b_i)}{\chi_{1i}} \quad (7)$$

and

$$f(w_i|\varepsilon_i) = \frac{\lambda \exp\{-\lambda w_i\} \Phi(w_i/\sigma_v + \beta_i)}{\chi_{2i}} \quad (8)$$

where $\lambda = \frac{1}{\sigma_u} + \frac{1}{\sigma_w}$, $\chi_{1i} = \Phi(b_i) + \exp\{\alpha_i - a_i\} \Phi(\beta_i)$, and $\chi_{2i} = \Phi(\beta_i) + \exp\{a_i - \alpha_i\} \Phi(b_i) = \exp\{a_i - \alpha_i\} \chi_{1i}$.

With these conditional distributions we derive the conditional expectation of u_i as

$$E(u_i|\varepsilon_i) = \frac{1}{\lambda} + \frac{\exp\{\alpha_i - a_i\} \sigma_v [\phi(-\beta_i) + \beta_i \Phi(\beta_i)]}{\chi_{1i}} \quad (9)$$

and the conditional expectation of w_i as

$$E(w_i|\varepsilon_i) = \frac{1}{\lambda} + \frac{\sigma_v [\phi(-b_i) + b_i \Phi(b_i)]}{\chi_{1i}} \quad (10)$$

which can be used to obtain observation-specific estimates of u_i and w_i , respectively.

Since the dependent variable in many regressions is in logarithmic form, one could interpret $E(u)$ and $E(w)$ —the point predictor of u and w —as the percentage reduction and increase in wage due to bargaining by the firm and worker, respectively, when u and w are small. To get an exact percentage measure of wage reduction due to a firm's ability to extract surplus, one could follow two alternative routes. First, use $100[e^z - 1]$, for $z = E(u|\varepsilon)$, $E(w|\varepsilon)$. However, $E(e^z) \neq e^{E(z)}$. Thus, one could use $E(\exp(-z))$ for $z = u$, w for computing the exact percentage decrease(increase) in wage due to firm's (worker's) bargaining.

To obtain the formula for computing observation-specific measures of $\exp(-u)$ and $\exp(-w)$, we need to derive the following conditional expectation, viz., $E(e^{-u_i}|\varepsilon_i)$ and $E(e^{-w_i}|\varepsilon_i)$, which are:

$$E(e^{-u_i}|\varepsilon_i) = \frac{\lambda}{1 + \lambda \chi_{2i}} \left[\Phi(b_i) + \exp\{\alpha_i - a_i\} \times \exp\{\sigma_v^2/2 - \sigma_v \beta_i\} \Phi(\beta_i - \sigma_v) \right] \quad (11)$$

and

$$E(e^{-w_i}|\varepsilon_i) = \frac{\lambda}{1 + \lambda \chi_{1i}} \left[\Phi(\beta_i) + \exp\{a_i - \alpha_i\} \times \exp\{\sigma_v^2/2 - \sigma_v b_i\} \Phi(b_i - \sigma_v) \right]. \quad (12)$$

These conditional expectations can be used as the point estimators of $\exp(-u)$ and $\exp(-w)$. The decomposition of ε into u and w suggests that the analyst does not need to make *a priori* assumptions about the bargaining power that a worker or firm has. The decomposition gives us a way to

assess the impact of bargaining on the overall wage, once the negotiations have taken place.

5 An empirical application

We operationalize the methods discussed in the preceding sections by estimating a wage function using the data from Blackburn and Neumark (1992). The outcome variable, following Blackburn and Neumark, is log wage and the x variables in the log wage regression are: education, work experience, tenure, squares of education, experience and tenure, age, a proxy variable for unmeasured ability (IQ), and dummy variables for working in an urban area, being married, and working in the south. To further control for unobserved, inherent correlates to wage variations, we also use the number of siblings the worker has, the birth order of the worker, and the years of mother's and father's education. Given that there are missing observations for mother's education, father's education, and the number of siblings, our dataset is reduced from 936 to 663 observations. Thus, we use a subsample of the original data used in Blackburn and Neumark (1992).

Table 1 presents results from the standard OLS wage regression that ignores the effect of bargaining on observed wages, except by incorporating a dummy variable for black workers. In this set-up the estimated coefficient of the black dummy suggests that, on average, black workers earn

Table 1 Estimates of log wage regression function (OLS)

Variable	Estimate	Variable	Estimate
Constant	3.429	IQ	0.004
	0.000		0.003
Education	3.039	Education ²	-1.220
	0.015		0.045
Experience	0.172	Experience ²	-0.022
	0.308		0.823
Tenure	0.161	Tenure ²	-0.068
	0.016		0.109
Age	0.508		
	0.011		
Married	0.198	South	-0.043
	0.000		0.169
Urban	0.199	Black	-0.109
	0.000		0.051
Number of siblings	0.009	Birth order	-0.017
	0.253		0.151
Mother's education	0.010	Father's education	0.005
	0.123		0.319

The natural logarithm of the monthly wage is used as the dependent variable in the regression and there are 663 observations. Asymptotic p values are reported beneath each estimate. The R^2 for this regression is 0.285

Table 2 Estimates of log wage regression (Two-tier frontier)

Variable	Estimate	Variable	Estimate
Constant	3.858	IQ	0.004
	0.000		0.000
Education	2.037	Education ²	−0.756
	0.071		0.171
Experience	0.282	Experience ²	−0.101
	0.071		0.276
Tenure	0.154	Tenure ²	−0.057
	0.016		0.136
Age	0.495		
	0.008		
Married	0.206	South	−0.035
	0.000		0.241
Urban	0.221	Black	−0.102
	0.000		0.052
Number of siblings	0.009	Birth order	−0.015
	0.223		0.147
Mother's education	0.008	Father's education	0.008
	0.147		0.131
σ_v	0.190	σ_u	0.221
	0.000		0.000
		σ_w	0.189
			0.000

The natural logarithm of the monthly wage is used as the dependent variable in the regression and there are 663 observations. Asymptotic p values are reported beneath each estimate

about 11% less than white workers, *ceteris paribus*. By construction this coefficient is group specific. Thus, if it represents the effects of bargaining, it is an average for the whole group of black workers. As a result, nothing can be said about the effect of bargaining on an individual worker's wage whether black or white. This drawback is eliminated in the two-tier frontier model where one can estimate the impact of bargaining on wage for each worker-firm pair. These observation-specific results can then be used, if desired, to examine whether a particular group (defined in any manner) has more (less) influence on wages, *ceteris paribus*, for the group as a whole.

Estimated parameters from the two-tier frontier function are presented in Table 2. We are not too concerned with the fact that our data does not come from a population of recently hired workers. The reason being that the impact of negotiations over wages can have effects throughout the tenure of a worker and as such we can still determine if wages are higher or lower that they should be due to negotiations at the time the job was offered/accepted.²⁰

²⁰ Calculations with a different data set, not reported here, suggest that there is an additional impact from being a new worker that lowers wages. The results are available upon request.

Also, our estimates are of surplus extracted due to unknown bounds on productivity on both sides of the match. As long as those bounds remain after workers accept a match, i.e. there is still some productive uncertainty after the worker has been at the firm t years after the initial match, then surplus extraction will take place and cause variations in wages which we can then attempt to estimate.

The deterministic part of the frontier model is the same as the OLS model. The parameter estimates from the OLS and frontier models are quite similar. This suggests that one can use either of these models if the objective is to determine the marginal effect of covariates. However, if the interest is to obtain the impact of bargaining on wages, it is necessary to use the two-tier frontier approach which provides deeper insights on the effect of bargaining on wages for each employee-employer pair.

From the estimates of σ_v , σ_u and σ_w , we find that the unexplained variation in log wage ($\sigma_v^2 + \sigma_u^2 + \sigma_w^2$) is 0.121. Of this unexplained variation, 70.4% is due to bargaining.²¹ From the estimate of $E(w - u) = \sigma_w - \sigma_u$, one can say whether, on average, bargaining affects wages or not, and if so, in what direction. On the other hand, if $\sigma_w - \sigma_u = 0$ then one would predict that, at the mean, wages are not affected by bargaining. This, however, does not mean absence of bargaining because a zero mean does not imply that the quartiles, for example, will be zero. To get an answer to this we need to resort to the worker-firm pair estimates from the two-tier frontier approach.

Details on surplus extraction results (the percentage change in wages), based on observation specific estimates of $E(u|e)$ and $E(w|e)$, are reported in Tables 3–5. These tables display percentage changes measured relative to the benchmark log wage estimated from $\log \text{wage} = x_i' \hat{\delta}$. Table 3 shows that at the mean, surplus extracted by firms decreased wages by 25.2%, while, surplus extracted by workers increased wages by 21.03%. These opposite effects led to a decrease in wages (estimated from $E((w - u)|e)$ by 3.33% relative to benchmark wages, *ceteris paribus*.²² The first quartile value of net surplus is −13.20% which means wages are at least 13.20% below the expected productive value of the match for 25% of the sample). The top (meaning a positive surplus extraction on behalf of the worker) 25% of the surplus extractions are at least 9.68% relative to the benchmark wage (meaning that wages for 25% of the sample, are increased by at least 9.68% above the expected productive value of the match). Thus, wages

²¹ In their 1987 (1996) paper, Polachek and Yoon found that 79.8% (98.5%) of the unexplained wage variation was due to incomplete information.

²² If the goal is to obtain an estimate of the mean of the net effect, one can use the estimated value of $E(w - u) = \sigma_w - \sigma_u$ which is −0.0339. This does not require use of the observation-specific estimates of w and u .

Table 3 Surplus extracted by firms and workers

	Mean (%)	Q1 (%)	Q2 (%)	Q3 (%)
Workers: $\hat{E}(w \varepsilon)$	18.9	11.4	14.4	21.2
Firms: $\hat{E}(u \varepsilon)$	22.1	12.2	16.4	25.0
Net surplus: $\hat{E}((w - u) \varepsilon)^a$	-3.2	-13.7	-2.0	9.0
Workers: $\hat{E}(1 - e^{-w} \varepsilon)$	15.9	10.3	12.8	18.1
Firms: $\hat{E}(1 - e^{-u} \varepsilon)$	18.1	10.9	14.4	21.1
Net surplus: $\hat{E}((e^{-u} - e^{-w}) \varepsilon)^a$	-2.2	-10.8	-1.6	7.2

Since the dependent variable is in logarithms we convert the estimates in the first panel in percentage form using $100[e^z - 1]$, where $z = \hat{E}(\cdot | \varepsilon)$. The second panel estimates are multiplied by 100 to express them in percentage form

^a The mean and quartiles of net surplus were constructed after calculating $\hat{E}((w - u) | \varepsilon)$ and $\hat{E}((e^{-u} - e^{-w}) | \varepsilon)$

Table 4 Surplus extracted by firms and workers across race

	Mean (%)	Q1 (%)	Q2 (%)	Q3 (%)
<i>White workers^a</i>				
Workers: $\hat{E}(w \varepsilon)$	18.9	11.4	14.3	21.2
Firms: $\hat{E}(u \varepsilon)$	22.1	12.2	16.5	24.9
Net surplus: $\hat{E}((w - u) \varepsilon)$	-3.2	-13.5	-2.2	9.0
<i>Black workers^b</i>				
Workers: $\hat{E}(w \varepsilon)$	18.4	11.1	15.0	20.6
Firms: $\hat{E}(u \varepsilon)$	21.3	12.3	15.6	27.4
Net surplus: $\hat{E}((w - u) \varepsilon)$	-2.9	-16.3	-0.6	8.3
<i>White workers^a</i>				
Workers: $\hat{E}(1 - e^{-w} \varepsilon)$	15.9	10.3	12.7	18.2
Firms: $\hat{E}(1 - e^{-u} \varepsilon)$	18.1	10.9	14.4	21.0
Net surplus: $\hat{E}((e^{-u} - e^{-w}) \varepsilon)$	-2.2	-10.7	-1.7	7.3
<i>Black workers^b</i>				
Workers: $\hat{E}(1 - e^{-w} \varepsilon)$	15.7	10.0	12.7	18.2
Firms: $\hat{E}(1 - e^{-u} \varepsilon)$	17.8	11.1	13.8	22.8
Net surplus: $\hat{E}((e^{-u} - e^{-w}) \varepsilon)$	-2.1	-12.8	-0.4	6.7

Since the dependent variable is in logarithms we convert the estimates in the first panel in percentage form using $100[e^z - 1]$, where $z = \hat{E}(\cdot | \varepsilon)$. The second panel estimates are multiplied by 100 to express them in percentage form

The mean and quartiles of net surplus were constructed after calculating $\hat{E}((w - u) | \varepsilon)$ and $\hat{E}((e^{-u} - e^{-w}) | \varepsilon)$

^a There are 609 observations for white workers

^b There are 54 observations for black workers

are not reduced for all workers, and in fact, some workers managed to negotiate for a substantial increase in their wages over their expected productive outcomes. The lower panel of Table 4 gives the same information. The only difference is in the calculation of the percentage figures. Note that the estimates are for each worker-firm pair. We provide a summary of these in Table 3.

In addition to determining the impact of bargaining on observed wages for all worker-firm pairs, one can analyze

Table 5 Surplus extracted by firms and workers across marital status

	Mean (%)	Q1 (%)	Q2 (%)	Q3 (%)
<i>Married workers^a</i>				
Workers: $\hat{E}(w \varepsilon)$	18.8	11.4	14.3	21.1
Firms: $\hat{E}(u \varepsilon)$	22.0	12.2	16.5	24.7
Net surplus: $\hat{E}((w - u) \varepsilon)$	-3.2	-13.3	-2.2	9.0
<i>Single workers^b</i>				
Workers: $\hat{E}(w \varepsilon)$	19.6	11.2	14.3	21.1
Firms: $\hat{E}(u \varepsilon)$	22.4	11.9	15.7	26.4
Net surplus: $\hat{E}((w - u) \varepsilon)$	-2.9	-15.2	-0.7	10.5
<i>Married workers^a</i>				
Workers: $\hat{E}(1 - e^{-w} \varepsilon)$	15.8	10.3	12.7	18.1
Firms: $\hat{E}(1 - e^{-u} \varepsilon)$	18.0	20.8	14.4	10.9
Net surplus: $\hat{E}((e^{-u} - e^{-w}) \varepsilon)$	-2.2	-10.5	-1.7	7.2
<i>Single workers^b</i>				
Workers: $\hat{E}(1 - e^{-w} \varepsilon)$	16.4	10.1	13.2	19.1
Firms: $\hat{E}(1 - e^{-u} \varepsilon)$	18.5	10.7	13.8	22.1
Net surplus: $\hat{E}((e^{-u} - e^{-w}) \varepsilon)$	-2.1	-12.0	-0.6	8.4

Since the dependent variable is in logarithms we convert the estimates in the first panel in percentage form using $100[e^z - 1]$, where $z = \hat{E}(\cdot | \varepsilon)$. The second panel estimates are multiplied by 100 to express them in percentage form

The mean and quartiles of net surplus were constructed after calculating $\hat{E}((w - u) | \varepsilon)$ and $\hat{E}((e^{-u} - e^{-w}) | \varepsilon)$

^a There are 597 observations for married workers

^b There are 66 observations for non-married workers

the impact of bargaining on wages across different groups. Here we focus on black versus white workers, and married versus single workers. These results are reported in Tables 4 and 5, respectively.

From the estimates of the percentage change in wage (net surplus extracted) it is clear that, on average, both whites and blacks are receiving a wage reduction, relative to the benchmark, after controlling for being black on expected productivity. Thus, at the mean, there is not a significant difference between the surplus extraction of black and white workers across both measures (-3.2% vs. -2.9% and -2.2% vs. -2.1%, respectively). What is clear however, is that looking at the tails of the surplus extraction distributions across measures, the lower quartile suggests that black workers have 2–3% more extracted from the benchmark, while at the upper quartile, white workers are able to extract about 1% more than black workers relative to the benchmark.

These results are not surprising. First, controlling for being black in the expected productivity regression, linked with the similar distributions of surplus extraction, suggests that while blacks may be paid lower due to expected productivity, additional surplus extraction, compared with whites is negligible. Second, assuming uncorrelatedness of model covariates with the error terms means that we should

not expect there to be significant differences between the surplus extraction distributions.

Switching to surplus extraction based on marital status, we see a similar picture. Both measures, across marital status, evaluated at the mean suggest there is no difference in surplus extraction based on marital status. What's more, the distributions of surplus extraction are quite similar as well. Again, this points to the fact that we are controlling for marital status in the expected productivity benchmark and the individual level surplus extraction measures are treated as uncorrelated with the covariates of the model. If one wanted to explicitly allow for the level of surplus extraction to depend on specific covariates, then the parameters of the two one-sided distributions could be modelled as such. Alternatively, if one believed that expected productivity did not depend on race or marital status, then a similar type of analysis would be better suited to reveal if surplus extraction depended on these worker features.

6 Conclusions

Firms and workers valuation of a job are inherently different, which leads room for negotiations over how much should be paid for the task at hand. A worker (firm) wants to extract as much of the surplus of the firm (worker) as possible. The surplus extracted by the firm reduces the wage while the surplus extracted by worker increases the wage. The net effect on the observed wage depends on the sum of these two opposing effects. In this paper we used a model that can identify workers' and firms' surplus extractions from the other party. The proposed technique allows us to estimate not only agent specific surplus extracted, but the net surplus extracted for each transaction as well. Once this net impact has been constructed comparisons across different strata of workers and/or firms may lead to a characterization of which qualities lead to better outcomes in a particular market.

We used the two-tier stochastic frontier technique to estimate the parameters of the model and to obtain observation-specific measures of extracted surplus by both the worker and the firm. This measure allows a secondary analysis of potential sources of bargaining power in the market. This secondary analysis could also be used to discern if particular groups of workers/firms are consistently being exploited in the market, in terms of extracting a smaller share of the extant surplus created from the match.

We provided an empirical application to examine the effect of worker and firm bargaining on wages, after controlling for worker characteristics. We found that, not only does a significant surplus exist, but the impact of

bargaining over this surplus has an asymmetric effect on wages at the mean. Indeed, at the mean, the net effect of surplus extraction (bargaining) by workers and firms decreased wages by 3.33% from the benchmark/market wage, while at the median, wages were reduced by 2.06% relative to the benchmark wage. Our ability to measure the effect of bargaining on wages for each worker-firm pair allowed us to correlate wage fluctuations to a worker's race. In our application, we found that, across the distribution of surplus extraction, *ceteris paribus*, there are no significant differences between either white and black workers and married/unmarried workers.

Although in this paper we discuss a labor market model that captures the idea of bargaining over the surplus generated due to worker and firm heterogeneity and match inefficiency, we believe that the modeling strategy is general enough to include many other markets. Some other examples where this technique may be of use are auctions, used car markets, and hedonic price models such as the residential housing market. Thus, although the two-tier frontier technique was originally conceived as a method to learn about the impact of incomplete information in the labor market, we trust the applicability of the model goes beyond its original intention.

Acknowledgements The authors thank two anonymous referees, Suqin Ge, Christopher Hanes, Daniel Henderson, Nicolai Kuminoff, Xiang Lie, Knox Lovell, and Solomon Polachek. Comments and suggestions from seminar participants at Syracuse University and SMU as well as Sandra Ahearn's help in proofreading the Appendix are gratefully acknowledged. The usual disclaimer applies.

Appendix

Derivations of selected equations

Derivation²³ of Eq. 5:

Beginning with the definition of the composed error term $\varepsilon_1 = v - u$, the marginal distribution of this is, following Kumbhakar and Lovell (2000),

$$f(\varepsilon_1) = (1/\sigma_u) \left(\Phi(-\varepsilon_1/\sigma_v - \sigma_v/\sigma_u) \exp\{\varepsilon_1/\sigma_u + \sigma_v^2/2\sigma_u^2\} \right). \quad (\text{A.1})$$

The three component error may then be written as $\varepsilon = \varepsilon_1 + w$, which implies that $\varepsilon_1 = \varepsilon - w$, yielding the following joint distribution, $g(\varepsilon, w) = g(\varepsilon_1, w) \cdot |d\varepsilon_1/d\varepsilon| = g(\varepsilon_1, w) = f(\varepsilon_1) \cdot f(w)$. Upon integrating out w one obtains the marginal distribution of ε . This is done below.

²³ To avoid notational clutter we dropped the i subscript in all the derivations.

$$\begin{aligned}
f(\varepsilon) &= \int_0^\infty \frac{1}{\sigma_u} \left(\Phi \left(-\frac{\varepsilon_1}{\sigma_v} - \frac{\sigma_v}{\sigma_u} \right) \exp \left\{ \frac{\varepsilon_1}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} \right\} \right) \\
&\quad \times \frac{1}{\sigma_w} \exp \left\{ -\frac{w}{\sigma_w} \right\} dw \\
&= \frac{1}{\sigma_u \sigma_w} \left[\exp \left\{ \frac{\varepsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} \right\} \int_0^\infty \Phi \left(\frac{w}{\sigma_v} - \left(\frac{\varepsilon}{\sigma_v} + \frac{\sigma_v}{\sigma_u} \right) \right) \right. \\
&\quad \left. \times \exp \left\{ -w \left(\frac{1}{\sigma_w} + \frac{1}{\sigma_u} \right) \right\} dw \right] \\
&= \frac{-1}{\sigma_u + \sigma_w} \left[\exp \left\{ \frac{\varepsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} \right\} \int_0^\infty \Phi \left(\frac{w}{\sigma_v} - \left(\frac{\varepsilon}{\sigma_v} + \frac{\sigma_v}{\sigma_u} \right) \right) \right. \\
&\quad \left. \times d \left(\exp \left\{ -w \left(\frac{1}{\sigma_w} + \frac{1}{\sigma_u} \right) \right\} \right) \right] \\
&= \frac{-\exp\{\alpha\}}{\sigma_u + \sigma_w} \left[\Phi(w/\sigma_v + \beta) \exp\{-w\lambda\} \Big|_0^\infty \right. \\
&\quad \left. - \int_0^\infty \phi(w/\sigma_v + \beta) \exp\{-w\lambda\} dw \right] \\
&= \frac{\exp\{\alpha\}}{\sigma_u + \sigma_w} \left[\Phi(\beta) + \exp\{-\alpha\} \exp \left\{ \frac{\sigma_v^2}{2\sigma_w^2} - \frac{\varepsilon}{\sigma_w} \right\} \right. \\
&\quad \left. \times \int_0^\infty \frac{1}{\sigma_v} \phi \left(\frac{w}{\sigma_v} - \left(\frac{\varepsilon}{\sigma_v} - \frac{\sigma_v}{\sigma_w} \right) \right) dw \right] \\
&= \frac{\exp\{\alpha\}}{\sigma_u + \sigma_w} \Phi(\beta) + \frac{\exp\{a\}}{\sigma_u + \sigma_w} \int_{-b}^\infty \phi(z) dz \\
&= \frac{\exp\{\alpha\}}{\sigma_u + \sigma_w} \Phi(\beta) + \frac{\exp\{a\}}{\sigma_u + \sigma_w} \Phi(b)
\end{aligned}$$

where $\alpha = \frac{\varepsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2}$; $\beta = -\left(\frac{\varepsilon}{\sigma_v} + \frac{\sigma_v}{\sigma_u}\right)$;
 $\lambda = \frac{1}{\sigma_u} + \frac{1}{\sigma_w}$; $a = \frac{\sigma_v^2}{2\sigma_w^2} - \frac{\varepsilon}{\sigma_w}$; $b = \frac{\varepsilon}{\sigma_v} - \frac{\sigma_v}{\sigma_w}$. (A.2)

Derivation of Eqs. 7 and 8:

$$\begin{aligned}
f(u|\varepsilon) &= \frac{f(u, \varepsilon)}{f(\varepsilon)} \\
&= \frac{(\exp\{a\}/\sigma_u \sigma_w) \exp\{-\lambda u\} \Phi(u/\sigma_v + b)}{(1/(\sigma_u + \sigma_w)) [\exp\{a\} \Phi(b) + \exp\{\alpha\} \Phi(\beta)]} \\
&= \frac{\lambda \exp\{a\} \exp\{-\lambda u\} \Phi(u/\sigma_v + b)}{[\exp\{a\} \Phi(b) + \exp\{\alpha\} \Phi(\beta)]} \\
&= \frac{\lambda \exp\{-\lambda u\} \Phi(u/\sigma_v + b)}{\chi_1}
\end{aligned} \tag{A.3}$$

where $\chi_1 = \Phi(b) + \exp\{a - \alpha\} \Phi(\beta)$. Similarly,

$$\begin{aligned}
f(w|\varepsilon) &= \frac{f(w, \varepsilon)}{f(\varepsilon)} \\
&= \frac{(\exp\{\alpha\}/\sigma_u \sigma_w) \exp\{-\lambda w\} \Phi(w/\sigma_v + \beta)}{(1/(\sigma_u + \sigma_w)) [\exp\{a\} \Phi(b) + \exp\{\alpha\} \Phi(\beta)]} \\
&= \frac{\lambda \exp\{\alpha\} \exp\{-\lambda w\} \Phi(w/\sigma_v + \beta)}{[\exp\{a\} \Phi(b) + \exp\{\alpha\} \Phi(\beta)]} \\
&= \frac{\lambda \exp\{-\lambda w\} \Phi(w/\sigma_v + \beta)}{\chi_2}
\end{aligned} \tag{A.4}$$

where $\chi_2 = \Phi(\beta) + \exp\{a - \alpha\} \Phi(b) = \exp\{a - \alpha\} \chi_1$.

Derivation of Eqs. 9 and 10:

$$\begin{aligned}
E(u|\varepsilon) &= \int_0^\infty u \frac{\lambda \exp\{-\lambda u\} \Phi(u/\sigma_v + b)}{\chi_1} du \\
&= \frac{-1}{\chi_1 \lambda} \left[\int_0^\infty \Phi(u/\sigma_v + b) d(\exp\{-\lambda u\}) \right. \\
&\quad \left. + \lambda \int_0^\infty \Phi(u/\sigma_v + b) d(u \exp\{-\lambda u\}) \right] \\
&= \frac{1}{\chi_1 \lambda} \left[\Phi(b) + \int_0^\infty \exp\{-\lambda u\} \phi(u/\sigma_v + b) du / \sigma_v \right. \\
&\quad \left. + \lambda \int_0^\infty u \exp\{-\lambda u\} \phi(u/\sigma_v + b) du / \sigma_v \right] \\
&= \frac{1}{\chi_1 \lambda} \left[\Phi(b) + \frac{\exp\{\alpha\}}{\exp\{a\}} \left[\int_0^\infty \phi(u/\sigma_v + b + \sigma_v \lambda) du / \sigma_v \right. \right. \\
&\quad \left. \left. + \lambda \int_0^\infty u \phi(u/\sigma_v + b + \sigma_v \lambda) du / \sigma_v \right] \right] \\
&= \frac{1}{\chi_1 \lambda} \left[\Phi(b) + \frac{\exp\{\alpha\}}{\exp\{a\}} \left[\int_{-\beta}^\infty \phi(z) dz + \sigma_v \lambda \right. \right. \\
&\quad \left. \left. \times \int_{-\beta}^\infty z \phi(z) dz + \lambda \sigma_v \beta \int_{-\beta}^\infty \phi(z) dz \right] \right] \\
&= \frac{1}{\chi_1 \lambda} [\Phi(b) + \exp\{\alpha - a\} \Phi(b) \\
&\quad + \sigma_v \lambda \exp\{\alpha - a\} [\phi(-\beta) + \beta \Phi(\beta)]] \\
&= \frac{1}{\lambda} + \frac{\sigma_v [\phi(-\beta) + \beta \Phi(\beta)]}{\chi_2}.
\end{aligned} \tag{A.5}$$

The derivation for $E(w|\varepsilon)$ follows similarly as:

$$\begin{aligned}
 E(w|\varepsilon) &= \int_0^\infty w \frac{\lambda \exp\{-\lambda w\} \Phi(w/\sigma_v + \beta)}{\chi_2} dw \\
 &= \frac{-1}{\chi_2 \lambda} \left[\int_0^\infty \Phi(w/\sigma_v + \beta) d(\exp\{-\lambda w\}) \right. \\
 &\quad \left. + \lambda \int_0^\infty \Phi(w/\sigma_v + \beta) d(w \exp\{-\lambda w\}) \right] \\
 &= \frac{1}{\chi_2 \lambda} \left[\Phi(\beta) + \int_0^\infty \exp\{-\lambda w\} \phi(w/\sigma_v + \beta) dw/\sigma_v \right. \\
 &\quad \left. + \lambda \int_0^\infty w \exp\{-\lambda w\} \phi(w/\sigma_v + \beta) dw/\sigma_v \right] \\
 &= \frac{1}{\chi_2 \lambda} \left[\Phi(b) + \exp\{a - \alpha\} \right. \\
 &\quad \times \left[\int_0^\infty \phi(w/\sigma_v + \beta + \sigma_v \lambda) dw/\sigma_v \right. \\
 &\quad \left. + \lambda \int_0^\infty w \phi(w/\sigma_v + \beta + \sigma_v \lambda) dw/\sigma_v \right] \Bigg] \\
 &= \frac{1}{\chi_2 \lambda} \left[\Phi(b) + \exp\{a - \alpha\} \left[\int_{-b}^\infty \phi(z) dz \right. \right. \\
 &\quad \left. \left. + \lambda \sigma_v \int_{-b}^\infty z \phi(z) dz + \lambda \sigma_v b \int_{-b}^\infty \phi(z) dz \right] \right] \\
 &= \frac{1}{\chi_2 \lambda} [\Phi(\beta) + \exp\{a - \alpha\} \Phi(b) + \sigma_v \lambda \exp\{a - \alpha\} \\
 &\quad \times [\phi(-b) + b \Phi(b)]] . \tag{A.6}
 \end{aligned}$$

Thus,

$$E(w|\varepsilon) = \frac{1}{\lambda} + \frac{\sigma_v [\phi(-b) + b \Phi(b)]}{\chi_1} . \tag{A.7}$$

Derivation of Eqs. 11 and 12:

$$\begin{aligned}
 E(e^{-u}|\varepsilon) &= \int_0^\infty e^{-u} \frac{\lambda e^{-\lambda u} \Phi(u/\sigma_v + b)}{\chi_1} du \\
 &= \frac{\lambda}{\chi_1} \int_0^\infty e^{-(1+\lambda)u} \Phi(u/\sigma_v + b) du \\
 &= \left(\frac{-\lambda}{\chi_1(1+\lambda)} \right) \int_0^\infty \Phi(u/\sigma_v + b) d(e^{-(1+\lambda)u}) . \tag{A.8}
 \end{aligned}$$

Using integration by parts, we get

$$\begin{aligned}
 E(e^{-u}|\varepsilon) &= \left(\frac{-\lambda}{\chi_1(1+\lambda)} \right) \left[\Phi(u/\sigma_v + b) e^{-(1+\lambda)u} \Big|_0^\infty \right. \\
 &\quad \left. - \int_0^\infty e^{-(1+\lambda)u} \phi(u/\sigma_v + b) du/\sigma_v \right] \\
 &= \left(\frac{\lambda}{\chi_1(1+\lambda)} \right) \left[\Phi(b) + e^{a-\alpha+5\sigma_v^2-\sigma_v\beta} \right. \\
 &\quad \times \left. \int_0^\infty \phi(u/\sigma_v + (b + \sigma_v(1+\lambda))) du/\sigma_v \right] , \tag{A.9}
 \end{aligned}$$

and using the change of variable, $z = \frac{u}{\sigma_v} + (b + \sigma_v(1+\lambda)) \Rightarrow dz = du/\sigma_v$, we have

$$E(e^{-u}|\varepsilon) = \left(\frac{\lambda}{\chi_1(1+\lambda)} \right) \left[\Phi(b) + e^{a-\alpha+5\sigma_v^2-\sigma_v\beta} \Phi(\beta - \sigma_v) \right] . \tag{A.10}$$

For the derivation of Eq. 12 we follow the same procedure as follows:

$$\begin{aligned}
 E(e^{-w}|\varepsilon) &= \int_0^\infty e^{-w} \frac{\lambda e^{-\lambda w} \Phi(w/\sigma_v + \beta)}{\chi_2} dw \\
 &= (\lambda/\chi_2) \int_0^\infty e^{-(1+\lambda)w} \Phi(w/\sigma_v + \beta) dw \\
 &= \left(\frac{-\lambda}{\chi_2(1+\lambda)} \right) \int_0^\infty \Phi(w/\sigma_v + \beta) d(e^{-(1+\lambda)w}) . \tag{A.11}
 \end{aligned}$$

Using integration by parts

$$\begin{aligned}
 E(e^{-w}|\varepsilon) &= \left(\frac{-\lambda}{\chi_2(1+\lambda)} \right) \left[\Phi(w/\sigma_v + \beta) e^{-(1+\lambda)w} \Big|_0^\infty \right. \\
 &\quad \left. - \int_0^\infty e^{-(1+\lambda)w} \phi(w/\sigma_v + \beta) dw/\sigma_v \right] \\
 &= \left(\frac{\lambda}{\chi_2(1+\lambda)} \right) \left[\Phi(\beta) + e^{a-\alpha-b\sigma_v+5\sigma_v^2} \right. \\
 &\quad \times \left. \int_0^\infty \phi\left(\frac{w}{\sigma_v} + (\beta + \sigma_v(1+\lambda))\right) dw/\sigma_v \right] . \tag{A.12}
 \end{aligned}$$

Finally, using the change of variable, $z = \frac{w}{\sigma_v} + (\beta + \sigma_v(1+\lambda)) \Rightarrow dz = dw/\sigma_v$, we have

$$E(e^{-w}|\varepsilon) = \left(\frac{\lambda}{\chi_2(1+\lambda)} \right) \left[\Phi(\beta) + e^{a-\alpha-b\sigma_v+5\sigma_v^2} \Phi(b - \sigma_v) \right] . \tag{A.13}$$

References

- Abowd J, Kramarz F, Margolis D (1998) High wage workers and high wage firms. *Econometrica* 67:251–333
- Acemoglu D (1999) Changes in unemployment and wage inequality: an alternative theory and some evidence. *Am Econ Rev* 89:1259–1278
- Acemoglu D, Shimer R (2000) Wage and technology dispersion. *Rev Econ Stud* 67:585–607
- Albrecht J, Axell B (1984) An equilibrium model of search employment. *J Polit Econ* 92:824–840
- Blackburn M, Neumark D (1992) Unobserved ability, efficiency wages, and interindustry wage differentials. *Q J Econ* 107:1421–1436
- Bontemps C, Robin J-M, Van den Berg GJ (1999a) An empirical equilibrium job search model with search on the job and heterogeneous workers and firms. *Int Econ Rev* 40:1039–1075
- Bontemps C, Robin J-M, Van den Berg GJ (1999b) Equilibrium search with continuous productivity dispersion: theory and nonparametric estimation. *Int Econ Rev* 41:305–358
- Bowlus A, Keifer N, Neumann G (1995) Estimation of equilibrium wage distributions with heterogeneity. *J Appl Econ* 10:S119–S131
- Burdett K, Judd L (1983) Equilibrium price distributions. *Econometrica* 51:955–970
- Burdett K, Mortensen D (1998) Wage differentials, employer size, and unemployment. *Int Econ Rev* 39:257–273
- Burdett K, Vishwanath T (1988) Balanced matching and labor market equilibrium. *J Polit Econ* 96:1048–1065
- Burdett K, Wright R (1998) Two-sided search and nontransferable utility. *Rev Econ Dyn* 1:220–245
- Butters G (1977) Equilibrium distributions of sales and advertising prices. *Rev Econ Stud* 44:465–491
- Cole H, Rogerson R (1999) Can the Mortensen-Pissarides matching model match the business-cycle facts? *Int Econ Rev* 40:933–959
- Dey M, Flinn C (2005) An equilibrium model of health insurance provision and wage determination. *Econometrica* 73:571–627
- Diamond P (1971) A model of price adjustment. *J Econ Theory* 3:156–168
- Eckstein Z, Wolpin K (1990) Estimating a market equilibrium search model from panel data on individuals. *Econometrica* 58:783–808
- Eckstein Z, Van den Berg G (forthcoming) Empirical labor search: a survey. *J Econom*
- Flinn C (1986) Wage and job mobility of young workers. *J Polit Econ* S88–S110
- Flinn C (2006) Minimum wage effects on labor market outcomes under search matching and endogenous contact rates. *Econometrica* 74:1013–1062
- Flinn C, Heckman J (1982a) Models for the analysis of labor force dynamics. In: Rhodes G, Basmann R (eds) *Advances in econometrics*. JAI Press, London, CT
- Flinn C, Heckman J (1982b) New methods for analyzing structural models of labor force dynamics. *J Econom* 18:115–168
- Gaumont D, Schindler M, Wright R (2006) Alternative theories of wage dispersion. *Eur Econ Rev* 50:831–848
- Hosios A (1990) On the efficiency of matching and related models of search and unemployment. *Rev Econ Stud* 57:279–298
- Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. *J Econom* 19:233–38
- Johnson W (1978) A theory of job shopping. *Q J Econ* 92:261–278
- Jovanovic B (1979a) Firm specific capital and turnover. *J Polit Econ* 87:1246–1260
- Jovanovic B (1979b) Job matching and the theory of turnover. *J Polit Econ* 87:972–990
- Kumbhakar SC, Lovell CAK (2000) *Stochastic frontier analysis*. Cambridge University Press, New York.
- Lippman S, McCall J (1976) The economics of job search: a survey, Part I. *Econ Inq* 14:155–189
- Marimon R, Zilibotti F (1999) Unemployment vs. mismatch of talent: reconsidering unemployment benefits. *Econ J* 109:266–291
- Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. *Int Econ Rev* 18:435–444
- Mortensen D, Pissarides C (1994) Job creation and job destruction in the theory of unemployment. *Rev Econ Stud* 61:397–415
- Nagypál É (2004) Learning-by-doing versus learning about match quality: can we tell them apart? Northwestern University, Mimeo
- Nelson P (1970) Information and consumer behavior. *J Polit Econ* 78(2):311–329
- Osbourne MJ, Rubinstein A (1990) *Bargaining and markets*. Academic Press, San Diego
- Pissarides CA (2000) *Equilibrium unemployment theory*. 2nd edn, MIT Press, Cambridge
- Polachek S, Yoon BJ (1987) A two-tiered earnings frontier estimation of employer and employee information in the labor market. *Rev Econ Stat* 69:296–302
- Polachek S, Yoon BJ (1996) Panel estimates of a two-tiered earnings frontier. *J Appl Econ* 11:169–178
- Postel-Vinay F, Robin J-M (2002) Equilibrium wage dispersion with heterogeneous workers and firms. *Econometrica* 70:2295–2350
- Postel-Vinay F, Robin J-M (2003) The distribution of earnings in an equilibrium search model with state-dependent offers and counter-offers. *Int Econ Rev* 43:989–1016
- Pries M (2004) Persistence of employment fluctuations: a model of recurring job loss. *Rev Econ Stud* 71:193–215
- Rothschild M (1973) Models of market organization with imperfect information: a survey. *J Polit Econ* 81(6):1283–1308
- Shapiro J (2006) Wage and effort dispersion. *Econ Lett* 92:163–169
- Shi S (2006) Wage differentials, discrimination and efficiency. *Eur Econ Rev* 50:849–875
- Tsionas EG (2008) Maximum likelihood estimation of non standard stochastic frontier models using the Fourier transform. Working paper, Athens University of Business and Economics
- Viscusi WK (1979) Job hazards and worker quit rates: an analysis of adaptive worker behavior. *Int Econ Rev* 20(1):29–58
- Viscusi WK (1980a) A theory of job shopping: a Bayesian perspective. *Q J Econ* 94(3):609–614
- Viscusi WK (1980b) Sex differences in worker quitting. *Rev Econ Stat* 62(3):388–398
- Viscusi WK (1980c) Self-selection, learning-induced quits, and the optimal wage structure. *Int Econ Rev* 21(3):529–546
- Viscusi WK (1983) Employment relationships with joint employer and worker experimentation. *Int Econ Rev* 24(2):313–322
- Wilde L (1980) An information-theoretic approach to job quits. In: Lippman S, McCall J (eds) *Studies in the economics of search*. Amsterdam, North Holland

Yujun Lian, Zhi Su, Yuedong Gu

Evaluating the Effects of Equity Incentives Using PSM: Evidence from China

© Higher Education Press and Springer-Verlag 2011

Abstract This paper investigates the effects of equity incentives on firm performance in Chinese listed firms. We address the sample selection problem by employing the propensity score matching methodology. Results show that, (1) On the whole, performance is positively related to equity incentives even after controlling for sample selection bias; (2) The final control rights have an important impact on the effects of equity incentives. The execution of equity incentives in privately owned firms can significantly decrease the agency costs between shareholders and managers, but such effects cannot be observed in state-owned firms; (3) Effects of equity incentives depend on the incentive type, that is, comparing to stock-based incentives, option-based incentives can reduce the agency costs significantly, thus are more effective; (4) Ownership structure also has important impacts on the effects of equity incentives. The agency costs decrease in firms with more decentralized ownership after introducing equity incentive, while in concentrated firms the effect is negligible.

Keywords equity incentives, firm performance, propensity score matching, bootstrap

Received June 9, 2010

Yujun Lian (✉)

Lingnan College, Sun Yat-sen University, Guangzhou 510275, China

E-mail: lianyj@mail.sysu.edu.cn

Zhi Su

School of Statistics, Central University of Finance and Economics, Beijing 100081, China

E-mail: suzhi1218@163.com

Yuedong Gu

School of Economics and Finance, Xi'an Jiaotong University, Xi'an 710063, China

E-mail: bigdoll@sohu.com

1 Introduction

Equity incentives are widely adopted in firms of the developed countries to align the interests of management with those of shareholders in order to improve firm performance. Up till January 2006, Chinese security regulations precluded listed firms from offering equity incentive plans to management (Ke, Rui and Yu, 2009). In late 2005, the China Securities Regulatory Commission (CSRC) made certain revisions to its Corporation Regulations and Security Regulations, and released new revisions in the document “Regulation of Equity Incentive Plans (trial)” (REIP) in January, 2006. Under the revised regulations, financial market conditions and the legal environment have improved dramatically. New conditions have made it possible for Chinese listed firms to adopt equity incentive plans. These include: (1) ongoing reform of non-tradable shares,¹ (2) changes in the regulations that allow firms to repurchase shares and allow management to trade shares within their terms, and (3) breakthroughs in the capital system that have made it possible for firms to adopt equity incentive plans.

At present, empirical studies regarding the effectiveness of such plans are limited and inconclusive. Based on 34 observations after the release of REIP, Yang and Li (2008) find that equity incentive plans cannot increase the value of Chinese listed firms. However, the study does not control for the sample selection bias and small sample bias. By focusing on Chinese listed firms that have adopted equity incentive plans from 2001 to 2006, Cheng and Xia (2008) find that equity incentive plans can indeed increase firm value. The effects are further enhanced by the ongoing reform of the split share structure. However, before 2006, Chinese listed firms could informally offer equity incentive plans to management. Thus, some of their observations are prior to the release of REIP. He (2008) finds that Chinese listed firms prefer to choose standardized incentive plans rather than individual plans for each executive. This tendency is more obvious in small firms or firms with highly concentrated ownership. Ke et al. (2009) compare firms cross-listed in both domestic and foreign markets with firms listed only in China. They find that the ability to offer equity incentive plans in the foreign market is an important determinant for Chinese firms to choose cross-listing. Additionally,

¹ Prior to the stock market reform, the Chinese domestic A-shares were divided into tradable and non-tradable shares with identical cash flow and voting rights. Non-tradable shareholders represent the government, holding about a two-thirds majority, and manage the firms, while tradable shareholders have little power to affect the decisions made by non-tradable shareholders. On April 29, 2005, the CSRC announced a program by which non-tradable shares would be converted into tradable shares. Holders of the non-tradable shares negotiated a compensation plan with holders of the tradable shares in order to make their shares tradable. By the end of 2006, the process was essentially complete, with over 95% of the affected companies completed the conversion (see Yeh, Shu, Lee and Su (2009), and Li, Wang, Cheung and Jiang (2010) for details).

adopting this plan can significantly increase shareholders' wealth, but it cannot improve firm performance, as proxied by ROA.

The literature has demonstrated theoretically that equity incentive plans can effectively improve firm performance (e.g., Baker, Jensen and Murphy, 1988; Shivdasani, 2002; among others). A potential problem in empirical studies on this topic regarding Chinese listed firms is that prior studies do not control for small sample bias. Since this plan is still relatively new, only a limited number of firms have adopted it. Thus, the sample size is small. The second concern, which is more important, is the sample selection bias documented in Heckman (1979). More specifically, firms with good prior performance are more inclined to adopt equity incentive plans. Consequently, observed improved performance may be driven by other reasons, rather than the plan itself. Surprisingly, there is little literature concerns this problem, yet little related evidence is provided.

In this study, we address the sample selection bias by applying the propensity score matching approach (PSM). Additionally, we use the Bootstrap method to control for the small sample bias. We investigate the following issues. First, can firms increase value by adopting equity incentive plans? Second, how does the type of the plan and the ownership structure change the effect of the plan on firm performance? Third, what drives the effectiveness of the plan? We find that equity incentive plans can indeed improve firm performance, as measured by ROE. However, this effectiveness is stronger in firms that offer option plans or in firms with a decentralized ownership structure. Additionally, equity incentive plans result in reduced agency costs and increased investment in fixed assets.

The rest of this paper proceeds as follows: Section 2 reviews the literature and develops the hypotheses. Section 3 presents the experimental design and econometric method. Section 4 describes the sample selection and variable measurement. Section 5 discusses the results, followed by the conclusion in Section 6.

2 Literature Review and Hypotheses Development

2.1 Equity Based Compensation and Firm Performance

The prior literature has shown that equity incentive plans can effectively improve firm performance in the United States (Murphy, 1999; Core, Guay and Larcker, 2003; Frydman and Saks, 2010). With an equity incentive plan in place, agency costs can be mitigated, since management not only tends to align their interests with those of shareholders, but also is motivated to focus, not on short term, but on long term firm performance.

The literature regarding the effectiveness of equity incentive plans in China is limited and inconclusive. One major concern of the financial market in China is

that the government has a large amount of control over it, however, there is still a lack of supervision of financial markets. Furthermore, the government generally not only controls a large portion of shares, but also nominates the management. This can result in corruption and management behavior of over spending and misuse of funds. Thus, agency costs are high in Chinese listed firms. Equity incentive plans seem to be a solution to this problem. With a promise of shares, management is motivated to make decisions for shareholders, including themselves. Therefore, Chinese listed firms should have better performance after adopting an equity incentive plan (Yu, 2006). Empirical studies generally support this view. The stock market generally responds positively to the announcement of the plan (Cui and Zhang, 2008; Zhang and Zheng, 2008). Meanwhile, a few studies argue that equity incentive plans can incur negative effects on corporations due to the unique governance structure in Chinese listed firms. For example, Yu and Gu (2001) find that equity incentive plans can increase the shareholdings of management and thus make the market less liquid. This can make the “insiders’ control” problem more severe. Lü and Zhao (2008) argue that with equity incentive plans, management has more incentive to manipulate earnings and utilize “window-dressing” of firm performance.

However, most empirical studies focus on horizons prior to the release of REIP in January, 2006. It is generally believed by market participants that the operation of corporations has become more efficient in recent years due to (1) REIP, the revised Corporation Regulations, and especially (2) reform of the split share structure, a major reform in the Chinese capital market that is expected to resolve problems prevailing in the market for a long time. Hence, under the new capital system and market environment, equity incentive plans are expected to motivate management to enhance shareholder value.

Based on the above analysis, we propose our first hypothesis:

H1 Equity incentive plans can improve firm performance in Chinese listed firms.

2.2 The Impact of Final Control Rights on Equity Based Compensation

Before entering the market as listed firms, the majority of Chinese firms are owned by the government, which also has final ownership control rights, even after those firms have been traded in the market. Along with the development of the economy, more and more private firms have been listed in the market. Consequently, the final control rights of those firms belong to the major shareholders. We argue that the variations of final ownership control rights could have varying impacts on the effectiveness of equity incentive plans.

In order to make an equity incentive plan effective, three necessary conditions have to be met (Yu, 2006). First, there is a competitive market for management,

from which all management is chosen. Second, management is motivated by economic incentives, not by other formats of incentives. Third, the board functions well. This means that management is under the supervision of board members. However, the above three conditions cannot be satisfied among Chinese listed firms controlled by the government. For example, unlike in the United States, there is not a competitive market for management in China. Most of those chosen for management are nominated by the government directly, resulting in management having a greater interest in satisfying the government, rather than shareholders. Thus, equity incentive plans will not be as effective in those firms as in firms controlled by shareholders, since the management in the latter, nominated by the board rather than the government, is expected to represent shareholders to increase shareholder value.

Based on the above analysis, we propose our second hypothesis:

H2 Equity incentive plans are more effective in private firms than in stated-owned firms.

2.3 The Effectiveness and Types of Equity Incentive Plans

A series of empirical studies have shown that the effectiveness of equity incentive plans depends on the type of the plan. For example, Goering (1996) demonstrates theoretically that (1) an optimum choice of a plan relies on the characteristics of a specific management, and (2) different types of plans are associated with varying results. Lazear (2000), Barron and Waddell (2003, 2008) also prove these points.

In China, firms can choose one of four types of equity incentive plans: (a) management stock options, (b) shares transferred to management from shareholders, (c) shares newly issued just for management, or (d) shares purchased for management from the market. In order to simplify the following discussion, we classify them into two categories: (1) option plans (consisting of (a) above), and (2) share plans (consisting of (b), (c) and (d) above).

In comparison with share plans, option plans can be more effective, due to the following reasons. First, option plans are more attractive to reputation management (Oyer and Schaefer, 2005). With option rewards, management has more incentive to improve firm performance in the long run. More specifically, if stock prices do not accelerate, option plans have no value to management, however, with share plans, management can retain certain value regardless of the stock price movement. Additionally, with option plans, management tends to choose more risky projects to boost up stock prices. By doing so, the potential insufficient investment problem, due to the risk adverse nature of management, could be resolved (Hall and Murphy, 2003). Second, option plans can increase the retention rate of management. Third, option plans can help alleviate the

financial constraints (Core and Guay, 2001; Kato, Lemmon, Luo and Schallheim, 2005). Given the fact that the profitability of listed firms in China is generally low, which greatly limits their internal financing capacity, and the Chinese capital market is far from perfect, which made it difficult for the majority of Chinese listed firms to issue new shares in the market (Lian and Chung, 2008), we thus expect the Chinese listed firms tend to adopt option plans relative to stock plans.

Based on the above analysis, we propose our third hypothesis:

H3 Option plans are more effective than share plans in improving firm performance.

2.4 Shareholding Concentration and the Effectiveness of Equity Incentive Plans

One purpose of equity incentive plans is to mitigate agency costs. However, agency costs vary with ownership structure. Variations in ownership structure can result in different effects of equity incentive plans.

Chinese listed firms have two distinctive characteristics. First, ownership is highly concentrated. This is more severe in firms controlled by the government. Second, insiders are powerful and have tight control over firms. The above two features could make equity incentive plans less effective than expected. First, when ownership is highly concentrated, big shareholders play a significant role in firms. This makes the role of both management and equity incentive plans less significant. Mehran (1995) finds that with the existence of large shareholders, firms offer fewer options. Ke, Petroni and Safieddine (1999) argue that large shareholders can supervise management at lower cost. This makes motivating management with option plans unimportant. Thus, large shareholders and option plans seem to be the replacement of each other. Second, when shares are highly concentrated, controlling shareholders (especially government as a major controller) generally have goals other than that of management. Under these conditions, option plans are useless (Ke et al., 1999). For example, the goal of government as the controlling shareholder might be to retain all employees to maintain the prosperity of the society, etc. This is not feasible if management has total control of the firm. Therefore, option plans work better when the ownership is more scattered. In those firms, the goals of management and the shareholders could be easily aligned, and thus agency costs could be reduced. Third, firms with highly concentrated ownership are vulnerable to “tunneling,” a kind of financial fraud in which a group of major shareholders of a publicly traded company orders that the company sell off its assets to a second company, owned by the group of shareholders, at unreasonably low prices. The shareholders typically own the second company outright, and thus profit from the otherwise disastrous sale. In the event such a scheme is underway, management has no incentive to enhance stockholder value, and equity incentive plans are, of course,

not applicable. Rational management would rather quit or take no efforts.

Based on the above analysis, we propose our fourth hypothesis:

H4 The effects of equity incentive plans are negatively related to the concentration of ownership.

3 Methodology

To empirically test the hypotheses proposed above, we classify the samples into two groups: (1) The incentive group, including firms with equity incentive plans, and (2) the control group, including firms without such plans. As mentioned above, it is necessary to control for sample selection bias. We apply the propensity score matching (PSM) method developed by Rosenbaum and Rubin (1983). Using this approach, we can obtain propensity scores (*PS*), which measure the extent of matching of the incentive group and the control group in multi-dimensions.

In the following, we briefly introduce how to calculate *PS* values, followed by discussions regarding the three matching approaches and the average effect of treatment on the treated (*ATT*).

3.1 Propensity Score

The propensity score is defined as “the conditional probability of receiving a treatment given pre-treatment characteristics” by Rosenbaum and Rubin (1983).

$$p(X) = \Pr[D = 1 | X] = E[D | X], \quad (1)$$

where X is the multidimensional vector of characteristics of the control group, D is the indicator variable, which equals 1 if a firm adopts an equity incentive plan and 0 otherwise. Theoretically, if we can get the estimates of propensity score $p(X_i)$ (we will discuss this issue in details in the next section), the *ATT* can be estimated by the differences of the potential outcomes of the incentive group and the control group (Becker and Ichino, 2002),

$$\begin{aligned} ATT &= E[Y_{1i} - Y_{0i} | D_i = 1] \\ &= E\{E[Y_{1i} - Y_{0i} | D_i = 1, p(X_i)]\} \\ &= E\{E[Y_{1i} | D_i = 1, p(X_i)] - E[Y_{0i} | D_i = 0, p(X_i)] | D_i = 1\}, \end{aligned} \quad (2)$$

where Y_{1i} and Y_{0i} represent the potential outcomes of the incentive group and the control group, respectively.

To estimate the *PS* score, we follow Dehejia and Wahba (2002) and Becker and Ichino (2002) and use the Logit model with the following steps.

We start with estimating probabilities using the Logit model,

$$p(X_i) = \Pr(D_i = 1 | X_i) = \frac{\exp(\beta X_i)}{1 + \exp(\beta X_i)}, \quad (3)$$

where X is the multidimensional vector of independent variables which may affect the propensity of firms to implement equity incentive plans, and β is the vector of coefficients. The propensity score (PS) is the predicted values of the Logit model.

3.2 Matching Methods

We cannot estimate the ATT of interest directly using (2) even though the propensity scores have been estimated. The reason is that $p(X)$ is a continuous variable, and thus it is impossible to find two units with identical propensity score. Several matching methods have been suggested in the literature to overcome this problem. Three of the most widely used are Nearest-Neighbor Matching, Radius Matching and Kernel Matching.²

The nearest neighbor matching method is to search the closest control sample, both backwards and forwards, from the estimated PS values of the incentive group. Let T and C represent the incentive group and the control group, and Y_i^T and Y_j^C represent the observed performance of incentive and control units, respectively. Meanwhile, let $C(i)$ represent the set of control units matched to the i^{th} incentive unit with an estimated value of the propensity score of p_i . Then, the nearest neighbor matching method can be described as follows,

$$C(i) = \min_j \|p_i - p_j\|. \quad (4)$$

The radius matching method is to search all units in the control group, and those with estimated propensity scores falling within a radius r from p_i are matched to the incentive unit i . r is a positive real number set beforehand. We can describe the radius matching method as follows,

$$C(i) = \{p_j \mid \|p_i - p_j\| < r\}. \quad (5)$$

After identifying the matching samples using nearest neighbor or radius matching, we can calculate ATT . Let the number of controls matched with observation $i \in T$ by N_i^C and define the weights $w_{ij} = 1/N_i^C$ if $j \in C(i)$ and $w_{ij} = 0$ otherwise. Also, suppose there are N^T numbers of observations in the incentive group. Then, according to Becker and Ichino (2002), the ATT can be estimated as follows,

² The methods and formulas introduced here are closely related to Becker and Ichino (2002).

$$\tau^M = \frac{1}{N^T} \sum_{i \in T} Y_i^T - \frac{1}{N^T} \sum_{j \in C} w_j Y_j^C, \quad (6)$$

where M represents the matching method, such as the nearest neighbor matching method or the radius matching method, and w_j is defined as $w_j = \sum_i w_{ij}$. Assume that the weights are held constant and the effectiveness of the equity incentive plan in each firm is independent. The variance of τ^M could be estimated as follows,

$$\text{Var}(\tau^M) = \frac{1}{N^T} \text{Var}(Y_i^T) + \frac{1}{(N^T)^2} \sum_{j \in C} (w_j)^2 \text{Var}(Y_j^C). \quad (7)$$

The idea of the Kernel matching method is somewhat different from the previous two methods. With Kernel matching, a fictitious unit will be constructed to match each incentive firm, that is, each incentive firm is matched with a weighted average of all controls with weights that are inversely proportional to the distance between the propensity scores of incentive and controls. When the Kernel matching method is used, the *ATT* can be estimated as follows,

$$\tau^K = \frac{1}{N^T} \sum_{i \in T} \left\{ Y_i^T - \frac{\sum_{j \in C} Y_j^C G((p_j - p_i)/h_n)}{\sum_{k \in C} G((p_k - p_i)/h_n)} \right\}, \quad (8)$$

where $G(\cdot)$ represents the Gaussian kernel function, and h_n represents bandwidth parameter. Under standard conditions on the bandwidth and kernel,

$$\frac{\sum_{j \in C} Y_j^C G((p_j - p_i)/h_n)}{\sum_{k \in C} G((p_k - p_i)/h_n)}$$

is a consistent estimator of the counterfactual outcome Y_{0i} . Since it is difficult to get the estimation of variance of τ^K , bootstrap is instead used. We apply the same approach and estimate the variance of τ^K based on bootstrap with 500 replications.

3.3 Obtaining Robust Variance of *ATT* Using Bootstrap

One problem of the statistical analysis in this study is that the sample size is small. To reduce the small sample size bias, we use the bootstrap approach to estimate the standard errors for further analysis.³ We follow Efron and Tibshirani

³ An important feature of this method is that we need not assume the distribution of the statistic of interest in prior (See Abadie, Drukker, Herr and Imbens (2004), and Abadie and Imbens (2006)).

(1993) and take steps as follows. First, we resample n observations with replacements from the original sample and get the so called empirical sample. Second, we use the above matching approaches to calculate ATT_i . Third, we repeat the first and the second step for K times ($K = 500$ in this study)⁴ and obtain $ATT_1, ATT_2, \dots, ATT_{500}$. Fourth, we calculate the standard deviation of $ATT_1, ATT_2, \dots, ATT_{500}$, thus get the standard errors of the ATT statistic.

4 Sample and Proxies

4.1 Sample Construction

Our data set is obtained from the China Stock Market and Accounting Research Database (CSMAR) developed by Shenzhen GTA Information Technology Company. CSMAR provides detailed information on CEOs as well as accounting and financial data for Chinese listed firm.

We identified firms with equity incentive plans in 2006 and 2007 (starting from December of 2005) and study the performance of those firms from 2008 to 2009. We deleted the observations without any clear documentation of the offering date and details. Our final dataset has 59 firms that adopt the equity incentives during 2006–2007, among which, 46 firms adopt option plans and the remaining 13 firms offer share plans. Those firms are in our Incentive group. Regarding the control group, we take the following steps to identify them. First, we eliminate firms in the financial sector (banks, insurance, and other financial firms), as they are subject to different disclosure requirements in China. ST/PT firms are also excluded because their financial conditions are abnormal.⁵ Second, we delete firms with leverage ratio greater than 100%. Third, we delete firms with sales (or total asset) growth rate greater than 150% as these firms may involves in large asset sales or mergers. Eventually, we find 1 346 firms in the Control group, among which we are going to identify those that can match firms in the Incentive group. Additionally, the key financial variables are winsorized at the 1st and 99th percentiles to avoid the influence of outliers.

⁴ Efron and Tibshirani (1993) suggest that, to get standard error using bootstrap, 200 replications are often good enough to give a good estimate of standard error.

⁵ According the Chinese Company Law, listed firms that have been making losses (negative net earnings) for two consecutive years are categorized as “special treatment” (ST), whereas companies that have been making losses for three consecutive years are to be put into “Particular Treatment” (PT) status and are suspended from the Exchanges. ST firms are limited to 5% share-price movements up or down daily. PT firms are given a maximum one-year grace period to return to profitability, failing which they will be permanently de-listed from the Stock Exchange. There are no ST/PT firms in our incentive sub-sample.

4.2 Measurement of Effectiveness of Equity Incentives

Following the literature, we use *ROE* to measure the effectiveness of equity incentive plans (e.g., Xia and Zhang, 2008). We do not use Tobin’s *Q* as a measure, due to two reasons. First, Tobin’s *Q* reflects long term investment opportunity (Gomes, 2001; Lian and Chung, 2008), whereas *ROE* measures short term performance. Since we focus on a relatively short period, *ROE* is a better indicator of firm performance than Tobin’s *Q*. Second, in our sample period, the Chinese stock market is highly volatile. Under this market, it is difficult to separate the performance of firms from that of the market. Thus, Tobin’s *Q* is not suitable in this study, since it is based on stock prices.

To measure the effectiveness of equity incentive plans, the first measure that we use is the ratio of management costs to total revenue (*AC*), a proxy for agency costs. We expect *AC* of the Incentive group to be less than that of the Control group. Second, we use both the ratio of investment expenditure over total assets (*INVT*) and the growth rate of total assets (*TAGR*) as other measures of the effectiveness of the plans. In other words, we measure the effectiveness of the plans from the point of view of both investment activities (*INVT*, *TAGR*) and agency cost decreasing (*AC*).

Table 1 shows summary statistics of variables.

Table 1 Summary Statistics of Variables

Variables	Mean	S.D.	Min	Max
<i>ROE</i>	0.041	0.167	−1.015	0.303
<i>AC</i>	0.089	0.084	0.008	0.573
<i>INVT</i>	0.061	0.059	0.000	0.519
<i>TAGR</i>	0.118	0.207	−0.331	0.845

Note: This table presents the summary statistics of the variables. *ROE* is calculated by dividing net income by total equity. *AC* equals the ratio of management costs over sales. *INVT* is calculated by dividing investment expenditure to total assets. *TAGR* represents the percentage change of total assets.

5 Empirical Results

5.1 Propensity to Perform Equity Incentive Plans

The first step to perform propensity score matching analysis is to estimate the propensity scores (*PS*). The *PS* values summarize several pretreatment firm characteristics of each subject into a single-index, which makes matching subjects on an *n*-dimensional vector of characteristics feasible for large *n*. Note that, multi-dimensional matching is typically unfeasible using traditional

methods.

Table 2 Summary Statistics for the Matching Variables

Variable	Mean	S.D.	Min	Max
<i>SIZE</i>	21.514	1.114	18.157	28.003
<i>LEV</i>	0.501	0.179	0.009	0.997
<i>TOBIN</i>	1.661	0.902	0.811	5.889
<i>TANG</i>	0.472	0.170	0.000	0.975
<i>PROF</i>	0.044	0.161	-0.861	0.460
<i>GPAY</i>	13.430	0.834	3.367	17.583
<i>MAGSTK</i>	0.028	0.103	0.000	0.784
<i>HEYI</i>	0.139	0.345	0.000	1.000
<i>HHI5</i>	0.175	0.121	0.004	0.760
<i>TOPONE</i>	0.367	0.153	0.035	0.864
<i>Zindex</i>	20.981	49.288	0.720	802.970
<i>Stated-owned</i>	0.644	0.479	0.000	1.000

Note: This table provides summary statistics for the variables in our sample of firm-years from Chinese publicly traded firms over the period 2005 to 2009. *SIZE* is the natural log of total assets. *LEV* is total debt divided by total assets. *TOBIN* is the ratio of market value to replacement costs, where market value of the firm is defined as the number of tradable shares multiplied by the stock’s closing price at the fiscal year-end plus number of non-tradable shares multiplied by the net asset per share, plus the total debt, and replacement costs is measured by book value of the firm. *TANG* is defined as the fixed assets divided by the total assets. *PROF* is defined as the retained profits divided by the sales income. *GPAY* is the natural log of annual salary of the top three managers. *MAGSTK* is defined as the number of shares managers hold divided by the number of the total shares. *HEYI* equals one if the broad chairman is also the top manager, and zero otherwise. *TOPONE* is the shareholding proportion of the top one (largest) shareholder. *HHI5* is the Herfindahl-Hirschman Index, defined as the sum of squared share proportions of top five shareholders. *Zindex* is calculated as the ratio of shares hold by the largest shareholders to the second large shareholder. *Stated-owned* is a dummy variable that equals one if a firm is stated-owned, and zero otherwise.

In line with previous literature, we estimate the Logit model (3) to get the *PS* values. To get a good specification of the Logit model, we follow Tzioumis (2008) among others, and add variables related to firm’s financial characters, corporate governance, and manager’s compensations in the model. We expect such a multi-dimensional matching can help us find the control firms that are as similar as possible to the incentive ones. The variables included in model (3) are: (1) financial variables, including firm size (*SIZE*), leverage (*LEV*), growth

opportunity (*TOBIN*), asset structure (*TANG*), profitability (*PROF*), and industry dummies; (2) corporate governance variables, including shareholding concentration (*HHI5*), the shareholding proportion of the top one shareholder (*TOPONE*), the balance of power of large shareholders (*Zindex*), whether the broad chairman is also the top manager (*HEYI*), and a dummy variable (*Stated-owned*) indicating the final control type of the firm; (3) manager's compensation variables, including CEO compensation (*GPAY*), and holding proportion of managers (*MAGSTK*). Table 2 presents the summary statistics of the variables mentioned above.

We estimate model (3) with various specifications, and the results are presented in Table 3. In order to control the industrial effects, we add industry dummies in all the models in Table 3. It is shown that the probability of implementing equity incentive plans is significantly positively related to firm size (*SIZE*), growth opportunity (*TOBIN*), profitably (*PROF*), and CEO compensation (*GPAY*), and is negatively correlated to leverage (*LEV*), shareholding concentration (*HHI5*, *TOPONE*), and the balance of power of large shareholders (*Zindex*). The probability to implement equity incentive plans is not significantly related to the asset structure (*TANG*) which is measured by the ratio of fixed assets to total assets, and the dummy variable indicating whether the broad chairman is also the top manager (*HEYI*). Moreover, the propensity to implement incentive plans is lower in stated-owned firms relative to other firms, as the *Stated-owned* dummies are negatively significant in all specifications.

Table 3 The Estimation Results of Logit Models

Variables	(1)	(2)	(3)	(4)	(5)
<i>SIZE</i>	0.846*** (10.42)	0.850*** (10.42)	0.818*** (10.37)	0.775*** (10.07)	0.843*** (10.45)
<i>LEV</i>	-1.277** (-2.35)	-1.280** (-2.40)	-1.159** (-2.19)	-1.050** (-1.99)	-1.284** (-2.41)
<i>TOBIN</i>	0.379*** (5.31)	0.380*** (5.34)	0.384*** (5.44)	0.407*** (5.83)	0.378*** (5.32)
<i>TANG</i>	0.016 (0.03)				
<i>PROF</i>	1.253* (1.80)	1.208* (1.74)	1.250* (1.81)	1.012 (1.50)	1.247* (1.80)
<i>GPAY</i>	0.410*** (4.10)	0.412*** (4.16)	0.412*** (4.17)	0.433*** (4.35)	0.415*** (4.20)

(To be continued)

(Continued)

Variables	(1)	(2)	(3)	(4)	(5)
<i>MAGSTK</i>		0.382 (0.60)			
<i>HEYI</i>	0.076 (0.38)				
<i>HHI5</i>	-2.729*** (-3.86)	-2.727*** (-3.86)			-2.725*** (-3.86)
<i>TOPONE</i>			-1.854*** (-3.56)		
<i>Zindex</i>				-0.007** (-2.00)	
<i>Stated-owned</i>	-1.371*** (-8.23)	-1.355*** (-7.96)	-1.370*** (-8.20)	-1.414*** (-8.48)	-1.376*** (-8.28)
Industry dummies	Yes	Yes	Yes	Yes	
Constant	-27.838*** (-14.43)	-27.975*** (-14.44)	-27.295*** (-14.21)	-27.678*** (-14.24)	-27.821*** (-14.51)
Pseudo- <i>R</i> ²	0.193	0.193	0.191	0.187	0.199
AUC	0.844	0.844	0.842	0.842	0.854
Observations	3 339	3 339	3 339	3 339	3 339

Note: 1. The dependent variable is “Incentive,” which is a discrete variable, equals one if a firm has equity incentive plans, and zero otherwise.
2. ***, ** and * represent significance at 1%, 5% and 10% level, respectively, with *t*-values in parentheses.
3. The AUC denotes the area under the ROC curve.

Our final goal is to estimate the propensity scores, according to which we can match the incentive firms to their control pairs. Obviously, the model specification is an important thing to ensure the validity of the matching procedure. Unfortunately, there is no straightforward criterion to properly specify the Logit model in the literature. We use two diagnostic proxies namely pseudo-*R*² which is widely used in Logit analysis, and the area under the ROC curve (AUC). The reason we use AUC is that, the dependent variable in the Logit model is a discrete variable (0/1), while the propensity scores (which is the predicted values of the Logit model) is a continuous variable, thus the traditional statistics (such as Pearson correlation coefficient) can not be use to analyze their correlation (see Hosmer and Lemeshow, 2000). In this case, the AUC which is widely used in the Receiver Operating Characteristic (ROC) literature can give

better inference.⁶

The pseudo- R^2 's are in the 0.187–0.199 range. This goodness-of-fit is higher than those reported in Villalonga (2004). Comparing the values of Pseudo- R^2 and AUC in Table 3, we can see that specification (5) is better than others. Stürmer, Joshi, Glynn, Avorn, Rothman and Schneeweiss (2006) find that, when we use Logit model to get propensity scores, an AUC value larger than 0.8 can be regarded as a good indicator that the model is well specified. In model (5), the AUC is 0.854, well above the value suggested by Stürmer et al. (2006). Therefore, we will use model (5) as our basic specification to calculate the propensity scores, and then compare the firm performances between incentive firms and control firms.

5.2 Sample Matching Results

The following discussion is based on the nearest neighbor matching approach. Fig. 1(a) and Fig. 1(b) show the kernel density functions of the Incentive group and the Control group, based on pre- and post-matching of the two groups, respectively. Clearly, the kernel density functions of the two groups are significantly different before matching. Prior studies use all firms in the Control group to compare with the Incentive group, and thus their results are biased. In contrast, we choose firms from the Control group to match those in the Incentive group, based on Propensity scores. After matching, as shown in Fig. 1(b), the kernel density functions of the two groups are a lot closer, indicating that the characteristics of the variables in the two groups are similar, after matching. We also match the two groups using radius matching and kernel matching. The results are similar and are not reported.

5.3 Analysis of the Effectiveness of Equity Incentive Plans

5.3.1 Results on H1: The Full Sample Effects of Equity Incentives

We use three approaches, as mentioned above, to estimate *ATTs*. The following discussion is based on the nearest neighbor matching approach. The results from the other two methods are used as robustness tests.

Table 4 shows the *ATTs* based on the nearest neighbor matching method, both pre- and post-matching. In the analysis of either pre or post matching, we find that ROE is significantly different from zero at the 5% level, indicating that equity based compensation can indeed improve firm performance. This is consistent with H1.

⁶ For details on ROC analysis, see Fawcett (2006) and Stein (2005).

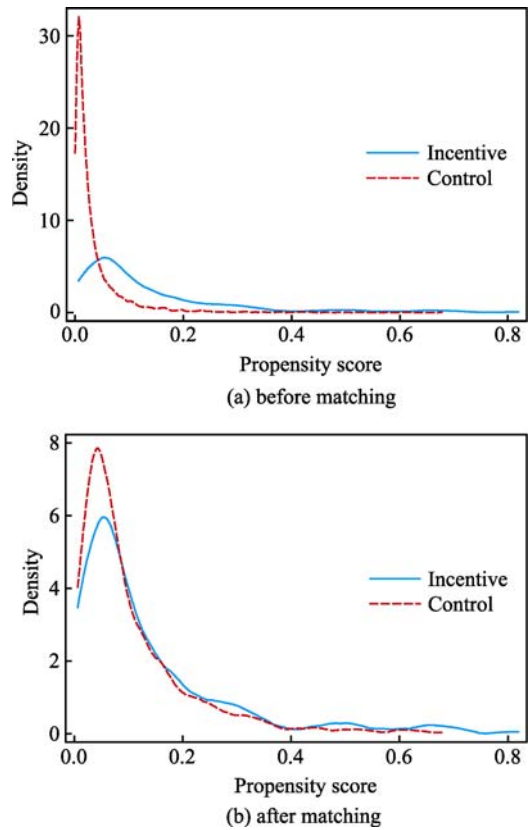


Fig. 1 Kernel Density of the Incentive and Control Groups

A natural question that comes up is: What drives equity incentive plans to be so effective? To answer this question, we further analyze several economic factors. Regarding agency costs, we do not find any significant differences of *ACs* between the Incentive group and the Control group after matching. This indicates that agency costs are not significantly reduced by equity incentive plans. In detail, before matching, *ACs* of the Incentive group and the Control group are 0.077 and 0.09, respectively, significant at the 5% level. This indicates that firms that adopt the plans have lower agency costs than average firms in the market. However, after matching, the *ACs* of the two groups are 0.077 and 0.083, respectively. Still, the *AC* of the Incentive group is lower than that of the Control group, but the differences are insignificant at the 10% level. Hence, at least in our sample range, we can not find evidence that it is the agency costs that drive the improvement of *ROE* caused by equity incentive plans.

Table 4 Comparison of *ATTs* (by Nearest Neighbor Matching Approach)

Variable	Sample	Incentive group	Control group	<i>ATT</i>	<i>s.e.</i>	<i>t</i> -value
<i>ROE</i>	Pre-matching	0.116	0.038	0.078	0.011	6.92***
	Post-matching	0.116	0.100	0.016	0.007	2.23**
<i>AC</i>	Pre-matching	0.077	0.090	-0.013	0.006	-2.18**
	Post-matching	0.077	0.083	-0.005	0.005	-0.99
<i>INVT</i>	Pre-matching	0.072	0.060	0.011	0.004	2.80***
	Post-matching	0.072	0.061	0.011	0.005	2.28**
<i>TAGR</i>	Pre-matching	0.263	0.112	0.151	0.014	10.80***
	Post-matching	0.263	0.192	0.071	0.019	3.65***

Note: 1. “Pre-matching” refers to the sample without matching the Incentive group with the Control group, and “Post-matching” refers the groups after matching.
2. “Incentive group” and “Control group” refer to firms with and without equity based compensation, respectively.
3. ***, ** and * represent significance at 1%, 5% and 10% level, respectively.
4. Standard errors are calculated using Bootstrap with 500 replications.

We thus investigate another mechanism through which firms can improve performance: changes in investment behavior. The comparison of *INVT* shows that *INVT* of the Incentive group, 0.072, is greater than that in the control group after matching, 0.061, at the 5% level. In other words, firms with equity incentive plans invest 18% more than those without such plans. A greater level of investment causes these firms to have a greater potential to grow, which is shown in the significant differences of *TAGR* between the Incentive group and the Control group, at the 1% significance level.

In sum, we find that firms with equity incentive plans can improve firm performance; the improvement is mainly via increasing investment, not via reducing agency costs. This seems to be inconsistent with the commonly accepted assertion which states that equity incentive plans motivate management to align its interest with that of shareholders, and thus reduce agency costs.

5.3.2 Results on H2: The Effects of Final Control Right on Equity Incentives

To examine the second hypothesis, we further classify the Incentive group into two subgroups: (1) The stated-owned (government controlled) subgroup, and, (2) the privately owned subgroup. Comparison of the two subgroups with respect to the effectiveness of equity incentive plans is shown in Table 5.

Table 5 The Impact of Final Control Rights on the Effectiveness of Equity Incentives

Variable	A. Full sample		B. Stated owned		C. Privately owned	
	ATT	t-value	ATT	t-value	ATT	t-value
Nearest neighbor matching						
ROE	0.016	2.23**	0.012	0.78	0.029	1.96**
AC	−0.005	−0.99	0.000	0.01	−0.014	−1.99**
INVT	0.011	2.28**	−0.003	−0.36	0.012	1.66*
TAGR	0.071	3.65***	0.058	1.69*	0.112	3.92***
Radius matching						
ROE	0.021	2.85***	0.008	0.53	0.035	2.76***
AC	−0.004	−0.81	−0.004	−0.61	−0.014	−1.57
INVT	0.008	1.69*	0.012	1.33	0.010	1.49
TAGR	0.075	4.15***	0.062	1.84*	0.109	4.07***
Kernel matching						
ROE	0.032	4.42***	0.020	1.56	0.037	3.55***
AC	−0.007	−1.41	−0.001	−0.13	−0.008	−1.97**
INVT	0.011	2.70***	0.001	0.09	0.019	3.13***
TAGR	0.092	5.27***	0.077	2.72***	0.117	4.59***

Note: 1. “Stated owned” and “private owned” refer stated-owned firms and private-owned firms, respectively.
2. ***, ** and * represent significance at 1%, 5% and 10% level, respectively.
3. Standard errors are calculated using Bootstrap with 500 replications.

Panel B and C of Table 5 show the results of *ATTs* of the two subgroups. The results generally support the second hypothesis. Using the nearest neighbor matching method, we find significant differences between the two subgroups. *ROEs* are not significantly different within the stated-owned subgroup, whereas they are different within the privately owned subgroup, significant at the 1% or 5% level. Thus, we argue that the differences of firm performance (*ROE*) in Table 4 between the Incentive group and the Control group are driven mainly by the privately owned subgroup. This further demonstrates that equity incentive plans can effectively motivate management to improve firm performance.

The examination of *AC*, *INVT* and *TAGR* can explain the source of the above difference. Panel B and C of Table 5 show the results of the above variables for the stated-owned subgroup and the privately owned subgroup, respectively. We find that none of the three variables are significant at 5% level in Panel B, whereas all of the three variables are significant at the 5% level in Panel C. In detail: (1) Agency costs in privately owned firms with equity incentive plans are significantly lower than firms without such plans, and (2) The average investment, as measured by *INVT* and the growth rate of total assets (*TAGR*), is

higher in privately owned firms with equity incentive plans than in firms without such plans. In other words, in the privately owned firms, equity based compensation can motivate management to reduce agency costs and increase investment.

Further, we use both radius matching and kernel matching as robustness tests to reexamine the first and second hypotheses. Focusing on the full sample in Panel A of Table 5, we find similar results, by radius matching and kernel matching methods, as those in Table 4. Hence, the robustness tests further confirm the results of the first hypothesis. Focusing on the subgroups in Panel B and Panel C, we also identify the similar patterns by these two approaches. Thus, the robustness of the results for the second hypothesis is also demonstrated.

5.3.3 Results on H3: The Effects of Incentive Type

We test the *ATTs* of the alternative type of the incentive plans: (1) option plans, and (2) share plans. The results, as shown in Panel A and B of Table 6, show that the types of the plans can produce different results in the effectiveness of equity incentive plans.

We find that the option plans can significantly improve firm performance as measured by *ROE*. This result is robust whether we calculate *ATTs* using nearest neighbor matching, radius matching, or kernel matching method. However, the effect of stock plans is much weaker. We observe significant performance improvement of stock plans only when the kernel matching method is used.

Further comparison of *AC*, *INVT* and *TAGR* in Panel A and B shows that the causal factors driving the effectiveness of equity incentive plans differ between the two styles of plans. First, the *ATTs* of *AC* representing agency costs under all three approaches are not significant in Panel A, indicating that stock plans cannot reduce agency costs. In comparison, results in Panel B shows that option plans can significantly reduce agency costs: With the nearest neighbor matching method and the radius matching method, *ATTs* of *AC* are less than 0 at the 5% significance level. Kernel matching is the strictest method among the three. With this method, we find that *ATTs* of *AC* are less than 0 at the 1% significance level. Meanwhile, the *AC* of the Incentive group has been reduced by 0.014, with an average reduction rate over 20%. Regarding *INVT*, in all matching methods, stock plans do not significantly increase firms' investment expenditure. In comparison, option plans can significantly increase firm investment (*INVT*) and especially total assets growth (*TAGR*) as shown in Panel B.

The above results are supportive of the theoretical analysis. Option plans are not going to bring benefits to management unless the future stock price is higher than the exercise price of the option. In contrast, share plans bring benefits to management immediately. Thus, option plans bring more incentive to

management and motivate action to improve firm performance. In reality, among the 95 firms that adopted equity incentive plans in 2008, only 16 firms chose the share plan. The value in 2009 is 87 vs 18. This indicates that option plans became more popular as more firms adopting equity incentive plans.

Table 6 The Impact of Reward Methods on the Effectiveness of Equity Compensation

Variable	A. Stock reward		B. Option reward	
	ATT	t-value	ATT	t-value
Nearest neighbor matching				
ROE	0.025	1.47	0.019	2.11**
AC	−0.031	−1.30	−0.005	−1.98**
INVT	0.004	0.40	0.005	0.95
TAGR	0.050	1.11	0.081	3.77***
Radius matching				
ROE	0.015	0.87	0.019	2.26**
AC	0.009	0.71	−0.008	−2.03**
INVT	0.015	1.40	0.009	1.70*
TAGR	0.020	0.51	0.093	4.66***
Kernel matching				
ROE	0.044	2.99***	0.038	5.21***
AC	0.017	1.25	−0.014	−3.26***
INVT	0.014	1.45	0.010	2.15**
TAGR	0.061	1.52	0.114	6.11***

Note: 1. The examples of “stock rewards” include: transferring shares from other shareholders to management, issuing new stocks to management, or using company funds to purchase shares for management from the market. “Option rewards” include issuing stock options to management.
2. ***, ** and * represent significance at 1%, 5% and 10% level, respectively.
3. Standard errors are calculated using Bootstrap with 500 replications.

The above analysis provides evidence to our third hypothesis. The findings suggest that different reward approaches are associated with variant effects. Especially in reducing agency costs, option plans can be very effective.

5.3.4 Results on H4: The Effects of Shareholding Concentration

We classify sample firms into two subgroups based on ownership concentration. In specifically, in each year, we split the sample into two groups according to the

sample mean of Herfindahl-Hirschman Index (*HHI5*). Firms with *HHI5* larger than sample mean are classified into the concentrated-ownership subgroup, while the remaining firms are classified into the scatted-ownership group. The *ATTs* are then calculated in each subgroup. The results are shown in Panel A and B in Table 7.

Table 7 The Impact of Shareholding Concentration on the Effectiveness of Equity Incentives

Variable	A. Concentrated ownership		B. Scattered ownership	
	<i>ATT</i>	<i>t</i> -value	<i>ATT</i>	<i>t</i> -value
Nearest neighbor matching				
<i>ROE</i>	0.003	0.23	0.015	1.98**
<i>AC</i>	0.002	0.20	−0.006	−1.84*
<i>INVT</i>	0.003	0.38	0.022	3.92***
<i>TAGR</i>	0.123	3.32***	0.065	2.77***
Radius matching				
<i>ROE</i>	0.003	0.27	0.024	2.22**
<i>AC</i>	0.011	1.25	−0.007	−2.12**
<i>INVT</i>	0.008	0.89	0.014	2.41**
<i>TAGR</i>	0.123	3.19***	0.069	3.16***
Kernel matching				
<i>ROE</i>	0.021	2.31**	0.030	2.59***
<i>AC</i>	0.000	−0.06	−0.009	−2.29**
<i>INVT</i>	0.005	0.67	0.018	3.28***
<i>TAGR</i>	0.133	3.91***	0.073	3.40***

Note: 1. Firms are classified into the subgroup with concentrated ownership when their largest shareholders' ownership is more than the median among all firms. The remaining firms are classified into the subgroup with scattered ownership.
2. ***, ** and * represent significance at 1%, 5% and 10% level, respectively.
3. Standard errors are calculated using Bootstrap with 500 replications.

Using nearest neighbor matching and radius matching, we find that in the concentrated ownership subgroup, the *ATTs* of *ROE* are not significant. They are significant only when the kernel matching method is used. Thus, we can not find strong evidence that equity incentive can improve firm performance when ownerships are highly concentrated. , However, the results in Panel B show that, in scattered ownership subgroup, the *ATTs* of *ROE* are greater than zero, at the 1% significance level, under all three matching methods. Thus, consistent with the fourth hypothesis, we find that the effect of equity based compensation

decreases with ownership concentration. Furthermore, *ATTs* of the *AC* are significantly negative in the scattered ownership subgroup, whereas they are insignificant in the concentrated ownership subgroup. This indicates that agency costs can be significantly reduced among firms with scattered ownership, in that when firms have scattered ownership the interests of management are more aligned with that of shareholders. The investment expenditure show similar patterns between two sub-groups, which indicates that managers in scattered ownership group invest more. Regarding growth rate of total assets (*TAGR*), the *ATTs* in both groups are significantly positive, indicating that equity incentive plans do enhance firm growth. This is consistent with results reported in Table 4 for the whole sample.

In summary, the effectiveness of equity incentive plans may be affected by the ownership of the largest shareholders. Equity incentive plans are more effective in firms with more scattered ownership. This is consistent with the fourth hypothesis.

6 Conclusion

On January 1st, 2006, new regulations regarding equity incentive plans were released in China. During year 2006–2007, 59 companies adopted equity incentive plans. In this study, we focus on those 59 firms to study the effectiveness of equity incentive plans in China.

Prior studies in the literature do not control for sample selection bias. We contribute to the literature by using the Propensity Score Matching (PSM) approach to control for sample selection bias. In specific, we use nearest neighbor matching, radius matching and kernel matching, to identify proper target firms for each of the observations in the sample. We further use the Bootstrap method to calculate the standard errors to control for the small sample bias.

We find that option plans can effectively improve firm performance, but such effectiveness can only be observed in firms controlled by major shareholders, not by the government. Further, option plans are more effective than share plans. Lastly, the effectiveness is greater in firms with more scattered ownership.

Additionally, we examine the causal factors that drive the effectiveness. We find that among firms controlled by major shareholders, equity incentive plans can effectively reduce agency costs and increase investment. Moreover, option plans can also significantly reduce agency costs. This effect does not appear among firms with share plans. Lastly, when shareholder ownership is more scattered, agency costs could be reduced through the adoption of an equity incentive plan. We do not find similar results among firms with more concentrated ownership.

In summary, equity incentive plans can effectively motivate managements to align their interest with that of shareholders. However, the effectiveness depends on the ownership structure and the style of the plan, such as option plans or share plans.

The present study has a number of limitations. Firstly, there are only 59 firms adopt equity incentives in our sample, which may induce the small sample bias, though we try to overcome this problem by using the bootstrap techniques. Secondly, Chinese listed firms begin to adopt the new accounting standards,⁷ which may have some effects on the definition of financial variables in our study. As our sample ranges from 2005 to 2009, which covers the accounting reform period, this may bias our estimation to some extent. Unfortunately, we can not find a proper method to handle this problem.

Acknowledgements This paper is supported by the National Science Foundation of China (No. 71002056, 70902071), Social Scientific Research Foundation of Ministry of Education of China (No. 09YJC790269), Natural Science Foundation of Guangdong Province (No. 9451027501002497), and the Fundamental Research Fund for the Central Universities. The authors would like to thank Dongdong Ding for his assistance with data collection and management.

References

- Abadie, A., Drukker, D., Herr, J., & Imbens, G. 2004. Implementing matching estimators for average treatment effects in stata. *The Stata Journal*, 4(3): 290–311.
- Abadie, A., & Imbens, G. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1): 235–267.
- Baker, G., Jensen, M., & Murphy, K. 1988. Compensation and incentives: Practice vs. Theory. *Journal of Finance*, 43(3): 593–616.
- Barron, J. M., & Waddell, G. R. 2003. Executive rank, pay and project selection. *Journal of Financial Economics*, 67(2): 305–349.
- Barron, J. M., & Waddell, G. R. 2008. Work hard, not smart: Stock options in executive compensation. *Journal of Economic Behavior & Organization*, 66(3–4): 767–790.
- Becker, S., & Ichino, A. 2002. Estimation of average treatment effects based on propensity scores. *The Stata Journal*, 2(4): 358–377.
- Cheng, Z. 程仲鸣, & Xia, Y. 夏银桂. 2008. 制度变迁、国家控股与股权激励 (Institutional change, state block-holder and managerial equity incentives). *南开管理评论 (Nankai Business Review)*, 11(4): 89–95.
- Core, J., Guay, W., & Larcker, D. 2003. Executive equity compensation and incentives: A survey. *Economic Policy Review*, 9(1): 27–50.
- Core, J. E., & Guay, W. R. 2001. Stock option plans for non-executive employees. *Journal of Financial Economics*, 61(2): 253–287.
- Cui, M. 崔明会, & Zhang, B. 张兵. 2008. 上市公司股权激励的短期财富效应研究 (The short-term wealth effect of equity incentive in listed companies). *经济研究导刊*

⁷ See Qu and Zhang (2010) for details.

- (*Economic Research Guide*), (14): 95–98.
- Dehejia, R. H., & Wahba, S. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1): 151–161.
- Efron, B., & Tibshirani, R. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.
- Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8): 861–874.
- Frydman, C., & Saks, R. 2010. Executive compensation: A new view from a long-term perspective, 1936–2005. *Review of Financial Studies*, 23(5): 2099–2138.
- Goering, G. E. 1996. Managerial style and the strategic choice of executive incentives. *Managerial and Decision Economics*, 17(1): 71–82.
- Gomes, J. 2001. Financing investment. *American Economic Review*, 91(5): 1263–1285.
- Hall, B. J., & Murphy, K. J. 2003. The trouble with stock options. *Journal of Economic Perspectives*, 17(3): 49–70.
- He, F. 何凡. 2008. 高管层激励股权分布结构及其成因 (Research on the structure and causes of top management equity incentives). *当代经济科学 (Modern Economic Science)*, 30(6): 98–103.
- Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica*, 47(1): 153–161.
- Hosmer, D., & Lemeshow, S. 2000. *Applied logistic regression*. New York: John Wiley & Sons, Inc.
- Kato, H., Lemmon, M., Luo, M., & Schallheim, J. 2005. An empirical examination of the costs and benefits of executive stock options: Evidence from Japan. *Journal of Financial Economics*, 78(2): 435–461.
- Ke, B., Petroni, K., & Safieddine, A. 1999. Ownership concentration and sensitivity of executive pay to accounting performance measures: Evidence from publicly and privately-held insurance companies. *Journal of Accounting and Economics*, 28(2): 185–209.
- Ke, B., Rui, O., & Yu, W. 2009. *Hong Kong stock listing and the sensitivity of managerial compensation to firm performance in state-controlled chinese firms*. Working Paper, Pennsylvania State University.
- Lazear, E. P. 2000. The power of incentives. *American Economic Review*, 90(2): 410–414.
- Li, K., Wang, T., Cheung, Y., & Jiang, P. 2010. Privatization and risk sharing: Evidence from the split share structure reform in china. *Review of Financial Studies*, forthcoming.
- Lian, Y., & Chung, C. F. 2008. *Are Chinese listed firms over-investing?* SSRN working paper, Available at SSRN: <http://ssrn.com/abstract=1296462>
- Lü, C. 吕长江, & Zhao, Y. 赵宇恒. 2008. 国有企业管理者激励效应研究 (A study on the effect of the incentive given to managers of state-owned enterprises). *管理世界 (Management World)*, (11): 99–109.
- Mehran, H. 1995. Executive compensation structure, ownership, and firm performance. *Journal of Financial Economics*, 38(2): 163–184.
- Murphy, K. 1999. Executive compensation. *Handbooks in Economics*, (5): 2485–2566.
- Oyer, P., & Schaefer, S. 2005. Why do some firms give stock options to all employees? An empirical examination of alternative theories. *Journal of Financial Economics*, 76(1): 99–133.
- Qu, X., & Zhang, G. 2010. Measuring the convergence of national accounting standards with international financial reporting standards: The application of fuzzy clustering analysis. *The International Journal of Accounting*, 45(3): 334–355.
- Rosenbaum, P., & Rubin, D. 1983. The central role of the propensity score in observational

- studies for causal effects. *Biometrika*, 70(1): 41–55.
- Shivdasani, A. 2002. The economics of executive compensation. *Journal of Finance*, 57(2): 1023–1028.
- Stürmer, T., Joshi, M., Glynn, R., Avorn, J., Rothman, K., & Schneeweiss, S. 2006. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59(5): 437.
- Stein, R. 2005. The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking and Finance*, 29(5): 1213–1236.
- Tzioumis, K. 2008. Why do firms adopt CEO stock options? Evidence from the United States. *Journal of Economic Behavior & Organization*, 68(1): 100–111.
- Villalonga, B. 2004. Does diversification cause the “diversification discount”? *Financial Management*, 33(2): 5–27.
- Xia, J. 夏纪军, & Zhang, Y. 张晏. 2008. 控制权与激励的冲突——兼对股权激励有效性的实证分析 (The conflicts between control rights and incentives: An empirical analysis on the effect of stock incentives in china). *经济研究 (Economic Research)*, (3): 87–98.
- Yang, H. 杨惠贤, & Li, L. 李丽瑛. 2008. 我国上市公司股权激励及其绩效的实证研究 (An empirical analysis of the effects of equity incentive plans in chinese listed firms). *中国管理信息化 (China Management Informationization)*, 11(23): 40–42.
- Yeh, Y. H., Shu, P. G., Lee, T. S., & Su, Y. H. 2009. Non-tradable share reform and corporate governance in the chinese stock market. *Corporate Governance: An International Review*, 17(4): 457–475.
- Yu, D. 于东智, & Gu, L. 谷立日. 2001. 上市公司管理层持股的激励效用及影响因素 (Influence factors of managerial ownership and effectiveness of stimulation). *经济理论与经济管理 (Economic Theory and Business Management)*, (9): 24–31.
- Yu, H. 俞鸿琳. 2006. 国有上市公司管理者股权激励效应的实证检验 (The effects of equity incentive in chinese state-owned listed firms: An empirical analysis). *经济科学 (Economy Science)*, (1): 108–116.
- Zhang, Y. 张颖, & Zheng, X. 郑学清. 2008. 股权激励的市场反应及其内幕交易的实证研究 (An empirical study on the market reaction of equity incentive and insider trading). *华东经济管理 (East China Economic Management)*, 22(12): 155–158.

Financial development and dynamic investment behavior: Evidence from panel VAR

Inessa Love^a, Lea Zicchino^{b,*}

^a World Bank, Research Department—Finance Group, 1818 Hst, NW, MC3-300,
Washington, DC 20433, United States

^b Bank of England, Monetary Instruments and Markets Division, HO-2,
Threadneedle Street, London EC2R 8AH, UK

Received 14 April 2004; received in revised form 31 October 2005; accepted 4 November 2005

Abstract

We apply vector autoregression (VAR) to firm-level panel data from 36 countries to study the dynamic relationship between firms' financial conditions and investment. By using orthogonalized impulse-response functions we are able to separate the 'fundamental factors' (such as marginal profitability of investment) from the 'financial factors' (such as availability of internal finance) that influence the level of investment. We find that the impact of financial factors on investment, which indicates the severity of financing constraints, is significantly larger in countries with less developed financial systems. Our finding emphasizes the role of financial development in improving capital allocation and growth.

© 2006 Board of Trustees of the University of Illinois. All rights reserved.

Keywords: Financial development; Vector autoregression; Dynamic investment behavior

1. Introduction

Unlike the neoclassical theory of investment, the literature based on asymmetric information emphasizes the role played by moral hazard and adverse selection problems in a firm's decision to invest in physical and human capital. The presence of asymmetric information means that the classical dichotomy between real and financial variables may no longer hold. Financial variables can have an impact on real variables, such as the level of investment and the real interest rate, as well as propagate and amplify the effects of exogenous shocks to the economy. For example, [Bernanke](#)

* Corresponding author. Tel.: +44 20 7601 5212; fax: +44 20 7601 3217.
E-mail address: lea.zicchino@bankofengland.co.uk (L. Zicchino).

and Gertler (1989) show that a firm's net worth (a financial variable) can be used as collateral in order to reduce the agency cost associated with the presence of asymmetric information between lenders and borrowers. In this model, firms' investment decisions are not only dependent on the present value of future marginal productivity of capital, as the *q*-theory predicts, but also on the level of collateral available to the firms when they enter a loan contract.

Since economists started to look at real phenomena abstracting from the Arrow-Debreu framework with its frictionless capital markets, a vast literature has been developed on the relationship between investment decisions and firms' financing constraints (see [Hubbard, 1998](#), for a review). Even though asymmetric information between borrowers and lenders may be not the only source of imperfection in the credit markets, firms seem to prefer internal to external finance to fund their investments. This observation leads to the prediction of a positive relationship between investment and internal finance. The first study on panel data by Fazzari, Hubbard, and Peterson (1988) finds that, after controlling for investment opportunities with Tobin's *q*, changes in net worth have a greater impact on investment by firms with higher costs of external financing.

The link between the cost of external financing and investment decisions not only sheds light on the dynamics of business cycles but also represents an important element in understanding economic development and growth. For instance, in the presence of moral hazard in the credit market, firms that need a bank loan may be induced to undertake risky investment projects with low expected marginal productivity. This corporate decision affects the growth path of the economy, which may even fall into a poverty trap (see [Zicchino, 2001](#)). [Rajan and Zingales \(1998\)](#), [Demirguc-Kunt and Maksimovic \(1998\)](#) and [Wurgler \(2000\)](#), among others, have investigated the link between finance and growth by asking whether underdeveloped legal and financial systems could prevent firms from investing in potentially profitable growth opportunities. Their empirical results show that an active stock market, developed financial intermediaries and the respect of legal norms are determinants of economic growth.

Estimation of the relationship between investment and financial variables is challenging because it is difficult for an econometrician to observe firms' net worth and investment opportunities. In theory, the measure of investment opportunities is the present value of expected future profits from additional capital investment, or what is commonly called marginal *q*. This is the shadow value of an additional unit of capital and, under certain conditions, it can be shown to be a sufficient statistic for investment ([Hayashi, 1982](#)). In other words, it is the 'fundamental' factor that determines investment policy of profit-maximizing firms in efficient markets. The difficulty in measuring marginal *q*, which is not observable, results in low explanatory power of the *q*-models and, typically, entails implausible estimates of the adjustment cost parameters.¹

Another challenge is finding an appropriate measure for the 'financial' factors that enter the investment equation in models with capital markets imperfections. A widely used measure for the availability of internal funds is cash flow (current revenues less expenses and taxes, generally scaled by capital). However, cash flow is likely to be correlated with future investment profitability.² This makes it difficult to distinguish the response of investment to the 'fundamen-

¹ See [Whited \(1998\)](#) and [Erickson and Whited \(2000\)](#) for a discussion of the measurement errors in investment models. Also see [Schiantarelli \(1996\)](#) and [Hubbard \(1998\)](#) for a review on methodological issues related to investment models with financial constraints.

² For example, the current realization of cash flow would proxy for future investment opportunities if the productivity shocks were positively serially correlated.

tal' factors, such as marginal profitability of capital, and 'financial' factors, such as net worth (see Gilchrist and Himmelberg, 1995, 1998, for further discussion of this terminology).

In this paper we use the vector autoregression (VAR) approach to overcome this problem and isolate the response of investment to financial and fundamental factors. Specifically, we focus on the orthogonalized impulse-response functions, which show the response of one variable of interest (i.e. investment) to an orthogonal shock in another variable of interest (i.e. marginal productivity or a financial variable). By orthogonalizing the response we are able to identify the effect of one shock at a time, while holding other shocks constant.

We use firm-level panel data from 36 countries to study the dynamic relationship between firms' financial conditions and investment levels. Our main interest is to study whether the dynamics of investment are different across countries with different levels of development of financial markets. We argue that the level of financial development in a country can be used as an indication of the different degrees of financing constraints faced by firms. After controlling for the shocks to 'fundamental' factors, we interpret the response of investment to 'financial' factors as evidence of financing constraints and we expect this response to be larger in countries with lower levels of financial development. To test this hypothesis we divide our data in two groups according to the degree of financial development of the country in which they operate. We document significant differences in the response of investment to 'financial' factors for the two groups of countries. Furthermore, splitting the sample based on different indicators of economic development does not produce significant differences, supporting our claim that the level of *financial* development is the main determinant of financing constraints.

We believe our paper contributes to a number of strands in the recent financial economics literature. We contribute to the literature on financial constraints and investment in several ways. First, by using vector autoregressions on panel data we are able to consider the complex relationship between investment opportunities and the financial situation of the firms, while allowing for a firm-specific unobserved heterogeneity in the levels of the variables (i.e. fixed effects). Second, thanks to a reduced-form VAR approach, our results do not rely on strong assumptions that are necessary in models that use the q -theory of investment or Euler equations. Third, by analyzing orthogonalized impulse-response functions we are able to separate the response of investment to shocks coming from fundamental or financial factors.

We also contribute to the finance and growth literature by presenting new evidence that investment in firms operating in financially underdeveloped countries exhibits dynamic patterns consistent with the presence of financing constraints. Our paper also adds to the recent work that used dynamic panel-data techniques to argue that there is a causal link between financial development and growth (see, for example, Beck and Levine, 2004). While most of the previous studies relied on country-level data, our paper uses firm-level data to demonstrate how the link between finance and growth operates on the level of the firm and to provide additional evidence on the channels behind this link. Specifically, we find that financial development has an immediate effect on efficient allocation of capital via investment that follows the most productive uses of capital. Our finding is also consistent with the evidence presented by Beck, Demircuc-Kunt, Levine, and Maksimovic (2001) who found that it is easier for firms' to access external financing in countries with a higher level of overall financial sector development.

Our paper also adds to the recent debate on bank-based versus market-based financial systems (see, for example, Demircuc-Kunt and Levine, 2001a, b, and Beck and Levine (2002), among others). This literature demonstrated that despite conflicting theoretical predictions, there is no empirical evidence of the relationship between financial structures and economic growth. However, the literature has found that it is the overall financial development that helps in explaining

cross-countries differences in economic performance. Our findings are consistent with this literature and expand the range of real effect previously studied to include the micro-level evidence of the effect of financial development on investment behavior and capital allocation.

Our paper is also related to Gilchrist and Himmelberg (1995, 1998), who were the first to analyze the relationship between investment, future capital productivity and firms' cash flow with a panel-data VAR approach. They use a two-stage estimation procedure to obtain measures of what they call 'fundamental' q and 'financial' q . These factors are then substituted in a structural model of investment, which is a transformation of the Euler-equation model. Unlike Gilchrist and Himmelberg, we do not estimate a structural model of investment, but instead study the unrestricted reduced-form dynamics afforded by the VAR (which is in effect the first-stage in their estimation). Gallegati and Stanca (1999) also investigate the relationship between firms' balance sheets and investment by estimating reduced-form VARs on company panel data for UK firms. Despite some differences in the specification of the empirical model and the estimation methodology, the approach and the results of their paper are similar to ours. However, they do not present an analysis of the impulse-response functions, which we consider to be the main tool in separating the role of financial variables in companies' investment decisions. In addition, the distinguishing feature of our paper is the focus on the differences in the dynamic behavior of firms in countries with different levels of financial development.

Finally, our paper is related to Love (2003) who uses the Euler-equation approach and shows that financing constraints are more severe in countries with lower levels of financial development, the same as we find in this paper. However, the interpretation of the results in the previous paper is heavily dependent on the assumptions and parameterization of the model, while the approach we use here imposes the bare minimum of restrictions on parameters and temporal correlations among variables.

The rest of the paper is as follows: Section 2 presents the empirical specification and the data description; Section 3 provides the results of our work; and Section 4 presents our conclusions.

2. Empirical methodology

We use a panel-data vector autoregression methodology. This technique combines the traditional VAR approach, which treats all the variables in the system as endogenous, with the panel-data approach, which allows for unobserved individual heterogeneity. We specify a first-order VAR model as follows:

$$z_{it} = \Gamma_0 + \Gamma_1 z_{it-1} + f_i + d_{c,t} + e_t \quad (1)$$

where z_t is either a three-variable vector $\{SKB, CFKB, IKB\}$ or a four-variable vector $\{SKB, CFKB, IKB, TOBINQ\}$; SKB , sales to capital ratio, is our proxy for the marginal productivity of capital.³ IKB is the investment to capital ratio, which is our main variable of interest, $CFKB$ is cash flow scaled by capital, and $TOBINQ$ is 'Tobin's q ', measured as market value of assets over book value of assets.

In this model sales to capital ratio and Tobin's q represent 'fundamental' factors, i.e. factors that capture the marginal productivity of capital. In the absence of market frictions, positive shocks to

³ See Gilchrist and Himmelberg (1998) for a derivation of the ratio of sales to capital as a measure of marginal productivity of capital.

these fundamental factors should lead to an increase in investment as firms will take advantage of better investment opportunities.

Cash flow is commonly used in investment models as an indicator for internally available funds (see [Hubbard, 1998](#), for a review). In our model, we consider cash flow also as a proxy for ‘financial factors’.⁴ Our analysis is implicitly based on an investment model in which, after controlling for the marginal profitability, the effect of the financial variables on investment is interpreted as evidence of financing constraints.⁵ We do this by relying on the orthogonalization of impulse responses. Because the shocks are orthogonalized, i.e. ‘fundamentals’ are kept constant, the impulse response of investment to cash flow isolates the effect of the ‘financial’ factors. We use this orthogonalized response of investment to ‘financial factors’ as a measure of market frictions and financing constraints.

The impulse-response functions describe the reaction of one variable to the innovations in another variable in the system, while holding all other shocks equal to zero. However, since the actual variance–covariance matrix of the errors is unlikely to be diagonal, to isolate shocks to one of the variables in the system it is necessary to decompose the residuals in a such a way that they become orthogonal. The usual convention is to adopt a particular ordering and allocate any correlation between the residuals of any two elements to the variable that comes first in the ordering.⁶ The identifying assumption is that the variables that come earlier in the ordering affect the following variables contemporaneously, as well as with a lag, while the variables that come later affect the previous variables only with a lag. In other words, the variables that appear earlier in the systems are more exogenous and the ones that appear later are more endogenous.⁷

In our specification we assume that current shocks to the marginal productivity of capital (proxied by sales to capital) have an effect on the contemporaneous value of investment, while investment has an effect on the marginal productivity of capital only with a lag. We believe this assumption is plausible for two reasons. First, the sales to capital ratio is likely to be the most exogenous firm-level variable since it depends on the demand for firms’ output, which often is outside of the firms’ control (of course, sales depend on firms’ actions as well, but most likely with a lag). Second, investment is likely to become effective with some delay since it requires time to become fully operational (the so-called “time-to-build” effect). We also argue that the effect of sales on cash flow is likely to be contemporaneous and if there is any feedback effect it is likely to happen with a lag. Finally, we assume that investment responds to cash flow contemporaneously, while cash flow responds to investment only with a lag.⁸ In the model with four variables, we

⁴ Although cash flow is the most commonly used proxy for net worth, it is closely related to operating profits, and therefore to the marginal productivity of capital. If the investment expenditure does not result in higher sales but in lower costs (i.e. more efficiency), the sales to capital ratio (our main measure of marginal productivity of capital) would not pick up this effect, while the cash flow would. Thus, cash flow may partly capture fundamental factors rather than the financial factors affecting investment. To reduce this effect, we included *Tobin*q as another fundamental variable to control for marginal productivity and investment opportunities.

⁵ See [Gilchrist and Himmelberg \(1998\)](#) for a more formal structural model that is behind their first-stage reduced VAR approach, which is similar to our approach.

⁶ The procedure is known as Choleski decomposition of variance–covariance matrix of residuals and is equivalent to transforming the system in a “recursive” VAR for identification purposes. See [Hamilton \(1994\)](#) for the derivations and discussion of impulse-response functions.

⁷ More formally, if a variable x appears earlier in the system than a variable y , then x is weakly exogenous with respect to y in the short run.

⁸ Our results are robust to changing the order of cash flow and investment.

assume that Tobin's q affects all other variables with a lag and is simultaneously affected by all other variables. As a result, TOBINQ is the most endogenous variable in the system, thus capturing all available information (i.e. all the contemporaneous shocks to other variables).

Our main objective is to compare the response of investment to financial factors in countries on a different level of financial development. To achieve this, we split our firms into two samples according to the level of financial development of the country in which they operate and we analyze the difference in impulse responses for the two samples. We refer to these two groups as 'high' (financial development) and 'low' (financial development). This distinction is relative and is based on the median level of financial development among countries in our sample.⁹

In applying the VAR procedure to panel data, we need to impose the restriction that the underlying structure is the same for each cross-sectional unit. Since this constraint is likely to be violated in practice, one way to overcome the restriction on parameters is to allow for "individual heterogeneity" in the levels of the variables by introducing fixed effects, denoted by f_i in the model. Since the fixed effects are correlated with the regressors due to lags of the dependent variables, the mean-differencing procedure commonly used to eliminate fixed effects would create biased coefficients. To avoid this problem we use forward mean-differencing, also referred to as the 'Helmert procedure' (see Arellano and Bover, 1995). This procedure removes only the forward mean, i.e. the mean of all the future observations available for each firm-year. This transformation preserves the orthogonality between transformed variables and lagged regressors, so we can use lagged regressors as instruments and estimate the coefficients by system GMM.¹⁰

Our model also allows for country-specific time dummies, $d_{c,t}$, which are added to model (1) to capture aggregate, country-specific macro shocks that may affect all firms in the same way. We eliminate these dummies by subtracting the means of each variable calculated for each country-year.

To analyze the impulse-response functions we need an estimate of their confidence intervals. Since the matrix of impulse-response functions is constructed from the estimated VAR coefficients, their standard errors need to be taken into account. We calculate standard errors of the impulse-response functions and generate confidence intervals with Monte Carlo simulations.¹¹ To compare the impulse responses across our two samples (i.e. 'high' and 'low' financial development) we simply take their difference. Because our two samples are independent, the impulse responses of the differences are equal to the difference in impulse responses (the same applies to the simulated confidence intervals).

Finally, we also present variance decompositions, which show the percent of the variation in one variable that is explained by the shock to another variable, accumulated over time. The variance decompositions show the magnitude of the total effect. We report the total effect accumulated over the 10 years, but longer time horizons produced equivalent results.

⁹ A recent paper by Powell, Ratha, and Mohapatra (2002) uses similar approach to ours (i.e. splitting the countries into two groups and estimating VARs separately for each group) to study the interrelationships between inflows and outflows of capital and other macro variables.

¹⁰ In our case the model is "just identified", i.e. the number of regressors equals the number of instruments, therefore system GMM is numerically equivalent to equation-by-equation 2SLS.

¹¹ In practice, we randomly generate a draw of coefficients Γ of model (1) using the estimated coefficients and their variance-covariance matrix and re-calculate the impulse-responses. We repeat this procedure 1000 times (we experimented with a larger number of repetitions and obtained similar results). We generate 5th and 95th percentiles of this distribution which we use as a confidence interval for the impulse-responses.

Table 1

Sample coverage across countries

Country	Country code	Number of observations	Percent of total observations	Number of firms	Financial development
Panel A: Low financial development sample					
Argentina	AR	250	0.005	39	−1.38
Belgium	BE	586	0.01	91	−0.82
Brazil	BR	894	0.02	143	−1.04
Chile	CL	507	0.01	74	−0.75
Colombia	CO	146	0.00	21	−1.6
Denmark	DK	1051	0.02	138	−0.49
Finland	FI	818	0.02	113	−0.41
Indonesia	ID	708	0.01	114	−1.17
India	IN	1856	0.03	294	−0.7
Italy	IT	1100	0.02	151	−0.64
Mexico	MX	522	0.01	76	−0.85
New Zealand	NZ	304	0.006	44	−0.53
Philippines	PH	406	0.008	68	−1.15
Pakistan	PK	546	0.01	88	−1.28
Portugal	PT	291	0.005	53	−0.67
Sweden	SE	1178	0.02	178	−0.31
Turkey	TR	248	0.005	54	−1.2
Venezuela	VE	92	0.002	13	−1.26
Group average		639	0.012	97	−1
Group total		11503		1752	
Panel B: High financial development sample					
Austria	AT	530	0.01	83	−0.27
Australia	AU	1383	0.03	184	0.42
Canada	CA	3136	0.06	443	0.03
Switzerland	CH	1087	0.02	151	2.2
Germany	DE	4092	0.08	582	1.68
Spain	ES	987	0.02	134	−0.14
France	FR	3338	0.06	524	0.1
United Kingdom	GB	8657	0.16	1165	1.68
Israel	IL	164	0.00	37	0.01
Japan	JP	6654	0.12	1271	3.3
South Korea	KR	1643	0.03	259	0.84
Malaysia	MY	1837	0.03	291	1.19
Netherlands	NL	1282	0.02	154	0.66
Norway	NO	878	0.02	148	−0.15
Singapore	SG	906	0.02	145	1.6
Thailand	TH	1233	0.02	185	0.36
USA	US	3399	0.06	356	1.35
South Africa	ZA	1189	0.02	244	0.25
Group average		2355	0.044	353	1
Group total		42395		6356	
Total sample		53898		8108	

Countries are split into two groups based on the median level of financial development.

Table 2
Variable definitions

Abbreviation	Description
Firm-level variables (from <i>Worldscope</i>)	
CAPEX	Capital expenditure
NETPEQ	Property plant and equipment
SALES	Net sales or revenues
IKB	Investment to capital ratio = $\text{CAPEX}/(\text{NETPEQ} - \text{CAPEX})$
SKB	Sales to capital ratio = $\text{SALES}/(\text{NETPEQ} - \text{CAPEX})$
CF	Cash flow (derived from <i>Worldscope</i> cash flow to sales ratio)
CFKB	Cash flow divided by $(\text{NETPEQ} - \text{CAPEX})$
RANK	Ranking based on size of PPENT (first ranked by year, then averaged over the years), largest firm in each country has rank equal to one
TOBINQ	Tobin's q , generated as market value of equity plus book value of total liabilities divided by book value of total assets
Country-level variables	
STKMKT	Stock market development is Index 1 from Demirguc-Kunt and Levine (1996) , equals to the sum of (standardized indices of) market capitalization to GDP, total value traded to GDP, and turnover (total value traded to market capitalization)
FININT	Financial intermediary development is Index 1 from Demirguc-Kunt and Levine (1996) , equals to the sum of (standardized indices of) ratio of liquid liabilities to GDP, and ratio of domestic credit to private sector to GDP
FD	Financial development = $\text{STKMKT} + \text{FININT}$
GDPPC	GDP per capita from World development indicators
HIGHINC	World bank classification category based on 2002 gross national income per capita

2.1. Data

Our firm-level data come from the *Worldscope* database, which contains standardized accounting information on large publicly traded firms and includes 36 countries with over 8000 firms for the years 1988–1998. [Table 1](#) gives the list of countries in the sample with the number of firms and observations per country, while details on the sample selection are given in Appendix 1. The number of firms included in the sample varies widely across the countries and the less developed countries are underrepresented. The US and UK have more than 1000 firms per country, while the rest of the countries have only 136 firms on average (Japan is the third largest with over 600 firms). Such a prevalence of US and UK companies might overweight these countries in the cross-country regressions and prevent smaller countries from influencing the coefficients.¹²

We constructed the index of financial development, FD by combining standardized measures of five indicators from [Demirguc-Kunt and Levine \(1996\)](#): market capitalization over GDP, total value traded over GDP, total value traded over market capitalization, ratio of liquid liabilities (M3) to GDP and credit going to the private sector over GDP. We split the countries into two groups based on the median of this indicator. We refer to these two groups as ‘high’ (financial development) and ‘low’ (financial development), but we remind the reader that this distinction

¹² To reduce the influence of countries with a large number of firms we also run our regressions with a subgroup of firms in each country, i.e. only including a fixed number of the largest firms. The inclusion criteria are based on firm ranking, where rank one is given to the largest firm in each country. The results obtained were very similar to the ones reported in the paper and are available on request.

Table 3
Summary statistics for main variables

	Low financial development sample					High financial development sample				
	Mean	Standard deviation	25th percentile	50th percentile	75th percentile	Mean	Standard deviation	25th percentile	50th percentile	75th percentile
SKB	3.39	3.54	1.06	2.31	4.38	4.12	4.05	1.41	2.92	5.33
IKB	0.21	0.15	0.10	0.17	0.28	0.21	0.14	0.11	0.18	0.27
CFKB	0.29	0.32	0.11	0.22	0.38	0.28	0.28	0.13	0.23	0.38
TOBINQ	1.35	0.78	0.89	1.11	1.51	1.46	0.76	1.00	1.22	1.63

Summary statistics by country for main variables. Variable definitions are given in Table 2. Countries are split into two groups based on the median level of financial development.

is relative and is based on the median level of financial development among countries in our sample.

Table 2 summarizes all the variables used in the paper (note that we normalize all the firm-level variables by the beginning-of-period capital stock), and Table 3 reports the summary statistics for the firm-level variables.

3. Results

We estimate the coefficients of the system given in (1) after the fixed effects and the country-time dummy variables have been removed. In Table 4 we report the results of the model with three variables {SKB, IKB, CFKB}, while in Table 5 we report the model with four variables {SKB, IKB, CFKB, TOBINQ}. We report the results that include all sample of firms in each

Table 4
Main results of a 3-variable VAR model

Response of	Response to		
	SKB(<i>t</i> -1)	CFKB(<i>t</i> -1)	IKB(<i>t</i> -1)
Panel A: Low financial development sample			
SKB(<i>t</i>)	0.571 (6.77)***	0.359 (1.54)	-1.528 (-7.03)***
CFKB(<i>t</i>)	0.025 (3.61)***	0.300 (9.68)***	-0.124 (-5.66)***
IKB(<i>t</i>)	-0.009 (-1.98)	0.129 (5.54)***	0.111 (5.83)***
<i>N</i> obs	7228		
<i>N</i> firms	1518		
Panel B: High financial development sample			
SKB(<i>t</i>)	0.462 (13.55)***	0.771 (5.56)***	-1.599 (-12.22)***
CFKB(<i>t</i>)	0.010 (3.75)***	0.361 (19.99)***	-0.104 (-7.89)***
IKB(<i>t</i>)	0.004 (-2.19)	0.084 (7.06)***	0.132 (9.99)***
<i>N</i> obs	26675		
<i>N</i> firms	5370		

Variable definitions are in Table 2. Three variable VAR model is estimated by GMM, country-time and fixed effects are removed prior to estimation (see Section 2 for details). Countries are split into two groups based on the median level of financial development. Reported numbers show the coefficients of regressing the row variables on lags of the column variables. Heteroskedasticity adjusted *t*-statistics are in parentheses. *** indicates significance at 1% level.

Table 5
Main results of a 4-variable VAR with Tobin's q

Response of	Response to			
	SKB($t-1$)	CFKB($t-1$)	IKB($t-1$)	TOBINQ($t-1$)
Panel A: Low financial development sample				
SKB(t)	0.589 (6.04)***	0.363 – 1.470	–1.610 (–6.82)***	0.208 (2.26)
CFKB(t)	0.023 (2.70)	0.275 (8.03)***	–0.111 (–4.47)***	0.024 (1.56)
IKB(t)	–0.012 (–2.12)	0.123 (4.98)***	0.118 (5.49)***	0.041 (2.77)
TOBINQ(t)	–0.0005 (–0.08)	0.039 (1.27)	–0.020 (–0.96)	0.449 (12.97)***
N obs	5813			
N firms	1381			
Panel B: High financial development sample				
SKB(t)	0.447 (11.93)***	0.578 (3.89)***	–1.442 (–10.25)***	0.330 (6.10)***
CFKB(t)	0.009 (2.9)**	0.329 (18.31)***	–0.097 (–6.89)***	0.070 (8.37)***
IKB(t)	0.005 (2.04)	0.065 (5.21)***	0.122 (9.33)***	0.055 (7.63)***
TOBINQ(t)	–0.004 (–1.75)	0.071 (4.36)***	–0.029 (–2.00)	0.464 (24.61)***
N obs	24253			
N firms	5032			

Variable definitions are in Table 2. Four variable VAR model is estimated by GMM, country-time and fixed effects are removed prior to estimation (see Section 2 for details). Countries are split into two groups based on the median level of financial development. Reported numbers show the coefficients of regressing the row variables on lags of the column variables. Heteroskedasticity adjusted t -statistics are in parentheses. *** and ** indicates significance at 1% and 5% level, respectively.

country.¹³ We present graphs of the impulse-response functions and the 5% error bands generated by Monte Carlo simulation. Fig. 1 reports graphs of impulse responses for the model with three variables estimated for a sample of countries with 'low' financial development, while Fig. 2 reports this model for countries with 'high' financial development. In Fig. 3 we show the differences in impulse responses of two samples (the difference is 'low' minus 'high'). Figs. 4–6 present similar graphs for the model with TOBINQ.

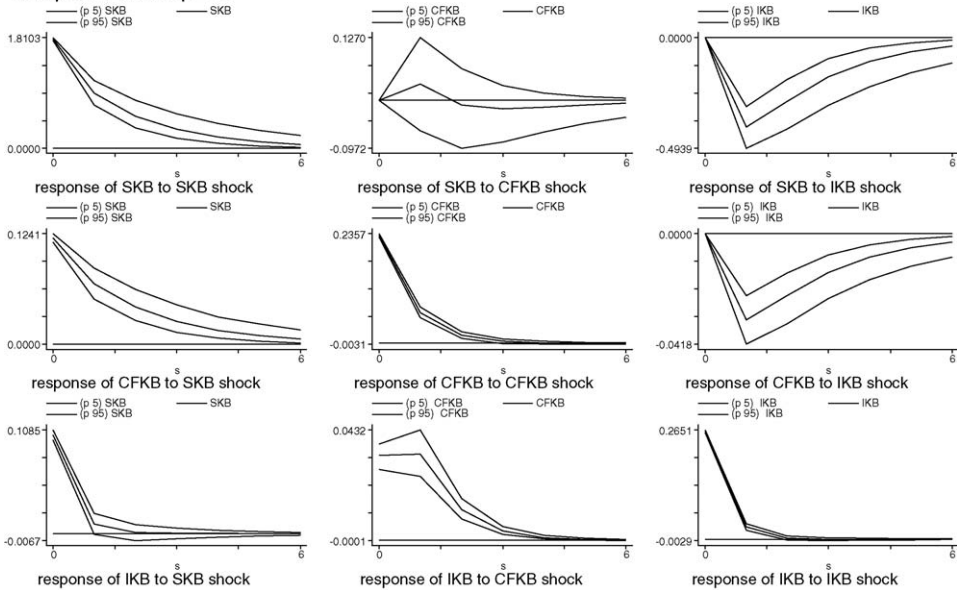
We discuss general results first before proceeding to the ones of our particular interest. We observe that the response of the sales to capital ratio to investment is negative in the estimated coefficients and impulse responses. This is expected as sales to capital is our proxy for marginal product of capital. A shock to investment increases the capital stock, which moves the firm along the production frontier. With diminishing returns to capital, the marginal product will decrease. A similar pattern is observed in the response of TOBINQ to investment shock (however, it is only significant in the 'high' development sample, suggesting that in less developed countries TOBINQ is less responsive to firms investment). Since TOBINQ is a measure of investment opportunities, an investment shock implies that the available opportunities have been acted upon and therefore the market to book value drops.

The investment shows an expected positive response to a shock in the sales to capital ratio (i.e. marginal profitability), both in the estimated coefficients and in the impulse responses. A similar

¹³ As mentioned above, we repeated our analysis with a sample including only up to 150 largest firms in each country using a rank-based approach described in the data section. We also considered models with different proxies for cash flow and different normalizations (for example, scaling by total assets instead of capital stock). Finally, we used different cutoff points—such as 50 or 100 firms. The results were similar to the ones reported and are available on request.

Impulse-responses for 1 lag VAR of SKB CFKB IKB

Sample : if develop==0

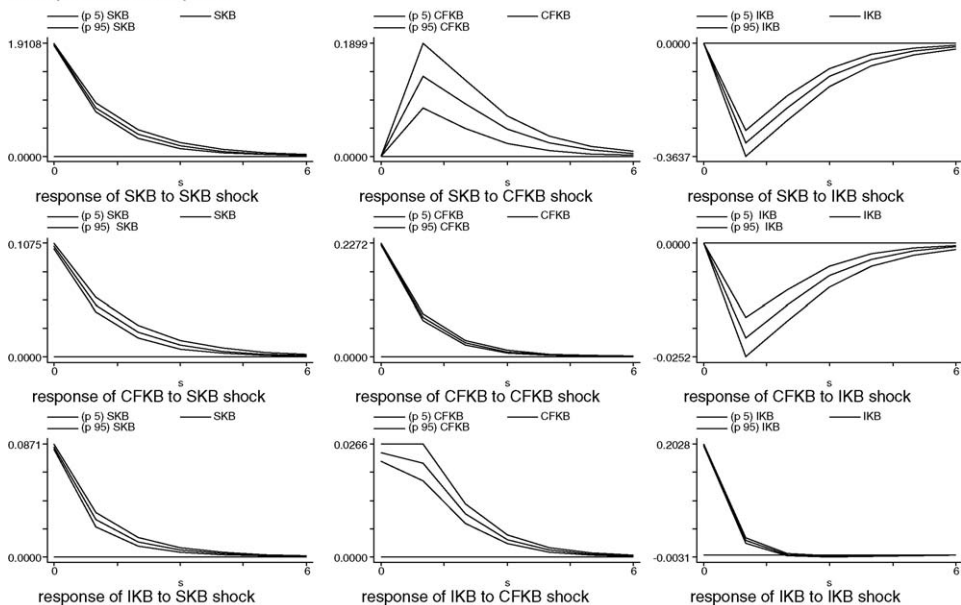


Errors are 5% on each side generated by Monte-Carlo with 1000 reps

Fig. 1. Impulse responses for low financial development sample (model with three variables: SKB, CFKB, IKB).

Impulse-responses for 1 lag VAR of SKB CFKB IKB

Sample : if develop==1



Errors are 5% on each side generated by Monte-Carlo with 1000 reps

Fig. 2. Impulse responses for high financial development sample (model with three variables: SKB, CFKB, IKB). Errors are 5% on each side generated by Monte-Carlo with 1000 reps.

Impulse-responses for 1 lag VAR of SKB CFKB IKB
Sample : Difference of Undeveloped - Developed

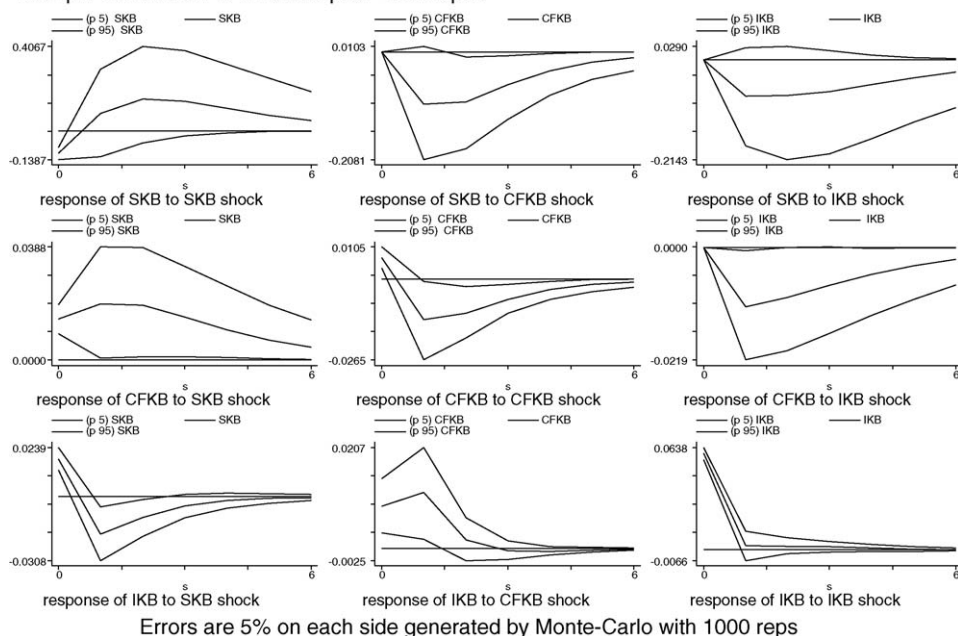


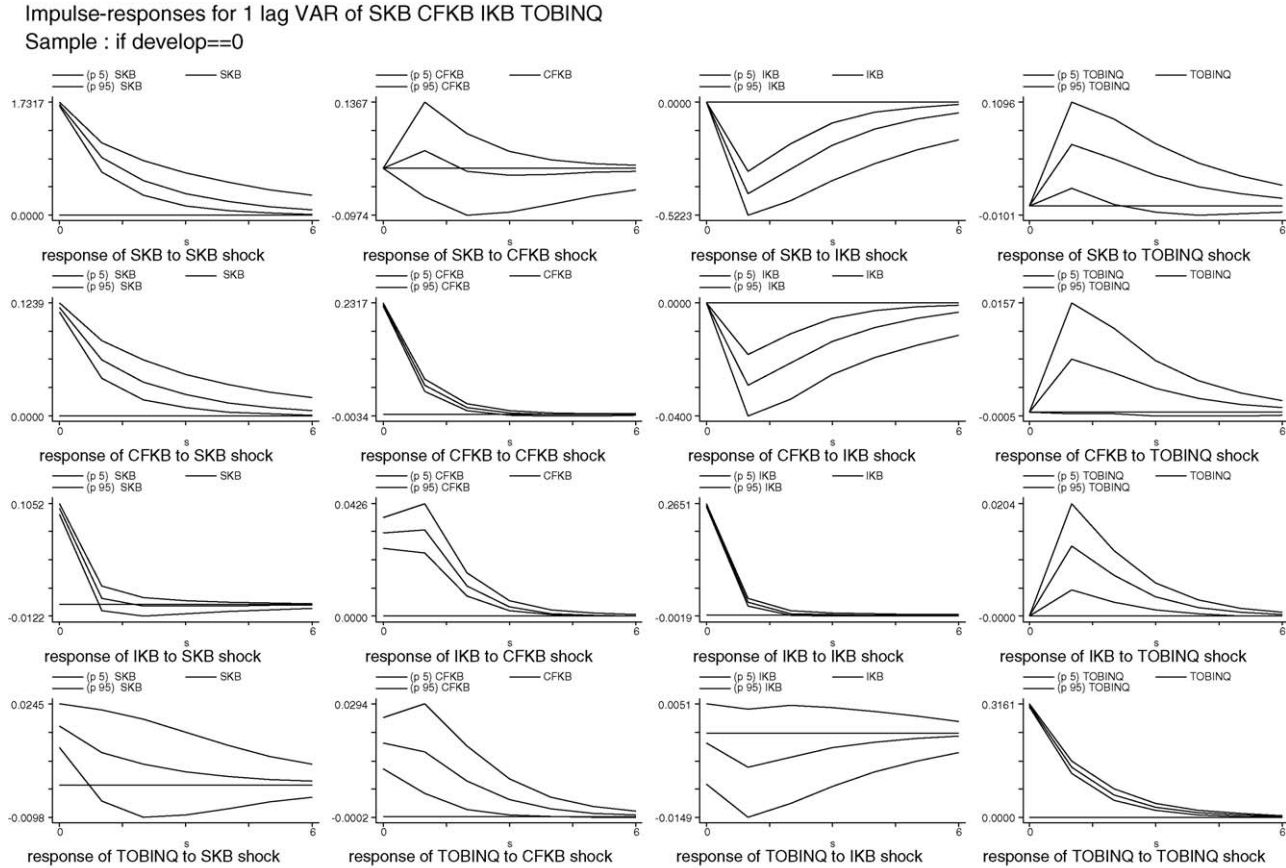
Fig. 3. Difference in impulse responses (low–high) for the model with three variables: SKB, CFKB, IKB.

pattern is observed in the response to TOBINQ in the model with four variables, confirming the prediction that TOBINQ captures a part of the “fundamental” shock.

Cash flow increases in response to a sales shock (higher revenues imply more cash), while it decreases in response to investment. Cash flow has no significant effect on sales to capital (and there is no reason to expect such an effect). These patterns are very similar across our two groups of countries.

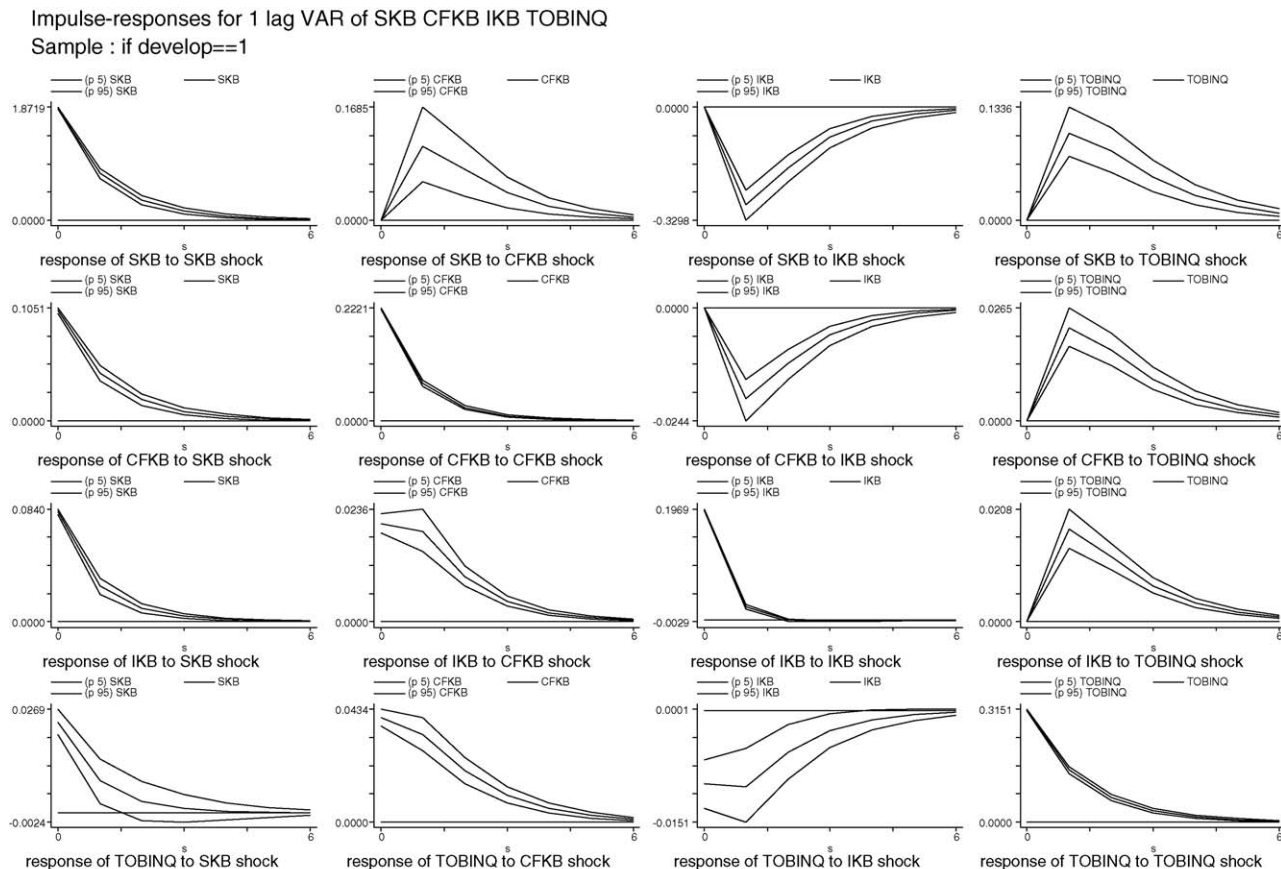
The result of our particular interest is the response of investment to our financial variable—the cash flow. We observe that the impact of the lagged cash flow on the level of investment is much larger in countries with ‘low’ financial development than it is in countries with ‘high’ financial development. This difference is most pronounced in the model with TOBINQ in which the cash flow coefficient is twice as large in the ‘low development’ sample than in the ‘high development’ one (i.e. 0.123 compared with 0.065—see second column in Table 5), and this difference is statistically significant. In the model with three variables, the coefficient is one and a half times larger in the ‘low’ sample than in the ‘high’ one.

The panels representing the impulse response of investment to a one standard deviation shock in cash flow clearly show a positive impact. We also notice that this response has a larger impact on the value of the investment for firms in the ‘low’ sample. This can be seen most clearly in Fig. 3 that reports the difference in two samples responses (i.e. ‘low’ minus ‘high’). The difference between two impulse responses is significant at better than 5% (i.e. the 5% lower band is quite above the zero line). The same is true when we use a model which includes TOBINQ (see Fig. 6). These results suggest that financial factors have larger effect on investment in countries with lower levels of financial development.



Errors are 5% on each side generated by Monte-Carlo with 1000 reps

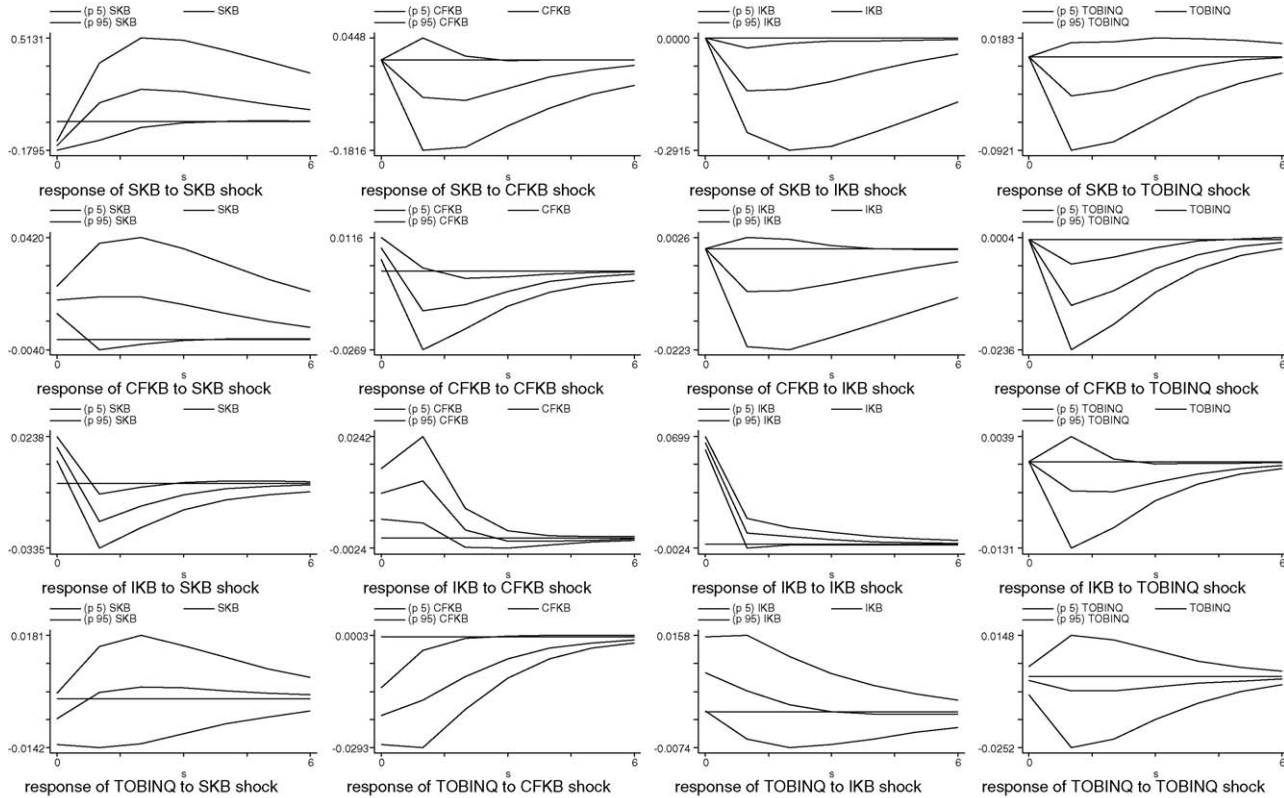
Fig. 4. Impulse responses for low financial development sample (model with four variables: SKB, CFKB, IKB, TOBINQ). Errors are 5% on each side generated by Monte-Carlo with 1000 reps.



Errors are 5% on each side generated by Monte-Carlo with 1000 reps

Fig. 5. Impulse responses for high financial development sample (model with four variables: SKB, CFKB, IKB, TOBINQ).

Impulse-responses for 1 lag VAR of SKB CFKB IKB TOBINQ
Sample : Difference of Undeveloped - Developed



Errors are 5% on each side generated by Monte-Carlo with 1000 reps

Fig. 6. Difference in impulse responses (low–high) for the model with four variables: SKB, CFKB, IKB, TOBINQ. Errors are 5% on each side generated by Monte-Carlo with 1000 reps.

Table 6
Variance decompositions

	SKB	CFKB	IKB	
Panel A: Low financial development sample				
SKB	0.940	0.000	0.061	
CFKB	0.263	0.713	0.024	
IKB	0.131	0.029	0.840	
Panel B: High financial development sample				
SKB	0.959	0.006	0.035	
CFKB	0.194	<u>0.796</u>	0.010	
IKB	0.162	0.024	0.814	
	SKB	CFKB	IKB	TOBINQ
Panel C: Low financial development sample				
SKB	0.923	0.0	0.075	0.002
CFKB	0.260	0.718	0.021	0.001
IKB	0.123	<u>0.027</u>	0.847	0.003
TOBINQ	0.004	0.006	0.001	0.989
Panel D: High financial development sample				
SKB	0.963	0.005	0.028	0.005
CFKB	0.188	0.791	0.008	0.013
IKB	0.158	<u>0.019</u>	0.813	0.010
TOBINQ	0.005	0.025	0.002	0.968

Percent of variation in the row variable (10 periods ahead) explained by column variable.

Panels A and B refer to the model with three variables (SKB, CFKB, IKB); panels C and D to the model with four variables (SKB, CFKB, IKB, TOBINQ).

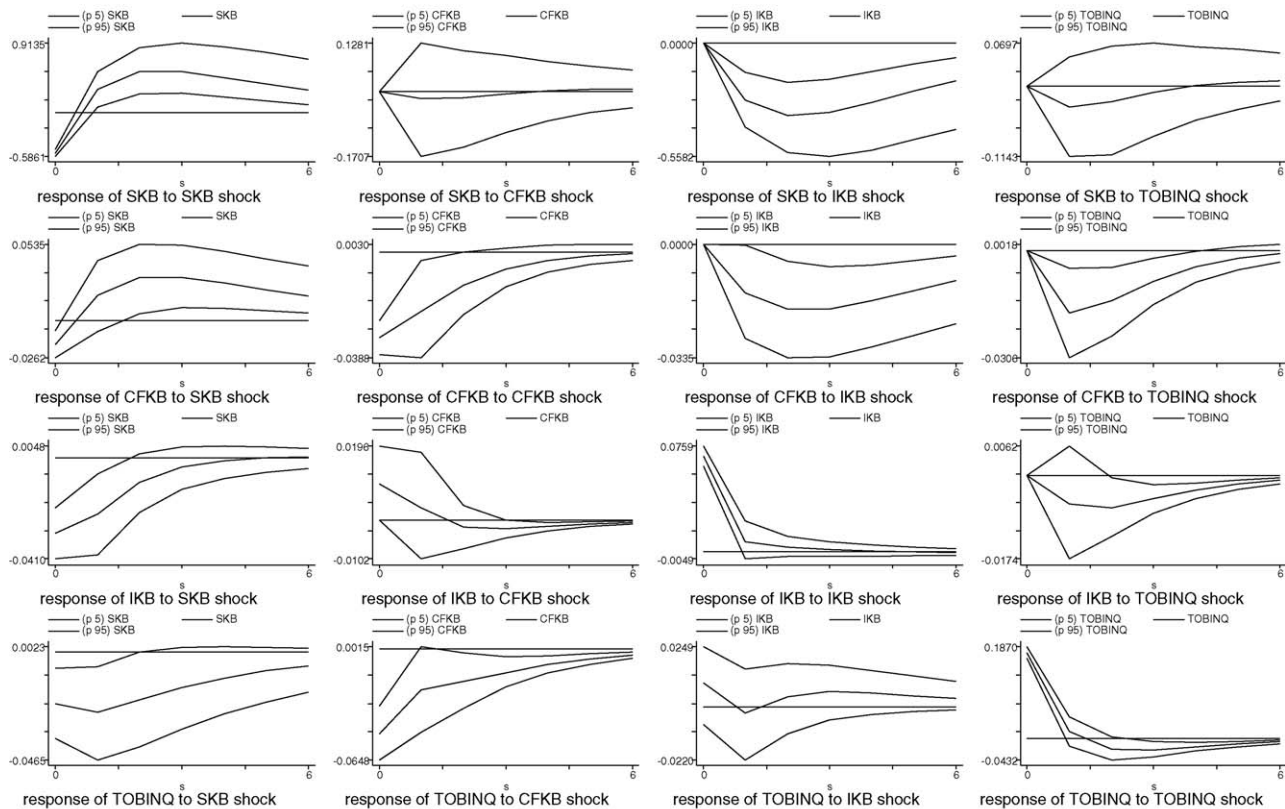
The variance decompositions for the different models, presented in Table 6, are in line with these results. Cash flow explains more of the investment variation 10 periods ahead in the sub-sample of countries characterised by low financial development. However, the magnitude of the effect is rather small, cash flow only explains about 2.7% of total variation in investment in low development sample and about 1.9% in high development sample (using the model with four variables).

The orthogonalization of the VAR residuals (discussed in Section 2) allows us to isolate the response of investment to ‘financial’ factors (cash flow) from the response to ‘fundamental’ factors (marginal productivity of capital). We interpret our results as evidence that the response of investment to ‘financial’ factors (and, therefore, the intensity of financing constraints) is significantly larger in countries with less developed financial markets.

A mirror image result is that ‘fundamental’ factors have less effect on investment in countries with low financial development sample. The impulse response of investment to sales to capital is significantly lower in the ‘low’ sample (but only after the contemporaneous response, which, surprisingly, is higher). However, over time the response of investment to sales to capital is significantly lower in the low development sample, as shown by the variance decomposition: sales to capital explains about 12% of variation in the ‘low’ sample and about 16% in the ‘high’ sample (using the model with four variables).¹⁴

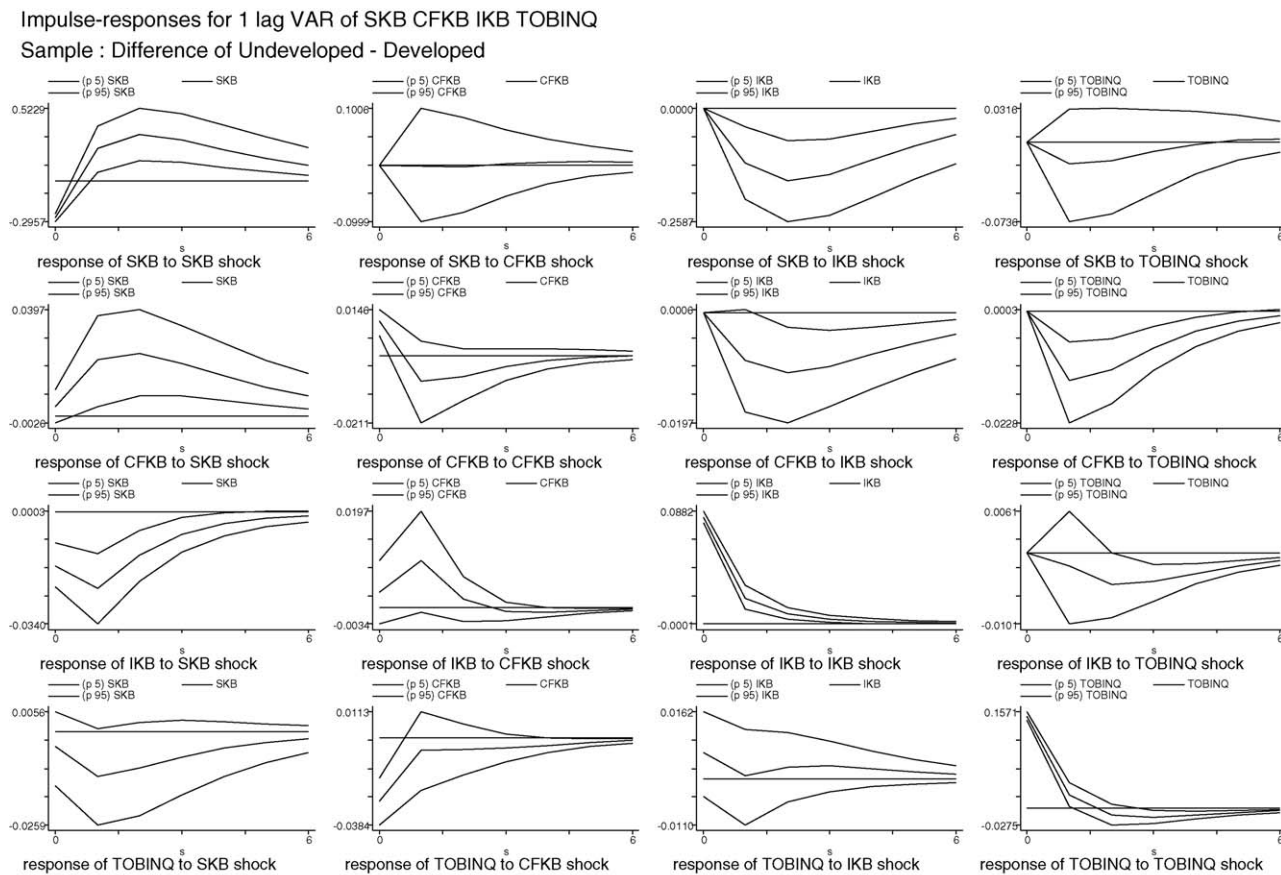
¹⁴ While impulse-responses show that investment responds less to *Tobinq* in the ‘low’ sample, this difference is not significant at 5% level. The percent of variation in investment explained by *Tobinq* is very small (1% in the ‘high’ sample and 0.3% in the ‘low’ sample). Thus, *Tobinq* has a negligible additional explanatory power in the model, which uses sales to capital as a proxy for the marginal product of capital.

Impulse-responses for 1 lag VAR of SKB CFKB IKB TOBINQ
Sample : Difference of Undeveloped - Developed



Errors are 5% on each side generated by Monte-Carlo with 1000 reps

Fig. 7. Difference in impulse responses (low–high) for a model with four variables and development defined over the median of GDP PC.



Errors are 5% on each side generated by Monte-Carlo with 1000 reps

Fig. 8. Difference in impulse responses (low–high) for a model with four variables and development defined as high income (using World Bank classification). Errors are 5% on each side generated by Monte-Carlo with 1000 reps.

To confirm that our main result of a significantly different impact of cash flow shocks on investment is capturing the level of financial development rather than economic development we ran two further tests. First, we split the countries according to their GDP per capita (using the sample median). Second, we use the World Bank classification of countries into different income categories based on GNI per capita, separating high-income countries from the rest. The graphs of the differences between the impulse responses for the model with four variables are shown in Figs. 7 and 8.¹⁵ There was no significant difference in the response of investment to a cash flow shock in either case, corroborating our finding that firms in less financially (rather than less economically) developed countries are more likely to need to rely on internal sources of finance in order to invest.

To conclude, the coefficient estimates, the impulse-response functions, and the variance decompositions resulting from the vector autoregressions support our claim that in the presence of financing constraints, which are more stringent in countries that do not have a well-developed financial system, the availability of liquid assets affects firms' investment decisions. Financing constraints manifest not only in higher response of investment to 'financial' factors but also in lower response of investment to 'fundamental' factors. Both of these effects imply that financial under-development adversely affects the dynamic investment behavior, thus leading to inefficient allocation of capital.

4. Conclusions

This paper uses a VAR approach to analyze the relationship between firms' investment decisions and the level of financial development in their hosting countries. It shows that the availability of internal funds is more important in explaining investment in countries with less developed financial systems. More specifically, the impact of a positive shock to cash flow on investment is significantly higher in countries with a 'low' level of financial development than in countries with a 'high' level of financial development. Symmetrically, we find that positive shock to marginal productivity has less impact on investment of firms in countries with low level of financial development.

Our paper complements earlier work in finance and growth literature by Demircuc-Kunt and Levine (2001a, b), Beck and Levine (2002, 2004) and others. While this literature did not find links between financial structure (i.e. market-based versus bank-based systems) and growth, it found strong casual links between overall financial development and growth. Our results contribute to this literature by showing that in countries with underdeveloped financial markets firms have inefficient allocation of capital and that they exhibit slower growth rates.

We believe our paper contributes to the literature on financial constraints and investment decisions by adopting a simple approach to separate the fundamental from the financial factors that influence the level of investment. The analysis of the impulse-response functions obtained from a reduced-form VAR model allowed us to obtain clear evidence of the importance of financial development for capital investment without having to impose the strong structural assumptions necessary in the q -theory or the Euler-equation approaches. In conclusion, while supporting earlier results, our paper also presents a simple methodology that could be used to further explore the differences in dynamic firm behaviour across different countries.

¹⁵ We used the model with three variable as well and obtained similar results.

Acknowledgment

The paper was completed while Lea Zicchino was at Columbia University, New York. The views presented here are the author's own and not necessarily those of the World Bank, its member countries or the Bank of England

Appendix A. Sample selection

All countries in the *Worldscope* database (May 1999 Global Researcher CD) with at least 30 firms and at least 100 firm-year observations are included in the sample (in addition we include Venezuela (VE), though it has only 80 observations); former socialist economies are excluded. This results in a sample of 40 countries. The sample does not include firms for which the primary industry is either financial (one digit SIC code of 6) or service (one digit SIC codes of 7 and above).

In addition we delete the following (see [Table 2](#) for variable definitions):

- All firms with 3 or less years of coverage;
- All firm-years with missing CAPEX, Sales, Netpeq, Compnumb or Cash;
- Outliers for the distributions of SKB, IKB, CFK, and TOBINQ.

The resulting data set has about 54,000 observations. The number of observations by country is given in [Table 1](#).

References

- Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error component models. *Journal of Econometrics*, 68, 29–51.
- Beck, T., Demirguc-Kunt, A., Levine, R., & Maksimovic, V. (2001). Financial structure and economic development: Firm, industry, and country evidence. In A. Demirguc-Kunt & R. RossLevine (Eds.), *Financial structure and economic growth: A cross-country comparison of banks, markets and development*. Cambridge, MA: MIT Press.
- Beck, T., & Levine, R. (2002). Industry growth and capital allocation: Does having a market- or bank-based system matter? *Journal of Financial Economics*, 64(2), 147–180.
- Beck, T., & Levine, R. (2004). Stock markets, banks, and growth: Panel evidence. *Journal of Banking and Finance*, 28, 423–442.
- Bernanke, B., & Gertler, M. (1989). Agency costs, net worth, and business fluctuations. *American Economic Review*, 79(1), 14–31.
- Demirguc-Kunt, A., & Levine, R. (1996). Stock market development and financial intermediaries: Stylized facts. *World Bank Economic Review*, 10, 291–321.
- Demirguc-Kunt, A., & Levine, R. (2001a). *Financial structure and economic growth: A cross-country comparison of banks, markets and development*. Cambridge, MA: MIT Press.
- Demirguc-Kunt, A., & Levine, R. (2001b). Bank-based and market-based financial systems: Cross-country comparisons. In A. Demirguc-Kunt & R. Levine (Eds.), *Financial structure and economic growth: A cross-country comparison of banks, markets and development*. Cambridge, MA: MIT Press.
- Demirguc-Kunt, A., & Maksimovic, V. (1998). Law, finance, and firm growth. *Journal of Finance*, 8(6), 2107–2137.
- Erickson, T., & Whited, T. (2000). Measurement Error and the relationship between investment and q . *Journal of Political Economy*, 108, 1027–1057.
- Fazzari, S., Hubbard, G., & Peterson, B. (1988). Financing constraints and corporate investment. *Brookings Papers on Economic Activity*, 78(2), 141–195.
- Gallegati, M., & Stanca, L. (1999). The dynamic relation between financial positions and investment: Evidence from company account data. *Industrial and Corporate Change*, 8(3), 551–572.

- Gilchrist, S., & Himmelberg, C. (1995). Evidence on the role of cash flow for investment. *Journal of Monetary Economics*, 36, 541–572.
- Gilchrist, S. Himmelberg C. (1998). Investment, fundamentals and finance. NBER Working Paper 6652.
- Hamilton, J. (1994). *Time series analysis*. Princeton University Press.
- Hayashi, F. (1982). Tobin's marginal q and average q : A neoclassical interpretation. *Econometrica*, 50(1), 213–224.
- Hubbard, G. (1998). Capital-market imperfections and investment. *Journal of Economic Literature*, 36(1), 193–225.
- Love, I. (2003). Financial development and financing constraints: International evidence from the structural investment model. *Review of Financial Studies*, 16, 765–791.
- Powell, A., Ratha, D., & Mohapatra, S. (2002). Capital inflows and outflows: On their determinants and consequences for developing countries. *Mimeograph*.
- Rajan, R. G., & Zingales, L. (1998). Financial development and growth. *American Economic Review*, 88(3), 559–586.
- Schiantarelli, F. (1996). Financial constraints and investment: Methodological issues and international evidence. *Oxford Review of Economic Policy*, 70–89.
- Wurgler, J. (2000). Financial markets and allocation of capital. *Journal of Financial Economics*, 58(1–2), 187–214.
- Zicchino, L. (2001). *Endogenous financial structure and business fluctuations in an economy with moral hazard*. Mimeo-graph: Columbia University.

The determinants and implications of corporate cash holdings[☆]

Tim Opler^a, Lee Pinkowitz^a, René Stulz^{a,*}, Rohan Williamson^b

^a*Fisher College of Business, The Ohio State University, Columbus, OH 43210, USA*

^b*McDonough School of Business, Georgetown University, Washington, DC 20057, USA*

Received 19 September 1997; received in revised form 1 June 1998

Abstract

We examine the determinants and implications of holdings of cash and marketable securities by publicly traded U.S. firms in the 1971–1994 period. In time-series and cross-section tests, we find evidence supportive of a static tradeoff model of cash holdings. In particular, firms with strong growth opportunities and riskier cash flows hold relatively high ratios of cash to total non-cash assets. Firms that have the greatest access to the capital markets, such as large firms and those with high credit ratings, tend to hold lower ratios of cash to total non-cash assets. At the same time, however, we find evidence that firms that do well tend to accumulate more cash than predicted by the static tradeoff model where managers maximize shareholder wealth. There is little evidence that excess cash has a large short-run impact on capital expenditures, acquisition spending, and payouts to shareholders. The main reason that firms experience large changes in excess cash is the occurrence of operating losses. © 1999 Elsevier Science S.A. All rights reserved.

* Corresponding author. Tel.: 614-292-1970; fax: 617-292-2359.

E-mail address: stulz@cob.ohio-state.edu (R. Stulz)

[☆]We thank participants at presentations at Dartmouth College, New York University, The Ohio State University, University of Florida, University of Lausanne, University of Maryland, The Wharton School, the Financial Management Association meetings in Hawaii, the American Finance Association meetings in Chicago, the NBER corporate finance meeting, Harry DeAngelo, Eugene Fama, Jarrad Harford, Laurie Hodrick, Glenn Hubbard, Anil Kashyap, Fred Schlingemann, Clifford Smith, Bill Schwert (the editor), Deon Strickland, Ralph Walkling, and, especially, Cathy Schrand, David Scharfstein, and the referee, Stewart Myers, for useful comments.

1. Introduction

On February 8, 1996 Chrysler Corporation's Chairman Robert J. Eaton and investor Kirk Kerkorian agreed to a 5-year standstill agreement, in which Kerkorian would cease attempts to take over Chrysler. An important element of the agreement was a commitment from Chrysler that liquid assets, defined as cash and marketable securities, in excess of a \$7.5 billion target be returned to shareholders in the form of share repurchases or dividends.

The Chrysler/Kerkorian story raises questions that have gone largely unexamined in the finance literature. Is there an optimal level of liquid asset holdings on a corporate balance sheet? And, if so, is the relatively large amount of liquid assets held by firms like Chrysler justified? This question is particularly relevant. The S&P 500 corporations reported a total of \$716 billion in cash and marketable securities on their balance sheets as of fiscal year 1994. The largest non-financial holders of liquid assets were Ford (\$13.8 billion), General Motors (\$10.7 billion), and IBM (\$10.5 billion).

Management that maximizes shareholder wealth should set the firm's cash holdings at a level such that the marginal benefit of cash holdings equals the marginal cost of those holdings. The cost of holding liquid assets includes the lower rate of return of these assets because of a liquidity premium and, possibly, tax disadvantages. There are two main benefits from holding liquid assets. First, the firm saves transaction costs to raise funds and does not have to liquidate assets to make payments. Second, the firm can use the liquid assets to finance its activities and investments if other sources of funding are not available or are excessively costly. Keynes (1934) describes the first benefit as the transaction cost motive for holding cash, and the second one as the precautionary motive. The costs considered in the literature have evolved from brokerage costs, in the classic paper by Miller and Orr (1966), to inefficient investment resulting from insufficient liquidity, emphasized in theoretical models such as Jensen and Meckling (1976), Myers (1977), and Myers and Majluf (1984), as well as in empirical papers that build on Fazzari et al. (1988).

Theories that focus on the tradeoff between the costs and benefits of cash holdings can make it possible to answer the question of whether a firm holds too much cash from the perspective of shareholder wealth maximization. In general, however, managers and shareholders view the costs and benefits of liquid asset holdings differently. Agency theory can therefore explain why firms do not hold the amount of cash that maximizes shareholder wealth, and help to identify firms that are likely to hold too much cash. Managers have a greater preference for cash, because it reduces firm risk and increases their discretion. This greater preference for cash can lead managers to place too much importance on the precautionary motive for holding cash. One would therefore expect firms where agency costs of managerial discretion are more important to hold more liquid assets than would be required to maximize shareholder wealth.

An alternative view to the tradeoff model of cash holdings is that there is no optimal amount of cash. With this view, cash holdings are an irrelevant sideshow. The argument is that nothing changes in a corporation if it has one more dollar of cash financed with one more dollar of debt. Hence, even if one believes that there is an optimal capital structure for a corporation, this optimal capital structure specifies an optimal amount of net debt, which is debt minus cash. As a result, there is no optimal amount of cash, because cash is simply negative debt. The same reasoning holds with the pecking order or financing hierarchy model. According to the pecking order model, a firm's leverage, defined using net debt, reacts passively to changes in the firm's internal funds. As a firm accumulates internal funds, its leverage falls. The firm avoids issuing equity because adverse selection costs make equity too expensive. As the firm maintains a surplus of internal funds, it accumulates cash and pays back debt when it becomes due. Faced with a deficit of internal funds, the firm decreases cash holdings and eventually raises debt. With this view, changes in internal resources are the driving force for changes in cash holdings, but it is a matter of indifference whether a firm uses the internal resources to accumulate cash or repay debt. A firm that is not constrained in its investment policy simply uses cash flow to increase cash, unless it has debt to repay.

Myers and Majluf (1984) provide a theoretical foundation for the pecking-order model that makes it consistent with shareholder wealth maximization. A challenge that arises with extending the financing hierarchy model to explain cash holdings is that the conditions under which this extension is consistent with shareholder wealth maximization are rather restrictive. As long as there is any cost to holding cash, a firm that simply accumulates cash will at some point have an excessive amount of cash, and shareholders would be better off if the firm used that cash to pay additional dividends or to repurchase shares. If management is reluctant to use cash in this way, for the reasons discussed in Jensen's (1986) free cash flow theory, empirical evidence will support the financing hierarchy view, even though there is an amount of cash that maximizes shareholder wealth.

This paper proceeds in three steps. We first examine simple dynamic models of changes in cash holdings to assess the success of the static trade-off and financing hierarchy views in explaining changes in cash holdings. Though Shyam-Sunder and Myers (1998) demonstrate that the financing hierarchy view is extremely successful at explaining changes in leverage, we find here that the static tradeoff theory of cash holdings cannot be dismissed as irrelevant, and that the theory makes important predictions that find support in the empirical evidence. In our second step, we show that the predictions of the static tradeoff theory for the determinants of cash holdings are empirically relevant. At the same time, some firms hold dramatically more cash than predicted by the static tradeoff theory. In our third step, we investigate these firms in detail to understand how these large cash holdings come about, and what these excessive

holdings imply about the future behavior of these firms. Jensen's free cash flow theory predicts that these firms will increase their investments, rather than return the cash to the shareholders. We find that firms with large amounts of excess cash acquired it through the accumulation of internal funds. Surprisingly, spending on new projects and acquisitions is only slightly higher for firms with excess cash. Firms typically lose excess cash by covering losses, rather than by spending on new projects or making acquisitions. There is little evidence, therefore, that excess cash 'burns a hole in management's pockets'. Further work will be required to find out whether shareholders are made better off by management's hoarding of cash.

Our results build on an extensive, but generally older, literature on corporate liquidity. Chudson (1945), for example, finds that cash-to-assets ratios tend to vary systematically by industry, and tend to be higher among profitable companies. Vogel and Maddala (1967) find that cash balances have been declining over time, and that larger firms tend to have lower cash-to-assets and cash-to-sales ratios. This finding suggests that there are economies of scale in the transaction motive for cash.¹ Baskin (1987) argues that firms may use cash holdings for competitive purposes. He concludes that '[t]he empirical evidence is entirely consistent with the model wherein liquid assets are employed both to signal commitment to retaliate against encroachment and to enable firms to rapidly preempt new opportunities' (Baskin, 1987, p. 319). A paper by John (1993) argues that firms wish to hold greater amounts of cash when they are subject to higher financial distress costs. Using a 1980 sample of 223 large firms, John finds that firms with high market-to-book ratios and low tangible asset ratios tend to hold more cash. This observation is consistent with the financial distress theory if one agrees that a high market-to-book ratio is a proxy for financial distress costs. Finally, in a contemporaneous paper, Harford (1998) explores the relation between a firm's acquisition policy and its liquid asset holdings. He finds that cash rich firms are more likely to make acquisitions, that these acquisitions are more likely to be diversifying acquisitions, and that they are more likely to decrease shareholder wealth. He views his evidence as strongly supportive of free cash flow theory.

The next section of this paper describes our empirical hypotheses. We present our data in Section 3. In Section 4, we report estimates from time-series and cross-sectional regressions. In Section 5, we investigate whether the investment and payout policies of firms with given investment opportunities are related to

¹ A number of early studies considered the question of whether there are economies of scale in holding cash, including Frazer (1964) and Meltzer (1963). Beltz and Frank (1996) provide evidence on these economies of scale that extends to the 1980s. Mulligan (1997) shows that cash balances fall with respect to sales, and that firms located in U.S. counties with higher wages hold more cash. He views his evidence to support the hypothesis that time can substitute for money in the provision of transaction services and to support the presence of economies of scale in cash holdings.

their liquid asset holdings in the short run. Section 6 examines how likely firms are to keep excess cash over a number of years, and examines the characteristics of firms that experience large changes in excess cash. Section 7 summarizes the findings, and suggests future directions for empirical research in this area.

2. Theory and empirical hypotheses

In a world of perfect capital markets, holdings of liquid assets are irrelevant. If cash flow turns out to be unexpectedly low, such that a firm has to raise funds to keep operating and to invest, it can do so at zero cost. Since there is no liquidity premium in such a world, holdings of liquid assets have no opportunity cost. Hence, if a firm borrows money and invests it in liquid assets, shareholder wealth is unchanged.

However, if it is costly for the firm to be short of liquid assets, the firm equates the marginal cost of holding liquid assets to the marginal benefit of holding those assets. Holding an additional dollar of liquid assets reduces the probability of being short of liquid assets, and decreases the cost of being short of cash, under the reasonable assumption that the marginal benefit of liquid assets declines as holdings of liquid assets increase. We define a firm to be short of liquid assets if it has to cut back investment, cut back dividends, or raise funds by selling securities or assets. A firm can make it less likely that it will be short of liquid assets in a particular state of the world by having lower leverage, or by hedging. Consequently, an optimal theory of liquid asset holdings has to address the issue of why it is more efficient for the firm to hold an additional dollar of liquid assets instead of decreasing leverage by some amount, or increasing hedging.

In the remainder of the section, we first address the role of transaction costs as a determinant of cash holdings, and then turn to the impact of information asymmetries and agency costs on cash holdings. The section concludes with a discussion of the financing hierarchy model.

2.1. *The transaction costs model*

Keynes' (1936) transaction motive for holding cash arises from the cost of converting cash substitutes into cash. Consider the effect of transaction costs on the irrelevance result within the framework we have just discussed. We now assume that there are costs to buying and selling financial and real assets. In particular, let us assume that there is a cost to raising outside funds that takes the form of a fixed cost, plus a variable cost which is proportional to the amount raised. In this case, a firm short of liquid assets has to raise funds in the capital markets, liquidate existing assets, reduce dividends and investment, renegotiate

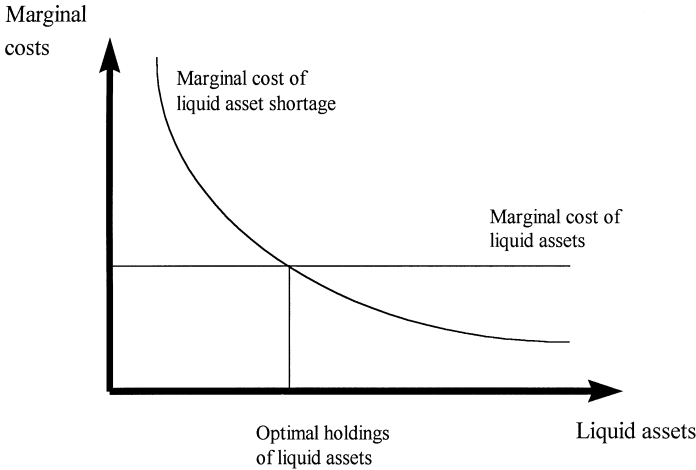


Fig. 1. Optimal holdings of liquid assets. The optimal amount of liquid assets is given by the intersection of the marginal cost of liquid assets curve and the marginal cost of liquid asset shortage curve. The marginal cost of liquid assets curve is non-decreasing while the marginal cost of liquid asset shortage curve is decreasing.

existing financial contracts, or some combination of these actions. Unless the firm has assets that can be liquidated at low cost, it prefers to use the capital markets. However, it is costly to raise funds, regardless of whether the firm does so by selling assets or using the capital markets. The fixed costs of accessing outside markets induce the firm to raise funds infrequently, and to use cash and liquid asset holdings as a buffer. As a result, for a given amount of net debt, there is an optimal amount of cash, and cash is not simply negative debt.

Fig. 1 shows the marginal cost curve of being short of liquid assets, and the marginal cost curve of holding cash. The marginal cost curve of being short of liquid assets is downward sloping and the marginal cost curve of holding liquid assets is assumed to be horizontal. With the transaction costs model, the cost of liquid assets is their lower pecuniary expected return, because part of the benefit from holding liquid assets is that they can be more easily converted into cash. There is no reason to think that this cost varies with the amount of liquid assets held. If the firm has a shortage of liquid assets, it can cope with the shortage by either decreasing investment or dividends, or by raising outside funds through security issuances or asset sales. A greater shortage has greater costs, because addressing a larger shortage involves decreasing investment more or raising more outside funds. For a given amount of liquid assets, an increase in the cost of being short of liquid assets, or an increase in the probability of being short of liquid assets, both shift the marginal cost curve to the right, and increase the firm's holdings of liquid assets.

With the assumptions that lead to Fig. 1, one would expect the marginal cost of being short of funds, and a related increase in holdings of liquid assets to respond to the following variables:²

Magnitude of transaction costs of raising outside funds. One would expect transaction costs to be lower for firms that have already accessed public markets. This expectation means that firms with a debt rating have less liquid assets. Firms could also raise outside funds more easily if they have credit lines outstanding, but credit lines may get canceled precisely when outside funds are the most valuable for a company.

Cost of raising funds through asset sales, dividend cuts, and renegotiation. Shleifer and Vishny (1993) discuss the role of assets sales as a source of financing. A firm with assets on its balance sheet that can be cheaply converted into cash can raise funds at low cost by selling these assets. Hence, firms with mostly firm-specific assets have higher levels of liquid assets. To the extent that diversified firms are more likely than specialized firms to have substantial assets that can be sold, because they can sell non-core segments, diversified firms have lower levels of liquid assets. Also, a firm that currently pays dividends can raise funds at low cost by reducing its dividend payments, in contrast to a firm that does not pay dividends, which has to use the capital markets to raise funds.

Investment opportunities. An increase in the number of profitable investment opportunities means that, if faced with a cash shortage, the firm has to give up better projects.

Cost of hedging instruments. By hedging with financial instruments, a firm can avoid situations where it has to seek funds in the capital markets because of random variation in cash flow. Hence, firms for which hedging is expensive are expected to hold more liquid assets.

Length of the cash conversion cycle. One would expect the cash conversion cycle to be short for firms in multiple product lines and firms with low inventory relative to sales. Consequently, these firms should have less liquid assets.

Cash flow uncertainty. Uncertainty leads to situations in which, at times, the firm has more outlays than expected. Therefore, one would expect firms with greater cash flow uncertainty to hold more cash.

Absence of economies of scale. Simple transaction costs models, such as Miller and Orr (1966), suggest that there are economies of scale in cash management.

In a world with significant transaction costs, one would expect assets that can be exchanged for cash, while incurring lower transaction costs, to have a lower

² In a contemporaneous paper, Kim et al. (1998) model the transaction costs motive to hold cash and make some similar predictions.

return to reflect this benefit (see Amihud and Mendelson, 1986). This expectation means that there is now a cost to holding liquid assets. We call this the liquidity premium. Note that this liquidity premium cannot be a risk premium. If liquid assets simply earn less because they have different risk characteristics, holding them does not entail a cost. One would expect this cost to be highest for cash, and to decrease for assets that are poor substitutes for cash. Consequently, a firm's liquid assets have an opportunity cost. For liquid assets held in the form of demand deposits, the opportunity cost increases with interest rates. To the extent that cash substitutes are deposited in short-maturity instruments, holding these cash substitutes becomes more expensive when the liquidity premium component of the term structure rises.

So far, our discussion has omitted taxes. Taxes increase the cost of holding liquid assets. The reason is that the interest income from liquid assets is taxed twice. It is taxed first at the corporate level, and then taxed again as it generates income for the shareholders. Consider the case of a shareholder that pays no capital gains taxes. Such a shareholder would prefer the firm to use excess liquid asset holdings to repurchase shares. By taking this action, the marginal tax rate on the liquid asset holdings for that investor would fall by the corporation's tax rate. This relation means that the cost of holding liquid assets increases with the firm's marginal tax rate.

In summary, the transaction costs model implies that liquid assets increase with (1) the volatility of cash flow divided by total assets, and (2) the length of the cash conversion cycle. The model also implies that liquid asset holdings decrease (1) with interest rates and the slope of the term structure, (2) with the cost of raising debt, (3) with the ease of selling assets, (4) with the cost of hedging risk, and (5) with the size of a firm's dividend. The inclusion of taxes has the additional implication that the cost of holding liquid assets increases with the firm's marginal tax rate.

2.2. *Information asymmetries, agency costs of debt, and liquid asset holdings*

We now extend the analysis to allow for information asymmetries and agency costs of debt. In this case, cash flow shortfalls might prevent a firm from investing in profitable projects if the firm does not have liquid assets, so that firms can find it profitable to hold cash to mitigate costs of financial distress. We call this motivation to hold liquid assets the precautionary motive for holding cash.

First, consider the role of information asymmetries. Information asymmetries make it harder to raise outside funds. Outsiders want to make sure that the securities they purchase are not overpriced, and consequently discount them appropriately. Since outsiders know less than management, their discounting may underprice the securities, given management's information (see Myers and Majluf, 1984). In fact, outsiders may require a discount that is large enough that

management may find it more profitable to not sell the securities, and reduce investment instead. Since information asymmetries make outside funds more expensive, the model with information asymmetries makes many predictions that are similar to the model with transaction costs discussed earlier. However, the model with information asymmetries provides an explicit reason why outside funds would be expensive, possibly prohibitively so. This model predicts that the cost of raising outside funds increases as securities sold are more information sensitive, and as information asymmetries are more important. It is important to note that information asymmetries can change over time, so that a firm for which these asymmetries are unimportant at one point in time may later find itself in a situation where these asymmetries become crucial. Myers and Majluf (1984) argue that shifting information asymmetries make it valuable to build up slack in periods when information asymmetries are small. Antunovich (1996) further argues that firms with higher information asymmetries will have a greater dispersion of slack, since these firms have more difficulty accessing capital markets. When information asymmetries are important, a cash flow shortfall forces firms to contract investment, and hence involves greater costs. One would expect this cost of financial distress to be larger for firms with high research and development (R&D) expenses, since R&D expenses are a form of investment where information asymmetries are most important (see Opler and Titman, 1994). Consequently, we would expect that firms with higher R&D expenses will hold more liquid assets.

We now turn to the role of agency costs of debt. These agency costs arise when the interests of the shareholders differ from the interests of the debtholders, and, possibly, when interests differ among various classes of debtholders. Because of these costs, highly leveraged firms find it difficult and expensive to raise additional funds. These firms also sometimes find it impossible to renegotiate existing debt agreements to prevent default and bankruptcy. Such firms have high incentives to engage in asset substitution, as argued by Jensen and Meckling (1976), so that debt will be expensive, both in terms of the required promised yield, and in terms of the covenants attached to the debt. They are also likely to face the underinvestment problem emphasized by Myers (1977), namely, that raising funds to invest may benefit debtholders but not shareholders, so that shareholders prefer not to invest, even though the firm has valuable projects.

Firms want to avoid situations where the agency costs of debt are so high that they cannot raise funds to finance their activities and invest in valuable projects. Obviously, one way to do so is to choose a low level of leverage. However, one would expect firms with valuable investment opportunities, for which the cost of raising additional outside funds is high, or even prohibitive, to hold more liquid assets, since the cost of being short of funds is higher. The market-to-book ratio is often used as a proxy for investment opportunities (see

Smith and Watts, 1992; Jung et al., 1996). Holding the degree of information asymmetry between managers and investors constant, one would expect firms with high market-to-book ratios to hold more cash, since the costs they incur if their financial condition worsens are higher. The problem is that such firms invest a lot, so that if investment expenditures occur discretely, they hold more cash, on average, in order to pay for investment expenditures. Hence, one would expect liquid assets to increase with the market-to-book ratio, controlling for the level of investment expenditures.

2.3. *Agency costs of managerial discretion*

In the presence of agency costs of managerial discretion, management may hold cash to pursue its own objectives at shareholder expense. First, management may hold excess cash simply because it is risk averse. More entrenched management would therefore be more likely to hold excess cash because it can avoid market discipline. Hence, one would expect firms with anti-takeover amendments to be more likely to hold excess cash. Second, management may accumulate cash to have more flexibility to pursue its own objectives. Cash is like free cash flow. Cash allows management to make investments that the capital markets would not be willing to finance. In this sense, cash is not negative debt for management. While management can spend the cash whenever it wants to, it may not be able to raise debt whenever it wants to. By enabling management to avoid the discipline of capital markets, investing in cash can therefore have an adverse effect on firm value. To put it another way, increasing a firm's holdings of liquid assets by one dollar may increase firm value by less than one dollar. The possibility that management could be using cash for its own objectives raises the costs of outside funds, because outsiders do not know whether management is raising cash to increase firm value or to pursue its own objectives. Third, management may accumulate cash because it does not want to make payouts to shareholders, and wants to keep funds within the firm. Having the cash, however, management must find ways to spend it, and hence chooses poor projects when good projects are not available. In general, the agency costs of managerial discretion are less important, and may be trivial for firms with valuable investment opportunities, because the objectives of management and shareholders are more likely to coincide.

When is it more likely that management will not be disciplined, so that it can afford to hold excess cash to pursue its own objectives? We hypothesize four conditions that increase the likelihood of holding excess cash. First, we expect that firms will hold excess cash where outside shareholders are highly dispersed. As argued by Shleifer and Vishny (1986), the existence of large independent shareholders makes a takeover or a proxy contest, or both, easier. Second, we expect large firms to hold excess cash. Firm size is a takeover deterrent. A larger target requires more resources to be husbanded by the bidder, and a large firm

can more easily use the political arena to its advantage. Third, we expect firms with low debt to hold excess cash. By having low debt, the firm is less subject to monitoring by the capital markets. Fourth, firms that are protected from the market for corporate control through anti-takeover charter amendments will also hold excess cash. These amendments make it less likely that the firm becomes a takeover target.

For entrenched management, accumulating liquid assets can be a double-edged sword. Holding excess cash makes it easier for management to remain independent from the capital markets, and to pursue its investment policies. At the same time, it increases the gain to a bidder from taking over the firm, since the bidder gains control of liquid assets that can help finance the acquisition.

To the extent that agency costs of managerial discretion are higher for low market-to-book firms than for high market-to-book firms, as argued in Stulz (1990), one expects low market-to-book firms with entrenched management to have excess liquid assets. To the extent that low market-to-book firms have poor investment opportunities, and management holds liquid assets to facilitate an investment program that it would find difficult to finance through the capital markets, one would expect low market-to-book firms with more liquid assets to invest more.

Management's holdings of shares help align its interests with those of shareholders. At the same time, however, these holdings protect management against outside pressures, and may make management more risk-averse (see Stulz, 1988). If holding cash is costly and management tends to hold more cash than is optimal from the perspective of maximizing shareholder wealth, then one would expect cash holdings to fall with managerial ownership. However, to the extent that managerial ownership makes management more risk averse, then one would expect cash holdings to increase with managerial ownership.

2.4. *The financing hierarchy theory*

Consider now the alternative hypothesis that there is no optimal amount of cash. For that to be the case, firms can issue securities at low cost to raise cash whenever they have insufficient cash to finance their plans. It may be that a firm has an optimal amount of net debt, but it is then a matter of indifference for the firm whether it has high cash holdings and high debt, or low cash holdings and low debt, as long as it has the optimal amount of net debt. However, there might not be an optimal amount of cash, because there is no optimal amount of net debt. This result is the case with the financing hierarchy model. Firms find equity expensive because of information asymmetries, so they do not raise funds in the form of equity under normal circumstances. They sell debt when they do not have sufficient resources, and they can do so. If they have sufficient resources

to invest in the profitable projects available, they repay debt that becomes due, and accumulate liquid assets otherwise. With this hypothesis, liquid assets rise and fall with the fortunes of the firm. If holding cash has no costs for the shareholders, there is no reason for them to object if the firm has large amounts of liquid assets at times.

The distinction between the financing hierarchy model and the static tradeoff model is not as clear-cut as one might want. The distinction becomes blurry as the cost of external capital is allowed to play more of a role in the financing hierarchy model. We will stick to a narrow view of the financing hierarchy model, according to which debt and cash increase mechanically as the firm has more funds available. Even though we focus on an extreme version of the financing hierarchy model, some of its empirical predictions are similar to those of the static tradeoff model, so that it is difficult to distinguish empirically between the two models. In the financing hierarchy model, firms with high cash flow will have more cash. However, as argued by Shyam-Sunder and Myers (1998), it is often the case that firms with high cash flow also have a high market-to-book ratio. This condition occurs because these firms can be expected to be profitable in the future. Hence, discovering that firms with a high market-to-book ratio have more cash is not inconsistent with the financing hierarchy model. With this model, firms that pay more dividends should have lower cash. Everything else equal, however, a firm that invests more should have fewer internal resources, and hence would accumulate less cash. In contrast, with the static tradeoff theory, firms with more capital expenditures have more liquid assets. The same argument applies to R&D investments. There seems to be no reason why the variables emphasized by the agency theory arguments, namely the proxies for managerial entrenchment, would have implications for cash holdings in the financing hierarchy model. Finally, with the financing hierarchy view, firms that are larger presumably have been more successful, and hence should have more cash, after controlling for investment. The static tradeoff model argues that there are economies of scale in liquid assets, so that one would expect firm size to have a negative impact on cash holdings.

3. Data

To investigate our hypotheses on the determinants of cash holdings, we construct a sample of firms for our empirical tests by merging the Compustat annual industrial and full coverage files with the research industrial file for the 1952–1994 period. These data include survivors and non-survivors that appeared on Compustat at any time in the sample period. We exclude financial firms, with Standard Industrial Classification (SIC) codes between 6000 and 6999, because their business involves inventories of marketable securities that

are included in cash, and because of their need to meet statutory capital requirements. We also exclude utilities, because their cash holdings can be subject to regulatory supervision in a number of states. We exclude firms with nonpositive sales for the years in which they have nonpositive sales. Finally, we exclude American Depository Receipts (ADRs), and firms designated as pre-FASB. We present regressions predicting cash and the persistence of cash holdings using the entire dataset. We also present a separate regression analysis of cash holdings in 1994 for the simple reason that data are available to us for the governance structure and risk management activities of firms for that year. Insider share ownership is measured as the fraction of shares outstanding held by officers and directors, as reported by Compact Disclosure. Firm diversification is measured using the Compustat segment tapes.

3.1. Measure of liquid asset holdings

We measure liquid asset holdings as the ratio of cash and marketable securities (Compustat item #1) to total assets (Compustat item #6) minus cash and marketable securities. We deflate liquid asset holdings by the book value of total assets, net of liquid assets, which we call net assets hereafter, with the view that a firm's ability to generate future profits is a function of its assets in place. While not reported in this paper, we also measure liquidity using the cash-to-sales ratio. This alternative measure does not affect our main conclusions in a material way.

We measure the likelihood that a firm will have positive net present value (NPV) projects in the future by using the ratio of the market value of a firm's assets to the book value of its assets. Since the book value of assets does not include future growth options, we would expect the ratio of the market value of the firm, relative to the book value, to be higher when a firm has a high preponderance of growth options. A variety of past papers find that the market-to-book ratio is an important determinant of corporate financing choices thought to depend on a firm's portfolio of growth options (see, Smith and Watts, 1992; Jung et al., 1996; Barclay and Smith, 1995).

We allow for possible effects of regulation by using a dummy variable for industries that are, or have been, subject to entry and price regulation. This variable is identical to that employed by Barclay and Smith (1995). Regulated industries include railroads (SIC code 4011) through 1980, trucking (SIC codes 4210, 4213) through 1980, airlines (SIC code 4512) through 1978, and telecommunications (SIC codes 4812, 4813) through 1982.

We measure firm size as the natural logarithm of the book value of assets in 1994 dollars. We measure leverage using the debt-to-assets ratio defined as (long-term debt + short-term debt)/book value of assets. To distinguish the effects of a firm's dividend payouts, we define a dummy set equal to one in years where a firm pays a dividend. Otherwise, the dummy variable equals zero.

Finally, we measure cash flow as earnings after interest, dividends, and taxes, but before depreciation, divided by net assets.

We measure cash flow riskiness using two measures. First, we use the standard deviation of industry cash flow computed as follows. For each firm, we compute the cash flow standard deviation for the previous 20 years, if available, using Compustat since 1950. We then take the average across the 2-digit SIC code of the standard deviations of firm cash flow (industry sigma). Second, we compute a firm's cash flow standard deviation for 1994 using the previous twenty years of data, if available.

We use the R&D expense-to-sales ratio as a measure of the potential for financial distress costs. Firms that do not report R&D expenses are considered to be firms with no R&D expenses.

Our hypotheses consider the agency costs of managerial discretion. It is difficult to measure the extent of conflict of interest between the managers of a corporation and its shareholders. In theory, the severity of this conflict is affected by a number of hard-to-measure concepts, including the efficiency of the managerial labor market, and the extent of product market discipline (Fama and Jensen, 1983). Nonetheless, there is a large body of literature that suggests that certain types of firms are more likely to suffer from agency conflicts. For example, firms with inside ownership in excess of 5%, but less than 25–40%, appear to trade at somewhat higher market valuations than other firms (Morck et al., 1988; McConnell and Servaes, 1990). We employ a dummy for whether insider ownership of a firm is in the 5–25% range, and a dummy for whether insider ownership is greater than 25%.

Firms may choose to insure themselves against losses by holding liquid assets besides cash, and by having credit lines available. For example, it is common for firms to sell off non-core assets in periods of economic distress (see Lang et al., 1994). It is also becoming increasingly frequent for firms to liquidate receivables through factoring or securitization as a means of raising liquidity. We use net working capital, minus cash, as a measure of liquid asset substitutes. In addition, we employ a count of the number of reported line of business segments to measure whether firms have non-core assets that could be liquidated in periods of economic distress. Unfortunately, we do not have data on credit lines.

Finally, to assess a firm's derivatives usage in 1994, we use the Corporate Risk Management Handbook from Risk Publications for that year. We collect information on whether an S&P 500 corporation uses derivatives, and on the total of the notional amount of the derivatives it reports.

Table 1 describes the main variables used in the study. There is wide variation in the ratio of cash and marketable securities to assets. The median firm has cash equal to approximately 6% of net assets, or total assets less cash. On a dollar basis, the median firm has cash holdings of \$6.28 million, a relatively small amount. This statistic reflects the size distribution of firms in our sample: The median firm in the sample has an asset base of \$90.1 million.

Table 1

Description of variables for the 1971–1994 Compustat sample

Descriptive statistics on key variables for our sample of firm years from the 1971–1994 sample of U.S.-based publicly traded firms. **Assets** in the denominators of variables are calculated as assets less cash and marketable securities. **Real variables** are deflated using the CPI into 1994 dollars. **Truncated cash** to assets is calculated such that, for any cash-to-assets ratio greater than one, it is given a ratio of one. **Size** is defined as the natural logarithm of assets. **The market-to-book ratio** is measured as the book value of assets, less the book value of equity, plus the market value of equity, divided by assets. **Cash flow** is defined as earnings before interest and taxes, but before depreciation and amortization, less interest, taxes, and common dividends. **Net working capital** is calculated without cash. **Payout** to shareholders is the sum of cash dividends over assets and stock repurchases over assets. Industry sigma is a measure of the volatility of an industry's cash flow for a 20-year period. **Industries** are defined by 2-digit SIC codes. **Total leverage** is total debt over total assets. Other variables displayed include measures of research and development (R&D) spending, capital expenditures, and acquisitions. N is the number of non-missing observations in the sample for each variable.

Variable	Mean	25th Percentile	Median	75th Percentile	N
Cash/assets	0.170	0.025	0.065	0.174	87,117
Truncated cash/assets	0.153	0.025	0.065	0.174	87,117
Real size	4.586	3.291	4.504	5.821	87,117
Market-to-book ratio	1.533	0.922	1.172	1.694	87,117
R&D/sales	0.027	0.000	0.000	0.019	87,117
Cash flow/assets	0.037	0.024	0.070	0.113	87,117
Net working capital/assets	0.176	0.029	0.192	0.345	87,117
Capital expenditures/assets	0.090	0.034	0.064	0.115	87,117
Acquisitions/assets	0.011	0.000	0.000	0.000	85,926
Payout to shareholders	0.017	0.000	0.006	0.024	85,095
Industry sigma	0.121	0.056	0.086	0.168	87,117
Total leverage	0.261	0.104	0.239	0.378	87,117

Fig. 2 shows the median cash-to-assets ratio in the 1952–1994 period for firms with real assets in the \$90-to-\$110 million range and in the \$900 million-to-\$1.1 billion range in 1994 dollars, adjusted for inflation using the Consumer Price Index (CPI) series. For small firms, cash holdings decline throughout the 1950s and the 1960s. Part of this trend may be due to firms having a surplus of cash at the end of WWII, and part of this trend may be the result of technological improvements in cash management. There was a strong decline of cash holdings in the second half of the 1960s. The other reason why cash holdings might be higher in the 1950s and early 1960s in our sample is that, since Compustat was started in the 1960s, all of these firms are survivors. Except for the 1950s and early 1960s, there is little evidence of dramatic changes in cash holdings over time. For small and large firms, there is little evidence of secular changes in cash to assets since the 1960s.

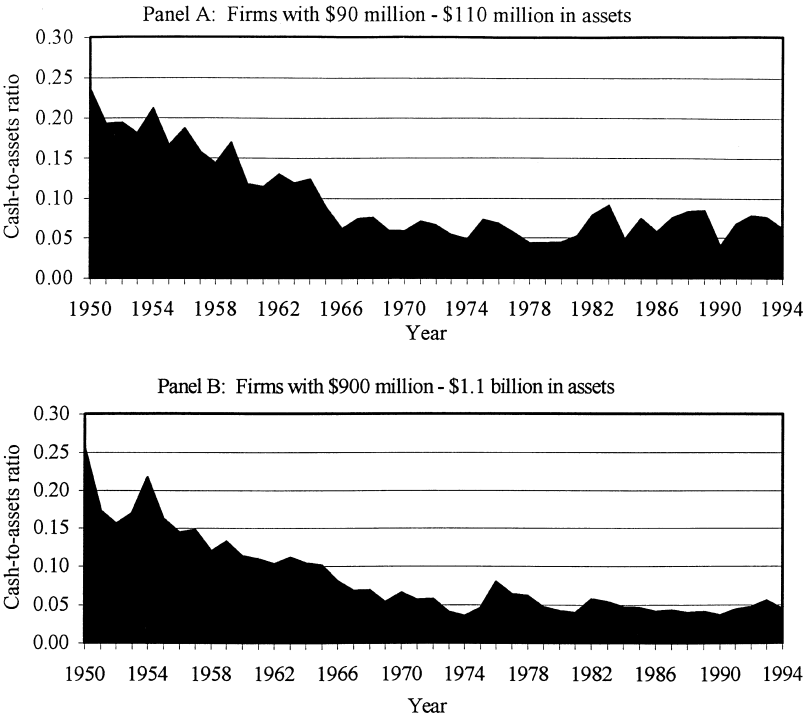


Fig. 2. Median cash-to-assets, 1950–1994. Median cash-to-assets ratio in the 1950–1994 period for Compustat firms with real assets in the \$90–110 million range, and in the \$900 million to \$1.1 billion range, in 1994 dollars (adjusted for inflation using the CPI). The ratio is calculated as cash plus marketable securities, over assets less cash plus marketable securities.

4. The determinants of cash balances

In this section, we first test whether firms have target cash levels. Finding that they do, we then estimate linear regression models where the logarithm of cash to net assets is a function of the variables that theory identifies as determinants of cash balances.

4.1. Do firms have target cash levels?

The first step in investigating whether firms have target cash levels is to examine whether cash holdings revert to the mean. If they do not, we can reject the hypothesis that firms have target cash levels. However, the financing hierarchy model is not inconsistent with mean reversion in cash holdings. In the financing hierarchy model, the time-series properties of changes in cash depend on the time-series properties of the firm’s growth in internal resources. Negative

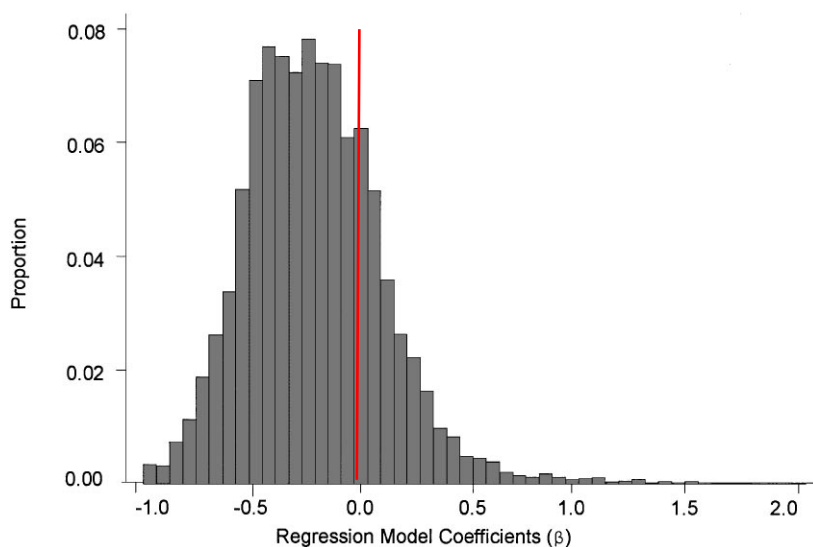


Fig. 3. Distribution of Coefficients on Lagged Change in Cash/Assets. Distribution of coefficients on lagged change in cash/assets from the firm-wise regression:

$$\Delta(\text{Cash/Assets})_t = \alpha + \beta \Delta(\text{Cash/Assets})_{t-1} + \varepsilon_t,$$

where Δ is a first difference operator, and time steps are annual. Cash/assets is defined as cash and marketable securities, over assets less cash and marketable securities. The chart includes information on 10,441 U.S. based firms included on Compustat with at least five years of data on cash holdings in the 1950–94 period. The median coefficient value is -0.242 .

autocorrelation in the growth in internal resources would lead to negative autocorrelation in cash holdings. We test the hypothesis that cash holdings are mean reverting by estimating a first order autoregressive model for each Compustat firm of the form

$$\Delta(\text{Cash/Assets})_t = \alpha + \beta \Delta(\text{Cash/Assets})_{t-1} + \varepsilon_t, \quad (1)$$

where ε_t is an independent and identically distributed disturbance with zero mean. Fig. 3 shows the distribution of the autoregressive coefficients (β) from this regression for all Compustat firms with more than five years of data in the 1950–94 period.³ The median coefficient is negative, indicating that cash balances are mean reverting. It appears that there are systematic factors that cause firms to not let cash balances rise too high or fall too low.

In Table 2, we attempt to distinguish more directly between the static tradeoff model and the financing hierarchy model. The sample used in this table is much

³ It should be noted that these coefficients are biased downwards in small samples, so that we may be underestimating the extent of mean-reversion (see Hamilton, 1994, p. 217).

Table 2
Time series analysis of liquid asset holdings

Regressions examining whether firms have target cash levels. The dependent variable is the change in level of cash/net assets from the prior year where net assets are assets net of cash and marketable securities. Target adjustment is the difference between the estimated target level of cash/net assets and the previous year's level. The target is estimated in three different ways. Mean target adjustment is an average of the prior five years of cash/net assets. Size and sigma target adjustment is calculated as the predicted value from a regression of cash/net assets on real size and industry sigma. Sophisticated target is calculated as the predicted value from the Fama-MacBeth regressions in Table 4. Pecking order is the flow of funds deficit, defined as cash dividends plus capital expenditures, change in net working capital (less cash) and current portion of long term debt due, less operating cash flow, where all variables are deflated by net assets. Pecking order * above target is an interactive dummy variable which equals Pecking order if the firm is above its target level of cash, and zero otherwise. Real size is the natural logarithm of assets, deflated using the CPI into 1994 dollars. Industry sigma is the mean of the standard deviation of the prior 20 years of cash flow divided by net assets for each industry. Industries are defined by 2-digit SIC code. The sample is restricted to 1048 firms for which cash data is available for every year of the sample. Heteroskedastic-consistent standard errors are used to calculate *t*-statistics, shown in parentheses.

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intercept	0.0002 (0.21)	− 0.0013 (− 1.70)	− 0.0037 (− 4.45)	0.0041 (5.12)	0.0056 (6.47)	0.0059 (8.01)	0.0005 (0.50)	0.0054 (6.42)	0.0055 (7.47)	0.0011 (1.12)
Mean target	− 0.3283 (− 8.69)				− 0.3519 (− 8.20)			− 0.3400 (− 7.87)		
Size and sigma target adjustment		− 0.2117 (− 11.54)				− 0.2270 (− 12.13)			− 0.2157 (− 11.37)	
Sophisticated target			− 0.2586 (− 13.22)				− 0.2555 (− 11.68)			− 0.2525 (− 11.41)
Pecking order				− 0.2195 (− 15.10)	− 0.2103 (− 12.60)	− 0.2204 (− 15.61)	− 0.2020 (− 16.83)	− 0.1310 (− 6.38)	− 0.1212 (− 6.78)	− 0.1705 (− 11.19)
Pecking order * above target								− 0.1573 (− 4.86)	− 0.2410 (− 8.65)	− 0.0673 (− 2.82)
<i>N</i>	19,912	22,851	21,582	16,086	12,028	15,351	15,078	12,028	15,351	15,078
Adjusted <i>R</i> ²	0.082	0.104	0.140	0.130	0.214	0.244	0.266	0.230	0.280	0.270

smaller than the sample used in the first test. We use only the 1048 firms for which flow-of-funds data are available every year from 1971 to 1994. The target adjustment model we posit states that changes in cash holdings in year $t + 1$ depend on the difference between actual cash holdings and the target cash holdings at the end of year t . The first model uses the average cash holdings of a firm during the five previous years as the firm's target cash holdings. This model is similar to the models tested in the capital structure literature discussed in Shyam-Sunder and Myers (1999). As shown in column (1) of Table 2, the adjustment coefficient is -0.3283 , with a t -statistic of -8.69 . The regression R^2 is 0.08. This regression indicates that a simple target adjustment model explains some of the change in cash holdings. The second regression presents estimates of a model where the target for a firm is obtained each year from the fitted values of a cross-sectional regression of cash holdings on real firm size and industry volatility. The target estimate for a given year is obtained without using information from subsequent years. The motivation for this model comes from theoretical models of cash holdings, the fact that there are returns to scale in cash holdings, and that the precautionary motive to hold cash specifies a negative relation between cash holdings and volatility. This model has a regression R^2 of 0.10, and the slope coefficient is highly significant. Finally, we use as the target the fitted values from the Fama–MacBeth cross-sectional model presented later in this paper. This model is estimated annually, so that all information required to estimate the target is available in the year in which the target is used in the regression. Using this target increases the regression R^2 to 0.14, as shown in column (3) of Table 2. All three target adjustment models are supported.

We then turn to the financing hierarchy model. We test that model by assuming that changes in cash holdings are given by the flow of funds deficit, measured as cash dividends plus capital expenditures plus the change in net working capital (less cash), plus the current portion of long-term debt due, minus the operating cash flow. Note that the flow of funds deficit is computed before financing, so that we are not estimating an identity. Both cash holdings and the flow of funds deficit are normalized by assets minus liquid assets. The coefficient on the flow of funds deficit is -0.2195 , with a t -statistic of -15.10 (see column (4) of Table 2). Consequently, there is support for the financing hierarchy model as well. When the financing hierarchy model is used for debt, the financing hierarchy model has an extremely high R^2 , in excess of 0.7, but this result is not true here. The R^2 of the financing hierarchy model is 0.13, which is slightly less than the result for the target adjustment model, using our cross-sectional model for the target.

The next three regressions of Table 2, columns (5)–(7), allow the change in cash to be influenced by both the target adjustment model and the financing hierarchy model. In all three regressions, both models are significant. It seems that the two models capture different aspects of the change in cash holdings of

firms. Adding the financing hierarchy model to the target adjustment model has little impact on the coefficient estimates or t -statistics of either model. Further, the R^2 s for the regressions that combine both models are almost equal to the sum of the R^2 s of the individual regressions.

Agency considerations suggest that managers who want to keep resources within the firm would let cash accumulate if the firm does well. However, this management would also take steps to remedy a situation where the firm has too little cash, relative to some target, even if the firm has a cash flow deficit. This reasoning makes it plausible that the financing hierarchy model would better predict changes in cash for firms that exceed their target. The last three regressions of Table 2, shown in columns (8)–(10), test this hypothesis. For all three target models, the coefficient of the flow of funds deficit is significantly higher in absolute value for firms that have liquid assets in excess of their target.

4.2. Univariate tests

Table 3 presents univariate comparisons of key descriptive variables by cash-to-assets quartile. The quartiles are constructed each year, which explains why the ranges of the cash-to-assets ratio overlap across quartiles. We are interested in whether the characteristics of companies which hold high cash balances, such as those companies in the fourth quartile, differ from those with low cash balances, such as those in the first quartile. We test the hypothesis that the fourth-quartile firms differ significantly from the first-quartile firms using a t -test. However, it turns out that firm characteristics do not always change monotonically with cash holdings, so that comparing the firms in the first and fourth quartiles of cash holdings is not sufficient to describe the relation between cash holdings and firm characteristics.

Firms in the fourth quartile of cash holdings differ significantly from firms in the first quartile of cash holdings at the 10% level, or better, for all variables we are considering. As expected, the firms with the most cash are smaller than the ones with the least cash. However, the univariate relation between cash and firm size is not monotonic. Firms in the first three quartiles of cash holdings are similar in size, but firms in the fourth quartile are substantially smaller. The market-to-book ratio increases monotonically with cash holdings. The same result holds for the R&D-to-sales ratio. The average ratio of cash flow-to-assets increases over the first three quartiles, and then falls dramatically so that the firms in the fourth quartile have the lowest average cash flow-to-assets ratio. Yet, the median cash-to-assets ratio increases monotonically, as predicted by the financing hierarchy model. Capital expenditures increase monotonically with the cash-to-assets ratio, which seems inconsistent with the financing hierarchy model. Somewhat surprisingly, the acquisition spending-to-assets ratio is lower

Table 3

Firm characteristics by cash/assets quartiles

Univariate comparison of means and medians of measures of firm characteristics of 87,135 firm years from the 1971–1994 sample of U.S.-based publicly traded firms. Median values are bracketed. Assets in the denominators of all variables are assets net of cash holdings. Real variables are deflated using the CPI into 1994 dollars. Size is defined as the natural log of assets. The market-to-book ratio is measured as the book value of assets, less the book value of equity, plus the market value of equity, divided by assets. Cash flow is defined as earnings before interest and taxes, but before depreciation and amortization, less interest, taxes, and common dividends. Net working capital is calculated without cash. Payout to shareholders is the sum of cash dividends over assets and stock repurchases over assets. Industry sigma is a measure of the volatility of an industry's cash flow for a 20-year period. Industries are defined by 2-digit SIC codes. Total leverage is total debt over total assets. Other variables are included to control for research and development (R&D) spending, acquisitions, and capital expenditures. Quartiles for cash-to-assets are determined each year. The *t*-statistic is for a difference of means test from the first to the fourth quartile. Each quartile contains approximately 21,780 firm years.

Variable	First quartile	Second quartile	Third quartile	Fourth quartile	<i>t</i> -statistic (<i>p</i> -value)
Cash/assets range	0.00 to 0.04	0.02 to 0.09	0.05 to 0.28	0.09 to 3.47	
Cash/assets	0.0129 [0.0125]	0.0427 [0.0412]	0.1142 [0.1034]	0.5082 [0.3437]	– 154.22 (0.0001)
Real size	4.773 [4.645]	4.801 [4.769]	4.687 [4.657]	4.083 [4.000]	39.32 (0.0001)
Market-to-book ratio	1.322 [1.090]	1.351 [1.102]	1.503 [1.177]	1.958 [1.458]	– 56.24 (0.0001)
R&D/sales	0.0152 [0.0000]	0.0158 [0.0000]	0.0213 [0.0000]	0.0545 [0.0000]	– 35.85 (0.0001)
Cash flow/assets	0.0324 [0.0586]	0.0390 [0.0638]	0.0488 [0.0769]	0.0288 [0.0930]	1.69 (0.0911)
Net working capital/assets	0.1828 [0.2064]	0.1783 [0.1948]	0.1729 [0.1813]	0.1686 [0.1870]	5.96 (0.0001)
Capital expenditures/assets	0.0805 [0.0578]	0.0847 [0.0603]	0.0918 [0.0666]	0.1023 [0.0736]	– 26.03 (0.0001)
Acquisitions/assets	0.0108 [0.0000]	0.0125 [0.0000]	0.0119 [0.0000]	0.0094 [0.0000]	4.14 (0.0001)
Payout to Shareholders	0.0131 [0.0048]	0.0150 [0.0063]	0.0180 [0.0082]	0.0233 [0.0056]	– 35.87 (0.0001)
Industry sigma	0.1111 [0.0801]	0.1147 [0.0819]	0.1213 [0.0866]	0.1362 [0.0988]	– 28.45 (0.0001)

for firms in the fourth quartile of cash holdings than for firms in the first quartile. In contrast, payouts to shareholders increase monotonically across quartiles of cash holdings. Neither of these results seems consistent with free cash flow theory. Finally, industry volatility increases monotonically across quartiles.

4.3. Regression tests on the 1971–1994 sample

Table 4 presents panel regressions predicting liquidity levels in the 1971–1994 period, using the independent variables described earlier. The variables predicting liquidity levels are observed in the same fiscal year as the liquidity levels. In most, but not all, regressions, firms are allowed to enter and leave the panel. We use the logarithm of liquidity as our dependent variable.⁴ In cases where we look at industry-adjusted variables, we include dummy variables for each industry, defined by the 2-digit SIC code.

The first column of Table 4 reports estimates using the method presented in Fama and MacBeth (1973), referred to hereafter as the Fama-MacBeth model. With this approach, a cross-sectional regression is estimated each year. This method eliminates the problem of serial correlation in the residuals of a time-series cross-sectional regression. The Fama-MacBeth model effectively treats each year as an independent cross-section. We find that cash holdings decrease significantly with size, net working capital, leverage, whether a firm pays dividends, and whether it is regulated. Cash holdings increase significantly with the cash flow-to-assets ratio, the capital expenditures-to-assets ratio, industry volatility, and the R&D-to-sales ratio. The coefficients are not only statistically significant but, in general, they are also economically significant. Going from small firms to large firms, increasing real assets by a factor of 100, for example, multiplies the log of the liquid assets ratio by about 0.80. An increase in the market-to-book ratio, from the first to the fourth quartile of the market-to-book distribution, more than doubles the cash-to-assets ratio.

With the Fama-MacBeth regressions, the coefficients of the market-to-book, cash flow-to-assets, and the dividend dummy variables are consistent with the static tradeoff theory, as well as with the financing hierarchy model. However, the coefficients of the size, capital expenditures, and R&D variables are more consistent with the static tradeoff theory than with the financing hierarchy model. It is not clear that the financing hierarchy model has predictions for the sign of working capital and industry volatility. The coefficients of these variables are consistent with the static tradeoff theory. Finally, the static tradeoff theory does not make clear predictions about the coefficient of leverage, but the result for this coefficient is consistent with the financing hierarchy model. To the extent that the evidence supports the static tradeoff model, it cannot be the case that cash holdings are a matter of indifference. Students of the determinants of leverage will notice that it is generally the case that the variables that affect cash holdings are also variables that affect leverage, but usually with the opposite sign, so that variables that are associated with more cash are variables that are

⁴Our qualitative results are not affected when using the level of liquidity as the dependent variable.

Table 4
Regressions predicting firm liquidity levels, 1971–1994

The dependent variable in all regressions is the natural log of cash/assets, which is calculated as cash divided by assets less cash holdings. In all the independent variable denominators, assets are net of cash. The year dummy regressions are run with a dummy variable for each year from 1972–1994. Real size is the natural log of assets, deflated using the CPI to 1994 dollars. The market-to-book ratio is measured as the book value of assets, less the book value of equity, plus the market value of equity, divided by assets. Cash flow is defined as earnings before interest and taxes, but before depreciation and amortization, less interest, taxes, and common dividends. Net working capital is calculated without cash. Total leverage is total debt over total assets. Industry sigma is the mean of standard deviations of cash flow over assets over 20 years, for firms in the same industry, as defined by 2-digit SIC code. Dividend dummy is a variable set to one if the firm paid a dividend in the year, and set to 0 if it did not. Regulation dummy is a variable set to 1 if the firm is in a regulated industry for the year, and set to 0 if it is not. Other variables are included to control for research and development (R&D) spending and capital expenditures. Industry dummy variables are constructed for each industry, defined by the 2-digit SIC code. The Fama-MacBeth model gives the average of the time series of coefficients from annual cross-sectional regressions. The cross-sectional regression uses the means of all variables for each firm. Only firms for which a full panel of data is available are used in the cross-sectional specification. All *t*-statistics are corrected for heteroskedasticity using White's (1980) correction. The adjusted R^2 of the fixed-effects model is computed without the fixed effects. The fixed-effects regression excludes firms with only one observation.

Independent variable	Fama-MacBeth model	Regressions using dummy variables for:		Cross-sectional regression	Fixed-effects regression
		Year	Year and industry		
Intercept	− 2.017 (− 35.35)	N.A.	N.A.	− 1.1247 (− 6.91)	N.A.
Market-to-book ratio	0.1515 (16.47)	0.1422 (27.60)	0.1328 (25.64)	0.3058 (6.58)	0.0998 (18.10)
Real size	− 0.0439 (− 6.79)	− 0.0402 (− 13.37)	− 0.0332 (− 10.77)	− 0.1214 (− 7.57)	− 0.0826 (− 10.14)
Cash flow/assets	0.6601 (3.71)	0.1618 (4.44)	0.0963 (2.65)	− 0.4337 (− 0.66)	0.0742 (1.93)
Net working capital/assets	− 0.9713 (− 11.71)	− 0.8136 (− 31.24)	− 0.7742 (− 25.84)	− 1.8038 (− 9.15)	− 0.5560 (− 16.95)
Capital expenditures/assets	0.0703 (0.32)	0.4850 (7.38)	0.6832 (10.11)	− 2.2110 (− 2.71)	0.6524 (10.52)
<u>Total leverage</u>	− 2.8145 (− 29.16)	− 3.0234 (− 101.61)	− 3.0504 (− 100.45)	− 3.3587 (− 15.40)	− 2.3395 (− 65.80)
<u>Industry sigma</u>	0.4533 (1.98)	1.1636 (14.92)	1.0194 (9.65)	1.0538 (2.25)	− 0.8903 (− 12.51)
R&D/sales	1.2783 (10.03)	1.6606 (19.81)	1.5452 (18.47)	− 0.4762 (− 0.52)	0.7631 (9.04)

Table 4. Continued.

Independent variable	Fama–MacBeth model	Regressions using dummy variables for:		Cross-sectional regression	Fixed-effects regression
		Year	Year and industry		
Dividend dummy	– 0.1001 (– 2.67)	– 0.1275 (– 11.35)	– 0.1247 (– 11.05)	– 0.1815 (– 2.11)	0.0403 (3.10)
Regulation dummy	– 0.1438 (– 2.59)	– 0.0968 (– 2.16)	– 0.2414 (– 4.06)	– 1.0230 (– 2.42)	– 0.0284 (– 0.60)
<i>N</i>	24	87,117	87,117	1,048	86,955
Adjusted <i>R</i> ²	0.223	0.219	0.234	0.381	0.101

associated with less debt.⁵ This reasoning suggests that variables that make debt costly make cash holdings advantageous. At the same time, however, our regressions do not imply that firms are indifferent between having one more dollar of cash or one less dollar of debt. If this were the case, one would expect the coefficient on debt to be insignificantly different from minus one. In our regressions, the coefficient on leverage is significantly different from minus one.⁶

We present four additional regression estimates in Table 4. First, we use a time-series cross-sectional regression with year dummies, and a time-series cross-sectional regression with year dummies where the variables are adjusted for industry, using dummy variables at the 2-digit SIC code level. These two regressions lead to the same results as the Fama-MacBeth regressions, but they have much higher absolute value *t*-statistics. Second, we estimate the regression using the average of the variables over the sample period for the firms used in the estimates of the target adjustment model in Table 2. The coefficient estimates in that regression are consistent with the estimates of the other regressions, except for cash flow, which is not significant, capital expenditures, which has a negative coefficient, and the R&D-to-sales ratio, which has a negative coefficient. Finally, we use a fixed-effects regression. Except for two variables, this regression has the same results as the time-series cross-sectional regressions. First, industry volatility has a significant negative coefficient. Second, the dividend dummy has a significant positive coefficient.

Table 5 addresses some concerns that arise from reviewing the results shown in Table 4. First, some of the variables in Table 4 may be determined for each

⁵ See Harris and Raviv (1991) for a review of capital structure theories and the empirical evidence.

⁶ Interestingly, Graham (1998) finds that the correlation between excess cash and excess debt capacity to be only 11.4%.

Table 5

Modified regressions predicting firm liquidity levels, 1971–1994

The dependent variable in all regressions is the natural log of cash/assets, which is calculated as cash divided by assets less cash holdings. In all the independent variable denominators, assets are net of cash. Panel A shows reduced form regressions that omit capital expenditures, leverage and dividends. Panel B shows regressions that include a measure for the difference in cash holdings. The year dummy regressions are run with a dummy variable for each year from 1972–1994. Real size is the natural log of assets, deflated using the CPI to 1994 dollars. The market-to-book ratio is measured as the book value of assets, less the book value of equity, plus the market value of equity, divided by assets. Cash flow is defined as earnings before interest and taxes, but before depreciation and amortization, less interest, taxes, and common dividends. Net working capital is calculated without cash. Total leverage is total debt over total assets. Industry sigma is the mean of standard deviations of cash flow over assets over 20 years, for firms in the same industry, as defined by the 2-digit SIC code. Dividend dummy is a variable set to one if the firm paid a dividend in the year, and set to 0 if it did not. Regulation dummy is a variable set to 1 if the firm is in a regulated industry for the year, and set to 0 if it is not. Other variables are included to control for research and development (R&D) spending and capital expenditures. Difference in cash is the change in cash over net assets from year t to year $t + 1$. Industry dummy variables are constructed for each industry, defined by 2-digit SIC code. The Fama–MacBeth model gives the average of the time series of coefficients from annual cross-sectional regressions. The cross-sectional regression uses the means of all variables for each firm. Only firms for which a full panel of data is available are used in the cross-sectional specification. All t -statistics are corrected for heteroskedasticity using White's (1980) correction. The adjusted R^2 of the fixed-effects model is computed without the fixed effects. The fixed-effects regression excludes firms with only one observation.

Independent variable	Fama–MacBeth model	Regressions using dummy variables for:		Cross sectional regression	Fixed-effects regression
		Year	Year and industry		
<i>Panel A: Reduced form regressions</i>					
Intercept	– 3.0135 (– 57.48)	N.A.	N.A.	– 2.7252 (– 19.00)	N.A.
Market-to-book ratio	0.2270 (20.62)	0.2411 (43.71)	0.2299 (41.51)	0.4512 (8.27)	0.1416 (24.57)
Real size	– 0.0727 (– 13.33)	– 0.0734 (26.00)	– 0.0666 (– 22.68)	– 0.1434 (– 8.91)	– 0.1518 (– 18.60)
Cash flow/assets	1.4205 (5.75)	0.7366 (16.94)	0.6289 (14.50)	1.2965 (2.23)	0.3762 (9.01)
Net working capital/assets	– 0.2174 (– 2.51)	– 0.1037 (– 4.01)	0.0613 (1.93)	– 0.6907 (– 3.95)	0.1442 (4.37)
Industry sigma	0.9554 (4.04)	1.6970 (20.13)	1.2452 (10.89)	1.2115 (2.23)	– 0.8924 (– 12.20)
R&D/sales	1.7285 (9.51)	2.3590 (24.45)	2.2972 (23.74)	1.2027 (1.30)	1.0643 (11.79)

Table 5. Continued.

Independent variable	Fama–MacBeth model	Regressions using dummy variables for:		Cross sectional regression	Fixed-effects regression
		Year	Year and industry		
Regulation dummy	(− 0.2184 (− 3.51)	− 0.2178 (− 4.89)	− 0.3038 (− 5.11)	− 1.1217 (− 2.55)	− 0.1540 (− 3.21)
<i>N</i>	24	87,117	87,117	1,047	86,955
Adjusted <i>R</i> ²	0.098	0.091	0.111	0.190	0.026
<i>Panel B: Regressions adding a measure for difference in cash holdings</i>					
Intercept	− 2.0311 (− 40.99)	N.A.	N.A.	− 1.2033 (− 7.86)	N.A.
Market-to-book ratio	0.1463 (16.34)	0.1335 (27.85)	0.1249 (25.94)	0.3501 (7.60)	0.0750 (16.32)
Real size	− 0.0455 (− 7.69)	− 0.0437 (− 15.37)	− 0.0374 (− 12.85)	− 0.1224 (− 7.84)	− 0.1447 (− 20.52)
Cash flow/assets	0.7182 (4.77)	0.3099 (8.65)	0.2383 (6.69)	− 0.7788 (− 1.23)	0.1625 (5.02)
Net working capital/assets	− 0.9952 (− 13.80)	− 0.8725 (− 35.23)	− 0.8098 (− 28.37)	− 1.7247 (− 9.19)	− 0.4739 (− 16.63)
Capital expenditures/assets	− 0.2863 (− 1.55)	0.0394 (0.64)	0.1842 (2.89)	− 1.5559 (− 2.06)	0.1012 (1.90)
Total Leverage	− 2.6744 (− 28.95)	− 2.8652 (− 100.34)	− 2.8798 (− 98.98)	− 3.2711 (− 15.63)	− 1.8138 (− 59.11)
Industry sigma	0.5908 (2.56)	1.2470 (16.75)	1.0366 (10.36)	0.6657 (1.47)	− 0.7739 (− 12.31)
R&D/sales	1.3082 (11.11)	1.7484 (21.17)	1.6423 (19.92)	− 0.9932 (− 1.10)	0.5932 (8.06)
Dividend dummy	− 0.1015 (− 2.78)	− 0.1267 (− 11.82)	− 0.1211 (− 11.27)	− 0.0826 (− 0.97)	0.0381 (3.31)
Regulation dummy	− 0.1453 (− 2.73)	− 0.1120 (− 2.56)	− 0.2528 (− 4.35)	− 1.1620 (− 2.73)	− 0.0548 (− 1.30)
Difference in cash	− 0.4356 (− 56.18)	− 0.4405 (− 85.83)	− 0.4394 (− 86.97)	2.6344 (7.66)	− 0.4792 (− 135.92)
<i>N</i>	24	81,819	81,819	1,047	81,775
Adjusted <i>R</i> ²	0.331	0.328	0.345	0.422	0.354

firm, jointly with their cash holdings. The static tradeoff theory would suggest that firms choose leverage, cash holdings, and investment policy simultaneously. This simultaneous determination could make our estimates inconsistent. We therefore re-estimate the regressions of Table 4 omitting

the capital expenditures, dividend, and leverage variables. These regressions, shown in Panel A of Table 5, have the interpretation of reduced-form regressions. The resulting regressions lead to the same conclusions as those of Table 4.

Another concern arising from a review of Table 4 is that some of the cash holdings are transitory, because a firm might have raised funds that it is waiting to spend, or the firm has raised funds simply because it is away from its target holdings. To allow for the existence of transitory cash holdings, we add next year's change in cash holdings as an explanatory variable. If a firm has unusually high cash because the firm just raised funds that will be spent next year, this variable should capture the part of cash holdings that is transitory. As shown in Panel B of Table 5, introducing this additional variable has little impact on the coefficients of the other variables, except that the coefficient of capital expenditures is no longer as reliable.

The fact that the univariate results indicate that firms in the fourth quartile of cash holdings are quite different from other firms, and that some variables do not change monotonically across quartiles of cash holdings, raises the issue that our results might be excessively influenced by firms that hold especially large amounts of cash. Although we do not report the results in the table, we investigate whether our results are driven by firms with large amounts of cash relative to net assets. For this investigation, we re-estimate the regressions in Table 4 after eliminating from the sample the firms that are in the top decile of cash holdings each year. The estimated coefficients in the regressions without the firms in the top decile of cash holdings have the same implications as the estimated coefficients in the regressions we report in Tables 4 and 5. The significance of our results does not depend on the firms that are in the top decile of cash holdings.

4.4. Regression tests on the 1994 sample

For 1994, we also have data on managerial ownership, derivatives usage, bond ratings, and anti-takeover charter amendments. We restrict the sample to firms for which the degree of diversification, as measured by the number of industry segments, is available. Table 6 estimates cross-sectional regressions using the explanatory variables from Tables 4 and 5, and additional explanatory variables available for 1994. The first two regressions in Table 6 use the full sample. The other two regressions use the subsample of S&P 500 firms for which derivatives usage information is available, and for which all our other variables are also available. Looking at the first two columns, we find that the explanatory variables that are in these regressions, as well as in the earlier regressions, lead to the same inferences. In most cases, the coefficient estimates are very similar. For instance, the coefficient of the market-to-book variable is 0.1445 in Table 6 and 0.1515 in the first column of Table 4.

Table 6

Derivative use and cash/assets, 1994

Two samples are used for these regressions. The full 1994 sample includes all firms on Compustat for 1994 for which we have data on insider ownership, bond rating, and the number of industry segments. The subsample of firms reporting derivatives includes only S&P 500 firms for which we have data on derivative usage. We measure derivatives as the actual value of the derivatives as reported by Risk Publications. The dependent variable in all regressions is the natural log of cash/assets, which is calculated as cash divided by assets less cash holdings. In all the independent variable denominators, assets are net of cash. Real size is the natural log of assets, deflated using the CPI to 1994 dollars. The market-to-book ratio is measured as the book value of assets, less the book value of equity, plus the market value of equity, divided by assets. Cash flow is defined as earnings before interest and taxes, but before depreciation and amortization, less interest, taxes, and common dividends. Net working capital is calculated without cash. Total leverage is total debt over total assets. Firm sigma is the standard deviation of cash flow over assets over 20 years. Dividend dummy is a variable set to one if the firm paid a dividend in the year, and set to 0 if it did not. Number of segments is measured using the Compustat segment tapes. INSIDE 0% to 5% equals inside ownership if inside ownership is less than 5% and 5% if inside ownership is greater than 5%. INSIDE 5% to 25% equals zero if inside ownership is less than 5%, equals inside ownership minus 5% if inside ownership is greater than 5% but less than 25% and equals 20% if inside ownership is greater than 25%. INSIDE over 25% equals zero if board ownership is less than 25% and equals inside ownership minus 25% if inside ownership is greater than 25%. The bond rating dummy is equal to 1 if the firm's debt has an investment grade rating (BBB or higher), and 0 if it is below investment grade (BBB- or lower), or it has no rating reported on Compustat for 1994. The anti-takeover dummy is equal to one if the firm had an anti-takeover amendment in place, and set to zero otherwise. Derivative use dummy is a variable set equal to one if the firm uses derivatives, and set to zero otherwise. Derivative use > 10% of assets is a variable set to one if the firm uses derivatives with a value greater than 10% of the firm's assets, and set to zero otherwise. Other variables are included to control for research and development (R&D) spending and capital expenditures. Due to some very significant outliers in the R&D-to-sales variable, we delete observations at the 1% tails.

Independent variable	Full 1994 sample		Firms reporting derivatives	
Intercept	– 2.3514 (– 16.06)	– 2.050 (– 15.14)	– 3.0222 (– 3.44)	– 2.9136 (– 3.35)
Market-to-book ratio	0.1445 (7.35)	0.1597 (8.17)	0.2351 (1.88)	0.2422 (1.98)
Real size	– 0.0360 (– 1.67)	– 0.0463 (– 2.14)	0.0388 (0.41)	0.0336 (0.35)
Firm sigma	0.4446 (3.65)	0.5127 (4.21)	4.7610 (2.08)	4.7900 (2.09)
R&D/sales	0.9018 (6.52)	1.0394 (7.61)	6.5442 (2.45)	7.6400 (3.22)
Cash flow/assets	0.6320 (3.93)	0.6676 (4.14)	– 1.9262 (– 1.01)	– 1.9811 (– 1.04)
Net working capital/assets	– 1.2330 (– 9.39)	– 1.2333 (– 9.35)	– 0.8547 (– 1.20)	– 0.9188 (– 1.30)

Table 6. Continued.

Independent variable	Full 1994 sample		Firms reporting derivatives	
Capital expenditure/assets	0.6426 (1.67)	0.4575 (1.19)	1.6739 (0.97)	1.5133 (0.89)
Total leverage	− 3.0598 (− 20.51)	− 3.1271 (− 20.92)	− 4.0950 (− 5.99)	− 4.1210 (− 6.04)
Number of segments	− 0.0234 (− 0.74)	− 0.0201 (− 0.63)	− 0.1011 (− 1.74)	− 0.1001 (− 1.72)
Industry sigma	1.2546 (5.23)		0.5488 (0.90)	
Dividend dummy	− 0.1422 (− 1.95)	− 0.1701 (− 2.33)	− 0.2718 (− 0.97)	− 0.2548 (− 0.91)
INSIDE 0% to 5%	3.8038 (1.85)	4.1918 (2.03)	3.5415 (0.66)	3.7571 (0.70)
INSIDE 5% to 25%	− 0.9004 (− 1.47)	− 1.0007 (− 1.63)	− 0.1298 (− 0.05)	− 0.1029 (− 0.04)
INSIDE over 25%	− 0.0870 (− 0.33)	− 0.1090 (− 0.40)	0.3899 (0.07)	0.5101 (0.09)
Bond rating dummy	− 0.5211 (− 4.51)	− 0.4770 (− 4.11)	− 0.1240 (− 0.62)	− 0.1005 (− 0.51)
Anti-takeover dummy			− 0.1870 (− 1.17)	− 0.1841 (− 1.15)
Derivative use dummy			0.0319 (0.11)	0.0107 (0.04)
Derivative use > 10% of assets			0.2822 (1.68)	0.3100 (1.88)
N	2400	2400	216	216
Adjusted R ²	0.286	0.278	0.364	0.364

The static tradeoff model implies that firms with a higher debt rating hold less cash, whereas the financing hierarchy model implies the contrary, since firms that have done well have less debt and hence a higher bond rating. The financing hierarchy model has no clear predictions for the other variables. In Table 6, firm volatility has a strong positive effect on cash holdings, even when we control for industry volatility using our industry sigma variable. Management ownership has a positive effect on cash holdings, significant at the 0.10 level for low ownership, but cash holdings do not increase further as ownership increases past 5%. This result is consistent with managerial risk aversion, insofar as managers may wish to protect their human capital with a cash buffer. Not surprisingly, from the perspective of the static tradeoff theory, firms that have an investment

grade bond rating hold less cash. Although the diversification variable has the predicted sign, it is not statistically significant.

In the last two columns of Table 6, we present regressions for the subsample of S&P 500 firms reporting derivatives usage. The results are largely similar to the ones for the full sample. One exception is that diversification has a negative coefficient, significant at the 0.10 level. Whether a firm pays dividends or not does not seem to matter, which may reflect the fact that most firms in this subsample pay dividends. The coefficients on the dummy variables denoting inside ownership, presence of anti-takeover amendments, and the bond rating level are not statistically significant. This result may be due to a lack of cross-sectional variation for these variables among S&P 500 firms. Cash holdings are unrelated to whether a firm uses derivatives, but not to the intensity of derivatives usage. A dummy variable that takes the value of one if a firm has derivatives, with a notional amount in excess of 10% of assets, has a significant positive coefficient. Consequently, the regressions in Table 6 do not provide support for the view that cash holdings and derivatives are substitutes, but are not inconsistent with the view that cash holdings and derivatives are complements.

5. Does excess cash affect spending?

We think of the regressions of Section 4 as providing a measure of the cash a firm should hold. We compute a measure of excess cash from the residuals from the Fama–MacBeth regression of Table 4. A company with positive excess cash is one that holds more cash than predicted by our model in that year. Since the cross-sectional regressions are estimated yearly, the regression model has no implication for the behavior of excess cash for a firm over time. Hence, the regression model does not imply that excess cash exhibits mean reversion, but it does imply that average excess cash across firms is equal to zero in a given year. It is interesting to note that Chrysler held \$5.145 billion of cash in 1994, of which over \$3.9 billion was excess cash, according to our model. Other large companies with excess cash are IBM, Procter and Gamble, and Ford. Companies with negative excess cash include DuPont, RJR, and Wal-Mart. Since theory provides only limited guidance for our empirical model, alternate specifications of our regressions might affect our estimates of excess cash. At the same time, however, our results do not seem to be sensitive to the alternate specifications we have explored.

To more fully understand how firms manage their cash, we show in Table 7 how spending patterns in year $t + 1$ are related to positive excess cash in year t . We use the Compustat flow of funds data to identify spending patterns on an annual, as well as on a cross-sectional, basis. Firm years are separated into quartiles on the basis of the market-to-book (MB) ratio. If the market-to-book

Table 7

Spending patterns based on market-to-book ratio and previous years excess cash

The sample includes only firm years in which the firm has positive lagged excess cash. Firm years are ranked into quartiles by the market-to-book ratio as measured by the book value of assets, less the book value of equity, plus the market value of equity, divided by assets in the current year. High (low) market-to-book firms are those ranked in the top (bottom) quartile. The firm years are also independently broken into quartiles based on the previous year's holdings of excess cash. The table shows the cross-tabulations of high and low market-to-book firm years and quartiles of excess cash holdings. The excess cash holding is the antilog of a residual from a first pass regression to predict the natural log of cash divided by assets less cash. The cash quartiles are generated for every year, and firms are regrouped each year. Panel A shows capital expenditures, Panel B shows expenditures on acquisitions, Panel C shows payments to shareholders, which is defined as stock repurchases plus cash dividends, and Panel D shows the operating cash flow. All variables are from the flow of funds statement, and are deflated by total assets less cash. Number of firm years of each quartile is in brackets. The *t*-statistic is generated from a difference of means test between the first and fourth quartiles of excess cash (column values) or the difference between high and low market-to-book (row values).

Market-to-book ratio performance	Quartiles of previous year excess cash holdings				
	First	Second	Third	Fourth	(<i>t</i> -statistic) <i>p</i> -value
Panel A: Capital expenditures					
High market-to-book firms	0.1027 [1411]	0.1019 [1971]	0.1075 [2601]	0.1166 [3539]	(− 4.64) 0.0001
Low market-to-book firms	0.0637 [3456]	0.0711 [2896]	0.0755 [2266]	0.0766 [1327]	(− 5.36) 0.0001
<i>t</i> -statistic (<i>p</i> -value)	(− 14.57) 0.0001	(− 13.28) 0.0001	(− 14.25) 0.0001	(− 14.39) 0.0001	
Panel B: Acquisitions					
High market-to-book firms	0.0125 [1397]	0.0128 [1954]	0.0139 [2588]	0.0166 [3478]	(− 3.17) 0.0015
Low market-to-book firms	0.0047 [3445]	0.0066 [2869]	0.0081 [2249]	0.0121 [1312]	(− 6.43) 0.0001
<i>t</i> -statistic (<i>p</i> -value)	(− 7.39) 0.0001	(− 5.95) 0.0001	(− 5.81) 0.0001	(− 3.22) 0.0013	
Panel C: Payments to shareholders					
High market-to-book firms	0.0206 [1375]	0.0260 [1933]	0.0228 [2490]	0.0228 [3277]	(− 1.90) 0.0577
Low market-to-book firms	0.0133 [3418]	0.0151 [2862]	0.0193 [2240]	<u>0.0265</u> [1267]	(− 12.30) 0.0001
<i>t</i> -statistic (<i>p</i> -value)	(− 8.27) 0.0001	(− 11.74) 0.0001	(− 9.45) 0.0001	(2.94) 0.0033	

Table 7. Continued.

Market-to-book ratio performance	Quartiles of previous year excess cash holdings				
	First	Second	Third	Fourth	(<i>t</i> -statistic) <i>p</i> -value
<i>Panel D: Operating cash flow</i>					
High market-to-book firms	0.1035 [793]	0.1150 [1105]	0.1180 [1432]	0.0519 [1788]	(4.92) 0.0001
Low market-to-book firms	0.0731 [2981]	0.0800 [2379]	0.0843 [1765]	0.0781 [832]	(− 0.69) 0.4886
<i>t</i> -statistic	(− 4.13)	(− 5.64)	(− 5.26)	(2.51)	
(<i>p</i> -value)	0.0001	0.0001	0.0001	0.0120	

ratio is a good proxy for the presence of profitable growth opportunities, then our discussion of the agency costs of managerial discretion predicts that these agency costs are small in high-MB firms. We therefore compare firms in the highest and lowest quartiles of the market-to-book measure for different quartiles of positive excess cash. The excess cash quartiles are computed separately across all firms each year, so that the number of firms in each cell varies, but firms in the same excess cash quartile have similar amounts of excess cash irrespective of their market-to-book ratio.

We find that capital expenditures increase monotonically in excess cash for both high-MB and low-MB firms. For all quartiles of excess cash, high-MB firms invest significantly more than low-MB firms but there is no evidence that capital expenditures increase faster for low-MB firms than for high-MB firms as excess cash increases.⁷ The increase in capital expenditures across excess cash quartiles is small compared to the increase in excess cash. Moving from the first quartile to the fourth quartile of excess cash, capital expenditures increase by about 1.4% of net assets for high-MB firms and 1.3% for low-MB firms. However, excess cash increases dramatically, since average excess cash is 1.2% of net assets in the first quartile and 58.05% in the fourth quartile. The increase in capital expenditures across quartiles seems therefore to be almost trivial relative to the increase in excess cash. Although we do not reproduce results for firms with negative excess cash, it is interesting to note that capital expenditures are U-shaped in excess cash, when we look both at positive and negative excess

⁷ The ratio of high-MB to low-MB capital expenditures is 1.57 for the first quartile of positive excess cash, 1.42 for the second, 1.43 for the third, and 1.54 for the fourth quartile.

cash. Low-MB firms in the lowest quartile of negative excess cash have capital expenditures that are similar to those of low-MB firms in the highest quartile of positive excess cash (0.0766 versus 0.0794). The same result holds for high-MB firms (0.1166 versus 0.1150). In summary, there is no evidence that the firms where one would expect the agency costs of managerial discretion to be the highest, namely low-MB firms, have a higher propensity to spend excess cash on capital expenditures than other firms.

□ Spending on acquisitions increases with excess cash, and is significantly greater for firms in the fourth quartile of excess cash than for firms in the first quartile. Hence, the evidence shown in Table 7 is consistent with Harford (1998), who predicts that more spending takes place on acquisitions as excess cash increases. When one looks at spending on acquisitions in relation to excess cash for both positive and negative amounts of excess cash, firms with negative excess cash spend less than half as much on acquisitions as do firms in the fourth quartile of positive excess cash. Again, however, spending increases much more slowly than excess cash across the excess cash quartiles.

Payments to shareholders, which are the sum of dividends and stock repurchases, do not seem to be related to excess cash for high-MB firms but are related for low-MB firms. As shown in Table 7, low-MB firms pay out less to shareholders in the first three quartiles of excess cash than do high-MB firms, but pay out more in the fourth quartile.

The bottom line from this analysis is that the spending of low-MB firms is more sensitive to excess cash. Rather surprisingly, in light of the predictions of free cash flow theory, the impact of excess cash on payouts to shareholders is of the same magnitude as the impact of excess cash on investment and spending on acquisitions. Not surprisingly, firms with negative excess cash have lower payouts than other firms.

Firms with more excess cash have higher capital expenditures, and spend more on acquisitions, even when they have poor investment opportunities. To investigate further the relation between excess cash and investment, we add excess cash to traditional investment equations. Table 8 reports such investment equations for firms in our sample. We find that, after controlling for the determinants of investment, it is still the case that greater excess cash leads firms to invest more, whether they have good investment opportunities or not. At the same time, however, it appears that the impact of excess cash on investment is significantly smaller for positive excess cash than negative excess cash. In other words, negative excess cash reduces investment more than positive excess cash increases investment. This relation could be viewed as evidence for credit constraints of the type discussed in Fazzari, Hubbard and Petersen (1988).

○ Overall, the results suggest that the propensity to spend positive excess cash is small. Table 8 also provides no evidence to support the view that it takes time for excess cash to affect investment, since most of the effect seems to take place within one year.

Table 8
Determinants of capital expenditures

Ordinary least squares (OLS) and fixed-effects regression results for uses of cash using a sample of firms for which excess cash can be calculated. A firm had to be observed for at least 2 years to be included in the regressions. The t subscripts indicate time periods. The dependent variable in all regressions is capital expenditures divided by assets in year t . Assets are net of cash holdings in all variables. All data for right-hand-side numerators come from the flow of funds statements. Cash flow is defined as earnings before interest and taxes, but before depreciation and amortization, less interest, taxes, and common dividends. Sales Growth is the natural log of sales in year t , minus the natural log of sales in year $t - 1$. (Excess Cash/Assets) $_{t-1}$ is the antilog of a lagged residual from a first pass regression to determine the natural log of cash divided by assets less cash. (Normal Cash/Assets) $_{t-1}$ is the antilog of a lagged predicted value from a first pass regression to determine the natural log of cash divided by assets less cash. **POSX** is a dummy variable which is given a value of 1 if there is positive excess cash in the firm year, and zero otherwise. For the industry-adjusted data, industry dummy variables are included in the specification. Industries are defined by 2-digit SIC codes. t -statistics are in parentheses, and are calculated using White's (1980) correction for heteroskedasticity. The adjusted R^2 for fixed-effects models are computed without the fixed effects.

Independent variable	Raw regression results:			Industry-adjusted results:			
	OLS	OLS	Fixed-effects	Fixed-effects	OLS	OLS	Fixed-effects
Intercept	0.0549 (52.67)	0.0543 (54.77)	N.A.	N.A.	N.A.	N.A.	N.A.
(Cash flow/assets) $_t$	0.0885 (23.61)	0.0872 (23.29)	0.0178 (4.96)	0.0217 (6.25)	0.0794 (22.40)	0.0787 (22.22)	0.0175 (4.96)
<u>Sales growth</u>	0.0165 (11.68)	0.0168 (11.86)	0.0183 (12.48)	0.0165 (11.54)	0.0171 (12.46)	0.0173 (12.54)	0.0169 (11.82)
Market-to-book ratio	0.0075 (14.17)	0.0075 (14.23)	0.0107 (17.17)	0.0100 (16.32)	0.0076 (15.38)	0.0075 (15.35)	0.0097 (15.83)
(Normal cash/assets) $_{t-1}$	0.0541 (3.59)	0.1196 (6.34)	0.0879 (4.69)	0.1267 (6.25)	0.0454 (3.69)	0.1127 (6.68)	0.0840 (4.67)

(Normal cash/assets) _{t-2}	0.0547 (4.18)	0.0531 (4.13)	0.0533 (3.71)	0.0524 (4.25)	0.0447 (4.27)	0.0419 (4.18)	0.0514 (3.73)	0.0521 (4.20)
(Normal cash/assets) _{t-3}	0.0052 (0.72)	0.0025 (0.35)	0.0149 (2.53)	0.0157 (2.44)	-0.0016 (-0.23)	-0.0037 (-0.56)	0.0157 (2.72)	0.0155 (2.64)
(Excess cash/assets) _{t-1}		0.0959 (3.94)		0.0720 (12.98)		0.0983 (4.85)		0.0479 (7.65)
(Excess cash/assets) _{t-2}		-0.0025 (-0.83)		0.0064 (2.30)		-0.0020 (-0.70)		0.0067 (2.44)
(Excess cash/assets) _{t-3}		-0.0079 (-3.59)		0.0042 (1.97)		-0.0055 (-2.63)		0.0043 (2.04)
POSX	0.0011 (0.91)	0.0020 (1.66)	-0.0026 (-5.96)	-0.0038 (-8.68)	-0.0028 (-2.53)	-0.0021 (-1.96)	0.0057 (9.39)	0.0016 (2.44)
POSX *	0.0326 (1.92)	-0.0260 (-1.27)	0.0338 (1.35)	-0.0130 (-0.55)	0.0787 (5.29)	0.0142 (0.77)	0.0400 (1.65)	0.0031 (0.13)
(Normal cash/assets) _{t-1}		-0.0885 (-3.62)		-0.0418 (-6.74)		-0.0906 (-4.46)		-0.0178 (-2.60)
(Excess cash/assets) _{t-1}		57,495	35,235	35,235	57,495	57,495	35,235	35,235
Sample size	57,495	0.070	0.061	0.079	0.236	0.237	0.073	0.085
Adjusted R ²	0.069							

6. What happens to excess cash?

In Section 5, we saw that an increase in excess cash leads to a surprisingly small increase in capital expenditures, acquisitions spending, and payouts to shareholders. This observation suggests that there is substantial persistence in excess cash. To examine this persistence, we divide the firm-years in our sample into quartiles of excess cash. In Table 9 we show the status in subsequent years of firms that have entered the fourth quartile of excess cash in our sample for the first time. 55.5% of these firms are in the same quartile the following year. This indicates that being in the fourth quartile is a transitory state for more than 40% of the firms. However, firms which are in the fourth quartile for more than one year tend to be in that quartile for a substantial amount of time. The percentage of firms that are in the fourth quartile five years after the first time that they are in the fourth quartile of excess cash is 38.8%. A similar result holds for the firms that enter the first quartile of excess cash. Hence, there is persistence both in the highest quartile of excess cash, and in the lowest quartile.

The counterpart of this persistence is that neither firms in the first quartile nor firms in the fourth quartile change their spending patterns dramatically. Table 10 shows spending patterns for firms in the first and fourth quartiles for five years. Comparing the results from Panel A to those in Panel B, firms in the fourth quartile spend more. They still spend more on acquisitions and shareholder payouts five years after having been identified as firms in the fourth quartile of excess cash.

Why is it that firms experience large changes in excess cash? We have seen that, on average, expenditure patterns of high excess cash firms are not such that they use up their excess cash quickly. We therefore look at firms that go from the top quartile of excess cash to the bottom quartile of excess cash in one year. We then look at the expenditure and cash flow patterns for these firms. The results are reproduced in Table 11. The clear result in that table is that firms that experience large changes in excess cash, on average, experience large negative operating cash flows. Note that in Table 11, the end of year 0 is used to assign a firm to an excess cash quartile. We then select the firms that go from quartiles 4 to 1. The largest swing in the ratios reported is the one for operating cash flow. This swing represents a change, on average, of more than 3.5% of assets. Neither capital expenditures nor acquisitions increase by as much as one percent of assets. The results in Table 11 also show that lumpiness of capital expenditures and acquisitions is not an important reason for large changes in excess cash. Firms that experience large increases in excess cash also experience them because of large swings in operating cash flow. In Panel A, firms that go from the first quartile of excess cash to the fourth quartile experience an average swing in operating cash flow of more than 15% of net assets. Strikingly, however, this dramatic shift in cash flow has a small impact on capital expenditures, acquisitions, and payments to shareholders. In other words, the firms that experience such a large increase in excess cash keep it.

Table 9

Persistence of excess cash

Persistence of levels of excess cash for firms selected based on the first time they enter the highest (lowest) quartile of excess cash. The excess cash holding is the antilog of a residual from a first pass regression to predict the natural log of cash divided by assets less cash. The firms are followed for the next five years to determine the quartile in which they belong in the subsequent years. Quartile 4 represents the highest excess cash quartile, and Year 0 is the measurement year. Numbers shown are percentages. The number of firm years in each quartile, each year, is in brackets.

	Quartile 4	Quartile 3	Quartile 2	Quartile 1
<i>Panel A. Persistence of excess cash for firms that are in the highest quartile of excess cash in year 0</i>				
Year 0	100.0 [6221]			
Year 1	55.5 [3010]	22.8 [1235]	8.8 [476]	12.9 [701]
Year 2	46.4 [2155]	24.6 [1140]	13.8 [641]	15.2 [706]
Year 3	41.4 [1664]	26.4 [1060]	14.5 [581]	17.7 [712]
Year 4	40.6 [1432]	25.8 [909]	15.7 [552]	18.0 [634]
Year 5	38.8 [1192]	27.3 [841]	15.9 [489]	18.0 [554]
<i>Panel B. Persistence of excess cash for firms that are in the lowest quartile of excess cash in year 0</i>				
Year 0				100.0 [6417]
Year 1	8.9 [494]	13.4 [741]	25.0 [1383]	52.7 [2912]
Year 2	11.8 [565]	16.6 [792]	28.0 [1335]	43.6 [2082]
Year 3	13.4 [563]	19.8 [829]	28.0 [1175]	38.8 [1629]
Year 4	14.8 [543]	20.9 [767]	27.1 [993]	37.1 [1359]
Year 5	15.8 [510]	20.8 [674]	26.5 [856]	37.0 [1196]

We also investigate firms that end up in the top or bottom quartile of excess cash for the five years before they end in that state. These results are not reported, although they are fully consistent with our other results. Firms that end up with excess cash are firms that have done well, and firms that end up with low excess cash are firms that have done poorly in the most recent years.

Table 10

Future cash disposition for firms with high and low excess cash

The sample includes firms that enter the highest or lowest excess cash quartile for only the first year that they were in the first or fourth excess cash quartile, designated as year zero. The firm years are independently broken into quartiles based on the previous year's holdings of excess cash. The excess cash holding is the antilog of a residual from a first pass regression to predict the natural log of cash divided by assets less cash. The cash quartiles are generated for every year, and firms are regrouped each year. Panel A shows firms in the fourth quartile, and Panel B shows firms in the first quartile. Panel C shows the *t*-statistics and the *p*-values for the tests of differences of means between the first and fourth quartile for each variable and each year. The table shows operating cash flow, capital expenditures, expenditures on acquisitions, and payments to shareholders, which is defined as stock repurchases plus cash dividends. All variables are from the flow of funds statement, and are deflated by total assets less cash. Number of firm years of each quartile is also included in brackets. *t*-statistics for the difference in means from year 0 are shown in parentheses in Panels A and B.

Variable	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
<i>Panel A. Firms entering the top quartile of excess cash holdings in year 0</i>						
Operating cash flow	− 0.0011 [3575]	0.0604 [3949] (− 11.94)	0.0673 [3351] (− 13.36)	0.0749 [2904] (− 14.70)	0.0839 [2522] (− 16.37)	0.0884 [2200] (− 17.44)
Capital expenditures	0.1127 [6342]	0.0968 [5458] (8.39)	0.0911 [4724] (11.38)	0.0874 [4137] (13.37)	0.0847 [3642] (14.61)	0.0834 [3224] (15.30)
Acquisitions	0.0123 [6203]	0.0105 [5370] (2.58)	0.0104 [4645] (2.60)	0.0094 [4068] (3.91)	0.0085 [3579] (5.15)	0.0094 [3177] (3.74)
Payments to shareholders	0.0121 [6136]	0.0126 [5332] (− 1.07)	0.0135 [4613] (− 3.06)	0.0149 [4044] (− 5.69)	0.0159 [3565] (− 7.44)	0.0174 [3160] (− 9.38)

Panel B. Firms entering the bottom quartile of excess cash holdings in year 0

Operating cash flow	0.0561 [3511]	0.0419 [3726] (1.95)	0.0732 [3114] (-2.43)	0.0817 [2680] (-3.69)	0.0990 [2284] (-6.16)	0.1041 [1869] (-6.69)
Capital expenditures	0.1149 [6231]	0.1132 [5423] (0.85)	0.0983 [4633] (8.56)	0.0911 [4007] (12.17)	0.0859 [3518] (14.88)	0.0851 [3078] (14.84)
Acquisitions	0.0104 [6114]	0.0159 [5309] (-6.68)	0.0136 [4559] (-4.01)	0.0141 [3945] (-4.22)	0.0115 [3485] (-1.25)	0.0114 [3054] (-1.11)
Payments to shareholders	0.0143 [5977]	0.0152 [5272] (-1.81)	0.0161 [4514] (-3.22)	0.0177 [3922] (-5.66)	0.0186 [3443] (-6.90)	0.0200 [3009] (-8.34)

Panel C. *t*-statistics and *p*-values for difference in means of firms entering the first and fourth quartiles in year 0

Operating cash flow	-7.85 0.0001	3.62 0.0003	-1.24 0.2161	-1.45 0.1459	-3.20 0.0016	-3.16 0.0016
Capital expenditures	-1.10 0.2718	-8.57 0.0001	-3.92 0.0001	-2.01 0.0443	-0.66 0.5120	-0.90 0.3703
Acquisitions	2.68 0.0074	-6.57 0.0001	-3.90 0.0001	-5.23 0.0001	-3.51 0.0004	-2.15 0.0317
Payments to shareholders	-4.51 0.0001	-5.22 0.0001	-4.68 0.0001	-4.53 0.0001	-4.12 0.0001	-3.42 0.0006

Table 11

Future cash disposition for firms going from the highest to the lowest quartile, or vice versa, in one year

The sample includes firms that enter the highest or lowest excess cash quartile for only the first year that they were in the first or fourth excess cash quartile, designated as year zero, and then the fourth, or first, excess cash quartile, respectively, the next year, designated as year one. The firm years are independently broken into quartiles based on the previous year's holdings of excess cash. The excess cash holding is the analog of a residual from a first pass regression to predict the natural log of cash divided by assets less cash. The cash quartiles are generated for every year, and firms are regrouped each year. Panel A shows firms which moved from the first to the fourth quartile, and Panel B shows firms which moved from the fourth to the first quartile. Panel C shows the *t*-statistics and the *p*-values for the tests of differences of means between the first and fourth quartile for each variable and each year. The table shows operating cash flow, capital expenditures, expenditures on acquisitions, and payments to shareholders, which is defined as stock repurchases plus cash dividends. All variables are from the flow of funds statement, and are deflated by total assets less cash. Number of firm years of each quartile is also included in brackets. *t*-statistics for the difference in means from year zero are shown in parentheses in Panels A and B.

Variable	Year - 1	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
<i>Panel A. Bottom quartile excess cash holdings in year 0, top quartile in year 1</i>							
Operating cash flow	- 0.0604 [215] (- 0.45)	- 0.0759 [248]	0.0744 [276] (- 5.25)	- 0.0012 [230] (- 2.63)	0.0346 [188] (- 3.99)	0.0211 [152] (- 3.40)	0.0438 [129] (- 4.32)
Capital expenditures	0.1182 [243] (- 0.84)	0.1105 [430]	0.1006 [430] (1.41)	0.1044 [338] (0.86)	0.0875 [286] (3.31)	0.0907 [254] (2.55)	0.0818 [215] (3.84)
Acquisitions	0.0107 [236] (1.28)	0.0152 [418]	0.0146 [424] (0.19)	0.0219 [331] (- 1.69)	0.0128 [281] (0.72)	0.0130 [252] (0.61)	0.0115 [213] (1.05)
Payments to shareholders	0.0146 [234] (- 2.42)	0.0092 [411]	0.0113 [418]	0.0126 [330] (- 1.94)	0.0136 [280] (- 2.29)	0.0159 [248] (- 3.06)	0.0143 [208] (- 2.40)

Panel B. Top quartile excess cash holdings in year 0, bottom quartile in year 1

Operating cash flow	-0.0150 [213] (-1.40)	-0.0637 [342]	-0.1015 [408] (1.37)	0.0043 [343] (-2.42)	0.0390 [284] (-3.59)	0.0188 [232] (-2.74)	0.0654 [171] (-4.18)
Capital expenditures	0.0989 [201] (5.04)	0.1435 [629]	0.1468 [629] (-0.44)	0.0976 [487] (6.86)	0.0904 [386] (7.52)	0.0848 [339] (8.49)	0.0885 [300] (7.62)
Acquisitions	0.0106 [195] (0.28)	0.0116 [614]	0.0204 [613] (-3.06)	0.0126 [482] (-0.38)	0.0144 [380] (-0.92)	0.0132 [338] (-0.55)	0.0116 [297] (0.01)
Payments to shareholders	0.0139 [193] (-0.68)	0.0123 [604]	0.0099 [616] (1.46)	0.0106 [481] (0.98)	0.0106 [384] (1.05)	0.0123 [333] (-0.02)	0.0124 [296] (-0.07)

Panel C. t-statistics for difference in means of firm groups

Operating cash flow	-1.21 0.2252	-0.39 0.6998	7.31 0.0001	-0.23 0.8207	-0.18 0.8542	0.09 0.9324	-0.81 0.4199
Capital expenditures	1.86 0.0636	-4.51 0.0001	-6.54 0.0001	1.04 0.2991	-0.44 0.6609	0.79 0.4296	-0.91 0.3652
Acquisitions	0.03 0.9777	1.20 0.2309	-1.91 0.0563	2.47 0.0137	-0.48 0.6297	-0.06 0.9534	-0.04 0.9722
Payments to shareholders	0.24 0.8103	-1.90 0.0578	0.82 0.4110	1.09 0.2755	1.56 0.1195	1.53 0.1262	0.79 0.4322

7. Conclusion

We examine the determinants of corporate holdings of cash and marketable securities among publicly traded US firms from 1971–1994, as well as how firms change their holdings over time. We find evidence supportive of a target adjustment model, but it is also clearly the case that firms that do well accumulate more cash than one would expect with the static tradeoff theory, where managers maximize shareholder wealth. Our results indicate that firms with strong growth opportunities, firms with riskier activities, and small firms hold more cash than other firms. Firms that have the greatest access to the capital market, such as large firms and those with credit ratings, tend to hold less cash. These results are consistent with the view that firms hold liquid assets to ensure that they will be able to keep investing when cash flow is too low, relative to investment, and when outside funds are expensive. Our analysis provides limited support for the view that positive excess cash leads firms to spend substantially more on investment or acquisitions. Whereas acquisitions increase with excess cash, payouts to shareholders increase with excess cash as well. However, in both cases, the propensity to use excess cash on investment and acquisitions is quite limited.

The evidence in this paper is consistent with the view that management accumulates excess cash if it has the opportunity to do so. The motivation for this behavior seems to be that the precautionary motive for holding cash is excessively strong. The result that the firm's flow of funds deficit has a stronger impact on changes in cash holdings for firms that have cash in excess of their target supports this conclusion. At the same time, however, using cross-sectional data for 1994, we are not successful in demonstrating that proxies for agency costs have an important impact on cash holdings. Therefore, our results suggest that more work needs to be done to explain why firms appear to hold excess cash, as well as to understand the cost of this practice. An important issue for further research is whether, when a firm runs into difficulties, excess cash allows management to avoid making required changes, using up the firms cash to finance losses. If this behavior is the case, it would not be surprising that management is not as concerned about hoarding excess cash as shareholders might be.

Our results provide support for a static tradeoff view. At the same time, however, it is also quite clear that variables that make debt costly for a firm are variables that make cash advantageous. Because the determinants of cash are so closely related to the determinants of debt in our analysis, it is important in future work to figure out, both theoretically and empirically, to what extent cash holdings and debt are two faces of the same coin.

References

- Antunovich, P., 1996. Optimal slack policy under asymmetric information. Unpublished manuscript. Northwestern University, Evanston, IL.

- Amihud, Y., Mendelson, H., 1986. Liquidity and stock returns. *Financial Analyst Journal* 42, 43–48.
- Barclay, M.J., Smith Jr., C.W., 1995. The maturity structure of corporate debt. *Journal of Finance* 50, 609–631.
- Baskin, J., 1987. Corporate liquidity in games of monopoly power. *Review of Economics and Statistics* 69, 312–319.
- Beltz, J., Frank, M., 1996. Risk and corporate holdings of highly liquid assets. Unpublished manuscript. University of British Columbia, Vancouver.
- Chudson, W., 1945. *The Pattern of Corporate Financial Structure*. National Bureau of Economic Research, New York.
- Fama, E.F., Jensen, M.C., 1983. Agency problems and residual claims. *Journal of Law and Economics* 26, 327–350.
- Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: empirical tests. *Journal of Political Economy* 81, 607–636.
- Fazzari, S.M., Hubbard, R.G., Petersen, B., 1988. Financing constraints and corporate investment. *Brookings Papers on Economic Activity* 19, 141–195.
- Frazer, W.J., 1964. Financial structure of manufacturing corporations and the demand for money: some empirical findings. *Journal of Political Economy*, 176–183.
- Graham, J., 1998. How big are the tax benefits to debt. Unpublished working paper. Duke University, Durham, NC.
- Hamilton, J.D., 1994. *Time Series Analysis*. Princeton University Press, NJ.
- Harris, M., Raviv, A., 1991. The theory of capital structure. *Journal of Finance* 46, 297–356.
- Harford, J., 1998. Corporate cash reserves and acquisitions. Unpublished working paper. University of Oregon, Eugene, OR.
- Jensen, M.C., Meckling, W.H., 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305–360.
- John, T.A., 1993. Accounting measures of corporate liquidity, leverage, and costs of financial distress. *Financial Management* 22, 91–100.
- Jung, K., Kim, Y., Stulz, R., 1996. Timing, investment opportunities, managerial discretion, and the security issue decision. *Journal of Financial Economics* 42, 159–185.
- Keynes, J.M., 1936. *The General Theory of Employment*. In: *Interest and Money*. Harcourt Brace, London.
- Kim, Chang-Soo, Mauer, D.C., Sherman, A.E., 1998. The determinants of corporate liquidity: theory and evidence. *Journal of Financial and Quantitative Analysis* 33, 305–334.
- Lang, L., Poulsen, A., Stulz, R., 1994. Asset sales, firm performance, and the agency costs of managerial discretion. *Journal of Financial Economics* 37, 3–37.
- McConnell, J., Servaes, H., 1990. Additional evidence on equity ownership and firm value. *Journal of Financial Economics* 27, 595–612.
- Meltzer, A.H., 1963. The demand for money: a cross-section study of business firms. *Quarterly Journal of Economics*, 405–422.
- Miller, M.H., Orr, D., 1966. A model of the demand for money by firms. *Quarterly Journal of Economics*, 413–435.
- Morck, R., Shleifer, A., Vishny, R.W., 1988. Management ownership and market valuation: an empirical analysis. *Journal of Financial Economics* 20, 293–315.
- Mulligan, C.B., 1997. Scale economies, the value of time, and the demand for money: longitudinal evidence from firms. *Journal of Political Economy* 105, 1061–1079.
- Myers, S.C., 1977. Determinants of corporate borrowing. *Journal of Financial Economics* 5, 147–175.
- Myers, S.C., Majluf, N., 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13, 187–221.
- Opler, T.C., Titman, S., 1994. Financial Distress and Corporate Performance. *Journal of Finance* 49, 1015–1040.

- Shleifer, A., Vishny, R., 1986. Large shareholders and corporate control. *Journal of Political Economics* 94, 461–488.
- Shleifer, A., Vishny, R., 1993. Liquidation values and debt capacity: a market equilibrium approach. *Journal of Finance* 47, 1343–1366.
- Shyam-Sunder, L., Myers, S.C., 1999. Testing static trade-off against pecking order models of capital structure. *Journal of Financial Economics* 51, 219–244.
- Smith, C.W., Watts, R.L., 1992. The investment opportunity set and corporate financing, dividend and compensation policies. *Journal of Financial Economics* 32, 263–292.
- Stulz, R., 1988. Managerial control of voting rights: financing policies and the market for corporate control. *Journal of Financial Economics* 20, 25–54.
- Stulz, R., 1990. Managerial discretion and optimal financing policies. *Journal of Financial Economics* 26, 3–27.
- Vogel, R.C., Maddala, G.S., 1967. Cross-section estimates of liquid asset demand by manufacturing corporations. *Journal of Finance* 22, 557–575.
- White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–838.

A Stochastic Frontier Analysis of Financing Constraints on Investment: The Case of Financial Liberalization in Taiwan

Hung-Jen WANG

The Institute of Economics, Academia Sinica, Taipei 115, Taiwan (hjwang@econ.sinica.edu.tw)

It is shown that investment under financing constraints can be modeled as a one-sided deviation from a frictionless investment level, and that effects of financing constraints can be identified and quantified by imposing a distributional assumption on the effects. Panel data on Taiwanese manufacturing firms between 1989 and 1996 are used in the estimation. It is found that (1) some of the sorting criteria used in the literature do not have significant and monotonic relationships with the degrees of financing constraint, resulting in problematic sample separations, and (2) the effects of financial liberalization in Taiwan are such that the investment efficiency improved over time for a typical firm, and the improvement was particularly large for smaller firms.

KEY WORDS: Financial liberalization; Financing constraints; Stochastic frontier.

1. INTRODUCTION

Over the past decades, many studies have been devoted to providing evidence for the hypothesis of financing constraints on investment (see Chirinko 1993 and Hubbard 1998 for extensive reviews). According to this hypothesis, the capital market is imperfect owing to informational problems, and as such, corporate capital investment is no longer determined solely by fundamentals such as user costs or Tobin's Q , but also by financial factors. In particular, investment is constrained if market imperfections exert difficulties for financing investment. Therefore, failure to take into account the effect of the financing constraint on investment leads to misspecified empirical equations.

Efforts in the empirical modeling, however, are to some extent hindered by the difficulty in specifying the structural relationships among the real and financial variables in both the firm and time dimensions. For example, in linear regression models (e.g., Carpenter, Fazzari, and Petersen 1994), it is not clear how financing constraints, liquidity variables, and investment spending should enter the equation and interact with each other (Chirinko 1997). For a structural Euler equation model (e.g., Whited 1992; Hubbard, Kashyap, and Whited 1995), doubts are sometimes raised as to whether the period-by-period perturbation method can pick up the effect of a firm for which the overall investment level is constrained in the entire sample period (Gilchrist and Himmelberg 1995; Hubbard 1998). Another problem common to either of the approaches is the use of ad hoc classification criteria to separate firms into a priori constrained and unconstrained groups. As argued by Hu and Schiantarelli (1998), the dependence on a single indicator to separate samples can be risky, the implication that a firm's financial status does not change over time may be unrealistic, and selection of the criterion may also give rise to endogenous selection problems.

This article proposes a new estimation strategy that circumvents some of the aforementioned problems. This approach does not separate samples a priori to test investment cash flow sensitivity, and it can provide not only cross-sectional, but also intertemporal comparisons of the effect of financing constraints. Using the experience of financial liberalization in Tai-

wan between 1989 and 1996 as a natural experiment, the article provides new evidence concerning the financing constraints on investment which is otherwise unavailable.

The key to the approach herein lies in the insight that financing constraints should have asymmetric effects on a frictionless level of investment; it forces realized investment to be below, but never above, the frictionless neoclassical level. Therefore, identification of the constraints is achieved by imposing a one-sided distributional assumption on the effect of financing constraints. With a neoclassical model characterizing the frontier investment, the level of financing-constrained investment is then estimated as a deviation from the frontier, with the option of modeling the one-sided deviation as a function of firm characteristics. The degree of financing constraints can also be calculated using the difference between the frontier and the actual levels of investment. The econometric technique is essentially the *stochastic frontier* estimation, which is well studied in the relevant literature.

This approach has several advantages. First, the structural relationship between the financial and real variables are relatively straightforward; that is, financing constraints have one-sided effects on investment, and financing constraints can be explained by a vector of observable variables. This is to be compared with the more traditional approach in which direct interactions between financing constraints, liquidity variables, and investment spending must be modeled more elaborately. Second, the sample does not have to be split a priori. Rather, ex post quantitative measures of the effect of financing constraints can be obtained for each observation, and comparisons of financing constraints can be made based on these measures. This lifts the sample separation problem altogether and makes comparisons possible not only across groups of firms in a given time period, but also across time for all or selected groups of firms.

The last-mentioned property (that the effects of financing constraints can be compared over time) is of special interest in the case of Taiwanese corporate investment in the late 1980s and the 1990s, on which the empirical study herein is based. Taiwan's financial market was heavily regulated before the late 1980s and has since undergone a series of liberalization reforms. For instance, the law that banned the establishment of new banks was lifted in 1991, and 18 new domestic banks had been set up by the end of 1996. Various interest rates were also deregulated during this period, including the bank loan rate in 1989 and the interbank call loan rate in 1992. The deposit rate was also deregulated, albeit in two phases, first in 1989 and then in 1995. As a result of the reforms in the banking industry, the ratio of loans to private enterprises to GDP increased from .42 at the end of 1989 to .55 at the end of 1996. The stock market also underwent a series of reforms, and one of the consequences is the increase in the numbers of publicly traded firms and securities brokers. For example, the number publicly traded firms rose from 161 at the end of 1988 to 382 by the end of 1996.

This experience of financial liberalization provides a unique opportunity for testing the financing constraint hypothesis. If this hypothesis were true, then one would expect to see that the problem of financing constraints attenuates over time for most of the firms, and for underprivileged firms in particular. To the author's knowledge, this is the first study that provides evidence of financing constraint from both cross-sectional and intertemporal comparisons.

The rest of the article is organized as follows. Section 2 shows how investment in an imperfect capital market can be represented as a combination of frictionless investment and a one-sided constraint effect. Section 3 details the econometrics, and Section 4 describes the data. Section 5 describes the estimation results. Section 6 reports on a postestimation analysis carried out based on the quantitative measures of the financing constraint effects. Section 7 concludes the article.

2. THE MODEL

This section motivates a stochastic frontier investment model from a firm's optimization problem. The model follows from Chirinko and Schaller (1995), who showed that coefficients in a liquidity-augmented Q model are nonlinear functions of the firm's characteristics and financial positions. With some modifications, we show that the same model also implies that constrained investment is a combination of frontier investment, represented by a Q investment model, and a one-sided constraint effect. Because of the lengthy derivation and the similarity with the cited one, the model is only sketched here; see referred to the work of Chirinko and Schaller (1995) for more information.

Denote a firm's Q at the beginning of period t as Q_t , which is defined as

$$Q_t = \frac{S_{t-1}(1+r) + B_{t-1} - L_{t-1} - p_t^I K_{t-1}}{p_t^I K_{t-1}}. \quad (1)$$

The four stock variables, S_{t-1} , B_{t-1} , L_{t-1} , and K_{t-1} , are the market values of the firm's equity shares, debts, liquid assets,

and capital stocks, measured at the beginning of period t . Because it is assumed that the firm accrues revenues and pays expenses at the end of period t , S_{t-1} is multiplied by 1 plus the interest rate r to get the value of the shares for period t . Therefore, the numerator measures the market value of the firm in exceeding the replacement cost of the firm at the beginning of period t ; the denominator, the replacement cost of the firm also measured at the beginning of period t .

The market value of equity, S_{t-1} , is obtained from the following optimization problem, in which managers maximize the expected discounted sum of dividends:

$$S_{t-1} = \max E_t \left\{ \sum_{s=t}^{\infty} \Delta^{s-t+1} \left\{ T[K_{s-1}, B_{s-1}, L_{s-1}, I_s, b_s] - p_s^I I_s - i[m[s], K_{s-1}, B_{s-1}, L_{s-1}, I_s] B_{s-1} + (j_s L_{s-1} - I_s + b_s) - \lambda_s \left[K_s - \sum_{u=-\infty}^{u=s} I_u \right] + \phi_s \left[B_s - \sum_{u=-\infty}^{u=s} b_u \right] - \psi_s \left[L_s - \sum_{u=-\infty}^{u=s} l_u \right] \right\} \right\}, \quad (2)$$

where Δ is the time-invariant discount factor; K and I are the capital stock and investment, and $K_s = \sum_{u=-\infty}^s I_u$; B and b are the stock and flow of debt, and $B_s = \sum_{u=-\infty}^s b_u$; L and l are the stock and flow of liquid assets, and $L_s = \sum_{u=-\infty}^s l_u$; j is the interest rate accrued from the stock of liquidity asset; p^I is the price of investment goods; $T[\cdot]$ is the net revenue function; $i[\cdot]$ is the interest rate function; $m[\cdot]$ is the function of liquidity, where $m[\cdot] = T[\cdot] + L$; λ is the shadow value of K ; ϕ is the shadow value of B ; and ψ is the shadow value of L .

Solutions to the foregoing maximization problem lead to

$$S_{t-1} = E_t \left\{ \frac{1}{1+r} (\Lambda_t K_{t-1} - \Phi_t B_{t-1} + \Psi_t L_{t-1}) \right\}, \quad (3)$$

where Λ_t is the discount sum of λ_s , $s = t, t+1, \dots, \infty$; Φ_t and Ψ_t are defined similarly. These three variables have the following values in a dividend-maximization equilibrium:

$$\Lambda_t = E\{p_t^I - T_t[t](1 - i_m[t]B_{t-1})\}, \quad (4)$$

$$\Phi_t = E\{1 + T_b[t](1 - i_m[t]B_{t-1})\}, \quad (5)$$

and

$$\Psi_t = 1. \quad (6)$$

Specifying the functional form of $T[\cdot]$ is also necessary to derive the investment equation,

$$T[K_{s-1}, B_{s-1}, L_{s-1}, I_s, b_s] = P[K_{t-1}, B_{t-1}, L_{t-1}] - p_t^I A[I_t, K_{t-1}, v_t] - F[b_t, B_{t-1}], \quad (7)$$

$$A[t] = \frac{\rho}{2} \left[\frac{I_t}{K_{t-1}} - \alpha - v_t \right]^2 K_{t-1}, \quad (8)$$

and

$$F[t] = \frac{\Gamma[Z_t]}{2} \left(\frac{b_t^2}{B_{t-1}} \right); \quad \Gamma[Z_t] \geq 0. \quad (9)$$

Equation (8) is a standard adjustment cost function in which v_t is a 0-mean production shock. Equation (9) is an information cost function in which $\Gamma[Z_t]$ measures the firm's proneness to information and incentive problems and is a function of firm characteristics Z_t . A firm more likely to suffer from information problems has a larger $\Gamma[Z_t]$ and, given b_t and B_{t-1} , incurs higher informational costs. Thus $F[\cdot]$ is a cost wedge between internal and external finance. Under the assumption of perfect capital markets, $\Gamma[Z_t]$ would be 0, and there is no informational cost of borrowing.

2.1 A Stochastic Frontier Model

With the functional forms in (4)–(9), (3) is substituted into (1) to obtain

$$\frac{I_t}{K_{t-1}} = \alpha + \frac{1}{\rho} Q_t - \frac{b_t \Gamma[Z_t]}{\rho p_t^I K_{t-1}} + v_t, \quad v_t \sim N(0, \sigma_v^2). \quad (10)$$

Equation (10) shows that the rate of investment depends on a constant, Q , and on an unspecified function of financial variables and firm characteristics.

The expected investment is compared with and without financing constraints. If the capital market is perfect, as is assumed in the neoclassical framework, the information problem does not exist ($\Gamma[Z_t] = 0$), and Q is a sufficient statistic of investment. It can be seen from (10) that

$$E\left(\frac{I_t}{K_{t-1}} \middle| Q_t = \bar{Q}_t; \Gamma[Z_t] = 0\right) - E\left(\frac{I_t}{K_{t-1}} \middle| Q_t = \bar{Q}_t; \Gamma[Z_t] > 0\right) > 0. \quad (11)$$

That is, other things being equal, financing constraints restrict investment below the neoclassical level. Thus the effect of capital market imperfection is one-sided; it forces investment to go below, but never above, the frictionless level.

The foregoing observation suggests that the neoclassical investment function effectively determines the *stochastic frontier* investment, $(I_t/K_{t-1})^{SF}$, which, by definition, defines the maximum permissible rate of investment,

$$\left(\frac{I_t}{K_{t-1}}\right)^{SF} = \alpha + \frac{1}{\rho} Q_t + v_t, \quad v_t \sim N(0, \sigma_v^2). \quad (12)$$

Equation (10) thus can be generalized as the difference between the frontier investment function and a nonnegative financing constraint effect u_t , with the latter being possibly parameterized by a function of a stochastic random error and variables affecting the firm's ability to finance. Therefore,

$$\frac{I_t}{K_{t-1}} = \left(\frac{I_t}{K_{t-1}}\right)^{SF} - u(Z_t, w_t), \quad (13)$$

where Z_t is a vector of nonstochastic variables and w_t is a random error.

3. ECONOMETRICS

The foregoing equation is expanded and generalized into the following empirical equations for a panel data:

$$y_{it} = \alpha + \tilde{\mathbf{X}}_{it} \tilde{\beta} + e_{it}, \quad (14)$$

$$e_{it} = v_{it} + f_i + \tau_t - u_{it}, \quad (15)$$

$$v_{it} \sim \text{iid } N(0, \sigma_v^2), \quad (16)$$

and

$$u_{it} \sim \text{nonnegative truncation of } N(\mu_{it}, \sigma_{it}^2). \quad (17)$$

Equation (14) says that y_{it} (i.e. the rate of investment) is a function of the explanatory variables $\tilde{\mathbf{X}}$ (i.e. Q variable) and an error term e_{it} . The composite error e_{it} consists of (a) a white noise error, v_{it} ; (b) the unobservable firm-specific effect, f_i ; (c) the time-specific effect, τ_t ; and (d) the negative of the financing constraint effect, u_{it} , which is a nonnegative truncation of a normal random variable with observation-specific mean and variance. The effects of f_i and τ_t are allowed to correlate with variables in $\tilde{\mathbf{X}}$, thus treating them as “fixed effects” in the panel model. The assumption of fixed effect, as opposed to the random effect, is widely acknowledged in the investment literature. Thus $\mathbf{X}_{it} = [1, \tilde{\mathbf{X}}_{it}, f_i, \tau_t]$ is defined, and the equations are rewritten as (purging each one of the f_i s and τ_t s to avoid multicollinearity)

$$y_{it} = \mathbf{X}_{it} \beta + \epsilon_{it} \quad (14a)$$

and

$$\epsilon_{it} = v_{it} - u_{it}, \quad (15a)$$

where v_{it} and u_{it} are independent among themselves and are also independent of \mathbf{X}_{it} . The model comprising (14a), (15a), (16), and (17) is essentially a stochastic frontier model.

The econometric estimation of a stochastic frontier model was introduced by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), who assumed a half-normal distribution for the one-sided error. But the assumed distribution is restrictive, because it implies that the mode of the one-sided error is at 0. Stevenson (1980) specified a model with a truncated normal distribution that has a mode not necessarily equal to 0. It is more general and includes the half-normal as a special case. Kumbhakar, Ghosh, and McGuckin (1991), Huang and Liu (1994), and Battese and Coelli (1995) further generalized the model so that the mode of the distribution shifts with observation-specific variables rather than being a constant.

Our foregoing model is based on work of Kumbhakar et al. (1991), Huang and Liu (1994), and Battese and Coelli (1995) with two further extensions designed to minimize model misspecifications. The first extension is to incorporate the firm-specific effect (f_i) and the aggregate time effect (τ_t) in the model. Without firm-specific effects, the model effectively treats multiple observations of the same firm as being obtained from independent samples, leaving the data's panel nature unexploited. As emphasized in studies by Kumbhakar (1991) and Kumbhakar and Hjalmarsen (1995), failure to include firm-specific effects in a panel stochastic frontier model is also likely to bias the estimate of the one-sided error, u_{it} , which is one

of the most important elements of the estimation. The reason for this is because the measure of u_{it} is based on the composite error term, which in turn is influenced by the parameter estimates of the frontier function. The need to use time effects can be argued similarly. The firm and time effects are treated as fixed effects in the model. Because of the truncated error distribution, one cannot take first differences or subtract means from the data to eliminate the effects; differenced truncated normal distributions do not result in a known distribution. Instead, the dummy variable approach is used as suggested in the formulation of Kumbhakar (1991). Although the dummy variable approach could be impractical for a very large number of cross-sections, with the N and T (184 and 8), one can include all of the dummies and still manage the estimation.

The second extension is to use a flexible approach to model heteroscedasticity in u_{it} . Whereas heteroscedasticity may affect estimation efficiency only in a linear regression model, it leads to biased estimates in a stochastic frontier model, in which a part of the error is distributed asymmetrically (Caudill and Ford 1993; Caudill, Ford, and Gropper 1995; Hadri 1999). Because u_{it} has a truncated normal distribution, its variance is a function of both μ_{it} and σ_{it}^2 ; therefore, heteroscedasticity of u_{it} can be modeled through a nonconstant, μ_{it} , a nonconstant, σ_{it}^2 , or both. The approach of Kumbhakar et al. (1991), Huang and Liu (1994), and Battese and Coelli (1995) makes μ_{it} observation-specific. Caudill et al. (1995) kept μ_{it} constant (and equal to 0, i.e., a half-normal model) but allows σ_{it}^2 to be observation-specific. The model here allows both μ_{it} and σ_{it}^2 to be observation-specific. As is shown in the estimation result, this added flexibility is not redundant, but rather improves the estimation significantly.

More specifically, variables in (17) are parameterized as

$$\mu_{it} = c0 + \mathbf{Z}_{it}\delta \quad (18)$$

and

$$\sigma_{it}^2 = \exp(c1 + \mathbf{Z}_{it}\gamma), \quad (19)$$

where $c0$ and $c1$ are constant and \mathbf{Z} is a vector of nonstochastic variables. The parameterizations specify μ_{it} and σ_{it}^2 to be a function of the same variables but allows for different intercepts and slopes. That is, it is assumed that variables affecting the mean of the pretruncated distribution also influence the (log of the) variance of the distribution, although the effects are not necessarily the same and could even have opposite signs.

The log-likelihood function of observation y_{it} can be written as

$$\begin{aligned} \ln[f(y_{it})] = & -\frac{1}{2} \ln(\sigma_v^2 + \sigma_{it}^2) + \ln \left[\phi \left(\frac{y_{it} - \mathbf{X}_{it}\beta + \mu_{it}}{\sqrt{\sigma_v^2 + \sigma_{it}^2}} \right) \right] \\ & - \ln \left[\Phi \left(\frac{\mu_{it}}{\sigma_{it}} \right) \right] + \ln \left[\Phi \left(\frac{\check{\mu}_{it}}{\check{\sigma}_{it}} \right) \right], \quad (20) \end{aligned}$$

where

$$\check{\mu}_{it} = \frac{\sigma_v^2 \mu_{it} - \sigma_{it}^2 (y_{it} - \mathbf{X}_{it}\beta)}{\sigma_v^2 + \sigma_{it}^2}, \quad (21)$$

$$\check{\sigma}_{it}^2 = \frac{\sigma_v^2 \sigma_{it}^2}{\sigma_v^2 + \sigma_{it}^2}, \quad (22)$$

and ϕ and Φ are the probability density and cumulated density functions of a standard normal distribution.

As mentioned, one of the advantages of the stochastic frontier estimation is the ability to obtain quantitative measures of the one-sided effects. Based on these measures, one can then calculate the investment efficiency index (IEI), which measures the extent to which a firm's rate of investment is close to the frictionless and deterministic level. This is defined as $IEI_{it} = (\mathbf{X}_{it}\beta - u_{it})/\mathbf{X}_{it}\beta$ if the dependent variable is in the original unit. If the dependent variable is in logarithms, then the measure is modified to

$$IEI_{it} = \frac{\exp(\mathbf{X}_{it}\beta - u_{it})}{\exp(\mathbf{X}_{it}\beta)} = \exp(-u_{it}). \quad (23)$$

This has a value between 0 and 1, with 0 ($u_{it} \rightarrow \infty$) indicating the least efficient, and 1 ($u_{it} = 0$) the most efficient. To make the measure operational, the expectation of the index conditional on the estimates is (Battese and Coelli 1988)

$$\begin{aligned} E(\exp(-u_{it})|\epsilon_{it} = \hat{\epsilon}_{it}) \\ = \exp(-.5(2\check{\mu}_{it} - \check{\sigma}_{it}^2)) \frac{\Phi(\frac{\check{\mu}_{it}}{\check{\sigma}_{it}} - \check{\sigma}_{it})}{\Phi(\frac{\check{\mu}_{it}}{\check{\sigma}_{it}})}. \quad (24) \end{aligned}$$

Estimations of the maximum likelihood function are carried out using Stata 6.0 computer software, which uses a combination of the steepest ascent and Newton–Raphson algorithms. The ordinary least squares results of the frontier-only function provide consistent estimates of β except for the intercept, and they are used as initial values; other parameters' initial values are set to 0. Small perturbations on the initial value vector were tried, and the results seem quite robust to the changes. Convergence is declared when either the change of the log-likelihood function value or the maximum relative change of the coefficient, defined later, is not larger than 10^{-6} (the default),

$$\max_k \left(\frac{|\mathbf{B}_{j+1}[k] - \mathbf{B}_j[k]|}{|\mathbf{B}_j[k]| + 1} \right).$$

In this expression, $\mathbf{B}_j[k]$ is the coefficient vector's k th element from the j th iteration. The number 1 is added to the denominator to avoid division by 0, as well as to provide a smooth transition between the percentage difference (when $\mathbf{B}_j[k] \rightarrow \infty$) and absolute difference (when $\mathbf{B}_j[k] \rightarrow 0$) of the coefficients. This algorithm is found to produce satisfying convergence property in terms of sizes of the first derivatives. For example, in model (ii) of Section 5 (the main model in the latter analysis), the largest gradient (in the absolute value) in the final iteration is 7.17×10^{-6} .

4. DATA AND MODEL SPECIFICATIONS

The empirical data are from the Taiwan Economic Journal Data Bank, which contains data of Taiwanese manufacturing firms publicly traded on the Taiwan Stock Exchange. The Data Bank is similar to the Compustat of the U.S., but most of the financial and asset data are collected only from 1981. The sample period covered in this study is from 1988 to 1996, but because lags were used in constructing estimation variables, the actual estimation period is from 1989 to 1996. Going further

back in the years is not likely to gain much, because the number of publicly traded firms is small before the stock market reform in 1988.

Firms in the construction industry and those with missing or unreasonable values (such as asset values equal 0) in the required variables are deleted. For each of the estimation samples, we require that each firm has at least three contiguous observations. The result is an unbalanced but contiguous panel of 184 firms and 1,220 observations. It is unbalanced because not every firm has the data beginning from 1989, but every firm has contiguous data that ends at 1996.

We pay great attention in constructing the replacement cost of capital variable (K), which is an important element in constructing the Q variable. Because the available data are relatively short term (only from 1981), the perpetual inventory method is inapplicable; most of the firms are likely to have been established before 1981, so 1981's book value of capital stock can be quite different from the replacement cost in that year. Instead the vintage structure method of Lewellen and Badrinath (1997), which does not require that data be available from the year of establishment, is used. (See the Appendix for a sketch of the application of this method on our data.)

The following variables for the model (14)–(19) are considered:

$$y_{it} : \ln(I_{it}/K_{it-1});$$

$$X_{it} : \ln(Q_{it}), \ln(Sales_{it}/K_{it-1}), \ln(Sales_{it-1}/K_{it-2}), f_i, \tau_i;$$

and

$$Z_{it} : (CF_{it}/K_{it-1}), \ln(Assets_{it}).$$

In the foregoing, I is capital investment measured by capital expenditures from cash flow statements. Variables in vector \mathbf{X}_{it} include the (log of) Q , the sales ratios, the firm fixed effect (f_i), and the time fixed effects (τ_i). As shown in the literature, Q is a sufficient statistic under the null of perfect capital markets (e.g., Hayashi 1985; Osterberg 1989; Chirinko 1993, p. 1892; Gilchrist and Himmelberg 1995, p. 551). To refine the estimation of the investment frontier, also considered are models incorporating the sales ratio variable, for which the numerator is the net sales from income statements. This variable captures the output effect (i.e., current sales predict future sales, prompting current investment), and may also capture effects of departures from constant returns to scale.

The output effect was also tested using the ratio of current to last period sales, which relates investment demand to changes rather than levels of sales. This variable turns out to be statistically insignificant, and hence the results are not reported.

The vector of \mathbf{Z}_{it} has two variables: the cash flow ratio variable (CF) and the total assets variable ($Assets$). The CF is from cash flow statements, based on the after-tax net income, and adjusts for noncash expenses and revenues to obtain the realized cash flow. $Assets$ is the value of total assets from the balance sheets. It is in billions of new Taiwan dollars at the 1991 price level. In that year, 25.75 new Taiwan dollars changed for 1 U.S. dollar.

✓ The use of the cash flow ratio and total assets to measure the degree of financing constraint warrants an explanation. As pointed out in the literature (e.g., Kaplan and Zingales 1997,

Table 1. Statistics of Regression Variables

	Mean	Median	Standard deviation
(I_{it}/K_{it-1})	.175	.107	.218
Q_{it}	3.696	3.024	2.345
$(Sales_{it}/K_{it-1})$	1.864	1.428	1.697
(CF_{it}/K_{it-1})	.127	.102	.253
$Assets_{it}$	8.522	4.447	11.287

* In billions of new Taiwan dollars at the 1991 price level. One new Taiwan dollar in 1991 exchanges for 1/25.75\$ U.S. dollars.

2000), two main factors cause firms to have different levels of investment in imperfect capital markets (holding investment demand identical). First, a firm's investment is higher (or less constrained) if it has a higher level of internal funds. The cash flow variable is used to capture this effect. Second, a firm's investment is less constrained if its informational problem is less severe. How severe the information problem is may depend on the firm's intrinsic characteristics, which may or may not be observable to economists. An oft-used proxy is the asset size (Gertler and Gilchrist 1994; Carpenter et al. 1994; Gilchrist and Himmelberg 1995). Firms with larger assets may have better means of providing collateral to mitigate the information problem. For a given industry, larger firms also tend to be older and more mature, so that the market usually has better access to, and assessment of, the firm's information.

The foregoing model has the unobservable firm and time effects (f_i , τ_i) in \mathbf{X}_{it} . Also considered is the alternative of excluding f_i and τ_i from \mathbf{X}_{it} and putting them instead in \mathbf{Z}_{it} , the function of inefficiency. This specification has a random-effects implication on f_i and τ_i , in the sense that they are in the model's composite error and are thus uncorrelated with \mathbf{X}_{it} by construction. (Having f_i and τ_i in both \mathbf{X}_{it} and \mathbf{Z}_{it} requires a maximum likelihood estimation of nearly 400 parameters, which exerts great numerical difficulty and thus is not considered as a practical alternative.) The nonnested hypothesis test of Vuong (1989) is used for this specification test. The test is structured in such a way that a positive test statistic indicates that the model of fixed effects in \mathbf{X}_{it} is favored, and a negative statistic favors the alternative. The test results in a positive statistic with a p value equal to .32, indicating a slight preference for the stated model.

Table 1 presents the summary statistics of the regression variables. Note that the average value of Q_{it} appears to be large, possibly due in part to the phenomenon of cross-shareholdings among publicly traded firms in Taiwan, a widely held view in the domestic market. Because the cross-held firms have stakes in one another, this creates misaligned incentives for the member firms to keep one another's share prices high.

5. ESTIMATION RESULTS

Table 2 gives the results of various model estimations. Estimates of the vast firm and time dummies are not listed. Models (i), (ii), and (iii) are the focus, and they estimate the unrestricted model of (14)–(19). Model (i) includes only the (log of) Q_{it} in the frictionless investment function; under the null of perfect capital markets, Q_{it} is a sufficient statistic. Models (ii)

Table 2. Estimation Results of Eq. (20)

	δ, γ unrestricted			$\delta = \gamma$	$\gamma = 0$	$\mu_{it} = 0$	$\gamma = \mu_{it} = 0$
	(i)	(ii)	(iii)	(iii-a)	(iii-b)	(iii-c)	(iii-d)
Investment function							
β_Q	1.213*** (.087)	.937*** (.105)	.819*** (.116)	.792*** (.116)	.801*** (.116)	.770*** (.118)	.753*** (.118)
β_S		.505*** (.111)	.051 (.153)	.079 (.153)	.078 (.152)	.107 (.154)	.164 (.153)
β_{S-1}			.793*** (.133)	.783*** (.134)	.804*** (.134)	.832*** (.134)	.790*** (.134)
Constraint function μ_{it}							
c_0	-.082 (1.336)	-.359 (1.488)	.130 (1.442)	-41.619 (71.926)	-4.589 (7.959)		
δ_C	-.358 (1.505)	.239 (1.711)	-.274 (1.547)	-.467** (.222)	-3.547 (4.444)		
δ_A	-3.826*** (1.409)	-4.090*** (1.552)	-3.613*** (1.380)	-.260*** (.061)	-4.498 (4.860)		
Constraint function σ_{it}^2							
c_1	.987** (.503)	1.010* (.527)	1.013* (.564)	3.738** (1.624)	2.046* (1.115)	.481 (.302)	-.201 (.321)
γ_C	-.492* (.294)	-.435 (.301)	-.487 (.322)	-.467** (.222)		-.159 (.609)	
γ_A	.350*** (.077)	.345*** (.078)	.278*** (.09)	-.260*** (.061)		-1.561*** (.637)	
σ_v	-1.313*** (.175)	-1.260*** (.164)	-1.458*** (.235)	-1.370*** (.240)	-1.351*** (.290)	-.734*** (.085)	-1.206*** (.278)
Log-likelihood value	-1,424.254	-1,414.028	-1,135.342	-1,142.568	-1,140.209	-1,156.766	-1,166.076

NOTE: The fixed firm and time effects in the investment function are not listed. Numbers in parentheses are standard errors. Because different numbers of lag variables are used, the number of observations of models (i) and (ii) are different from the rest, and therefore the log likelihood values are not directly comparable. A basic model without the one-sided effect is also estimated. The explanatory variables include the Q and the firm and time effects. The Q coefficient is 1.168, which is significant at the 1% level, and the corresponding log-likelihood value is -1,465.909.

*Significant at the 10% level. **Significant at the 5% level. ***Significant at the 1% level.

and (iii) also consider the effects of sales ratio variables on the frontier function, as discussed earlier.

Models (iii-a)–(iii-d) build on model (iii) by imposing restrictions on the modeling of heteroscedasticity. In particular, models (iii-b), (iii-c), and (iii-d) each correspond to an existing model specification in the literature. Model (iii-a) assumes $\delta = \gamma$, requiring variables in Z_{it} to have the same coefficients in both the mean and the variance functions of the pretruncated distribution. Model (iii-b) assumes $\gamma = 0$, which corresponds to the specification of Battese and Coelli (1995). Model (iii-c) assumes $c_0 = \delta_C = \delta_A = 0$, which is the half-normal model with heteroscedastic variances proposed by Caudill, Ford, and Gropper (1995). Model (iii-d) is the original half-normal model proposed by Aigner et al. (1977). Note that this model does not permit exogenous influences (i.e., cash flow and total asset) to directly explain the one-sided deviation. Results from these models are to be compared with those from model (iii).

As shown in Table 2, the Q and Sales ratio variables in the investment functions are usually highly significant, except when a lag of the Sales ratio variable is also included, in which case, the contemporaneous Sales ratio does not appear to be important. In five of the seven models, the elasticity of investment with respect to Q is slightly less than 1; a 1% increase in Q leads to about .8%–.9% increases in the rate of investment. The effect of sales ratios on investment is also important, exerting about .5–.8 percentage points in terms of elasticity.

As a benchmark, a Q model without the one-sided deviation effect is also estimated (i.e., $\mu_{it} = \sigma_{it}^2 = 0$). This is essentially a

linear model with Q and fixed firm and time effects. The result (not shown in Table 2) has a coefficient of Q equal to 1.168, which is significant at the 1% level. The adjusted R^2 is .472, and the corresponding log-likelihood value is -1,465.909. That the Q coefficient is quite close to that of model (i) is not surprising, because estimates of the linear model are consistent estimates of the stochastic frontier model. The smaller log-likelihood value, in contrast, reveals the importance of allowing for one-sided deviations in the model. Discussions on this point ensue later in this section.

We are interested mainly in the estimates of the financing constraint effect u_{it} , which is decomposed into the mean function (μ_{it}) and the variance function (σ_{it}^2) of its pretruncated distribution. Except in model (iii-d), the mean and variance functions are parameterized by cash flow and total asset variables. The reported coefficients are not very informative, however, because they are not the marginal effects due to the model's nonlinearity. Even the sign of a variable's latent marginal effect is difficult to tell from the slope coefficients, because the marginal effect depends on estimates in both the μ_{it} and σ_{it}^2 functions. Therefore, the marginal effects and bootstrap bias-corrected confidence intervals are calculated for all of the models except (iii-d). The results, along with some other hypothesis test results, are given in Table 3. Formulas of the marginal effects are provided in the Appendix. The focus is on the signs of the marginal effects.

Table 3 shows that the point estimates of the marginal effects of cash flow and total asset are all negative on both the

Table 3. Marginal Effects and Significance Testings

	(i)	(ii)	(iii)	(iii-a)	(iii-b)	(iii-c)
Marginal effect on $E(u_{it})$						
Cash flow	-.305	-.207	-.287	-.285	-.174	-.028
95% CI ^a	[-.779 .006]	[-.678 .068]	[-1.055 -.001]	[-1.106 -.010]	[-.599 -.003]	[-.276 .231]
90% CI	[-.724 -.043]	[-.622 .026]	[-.833 -.048]	[-1.007 -.051]	[-.496 -.035]	[-.206 .208]
85% CI	[-.664 -.081]	[-.594 -.013]	[-.760 -.083]	[-.947 -.085]	[-.468 -.052]	[-.164 .168]
Total assets	-.205	-.201	-.238	-.159	-.220	-.274
95% CI	[-.404 -.030]	[-.383 -.003]	[-.470 -.021]	[-.334 .141]	[-.409 .060]	[-.473 -.026]
90% CI	[-.368 -.048]	[-.340 -.027]	[-.426 -.042]	[-.291 .141]	[-.345 .034]	[-.449 -.058]
85% CI	[-.339 -.055]	[-.311 -.039]	[-.392 -.067]	[-.257 .067]	[-.313 -.001]	[-.437 -.076]
Marginal effect on $V(u_{it})$						
Cash flow	-.330	-.223	-.305	-.306	-.202	-.017
95% CI	[-1.248 -.009]	[-1.102 .041]	[-1.286 -.014]	[-1.408 -.014]	[-1.056 -.041]	[-.255 .311]
90% CI	[-1.128 -.053]	[-.934 .007]	[-1.195 -.057]	[-1.331 -.037]	[-1.056 -.067]	[-.159 .203]
85% CI	[-1.035 -.091]	[-.831 -.019]	[-1.090 -.094]	[-1.292 -.062]	[-1.056 -.096]	[-.124 .160]
Total assets	-.161	-.158	-.220	-.220	-.256	-.171
95% CI	[-.396 -.035]	[-.432 -.034]	[-.531 -.067]	[-.343 -.042]	[-.596 -.138]	[-.539 .003]
90% CI	[-.350 -.055]	[-.372 -.053]	[-.453 -.093]	[-.318 -.058]	[-.520 -.164]	[-.446 -.021]
85% CI	[-.315 -.068]	[-.329 -.070]	[-.404 -.104]	[-.295 -.071]	[-.459 -.185]	[-.405 -.037]
LR tests ($\chi^2_{(f)}$) ^b						
Absence of cash flow in u_{it}	7.063 $p = .029$	3.877 $p = .144$	5.069 $p = .079$	4.730 $p = .099$	2.485 $p = .115$.068 $p = .734$
Absence of total assets in u_{it}	35.850 $p = 0$	37.151 $p = 0$	36.308 $p = 0$	21.856 $p = 0$	28.931 $p = 0$	16.016 $p = 0$

^aThe bias-corrected confidence intervals are bootstrapped from 1,000 replications.

^b f is the degree of freedom, which equals 2 for models (i), (ii), (iii), and (iii-a) and 1 for models (iii-b) and (iii-c).

mean and the variance of the financing constraints. The statistical significance is revealed by the confidence intervals. For the unrestricted models, model (iii)'s negative marginal effects are all significant at the 5% level, model (i)'s are significant at the 10% level, and model (ii)'s are at least significant at the 15% level. In general, the total asset's effects on the mean and the variance are both very robust, and the cash flow's effects are more robust on the variance than on the mean.

Bearing in mind some of the marginal cases, the negative effects of cash flow on the mean and variance are of particular interest. The results indicate that in general, acquiring additional cash not only is likely to reduce levels of financing constraints, but also decreases the uncertainty of the constraints. The latter effect tends to be more significant than the former according to the confidence intervals. This is the first study in the literature that documents the cash flow's second-order effect on the financing constraint. Unlike the first-order effect of increasing the level of investment, this second-order effect cannot be explained by the cash flow's expectation factor. This provides appealing evidence for the financing constraint hypothesis. Similarly, the negative marginal effects of assets state that the level and variance of financing constraints are smaller for larger firms.

The significance of the cash flow and total asset variables in the inefficiency function was also evaluated. Unlike the confidence interval analysis, this tests the variable's overall significance on financing constraint u_{it} , irrespective of whether the effect is on the mean or on the variance. The null hypothesis is that the variable has no effect on u_{it} . A generalized likelihood ratio (LR) test, defined as $-2[L(H_0) - L(H_1)]$, is used for this purpose, where $L(H_0)$ and $L(H_1)$ are values of the log-likelihood functions under the null and the alternative hypotheses. The statistic has an asymptotic chi-squared distribution with the degrees of freedom equal to the number of re-

strictions. The results are given in the lower panel of Table 3. For the cash flow ratio variable, the worst case is model (iii-c), for which the variable has little statistical justification. As discussed later, however, restrictions imposed on model (iii-c) are overwhelmingly rejected, suggesting that model (iii-c) is misspecified. For the other five cases, three of them have p values below .10. For the total asset variable, the p values are all below .01.

All of the foregoing discussions are based on the presumption that the one-sided effect of u_{it} is significant in the model, which is a testable hypothesis. If u_{it} is not significant, then it should be dropped from (15), and the model reduces to a neoclassical one with classical disturbances. Furthermore, given the maintained hypothesis that u_{it} captures financing constraints, the test of u_{it} is also a test of whether the financing constraint is a statistically important phenomenon in explaining the investment behavior. Again, a LR test is used. The model under the null of no effect from u_{it} is composed of (14)–(16) with no u_{it} in (15), which can be estimated by the traditional panel data method, such as the least squares dummy variable estimator. Because the null hypothesis implies that the variance of u_{it} is 0, which is on the boundary of the admissible parameter space, the correct asymptotic of the LR statistic is a mixture of chi-squared distributions (Coelli 1995). Following Coelli and Battese (1996), critical values of the mixed chi-squared distribution are obtained from table 1 of Kodde and Palm (1986). The degrees of freedom of this statistic equals the number of parameters used to parameterize the distribution of u_{it} . For models (i)–(iii-a), this number is 6, which has a critical value of 16.074 at the 1% significance level. For model (iii-b), the number is 4, and the critical value is 12.483. For model (iii-c), the number is 3, and the critical value is 10.501. For model (iii-d), the number is 1, and the critical value is 5.412. The LR test results in the statistics equal to 83.311, 77.633, 86.114, 71.662, 76.381, 43.266,

and 24.646 for models (i)–(iii-d). The null hypotheses of no effect from u_{it} and, equivalently, no financing constraint effect on investment behavior, are convincingly rejected for all of the models.

Finally, LR tests are performed for four separate hypotheses: $\delta = \gamma$, $\gamma = 0$, $\mu_{it} = 0$, and $\gamma = \mu_{it} = 0$, which are the restrictions on models (iii-a)–(iii-d). On the one hand, these test the validity of the restrictions. On the other hand, they amount to testing whether the cash flow and total asset variables are important for μ_{it} and σ_{it}^2 . The null hypothesis is that the restriction is valid, and the alternative is the unrestricted specification of model (iii). The tests result in p values equal to .001, .008, 0, and 0; therefore, the hypotheses are decisively rejected. This had two implications. In terms of model specifications, an unrestricted, flexible model on heteroscedasticity is preferred over the four alternatives. In terms of financing constraint effects, the cash flow and total asset variables appear to be important for each and both of μ_{it} and σ_{it}^2 of the pretruncated distribution.

5.1 Alternative Model Specifications

This section examines the possibility that the estimated results may be sensitive to different model specifications. In particular, two types of alternative model specifications are considered. For the first type, only lagged variables of Q and $Sales$ are included in the investment function. Had there been serious endogeneity problems among the contemporaneous variables of the investment, Q , and $Sales$, then the lagged variable approach may have been preferred. Whether this modification has any significant impact on the estimation results is investigated.

For the second type of alternative specification, it is assumed that the one-sided financing constraint effect, u_{it} , follows an exponential distribution. Together with half-normal and truncated-normal distributions already used in the models of Table 2, the exponential is one of the most often assumed distributions in a stochastic frontier analysis. In terms of a prior, note that Kumbhakar and Lovell (2000, p. 90) provided evidence that alternative distributions have nonconsequential effects on the stochastic frontier analysis; whether this holds true in the case herein will be investigated.

The single parameter of an exponential distribution is denoted by η ($\eta > 0$). Parallel to the cases of truncated- and half-normal distributions, η is parameterized as a function of observation-specific factors by $\eta_{it}^2 = \exp(c2 + \mathbf{Z}_{it}\theta)$. The log-likelihood function of observation y_{it} is (e.g., Kumbhakar and Lovell 2000)

$$\ln[f(y_{it})] = -\ln(\eta_{it}) + \ln\left[\Phi\left(-\frac{y_{it} - \mathbf{X}_{it}\beta}{\sigma_v} - \frac{\sigma_v}{\eta_{it}}\right)\right] + \frac{y_{it} - \mathbf{X}_{it}\beta}{\eta_{it}} + \frac{\sigma_v^2}{2\eta_{it}^2}. \quad (25)$$

The estimation results are given in Table 4. Models (L1) and (L2) are cases with lagged variables in the investment function. As is shown, estimates in the μ_{it} and σ_{it}^2 functions are in line with those in Table 2. The marginal effect analysis (not shown) also paints a very similar (if not better) picture; all of

the marginal effects of cash flow and total assets on the mean and the variance of u_{it} are significantly negative at the 5% level.

Models (E1)–(E4) are cases with exponential distributions. Note that in the case of exponential distributions, the mean and the variance of u_{it} are η_{it} and η_{it}^2 . Therefore, signs of the marginal effects need not be recalculated. Again, the negative and significant marginal effects are consistent with the results of previous models.

It is also of interest to see whether rankings of the observation-specific IEI are sensitive to model specifications. The IEI is calculated according to (24), and it quantitatively measures the extent to which financing constraints affect the rate of investment. This issue is particularly relevant to the analysis in the next section, because the estimated IEI is to be compared across firms and over time periods. If rankings of the IEI vary substantially in different models, then the results of the comparisons will be sensitive to the choice of models. To this end, the Spearman rank correlation coefficients between the IEI from the models of Table 4 and from the unrestricted models of Table 2 are calculated. The results are given in Table 5. As shown, there is high agreement between the rankings from different models, with most of the correlation coefficients well above .90. Note that model (E4) has both of the alternative specifications; it uses lagged variables in the investment function and adopts the exponential distribution. Even for this case, the correlation coefficients with the compared models are no lower than .88.

The foregoing evidence ensures that results of this study's analysis are not likely to be significantly altered if different model specifications were to be adopted.

6. POSTESTIMATION ANALYSIS

This section uses the estimated IEI in two different analyses. In the first analysis, the IEI is used to investigate the validity of some of the sorting criteria used in the literature. In the second analysis, the IEI is used to evaluate the effects of financial liberalization in Taiwan during the sample period. The results can be inferred to provide additional evidence of the financing constraint hypothesis.

The analyses are based on the estimated IEI of model (ii). It can be shown that model (ii) is statistically preferable to model (i) from a LR test. It also retains one more year of data than model (iii), because the latter uses an additional lag variable. The conclusions would change little if the IEI from either model (i) or (iii) were used. Indeed, the IEI from the three models are very close; the Spearman rank correlation coefficients between the three IEIs are .991 for models (i) and (ii), .955 for models (ii) and (iii), and .964 for models (i) and (iii). Models (iii-a)–(iii-d) are not considered, because, as shown previously, the simplifying assumptions on heteroscedasticity used in these models are rejected by the data.

Figure 1 plots the frequency distribution of the IEI. The distribution is skewed to the right with the mean equal to .606 and the standard deviation equal to .190. The mode of the distribution is around .7–.8, indicating a loss of 20%–30% of the rate of investment due to financing constraints.

Table 4. Results of Alternative Model Specifications

	Lagged variables in Investment function		Exponential distributions			
	(L1)	(L2)	(E1)	(E2)	(E3)	(E4)
Investment function						
β_Q			1.195*** (.087)	.923*** (.105)	.791*** (.115)	
β_{Q-1}	.893*** (.098)	.420*** (.116)				.414*** (.117)
β_S				.505*** (.112)	.080 (.153)	
β_{S-1}		.922*** (.128)			.783*** (.134)	.922*** (.129)
Constraint function μ_{it}						
c_0	.604 (1.237)	.138 (1.429)				
δ_C	-.458 (1.137)	-.292 (1.435)				
δ_A	-2.923*** (1.077)	-3.417** (1.378)				
Constraint function δ_{it}^2						
c_1	.795 (.568)	.941 (.579)				
γ_C	-.808** (.328)	-.628* (.330)				
γ_A	.297*** (.100)	.267*** (.093)				
Constraint function η_{it}^2						
c_2			-.229 (.216)	-.294 (.221)	-.116 (.272)	-.186 (.261)
θ_C			-.935** (.383)	-.693* (.381)	-.938** (.451)	-1.177** (.485)
θ_A			-.412*** (.107)	-.434*** (.115)	-.524*** (.125)	-.558*** (.146)
σ_v	-1.295*** (.178)	-1.274*** (.173)	-1.225*** (.148)	-1.198*** (.154)	-1.335*** (.218)	-1.150*** (.158)
Log-likelihood value	-1,185.010	-1,159.742	-1,433.788	-1,423.689	-1,142.581	-1,165.882

NOTE: Models (L1) and (L2) estimate (20), and models (E1)–(E4) estimate (25); see also the footnotes of Table 2.

6.1 Investment Efficiency Index by Various Sorting Criteria

As discussed in Section 1, the literature often chooses a single criterion, such as the asset size, and then picks a critical value of the criterion to separate samples into groups believed to have different characterizations of financing constraints. The evidence of financing constraints is then inferred from the comparisons of cash flow coefficients across the groups.

Implicitly assumed in this approach, however, is that the severity of the financing constraint problem is significantly and monotonically related to the sorting criterion. This is necessary

so that no matter how the critical value is chosen to separate the samples, one of the two groups would be consistently more financially constrained than the other. If, however, the relationship between the financing constraint and the sorting criterion

Table 5. Spearman Rank Correlation Coefficients

	Models of Table 4					
	(L1)	(L2)	(E1)	(E2)	(E3)	(E4)
Unrestricted models of Table 2						
(i)	.928	.909	.980	.969	.936	.886
(ii)	.922	.918	.972	.979	.945	.895
(iii)	.944	.971	.935	.942	.981	.947

NOTE: The Spearman rank correlation coefficients between the estimated financing constraint effects (\hat{u}_{it}) from different models.

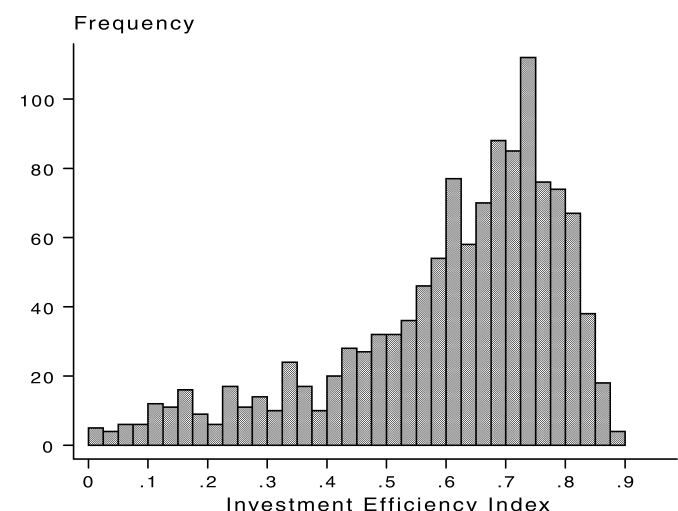


Figure 1. The Distribution of IEI.

Table 6. Investment Efficiency Index by Various Sorting Criteria:
The Average Statistics

Quintile	E(IEI)	Standard deviation	$\frac{\bar{C}}{K}$	Average assets	No. of observations
<i>By Assets^a</i>					
1st	.501	.084	.051	1.682	248
2nd	.613	.102	.129	3.110	245
3rd	.631	.116	.165	4.735	244
4th	.638	.131	.127	7.767	244
5th	.647	.139	.164	25.802	239
<i>By RR^b</i>					
1st	.635	.114	.228	10.206	250
2nd	.633	.114	.191	12.528	240
3rd	.604	.108	.117	7.760	244
4th	.597	.100	.109	7.289	244
5th	.558	.108	-.013	4.819	242
<i>By IIR^c</i>					
1st	.574	.107	.048	5.040	248
2nd	.609	.107	.272	8.256	245
3rd	.623	.097	.174	9.799	242
4th	.616	.116	.097	9.407	248
5th	.606	.119	.043	10.208	237
<i>By DAR^d</i>					
1st	.580	.101	.234	4.161	247
2nd	.582	.100	.142	6.251	242
3rd	.633	.106	.153	9.566	245
4th	.623	.119	.098	8.868	243
5th	.610	.122	.005	13.817	243

[‡] Average cash flow ratios.

^a Total assets, in billions of new Taiwan dollars at the 1991 price level.

^b Retention ratio: $100 \times (\text{net income after dividend})/(\text{net income})$.

^c Interest coverage ratio: $100 \times (\text{interest expenses})/(\text{net income} + .75 \times \text{interest expenses})$.

^d Debt-to-asset ratio: $100 \times (\text{total liability})/(\text{asset})$.

is inconsequential or nonmonotonic, then a chosen cutoff point is not guaranteed to separate samples into groups that have distinct financing constraint characteristics.

In this section the Taiwanese data are used to investigate the validity of some of the sorting criteria used in the literature. The approach is simple. Because IEI is a measure of financing constraints, the samples are simply classified into groups by the criterion, and then the sample mean of IEI of each groups is calculated. Comparing the averaged IEI and the associated statistics across the groups reveals whether the differences are significant and whether the relationships are monotonic. Because of the different data sources, the results are not a direct assessment of the current literature, for which the bulk of studies are based on U.S. data.

When sorting the samples, first the criterion's average values for each firm are calculated, and then the samples are classified into five quintiles with roughly equal numbers of firms. Because firms may have different numbers of observations (unbalanced panel), the numbers of observations in the quintiles are not exactly the same. Table 6 lists the results with some related statistics. Table 6 presents the standard deviations of the sample average figures, which are bootstrapped from 1,000 replications. For an easier presentation, the numbers are plotted and the graphs presented in Figure 2.

6.1.1 Asset Sizes. The first panel of Table 6 uses the asset size as the sorting criterion (e.g., Gertler and Gilchrist 1994; Carpenter et al. 1994; Gilchrist and Himmelberg 1995). Larger firms are assumed to be less likely to face obstacles in the capital market, because they tend to be older and more mature, and also to have better market recognition. Larger assets may also

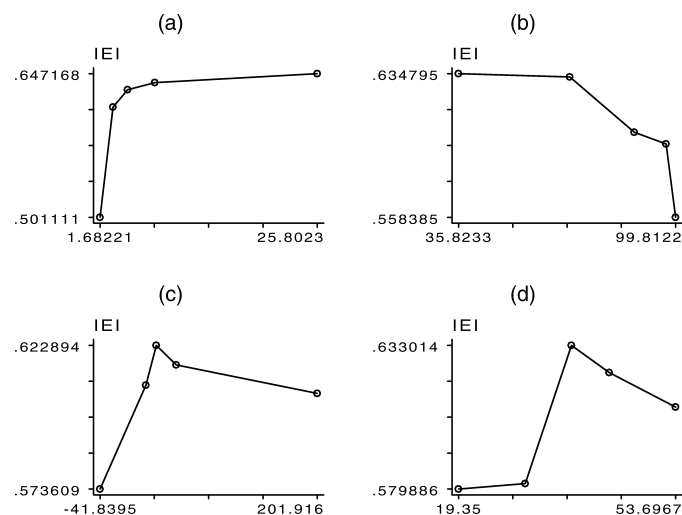


Figure 2. The Monotonicity of Various Sorting Criteria: (a) Size Quantiles; (b) Retention Ratio Quantiles; (c) Interest Income Ratio Quantiles; (d) Debt Asset Ratio Quantiles. The graphs plot the IEI against quantiles of the sorting criteria based on the numbers in Table 6. The literature's presumption is that asset sizes should be positively related to the IEI, whereas all of the other three criteria should be negatively related to the IEI.

enable them to provide collateral to mitigate information problems.

The result shows that the IEI is monotonically related to the size, increasing from .501 for firms in the first quintile to .647 for firms in the fifth quintile. This difference (.146) appears to be significantly different from 0; the bootstrapped 90% confidence interval, based on 1,000 replications, is between .014 and .366. The monotonic relationship can also be easily seen from the first graph in Figure 2(a). This confirms the findings of many related studies that larger firms tend to have higher investment efficiency.

Note that even for the most investment-efficient group (the first quintile), the average investment efficiency is only .647. This means that the group's average investment rate is about 64.7% of that of the hypothetically best practice firm. The deficiency is apparent and indicates the wide-spread financing constraint problem among all classes of firms.

6.1.2 Retention Ratios. The second panel of Table 6 uses the retention ratio as the sorting criterion (e.g., Fazzari, Hubbard, and Petersen 1988; Bond and Meghir 1994). A high retention ratio is likely to result from high opportunity costs of dividends, and it can be inferred that firms with higher opportunity costs of dividends are those facing more serious financing constraints.

The average IEI in the first quintile is .635, and the number declines monotonically over the quintiles to the lowest of .558; the graph in Figure 2 shows the monotonicity. The bootstrapped 95% confidence interval of the difference (-.076) is from -.163 to -.008. This result is in accordance with the prediction that high retention ratios are associated with financing constraint problems.

An oft-raised question, however, is whether the discriminatory power of the ratio results from its close approximation to firms' asset sizes. This concern seems to be partially supported by Table 6; the fifth column shows that higher retention firms

tend to be smaller ones, although an exception in the first two quantiles is also noted.

We further investigate this issue by estimating model (ii) with the retention ratio variable added into the functions of μ_{it} and σ_{it}^2 . The estimated coefficients of the variable are .002 and .004 in the functions of μ_{it} and σ_{it}^2 , both of which are statistically insignificant. An LR test of the joint significance of the ratio variable yields a chi-squared statistic equal to .403, and the null hypothesis of no effect cannot be rejected. Therefore, on controlling for the size effect, it is found retention ratios do not provide additional information in distinguishing constrained and unconstrained firms in the data.

6.1.3 Interest-to-Income Ratios. The third panel of Table 6 classifies samples by interest-to-income ratios (e.g., Whited 1992). The ratio indicates the likelihood of a firm in financial distress. If a firm had sufficient internal finance, then it would not have a great need to borrow.

Unlike in the previous cases, the ratio here does not seem to be monotonically related to the average IEI. Instead, a look at the graph in Figure 2 suggests a concave relationship, with the first quintile having the lowest value of investment efficiency (.574, probably owing to underutilized borrowing capacities) and the third quintile having the highest (.623). Based on these point estimates, moderate ratios are favored in terms of investment efficiency. For statistical significance, the difference between the first and the third quintiles is marginally significant at the 15% level (again from bootstrapping), but the difference between the third and the fifth quintiles is not appreciable at reasonable significant levels.

What does this result imply? As noted, the literature hypothesizes that firms with higher interest-to-income ratios have greater problems financing investment, implying a negatively sloped line in the IEI to the interest-to-income ratio plot. The results herein, however, show that the empirical line is likely to be *positively* sloped when the ratio is relatively low and that the variable may not be able to discern degrees of financing constraints when the ratio is relatively high. This makes the interest-to-income ratio an unappealing sorting criterion when applied to the Taiwanese data.

6.1.4 Debt-to-Asset Ratios. The fourth panel of Table 6 uses the debt-to-asset ratio as the sorting criterion (e.g., Whited 1992). This ratio is a measure of current demand for borrowing relative to the firm's debt capacity.

Similar to the interest-to-income ratio, the average IEI is at the highest, .633, in the middle quintile, indicating a possible concave relationship between the criterion and the degree of financing constraint. Again, bootstrapped confidence intervals show that the difference is significant at the 15% level between the first and the third quintiles, but statistically insignificant between the third and the fifth quintiles. The findings are similar to those of the previous case. The literature has presumed a negative relationship between IEI and debt-to-asset ratios, but the results indicate that the relationship is either positive (quintiles 1–3) or inconsequential (quintiles 3–5).

The foregoing analysis points to asset sizes and retention ratios as better sorting criteria for Taiwanese data, although it also shows that retention ratios obtain the discriminatory power mainly from their associations with asset sizes. The other criteria (the interest-to-income ratio and the debt-to-asset ratio) do

not seem to have the predicated relationships with investment efficiency.

This section concludes with a few more regressions of the augmented linear investment model as used by Carpenter et al. (1994). As we stated in Section 1, at least two different aspects of this approach have drawn criticism. One aspect is that concerning the fundamentals of the unstructured linear models, and the other is the justification of using various sorting criteria to split samples. In light of the results of the better sorting criteria, assessing the linear model's ability to capture financing constraint effects after the issue of sorting can be set aside would be desirable.

The samples are split into three groups, based on asset size and retention ratios. A standard fixed-effect panel data model is estimated for each of the groups. This model is

$$\begin{aligned} \ln\left(\frac{I_{it}}{K_{it-1}}\right) = & \beta_Q \ln(Q_{it}) + \beta_S \ln\left(\frac{Sales_{it}}{K_{it-1}}\right) \\ & + \beta_{S-1} \ln\left(\frac{Sales_{it-1}}{K_{it-2}}\right) + \beta_C \left(\frac{CF_{it}}{K_{it-1}}\right) \\ & + \beta_{C-1} \left(\frac{CF_{it-1}}{K_{it-2}}\right) + \Psi' \xi + v_{it}, \end{aligned} \quad (26)$$

where Ψ is a vector of firm and time dummies and ξ is the corresponding coefficient vector. The error term v_{it} is symmetrically distributed with mean 0. According to Carpenter et al. (1994), the cash flow coefficients should indicate the severity of financing constraints. The estimation results are given in Table 7.

For samples classified by firm sizes, Table 7 shows that larger firms' investment tends to respond more strongly to cash flows. This result matches the findings of Kaplan and Zingales (1997, 2000), and it is the opposite of Carpenter et al.'s hypothesis. For samples classified by retention ratios, the results also do not support Carpenter et al.'s hypothesis.

Linear investment models with augmented cash flows are not adequate to capture the effects of financing constraints on Taiwanese firms. The stochastic frontier approach gives evidence of financing constraints on smaller firms and on high-retention ratio firms; the same cannot be obtained using the linear regression approach.

6.2 Effects of Financial Liberalization

As described in Section 1, Taiwan's financial markets underwent liberalization reforms in the sample period. The theory predicts that the process should help release financing constraints for most of the firms, and for underprivileged firms in particular. Table 8 shows the average IEI across time for all of the firms, as well as for three groups of firms classified by asset sizes. The classification is based on the 33rd and 66th quantiles of the average size distribution.

The IEI of the total samples is the lowest in 1990 (.578); it rises in the next 2 years before experiences a small setback in 1993. The number is up again in 1994 and 1995, reaching the highest level of .631 in the sample period. This is to say that the loss of the rate of investment for an average firm was reduced by 5.3 percentage points from 1990 to 1995. This period chronicles

Table 7. Linear Investment Model of (26)

	By asset sizes			By retention ratios		
	Small	Median	Large	Small	Median	Large
β_Q	.729 (.260)***	.845 (.231)***	.71 (.249)***	1.335 (.226)***	.154 (.246)	.776 (.240)***
β_S	.343 (-.314)	.108 (-.293)	-.223 (-.315)	-.370 (.320)	.226 (.305)	.306 (.288)
β_{S-1}	.538 (.284)*	.709 (.244)***	1.275 (.273)***	.637 (.296)**	1.109 (.275)***	.555 (.253)**
β_C	.087 (-.285)	-.262 (-.277)	.431 (-.268)	.251 (.260)	.309 (.255)	-.102 (.361)
β_{C-1}	-.487 (-.313)	.102 (-.304)	.723 (.338)**	-.401 (.296)	.596 (.340)*	-.131 (.326)
Sum of cash flow coefficients	-.400 (.439)	-.160 (.443)	1.153 (.514)**	-.150 (.415)	.905 (.512)	-.233 (.501)
\bar{R}^2	.181	.252	.158	.107	.232	.163
Average criterion	2.056 (bill.)	4.846	15.909	.429	.874	.992
No. of observations	338	342	334	341	336	337

NOTE: The fixed firm and time effects are not listed. Numbers in parentheses are standard errors. Also see the footnotes in Table 2.

some important financial reforms in the Taiwanese markets, as discussed in Section 1. The average IEI dropped substantially, however, to .603 in 1996, the last year in the sample. This setback is likely to be triggered by the political and military intimidations of the Beijing government toward Taiwan in late 1995 and early 1996. The tension began from then-Taiwan President Lee Teng-Hui's visit to the United States in June 1995, and escalated with Beijing's missile tests off the coast of Taiwan in March 1996, around the day of the first direct presidential election in Taiwan. Beijing's military threat led the United States to send forces into the region. Undoubtedly, this event had adversely affected investors' and financiers' confidence.

The average IEI for firms of different sizes are given in the third, fourth, and fifth columns of Table 8, and these numbers are plotted in Figure 3 for easier presentation. The lines in Figure 3 are drawn through the medians of pairs of the data points, thus representing biannual trends of IEI.

The average IEI of the small firms was always lower than those of median- and large-sized firms. For the small firms, the number was at its lowest (.497) in 1989 and reached its highest level of .604 in 1995, which is an increase of about 10 percentage points in term of investment efficiency. Median-sized firms also showed continuous efficiency gains in the early 1990s, with

a gain of about 6.7 percentage points from the lowest to the highest. The group of large firms also had a maximum gain of 6.1 percentage points in the IEI in this period.

The combined graph in Figure 3(d) shows interesting comparisons. The lines of pairwise medians indicate an increases in the biannual trend of IEI for small firms in the sample period. The trend was also in the rise for medium-sized firms from 1990 to 1994, but the same cannot be said for the group of large firms. Indeed, the increasing trend of small firms has such an effect that the difference between large firms' and small firms' IEI narrowed from .138 in 1989 to .062 in 1995, although the number widened again in 1996.

The evidence presented in this section is favorable to the financing constraint hypothesis. In particular, the finding of the disproportional gain for smaller firms in the financial liberalization process is new and strong evidence of financing constraint that has not yet been documented in the literature.

Table 8. The Effects of Financial Liberalization on Firms of Different Sizes: A Comparison of Investment Efficiency Index Across Time

Year	All firms	Small firms	Medium firms	Large firms	No. of observations
1989	.580	.497	.596	.635	101
1990	.578	.525	.596	.605	115
1991	.604	.537	.620	.646	131
1992	.606	.546	.625	.644	148
1993	.598	.512	.663	.623	173
1994	.622	.572	.662	.638	184
1995	.631	.604	.627	.666	184
1996	.603	.527	.625	.666	184
No. of observations	1220	412	402	406	

NOTE: Samples are classified by the average asset sizes at the 33rd and 66th quantiles. They show that the IEI of small firms has increased more significantly over the period.

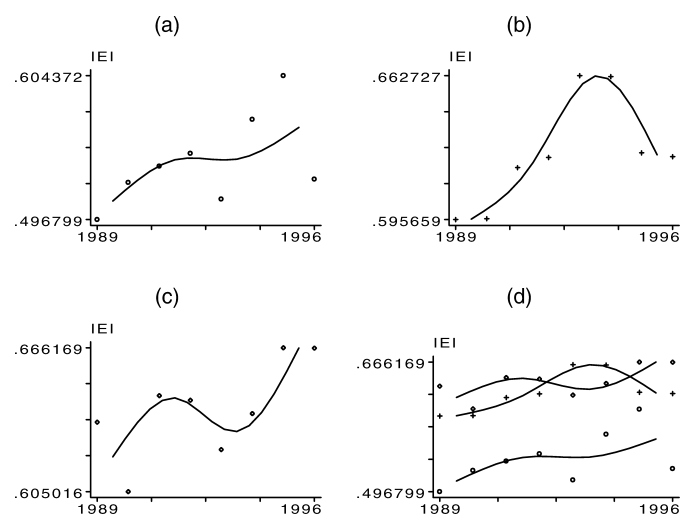


Figure 3. Small Firms Gained More From Financial Liberalization. The biannual trend lines are drawn through the medians of pairs of the data points. (a) Small-size firms; (b) medium-size firms; (c) large-size firms; (d) all firms.

7. CONCLUSION

This article has shown that investment of financing-constrained firms can be modeled as a one-sided deviation from a frictionless model. By imposing the distribution assumption on the constraint, the effect of financing constraints can be identified and quantified and some of the shortcomings of the regression method can be avoided.

The results are very encouraging. For cash flow, not only is it likely to promote the rate of investment in an environment of financing constraints, but it also has a strong effect on reducing the variance of financing constraints. This second-order effect is appealing evidence for the financing constraint hypothesis, because the oft-alleged role of cash's investment expectation cannot be applied here to explain this result.

This article has also investigated some of the oft-used criteria in the literature for splitting samples into a priori constrained and unconstrained groups. Of particular interest was whether the criteria are significantly and monotonically related to the degree of financing constraints. For the Taiwanese data, it was found that the asset size and the retention ratio are better criteria, and that the interest coverage ratio and the debt-to-asset ratio appear to be more problematic.

Comparisons of the investment efficiency index across time for different groups of firms are also revealing. Using Taiwan's financial liberalization as a natural experiment, it was shown that the investment efficiency improved for a typical firm during the process, and that the improvement was particularly significant for firms of smaller asset sizes.

ACKNOWLEDGMENTS

The author thanks Cliff Huang and Peter Schmidt for helpful discussions and Fung-Mey Huang, Ming Liu, James Tybout, and two referees for valuable comments. Financial support was provided by the National Science Council of Taiwan grant 89-2415-H-001-064.

APPENDIX: Replacement Cost of Capital and Marginal Effects

A.1 Replacement Cost of Capital

The method of Lewellen and Badrinath (1997) is used to uncover the vintage structure of period t 's net capital stock. It is sketched as follows (see Lewellen and Badrinath 1997) for more details. First, period t 's gross capital stock is obtained by adding up period t 's accumulated depreciation of capital and the net capital stock. Because the gross capital stock is the accumulation of past periods' gross capital investment, the gross capital investment are added up *backward* in time starting from period t until the sum of the investment equals the gross capital stock of period t . Equivalence occurring at period $t - N$ implies that period t 's capital stock is accumulated from period $t - N$. With assumed depreciation rates, the current value of each vintage of period t 's net capital stock can then be calculated.

The firms' periodic reports on revaluations of selected assets are also used to aid calculation of the market value of capital. In particular, because local regulation permits asset revaluations only when the price level has increased by 25% relative

to the purchasing year, this information is used in the interpolation of asset prices. Also, because land prices in Taiwan have increased disproportionately relative to other asset prices in the past decades, the land value is calculated separately from other fixed assets.

A.2 Marginal Effects on $E(u_{it})$ and $V(u_{it})$

For models (i), (ii), and (iii), the unconditional mean and variance of u_{it} are

$$E(u_{it}) = \sigma_{it} \left[\Lambda + \frac{\phi(\Lambda)}{\Phi(\Lambda)} \right] \quad (\text{A.1})$$

and

$$V(u_{it}) = \sigma_{it}^2 \left[1 - \Lambda \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right] - \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right]^2 \right], \quad (\text{A.2})$$

where μ_{it} and σ_{it}^2 are as specified in (18) and (19) and $\Lambda = \mu_{it}/\sigma_{it}$. The marginal effects of $\mathbf{Z}[k]$, the k th variable of the \mathbf{Z} vector in (18) and (19), can be derived by taking the derivatives of the foregoing functions with respect to $\mathbf{Z}[k]$. Tedious manipulations lead to

$$\begin{aligned} \frac{\partial E(u_{it})}{\partial z[k]} &= \delta[k] \left[1 - \Lambda \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right] - \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right]^2 \right] \\ &+ \gamma[k] \frac{\sigma_{it}}{2} \left[(1 + \Lambda^2) \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right] + \Lambda \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right]^2 \right], \quad (\text{A.3}) \end{aligned}$$

$$\begin{aligned} \frac{\partial V(u_{it})}{\partial z[k]} &= \frac{\delta[k]}{\sigma_{it}} \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right] ([E(u_{it})]^2 - V(u_{it})) \\ &+ \gamma[k] \sigma_{it}^2 \left\{ 1 - \frac{1}{2} \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right] \left(\Lambda + \Lambda^3 + (2 + 3\Lambda^2) \right. \right. \\ &\quad \left. \left. \times \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right] + 2\Lambda \left[\frac{\phi(\Lambda)}{\Phi(\Lambda)} \right]^2 \right) \right\}, \quad (\text{A.4}) \end{aligned}$$

where $\delta[k]$ and $\gamma[k]$ are the k th elements of the respective coefficient vectors. The formula is evaluated at every observation, and the sample average is reported. The marginal effects for models (iii-a)–(iii-c) can be derived similarly.

It is interesting to note that in (A.3) a nonparameterized σ^2 would imply a monotonic effect of $\mathbf{Z}[k]$ on $E(u_{it})$. To see this, note that $\gamma[k] = 0$ if σ^2 is constant, and that the function in the big square bracket of the remaining term is positive (more precisely, between 0 and 1, as shown in the literature of limited dependent variables). Thus the marginal effect takes the sign of $\delta[k]$, which is fixed for all values of $\mathbf{Z}[k]$. This in turn implies that $E(u_{it})$ is monotonic in $\mathbf{Z}[k]$. On the other hand, nonmonotonic relationships are accommodated if σ^2 is parameterized by \mathbf{Z} and $\gamma[k]$ is not 0.

[Received April 2001. Revised October 2001.]

REFERENCES

- Aigner, D., Lovell, C. A. K., and Schmidt, P. (1977), "Formulation and Estimation of Stochastic Frontier Production Function Models," *Journal of Econometrics*, 6, 21–37.
- Battese, G. E., and Coelli, T. J. (1988), "Prediction of Firm-Level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data," *Journal of Econometrics*, 38, 387–399.
- (1995), "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data," *Empirical Economics*, 20, 325–332.
- Bond, S., and Meghir, C. (1994), "Dynamic Investment Models and the Firm's Financial Policy," *Review of Economic Studies*, 61, 197–222.
- Carpenter, R. E., Fazzari, S. M., and Petersen, B. C. (1994), "Inventory Investment, Internal-Finance Fluctuations, and the Business Cycle," *Brookings Papers on Economic Activity*, 2, 75–137.
- Caudill, S. B., and Ford, J. M. (1993). "Biases in Frontier Estimation due to Heteroscedasticity," *Economics Letters*, 41, 17–20.
- Caudill, S. B., Ford, J. M., and Gropper, D. M. (1995), "Frontier Estimation and Firm-Specific Inefficiency Measures in the Presence of Heteroscedasticity," *Journal of Business and Economic Statistics*, 13, 105–111.
- Chirinko, R. S. (1993), "Business Fixed Investment Spending: Modeling Strategies, Empirical Results, and Policy Implications," *Journal of Economic Literature*, 31, 1875–1911.
- (1997), "Finance Constraints, Liquidity, and Investment Spending: Theoretical Restrictions and International Evidence," *Journal of the Japanese and International Economies*, 11, 185–207.
- Chirinko, R. S., and Schaller, H. (1995), "Why Does Liquidity Matter in Investment Equations?," *Journal of Money, Credit, and Banking*, 27, 527–548.
- Coelli, T. (1995), "Estimators and Hypothesis Tests for a Stochastic Frontier Function: A Monte Carlo Analysis," *Journal of Productivity Analysis*, 6, 247–268.
- Coelli, T., and Battese, G. (1996), "Identification of Factors Which Influence the Technical Inefficiency of Indian Farmers," *Australian Journal of Agricultural Economics*, 40, 103–128.
- Fazzari, S. M., Hubbard, R. G., and Petersen, B. C. (1988), "Financing Constraints and Corporate Investment," *Brookings Papers on Economic Activity*, 1, 141–195.
- Gertler, M., and Gilchrist, S. (1994), "Monetary Policy, Business Cycles, and the Behavior of Small Manufacturing Firms," *Quarterly Journal of Economics*, 109, 309–340.
- Gilchrist, S., and Himmelberg, C. P. (1995), "Evidence on the Role of Cash Flow for Investment," *Journal Of Monetary Economics*, 36, 541–572.
- Hadri, K. (1999), "Estimation of a Doubly Heteroscedastic Stochastic Frontier Cost Function," *Journal of Business and Economic Statistics*, 17, 359–363.
- Hayashi, F. (1985), "Corporate Finance Side of the Q Theory of Investment," *Journal of Public Economics*, 27, 261–280.
- Hu, X., and Schiantarelli, F. (1998), "Investment and Capital Market Imperfections: A Switching Regression Approach Using U.S. Firm Panel Data," *Review of Economics and Statistics*, 80, 466–479.
- Huang, C. J., and Liu, J. T. (1994), "Estimation of a Non-Neutral Stochastic Frontier Production Function," *Journal of Productivity Analysis*, 5, 171–180.
- Hubbard, R. G. (1998), "Capital-Market Imperfections and Investment," *Journal of Economic Literature*, 36, 193–225.
- Hubbard, R. G., Kashyap, A. K., and Whited, T. M. (1995), "Internal Finance and Firm Investment," *Journal of Money, Credit and Banking*, 27, 683–701.
- Kaplan, S. N., and Zingales, L. (1997), "Do Investment-Cash Flow Sensitivities Provide Useful Measures of Financing Constraints," *Quarterly Journal of Economics*, 112, 169–215.
- (2000), "Investment-Cash Flow Sensitivities Are Not Valid Measures of Financing Constraints," *Quarterly Journal of Economics*, 115, 707–712.
- Kodde, D. A., and Palm, F. C. (1986), "Wald Criteria for Jointly Testing Equality and Inequality Restrictions," *Econometrica*, 54, 1243–1248.
- Kumbhakar, S. (1991), "Estimation of Technical Inefficiency in Panel Data Models With Firm- and Time-Specific Effects," *Economics Letters*, 36, 43–48.
- Kumbhakar, S., and Lovell, C. A. K. (2000), *Stochastic Frontier Analysis*, Cambridge, U.K.: Cambridge University Press.
- Kumbhakar, S. C., Ghosh, S., and McGuckin, J. T. (1991), "A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in U.S. Dairy Farms," *Journal of Business and Economic Statistics*, 9, 279–286.
- Kumbhakar, S. C., and Hjalmarsson, L. (1995), "Labour-Use Efficiency in Swedish Social Insurance Offices," *Journal of Applied Econometrics*, 10, 33–47.
- Lewellen, W. G., and Badrinath, S. G. (1997), "On the Measurement of Tobin's q," *Journal of Financial Economics*, 44, 77–122.
- Meeusen, W., and van den Broeck, J. (1977), "Technical Efficiency and Dimension of the Firm: Some Results on the Use of Frontier Production Functions," *Empirical Economics*, 2, 109–122.
- Osterberg, W. P. (1989), "Tobin's q, Investment, and the Endogenous Adjustment of Financial Structure," *Journal of Public Economics*, 40, 293–318.
- Stevenson, R. E. (1980), "Likelihood Functions for Generalized Stochastic Frontier Estimation," *Journal of Econometrics*, 13, 57–66.
- Vuong, Q. H. (1989), "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.
- Whited, T. M. (1992), "Debt, Liquidity Constraints, and Corporate Investment: Evidence from Panel Data," *The Journal of Finance*, 47, 1425–1460.

融资约束、不确定性与上市公司投资效率

连玉君¹ 苏 治²

(1. 中山大学岭南学院、中山大学经济研究所, 广州 510275;

2. 清华大学经济管理学院, 北京 100084)

摘要:本文以异质性随机前沿模型为基础, 定量测算了中国上市公司在融资约束情况下的投资效率。结果表明: (1) 融资约束的存在使得中国上市公司的投资支出比最优水平低了约 20-30%, 平均投资效率仅为 72%。 (2) 在上市公司的三种主要融资方式中, 现金流量的增加不但能缓解融资约束, 还能降低后续融资的不确定性; 而股权融资和债务虽然能够有效缓解融资约束, 但前者无法降低融资不确定性, 而后者会显著加剧融资不确定性。 (3) 大规模公司和东部地区上市公司面临的融资约束和融资不确定性较低, 而小规模公司和西部地区上市公司的融资约束有逐渐加剧的倾向。

关键词:投资效率; 融资约束; 现金流; 随机前沿模型

引 言

中国的转型经济特征使得资本市场虽然初具规模, 但仍然存在结构性缺陷, 如股市缺乏有效性、公司债券市场畸形发展、银行贷款的信贷歧视等。这使得生存于其中的上市公司往往面临融资约束, 进而在很大程度上降低了投资效率。从理论上讲, 上述因素都可以归结为资本市场缺陷, 有悖于传统投资理论(如 Q 投资理论)的基本假设——资本市场完美无缺。因此, 在研究中国上市公司投资行为的过程中, 我们必须纳入融资约束的考量。相对前期文献仅仅探讨融资约束是否影响上市公司投资行为这一问题, 我们更为关注的是, 它在多大程度上影响着上市公司的投资行为? 而目前的金融体系又是如何影响企业的投资行为的? 对这些问题的分析将为转轨时期金融体系的改革和创新提供相应的微观基础。

对于融资约束是否会影响公司投资行为这一问题, 国外最具代表性的研究当属 Fazzari 等(1988), 其基本检验策略是在分组的基础上考察投资—现金流量敏感性差异。国内学者采用相似的方法对中国上市公司的投资行为进行了研究, 如冯巍(1999)、郑江淮等(2001)、梅丹(2005)、连玉君和程建(2007), 但观点并不一致。另一些学者则试图从现金—现金流敏感性角度进行研究, 如章晓霞和吴冲锋(2006)、李金等(2007)、连玉君等(2008), 但同样未达成一致看法。虽然样本筛选、估计方法上的差异可能导致观点分歧, 但上述研究的局限性也非常明显: 其一, 在对样本进行分组过程中, 单一分组指标可能无法区分不同公司所面临的融资约束差异, 而采用多变量分组又容易产生内生性问题; 其二, 多数研究都依据投资—现金流量敏感性这一现象来判断融资约束的存在性, 但大量研究表明融资约束并非导致这一现象的唯一原因, 当代理问题比较严重时, 公司同样

收稿日期: 2008-03-24

基金项目: 中山大学文科青年教师科研基金项目(3171913); 中国博士后科学基金项目(20070410539); 国家自然科学基金项目(70573040); 国家社会科学基金项目(06CJL006)。

作者简介: 连玉君, 中山大学岭南学院讲师; 苏治, 清华大学经济管理学院博士后。

会表现出投资-现金流敏感性(Pawlina and Renneboog, 2005; 连玉君和程建, 2007); 其三, 也是更为重要的是, 上述研究方法都无法对融资约束造成的投资效率损失进行定量估算, 而同时也都没有考察不确定性对投资行为的影响。

为此, 本文在异质性随机前沿模型框架下同时进行了定性和定量两个层面的分析。不同于前期从投资-现金流敏感性角度入手的实证研究, 该方法无需对样本公司进行分组, 同时又可以避免前期研究判断标准过于模糊的缺陷。实证结果表明, 在 2001-2006 年样本区间内, 融资约束的存在使得中国上市公司的投资支出比最优水平低了 20-30%, 平均投资效率仅为 72%。进一步分析表明, 小规模和西部上市公司面临的融资约束问题有日益加重的趋势, 地区金融发展水平, 尤其是信贷发展水平对上市公司的投资效率有显著影响。

文章的后续安排如下: 第二部分建立融资约束假设下的随机前沿模型, 第三部分介绍实证检验方法, 第四部分呈现实证结果, 最后做出总结。

异质性随机前沿投资模型

传统 Q 投资理论表明, 在资本市场完美假设下, 公司的投资支出仅决定于投资机会 (Hayashi, 1982), 其最优投资支出可表示为:

$$I_{it}^* = \beta_0 + (1/\alpha) Q_{it} + v_{it} \quad (1)$$

其中, I_{it} 为投资支出, $(1/\alpha)$ 为资本的调整系数, Q_{it} 为投资机会, v_{it} 为来自外部的技术冲击。然而, 中国资本市场并不完美, 结构性的缺陷导致公司在外部融资过程中面临各种限制, 此时公司的投资行为可以采用 Chrinko and Schaller (1995) 设定的投资模型加以描述:

$$I_{it} = \beta_0 + (1/\alpha) Q_{it} - F(\mathbf{z}_{it}) + v_{it} \quad (2)$$

其中, $F(\mathbf{z}_{it})$ 表示由于资本市场不完善导致的融资约束的大小, 它是一系列反映公司特征的财务变量的非线性函数。根据 (1) 和 (2) 式, 公司在不存在融资约束和存在融资约束两种情况下的投资支出之间存在如下关系:

$$E[I_{it} | Q_{it}, F(\mathbf{z}_{it}) = 0] > E[I_{it} | Q_{it}, F(\mathbf{z}_{it}) > 0] \quad (3)$$

因此, 融资约束的存在只会使公司的投资支出降低, 具有单边 (one-sided) 分布的特征。若设 $F(\mathbf{z}_{it}) = u_{it}$, 则实际投资支出 I_{it} 与最优水平 I_{it}^* 之间存在如下关系:

$$I_{it} = I_{it}^* - u_{it} = \beta_0 + (1/\alpha) Q_{it} + v_{it} - u_{it} \quad (4)$$

模型 (4) 是一个典型的随机前沿模型。为了反映面板数据的特征以及不同公司所面临的融资约束的异质性, 本文对模型 (4) 作了如下设定:

$$I_{it} = \mathbf{x}_{it}'\beta + \varepsilon_{it} \quad \varepsilon_{it} = v_{it} - u_{it} \quad (5)$$

其中 $\mathbf{x}_{it} = (1, Q_{it}, D_i, D_t)'$, β 为相应的系数向量, D_i 和 D_t 分别为反映个体效应和时间效应的虚拟变量。混合干扰项 ε_{it} 包括两个部分: v_{it} 和 u_{it} 。其中, v_{it} 为通常意义上的随机干扰项, 假设其服从正态分布且彼此独立, 即 $v_{it} \sim \text{i.i.d.} N(0, \sigma_v^2)$; u_{it} 表示融资约束效应, 由于其具有单边分布的特征, 我们假设其服从非负的截断型半正态分布, 即 $u_{it} \sim N^+(\omega_{it}, \sigma_u^2)$ 。 u_{it} 的异质性设定如下:

$$\omega_{it} = \exp(b_0 + \mathbf{z}_{it}'\delta) \text{ 和 } \sigma_{it}^2 = \exp(b_1 + \mathbf{z}_{it}'\gamma) \quad (6)$$

其中, b_0 和 b_1 均为常数项。需要指出的是, (5)-(6) 式构成了异质性的随机前沿模型, 这一设定使得本文的后续分析具有很大的灵活性: 其一, 我们可以同时分析外生变量对融资约束效应本身 (ω_{it}) 及其不确定性 (σ_{it}^2) 的影响, 而文献中常见的随机前沿模型事实上都是这一模型的特例; 其二, 借助这一模型, 我们可以定量分析融资约束导致的投资效率损失, 这是前期研究中基于线性回归分析无法实现的。

实证检验方法

1、检验策略与投资效率的衡量

由 (5)-(6) 式构成的异质性随机前沿模型可采用最大似然法估计, 对数似然函数为:

$$\ln L = -0.5 \ln(\sigma_v^2 + \sigma_u^2) + \ln[\phi(\varepsilon_u + \omega_u) / \sqrt{\sigma_v^2 + \sigma_u^2}] - \ln[\Phi(\omega_u / \sigma_u)] + \ln[\Phi(\tilde{\omega}_u / \tilde{\sigma}_u)] \quad (7)$$

其中, $\tilde{\omega}_u = (\sigma_v^2 \omega_u - \sigma_u^2 \varepsilon_u) / (\sigma_v^2 + \sigma_u^2)$, $\tilde{\sigma}_u = (\sigma_v^2 \sigma_u^2) / (\sigma_v^2 + \sigma_u^2)$, $\phi(\cdot)$ 和 $\Phi(\cdot)$ 分别为标准正态分布的密度函数和累积分布函数。

我们从两个方面来分析融资约束对投资行为的影响。其一, 采用似然比检验进行定性分析。原假设为 $H_0: u_u = 0$, 即不存在融资约束, 相应的备择假设为 $H_1: u_u \neq 0$ 。似然比统计量为 $LR = -2[L(H_0) - L(H_1)]$, 其中, $L(H_0)$ 和 $L(H_1)$ 分别为原假设和备择假设下的似然函数值。LR 统计量渐进地服从卡方分布, 自由度为约束的个数。同时, 我们也可以采用似然比检验来考察模型的异质性设定是否正确。其二, 构造“投资效率指数”(IEI_{it}) 进行定量分析。它表示公司的实际投资支出与最优投资支出的偏离程度, 定义如下:

$$IEI_{it} = \frac{\exp(\mathbf{x}'_{it}\beta - u_{it})}{\exp(\mathbf{x}'_{it}\beta)} = \exp(-u_{it}) \quad (8)$$

显然, IEI_{it} 介于 0 和 1 之间, 当 IEI_{it} = 0 时 ($u_{it} \rightarrow \infty$), 投资效率最低, 公司面临的融资约束最为严重; 当 IEI_{it} = 1 时 ($u_{it} \rightarrow 0$), 投资效率最高, 融资约束几乎不存在。采用最大似然法获得模型的参数估计值后, 可以进一步得到的估计式 (Battese and Coelli, 1988):

$$IEI_{it} = E[\exp(-u_{it} | \varepsilon_{it} = \hat{\varepsilon}_{it})] = \exp(-\tilde{\omega}_{it} + 0.5\tilde{\sigma}_{it}) \frac{\Phi(\tilde{\omega}_{it}/\tilde{\sigma}_{it} - \tilde{\sigma}_{it})}{\Phi(\tilde{\omega}_{it}/\tilde{\sigma}_{it})} \quad (9)$$

这里, $\tilde{\omega}_{it}$ 和 $\tilde{\sigma}_{it}$ 的定义同前, 只是将所有参数都替换成其估计值。遵循随机前沿文献中通常的做法, 我们采用 I_{it} 的对数形式作为被解释变量。因此, (9) 式中的 IEI 指数表示公司的实际投资支出相对于最优水平 (不存在融资约束时的投资支出) 偏离的百分比。

2、参数设定

对 (5)–(6) 式中各变量代理指标的设定如下: 投资机会 Q_{it} 采 Tobin's Q 加以衡量, $\mathbf{z}_{it} = (CF_{it}, EQUI_{it}, DBET_{it}, SIZE_{it})'$, 其中, CF_{it} 为现金流量, 用于衡量公司的内部融资能力, $EQUI_{it}$ 和 $DBET_{it}$ 分别为股权融资增加额和债务融资增加额, 二者都用于衡量公司的外部融资能力, 但分别反映了股票市场和银行体系的影响。根据信息不对称理论, 资本市场的缺陷会导致外部融资成本明显高于内部融资成本, 并进而导致那些内部融资能力差的公司面临融资约束。因此, 通过考察上述三种不同融资渠道对融资约束的影响, 我们便可对影响融资约束的因素进行分析。公司规模 $SIZE_{it}$ 主要作为控制变量。通常认为小规模公司面临更为严重的融资约束, 因为小规模公司的上市时间较短, 使得外界对公司的信誉情况缺乏了解, 相对于其总资产而言, 其贷款抵押品价值通常也较低。

3、样本筛选

本文数据来自于国泰安数据库, 样本区间为 2000–2006 年。筛选原则如下: (1) 为了防止兼并或重组的影响, 剔除了样本区间内总资产成长率或销售成长率大于 100% 的公司; (2) 剔除金融类上市公司和样本区间内被 ST 或 PT 的公司; (3) 为了避免异常值的影响, 剔除 Tobin's Q 大于 10 或小于 0 以及投资支出为负的公司。Tobin's Q 的计算方法同连玉君和程建 (2007), 即公司的总市值为负债与权益的市场价值之和, 流通股市值为流通股年平均股价与流通股本之积, 而非流通股市值为其股本数与每股净资产之积。负债的市值用其账面价值代替, 资产重置成本用公司总资产的账面价值代替。最终我们选择了 702 家上市公司, 共 4212 个观察值。代理变量的定义方法和基本统计量见表 1。数据处理和估计均采用 STATA9.2 完成。

表 1 变量的基本统计量和计算方法

变量	均值	标准差	最小值	最大值	计算方法
投资支出(I)	0.322	0.736	0.000	14.704	构建固定资产、无形资产和其他长期资产所支付的现金/期初固定资产净额
投资机会(Tobin)	1.326	0.382	0.117	4.879	公司总市值/资产重置成本
现金流量(CF)	0.231	0.801	-3.761	4.719	经营活动产生的现金流量净额/期初固定资产净额
公司规模(SIZE)	21.359	0.861	18.742	25.741	总资产的自然对数
股权融资(EQUI)	0.015	0.056	-0.101	0.350	$\Delta(\text{股本} + \text{资本公积金})/\text{总资产}$
债务融资(DEBT)	0.077	0.152	-0.265	0.667	$\Delta(\text{负债融资})/\text{总资产}$

结果及分析

1、异质性随机前沿模型估计结果

表2列示了在多种设定下的估计结果。模型1是我们讨论的重点,它没有对异质性随机前沿模型的参数施加任何约束。模型2-模型5则是通过在模型1的基础上施加各种约束条件后得到的。模型2假设现金流、公司规模等变量对融资约束的不确定性没有影响,对应着 Battese and Coelli(1995)的设定方式;而模型3则假设这些变量对融资约束效应本身没有影响。模型4对应着 Caudill et al.(1995)的模型,假设融资约束效应服从在零处截断的异质性半正态分布。作为对比,我们还估计了传统Q模型(1),即表2中模型5对应的结果。

整体而言,在所有设定方式下,投资机会(LnTobin)都在5%水平上显著,而个体效应和时间效应也都非常显著(受限于篇幅,个体效应和时间效应的估计结果未能列出)。这表明中国上市公司的投资行为一方面决定于投资机会的多寡,同时也受到资本市场发展状况的影响。从表2中最后四行的似然比检验(LR test)结果来看,无论将检验的原假设设定为“不存在融资约束”(对应于LR1)还是设定为“存在异质性融资约束”(对应于LR2),最终的检验结果都表明异质性随机前沿模型1显著优于其它四个模型。尤其是模型1显著优于模型5,表明融资约束及其不确定性对中国上市公司的投资支出具有显著的影响。因此,本文随后的分析都将基于模型1展开。

从表2模型1列示的结果来看,内源融资(CF)在融资约束方程和融资不确定性方程中都在1%水平上显著为负,表明现金流的增加不但可以缓解融资约束,还可以明显降低公司后续融资的不确定性。这一结果与屈耀辉和傅元略(2007)对中国上市公司融资序位的研究结论一致,即上市公司在融资过程中会优先选择内部融资。同时,这也与前期多数文献基于投资-现金流敏感性分析得到的结论一致,如梅丹(2005)、连玉君和程建(2007)等均发现中国上市公司的投资支出对现金流量非常敏感。这表明中国上市公司对内部融资有很强的依赖性,意味着外部融资约束的作用相当显著。

就两种外部融资来源而言,股权融资(EQUI)和债务融资(DEBT)都能够在1%显著水平上缓解融资约束,考虑到银行贷款是多数上市公司的主要融资来源,而中国上市公司整体上又具有股权融资偏好,得到这一结果并不奇怪。但我们注意到,二者对融资不确定性具有截然不同的影响。股权融资对融资不确定性并没有显著影响,而债务融资的增加却会显著加剧未来融资的不确定性。究其原因,我国上市公司能否获得增发和配股,主要决定于其盈利能力是否能够达到证监会的要求,因此能够获得权融资的公司通常具有较高的经营业绩,其融资不确定性相对较低。与之不同的是,中国上市公司的债务融资主要以短期负债为主(在本文样本区内,短期负债占总负债的比例约为92%),而负债率较高的公司多集中于竞争激烈、盈利能力较低的行业中(如批发零售业)。在这种情况下,负债融资往往是公司延续经营的一种被动融资方式,更多地出于摆脱当前经营困境的目的,而非为长期投资融资。负债率的持续增加会产生“债务悬置效应”(Myers, 1977),使其后续融资更加困难,进而加大了公司未来融资的不确定性。

同时,我们注意到,公司规模与融资约束和融资不确定性都显著负相关,表明大规模公司面临的融资约束程度较低,未来融资实现的不确定性也相对较小。这与我国多数大规模上市公司的国企转型背景是一致的。在样本公司中,多数大规模公司都归属于能源、电力行业,稳定的收益使它们更容易获得股权融资,而深厚的国企背景和特殊的行业特征又使它们备受银行的青睐。

2、地区差异分析

地区金融体系的发展状况能够从一定程度上反映融资约束程度,为此,我们将上市公司按地区分成了三组,并分别估计了异质性随机前沿模型(5)-(6),结果见表3。

对比三个地区的估计结果可以发现,内源融资和外源融资在不同地区上市公司中发挥的作用存在较大的差异。就内部融资而言,现金流量(CFLOW)的增加能够显著缓解中部和东部上市公司的融资约束,而对西部上市公司的影响则不显著。这主要源于西部上市公司的整体盈利能力相对较低,致使其内源融资有限。同时,现金流对三地区上市公司的融资不确定性也具有相似的影响。就外部融资而言,股权融资(EQUI)能够显著缓解中部和东部上市公司的融资约束,但对西部公司的影响有限。类似于对所有上市公司的回归结果(见表2),

表2 异质性随机前沿模型估计及检验结果

	模型 1: 无约束	模型 2: $\gamma=0$	模型 3: $\delta=0$	模型 4: $\omega_e=0$	模型 5: $u_e=0$
投资函数					
LnTobin	0.421*** (4.56)	0.458*** (4.85)	0.354*** (3.82)	0.354*** (3.78)	0.313*** (2.57)
年度效应	控制	控制	控制	控制	控制
Cons	-0.125 (-1.31)	-0.249*** (-2.99)	-0.825*** (-11.30)	-0.750*** (-9.91)	-0.641*** (-10.20)
融资约束 ω_a					
CF	-0.288*** (-7.67)	-0.327*** (-7.14)			
SIZE	-0.048*** (-3.12)	-0.402*** (-7.85)			
EQUI	-4.785*** (-4.90)	-3.706*** (-4.96)			
DBET	-5.345*** (-14.36)	-5.093*** (-12.63)			
Cons	2.744*** (2.84)	9.961*** (9.70)	-1.475*** (-2.58)		
融资不确定性 σ_a^2					
CF	-0.105*** (-4.14)		-0.328*** (-5.75)	-0.402*** (-5.84)	
SIZE	-0.459*** (-7.33)		-0.370*** (-8.72)	-0.445*** (-9.39)	
EQUI	1.118 (1.62)		1.447 (1.38)	1.476 (1.47)	
DBET	1.295*** (6.28)		4.386*** (9.32)	5.421*** (10.12)	
Cons	9.958*** (7.58)	0.450*** (5.96)	9.142*** (10.92)	10.213*** (10.33)	
对数似然值					
LR1	-6371.6	-6414.5	-6469.0	-6480.4	—
P 值	0.000	0.000	0.000	0.000	—
LR2	—	85.87	194.81	217.64	871.05
P 值	—	0.000	0.000	0.000	0.000

注: (1) ***, ** 和 * 分别表示在 1%, 5% 和 10% 水平上显著, 括号中为 t 值, 样本数均为 4212; (2) LR1 和 LR2 分别为相应模型针对模型 5 和模型 1 进行似然比检验得到的卡方值。

股权融资在分地区回归结果中同样不能显著降低融资不确定性。债务融资(DEBT)在分地区回归中的差异最为明显,对于中部和东部上市公司而言,债务融资的增加能够显著缓解融资约束,而在西部上市公司中,却会产生相反的作用。同时,债务融资的增加会加剧未来融资不确定性,但这一影响效果在西部上市公司中最为显著,而在中部上市公司中则不显著。

整体而言,由于西部上市公司盈利能力相对较低,使得内部现金流对融资约束的缓解作用非常有限,虽然其融资需求主要通过债务融资来实现,但债务融资的增加会使其面临更强的融资约束,并进一步加大后续融资的不确定性。这可以从两个角度来理解:一方面,不同于股权融资,债务融资的实现在很大程度上受到地域限制,因此,上述结果表明区域性银行体系和金融中介在西部上市公司的融资过程中作用有限;另一方面,从公司层面来看,这反映出西部上市公司增加债务融资的行为往往是被动的,更多地出于摆脱当前经营困境的目的,而非为长期投资融资。下面对投资效率的分析进一步证实了这一观点。

表 3 分地区估计结果

	西部	中部	东部
投资函数			
LnTobin	0.421** (1.99)	0.435* (1.90)	0.386*** (3.36)
Cons	0.179 (0.92)	0.080 (0.48)	-0.13 (-1.08)
年度效应	控制	控制	控制
融资约束 ω_i			
CF	-0.133* (-1.84)	-1.399*** (-6.18)	-0.286*** (-7.18)
SIZE	-0.141* (-1.77)	-0.010 (-0.12)	-0.050 (-1.06)
EQUI	-3.210* (-1.69)	-5.441*** (-3.37)	-4.750*** (-4.07)
DEBT	5.226** (1.91)	-8.614*** (-10.83)	-4.689*** (-10.69)
Cons	4.994*** (2.82)	2.506 (1.44)	2.902*** (2.68)
融资不确定性 σ_u^2			
CF	-0.173* (-1.79)	-1.193*** (-5.17)	-0.119*** (-3.04)
SIZE	-0.622*** (-4.87)	-0.598*** (-4.94)	-0.397*** (-5.47)
EQUI	2.135 (1.42)	1.836 (1.53)	0.707 (0.72)
DEBT	2.104*** (5.40)	0.881* (1.67)	1.520*** (5.68)
Cons	13.252*** (5.01)	13.232*** (5.21)	8.516*** (5.54)
对数似然值(LL)	-1361.3	-1413.1	-3557.6
样本数	882	912	2418

注:***, ** 和 * 分别表示在 1%, 5% 和 10% 水平上显著, 括号中为 t 值。

3、投资效率分析

采用随机前沿分析的一个重要特点在于我们可以定量分析每家公司的投资效率, 它间接反映了公司所面临的融资约束程度。图 1 绘制了投资效率指数(IEI)的频数分布图, 呈现右偏的特征, 表明少数公司面临非常严重的融资约束问题。IEI 的样本均值和标准误差分别为 0.719 和 0.067, 从整个分布来看, 多数上市公司的 IEI 值都集中在 0.7-0.8 之间, 表明融资约束的存在使得我国上市公司整体上的投资支出比最优水平低了约 20-30%。

中小企业融资难是目前备受关注的问题, 那么随着我国资本市场的不断发展和完善, 这些公司的融资状况是否得到了明显的改善呢? 为此, 我们按照总资产将所有样本公司等分为三组, 依次定义为大规模、中等规模和小规模公司, 进而分年度估算了这三类公司的平均投资效率指数 IEI。从图 2(a)中绘制的时序图来看, 在 2001-2006 样本区间内, 样本总体的投资效率呈现先升后降的趋势, 但基本介于 70%-72% 之间, 这似乎表明资本市场的发展是一个缓慢的过程。

然而, 对比不同规模公司的投资效率, 我们发现大规模公司的投资效率最高, 整体上呈现上升趋势; 小规模公司的投资效率最低, 整体上呈现下降趋势; 而中等规模公司的投资效率则介于二者之间。可见, 真正从资本市场发展中受益的是大规模公司, 而小规模公司的融资约束程度非但没有减轻, 反而有日趋加重的倾向。虽然小规模公司抵押品少、信誉记录短以及单位融资成本高等因素都会导致它们比大规模公司面临更为严重的融资约束, 但我国特殊的转型经济背景或许是这类公司融资状况始终未能得到明显改善的根本原因。我国股市建立的初衷是帮助国企解困, 从上述结果来看, 这一股市最初的定位思想至今仍对上市公司的投融资行为产生着重要影响。银行改制虽然取得了很大的进步, 但由于产权问题始终未能得到根本的解决, 而风险控制机制又缺乏有效性, 所以就债务融资而言, 银行对中小规模公司仍然存在着明显的信贷歧视。

从图 2(b)来看, 上市公司的投资效率也存在明显的地区差异。从时间上来看, 在 2003 年以前, 地区间的投资效率似乎不存在显著差异, 但在此之后, 东部上市公司的投资效率始终高于样本平均值, 而中部和西部则

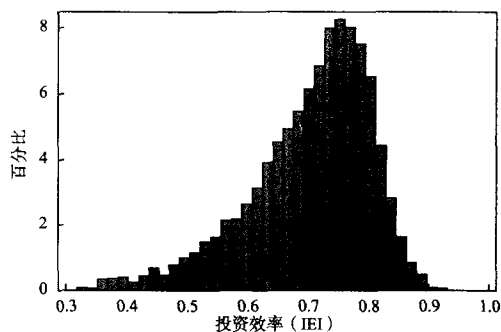
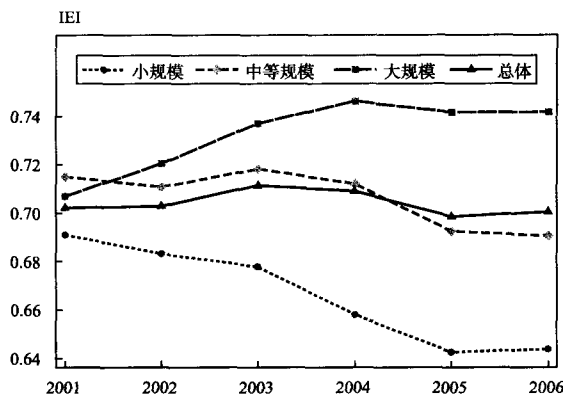
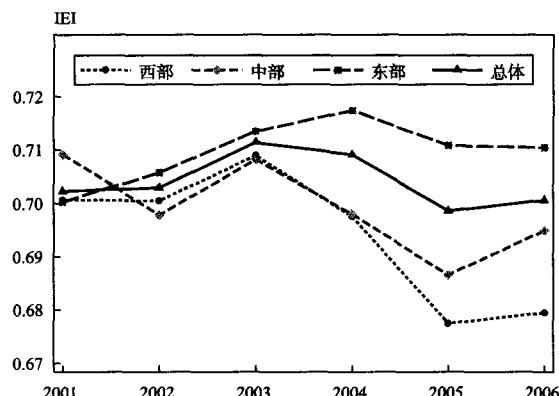


图 1 投资效率指数(IEI)的频数分布



(a)不同规模上市公司投资效率指数



(b)不同地区上市公司投资效率指数

图2 上市公司投资效率指数对比

较低,尤其是西部上市公司的投资效率甚至呈现下降的趋势。我们的统计分析表明,西部上市公司的规模明显小于东部公司。因此,这一结果与我们前文按公司规模分组得到的结果是一致的。一个可能的解释是,虽然近年来银行改制和地区金融中介都在不断发展,但东部地区的发展速度明显快于西部,致使前者面临的融资约束程度明显低于后者。从截面来看,广东和上海两地上市公司的投资效率最高,而宁夏和甘肃两地的投资效率则最低。同时,相对于广东、上海等融资状况较好的省份而言,甘肃、宁夏等金融发展落后地区的投资效率存在较大的波动,表明处于这些地区的公司在融资过程中更多地受到宏观经济状况的影响(受限于篇幅,这些结果未能列出)。在上市公司盈利能力普遍较低的情况下,外部融资成为上市公司的主要融资来源。显然,股权融资的实现基本上不受地域的限制,因此,上述投资效率差异进一步证实了前文的分析,即地区金融发展水平,尤其是信贷发展水平对上市公司的投资效率有重要影响。

结 论

资本市场的发展状况对上市公司的投资行为有着重要的影响。在资本市场完美假设下,公司的投资支出处于最优水平上,而融资约束的存在使得公司的实际投资支出会单边偏离这一最优水平。本文便利用这一特点,采用异质性随机前沿模型对中国上市公司的投资效率进行了研究。采用这一方法使我们一方面可以克服前期基于投资—现金流敏感性分析之实证研究存在的诸多缺陷,同时又可以定量地分析融资约束对投资效率的影响。

我们的实证结果表明:(1)融资约束的存在使得中国上市公司的投资支出比最优水平低了约20-30%,平均投资效率仅为72%。(2)在上市公司的三种主要融资方式中,现金流量的增加不但能缓解融资约束,还能降低后续融资的不确定性;而股权融资虽然能够有效缓解融资约束,但却不能降低融资不确定性;债务融资虽然也能够缓解融资约束,但却会显著加剧融资不确定性。(3)大规模公司和东部地区上市公司面临的融资约束和融资不确定性较低,而小规模公司和西部地区上市公司的融资约束有逐渐加剧的倾向。

本文研究结论的政策含义可概括如下:(1)由于现金流能够有效缓解融资约束并降低融资不确定性,因此就短期而言,为提高投资效率,上市公司应当努力提高自己的盈利能力,减少对外部融资的依赖。(2)就长期而言,由于股权融资能有效缓解上市公司面临的融资约束,充分发挥了其融资功能,而债务融资则会加剧公司的融资困境,因此应该在适当降低上市公司融资门槛的基础上,建立多层次的资本市场,降低银行贷款的信息非对称性,改变银行片面注重抵押价值融资的商业模式和风险控制方式,进而降低债务融资在加剧公司融资困境方面的消极作用。如果我们进一步考虑数量庞大的非上市公司,那么建立区域性多层次资本市场对地区(尤其是西部地区)的经济发展将具有更为深远的意义。

参考文献:

- [1] 冯巍. 内部现金流量和企业投资. 经济科学[J], 1999,(1):51-57

- [2] 李金,李仕明,严整. 融资约束与现金-现金流敏感度[J]. 管理评论, 2007(3):53-57
- [3] 连玉君,程建. 投资-现金流敏感性:融资约束还是代理成本[J]. 财经研究, 2007,(2):36-45
- [4] 连玉君,苏治,丁志国. 现金-现金流敏感性能检验融资约束假说吗[J]. 统计研究, 2008,(10):92-99
- [5] 刘星,魏锋,詹宇,B.Y.Tai. 我国上市公司融资序位的实证研究[J]. 会计研究, 2004,(6):66-72
- [6] 梅丹. 我国上市公司固定资产投资规模财务影响因素研究[J]. 管理科学, 2005,(5):80-86
- [7] 屈耀辉,傅元略. 优序融资理论的中国上市公司数据验证[J]. 财经研究, 2007(2):108-118
- [8] 章晓霞,吴冲锋. 融资约束影响我国上市公司的现金持有政策吗[J]. 管理评论, 2006,(10):59-62
- [9] 郑江淮,何旭强,王华. 上市公司投资的融资约束[J]. 金融研究, 2001,(11):92-99
- [10] Abel A.B. and J.C. Eberly. Q theory without adjustment costs and cash flow effects without financing constraints[J]. 2004 Meeting Papers, Society for Economic Dynamics, No.205
- [11] Battese, G. E., and T. J. Coelli, Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data[J]. Journal of Econometrics, 1988, (38): 387-399
- [12] Battese G.E. and T.J. Coelli. A model for technical inefficient effects in a stochastic frontier production function for panel data[J]. Empirical Economics, 1995,(20): 325-332
- [13] Caudill S.B., J.M. Ford, and D.M. Gropper. Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity[J]. Journal of Business and Economic Statistics, 1995,(13): 105-111
- [14] Chirinko R.S. and H. Schaller. Why does liquidity matter in investment equations?[J]. Journal of Money, Credit and Banking, 1995,(27): 527-548
- [15] Cummins J. G., K. A. Hassett, and S. D. Oliner. Investment behavior: Observable expectations, and internal funds[J]. American Economic Review. 2006,96(3): 796-810
- [16] Hayashi F. Tobin's marginal q and average q: A neoclassical interpretation[J]. Econometrica, 1982,(50): 224-313
- [17] Fazzari S., G. Hubbard, and B. Peterson. Financing constraints and corporate investment[J]. Brookings Papers on Economic Activity, 1988, (1): 141-195
- [18] Myers, S.C. Determinants of corporate borrowing[J]. Journal of Financial Economics, 1977,(5): 147-175
- [19] Pawlina G., and L. Renneboog. Is investment-cash flow sensitivity caused by agency costs or asymmetric information? Evidence from the UK[J]. European Financial Management. 2005,(11): 483-513

Financial Constraints, Uncertainty and Firms' Investment Efficiency

Lian Yujun¹ and Su Zhi²

(1. Lingnan College, Sun Yat-Sen University, Guangzhou 510275;

2. School of Business and Economics, Tsinghua University, Beijing 100084)

Abstract: This paper estimates the investment efficiency of Chinese listed firms by using the heteroscedastic stochastic frontier model. The results show that, (1) the efficiency of Chinese listed firms' investment declines 20-30% due to the financial constraints, with average efficiency at 72%; (2) among three main kinds of financing channel, the increase of cash flow can reduce both financial constraints and uncertainties, and equity financing can only reduce financial constraints, while although debt financing can reduce financial constraints, it will enlarge the uncertainties; (3) large firms and eastern firms face less financing constraints and uncertainties, while small firms and western firms suffer increasingly severe financial constraints.

Key Words: investment efficiency; financial constraints; cash flow; stochastic frontier model

中国医疗服务市场中的信息不对称程度测算^{*}

卢洪友 连玉君 卢盛峰

内容提要:本文构建了一个医疗服务市场上信息不对称程度的测度模型,并基于“中国健康与营养调查”(CHNS)中微观个体调查数据,对医疗服务市场上医患双方的信息程度及其对最终的医疗服务价格的影响效应进行了实证测度。研究表明:(1)医患双方所掌握的信息因素对最终医疗服务价格的形成具有重要影响,同时医生相对于患者掌握着更多的信息并具有更强的议价能力;(2)几乎所有的患者都将被迫接受一个高于公正基准价格的价格,平均而言达成的医疗服务价格相对于公正的基准价格要高出26.61%;(3)年度效应分析发现,1989—2006年,各年度的医疗服务市场价格大致都高于公正基准价格26%左右,换言之,改革开放以来中国的医疗服务体制改革,并未有效起到解决“看病难、看病贵”的作用;(4)患者在城乡因素、医疗保险、工作状况、年龄以及受教育程度等因素上的异质性,对医患双方最终价格的作用是有限的。本文的政策含义是:强调通过引进竞争,强化市场机制在医疗服务市场中调节作用的改革思路,是否适合中国值得反思。解决现实中普遍存在的医疗服务价格虚高问题,回归医疗服务的公益性,需要政府更多地参与其中,并有效发挥价格规制、市场监管以及外部性矫正等功能。

关键词:医疗价格 信息不对称 双边随机前沿分析 基准价格

一、引言与文献回顾

近年来,国内医疗卫生领域矛盾日益加剧,已经成为影响社会和谐发展的关键因素,不仅影响到国民健康,也带来了诸如贫困、公众情绪不满、医患关系紧张等一系列社会问题,并进一步影响着经济发展,甚至危及社会稳定以及公众对改革的支持力度。在这种背景下,2009年中央新医疗改革方案正式出台,2010年“两会”也明确提出要明显提高基本医疗服务可及性,有效减轻居民就医费用负担,缓解“看病难、看病贵”问题。显然,在影响居民医疗服务需求的诸多因素中,卫生医疗价格是最为关键的(Akin, 1986; Gupta et al., 2002; 王俊等, 2008)。因此,在当前环境下,分析医疗服务市场中的实际医疗费用负担问题显得尤为迫切,具有重要的现实意义。

表面上竞争激烈的医疗服务市场达成的实际价格却存在很大差异,而这种巨大的价格差异不能完全用医生或患者的特性和服务质量的不同来解释,信息不对称绝对是其中相当重要的因素之一。前期相关研究表明,在严格的假定条件下,医疗服务市场将自动形成一个有效的医疗服务价格。Wolinsky (1993) 研究发现,向不同医疗专家咨询可以有效地减少甚至消除医务人员的欺骗行为,但该文暗含的假设条件是信息搜集成本为零。Alger & Salanie (2006) 证明,在医疗市场充分信息的完全竞争假设下,消费者可以低成本获取最低治疗价格的信息,从而有效消除专家欺骗行为。

^{*} 卢洪友,武汉大学经济与管理学院,邮政编码:430072,电子信箱:hongylu@sohu.com;连玉君,中山大学岭南学院,邮政编码:510275,电子信箱:arlionn@163.com;卢盛峰,武汉大学经济与管理学院,邮政编码:430072,电子信箱:lsfjinlin@yahoo.com.cn。本文的研究得到了国家社会科学基金项目(批准号:08BJY132)、国家自然科学基金项目(批准号:70673073、71002056)、教育部人文社会科学基金项目(09YJC790269)以及中央高校基本科研业务费专项资金的资助。作者感谢武汉大学经济与管理学院财政税收系讨论会及武汉大学经济发展研究中心(CEDR)“增长与发展工作室”全体成员的有益讨论,感谢匿名审稿专家的宝贵意见,当然文责自负。

黄涛和颜涛(2009)从信任商品角度,构建信号博弈模型分析了医疗市场中的过度治疗现象,并指出特定条件下通过引入消费者知识搜寻决策及专家欺骗处罚机制可以起到遏制作用。

然而,医疗服务市场具有非常显著的信息不对称特征(Phelps,1997)。由于医疗服务提供方(医院)具有信息优势,这将诱发医疗服务供应方的道德风险——诱导需求(Feldstein,1970)。Rice(1983)发现医师服务费用下降后,医疗服务量随之上升。Emons(1997)对瑞士Ticino州医生安排七项重要手术患者结构数据的分析结果显示,普通患者的比重比身份为医生或者医生家属的患者比重要高出33%。Yip(1998)根据诱发需求假说推断:当医师的职业项目受收入效应影响时会通过增加服务量来弥补收入的减少。朱恒鹏(2007)指出,医疗服务市场信息不对称现象尤为严重,且医生和患者的经济利益并不一致甚至是冲突的,这为医生的道德风险行为提供了实施的可能性和空间:医生可以追求自身经济利益而不顾患者利益,利用信息优势诱使患者消费过多的医疗服务及药品。制度安排中的激励机制使医生收入与其诊疗收入密切相关,他们将充分利用信息不对称和不确定性,诱导更多医疗需求,增加病患或医疗体系的卫生支出(高春亮等,2009)。

鉴于信息因素在医疗服务定价中的重要作用,有必要从研究市场行为入手,强化政府的规制功能,才能切实解决医疗服务价格虚高问题。尽管我们迫切希望通过医疗服务体制改革来缓解医疗服务市场上的信息不完善,进而提高患者对医疗服务的可及性,但是我们还没有找到有关度量中国医疗服务市场信息不对称程度的研究文献。

本研究的贡献可以归结为如下三个方面:(1)首次采用微观数据实证测度了中国医疗服务市场价格形成中的信息不对称程度;(2)文章提出的双边随机前沿分析法,对定量估算市场参与主体中的信息问题提供一种全新的思路,这一研究思路将为后续研究提供一个新的视角;(3)研究丰富了信息经济学类研究文献,并提供了宝贵的经验证据。

后文结构安排如下:第二部分构建一个医疗服务市场价格形成中信息程度测度模型;第三部分介绍数据与指标;第四部分呈现实证分析结果;最后是文章的主要结论及政策性建议。

二、医疗服务市场价格形成中信息不对称程度测度模型

假定在一个典型的医疗服务市场中,存在众多医疗服务供给方和需求方,医生和患者都掌握着一定的信息。^① 医疗服务的最终定价(P)可表述为如下形式:^②

$$P = \underline{P} + \eta(\bar{P} - \underline{P}) \quad (1)$$

其中 \underline{P} 为医生所可能接受的最低医疗服务价格, \bar{P} 为患者所愿意支付的最高医疗服务价格。 $\eta(0 \leq \eta \leq 1)$ 用于衡量医生在定价过程中掌握的信息程度,因此 $\eta(\bar{P} - \underline{P})$ 反映了在医疗服务价格达成过程中医生所掠取的剩余。

为了在模型中同时体现出医生和患者在定价过程中掌握的信息程度,需要对(1)式进一步分解。我们首先描述在个体基本特征 x 给定条件下的“公正”医疗服务价格 $\mu(x) = E(\theta|x)$,这里 θ 是实际存在的,但是无法获知,并且总满足: $\underline{P} \leq \mu(x) \leq \bar{P}$ 。^③ 因此 $(\bar{P} - \mu(x))$ 代表着在医疗服务价

① 在特定的医疗服务机构内,医生在医学知识、药物疗效及价格等诸多方面存在信息优势,对特定患者的偏好特征存在信息劣势;患者在收集特定医疗服务机构的价格信息及相关医生的治疗习性方面具有信息优势,并基于此对医疗机构和医生个体进行选择。

② 作者的这一思路主要受到Polachek & Yoon(1987,1996)、Gaynor & Polachek(1994)以及Kumbhaka & Parmeter(2009)中双边随机前沿方法的启发。

③ 许多国外研究都分析了这种有效的配比价格问题(如Acemoglu & Shimer,2000;Flinn,2006等),但基本都被设定为已知或服从某种分布。由于本文分析的医疗服务市场的特殊性,很难先验性地找到一个“公正合理”的价格,因此,我们设定其事先不可获知,但客观存在。该价格由患者的基本特征以及社会一般医疗服务价格水平决定,也即文章后面所求解出来的基准价格。

格达成过程中患者的预期剩余; $(\mu(x) - \underline{P})$ 代表医生的预期剩余。而哪一方能够“掠取”更多的剩余将依赖于他们所占有的信息程度以及基于此的讨价还价能力 (Osbourne & Rubinstein, 1990)。我们可以用这些剩余的定義將式(1)重新表述为:

$$\begin{aligned} \underline{P} &= \mu(x) + [\underline{P} - \mu(x)] + \eta[\bar{P} - \mu(x)] - \eta[\underline{P} - \mu(x)] \\ &= \mu(x) + \eta[\bar{P} - \mu(x)] - (1 - \eta)[\mu(x) - \underline{P}] \end{aligned} \quad (2)$$

(2) 式表明, 医生可以通过掠取患者预期剩余的一部分来提高医疗服务价格, 所掠取剩余规模为 $[\bar{P} - \mu(x)]$; 同样患者可以通过掠取医生剩余的一部分来实现降低医疗服务价格, 所掠取剩余规模为 $(1 - \eta)[\mu(x) - \underline{P}]$ 。而医生能够掠取的剩余取决于医生掌握的信息程度 η 和患者的总预期剩余 $\bar{P} - \mu(x)$, 这意味着, 医生可以依靠其掌握的信息程度来提高价格; 同理, 患者获得剩余的多寡则依赖于患者掌握的信息程度 $(1 - \eta)$ 和医生的总预期剩余 $\mu(x) - \underline{P}$, 而患者也可以通过其掌握的信息程度来压低医疗服务价格。

定价方程式(2)由三部分组成: 第一部分 $\mu(x)$ 表示在给定个体特征 x 情况下“公正”的医疗服务价格, 我们称之为基准价格; 第二部分 $\eta[\bar{P} - \mu(x)]$ 体现了医生通过掌握的信息程度所掠取的剩余; 第三部分 $(1 - \eta)[\mu(x) - \underline{P}]$ 是患者通过掌握的信息程度所获得的剩余。净剩余 $NS = \eta[\bar{P} - \mu(x)] - (1 - \eta)[\mu(x) - \underline{P}]$ 可用以描述医疗服务价格形成过程中信息不对称程度的综合效应。

因此在本模型框架下, 医生信息因素对于达成的医疗服务价格具有一个正效应, 患者信息因素具有一个负效应, 即信息因素对于医疗价格形成的影响是双边的, 我们可以将医疗服务价格模型(2)简写为如下形式:

$$P_i = \mu(x_i) + \xi_i, \quad \xi_i = w_i - u_i + v_i \quad (3)$$

该模型是一个典型的双边随机前沿模型 (Kumbhakar & Christopher 2009)。其中 $\mu(x_i) = x_i' \beta$ β 为待估计参数向量, x_i 为样本个体特征, 本文包括了病情状况、健康价值期望以及其它方面特征因素; $w_i = \eta_i[\bar{P} - \mu(x_i)] \geq 0$; $u_i = (1 - \eta_i)[\mu(x_i) - \underline{P}] \geq 0$; v_i 为一般意义上的随机干扰项。医生可以通过掠取患者的预期剩余来提高医疗服务价格, 这可以通过 w_i 体现, 而患者可以通过获得一部分医生剩余来降低所支付的医疗服务价格, 这由 u_i 描述。而这些掠取所得剩余的规模取决于医患双方掌握的信息程度 η 、患者预期剩余 $\bar{P} - \mu(x)$ 和医生预期剩余 $\mu(x) - \underline{P}$ 。

为了同时测度 β 参数向量和医患双方掠取剩余部分, 我们采用最大似然估计方法 (MLE) 来估计模型(3)。由前述分析和模型(3)的设定可知, 干扰项 w_i 和 u_i 都具有单边分布 (one-sided distribution) 的特征, 为此, 我们假设二者均服从指数分布, 即 $u_i \sim i. i. d. Exp(\sigma_u, \sigma_u^2)$, $w_i \sim i. i. d. Exp(\sigma_w, \sigma_w^2)$ 。^① 对于干扰项 v_i , 假设其服从正态分布, 即 $v_i \sim i. i. d. N(0, \sigma_v^2)$ 。同时, 我们假设 v_i 、 u_i 和 w_i 之间彼此独立, 且均独立于个体特征 x_i 。基于上述设定, 可推导出复合干扰项 ξ_i 的概率密度函数如下:^②

$$f(\xi_i) = \frac{\exp(a_i)}{\sigma_u + \sigma_w} \Phi(c_i) + \frac{\exp(b_i)}{\sigma_u + \sigma_w} \int_{-h_i}^{\infty} \phi(z) dz = \frac{\exp(a_i)}{\sigma_u + \sigma_w} \Phi(c_i) + \frac{\exp(b_i)}{\sigma_u + \sigma_w} \phi(h_i) \quad (4)$$

其中 $\phi(\cdot)$ 和 $\Phi(\cdot)$ 分别为标准正态分布的概率密度函数和累积分布函数, 其它参数设定如下:

$$a_i = \frac{\sigma_v^2}{2\sigma_u^2} + \frac{\xi_i}{\sigma_u}; b_i = \frac{\sigma_v^2}{2\sigma_w^2} - \frac{\xi_i}{\sigma_w}; h_i = \frac{\xi_i}{\sigma_v} - \frac{\sigma_v}{\sigma_w}; c_i = -\frac{\xi_i}{\sigma_v} - \frac{\sigma_v}{\sigma_u}$$

对于包含 n 个观测值的样本而言, 对数似然函数可表述如下:

① 当然, 也可以假定 u_i 和 w_i 服从其它类型的单边分布, 如半正态分布、伽马分布等。Kumbhakar & Lovell (2000) 研究表明, 采用不同的分布假设对结果并没有实质性的影响, 为此, 本文采用形式最为简单的指数分布。

② 详情可参见 Kumbhakar & Christopher (2009)。

$$\ln L(X; \theta) = -n \ln(\sigma_u + \sigma_w) + \sum_{i=1}^n \ln [e^{a_i} \Phi(c_i) + e^{b_i} \Phi(h_i)] \quad (5)$$

其中 $\theta = [\beta \ \sigma_v \ \sigma_u \ \sigma_w]'$ 。通过对数似然函数的最大化,可获得所有参数的极大似然估计值。

本文重点关注的是患者和医生通过掌握的信息程度所获得的剩余,为此,我们需要进一步推导出 u_i 和 w_i 的条件分布,分别记为 $f(u_i | \xi_i)$ 和 $f(w_i | \xi_i)$,则有:

$$f(u_i | \xi_i) = \frac{\lambda \exp(-\lambda u_i) \Phi(u_i/\sigma_v + h_i)}{\Phi(h_i) + \exp(a_i - b_i) \Phi(c_i)} \quad (6a)$$

$$f(w_i | \xi_i) = \frac{\lambda \exp(-\lambda w_i) \Phi(w_i/\sigma_v + c_i)}{\exp(b_i - a_i) [\Phi(h_i) + \exp(a_i - b_i) \Phi(c_i)]} \quad (6b)$$

其中 $\lambda = 1/\sigma_u + 1/\sigma_w$ 。以(6)式确定的条件分布为基础,可以分别得到医疗服务价格形成过程中的 u_i 和 w_i 的条件期望,^①我们直接给出两者的估计式:

$$E(1 - e^{-u_i} | \xi_i) = 1 - \frac{\lambda}{1 + \lambda} \frac{[\Phi(h_i) + \exp(a_i - b_i) \exp(\sigma_v^2/2 - \sigma_v c_i) \Phi(c_i - \sigma_v)]}{\Phi(h_i) + \exp(a_i - b_i) \Phi(c_i)} \quad (7a)$$

$$E(1 - e^{-w_i} | \xi_i) = 1 - \frac{\lambda}{1 + \lambda} \frac{[\Phi(c_i) + \exp(b_i - a_i) \exp(\sigma_v^2/2 - \sigma_v h_i) \Phi(h_i - \sigma_v)]}{\exp(b_i - a_i) [\Phi(h_i) + \exp(a_i - b_i) \Phi(c_i)]} \quad (7b)$$

进一步,可以将议价过程中的净剩余 NS 表示为:

$$NS = E(1 - e^{-w_i} | \xi_i) - E(1 - e^{-u_i} | \xi_i) = E(e^{-u_i} - e^{-w_i} | \xi_i) \quad (8)$$

这里需要特别强调的是,由于参数 σ_u 仅出现在 a_i 和 c_i 中,而 σ_w 则仅出现在 b_i 和 d_i 中,所以二者即可识别(identificate)。因此,在后续检验过程中,我们无需事先假定医患双方掌握信息程度的相对大小,而完全由估计结果决定,这也是本文分析方法有别于传统回归分析方法的根本优势所在。

三、数据与指标

(一) 数据来源

本文数据来源于“中国健康与营养调查”(CHNS)数据库。该数据库是由中国疾病预防控制中心营养和食品安全研究所与美国北卡罗纳大学教堂山分校卡罗纳人口中心联合调查并创建的。它涵盖了地理特征、经济发展水平、公共资源和健康指标差异较大的辽宁、黑龙江、江苏、山东、河南、湖北、湖南、广西和贵州 9 个省份,并分别于 1989 年、1991 年、1993 年、1997 年、2000 年、2004 年以及 2006 年进行了七次调查,每次调查大约访问 4400 个左右的家庭,包含 19000 笔个体样本以及部分社区统计数据。目前该数据库主要用于对中国城乡居民的医疗、健康、劳动等方面的研究。

CHNS 数据库中的“个体医疗服务”和“医疗保险”子数据库提供了较为全面的医疗服务数据,这构成了本文实证分析的原始样本。我们基于如下原则进行了样本筛选:(1)为了保证消费医疗服务的患者的完全理性,我们剔除了未满 18 周岁(阳历)的个体样本。这是因为未成年人的医疗支出主要由父母承担,并且这种支出往往趋向于非理性,使其难以真实反映医疗服务市场的价格特征。(2)根据患者对于医疗支出的回答情况对样本进行筛选,剔除未回答的、全部由保险支付的,以及答案为“不知道”的样本。(3)删除了部分其它变量观察值缺失的样本。最终,我们得到了 1806 笔观察值,分布特征如表 1 所示。从地区分布上看,样本在 9 个省份中的分布大致平衡,意味着样本具有广泛的代表性。同时,辽宁、广西、江苏和湖南的样本数位居前四位,表明样本能够较好地体现经济发展水平的地域差异。从时序上看,样本量逐年上升,2000 年以来的三次调研样本权重约为 75%,表明本文的样本可以大致反映中国当前的基本状况。此外,样本在城乡分布和性别

^① 限于文章篇幅,作者并未列出条件期望表达式具体形式,感兴趣的读者可向作者索取。

分布上也基本平衡,而超过 65% 的被调查个体未参加医疗保险。

表 1 样本分布状况(样本总量:1806 个)

		观测样本	占比(%)	城乡分布		性别分布		是否有医疗保险	
				城市(%)	农村(%)	男(%)	女(%)	有(%)	无(%)
地区分布状况	辽宁	296	16.39	8.03	8.36	6.76	9.63	6.37	10.02
	黑龙江	139	7.70	4.60	3.10	3.49	4.21	3.16	4.54
	江苏	254	14.06	5.76	8.31	5.92	8.14	8.53	5.54
	山东	108	5.98	2.93	3.05	2.82	3.16	2.33	3.65
	河南	166	9.19	4.60	4.60	3.99	5.20	1.88	7.31
	湖北	178	9.86	5.48	4.37	5.26	4.60	3.27	6.59
	湖南	213	11.79	5.43	6.37	5.70	6.09	3.77	8.03
	广西	258	14.29	5.15	9.14	6.37	7.92	3.16	11.13
	贵州	194	10.74	4.37	6.37	4.71	6.04	1.88	8.86
年份分布状况	1989	56	3.10	0.61	2.49	1.33	1.77	0.00	3.10
	1991	73	4.05	1.00	3.05	1.50	2.55	0.89	3.16
	1993	138	7.64	2.93	4.71	3.99	3.65	1.66	5.98
	1997	187	10.35	5.09	5.26	5.15	5.20	3.60	6.76
	2000	324	17.94	7.20	10.74	7.42	10.52	3.93	14.01
	2004	618	34.22	16.39	17.83	15.17	19.05	11.35	22.87
	2006	410	22.70	13.12	9.58	10.47	12.24	12.90	9.80
合 计		1806	100%	46.35	53.65	45.02	54.98	34.33	65.67

注:表中所有比重均为占样本总数比值,四舍五入下局部可能存在加总同总比重不等情形。

(二) 变量指标的选取

在医疗服务价格的衡量上,我们选取成人调查表(18 周岁以上居民调查表)中的“卫生保健和医疗服务的利用”指标,并从中筛选出专门针对过去四周生过病、或者受过伤、或者长期患有慢性病、或急性病患者提问的问题 M39“您为治这病或伤花了多少钱”指标,该指标能够在很大程度上衡量患者消费医疗服务的价格。^①

为了衡量式(2)中的“基准价格” $\mu(x)$,我们选择如下个体特征变量:

1. 病情严重程度(Symptoms)

我们采用问卷中问题 M25“疾病的严重程度”获取的被调查个体病情信息描述患者的消费紧迫性状况,^②该项指标数值越大,表明消费行为越紧迫。

2. 健康和生命价值期望(Endurance)

居民个人由于对自己健康状况和生命价值的期望不尽相同,因此在部分医疗服务的选择上也不尽相同。问卷中 M26“当你感到不舒服时,你怎么做?”可以较好地描述该因素的影响,对于健康和生命价值期望较高的居民会更加倾向于选择就医,而期望较低居民则很有可能选择自己治疗或者不理睬,这里通过虚拟变量方法衡量对于健康和生命价值期望较低的被调查个体对于医疗服务价格的影响。

3. 其他变量

我们考虑了患者的医疗服务消费偏好,包括个体的年龄、性别、所处的地区、婚姻状况、受教育

① 从严格意义上讲,这里的价格大致相当于患者消费该项医疗服务的总支出,即包括了诊疗费和药物费等,该项指标的确定受限于问卷本身问题的设定且不可分。

② 我们也尝试采用问卷中的问题 M40“医生的诊断结果”来衡量病情的状况,但是由于问卷中病情被细化为 22 种,且不利于合并,同时大量的被调查个体未作回答,我们最终放弃了这种思路。

程度、是否有固定工作、^①是否有医疗保险等因素的影响。^②最后,控制了不同省份的地区因素以及不同年度的时间因素影响。需要进一步说明的是:(1)问卷调查表中涉及到被调查个体农村或城市地区的指标有两个,其一为 A8b1 中的户籍栏目,另一个是调查户编号中 T2“农村调查点/城市调查点”,我们采用 T2 指标衡量城乡地区医疗服务差异^③;(2)分析过程中我们分别采用赋值法和虚拟变量两种方法衡量受教育程度变量,赋值法下小学、初中、高中、中职、大专及本科、研究生以上分别赋值为 1—6;在虚变量法下,设定两个虚拟变量:Education1 高中及以下学历为 1,其他为 0;Education2 中职、大专及本科学历为 1,其他为 0。

表 2 列示了上述变量的界定方法和基本统计量。

表 2 变量的统计性描述

变 量	变量名称	问卷问题	平均值	最大值	最小值	标准差
第一部分:医疗服务价格						
治病或伤花了多少钱(元)	Price	M39	239.86	9999.0	1.0000	824.29
第二部分:病情状况						
消费紧迫性(不严重、一般、相当严重分别被赋值为 1、2 和 3)	Symptoms	M25	1.6368	3.0000	1.0000	0.6547
第三部分:健康和生命价值期望						
感觉不舒服时怎么做的(1 为选择自己治疗或不理会;0 为去找当地卫生员或看医生)	Endurance	M26	0.6346	1.0000	0.0000	0.4817
第四部分:个体基本特征						
年龄(阳历生日计算,取整数)	Age	U1	52.01	93.00	18.00	15.96
性别(1 男 0 女)	Sex	U1B	0.4502	1.0000	0.0000	0.4976
农村还是城市(1 城市 0 农村)	Urban	T2	0.4635	1.0000	0.0000	0.4988
你目前的婚姻状况(1 为未婚 0 为已婚<含离婚>)	Married	A8	0.0825	1.0000	0.0000	0.1752
受教育程度(赋值法下)	Education	A12	1.8721	6.0000	0.0000	1.1977
现在有工作吗(1 是 0 否)	Job	B2	0.5199	1.0000	0.0000	0.4997
是否有医疗保险(1 有 0 无)	Insurance	M1	0.3433	1.0000	0.0000	0.4749
第五部分:控制变量						
地区因素:调查省份(九省份)	Area	T1	—	—	—	—
时间因素:调查年份(七年份)	year	T7	—	—	—	—

① 考虑到收入因素对患者所接受医疗服务价格的影响,本文也尝试将该变量纳入分析。但是问卷中问题 B2 专门针对“有固定工作的被调查个体”询问的月工资收入,无工作个体该项指标统计为 0,我们最终放弃该变量。

② 需要说明的是,该调研数据表中未涉及给患者提供医疗服务的医生个体特征。针对患者所选择的医疗服务机构特征方面设计的问题只有 M27b“您在哪个医院看的病”一项指标,但是由于该项指标中被调查者回答状况不理想,同时以做出回答的样本部分回归分析发现,该因素不够显著,最终未采用。

③ 对比这两项调研数据发现两者之间存在着一定的出入,我们认为采用个体当前所居住及享受医疗服务的地域相对于户籍所在地更加合理,因为即便是农村户口居民在城市居住,其实际享受医疗服务的支付成本也与同等情况下该地区城市居民医疗服务的价格大致相当。

四、实证结果分析

本部分在模型设定和基准价格因素分析基础上,对模型总方差分解,并对医生和患者的信息不对称程度及议价能力差异所带来的获得剩余规模进行测度,并进一步分析各因素的影响差异。

(一) 模型设定及基准价格的影响因素

基于上述信息不对称下医疗服务市场价格形成机制及定量测度技术方法,我们对医生和患者在医疗服务价格形成中掌握信息程度的效应进行了分析。这里采用双边随机前沿方法进行测度,表3给出了基于双边随机前沿方法估计得到的回归结果。^①

表3 议价能力效应模型估计

因变量	<i>L-OLS</i>	<i>L-MLE</i>	<i>SEA</i>	<i>lnprice</i>			
	模型 1	模型 2	模型 3	模型 4	模型 5	模型 6	模型 7
<i>lnage</i>	1.018 *** (7.603)	1.056 *** (7.850)	1.007 *** (7.755)	0.751 *** (5.337)	0.796 *** (6.809)	0.839 *** (7.334)	0.598 *** (5.228)
<i>symptoms</i>	0.867 *** (14.350)	0.897 *** (14.625)	0.849 *** (14.378)	0.770 *** (13.495)	0.771 *** (13.534)	0.755 *** (13.437)	0.746 *** (13.600)
<i>urban</i>	0.287 *** (3.506)	0.296 *** (3.547)	0.275 *** (3.457)	0.260 *** (3.279)	0.260 *** (3.267)	0.245 *** (3.133)	0.205 *** (2.693)
<i>sex</i>	0.010 (0.127)	-0.001 (-0.011)	0.005 (0.069)	0.050 (0.665)	—	—	—
<i>married</i>	0.160 (0.964)	0.185 (1.121)	0.187 (1.161)	-0.091 (-0.559)	—	—	—
<i>education</i>	0.100 *** (2.809)	0.112 *** (3.064)	0.106 *** (3.074)	0.114 *** (3.267)	0.117 *** (3.401)	0.099 *** (2.886)	0.063 * (1.815)
<i>job</i>	—	—	—	-0.386 *** (-4.518)	-0.367 *** (-4.512)	-0.318 *** (-3.988)	-0.293 *** (-3.767)
<i>endurance</i>	—	—	—	-0.783 *** (-9.942)	-0.780 *** (-9.925)	-0.819 *** (-10.032)	-1.016 *** (-11.804)
<i>insurance</i>	—	—	—	0.207 ** (2.438)	0.209 ** (2.464)	0.142 (1.641)	0.099 (1.155)
<i>Constant</i>	-2.025 *** (-3.726)	-2.302 *** (-4.221)	-2.908 *** (-5.490)	-1.174 ** (-1.973)	-1.356 *** (-2.700)	-1.493 *** (-2.994)	-1.368 *** (-2.538)
<i>Area dummies</i>	—	—	—	—	—	控制	控制
<i>Year dummies</i>	—	—	—	—	—	—	控制
adj-R ²	0.151	—	—	—	—	—	—
Log likelihood	—	-3490.63	-3462.30	-3405.83	-3406.17	-3358.07	-3311.41
LR (chi2)	—	—	56.64	169.59	168.92	265.11	358.42
p-value	—	—	0.000	0.000	0.000	0.000	0.000
N	1806	1806	1806	1806	1806	1806	1806

注: *、**、***分别表示 10%、5% 和 1% 水平下显著,括号内为 t 值。

① 因变量对数化处理时为了避免数值为 0,我们将 price 变量中 37 个观测值为 1 的数值替换成了 1.01。

在表3中,模型1采用OLS估计,模型2—6均采用双边随机前沿下MLE估计,同时模型2中附加了约束条件 $\ln\sigma_u = \ln\sigma_w = 0$ 。^① 模型3和4中通过考虑增加工作、医疗保险和健康价值期望因素拟合状况得到明显改善;进一步地在模型5、模型6以及模型7中,剔除不显著的性别和婚姻状况因素并逐步增加控制了地区因素和年份因素,发现模型的拟合效果得到很大改善,本文后续分析主要基于模型7下的变量以及测度结果进行。

估计结果显示:年龄因素、城乡因素、病情状况、受教育程度以及有医疗保险等因素对价格具有正向效应。年龄越大、城市地区、病情越紧迫、受教育程度越高以及具有医疗保险的居民,更有可能接受一个较高的医疗服务价格;而对自己健康和生命价值期望较低的患者以及具有稳定工作的居民^②更倾向于接受相对低的医疗服务价格。

(二) 方差分解:医疗价格形成中信息程度测度模型的解释能力

表4汇报了掌握信息程度因素效应的分析结果。我们发现,掌握信息程度对医疗服务价格的形成具有相当重要的影响,其中,医生相对于患者具有更强的信息优势,这将导致医患信息因素对于医疗服务价格形成的综合影响为正, $E(w - u) = \sigma_w - \sigma_u = 0.6954$,表明综合而言,讨价还价将形成一个相对于基准价格更高的价格。同时, $\ln price$ 无法解释部分总方差 ($\sigma_v^2 + \sigma_u^2 + \sigma_w^2$) 为 2.3915,这其中 46.1% 由医患信息因素所贡献;而在信息因素对价格的总影响中,医生相对于患者几乎处于一个绝对的信息优势地位,达到 91.38%;患者掌握信息程度在信息因素的总影响中仅为 8.62%。这表明,虽然在医疗服务价格形成过程中,患者具有一定的信息和议价能力,但是价格的形成更取决于医生。为了分析特定“医生—患者”在讨价还价中各自所掠取的剩余以及净剩余,我们进一步对医患双方做了单边效应估计。

表4 议价能力因素的医疗价格效应分析

	变量含义	符号	测度系数
议价机制	随机误差项	σ_v	1.1353
	患者议价能力	σ_u	0.3083
	医生议价能力	σ_w	1.0037
方差分解	随机项总方差	$\sigma_v^2 + \sigma_u^2 + \sigma_w^2$	2.3915
	总方差中讨价还价因素影响比重	$(\sigma_u^2 + \sigma_w^2) / (\sigma_v^2 + \sigma_u^2 + \sigma_w^2)$	46.1%
	患者议价能力影响比重	$\sigma_u^2 / (\sigma_u^2 + \sigma_w^2)$	8.62%
	医生议价能力影响比重	$\sigma_w^2 / (\sigma_u^2 + \sigma_w^2)$	91.38%

(三) 患者剩余与医生剩余的估计

1. 样本总体估计结果

本部分研究的重点是估算医患双方在信息不对称下各自所获得的剩余,即 $E(u|\xi)$ 和 $E(w|\xi)$,相应的估计式为(7a)和(7b),其含义是医生和患者在信息因素下各自获得的剩余相对于基准价格 $\ln P = x_i' \beta$ 价格变动的百分比。表5呈现了针对全样本的估计结果。平均而言,医生在信息不对称下所获得的剩余将使医疗服务价格高出基准价格 50.17%;而患者剩余则仅能使医疗服务价格降低 23.56%。这种相对悬殊的掌握信息程度使得实际医疗服务价格比基准价格高出了 26.61%。换言之,由于医患双方信息不对称的存在和议价能力的差异,对于公平市场上 100 元的医疗服务,患者需要支出 126.61 元。

表5后三列(Q1—Q3)更为细致地呈现了医患剩余的分布特征,表明医患双方掌握的信息程度

^① 在随机前沿模型估计过程中,为了保证 σ_v 、 σ_u 和 σ_w 的估计值为正数(这也是模型设定中所要求的),我们在估计过程中采用了上述三个参数的对数形式,并在完成估计后通过指数化得到相应参数的原始估计值。

^② 这里一个可能的解释在于公司内部的医疗服务对消费者自身的需求在一定程度上起到了替代作用。

的差异具有较强的异质性,但是在医患双方的议价过程中几乎所有患者都处于劣势地位。具体而言,由第一四分位(Q1)的统计结果可知,有1/4的患者,医患议价的结果是医疗价格相对于基准价格有不超过12%的上升。然而,从第三四分位(Q3)的统计结果来看,另有1/4的患者,医患议价的结果则是医疗服务价格相对于基准价格上涨幅度高达近40%。

表5 议价中医生和患者获得的总剩余

变 量	平均值 (%)	标准差 (%)	Q1 (%)	Q2 (%)	Q3 (%)
医生: $E(1 - e^{-u} \varepsilon)$	50.17	16.73	37.44	46.21	59.87
患者: $E(1 - e^{-u} \varepsilon)$	23.56	4.07	20.50	22.61	25.49
净剩余: $E(e^{-u} - e^{-w} \varepsilon)$	<u>26.61</u>	20.16	11.95	23.60	39.37

注:Q1、Q2、Q3 分别表示第1、2、3 四分位,即第25、50 和75 百分位,下同。

图1—3 更为直观地呈现了医生、患者以及二者净剩余的分布特征。由图1 和图2 可知,无论是医生剩余还是患者剩余,其分布都呈现出向右拖尾的特征,意味着只有少数医生或患者的议价能力处于绝对强势地位。由图3 中净剩余的分布特征可以看出,实际上,并非所有患者在议价过程中都处于绝对的劣势地位。我们的统计分析表明,大约有不超过10% 的患者的净剩余小于零,意味着他们事实上具有较强的掌握信息的程度并压低了医疗价格。这同时也意味着,90% 以上的患者都被迫接受了不合理价格的事实。整体而言,我们的分析表明在医疗服务市场的医患议价过程中,医生相对于患者具有更强的信息优势,并最终依赖这种能力在制定服务价格过程中实施了“高价”策略。

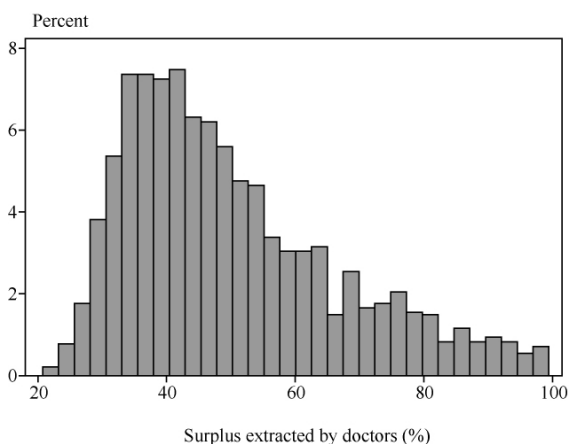


图1 医生获得剩余的频数分布

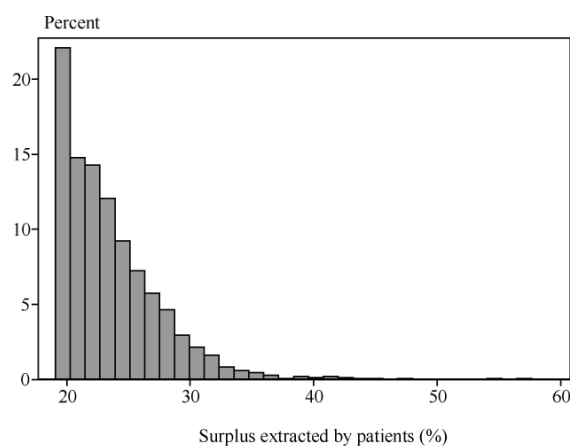


图2 患者获得剩余的频数分布

格”策略。

自改革开放以来,伴随着各个领域经济体制改革的推进,中国的医疗服务体制改革也向纵深发展,鼓励私人医疗服务机构发展、医院产权改革等措施在一定程度上将竞争机制引入到医疗服务市场。那么,医患的相对信息不对称程度是否随着医疗服务体制改革进展而有所改善呢?为此,我们分年度统计了医生和患者净剩余的分布特征,呈现于表6。令人吃惊的是,医生—患者净剩余从1989 年到2006 年几乎都位于26% 左右,这意味着在过去的近20 年时间

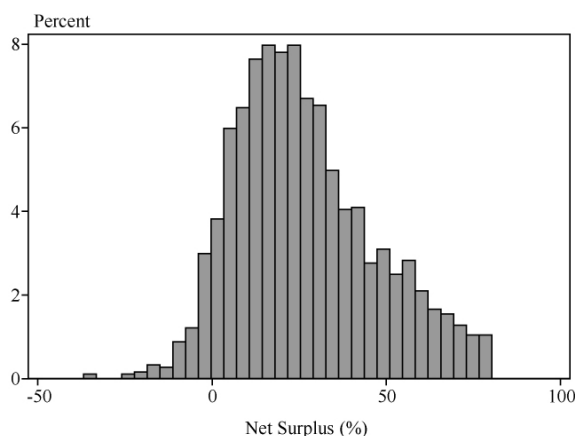


图3 净剩余的频数分布

里,通过医疗服务市场化改革,引入竞争机制,并未有效起到预期的扭转患者信息劣势,增强患者相对议价能力的作用。这也表明,通过单纯地向医疗服务机制中引入市场因素以增强竞争机制发挥作用的思路,无法有效解决医疗服务市场信息不对称和价格虚高问题。

表 6 医生和患者净剩余的年度分布特征

年份	平均值 (%)	标准差 (%)	Q1 (%)	Q2 (%)	Q3 (%)
1989	26.66	20.07	12.68	20.14	44.06
1991	26.89	20.87	14.12	22.33	40.47
1993	26.59	20.02	10.52	23.84	40.08
1997	25.36	16.99	12.70	22.88	37.46
2000	26.18	19.02	11.67	24.88	40.70
2004	27.45	22.19	11.88	23.58	39.83
2006	26.18	19.13	12.08	23.15	37.30

2. 个体特征对掌握信息程度的影响

在前文的分析中,我们发现医患双方掌握信息的程度具有很强的异质性。为了探求其根源,我们进一步从城乡差异、医疗保险状况、工作稳定程度以及年龄和学历特征等方面分组统计和分析医生和患者剩余分布特征。

由表 7 的统计结果可知,在城乡因素方面,城市地区医生和患者都相对具备更强的掌握信息能力,但是医患议价的净剩余方面城乡差异不大,几乎所有患者都面临着接受一个大致高于基准价格 26.6% 的均衡价格,同时不同分位的患者所面临的上涨幅度不尽相同。

表 7 城乡因素对医患双方获得剩余的效应

变 量	平均值 (%)	标准差 (%)	Q1 (%)	Q2 (%)	Q3 (%)
□ 农村 (urban = 0)					
医生: $\hat{E}(1 - e^{-u} \varepsilon)$	50.10	16.65	37.43	46.03	60.25
患者: $\hat{E}(1 - e^{-u} \varepsilon)$	23.53	3.85	20.47	22.65	25.49
净剩余: $\hat{E}(e^{-u} - e^{-w} \varepsilon)$	26.57	19.97	11.94	23.38	39.79
□ 城市 (urban = 1)					
医生: $\hat{E}(1 - e^{-u} \varepsilon)$	50.24	16.82	37.69	46.30	59.49
患者: $\hat{E}(1 - e^{-u} \varepsilon)$	23.60	4.31	20.54	22.59	25.37
净剩余: $\hat{E}(e^{-u} - e^{-w} \varepsilon)$	26.64	20.38	12.32	23.71	38.95

在受教育程度因素方面,同样存在着医生的价格歧视行为。表 8 反映了医生和患者讨价还价中所获得剩余及最终净剩余在不同学历方面的对比结果,结果表明,大学及以上学历文化的患者在不同分位上均面临着相对最低的医疗服务价格上涨幅度。这一方面在于患者通过掌握的信息和议价能力获得了相对更多的剩余,同时医生相对于在其他学历下“掠取”了相对更少的剩余。这意味着,一方面患者利用自己的知识水平掌握了更多的信息,并在讨价还价中实现了“自保”效果,同时医生对高学历群体实施了相对较低的医疗服务价格策略。

同样,我们针对患者是否有医疗保险、是否具有稳定工作以及所处的年龄段等方面的异质性特征分别对样本进行了分组对比分析,结果都无一例外地表明患者被迫接受了一个相对于基准价格

更高的议价后均衡价格,只是不同分位患者之间所面临的涨幅有所差别。^①因此,从某种程度上而言,相对于医生的信息优势,患者的异质性能力对于医疗价格议价的作用十分有限。

表 8 学历因素对医患双方获得剩余的效应

变 量	平均值 (%)	标准差 (%)	Q1 (%)	Q2 (%)	Q3 (%)
$\text{education} < = 1$	<u>小学及文盲</u>				
医生: $\hat{E}(1 - e^{-u} \varepsilon)$	50.42	17.38	36.71	45.52	61.42
患者: $\hat{E}(1 - e^{-u} \varepsilon) +$	23.63	4.10	20.35	22.78	25.83
净剩余: $\hat{E}(e^{-u} - e^{-w} \varepsilon)$	26.79	20.89	10.87	22.74	41.07
$1 < \text{education} < = 4$	<u>中学文化</u>				
医生: $\hat{E}(1 - e^{-u} \varepsilon)$	49.94	15.71	38.19	46.80	58.44
患者: $\hat{E}(1 - e^{-u} \varepsilon)$	23.38	3.72	20.65	22.48	25.15
净剩余: $\hat{E}(e^{-u} - e^{-w} \varepsilon)$	26.57	18.86	13.04	24.32	37.79
$4 < \text{education} < = 6$	<u>大学以上文化</u>				
医生: $\hat{E}(1 - e^{-u} \varepsilon)$	49.57	17.64	38.89	45.20	58.75
患者: $\hat{E}(1 - e^{-u} \varepsilon)$	24.21	5.64	20.62	22.86	24.86
净剩余: $\hat{E}(e^{-u} - e^{-w} \varepsilon)$	<u>25.36</u>	22.08	14.03	22.34	38.13

五、结论与政策性建议

本文构建了一个医疗服务市场上信息不对称程度的测度模型,基于“中国健康与营养调查(CHNS)”中微观个体数据,对医疗服务市场上医患双方的信息程度及其对最终的医疗服务价格的影响效应进行了实证测度。实证检验结果表明:

(一) 医患双方所掌握的信息因素对最终医疗服务价格的形成具有重要影响,同时,医生相对于患者掌握着更多的信息并具有更强的议价能力。信息不对称因素对于最终形成的医疗服务价格的综合影响为正 0.6954,表明综合而言,医患信息因素将形成一个相对于基准价格更高的价格。

(二) 对医患双方单边效应全样本分析发现,就平均而言,在医疗服务价格形成中,医生凭借其掌握的信息将以 50.17% 的幅度提高医疗服务价格;而患者凭借其掌握的信息将以 23.56% 的幅度降低医疗服务价格。这两种相反的作用,使得达成的医疗服务价格相对于基准价格上涨了 26.61%。分位分析进一步表明,在医疗服务市场价格形成过程中,几乎所有患者都将被迫接受一个高于基准价格的价格,而异质患者面对的上涨幅度则有所不同。

(三) 在 1989 年到 2006 年期间,医生凭借其信息优势“主持”达成的价格,大致都高于公正基准价格 26% 左右。这说明中国的医疗服务体制改革,并未有效起到缓解医疗服务市场信息不对称问题。平均而言,医患议价所形成的价格相对于公正的基准价格都要高出 26% 左右,因此,医疗服务体制改革也未能真正解决“看病难、看病贵”问题。

(四) 进一步分析患者在城乡因素、医疗保险、工作状况、年龄因素以及受教育程度因素上的异

^① 限于篇幅,本文未报告出是否有医疗保险、是否有稳定工作以及年龄因素对于医患双方剩余的效应测度结果,感兴趣的读者可以向作者索取。

质性对医患双方最终价格的作用效应,同样表明:几乎所有患者都不同程度地面临着被迫接受一个高于基准价格的医疗服务价格,并且医生可以有效地实施歧视性定价策略。

本文的分析结论表明:改革开放以来,中国的医疗服务体制改革并未有效地起到缓解医疗服务市场信息不对称问题,也没有起到解决“看病难、看病贵”的作用。换言之,这种过于强调通过引进竞争,强化市场机制在医疗服务市场中调节作用的改革思路是否适合中国值得反思。解决中国现实中普遍存在的医疗服务价格虚高、医患关系紧张等突出问题,必须回归医疗服务的公益性,需要政府更多地参与其中,并有效发挥价格规制、市场监管以及外部性矫正等功能。

参考文献

- 高春亮、毛丰付、余晖,2009:《激励机制、财政负担与中国医疗保障制度演变——基于建国后医疗制度相关文件的解读》,《管理世界》第4期。
- 黄涛、颜涛,2009:《医疗信任商品的信号博弈分析》,《经济研究》第8期。
- 王俊、吕忠泽、刘宏,2008:《中国居民卫生医疗需求行为研究》,《经济研究》第7期。
- 朱恒鹏,2007:《医疗体制弊端与药品定价扭曲》,《中国社会科学》第4期。
- Acemoglu, D. and R. Shimer, 2000, “Wage and Technology Dispersion”, *Review of Economics Studies*, Vol. 67, pp. 585—607.
- Akin J. S., C. C. Griffin, D. K. Guilkey and B. M. Popkin, 1986, “The Demand for Primary Health Care Services in the Bicol Region of the Philippines”, *Economic Development and Cultural Change*, Vol. 34, No. 4, pp. 755—782.
- Alger, I. and F. Salanie, 2006, “A Theory of Fraud and Over-treatment in Experts Markets”, *Journal of Economics and Management Strategy*, Vol. 15, No. 4, pp. 853—881.
- Flinn, C., 2006, “Minimum Wage Effects on Labor Market Outcomes under Search Matching and Endogenous Contact Rates”, *Econometrica*, Vol. 74, pp. 1013—1062.
- Gupta, S., M. Verhoeven and E. R. Tiongson, 2002, “The Effectiveness of Government Spending on Education and Health Care in Developing and Transition Economies”, *European Journal of Political Economy*, Vol. 18, No. 4, pp. 717—737.
- Kumbhakar S. C. and C. F. Parmeter, 2009, “The Effects of Match Uncertainty and Bargaining on Labor Market Outcomes: Evidence from Firm and Worker Specific Estimates”, *Journal of Productivity Analysis*, Vol. 31, No. 1, pp. 1—14.
- Kumbhakar S. C. and Lovell C. A. K., 2000, *Stochastic Frontier Analysis*, Cambridge University Press, New York, pp. 90.
- Martin Gaynor and Solomon W. Polachek, 1994, “Measuring Information in the Market: An Application to Physician Services”, *Southern Economic Journal*, Vol. 60, No. 4, pp. 815—831.
- Martin S. Feldstein, 1970, “The Rising Price of Physicians’ Services”, *Review of Economics and Statistics*, Vol. 52, pp. 121—133.
- Rice, T., 1983, “The Impact of Changing Medicare Reimbursement Rates on Physician-induced Demand”, *Medical Care*, Vol. 21, pp. 803—815.
- Solomon W. Polachek and Bong Joon Yoon, 1987, “A Two-Tiered Earnings Frontier Estimation of Employer and Employee Information in the Labor Market”, *Review of Economics and Statistics*, Vol. 69, No. 2, pp. 296—302.
- Solomon W. Polachek and Bong Joon Yoon, 1996, “Panel Estimates of a Two-Tiered Earnings Frontier”, *Journal of Applied Econometrics*, Vol. 11, No. 2, pp. 169—178.
- Osbourne M. J. and A. Rubinstein, 1990, *Bargaining and Markets* (Chapter 5), Academic Press, San Diego.
- Phelps C. E., 1997, *Health Economics*, New York, NY: Addison-Wesley Educational Publishers Inc.
- Winand Emons, 1997, “Credence Goods Monopolists”, Berkeley Olin program in Law & Economics, Working Paper Series 1060.
- Wolinsky A., 1993, “Competition in a Market for Informed Experts’ Services”, *Rand Journal of Economics*, Vol. 24, pp. 380—398.
- Yip W. C., 1998, “Physician Response to Medicare Fee Reductions: Changes in the Volume of Coronary Artery Bypass Graft (CABG) Surgeries in the Medicare and Private Sectors”, *Journal of Health Economics*, Vol. 17, pp. 675—699.

Measurement of the Information Asymmetric in Medical Service Market of China

Lu Hongyou ,Lian Yujun and Lu Shengfeng
(Wuhan University;Sun Yat-Sen University;Wuhan University)

Abstract: This paper builds up a model for measuring the information asymmetric in the medical service market of China , and analyses the effects on attaining the medical service price caused by the asymmetric information between doctors and patients , based on the data in the CHNS database. The result shows that: (1) the degree of asymmetric information does have an important effect on forming the ultimate service price , and doctors possess more information and stronger bargaining ability; (2) in the medical service market , almost all patients have to receive a service price which is 26. 61% higher than the standard one on average; (3) estimating the effect between years shows the final market price is about 26% higher than the standard one every year during 1989 to 2006 , that means the reform of medical service in China is not helpful to solve the problem of “difficult and expensive medical treatment”; (4) the divergence is depended on various causes: living position; medical insurance possession; job; age and education. Our suggestions are as follows: it is necessary to reflect whether the ideas of current medical service reform fit for China since the reform and opening up. The government is required to take more part in price-making process and plays an important role in price setting and market monitoring , so as to solve the hypocritical expensive service price and realize the public welfare.

Key Words: Medical Service Price; Information Asymmetric; Two-tier Stochastic Model; Standard Price

JEL Classification: I18 ,H41 ,C78

(责任编辑:詹小洪)(校对:昱 莹)

~~~~~  
(上接第 93 页)

## The Impact of Elder-care Patterns on Chinese Elderly's Health and Well-being

Liu Hong , Gao Song and Wang Jun  
(Central University of Finance and Economics)

**Abstract:** The development trend of China's population aging brings great pressure on social security , health care and social welfare. Using Chinese Longitudinal Healthy Longevity Survey (2002—2005) data , this paper (1) defines elder-care patterns in China from two aspects: living arrangements and means of financial support , and (2) examines how different elder-care patterns affect health and subjective well-being of Chinese elderly , and (3) investigates gender difference , urban-rural difference and cohort difference in the association of elder-care patterns and outcomes. It is found that married couples living alone with financial independence are the most advantaged group in terms of health and life quality. The evidence also suggests that the elderly may benefit from the development of institutionalized care.

**Key Words:** Elder Care; Health; Subjective Well-being; Empirical Analysis

**JEL Classification:** I12 , I18 , J14

(责任编辑:松木)(校对:梅 子)

## 消费文化、认知偏差与消费行为偏差\*

叶德珠 连玉君 黄有光 李东辉

**内容提要：**文化影响消费是一个基本共识，但对该命题的理论分析和经验支持还较为有限。本文放松了理性经济人假设，在行为经济学双曲线贴现模型框架下，以“自我控制”认知偏差及相应的模型参数设定对东西方消费文化差异进行了技术表达，进而阐明了消费过度（欧美国家）和消费不足（东亚国家）这两类消费行为偏差的形成机制。最后，我们采用全球48 个国家和地区 1978-2007 年的面板数据，以儒家虚拟变量和性生活指数作为消费文化的替代变量检验了文化与消费的关系。结果表明，在解释东西方消费率差异时，预防性储蓄等传统理论的解释力远低于不可观测的国家个体效应。儒家虚拟变量和性生活指数能分别解释国家个体效应的 28%和 58%，这表明消费文化等不随时间改变的个体因素比传统变量更能解释各国居民的消费差异。实践层面上，双曲线贴现模型中锁定技术能有效纠正“自我控制”认知偏差，从而消解儒家文化对消费的深度抑制，可为扩大内需政策创新提供思路启发和技术支撑。

**关键词：**消费文化 自我控制认知偏差 双曲线贴现 扩大内需

### 一、引言

对于各国在消费率/储蓄率上的显著差异，前期学者从多个角度进行了研究，主要集中于经济、人口和制度等方面。其一，经济因素（如收入、利率等）。根据凯恩斯消费函数，消费率是收入的减函数。利率一般被认为是储蓄的补偿，因此会与消费率负相关（Summers, 1984）。其二，人口因素。“生命周期假说”认为，消费者根据一生预期总收入来平滑各期消费，因此会在工作期内进行净储蓄，而在其他生命阶段进行净消费（Modigliani and Brumberg, 1954）。因此，一国的赡养率越高则消费率越低，反之亦然。其三，制度因素。预防性储蓄理论（Leland, 1968）认为，人们会因为未来的不确定性而进行谨慎性储蓄，因此，完善的社会保障体系有助于刺激消费（Hubbard et al., 1995）。流动性约束理论认为，金融市场发展滞后会限制消费者借贷，刺激储蓄（Deaton, 1991）。此外，影响消费的因素还包括保险、习惯性坚持、相对消费等因素（Harbaugh, 2003）。

在上述理论中，收入和人口因素可以较好地解释日本、新加坡等国的居民消费行为，但却无法解释中国与欧美发达国家之间的消费行为差异。预防性储蓄理论和流动性约束理论能够同时解释欧美居民的过度消费和中国居民的消费不足（龙志和和周浩明, 2000；万广华等, 2001；罗楚亮, 2004），却难以解释为什么社会保障体系和金融市场发展都较发达的日本、新加坡等国，仍然面临“低消费、高储蓄”的困境。因此，要更全面地理解消费/储蓄行为，或许还需要补充对其他因素的分析。

---

\* 叶德珠，暨南大学金融研究所及金融系，邮政编码：510632，电子邮箱：gzydz@126.com；连玉君，中山大学岭南学院，邮政编码：510275，电子邮箱：15889968888@163.com；黄有光，澳大利亚莫纳什大学，电子邮箱：kwang.ng@monash.edu；李东辉，澳大利亚新南威尔士大学，电子邮箱：donghui@unsw.edu.au。本文受国家自然科学基金项目（71002056）、国家社会科学基金项目（10CJL010、11AJY013）、教育部人文社会科学研究基金项目（09YJC790269），以及中山大学经济研究所基地建设经费资助。感谢匿名审稿人的宝贵建议，使本文得到实质性改进。当然，文责自负。

一个值得注意的现象是，消费不足主要集中在东亚儒家思想圈国家，而消费过度则主要集中在欧美国家。因此，一个朴素而直观的猜想就是消费文化导致了消费行为的国别差异（Harbaugh, 2003）。然而，由于技术上的困难，延循这一研究路线的文献还很有限。相关研究主要集中在市场营销等微观层面（Briley et al., 2000; Johar et al., 2006），宏观层面研究较少，且都局限于个别国家，对文化是否影响消费也未有定论（Carroll et al., 2000; Mouawiya and Elhiraika, 2003）。更为重要的是，文化对消费行为的作用机制仍然处于黑箱状态，缺乏规范的政治学理论分析。主要原因可能是对消费文化的政治学技术表达存在瓶颈。跨期消费/储蓄决策涉及消费者的风险偏好和时间偏好，而消费文化可能同时对这两个偏好产生影响。鉴于已有大量基于预防性储蓄理论的文献从风险偏好角度研究了东西方消费/储蓄行为的差异（朱信凯和骆晨，2011），本文着重分析消费文化对消费者时间偏好的影响。

消费文化主要通过影响消费者的自我控制力并进而影响其消费行为。自我控制力具有明显的文化特征，同时又是消费者时间偏好的重要决定因素（Fisher, 1930），但这种文化及相应的消费行为差异在以传统的新古典时间偏好理论为基础的指数贴现模型中很难得到技术表达。该模型以“理性经济人、固定贴现率”为前提，主要通过贴现率来刻画消费行为的差别。要解释东西方消费差异，就需要假设欧美消费者的贴现率很高、东亚消费者的贴现率很低，但这往往难以让人信服。

行为经济学中的双曲线贴现模型能够在一定程度上克服上述困扰。它放松了“理性经济人”假设，认为消费者存在自我控制认知偏差，因而会形成短期和长期不一致的贴现率结构。为此，该模型无需设定极端的贴现率就可以很好地解释吸毒、过度饮食等消费行为偏差（Akerlof, 1991; Gruber and Koszegi, 2001）。这为解释东西方消费行为差异提供了一个可能的技术途径，但目前还仅用于解释一些较为具体的消费行为，很少上升到消费文化层面。

本文采用双曲线贴现模型来分析消费文化对东西方消费行为差异的影响。本文的理论逻辑是，在消费文化影响下，消费者会形成程度各异的自我控制认知偏差，这在双曲线贴现模型中可以用短期贴现因子 $\beta$ 来表达。 $\beta < 1$ 表示消费者具有自我控制不足认知偏差，主要反映欧美消费文化； $\beta > 1$ 表示过度自我控制，主要反映儒家消费文化。自我控制认知偏差使得消费者在制定长期消费计划时较为理性，但在实际消费时却经常偏离原有计划，从而出现消费过度或消费不足。本文基本理论预期是：居民受儒家文化影响越大，自我控制力越强，过度自我控制认知偏差越严重，则消费率越低，反之亦然。

我们采用全球 48 个国家和地区在 1978-2007 年期间的面板数据，对上述理论预期进行了实证检验。本文以自我控制力作为消费文化的替代变量，并用两组代理变量来表示：一是儒家文化圈虚拟变量，二是性生活指数。我们发现，前期文献中提及的经济、人口和制度等因素（如收入、社会保障等）对消费率国别差异的解释力不足 5%，而不随时间改变的国家个体因素（如风俗、文化等）则能够解释约 79% 的消费差异。这些个体因素中，约有 28% 可以藉由儒家文化虚拟变量来解释，有 58% 可以归因为以性生活指数为代表的文化因素。具体来说，消费率与代表文化的自我控制力变量显著负相关；消费行为偏差程度与自我控制认知偏差程度显著正相关，与教育水平显著负相关。这些结果表明，文化是造成消费行为偏差的主要原因，而一国理性居民比例越低，消费文化的影响越大，消费行为偏差越严重。

本文的主要贡献可以归结如下：其一，在理论层面上，用自我控制认知偏差对两类对立的消费文化进行了统一的技术表达，尤其是用规范的政治学模型分析了儒家消费文化抑制居民消费的作用机制。其二，在实证层面上，采用儒家文化虚拟变量和性生活指数来衡量消费文化，并基于跨国面板数据，对文化与消费之间的关系进行了检验，凸显了文化因素在解释消费率国别差异上相对于收入等传统变量的优越性。其三，在政策干预方面，本文的分析为各国纠正消费行为偏差尤其是中国的扩大内需政策提供了微观基础及技术创新的思路。



后文结构安排如下：第二部分对消费文化与认知偏差之间的关系进行梳理和技术表达。第三部分建立双曲线贴现模型，剖析消费不足和消费过度的形成机制。第四部分建立实证模型对消费文化与消费率之间关系进行回归分析。第五部分是政策讨论。最后是结论。

## 二、消费文化与自我控制认知偏差

### (一) 东西方消费文化的典型特征

欧美国家消费文化的典型特征是超前消费，居民往往利用金融市场的便利，将未来的收入提前透支到当期进行消费；东亚国家的消费文化则表现为谨慎消费和节俭消费。

东西方消费文化的形成具有深刻的社会根源。在 20 世纪以前，由于生产力低下和基督教义的影响，欧美国家居民也同样尊崇节俭。但随着消费信贷金融创新的普及，以及二战后物质水平的大幅提高，人们生活质量得以提高。尤其是为应对 20 世纪 30 年代经济大危机而兴起的凯恩斯主义理论更是强调刺激消费需求的重要性，提出“消费即是爱国”的口号，从而使消费主义开始盛行，认为人的满足和快乐的第一位要求是占有和消费物质产品。在这种思潮的影响下，无节制地消耗物质财富得到社会认可甚至鼓励。日益发达的金融市场更是推波助澜，将这种寅吃卯粮的消费模式引向极致。东亚儒家思想在对待消费与储蓄问题上则一直非常内敛，一直保持着“崇俭黜奢”的禁欲倾向。孔子指出“奢则不孙，俭则固。与其不孙也，宁固”，并把“俭”和温、良、恭、让同视为重要的德目。儒家思想的长期熏陶使人们在消费生活必需品时心安理得，而在消费奢侈品时则容易产生负罪感，并将改善消费的希望寄望于将来，形成对消费的过度抑制。

### (二) 消费文化的技术表达——自我控制认知偏差及短期贴现因子

道德文化历来被认为是时间偏好形成与决定的重要因素（Fisher, 1930; Becker and Mulligan, 1997）。因此，要对东西方消费文化进行技术表达，时间偏好理论可以作为一个合适的分析框架。对于不同消费行为差异，在时间偏好分析技术上有两种表达方法：一种是设定时间偏好率（贴现率）的绝对数值，另一种是设定时间偏好率的结构特征。

在以新古典经济学时间偏好理论为基础的指数贴现模型中，时间偏好率（贴现率）是一个固定常数，消费者的消费行为主要通过贴现率的高低来刻画。因此要解释东西方消费行为差异，就需要为欧美消费者设定很高的贴现率，而为东亚消费者设定极低的（甚至是负的）贴现率（Lu and McDonald, 2006），但这并不符合经济学直觉。

以行为经济学时间偏好理论为基础的双曲线贴现模型采用的是允许时间偏好率发生结构变化的思路。该模型确认了消费者存在的自我控制认知偏差，并增加了一个短期贴现因子（贴现率）对此进行了刻画。在该模型框架下，消费者的跨期效用为（Laibson, 1997）：

$$U(t, s) = u_t + \beta \sum_{s=t+1}^{\infty} \delta^{s-t} u_s \quad (1)$$

其中，消费者的贴现因子结构设定为  $\{1, \beta\delta, \beta\delta^2, \dots, \beta\delta^t\}$ ，消费者在未来  $t$  期与  $t+1$  期（长期）之间使用的长期贴现因子为  $\delta$ ，在 0 期与 1 期（短期）之间的短期贴现因子为  $\beta\delta$ 。其中， $\beta$  用来刻画消费者短期贴现时存在的自我控制认知偏差。作为特例， $\beta = 1$  表示消费者没有认知偏差，此时模型退化为指数贴现模型。当  $\beta \neq 1$  时，表示消费者存在认知偏差，长期贴现率与短期贴现率不同，从而可能导致消费异常。 $\beta < 1$  代表着自我控制不足认知偏差， $\beta > 1$  代表着自我控制过度认知偏差（Krusell et al., 2002）。

为表达消费者存在自我控制不足认知偏差，欧美学者多通过设定  $\beta < 1$  以得到“短期高、长期低”的贴现率结构，可以刻画出欧美国家居民易受到短期诱惑而形成的过度消费行为。

儒家消费文化认为冲动型消费是不可取的，强调消费者要“禁奢崇俭”，使消费者逐渐形成过强的自我控制，在双曲线贴现模型中相应的技术表达是  $\beta > 1$ 。此时，会出现“短期低、

长期高”的贴现率结构，从而可以较好地与儒家消费文化“重未来、轻现在”倾向性态度相契合。虽然 $\beta > 1$ ，但由于长期贴现因子 $\delta < 1$ ， $\beta\delta'$ 还是会小于1，消费者的跨期效用之和仍然可以收敛，符合跨期贴现模型的技术要求。因此，本文对 $\beta > 1$ 的设定在逻辑上是前后一致的，在技术上也是可行的。

### 三、消费文化与消费异常的模型分析

为解构消费文化对消费行为的作用机制，本文将 [Koszegi \(2005\)](#)等文献中使用的双曲线贴现模型引入消费文化分析，以自我控制认知偏差 $\beta$ 来表征消费文化。技术上的改进之处在于：其一，对 $\beta$ 进行两个方向的设定，以方便将儒家消费文化及相应的消费拖延行为纳入分析框架。其二，根据本文分析对象重新设定消费者效用函数。其三，对消费者根据其理性程度进行分类。本文的讨论主要集中在创造性奢侈品领域，<sup>①</sup>与生活必需品相比，创造性奢侈品的消费更容易受到消费文化和心理认知等因素的影响，因此是本文的分析重点。

#### （一）基本模型

设消费者生存三期， $t = 0, 1, 2$ 。消费行为只在 $t = 1$ 期发生，消费者需在 $t = 0$ 时决定最优的消费数量。对创造性奢侈品的消费在 $t = 1$ 期会产生即期效用，在 $t = 2$ 时发生成本（如分期付款）。例如，假设消费者决定在 $t = 1$ 时消费 $x$ 单位的奢侈品 $X$ ，则他可以得到的即期效用为 $U(x)$ ， $U(x)$ 是凹函数，即 $U'(x) > 0$ ， $U''(x) < 0$ ；付出的成本是 $xc$ （ $c$ 是单位成本），该成本在 $t = 2$ 时才支付。

假设消费者在 $t = 1$ 期的总收入为 $I$ ，除去奢侈品消费之外剩余的部分用来消费生活必需品 $y$ 。<sup>②</sup>由于奢侈品的消费延迟至 $t = 2$ 时才付款 $xc$ ，因此，若根据市场利率水平假设 $t = 2$ 到 $t = 1$ 期之间的贴现因子为 $\gamma$ ，则消费者在时期1除消费奢侈品之外剩余部分为： $I - \gamma xc$ 。为便于分析，在比较奢侈品 $x$ 和必需品 $y$ 的效用时，可将必需品 $y$ 的消费效用单位化，令 $U(x, y) = U(x) + U(y) = U(x) + y$ ，即令 $U(y) = y$ ，则在 $t = 1$ 时，消费者消费必需品的效用为 $U[(I - \gamma xc)] = I - \gamma xc$ 。

消费者具有自我控制认知偏差，根据双曲线贴现模型，消费者在 $t = 0$ 期的贴现因子结构是 $\{1, \beta\delta, \beta\delta^2\}$ ，消费者时期0的效用最大化的目标函数是：

$$\text{Max}_x \beta\delta[(I - \gamma xc) + U(x)] - \beta\delta^2 xc \quad (2)$$

由最优化一阶条件可得：

$$U'(X^*) = \gamma c + \delta c \quad (3)$$

其中， $X^*$ 为最优的奢侈品消费数量。到了 $t = 1$ 时消费者再次考虑这个消费计划时，从 $t = 2$ 贴现到 $t = 1$ 的贴现因子是 $\beta\delta$ ，因此消费者的效用最大化问题又变成：

$$\text{Max}_x (I - \gamma xc) + U(x) - \beta\delta xc \quad (4)$$

此时的一阶条件是

$$U'(X'') = \gamma c + \beta\delta c \quad (5)$$

$X''$ 为 $t = 1$ 期消费者在受到过度自我控制的干扰而又没有采取任何措施的情况下，实际愿意消费的奢侈品数量。显然，此时消费者的最优条件发生了改变，其结果是消费者的后期消费实践与其早期消费计划相背离，表现为消费异常。

对比（3）式和（5）式可知，当 $\beta > 1$ 时， $U'(X'') > U'(X^*)$ ，由于效用函数是凹函数，因此可以推知 $X'' < X^*$ ，即实际消费量少于最优消费量，出现消费不足；当 $\beta < 1$ 时， $X'' > X^*$ ，出现消费过度。综上所述，通过 $\beta \neq 1$ 的技术设定，本文的模型能够在统一

① 奢侈品可分为创造性奢侈品和浪费性奢侈品（拉茨勒，2003）。前者指可以推动技术进步，提升生活质量的奢侈品如旅游消费等。后者则指纯粹为炫耀性质的高消费，如“人情面子”消费等。

② 剩余的收入如果用来必需品消费还有剩余，就以现金储蓄形式出现，现金某种程度上也相当于必需品。

的模型框架下解释消费过度和消费不足行为。

## (二) 扩展模型

虽然东西方消费文化对各自区域的消费行为影响深远，但即使在一国之内，消费者也还是会存在差别，因此消费文化影响程度也会各有不同。按理性程度可将消费者分为两类，一类是理性消费者， $\beta=1$ ，比例为 $\theta$ ，另一类是存在认知偏差的消费者， $\beta \neq 1$ ，比例为 $1-\theta$ 。则在 $t=0$ 期进行计划时的社会总效用函数是：

$$\text{Max}_x (1-\theta) \{ \beta \delta [(I - \gamma xc) + U(x)] - \beta \delta^2 xc \} + \theta \{ \delta [(I - \gamma xc) + U(x)] - \delta^2 xc \} \quad (6)$$

相应的一阶最优条件为：

$$U'(X^*) = \gamma c + \delta c \quad (7)$$

在 $t=1$ 期，实际消费购买时的总效用函数是：

$$\text{Max}_x (1-\theta) [(I - \gamma xc) + U(x) - \beta \delta xc] + \theta [(I - \gamma xc) + U(x) - \delta xc] \quad (8)$$

相应的一阶最优条件变为：

$$U'(X'') = \gamma c + \delta c + (1-\theta)(\beta-1)\delta c \quad (9)$$

不考虑消费者分层，比较式(3)和(5)可知，消费最优条件的差别是 $U'(X'') - U'(X^*) = (\gamma c + \beta \delta c) - (\gamma c + \delta c) = (\beta-1)\delta c$ ；考虑消费者分层之后，比较(7)、(9)式可知该差别变成了 $U'(X'') - U'(X^*) = (1-\theta)(\beta-1)\delta c$ ，偏差程度缩小到原来的 $(1-\theta)$  ( $\theta < 1$ )倍。 $\theta$ 值越小，偏差程度越大，消费异常程度越严重。

综合上述分析，可以得到如下两个推论：

**推论 1:** 消费率与自我控制力负相关，即自我控制力越强， $\beta$ 越大，消费率越低。

**推论 2:** 理性消费者占比 $\theta$ 越小，异常消费程度越严重。

## 四、实证分析：消费文化与消费率

### (一) 核心变量的衡量指标

#### 1. 消费文化（自我控制力）的衡量指标

本文的分析逻辑是，在消费文化影响下，消费者会形成不同方向和程度的自我控制认知偏差，进而导致不同的消费行为异常。模型的基本结论是：自我控制能力越强，消费率越低。因此，一个合理的消费文化替代变量，既要能够代表消费者的自我控制力，又要具有明显的文化特征。本文采用儒家虚拟变量和性生活指数作为自我控制力的替代变量。

(1) 儒家虚拟变量 (*Rujia*)。前文已经反复提到，欧美文化则强调自我实现，而儒家文化则使人们过度自我克制。这意味着，儒家虚拟变量可以定性地（虽然较为粗略）衡量两类自我控制认知偏差：儒家国家表现为过度自我控制，而非儒家国家则表现为自我控制不足。

(2) 性生活指数，包括三个指标：一年中的性生活频率 (*Sex\_Freq*)、不采取安全措施的行为占总体性行为的比例 (*Sex\_Unsafe*)、初次性生活的年龄 (*Sex\_Age*)。前两个指标与自我控制力负相关，*Sex\_Age* 与自我控制力正相关。选择性生活指数作为自我控制力的替代指标主要是因为性行为与自我控制力密切相关，这在心理学和性科学文献中得到了大量的经验支持。例如，在自我控制力低的人群中，性生活频率明显较高 (Gailliot and Baumeister, 2007)，经历初次性行为的年龄相对较低 (Toates, 2009)，在性行为过程中更倾向于不采取安全措施 (Trost et al, 2002, Quinn and Fromme, 2010)。更为重要的是，Gailliot and Baumeister (2007) 发现，在其他行为中（如理财、消费、按期完成任务等）表现出较强的自我控制力的人，也能够较好地控制其性行为。

#### 2. 理性消费者占比的替代指标

O'Donoghue and Rabin (2001) 在双曲线贴现模型框架下将消费者分为两类，一类完全不

知道自己认知偏差程度，称为无知的消费者；另一类知道自己存在认知偏差，并会通过各种途径来纠正自己的消费偏差行为，称为精明的消费者，即本文中的理性消费者。Lawrance (1991) 和 Sourdin (2008) 指出，教育水平是影响消费者理性程度的重要变量。Gailliot and Baumeister (2007) 将自控能力分为两种：特质型和状态型。前者更多地归因为先天因素，例如，有些人在很小的时候便表现出很强的自我控制力。后者则主要强调外部环境的影响，如文化、法律、教育等。因此，虽然我们无法获取理性消费者占比的一手资料，但可以采用教育水平作为替代指标，因为教育能有效改变个体的状态型自控能力（Pongratz, 2006）。

### 3. 异常消费率

异常消费率定义为实际消费率与正常消费率之间的偏差，可以采用两步法估算。第一步，我们选取了一些前期文献中已经确认的重要变量，并采用它们的线性拟合值来衡量正常消费率。具体而言，我们首先估计了如下线性模型：

$$Consum_{it} = \alpha + Z_{it}\beta + \varepsilon_{it} \quad (10)$$

其中， $Consume$  为居民的最终消费率，定义为居民最终消费占 GDP 的比例， $Z$  为影响消费率的变量：收入水平 ( $GDPper$ )、真实利率 ( $Real\_i$ )、社会保障水平 ( $SocialSecu$ )、赡养率 ( $Depend$ )、金融发展水平 ( $FinanDev$ )。第二步，利用模型 (10) 的 OLS 估计值  $\hat{\alpha}$  和  $\hat{\beta}$  可以得到  $Consume$  的拟合值  $NormConsum_{it} = \hat{\alpha} + Z_{it}\hat{\beta}$ ，异常消费率可以用残差衡量，即：

$$Ex\_Consume_{it} = Consume_{it} - NormConsum_{it} \quad (11)$$

(二) 消费率与消费文化的实证模型设定

为验证推论 1，本文设定了如下线性回归模型：

$$Consum_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Controls_{it} + \varepsilon_{it} \quad (12)$$

其中， $Consume$  的定义同前， $\varepsilon_{it}$  为随机干扰项。 $X$  为消费文化（自我控制力）的代理变量，分别由上一小节的儒家文化虚拟变量 ( $Rujia$ ) 和性生活指数 ( $Sex\_Freq$ 、 $Sex\_Age$ 、 $Sex\_Unsafe$ ) 来衡量。 $Controls$  表示前期文献中提及的一系列可能影响消费率的控制变量，包括：居民收入 ( $GDPper$ )、社会保障支出 ( $SocialSecu$ )、实际利率水平 ( $Real\_i$ )、长短期利差 ( $Rategap$ )、儿童和老年负担系数 ( $Depend$ )、金融市场发展程度 ( $FinanDev$ )。

为验证推论 2，本文设定了如下模型：

$$|Ex\_Consume_{it}| = \alpha + \beta_1 Edu_{it} + \beta_2 |Ex\_Sex_{it}| + \varepsilon_{it} \quad (13)$$

其中， $Ex\_Consume$  为(12)式中对应的异常消费率。由于这里重点关注的是实际消费率与正常消费率的偏离程度（不再区分消费不足或消费过度），在(13)式中采用  $Ex\_Consume$  的绝对值 ( $|Ex\_Consume|$ ) 作为被解释变量。 $Edu$  表示教育水平，用大学入学率来衡量，该值越高则理性消费者比重越高。 $Ex\_Sex$  表示自我控制认知偏差，用第  $i$  个国家在第  $t$  年的性生活频率与样本均值的离差来衡量，即  $Ex\_Sex_{it} = Sex\_Freq_{it} - \overline{Sex\_Freq}$ 。前文已经提到，自我控制力可以分为特质型和状态型两类，而文化和教育等外部因素只能影响后者，而无法改变前者。我们假设特质型自控能力不存在国别和时间差异，采用  $\overline{Sex\_Freq}$  来衡量，因此， $Ex\_Sex_{it}$  主要衡量了状态型自控能力。此外，类似于  $|Ex\_Consume|$  的设定思路，同样采用  $Ex\_Sex_{it}$  的绝对值，即  $|Ex\_Sex_{it}|$  作为解释变量。

若推论 2 是合理的，则可以预期，在控制自我控制认知偏差程度 ( $|Ex\_Sex_{it}|$ ) 的前提下， $Edu$  的系数  $\beta_1$  应显著为负。

(三) 数据来源和基本统计分析

本文的数据主要来源于世界银行发展数据库，包含 48 个国家和地区在 1978-2007 年期间的面板数据。<sup>①</sup>性生活指数的相关数据来源于 Durex 公司针对全球范围进行的调查报告。

① 这 48 个国家和地区分别是：阿根廷、澳大利亚、奥地利、比利时、巴西、加拿大、智利、中国、哥伦比亚、捷克、埃及、丹麦、芬兰、法国、德国、希腊、香港、匈牙利、冰岛、印度、印尼、伊朗、爱尔兰



表 1 列示了文中主要变量的计算方法、基本统计量和数据来源。

| 变量名称              | 平均值    | 标准差   | 最小值    | 最大值    | 定义及数据来源                                |
|-------------------|--------|-------|--------|--------|----------------------------------------|
| <i>Netsaving</i>  | 0.115  | 0.076 | -0.290 | 0.420  | 净的国民储蓄占 GDP 比例 <sup>a</sup>            |
| <i>Domsaving</i>  | 0.233  | 0.075 | -0.040 | 0.540  | 居民国内储蓄占 GDP 比例 <sup>a</sup>            |
| <i>Consume</i>    | 0.771  | 0.082 | 0.480  | 1.047  | 居民最终消费占 GDP 比例 <sup>a</sup>            |
| <i>GDPper</i>     | 3.912  | 0.511 | 2.453  | 4.750  | 收入：人均 GDP 取对数 <sup>a</sup>             |
| <i>Edu</i>        | 0.046  | 0.015 | 0.010  | 0.080  | 教育：大学生入学率 <sup>a</sup>                 |
| <i>Real_i</i>     | -0.008 | 0.119 | -0.928 | 0.176  | 真实利率：一年期储蓄率-CPI 增长率 <sup>a</sup>       |
| <i>Rategap</i>    | 0.023  | 0.046 | -0.328 | 0.365  | 利差：市场长期利率-短期利率 <sup>b</sup>            |
| <i>SocialSecu</i> | 0.276  | 0.152 | 0.007  | 0.565  | 社会保障：社会保障支出/政府总支出 <sup>c</sup>         |
| <i>Depend</i>     | 0.563  | 0.116 | 0.379  | 0.968  | 负担系数：≤15 岁和≥65 岁人口数/总就业人数 <sup>a</sup> |
| <i>FinanDev</i>   | 0.109  | 0.336 | 0.000  | 1.417  | 金融市场发展：ln(国内信贷总额/GDP) <sup>a</sup>     |
| <i>Rujia</i>      | 0.158  | 0.365 | 0.000  | 1.000  | 儒家思想圈国家为 1，其他为 0 <sup>d</sup>          |
| <i>Sex_Freq</i>   | 0.294  | 0.030 | 0.203  | 0.384  | 性生活频率：一年当中性生活次数/360 天 <sup>e</sup>     |
| <i>Sex_Unsafe</i> | 0.510  | 0.124 | 0.245  | 0.735  | 不采取安全措施性生活的比例 <sup>e</sup>             |
| <i>Sex_Age</i>    | 17.352 | 1.162 | 15.600 | 23.000 | 最早开始性生活的年龄 <sup>e</sup>                |

注：(1) *Sex\_Freq*、*Sex\_Unsafe* 和 *Sex\_Age* 的样本数分别为 590、590 和 624，其它变量的样本数均为 822。(2) 数据来源简写如下，a：世界银行发展数据库；b：IMF 国际金融数据库；c：IMF 政府财政年鉴；d：本文整理，儒家文化圈包括中国、香港、印度尼西亚、日本、韩国、马来西亚、菲律宾、新加坡、泰国和越南；e：Durex 全球性调查报告，下载地址为 <http://www.durexnetwork.org/en-GB/research>。

#### (四) 回归结果及分析

##### 1. 消费率与自我控制力：推论 1 的检验结果

为验证推论 1，本文采用 OLS 估计了模型(12)，回归结果呈现于表 2。在第(1)列中，我们重点考察了前期文献中提及的主要变量对居民消费率的影响。其中，收入水平(*GDPper*)和利差(*Rategap*)均显著为负，与凯恩斯绝对收入假说的理论预期一致；社会保障程度(*SocialSecu*)显著为正，与预防性储蓄理论一致。负担系数(*Depend*)显著为正，与生命周期理论相悖。真实利率水平(*Real\_i*)以及金融发展水平(*FinanDev*)并不能很好地解释国家之间的消费率差异。我们注意到，该模型的  $R^2$  仅为 0.044，意味着在本文的样本中，前期文献中所强调的经济因素的解釋能力非常有限。如前文所述，消费文化等不可观测个体因素在很大程度上影响着居民的消费行为。为此，在第(2)列中，我们进一步加入了 47 个反映国家个体效应的虚拟变量。此时，模型的  $R^2$  提高为 0.828，这意味着，不随时间改变的个体因素（如风俗、文化等）能够进一步解释约 79% (0.83-0.04) 的消费行为的变动。<sup>①</sup>在第(3)列中，我们进一步加入了反映时间特征的年度虚拟变量，发现在多数年份中，相应的虚拟变量并不显著。<sup>②</sup>这表明，在本文的样本区间内，居民的消费行为具有高度的稳定性。

从上面的分析可以看出，前期文献中所强调的经济变量似乎无法很好地解释不同国家之间的消费差异。虽然第(2)列的结果表明，个体效应能在很大程度上解释消费率的截面差异，但我们更为关心的是：这些个体效应到底是什么？

兰、意大利、荷兰、新西兰、挪威、基斯丹、日本、菲律宾、波兰、葡萄牙、罗马尼亚、俄罗斯、韩国、新加坡、南非、卢森堡、马来西亚、墨西哥、西班牙、瑞典、瑞士、泰国、土耳其、英国、美国、越南。

① 我们进一步采用似然比检验 (LR test) 检验了个体效应的联合显著性(原假设为  $H_0: \alpha_i = 0$ )，得到的  $\chi^2$  值为 1457.0，相应的  $p$  值为 0.000。

② 采用似然比检验对第(3)栏中反映年度效应的虚拟变量执行联合显著性检验(原假设为  $H_0: \lambda_t = 0$ )，得到的  $\chi^2$  值为 35.2，相应的  $p$  值为 0.197。

表 2 推论 1 检验结果：自我控制认知偏差对最终消费率的影响

|                          | (1)                  | (2)                   | (3)                   | (4)                   | (5)                 | (6)                 | (7)                  | (8)                  |
|--------------------------|----------------------|-----------------------|-----------------------|-----------------------|---------------------|---------------------|----------------------|----------------------|
| <i>GDPper</i>            | -0.026***<br>(-3.41) | -0.206***<br>(-16.60) | -0.203***<br>(-15.85) | -0.055***<br>(-7.79)  | -0.000<br>(-0.07)   | -0.003<br>(-0.35)   | -0.025***<br>(-3.21) | -0.039***<br>(-5.43) |
| <i>Real_i</i>            | 0.002<br>(0.08)      | 0.082***<br>(5.94)    | 0.085***<br>(6.10)    | 0.083***<br>(3.79)    | 0.179***<br>(3.79)  | 0.197***<br>(3.08)  | 0.048<br>(1.51)      | 0.186***<br>(4.18)   |
| <i>Rategap</i>           | -0.166***<br>(-2.70) | 0.077**<br>(2.24)     | 0.079**<br>(2.27)     | -0.088<br>(-1.63)     | 0.044<br>(0.89)     | -0.093<br>(-1.41)   | -0.098<br>(-1.64)    | 0.104**<br>(2.19)    |
| <i>SocialSecu</i>        | 0.140***<br>(5.58)   | 0.005<br>(0.26)       | 0.001<br>(0.06)       | -0.025<br>(-1.02)     | 0.022<br>(1.09)     | 0.156***<br>(5.75)  | 0.124***<br>(5.23)   | -0.057***<br>(-2.70) |
| <i>Depend</i>            | 0.058**<br>(2.20)    | 0.003<br>(0.17)       | 0.013<br>(0.48)       | -0.045*<br>(-1.88)    | 0.067**<br>(2.48)   | 0.044<br>(1.20)     | 0.042<br>(1.40)      | 0.033<br>(1.27)      |
| <i>FinanDev</i>          | 0.013<br>(1.54)      | 0.008<br>(0.16)       | 0.005<br>(0.11)       | 0.011<br>(1.42)       | 0.044***<br>(7.51)  | 0.017**<br>(2.25)   | 0.015**<br>(2.05)    | 0.032***<br>(5.68)   |
| <i>Rujia</i>             |                      |                       |                       | -0.140***<br>(-15.57) |                     |                     |                      | -0.113***<br>(-8.76) |
| <i>Sex_Freq</i>          |                      |                       |                       |                       | 1.752***<br>(22.07) |                     |                      | 1.149***<br>(11.31)  |
| <i>Sex_Unsafe</i>        |                      |                       |                       |                       |                     | 0.085***<br>(3.31)  |                      |                      |
| <i>Sex_Age</i>           |                      |                       |                       |                       |                     |                     | -0.025***<br>(-9.69) |                      |
| <i>Constant</i>          | 0.805***<br>(23.57)  | 1.622***<br>(32.77)   | 1.606***<br>(29.25)   | 1.041***<br>(30.97)   | 0.204***<br>(5.86)  | 0.664***<br>(17.85) | 1.251***<br>(19.34)  | 0.598***<br>(10.74)  |
| 个体效应                     | No                   | Yes                   | Yes                   | No                    | No                  | No                  | No                   | No                   |
| 年度效应                     | No                   | No                    | Yes                   | No                    | No                  | No                  | No                   | No                   |
| <i>N</i>                 | 822                  | 822                   | 822                   | 822                   | 590                 | 590                 | 624                  | 590                  |
| <i>adj-R<sup>2</sup></i> | 0.044                | 0.828                 | 0.829                 | 0.262                 | 0.506               | 0.110               | 0.207                | 0.563                |

注：(1) \*\*\*、\*\* 和 \* 分别表示在 1%、5% 和 10% 水平上显著，括号中基于 White 异方差稳健型标准误计算而得的 t 值。(2) 被解释变量均为最终消费率 (Consume)。

为此，在第 (4)–(6) 列中，我们分别在第 (1) 列中的模型设定基础上，依次加入了反映消费文化的代理变量。<sup>①</sup>由第 (4) 列的结果来看，*Rujia* 虚拟变量的系数为 -0.14，且在 1% 水平上显著异于零。在本文的样本中，最终消费率 (Consume) 的平均值为 0.771 (见表 1)，这意味着，在其他条件相同的情况下，儒家文化圈国家的消费率比样本平均水平低了约 18% (=0.14/0.771)。从模型的整体拟合优度来看，第 (4) 列的  $R^2$  比第 (1) 列提高了 21.8% (=0.262-0.044)，表明在上文提及的个体效应中，约有 28% (= 21.8%/79%) 可以藉由儒家文化来解释。

在第 (5) 列中，我们采用了另一个能够更直观地反映自我控制力的代理变量——性生活频率 (*Sex\_Freq*)，其系数估计值在 1% 水平上显著为正，表明自我控制力越弱，消费率越高。此时模型的拟合优度  $R^2$  为 0.506，比第 (1) 列提高了约 46% (=0.506-0.044)，这意味着约有 58.2% (46%/79%) 的个体效应可以归因为自我控制力所代表的消费文化。<sup>②</sup>

① 需要说明的是，在第 (4)–(6) 列的回归分析中，并未加入反映个体效应的国家虚拟变量，原因有二：其一，我们分析的重点是揭示个体效应的具体构成；其二，由于儒家经济圈虚拟变量 (*Rujia*) 不随时间改变，它会与个体效应虚拟变量完全共线性，而反映性生活指数的指标虽然随时间变化，但年度之间的差异很小，这使得性生活指数与个体效应虚拟变量高度共线性，若同时加入会大幅降低统计推断的有效性。

② 注意到第 (5) 栏和第 (4) 栏中所使用的观察值个数有所差异，但这并不影响上述结论。我们用第 (5) 栏中的样本重新估计了第 (4) 栏中的模型，得到的  $R^2$  为 0.253。

作为稳健性检验,在第(6)和第(7)列中,我们分别采用不采取安全措施性生活的比例(*Sex\_unsafe*)和初次性生活的年龄(*Sex\_Age*)作为自我控制力的替代指标。二者的估计系数分别为 0.085 和 -0.025,且均在 1%水平上显著异于零,与第(5)列中基于 *Sex\_Freq* 指标得到的结论一致。<sup>①</sup>当然,在第(6)和第(7)列中, $R^2$  分别为 0.110 和 0.207,均低于第(5)列。<sup>②</sup>为此,后续分析将主要采用 *Sex\_Freq* 来衡量自我控制力及相应的消费文化。

在第(8)列中,我们在第(1)列中模型设定的基础上,同时加入反映消费文化的虚拟变量 *Rujia* 和 *Sex\_Freq*,二者的系数分别为 -0.113 和 1.149,且均在 1%水平上显著异于零。此时的  $R^2$  为 0.563,略高于第(5)列中仅加入 *Sex\_Freq* 时的 0.506。这似乎表明,在控制了自我控制力(*Sex\_Freq*)后,儒家文化虚拟变量(*Rujia*)对国家之间消费率差异的解释能力明显下降了。进一步分析表明,二者之间的 Spearman 相关系数高达 -0.67。这意味着,居民自我控制认知偏差主要根源于其所处的文化环境,甚至已经内化于所处的文化氛围之中。

上述分析表明,本文的推论 1——自我控制力越强消费率越低,得到了较为稳健的经验支持。图 1 中的散点图更为直观地反映了上述结论。图 1 中的纵轴为(12)式中定义的异常消费率(*Ex\_Consume*)。从计量经济学的角度来看,*Ex\_Consume* 可视为实际消费中无法用传统经济变量(即(10)式中的 *Z* 向量)解释的部分。若模型(12)的设定是正确的,则 *Ex\_Consume* 中主要包含了自我控制力对消费率的影响。图 1 中的横轴为反映自我控制力的变量 *Sex\_Freq* (为了更为直观,这里的横轴变量是 *Sex\_Freq*×360,即每年性生活次数)。<sup>③</sup>因此,图 1 中散点图的斜率实际上可以视为自我控制力(*Sex\_Freq*)对消费率的(条件)边际影响。换个角度来看,若把 *Ex\_Consume* 视为异常消费(与  $|Ex\_Consume|$  略有差异),则图 1 反映了异常消费与文化变量之间的显著相关关系。

可以看出,归属于儒家文化圈的国家都表现为消费不足( $Ex\_Consume < 0$ ),其中消费不足程度最大的三个国家是:新加坡、马来西亚和中国;而多数归属于非儒家文化圈的国家则倾向于过度消费,倾向最强的三个国家是:希腊、美国和英国。<sup>④</sup>同时,我们也发现了一个非常值得深入探讨的问题。在现有文献中,谈及“高储蓄率”或“低消费率”问题时,学者们往往将关注的焦点集中于中国,如 Modigliani and Cao (2004)、Kroeber (2011)等。然而,图 1 表明,在控制了收入水平(*GDPper*)、社会保障水平(*SocialSecu*)等因素后,新加坡居民的消费偏差明显大于中国,其消费不足的程度更为严重。这显然无法用传统的预防性储蓄理论或流动性约束理论来解释,而本文强调的自我控制认知偏差则具有很好的解释力。

在稳健性检验中,我们进一步采用国内储蓄率(*Domsaving*)和净储蓄率(*Netsaving*)作为被解释变量,分别采用表 2 中第(4)列和第(5)列的模型设定形式,重新估计了模型(12)。<sup>⑤</sup>结果表明,当采用 *Domsaving* 作为被解释变量时,*Rujia* 和 *Sex\_Freq* 的系数分别为 0.108 和 -1.423,且均在 1%水平上显著,表明儒家国家的储蓄率明显较高,自我控制力越低则储蓄率越低。采用 *Netsaving* 作为被解释变量时得到的结果与此相似。由于储蓄率与消费率高度负相关,因此,上述结果与表 2 引申出的结论是一致的。

① 需要说明的是,初次性生活的年龄(*Sex\_Age*)越小,表明消费者的冲动行为越严重,因此,基于 *Sex\_Age* 得到的估计结果与基于 *Sex\_Freq* 和 *Sex\_Unsafe* 得到的结果具有相反的符号。

② 一个可能的原因是,相对于性生活次数(*Sex\_Freq*),初次性生活的年龄(*Sex\_Age*)和不安全性生活的比例(*Sex\_Unsafe*)可能存在较为严重的衡量偏差。

③ 图 1 中的横轴和纵轴变量均为各个国家在 1978-2007 年期间的平均值。我们也分年度绘制了二者的散点图,所得到的结论并不存在明显差异。有兴趣的读者可以向我们索要相关结果。

④ 在图 1 中,日本居民性生活次数均值仅为 45 次/年,远低于样本均值 101 次/年。我们查阅的十余篇相关文献都报告了相似的结果,这表明本文所使用的 Durex 公司发布的数据并不存在明显的衡量偏差。

⑤ 受限于篇幅,相关估计结果未能呈现,有兴趣的读者可以向我们索要。

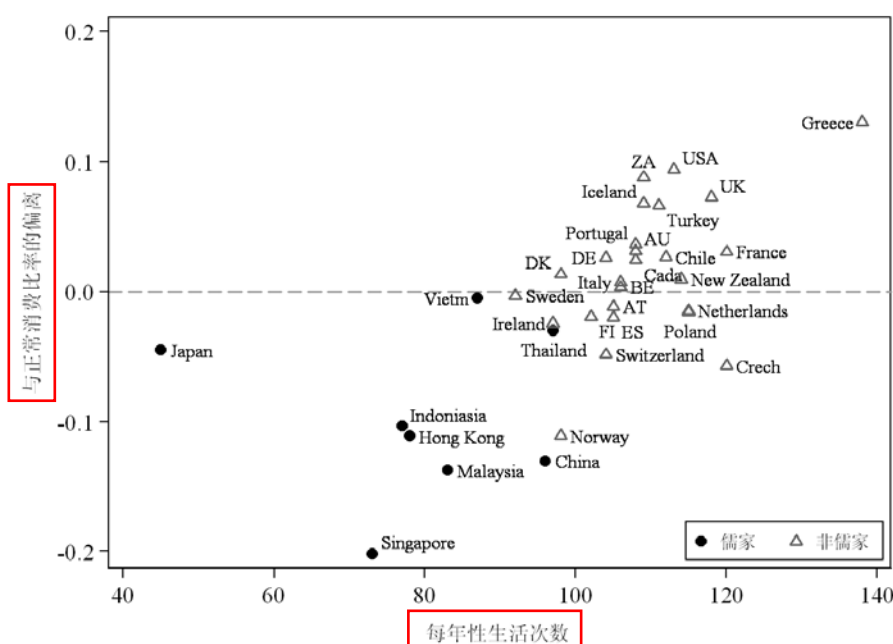


图 1 消费偏差与每年性生活次数的散点图（1998-2007）

注：部分国家名称采用了简写，AU (Australia), AT (Austria), BE (Belgium), DK (Denmark), FI (Finland), DE (Germany), ZA (South Africa), ES (Spain), UK (United Kingdom), USA (United State)。

## 2. 异常消费与理性消费者占比：推论 2 的检验结果

表 3 呈现了模型(13)的估计结果。在 A 栏中，我们采用  $|Ex\_consume|$  作为被解释变量。在第(1)列中，我们仅加入了  $|Ex\_Sex|$ ，其系数估计值为 0.598，在 1%水平上显著，表明自我控制认知偏差越大，则异常消费越严重，换言之，消费文化差异程度与消费异常程度一一对应，这进一步验证了推论 1 的合理性。由第(2)列中的结果可知， $Edu$  在 1%水平上显著为负，表明理性消费者比重的提高有助于降低异常消费率，初步证实了推论 2。第(3)列是本文最为关注的，我们同时加入了  $Edu$  和  $|Ex\_Sex|$ ，此时  $Edu$  仍然显著为负，这表明在控制了自我控制认知偏差的前提下，理性消费者比重的提高仍然有助于降低异常消费率。换言之，上述结果在印证了本文的推论 2 的同时，也表明推论 2 是推论 1 的进一步深化。

在 B 栏和 C 栏中，我们分别采用异常国内储蓄率 ( $|Ex\_Domsaving|$ ) 和异常净储蓄率 ( $|Ex\_Netsaving|$ ) 作为被解释变量（二者的估算方法与  $|Ex\_consume|$  相似），得到的结论与 A 栏中一致。这一方面表明本文对推论 2 的经验分析结论是稳健的，另一方面也表明代表消费文化的自我控制认知偏差不仅适于解释消费行为，也能够解释储蓄行为。

## 五、政策含义：扩大内需政策的理论基础与技术创新

由本文的理论分析可知，影响消费行为的参数主要有消费收益  $b$ 、成本  $c$ 、长期贴现因子  $\delta$  和短期贴现因子  $\beta$ 。在新古典时间偏好理论框架下，居民没有自我控制认知偏差， $\beta=1$ ，因此不会出现消费偏差行为，无需政府进行干预。但在现实中，各国政府实际上是频繁地进行干预（如公共支出和税收等），其目的是影响  $b$  和  $c$ 。在行为经济学框架下，居民具有程度不一的自我控制认知偏差， $\beta \neq 1$ ，并会通过杠杆作用，使得  $b$  和  $c$  发生倍数效应的变化，形成消费异常，需要外部干预，这就为“扩大内需”等政府干预行为提供了微观基础。

从本文实证结果来看，社会保障支出、金融市场发展与消费率显著正相关，因此加大社会保障支出，加快金融市场发展能够从一定程度上刺激消费。收入与消费率显著负相关，与凯恩斯消费函数性质吻合。其政策含义在于，虽然增加居民收入无法促进消费率的增长，但



表 3 推论 2 的检验结果：异常消费与理性消费者占比

|                                   | (1)              | (2)               | (3)               |
|-----------------------------------|------------------|-------------------|-------------------|
| <b>A:   <i>Ex_Consume</i>  </b>   |                  |                   |                   |
| <i>Ex_Sex</i>                     | 0.598*** (8.40)  |                   | 0.495*** (6.68)   |
| <i>Edu</i>                        |                  | -0.640*** (-5.83) | -0.520*** (-4.30) |
| <i>Constant</i>                   | 0.040*** (14.88) | 0.090*** (17.28)  | 0.068*** (9.63)   |
| <i>N</i>                          | 608              | 854               | 608               |
| <i>adj-R</i> <sup>2</sup>         | 0.103            | 0.037             | 0.128             |
| <b>B:   <i>Ex_Domsaving</i>  </b> |                  |                   |                   |
| <i>Ex_Sex</i>                     | 0.475*** (6.90)  |                   | 0.318*** (4.54)   |
| <i>Edu</i>                        |                  | -0.761*** (-7.99) | -0.800*** (-7.01) |
| <i>Constant</i>                   | 0.037*** (14.07) | 0.088*** (19.55)  | 0.080*** (11.97)  |
| <i>N</i>                          | 609              | 864               | 609               |
| <i>adj-R</i> <sup>2</sup>         | 0.071            | 0.068             | 0.139             |
| <b>C:   <i>Ex_Netsaving</i>  </b> |                  |                   |                   |
| <i>Ex_Sex</i>                     | 0.252*** (3.45)  |                   | 0.066 (0.89)      |
| <i>Edu</i>                        |                  | -0.740*** (-7.57) | -0.947*** (-7.92) |
| <i>Constant</i>                   | 0.042*** (15.31) | 0.085*** (18.38)  | 0.093*** (13.36)  |
| <i>N</i>                          | 609              | 864               | 609               |
| <i>adj-R</i> <sup>2</sup>         | 0.018            | 0.061             | 0.108             |

注：\*\*\*、\*\* 和 \* 分别表示在 1%、5% 和 10% 水平上显著，括号中基于 White 异方差稳健型标准误计算而得的 *t* 值。

能在消费支出的绝对增长上发挥作用。实际利率与消费率显著正相关，与储蓄理论预期不一致。但由于利率政策与货币政策密切关联，因此很难作为扩大内需的政策工具。

本文的重要发现在于，消费文化是解释消费率国别差异的主要因素，儒家文化影响力越强，消费率越低。消费文化主要通过影响居民的自我控制力来影响消费，自我控制力越强，消费率越低。因此从文化角度进行扩大内需政策设计，可以从两方面着手，一方面是直接从文化角度切入，推行理性消费观念。另一方面，可间接地通过影响居民的自我控制力入手，纠正儒家文化对居民消费的过度抑制。

在文化宣传方面，应通过各种途径加强对居民理性消费观的培养。欧美国家消费者之所以形成目前的过度消费的消费文化，很大程度上是因为 20 世纪 30 年代经济大危机之后，对凯恩斯主义扩张经济政策的持续推行和长期宣传“消费即是爱国”理念的结果。

影响我国居民的自我控制力有两种可能的思路。一种是削弱居民的自我控制力，例如，加快资产证券化进程，完善金融市场功能，缓解居民消费的流动性约束，弱化居民的过度自我控制倾向（Laibson, 1997）。另一种思路是利用双曲线贴现模型提供的能有效纠正自我控制认知偏差的“锁定”技术。锁定是指为防止未来的相机抉择行为，消费者自身或借助外力提前进行投资或消费支付，利用惩罚机制来强制消费者实施最初的计划（叶德珠，2010）。

在目前中国为刺激消费而出台的政策中，降低税率、完善社会保障体系、增加农民收入、增加公共基础设施投资等措施被寄予厚望。这些政策属于传统的成本-收益干预范畴，虽然具有一定的短期效果，但却存在长期瓶颈，因为即使公共基础设施建设达到日本的水平，也仍然难以避免多年的消费低迷。相比之下，反而是一些技术性规定对刺激消费的作用更加直接明显。比如，政府制定的“双休日”和“黄金周”政策就较强地促进了居民的旅游及相关消费。从本文模型逻辑来看，黄金周长假期强制规定了居民旅游休闲消费的下限，实际上是相当于一种锁定技术。这种锁定技术在财政资源占用上不大，但却能有效地撬动消费，“四

两拨千斤”式的杠杆作用较为显著。

目前,中国的出口导向型发展战略面临巨大挑战,扩大内需的压力剧增,这种压力对于纠正中国国民保守消费心态、培养理性消费理念未尝不是一个机遇。中国政府应该把握这个机遇,有意识地围绕锁定技术来进行制度设计,以有效地扩大内需,获得可持续发展的能力。

## 六、结 论

文化影响消费是一个基本共识,但该命题在理论和实证分析上的支撑却非常有限。本文采用行为经济学双曲线贴现模型,在思想路线上,用自我控制认知偏差来描述消费文化对居民消费认知态度的倾向性影响;在技术路线上,用短期贴现因子  $\beta$  来表达消费文化的非线性特征。从而保证了在同一个模型框架下,以对短期贴现因子  $\beta$  不同方向的赋值,可分别刻画出消费过度和消费不足,逻辑一致地解释东西方消费行为的差异。本文的主要理论结论是:居民受儒家文化影响越深,自我控制力越强,则消费率越低;受欧美文化影响越深,自我控制力越弱,则消费率越高。

在实证层面上,本文用 48 个国家和地区在 1978-2007 年间的面板数据,在控制了收入等传统变量的前提下,以儒家文化虚拟变量和性生活指数作为消费文化的主要替代变量,对消费率进行回归。结果表明:其一,文化确实是影响消费的主要因素。收入等传统变量对消费率国别差异解释力不足 5%,而不随时间改变的个体因素则能解释约 79% 的消费差异。在这些个体因素中,约有 28% 可由儒家文化虚拟变量来解释,有 58% 可以归因为以性生活指数为代表的文化因素。其二,消费偏差程度与自我控制认知偏差程度显著正相关,说明文化差别与消费行为差别存在一一对应关系。其三,消费偏差程度与代表理性消费者占比的教育变量显著负相关,说明一国理性消费者越少,消费偏差程度越严重。

在政策操作层面,本文模型结论表明,作为消费行为偏差的根本内因,消费认知偏差会放大居民消费时的成本收益比较,因此传统的仅仅针对消费收益与成本进行的政策措施会由于认知偏差因子的杠杆作用而事倍功半。双曲线贴现模型框架为我们提供了纠正认知偏差因子的锁定技术,围绕锁定技术进行政策设计可起到“四两拨千斤”作用。这个模型结论可以解释传统政策的困境和像“黄金周”这类措施的意想不到的效果,也为进一步的扩大内需政策创新提供了新思路。

## 参考文献

- 龙志和、周浩明, 2000:《中国城镇居民预防性储蓄实证研究》,《经济研究》第 11 期。
- 拉茨勒·沃尔冈, 2003:《奢侈带来富足》,北京:中信出版社。
- 罗楚亮, 2004:《经济转轨、不确定性与城镇居民消费行为》,《经济研究》第 4 期。
- 万广华、张茵、牛建高, 2001:《流动性约束、不确定性与中国居民消费》,《经济研究》第 11 期。
- 叶德珠, 2010:《和谐社会构建与政府干预的路径选择》,《经济学季刊》第 1 期。
- 朱信凯、骆晨, 2011:《消费函数的理论逻辑与中国化:一个文献综述》,《经济研究》第 1 期。
- Akerlof, G. A., 1991, “Procrastination and Obedience”, *American Economic Review* 81 (2): 1-19.
- Becker, G. S., C. Mulligan, 1997, “The Endogenous Determination of Time Preference”, *Quarterly Journal of Economics* 112 (3): 729-758.
- Briley, D. A., M. W. Morris, I. Simonson, 2000, “Reasons as Carriers of Culture: Dynamic Vs. Dispositional Models of Cultural Influence on Decision Making”, *Journal of Consumer Research* 27 (2): 157-178.
- Carroll, C. D., J. Overland, D. N. Weil, 2000, “Saving and Growth with Habit Formation”, *American Economic Review* 90 (3): 341-355.
- Deaton, A., 1991, “Saving and Liquidity Constraints”, *Econometrica* 59 (5): 1221-1248.

- Fisher, I., 1930, "The Theory of Interest", New York: Macmillan.
- Gailliot, M. T., R. F. Baumeister, 2007, "Self-Regulation and Sexual Restraint: Dispositionally and Temporarily Poor Self-Regulatory Abilities Contribute to Failures at Restraining Sexual Behavior", *Personality and Social Psychology Bulletin* 33 (2): 173-186.
- Gruber, J., B. Koszegi, 2001, "Is Addiction "Rational": Theory and Evidence", *Quarterly Journal of Economics* 116 (4): 1261-1303.
- Harbaugh, R., 2003, "China's High Savings Rates", *Working Paper*.
- Hubbard, R. G., J. Skinner, S. P. Zeldes, 1995, "Precautionary Saving and Social Insurance", *Journal of Political Economy* 103 (2): 360-399.
- Johar, G. V., D. Maheswaran, L. A. Peracchio, 2006, "Mapping the Frontiers: Theoretical Advances in Consumer Research on Memory, Affect, and Persuasion", *Journal of Consumer Research: An Interdisciplinary Quarterly* 33 (1): 139-149.
- Koszegi, B., 2005, "On the Feasibility of Market Solutions to Self-Control Problems", *Swedish Economic Policy Review* 12 (2): 71-94.
- Kroeber, A., 2011, "China's Consumption Paradox: Causes and Consequences", *Eurasian Geography and Economics* 52 (3): 330-346.
- Krusell, P., B. Kuruscu, A. A. Smith, 2002, "Equilibrium Welfare and Government Policy with Quasi-Geometric Discounting", *Journal of Economic Theory* 105 (1): 42-72.
- Laibson, D., 1997, "Golden Eggs and Hyperbolic Discounting", *Quarterly Journal of Economics* 112 (2): 443-477.
- Lawrance, E. C., 1991, "Poverty and the Rate of Time Preference: Evidence from Panel Data", *Journal of Political Economy* 99 (1): 54-77.
- Leland, H. E., 1968, "Saving and Uncertainty: The Precautionary Demand for Saving", *Quarterly Journal of Economics* 82 (3): 465-473.
- Lu, L., and I. McDonald, 2006, "Does China Save Too Much?", *Singapore Economic Review* 51 (3): 283-301.
- Modigliani, F., R. Brumberg, 1954, "Utility Analysis and the Consumption Function: An Attempt at Integration", in K. Kurihara ed, *Post-Keynesian Economics* (Rutgers University Press, New Brunswick, NJ) 388-436.
- Modigliani, F., S. L. Cao, 2004, "The Chinese Saving Puzzle and the Life-Cycle Hypothesis", *Journal of Economic Literature* 42 (1): 145-170.
- Mouawiya, A., A. Elhiraika, 2003, "Cultural Effects and Savings: Evidence from Immigrants to the United Arab Emirates", *Journal of Development Studies* 39 (5): 139-151.
- O'Donoghue, T., M. Rabin, 2001, "Choice and Procrastination", *Quarterly Journal of Economics* 116 (1): 121-160.
- Pongratz, L. A., 2006, "Voluntary Self-Control: Education Reform as a Governmental Strategy", *Educational Philosophy and Theory* 38 (4): 471-482.
- Quinn, P. D., K. Fromme, 2010, "Self-Regulation as a Protective Factor against Risky Drinking and Sexual Behavior", *Psychology of Addictive Behaviors* 24 (3): 376-385.
- Sourdin, P., 2008, "Pension Contributions as a Commitment Device: Evidence of Sophistication among Time-Inconsistent Households", *Journal of Economic Psychology* 29 (4): 577-596.
- Summers, L. H., 1984, "The after-Tax Rate of Return Affects Private Savings", *American Economic Review* 74 (2): 249-253.
- Toates, F., 2009, "An Integrative Theoretical Framework for Understanding Sexual Motivation, Arousal, and

Behavior”, *Journal of Sex Research* 46 (2): 168-193.

Trobst, K. K., J. H. Herbst, H. L. Masters, 2002, “Personality Pathways to Unsafe Sex: Personality, Condom Use, and Hiv Risk Behaviors”, *Journal of Research in Personality* 36 (2): 117-133.

## Consumption Culture, Cognitive Bias and Consumption Anomalies

Ye Dezhu<sup>a</sup>, Lian Yujun<sup>b</sup>, Ng Yew-Kwang<sup>c</sup>

(a: Jinan University; b: Sun Yat-Sen University; c: Monash University)

**Abstract:** It is widely acknowledged that national cultures have effects on consumption behavior. But there is little evidence to support this argument. This paper loose the hypothesis of ‘rational agent’, expresses the culture with cognitive bias of self control in the frame of behavioral hyperbolic discounting model, and explains the mechanism of insufficient consumption(in Europe and America) and excessive consumption(in East Asia). We regress between culture and consumption with a panel data covering 48 countries over year 1978 to 2007. The results show that traditional explanatory variables such as precautionary saving are less powerful than the unobservable country individual effects in explaining consumption rate difference, and Confucianism dummy variable and sex indices which proxy culture can explain 28% and 58% of those unobservable country individual effects. This indicates that culture which unchangeable over time is stronger than traditional variables in explaining consumption rate difference across countries. In practice, consumption commitment technology originated from the hyperbolic discounting model, can effectively correct consumption bias due to the cognitive bias induced by consumption culture, hence can make the intervention policy more effectively which is aimed at increasing internal demand in China.

**Key Words:** Consumption Culture; Self-Control; Hyperbolic Discounting; Increasing Internal Demand



# 中国上市公司资本结构动态调整机制研究

连玉君 钟经樊\*

**内容摘要** 本文从动态角度对中国上市公司资本结构的调整行为进行了研究。结果表明,我国上市公司存在最优资本结构,整体上表现为负债不足,由于调整成本的存在使得公司在偏离最优水平后只能进行部分调整。调整速度会受到公司规模、成长性和偏离最优水平的程度等因素影响,而且会因时间、行业和公司规模的不同而存在显著差异。

**关键词** 资本结构 动态调整 调整成本 优化比率

**JEL分类:**G32,G38 **中图分类号:**F830.91 **文献标识码:**A **文章编号:**1000-6249(2007)01-0023-016

## 一 引言

资本结构理论是财务经济学的一个重要课题。Modigliani 和 Miller (1958)在一系列严格假设条件下提出了资本结构与公司价值无关的 MM 定理。随后的研究通过从不同角度放松 MM 定理的严格假设提出了静态权衡理论,认为债务融资的利弊相互权衡可以决定出使公司价值极大的最优资本结构。这其中包括:负债的税收利益与破产成本之间的权衡 (Modigliani 和 Miller, 1963); 负债和权益融资产生的各种代理成本之间的权衡 (如, Jensen 和 Meckling, 1976; Hart 和 Moore, 1995); 以及与资本结构所发挥的信号功能相关的成本和收益之间的权衡 (Ross, 1977)。但影响最优资本结构的各种因素往往是不断变化的, 因此从动态角度的研究表明, 最优资本结构也是随时间变化的, 如 Bradley、Jarrell 和 Kim (1984), Goldstein 等 (2001), Dangl 和 Zechner (2004)。同时, 实际资本结构往往会偏离最优资本结构, 且调整过程往往比较缓慢, 如 Jalilvand 和 Harris (1984), Hovakimian 等 (2001), Lööf (2004)。调整成本的存在是导致实际负债率长期偏离最优值一个主要因素, 这使得公司的资本结构变化遵循一个部分调整的过程。Myers (1984) 认为相似的公司之所以会在资本结构的选择上表现出显著的差异, 很可能是因为调整成本的存在。Fischer 等 (1989)、Leland (1994, 1998) 将调整成本纳入动态资本结构模型中, 发现即使很小的调整成本也会使公司的负债率与最优水平发生较大的偏离, 但就长期来看, 公司会间断地向最优水平调整其资本结构以平衡负债融资的利弊。

既然公司的最优资本结构会随着各种因素的变化而变化, 而调整成本的存在又导致公司的实际资本结构向最优值调整的过程中存在时滞, 那么对于处于转型阶段的中国上市公司而言, 是否存在目标资本

\* 连玉君: 西安交通大学金禾经济研究中心 西安 710049 电子邮箱: arlion@stu.xjtu.edu.cn; 钟经樊: 台湾中央研究院经济研究所。

本文受西安交通大学人文社会科学基金 (2400-573001) 资助, 特此致谢。作者感谢两位匿名审稿人提出的宝贵意见和金禾经济研究中心博士生程建和朱晓明在论文写作过程中给予的帮助。当然, 文责自负。

结构? 资本结构的调整是否也呈现一个部分调整的过程?

在这方面, 针对中国上市公司的研究主要集中在静态分析上, 如陆正飞和辛宇 (1998)、冯根福等 (2000)、黄晓莉 (2002)、Chen (2004)、洪正 (2005)。<sup>①</sup>这些研究对影响我国上市公司资本结构的各种因素进行了较为全面的分析, 但存在以下两方面的局限: 其一, 由于最优资本结构不可观测, 因此以上研究都采用实际观测值作为替代指标。但实际负债率往往与最优值有较大的偏差, 使用这种替代方法可能会导致严重的偏误 (Fischer 等, 1989)。其二, 采用静态模型进行分析, 无法捕捉资本结构的动态调整特征。肖作平 (2004) 首先注意到了这个问题, 他采用了一个部分调整模型对上市公司最优资本结构的影响因素及调整成本进行了分析, 认为与发达国家相比, 中国上市公司的调整成本较低。<sup>②</sup>该文的主要局限在于假设调整成本不随时间和公司而变化。考虑到不同的公司在收益能力、外部融资能力等方面的差异, 这样的假设显然有些过于严格了。同时, 调整成本在很大程度上决定于资本市场的发展程度, 对于市场经济体制还不完善, 金融体系还没有完全建立的中国而言, 我们认为中国上市公司只可能面临较高的调整成本。

因此, 本文通过放松肖作平 (2004) 文中调整成本固定不变的假设, 将其设定为一个受公司特征因素影响的内生变量, 允许其随公司和时间发生变化。相比于静态模型, 动态模型有以下两个方面的优点: 其一, 通过将最优负债率和调整速度内生化的, 我们可以研究上市公司资本结构的动态调整机制, 并分析影响公司最优资本结构 (而非实际资本结构) 的因素。其二, 利用动态模型, 我们可以分析调整的快慢和影响因素, 从而间接地考察调整成本问题。

我们的研究表明, 相对于静态模型, 动态调整模型可以更好地解释我国上市公司的资本结构决定行为, 表明上市公司的资本结构是动态调整的。整体而言, 我国上市公司的实际负债率普遍低于最优水平, 而调整成本的存在使得公司向最优水平的调整过程比较缓慢。不同于肖作平 (2004) 的估计结果 0.8, 我们估计出的调整速度均值为 0.311, 表明上市公司面临较高的调整成本。同时, 调整速度存在明显的时间、行业和公司规模差异, 因此放松调整速度固定不变的假设是非常必要的。

文章的结构安排如下: 第二部分介绍资本结构动态调整模型, 第三部分为代理变量的选择, 第四部分介绍样本选择和估计方法, 第五部分为实证结果和分析, 第六部分做出总结。

## 二 资本结构动态调整模型

权衡理论认为公司的最优资本结构是在负债带来的好处与成本之间权衡的结果。比如负债的免税效应与破产成本之间的权衡; 负债使得公司的自由现金流量减少从而减少的代理成本和因此而导致的投资不足之间的权衡。因此, 当外界条件 (如宏观经济、税收制度、市场结构等) 发生改变时, 影响公司资本结构的因素也会随之改变, 进而引起最优资本结构的变化。从这个角度来看, 公司的最优负债率应该是随时间变化而不断调整的。然而, 由于资本市场的不完善会导致公司的融资行为受到多种因素的限制, 致使公司在偏离最优资本结构时只能做出部分调整, 而调整的程度和快慢则取决于调整成本的大小。因此, 我们可以用如下部分调整模型来描述公司资本结构的动态调整过程:

$$TL_{it} - TL_{it-1} = \delta_{it} (TL_{it}^* - TL_{it-1}) \quad (1)$$

<sup>①</sup> 李善民和刘智 (2003) 对这方面的文献作了较为详细的评述。

<sup>②</sup> 肖作平 (2004) 估计出的调整系数为 0.8, Jalilvand 和 Harris (1984), Ozkan (2001) 针对美国 and 英国上市公司的研究估计出的调整系数分别为 0.617 和 0.705。

其中,  $TL_{it}^*$  和  $TL_{it}$  分别表示公司  $i$  在第  $t$  年的最优资本结构和实际资本结构。 $\delta_{it}$  为调整系数, 表示在一个年度内公司的资本结构向最优水平调整的快慢, 可以间接反映调整成本的大小。若  $\delta_{it}=1$ , 则表明公司可以在一个期间内完成全部调整, 即不存在调整成本, 那么公司在第  $t$  年的资本结构处于最优水平上; 若  $\delta_{it}=0$ , 则表明调整成本大于经由调整而获得的收益, 以至于公司不做任何调整, 其第  $t$  年的资本结构仍然保持在前一年的水平上。如果  $0 < \delta_{it} < 1$ , 则说明在存在调整成本的情况下, 公司只进行了部分调整。

虽然最优资本结构无法直接观测, 但上述分析表明我们可以将最优资本结构设定为一组能够反映负债融资的成本和收益并最终通过相互抵换决定出最优资本结构的变量的函数(如 Fischer 等, 1989),<sup>①</sup> 即:

$$TL_{it}^* = F(Y_{it}, D_i, D_t) \quad (2)$$

其中  $Y_{it}$  是影响公司最优资本结构的一组变量,  $D_i$  和  $D_t$  分别为行业和时间虚拟变量, 用于反映行业和宏观经济要素的影响。前面已经提到, 调整速度也是随时间变化和公司的不同而有所差异的, 因此, 我们将调整系数设定为:

$$\delta_{it} = G(Z_{it}, D_i, D_t) \quad (3)$$

其中,  $Z_{it}$  为一组影响调整速度的变量。(1)式可以变形为:

$$TL_{it} = \delta_{it} TL_{it}^* + (1 - \delta_{it}) TL_{it-1} \quad (4)$$

为了能够进行实证检验, 必须设定(2)式的具体形式。这里我们沿用 Nivorozhkin (2004) 的设定方法, 将最优资本结构设定为如下线性函数形式:<sup>②</sup>

$$TL_{it}^* = \alpha_0 + \sum_j \alpha_j Y_{jit} + \sum_s \alpha_s D_s + \sum_t \alpha_t D_t \quad (5)$$

而对于调整速度, 我们也采用类似的处理方法:

$$\delta_{it} = \beta_0 + \sum_k \beta_k Z_{kit} + \sum_s \beta_s D_s + \sum_t \beta_t D_t \quad (6)$$

最终, 我们可以用由(4)-(6)式构成的模型来描述上市公司资本结构的动态调整行为。

### 三 代理变量的选择

#### (一) 资本结构的度量

用于衡量公司资本结构的指标主要有市值负债率和账面负债率两种。前面已经提到, 最优资本结构是公司在负债的节税效应和相关成本之间权衡的结果, 因此公司在举债后, 负债市场价值的改变并不会直接影响公司利用税盾效应而得到的收益。而且, 当公司面临破产时, 债权人的债务是按照负债的账面价值而不是市场价值来衡量的, 因为此时公司的价值更接近其账面价值。另一方面, 公司的市场价值有时波动幅度很大, 使得我们无论在实证分析还是在具体的公司管理中都难以应用市场价值(Jalivand 和

<sup>①</sup> 另外两种常用的最优资本结构的代理指标分别为行业均值(如 Bowen 等, 1982)和公司自身负债率的移动平均(如 Jalivand 和 Harris, 1984; Shyam-Sunder 和 Myers, 1999)。相对而言, 本文的处理方式更为符合权衡理论的基本思想。

<sup>②</sup> Banerjee 等(2004)和 Hovakimian 等(2001)在研究最优资本结构时也都采用相似的处理方法。其基本思想在于, 权衡理论认为负债的成本和利益相互权衡可以决定出公司的最优资本结构, 因此我们只要恰当地选择能够反映权衡的正负面影响的变量, 那么就可以近似地拟合出最优负债率。Hovakimian 等(2001)发现, 通过这种方式来拟合最优资本结构对代理变量的选取有较强的稳健性。

Harris, 1984)。因此本文采用账面负债率作为公司资本结构的代理变量。

## (二)最优资本结构的拟合变量

1. 公司规模。大规模公司一般具有较强的风险分散能力,因此破产的风险也相对较低。对于大公司而言,破产的固定成本占其总资产的比例很小,这使得其负债的成本相对较低(Titman 和 Wessels, 1988)。另一方面,大公司能够更好地做到信息的公开化,从而有效降低信息的不对称程度,所以大公司更容易得到银行的贷款。对于处于转型阶段的发展中国家而言,大公司还往往被政府赋予重要的社会责任。在“国家信用”作为担保的情况下,债权人往往对大公司有更强的信心。因此,公司规模应该与负债率正相关。

2. 资产结构。当公司面临破产时,相对于很快就消失掉的无形资产,有形资产更容易变现,从而降低了破产成本。同时,有形资产的担保能在一定程度上降低债务的代理成本。从这两个角度来讲,有形资产的比例应与负债率正相关。但在研究中国上市公司的资本结构时,我们还需考虑由于市场发育不完善而产生的影响。首先,法律制度的不完善有可能导致在违约发生的情况下,债权人对抵押品的追索成本相当高。其次,由于次级市场的规模很小,资产的流动性较差,致使公司资产的抵押价值存在很大的不确定性。在这种情况下,二者的关系可能不显著甚至负相关。

3. 成长性。根据权衡理论,对于成长速度较快的公司而言,负债的代理成本也相对较高。因为股东此时具有更大的投资灵活性,他们可以通过投资在次优项目上来从银行或债权人的手中攫取财富(Titman 和 Wessels, 1988)。这表明公司的成长性应该与负债率负相关。另一方面,由于高成长性的公司多数属于新兴行业,经营上具有较大的风险,所以难以获得足够的长期贷款。为了弥补其大量的资金需求,短期贷款成为这些公司的主要选择。那么总负债率和成长性之间的关系也可能不显著或正相关。

4. 非负债类税盾。根据权衡理论,较高的非负债类税盾(如折旧)会部分抵消负债带来的税盾效应。所以在其他条件相同的情况下,拥有较多非负债类税盾的公司会更少的使用债务,即二者负相关。Bradley 等(1984)研究发现非负债类税盾与负债率正相关。Titman 和 Wessels(1988)的研究发现二者关系不显著。

5. 盈利能力。根据权衡理论,盈利能力强的公司会提高负债率从而更好的利用税盾效应。从代理成本理论来看,外部股东也会强制经理人提高负债率以减少公司的自由现金流量从而降低代理成本。据此,盈利能力应该与负债率正相关。而优序融资理论则依据内部人和外部人之间的信息不对称,认为公司会优先使用内部资金,所以二者负相关。

6. 资产流动性。资产流动性对公司负债率的影响既有正面的也有负面的。一方面,资产流动性高的公司支付短期债务的能力较强,因此应有较高的负债率。但另一方面,具有较多流动资产的公司也许会用其为投资融资,那么资产流动性就会对负债率负相关。

7. 公司的成熟度。成立时间较长的公司往往具有相对详尽的经营记录和较高的品牌价值,因此会具有较高的负债率。但是,Scholes 和 Wolfson(1989)则认为公司进行资本结构调整的成本会随着公司成立时间的增加而增加,即二者负相关,因为其决策会受到较多的约束。

8. 股权流通性。由于中国上市公司存在特殊的股权割裂现象,因此股权流通性也就成为一个重要的影响因素。Jensen 和 Meckling(1976)认为,由于股权代理成本和债券代理成本的存在,企业现金流量的概率分布并不独立于它的所有权结构,企业资本结构的不同会引起股权代理成本和债权代理成本之间的转移。我国上市公司的股权结构有两个特殊之处,一是存在着独特的非流通股股权结构安排,二是



在非流通股中,国家股和法人股占有重要地位。非流通股的存在造成了同股不同权、同股不同利。由于非流通股股东具有信息优势,因此上市公司的配股过程中,非流通股股东放弃配股权的现象普遍存在,而流通股股东则往往积极配合。因此,对于流通股比例较高的公司而言,其通过股票市场获得的资金会较多,负债率则相对较低。另一方面,由于流通股集中度非常低,流通股股东很少干预企业的经营活动,所以上市公司的控制权一般为国家股和法人股所掌握。这会产生预算“软约束”问题,使得国家控股的上市公司很可能像国有企业那样拥有较高的负债率。鉴于以上分析,我们认为股权流通性应与负债率负相关。

9. 宏观经济状况和行业因素。考虑到中国处于经济发展的转型阶段,而股市的发展时间也较短,我们在模型中加入时间虚拟变量来控制经济结构调整和政策变化对公司资本结构的影响。前期的许多研究都发现行业是影响公司资本结构的重要因素(如,陆正飞和辛宇,1998;郭鹏飞和孙培源,2003),因此我们加入行业虚拟变量来控制行业差异的影响。

### (三)调整速度的影响因素

公司向目标资本结构调整的速度主要取决于调整成本的大小,主要包括固定成本和制度成本。前者主要指进行调整所需的会计费用、律师费用、资产评估费用等成本;后者则视公司的经营绩效和资本市场的发展状况而定。对于不同的公司而言,进行调整的固定成本的绝对数量差别不大,而其相对大小则会因公司规模、盈利能力的差异而有所区别。制度成本主要归因于资本市场不完善或公司治理效率低下等因素,这会导致公司无法及时获得融资资金或融资环节过于复杂,从而使融资的机会成本增加。我们认为,对中国上市公司而言,制度成本是资本结构调整成本的主要组成部分。我们选取以下几个对调整成本有显著影响的变量来拟合由(6)式决定的调整速度模型。

1. 偏离最优负债率的程度。在资本市场较为完善的情况下,调整成本中固定成本占主要比重,那么公司只有在目前的资本结构与最优水平之间有足够的大偏差时,才会对其进行调整,调整速度应当和偏离最优水平的程度正相关。但对于中国而言,由于制度成本占很大的比重,二者可能负相关。这是因为,当公司的资本结构偏离最优水平的程度不大时,可以利用内源融资实现调整,但当偏离的程度较大时,就只能借助外部资本市场融资来实现调整,资本市场的完善程度决定了调整成本的大小和调整周期的长短。考虑到我们目前资本市场的发展状况,我们认为对中国上市公司而言,公司的负债率偏离最优值的程度可能和调整速度负相关。

2. 公司规模。公司规模对调整速度可能会产生正反两个方面的影响。一方面,大公司进行调整所需的资金规模往往较大,这使得融资的固定成本所占的比例较小(Jalilvand 和 Harris, 1984);同时,中国多数大规模上市公司往往都属于垄断性行业,收益稳定,借助国家信用的支持他们比小规模公司更容易获得银行贷款。这使得公司规模和调整速度正相关。另一方面,由于大规模公司进行调整所需的资金量往往较大,往往需要借助资本市场进行外部融资,而公司的融资决策也往往要经过多方利益的权衡,即调整成本中可变成本的比例较高,从而导致其调整速度与公司规模之间的关系不显著甚至负相关。

3. 成长性。成长性强的公司可以通过改变其新的融资来源来实现资本结构的快速调整。而对于非成长性公司而言,则只能通过发行新股来偿还银行贷款或进行相反的操作来调整其资本结构。在信息不对称的情况下,以上两种操作都会给市场传递负面信号,从而降低公司的市场价值。因此我们预期调整速度和成长性正相关。

## 四 样本选取和估计方法

### (一)样本的选取和指标定义

本文数据取自仅发行 A 股的上市公司 1998-2003 年报。我们遵循以下原则对样本进行了筛选:(1)不考虑金融类上市公司;(2)在 1998-2003 年连续 6 年可以获得相关数据的公司;(3)剔除 1998-2003 年内被 ST 和 PT 的上市公司;(4)为防止兼并或重组的影响,剔除样本区间内总资产成长率或主营业务收入成长率大于 100%的公司;(5)剔除负债率大于 100%及净利润率大于 100%或小于-100%的含有奇异值的公司。基于以上原则,本文选取了 1998 年 1 月 1 日以前上市的 427 家公司作为最终的研究对象。

表 1 中列示了文中第三部分所涉及变量的定义方法和基本统计量。在计算 Tobin's Q 时需要使用公司的市场价值,考虑到中国特殊的股权结构,我们采用冯根福等(2000)所建议的计算方法。具体而言,公司的市场价值为总负债的帐面价值与股票的市场价值之和。流通股的市价为流通股年平均股价与流通股股本数之积,而非流通股市价为其股本数与每股净资产之积。对于公司成熟度,文献中通常采用公司的成立年数((观察年份(公司成立时间)加以衡量,但依此定义出的成熟度很可能含有时间趋势。<sup>①</sup>为此,我们采用序别化变量来定义公司的成熟度 AGE。具体而言,以公司成立年份的第 33 和 66 分位值为分界点将样本公司分成三组,进而将成立时间小于第 33 分位值的公司定义为“成熟公司”,其 AGE 变量取值为 3;将成立时间大于第 66 分位值的公司定义为“年轻公司”,其 AGE 变量的取值为 1;而其它公司的则被定义为“中等公司”,其 AGE 变量取值为 2。这种定义方法一方面避免了前期文献的缺陷,同时也能够克服离群值的影响。

表 1 样本描述性统计量(1998-2003, N=427 家, T=6 年, NT=2562)

| 变量名称      | 变量含义     | 计算方法                      | 平均值   | 标准差  | 最小值   | 最大值   |
|-----------|----------|---------------------------|-------|------|-------|-------|
| TL        | 资本结构     | 总负债/总资产                   | 0.42  | 0.16 | 0.01  | 0.96  |
| FR        | 资产流动性    | 流动资产/流动负债                 | 1.83  | 1.80 | 0.15  | 41.97 |
| SIZE      | 公司规模     | 总资产的自然对数                  | 21.00 | 0.85 | 18.59 | 24.33 |
| NDTS      | 非负债类税盾   | 累计折旧/总资产                  | 0.13  | 0.11 | 0.00  | 0.91  |
| TANG      | 资产结构     | (固定资产+存货)/总资产             | 0.50  | 0.17 | 0.02  | 0.94  |
| NPR       | 盈利能力     | 净利润/主营业务收入                | 0.09  | 0.14 | -0.93 | 0.96  |
| TSHR      | 股权流通性    | 流通股数/总股本数                 | 0.38  | 0.13 | 0.05  | 1.00  |
| AGE       | 成熟度      | 1 年轻;2 中等;3 成熟            | 2.08  | 0.79 | 1     | 3     |
| Tobin's Q | 成长性      | 公司市值/公司帐面值                | 1.66  | 0.56 | 0.87  | 5.88  |
| DIST      | 偏离最优值的程度 | $ TL_{it}^* - TL_{it-1} $ | -     | -    | -     | -     |

表 2 中列示了样本公司的行业分布,这里我们采用中国证监会 2001 年 4 月发布的《上市公司行业分类指引》按行业门类对样本公司进行行业划分。在定义虚拟变量时,为了避免由于部分行业内公司数目过少而造成统计检验量的偏误,我们对公司数目小于 10 家的行业进行了合并,最终得到 8 个行业门类,因此定义了表 2 中列示的 7 个行业虚拟变量,以综合类为对比基础。

① 我们感谢匿名审稿人对此的提示。

表 2

样本行业分布及行业虚拟变量的定义

| 行业门类名称(代码)        | 公司数目(家) | 百分比(%) | 行业虚拟变量 | 合并方法 |
|-------------------|---------|--------|--------|------|
| 农、林、牧、渔业(A)       | 6       | 1.41   | -      | -    |
| 采掘业(B)            | 2       | 0.47   | -      | -    |
| 制造业(C)            | 244     | 57.14  | SIC1   | C+B  |
| 电力、煤气及水的生产和供应业(D) | 19      | 4.45   | SIC2   | D    |
| 建筑业(E)            | 7       | 1.64   | -      | -    |
| 交通运输、仓储业(F)       | 14      | 3.28   | SIC3   | F    |
| 信息技术业(G)          | 22      | 5.15   | SIC4   | G    |
| 批发和零售贸易(H)        | 50      | 11.71  | SIC5   | H    |
| 房地产业(J)           | 12      | 2.81   | SIC6   | J+E  |
| 社会服务业(K)          | 12      | 2.81   | SIC7   | K+L  |
| 传播与文化产业(L)        | 3       | 0.70   | -      | -    |
| 综合类(M)            | 36      | 8.43   | -      | M+A  |
| 合计                | 427     | 100    |        |      |

## (二)估计方法

我们用于实证分析的模型设定为:

$$TL_{it} = \delta_{it} TL_{it}^* + (1 - \delta_{it}) TL_{it-1} + \varepsilon_{it} \quad (7)$$

其中,  $\varepsilon_{it}$  为随机干扰项, 我们假设其服从均值为零、方差有限的正态分布,  $TL_{it}^*$  和  $\delta_{it}$  分别由(5)式和(6)式确定。由于整个模型是非线性的, 我们采用非线性最小二乘法(Nonlinear OLS)进行估计, 迭代方式为高斯—牛顿法。迭代前必须给出参数的初始值, 步骤如下: 第一步, 用做被解释变量, 估计如下静态模型:

$$TL_{it} = \alpha_0 + \sum_j \alpha_j Y_{jit} + \sum_s \alpha_s D_s + \sum_t \alpha_t D_t + \varepsilon_{it} \quad (8)$$

由此得到的参数估计值作为(5)式中相应参数的初始值, 同时我们还将得到的线性拟合值, 记为  $TL_{it}$ ; 第二步  $TL_{it}$ , 把作为最优负债率  $TL_{it}^*$  的初始值代入(1)式, 计算得到一组  $\delta_{it}$  的初始值, 即  $\delta_{it} = \Delta TL_{it} / \Delta TL_{it}^*$ , 其中,  $\Delta TL_{it} = TL_{it} - TL_{it-1}$ ,  $\Delta TL_{it}^* = TL_{it}^* - TL_{it-1}^*$ 。<sup>①</sup>第三步, 利用从第二步中得到的  $\delta_{it}$  的初始值估计(6)式, 得到其中参数估计的初始值。

作为对比, 我们还估计了由(8)式确定的静态模型以及由(9)式确定的准动态模型:

$$TL_{it} = \delta_0 TL_{it}^* + (1 - \delta_0) TL_{it-1} + \varepsilon_{it} \quad (9)$$

显然, 静态模型和准静态模型都是本文设定的动态模型的特例。若假设公司的资本结构始终都处于最优水平上, 即调整系数  $\delta_{it} = 1$ , 动态模型便转化为静态模型, 这也是目前多数文献在分析中国上市公司资本结构影响因素时所采用的模型。若假设调整系数不随时间和公司的特征而变化, 即  $\delta_{it} = \delta_0 = \text{Constant}$ , 便对应着肖作平(2004)设定的准动态模型。当然, 至于哪一种模型能够更好地描述中国上市公司的资本结构调整行为, 我们还需在后续的分析中进行假设检验。

本文的数据处理和模型估计均采用 STATA9.1 软件包完成。

<sup>①</sup> 这是因为在估计动态模型之前, 最优负债率  $TL_{it}^*$  和调整速度  $\delta_{it}$  都是不可观测的。

## 五 实证结果及分析

### (一)模型识别检验

相对于静态模型和准动态模型(肖作平,2004),本文模型的主要差别在于放松了调整系数固定不变的假设。下面,我们通过两个途径来检验本文模型的设定是否合理:其一,检验调整系数固定不变的假设是否成立;其二,分析经由估计动态模型得到的残差是否满足正态分布和不存在序列相关的假设。

为检验调整系数固定不变的假设,我们针对原假设  $H_0: \delta_{it} = \text{Constant}$  进行了检验,相应 F 统计量为:<sup>①</sup>

$$F(J, n-K) = \frac{[S(b^*) - S(b)]/J}{S(b)/(n-K)} \quad (10)$$

其中,  $S(b^*)$  和  $S(b)$  分别为准动态模型和动态模型设定下估计得到的残差平方和,分别为 11.4 和 10.2,  $J$  为约束的个数,  $K$  为动态模型中参数的个数,  $n$  为有效样本数。根据(10)式计算得到  $F(14, 2100) = 17.26$ , 明显大于 1% 显著水平下的临界值 2.22, 无法接受调整系数为常数  $\delta_{it}$  的原假设, 这同时也就拒绝了  $\delta_{it} = 1$ , 即静态模型的基本假设。

表 3 列示了动态模型残差的基本统计量、正态分布检验和一阶序列相关检验的结果。从 Panel A 列示的结果来看, 残差的均值为 -0.0004, 偏度和峰度分别为 -0.066 和 3.276, 非常接近标准正态分布。Panel B 中参考 D'Agostino 等(1990)介绍的方法检验表明, 无论是针对偏度、峰度还是联合检验都无法拒绝残差服从正态分布的原假设。Panel C 中的结果表明, 残差并不存在一阶序列自相关。这些结果一方面表明本文的模型设定不存在严重偏误, 同时也说明在假设残差服从正态分布的前提下采用非线性最小二乘法估计动态模型并进行相应的统计推断是合理的。

表 3 动态调整模型的残差分析

| Panel A: 基本统计量                 |        |             |        |                    |             |       |
|--------------------------------|--------|-------------|--------|--------------------|-------------|-------|
| 均值                             | 标准差    | 最小值         | 中位数    | 最大值                | 偏度          | 峰度    |
| -0.0004                        | 0.0643 | -0.2324     | 0.0001 | 0.2097             | -0.066      | 3.276 |
| Panel B: 正态分布性检验 <sup>a</sup>  |        |             |        |                    |             |       |
| 偏度检验                           |        | 峰度检验        |        | 联合检验               |             |       |
| p 值 = 0.573                    |        | p 值 = 0.221 |        | $\chi^2(2) = 1.82$ | p 值 = 0.402 |       |
| Panel C: 一阶序列相关检验 <sup>b</sup> |        |             |        |                    |             |       |
| F (1, 426) = 0.007             |        | p 值 = 0.936 |        |                    |             |       |

说明: a、参见 D'Agostino 等(1990); b、参见 Wooldridge(2002, 第 7 章)。

### (二)最优资本结构

静态模型、准动态模型和动态模型的对比结果见表 4。相对于静态模型, 两个动态模型有以下两个显著的变化: 其一, 动态模型的解释能力明显高于静态模型, 表明上市公司的资本结构的确是围绕最优水平进行动态调整的, 调整成本对资本结构的调整行为有显著的影响。其二, 在静态模型中资产结构与资

<sup>①</sup> 参见 Greene(2000, pp.438-439)。



本结构正相关,但不显著,而在两个动态模型中二者均显著负相关。这与我们前面的分析一致,说明法律制度的不完善和次级市场规模较小导致了公司固定资产抵押价值的大幅降低。这与多数针对发展中国家的实证结果是一致的。对比准动态模型和动态模型,我们发现主要变量的符号和显著水平都基本一致,二者的主要区别在于,后者的拟合程度明显高于前者,分别为 0.8206 和 0.6994,表明放松调整系数固定不变假设下的动态模型具有更强的解释能力。更为重要的是,我们发现,经由准动态模型估计出的调整系数为  $\hat{\delta}_0 = 0.7256$ ,非常接近于肖作平(2004)的估计结果 0.8,而通过动态模型估计出的平均调整系数为 0.311(见表 5)。考虑到本文与肖作平(2004)在样本筛选和变量的选择上的差异,我们可以认为在调整系数固定不变假设下得到的估计值可能过度偏高。

就最优资本结构的影响因素而言,与最优资本结构显著负相关的因素有:资产流动性 FR、非负债类税盾 NDTS、资产结构 TANG、盈利能力 NPR、股权流通性 TSHR 和成长性 Tobin's Q;正相关的因素有:公司规模 SIZE 和公司的成熟度 AGE。

资产流动性与资本结构负相关表明短期偿债能力强的公司并没有利用该优势增加短期负债,而是更多地再进行再投资。成长性与资本结构的负相关与权衡理论是一致的,表明高成长性的公司负债的代理成本也相对较高。这是因为中国上市公司中高成长性的公司多属于风险较高的新兴行业,而且规模都较小,因此难以获得银行贷款。盈利能力与资本结构负相关与权衡理论相矛盾。我们认为,虽然从理论上讲,中国上市公司的所得税相对于 GDP 而言较高,而破产成本很低,所以盈利性公司负债的税收优势非常明显,应当有较高的负债率。但是,中国上市公司的破产成本低是因为法律制度不健全,以至于公司对债权人的违约行为在多数情况下都不会导致破产。在这种情况下,债权人(如银行)的财务危机成本增加了。在利率无法通过市场来调节的情况下,银行就会通过限制放贷数量来达到降低风险的目的。这种状况表现为银行的普遍“惜贷”行为,因此盈利性公司的负债率反而较低。股权流通性、公司规模及成熟度与资本结构的关系与我们的理论预期相一致,这里不再赘述。

同时,我们发现时间虚拟变量和行业虚拟变量对公司的最优资本结构都有显著的影响。前者表明,对处于转型阶段的中国而言,宏观经济状况的变化对公司的融资行为会产生重要的影响;后者表明不同的行业的最优资本结构也存在很大的差异。

### (三)调整速度

1. 影响因素。表 4 中 B 栏列示了调整速度影响因素的估计结果,基本上与我们的理论预期一致。公司的成长性和调整速度正相关,这与我们前面的理论预期是一致的,表明成长性较强的公司在资本结构调整方面有更大的灵活性。

偏离最优负债率的程度与调整速度负相关,表明当公司的负债率接近最优水平时,公司可以快速实现调整,而偏离最优水平的差距较大时反而表现出“调整惰性”。这间接说明我国上市公司的外部融资成本较高。公司对资本结构的调整主要通过三种手段:内源融资、权益融资和债务融资。对于多数内部资金有限的公司而言,通过内源融资可以在短时间内实现资本结构的小幅调整,但当公司的资本结构偏离最优水平较大时,这种方式即使可行也是非常缓慢的。而权益融资不仅要合乎证监会的发行条件,而其发

① 证监会规定:公司一次配股发行股份总数,不得超过该公司前一次发行并募足股份后其普通股股份总数的 30%,公司将本次配股募集资金用于国家重点建设项目和技改项目的,在发起人承诺足额认购其可配股份的情况下,可不受 30% 比例的限制。见《关于 1996 年上市公司配股工作的通知》证监发字[1996]17 号。

② Banerjee 等(2004)和 Nivorozhkin(2004)对美国 and 保加利亚的估计值分别为 0.529 和 0.434。

## 中国上市公司资本结构动态调整机制研究

表 4

静态模型、准动态模型和动态模型的回归结果对比

| 参数               | 静态模型       |        | 准动态模型      |        | 动态模型       |        |
|------------------|------------|--------|------------|--------|------------|--------|
|                  | 估计值        | 标准误    | 估计值        | 标准误    | 估计值        | 标准误    |
| A、最优负债率：         |            |        |            |        |            |        |
| $\alpha_0$       | -1.0469*** | 0.1135 | 0.2996***  | 0.0719 | 0.5912***  | 0.0241 |
| $\alpha_{FR}$    | -0.0251*** | 0.0012 | -0.0576*** | 0.0037 | -0.1125*** | 0.0045 |
| $\alpha_{SIZE}$  | 0.0766***  | 0.0052 | 0.0234***  | 0.0037 | 0.0119***  | 0.0013 |
| $\alpha_{NDTS}$  | -0.1922*** | 0.0323 | -0.5057*** | 0.0536 | -0.4458*** | 0.0397 |
| $\alpha_{TANG}$  | 0.0075     | 0.0164 | -0.0277*** | 0.0070 | -0.0412*** | 0.0090 |
| $\alpha_{NFR}$   | -0.1567*** | 0.0139 | -0.4936*** | 0.0475 | -0.3853*** | 0.0359 |
| $\alpha_{TSHR}$  | -0.2032*** | 0.028  | -0.0034*   | 0.0021 | -0.0555*** | 0.0113 |
| $\alpha_{AGE}$   | 0.0334***  | 0.0067 | 0.0176***  | 0.0033 | 0.0066***  | 0.0015 |
| $\alpha_{TOBIN}$ | -0.0150*** | 0.0045 | -0.0608*** | 0.0075 | -0.0415*** | 0.0042 |
| $\alpha_{1999}$  | 0.0038     | 0.0048 | -          | -      | -          | -      |
| $\alpha_{2000}$  | 0.0129**   | 0.0056 | 0.0166**   | 0.0066 | 0.0258***  | 0.0066 |
| $\alpha_{2001}$  | 0.0193***  | 0.0059 | 0.0428***  | 0.0107 | 0.0342***  | 0.0077 |
| $\alpha_{2002}$  | 0.0198***  | 0.0061 | 0.0087*    | 0.0046 | 0.0223***  | 0.0069 |
| $\alpha_{2003}$  | 0.0301***  | 0.0064 | 0.0356***  | 0.0095 | 0.0286***  | 0.0062 |
| $\alpha_{SIC1}$  | -0.0475*** | 0.0181 | -0.0431*** | 0.0082 | -0.0062*** | 0.0020 |
| $\alpha_{SIC2}$  | -0.1069*** | 0.0301 | -0.0083    | 0.0111 | -0.0228    | 0.0164 |
| $\alpha_{SIC3}$  | -0.0736**  | 0.0335 | 0.0037     | 0.0071 | -0.0021    | 0.0019 |
| $\alpha_{SIC4}$  | -0.0013    | 0.028  | 0.0010     | 0.0009 | 0.0213***  | 0.0075 |
| $\alpha_{SIC5}$  | -0.0194    | 0.0224 | -0.0933*** | 0.0186 | -0.0906*** | 0.0118 |
| $\alpha_{SIC6}$  | 0.0034     | 0.0296 | -0.0148*   | 0.0090 | 0.0178*    | 0.0097 |
| $\alpha_{SIC7}$  | -0.0697**  | 0.0322 | -0.0732    | 0.0699 | -0.0732    | 0.0720 |
| B、调整速度：          |            |        |            |        |            |        |
| $\delta_0$       | -          | -      | 0.7256***  | 0.0214 | -          | -      |
| $\beta_0$        | -          | -      | -          | -      | 0.8240***  | 0.1670 |
| $\beta_{SIZE}$   | -          | -      | -          | -      | -0.0254*** | 0.0080 |
| $\beta_{DISTA}$  | -          | -      | -          | -      | -0.0931*** | 0.0076 |
| $\beta_{TOBIN}$  | -          | -      | -          | -      | 0.0227***  | 0.0072 |
| $\beta_{2000}$   | -          | -      | -          | -      | 0.0047     | 0.0050 |
| $\beta_{2001}$   | -          | -      | -          | -      | -0.0275**  | 0.0136 |
| $\beta_{2002}$   | -          | -      | -          | -      | -0.0509*** | 0.0170 |
| $\beta_{2003}$   | -          | -      | -          | -      | 0.0331     | 0.0307 |
| $\beta_{SIC1}$   | -          | -      | -          | -      | 0.0043     | 0.0034 |
| $\beta_{SIC2}$   | -          | -      | -          | -      | -0.0024    | 0.0021 |
| $\beta_{SIC3}$   | -          | -      | -          | -      | -0.0763**  | 0.0321 |
| $\beta_{SIC4}$   | -          | -      | -          | -      | 0.0784*    | 0.0416 |
| $\beta_{SIC5}$   | -          | -      | -          | -      | 0.0776*    | 0.0450 |
| $\beta_{SIC6}$   | -          | -      | -          | -      | -0.0369    | 0.0238 |
| $\beta_{SIC7}$   | -          | -      | -          | -      | -0.0439    | 0.0398 |
| 调整后的 $R^2$       | 0.3328     |        | 0.6994     |        | 0.8206     |        |

说明：\*\*\*、\*\*和\*分别表示在1%、5%和10%水平上显著。

行数量也受到限制(张军等,2005),<sup>①</sup>使得融资实现的周期较长。至于债务融资,受限於发展严重滞后的公司债券市场,多数上市公司只能通过银行贷款获得资金,而这一资金来源很大程度上受制于银行的信贷计划,表现为优先满足大型国有企业的歧视性政策。简言之,我国资本市场的发育不足在很大程度上限制了上市公司进行外部融资的能力和灵活性。

公司规模与调整速度在 1%水平上显著负相关。这表明对于大规模公司而言,调整成本中制度成本占了更大的比重。我们认为有两个原因:一是小公司的资本结构调整往往可以伴随着公司的日常经营活动来完成,而大公司的调整由于所需的资金规模较大,需要更多地依赖外部融资来实现调整。另一方面,由于大规模公司多为国企转型而来,相对于规模较小的民营上市公司而言,其治理机制要复杂得多,因此在此在融资的过程中所涉及的利益群体也相对较多,致使调整的机会成本较高。

如果我们结合公司规模和偏离最优水平的程度这两个因素对调整速度的影响就会发现,在调整所需的资金规模不大的情况下(如小规模公司和偏离最优水平较少的公司),调整速度较快,反之则较慢。这表明,上市公司在条件允许的情况下会积极地向最优负债水平调整。但当调整所需的资金规模较大时(如大规模公司和偏离最优水平较多的公司),调整速度会明显放缓。

2. 时间、行业和公司规模差异。表 5 第 I 栏列示了调整速度按时间、行业和公司规模的分类统计结

表 5

调整速度和最优比率的分类统计描述

|                    | I 调整速度 |       | II 最优比率 |       |
|--------------------|--------|-------|---------|-------|
|                    | 均值     | 中位数   | 均值      | 中位数   |
| A:按年份              |        |       |         |       |
| 1999               | 0.332  | 0.330 | 1.086   | 1.059 |
| 2000               | 0.340  | 0.335 | 1.130   | 1.098 |
| 2001               | 0.301  | 0.297 | 1.190   | 1.128 |
| 2002               | 0.269  | 0.265 | 1.210   | 1.133 |
| 2003               | 0.312  | 0.307 | 1.187   | 1.102 |
| B:按行业门类            |        |       |         |       |
| 采掘业和制造业(B+C)       | 0.304  | 0.305 | 1.160   | 1.126 |
| 电力、煤气及水的生产和供应业(D)  | 0.249  | 0.249 | 1.186   | 1.118 |
| 交通运输、仓储业(F)        | 0.218  | 0.219 | 1.174   | 1.190 |
| 信息技术业(G)           | 0.391  | 0.385 | 1.137   | 1.062 |
| 批发和零售贸易(H)         | 0.385  | 0.387 | 1.100   | 1.035 |
| 建筑业和房地产业(E+J)      | 0.275  | 0.279 | 1.025   | 1.009 |
| 社会服务业和传播与文化产业(K+L) | 0.279  | 0.276 | 1.502   | 1.359 |
| 农、林、牧、渔业和综合类(A+M)  | 0.304  | 0.305 | 1.168   | 1.070 |
| C:按公司规模            |        |       |         |       |
| 小型                 | 0.350  | 0.350 | 1.149   | 1.103 |
| 较小型                | 0.333  | 0.332 | 1.166   | 1.103 |
| 中等型                | 0.312  | 0.312 | 1.145   | 1.097 |
| 较大型                | 0.295  | 0.296 | 1.190   | 1.125 |
| 大型                 | 0.264  | 0.263 | 1.154   | 1.068 |
| 样本总体               | 0.311  | 0.309 | 1.161   | 1.098 |

## 中国上市公司资本结构动态调整机制研究

果。调整速度的平均值为 0.311, 低于对发达国家和发展中国家的估计值, 表明我国上市公司面临较高的调整成本。

从时间上来看, 从 1999 年到 2002 年上市公司总体的调整速度呈现先增后减的趋势, 在 2000 年达到最大。我们认为政策因素对这一转变过程具有重要作用。1999 年 3 月, 证监会对配股的规定作了调整, 放松了净资产收益率标准, 使得一些公司可以通过股权融资来获取资金。同年 7 月, 《证券法》出台, 出于对监管机制不断严格的预期, 资本结构不合理的公司进行了积极的调整, 这使得我们观察到 2000 年的调整速度加快。但随后两年调整速度的减慢则进一步说明了上市公司在进行调整时面临着结构性的障碍。从公司规模来看, 小规模公司具有相对快的调整速度, 对此我们在前文已经作了解释。

从行业分类来看, 信息技术业和批发零售业的调整速度是所有行业中最快的, 分别为 0.391 和 0.385; 而交通运输业和电力、煤气及水的生产和供应业(以下简称电力行业)的调整速度则是最慢的, 分别为 0.218 和 0.249。这里似乎会产生一个疑问, 交通运输业和电力行业上市公司大多由国企转型而来, 这类公司与国有银行的特殊关系使其相对于其它公司更容易获得银行贷款, 而其稳定的高收益又使其容易达到增配的条件, 所以整体看来这类公司的外部融资受到的限制最小, 应当有较快的调整速度。为此, 我们进一步比较分析了这两类公司的债务期限结构和收益能力, 结果见表 6。易于看出, 信息技术业和批发零售业公司的平均净收益率均较低, 分别为 0.093 和 0.036, 而总负债中长期负债所占的比重 LLR 分别为 8% 和 6.5%, 也是所有行业中最底的。相对而言, 电力行业和交通运输业公司则具有较强的盈利能力(净利润率分别为 0.18 和 0.179)和较多的长期负债(长期负债占总负债的比重分别为 27.8% 和 26.1%)。可见, 电力行业和交通运输业公司倾向于更多地使用股权融资和长期债务融资,<sup>①</sup>而这两种融资方式的实现时间通常较长, 从而使其调整速度较慢。依据相同的逻辑我们也就不难解释信息技术业和批发零售业调整速度较快这一现象了。

表 6

不同行业的财务特征和竞争强度

| 行业门类名称(代码)         | 调整速度  | 最优比率  | TL    | LLR   | NPR   | HHI <sub>a</sub> | HHI <sub>s</sub> |
|--------------------|-------|-------|-------|-------|-------|------------------|------------------|
| 采掘业和制造业(B+C)       | 0.304 | 1.160 | 0.417 | 0.131 | 0.069 | 0.002            | 0.003            |
| 电力、煤气及水的生产和供应业(D)  | 0.249 | 1.186 | 0.354 | 0.278 | 0.180 | 0.026            | 0.027            |
| 交通运输、仓储业(F)        | 0.218 | 1.174 | 0.364 | 0.261 | 0.179 | 0.025            | 0.083            |
| 信息技术业(G)           | 0.391 | 1.137 | 0.459 | 0.080 | 0.093 | 0.020            | 0.032            |
| 批发和零售贸易(H)         | 0.385 | 1.100 | 0.471 | 0.065 | 0.036 | 0.012            | 0.006            |
| 建筑业和房地产业(E+J)      | 0.275 | 1.025 | 0.522 | 0.128 | 0.120 | 0.018            | 0.023            |
| 社会服务业和传播与文化产业(K+L) | 0.279 | 1.502 | 0.350 | 0.116 | 0.123 | 0.016            | 0.034            |
| 农、林、牧、渔业和综合类(A+M)  | 0.304 | 1.168 | 0.478 | 0.115 | 0.102 | 0.007            | 0.011            |
| 合计                 | 0.311 | 1.161 | 0.429 | 0.129 | 0.082 | 0.008            | 0.012            |

说明: LLR = 长期负债合计/总负债; HHI<sub>a</sub> 和 HHI<sub>b</sub> 分别基于所有上市公司和本文样本公司计算出的赫芬因德指数, 用于反映市场竞争强度, 具体计算方法参见刘志彪等(2003, pp.64)。

为了验证调整速度在时间、行业和公司规模上的差异是否具有统计上的显著性, 我们依照表 5 的分

① 受限于数据, 我们未能对各行业的配股和增发情况进行统计分析, 不过阎达五等(2001)研究表明, 多数达到增配标准(以盈利能力为基本标准)的上市公司都提出了增配要求, 这使得我们可以间接推断盈利能力强的电力行业和交通运输业公司更倾向于使用股权融资方式。

组方式进行了 Kruskal-Wallis H 差异性检验,结果见表 7。显然,无论采取那种分组方式,调整速度都在 1%以上的水平上存在显著差异,这表明处于不同行业,规模不同的公司所面临的调整成本不同,同时也说明本文的模型设定中放松调整速度不随时间和公司变化这一假设是非常必要的。

表 7 调整速度行业和公司规模差异的 Kruskal-Wallis H 检验

| 年份   | 1999      | 2000      | 2001      | 2002      | 2003      | 自由度 |
|------|-----------|-----------|-----------|-----------|-----------|-----|
| 行业差异 | 230.03*** | 222.92*** | 218.79*** | 232.24*** | 238.64*** | 7   |
| 规模差异 | 166.16*** | 174.35*** | 179.44*** | 164.85*** | 154.40*** | 4   |

说明:表中统计量服从相应自由度下的卡方分布;\*\*\*表示在 1%的水平上显著;针对年度差异检验得到检验值为 455.91,自由度为 4。

#### (四)资本结构的优化程度

我们将最优负债率与实际负债率之间的比值( $TL_{it}^*/TL_{it}$ )定义为“最优比率”,用以衡量上市公司资本结构的优化程度。显然,当公司处于最优资本结构水平上时,其值为 1,最优比率与 1 之间的偏离程度越大表明公司的资本结构优化程度越低。表 5 第 II 栏列示按时间、公司规模和行业归属进行分类后的统计结果。

整体而言,我国上市公司负债不足,最优比率的均值为 1.161。究其原因,一方面,上市公司治理结构的固有缺陷以及二级市场上有限的监督功能,使得股权融资的实际成本明显低于债务融资,从而形成了中国上市公司的特有的股权融资偏好现象。另一方面,随着银行改制的不断深入,近年来债务还本付息的刚性特点正在不断加强,破产机制的引入使得负债融资的治理功能得以逐渐发挥出来,从而使上市公司很可能会为了逃避监督而更多地依赖股权融资。

从时间上来看,除 1999 和 2000 年最优比率较低外,其它年度的最优比率并不存在显著差异,基本维持在 1.19 左右,这从一定程度上反映出资本结构的整体优化是一个缓慢的过程,而我国的上市公司治理和资本市场的完善都还有很长的路要走。从行业分类来看,除社会服务业外,具有垄断特性的电力行业的优化程度最低,最优比率为 1.186,而竞争比较激烈的建筑和房地产业的优化程度最高,最优比率为 1.025,非常接近于 1。这一结果似乎表明行业竞争越激烈公司的资本结构优化程度越高 (James 和 Lewis, 1986),然而表 6 中列示的两个衡量行业竞争强度的指标(HHI<sub>a</sub> 和 HHI<sub>s</sub>)与最优比率之间并不存在明显的单调关系。这一结果或许可以从代理成本角度进行解释,我们发现相对于优化程度低的行业,优化程度高的行业具有较高的负债率,而债务融资本身能够对经理人行行为产生激励作用从而降低代理成本。当然,这一猜测还需在后续研究中做进一步检验。

## 六 结语

现代资本结构理论的发展在很大程度上是一个不断放松前期研究假设,进而拓展的过程。本文中,我们从动态角度对我国上市公司的资本结构调整行为进行了分析。相比于前期研究中所使用的静态模

① 相对于另一种常用的检验分组差异性的参数方法—单因素方差分析 (One-Way ANOVA), Kruskal-Wallis H 检验属于非参数方法,并不要求被检验变量满足正态分布和各组同方差两个假设条件,因此具有较高的检验力。在不存在组间差异的原假设下,检验统计量服从卡方分布,自由度为分组数减 1。



## 中国上市公司资本结构动态调整机制研究

型,本文所采用的动态模型具有更强的解释能力,说明我国上市公司存在最优资本结构,而由于调整成本的存在,公司无法始终维持在最优水平上,而是遵循一个部分调整的过程。通过放松肖作平(2004)的部分调整模型中调整速度不随时间和公司改变的假设,我们发现了以下两个非常有趣的结果:其一,调整速度会受到公司特征变量的影响。具体而言,调整速度与公司规模和偏离最优值的程度负相关,而与公司的成长性正相关。其二,调整速度会因时间、行业和公司规模的不同而呈现出显著的差异性,即调整速度固定不变的假设过于严格。我们估计出的调整系数平均值为 0.311,低于针对发达国家和发展中国家上市公司的估计值,表明我国上市公司面临较高的调整成本。整体而言,我国上市公司负债不足,相对于垄断性行业,竞争性行业的负债率更接近于最优水平,小规模公司的资本结构优化程度也高于大规模公司。

从实务的角度来看,我们的实证结果与陆正飞和高强(2003)对深市上市公司的问卷调查结果较为一致。他们发现,88%的样本公司认为应该设定一个“合理”的目标资本结构,而在这些公司中,又有 44% 的公司目前的负债率未达到自己认为的“合理”资本结构区间。这一方面表明上市公司存在最优资本结构,另一方方面也说明公司在向最优水平调整其资本结构的过程中面临着障碍。

本文的结果对我们进一步的研究指明了以下几个可能的方向。其一,对调整速度的估计仅能间接地反映公司在资本结构调整过程中所面临的调整成本,而调整成本显然会同时受到公司内部的经营状况和外部资本市场发展状况的影响。对于中国上市公司而言,政府的介入进一步增加了问题的复杂性。因此,对调整成本做进一步的分析和度量显得尤为必要。其二,既然最优资本结构是动态变化的,那么股票价格或收益的变动自然也会影响到公司的资本结构调整行为,Welch(2004)在这方面的开创性研究为我们提供了很好的借鉴。其三,本文研究表明不同行业的资本结构调整速度和优化程度存在显著的差异,通过简单地对比行业间的竞争强度和负债率等因素似乎还无法对这一结果做出令人满意的解释。从公司治理结构和行业竞争强度等角度对中国上市公司资本结构的调整和优化问题做进一步的探讨是我们后续研究的一个重要方向。

## 参考文献:

- 李善民、刘智:上市公司资本结构影响因素评述,《会计研究》,2003 年第 8 期。
- 冯根福、吴林江、刘世彦:我国上市公司资本结构形成的影响因素分析,《经济学家》,2000 年第 5 期。
- 郭鹏飞、孙培源:资本结构的行业特征:基于中国上市公司的实证研究,《经济研究》,2003 年第 5 期。
- 洪正:论中国上市公司资本结构的主导决定因素及其变迁逻辑,《经济评论》,2005 年第 3 期。
- 黄晓莉:我国上市公司资本结构影响因素实证分析,《数理统计与管理》,2002 年第 2 期。
- 刘志彪、姜付秀、卢二坡:资本结构与产品市场竞争强度,《经济研究》,2003 年第 7 期。
- 陆正飞、高强:中国上市公司融资行为研究——基于问卷调查的分析,《会计研究》,2003 年第 10 期。
- 陆正飞、辛宇:上市公司资本结构主要影响因素之实证研究,《会计研究》,1998 年第 8 期。
- 肖作平:资本结构影响因素和双向效应动态模型,《会计研究》,2002 年第 2 期。
- 阎达五、耿建新、刘文鹏:我国上市公司配股融资行为的实证研究,《会计研究》,2001 年第 9 期。
- 张军、郑祖玄、赵涛:中国上市公司资本结构:股权融资偏好、最优资本结构、还是过度融资?,《世界经济文汇》,2005 年第 6 期。
- Banerjee, S., A. Heshmati, and C. Wihlborg, 2004, "The dynamics of capital structure," *Research in Banking and Finance*, 4, pp. 275-297.
- Bowen, R.M., L.A. Daly, and C.C. Huber, Jr., 1982, "Evidence on the Existence and Determinations of Inter-Industry Differences in Leverage," *Financial Management*, 11, 10-20.
- Bradley, M., Jarrell, G. and E. H. Kim, 1984, "On the existence of an optimal capital structure: Theory and evidence," *Journal of Finance*, 39, pp. 857-878.

- Chen, J. J., 2004, "Determinants of capital structure of Chinese-listed companies," *Journal of Business Research*, 57, pp. 1341-1351.
- D'Agostino, R. B., A. Balanger, and R. B. D'Agostino, Jr, 1990, "A suggestion for using powerful and informative tests for normality," *The American Statistician*, 44, 316-321.
- Dangl, T., and J. Zechner, 2004, "Credit risk and dynamic capital structure choice," *Journal of Financial Intermediation*, 13, pp. 183-204.
- Drukker, D. M., 2003, "Testing for serial correlation in linear panel-data models," *The Stata Journal* (3)2, 1-10.
- Fischer, E. O., R. Heinkel, and J. Zechner, 1989, "Dynamic capital structure choice: Theory and tests," *Journal of Finance*, 44, pp. 19-40.
- Goldstein, R., N. J. Ju, and H. Leland, 2001, "An EBIT-Based model of dynamic capital structure," *Journal of Business*, 74(4), pp. 483-512.
- Greene, W. H., 2000, "Econometric Analysis" (4th), Prentice Hall international, Inc.
- Hart, O., Moore, J., 1995, "Debt and seniority: an analysis of the role of hard claims in constraining management," *American Economic Review*, 85, pp. 567-585.
- Hovakimian, A., T. Opler, and S. Titman, 2001, "The debt-equity choice," *Journal of Financial and Quantitative Analysis*, 36, pp. 1-24.
- Jalilvand, A. and R. Harris, 1984, "Corporate behavior in adjustment to capital structure and dividend targets: An econometric study," *Journal of Finance*, 39(1), pp. 127-145.
- James, B. and J. Lewis, 1986, "Oligopoly and financial structure: the limited liability effect", *American Economic Review*, 76, pp. 956-970.
- Jensen M. C. and W. H. Meckling, 1976, "Theory of the firm: Managerial behavior, agency costs, and ownership structure," *Journal of Financial Economics*, 3, pp. 305-360.
- Leland, H. E., 1994, "Corporate debt value, bond covenants, and optimal capital structure," *Journal of Finance*, 49, pp. 1213-1252.
- Leland, H. E., 1998, "Agency costs, risk management, and capital structure," *Journal of Finance*, 53, pp. 1213-1243.
- Löf, H., 2004, "Dynamic optimal capital structure and technical change," *Structural Change and Economic Dynamics*, 15, pp. 449-468.
- Modigliani, F. and M. H. Miller, 1958, "The cost of capital corporation finance and the theory of investment," *American Economic Review*, 48, pp. 261-297.
- Modigliani, F. and M. H. Miller, 1963, "Corporate income taxes and the cost of capital: A correction," *American Economic Review*, 53(3), pp. 433-443.
- Myers, Stewart C., 1984, "The capital structure puzzle," *Journal of Finance*, 39, pp. 575-592.
- Nivorozhkin, E., 2004, "The dynamics of capital structure in transition economies," *Economics of Planning*, 37 (1), pp. 25-45.
- Ross, S., 1977, "The determination of financial structure: the incentive-signaling approach," *Bell Journal of Economics*, 8, pp. 23-40.
- Scholes, M. S., and M. A. Wolfson, 1989, "Issues in the theory of optimal capital structure," In *frontiers of Modern Finance*, edited by S. Bhattacharya, and G. Constantinides. New York, N.Y.:Rowman & Littlefield.
- Shyam-Sunder, L. and S. C. Myers, 1999, "Testing Static Tradeoff against Pecking Order Models of Capital Structure," *Journal of Financial Economics*, 51, 219-244.
- Titman, S. and R. Wessels, 1988, "The determinants of capital structure choice," *Journal of Finance*, 43(1), pp. 1-19.
- Welch, I., 2004, "Capital structure and stock returns," *Journal of Political Economy*, 112, pp. 106-131.
- Wooldridge, J. M., 2002, "Econometric Analysis of Cross Section and Panel Data," Cambridge, MA: The MIT Press.

## The Dynamic Adjustment of Firms' Capital Structure in China

*Yujun Lian Ching-Fan Chung*

**Abstracts:** This paper investigates the dynamic adjustment of capital structure in the listed firms of China. The results show that, there are optimal capital structures, while the firms can only adjust their capital structure to the optimal level partially because the adjustment is costly. As a whole, the listed firms in China are under-leveraged. Moreover, the adjustment speed is influenced significantly by the size of firm, growth, and the distance between optimal and observed capital structure, and there are significant differences in adjustment speed of firms in year, industry and firm size.

**Keywords:** Capital Structure; Dynamic Adjustment; Adjustment Cost; Optimal Ratio

(责任编辑:林鲁东)



# 计量分析与 STATA 应用

---

钟经樊 连玉君

关于作者：钟经樊 台湾中央研究院 经济研究所

连玉君 中山大学 岭南学院 金融系

中文版本：版本 2.0，二〇一〇年六月

钟经樊和连玉君拥有版权 © 2007 – 2010。保留所有权利。

这份文档是我们即将出版的书稿，目前免费提供给中山大学岭南学院的师生使用。

发布这份文档的目的有二：

其一，用做授课讲义，帮助岭南学院的同学们学习 STATA；

其二，恳请大家对书稿提出修改意见，包括书稿的结构安排、表述错误，以及错别字等细节。

书稿的使用仅限于岭南学院范围内，请勿外传或散布于网络。

# 目录

|                             |    |
|-----------------------------|----|
| 第八章 面板模型及 <b>STATA</b> 应用   | 1  |
| 8.1 简介                      | 1  |
| 8.2 静态面板数据模型                | 2  |
| 8.2.1 固定效应模型                | 4  |
| 8.2.2 随机效应模型                | 10 |
| 8.2.3 假设检验                  | 13 |
| 8.3 STATA 实现 I: 静态面板模型      | 16 |
| 8.3.1 简介                    | 16 |
| 8.3.2 基本设定                  | 16 |
| 8.3.3 面板数据的处理               | 18 |
| 8.3.4 面板模型的估计               | 20 |
| 8.4 非均齐方差                   | 30 |
| 8.4.1 异方差                   | 30 |
| 8.4.2 序列相关                  | 34 |
| 8.4.3 方差形式未知时的稳健性估计         | 42 |
| 8.5 内生性问题与 IV/GMM 估计        | 55 |
| 8.6 动态面板模型                  | 55 |
| 8.6.1 简介                    | 56 |
| 8.6.2 IV 估计                 | 57 |
| 8.6.3 一阶差分 GMM (FD-GMM) 估计量 | 58 |
| 8.6.4 假设检验                  | 63 |
| 8.6.5 包含其它解释变量的动态面板模型       | 64 |
| 8.6.6 系统 GMM (SYS-GMM) 估计量  | 65 |
| 8.7 面板门槛模型                  | 66 |
| 8.7.1 简介                    | 66 |
| 8.7.2 单一门槛模型                | 66 |
| 8.7.3 多重门槛模型                | 71 |
| 8.7.4 STATA 实现              | 73 |



## 第八章

# 面板模型及 STATA 应用

### 8.1 简介

面板数据 (Panel Data)，简言之，是时间序列和截面数据的混合。严格地讲是指对一组个体 (如居民、国家、公司等) 连续追踪观察多期得到的资料。所以很多时候也称其为“追踪资料”。相对于单纯的截面资料和时序资料，这种特殊的资料结构，使得我们可以建立更为符合实际的计量模型。当然，由于资料结构的复杂性，也对模型的估计和分析提出了更高的要求。例如，由于面板资料是对特定的个体追踪多年得到的，此时，观察值之间彼此独立的假设可能不再成立。这会在很大程度上增加分析的难度，在非线性模型或动态模型中更是如此。

近年来，由于面板数据获得变得相对容易，使得其应用范围也不断扩大。而关于面板数据模型的计量理论也几乎涉及到了以往截面分析和时间序列分析中所有可能出现的主题，如近年来发展出的面板数据向量自回归模型 (Panel VAR)、面板数据单位根检验 (Panel Unit Root test)、面板数据协整分析 (Panel Cointegration)、面板数据门槛模型 (Panel Threshold) 等，都是在现有截面分析和时间序列分析中的热点主题的基础上发展起来的。

使用面板数据主要有以下几方面的优点：

- 便于控制个体的异质性。比如，我们在研究全国 30 个省份居民人均消费青岛啤酒的数量时，可以选取居民的收入、当地的啤酒价格、上一年的啤酒消费量等变量作为解释变量。但同时我们也会认为民族习惯、<sup>1</sup> 风俗文化、<sup>2</sup> 广告投放等因素也会显著地影响居民的啤酒消费量。对于特定的个体而言，前两种因素不会随时间的推移而有明显的变化，通常称为个体效应。而广告的投放往往通过电视或广播，我们可以认为在特定的年份所有省份所接受的广告投放量是相同的，通常称为“时间效应”。这些因素往往因为难以

---

<sup>1</sup>如宁夏属于回族自治区，那里的回民因为信仰伊斯兰教，所以不允许饮酒的，而生活在宁夏的许多汉民也往往因为自己的回民朋友无法饮酒而无形中减少了啤酒的消费量。

<sup>2</sup>如中国南部地区啤酒的消费量比较大，而北方很多地区只有在夏天才会饮用较多的啤酒，冬天他们一般只喝白酒。

获得数据或不易衡量而无法进入我们的模型，在截面分析中者往往会引起遗漏变量的问题。而面板数据模型的主要用途之一就在于处理这些不可观测的个体效应或时间效应。

- 包含的信息量更大，降低了变量间共线性的可能性，增加了自由度和估计的有效性。
- 便于分析动态调整。

本章主要介绍目前文献中常用的面板数据模型。第 8.2 节介绍两种基本的静态面板模型：固定效应模型和随机效应模型。第 8.4 节介绍异方差和序列相关稳健性估计量。第 8.6 节介绍动态面板模型，包括 FD-GMM 和 SYS-GMM 两种估计方法。有关面板数据模型的更为详尽的介绍，请参考 Baltagi (2001), Wooldridge (2002), Hsiao (2003) 以及 Arellano (2003)。

## 8.2 静态面板数据模型

我们一般所说的静态面板数据模型，是指解释变量中不包含被解释变量的滞后项 (通常为一阶滞后项) 的情形。但严格地讲，随机干扰项服从某种序列相关 (如 AR(1), AR(2), MA(1) 等) 的模型也不是静态模型。动态模型和静态模型在处理方法上往往有较大的差异。本节中我们重点介绍两种最为常用的静态模型 — 固定效应模型 (Fixed Effect Model) 和随机效应模型 (Random Effect Model)。

对于面板数据，在模型设定过程中，我们通常采用  $i$  表示个体<sup>3</sup> ( $i = 1, 2, \dots, N$ )，用  $t$  表示时间 ( $t = 1, 2, \dots, T$ )。最直接的想法可能是设定如下线性模型：

$$y_{it} = \alpha_{it} + \mathbf{x}_{it}'\boldsymbol{\beta}_{it} + \varepsilon_{it}$$

其中， $\boldsymbol{\beta}_{it}$  用于衡量个体  $i$  在第  $t$  时点， $\mathbf{x}_{it}$  对  $y_{it}$  的边际影响。显然，这个模型的设定过于一般化，因为我们假设  $\alpha_{it}$  和  $\boldsymbol{\beta}_{it}$  都会随着个体  $i$  和时点  $t$  发生变化。为此，需要对上述模型做进一步限定，例如，假设  $\boldsymbol{\beta}_{it}$  为常数，即  $\boldsymbol{\beta}_{it} = \boldsymbol{\beta}$ ，但常数项可以随着个体的不同而有所差异，可以表示如下：

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \quad (8-1)$$

$\mathbf{x}_{it}$  为  $K \times 1$  列向量 (不包含常数项)， $K$  为解释变量的个数， $\boldsymbol{\beta}$  为  $K \times 1$  系数列向量。这意味着，对于所有的个体和时点， $x$  的边际效果都相同，但个体  $i$  的平均水平不同于个体  $j$ 。对于特定的个体  $i$  而言， $\alpha_i$  表示那些不随时间改变的影响因素，而这些因素在多数情况下都是无法直接观测或难以量化的，如个人的消费习惯、企业文化和经营风格、国家的社会制度等，我们一般称其为“个体效应” (individual effects)。一般情况下，我们假设  $\varepsilon_{it}$  具有独立同分布的特征，<sup>4</sup> 均值为 0，方差为  $\sigma_\varepsilon^2$ 。由于模型 (8-1) 中， $\alpha_i$  可以看做随个体变化的截距项，那么也

<sup>3</sup>这里的个体可以是公司，国家，行业，或个人。

<sup>4</sup>也就是说，不同个体之间，以及同一个体的不同时间点上，干扰项  $\varepsilon_{it}$  都是不相关的。

就可以进一步将  $\alpha_i$  视为  $N$  个未知参数。也正因为如此, 模型 (8-1) 通常被称为“固定效应模型”(Fixed effects model)。

与固定效应模型相对应的另一种设定方式是所谓的“随机效应模型”(Random effects model)。该模型假设个体的截距项虽然有差异, 但不是固定的, 而是从一个服从均值为  $\mu$ , 方差为  $\sigma_\mu^2$  的分布中随机抽取的。模型设定如下:

$$y_{it} = \mu + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \quad (8-2)$$

该模型的干扰项包含两个部分: 不随时间改变的干扰项  $\alpha_i$  和通常意义上的(可以随时间改变的)干扰项  $\varepsilon_{it}$ 。<sup>5</sup> 需要说明的是, 由于我们在模型 (8-2) 中增加了截距项  $\mu$ , 此时  $\alpha_i$  的均值为 0。

对比二者的设定方式可知, 两种模型的差异主要反映在对“个体效应”的处理上。固定效应模型假设个体效应在组内是固定不变的, 个体间的差异反映在每个个体都有一个特定的截距项上; 随机效应模型则假设所有的个体具有相同的截距项, 个体间的差异是随机的, 这些差异主要反应在随机干扰项的设定上。基于此, 一种常见的观点认为, 当我们的样本来自一个较小的母体时, 我们应该使用固定效应模型, 而当样本来自一个很大的母体时, 应当采用随机效应模型。比如在研究中国地区经济增长的过程中, 我们以全国 28 个省区为研究对象, 可以认为这 28 个省区几乎代表了整个母体。同时也可以假设在样本区间内, 各省区间的经济结构、人口素质等不可观测的特质性因素是固定不变的, 因此采用固定效应模型是比较合适的。而当我们研究西安市居民的消费行为时, 即使样本数为 10000 人, 相对于西安市 600 万人口的母体而言仍然是个很小的样本。此时, 可以认为不同的居民在个人能力、消费习惯等方面的差异是随机的, 此时采用随机效应模型较为合适。

遗憾的是, 很多情况下, 我们并不能明确地区分我们的样本来自一个较大母体还是较小的母体。因此有些学者认为, 区分固定效应模型和随机效应模型应当看使用二者的假设条件是否满足。由于随机效应模型把个体效应  $\alpha_i$  设定为干扰项的一部分, 所以就要求解释变量与个体效应不相关。而在固定效应模型中, 个体效应  $\alpha_i$  被视为  $N$  个待估参数, 并不受这个假设条件的限制。因此, 如果我们的检验结果表明该假设满足, 那么就应采用随机效应模型, 因为它更为有效(所需估计的参数较少), 反之, 就需要采用固定效应模型。

另外, 有些学者认为具体采用哪一种模型主要决定于我们的分析目的。如果主要目的在于估计模型的参数, 而模型中个体的数目又不是很大的情况下, 采用固定效应模型是个不错的选择, 因为它非常容易估计。但当我们需要对模型的误差成分进行分析时(通常分解为长期效果和短期效果), 就只能采用随机效应模型。在这种情况下, 即使模型中的部分解释变量与个体效应相关, 我们仍然可以通过工具变量法对模型进行估计。

简言之, 两种模型有各自的优缺点和适用范围, 在实证分析的过程中, 我们一方面要根据分析的目的选择合适的模型, 同时也要以 8.2.3 节中介绍的假设检验方法为基础进行模型筛选。

<sup>5</sup>也正因为如此, 该模型也被称为“误差成分模型”(error components model)。

### 8.2.1 固定效应模型

#### 模型的假设条件

在估计模型 (8-1) 时，通常需要设定如下两个基本假设：<sup>6</sup>

假设 1：

$$E(\varepsilon_{it} | \mathbf{x}_{it}, \alpha_i) = 0$$

假设 2：

$$\text{Var}(\varepsilon_{it} | \mathbf{x}_{it}, \alpha_i) = \sigma^2$$

假设 1 表明干扰项  $\varepsilon$  与解释变量  $\mathbf{x}$  的当期观察值、前期观察值以及未来的观察值均不相关，也就是说模型中所有的解释变量都是严格外生的。假设 2 就是一般的同方差假设，在此假设下模型 (8-1) 的 OLS 估计是 BLUE (Best Linear Unbiased Estimator) 的。当此假设无法满足时，我们就需要处理异方差或序列相关以便得到稳健性估计量。

#### 最小二乘虚拟变量估计量

在假设 1 和假设 2 同时成立的情况下，我们可以采用虚拟变量的方式将模型 (8-1) 重新表述如下：

$$y_{it} = \sum_{j=1}^N \alpha_j d_{ij} + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it} \quad (8-3)$$

其中，若  $i = j$  时， $d_{ij} = 1$ ，否则为 0，即模型中包含了  $N$  个反应个体特征的虚拟变量。参数  $\alpha_1, \alpha_2, \dots, \alpha_N$  以及  $\boldsymbol{\beta}$  可以采用普通最小二乘法 (OLS) 估计得到。由此得到的  $\boldsymbol{\beta}$  系数称为“最小二乘虚拟变量估计量” (least squares dummy variable (LSDV) estimator)。当  $N$  较小时，采用这种方法非常简便，所有能执行 OLS 估计的计量软件都可以完成固定效应模型的估计。然而，当  $N$  比较大时，模型中将包含  $N + K$  个解释变量，计算的工作量往往很大，对于  $N$  相当大的情况 (如  $N = 100,000$ )，一般的计算机都无法胜任。因此，有必要先进行一些变换以消除固定效应，进而对简化的模型进行估计，随后三个小节介绍的法都是基于此目的进行的。

从模型 (8-3) 的设定形式可知，对于固定效应模型而言，由于  $\alpha_i$  不随时间变化，所以  $\mathbf{x}_{it}$  中不能包含不随时间改变的变量，如性别、种族、出生地等，因为这些变量都会与  $\alpha_i$  存在完全共线性。<sup>7</sup>

#### 组内估计量

##### 1. 基本思想

<sup>6</sup>一般应用中，我们也常采用如下两个相对较弱的假设。假设 1':  $E(\varepsilon_i | \mathbf{x}_i) = 0$  和假设 2':  $\text{Var}(\varepsilon_i | \mathbf{x}_i) = \sigma^2 \mathbf{I}_T$ 。

<sup>7</sup>若  $\mathbf{x}_{it}$  包含了任何不随时间变化的变量，STATA 会自动将这些变量删除。



给定 (8-1) 式, 我们可以进一步得到如下模型:

$$\bar{y}_i = \alpha_i + \bar{\mathbf{x}}_i \boldsymbol{\beta} + \bar{\varepsilon}_i \quad (8-4)$$

其中,  $\bar{y}_i = (1/T_i) \sum_{t=1}^{T_i} y_{it}$ ,  $T_i$  表示第  $i$  个个体的观察区间 (如  $T_1 = 5$  年,  $T_2 = 3$  年)。 $\bar{\mathbf{x}}_i$  和  $\bar{\varepsilon}_i$  的定义方式与此相同。换言之, 模型 (8-4) 表示个体  $i$  在样本观察区间内的平均值之间的关系。

模型 (8-1) 与模型 (8-4) 相减可以去除个体效应  $\alpha_i$ :<sup>8</sup>

$$(y_{it} - \bar{y}_i) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad (8-5)$$

若设定  $\dot{y}_{it} = (y_{it} - \bar{y}_i)$ ,  $\dot{\mathbf{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ , 以及  $\dot{\varepsilon}_{it} = (\varepsilon_{it} - \bar{\varepsilon}_i)$  则我们只需对如下模型执行 OLS 估计即可得到  $\boldsymbol{\beta}$  的估计值:

$$\dot{y}_{it} = \dot{\mathbf{x}}_{it}' \boldsymbol{\beta} + \dot{\varepsilon}_{it} \quad (8-6)$$

简言之, 要得到固定效应模型 (8-1) 的估计系数, 只需要从原始数据中间去其组内平均值, 进而对变换后的组内差分模型 (8-6) 执行 OLS 估计即可。为此, 该估计量也成为“组内估计量” (within group estimator), 记为  $\hat{\boldsymbol{\beta}}_{WG}$ 。

## 2. 更为严格的推导过程<sup>9</sup>

模型 (8-1) 可以采用向量的形式表示为:

$$\mathbf{y}_i = \alpha_i \mathbf{1}_T + \mathbf{x}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad (8-7)$$

其中,  $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iT})'$ ,  $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT})'$ ,  $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT})'$ ,  $\mathbf{1}_T$  是一个所有元素都为 1 的  $T \times 1$  列向量。将所有观察值进行堆叠, 模型 (8-1) 可用矩阵形式表示为:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (8-8)$$

其中,  $\mathbf{y} = (\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_N)'$ ,  $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \boldsymbol{\varepsilon}'_2, \dots, \boldsymbol{\varepsilon}'_N)'$ , 均为  $NT \times 1$  向量,  $\mathbf{D} = \mathbf{I}_N \otimes \mathbf{1}_T$ ,  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_N)'$ 。

需要注意的是, 在模型 (8-8) 中,  $\mathbf{D}$  项实际上对应着  $N$  个虚拟变量, 因此, 模型 (8-8) 等价于在模型  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  中加入  $N$  个虚拟变量。为了避免共线性问题, 解释变量  $\mathbf{X}$  中不应再包含常数项。<sup>10</sup>

<sup>8</sup>多数面板模型都具有大  $N$  小  $T$  结构, 此时我们重点关心的仍然是系数  $\boldsymbol{\beta}$ 。在  $N$  较小的面板模型中,  $\alpha_i$  的估计值可能成为分析的重点。此时可以采用 LSDV 进行估计, 或采用 (8-13) 式获得  $\alpha_i$  的估计值。

<sup>9</sup>对于矩阵运算不熟悉的读者可以跳过此节, 这并不影响你对固定效应模型估计方法的理解。

<sup>10</sup>当然, 我们也可以在  $\mathbf{X}$  中加入常数项, 但此时要同时加入约束条件:  $\sum_{i=1}^N \alpha_i = 0$ 。这样我们估计出的个体效应  $\hat{\alpha}_i$  就应当解释为个体  $i$  的相对截距项, 而不是前面得到的绝对截距项。STATA8.0 就采取了在  $\mathbf{X}$  中包含常数项的处理方式。

在正式估计模型之前，我们先定义一些有用的矩阵运算，它们将在后面的分析中反复使用。定义  $\mathbf{D}\mathbf{D}' = \mathbf{I}_N \otimes \mathbf{J}_T$ ，其中， $\mathbf{J}_T = \mathbf{1}_T \mathbf{1}_T'$  为  $T \times T$  维矩阵，每个元素均为 1。同时，定义  $\mathbf{P} = \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' = \mathbf{I}_N \otimes \bar{\mathbf{J}}_T$ ， $\bar{\mathbf{J}}_T = (1/T)\mathbf{J}_T$  是  $T \times T$  维矩阵，每个元素均为  $1/T$ ； $\mathbf{Q} = \mathbf{I}_{NT} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' = \mathbf{I}_{NT} - \mathbf{P}$ 。矩阵  $\mathbf{P}$  和  $\mathbf{Q}$  都具有如下性质：

- (1) 对称、幂等性:  $\mathbf{P}' = \mathbf{P}$ ，且  $\mathbf{P}^2 = \mathbf{P}$ ；
- (2) 正交性:  $\mathbf{P}\mathbf{Q} = \mathbf{0}$ ；
- (3) 和为单位矩阵:  $\mathbf{P} + \mathbf{Q} = \mathbf{I}_{NT}$ 。

我们可以从上述三个性质中的任意两个推导出第三个。易于证明， $\mathbf{Q}\mathbf{D} = \mathbf{0}$ ，因此，我们可以通过在等式 (8-8) 两边同时左乘  $\mathbf{Q}$  以消除固定效应：

$$\mathbf{Q}\mathbf{y} = \mathbf{Q}\mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\boldsymbol{\varepsilon} \quad (8-9)$$

易于证明，(8-9) 式变换后的结果其实就是前文提到的 (8-5) 式。变换后的模型的 OLS 估计量为：

$$\hat{\boldsymbol{\beta}}_{WG} = (\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}\mathbf{y} \quad (8-10)$$

方差估计量为：

$$\text{Var}(\hat{\boldsymbol{\beta}}_{WG}) = \sigma^2(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1} \quad (8-11)$$

显然， $\sigma^2$  的一致估计量为：

$$\hat{\sigma}^2 = \frac{1}{NT - N - K}(\mathbf{Q}\mathbf{y} - \mathbf{Q}\mathbf{X}\hat{\boldsymbol{\beta}}_{WG})'(\mathbf{Q}\mathbf{y} - \mathbf{Q}\mathbf{X}\hat{\boldsymbol{\beta}}_{WG}) \quad (8-12)$$

个体效应的估计值为：

$$\hat{a}_i = \bar{y}_i - \bar{\mathbf{x}}_i \hat{\boldsymbol{\beta}}_{WG} \quad (8-13)$$

### 一阶差分估计量

除了上述通过“组内去心”的办法消除固定效应外，还可以通过一阶差分的方式去除固定效应。对 (8-1) 式取一阶差分，得到

$$\begin{aligned} \Delta y_{i2} &= \Delta \mathbf{x}_{i2} \boldsymbol{\beta} + \Delta \boldsymbol{\varepsilon}_{i2} \\ &\vdots \\ \Delta y_{iT} &= \Delta \mathbf{x}_{iT} \boldsymbol{\beta} + \Delta \boldsymbol{\varepsilon}_{iT} \end{aligned} \quad (8-14)$$

采用矩阵形式可表示为

$$\mathbf{B}\mathbf{y}_i = \mathbf{B}\mathbf{x}_i \boldsymbol{\beta} + \mathbf{B}\boldsymbol{\varepsilon}_i \quad (8-15)$$

其中,

$$\mathbf{B} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}_{(T-1) \times T} \quad (8-16)$$

对所有观察值进行堆叠, 得到

$$(\mathbf{I}_N \otimes \mathbf{B})\mathbf{y} = (\mathbf{I}_N \otimes \mathbf{B})\mathbf{X} + (\mathbf{I}_N \otimes \mathbf{B})\boldsymbol{\varepsilon} \quad (8-17)$$

设  $\mathbf{Q}_B = \mathbf{I}_N \otimes \mathbf{B}$ , 则相应的 OLS 的估计量为:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{Q}_B\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}_B\mathbf{y} \quad (8-18)$$

根据假设 1 可知,  $E(\boldsymbol{\varepsilon}\mathbf{X}) = 0$ , 所以  $\hat{\boldsymbol{\beta}}_{OLS}$  是  $\hat{\boldsymbol{\beta}}$  的无偏估计量, 在  $N$  较大的情况下,  $\hat{\boldsymbol{\beta}}_{OLS}$  也是一致的。由假设 2 可知,  $\boldsymbol{\varepsilon}$  满足同方差假设, 且不存在序列相关。但变换后的干扰项  $\mathbf{B}\boldsymbol{\varepsilon}$  却并不满足同方差的假设,

$$\text{Var}(\mathbf{Q}_B\boldsymbol{\varepsilon}) = \sigma^2\mathbf{Q}_B\mathbf{Q}_B' \quad (8-19)$$

根据第四章中介绍的 GLS 理论可知, 模型 (8-17) 的 GLS 估计量是 BLUE 的,

$$\hat{\boldsymbol{\beta}}_{FD} = [\mathbf{X}\mathbf{Q}_B(\mathbf{Q}_B\mathbf{Q}_B')^{-1}\mathbf{Q}_B\mathbf{X}]^{-1}\mathbf{X}\mathbf{Q}_B(\mathbf{Q}_B\mathbf{Q}_B')^{-1}\mathbf{Q}_B\mathbf{y}. \quad (8-20)$$

易于证明  $\mathbf{Q}_B(\mathbf{Q}_B\mathbf{Q}_B')^{-1}\mathbf{Q}_B = \mathbf{Q}$ 。<sup>11</sup> 因此,

$$\hat{\boldsymbol{\beta}}_{GLS} \sim \hat{\boldsymbol{\beta}}_{WG}$$

也就是说, 采用一阶差分去除“固定效应”后, 再用 GLS 估计差分后的模型得到的 GLS 估计量与我们前面介绍的组内估计是等价的。由于二者都满足经典回归模型的基本假设, 所以都是 BLUE 的。

<sup>11</sup> 利用矩阵直乘的性质:  $(\mathbf{A} \otimes \mathbf{F})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{FD})$ , 可以得到  $\mathbf{Q}_B(\mathbf{Q}_B\mathbf{Q}_B')^{-1}\mathbf{Q}_B = \mathbf{I}_N \otimes \mathbf{B}'(\mathbf{BB}')^{-1}\mathbf{B}$ 。进一步, 可以证明  $\mathbf{B}'(\mathbf{BB}')^{-1}\mathbf{B} = \mathbf{I}_T - \bar{\mathbf{J}}_T$ : 由于矩阵

$$\mathcal{H} = \begin{bmatrix} T^{-1/2}\mathbf{1}_T' \\ (\mathbf{BB}')^{-1/2}\mathbf{B} \end{bmatrix}$$

满足  $\mathcal{H}\mathcal{H}' = \mathbf{I}_T$ , 所以  $\mathcal{H}'\mathcal{H} = \mathbf{I}_T$ , 即

$$\mathbf{1}_T'\mathbf{1}_T/T + \mathbf{B}'(\mathbf{BB}')^{-1}\mathbf{B} = \mathbf{I}_T$$

因此,  $\mathbf{Q}_B(\mathbf{Q}_B\mathbf{Q}_B')^{-1}\mathbf{Q}_B = \mathbf{I}_N \otimes (\mathbf{I}_T - \bar{\mathbf{J}}_T) = \mathbf{I}_{NT} - \mathbf{P} = \mathbf{Q}$ 。

### 前向正交分解

“前向正交分解” (forward orthogonal deviations, FOD) 法由 Arellano and Bover (1995) 提出。类似于上面介绍的一阶差分法。它也可以去除个体效果，但却不会在变换后的干扰项中引入序列相关问题。虽然在处理静态模型时这种方法略显繁复，但在动态面板数据模型的分析中，该方法显得格外重要。

正交变换基于如下  $(T-1) \times T$  矩阵

$$\mathbf{A} = (\mathbf{B}\mathbf{B}')^{-1/2}\mathbf{B}$$

如果将  $(\mathbf{B}\mathbf{B}')^{-1/2}$  视为裘拉斯基 (Cholesky) 分解的上三角阵，那么  $\mathbf{A}$  矩阵可表示为

$$\mathbf{A} = \text{diag}[(T-1)/T(T-2)/(T-1), \dots, 1/2]^{-1/2}\mathbf{A}^\dagger$$

其中，

$$\mathbf{A}^\dagger = \begin{bmatrix} 1 & (1-T)^{-1} & (1-T)^{-1} & \dots & (1-T)^{-1} & (1-T)^{-1} & (1-T)^{-1} \\ 0 & 1 & (2-T)^{-1} & \dots & (2-T)^{-1} & (2-T)^{-1} & (2-T)^{-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1/2 & -1/2 \\ 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{bmatrix} \quad (8-21)$$

因此，干扰项  $\varepsilon_i$  经矩阵  $\mathbf{A}$  转换后得到的  $\varepsilon_i^* = \mathbf{A}\varepsilon_i$  将具有如下  $T-1$  个元素：

$$\varepsilon_{it}^* = c_t \left[ \varepsilon_{it} - \frac{1}{T-t} (\varepsilon_{it+1} + \dots + \varepsilon_{iT}) \right] \quad (8-22)$$

其中， $c_t^2 = (T-t)/(T-t+1)$ 。显然， $\mathbf{A}'\mathbf{A} = \mathbf{I}_T - \bar{\mathbf{J}}_T$ ， $\mathbf{A}\mathbf{A}' = \mathbf{I}_{T-1}$ 。进一步，我们可以得到  $\mathbf{Q} = \mathbf{I}_N \otimes \mathbf{A}'\mathbf{A}$ 。采用 OLS 估计经过这种“前向正交分解”变换后的模型同样可以得到组内估计量  $\hat{\beta}_{WG}$ 。因此，正交分解可以视为一种类似于“一阶差分”的处理方式，其优点在于不会在转换后的干扰项中引入序列相关问题。

简言之，无论采用“组内去心”、“一阶差分”还是“正交分解”，我们都可以得到组内估计量。在第 8.6 节的动态面板数据模型中我们将主要应用这种转换方法来去除个体效应。

### 时间效应

前面介绍的固定效应模型着重在于考虑不可观测的个体效应，按照同样的思路，我们还可以在某些分析中考虑不可观测的时间效应。如在研究区域经济增长的过程中，全球石油价格的上涨、金融危机的爆发都会对所有研究对象在特定年份的产出有所影响。我们注意到，这些因素在特定的年份会对经济体中的所有个体产生影响，这启发我们可以通过设定时间虚拟变量来反映这些时间效应的影响。

#### 1. 直觉的解释和估计方法

我们可以在模型第 4 页中介绍的 LSDV 模型 (8-3) 的基础上进一步增加  $T - 1$  个时间虚拟变量  $\mathbf{s}_{it}$  来反映时间效应的影响:

$$y_{it} = \sum_{j=1}^N \alpha_j d_{ij} + \sum_{\tau=2}^T \lambda_{\tau} s_{it\tau} + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it} \quad (8-23)$$

其中, 若  $t = p$  时,  $s_{it\tau} = 1$ , 否则为 0。采用 OLS 即可获得所有系数的无偏估计量。即使在  $N$  较大的情况下, 我们仍然可以在经过组内去心的模型 (8-5) 中加入  $T - 1$  个时间虚拟变量来控制时间效应。简言之, 对于时间效应, 我们完全可以将其视为  $\mathbf{x}_{it}$  的一部分。

## 2. 更为严格的推导过程

模型的基本设定为:

$$y_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + u_{it} \quad (8-24)$$

$$u_{it} = a_i + \lambda_t + \varepsilon_{it}$$

其中,  $i = 1, 2, \dots, N$ ;  $t = 1, 2, \dots, T$ 。相应的向量形式为:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (8-25)$$

$$\mathbf{u} = (\mathbf{I}_N \otimes \mathbf{1}_T)\mathbf{a} + (\mathbf{1}_N \otimes \mathbf{I}_T)\boldsymbol{\lambda} + \boldsymbol{\varepsilon}$$

其中,  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_T)'$ ,  $\mathbf{a}$  和  $\boldsymbol{\varepsilon}$  的定义同前。假设对于任何  $i$  和  $t$  而言,  $\mathbf{x}_{it}$  均不与  $\varepsilon_{it}$  相关。为了分析的方便, 令  $\mathbf{D}_a = \mathbf{I}_N \otimes \mathbf{1}_T$ ,  $\mathbf{D}_{\lambda} = \mathbf{1}_N \otimes \mathbf{I}_T$ 。那么模型的矩阵形式可表示为:

$$\mathbf{y} = \mathbf{D}_a \mathbf{a} + \mathbf{D}_{\lambda} \boldsymbol{\lambda} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (8-26)$$

我们注意到,  $\mathbf{D}_a$  和  $\mathbf{D}_{\lambda}$  分别为  $(NT \times N)$  和  $(NT \times T)$  维矩阵, 当  $N$  或  $T$  较大时, 运算量都会很大。因此, 我们需要事先进行一些简单的运算以去除个体效应和时间效应。类似于前面  $\mathbf{J}_T$  的定义方式, 设  $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}'_N$ , 于是,  $\mathbf{D}_{\lambda} \mathbf{D}'_{\lambda} = \mathbf{J}_N \otimes \mathbf{I}_T$ 。同时, 定义  $\mathbf{E}_N = \mathbf{I}_N - \bar{\mathbf{J}}_N$ , 其中  $\bar{\mathbf{J}}_N = (1/N)\mathbf{J}_N$ 。进一步, 定义转换矩阵:

$$\mathbf{Q} = \mathbf{E}_N \otimes \mathbf{E}_T = \mathbf{I}_N \otimes \mathbf{I}_T + \mathbf{I}_N \otimes \bar{\mathbf{J}}_T - \bar{\mathbf{J}}_N \otimes \mathbf{I}_T + \bar{\mathbf{J}}_N \otimes \bar{\mathbf{J}}_T \quad (8-27)$$

该转换矩阵可以去除个体效应  $a_i$  和时间效应  $\lambda_t$ 。如,  $\tilde{\mathbf{y}} = \mathbf{Q}\mathbf{y}$  中的特定元素为:  $\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$ , 其中,  $\bar{y}_i = (1/N) \sum_{t=1}^T y_{it}$ ,  $\bar{y}_t = (1/T) \sum_{i=1}^N y_{it}$ ,  $\bar{y} = (1/NT) \sum_{i=1}^N \sum_{t=1}^T y_{it}$ 。因此, 我们可以用  $\tilde{\mathbf{y}} = \mathbf{Q}\mathbf{y}$  对  $\tilde{\mathbf{X}} = \mathbf{Q}\mathbf{X}$  进行 OLS 回归, 得到模型 (8-24) 的组内估计量为:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1} \mathbf{X}'\mathbf{Q}\mathbf{y} \quad (8-28)$$

个体效应和时间效应的估计量分别为:

$$\hat{a}_i = (\bar{y}_i - \bar{y}) - \tilde{\boldsymbol{\beta}}(\bar{x}_i - \bar{x}) \quad (8-29)$$

$$\hat{\lambda}_t = (\bar{y}_t - \bar{y}) - \tilde{\boldsymbol{\beta}}(\bar{x}_t - \bar{x}) \quad (8-30)$$

这里有两点需要注意：其一，模型 (8-24) 中不能包含不随时间或不随个体变化的解释变量，因为这些变量在转换过程中都被消除了；其二，我们没有特意强调模型中是否包含常数项，事实上只要保证不出现完全共线性问题即可，即，如果要加入常数项，那么就必须同时约束  $\sum_{i=1}^N a_i = 0$  和  $\sum_{t=1}^T \lambda_t = 0$ ；如果不加常数项，那么就无需作任何约束了。但加入常数项与否将影响到  $\hat{a}_i$  和  $\hat{\lambda}_t$  的含义，在解释系数的经济含义时需要注意。

### 8.2.2 随机效应模型

#### 模型的基本设定

当  $N$  很大时，采用固定效应模型往往会使参数的数目迅速增加，自由度的损失往往较大。前文已经提到，在固定小模型的设定中， $\mathbf{x}_{it}$  中不能包含不随时间改变的变量，如性别、种族等。然而，在有些研究中，我们研究的重点可能恰恰是这些变量。此时，随机效应模型可能更为适用。模型的基本设定同 (8-2):<sup>12</sup>

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it} \quad (8-31)$$

$$u_{it} = \alpha_i + \varepsilon_{it}$$

随机效应模型可以视为固定效应模型的一个扩展，这需要在上一节中假设 1 和假设 2 的基础上再增加如下假设：

假设 3：

$$\alpha_i \sim i.i.d(0, \sigma_\alpha^2)$$

假设 4：

$$\text{Cov}(\alpha_i, \mathbf{x}_{it}) = 0$$

假设 5：

$$\mathbf{u}_i | \mathbf{x}_i \sim i.i.d(0, \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T')$$

其中，假设 3 将个体效应  $\alpha_i$  设定为服从均值为 0，方差为  $\sigma_\alpha^2$  的随机变数，而我们在固定效应模型的设定中  $\alpha_i$  只是一个普通的解释变量，因此无需对它作任何限制；假设 4 非常显然，因为此时我们将  $\alpha_i$  视为随机干扰项的一部分，所以它不能与解释变量相关；假设 5 表明  $\alpha_i$  与  $\varepsilon_{it}$  相互独立。

#### 序列相关性

易于证明：

$$\text{Cov}(u_{it}, u_{js}) = \begin{cases} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \text{for } i = j, t = s \\ \sigma_\alpha^2 & \text{for } i = j, t \neq s \\ 0 & \text{for } i \neq j, t \neq s \end{cases} \quad (8-32)$$

<sup>12</sup>这里，为了表述的方便，我们把模型 (8-2) 中的常数项  $\mu$  放在了  $\mathbf{x}_{it}$  中。

和

$$\rho = \text{Corr}(u_{it}, u_{js}) = \begin{cases} 1 & \text{for } i = j, t = s \\ \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2) & \text{for } i = j, t \neq s \\ 0 & \text{for } i \neq j, t \neq s \end{cases} \quad (8-33)$$

从 (8-33) 式可以看出, 由于随机效应的引入使得组内不同时期的观察值之间存在固定不变的自相关关系, 相关系数为  $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2)$ 。这很容易理解, 因为尽管个体效应是随机的, 但在组内并不随时间改变, 组内不同期间固定的相关性也就必然存在。从另一个角度来看,  $\rho$  的含义在于, 模型中的总方差中有  $\sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2)$  来自于不随时间改变的干扰, 而余下的部分则归因于随个体和时间改变的干扰。当然, 在某些情况下这个假设显得过于严格。如在研究投资或消费时, 我们往往会假设组内不同期间的相关性是随时间逐渐减弱的。

### GLS 估计

基于以上设定, 可以写出干扰项的方差-协方差矩阵:

$$\mathbf{\Omega} = E(\mathbf{uu}') = \mathbf{I}_N \otimes (\sigma_\varepsilon^2 \mathbf{I}_T + \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T') = \mathbf{I}_N \otimes \mathbf{\Sigma} \quad (8-34)$$

其中,  $\mathbf{\Sigma} = \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T'$ , 具体形式为:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{bmatrix} \quad (8-35)$$

那么,  $\beta$  的 GLS 估计量为:

$$\hat{\beta}_{GLS} = [\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y} \quad (8-36)$$

方差估计量为:

$$\text{Var}(\hat{\beta}_{GLS}) = [\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}]^{-1} \quad (8-37)$$

为了进一步说明上述 GLS 估计量与前面介绍的组内估计量之间的关系, 我们可以对  $\mathbf{\Sigma}$  矩阵进行分解, 得到  $\mathbf{G} = \mathbf{\Omega}^{-1/2} = [\mathbf{I}_n \otimes \mathbf{\Sigma}]^{-1/2}$ , 进而采用  $\mathbf{G}$  矩阵对模型 (8-31) 进行转换。显然, 我们只需要求出  $\mathbf{\Sigma}^{-1/2}$  即可,

$$\mathbf{\Sigma}^{-1/2} = \frac{1}{\sigma_\varepsilon} \left[ \mathbf{I} - \frac{\theta}{T} \mathbf{1}_T \mathbf{1}_T' \right]$$

其中,

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}$$

于是我们可以对原始数据作如下转换：

$$\Sigma^{-1/2} \mathbf{y}_i = \frac{1}{\sigma_\varepsilon} \begin{bmatrix} y_{i1} - \theta \bar{y}_i \\ y_{i2} - \theta \bar{y}_i \\ \vdots \\ y_{iT} - \theta \bar{y}_i \end{bmatrix} \quad (8-38)$$

按照同样的方法我们可以对  $\mathbf{x}_i$  进行转换，对模型 (8-2) 转换后可得：

$$(y_{it} - \theta \bar{y}_i) = (\mathbf{x}_{it} - \theta \bar{\mathbf{x}}_i)' \boldsymbol{\beta} + (1 - \theta) \alpha_i + (\varepsilon_{it} - \theta \bar{\varepsilon}_i) \quad (8-39)$$

对模型 (8-39) 执行 OLS 估计即可得到与 (8-36) 式相同的结果。<sup>13</sup> 我们注意到，如果 (8-38) 式中的  $\theta = 1$ ，则上述变换就是我们前面讲到的“组内去心”，得到的就是固定效应模型对应的组内估计量 (8-10)。事实上，我们可以证明  $\hat{\boldsymbol{\beta}}_{GLS}$  可以表示为组内估计量和组间估计量的加权平均，详细过程请参考 Greene (2000, pp.295-296)。

### FGLS 估计

上面介绍的 GLS 估计是在假设方差成分已知的前提下进行了，但多数情况下我们并不知道  $\sigma_\varepsilon^2$  和  $\sigma_\alpha^2$ ，因此需要先估计这两个未知参数，继而用它们去代替 (8-35) 式中的真实值并采用 GLS 估计即可。基本思路是：先估计固定效应模型，得到  $\sigma_\varepsilon^2$  的估计值  $\hat{\sigma}_\varepsilon^2$ ，继而估计混合 OLS 模型，利用其残差和第一步得到的  $\hat{\sigma}_\varepsilon^2$  即可估计出  $\hat{\sigma}_u^2$ 。

由于组内估计量是无偏且一致的，所以我们可以利用固定效应模型的残差来估计  $\sigma_\varepsilon^2$ ，因为在估计固定效应模型的过程中我们已经去除了个体效应。设  $e_{it} = (y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \hat{\boldsymbol{\beta}}_{WG}$  为固定效应模型的残差，则

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T e_{it}^2}{nT - n - K} \quad (8-40)$$

接下来需要估计  $\sigma_\alpha^2$ 。模型 (8-31) 的 OLS 估计仍然无偏且一致的。设  $\tilde{e}_{it}$  为模型 (8-31) 的 OLS 残差，则

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T \tilde{e}_{it}^2}{nT - K - 1} = \hat{\sigma}_\varepsilon^2 + \hat{\sigma}_\alpha^2 \quad (8-41)$$

由此，我们可以得到：

$$\hat{\sigma}_\alpha^2 = \hat{\sigma}_u^2 - \hat{\sigma}_\varepsilon^2$$

由于该估计量可能为负值，所以我们可以略去 (8-40) 式和 (8-41) 式中对自由度的调整。这样就可以保证  $\hat{\sigma}_u^2$  一定是大于  $\hat{\sigma}_\varepsilon^2$  的，因为前者是后者在附加约束条件下的估计量。这种处理方法的依据在于我们只需要  $\sigma_\varepsilon^2$  和  $\sigma_\alpha^2$  的一致估计即可，至于是否无偏并不影响大样本性质。

<sup>13</sup>当然，在  $\theta$  未知的情况下，这一看似简单的估计方法是无法执行的。通过下一小节的介绍可以发现，事实上， $\theta$  中的参数  $\sigma_\varepsilon^2$  和  $\sigma_\alpha^2$  可以通过组内估计量和 Pooled OLS 估计量获得。因此，从实际操作的角度来讲，任何能执行 OLS 操作的软件，都可以用来估计随机效应模型。



上述估计方法虽然简单易行，但是当随机效应模型中包含不随时间改变的变量，如性别、种族等，我们就无法通过估计固定效应模型来估计  $\sigma_\varepsilon$  了。不过此时我们可以沿袭上面的思路，利用组间估计和混合 OLS 估计的残差来估计  $\sigma_\varepsilon^2$  和  $\sigma_\alpha^2$ 。采用 OLS 估计模型 (8-2) 的组内平均模型：

$$\bar{y}_i = \bar{\mathbf{x}}_i' \boldsymbol{\beta} + \bar{\mu}_i \quad (8-42)$$

可以得到一致估计量  $m^* = \hat{\sigma}_a^2 + (\hat{\sigma}_\varepsilon^2/T)$ ，结合  $m^*$  和  $\hat{\sigma}_u^2$  我们可以得到：

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 &= \frac{T}{T-1} (\hat{\sigma}_u^2 - m^*) \\ \hat{\sigma}_a^2 &= \frac{T}{T-1} m^* - \frac{1}{T-1} \hat{\sigma}_u^2 \end{aligned}$$

那么以上介绍的各种 FGLS 估计量哪个更为有效呢？我们知道，对于随机效应模型而言，针对方差成分的真实值进行 GLS 估计将得到 BLUE 估计量。而以上介绍的 FGLS 估计量在  $N \rightarrow \infty$  或  $T \rightarrow \infty$  或二者都成立的情况下，都是渐进有效的。Maddala 和 Mount(1973) 采用蒙特卡罗模拟方法对各种 FGLS 估计量的比较表明，在小样本下各种估计方法难分伯仲，所以建议采用简单易行的方法进行估计。Taylor (1980) 比较了小样本下随机效应的 FGLS 估计和固定效应的 LSDV 估计，结果表明：

- (1) 相对于 LSDV，FGLS 更具有效性，且具有较小的自由度；
- (2) FGLS 的方差不会大于 Cramer-Rao 下限的 17%。
- (3) 选择相对有效的方差成分估计量并不必然能够提高 FGLS 估计量的有效性。

### 8.2.3 假设检验

根据前面的介绍，我们大体可以采用三种方法估计面板数据模型：混合 OLS、固定效应模型和随机效应模型。那么如何对这三种模型进行区分和筛选呢？这就需要进行假设检验。显然，如果个体效应 (固定效应或随机效应) 显著异于零，那么就需要采用固定效应或随机效应模型。对于随机效应模型，它要求  $\text{Cov}(\alpha_i, \mathbf{x}_i) = 0$ ，而固定效应模型则没有这一限制，所以如果这一假设无法满足，我们就只能采用固定效应模型，或采用工具变量法来估计随机效应模型。

#### 固定效应的检验

由 8.2.1 小节的分析可知，固定效应模型的本质是通过个体间截距项的差异来捕捉不可观测的个体效果。但是，如果个体间 (组间) 并不存在统计意义上的显著差异，我们只需对混合数据执行 OLS 估计即可。<sup>14</sup> 检验的基本思路为，在个体效应不显著的原假设下，应当有如下关系成立：

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_n$$

<sup>14</sup>此时，模型设定为  $y_{it} = \alpha + \mathbf{x}_{it}' \boldsymbol{\beta} + \varepsilon_{it}$ 。文献中也将此模型成为“混合数据模型”，相应的估计量称为混合 OLS (Pooled OLS) 估计量。

我们可以采用 F 统计量来检验上述假设是否成立,

$$F = \frac{(R_u^2 - R_r^2)/(n-1)}{(1 - R_u^2)/(nT - n - K)} \sim F(n-1, nT - n - K) \quad (8-43)$$

其中,  $u$  表示不受约束的模型, 即我们的固定效应模型;  $r$  表示受约束的模型, 即混合数据模型, 仅有一个公共的常数项。

同理, 我们可以构造相应的 F 统计量来检验时间效应的显著性, 以及个体效应和时间效应的联合显著性。

### 检验随机效应

Breusch and Pagan (1980) 建议基于模型 (8-2) OLS 估计的残差构造 LM 统计量, 针对如下原假设来检验随机效应,

$$H_0: \sigma_\alpha^2 = 0 \quad v.s. \quad H_1: \sigma_\alpha^2 \neq 0$$

相应的检验统计量为:

$$LM = \frac{nT}{2(T-1)} \left[ \frac{\sum_{i=1}^n \left[ \sum_{t=1}^T e_{it} \right]^2}{\sum_{i=1}^n \sum_{t=1}^T e_{it}^2} - 1 \right]^2 \quad (8-44)$$

在原假设下, LM 统计量服从一个自由度为 1 的卡方分布。如果拒绝原假设则表明存在随机效应。如果采用矩阵的形式, 该 LM 统计量可以表示为:

$$LM = \frac{nT}{2(T-1)} \left[ \frac{\mathbf{e}'\mathbf{D}\mathbf{D}'\mathbf{e}}{\mathbf{e}'\mathbf{e}} - 1 \right]^2 \quad (8-45)$$

需要说明的是, 该检验假设模型的设定是正确的, 即  $\alpha_i$  与解释变量不相关, 而这一假设是否正确 还需要作进一步的检验, 这是我们下面要分析的内容。

### 固定效应还是随机效应? Hausman 检验

在前面的分析中, 我们从不同角度比较了固定效应模型和随机效应模型的差别, 那么在实际分析中应该使用哪个模型呢? 某些学者指出, 试图区分固定效应和随机效应本身就是错误的, 二者似乎不具可比性。Mundlak (1978) 指出, 一般情况下, 我们都应当把个体效应视为随机的。如果从单纯的实际操作角度来考虑, 固定效应模型往往会耗费很大的自由度, 尤其是对于截面数目很大的面板数据, 随机效应模型似乎更合适。但另一方面, 固定效应模型有一个独特的优势, 我们无须做个体效应与其它解释变数不相关的假设, 而在随机效应模型中, 这个假设是必须的, 否则就会导致内生性问题, 并进而导致参数估计的非一致性。

因此, 我们可以通过检验固定效应  $\alpha_i$  与其它解释变量是否相关作为进行固定效应和随机效应模型筛选的依据。此时, 我们可以采用 Hausman 检验。其基本思想是, 在  $\alpha_i$  与其他解释变量不相关的原假设下, 我们采用 OLS 估计固定效应模型和采用 GLS 估计随机效应模型得到的

参数估计都是无偏且一致的，只是前者不具有有效性。若原假设不成立，则固定效应模型的参数估计仍然是一致的，但随机效应模型却不是。因此，在原假设下，二者的参数估计应该不会有显著的差异，我们可以基于二者参数估计的差异构造统计检验量。

假设  $\mathbf{b}$  和  $\hat{\boldsymbol{\beta}}$  分别为固定效应模型的 OLS 估计和随机效应模型的 GLS 估计，则

$$\text{Var}(\mathbf{b} - \hat{\boldsymbol{\beta}}) = \text{Var}(\mathbf{b}) + \text{Var}(\hat{\boldsymbol{\beta}}) - \text{Cov}(\mathbf{b}, \hat{\boldsymbol{\beta}}) - \text{Cov}(\mathbf{b}, \hat{\boldsymbol{\beta}})' \quad (8-46)$$

基于上述 Hausman 检验的思想，有效估计量与它和非有效估计量之差的协方差应当为零，即

$$\text{Cov}[(\mathbf{b} - \hat{\boldsymbol{\beta}}), \hat{\boldsymbol{\beta}}] = \text{Cov}(\mathbf{b}, \hat{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{0} \quad (8-47)$$

由此我们可以得到：

$$\text{Cov}(\mathbf{b}, \hat{\boldsymbol{\beta}}) = \text{Var}(\hat{\boldsymbol{\beta}}) \quad (8-48)$$

将 (8-48) 式代入 (8-46) 式得到：

$$\text{Var}(\mathbf{b} - \hat{\boldsymbol{\beta}}) = \text{Var}(\mathbf{b}) - \text{Var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Psi} \quad (8-49)$$

Hausman 检验基于如下 Wald 统计量：

$$W = [\mathbf{b} - \hat{\boldsymbol{\beta}}]' \hat{\boldsymbol{\Psi}}^{-1} [\mathbf{b} - \hat{\boldsymbol{\beta}}] \sim \chi^2(K - 1) \quad (8-50)$$

其中， $\hat{\boldsymbol{\Psi}}$  采用固定效应和随机效应模型的协方差矩阵进行计算。如果拒绝了原假设，就表明个体效应  $\alpha_i$  和解释变量  $\mathbf{x}_{it}$  是相关的，此时我们有两种处理办法：一是采用固定效应模型，某些情况下这是一种无奈的选择；<sup>15</sup> 二是采用工具变量法来处理内生问题。<sup>16</sup>

<sup>15</sup> 因为有时我们通过 B-P 检验发现存在随机效应，但 Hausman 检验又表明使用随机效应模型的前提假设得不到满足，而我们又往往很难找到合适的工具变量，所以只能采用固定效应模型。

<sup>16</sup> 在 STATA 中可以采用 xthtaylor 和 xtivreg 命令来完成相应的估计。但这两个命令的侧重点还是有所差别的，前者重点处理的是模型 (8-31) 中  $\alpha_i$  与  $\mathbf{x}_i$  之间的相关性，而后者则重点处理通常意义上的内生性问题，即  $\varepsilon_i$  与  $\mathbf{x}_i$  之间的相关性。

## 8.3 STATA 实现 I: 静态面板模型

### 8.3.1 简介

在目前比较流行的计量软件中，STATA 在面板数据处理方面的优势比较明显。这一方面得益于 STATA 自身强大的数据处理功能，另一方面则归因于其快捷的更新速度。目前，STATA 不但能估计此前介绍的两种基本的静态面板模型 (xtreg 命令)，还能估计随后将要介绍的多种动态面板模型 (如 xtabond, xtdpdsys, xtdpd 命令)。对于内生性问题的处理也是 STATA 的一个强项 (如 xtivreg, xtivreg2 命令)。在 STATA11 中，我们还可以很方便的完成面板单位根检验 (xtunitroot 命令)、面板协整分析 (nharvey, xtwest 命令)，以及面板误差修正模型的分析 (xtpmg 命令)。与此同时，全球大量的 STATA 用户还提供了新近发展的面板门槛模型 (xtthres, xtptm 命令)，面板 VAR 模型 (pvar, xtvar) 等命令的估计程序。<sup>17</sup>

### 8.3.2 基本设定

#### 截面变量和时间变量的设定

我们首先通过一份简单的数据来说明如何在 STATA 中设定面板数据的结构。

```
. use http://www.stata-press.com/data/r11/invest2.dta, clear
. rename company id
. rename time year
. order id year
. replace year = year + 1990
(100 real changes made)
. list if (id<=3 & year<=1995), noobs
```

| id | year | invest | market | stock |
|----|------|--------|--------|-------|
| 1  | 1991 | 317.6  | 3078.5 | 2.8   |
| 1  | 1992 | 391.8  | 4661.7 | 52.6  |
| 1  | 1993 | 410.6  | 5387.1 | 156.9 |
| 1  | 1994 | 257.7  | 2792.2 | 209.2 |
| 1  | 1995 | 330.8  | 4313.2 | 203.4 |
| 2  | 1991 | 40.29  | 417.5  | 10.5  |
| 2  | 1992 | 72.76  | 837.8  | 10.2  |
| 2  | 1993 | 66.26  | 883.9  | 34.7  |
| 2  | 1994 | 51.6   | 437.9  | 51.8  |
| 2  | 1995 | 52.41  | 679.7  | 64.3  |
| 3  | 1991 | 33.1   | 1170.6 | 97.8  |
| 3  | 1992 | 45     | 2015.8 | 104.4 |
| 3  | 1993 | 77.2   | 2803.3 | 118   |

<sup>17</sup>若想全面了解 STATA 中有关面板数据的命令，可输入 help xt 命令。若想搜索有关面板数据的外部命令，可输入 findit panel data 命令。

|   |      |      |        |       |
|---|------|------|--------|-------|
| 3 | 1994 | 44.6 | 2039.7 | 156.2 |
| 3 | 1995 | 48.1 | 2256.2 | 172.6 |

这份数据共包含五个变量，其中，`id` 和 `year` 分别为截面变量和时间变量，分别对应于模型 (8-1) 中的下标  $i$  和  $t$ 。显然，通过这两个变量我们可以非常清楚地确定 **panel data** 的数据存储格式。因此，在使用 **STATA** 估计模型之前，我们必须告诉它截面变量和时间变量分别是什么，所用的命令为 `xtset`：<sup>18</sup>

```
. xtset id year
      panel variable:  id (strongly balanced)
      time variable:  year, 1991 to 2010
                delta:  1 unit
```

这里，**STATA** 确认了我们所设定的截面变量 (`id`) 和时间变量 (`year`)，并提示说我们的数据结构为 “strongly balanced”。这表示，在样本中，每个公司都有相同的年度观察值 (1991-2010)。在多数情况下，我们所收集的数据都是非平行数据 (**unbalanced panel data**)。<sup>19</sup>

若想面板数据的结构有更详细的了解，可以输入 `xtides` 命令：

```
. xtides
      id:  1, 2, ..., 5
      year: 1991, 1992, ..., 2010
      Delta(year) = 1 unit
      Span(year)  = 20 periods
      (id*year uniquely identifies each observation)

Distribution of T_i:  min      5%      25%      50%      75%      95%      max
                    20       20       20       20       20       20       20

      Freq.  Percent  Cum. | Pattern
-----|-----
      5      100.00  100.00 | 11111111111111111111
      5      100.00      | xxxxxxxxxxxxxxxxxxxxxx
```

在我们的样本中，共包含  $n = 5$  家公司，每家公司有  $T = 20$  年的观察值。同时，由于该样本是平行面板，所以  $T_i$  (每个公司对应的观察年数) 的分布在各个分位上都为 20。

## 统计描述

在正式进行模型的估计之前，我们必须对样本的基本分布特性有一个总体的了解。也要大体了解主要变量的均值、标准差、最大值、最小值等情况。显然，**STATA** 中有关统计描述

<sup>18</sup>另一个与该命令功能相似的命令是 `tsset`。需要注意的是，若同一家公司有两个以上相同年度的观察值，则截面变量和时间变量将无法 (联合起来) 唯一标示样本中的每一个观察值，此时，执行 `xtset` 命令时，**STATA** 将报告错误信息 “repeated time values within panel”。解决办法是在执行 `xtset` 命令前，先删除重复的观察值，命令为 “`duplicates drop id year, force`”。

<sup>19</sup>对于上市公司而言，有些公司上市较晚，有些公司中途退市，都可能导致个体间时间跨度的差异，从而使我们的数据是非平行的。随后我们会介绍这种两种数据结构对面板分析的影响，以及二者的转换。

的命令 (summarize, tabstat, histogram, kdensity 等) 仍然适用于面板数据的分析。同时, STATA 也专门为面板数据定制了一些进行描述性统计分析的命令, 如 xtsum、xttab, 以及 xttrans 等。<sup>20</sup>

xtsum 命令事实上是我们经常使用的命令 summarize 的扩展, 各个统计量都分别在样本总体、组内和组间三个层次上进行计算。二者的对比如下:

| . sum invest   |         |          |           |           |        |              |
|----------------|---------|----------|-----------|-----------|--------|--------------|
| Variable       |         | Obs      | Mean      | Std. Dev. | Min    | Max          |
| invest         |         | 100      | 248.957   | 267.8654  | 12.93  | 1486.7       |
| . xtsum invest |         |          |           |           |        |              |
| Variable       |         | Mean     | Std. Dev. | Min       | Max    | Observations |
| invest         | overall | 248.957  | 267.8654  | 12.93     | 1486.7 | N = 100      |
|                | between | 246.9354 | 42.8915   | 608.02    |        | n = 5        |
|                | within  | 149.9249 | -101.363  | 1127.637  |        | T = 20       |

相比于 summarize 命令, xtsum 提供了更为详细的信息。它将变量  $x_{it}$  分解成“组间”( $\bar{x}_i$ ) 和“组内”( $x_{it} - \bar{x}_i - \bar{\bar{x}}$ ) 两个部分。<sup>21</sup>

### 8.3.3 面板数据的处理

#### 1. 产生滞后项和差分项

由于 Panel Data 兼具截面数据和时间序列二者的特性, 所以对时间序列进行操作的运算同样可以应用于 Panel Data。这使得某些数据的处理变得非常方便。例如, 对于上述数据, 我们想产生一个新的变量  $invest_{it-1}$ , 也就是变量 invest 的一阶滞后项, 那么我们可以采用如下命令:

```
gen Lag_invest = L.invest
```

按照这样的思路, 还可以产生某个变量的移动平均、差分等。总之, 凡是应用到时间序列上的操作, 基本上都可以应用到 Panel Data 中来, 例如:

```
gen F_invest = F.invest // 超前项
gen D_invest = D.invest // 一阶差分
gen D2_invest = D2.invest // 二阶差分
```

#### 2. 产生组内均值

```
bysort id: egen mi_stock = mean(stock) // 每个公司的平均值
bysort year: egen mt_stock = mean(stock) // 每个年度的平均值
```

<sup>20</sup>另外一些用于面板数据统计性描述的命令可以从网上下载 (使用 findit 命令), 包括: xtcoun, countby, xtpattern, panels 等。

<sup>21</sup>这里所谓的“组间”其实就是个体的平均值。此外, 真正意义上的“组内”统计量应该是  $x_{it} - \bar{x}_i$ , 即 (8-5) 式的变换。这里之所以再加上总样本的平均值  $\bar{\bar{x}}$ , 是为了保证上述各统计量之间的可比性。由下文的分析可知, STATA 在估计固定效应模型时, 采用也是这一变换, 而不是理论推导公式 (8-5) 或 (8-9)。

## 3. 产生观察期末变量

```
bysort id: gen end_year = year[_n+1]
bysort id: replace end_year = year if _n==_N
```

或

```
bysort id: egen end_year2 = max(year)
```

## 4. 将非平行面板数据转换为平行面板数据

有时候我们的数据在经过初步处理后并非是平行面板数据，即每个截面的观察期数可能不同，可是许多情况下我们又必须使用平行数据，这就需要把非平行数据“削平”后转化为平行数据。STATA 官方发布的命令并不能快捷地处理这个看似简单问题。为此，笔者自行编写 `xtbalance` 命令来处理这一问题。在使用之前，读者需要输入如下命令安装 `xtbalance`：

```
ssc install xtbalance, replace
```

完成安装后，可输入 `help xtbalance` 命令查看其帮助文件。为了说明 `xtbalance` 的使用方法，我们先调入一份非平行面板数据：

```
. use http://www.stata-press.com/data/r11/abdata.dta, clear
. qui xtset id year
. xtides
```

```

      id:  1, 2, ..., 140              n =          140
    year: 1976, 1977, ..., 1984        T =           9
      Delta(year) = 1 unit
      Span(year)  = 9 periods
      (id*year uniquely identifies each observation)

```

| Distribution of T_i: |     |    |     |     |     |     |
|----------------------|-----|----|-----|-----|-----|-----|
|                      | min | 5% | 25% | 50% | 75% | 95% |
|                      | 7   | 7  | 7   | 7   | 8   | 9   |
|                      |     |    |     |     |     | max |
|                      |     |    |     |     |     | 9   |

| Freq. | Percent | Cum.   | Pattern   |
|-------|---------|--------|-----------|
| 62    | 44.29   | 44.29  | 1111111.. |
| 39    | 27.86   | 72.14  | .1111111. |
| 19    | 13.57   | 85.71  | .11111111 |
| 14    | 10.00   | 95.71  | 111111111 |
| 4     | 2.86    | 98.57  | 11111111. |
| 2     | 1.43    | 100.00 | ..1111111 |
| 140   | 100.00  |        | XXXXXXXXX |

假设我们想保留 1977-1983 年样本区间内的平行面板，便可输入如下命令：

```
. xtbalance, range(1977 1983)
(115 observations deleted due to out of range)
(384 observations deleted due to discontinues)
```

可见，有 115 个观察值由于在 1977-1983 年区间以外而被删除，另有 384 个观察值则因为观察年份不连续而被删除。读者可以输入 `xtides` 命令验证上述命令的效果。需要注意的是，虽然从表明上看，上述处理似乎使数据变成“balanced”，但由于部分变量仍包含缺漏值(可以

输入 `sum` 命令查验), 致使其实际上并非平行数据。更为稳妥的处理方法是附加 `miss(_all)` 选项, 以便在删除所有变量的缺漏值后, 再执行平行面板转换。<sup>22</sup>

### 8.3.4 面板模型的估计

#### STATA 面板模型命令概览

STATA 11 主要提供了多种面板模型的估计方法, 如表 8-1 所示 (参见 [xt] `xtreg`)。其中多数模型的估计方法我们都会在随后的章节中陆续讲到。

表 8-1: STATA 11.0 中用于估计 Panel Data 模型的主要命令一览

| 命令                      | 模型                                                                          |
|-------------------------|-----------------------------------------------------------------------------|
| <code>xtreg</code>      | Fixed-, between- and random-effects, and population-averaged linear models  |
| <code>xtregar</code>    | Fixed- and random-effects linear models with an AR(1) disturbance           |
| <code>xtgls</code>      | Panel-data models using GLS                                                 |
| <code>xtpcse</code>     | OLS or Prais-Winsten models with panel-corrected standard errors            |
| <code>xtrc</code>       | Random coefficients models                                                  |
| <code>xtivreg</code>    | Instrumental variables and two-stage least squares for panel-data models    |
| <code>xtivreg2</code>   | IV/2SLS, GMM and AC/HAC, LIML regression for panel data models (需要下载)       |
| <code>xtabond</code>    | Arellano-Bond linear dynamic panel-data estimator                           |
| <code>xtdpdsys</code>   | Arellano-Bond/Blundell-Bond estimation                                      |
| <code>xtdpd</code>      | Linear dynamic panel-data estimation                                        |
| <code>xtabond2</code>   | Arellano-Bond system dynamic panel data estimator (需要下载)                    |
| <code>xttobit</code>    | Random-effects tobit models                                                 |
| <code>xtintreg</code>   | Random-effects interval data regression models                              |
| <code>xtlogit</code>    | Fixed-effects, random-effects, population-averaged logit models             |
| <code>xtprobit</code>   | Random-effects and population-averaged probit models                        |
| <code>xtcloglog</code>  | Random-effects and population-averaged cloglog models                       |
| <code>xtpoisson</code>  | Fixed-effects, random-effects, population-averaged Poisson models           |
| <code>xtmixed</code>    | Multilevel mixed-effects linear regression                                  |
| <code>xtnbreg</code>    | Fixed-effects, random-effects, population-averaged negative binomial models |
| <code>xtfrontier</code> | Stochastic frontier models for panel-data                                   |
| <code>xthtaylor</code>  | Hausman-Taylor estimator for error-components models                        |
| <code>xtunitroot</code> | Panel-data unit-root tests                                                  |
| <code>xtwest</code>     | Westerlund error correction based panel cointegration tests                 |
| <code>xtpmg</code>      | Pooled mean-group, mean-group, and dynamic fixed-effects models (需要下载)      |

<sup>22</sup>当然, 也可以在 `miss()` 选项中指定变量的名称, `xtbalance` 将只删除这些变量中包含的缺漏值。



### 固定效应模型和随机效应模型的估计

第 8.2 节介绍的固定效应模型和随机效应模型 (以下分别简称 FE 和 RE)，可以采用 `xtreg` 命令估计，基本语法格式如下：

```
xtreg depvar [indepvars] [if] [in] [weight] [, model_type other_options ]
```

其中，`model_type` 选项用于指定需要估计的模型，对应关系如表 8-2 所示。这里有三点需要说明：其一，如果不填 `model_type` 选项，则 STATA 默认采用第 8.2.2 小节介绍 GLS 方法估计 RE 模型；其二，若设定 `mle` 选项，则 STATA 会采用 MLE 估计 RE 模型，相应的似然函数参见 [xt] `xtreg` (pp.466)；其三，上述命令格式只是一个基本形式，对于不同模型，还有一些相当灵活的控制选项，读者可以参考相应的帮助。

表 8-2: `xtreg` 命令中选项的含义

| model_type | 模型                                          |
|------------|---------------------------------------------|
| be         | Between-effects estimator                   |
| fe         | Fixed-effects estimator                     |
| re         | GLS Random-effects estimator                |
| pa         | GEE population-averaged estimator           |
| mle        | Maximum-likelihood Random-effects estimator |

下面，我们通过一个具体实例来说明上述命令的使用方法。我们仍然采用第 8.3.2 小节的 `invest2.dta` 数据，除了第 17 页中提到的 `id` 和 `year` 变量，另外三个变量分别是：`invest` 表示投资支出，`market` 表示市场价值，`stock` 表示资本存量。我们的目的是研究公司的投资额和资本存量如何影响其市场价值，为此建立了如下实证模型：

$$market_{it} = \mu + invest_{it} + stock_{it} + \alpha_i + \varepsilon_{it}$$

#### 1. FE 模型

若假设  $\alpha_i$  为固定效应，则上述模型就是一个典型的 FE 模型，估计结果如下：

```
. xtreg market invest stock, fe
```

|                                   |                      |   |        |
|-----------------------------------|----------------------|---|--------|
| Fixed-effects (within) regression | Number of obs        | = | 100    |
| Group variable: id                | Number of groups     | = | 5      |
| R-sq: within = 0.4168             | Obs per group: min = |   | 20     |
| between = 0.6960                  | avg =                |   | 20.0   |
| overall = 0.6324                  | max =                |   | 20     |
|                                   | F(2,93)              | = | 33.23  |
| corr(u_i, Xb) = 0.5256            | Prob > F             | = | 0.0000 |

| market | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|--------|-------|-----------|---|------|----------------------|
|        |       |           |   |      |                      |

|                        |           |                                   |       |                   |           |          |
|------------------------|-----------|-----------------------------------|-------|-------------------|-----------|----------|
| invest                 | 3.05273   | .4577368                          | 6.67  | 0.000             | 2.143756  | 3.961705 |
| stock                  | -.6763434 | .2216246                          | -3.05 | 0.003             | -1.116446 | -.236241 |
| _cons                  | 1372.613  | 76.96444                          | 17.83 | 0.000             | 1219.776  | 1525.449 |
|                        |           |                                   |       |                   |           |          |
| sigma_u                | 1023.5914 |                                   |       |                   |           |          |
| sigma_e                | 370.9569  |                                   |       |                   |           |          |
| rho                    | .88390837 | (fraction of variance due to u_i) |       |                   |           |          |
|                        |           |                                   |       |                   |           |          |
| F test that all u_i=0: |           | F(4, 93) =                        | 97.68 | Prob > F = 0.0000 |           |          |

其中, 选项 fe 表明我们采用的是固定效应模型。表头部分的前两行呈现了模型的估计方法 (Fixed-effects (within) regression)、截面变量的名称 (id)、以及估计中使用的样本数目和个体的数目。第 3 行到第 5 行列示了模型的拟合优度  $R^2$ , 分为组内、组间和样本总体三个层次 (详见第 8.3.4 页的解释)。第 6 行和第 7 行分别列示了针对模型中所有非常数变量执行联合检验得到的 F 统计量以及相应的 P 值, 本例中分别为 33.23 和 0.0000, 表明参数整体上相当显著。<sup>23</sup>第 8-11 行列示了解释变量的估计系数、标准误、t 统计量和相应的 P 值, 以及 95% 置信区间, 这和我们在进行截面回归时得到的结果是一样的。最后四行列示了固定效应模型中个体效应 ( $\alpha_i$ ) 和随机干扰项 ( $\varepsilon_{it}$ ) 的方差估计值、<sup>24</sup>以及二者在总方差中的相对比例, 即  $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2) = 0.8839$ 。<sup>25</sup>

需要注意的是, 表中最后一行列示了检验固定效应是否显著的 F 统计量 (参见 (8-43) 式) 和相应的 P 值。显然, 本例中固定效应非常显著。

细心的读者可能会产生如下疑问: 通过 (8-5) 或 (8-9) 式进行组内变换后的模型都不再包含常数项, 但为何上面的估计结果中还有常数项呢? 这是因为, STATA 中的“组内变换”与 (8-5) 式略有差异:

$$(y_{it} - \bar{y}_i + \bar{y}) = \alpha + (\mathbf{x}_{it} - \bar{\mathbf{x}}_i + \bar{\mathbf{x}})' \boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i + \bar{\varepsilon}) \quad (8-51)$$

其中,  $\bar{y} = (1/N) \sum_{i=1}^N \bar{y}_i$ , 即  $y_{it}$  的样本平均值,  $\bar{\mathbf{x}}$  和  $\bar{\varepsilon}$  的定义与此相似。相对于 (8-5) 式, 该模型中增加了一个常数项, 它表示个体效应  $\alpha_i$  的样本平均值。显然,  $\boldsymbol{\beta}$  系数的估计值和标准误在两种变换下并不存在任何差异。<sup>26</sup>

前文已经提到, 除了采用组内变换的方式, 亦可采用最小二乘虚拟变量法估计固定效应模型, 即 (8-3) 式。STATA 中有多种方法可以执行这一分析。最为简单的方法莫过于在 Pooled OLS 中增加  $N$  个反映个体效应的虚拟变量:

<sup>23</sup>F 统计量服从自由度分别为 2 和 93 的 F 分布。其中, 2 表示除常数项外, 模型中有两个解释变量,  $93 = NT - N - k = 5 \times 20 - 5 - 2$ 。

<sup>24</sup>STATA 列出的是二者的标准差, 分别为 sigma\_u 和 sigma\_e。其中, u 和 e 分别表示个体效应  $\alpha_{it}$  和干扰项  $\varepsilon_{it}$ 。

<sup>25</sup>这里的  $\sigma_\alpha^2$  经由个体效应的估计值  $\hat{\alpha}_i$  (由 (8-13) 式估得) 的标准差计算而得。显然, 它与 (8-33) 式中呈现的随机效应模型中的  $\sigma_\alpha^2$  具有不同的含义。

<sup>26</sup>STATA 之所以采用 (8-51) 式的变换, 一方面是为了保证不同模型之间的可比性, 另一方面则因为在模型中附加了常数项而保证了  $R^2$  仍然是有意义的。

```
. qui tabulate id, gen(dum_a)
```

```
. reg market invest stock dum_a*, noconstant
```

| Source   | SS        | df  | MS         | Number of obs = | 100    |
|----------|-----------|-----|------------|-----------------|--------|
| Model    | 556540367 | 7   | 79505766.7 | F( 7, 93) =     | 577.77 |
| Residual | 12797639  | 93  | 137609.021 | Prob > F =      | 0.0000 |
|          |           |     |            | R-squared =     | 0.9775 |
|          |           |     |            | Adj R-squared = | 0.9758 |
| Total    | 569338006 | 100 | 5693380.06 | Root MSE =      | 370.96 |

| market | Coef.     | Std. Err. | t     | P> t  | [95% Conf. Interval] |
|--------|-----------|-----------|-------|-------|----------------------|
| invest | 3.05273   | .4577368  | 6.67  | 0.000 | 2.143756 3.961705    |
| stock  | -.6763434 | .2216246  | -3.05 | 0.003 | -1.116446 -.236241   |
| dum_a1 | 2916.289  | 194.5067  | 14.99 | 0.000 | 2530.037 3302.54     |
| dum_a2 | 512.3015  | 85.89466  | 5.96  | 0.000 | 341.7317 682.8712    |
| dum_a3 | 1899.707  | 99.82705  | 19.03 | 0.000 | 1701.47 2097.944     |
| dum_a4 | 597.8959  | 83.66869  | 7.15  | 0.000 | 431.7464 764.0453    |
| dum_a5 | 936.87    | 158.2156  | 5.92  | 0.000 | 622.6851 1251.055    |

这里，我们首先采用 `tabulate` 命令附加 `gen()` 选项，生成了 5 个虚拟变量，<sup>27</sup>进而采用 `regress` 命令执行 OLS 估计。与此前的理论分析一致，此时得到的 `invest` 和 `stock` 变量的系数估计值与 `xtreg, fe` 的结果完全相同。然而，两种方法下得到的  $R^2$  存在显著差异：LSDV 下得到的  $R^2$  (0.9775) 明显高于 `xtreg, fe` 命令得到的  $R^2$  (within R-sq=0.4168)，这是因为后者在进行组内变换过程中去除个体效应  $\alpha_i$ ，致使 within R-sq 中并未包含  $\alpha_i$  对方差的贡献。

我们亦可用 `areg` 命令，或在 `regress` 命令前附加 `xi:` 前缀的方式估计 LSDV 模型：

```
. qui areg market invest stock, absorb(id) // areg
```

```
. est store fe_areg
```

```
. qui xi: reg market invest stock i.id // xi: reg
```

```
. est store fe_xi_reg
```

```
. local m "fe_areg fe_xi_reg"
```

```
. esttab `m', mtitle(`m') nogap scalar(N r2 r2_a) star(* 0.1 ** 0.05 *** 0.01)
```

```
>
```

|        | (1)<br>fe_areg       | (2)<br>fe_xi_reg     |
|--------|----------------------|----------------------|
| invest | 3.053***<br>(6.67)   | 3.053***<br>(6.67)   |
| stock  | -0.676***<br>(-3.05) | -0.676***<br>(-3.05) |
| _Iid_2 |                      | -2404.0***           |

<sup>27</sup>为了防止共线性问题，这里附加了 `noconstant` 选项。当然，也可以仅放入  $N - 1$  个虚拟变量，保留一个公共的常数项，此时无需再附加 `noconstant` 选项。

```

              (-12.40)
    _Iid_3      -1016.6***
              (-4.59)
    _Iid_4      -2318.4***
              (-11.30)
    _Iid_5      -1979.4***
              (-15.50)
    _cons       1372.6***      2916.3***
              (17.83)      (14.99)
-----
N                100          100
r2              0.936          0.936
r2_a            0.932          0.932
-----
t statistics in parentheses
* p<0.1, ** p<0.05, *** p<0.01

```

可以看出，两种方式得到的估计值和相同的统计量并不存在任何差异。需要说明的是，对于  $N$  较小的面板而言，采用第二种方法比较方便，能够直接得到个体效应的估计值  $\hat{\alpha}_i$ 。但当  $N$  较大时，这种方法将不再奏效，而 `areg` 和 `xtreg, fe` 命令则较好。<sup>28</sup> 相比于 `xtreg, fe`，虽然 `areg` 也同样未呈现个体效应的估计值，但在计算  $R^2$  时，它却考虑了个体效应对整个模型的方差贡献，因此得到的  $R^2$  相对较高。

## 2. RE 模型

若假设本例的样本公司是从一个很大的母体中随机抽取的，且  $\alpha_i$  与解释变量 `invest` 和 `stock` 均不相关，则我们可以将  $\alpha_i$  视为随机干扰项的一部分。此时，设定随机效应模型 (8-31) 更为合适。估计过程相当简单，仅需把上例中的 `fe` 选项去掉或附加 `re` 选项即可，例如：

```

. xtreg market invest stock, re

Random-effects GLS regression              Number of obs   =       100
Group variable: id                        Number of groups  =        5

R-sq:  within  = 0.4163                    Obs per group: min =       20
       between = 0.7054                      avg       =      20.0
       overall  = 0.6380                      max       =       20

Random effects u_i ~ Gaussian              Wald chi2(2)      =      95.98
corr(u_i, X)      = 0 (assumed)            Prob > chi2      =      0.0000

```

| market  | Coef.     | Std. Err.                         | z     | P> z  | [95% Conf. Interval] |           |
|---------|-----------|-----------------------------------|-------|-------|----------------------|-----------|
| invest  | 3.847014  | .4834565                          | 7.96  | 0.000 | 2.899457             | 4.794572  |
| stock   | -.7981618 | .256522                           | -3.11 | 0.002 | -1.300936            | -.2953879 |
| _cons   | 1212.764  | 154.6209                          | 7.84  | 0.000 | 909.7122             | 1515.815  |
| sigma_u | 223.80826 |                                   |       |       |                      |           |
| sigma_e | 370.9569  |                                   |       |       |                      |           |
| rho     | .26686395 | (fraction of variance due to u_i) |       |       |                      |           |

<sup>28</sup>在 STATA S.E. 版本中，矩阵的最大维度为 11000，因此，当  $N > 11000$  时，在 `regress` 命令中附加虚拟变量来估计 LSDV 模型的方法就不再适用了。

从列表形式上来看，此时得到的结果与 FE 模型并无大异。细心的读者可能注意到表头中呈现了如下信息 “ $\text{corr}(u_i, X) = 0$  (assumed)” ，这其实就是第 8.2.2 节中的假设 4，我们曾反复强调，该假设是保证 RE 模型估计结果无偏的基本前提。至于其他方面的差异，留待读者自行品味。

### 3. 模型的筛选和检验

这是模型设定过程中最为关键同时也是最难的一步，主要涉及使用混合 OLS 模型、FE 模型还是 RE 模型，更进一步还可能包括序列相关和异方差的检验等问题。在这方面功力的提高需要大量的实践经验和对理论的深入理解。

#### (1) 检验个体效应

对于固定效应模型而言，回归结果中最后一行汇报的 F 统计量便在于检验所有的个体效应整体上是否显著。在我们的例子中，上面的检验结果表明固定效应模型优于混合 OLS 模型。

#### (2) 检验随机效应

我们可以采用 (8-44) 式的 LM 统计量来检验随机效应是否显著，相应的命令为 `xttest0`：

```
. qui xtreg market invest stock, re
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

market[id,t] = Xb + u[id] + e[id,t]

Estimated results:

```

|        | Var      | sd = sqrt(Var) |
|--------|----------|----------------|
| market | 2018625  | 1420.783       |
| e      | 137609   | 370.9569       |
| u      | 50090.14 | 223.8083       |

```

Test:   Var(u) = 0
              chi2(1) =    325.74
              Prob > chi2 =    0.0000

```

这里，`qui` 命令的作用在于不把估计结果输出到屏幕上。LM 检验得到的 P 值为 0.0000，表明随机效应非常显著。可见，随机效应模型也优于混合 OLS 模型。

#### (3) Hausman 检验

虽然通过上面的分析，我们可以确认在模型中加入个体效应  $\alpha_i$ ，将显著优于  $\alpha_i$  为常数假设下的混合 OLS 模型，但还无法明确区分 FE 和 RE 的优劣。此时需要执行第 14 页中介绍的 Hausman 检验，具体步骤为：

- step1: 估计固定效应模型，存储估计结果；
- step2: 估计随机效应模型，存储估计结果；
- step3: 进行 Hausman 检验；

相应的 STATA 命令为:

```
. qui xtreg market invest stock, fe
. est store fe
. qui xtreg market invest stock, re
. est store re
. hausman fe re
```

|        | Coefficients |           | (b-B)<br>Difference | sqrt(diag(V_b-V_B))<br>S.E. |
|--------|--------------|-----------|---------------------|-----------------------------|
|        | (b)<br>fe    | (B)<br>re |                     |                             |
| invest | 3.05273      | 3.847014  | -.794284            | .                           |
| stock  | -.6763434    | -.7981618 | .1218184            | .                           |

```

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test:  Ho:  difference in coefficients not systematic

      chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =  -47.57    chi2<0 ==> model fitted on these
                        data fails to meet the asymptotic
                        assumptions of the Hausman test;
                        see suest for a generalized test

```

这里我们仍然采用 `qui` 命令屏蔽了结果的输出, 进而采用 `est store` 命令分别把 FE 和 RE 的估计结果存储到名称为 `fe` 和 `re` 的临时性文件中, 并最终用 `hausman` 命令调用二者的结果得到 (8-50) 式中的 Wald 统计量和相应的 P 值。

我们注意到, `sqrt(diag(V_b-V_B))` 全为缺失值, 而更为令人疑惑的是, Hausman 检验得到的统计量  $\chi^2(2) = -47.57$ , 是一个小于零的数值。理论上讲,  $\chi^2$  统计量一定为正数, 这是在进行 Hausman 检验过程中经常遇到的问题。产生这些情况的原因可能有多种, 但我认为一个主要的原因是模型设定有问题, 导致 Hausman 检验的基本假设得不到满足。<sup>29</sup>这时, 最好先重新审视一下模型设定是否合理, 看看是否遗漏了重要的解释变量, 或者某些变量是非平稳的等等。在确定模型的设定没有问题的情况下再进行 Hausman 检验, 如果仍然拒绝原假设或是出现上面的问题, 那么我们就认为随机效应模型的基本假设 (个体效应与解释变量不相关) 得不到满足。此时, 需要采用工具变量法或是使用固定效应模型。

对于采用 STATA 9.0 或以上版本的读者而言, 使用 `hausman` 命令中新增的 `sigmaless` 和 `sigmamore` 两个选项可以大大缓解  $\chi^2$  值为负的问题。看下面的例子:

```
. hausman fe re, sigmamore
```

|  | Coefficients |           | (b-B)<br>Difference | sqrt(diag(V_b-V_B))<br>S.E. |
|--|--------------|-----------|---------------------|-----------------------------|
|  | (b)<br>fe    | (B)<br>re |                     |                             |

<sup>29</sup>不过, STATA 手册的观点恰好相反, 认为当  $\chi^2$  统计量为负时, 意味着无法拒绝原假设。参见 [U] `hausman`, pp.642。

|        |           |           |          |          |
|--------|-----------|-----------|----------|----------|
| invest | 3.05273   | 3.847014  | -.794284 | .3300321 |
| stock  | -.6763434 | -.7981618 | .1218184 | .1205094 |

```

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

      chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =          38.91
      Prob>chi2 =          0.0000

. hausman fe re, sigmaless

      _____ Coefficients _____
            (b)      (B)      (b-B)      sqrt(diag(V_b-V_B))
            fe      re      Difference      S.E.
-----
invest      3.05273      3.847014      -.794284      .2580749
stock      -.6763434      -.7981618      .1218184      .0942346

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

      chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =          63.63
      Prob>chi2 =          0.0000

```

我们注意到，虽然通过设定 `sigmaless` 或 `sigmamore` 选项可以保证得到的 `chi2` 统计量为正数，但相应的 `P` 值却表明，`FE` 和 `RE` 不存在显著差异的原假设被高度拒绝了。这印证了我们的观点，即 `chi2` 为负是 **Hausman** 检验的原假设被拒绝的征兆。

### 时间固定效应

如果希望进一步在上述模型中加入时间效应，以便估计模型 (8-23)，那么可以采用时间虚拟变量来实现。首先，我们需要定义  $T - 1$  个时间虚拟变量：

```

tab year, gen(dumt)
drop dumt1

```

这里，`tab` 命令用于列示变量 `year` 的组类别，选项 `gen(dumt)` 用于产生  $T$  个以 `dumt` 开头的年度虚拟变量。第二条命令的作用在于去掉第一个虚拟变量以避免完全共线性。若在固定效应模型中加入时间虚拟变量，则估计模型 (8-23) 的命令为：

```

xtreg market invest stock dumt*, fe

```

若估计随机效应模型 (8-31) 中进一步控制时间效应，则只需将上述命令中的 `fe` 选项修改为 `re` 即可。

另一个我们非常关心的问题可能是时间效应的联合显著性，即采用类似于 (8-43) 式的 `F` 统计量来执行 **Wald** 检验。这可以采用 `test` 命令来完成。假设我们想检验 1992-1995 年的时间效应整体上是否显著，则可以执行如下命令：

```
. qui xtreg market invest stock dunt*, fe
. test dunt2 = dunt3 = dunt4 = dunt5 = 0

( 1)  dunt2 - dunt3 = 0
( 2)  dunt2 - dunt4 = 0
( 3)  dunt2 - dunt5 = 0
( 4)  dunt2 = 0

      F( 4,      74) =      9.20
      Prob > F =      0.0000
```

### 拟合值和残差的获取

完成上述模型的估计后, 可以采用 `predict` 命令获取被解释变量的拟合值, 以及模型的残差估计值。使用 `xtreg` 命令完成模型的估计后, `predict` 命令的语法格式如下:

```
predict newvar [if] [in] [, statistic]
```

表 8-3 呈现了 `statistic` 选项中各个统计量的代码及其含义。

表 8-3: `predict` 命令中 `statistic` 选项的含义

| 统计量               | 含义                                                                       |
|-------------------|--------------------------------------------------------------------------|
| <code>xb</code>   | $\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}$ , $y_{it}$ 的拟合值, 默认选项         |
| <code>ue</code>   | $\hat{\alpha}_i + e_{it}$ , 复合残差项                                        |
| <code>*xbu</code> | $\mathbf{x}_{it}'\hat{\boldsymbol{\beta}} + \hat{\alpha}_i$ , 包含个体效应的拟合值 |
| <code>*u</code>   | $\hat{\alpha}_i$ , 固定效应估计值或随机效应的误差成分                                     |
| <code>*e</code>   | $e_{it}$ , 残差                                                            |

注: 不带星号的统计量会针对所有样本进行计算, 若只需计算参与回归的样本对应的统计量, 可以采用 `predict...if e(sample)...` 命令; 带星号的统计量仅针对参与回归的样本进行计算。

在下面的例子中, 我们首先采用 1997-2010 年期间的样本估计了固定效应模型 ([L1]), 进而采用 `predict` 命令计算了被解释变量 `market` 的拟合值 ([L2])。由于在 [L2] 行的命令中, 我们附加了 `if e(sample)` 副指令, 计算出的拟合值 `market_hat` 仅包含在 [L1] 行中参与回归分析的样本对应的观察值, 即 1997-2010 年样本区间。因此, `market_hat` 变量中会包含 30 个缺漏值 (5 家公司在 1991-1996 年期间的观察值)。在 [L3] 和 [L4] 行, 我们采用两种方式计算了  $\hat{\alpha}_i + e_{it}$ , 从最后一行 `sum` 命令呈现的结果来看, 这两种计算方法完全等价。第 [L5] 和 [L6] 行的命令, 分别通过附加 `e` 和 `u` 选项, 计算出了个体效应  $\hat{\alpha}_i$  和残差  $e_{it}$ 。

```
. qui xtreg market invest stock if year>1996, fe // [L1]
. predict market_hat if e(sample), xb // y_hat = Xb [L2]
(30 missing values generated)
. gen ae_my = market - market_hat // y - y_hat = a_i + e_it [L3]
(30 missing values generated)
. predict ae_stata if e(sample), ue // a_i + e_it [L4]
```



```

(30 missing values generated)

. predict e, e                                // e_it                [L5]
(30 missing values generated)

. predict a, u                                // a_i                  [L6]
(30 missing values generated)

. sum ae_my ae_stata e a

```

| Variable | Obs | Mean      | Std. Dev. | Min       | Max      |
|----------|-----|-----------|-----------|-----------|----------|
| ae_my    | 70  | 1.30e-13  | 890.2898  | -997.8857 | 1991.987 |
| ae_stata | 70  | 1.30e-13  | 890.2898  | -997.8857 | 1991.987 |
| e        | 70  | -2.92e-14 | 290.6643  | -646.4621 | 863.0408 |
| a        | 70  | 1.59e-13  | 841.5047  | -783.008  | 1407.327 |

最后需要说明的是，对于固定效应模型而言，由于原始数据经过了组内变换，所以最终得到的残差序列不但在总样本中均值为零 ( $\sum_{i=1}^N \sum_{t=1}^T e_{it} = 0$ )，而且在每个截面内部，其均值同样为零 ( $\sum_{t=1}^T e_{it} = 0$ )。这一看似非常明了的结论在有些分析中显得尤为重要。例如，在研究资本结构文献中，学者们经常采用模型的残差来衡量公司的实际负债率与目标负债率之间的偏离程度；在现金持有文献中，残差通常用来衡量所谓的“超额现金持有 (excess cash)”；而在企业投资行为相关的文献中，残差则用以衡量投资不足或过度投资。<sup>30</sup>在这种情况下，若采用固定效应模型进行估计，在计算残差时，应采用表 8-3 中的“复合残差”，即在 predict 命令中附加 ue 选项。换言之，个体效应  $\hat{\alpha}_i$  也应考虑在内。

### 拟合优度—— $R^2$

虽然在传统的线性回归模型中，大家经常采用  $R^2$  来比较不同模型的优劣，然而，对于面板模型而言， $R^2$  的这一功能大为减弱。一方面，由于面板数据兼顾了截面资料和时序资料的特征，在有些研究中人们更关注组间 (between) 方差，而有些研究中则更为关注组内方差，而二者之间往往不具可比性。另一方面，只有在采用 OLS 进行估计，且模型中包含常数项的情况下， $R^2$  才能作为模型比较的标准，而对于多数面板模型，基本上都是采用 GLS、MLE 或 GMM 来估计的。尤其是在后两种情况下， $R^2$  几乎没有任何意义。

除了多数教科书讲到方差分解的方式获取  $R^2$  (参见第三章)， $R^2$  亦可定义为被解释变量的实际观察值与拟合值之间相关系数的平方。这里所言的拟合值意指上一小节中的  $\mathbf{x}'_i \hat{\boldsymbol{\beta}}$ ，即忽略  $\hat{\alpha}_i$ 。在上述例子中，xtreg 命令都会报告三个  $R^2$ ，下面解释其含义。

设  $\hat{\beta}$  为经由 xtreg 命令得到的系数估计值 (可以附加 be、fe 或 re 选项)，同时，用  $\rho^2(x, y)$  表示  $x$  和  $y$  之间相关系数的平方，则

<sup>30</sup>代表性的文献包括：资本结构 (Flannery and Rangan, 2006; Harford et al., 2009)、现金持有 (Opler et al., 1999; Mikkelsen and Partch, 2003; Frésard and Salva, 2010)、投资支出 (Richardson, 2006)。

$$\begin{aligned}
\text{Within } R^2 &: \rho^2 \left\{ (y_{it} - \bar{y}_i), (\mathbf{x}'_{it}\hat{\boldsymbol{\beta}} - \bar{\mathbf{x}}'_i\hat{\boldsymbol{\beta}}) \right\} \\
\text{Between } R^2 &: \rho^2(\bar{y}_i, \bar{\mathbf{x}}'_i\hat{\boldsymbol{\beta}}) \\
\text{Overall } R^2 &: \rho^2(y_{it}, \mathbf{x}'_{it}\hat{\boldsymbol{\beta}})
\end{aligned}$$

因此，对于固定效应模型而言，只有 **Within**  $R^2$  是真正意义上的  $R^2$ ，而对于组间效应模型而言，只有 **Between**  $R^2$  是真正意义上的  $R^2$ ，对于随机效应模型而言，并不存在真正意义上的  $R^2$ 。<sup>31</sup>

## 8.4 非均齐方差

在 8.2.1 和 8.2.2 小节的模型设定中，我们假设干扰项  $\varepsilon_{it}$  具有独立同分布的特征 (假设 2)。在很多情况下，这一假设显得过于严格。例如，对于大  $N$  小  $T$  型面板 (多见于微观个体或企业资料)，由于它主要表现出截面资料的特征，异方差便是一个需要重点考虑的问题；而对于大  $T$  小  $N$  型的面板 (多见于宏观资料)，截面内的时序特征往往是分析的重点，此时需要谨慎处理序列相关问题。除此之外，在两种资料形态中，截面相关也是一个不可忽略的问题。当模型中存在异方差、序列相关或截面相关时，在同方差假设下得到的估计量虽然仍旧是无偏且一致的，但不具有有效性。

在公司金融和资产定价领域，面板模型的应用越来越广泛，虽然多数学者多采用 White (1980) 的方法计算了异方差稳健型标准误，但对于序列相关和截面相关却并未给予足够的重视。Petersen (2009) 收集了在 2001-2004 年发表于 JF, JFE 和 RFS<sup>32</sup> 中的 207 篇文献，发现其中有 42% 并未针对可能存在的序列相关或组间相关调整其标准误。根据 Petersen 的理论分析和模拟分析，这可能导致严重的统计推断偏误。

### 8.4.1 异方差

在 FE 模型中，异方差主要源于  $\varepsilon_{it}$ ，即  $Var(\varepsilon_{it}) = \sigma_i^2$ ；而对于 RE 模型而言，由于其干扰项包含了  $\varepsilon_{it}$  和  $\alpha_i$  两个部分，二者都可能导致异方差。本节中仅介绍第一种情况，至于后两种情况，文献中应用有限，有兴趣的读者可以参考 Baltagi (2001)。

#### 组间异方差：FGLS 估计

模型的基本设定同 8.2.1 小节，考虑模型 (8-8)，

$$\mathbf{y} = \mathbf{D}\mathbf{a} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (8-52)$$

<sup>31</sup>有关这一主题的更多介绍，请参见 STATA11 Manual [xt], p.448，以及 Verbeek (2004, section 10.2.4)。

<sup>32</sup>这是金融领域最顶尖的三本期刊：Journal of Finance (JF), Journal of Financial Economics (JFE), Review of Financial Studies (RFS)。

其中,  $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \boldsymbol{\varepsilon}'_2, \dots, \boldsymbol{\varepsilon}'_n)'$ 。这里我们将第 4 页中的假设 2 放松为:

$$\text{Var}(\boldsymbol{\varepsilon}_i | \mathbf{x}_i, \alpha_i) = \sigma_i^2 \mathbf{I}_T \quad (8-53)$$

令  $\boldsymbol{\Sigma} = \text{diag}[\sigma_i^2]$ , 为  $N \times N$  矩阵, 则

$$\boldsymbol{\Omega}_0 = \text{Var}(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\Sigma} \otimes \mathbf{I}_T. \quad (8-54)$$

在 (8-52) 两边同乘  $\mathbf{Q}$  以消除个体效应, 得到

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \quad (8-55)$$

其中,  $\mathbf{y}^* = \mathbf{Q}\mathbf{y}$ ,  $\mathbf{X}^* = \mathbf{Q}\mathbf{X}$ ,  $\boldsymbol{\varepsilon}^* = \mathbf{Q}\boldsymbol{\varepsilon}$ 。干扰项的方差-协方差矩阵可以表示为:

$$\boldsymbol{\Omega} = \text{Var}(\boldsymbol{\varepsilon}^*) = E[\mathbf{Q}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{Q}'] = \mathbf{Q}\boldsymbol{\Omega}_0\mathbf{Q}' \quad (8-56)$$

于是, 模型 (8-55) 的 GLS 估计量为:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{GLS} &= [\mathbf{X}^{*'} \boldsymbol{\Omega}^{-1} \mathbf{X}^*]^{-1} \mathbf{X}^{*'} \boldsymbol{\Omega}^{-1} \mathbf{y}^* \\ &= [\mathbf{X}' \boldsymbol{\Omega}_0^{-1} \mathbf{X}]^{-1} \mathbf{X}' \boldsymbol{\Omega}_0^{-1} \mathbf{y} \\ &= \left[ \sum_{i=1}^N \frac{1}{\sigma_i^2} \mathbf{X}'_i \mathbf{X}_i \right]^{-1} \left[ \sum_{i=1}^N \frac{1}{\sigma_i^2} \mathbf{X}'_i \mathbf{y}_i \right] \end{aligned} \quad (8-57)$$

由此可以看出,  $\hat{\boldsymbol{\beta}}_{GLS}$  之所以比  $\hat{\boldsymbol{\beta}}_{WG}$  更为有效, 主要归因为采用  $1/\sigma_i^2$  作为权重, 从而为波动性较大的个体分配较小的权重。

进一步, 可以得到  $\hat{\boldsymbol{\beta}}_{GLS}$  的方差估计量为:

$$\text{Var}(\hat{\boldsymbol{\beta}}_{GLS}) = [\mathbf{X}^{*'} \boldsymbol{\Omega}^{-1} \mathbf{X}^*]^{-1} = [\mathbf{X}' \boldsymbol{\Omega}_0^{-1} \mathbf{X}]^{-1} \quad (8-58)$$

要获得相应的 FGLS 估计量, 需要首先估计出  $\boldsymbol{\Sigma}$  中的未知参数  $\sigma_i^2$ 。由于组内估计量在异方差设定下仍然是无偏且一致的, 所以我们可以基于组内估计的残差来估计  $\sigma_i^2$ 。令  $e_{it} = y_{it} - \mathbf{x}'_{it} \hat{\boldsymbol{\beta}}_{WG}$ , 其中,  $\hat{\boldsymbol{\beta}}_{WG}$  为模型 (8-52) 在同方差设定下的组内估计量, 即 (8-10)。由此可以得到  $\sigma_i^2$  的一致估计量:

$$\hat{\sigma}_i^2 = \frac{1}{T} \sum_{t=1}^T e_{it}^2 \quad (8-59)$$

进而可以得到  $\hat{\boldsymbol{\Sigma}} = \text{diag}[\hat{\sigma}_i^2]$ , 以及  $\hat{\boldsymbol{\Omega}}_0 = \hat{\boldsymbol{\Sigma}} \otimes \mathbf{I}_T$ 。用  $\hat{\boldsymbol{\Omega}}_0$  代替 (8-57) 和 (8-58) 式中的  $\boldsymbol{\Omega}_0$  即可得到相应的 FGLS 估计量。

## 组间异方差：Wald 检验

Greene (2000, pp.598) 建议采用如下修正后的 Wald 统计量来检验组间异方差：

$$W' = \sum_{i=1}^N \frac{(\hat{\sigma}_i^2 - \hat{\sigma}^2)^2}{V_i} \quad (8-60)$$

其中，

$$V_i = \frac{1}{T} \frac{1}{T-1} \sum_{t=1}^T (e_{it}^2 - \hat{\sigma}_i^2)^2 \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N \sigma_i^2$$

在原假设 ( $H_0: \sigma_i^2 = \sigma^2$ ) 成立的情况下，该统计量的渐进分布为：

$$W' \xrightarrow{d} \chi^2(N)$$

## ► Example

易于证明，给定假设条件 (8-53)，模型 (8-52) 的 GLS 估计量 (8-55) 与如下不考虑个体效应，但假设组间存在异方差的混合模型的 GLS 估计量完全相同：

$$\mathbf{y}_i = \mathbf{X}_i' \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad \text{Var}(\boldsymbol{\varepsilon}_i | \mathbf{x}_i, \alpha_i) = \sigma_i^2 \mathbf{I}_T$$

首先，需要采用 (8-60) 式的  $W'$  统计量来检验模型中是否存在组间异方差。Baum(2001) 编写的 xttest3 命令可以很方便地完成这一任务：

```
. qui xtreg market invest stock, fe
. xttest3
Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model
H0: sigma(i)^2 = sigma^2 for all i
chi2 (5) = 862.08
Prob>chi2 = 0.0000
```

显然，原假设被拒绝了。此时，需要采用 (8-57) 获得  $\boldsymbol{\beta}$  的 GLS 估计量，命令为 xtglsls：

```
. xtglsls market invest stock, panels(heteroskedastic)
Cross-sectional time-series FGLS regression
Coefficients: generalized least squares
Panels: heteroskedastic
Correlation: no autocorrelation

Estimated covariances = 5 Number of obs = 100
Estimated autocorrelations = 0 Number of groups = 5
Estimated coefficients = 3 Time periods = 20
Wald chi2(2) = 159.43
```

Prob > chi2 = 0.0000

| market | Coef.    | Std. Err. | z    | P> z  | [95% Conf. Interval] |          |
|--------|----------|-----------|------|-------|----------------------|----------|
| invest | 3.616278 | .5102399  | 7.09 | 0.000 | 2.616226             | 4.61633  |
| stock  | .7711491 | .378626   | 2.04 | 0.042 | .0290557             | 1.513243 |
| _cons  | 543.555  | 78.41464  | 6.93 | 0.000 | 389.8652             | 697.2449 |

其中，组间异方差通过 `panels()` 选项来设定。有兴趣的读者可以对比一下 GLS 估计值与第 21 页中不考虑异方差时的估计结果。

上述结果是采用两步法获得，即，先采用 OLS 估计不考虑异方差的模型 (8-52)，进而利用其残差计算 (8-59) 式中的  $\hat{\sigma}_t^2$ ，并最终得到 FGLS 估计量。

当然，在完成上述 FGLS 估计后，我们可以采用其残差重新计算  $\hat{\sigma}_t^2$ ，并带入 (8-57) 式，计算新一轮的估计系数。这一过程可以反复执行，直到两次估计值之间的差异小于某一预设的临界值为止。这一过程称为迭代 GLS 法，只需在上述命令中进一步附加 `igls` 选项即可得到相应的结果：

```
. xtglm market invest stock, p(het) igls
Iteration 1: tolerance = .15864396
(output omitted)
Iteration 45: tolerance = 8.265e-08

Cross-sectional time-series FGLS regression
Coefficients: generalized least squares
Panels:      heteroskedastic
Correlation: no autocorrelation

Estimated covariances      =      5      Number of obs      =      100
Estimated autocorrelations =      0      Number of groups   =      5
Estimated coefficients      =      3      Time periods      =      20
                                Wald chi2(2)      =      236.61
Log likelihood              = -759.5837          Prob > chi2          =      0.0000
```

| market | Coef.    | Std. Err. | z     | P> z  | [95% Conf. Interval] |          |
|--------|----------|-----------|-------|-------|----------------------|----------|
| invest | 3.214628 | .3395728  | 9.47  | 0.000 | 2.549078             | 3.880179 |
| stock  | .5016944 | .305698   | 1.64  | 0.101 | -.0974626            | 1.100851 |
| _cons  | 467.0456 | 36.60579  | 12.76 | 0.000 | 395.2995             | 538.7916 |

可见，经过 45 轮迭代，模型最终达到了收敛，系数估计值和标准误都有所变化，尤其是 `stock` 变量的标准误明显增大了，此时其系数估计值并不显著。

◀

### 8.4.2 序列相关

在此前的分析中，我们假设 FE 或 RE 模型中的个体效应  $\alpha_i$  可以有效捕捉截面内的跨期相关性，<sup>33</sup>并进而假设干扰项  $\varepsilon_{it}$  是 *i.i.d* 的，即，对于不同的个体，以及同一个个体的不同时间点上， $\varepsilon_{it}$  均不相关。然而，对于  $T$  较大的面板而言， $\alpha_i$  往往无法完全反映时序相关性，此时  $\varepsilon_{it}$  便可能存在序列相关，在多数情况下被设定为  $AR(1)$  过程。

这里，我们重点介绍 RE 模型中序列相关的处理方法，对于 FE 模型而言，处理方法相对简单，有兴趣的读者可以参考 Bhargava, Franzini and Narendranathan (1982)，以及 Stata 11 手册 [xt] **xtregar** 中的相关说明。

这里，我们仅介绍  $\varepsilon_{it}$  服从  $AR(1)$  过程时的估计方法，<sup>34</sup>模型的基本设定如下：

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it} \quad (8-61a)$$

$$u_{it} = \alpha_i + \varepsilon_{it} \quad (8-61b)$$

$$\varepsilon_{it} = \rho\varepsilon_{it-1} + v_{it} \quad (8-61c)$$

其中， $\alpha_i \sim i.i.d(0, \sigma_u^2)$ ， $v_{it} \sim i.i.d(0, \sigma_v^2)$ ，同时我们假设  $\alpha_i$  与所有  $\varepsilon_{it}$  均不相关，且满足稳定性条件  $|\rho| < 1$ 。处理的基本思路很简单，我们首先采用传统的处理时间序列模型的方法消除序列相关，进而采用 GLS 估计变换后的模型。对于第一期观察值，同样有两种处理方法：一是采用 Cochrane-Orcutt (1949) 建议的方法，舍弃第一期观察值；二是依据 Prais-Winsten (1956) 的方法，对第一期观察值进行特别处理。

#### Cochrane-Orcutt 估计

我们对模型 (8-61) 进行准差分处理，得到

$$(y_{it} - \rho y_{it-1}) = (\mathbf{x}_{it} - \rho \mathbf{x}_{it-1})'\boldsymbol{\beta} + (1 - \rho)\alpha_i + v_{it}$$

对应的向量形式为：

$$\mathbf{C}y_i = \mathbf{C}x_i + \mathbf{C}_1\alpha_i + \mathbf{C}v_i \quad (8-62)$$

其中， $\mathbf{C}_1 = (1 - \rho)\mathbf{1}_{T-1}$  为  $(T-1) \times 1$  列向量，

$$\mathbf{C} = \begin{bmatrix} -\rho & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}_{(T-1) \times T} \quad (8-63)$$

<sup>33</sup>例如，在 8.2.2 小节中已经提到，在 RE 模型中，由于  $\alpha_i$  不随时间变化，所以组内观察值会存在不随时间改变的序列相关性 (参见 (8-32) 式)。当然，在有些情况下，这显然是一个过于严格的假设条件。

<sup>34</sup>至于  $\varepsilon_{it}$  服从  $AR(2)$ 、 $AR(4)$  或  $MA(1)$  过程时的估计方法，有兴趣的读者可以参考 Baltagi (2001) 第五章相关内容。

我们可进一步写出 (8-62) 的矩阵形式:

$$(\mathbf{I}_N \otimes \mathbf{C})\mathbf{y} = (\mathbf{I}_N \otimes \mathbf{C})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_N \otimes \mathbf{C}_1)\mathbf{a} + (\mathbf{I}_N \otimes \mathbf{C})\mathbf{v} \quad (8-64)$$

其中,  $\mathbf{a} = (\alpha_1, \alpha_2, \dots, \alpha_n)'$ 。令  $\mathbf{u}^* = (\mathbf{I}_N \otimes \mathbf{C}_1)\mathbf{a} + (\mathbf{I}_N \otimes \mathbf{C})\mathbf{v}$ , 则模型 的方差-协方差矩阵为:<sup>35</sup>

$$\boldsymbol{\Omega} = E[\mathbf{u}^* \mathbf{u}^{*'}] = (1 - \rho)^2 \sigma_\alpha^2 (\mathbf{I}_N \otimes \mathbf{J}_{T-1}) + \sigma_v^2 (\mathbf{I}_N \otimes \mathbf{C}\mathbf{C}') \quad (8-65)$$

其中,  $\mathbf{J}_{T-1}$  为  $(T-1) \times (T-1)$  方阵, 所有元素都为 1。于是, 模型 (8-64) 的 GLS 估计为:

$$\hat{\boldsymbol{\beta}}_{GLS} = [\mathbf{X}^{*'} \boldsymbol{\Omega}^{-1} \mathbf{X}^*]^{-1} \mathbf{X}^{*'} \boldsymbol{\Omega}^{-1} \mathbf{y}^* \quad (8-66)$$

系数的方差-协方差矩阵为:

$$\text{Var}(\hat{\boldsymbol{\beta}}_{GLS}) = [\mathbf{X}^{*'} \boldsymbol{\Omega}^{-1} \mathbf{X}^*]^{-1} \quad (8-67)$$

其中,  $\mathbf{y}^* = (\mathbf{I}_N \otimes \mathbf{C})\mathbf{y}$ ,  $\mathbf{X}^* = (\mathbf{I}_N \otimes \mathbf{C})\mathbf{X}$ 。

由于  $\boldsymbol{\Omega}$  中含有未知参数  $\rho$ ,  $\sigma_u^2$  和  $\sigma_v^2$ , 所以要进行 FGLS 估计就需要先获得这三个参数的一致估计量。由于模型 (8-61) 的组内估计量是无偏且一致的, 所以我们可以利用其残差  $e_{it}$  来估计  $\rho$ :

$$\hat{\rho} = \sum_{i=1}^n \hat{\rho}_i \quad \text{其中,} \quad \hat{\rho}_i = \frac{\sum_{t=2}^T e_{it} e_{it-1}}{\sum_{t=2}^T e_{it-1}^2} \quad (8-68)$$

同时, 由 (8-61b) 式可知

$$\text{Var}[u_{it}] = \sigma_\alpha^2 + \sigma_v^2 / (1 - \rho^2) \quad (8-69)$$

而由 (8-61c) 式可知

$$\text{Var}[\varepsilon_{it}] = \sigma_v^2 / (1 - \rho^2) \quad (8-70)$$

这提示我们可以采用模型 (8-61) 的混合最小二乘估计残差和组内估计残差来估计  $\sigma_u^2$  和  $\sigma_v^2$ 。

设  $\hat{e}_{it}$  为对 (8-61a) 采用 OLS 估计得到的残差,<sup>36</sup>那么依据 (8-69) 有如下关系成立:

$$\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \hat{e}_{it}^2 = \hat{\sigma}_a^2 + \frac{\hat{\sigma}_v^2}{1 - \hat{\rho}^2} \quad (8-71)$$

同时, 依据 (8-70) 有如下关系成立:

$$\frac{1}{NT - N - K} \sum_{i=1}^N \sum_{t=1}^T e_{it}^2 = \frac{\hat{\sigma}_v^2}{1 - \hat{\rho}^2} \quad (8-72)$$

<sup>35</sup>由于前面假设  $\alpha_i$  与所有  $\varepsilon_{it}$  均不相关, 所以二者的交乘项均为零。同时, 计算中要利用矩阵直乘的性质  $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ 。

<sup>36</sup>注意, 这里我们忽略模型中的个体效应和序列相关的设定。

因此，联立 (8-68)、(8-71) 和 (8-70) 三式，我们可以得到  $\hat{\rho}$ 、 $\hat{\sigma}_v^2$  和  $\hat{\sigma}_a^2$ ，并进而得到

$$\hat{\Omega} = (1 - \hat{\rho})^2 \hat{\sigma}_a^2 (\mathbf{I}_N \otimes \mathbf{J}_{T-1}) + \hat{\sigma}_v^2 (\mathbf{I}_N \otimes \mathbf{C}\mathbf{C}') \quad (8-73)$$

最终我们得到模型的 FGLS 估计量为：

$$\hat{\beta}_{FGLS} = [\mathbf{X}^{*'} \hat{\Omega}^{-1} \mathbf{X}^*]^{-1} \mathbf{X}^{*'} \hat{\Omega}^{-1} \mathbf{y}^* \quad (8-74)$$

和

$$\text{Var}[\hat{\beta}_{FGLS}] = [\mathbf{X}^{*'} \hat{\Omega}^{-1} \mathbf{X}^*]^{-1} \quad (8-75)$$

#### Prais-Winsten 估计

当  $T$  较小而  $N$  较大时，采用 Cochrane-Orcutt 的方法舍弃第一期观察值会在很大程度上影响估计结果。<sup>37</sup> lillard 和 willis (1978) 采用 Prais 和 Winsten (1956) 处理时间序列模型的方法将我们在 8.2.2 小节中介绍的随机效应模型扩展为允许干扰项服从自相关的情形。由于此时需要考虑第一期观察值，所以我们增加对于干扰项初始值的假设  $v_{i0} \sim (0, \sigma_\varepsilon^2)$ ，即要求干扰项是从一个平稳的均衡态开始的。

首先，我们采用 Prais-Winsten (PW) 转换矩阵

$$\mathbf{C} = \begin{bmatrix} \sqrt{1 - \rho^2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}_{T \times T} \quad (8-76)$$

去除模型中的存在的序列相关性。对于面板数据而言，该转换需要应用到  $N$  个截面中。转换后干扰项的向量形式为：<sup>38</sup>

$$\mathbf{u}^* = (\mathbf{I}_N \otimes \mathbf{C})\mathbf{u} = [\mathbf{I}_N \otimes (\mathbf{C}\mathbf{1}_T)]\mathbf{a} + (\mathbf{I}_N \otimes \mathbf{C})\boldsymbol{\varepsilon} \quad (8-77)$$

我们注意到  $\mathbf{C}\mathbf{1}_T = (1 - \rho)\mathbf{1}_T^a$ ，其中， $\mathbf{1}_T^a = (\delta, \mathbf{1}_{T-1}')'$ ，而  $\delta = \sqrt{(1 + \rho)/(1 - \rho)}$ 。因此，(8-77) 式可表示为：

$$\mathbf{u}^* = (1 - \rho)(\mathbf{I}_N \otimes \mathbf{1}_T^a)\mathbf{a} + (\mathbf{I}_N \otimes \mathbf{C})\boldsymbol{\varepsilon} \quad (8-78)$$

于是变换后干扰项的方差-协方差矩阵为：<sup>39</sup>

$$\boldsymbol{\Omega}^* = E[\mathbf{u}^* \mathbf{u}^{*'}] = \sigma_a^2 (1 - \rho)^2 [\mathbf{I}_N \otimes (\mathbf{1}_T^a \mathbf{1}_T^{a'})] + \sigma_v^2 \mathbf{I}_{NT} \quad (8-79)$$

<sup>37</sup>虽然 Cochrane-Orcutt 估计在大样本下能够得到一致性的估计量，但一般的面板数据 ( $T$  较小,  $N$  较大) 显然 无法满足获得渐进性估计量的条件。我们一般采用 Cochrane-Orcutt 估计多出于计算的方便。

<sup>38</sup>推导过程中需要注意， $(\mathbf{I}_N \otimes \mathbf{C})(\mathbf{I}_T \otimes \mathbf{1}_T) = \mathbf{I}_T \otimes \mathbf{C}\mathbf{1}_T$ 。

<sup>39</sup>我们可以证明  $E[(\mathbf{I}_N \otimes \mathbf{C})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\mathbf{I}_N \otimes \mathbf{C})'] = \sigma_v^2 \mathbf{I}_{NT}$ 。因为， $\mathbf{C}\boldsymbol{\varepsilon}_i = (\sqrt{1 - \rho^2}\varepsilon_{i1}v_{i1}v_{i2}, \dots, v_{iT})'$ ，而由 (8-70) 式可知  $\text{Var}[\varepsilon_{i1}] = \sigma_v^2/(1 - \rho^2)$ 。所以， $E[\mathbf{C}\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \mathbf{C}'] = \sigma_v^2 \mathbf{I}_T$ 。由此，我们就可以很轻易地得到上述结果。



得到  $\Omega^*$  后, 我们就可以进一步得到类似于 (8-66) 式和 (8-67) 式的 GLS 估计量, 并可以利用前面得到的  $\hat{\rho}$ 、 $\hat{\sigma}_v^2$  和  $\hat{\sigma}_\alpha^2$  来获得相应的 FGLS 估计量。

下面我们说明上述变换的方差分解过程, 并借此说明序列相关模型与前面提到的一般性随机效应模型之间的关系。

令  $d^2 = \mathbf{1}_T' \mathbf{1}_T$ ,  $\mathbf{J}_T^a = \mathbf{1}_T \mathbf{1}_T'$  及  $\bar{\mathbf{J}}_T^a = \mathbf{J}_T^a / d^2$ 。同时我们定义  $\mathbf{I}_T = \mathbf{E}_T^a + \bar{\mathbf{J}}_T^a$ , 其中,  $\mathbf{E}_T^a = \mathbf{I}_T - \bar{\mathbf{J}}_T^a$ 。那么 (8-79) 可重新表示为:

$$\Omega^* = \sigma_\tau^2 (\mathbf{I}_N \otimes \bar{\mathbf{J}}_T^a) + \sigma_v^2 (\mathbf{I}_N \otimes \mathbf{E}_T^a) \quad (8-80)$$

其中,  $\sigma_\tau^2 = d^2 \sigma_\alpha^2 (1 - \rho)^2 + \sigma_v^2$ 。因此,

$$\sigma_v^2 \Omega^{*-1/2} = (\sigma_v / \sigma_\tau) (\mathbf{I}_N \otimes \bar{\mathbf{J}}_T^a) + (\mathbf{I}_N \otimes \mathbf{E}_T^a) = \mathbf{I}_{NT} - \theta_\tau (\mathbf{I}_N \otimes \bar{\mathbf{J}}_T^a) \quad (8-81)$$

其中,  $\theta_\tau = 1 - (\sigma_v / \sigma_\tau)$ 。

对经过 PW 转换的观察值  $\mathbf{y}^* = (\mathbf{I}_N \otimes \mathbf{C})\mathbf{y}$  左乘  $\sigma_v^2 \Omega^{*-1/2}$  得到  $\mathbf{y}^{**} = \sigma_v^2 \Omega^{*-1/2} \mathbf{y}^*$ , 其特定元素为:

$$\mathbf{y}_i^{**} = (y_{i1}^* - \theta_\tau \delta b_i, y_{i2}^* \theta_\tau b_i, \dots, y_{iT}^* \theta_\tau b_i)' \quad (8-82)$$

其中,  $b_i = [(\delta y_{i1}^* + \sum_{t=2}^T y_{it}^*) / d^2]$  ( $i = 1, 2, \dots, n$ )。可以看出, 我们在 AR(1) 随机效应模型中对第一期观察值进行了非常特殊的处理。同时我们注意到:

1. 当  $\rho = 0$ , 即不存在序列相关时,  $\Omega^* = T \sigma_\alpha^2 (\mathbf{I}_N \otimes \mathbf{J}_T) + \sigma_v^2 (\mathbf{I}_N \otimes \mathbf{I}_T)$ , 等价于 (8-34) 式, 这表明此时模型转化为普通的随机效应模型。
2. 当  $\sigma_\alpha^2 = 0$ , 即不存在随机效应, 我们将得到  $\sigma_\tau^2 = \sigma_v^2$  及  $\theta_\tau = 0$ , 这表明  $\mathbf{y}_{it}^{**}$  将转化为仅作 PW 转换得到的  $y_{it}^*$ 。

作上述分解的另一个好处是使我们可以基于对  $\Omega^*$  的分解来估计方差成分  $\sigma_\alpha^2$  和  $\sigma_v^2$ 。可以证明:<sup>40</sup>

$$(\mathbf{I}_N \otimes \mathbf{E}_T^a) \mathbf{u}^* \sim (0, \sigma_v^2 [\mathbf{I}_N \otimes \mathbf{E}_T^a]) \quad (8-83)$$

和

$$(\mathbf{I}_N \otimes \bar{\mathbf{J}}_T^a) \mathbf{u}^* \sim (0, \sigma_\tau^2 [\mathbf{I}_N \otimes \bar{\mathbf{J}}_T^a]) \quad (8-84)$$

由此, 我们可以得到  $\sigma_v^2$  和  $\sigma_\tau^2$  的如下一致估计量:

$$\hat{\sigma}_v^2 = \frac{\mathbf{u}^{*'} (\mathbf{I}_N \otimes \mathbf{E}_T^a) \mathbf{u}^*}{N(T-1)} \quad \text{和} \quad \hat{\sigma}_\tau^2 = \frac{\mathbf{u}^{*'} (\mathbf{I}_N \otimes \bar{\mathbf{J}}_T^a) \mathbf{u}^*}{N} \quad (8-85)$$

<sup>40</sup>  $\text{Var}[(\mathbf{I}_N \otimes \mathbf{E}_T^a) \mathbf{u}^*] = (\mathbf{I}_N \otimes \mathbf{E}_T^a) \Omega^* (\mathbf{I}_N \otimes \mathbf{E}_T^a)' = \sigma_\alpha^2 (1 - \rho)^2 [\mathbf{I}_N \otimes (\mathbf{E}_T^a \bar{\mathbf{J}}_T^a \mathbf{E}_T^a)] + \sigma_v^2 [\mathbf{I}_N \otimes (\mathbf{E}_T^a \mathbf{I}_T \mathbf{E}_T^a)]$ 。由于  $\mathbf{E}_T^a$ 、 $\bar{\mathbf{J}}_T^a$  和  $\mathbf{J}_T^a$  均为幂等矩阵, 且  $\mathbf{E}_T^a \bar{\mathbf{J}}_T^a = (\mathbf{I}_T - \bar{\mathbf{J}}_T^a) \mathbf{J}_T^a = \mathbf{0}$ , 所以  $\text{Var}[(\mathbf{I}_N \otimes \mathbf{E}_T^a) \mathbf{u}^*] = \sigma_v^2 [\mathbf{I}_N \otimes \mathbf{E}_T^a]$ 。同理, 可得 (8-84) 式。

这里，我们先可以采用  $\rho$  的一致估计量 (如  $\tilde{\rho}$ ) 得到 PW 转换矩阵的一致估计  $\hat{\mathbf{C}}$ ，然后用  $(\mathbf{I}_N \otimes \hat{\mathbf{C}})\mathbf{y}$  对  $(\mathbf{I}_N \otimes \hat{\mathbf{C}})\mathbf{X}$  作 OLS 回归，得到的残差  $\hat{\mathbf{u}}^*$  便是 (8-85) 式中  $\mathbf{u}^*$  的一致估计量。这样我们便可得到  $\hat{\sigma}_v^2$  和  $\hat{\sigma}_\tau^2$  的估计值。

因此，AR(1) 随机效应的模型的估计过程可总结如下：<sup>41</sup>

1. 对原始模型进行 PW 转换，这类似于我们处理一般时间序列模型的方法；
2. 从经过 PW 转换后的观察值中减去“伪均值”，如 (8-82) 式的设定；
3. 对经过第二步转换后的模型进行 OLS 估计。

### 序列相关检验

#### 1. Wooldridge 检验

对于固定效应模型 (8-1)，其一阶差分的形式为：

$$\Delta y_{it} = \Delta \mathbf{x}_{it} \boldsymbol{\beta} + \Delta \varepsilon_{it} \quad (8-86)$$

若设定  $\varepsilon_{it} = \rho \varepsilon_{it-1} + u_{it}$ ，则  $\Delta \varepsilon_{it} = \rho \Delta \varepsilon_{it-1} + \Delta u_{it}$ 。那么序列相关的原假设为：

$$H_0 : \rho = 0 \quad v.s. \quad \rho \neq 0$$

在原假设  $H_0$  成立的情况下，易于证明有如下关系成立：

$$Corr(\Delta \varepsilon_{it}, \Delta \varepsilon_{it-1}) = -0.5 \quad (8-87)$$

即使存在序列相关，(8-86) 式的 OLS 估计量仍然是一致的，假设其残差估计值为  $\tilde{\varepsilon}_{it}$ ，设用  $\tilde{\varepsilon}_{it}$  对  $\tilde{\varepsilon}_{it-1}$  进行 OLS 回归得到的系数估计值为  $\hat{\theta}$ ，那么上述序列相关检验就转化为检验  $\hat{\theta}$  是否显著异于  $-0.5$ ，这采用一般的  $t$  检验即可完成。<sup>42</sup>

至于随机效应模型设定下的序列相关检验就要相对复杂一些，有兴趣的读者可以参考 Baltagi (2001, Section 5.2)，以及 Sosa-Escudero and Bera (2008)。

#### 2. Durbin-Waston 检验

Bhargava, Franzini and Narendranathan(1982) 将时间序列中广泛应用的 Durbin-Waston 检验扩展到了面板情形下。原假设为：

$$H_0 : \rho = 0 \quad v.s. \quad \rho > 0 \quad \text{or} \quad \rho < 0$$

检验统计量为：

$$DW_p = \frac{\sum_{i=1}^N \sum_{t=2}^T (e_{it} - e_{it-1})^2}{\sum_{i=1}^N \sum_{t=1}^T e_{it}^2} \quad (8-88)$$

<sup>41</sup> STATA 11 手册 [XT] `xtregar` (pp.485-486)，对这一转换过程进行了非常简洁明了的说明。

<sup>42</sup> 对于这部分内容的详细介绍，请参考 Wooldridge (2002, pp.282)。Drukker(2003) 详细介绍了这一检验的基本思路，并编写一个 STATA 命令 (`xtserial`) 来执行这一检验。

其中,  $e_{it}$  为固定效应模型 (8-3) 或 (8-5) 的残差。Bhargava et al. (1982) 沿袭 Durbin and Waston 的思路, 进一步推导出在不同  $N, T$  和  $K$  取值下,  $DW_p$  的临界值的上限和下限, 见表 8-4。

在时间序列分析中, D-W 统计量可能落入所谓的“不确定区域 (inconclusive region)”, 这是该统计量的主要局限。然而, Bhargava et al. (1982) 发现, 基于面板数据得到的  $DW_p$  统计量的不确定区域要窄得多, 在  $N$  较大的面板中尤其如此, 这从见表 8-4 中的结果可以得到清晰的印证。

对于一个  $N = 100, T = 6$ , 且包含三个解释变量 ( $K = 3$ ) 的模型而言, 若  $DW_p < 1.859$ , 则我们可以在 5% 显著水平上拒绝  $H_0: \rho = 0$ 。对于  $N$  非常大的面板而言, Bhargava et al. (1982) 建议了一个简单的判断准则: 若  $DW_p$  小于 2, 则认为存在显著为正的序列相关。虽然 Bhargava et al. (1982) 仅针对 FE 模型推导了  $DW_p$  统计量, 但由于组内估计量也是 RE 模型的一致估计量, 所以对于 RE 模型而言,  $DW_p$  统计量仍然适用。

表 8-4: 面板 Durbin-Waston 统计量的 5% 上限和下限 (Bhargava et al., 1982)

|          |         | $N = 100$ |       | $N = 500$ |       | $N = 1000$ |       |
|----------|---------|-----------|-------|-----------|-------|------------|-------|
|          |         | $d_L$     | $d_U$ | $d_L$     | $d_U$ | $d_L$      | $d_U$ |
| $T = 6$  | $K = 3$ | 1.859     | 1.880 | 1.939     | 1.943 | 1.957      | 1.959 |
|          | $K = 9$ | 1.839     | 1.902 | 1.935     | 1.947 | 1.954      | 1.961 |
| $T = 10$ | $K = 3$ | 1.891     | 1.904 | 1.952     | 1.954 | 1.967      | 1.968 |
|          | $K = 9$ | 1.878     | 1.916 | 1.949     | 1.957 | 1.965      | 1.970 |

Bhargava et al. (1982) 提出的  $DW_p$  统计量仅适用于平行面板数据, Baltagi and Wu (1999) 进一步将其扩展至非平行面板, 并提出另一种统计量。在随后将要介绍的 `xtregar` 命令中附加 `lbi` 选项可以很方便地计算出这两个统计量。

除了上述两种检验方法外, Arellano and Bond (1991) 还提出了一个基于 GMM 估计的更为一般化的序列相关检验方法。Roodman (2009, pp.119-120) 对这一检验方法进行较为详细的论述, 并编写了 `abar` 命令来执行这一检验。由于该统计量是在 GMM 估计的基础上提出来的, 而 OLS 又是 GMM 的特例, 因此, 该统计量不但适用于一般的线性回归模型, 是适用于 IV 估计活 GMM 估计。至于其用法, 我们会在后续章节中介绍。

对于 RE 模型, 序列相关检验相对复杂一些。Sosa-Escudero and Bera (2008) 非常详细地探讨了四种相关的检验方法, 并编写了 `xttest1` 命令来执行这些检验。

## ► Example

### 1. FE 模型的序列相关检验

对于固定效应模型, 可以采用第 38 页中介绍的 Wooldridge 检验法, 命令为 `xtserial`:

```
. xtserial market invest stock
```

```
Wooldridge test for autocorrelation in panel data
H0: no first order autocorrelation
      F( 1,      4) =      4.442
      Prob > F =      0.1028
```

这里  $p = 0.0448$ ，我们可以在 5% 的显著水平上拒绝不存在序列相关的原假设。考虑到本例中样本的时间跨度为 20 年，这个结论应该在预料之中。该检验的最后一步是用  $\tilde{e}_{it}$  对  $\tilde{e}_{it-1}$  进行 OLS 回归，因此，输入 `mat list e(b)` 命令可以得到  $\hat{\theta} = -0.319$ 。

我们尚可采用 Bhargava et al. (1982) 提出的  $DW_p$  统计量，即 (8-88) 式，来检验序列相关是否存在，以便得到更为稳健的结论。命令如下：

```
. qui xtreg market invest stock dunt*, fe // FE estimation
. predict e, e // e_it
. qui gen de = D.e // D.e_it
. egen sum_e_sq = sum(e^2) // DW_p 分母
. egen sum_de_sq = sum(de^2) // DW_p 分子
. gen dw = sum_de_sq/sum_e_sq // DW_p
. dis "D-W = " dw[1]
D-W = 1.4965359
```

这里得到的  $DW_p = 1.497$ ，明显小于表 8-4 中相应的临界值 1.878。因此，基于  $DW_p$  统计量的检验同样在 5% 水平上拒绝了原假设。我们亦可采用 `xtregar` 并附加 `lbi` 选项，更为快捷地计算出  $DW_p$  统计量：

```
. qui xtregar market invest stock dunt*, fe lbi // 注意附加 [lbi] 选项
. dis "D-W = " e(d1)
D-W = 1.4965359
```

## 2. RE 模型的序列相关检验

对于 RE 模型，可以采用 `xttest1` 命令来执行 Sosa-Escudero and Bera (2008) 检验：<sup>43</sup>

```
. qui xtreg market invest stock dunt*, re
. xttest1

Tests for the error component model:

      market[id,t] = Xb + u[id] + v[id,t]
      v[id,t] = lambda v[id, (t-1)] + e[id,t]

Estimated results:

      _____
      |              Var      sd = sqrt(Var)
      |_____
      | market      2018625      1420.783
      | e           88403.93      297.32798
```

<sup>43</sup>该命令需要下载，可以在命令窗口中输入 `net install sg164_1.pkg` 直接下载，亦可输入 `findit xttest1` 命令，然后依据弹出页面的提示下载之。

```

          u |      52550.39      229.23873

Tests:
  Random Effects, Two Sided:
    ALM(Var(u)=0)      =   163.26 Pr>chi2(1) =   0.0000

  Random Effects, One Sided:
    ALM(Var(u)=0)      =   12.78 Pr>N(0,1) =   0.0000

  Serial Correlation:
    ALM(lambda=0)      =   15.94 Pr>chi2(1) =   0.0001

  Joint Test:
    LM(Var(u)=0,lambda=0) =  234.95 Pr>chi2(2) =   0.0000

```

这里汇报了 4 个统计量，分别用于检验 RE 模型中随机效应 (单尾和双尾)、序列相关以及二者的联合显著性。检验结果表明存在随机效应和序列相关，而且对随机效应和序列相关的联合检验也非常显著。

### 3. 估计

上述结果表明，无论是 FE 还是 RE 模型，干扰项中都存在显著的序列相关。为此，我们进一步采用 `xtregar` 命令来估计模型 (8-61)。首先考虑固定效应模型：

```

. xtregar market invest stock dumt*, fe lbi

FE (within) regression with AR(1) disturbances   Number of obs   =       95
Group variable: id                               Number of groups =        5

R-sq:  within = 0.6588                           Obs per group: min =       19
       between = 0.6379                             avg =      19.0
       overall = 0.5554                             max =       19

                                           F(20,70)      =       6.76
corr(u_i, Xb) = 0.4674                         Prob > F      =      0.0000

```

| market           | Coef.     | Std. Err.                             | t     | P> t  | [95% Conf. Interval] |           |
|------------------|-----------|---------------------------------------|-------|-------|----------------------|-----------|
| invest           | 2.547768  | .4751971                              | 5.36  | 0.000 | 1.600017             | 3.495519  |
| stock            | -.7422344 | .2788278                              | -2.66 | 0.010 | -1.298339            | -.1861299 |
| (output omitted) |           |                                       |       |       |                      |           |
| _cons            | 1916.31   | 158.156                               | 12.12 | 0.000 | 1600.878             | 2231.742  |
| rho_ar           | .26817919 |                                       |       |       |                      |           |
| sigma_u          | 1145.2561 |                                       |       |       |                      |           |
| sigma_e          | 289.77582 |                                       |       |       |                      |           |
| rho_fov          | .93983152 | (fraction of variance because of u_i) |       |       |                      |           |

```

F test that all u_i=0:      F(4,70) =      74.38          Prob > F = 0.0000
modified Bhargava et al. Durbin-Watson = 1.4965359
Baltagi-Wu LBI = 1.6364756

```

这里，我们通过附加 `fe` 选项来估计固定效应模型，此时 STATA 会采用 Cochrane-Orcutt 转换来消除序列相关。由于样本中有 5 家公司，在完成转换后，观察值从最初的 100 减少为 95。由于附加了 `lbi` 选项，最后两行报告了 Bhargava et al. (1982) 提出的  $DW_p$  统计量，同时还报

告了 Baltagi and Wu (1999) 的修正统计量 Baltagi-Wu  $LBI = 1.636$ ，略高于前者。遗憾的是，Baltagi and Wu (1999) 并未提供该统计量的临界值，虽然他们提供了标准化后的统计量，但并不适用于  $N$  较小的面板数据。同时，我们也注意到，这里的估计结果与此前不考虑序列相关时得到的结果存在一定的差异。

若需估计包含 AR(1) 干扰项的 RE 模型，只需在 `xtregar` 命令后附加 `re` 选项即可。此时，STATA 会采用相对有效的 Prais-Winsten 转换来消除序列相关，命令如下：

```
. xtregar market invest stock dunt*, re lbi
RE GLS regression with AR(1) disturbances      Number of obs      =      100
Group variable: id                             Number of groups    =       5
R-sq:  within  = 0.6906                        Obs per group: min =       20
       between = 0.7060                        avg              =      20.0
       overall  = 0.6374                        max              =       20
Wald chi2(22)      =      141.60
corr(u_i, Xb)      = 0 (assumed)                Prob > chi2        =      0.0000
```

| market           | Coef.     | Std. Err.                               | z     | P> z  | [95% Conf. Interval] |           |
|------------------|-----------|-----------------------------------------|-------|-------|----------------------|-----------|
| invest           | 2.990253  | .5047607                                | 5.92  | 0.000 | 2.00094              | 3.979565  |
| stock            | -.6175912 | .3100871                                | -1.99 | 0.046 | -1.225351            | -.0098317 |
| (output omitted) |           |                                         |       |       |                      |           |
| _cons            | 897.5951  | 250.6101                                | 3.58  | 0.000 | 406.4084             | 1388.782  |
| rho_ar           | .26817919 | (estimated autocorrelation coefficient) |       |       |                      |           |
| sigma_u          | 514.27681 |                                         |       |       |                      |           |
| sigma_e          | 424.74901 |                                         |       |       |                      |           |
| rho_fov          | .59448231 | (fraction of variance due to u_i)       |       |       |                      |           |
| theta            | .7594229  |                                         |       |       |                      |           |

```
modified Bhargava et al. Durbin-Watson = 1.4965359
Baltagi-Wu LBI = 1.6364756
```

由于 Prais-Winsten 转换保留第一期观察值，因此这里参与回归的样本是 100。我们也注意到，此时得到的  $DW_p$  和  $LBI$  统计量与 `fe` 模型中完全相同，这是因为，这两个统计量的计算公式并不依赖于我们所估计的模型 (FE 或 RE)。

◀

### 8.4.3 方差形式未知时的稳健性估计

在上述分析中，我们通过特定的假设设定了异方差和序列相关的具体形式。然而，在实证分析过程中，经常需要作出如下权衡：若上述估计方法的前提假设是正确的，则相应的估计量较为有效；然而，当这些假设条件不满足，或曰我们错误地设定了干扰项的协方差矩阵，则上述估计量反而不及普通的 FE 或 RE 估计量。

由于在多数情况下，我们都不知道真实的数据生成过程 (GDP)，这就使得出现后一种偏误的可能性较大。为此，在多数文献中，学者们都沿袭了 White (1980) 和 Newey and West (1987)

的思路，仍然采用 Pooled OLS, FE 或 RE 得到系数估计值，但需要根据可能存在的异方差、序列相关或截面相关来调整标准误。

根据不同的面板数据结构，STATA 提供了多种获取稳健型标准误的估计命令，见表 8-5。总体而言，对于“大  $T$  小  $N$ ”型面板而言，xtgls, xtpscse, xtsur 都是不错的选择，而对于“大  $N$  小  $T$ ”型面板而言，则应使用 xtreg 命令附加 vce(robust) 或 vce(cluster) 选项，或使用 xtscs 命令。在公司金融领域，Fama and MacBeth (1973) 提出的两步估计法也得到了广泛的应用，使用 xtfmb 命令可以很方便地实现这一估计。此外，近十年来，得益于计算机技术的发展，以重复抽样为基础的 bootstrap 法也获得了快速的发展。采用该方法计算标准误的好处在于，无需预先假设干扰项的分布特征，而完全基于样本的特征，通过重复抽样重现其经验分布 (empirical distribution)。对于样本较大的面板数据而言，这是个不错的选择。

表 8-5: 面板模型中计算稳健型标准误的 STATA 命令

| Command    | Option              | SE estimates are robust to disturbances that are                                                  | Notes                                                                      |
|------------|---------------------|---------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| reg, xtreg | robust              | heteroscedastic                                                                                   |                                                                            |
| reg, xtreg | cluster()           | heteroscedastic and autocorrelated                                                                |                                                                            |
| xtregar    |                     | autocorrelated with AR(1) <sup>a</sup>                                                            |                                                                            |
| newey2     |                     | heteroscedastic and autocorrelated of type MA( $q$ ) <sup>b</sup>                                 | can perform IV regression                                                  |
| xtgls      | panels(),<br>corr() | heteroscedastic, contemporaneously cross-sectionally correlated, and autocorrelated of type AR(1) | $N < T$ required for feasibility; tends to produce optimistic SE estimates |
| xtpscse    | correlation()       | heteroscedastic, contemporaneously cross-sectionally correlated, and autocorrelated of type AR(1) | large-scale panel regressions with xtpscse take a lot of time              |
| xtscs      |                     | heteroscedastic, autocorrelated with MA( $q$ ), and cross-sectionally dependent                   |                                                                            |

<sup>a</sup> AR(1) refers to first-order autoregression

<sup>b</sup> MA( $q$ ) denotes autocorrelation of the moving average type with lag length  $q$ .

\* Source: Hoechle (2007), Table 1.

### “异方差-序列相关”稳健型标准误

当第 8.2.1 小节中的假设 1 成立，但假设 2 不成立时，<sup>44</sup>基于组内估计量的方差矩阵 (参见

<sup>44</sup>若使用 (8-6) 式中的符号，可表示为  $E(\hat{\epsilon}_i | \mathbf{x}_i) = 0$ ，但  $Var(\hat{\epsilon}_i | \mathbf{x}_i) \neq \sigma^2 \mathbf{1}_T$ 。

(8-11) 式) 得到的系数标准误将是非一致的。我们可以将这一估计式重新表述为:

$$\widehat{Var}(\hat{\beta}_{WG}) = \hat{\sigma}^2(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1} \quad (8-89)$$

其中,  $\dot{\mathbf{X}} = \mathbf{QX}$ ,  $\dot{\mathbf{y}} = \mathbf{Qy}$ ,  $\hat{\sigma}^2$  由 (8-12) 式给出。然而, 由于  $E(\dot{\mathbf{X}}_i'\dot{\varepsilon}_i) = 0$ , 且

$$\left(\frac{1}{N}\dot{\mathbf{X}}'\dot{\mathbf{X}}\right)\sqrt{N}(\hat{\beta}_{WG} - \beta) = \frac{1}{N}\sum_{i=1}^N\dot{\mathbf{X}}_i'\dot{\varepsilon}_i$$

上式右侧可以视为以均值为 0 的随机数 ( $\dot{\varepsilon}_i$ ) 为权重的样本加权平均, 当  $T$  固定, 而  $N$  趋于无穷时, 利用中央极限定理可得:

$$\frac{1}{N}\sum_{i=1}^N\dot{\mathbf{X}}_i'\dot{\varepsilon}_i \xrightarrow{d} N[0, E(\dot{\mathbf{X}}_i'\dot{\varepsilon}_i\dot{\varepsilon}_i'\dot{\mathbf{X}}_i)]$$

因此, 对于大  $N$  小  $T$  型面板, 组内估计量的“异方差-序列相关”稳健型方差-协方差渐进估计量为:

$$\begin{aligned} \widetilde{Var}(\hat{\beta}_{WG}) &= (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1} \left( \sum_{i=1}^N \dot{\mathbf{X}}_i' \dot{\varepsilon}_i \dot{\varepsilon}_i' \dot{\mathbf{X}}_i \right) (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1} \\ &= \left[ \sum_{i=1}^N \dot{\mathbf{X}}_i' \dot{\mathbf{X}}_i \right]^{-1} \left( \sum_{i=1}^N \dot{\mathbf{X}}_i' \dot{\varepsilon}_i \dot{\varepsilon}_i' \dot{\mathbf{X}}_i \right) \left[ \sum_{i=1}^N \dot{\mathbf{X}}_i' \dot{\mathbf{X}}_i \right]^{-1} \\ &= \left[ \sum_{i=1}^N \sum_{t=1}^T \dot{\mathbf{x}}_{it}' \dot{\mathbf{x}}_{it} \right]^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T \dot{e}_{it} \dot{e}_{is} \dot{\mathbf{x}}_{it}' \dot{\mathbf{x}}_{is}' \right) \left[ \sum_{i=1}^N \sum_{t=1}^T \dot{\mathbf{x}}_{it}' \dot{\mathbf{x}}_{it} \right]^{-1} \end{aligned} \quad (8-90)$$

其中,  $\dot{\varepsilon}_i = \dot{\mathbf{y}}_i - \dot{\mathbf{X}}_i \hat{\beta}_{WG}$  (Arellano, 1987), 即组内估计量的残差。然而, 对于大  $T$  小  $N$  型面板而言, 该估计量不再具有一致性, 此时要采用下一小节介绍的 Driscoll and Kraay (1998) 稳健型估计量。

对于随机效应模型而言, 若  $\alpha_i$  与  $\mathbf{X}_{it}$  不相关, 则只需采用 (8-39) 式转换原始变量, 进而用转换后的变量替代上述公式中相应的变量即可。

#### ► Example

要获得“异方差-序列相关”稳健型标准误, 只需在 `xtreg` 命令中附加 `vce(robust)` 或 `vce(cluster)` 选项即可。例如, 对于 FE 模型, 我们可以执行如下命令:

```
. xtreg market invest stock, fe vce(robust)
Fixed-effects (within) regression      Number of obs   =      100
Group variable: id                    Number of groups =       5
R-sq:  within  = 0.4168                Obs per group: min =      20
```



```

    between = 0.6960          avg =      20.0
    overall = 0.6324          max =       20
                                F(2,4)      =    38.64
                                Prob > F      =    0.0024

corr(u_i, Xb) = 0.5256
                                (Std. Err. adjusted for 5 clusters in id)

```

| market  | Coef.     | Robust<br>Std. Err.               | t     | P> t  | [95% Conf. Interval] |          |
|---------|-----------|-----------------------------------|-------|-------|----------------------|----------|
| invest  | 3.05273   | 1.13323                           | 2.69  | 0.054 | -.0936203            | 6.199081 |
| stock   | -.6763434 | .501297                           | -1.35 | 0.249 | -2.068167            | .7154801 |
| _cons   | 1372.613  | 130.4248                          | 10.52 | 0.000 | 1010.495             | 1734.73  |
| sigma_u | 1023.5914 |                                   |       |       |                      |          |
| sigma_e | 370.9569  |                                   |       |       |                      |          |
| rho     | .88390837 | (fraction of variance due to u_i) |       |       |                      |          |

相比于此前未附加 `vce(robust)` 选项时的结果 (第 21 页), 虽然系数的估计值未发生变化, 但此时得到的标准误明显增大了, 致使最终的统计推断也更为保守 (`invest` 仅在 10% 水平上显著, 而 `stock` 则不在显著)。同时, 表头最后一行增加了一条信息 “Std. Err. adjusted for 5 clusters in id”, 这意味着, 对于面板模型而言, STATA 在计算所谓的 “robust” 标准误时, 是以个体为单位调整标准误的。因此, 我们得到的 “robust” 标准误其实是同时调整了异方差和序列相关后的标准误。换言之, 上述结果与设定 `vce(cluster)` 选项的结果完全相同:

```
. xtreg market invest stock, fe vce(cluster id)
```

(output omitted)

(Std. Err. adjusted for 5 clusters in id)

| market | Coef.     | Robust<br>Std. Err. | t     | P> t  | [95% Conf. Interval] |          |
|--------|-----------|---------------------|-------|-------|----------------------|----------|
| invest | 3.05273   | 1.13323             | 2.69  | 0.054 | -.0936203            | 6.199081 |
| stock  | -.6763434 | .501297             | -1.35 | 0.249 | -2.068167            | .7154801 |
| _cons  | 1372.613  | 130.4248            | 10.52 | 0.000 | 1010.495             | 1734.73  |

(output omitted)

◀

### “异方差-序列相关-截面相关” 稳健型标准误

虽然上述估计方法在估计方差-协方差矩阵时考虑了异方差和序列相关的影响, 但都未考虑截面之间的相关性。在早期研究中, Parks (1967) 提出了一个同时考虑异方差和截面相关性的可行性广义最小二乘 (FGLS) 估计量。<sup>45</sup>然而, 该模型存在两个主要的局限: 其一, 虽然多数面板

<sup>45</sup>Kmenta (1986) 在其基础上进行了扩展。Greene (2000, Section 15.2.2, pp.599-601) 对这一方法的推导过程和相关的检验方法进行了较为详细的介绍。在 STATA 中, 要估计该模型, 可以使用 `xtgls` 命令并附加 `panels(correlated)` 选项。

数据都是“大  $N$  小  $T$ ”型结构，但该模型却仅适用于“大  $T$  小  $N$ ”型面板；其二，Beck and Katz (1995) 研究表明，该方法估得的标准误严重偏低。

鉴于 Park-Kmenta 方法的局限，Beck and Katz (1995) 建议仍然采用 Pooled OLS 估计系数，但标准误则采用经过截面相关调整后的标准误 (panel-corrected standard errors, PCSEs)。在 STATA 中，可以采用 `xtpcse` 命令估计这一模型。由于这一模型需要以  $N \times N$  维协方差矩阵的估计值为基础，因此，对于  $N$  较大的面板数据而言，其参数估计值将变得非常不准确。

由于多数面板数据都是“大  $T$  小  $N$ ”型的，截面相关系数的数目会以  $N^2$  的速度增长，而观察值的数目的增长速度则仅为  $N$ 。因此，为了保证模型可以估计，多数研究都会假设样本中的所有个体之间都具有相同的截面相关性，此时，在模型中附加  $T - 1$  个时间虚拟变量即可控制截面相关性。然而，这一处理方法的假设条件在多数情况下都显得过于严格了。

为此，Driscoll and Kraay (1998) 在第五章中介绍的 Newey and West (1987) 估计量的基础上，提出了一个基于非参数方法的协方差估计量，在考虑序列相关的基础上，进一步控制了截面相关的影响。<sup>46</sup>通过这种方法获取稳健型标准误的好处在于，所得到的干扰项的方差-协方差矩阵的估计量并不依赖于截面数目  $N$ 。这使得该方法能够有效克服上文提到的 Parks (1967)、Kmenta (1986)、Beck and Katz (1995) 估计量的缺陷 (只有在  $T$  远大于  $N$  的情况下，这些估计量才是一致的)。

Driscoll and Kraay (1998) 最初提出的估计量是仅适用于平行面板，Hoechle (2007) 进一步将其扩展为非平行面板。当第 8.2.1 小节中的假设 1 成立，但假设 2 不成立时，组内估计量  $\hat{\beta}_{WG}$  仍然是  $\beta$  的无偏估计量，但 Driscoll and Kraay (1998) 建议采用如下“异方差-序列相关-截面相关”稳健型协方差矩阵来估计系数的标准误：

$$\widehat{Var}(\hat{\beta}_{WG}) = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{S}}_T(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \quad (8-91)$$

其中， $\hat{\mathbf{S}}_T$  具有与 Newey and West (1987) 相似的形式：

$$\hat{\mathbf{S}}_T = \hat{\Omega}_0 + \sum_{j=1}^{m(T)} w(j, m) [\hat{\Omega}_j + \hat{\Omega}_j'] \quad (8-92)$$

其中， $m(T)$  表示残差可能存在序列相关的滞后阶数，<sup>47</sup> $w(j, m)$  则表示修正后的 Bartlett 权重，定义为  $w(j, m) = 1 - j/(m(T) + 1)$ ，其作用在于保证  $\hat{\mathbf{S}}_T$  为半正定矩阵，同时对高阶滞后项赋以较小的权重。 $(K + 1) \times (K + 1)$  维矩阵  $\hat{\Omega}_j$  定义为 ( $K$  表示  $x_{it}$  中非常数项解释变量的个数)：

$$\hat{\Omega}_j = \sum_{t=j+1}^T \mathbf{h}_t(\hat{\beta}_{WG}) \mathbf{h}_{t-j}(\hat{\beta}_{WG})' \quad \text{其中} \quad \mathbf{h}_t(\hat{\beta}_{WG}) = \sum_{i=1}^{N(t)} \mathbf{h}_{it}(\hat{\beta}_{WG}) \quad (8-93)$$

<sup>46</sup>Arellano (2003, pp.19) 也对这一估计方法进行了较为详细的介绍。

<sup>47</sup>在随后要介绍的 `xtscc` 命令中， $m(T)$  的默认值为  $m(T) = \text{floor}[4(T/100)^{2/9}]$ 。更为详细的介绍请参见 Hoechle (2007, p.259)。

其中,  $\mathbf{h}_{it}(\hat{\boldsymbol{\beta}}_{WG})$  表示第  $t$  年的矩条件, 而  $\mathbf{h}_t(\hat{\boldsymbol{\beta}}_{WG})$  则表示第  $t$  年所有个体 (从 1 到  $N(t)$ ) 的矩条件之加总,  $N(t)$  表示第  $t$  年的公司数目。这意味着 (8-93) 适用于非平行面板。采用组内估计量获得  $\boldsymbol{\beta}$  的估计值后, (8-93) 式中的矩条件  $\mathbf{h}_{it}(\hat{\boldsymbol{\beta}}_{WG})$  可基于组内估计的残差 ( $\hat{e}_{it}$ ) 表示为如下  $(K+1) \times 1$  维矩条件:

$$\mathbf{h}_{it}(\hat{\boldsymbol{\beta}}_{WG}) = \mathbf{x}_{it}\hat{e}_{it} = \mathbf{x}_{it}(y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{WG})$$

由 (8-92) 和 (8-93) 可以看出, Driscoll and Kraay (1998) 将同一个年度 ( $t$ ) 内的所有个体的矩条件  $\mathbf{h}_{it}(\hat{\boldsymbol{\beta}})$  进行平均化处理, 从而将 Newey and West (1987) 基于时间序列提出的异方差-序列相关稳健型方差矩阵估计量适用于面板数据。由于  $\mathbf{h}_t(\hat{\boldsymbol{\beta}})$  定义为所有个体的平均值, 经由上述方法得到的标准误始终是一致的, 与  $N$  的大小无关。更为重要的是, 通过这种方法得到的标准误对于一般形式的截面相关都是稳健的。

显然, 在上述推导过程中, 若不对模型进行组内变换, 我们便可获得针对 Pooled OLS 估计量的 Driscoll and Kraay (1998) “异方差-序列相关-截面相关” 稳健型标准误。

有关截面相关检验的介绍, 请参见 Sarafidis and De Hoyos (2006), 以及 Greene(2000, pp.598-601)。

## ► Example

### 1. 截面相关检验

对于 FE 模型, 可以利用 Baum(2001) 编写的 xttest2 命令来检验截面相关性:<sup>48</sup>

```
. qui xtreg market invest stock, fe
. xttest2

Correlation matrix of residuals:
      ___e1      ___e2      ___e3      ___e4      ___e5
___e1  1.0000
___e2  0.7939  1.0000
___e3  0.6092  0.5348  1.0000
___e4  0.2504  0.4066  0.7326  1.0000
___e5  0.3103  0.1165  0.3728 -0.1097  1.0000

Breusch-Pagan LM test of independence: chi2(10) =    46.258, Pr = 0.0000
Based on 20 complete observations
```

显然, 截面之间不存在相关性的原假设被高度拒绝了, 为此, 在计算系数标准误时, 我们需要考虑这一问题, 其处理方式将在下一小节中介绍。

对于 RE 模型而言, 我们可以采用 Sarafidis and De Hoyos (2006) 编写的 xtcsd 命令来检验截面相关性是否显著:<sup>49</sup>

<sup>48</sup>下载方式为: `ssc install xttest2, replace`。由于该方法基于 SUR 估计进行检验, 因此要求面板资料为  $N < T$  型结构。

<sup>49</sup>下载方式为: `ssc install xtcsd, replace`。

```
. qui xtreg market invest stock, re // or `fe' for FE model
. xtcsd, pesaran // Pesaran's Test

Pesaran's test of cross sectional independence =      4.385, Pr = 0.0000

. xtcsd, frees // Frees's Test

Frees' test of cross sectional independence =      0.508
|-----|
Critical values from Frees' Q distribution
          alpha = 0.10 :    0.1294
          alpha = 0.05 :    0.1695
          alpha = 0.01 :    0.2468
```

这里，我们首先估计了 RE 模型，进而在 `xtcsd` 命令中分别附加 `pesaran` 和 `frees` 选项，采用两种不同的方法检验了截面相关性。

## 2. 获取稳健型标准误

对于 FE 模型，在确认存在截面相关的情况下，我们可以采用 [Hoechle \(2007\)](#) 编写的 `xtscc` 命令获取 [Driscoll and Kraay \(1998\)](#) 提出的“异方差-序列相关-截面相关”稳健型标准误：

```
. xtscc market invest stock, fe

Regression with Driscoll-Kraay standard errors   Number of obs   =      100
Method: Fixed-effects regression                 Number of groups =       5
Group variable (i): id                          F( 2, 4)        =     51.52
maximum lag: 2                                  Prob > F        =     0.0014
                                                within R-squared =     0.4168
```

| market | Drisc/Kraay |           |       | P> t  | [95% Conf. Interval] |          |
|--------|-------------|-----------|-------|-------|----------------------|----------|
|        | Coef.       | Std. Err. | t     |       |                      |          |
| invest | 3.05273     | .5832634  | 5.23  | 0.006 | 1.433331             | 4.672129 |
| stock  | -.6763434   | .3666318  | -1.84 | 0.139 | -1.694276            | .3415896 |
| _cons  | 1372.613    | 102.5325  | 13.39 | 0.000 | 1087.937             | 1657.289 |

这里，`xtscc` 根据脚注 47 中的公式自动选择的滞后阶数为 2。系数估计值和  $\text{Within-}R^2$  与 `xtreg, fe` 的结果完全相同，但标准误存在较大差异。可见，在本例中，截面相关对统计推断有较大的影响。

若读者有更好的方法来确定自相关的滞后阶数，则可以通过 `lag()` 选项设定之。当然，在多数情况下，这很难做到。不过，我们可以通过附加 `lag(0)` 选项，来估计仅考虑异方差和截面相关的稳健型标准误，命令如下：

```
xtscc market invest stock, fe lag(0)
```

对于 RE 模型而言，若希望计算 [Driscoll and Kraay \(1998\)](#) 稳健型标准误，可以先采用

`xtdata`, `re` 命令进行 (8-39) 式的 RE 转换, 进而对转换后的数据执行 `xtsc` 命令即可。<sup>50</sup>

◀

### 采用 Bootstrap 获取标准误

Bootstrap 由 Efron (1979) 提出,<sup>51</sup>其基本思想在于, 通过从原始样本中执行多次可重复抽样来计算统计量的标准误。相对于传统的统计推断方法 (通常会假设干扰项的分布特征), 该方法只需要假设观测样本是从母体中随机抽取的即可。

为了说明 Bootstrap 的基本原理, 考虑如下简单线性回归模型:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, 2, \dots, N)$$

只要  $x_i$  是严格外生的,  $\beta$  的 OLS 估计量  $\hat{\beta}$  就是无偏且一致的。然而, 在无法确知干扰项  $\varepsilon_i$  的分布特征或样本数较小的情况下, 基于渐进分布得到的标准误往往缺乏有效性, 致使传统的统计推断方法可能变得非常不准确。<sup>52</sup>

此时, 我们可以采用如下三个步骤计算出  $\hat{\beta}$  的 Bootstrap 标准误:

(1) 从原始样本中可重复地 (with replacement) 抽取  $N$  个观察值, 称之为经验样本 (empirical sample)。由于是可重复抽样, 有些观察值会被抽中 1 次, 有些被抽中 2 次, 而有些则可能一次都未被抽中;

(2) 采用第一步得到的经验样本估计上述模型, 得到  $\beta$  的 OLS 估计值  $\hat{\beta}_1^{bs}$ ;

(3) 将上述两个步骤进行  $k$  次, 可以得到  $\hat{\beta}_1^{bs}, \hat{\beta}_2^{bs}, \dots, \hat{\beta}_k^{bs}$ 。

完成上述估计后, 我们得到了  $k$  个  $\beta$  的经验估计, 它们都是无偏且一致估计量。在计算  $\beta$  的标准误之前, 我们需要简单回顾一下“标准误”这个看似简单的概念。对于随机数  $z$  而言, 我们往往采用标准差 (standard deviation) 来描述其波动性, 而对于统计量 (如这里的  $\hat{\beta}$ ), 我们同样可以采用标准差来衡量其准确程度, 并称之为标准误 (standard error)。换言之, 标准误其实就是统计量的标准差。在传统的统计推断中, 由于我们只能观察到从母体中随机抽取的  $N$  个特定的观察值, 由此只能计算出一个统计量, 这使得我们无法计算“统计量的标准差 (这需要多个统计量的观察值)”, 而只能依据特定假设条件推导出该统计量的分布特征。<sup>53</sup>然而, 在 Bootstrap 中, 由于可以通过重复抽样来获得  $k$  个经验样本, 并进而得到  $k$  个  $\beta$  的经验估计值  $\hat{\beta}_1^{bs}, \hat{\beta}_2^{bs}, \dots, \hat{\beta}_k^{bs}$ , 从而使得我们能够计算出“统计量的标准差”, 即所谓的标准误。显然, 采用这种方法,

<sup>50</sup>需要说明的是, 这种处理方法得到的估计结果存在微小的偏差。详情请参阅 [XT] `xtdata`, pp.63。

<sup>51</sup>Efron and Tibshirani (1993) 非常详细地介绍了 Bootstrap 的相关理论和应用, Cameron and Trivedi (2009, chap. 13) 介绍了 STATA 中的相关操作方法。

<sup>52</sup>回顾一下我们在 OLS 部分的介绍中是如何获得  $\text{s.e.}(\hat{\beta})$  的。在小样本下, 我们往往假设  $\varepsilon_i \sim N(0, \sigma^2)$ , 进而得到  $\text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ , 由此计算出  $\text{s.e.}(\hat{\beta})$  后, 进一步计算出  $t = \hat{\beta}/\text{s.e.}(\hat{\beta})$ 。在  $\varepsilon_i \sim N(0, \sigma^2)$  的假设下, 该统计量服从  $t$  分布。在大样本下, 可以采用中央极限定理, 因此上述推断无需假设  $\varepsilon_i$  服从正态分布。显然, 当  $\varepsilon_i$  的分布特征与上述假设不一致时,  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  不再是  $\text{Var}(\hat{\beta})$  的有效估计量。

<sup>53</sup>如, 在 OLS 中, 假设  $\varepsilon_i \sim N(0, \sigma^2)$ , 我们便可推导出  $\hat{\beta} \sim N(\beta, \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1})$ 。

仅需假设  $x_i$  是严格外生的, 以保证  $\hat{\beta}_j^{bs}$ , ( $j = 1, 2, \dots, k$ ) 是  $\beta$  的无偏估计量即可。换言之, 采用这种方法计算出的标准误将是“异方差-序列相关”稳健型估计量。

基于上述分析, 我们可以计算出  $\{\hat{\beta}_1^{bs}, \hat{\beta}_2^{bs}, \dots, \hat{\beta}_k^{bs}\}$  的样本标准差, 从而得到  $\hat{\beta}$  的 Bootstrap 标准误:

$$\widehat{se}_{bs} = \left\{ \frac{1}{k-1} \sum_{j=1}^k (\hat{\beta}_j^{bs} - \bar{\beta}^{bs})^2 \right\}^{1/2} \quad (8-94)$$

其中,  $\bar{\beta}^{bs}$  是  $\hat{\beta}_j^{bs}$  的样本平均值。

STATA 中的多数回归分析命令都支持 `vce(bootstrap)` 选项, 以便计算出 Bootstrap 稳健型标准误。<sup>54</sup>对于面板数据模型而言, 此前介绍的多数命令 (`xtreg`, `xtregar`, `xtgls` 等) 都支持这一选项。需要说明的是, 在针对面板数据进行抽样时, 为了保持每个截面内的时序特征, Bootstrap 是以个体(公司、国家或地区)为单位进行抽样的。换言之, 我们可以把面板中的每家公司视为一个“数据块(block)”, 而 Bootstrap 的抽样单位是这些数据块, 而非单个的观察值。<sup>55</sup>

Bootstrap 的精度决定于抽样的次数 ( $k$ ) 和原始样本 ( $N$ ) 的大小。STATA 默认的抽样次数为  $k = 50$ , 多数情况下都能够提供比较可信的统计推断。在模型筛选阶段, 这有助于节省抽样时间。Efron and Tibshirani (1993) 建议, 若采用 Bootstrap 获取标准误, 抽样 50-200 次已经可以达到比较稳定的效果。但在样本较大的情况下, 尤其是在报告论文的最终结果时, 选择 500 或 1000 次抽样可能更为稳妥, 这也是多数文献中的做法。

### ► Example

对于 FE 模型, 我们可以执行如下命令来获取 Bootstrap 稳健型标准误:

```
. xtreg market invest stock, fe vce(bootstrap, reps(200) seed(123))
(running xtreg on estimation sample)

Bootstrap replications (200)
-----|-----|-----|-----|-----|-----|
| 1 | 2 | 3 | 4 | 5 |
|-----|-----|-----|-----|-----|
..... 50
..... 100
..... 150
..... 200

Fixed-effects (within) regression               Number of obs   =       100
Group variable: id                             Number of groups =        5

R-sq:  within = 0.4168                          Obs per group:  min =       20
        between = 0.6960                             avg   =      20.0
        overall = 0.6324                             max   =       20

                                           Wald chi2(2)      =       6.09
```

<sup>54</sup>需要说明的是, 设定 `vce(bootstrap)` 选项只会改变标准误, 而系数估计值则仍然采用相应模型的估计公式。从直觉上我们可能认为  $\bar{\beta}^{bs}$  是一个优于  $\hat{\beta}_{OLS}$  的估计量, 但事实并非如此。详情请参阅 [U] **bootstrap**, pp.202。

<sup>55</sup>参见 Cameron and Trivedi(2005, section 11.6.2)。

corr(u\_i, Xb) = 0.5256

Prob > chi2 = 0.0475

(Replications based on 5 clusters in id)

| market  | Observed<br>Coef. | Bootstrap<br>Std. Err.            | z     | P> z  | Normal-based<br>[95% Conf. Interval] |          |
|---------|-------------------|-----------------------------------|-------|-------|--------------------------------------|----------|
| invest  | 3.05273           | 1.265414                          | 2.41  | 0.016 | .5725642                             | 5.532896 |
| stock   | -.6763434         | .4211295                          | -1.61 | 0.108 | -1.501742                            | .1490553 |
| _cons   | 1372.613          | 364.8994                          | 3.76  | 0.000 | 657.4229                             | 2087.802 |
| sigma_u | 1023.5914         |                                   |       |       |                                      |          |
| sigma_e | 370.9569          |                                   |       |       |                                      |          |
| rho     | .88390837         | (fraction of variance due to u_i) |       |       |                                      |          |

在上述命令中，我们通过 `reps()` 选项将抽样的次数设定为 200。为了使估计过程不至于太枯燥，<sup>56</sup>每完成一次抽样和相应的估计，STATA 会在屏幕上打印一个点，为此，本例中共打印出 200 个点。设定 `nodots` 选项可以避免在屏幕上打点。

这里，我们还通过附加 `seed()` 选项将 Bootstrap 的种子值 (`seed`) 设定为 135，其作用在于保证每次执行上述命令得到的结果都相同。<sup>57</sup>在撰写学术论文过程中，为了保证日后修改时结果不因随机抽样而发生变化，或保证其他学者可以验证我们的结论，设定这一选项显得尤为重要。

相比于不附加 `vce(bootstrap)` 选项的结果而言，此时得到的标准误较大，从而使统计推断更为保守 ( $t$  值较小)。表头下方的 “(Replications based on 5 clusters in id)” 信息表明，对于面板模型而言，STATA 是以公司为单位进行抽样的，这有助于保持截面内的时序特征。<sup>58</sup>

◀

2. 采用 Bootstrap 执行 Hausman 检验<sup>59</sup>

传统的 Hausman 检验 (第 8.2.3 小节) 要求在原假设下，RE 模型是完全有效的估计量 (fully efficient)，采用 Bootstrap 可以在该假设不满足的情况下执行 Hausman 检验。

► Example

首先，我们需要编写一个简单的程序，以便得到每一轮抽样过程中重点关注的统计量：FE

<sup>56</sup>在有些情况下，Bootstrap 会非常耗时，这种看似“单调”的娱乐方式的确有助于缓解你的急躁心情。

<sup>57</sup>Bootstrap 抽样是一个随机的过程，若不附加 `seed()` 选项，则每次得到的标准误都不会完全相同。在每一轮 Bootstrap 抽样过程中，STATA 都会调用一个所谓的随机数发生器函数，并根据这些随机数从原始样本中抽样。种子值其实就是这个随机数发生器的初始值。详情请参阅 STATA 11 手册 [D] **functions** (pp.223-224)。种子值可以设定为任意正整数，选取不同的种子值仅会导致标准误的微小变动，但统计推断的结论通常不会发生实质性的变化。

<sup>58</sup>显然，采用这种方法得到的标准误将是“异方差-截面相关”稳健型估计量。若希望进一步考虑序列相关，则可以以单个观察值为抽样单位，但同时考虑个体效应，可以采用第 23 页中介绍的 LSDV 法或 `areg` 命令，并附加 `vce(bootstrap)` 选项。

<sup>59</sup>这一小节的分析深受 Cameron (2009, p.430) 的启发。有关这一部分的理论分析，请参见 Cameron (2005, section 21.4.3, pp.717-718)。



和 RE 估计系数的差异：

```
. * Program to return (b1-b2) for Hausman test of FE v.s. RE
. cap program drop Hausman_FE_RE
. program define Hausman_FE_RE, eclass
1.   version 10
2.   tempname b bfe bre
3.   xtreg market invest stock, fe
4.   matrix `bfe' = e(b)           // FE coefficient
5.   xtreg market invest stock, re
6.   matrix `bre' = e(b)           // RE coefficient
7.   matrix `b' = `bfe' - `bre'    // Difference between FE and RE
8.   ereturn post `b'
9.   end
```

我们将该程序命名为 Hausman\_FE\_RE，并定义为 eclass 类型，以便将该程序的计算结果传递给后续程序。这主要通过最后一行的 ereturn post 命令来实现的。

接下来，就可以执行 Bootstrap 估计了：<sup>60</sup>

```
. * Bootstrap estimates for robust Hausman test
. bootstrap _b, reps(500) seed(135) nodots nowarn: Hausman_FE_RE

Bootstrap results                                Number of obs    =      100
                                                Replications      =      500
```

|        | Observed<br>Coef. | Bootstrap<br>Std. Err. | z     | P> z  | Normal-based<br>[95% Conf. Interval] |          |
|--------|-------------------|------------------------|-------|-------|--------------------------------------|----------|
| invest | -.794284          | .4802484               | -1.65 | 0.098 | -1.735554                            | .1469854 |
| stock  | .1218184          | .2707035               | 0.45  | 0.653 | -.4087508                            | .6523875 |
| _cons  | 159.8489          | 165.7438               | 0.96  | 0.335 | -165.0031                            | 484.7008 |

这里，reps() 选项用于设定 Bootstrap 的次数，nodots 选项用于告知 STATA 无需在屏幕上打点，而 nowarn 则可以屏蔽所有警示信息。需要注意的是，根据 Hausman\_FE\_RE 的定义，这里报告的系数估计值其实是 FE 和 RE 估计的系数差异，显然，除了 invest 变量的系数差异在 10% 水平上显著异于零外，其他两个变量并不存在显著差异。

我们可以进一步执行如下 Wald 检验，以便确认所有变量是否在 FE 和 RE 估计之间存在显著差异，这其实就是稳健型 Hausman 检验：

```
. * Perform robust Hausman test for FE v.s. RE
. test invest = stock = _cons = 0

( 1)  invest - stock = 0
( 2)  invest - _cons = 0
( 3)  invest = 0
```

<sup>60</sup>在执行 Bootstrap 命令之前，需要先将上述 Hausman\_FE\_RE 程序另存为一个同名的 ado 文件，并保存到 personal 文件夹下。另一种方式是，选中上面定义的 Hausman\_FE\_RE 程序代码，执行快捷键 Ctrl+R，以便将 Hausman\_FE\_RE 程序定义到内存中。



```
chi2( 3) =      2.90
Prob > chi2 =    0.4080
```

可见，若依据 Bootstrap 的检验结果，我们无法拒绝 Hausman 检验的原假设，应该选择相对更为有效的 RE 模型。

◀

### Fama-MacBeth 两步法

在公司财务领域，Fama and MacBeth (1973) 两步估计法具有非常广泛的应用。<sup>61</sup>

Fama-MacBeth 估计法分为两步：(1) 在各个年度上分别针对所有样本公司执行 OLS (截面) 回归，得到分年度系数估计值  $\hat{\beta}_t$  ( $t = 1, 2, \dots, T$ )；(2) 计算上述  $T$  次回归的平均系数，得到全样本的系数估计值，即

$$\begin{aligned}\hat{\beta}_{FM} &= \sum_{t=1}^T \frac{\hat{\beta}_t}{T} \\ &= \frac{1}{T} \sum_{t=1}^T \left( \frac{\sum_{i=1}^N X_{it} y_{it}}{\sum_{i=1}^N X_{it}^2} \right) = \beta + \frac{1}{T} \sum_{t=1}^T \left( \frac{\sum_{i=1}^N X_{it} \varepsilon_{it}}{\sum_{i=1}^N X_{it}^2} \right)\end{aligned}\quad (8-95)$$

进一步可以计算出其标准误：

$$s.e.(\hat{\beta}_{FM}) = \sqrt{\frac{1}{T} \sum_{t=1}^T \frac{(\hat{\beta}_t - \hat{\beta}_{FM})^2}{T-1}} \quad (8-96)$$

由 (8-95) 式可知，在  $X_{it}$  严格外生的假设下， $E(X_{it}\varepsilon_{it}) = 0$ ，此时  $\hat{\beta}_{FM}$  是其真实值的无偏估计量。对于其标准误而言，(8-96) 成立的前提是  $\beta_t$  彼此独立，二者要要求  $\text{Corr}(X_{it}\varepsilon_{it}, X_{is}\varepsilon_{is}) \neq 0$  ( $t \neq s$ )。由于 Fama-MacBeth 两步法并未考虑个体效应，当个体效应对  $y_{it}$  有显著影响时，将主要反映在干扰项  $\varepsilon_{it}$  中。此时，上述独立性假设不再成立，这会导致  $s.e.(\hat{\beta}_{FM})$  严重偏低 (Petersen, 2009, pp.446)，进而导致过度拒绝原假设 (因为此时计算出来的  $t$  值偏大)。

### ► Example

我们可以采用 STATA 用户编写的 `xtfmb` 命令实现 Fama-MacBeth 两步法：

```
. xtfmb market invest stock
Fama-MacBeth (1973) Two-Step procedure      Number of obs      =      100
                                           Num. time periods =      20
```

<sup>61</sup>虽然，Fama and MacBeth (1973) 提出这一方法的初衷是为了克服干扰项的序列相关问题，但 Petersen (2009) 在对比了该方法与此前介绍的 Pooled OLS、FE，以及考虑异方差、序列相关和截面相关的稳健型估计量的统计性质后，发现在多数情况下，该方法都存在比较严重的偏误。因此，就笔者的观点来看，使用该方法只能作为与前期相关研究进行对比分析时使用。

|        |              |           |      | F( 2, 19)      | =                    | 28.59    |
|--------|--------------|-----------|------|----------------|----------------------|----------|
|        |              |           |      | Prob > F       | =                    | 0.0000   |
|        |              |           |      | avg. R-squared | =                    | 0.8954   |
| market | Fama-MacBeth |           |      |                |                      |          |
|        | Coef.        | Std. Err. | t    | P> t           | [95% Conf. Interval] |          |
| invest | 3.61214      | .6539737  | 5.52 | 0.000          | 2.243357             | 4.980922 |
| stock  | 4.91288      | 1.334844  | 3.68 | 0.002          | 2.119019             | 7.70674  |
| _cons  | 185.2234     | 36.44294  | 5.08 | 0.000          | 108.9475             | 261.4994 |

对比此前的 FE 估计，可以发现此时 stock 变量的系数显著为负，这是因为，Fama-MacBeth 两步法主要反映了数据的横截面特征，而 FE 模型则重点关注组内差异。从这个意义上来讲，Fama-MacBeth 两步法的估计结果与组间效应模型的估计结果更加接近。<sup>62</sup>同时，我们也发现，采用 Fama-MacBeth 两步法得到的  $R^2$  远高于 FE 或 RE 估计，甚至明显高于 Pooled OLS。当然，在本例中，每个年度上只有  $N = 5$  个观察值，这可能导致 Fama-MacBeth 两步法的估计结果很不准确。

为了使大家更为深入地理解 Fama-MacBeth 两步法的估计过程，我们可以通过如下简单程序来估计其系数：

```
. qui tsset id year
. egen t = group(year) // t=1,2,3,...,T
. qui tsset id t
. qui sum t
. local T = r(max)
. qui gen b_cons = . // 用于存储各年度的系数估计值
. qui gen b_invest = .
. qui gen b_stock = .
. qui gen r2 = . // 用于存储各年度的 R-sq
. forvalues t = 1/`T'{ // 分年度估计
2.   qui{
3.       reg market invest stock if (t==`t')
4.       replace b_cons = _b[_cons] in `t'
5.       replace b_invest = _b[invest] in `t'
6.       replace b_stock = _b[stock] in `t'
7.       replace r2 = e(r2) in `t'
8.   }
9. }
. sum b_* r2 // 呈现估计结果
```

| Variable | Obs | Mean     | Std. Dev. | Min       | Max      |
|----------|-----|----------|-----------|-----------|----------|
| b_cons   | 20  | 185.2234 | 162.9778  | -15.99471 | 682.8146 |

<sup>62</sup>STATA 命令为：xtreg market invest stock, be。

|          |    |          |          |           |          |
|----------|----|----------|----------|-----------|----------|
| b_invest | 20 | 3.61214  | 2.924659 | -.6602821 | 9.543836 |
| b_stock  | 20 | 4.91288  | 5.969603 | -.2354539 | 26.91374 |
| r2       | 20 | .8954157 | .1047199 | .6821356  | .9991304 |

当然，我们也可以采用更为简洁的 `statsby` 命令来完成上述操作：

```
. preserve
. qui statsby _b e(r2), by(year) clear: reg market invest stock
. sum _b* _eq2
```

| Variable    | Obs | Mean     | Std. Dev. | Min       | Max      |
|-------------|-----|----------|-----------|-----------|----------|
| _b_invest   | 20  | 3.61214  | 2.924659  | -.6602821 | 9.543836 |
| _b_stock    | 20  | 4.91288  | 5.969603  | -.2354539 | 26.91374 |
| _b_cons     | 20  | 185.2234 | 162.9778  | -15.99471 | 682.8146 |
| _eq2_stat_1 | 20  | .8954157 | .1047198  | .6821356  | .9991304 |

```
. restore
```

执行 `statsby` 命令后，内存中现有的数据会被自动清除，代之以各个年度的平均估计系数估计值。为了避免当前数据被篡改，我们首先使用了 `preserve` 命令，以便在执行 `statsby` 命令之前备份当前数据，待执行完 `statsby` 命令，并采用 `sum` 命令在屏幕上呈现出所需的估计结果后，我们进一步采用 `restore` 命令重新调入此前执行 `preserve` 命令时自动备份的数据。

◀

## 8.5 内生性问题与 IV/GMM 估计

参见 Cameron (2005)

Hausman-Taylor estimator [Sun, 2004, pp.37]

Sun(2004) An Introduction to Panel Data

to be finished

## 8.6 动态面板模型

动态面板数据模型的典型特征是解释变量中包含被解释变量的滞后项。Nerlove (1971) 和 Nickell (1981) 研究表明，对于  $T$  很小的数据而言，常用的针对固定效应模型的最小平方差估计量 (LSDV) 是有偏的，即使  $N$  很大甚至趋于无穷大，我们仅能通过增加  $T$  来降低偏误。

对于  $T$  很小这一类的动态面板数据而言，文献中提供了多种获得一致估计量的方法，这其中的多数方法都是在 GMM 框架下进行的。包括 Anderson and Hsiao (1981), Holtz-Eakin et al. (1988), Arellano and Bond (1991), 以及 Ahn and Schmidt (1995, 1997)。

## 8.6.1 简介

考虑如下动态面板模型：

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \quad (8-97)$$

假设  $\varepsilon_{it} \sim i.i.d.(0, \sigma_\varepsilon^2)$ 。相比于此前介绍的模型，该模型的解释变量中包含了被解释变量一阶滞后项  $y_{i,t-1}$ ，在很多情况下，其系数  $\gamma$  是分析的重点。

为了便于说明，我们先不考虑外生变量  $\mathbf{x}_{it}$ ，重点分析如下简化模型：

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it}, \quad |\gamma| < 1 \quad (8-98)$$

依据该模型的设定，我们可以将  $y_{i,t-1}$  表示为：

$$y_{i,t-1} = \gamma y_{i,t-2} + \alpha_i + \varepsilon_{i,t-1} \quad (8-99)$$

这一表述对于理解动态面板模型很有帮助。

下面要说明的是，模型 (8-98) 的 Pooled OLS、组内估计量，以及一阶差分估计量都是有偏且非一致的。

若采用 Pooled OLS 估计模型 (8-98)，则可将模型表述为：

$$y_{it} = \gamma y_{i,t-1} + u_{it}, \quad u_{it} = \alpha_i + \varepsilon_{it}$$

显然， $\text{Corr}(y_{i,t-1}, u_{it}) \neq 0$ ，因为二者都包含  $\alpha_i$ ，这意味着  $y_{i,t-1}$  是一个内生变量，Pooled OLS 估计量是有偏的。

事实上，即使采用组内变换或差分变换去除个体效应  $\alpha_i$ ，仍然无法解决动态面板模型 (8-98) 的内生性问题。先考虑组内估计量，

$$y_{it} - \bar{y}_i = \gamma(y_{i,t-1} - \bar{y}_{i,-1}) + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad (8-100)$$

其中， $\bar{y}_i = (1/T) \sum_{t=1}^T y_{it}$ ， $\bar{y}_{i,-1} = [1/(T-1)] \sum_{t=2}^T y_{i,t-1}$ 。由于  $\bar{\varepsilon}_i$  中包含  $\varepsilon_{i,t-1}$ ，而根据 (8-99)，它与  $y_{i,t-1}$  是相关的，这意味着  $\text{Corr}(y_{i,t-1} - \bar{y}_{i,-1}, \varepsilon_{it} - \bar{\varepsilon}_i) \neq 0$ ，即， $\gamma$  的组内估计量也是有偏的。<sup>63</sup>

接下来，我们考虑一阶差分估计量，对模型 (8-98) 执行一阶差分变换可得：

$$\Delta y_{it} = \gamma \Delta y_{i,t-1} + \Delta \varepsilon_{it}, \quad t = 2, 3, \dots, T \quad (8-101)$$

<sup>63</sup>只有在  $N \rightarrow \infty$  和  $T \rightarrow \infty$  的情况下，组内估计量才是无偏的。对于多数大  $N$  小  $T$  型面板而言，组内估计量都存在严重的下偏偏误。例如，Verbeek (2004, pp.361) 发现，若  $\gamma$  的真实值为 0.5，即使  $N \rightarrow \infty$ ，当  $T = 2$  时， $\text{plim } \hat{\gamma}_{WG} = -0.25$ ；当  $T = 3$  时， $\text{plim } \hat{\gamma}_{WG} = -0.04$ ；当  $T = 10$  时， $\text{plim } \hat{\gamma}_{WG} = 0.33$ 。Hisao (2003, Section 4.2) 以及 Canova (2007) 也对这一问题进行了非常细致的探讨。

虽然去除了个体效应  $\alpha_i$ ，但  $\text{Corr}(\Delta y_{i,t-1}, \Delta \varepsilon_{it}) \neq 0$ ，<sup>64</sup>因此，一阶差分估计量也是有偏的。

事实上，我们也不难证明，RE 估计量同样面临上述问题。

这里需要说明的是，虽然在模型 (8-98) 中， $\gamma$  的 Pooled OLS 估计量和组内估计量分别上偏 (upward bias) 和下偏 (downward bias) 于其真实值，但二者却构成了  $\gamma$  估计值的上限和下限 (Roodman, 2009)。因此，在后续分析中，我们可以用 OLS 和 FE 估计值作为评估估计结果好坏的一个粗略标准。

### 8.6.2 IV 估计

通过上述分析可以看出，对于动态面板模型而言，此前在静态面板模型使用的估计方法都不再适用。然而，我们注意到，问题的关键在于  $y_{i,t-1}$  作为解释变量导致的内生性问题。为此，可以采用 IV 或 GMM 估计得到  $\gamma$  的一致估计量。

IV 估计量由 Anderson and Hisao (1981) 提出。对于模型 (8-98) 而言，若  $\varepsilon_{it}$  不存在序列相关，则可以先通过一阶差分去除个体效应  $\alpha_i$ 。在模型 (8-101) 中， $y_{i,t-2}$  可以作为  $\Delta y_{i,t-1}$  的工具变量，因为  $y_{i,t-2}$  与  $\Delta y_{i,t-1}$  相关，但与  $\Delta \varepsilon_{it}$  不相关。基于此，我们可以得到如下 IV 估计量：

$$\hat{\gamma}_{IV}^1 = \frac{\sum_{i=1}^N \sum_{t=2}^T y_{i,t-2} \Delta y_{it}}{\sum_{i=1}^N \sum_{t=2}^T y_{i,t-2} \Delta y_{i,t-1}} \quad (8-102)$$

采用矩阵的形式可表述为：

$$\hat{\gamma}_{IV}^1 = (\mathbf{y}'_{-2} \Delta \mathbf{y}_{-1})^{-1} (\mathbf{y}'_{-2} \Delta \mathbf{y}) \quad (8-103)$$

上述估计量具有一致性的前提条件 (矩条件) 是：

$$E(y_{i,t-2} \Delta \varepsilon_{it}) = 0 \quad (8-104)$$

按照相似的思路，Anderson and Hisao (1981) 建议亦可采用  $\Delta y_{i,t-2}$  作为  $\Delta y_{i,t-1}$  的工具变量，相应的 IV 估计量为：

$$\hat{\gamma}_{IV}^2 = \frac{\sum_{i=1}^N \sum_{t=2}^T \Delta y_{i,t-2} \Delta y_{it}}{\sum_{i=1}^N \sum_{t=2}^T \Delta y_{i,t-2} \Delta y_{i,t-1}} \quad (8-105)$$

采用矩阵的形式可表述为：

$$\hat{\gamma}_{IV}^2 = (\Delta \mathbf{y}'_{-2} \Delta \mathbf{y}_{-1})^{-1} (\Delta \mathbf{y}'_{-2} \Delta \mathbf{y}) \quad (8-106)$$

对应的矩条件为：

$$E(\Delta y_{i,t-2} \Delta \varepsilon_{it}) = 0 \quad (8-107)$$

对于上述两个 IV 估计量的优劣，可以从两个角度来看。其一，相比于  $\hat{\gamma}_{IV}^1$  估计量，在  $\hat{\gamma}_{IV}^2$  中，为了构造工具变量  $\Delta y_{i,t-2}$ ，不得不多损失一年的资料。换言之，要获得  $\hat{\gamma}_{IV}^1$  至少需要 3 年的观察值，而获得  $\hat{\gamma}_{IV}^2$  则至少需要 4 年的观察值。其二，Arellano (1989) 发现，若使用

<sup>64</sup>这是因为， $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$ ， $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$ ，而根据 (8-99)， $y_{i,t-1}$  与  $\varepsilon_{i,t-1}$  是相关的。

$\Delta y_{i,t-2}$  作为工具变量, 会导致估计量的方差异常增大, 从而使统计推断变得相当不准确。因此, Arellano 建议采用  $y_{i,t-2}$  做工具变量来执行上述 IV 估计, 即推荐使用  $\hat{\gamma}_{IV}^1$ 。

不过, 自从 Hansen (1982) 提出广义矩估计 (GMM) 方法后, 上述 IV 估计量很少用于估计动态面板模型, 致使以上比较其实上不再是个重要的议题。例如, Ahn and Schmidt (1995) 认为, 虽然上述 IV 估计量是一致的, 但并非最有效的, 因为二者都未充分利用所有可用的矩条件, 同时, IV 估计量也未充分考虑差分后的干扰项 ( $\Delta \varepsilon_{it}$ ) 的方差结构。<sup>65</sup>

### 8.6.3 一阶差分 GMM (FD-GMM) 估计量

#### 基本思路

我们可以从 GMM 的角度重新解读 Anderson and Hisao (1981) 提出的两个 IV 估计量  $\hat{\gamma}_{IV}^1$  和  $\hat{\gamma}_{IV}^2$ , 事实上, 二者分别对应了矩条件 (8-104) 和 (8-107)。沿袭这一思路, 若我们可以找到更多合理的矩条件,<sup>66</sup>相应估计量将更为有效。Arellano and Bond (1991) 正是沿着这一思路提出了在文献中广泛应用的一阶差分 GMM 估计量 (以下简称 FD-GMM)。例如, 对于差分方程 (8-101), 当  $t = 3$  时, 我们能够观察到的第一期观察值为:

$$y_{i3} - y_{i2} = \gamma(y_{i2} - y_{i1}) + (\varepsilon_{i3} - \varepsilon_{i2}) \quad (8-108)$$

显然, 此时  $y_{i1}$  是合理的工具变量。因为, 只要  $\varepsilon_{it}$  不存在序列相关,  $y_{i1}$  便与  $(\varepsilon_{i3} - \varepsilon_{i2})$  不相关, 而  $y_{i1}$  与  $(y_{i2} - y_{i1})$  则是高度相关的。我们可以将上述分析归结为如下矩条件:

$$E[y_{i1}(\varepsilon_{i3} - \varepsilon_{i2})] = 0$$

进一步分析  $t = 4$  的情形, 在 (8-101) 中, 我们能够进一步观察到第二期观察值:

$$y_{i4} - y_{i3} = \gamma(y_{i3} - y_{i2}) + (\varepsilon_{i4} - \varepsilon_{i3}) \quad (8-109)$$

此时,  $y_{i2}$  和  $y_{i1}$  都是  $(y_{i3} - y_{i2})$  的合理工具变量, 因为二者与  $(\varepsilon_{i4} - \varepsilon_{i3})$  的相关系数均为零。换言之, 此时可以进一步得到如下两个矩条件:

$$E[y_{i1}(\varepsilon_{i4} - \varepsilon_{i3})] = 0$$

$$E[y_{i2}(\varepsilon_{i4} - \varepsilon_{i3})] = 0$$

显然, 上述分析思路可以一直延续到样本中的所有观察区间, 随着  $t$  的增加, 工具变量的数目也会逐渐增加, 当  $t = T$  时, 合理的工具变量集为  $(y_{i1}, y_{i2}, \dots, y_{i,T-2})$ 。

<sup>65</sup>若假设  $\varepsilon_{it} \sim i.i.d.(0, \sigma_\varepsilon^2)$ , 则  $\Delta \varepsilon_{it}$  不再服从  $i.i.d.$  分布。详见 (8-121) 式。

<sup>66</sup>其实, 一个非常直接的 GMM 估计量就是把上述两个矩条件联合起来构造相应的 GMM 估计量。

相比于上一节介绍的 IV 估计量, GMM 的主要差异在于可以为不同的观察期设定不同的工具变量集。<sup>67</sup>因此, 对于个体  $i$  而言, 各期的工具变量集合可汇总为如下矩阵:

$$Z_i = \begin{bmatrix} [y_{i1}] & 0 & \cdots & 0 \\ 0 & [y_{i1}, y_{i2}] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [y_{i1}, y_{i2}, \cdots, y_{i,T-2}] \end{bmatrix} \quad (8-110)$$

由于模型 (8-98) 中仅包含了  $y_{it}$  的一阶滞后项, 设  $p = 1$ , 则  $Z_i$  矩阵包含  $T - p - 1$  行,  $\sum_{m=p}^{T-2} m$  列 (这也是矩条件的数目)。样本中所有个体工具变量构成的矩阵为

$$\mathbf{Z} = [Z'_1, Z'_2, \cdots, Z'_N]' \quad (8-111)$$

同时, 上述矩条件简写如下:

$$E(Z'_i \Delta \varepsilon_i) = \mathbf{0} \quad (8-112)$$

为了推导出 GMM 估计量, 可将该矩条件重新表述为对应的样本矩条件:

$$E[Z'_i (\Delta y_i - \gamma \Delta y_{i,-1})] = \mathbf{0} \quad (8-113)$$

由于工具变量的数目可能远远多于未知参数的个数, 我们无法保证上述矩条件严格等于零, 此时需要极小化如下目标函数 (参见第 7 章):

$$\min_{\gamma} \left[ \frac{1}{N} \sum_{i=1}^N (Z'_i (\Delta y_i - \gamma \Delta y_{i,-1})) \right]' \mathbf{W}_N \left[ \frac{1}{N} \sum_{i=1}^N (Z'_i (\Delta y_i - \gamma \Delta y_{i,-1})) \right] \quad (8-114)$$

其中,  $\mathbf{W}_N$  是一个对称且正定的权重矩阵。<sup>68</sup>对  $\gamma$  求一阶偏导数, 可以得到  $\gamma$  的 GMM 估计量:

$$\hat{\gamma}_{GMM} = \left[ \left( \sum_{i=1}^N \Delta \mathbf{y}'_{i,-1} Z_i \right) \mathbf{W}_N \left( \sum_{i=1}^N Z'_i \Delta \mathbf{y}_{i,-1} \right) \right]^{-1} \times \left[ \left( \sum_{i=1}^N \Delta \mathbf{y}'_{i,-1} Z_i \right) \mathbf{W}_N \left( \sum_{i=1}^N Z'_i \Delta \mathbf{y}_i \right) \right] \quad (8-115)$$

显然,  $\hat{\gamma}_{GMM}$  的性质决定于权重矩阵  $\mathbf{W}_N$  的选择。例如, 若选择  $\mathbf{W}_N = \mathbf{I}$ , 则上述 GMM 估计量便转化为普通的 IV 或 2SLS 估计量  $\hat{\gamma}_{2SLS}$ 。<sup>69</sup>虽然当  $N \rightarrow \infty, T \rightarrow \infty$  时,  $\hat{\gamma}_{IV}$  是  $\hat{\gamma}$  的渐进一致

<sup>67</sup>有关面板 GMM 的详细介绍, 请参见 Cameron and Trivedi(2005, Chp 22)。

<sup>68</sup>需要说明的是,  $\mathbf{W}_N$  的下标  $N$  仅表明该权重矩阵依赖于样本数  $N$ , 并非用于标示矩阵的维度。

<sup>69</sup>其估计式为:  $\hat{\gamma}_{2SLS} = [\Delta \mathbf{y}'_{-1} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \Delta \mathbf{y}_{-1}]^{-1} [\Delta \mathbf{y}'_{-1} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \Delta \mathbf{y}]$ 。

性估计量, 但如同我们在第 58 页脚注 65 中所强调的, 该估计量并未考虑干扰项的方差结构, 因此并不是  $\hat{\gamma}$  的有效估计量。

对于小样本而言, 我们必须慎重选择  $\mathbf{W}_N$  以便获得最为有效 ( $\text{Var}(\hat{\gamma}_{GMM})$  最小) 的估计量。根据第 7 章中有关 GMM 理论的介绍, 我们知道, 最优权重矩阵与样本矩 (sample moments) 的协方差矩阵的逆矩阵成正比。这意味着, 最优权重矩阵应该满足如下条件:

$$\text{plim}_{N \rightarrow \infty} \mathbf{W}_N = V(Z_i' \Delta \varepsilon_i)^{-1} = E(Z_i' \Delta \varepsilon_i \Delta \varepsilon_i' Z_i)^{-1} \quad (8-116)$$

这意味着, 只要能获得  $\gamma$  的一致估计量, 进而得到  $\hat{\varepsilon}_{it}$ , 即可用样本均值代替 (8-116) 的期望运算, 以便获得最优权重的估计值:

$$\hat{\mathbf{W}}_N^{opt} = \left( \frac{1}{N} \sum_{i=1}^N Z_i' \Delta \hat{\varepsilon}_i \Delta \hat{\varepsilon}_i' Z_i \right)^{-1} \quad (8-117)$$

例如, 我们可以选择  $\mathbf{W}_N = \mathbf{I}$ , 经由 (8-115) 估得  $\hat{\gamma}_{GMM}$  后, 进一步得到  $\Delta \hat{\varepsilon}_i$ , 由此计算出  $\hat{\mathbf{W}}_N^{opt}$  后重新代入 (8-115) 式, 即可得到  $\hat{\gamma}_{GMM}$  的有效估计量。然而, 这一处理方法并未考虑  $\Delta \varepsilon_i$  的方差结构, 因此, 并非最佳选择。

下面, 我们首先利用工具变量集  $\mathbf{Z}$ , 并采用 GLS 估计差分模型 (8-101), 从而得到  $\gamma$  的一步估计量 (one-step estimator)  $\hat{\gamma}_1$ , 并将其残差代入 (8-117) 式, 得到最优权重矩阵  $\hat{\mathbf{W}}_N^{opt}$ , 进而设  $\mathbf{W}_N = \hat{\mathbf{W}}_N^{opt}$ , 由 (8-115) 式估计出  $\gamma$  的两步估计量 (two-step estimator)  $\hat{\gamma}_2$ 。

### 一步 GMM 估计量

为了便于推导, 采用向量形式将 (8-98) 式重新表示如下:

$$\mathbf{y}_i = \gamma \mathbf{y}_{i,-1} + \mathbf{1}_T \alpha_i + \varepsilon_i \quad (8-118)$$

采用差分矩阵  $\tilde{\mathbf{B}}$  左乘 (8-118) 以去除个体效应  $\alpha_i$ :

$$\tilde{\mathbf{B}} \mathbf{y}_i = \gamma \tilde{\mathbf{B}} \mathbf{y}_{i,-1} + \tilde{\mathbf{B}} \varepsilon_i \quad (8-119)$$

其中,<sup>70</sup>

$$\tilde{\mathbf{B}} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}_{(T-2) \times T} \quad (8-120)$$

干扰项  $\tilde{\mathbf{B}} \varepsilon_i$  的方差-协方差矩阵为:

$$\text{Var}(\tilde{\mathbf{B}} \varepsilon_i) = \sigma_\varepsilon^2 \tilde{\mathbf{B}} \tilde{\mathbf{B}}' = \sigma_\varepsilon^2 \mathbf{G} \quad (8-121)$$

<sup>70</sup>细心的读者会发现, 这里的  $\tilde{\mathbf{B}}$  矩阵与 (8-16) 式的  $\mathbf{B}$  矩阵非常相似, 只是后者的维度为  $(T-1) \times T$ 。这是因为, 对于模型 (8-101) 而言, 只有在  $t=3$  时, 我们才能观察到第一期观察值。



其中,

$$\mathbf{G} = \tilde{\mathbf{B}}\tilde{\mathbf{B}}' = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix}_{(T-2) \times (T-2)} \quad (8-122)$$

将所有观察值累叠后可将 (8-119) 表示为:

$$(\mathbf{I}_N \otimes \tilde{\mathbf{B}})\mathbf{y} = \gamma(\mathbf{I}_N \otimes \tilde{\mathbf{B}})\mathbf{y}_{-1} + (\mathbf{I}_N \otimes \tilde{\mathbf{B}})\boldsymbol{\varepsilon} \quad (8-123)$$

亦可表示为:

$$\Delta\mathbf{y} = \gamma\Delta\mathbf{y}_{-1} + \Delta\boldsymbol{\varepsilon} \quad (8-124)$$

其中,  $\Delta\mathbf{y}_{-1} = (\mathbf{I}_N \otimes \tilde{\mathbf{B}})\mathbf{y}_{-1}$ , 相对而言, 这种表述方法更容易理解。两边同时左乘  $\mathbf{Z}'$  ((8-111) 式) 可得:

$$\mathbf{Z}'\Delta\mathbf{y} = \gamma\mathbf{Z}'\Delta\mathbf{y}_{-1} + \mathbf{Z}'\Delta\boldsymbol{\varepsilon} \quad (8-125)$$

干扰项的方差-协方差矩阵为:<sup>71</sup>

$$\text{Var}(\mathbf{Z}'\Delta\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{A}_1 \quad (8-126)$$

其中,

$$\mathbf{A}_1 = \mathbf{Z}'(\mathbf{I}_N \otimes \mathbf{G})\mathbf{Z} = \sum_{i=1}^N \mathbf{Z}_i' \mathbf{G} \mathbf{Z}_i$$

采用 GLS 估计模型 (8-126) 即可得到 Aellano and Bond (1991) 提出的一步 GMM 估计量:

$$\hat{\gamma}_1 = [\Delta\mathbf{y}_{-1}' \mathbf{Z} \mathbf{A}_1^{-1} \mathbf{Z}' \Delta\mathbf{y}_{-1}]^{-1} [\Delta\mathbf{y}_{-1}' \mathbf{Z} \mathbf{A}_1^{-1} \mathbf{Z}' \Delta\mathbf{y}] \quad (8-127)$$

若设  $\mathbf{W}_N = \mathbf{A}_1^{-1}$ , 则可用 (8-115) 式的形式将 (8-127) 式重新表示如下:

$$\hat{\gamma}_1 = \mathbf{Q}_1^{-1} \left( \sum_{i=1}^N \Delta\mathbf{y}_{i,-1}' \mathbf{Z}_i \right) \mathbf{A}_1^{-1} \left( \sum_{i=1}^N \mathbf{Z}_i' \Delta\mathbf{y}_i \right) \quad (8-128)$$

其中,

$$\mathbf{Q}_1 = \left( \sum_{i=1}^N \Delta\mathbf{y}_{i,-1}' \mathbf{Z}_i \right) \mathbf{A}_1^{-1} \left( \sum_{i=1}^N \mathbf{Z}_i' \Delta\mathbf{y}_{i,-1} \right)$$

<sup>71</sup>推导过程为:  $\text{Var}(\mathbf{Z}'\Delta\boldsymbol{\varepsilon}) = \text{Var}(\mathbf{Z}'(\mathbf{I}_N \otimes \tilde{\mathbf{B}})\boldsymbol{\varepsilon}) = \mathbf{Z}'(\mathbf{I}_N \otimes \tilde{\mathbf{B}})\text{Var}(\boldsymbol{\varepsilon})(\mathbf{I}_N \otimes \tilde{\mathbf{B}}')\mathbf{Z} = \mathbf{Z}'(\mathbf{I}_N \otimes \tilde{\mathbf{B}})(\sigma_\varepsilon^2 \mathbf{I}_{NT})(\mathbf{I}_N \otimes \tilde{\mathbf{B}}')\mathbf{Z} = \sigma_\varepsilon^2 \mathbf{Z}'(\mathbf{I}_N \otimes \tilde{\mathbf{B}})(\mathbf{I}_N \otimes \tilde{\mathbf{B}}')\mathbf{Z} = \sigma_\varepsilon^2 \mathbf{Z}'(\mathbf{I}_N \otimes \mathbf{G})\mathbf{Z} = \sigma_\varepsilon^2 \mathbf{A}_1$ 。可利用如下矩阵运算:  $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$ ,  $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$ 。

显然, (8-128) 式中未包含任何未知参数。此外, 从 (8-126) 式的变换过程来看,  $\hat{\gamma}_1$  与 Anderson and Hisao 提出的 IV 估计量非常相似, 区别在于, 它使用了更多的工具变量, 更为重要的式, 它考虑了  $\Delta\epsilon_i$  的方差-协方差结构, 因此更为有效。

估计出  $\hat{\gamma}_1$  后, 可以进步得到一步估计量的残差:

$$\Delta\hat{\epsilon}_i = \Delta\mathbf{y}_i - \hat{\gamma}_1\Delta\mathbf{y}_{i,-1} \quad (8-129)$$

若假设  $\epsilon_{it}$  是同方差的, 则  $\hat{\gamma}_1$  的方差-协方差矩阵为:

$$\hat{V}_1 = \hat{\sigma}_1^2 \mathbf{Q}_1^{-1} \quad (8-130)$$

其中,

$$\hat{\sigma}_1^2 = \frac{1}{NT - K} \sum_{i=1}^N \Delta\hat{\epsilon}_i' \Delta\hat{\epsilon}_i$$

稳健型估计量为:

$$\hat{V}_{1r} = \mathbf{Q}_1^{-1} \left( \sum_{i=1}^N \Delta\mathbf{y}_{i,-1}' Z_i \right) \mathbf{A}_1^{-1} \mathbf{A}_2 \mathbf{A}_1^{-1} \left( \sum_{i=1}^N Z_i' \Delta\mathbf{y}_{i,-1} \right) \mathbf{Q}_1^{-1} \quad (8-131)$$

其中,

$$\mathbf{A}_2 = \sum_{i=1}^N Z_i' \Delta\hat{\epsilon}_i' \Delta\hat{\epsilon}_i Z_i$$

### 两步 GMM 估计量

将 (8-129) 式中由一步估计量得到的残差  $\Delta\hat{\epsilon}_i$  代入 (8-117) 式, 即可得到 GMM 估计的最优权重,  $\hat{\mathbf{W}}_N^{opt} = \mathbf{A}_2$ 。因此, Arellano and Bond (1991) 提出的两步估计量 (一个真正意义上的 GMM 估计量) 可表示为:

$$\hat{\gamma}_2 = \mathbf{Q}_2^{-1} \left( \sum_{i=1}^N \Delta\mathbf{y}_{i,-1}' Z_i \right) \mathbf{A}_2^{-1} \left( \sum_{i=1}^N Z_i' \Delta\mathbf{y}_i \right) \quad (8-132)$$

其中,

$$\mathbf{Q}_2 = \left( \sum_{i=1}^N \Delta\mathbf{y}_{i,-1}' Z_i \right) \mathbf{A}_2^{-1} \left( \sum_{i=1}^N Z_i' \Delta\mathbf{y}_{i,-1} \right)$$

此时,  $\hat{\gamma}_2$  的方差-协方差矩阵为:

$$\hat{V}_2 = \mathbf{Q}_2^{-1} \quad (8-133)$$

Arellano and Bond (1991) 的 Monte Carlo 模拟分析表明, 对于小样本而言, 两步估计量的标准误存在严重的下偏偏误 (downward bias)。因此, 他们建议在进行系数统计推断时, 仍然采用一步 GMM 估计量, 因为一步估计量的权重矩阵与待估参数无关, 而两步估计量的权重矩阵则依赖于一步估计得到的残差。在最近的一篇文章中, Windmeijer (2005) 提出了一种在小样本下修正这种偏误方法, 其 Monte Carlo 模拟分析表明, 修正后的两步估计量的标准误具有很好的统计性质。在 STATA 中, 两步估计量的稳健型标准误便是采用 Windmeijer (2005) 的修正公式计算的。

### 8.6.4 假设检验

#### 序列相关检验

前面已经提到, 无论是 IV 估计还是 GMM 估计, 工具变量的合理性都依赖于一个重要的前提假设:  $\varepsilon_{it}$  不存在序列相关。为此, Arellano and Bond (1991) 提出了如下统计量, 以检验差分方程的干扰项  $\Delta\varepsilon_i$  的  $m$  阶序列相关是否显著:

$$AR_m = \frac{\sum_{i=1}^N \Delta\hat{\varepsilon}'_{mi} \Delta\hat{\varepsilon}_i}{B_1^{1/2}} \quad (8-134)$$

其中,

$$\Delta\hat{\varepsilon}_{mi} = L_m(\Delta\hat{\varepsilon}_i)$$

这里,  $L_m$  表示  $m$  阶滞后算子。同时,

$$\begin{aligned} B_1 = & \sum_{i=1}^N \Delta\hat{\varepsilon}'_{mi} \mathbf{G} \Delta\hat{\varepsilon}_{mi} - 2 \left( \sum_{i=1}^N \Delta\hat{\varepsilon}'_{mi} \Delta\mathbf{y}_{i,-1} \right) \mathbf{Q}_1^{-1} \left( \sum_{i=1}^N \Delta\mathbf{y}'_{i,-1} \mathbf{Z}_i \right) \mathbf{A}_1^{-1} \left( \sum_{i=1}^N \mathbf{Z}'_i \mathbf{G} \Delta\hat{\varepsilon}_{mi} \right) \\ & + \left( \sum_{i=1}^N \Delta\hat{\varepsilon}'_{mi} \Delta\mathbf{y}_{i,-1} \right) \hat{\mathbf{V}}_1 \left( \sum_{i=1}^N \Delta\mathbf{y}'_{i,-1} \Delta\hat{\varepsilon}_{mi} \right) \end{aligned}$$

需要说明的是, 即使  $\varepsilon_{it}$  不存在序列相关, 其差分项  $\Delta\varepsilon_{it}$  也必然存在一阶序列相关,<sup>72</sup>但不存在二阶序列相关。因此, 在实证分析过程中, 我们重点关注的是  $\Delta\varepsilon_{it}$  是否存在二阶序列相关, 即  $AR_2$  是否显著。

上述  $AR_m$  统计量是在  $\varepsilon_{it}$  为同方差假设下得到的。若考虑异方差并采用 (8-131) 式计算稳健型标准误, 在计算  $AR_m$  统计量时, 需要将 (8-135) 式  $B_1$  中的  $\mathbf{G}$  替换为  $\Delta\hat{\varepsilon}'_i \Delta\hat{\varepsilon}_i$ ,  $\mathbf{A}_1$  替换为  $\mathbf{A}_2$ ,  $\hat{\mathbf{V}}_1$  替换为  $\hat{\mathbf{V}}_{1r}$ 。对于两步 GMM 估计量而言, 在计算  $AR_m$  统计量时, 要将 (8-135) 式  $B_1$  中的  $\mathbf{Q}_1$  替换为  $\mathbf{Q}_2$ ,  $\mathbf{A}_1$  替换为  $\mathbf{A}_2$ ,  $\hat{\mathbf{V}}_1$  替换为  $\hat{\mathbf{V}}_2$ 。同时, 还需将  $\Delta\hat{\varepsilon}_{mi}$  替换为  $\Delta\hat{\varepsilon}_{mi} = \Delta\mathbf{y}_i - \hat{\gamma}_2 \Delta\mathbf{y}_{i,-1}$ , 并将  $\mathbf{G}$  替换为  $\Delta\hat{\varepsilon}'_{mi} \Delta\hat{\varepsilon}_{mi}$ 。详见 Stata 11 Manual, [XT] xtabond。

#### 过度识别检验: Sargan 检验

在上述 GMM 估计过程中, 工具变量的数目 ( $r$ ) 通常远远大于未知参数的数目 ( $k$ ), 这意味着 GMM 估计中包含了  $(r - k)$  个过度识别约束 (over-identifying restrictions)。Arellano and Bond (1991) 建议采用 Sargan (1958) 提出的检验统计量。对于一步 GMM 估计量而言, 相应的 Sargan 统计量为:

$$S_1 = \left( \sum_{i=1}^N \Delta\hat{\varepsilon}'_i \mathbf{Z}_i \right) \mathbf{A}_1 \left( \sum_{i=1}^N \mathbf{Z}'_i \Delta\hat{\varepsilon}_i \right) \left( \frac{1}{\hat{\sigma}_1^2} \right) \sim \chi_{r-k}^2 \quad (8-135)$$

<sup>72</sup>因为  $\Delta\varepsilon_{it}$  和  $\Delta\varepsilon_{i,t-1}$  中都包含  $\varepsilon_{i,t-1}$ 。

对于两步 GMM 估计量而言, 相应的 Sargan 统计量为:

$$S_2 = \left( \sum_{i=1}^N \Delta \hat{\varepsilon}_i' Z_i \right) A_2 \left( \sum_{i=1}^N Z_i' \Delta \hat{\varepsilon}_i \right) \sim \chi_{r-k}^2 \quad (8-136)$$

其中,  $\Delta \hat{\varepsilon}_{mi} = \Delta y_i - \hat{\gamma}_2 \Delta y_{i,-1}$ 。如果  $S_1$  或  $S_2$  大于给定置信水平 (如 5%) 上的临界值, 则过度识别的矩条件被拒绝, 这意味着部分  $Z_i$  中的工具变量与干扰项相关, 并非合理的工具变量。此时需要重新设定模型或选择新的工具变量。

需要特别说明的是, 虽然我们同时列出了  $S_1$  或  $S_2$  两个统计量, 但 Arellano and Bond (1991) 以及后续大量研究都表明, 应该使用  $S_2$  进行 Sargan 检验。因为,  $S_1$  并未考虑异方差问题, 因此存在严重的偏误。

### 8.6.5 包含其它解释变量的动态面板模型

考虑如下一般化的动态面板模型:

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}' \boldsymbol{\beta} + \alpha_i + \varepsilon_{it} \quad (8-137)$$

首先, 需要通过一阶差分去除个体效应,

$$\Delta y_{it} = \gamma \Delta y_{i,t-1} + \Delta \mathbf{x}_{it}' \boldsymbol{\beta} + \Delta \varepsilon_{it} \quad (8-138)$$

我们仍然可以采用前面介绍的 GMM 进行估计, 差别仅在于 (8-110) 式的工具变量集会有所变化。在正式分析之前, 有必要对  $\mathbf{x}_{it}$  中包含的变量类型进行区分:

(1) 外生变量 (exogenous variables)。若对于任意的  $s$  和  $t$ , 都有  $E(x_{it}\varepsilon_{is}) = 0$  成立, 则称  $x_{it}$  为严格外生变量。换言之, 任何时点的外来干扰都不会影响  $x_{it}$ 。

(2) 内生变量 (endogenous variables)。若当  $s \leq t$  时,  $E(x_{it}\varepsilon_{is}) \neq 0$ , 则称  $x_{it}$  为内生变量。

(3) 先决变量 (predetermined variables)。若当  $s < t$  时,  $E(x_{it}\varepsilon_{is}) \neq 0$ , 但当  $s \geq t$  时,  $E(x_{it}\varepsilon_{is}) = 0$ , 则称  $x_{it}$  为先决变量。对于此类变量而言, 虽然当期干扰  $\varepsilon_{it}$  不会影响当期  $x_{it}$ , 但会影响其随后各期的观察值, 即  $x_{i,t+p}$  ( $p > 0$ )。

完成上述界定后, 我们可以依据如下原则来设定模型 (8-138) 中的工具变量:

(1) 若  $x_{it}$  为严格外生变量, 则  $\Delta x_{it}$  可以作为自身的工具变量;

(2) 若  $x_{it}$  为内生变量, 其处理方法与  $y_{i,t-1}$  完全相同, 可以采用滞后二阶以上的水平变量 ( $x_{it-s}, s \geq 2$ ) 作为  $\Delta x_{it}$  的工具变量;

(3) 若  $x_{it}$  为先决变量, 则  $x_{i,t-1}$  也是  $\Delta x_{it}$  的合理工具变量, 为此, 可以采用滞后一阶以上的水平变量 ( $x_{it-s}, s \geq 1$ ) 作为  $\Delta x_{it}$  的工具变量。

例如, 若假设  $\mathbf{x}_{it}$  中包含的所有变量都是严格外生的, 则工具变量集可以表述为:

$$Z_i = \begin{bmatrix} [y_{i1}] & 0 & \cdots & 0 & \Delta \mathbf{x}_{i3}' \\ 0 & [y_{i1}, y_{i2}] & \cdots & 0 & \Delta \mathbf{x}_{i4}' \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & [y_{i1}, y_{i2}, \cdots, y_{i,T-2}] & \Delta \mathbf{x}_{iT}' \end{bmatrix} \quad (8-139)$$

此时,  $Z_i$  的行数与 (8-110) 相同, 仍然是  $T - p - 1$ , 而其列数则增加为  $\sum_{m=p}^{T-2} m + k$ , 其中,  $k$  表示  $\mathbf{x}$  中包含的变量的个数。

在实际分析过程中,  $\mathbf{x}$  中可能同时包含外生变量、内生变量和先决变量, 我们只需按照上述原则设定工具变量矩阵  $Z_i$  即可。当然, 若能够获得其它的工具变量, 亦可将其加入  $Z_i$  矩阵。

### 滞后阶数的选择 (弱工具变量问题)

显然, 当  $\mathbf{x}$  中包含一个或数个内生变量或先决变量时, 工具变量的数目随着样本区间  $T$  的增大会迅速增加。相对于 Anderson and Hisao (1981) 提出的 IV 估计量, Arellano and Bond (1991) 的 FD-GMM 估计量, 以及随后将要介绍的 Blundell and Bover (1998) 的 SYS-GMM 估计量都采用了大量的工具变量 (工具变量的数目随着样本区间  $T$  的增大会迅速增加), 以便提高 GMM 估计的有效性。虽然从理论上讲, 对于内生变量  $\Delta w_{it}$ , 滞后两阶以上的水平变量 ( $w_{it-s}, s \geq 2$ ) 都是合理的工具变量, 但当  $s > 4$  时,  $w_{it-s}$  与  $\Delta w_{it}$  之间的相关性可能很低, 此时便可能出现所谓的“弱工具变量 (weak instruments)”问题。虽然文献中并未明确给出弱工具变量的定义, 但基本的共识是, 当工具变量与内生变量 (或先决变量) 之间的相关性较低时, 便会出现弱工具变量问题。此时, GMM 估计量将不再具有一致性。近期的研究发现 (Ziliak, 1997; Judson and Owen, 1999), 使用过度的工具变量会导致 GMM 估计量的小样本性质非常糟糕。

工具变量过多还会对假设检验产生重要的影响,<sup>73</sup>例如, Anderson and Sorenson (1996) 以及 Bowsher (2002) 均发现, 这可能大幅弱化 Hansen 检验的检定力, 表现为 Hansen  $J$  统计量对应的  $p$  值为 1。Sargan (1958) 在提出 Sargan 统计量时, 也特别指出该统计量的偏误与工具变量的个数成正比, 因此, 若使用 Sargan 统计量的渐进分布进行统计推断, 则工具变量的个数不能过大。<sup>74</sup> 因此, 对于一个  $T = 10$  的面板而言, 选择  $s \leq 3$  或  $s \leq 4$  比选择  $s \leq T - 2$  要更合理一些。

### 估计和检验方法

在 8.6.3 小节介绍的估计和检验方法同样适用于模型 (8-138)。设  $\Delta \mathbf{x}_i^* = [\Delta \mathbf{y}_{i,-1}, \mathbf{x}_i']'$ , 则只需用  $\Delta \mathbf{x}_i^*$  替换 8.6.3 小节中的  $\Delta \mathbf{y}_{i,-1}$ , 并根据  $x_{it}$  是内生变量、外生变量还是先决变量调整  $Z_i$  矩阵的定义, 便可利用 8.6.3 小节中介绍的公式得到模型 (8-138) 的各个估计量。

### 8.6.6 系统 GMM (SYS-GMM) 估计量

to be finished

<sup>73</sup>当然, 对于何谓“过多”这一问题, 文献中并未形成一致观点。或许是因为即使在工具变量数目很少的情况下, GMM 估计量仍然可能存在偏误 (Roodman, 2009, pp.99)。

<sup>74</sup>“They were found to be proportional to the number of instrumental variables, so that, if the asymptotic approximations are to be used, this number must be small.” 参见 Sargan (1958, pp.393)。

## 8.7 面板门槛模型

### 8.7.1 简介

自从 Tong (1978) 提出门限自回归模型 (Threshold Auto-regression, 简称 TAR) 后, 这种非线性时间序列模型在经济和金融领域得到了广泛的应用。虽然 TAR 模型大部分被应用于时间序列资料, 但 Tiao and Tsay (1994)、Potter (1995)、Marterns, Kofman and Vorst (1998) 也利用此方法分析横截面资料或面板资料。门限自回归模型在计量方法上有较客观地研究方式, 利用门限变量 (Threshold variable) 来决定不同的分界点, 进而利用门限变量的观察值估计出适合的门限值, 这可以有效避免了一般研究者所使用的主观判定分界点法所造成的偏误。

在估计门限自回归模型时, 必须首先检验是否存在门限效应。由于未知参数 (nuisance) 的存在将导致检验统计量的分布是非标准的, 即出现了所谓的“Davies Problem”。因此, Hansen (1999) 建议采用“自体抽样法” (Bootstrap, 也称为“拔靴法”) 来计算检验统计量的渐进分布, 以便检验门限效应的显著性。在拒绝原假设, 即存在门限效应的情况下, Chan (1993) 研究表明, 门限自回归模型的 OLS 估计量具有超一致性, Chan 进而推导出了 OLS 估计量的渐进分布。不幸的是, 未知参数的存在会导致该分布呈现非标准态。Hansen (1999) 通过似然比检验 (Likelihood Ratio test) 构造“非拒绝域”的方法解决了这个问题。

Hansen (1999) 建议采用两阶段最小二乘法来估计门限面板数据模型。第一步, 对于给定的门限值 ( $\gamma$ ), 计算相应的残差平方和 (SSR), 进而所有 SSR 中的最小值所对应的  $\gamma$  值  $\hat{\gamma}$ 。第二步, 利用  $\hat{\gamma}$  值来估计模型中不同区间 (Regime) 的系数并作相关的分析。

### 8.7.2 单一门槛模型

#### 模型的基本设定

基本模型设定如下:

$$y_{it} = u_i + x'_{it}\beta_1 \cdot I(q_{it} \leq \gamma) + x'_{it}\beta_2 \cdot I(q_{it} > \gamma) + \varepsilon_{it}, \quad (8-140)$$

其中,  $i = 1, 2, \dots, N$  表示不同的个体,  $t = 1, 2, \dots, T$  表示时间,  $q_{it}$  为门槛变量,  $y_{it}$  和  $X_{it}$  分别为被解释变量和解释变量,  $I(\cdot)$  为一个指标函数, 相应的条件成立时取值为 1, 否则取值为 0。用下面的方式表示 (8-140) 式可能更为清晰:

$$y_{it} = \begin{cases} \mu_i + x_{it}\beta'_1 + \varepsilon_{it} & q_{it} \leq \gamma \\ \mu_i + x_{it}\beta'_2 + \varepsilon_{it} & q_{it} > \gamma \end{cases} \quad (8-141)$$

为了下面分析的方便, 我们采用一种更为紧凑的方式来表示 (8-140) 式, 设

$$x_{it}(\gamma) = \begin{cases} x_{it}I(q_{it} \leq \gamma) \\ x_{it}I(q_{it} > \gamma) \end{cases} \quad (8-142)$$

且  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1 \ \boldsymbol{\beta}'_2)'$ ，于是 (8-140) 式等价于：

$$y_{it} = \mu_i + x'_{it}(\gamma)\boldsymbol{\beta} + \varepsilon_{it} \quad (8-143)$$

依据门限变量  $q_{it}$  与门限值  $\gamma$  的相对大小，我们可以将样本观察值分成两个区间。区间的差异表现在回归系数  $\boldsymbol{\beta}_1$  和  $\boldsymbol{\beta}_2$  的不同上。为了保证  $\boldsymbol{\beta}_1$  和  $\boldsymbol{\beta}_2$  可以识别， $x_{it}$  中不能包含诸如性别、国籍等不随时间改变的变量。同时，我们也要求门限变量  $q_{it}$  是不随时间改变的。假设  $\varepsilon_{it}$  服从均值为 0，方差  $\sigma_\varepsilon^2$  有限的独立、同分布，即  $\varepsilon_{it} \sim i.i.d \ N(0, \sigma_\varepsilon^2)$ 。*iid* 的假设使得我们的解释变量中不能包含被解释变量的滞后期，也就是说，我们的模型是一个静态模型。当然，如何将此模型扩展到动态模型或包含异方差的模型是一个有待进一步研究的问题。

### 模型的估计

我们需要先去除个体效应  $\mu_i$ ，采用的方法就是去除组内平均值，这与我们处理一般的固定效应模型所采用的方法是相同的。对 (8-140) 式中所有截面取组内平均得到：

$$\bar{y}_i = \mu_i + \bar{\mathbf{x}}_i(\gamma) + \bar{\varepsilon}_i \quad (8-144)$$

其中，

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}, \quad \bar{\varepsilon}_i = \frac{1}{T} \sum_{t=1}^T \varepsilon_{it}, \quad \bar{\mathbf{x}}_i = \frac{1}{T} \sum_{t=1}^T x_{it}(\gamma) = \begin{pmatrix} \frac{1}{T} \sum_{t=1}^T x_{it} I(q_{it} \leq \gamma) \\ \sum_{t=1}^T x_{it} I(q_{it} > \gamma) \end{pmatrix}.$$

用 (8-143) 式减去 (8-144) 式，得到

$$y_{it}^* = x_{it}^*(\gamma)\boldsymbol{\beta} + \varepsilon_{it}^*. \quad (8-145)$$

其中， $y_{it}^* = y_{it} - \bar{y}_i$ ， $x_{it}^*(\gamma) = x_{it} - \bar{\mathbf{x}}_i$ ， $\varepsilon_{it}^* = \varepsilon_{it} - \bar{\varepsilon}_i$ 。进一步将个体观察值进行累叠<sup>75</sup>

$$y_i^* = \begin{pmatrix} y_{i2}^* \\ \vdots \\ y_{iT}^* \end{pmatrix}, \quad x_i^*(\gamma) = \begin{pmatrix} x_{i2}^*(\gamma) \\ \vdots \\ x_{iT}^*(\gamma) \end{pmatrix}, \quad \varepsilon_i^* = \begin{pmatrix} \varepsilon_{i2}^* \\ \vdots \\ \varepsilon_{iT}^* \end{pmatrix}$$

我们进而可以对所有个体的观察值进行累叠，分别表示为  $\mathbf{Y}^*$ 、 $\mathbf{X}^*(\gamma)$  和  $\boldsymbol{\varepsilon}^*$ ，例如，

$$\mathbf{X}^*(\gamma) = \begin{pmatrix} \mathbf{x}_1^*(\gamma) \\ \vdots \\ \mathbf{x}_i^*(\gamma) \\ \vdots \\ \mathbf{x}_n^*(\gamma) \end{pmatrix}$$

<sup>75</sup> 由于在得到 (8-145) 式的过程中，我们去除了组内平均，所以第一期的观察值已经无用，不过如果不去除第一期的观察值，结果也不会受到影响。笔者曾写信给 Hansen 教授确认过这个问题。



采用这种表示方式, (8-145) 式等价于

$$\mathbf{Y}^* = \mathbf{X}^*(\gamma)\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \quad (8-146)$$

对于给定的  $\gamma$  值, 我们可以采用普通最小二乘法 (OLS) 得到参数  $\boldsymbol{\beta}$  的一致估计量, 即

$$\hat{\boldsymbol{\beta}}(\gamma) = [\mathbf{X}^*(\gamma)' \mathbf{X}^*(\gamma)]^{-1} \mathbf{X}^*(\gamma)' \mathbf{Y}^*, \quad (8-147)$$

相应的残差向量为,

$$\hat{\mathbf{e}}^*(\gamma) = \mathbf{Y}^* - \mathbf{X}^*(\gamma)\hat{\boldsymbol{\beta}}^*(\gamma), \quad (8-148)$$

残差平方和为,

$$S_1(\gamma) = \hat{\mathbf{e}}^*(\gamma)' \hat{\mathbf{e}}^*(\gamma) \quad (8-149)$$

Chan (1993) 和 Hansen (1997) 建议采用最小二乘法来估计  $\gamma$ 。我们可以通过最小化 (8-149) 式对应的  $S_1(\gamma)$  来获得  $\gamma$  的估计值, 即

$$\hat{\gamma} = \arg \min_{\gamma} S_1(\gamma). \quad (8-150)$$

一旦我们得到了  $\hat{\gamma}$ , 便可进而得到  $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\gamma})$ 、残差向量  $\hat{\mathbf{e}}^* = \hat{\mathbf{e}}^*(\hat{\gamma})$  以及残差的方差

$$\hat{\sigma} = \hat{\sigma}^2(\hat{\gamma}) = \frac{1}{n(T-1)} \hat{\mathbf{e}}^{*'} \hat{\mathbf{e}}^* = \frac{1}{n(T-1)} S_1(\hat{\gamma}) \quad (8-151)$$

其中,  $n$  表示公司数目,  $T$  表示时间跨度。

#### 计算相关问题

在我们对 (8-150) 式极小化从而得到门限值  $\gamma$  之 OLS 估计量的过程中, 由于残差平方和  $S_1(\gamma)$  仅决定于  $\gamma$ , 所以它是  $\gamma$  的一个非连续函数, 其中至多有  $nT$  个间断点 (因为样本容量为  $nT$ )。但事实上, 我们在极小化 (8-150) 式的过程中, 仅需要对那些非重复的门槛值  $q_{it}$  进行搜索即可。因此, 一般情况下搜索次数会小于  $nT$  次。我们可以采用以下步骤对 (8-150) 式进行求解:

- 第一步, 对门限变量  $q_{it}$  中非重复的观察值进行排序, 去掉其中最大和最小的观察值各  $\eta\%$  个 ( $\eta > 0$ )。
- 第二步, 用余下的  $N^*$  个观察值作为估计样本, 将  $q_{it}$  值从小到大依次代入 (8-146) 式进行估计, 得到相应的残差平方和 (8-149), 后者的最小值便对应着  $\gamma$  的估计值  $\hat{\gamma}$ 。

在实际操作过程中,  $N$  可能是一个很大的数值, 上面介绍的方法可能相当费时。我们可以采用一种更为简洁的等价方式, 使得需要搜索的  $\gamma$  值的个数大幅减少。我们不必搜索所有介于第  $\eta\%$  和  $(1 - \eta)\%$  百分位之间的  $q_{it}$  值, 而是仅仅搜索特定百分位上的数值。采用这种方法在多数实证分析中都能达到我们所期望的精度。在笔者的一篇文章中 (连玉君和程建, 2006), 我们按照 {1.00%, 1.25%, 1.50%, 1.75%, 2%, ..., 99.0%} 栅格进行搜索, 共包括 393 个分位值。



## 假设检验

### 1. 门槛效应的检验

虽然在模型的设定中我们假设存在门限效应，但是其是否具有统计上的显著性，还需要做进一步的检验。原假设为不存在门限效应，可以表示为：

$$H_0 : \beta_1 = \beta_2$$

相应的备择假设为：

$$H_1 : \beta_1 \neq \beta_2$$

在原假设  $H_0$  下，门限值  $\gamma$  是无法识别的，<sup>76</sup> 此时传统检验统计量的分布是非标准的，这就是所谓的“Davies Problem”。<sup>77</sup> Hansen (1996) 建议采用“自体抽样法”(Bootstrap) 来模拟似然比检验的渐进分布。在不存在门限效应的原假设下，模型 (8-146) 转化为，

$$y_{it} = \mu_i + x'_{it}\beta_1 + \varepsilon_{it}, \quad (8-152)$$

去除个体效应后，可得，

$$y_{it}^* = x_{it}^*\beta_1^* + \varepsilon_{it}^* \quad (8-153)$$

我们可以采用 OLS 得到  $\beta_1$  的估计量，表示为  $\tilde{\beta}_1$ ，相应的残差为  $\tilde{\varepsilon}_{it}^*$ ，残差平方和为  $S_0 = \tilde{\varepsilon}_{it}^{*\prime}\tilde{\varepsilon}_{it}^*$ 。似然比检验 (LR test) 基于如下统计量，

$$F_1 = \frac{S_0 - S_1(\hat{\gamma})}{\sigma_\varepsilon^2} = \frac{S_0 - S_1(\hat{\gamma})}{S_1(\hat{\gamma})/n(T-1)} \quad (8-154)$$

$F_1$  的渐进分布是非标准的，完全不同于卡方分布。而且，一般而言其分布依赖于样本的矩 (如均值、方差、峰度和偏度等)，所以临界值无法查表得到。Hansen (1996) 研究表明，采用“自体抽样法”可以获得其一阶渐进分布，基于此构造的  $p$  值也将是渐进有效的 (asymptotically valid)。考虑到面板模型的数据结构特征，我们建议采用如下步骤进行自体抽样。

- 第一步，在反复抽样的过程中，假设解释变量  $x_{it}$  和门槛变量  $q_{it}$  都是固定不变的。将估计备择假设对应的模型 (8-145) 得到的残差  $\hat{\varepsilon}^*$  按个体分组： $\hat{\varepsilon}_i^* = (\hat{\varepsilon}_{i1}^*, \hat{\varepsilon}_{i2}^*, \dots, \hat{\varepsilon}_{iT}^*)$ ，把由此得到的样本观察值  $\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \dots, \hat{\varepsilon}_n^*$  视为“自体抽样”的实证分布 (empirical distribution)。
- 第二步，从实证分布中随机抽取 (可重复)  $n$  个样本观察值，构造出原假设  $H_0$  下的“自体抽样”样本，即， $\mathbf{Y}_{bs} = \mathbf{X}^*(\gamma)\boldsymbol{\beta} + \mathbf{e}_{bs}$ 。<sup>78</sup>

<sup>76</sup>因为在原假设  $H_0$  下，我们事实上对 (8-140) 式施加了线性约束  $\beta_1 - \beta_2 = 0$ ，所以不存在一个唯一的  $\gamma$  值，使 (8-140) 式成立。

<sup>77</sup>Davies Problem 是指由于未知参数的存在使得检验统计量服从非标准分布的问题 (Davies, 1977, 1987)。在近期的研究中，Andrews and Ploberger (1994) 以及 Hansen (1996) 又重新审视了这个问题，并在处理方法上做了一些改进。

<sup>78</sup>需要说明的是，在原假设  $H_0$  下， $F_1$  统计量与参数  $\beta_1$  无关，所以任何的  $\beta_1$  值都可以使用。

- 第三步, 利用第二步构造的“自体抽样”样本分别估计原假设对应的模型 (8-153) 和备择假设对应的模型 (8-145), 计算由 (8-154) 式决定的似然比统计量。
- 第四步, 重复以上过程多次 (如 500 次), 计算模拟值大于真实值的概率, 这便是我们采用“自体抽样”法得到的原假设  $H_0$  下  $F_1$  统计量的渐进 P 值。

如果得到的 P 值小于我们设定的临界值 (如 5%), 那么就拒绝原假设, 从而认为存在门槛效应。

## 2. 门槛估计值的渐进分布特征

在已经确认存在门槛效应的情况下 ( $\beta_1 \neq \beta_2$ ), Chan (1993) 以及 Hansen (1997) 研究表明  $\hat{\gamma}$  是  $\gamma_0$  ( $\gamma$  的真实值) 的一致估计量, 然而其渐进分布是高度非标准的。Hansen (1997) 指出, 构造  $\gamma$  的置信区间的最佳方法是利用似然比统计量构造出“非拒绝域”。对于原假设  $H_0: \gamma = \gamma_0$  而言, 似然比统计量为:

$$LR_1(\gamma) = \frac{S_1(\gamma) - S_1(\hat{\gamma})}{\hat{\sigma}^2} \quad (8-155)$$

如果  $LR_1(\gamma_0)$  的值足够大, 那么我们就拒绝原假设。需要注意的是, (8-155) 式的统计量和 (8-154) 式的统计量所检验的原假设是不同的。 $LR_1(\gamma_0)$  用于检验  $H_0: \gamma = \gamma_0$ , 而  $F_1$  用于检验  $H_0: \beta_1 = \beta_2$ 。

在满足一系列假设条件及  $H_0: \gamma = \gamma_0$  的情况下, Hansen (1999) 导出  $LR_1(\gamma)$  的渐进分布满足定理 1:

$$LR_1(\gamma) \rightarrow_d \xi \quad (n \rightarrow \infty) \quad (8-156)$$

其中,  $\xi$  是一个具有如下分布函数的随机变量,

$$p(\xi \leq x) = (1 - \exp(-x/2))^2 \quad (8-157)$$

这表明似然比统计量具有非标准分布, 但不存在“未知参数”问题 (free of nuisance parameters)。由于定理 1 中的渐进分布是枢轴的 (pivotal), 因此可用来构造一个有效渐进的置信区间。进一步地, 分配函数 (8-157) 的反函数可表示为,

$$c(\alpha) = -2 \ln(1 - \sqrt{1 - \alpha}) \quad (8-158)$$

据此我们可以很方便的计算出临界值。比如, 10% 显著水平下的临界值为 6.53, 5% 临界值为 7.35, 1% 临界值为 10.59。如果  $LR_1(\gamma_0)$  的值大于  $c(\alpha)$ , 那么我们就在  $\alpha$  显著水平上拒绝原假设  $H_0: \gamma = \gamma_0$ 。

为了得到  $\gamma$  的渐进置信区间, 我们可以引入“非拒绝域”的概念。在  $1 - \alpha$  置信水平上的“非拒绝域”是指一系列满足  $LR_1(\gamma) \leq c(\alpha)$  的  $\gamma$  值, 二者分别由 (8-155) 式和 (8-158) 式确定。作为一种简易的直观判断方法, 我们可以绘制出以  $LR_1(\gamma)$  为纵坐标,  $\gamma$  为横坐标的二维图, 同时在  $c(\alpha)$  处画一条水平线, 以确定置信区间。

## 3. 系数估计值的渐进分布特征

估计值  $\hat{\beta} = \hat{\beta}(\hat{\gamma})$  与门槛估计值  $\hat{\gamma}$  相关联, 这似乎使得对  $\beta$  的统计推论变得相当复杂。Chan (1993) 和 Hansen (1997) 指出这种相关性并不会显著影响一阶渐进性质, 所以对  $\beta$  的统计推论可以在假设门槛估计值  $\hat{\gamma}$  就是其真实值的情况下进行。因此,  $\hat{\beta}$  的渐进分布将是正态分布, 其协方差矩阵可以用下式估计得到:

$$\hat{V} = \hat{\sigma}^2 \left( \sum_{i=1}^n \sum_{t=1}^T x_{it}^*(\hat{\gamma}) x_{it}^{*\prime}(\hat{\gamma}) \right)^{-1}. \quad (8-159)$$

虽然在构造门槛  $\gamma$  的置信区间时, 我们需要假设残差项为独立同分布的 (iid), 但是在此处构造估计系数的置信区间时, 我们完全可以放松这个假设。在允许残差项存在条件异方差的情况下,  $\hat{\beta}$  的方差-协方差矩阵的估计式为:

$$\hat{V}_h = \left( \sum_{i=1}^n \sum_{t=1}^T x_{it}^*(\hat{\gamma}) x_{it}^{*\prime}(\hat{\gamma}) \right)^{-1} \left( \sum_{i=1}^n \sum_{t=1}^T x_{it}^*(\hat{\gamma}) x_{it}^{*\prime}(\hat{\gamma}) (\hat{\varepsilon}_{it}^*)^2 \right) \left( \sum_{i=1}^n \sum_{t=1}^T x_{it}^*(\hat{\gamma}) x_{it}^{*\prime}(\hat{\gamma}) \right)^{-1}. \quad (8-160)$$

这事实上是 White 估计式。

### 8.7.3 多重门槛模型

模型 (8-140) 中仅有一个门槛, 但在许多情况下门槛的个数可能不止一个。如双重门槛模型的设定如下:

$$y_{it} = u_i + x_{it}' \beta_1 \cdot I(q_{it} \leq \gamma_1) + x_{it}' \beta_2 \cdot I(\gamma_1 < q_{it} \leq \gamma_2) + x_{it}' \beta_3 \cdot I(q_{it} > \gamma_2) + \varepsilon_{it}, \quad (8-161)$$

其中  $\gamma_1 < \gamma_2$ 。这里我们集中讨论双重门槛模型, 因为由此可以很方便地扩展到跟多门槛的情形。下面我们重点讨论三个方面的问题: (1) 双重门槛模型的估计; (2) 检验双重门槛效应的显著性; (3) 构造门槛参数  $\gamma_1$  和  $\gamma_2$  的置信区间。

#### 估计

对于给定的  $(\gamma_1, \gamma_2)$ , (8-161) 式是关于  $(\beta_1, \beta_2, \beta_3)$  的线性模型, 所以我们可以用 OLS 进行估计。因此, 对于每一组给定的  $(\gamma_1, \gamma_2)$ , 我们可以首先计算出相应的残差平方和  $S(\gamma_1, \gamma_2)$ , 这与我们对单一门槛模型的处理相似。 $(\gamma_1, \gamma_2)$  的联合 OLS 估计值就是使得  $S(\gamma_1, \gamma_2)$  最小时对应的值  $(\hat{\gamma}_1, \hat{\gamma}_2)$ 。虽然从理论上讲, 同时估计  $(\hat{\gamma}_1, \hat{\gamma}_2)$  是可行的, 但是实际操作的过程中这种做法是非常费时的。我们对  $(\gamma_1, \gamma_2)$  进行逐个搜索大约需要进行  $N^2 = (nT)^2$  次回归。

为了减小运算量, 我们采用“循环法”进行估计。在含有多个结构突变点的模型中, 这种方法可以得到参数的一致估计量, 如 Chong (1994)、Bai (1997) 以及 Bai and Perron (1998)。在我们估计多重门槛模型时, 也可以采用同样的方法。第一步, 设  $S_1(\gamma)$  为由 (8-149) 式所定义的单一门槛设定下的残差平方和,  $\gamma_1$  为使得  $S_1(\gamma)$  最小时对应的门槛估计值。Chong (1994) 以及 Bai (1997) 的研究表明无论对于  $\gamma_1$  还是  $\gamma_2$  而言,  $\hat{\gamma}_1$  都是  $\gamma_1$  的一致估计量。固定第一步得到的  $\hat{\gamma}_1$ , 估计 (8-161) 式, 则第二步的筛选标准为,

$$S_2^\gamma(\gamma_2) = \begin{cases} S(\hat{\gamma}_1, \gamma_2) & \text{若 } \hat{\gamma}_1 < \gamma_2 \\ S(\gamma_2, \hat{\gamma}_1) & \text{若 } \gamma_2 < \hat{\gamma}_1 \end{cases} \quad (8-162)$$

进而得到第二步的门槛估计值为,

$$\hat{\gamma}_2^\gamma = \arg \min_{\gamma_2} S_2^\gamma(\gamma_2). \quad (8-163)$$

由于若个别区间内的观察值过少我们无法进行有效的估计, 所以我们可以搜索 (8-162) 式的过程中限制每个区间内的最小观察值个数, 即, 若某个以门槛值划分的一个区间内的观察值少于 30 个, 我们就跳过此门槛值进行下一个门槛值的搜索。

Bai (1997) 的研究表明  $\hat{\gamma}_2^\gamma$  是渐进有效的, 但  $\hat{\gamma}_1$  的估计却不具有此性质。这是因为在估计  $\hat{\gamma}_1$  的过程中, 残差平方和中包含了我们所忽略的区间造成的影响。但由于  $\hat{\gamma}_2^\gamma$  是渐进有效的, 所以我们可以固定  $\hat{\gamma}_2^\gamma$ , 然后重新估计, 此时参数的筛选标准为,

$$S_1^\gamma(\gamma_1) = \begin{cases} S(\gamma_1, \hat{\gamma}_2^\gamma) & \text{若 } \gamma_1 < \hat{\gamma}_2 \\ S(\hat{\gamma}_2, \gamma_1) & \text{若 } \hat{\gamma}_2 < \gamma_1 \end{cases} \quad (8-164)$$

得到更新后的  $\hat{\gamma}_1$  的估计值为,

$$\hat{\gamma}_1 = \arg \min_{\gamma_1} S_1^\gamma(\gamma_1). \quad (8-165)$$

Bai (1997) 指出在结构突变模型中, 更新后的估计值  $\hat{\gamma}_1^\gamma$  是渐进有效的, 所以我们认为在门槛模型中这个结论仍然成立。

### 门槛个数的确定

在模型 (8-161) 的设定中, 可能不存在门槛, 也可能存在一个或两个门槛值, 因此我们需要对此进行检验。在上一节中, 我们采用  $F_1$  统计量来检验单一门槛效应的显著性, 并采用“自体抽样”法获得了该统计量的置信区间。如果  $F_1$  拒绝了原假设, 即存在一个门槛, 那么在模型 (8-161) 的设定中, 我们就需要作进一步的检验以便区分单一门槛和双重门槛。

我们从第二步中估计出的最小残差平方和为  $\hat{\sigma}^2 = S_2^\gamma/n(T-1)$ , 因此我们可以基于下面的统计量来检验单一门槛与双重门槛那个更为显著:

$$F_2 = \frac{S_1(\hat{\gamma}_1) - S_2^\gamma(\hat{\gamma}_2^\gamma)}{\hat{\sigma}^2} \quad (8-166)$$

该检验量事实上是一个近似的似然比检验。如果  $F_2$  的值较大, 那么我们就拒绝仅存在一个门槛值的原假设。由于在原假设下, 似然比统计检验量的渐进分布是非枢轴的 (non-pivotal), Hansen (1999) 建议采用“自体抽样法”来近似其样本分布。抽样方法和检验统计量的构造方法与我们在上一节中介绍的  $F_1$  统计量的构造方法相似。区别在于, 此时的我们的被择假设对应的模型为 (8-161) 式, 而原假设对应的模型为 (8-140)。因此, 我们产生的自抽样样本为,

$$y_{it}^\# = x'_{it}\hat{\beta}_1 \cdot I(q_{it} \leq \hat{\gamma}) + x'_{it}\hat{\beta}_2 \cdot I(q_{it} > \hat{\gamma}) + \varepsilon_{it}^\# \quad (8-167)$$

从 (8-167) 式我们可以看出,  $F_2$  的抽样分布依赖于回归参数  $\beta_1 - \beta_2$  和  $\gamma$ 。这不同于  $LR_1$  的抽样分布,  $LR_1$  并不依赖于任何回归参数。所以,  $F_2$  的抽样分布是非枢轴的, 使得我们无法保证“自体抽样法”能产生很好的近似效果。

### 置信区间的构造

最后，我们来构造两个门槛参数  $(\gamma_1, \gamma_2)$  的置信区间。Bai (1997) 研究表明，我们在本节第一部分得到的更新后的  $\hat{\gamma}_1$  的估计值与在单一门槛模型中得到的估计值具有相同的渐进分布。这启发我们可以采用与上一节相同的方法来构造置信区间。设

$$LR_2^\gamma(\gamma) = \frac{S_2^\gamma - S_2^\gamma(\hat{\gamma}_2^\gamma)}{\hat{\sigma}^2} \quad (8-168)$$

及

$$LR_1^\gamma(\gamma) = \frac{S_1^\gamma - S_1^\gamma(\hat{\gamma}_1^\gamma)}{\hat{\sigma}^2} \quad (8-169)$$

其中， $S_2^\gamma(\gamma)$  和  $S_1^\gamma(\gamma)$  分别由 (8-162) 和 (8-164) 式定义。 $\gamma_2$  和  $\gamma_1$  在  $(1 - \alpha)\%$  显著水平上的置信区间分别为一系列满足  $LR_2^\gamma(\gamma) \leq c(\alpha)$  和  $LR_1^\gamma(\gamma) \leq c(\alpha)$  的  $\gamma$  值。

#### 8.7.4 STATA 实现

Hansen (1999) 在其个人网站上发布了估计面板门槛模型的 Gauss 程序，附带有程序的说明和其实证分析中所用的数据。笔者采用 STATA8.0 软件包完成了该程序的估计、检验和绘图程序，主要包括两个部分，分别为 xtthres.ado 和 xttr\_graph.ado。由于程序代码较长，所以这里仅呈现程序的帮助文件。需要程序源文件的读者可以向笔者索取 (arlionn@163.com)。

作为这一模型的应用，国外已有数十位学者采用它进行实证分析，国内使用该模型的文献十分有限，就我所知，目前仅有两篇：魏锋和孔煜 (2005)、连玉君和程建 (2006)。



## 参考文献

- [1] Arellano, M. 1987. "Computing Robust Standard Errors for within-Groups Estimators." *Oxford Bulletin of Economics and Statistics*, 49(4), pp. 431-34.
- [2] Arellano, M. 2003. *Panel Data Econometrics*. New York: Oxford University Press.
- [3] Arellano, M and S Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies*, 58(2), pp. 277-97.
- [4] Arellano, M and O Bover. 1995. "Another Look at the Instrumental Variable Estimation of Error-Components Models." *Journal of Econometrics*, 68(1), pp. 29-51.
- [5] Arellano, M and B Honoré. 2001. "Panel Data Models: Some Recent Developments." *Handbook of Econometrics*, 5, pp. 3229-96.
- [6] Baltagi, BH. 2001. *Econometric Analysis of Panel Data*. Chichester: John Wiley & Sons.
- [7] Baum, CF. 2001. "Residual Diagnostics for Cross-Section Time Series Regression Models." *STATA JOURNAL*, 1(1), pp. 101-04.
- [8] Baum, CF; ME Schaffer and S Stillman. 2003. "Instrumental Variables and Gmm: Estimation and Testing." *STATA JOURNAL*, 3(1), pp. 1-31.
- [9] Bhargava, A; L Franzini and W Narendranathan. 1982. "Serial Correlation and the Fixed Effects Model." *The Review of Economic Studies*, 49(4), pp. 533-49.
- [10] Bruno, Giovanni S. F. 2005. "Approximating the Bias of the Lsdv Estimator for Dynamic Unbalanced Panel Data Models." *Economics Letters*, 87(3), pp. 361-66.
- [11] Cameron, AC and PK Trivedi. 2009. *Microeconometrics Using Stata*. Stata Press.
- [12] Cameron, AC and PK Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.
- [13] Canova, F. 2007. *Methods for Applied Macroeconomic Research*. Princeton University Press.
- [14] Cornelissen, T. 2008. "The Stata Command Felsdvg to Fit a Linear Model with Two High-Dimensional Fixed Effects." *STATA JOURNAL*, 8(2), pp. 170-89.
- [15] Drukker, DM. 2003. "Testing for Serial Correlation in Linear Panel-Data Models." *STATA JOURNAL*, 3(2), pp. 168-77.
- [16] Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The annals of statistics*, 7(1), pp. 1-26.
- [17] Efron, B and RJ Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

- [18] Flannery, Mark J. and Kasturi P. Rangan. 2006. "Partial Adjustment toward Target Capital Structures." *Journal of Financial Economics*, 79(3), pp. 469-506.
- [19] Frésard, Laurent and Carolina Salva. 2010. "The Value of Excess Cash and Corporate Governance: Evidence from U.S. Cross-Listings." *Journal of Financial Economics*, Forthcoming.
- [20] Greene, W H. 2000. *Econometric Analysis*, 4th Edition. New Jersey: Prentice Hall.
- [21] Hansen, LP. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, 50(4), pp. 1029-54.
- [22] Harford, Jarrad; Sandy Klasa and Nathan Walcott. 2009. "Do Firms Have Leverage Targets? Evidence from Acquisitions." *Journal of Financial Economics*, 93(1), pp. 1-14.
- [23] Hoechle, D. 2007. "Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence." *STATA JOURNAL*, 7(3), pp. 281-312.
- [24] Hsiao, C. 2003. *Analysis of Panel Data*. Cambridge University Press.
- [25] Judson, RA and AL Owen. 1999. "Estimating Dynamic Panel Data Models: A Guide for Macroeconomists." *Economics Letters*, 65(1), pp. 9-15.
- [26] Mikkelsen, WH and MM Partch. 2003. "Do Persistent Large Cash Reserves Hinder Performance?" *Journal of Financial and Quantitative Analysis*, 38(2), pp. 275-94.
- [27] Newey, WK and KD West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55(3), pp. 703-08.
- [28] Opler, T; L Pinkowitz; R Stulz and R. Williamson. 1999. "The Determinants and Implications of Corporate Cash Holdings." *Journal of Financial Economics*, 52(1), pp. 3-46.
- [29] Petersen, Mitchell A. 2009. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *Review of Financial Studies*, 22(1), pp. 435-80.
- [30] Richardson, S. 2006. "Over-Investment of Free Cash Flow." *Review of Accounting Studies*, 11(2), pp. 159-89.
- [31] Roodman, David. 2009. "How to Do Xtabond2: An Introduction to Difference and System Gmm in Stata." *STATA JOURNAL*, 9(1), pp. 86-136.
- [32] Sarafidis, V and RE De Hoyos. 2006. "Testing for Cross-Sectional Dependence in Panel-Data Models." *STATA JOURNAL*, 6, pp. 482-96.
- [33] Schaffer, M.E. 2010. "Xtivreg2: Stata Module to Perform Extended Iv/2sls, Gmm and Ac/Hac, Liml and K-Class Regression for Panel Data Models." <http://ideas.repec.org/c/boc/bocode/s456501.html>.
- [34] Sosa-Escudero, W and AK Bera. 2008. "Tests for Unbalanced Error-Components Models under Local Misspecification." *STATA JOURNAL*, 8(1), pp. 68-78.



- 
- [35] Verbeek, M. 2004. A Guide to Modern Econometrics. Wiley.
- [36] Wooldridge, JM. 2002. Econometric Analysis of Cross Section and Panel Data. MIT press.
- [37] Ziliak, JP. 1997. "Efficient Estimation with Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators." Journal of Business & Economic Statistics, 15(4), pp. 419-31.



# 一份不太长的 Stata 简介

[连玉君](#)

中山大学 岭南学院

[arlionn@163.com](mailto:arlionn@163.com)

2010-7-14

## 目 录

|   |                             |    |
|---|-----------------------------|----|
| 1 | Stata概貌 .....               | 1  |
| 2 | 为何选择Stata? .....            | 2  |
| 3 | 如何学习Stata? .....            | 4  |
| 4 | 最后的话 .....                  | 7  |
|   | 参考文献 .....                  | 7  |
|   | 附录A: 一些有用的Stata链接 .....     | 9  |
|   | 附录B: 43 个不可不知的Stata命令 ..... | 12 |
|   | 附录C: Stata视频教程.....         | 13 |

## 1 Stata 概貌

自从 2003 年开始使用 Stata 以来，我一直把“Stata”读为“Stay-ta”。有一次和一个从日本回来的朋友聊天，她把 Stata 读为“Star-ta”，让我甚感不适。经查阅，方才发现，原来“Stata”并非数个单词的缩写（因此其正确拼写为 Stata 而非 STATA），而是由“statistics”和“data”合成的一个新词，Stata 公司的员工都将其读做“Stay-ta”。从这个小小的趣闻中，可以看出 Stata 在问世之初（1985 年）的主要功能在于统计分析和数据处理。经历了二十余年的发展，Stata 已经升级到第 11.1 版（表 1），在不断强化上述功能的同时，Stata 在矩阵运算、绘图、编程等方面的功能也在不断加强。

表 1 Stata 发展历程

|      |                |      |                |
|------|----------------|------|----------------|
| 1.0  | January 1985   | 6.0  | January 1999   |
| 1.1  | February 1985  | 7.0  | December 2000  |
| 1.2  | March 1985     | 8.0  | January 2003   |
| 1.4  | August 1986    | 8.1  | July 2003      |
| 1.5  | February 1987  | 8.2  | October 2003   |
| 2.0  | June 1988      | 9.0  | April 2005     |
| 2.05 | June 1989      | 9.1  | September 2005 |
| 2.1  | September 1990 | 9.2  | April 2006     |
| 3.0  | March 1992     | 10.0 | June 2007      |
| 3.1  | August 1993    | 10.1 | August 2008    |
| 4.0  | January 1995   | 11.0 | July 2009      |
| 5.0  | October 1996   | 11.1 | June 2010      |

Source: <http://www.Stata.com/support/faqs/res/history.html>

Stata 擅长数据处理、面板数据分析、时间序列分析、生存分析，以及调查数据分析，但其它方面的功能也并不逊色（表 2）。

表 2 Stata 的功能一览

|                                         |                                       |                                           |                             |
|-----------------------------------------|---------------------------------------|-------------------------------------------|-----------------------------|
| 数据处理和绘图                                 |                                       |                                           |                             |
| <a href="#">Data management</a>         | <a href="#">Graphics</a>              |                                           |                             |
| 统计分析和检验                                 |                                       |                                           |                             |
| <a href="#">Basic statistics</a>        | <a href="#">Nonparametric methods</a> | <a href="#">Exact statistics</a>          |                             |
| <a href="#">ANOVA/MANOVA</a>            | <a href="#">其它检验方法和函数</a>             |                                           |                             |
| 回归分析                                    |                                       |                                           |                             |
| <a href="#">Linear models</a>           | <a href="#">GLM</a>                   | <a href="#">MLE</a>                       | <a href="#">GMM</a>         |
| <a href="#">Multilevel mixed models</a> | <a href="#">Panel data</a>            | <a href="#">Probit/Logit/Count</a>        | <a href="#">Time series</a> |
| 多变量模型（多元统计）                             |                                       | 抽样和模拟分析                                   |                             |
| <a href="#">Multivariate methods</a>    | <a href="#">Cluster analysis</a>      | <a href="#">Resampling and simulation</a> |                             |
| 调查分析和生存分析                               |                                       |                                           |                             |
| <a href="#">Survey methods</a>          | <a href="#">Survival analysis</a>     | <a href="#">Epidemiologists</a>           |                             |
| 编程                                      |                                       |                                           |                             |
| <a href="#">Programming language</a>    | <a href="#">Mata</a>                  | <a href="#">User-written commands</a>     |                             |

## 2 为何选择 Stata?

这是个不太容易回答的问题。Stata网站列举了数条[可能的原因](#)。Edwards (2005) 曾经非常细致地对比了Stata, SPSS和SAS的优劣。Princeton大学的Torres-Reyna博士则将四种常用软件的特征总结为表 3。整体而言, Stata具有较强的优势。

表 3 四款统计软件的对比分析

| Features          | Stata                           | SPSS                      | SAS                | R                  |
|-------------------|---------------------------------|---------------------------|--------------------|--------------------|
| Learning curve    | Steep/gradual                   | Gradual/flat              | Pretty steep       | Pretty steep       |
| User interface    | Programming/<br>point-and-click | Mostly<br>point-and-click | Programming        | Programming        |
| Data manipulation | Very strong                     | Moderate                  | Very strong        | Very strong        |
| Data analysis     | Powerful                        | Powerful                  | Powerful/versatile | Powerful/versatile |
| Graphics          | Very good                       | Very good                 | Good               | Good               |

Source: <http://dss.princeton.edu/training/StataTutorial.pdf>, p.3.

### 我为何钟情于 Stata?

就我个人的经历而言, 如下几个原因使我自 2003 年以来一直钟情于 Stata。

**Stata的数据处理功能很强大。**由于将数据导入内存后进行运算, 其速度非常快。在多个数据文件的合并和追加, 以及文字资料、时序资料, 以及调查资料的处理方面, Stata 总能以极为简洁的[命令](#)完成分析。虽然Stata管方命令仅能支持txt和xml格式数据文件的导入和导出, 但借助[Stat/Transfer](#)软件, 我们可以非常方便地实现不同软件数据格式的转换, 如Excel, Access, SPSS, SAS, Eviews, Gauss, Limdep, S-Plus, R等。我是做公司财务的, 每年 5 月, 在GTA、CCER、Wind等数据库提供商提供了最新的数据后, 我也需要更新自己的Stata数据库(我把这些数据库提供的几十个子库合并为一个名为“Arlion\_data.dta”的Stata数据文件, 并与我的合作者们分享)。借助Stata的数据处理功能, 我只需在上一年度已经完成的do-files中稍作修改即可完成数据的更新工作。整个过程仅需 2 天的时间。我无法想象, 如果没有Stata提供的merge、append、forvalues等命令, 这个数据更新的过程将会有多么痛苦。

**Stata 的 do-files 带来的便利。**我很少点击 Stata 的菜单, 也很少在命令窗口中输入命令, 我使用 do-files (当然, 每天要在这个窗口中敲入几十次 help 命令)。简单而言, Stata 的 do-files 只是一个包含了多行 Stata 命令的文本文件而已 ([U]16 Do-files, Long (2009))。有些时候, 要完成一篇文章的数据处理过程需要数周的时间, do-files 就显得格外重要, 它使得我们很容易对此前的处理过程进行修改。更为重要的是, 后续文章都可以在这个 do-files 的基础上扩展。我与搭档合作时, 每天只需通过电子邮件发送只有几 k 大小的 do-file 即可; 而我的学生们则可以通过 do-files 重现我上课时讲解的每一个估计命令; 很多学生的第一篇实证分析的论文都是在我已经完成的 do-files 基础上完成的。

**Stata 绘制的图形非常精美。**这也为回归分析提供了一种可视化的分析工具, 自

Stata10 发布以来, Stata 增加了图形编辑、多种字体支持, 以及数学符号支持等功能。Stata 可以输出十余种图片格式, 可以非常方便地插入 Word、LaTeX 等文字排版软件。即使采用点击鼠标的方式绘制图形, Stata 也会自动生成命令代码, 为图形的修改提供了极大的便利。

**Stata在编程方面提供了良好的平台。**比如, 做非线性最小二乘(NLS)、最大似然估计(MLE)、广义矩估计(GMM), 只需要设定函数形式, 编写一些简单的程序即可完成, 至于数值算法等比较复杂的技术问题, Stata都已帮你做好了。例如, 我完成的第一篇实证分析的论文便是以NLS为基础的(连玉君 and 钟经樊(2007)), 随后, 我又采用MLE完成了异质性随机边界模型(连玉君 and 苏治(2009))和双边随机边界模型(Lian and Chung(2008); 连玉君(2009))的估计。自Stata11 发布以来, [GMM](#)的实现也变得非常简单了, 你只需设定残差方程、指定工具变量, 并选择何时得权重矩阵即可完成估计。

**Stata具有良好的扩展性。**Stata具有自己的编程语言, 其所有命令都对应着一个以“.ado”为后缀的同名程序文件。对于Stata用户而言, 我们可以使用viewsource或doedit命令查看这些程序的代码。更为重要的是, 我们可以非常方便地自行编写命令, 以实现Stata官方命令的补充和扩展。这种特殊的扩展功能赋予了Stata用户极大的灵活性, 我们可以用[findit](#)命令下载到大量的[外部命令](#), 以便适时跟进最新的统计方法。这同时也推动了Stata自身的发展, 例如, Stata用户开发出的可绘制地图的命令[tamp](#), [spmap](#), [china\\_map](#)等就是一个很好的例证; 由[David Roodman](#)编写的[xtabond2](#)命令则被Stata11 设定为估计动态面板模型的官方命令([xtdpd](#), [xtdpdsys](#)); 同样, 由F. Bornhorst and C.F. Baum编写的[ipshin](#)、[levinlin](#)命令, C.F. Baum编写的[hadri](#)命令, 以及S. Merryman编写的[xtfisher](#)等用于执行面板单位根检验的命令都被Stata11 设定为官方命令[xtunitroot](#)。饮水思源, 我自己也贡献了[xtbalance](#)等命令。若想发布自己编写的Stata命令, 只需发邮件给波士顿大学的[C.F. Baum](#)教授即可。

最后, 从我身边这些老师和朋友的经验来看, Stata受到了越来越多的关爱。[我的导师](#)使用Gauss十年有余, 在 2001 年接触Stata后, 毅然改用Stata。还有很多国外的朋友, 基本上都在使用Stata。当越来越多的人开始使用Stata时, 我们的交流成本会迅速下降。

当然, 软件本身并无好还之分, 只是一个习惯的问题。关键的问题还是对统计和计量理论的掌握, 这是决定你是否能驾驭软件的关键。

## 正在消弭的 Stata 缺陷

Stata并不完美, 但她正在趋近完美。[Evan Stark](#)博士非常精辟的概括了这一特征: “You get the sense that at Stata they thought of everything, and when they or a user points out that they didn’t, they quickly provide a fix or new functionality. Although it did take me a while to understand its syntax [switching from SPSS], I did master it and statistical life became thereafter very [enjoyable](#).”

诚如[MacStats网站的评价](#)，Stata结果似乎无法像SPSS或Eviews那样非常美观地输出（或粘贴）到Word/Excel文档中。然而，得益于广大Stata用户的努力，这不再是个问题，我们可以使用[tabout](#) (Watson (2007)), [esttab](#) (Jann and Long (2010)), `logout`, `outreg2` (Jann (2005), Jann (2007)), `xml_tab` (Lokshin and Sajaia (2008)) 等命令非常方便的把Stata结果输出到Excel, [Word](#), [LaTeX](#)和HTML (Gini and Pasquini (2006)) 等文件中。连玉君博士制作的视频文件[Stata与Word、Excel、LaTeX的亲密接触](#) 非常细致地介绍了这一主题。他的另一份文档[Stata与LaTeX的完美结合](#) 则较为全面的介绍了如何将Stata结果输出到LaTeX。

在早期版本中，Stata的do-files编辑器[过于简单](#)。Stata11 发布后，其do-files编辑器已然从灰姑娘变成了[白雪公主](#)，具有了语法高亮显示、结构代码折叠、书签设定等功能，而且，对于书写大型do-files的用户而言，命令的行数也不再受到任何限制。对于中文用户而言，只需稍作[调整](#)，即可获得很好的显示效果。

Stata9 以前的版本无法对图形进行二次编辑，且图形中的可供选择的字体也非常有限。自从Stata10 和Stata11 发布以来，这两个问题得到了很好的[解决](#)。图形中的文字可以是粗体、斜体，亦可包含多种数学符号；在用户手动编辑图形时，相应的命令会自动显示在屏幕上，进而用于处理其他类似的图形。

不同于SAS等从硬盘上读取数据的统计软件，Stata是将数据调入内存后执行运算的，这使得其运算速度非常快。然而，对于经常处理高频数据和大型调查数据的用户而言，Stata的这种运算机制反而成了其缺陷——它能够处理的数据量受限于计算机的内存容量。虽然在既有的[多个Stata版本](#)中，Stata11 家族中进一步增加了[Stata/MP](#)，使其在配有多核处理器的计算机中运算速度进一步得到[提升](#)，但数据容量的限制问题仍然未能得到实质性的改进。

### 3 如何学习 Stata?

我经常会被问到“Stata 好学吗”、“我多长时间能学会 Stata”，诸如此类的问题。诚然，相比于 SPSS 和 Eviews 等软件，Stata 的门槛的确要高一些。然而，问题的关键并不在于 Stata 本身有多么难学，而在于你在统计和计量方面花费了多少时间，这与学习 Stata 所需的时间显著负相关。因此，我的回答往往是：“哦，这个不好说，如果……，其实很简单……”。

相比于十年前，现在学习 Stata 的资料已经非常丰富了。虽说殊途同归，但不同的学习路径却存在着巨大的效率差异。对于初学者而言，我的建议是，首要的问题是知道“Stata 能做什么”，继而才是“Stata 如何做什么”。

第一个问题之所以重要，是因为从本质上讲，Stata只是我们完成统计分析的工具而已，因此，其基本平台是否宽广、是否有扩展潜力，以及它提供的分析工具是否能满足你的专业需求，都是你在选择Stata之前需要深入了解的。[Stata User's Guide](#)（400 页，[中](#)

文)对这些问题做出了很好的解答,是一幅绝佳的导航图,能帮助你在短时间内了解Stata的基本架构、语法特征和核心功能。对于第二个问题,则有众多的资料可供参考:

### (1) 网络资源

在附录A中,我精选了一些链接。值得一提的有如下几个:

- **Stata官方网站。**Stata公司提供的[Web resources](#),涵盖了大量相关网络资源;其[FAQ](#)则提供了各种常见问题的解答;[Statalist](#)则是一个类似于人大经济论坛的免费的讨论区。[加入 Statalist 的方法](#)很简单,你只需要发送邮件至[majordomo@hsphsun2.harvard.edu](mailto:majordomo@hsphsun2.harvard.edu),邮件内容无需任何称谓,只需写上“**subscribe Statalist**”的字样即可。接到确认信息后,你便成为一名Statalist的成员了。当然,即使不加入,你仍然可以浏览,但不能提问。
- **UCLA(加州大学洛杉矶分校)提供的网络教程。**该网站提供的[Data Management](#)、[Graphics](#)、[Regression](#)、[Logistic Regression](#)、[Multilevel Modeling](#)、[Survey Data Analysis](#)等模块都非常出色;其[Web Books](#)、[Textbook Examples](#)模块则非常细致地呈现了几十本非常流行的统计和计量教材的Stata实例;对于LaTeX感兴趣的朋友,则可以通过[Stata Tools for LaTeX](#)模块获得诸多有用的信息;在[Graph examples](#)模块中,则列举了四十余种图形的绘制方法;最后,在[Classes and Seminars](#)模块中,你可以在线观看数十个Stata教学视频。
- **人大经济论坛。**若从人数上来讲,[人大经济论坛](#)或许是全球最大的经济类论坛了。目前,其[计量经济学板块](#)又细分出多个计量软件专题讨论区。在[Stata专版](#)已发布了4000余个讨论主题(18000余条回复),而[Stata上传下载区](#)则汇集了大量学习资料。在[统计软件培训班VIP答疑区](#)中,Stata培训班的学员所提出的问题,可以在24小时内得到详尽的回复。

### (2) 相关的书籍

自从Hamilton(1990)出版*Statistics with Stata*后,一系列将计量理论与软件操作结合起来的书籍开始相继面世,而在此之前,人们似乎都认为软件操作是件非常简单的事情。也正因为如此,很多学生在修习完了一个学年的计量经济学课程后,仍然不知道该如何完成OLS估计。为此,我列举的书籍多附有Stata实例(\*表示我的推荐程度),多数书中的范例数据都可通过Stata官方网站[下载](#)。

- **一份详细的书单。**UCLA提供了的书单:[Statistics Books for Loan](#)。
- **入门教材:**Baum(2006)\*、Newton and Cox(2009)、Chen et al.(2005)、Adkins and Hill(2008)\*;Wooldridge(2009)\*,波士顿大学的网站上提供了该书所有章节的[Stata范例](#),是一套非常好的学习资料。
- **综合性教材:**Cameron and Trivedi(2005)撰写的*Microeconometrics: Methods and applications*一书全面介绍了微观计量中的基本分析工具,其中不乏最近十年中得到广泛应用的Bootstrap、Monte Carlo模拟,以及非参数估计法。二人于2009



年出版的另一力作(Cameron and Trivedi (2009)\*)是这本书的姊妹篇,重点介绍了常用计量模型的 Stata 实现方法。

- **Stata手册**。我一直非常佩服撰写Stata手册的那些人([我的导师](#)也有相似的感觉),他们总能以最简洁的语言说清楚困扰我很久的问题。Stata11 附有 16 本电子手册,仅需统一放置于D:\stata11\utilities目录下,即可从Stata内部的帮助文件中的Also see部分直接链接到相应的PDF说明书中。作为初学者,我强烈建议你将来将 [U] 和 [D] 打印出来,反复研读。
- **统计方法**: Rabe-Hesketh and Everitt (2006)。
- **Stata 绘图**: Mitchell (2008), 非常细致地介绍了各种图形的绘制方法。
- **Stata 数据处理**: Kohler and Kreuter (2005)\*、Long (2009)\*、杨菊华 (2008)。
- **Stata 编程**: Baum (2009), 当然,该书有关数据处理的介绍也非常精彩。
- **Logit/Probit模型**: Hosmer and Lemeshow(2000)\*对相关的理论进行非常细致的介绍,是我学习Logit模型的入门教材; Long and Freese(2001)\*、Long and Freese(2006)、Hilbe(2009)则涉及了大量的Stata实例,对解读Logit/Probit模型的结果很有帮助; Rabe-Hesketh et al.(2004)提供了在GLLAMM架构下估计xtlogit, xtprobit, xtmelogit, 以及xtmepoisson模型的方法。
- **Panel Data 和多层次模型**: Stata11 手册[XT]\*, 简洁明了, 附有大量实例; Cameron and Trivedi (2009)\*、王志刚 (2008)、Rabe-Hesketh and Skrondal (2008)。
- **Mata**: Schmidheiny (2008)\*, 简洁明了介绍了 Mata 的基本用法; 详情则可参考 Stata11 手册 [M]。
- **GLLAMM**: Rabe-Hesketh et al. (2004) ([下载](#))。
- **Meta**: Sterne (2009)。
- **GLM**: Hardin et al. (2007)。
- **MLE**: Harrison (2008) (Lectures)、Gould et al. (2006)。
- **生存分析**: Cleves et al. (2008)。

### (3) Stata 视频

相比于网络教程和纸本教材,通过视频学习Stata可能是最快捷的方式了。坊间流传有两套Stata视频教程:一套是UCLA免费发布的视频教程,内容涉及Stata入门、数据处理和绘图等。[该视频教程](#)采用英文讲解,思路清晰。局限在于所涉及内容不够系统,但对于想快速入门的学生则是一份不错的参考资料。同时,藉由这份资料也可以练习一下英语听力。另一套是由中山大学岭南学院的连玉君博士制作的[Stata 视频教程](#)。该教程分为初级(36学时)、高级(48学时)和Panel data专题(12学时)三个部分。该视频教程涵盖了Stata简介、数据处理、矩阵、绘图、编程等基本操作,同时还包含了OLS、GLS、MLE、GMM、Bootstrap、Monte Carlo模拟、时间序列分析、面板数据模型等分析工具。详见[附录C](#)。

## 4 最后的话

- (1) **好脑瓜不如烂笔头。**这是一个适用于学习任何新知识的“秘诀”，对于功能强大，以敲命令为基础的Stata软件而言尤其如此。因此，你要时刻记录新学到的命令、方法和技巧，并定期整理。若能将这些手记与其他Stata用户分享，你会有更多的收获。[我的博客](#)中便提供了不少这样的笔记。
- (2) **学以致用。**在了解了 Stata 的基本功能和架构后，想要进一步提升自己的最佳途径就是动手写一篇实证分析的论文，并自始至终用 Stata 解决所有问题。这项工作的起点是一份以 txt 或 Excel 格式存储的原始数据文件，中间过程完整地记录于一个 do-files 文档中，最终的分析结果要自动输出到 Word, Excel 或 LaTeX 文档中。
- (3) **不耻下问。**这个不用多言了，你只需克服“不耻”，进而多花些精力考虑考虑该如何提问即可（注：很多人不会提问）。

## 参考文献

- Adkins, L., R. Hill. Using stata for principles of econometrics[M]. Wiley, 2008.
- Baum, C. An Introduction to Modern Econometrics using Stata[M]. Stata Corp, 2006.
- Baum, C. An Introduction to Stata Programming[M]. Stata Press, 2009.
- Cameron, A., P. Trivedi. Microeconometrics: methods and applications[M]. Cambridge University Press, 2005.
- Cameron, A., P. Trivedi. Microeconometrics Using Stata[M]. Stata Press, 2009.
- Chen, X., P. Ender, M. Mitchell, C. Wells, 2005, Stata Web Books: Regression with Stata (<http://www.ats.ucla.edu/stat/stata/webbooks/reg/default.htm>).
- Cleves, M., W. Gould, R. Gutierrez, Y. Marchenko. An introduction to survival analysis using Stata[M]. Stata Press, 2008.
- Edwards, M., **2005**, SPSS, STATA, and SAS: Flavours of Statistical Software, *URI*: <http://hdl.handle.net/1873/250>.
- Gini, R., J. Pasquini, **2006**, Automatic generation of documents, *STATA JOURNAL*, 6 (1): 22-39.
- Gould, W., J. Pitblado, W. Sribney. Maximum likelihood estimation with Stata[M]. Stata Press, 2006.
- Hamilton, L. Statistics with Stata[M]. Brooks/Cole, 1990.
- Hardin, J., J. Hilbe, J. Hilbe. Generalized linear models and extensions[M]. Stata Press, 2007.
- Harrison, G. Maximum Likelihood Estimation of Utility Functions Using Stata[M]. University of Central Florida, <http://web.bus.ucf.edu/documents/economics/workingpapers/2006-12.pdf>, 2008.
- Hilbe, J. Logistic regression models[M]. Chapman & Hall/CRC Press, 2009.
- Hosmer, D., S. Lemeshow. Applied Logistic Regression[M]. New York: John Wiley & Sons, Inc, 2000.
- Jann, B., **2005**, Making regression tables from stored estimates, *STATA JOURNAL*, 5 (3): 288-308.
- Jann, B., **2007**, Making regression tables simplified, *STATA JOURNAL*, 7 (2): 227-244.

- Jann, B., J. Long, **2010**, Tabulating SPost results using estout and esttab, *STATA JOURNAL*, 10 (1): 46-60.
- Kohler, U., F. Kreuter. Data Analysis Using Stata[M]. Stata Press, 2005.
- Lian, Y., C.-F. Chung, **2008**, Are Chinese Listed Firms Over-Investing?, *SSRN working paper*, Available at SSRN: <http://ssrn.com/abstract=1296462>.
- Lokshin, M., Z. Sajaia, **2008**, Creating print-ready tables in Stata, *STATA JOURNAL*, 8 (3): 374-389.
- Long, J. The workflow of data analysis using Stata[M]. Stata Press, 2009.
- Long, J., J. Freese. Regression models for categorical dependent variables using Stata[M]. Stata press, 2001.
- Long, J., J. Freese. Regression Models for Categorical Dependent Variables using Stata[M]. Stata press, 2006.
- Mitchell, M. A visual guide to Stata graphics[M]. Stata Corp, 2008.
- Newton, H., N. Cox. Seventy-six Stata Tips[M]. Stata Press, 2009.
- Rabe-Hesketh, S., B. Everitt. A Handbook of Statistical Analyses Using Stata[M]. Chapman & Hall/CRC, 2006.
- Rabe-Hesketh, S., A. Skrondal. Multilevel and Longitudinal Modelling Using Stata (Second Edition)[M]. Stata Press, 2008.
- Rabe-Hesketh, S., A. Skrondal, A. Pickles, **2004**, GLLAMM manual, *UC Berkeley Division of Biostatistics working paper series 160*, <http://www.bepress.com/ucbbiostat/paper160/>.
- Schmidheiny, K., **2008**, Coding with Mata in Stata, *Lectures in Universitat Pompeu Fabra*, <http://kurt.schmidheiny.name/teaching/statamata.pdf>.
- Sterne, J. Meta-analysis in stata: An updated collection from the stata[M]. Stata Press, 2009.
- Watson, I., **2007**, Publications quality tables in Stata: a tutorial for the tabout program, *Working Paper*, [http://fmwww.bc.edu/repec/bocode/t/tabout\\_tutorial.pdf](http://fmwww.bc.edu/repec/bocode/t/tabout_tutorial.pdf).
- Wooldridge, J. Introductory econometrics: A modern approach[M]. South Western Cengage Learning, 2009.
- 王志刚. 面板数据模型及其在经济分析中的应用[M]. 北京: 经济科学出版社, 2008.
- 杨菊华. 社会统计分析与数据处理技术——STATA 软件的应用[M]. 北京: 中国人民大学出版社, 2008.
- 连玉君. 中国上市公司投资效率研究[M]. 北京: 经济管理出版社, 2009.
- 连玉君, 苏治, **2009**, 融资约束、不确定性与上市公司投资效率, *管理评论*, (01): 19-26.
- 连玉君, 钟经樊, **2007**, 中国上市公司资本结构动态调整机制研究, *南方经济*, (01): 23-38.

## 附录 A：一些有用的 Stata 链接

### I. Websites of Stata CP

Stata website: <http://www.Stata.com> [导航图](#)  
Sata resources: <http://www.Stata.com/links/resources1.html> (大量网络教程链接)  
Stata journal: <http://www.Stata.com/support/faqs/res/sj.html> [中](#)  
Stata library: <http://www.ats.ucla.edu/stat/Stata/library/>  
Statalist archive: <http://www.hsph.harvard.edu/cgi-bin/lwgate/STATALIST/archives/>  
Stata FAQs: <http://www.Stata.com/support/faqs/>  
Stata statistics FAQs: <http://www.Stata.com/support/faqs/stat/>  
Stata listserver: <http://www.Stata.com/support/Statalist/>  
Stata discussion list: [Statalist@hsphsun2.harvard.edu](mailto:Statalist@hsphsun2.harvard.edu)  
Stata bookstore: <http://www.Stata.com/bookstore/> [Example Datasets](#) [中](#)  
Stata Manual: <http://www.Stata-press.com/manuals/> [Example Datasets](#) [中](#)

### II. Websites in China

- 人大经济论坛（国内最大的经济论坛）
  - 人大经济论坛Stata专版: <http://www.pinggu.org/bbs/forum-67-1.html>
  - 人大经济论坛Stata上传下载区:  
<http://www.pinggu.org/bbs/forum-121-1.html> [汇](#)
  - 人大经济论坛统计软件培训班VIP答疑区（针对[Stata视频教程](#)学员）:  
<http://www.pinggu.org/bbs/forum-114-1.html> (所有Stata问题 24 小时内回复)

### III. [UCLA](#) Academic Technology Services (极力推荐)

- [Classes and Seminars](#)
  - [Introduction to Stata 10 with movies](#)
  - [Introduction to Stata](#) (for Stata version 8 and 9) *with movies*
  - [Regression with Stata](#)
  - [Logistic Regression with Stata](#) *with movies*
  - [Beyond Binary Logistic Regression with Stata](#) *with movies*
  - [Factor Variables and Interactions in Stata 11](#)
  - [Main Effects and Interactions for Logit Models in Stata](#) *with movies*
  - [Multiple Imputation in Stata, Part 1](#)
  - [Multiple Imputation in Stata, Part 2](#)
  - [Survey Data Analysis with Stata 8](#) *with movies*
  - [Introduction to Survey Data Analysis with Stata 9](#)
  - [Applied Survey Data Analysis with Stata 9](#) with [movie](#) and [mp3](#)
  - [Survival Analysis Using Stata](#)
  - [Graphics using Stata 8](#) *with movies*
  - [Introduction to Programming in Stata](#)
  - [What's New in Stata 8](#)
  - [What's New in Stata 9](#)
  - [What's New in Stata 10](#)

- Links by Topic
  - [Data Management](#)
  - [Graphics](#)
  - [ANOVA](#)
  - [Regression](#)
  - [Logistic \(and Categorical\) Regression](#)
  - [Count Models](#)
  - [Multilevel Modeling](#)
  - [Survival Analysis](#)
  - [Survey Data Analysis](#)
- [Frequently Asked Questions](#) (FAQ)
- Statistical Analysis
  - [Data Analysis Examples](#) (绝佳的数据处理专题)
  - [Annotated Output](#) (详细解读Stata输出结果)
  - [Textbook Examples](#) (包含十余本教科书的Stata实例)
  - [Web Books](#) (两本Stata网络教程)
  - [What statistical analysis should I use?](#) (常用统计分析的Stata实例)

#### IV. [Stata Portal](#) (a comprehensive links)

- 不错的入门资料：
  - [Getting Started with Stata.](#)
  - [Introduction to STATA with Econometrics in Mind](#)
- Stata and Related Resources
  - [Stata Web Site](#) (including)
    - [Links to other Stat Resources](#), [NetCourses](#), [Stata Press](#), [Product Information](#), [sample Stata Session](#), [Capabilities](#), [The Stata Journal](#)
  - [Frequently Asked Questions](#) from Stata Corporation
  - [Statalist](#), the Stata listserv
    - [Information from Stata Corp](#)
    - Weblog: <http://Statalist.blogspot.com>
    - RSS Feed: <http://feeds.feedburner.com/Statalist>
  - [Stata Tips](#) (几十个Stata应用的小贴士)
- Course Notes
  - [ED 230A: Introduction to Research Design and Statistics](#) from Phil Ender
  - [ED 230BC: Linear Statistical Models](#) from Phil Ender
  - [ED 231A: Multivariate Analysis](#) from Phil Ender
  - [ED 231C: Applied Categorical & Nonnormal Data Analysis](#) from Phil Ender
  - [e-Tutorial on Stata for Econ 508](#) from University of Illinois
- Textbook Examples
  - [Stata Textbook Examples: Introductory Econometrics by Jeffrey Wooldridge](#) from Boston College
- Beginning Stata Tutorials
  - [Introduction to Stata](#), Jeroen Weesie, Utrecht University, Netherlands

- [Getting Started with Stata for MS Windows: A Brief Introduction](#), Robert Yaffee, New York University, USA
- [Online Help for Stata](#) from DSS at Princeton University
- [Introduction to Stata](#) from the Department of Economics at Princeton University
- [Publications on Statistical Software](#), from University of Wisconsin-Madison Social Science Computing Cooperative
- [Analysis of Survey Data for Social Science Research](#) (introduces Stata), a collaborative project of the University of Cape Town and the University of Michigan.
- [Stata Programming: Data Management](#) from The Carolina Population Center at UNC Chapel Hill
- Data Files
  - [Data in Stata format from the Center for International Development](#), Harvard University, USA
- Statistical
  - [Survival Analysis with Stata: Course EC968](#), Institute for Social and Economic Research, University of Essex, UK
  - [Generalized Linear Models](#) from Princeton University.
  - [S-Post: Post Estimation Commands in Stata](#) from J. Scott Long and J. Freese
  - [gllamm for complex problems](#) by [Stas Kolenikov](#)
  - [Statistics and Social Science support](#) from NYU Information Technology Services
- Stata Programs
  - [SSC Stata Program Archive](#) Boston College Department of Economics
  - [Statistical Software](#) from Gary King
  - [Stata Code by Christopher Ferrall](#), Queen's University, Canada.
  - [Stata program by Tony Brady](#), Sealed Envelope Ltd
  - [Generalized Linear Latent and Mixed Models \(GLLAMM\)](#) by Sophia Rabe-Hesketh, Kings College London
  - [Stata Software](#) by Nicola Orsini, Institute of Environmental Medicine, Karolinska Institutet
- Other
  - [Some notes on text editors for Stata users](#), Statalist members
  - [StyleRules - Suggestions on Programming Style](#) by Nicholas J. Cox
  - [A SAS User's Guide to Stata](#) courtesy of The Carolina Population Center at UNC Chapel Hill
  - [A review of random effects modelling in Stata](#) from the Centre for Multilevel Modeling

## 附录 B: 43 个不可不知的 Stata 命令

虽然Stata已经历了二十余年的发展, 命令不断增加, 但牢记如下 43 个基本命令却是作为一个Stata用户的立身之本 (Source: [Stata 11 Manual](#), [U]27, p.375):

### Getting online help

help, hsearch,  
net search, search [U] 4 Stata's help and search facilities

### Keeping Stata up to date

ado, net, update [U] 28 Using the Internet to keep up to date  
adoupdate [R] adoupdate

### Operating system interface

pwd, cd [D] cd

### Using and saving data from disk

save [D] save  
use [D] use  
append, merge [U] 22 Combining datasets  
compress [D] compress

### Inputting data into Stata

input [D] input  
edit [D] edit  
infile [D] infile (free format); [D] infile (fixed format)  
infix [D] infix (fixed format)  
insheet [D] insheet

### Basic data reporting

describe [D] describe  
codebook [D] codebook  
list [D] list  
browse [D] edit  
count [D] count  
inspect [D] inspect  
table [R] table  
tabulate [R] tabulate oneway and [R] tabulate twoway

### Data manipulation [U] 13 Functions and expressions

generate, replace [D] generate  
egen [D] egen  
rename [D] rename  
drop, keep [D] drop  
sort [D] sort  
encode, decode [D] encode  
order [D] order  
by [U] 11.5 by varlist: construct  
reshape [D] reshape

### Keeping track of your work

log [U] 15 Saving and printing output—log files  
notes [D] notes

### Convenience

display [R] display

## 附录 C: Stata 视频教程

自 2007 年以来,人大经济论坛陆续推出了SAS、Eviews、Stata等软件的[视频教程](#)。相比于传统的教科书和课堂授课方式,视频教学大大降低了学习统计软件的门槛,因而受到了广大学员的一致好评。

[Stata视频教程](#)由中山大学岭南学院的连玉君博士制作,分为初级、高级和Panel data专题三个部分,是一套学习计量经济学和Stata应用的绝佳教程。

**(a) Stata 初级视频教程。**主要介绍 Stata 的操作方法,包括 Stata 入门、数据处理、绘图、矩阵和编程初步五个部分,共计 36 个学时,全面介绍了 Stata 的基本操作方法。

相关链接: [说明书](#) <http://baoming.pinggu.org/Default.aspx?id=16> (大纲和试听视频)

**(b) Stata 高级视频教程。**主要介绍各种计量模型的基本思想及其在 Stata 中的实现方法,包括 OLS、GLS、MLE、IV-GMM、时间序列分析、面板模型、Stata 高级编程、Bootstrap 和 Monte Carlo 模拟等内容,共计 48 个学时,全面的覆盖了计量经济学的核心内容。该视频的特点是以实 Stata 证分析为导向,视频中介绍了大量的应用实例。

相关链接: [说明书](#) <http://baoming.pinggu.org/Default.aspx?id=25> (大纲和试听视频)

**(c) Panel Data 专题。**重点介绍近十年应用非常广泛的各种面板模型,包括固定效应模型、随机效应模型、IV-GMM 估计、异方差和序列相关、动态面板模型、随机边界面板模型等,并采用模拟的方式非常细致地呈现了各种模型的小样本性质,对于深入理解面板模型的理论基础颇有裨益。该视频包含 12 个视频文件,每个文件 40-60 分钟。

相关链接: [目录](#) <http://baoming.pinggu.org/Default.aspx?id=26> (大纲和试听视频)