

# 从零开始的加密货币之路

ABSTRACT. 根据本人使用五百块钱自从去年十二月起的交易来看，区块链的加密货币就好比游戏氪金，纯纯的傻呗。

## 1. 引言

回归分析作为统计学和机器学习中的基础工具，在建模与预测中应用广泛。然而，回归结果的可靠性往往受数据质量和变量特征的影响。影响性分析 (Influence Analysis) 和复共线性分析 (Collinearity Analysis) 是两个重要的回归诊断领域，用以识别可能扭曲模型估计的异常观测点和变量相关问题。

影响性分析主要针对观测值，以检测个别数据点对回归结果的影响；复共线性分析则关注自变量间的线性相关性，以评估多重共线性对参数估计精度的影响。

在计量经济学、金融风险评估、生物医学统计等领域，这两种分析方法被广泛运用。例如，在金融时间序列建模中，离群点和强相关自变量会严重影响模型预测能力；在医学统计中，多重共线性可能导致回归系数难以解释。在本报告中，我们将从定义、方法、数学原理和应用注意事项等方面，系统阐述影响性分析与共线性分析的理论基础，并通过实际案例分析其应用效果与局限性。

## 2. 影响性分析

**2.1. 定义.** 在回归分析中，如果删除某个观测点后，模型参数估计或拟合结果发生显著变化，则该观测点被称为“影响观测”。换言之，影响观测是指其包含与否会“明显地改变”回归模型结果的数据点。

影响性分析 (Influence Analysis) 即旨在量化并检测这些异常观测，常结合离群值分析 (检测残差较大的观测) 和杠杆值分析 (检测自变量值极端的观测) 来全面评估观测对模型的影响。”影响观测“往往具有较大的残差 (离群点) 或较高的杠杆值 (极端自变量)，两者同时存在时对模型影响尤为显著。

**2.2. 主要方法.** 影响性分析中常用的统计量包括 Cook' s 距离 (Cook' s Distance)、DFBETAS、DFFITs 等。它们基于“删除观测后模型变化”的思想来衡量每个观测的影响程度：

Cook' s 距离：由 Cook (1977) 提出，用于综合衡量第  $i$  个观测对所有回归系数的影响。形式上， $D_i$  可定义为删除第  $i$  个观测前后预测值差异的归一化度量。常用公式为：

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{p MSE}$$

其中  $\hat{y}_j$  为使用全量数据拟合模型预测的第  $j$  个值， $\hat{y}_{j(i)}$  为删除第  $i$  个观测后重新拟合得到的预测， $p$  为参数个数， $MSE$  为均方误差。等价地， $D_i$  也可写为：

$$D_i = \frac{e_i^2}{p \hat{\sigma}^2} \frac{h_{ii}}{(1 - h_{ii})^2},$$

其中  $e_i$  是第  $i$  个观测的残差， $h_{ii}$  是该观测的杠杆值 (Hat 值)， $\hat{\sigma}^2$  为残差方差估计。Cook' s 距离值越大，说明第  $i$  个观测对模型的整体拟合影响越大。实践中常用经验阈值，如  $D_i > 1$  或  $D_i > 4/(n - p)$ ，来判定影响点。

DFBETAS：衡量删除单个观测前后，各回归系数的变化程度。对于第  $i$  个观测和第  $j$  个系数，定义：

$$DFBETA_{ij} = \frac{\beta_j - \beta_{j(i)}}{s_{(i)} \sqrt{c_{jj}}},$$

其中  $\beta_j$  为使用全数据拟合的第  $j$  个系数估计， $\beta_{j(i)}$  为删除第  $i$  个观测后的估计， $s_{(i)}$  为不含第  $i$  观测时的残差标准误， $c_{jj}$  为  $(X^T X)^{-1}$  的第  $j$  个对角元素。DFBETAS 是标准化的系数差值，对观测  $i$

影响第  $j$  个系数的程度进行量化。常用规则为：若  $|\text{DFBETA}_{ij}| > 2/\sqrt{n}$ （或更保守的  $2/\sqrt{n}$ ），说明第  $i$  个观测在第  $j$  个系数上有较大影响。

DFFITS：衡量删除观测对自身预测值的影响。定义为：

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{h_{ii}}},$$

其中  $\hat{y}_i$  为全数据下对第  $i$  点的预测， $\hat{y}_{i(i)}$  为删除该点后重新拟合得到的预测值， $h_{ii}$  为杠杆。DFFITS 衡量第  $i$  点对其自身拟合值的影响程度。一般经验准则是，当  $|\text{DFFITS}_i| > 2\sqrt{p/n}$  时，该点被认为具有显著影响。

此外，还可利用标准化残差和学生化残差检测离群点，以及 CovRatio 检验删除观测对协方差矩阵的影响。以上多种影响统计量相互补充，通常一起使用可全面识别潜在影响点。

影响性分析的数学原理基于“逐点删除”的思想，考察删除观测  $i$  后参数估计和预测值的变化。设原模型估计为  $\hat{\beta} = (X^T X)^{-1} X^T y$ ，删除第  $i$  个观测后重新估计得  $\hat{\beta}_{(i)}$ 。通过矩阵运算可推导：

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{e_i}{1 - h_{ii}} (X^T X)^{-1} x_i,$$

其中  $x_i$  为第  $i$  个观测的自变量向量， $e_i$  为残差， $h_{ii} = x_i^T (X^T X)^{-1} x_i$ 。由此可得到 Cook' s 距离的另一种表达式：

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (X^T X) (\hat{\beta} - \hat{\beta}_{(i)})}{p \hat{\sigma}^2}.$$

展开上述公式，可得到上述关于  $e_i$  和  $h_{ii}$  的表达形式。类似地，DFBETAS 和 DFFITS 可分别通过  $\hat{\beta} - \hat{\beta}_{(i)}$  和  $\hat{y} - \hat{y}_{(i)}$  推导而来。关键在于，删除高杠杆或大残差的观测可对  $\hat{\beta}$  和拟合值产生显著偏移，使这些统计量异常增大。因此，从几何上看，影响性分析结合了残差（离群程度）和杠杆（位置极端性）两个因素，通过上述公式精确度量每点的综合影响力。

**2.3. 应用场景及注意事项.** 影响性分析通常用于回归诊断和数据清洗中。在有时序或截面金融数据中，一些异常波动日或极端交易量可能成为影响点；在医疗数据中，测量误差或特殊病例也可能引入影响观测。通过 Cook' s 距离、DFBETAS 和 DFFITS 可以识别出这些影响点，研究人员可进一步检查数据质量，或选择稳健回归等替代方法降低其影响。需要注意的是，影响性统计量本身也受数据结构影响。例如，当自变量高度共线性时，Cook' s 距离等量可能会错误标记多个点为影响点。因此，通常在模型诊断时需同时考虑复共线性状况。在使用阈值判断影响点时，应结合研究背景和经验（如 Cook' s D 常用  $D_i > 4/(n - p)$  或 DFBETAS 标准）。此外，删除影响观测前需谨慎判断，确保它们确实是异常，而非承载关键信息的有效数据。如是测量误差，可考虑删除；若反映某种潜在机制，应考虑使用分组或混合效应模型而非盲目删除。

### 3. 复共线性分析

**3.1. 定义.** 复共线性（Multicollinearity）是指多个自变量之间存在高度线性相关的现象。通俗地说，当自变量相互近似线性组合时，回归模型的参数估计会变得不稳定，其方差大幅增大，导致估计结果缺乏统计意义。共线性并不降低模型的拟合优度，但会使个别系数的置信区间变宽，t 检验失效，变量重要性难以确定。因此，检测和处理共线性是回归分析的关键步骤。

**3.2. 判别方法.** 常用的共线性诊断方法包括方差膨胀因子（VIF）和条件数（Condition Number）等。

方差膨胀因子（VIF）：对每个自变量  $i$ ，将其作为因变量，其余自变量为自变量，做辅助回归并计算  $R_i^2$ 。定义

$$\text{VIF}_i = \frac{1}{1 - R_i^2},$$

其中  $R_i^2$  为第  $i$  个自变量与其他自变量回归的决定系数。 $\text{VIF}_i$  衡量第  $i$  个系数方差因共线性而被放大的倍数。如果自变量之间无相关，则  $R_i^2 = 0$ ， $\text{VIF}_i = 1$ ；若  $R_i^2$  接近 1， $\text{VIF}_i$  趋于无穷大，或者当某个  $\text{VIF} > 10$ （或更保守的 5）时，该变量存在严重复共线性问题，需要处理。

条件数 (Condition Number): 考虑自变量矩阵  $X$  的奇异值或  $X^T X$  的特征值。令  $\lambda_{\max}, \lambda_{\min}$  分别为  $X^T X$  的最大和最小特征值, 则条件数定义为  $\kappa = \sqrt{\lambda_{\max}/\lambda_{\min}}$ 。条件数越大, 说明  $X$  越“接近病态”, 即存在线性依赖。当  $\kappa$  超过某阈值 (如 10 30) 时, 表明存在中度以上复共线性。Belsley 等人建议, 当条件数约为 10 时即可注意共线性, 而超过 100 时将导致严重的数值不稳定 (David Belsley 这个比是谁呢, 我也不知道啊, 查了一下是 MIT 毕业的在 Boston College 工作的数据分析/经济学高手, 所以这笔的话应该很可信了)。此外, 利用特征分解还可得到每个特征向量对应自变量方差的方差比例 (variance decomposition proportion), 用于定位参与共线性的具体变量组合。

**3.3. 原理.** VIF 的推导基于协方差矩阵: 在普通最小二乘中, 参数估计的方差矩阵为  $\text{Var}(\hat{\beta}) = \sigma^2(X^T X)^{-1}$ 。对于第  $i$  个自变量, 其系数  $\beta_i$  的方差是  $\sigma^2 c_{ii}$ , 其中  $c_{ii}$  为  $(X^T X)^{-1}$  的第  $i$  个对角元。若将该自变量单独回归, 系数的方差为  $\sigma^2/(n-1)\text{Var}(x_i)$ 。通过多元回归的方差膨胀因子定义可导出  $\text{VIF}_i = c_{ii}(n-1)\text{Var}(x_i)/\sigma^2 = 1/(1-R_i^2)$ 。

条件数的计算源自矩阵理论: 将  $X$  列中心化并标准化 (使每列单位方差) 后,  $X^T X$  的特征值  $\{\lambda_j\}$  刻画了自变量的独立性程度。若某些特征值  $\lambda_k$  接近 0, 则  $X$  列几乎线性相关, 回归方程不稳定。条件数  $\sqrt{\lambda_{\max}/\lambda_{\min}}$  即衡量了这种接近相关的程度; 如条件数大, 表示最小特征值极小, 即存在高度线性依赖。通过特征向量分解, 还可计算每个自变量在各特征向量下的方差比例, 用于诊断共线性的来源。

**3.4. 处理方法.** 遇到严重复共线性时, 常用的方法包括岭回归 (Ridge Regression) 和主成分回归 (Principal Component Regression, PCR) 等 (唉唉, 这俩玩意儿好像数理统计学学过啊, 虽然我都忘了, 我觉得这个数理统计这课就离谱, 扯了一堆蛋不让我们知道这些什么回归干什么用)。

岭回归 (Ridge Regression): 在最小二乘目标函数中加入二次惩罚项, 即最小化  $\|y - X\beta\|^2 + k\|\beta\|^2$ , 其中  $k > 0$  为岭参数。其解为

$$\hat{\beta}_{\text{RR}} = (X^T X + kI)^{-1} X^T y.$$

通过引入  $kI$ , 避免了  $(X^T X)$  奇异或接近奇异 (多重共线) 时的数值问题。岭回归获得的是有偏估计, 但可以显著降低估计方差。已有研究指出, 在适当选择  $k$  的情况下, 岭估计的均方误差 (MSE) 通常低于普通最小二乘估计。具体而言, 当复共线导致 OLS 方差急剧增大时, 岭回归通过引入偏差换取更小的方差, 从而总体上提高预测精度。

主成分回归 (PCR)\*\*: 先对自变量矩阵  $X$  进行主成分分析, 将其分解为正交的主成分。通常选取前几个拥有最大方差 (对应最大特征值) 的主成分  $\{Z_1, Z_2, \dots, Z_m\}$  作为新的预测变量, 然后对  $y$  做回归:  $y \sim Z_1 + \dots + Z_m$ 。由于主成分彼此正交, PCR 消除了自变量间的相关性。选取的主成分数  $m$  应小于原变量数, 以滤除引起共线性的低方差方向。主成分回归保持了主要信息, 但舍弃了与响应变量贡献较小的方向, 从而缓解了复共线性问题。

除以上两种方法外, 还有偏最小二乘回归 (Partial Least Squares)、弹性网、逐步变量选择等策略用于处理共线性。一般原则是在提高模型稳定性的同时, 兼顾解释性, 并结合领域知识决定是否剔除或合并高度相关的变量。

## 4. 应用案例分析

**4.1. 复共线性分析.** 论文背景: 张鹏程等人 (2024) 在《中国假期效应对加密货币市场投资者情绪的影响》一文中, 使用固定效应模型研究中国法定节假日对加密货币收益率的影响。他们收集了 2017 至 2022 年 1 月 1 日间前 100 大市值加密货币的每日交易数据, 并构建了包含节假日哑变量和投资者情绪指标的面板回归模型。该研究背景为: 由于中国法定节假日市场关闭, 可能导致资金外流至加密货币市场; 同时构建了社交媒体情绪指标, 考察情绪对节假日效应的调节作用。

RESEARCH

Open Access

# Investor sentiment and the holiday effect in the cryptocurrency market: evidence from China



Pengcheng Zhang<sup>1†</sup>, Kunpeng Xu<sup>2†</sup>, Jian Huang<sup>3</sup> and Jiayin Qi<sup>4\*</sup>

FIGURE 1.

复共线性分析应用：研究者在模型构建中注意到可能的多重共线性问题，例如将宏观经济政策不确定性（EPU）等宏观时间序列与加密货币面板数据混合时可能产生共线性。为此，他们首先对自变量做了 Pearson 相关分析，发现部分预测变量之间存在潜在相关性。随后，作者进一步进行了方差膨胀因子（VIF）测试：对研究假设（H1–H3）所用的回归模型逐一计算 VIF 值。结果显示，所有自变量的 VIF 都较低（一般远小于常用阈值 5 或 10），说明回归模型中的多重共线性“不是问题”。因此，作者得出结论，在其设定的模型中，共线性程度有限，对估计无显著影响。

As elucidated in the correlation analysis presented in Table 3, the highest correlation among the independent variables was established between Cap and Volume ( $r=0.824$ ), while the lowest correlation was set between Cap and BTCD ( $r=-0.413$ ). Because none of the Pearson coefficients exceeds 0.9, multicollinearity poses no problem for subsequent econometric estimation (Batranea 2021a). We conducted variance inflation factor (VIF) tests for the regression models corresponding to H1–H3 to further assess the potential presence of multicollinearity in our regression models. The results show that the VIF values of all variables do not exceed 5, indicating that the multicollinearity problem in this study is not a problem.<sup>9</sup>

FIGURE 2.

评价与启示：该论文的共线性分析表明，在包含多个市场和宏观变量的加密货币面板数据模型中，进行共线性诊断是必要的步骤。它的优点是使用了直观的相关系数矩阵和 VIF 相结合的方法，对潜在的线性相关问题进行验证。然而，VIF 仅检测整体相关性，其判断依赖于阈值设置；在该研究中 VIF 值较低，但并未说明具体阈值或敏感性分析。此外，如果某些变量天然相关（例如多种市场情绪指标），可以考虑应用主成分回归或正则化回归以提取共同信息。总体而言，该案例提醒研究者：即使在数据量大、维度多的加密货币研究中，也应检查并报告复共线性诊断结果，以确保回归系数解释的稳健性。

**4.2. 影响性分析.** 论文背景：Henaff 等人（2022）在《Frontiers in Blockchain》上发表了题为“社区影响力：狗狗币与比特币的推特对比分析”的文章，研究社交媒体信息对加密货币价格的影响。在分析使用多种回归模型预测加密货币价格时，作者关注了数据质量问题，并进行了回归诊断。该研究使用线性模型（包含推文数量等变量）进行价格预测，重点考察社交媒体指标对不同币种的影响。



# Community Impact on a Cryptocurrency: Twitter Comparison Example Between Dogecoin and Litecoin

Edouard Lansiaux<sup>1\*</sup>, Noé Tchagaspian<sup>1</sup> and Joachim Forget<sup>1,2</sup>

<sup>1</sup>Global Variations, Geneva, Switzerland, <sup>2</sup>Assemblée Nationale, Paris, France

FIGURE 3.

影响性分析应用：论文中在回归结果部分进行了残差与杠杆诊断。作者绘制了“残差对杠杆”图，并发现：每个样本（狗狗币和莱特币）数据集中均存在一个 Cook's 距离大于 1 的点。具体而言，对于这两组数据，他们定义“可疑点”为学生化残差绝对值大于 2 或 Cook's 距离大于 1 的观测。该点对应的观测对模型系数的确定产生了“非常强的作用”。研究者强调，对这些观察到的影响点，没有统一的解决方法，因此在后续分析中转而使用机器学习方法（如 LSTM 神经网络）来进行价格预测。

We have also identified the suspect points, which are the points whose studentized residual is greater than 2 in absolute value and/or the Cook's distance is greater than 1 (**Figure 7B, Figure 8B**). In the latter case, the point contributes very/too strongly to the determination of coefficients of the model compared to others. However, there is no one-size-fits-all method for dealing with these types of stitches. Thus, a machine learning modelization, as performed then, is required.

FIGURE 4.

评价与启示：该案例中的影响性分析展示了实际经济金融数据中异常值识别的过程。优点在于作者使用了 Cook's 距离和学生化残差相结合的方法，明确识别出了数据集中的极端影响点。他们直观地说明了该点“使回归系数倾斜”，并对其进行了讨论。缺点是分析较为描述性，未进一步量化删除该点对模型性能的具体影响，也未探讨如何处理（仅提及使用其他模型）。此外，仅依靠单一阈值（Cook's  $s > 1$ ）可能比较粗糙，在样本量变化时需要调整。总体来说，该案例表明即使在加密货币领域，传统的回归诊断方法也能有效识别出异常点，为后续模型选择或稳健化提供依据。研究者应根据分析目的，决定是否删除或调整此类影响观测，并考虑使用稳健统计或替代模型来缓解影响。

## 5. 结论与个人思考

影响性分析和复共线性分析是回归诊断中必不可少的工具。影响性分析通过 Cook's 距离、DFBETAS、DFITS 等量度，识别出对模型影响过大的个体观测；复共线性分析则通过 VIF、条件数等指标，揭示自变量的多重线性相关问题。两者在理论上互补：前者强调数据层面的异常，后者关

注变量间的结构关联。实践中，应当同时进行这两方面的检查，以确保模型估计稳健可靠。如加密货币市场和医疗等领域的案例所示，诊断结果为模型修正提供了指导。例如剔除极端点、采用岭回归或PCR以解决共线性问题，或转用稳健回归和机器学习方法以提高预测性能。

以及这个加密货币啊，有闲钱了当游戏氪金就行，几千块都没事，反正就当游戏氪金，真赚上钱了那说明运气好我建议考个公务员，毕竟公务员考试也是筛选运气好的人提升国运（doge）。

## 6. 小组分工

本次小组作业由三组人员（庞云涵、罗嘉宝、徐梓乔、熊高贤、李玉焜、孙俊昊、任行、牛志远）共同完成，其中后两位同学完成了报告的构思和内容的筛选，前三位同学进行了报告的编排，中间三位同学进行了内容的查找和文章主要内容的撰写，大家都收获颇丰。

## REFERENCES

- [1] Edouard Lansiaux, Noé Tchagaspian, *Community Impact on a Cryptocurrency: Twitter Comparison Example Between Dogecoin and Litecoin*, (Frontiers in Blockchain, 19 April 2022), doi: 10.3389/fbloc.2022.829865.
- [2] Pengcheng Zhang etc., *Investor sentiment and the holiday effect in the cryptocurrency market: evidence from China*, (Financial Innovation(2024) 10:113), <https://doi.org/10.1186/s40854-024-00639-x>.

Email address: [zqiaoxu678@gmail.com](mailto:zqiaoxu678@gmail.com)