第7章 最小二乘法（OLS）及其Stata实现

华北电力大学 法政系
公管教研室　马冉
ncepumaran@163.com

- **一、回归基本概念**
  - （一）变量间关系
  - （二）回归分析的基本概念
  - （三）回归的函数形式
  - （四）随机扰动项
  - （五）样本回归函数
- **二、最小二乘法**
  - （一）经典假设
  - （二）最小二乘法
- **三、最小二乘法的STATA操作**

一、回归基本概念

（一）变量间关系

经济变量之间的关系，大体可分为两类：

a.确定性关系或函数关系：研究的是确定现象、非随机变量间的关系。

b.统计依赖或相关关系：研究的是非确定现象、随机变量间的关系。

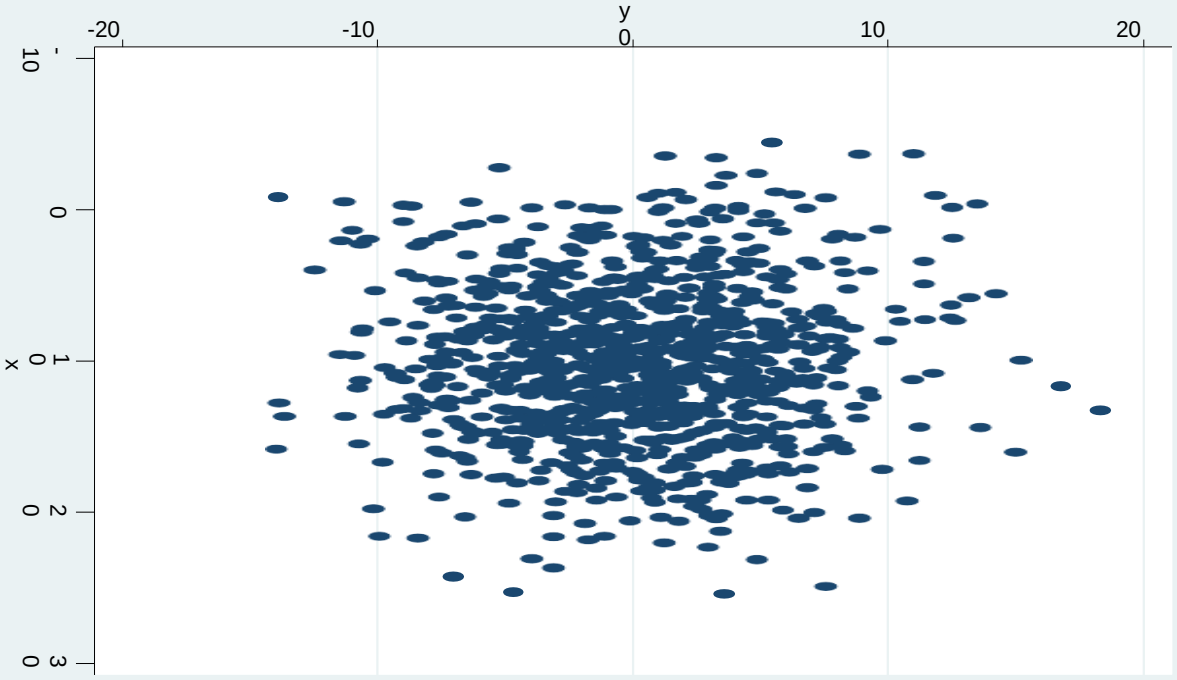
- 例如：
  - 函数关系：

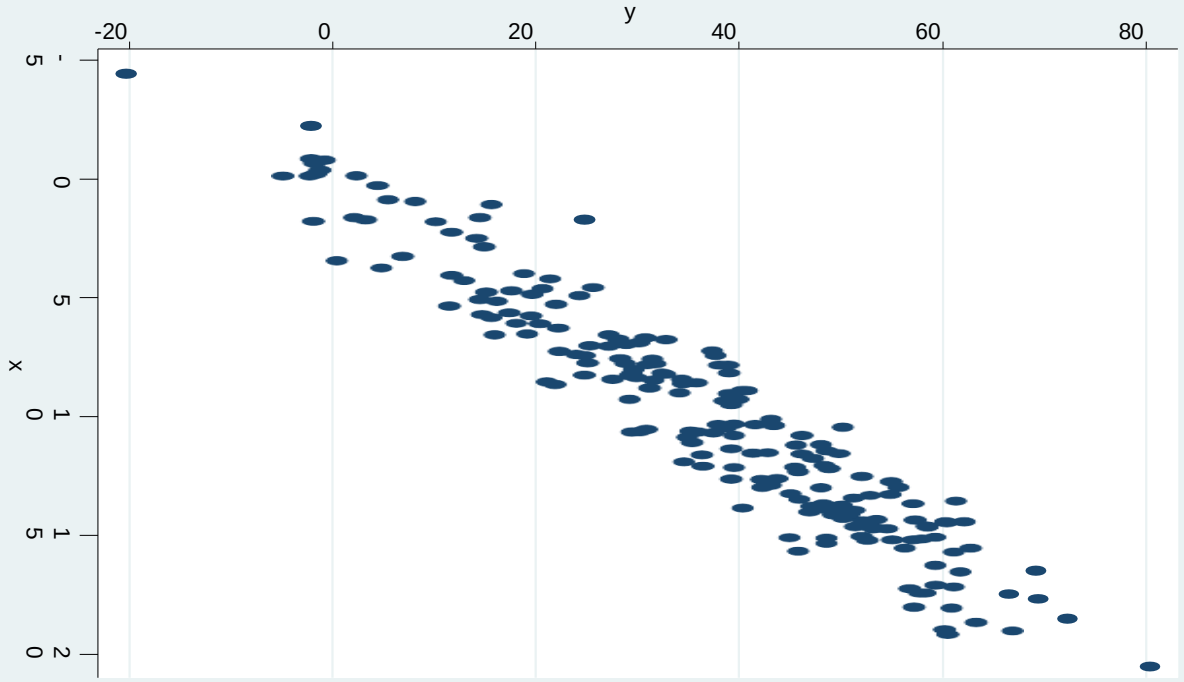
    圆面积 $=f(\pi,$ 半径$)=\pi\cdot$ 半径^2

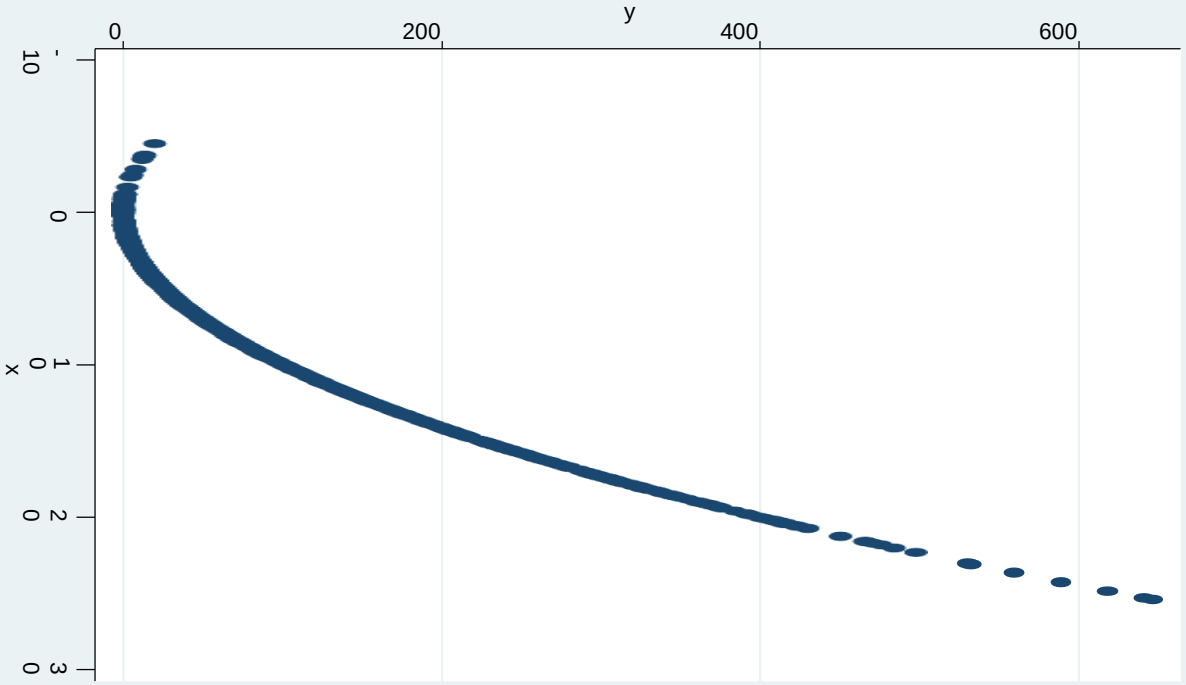
  - 统计依赖关系/统计相关关系：

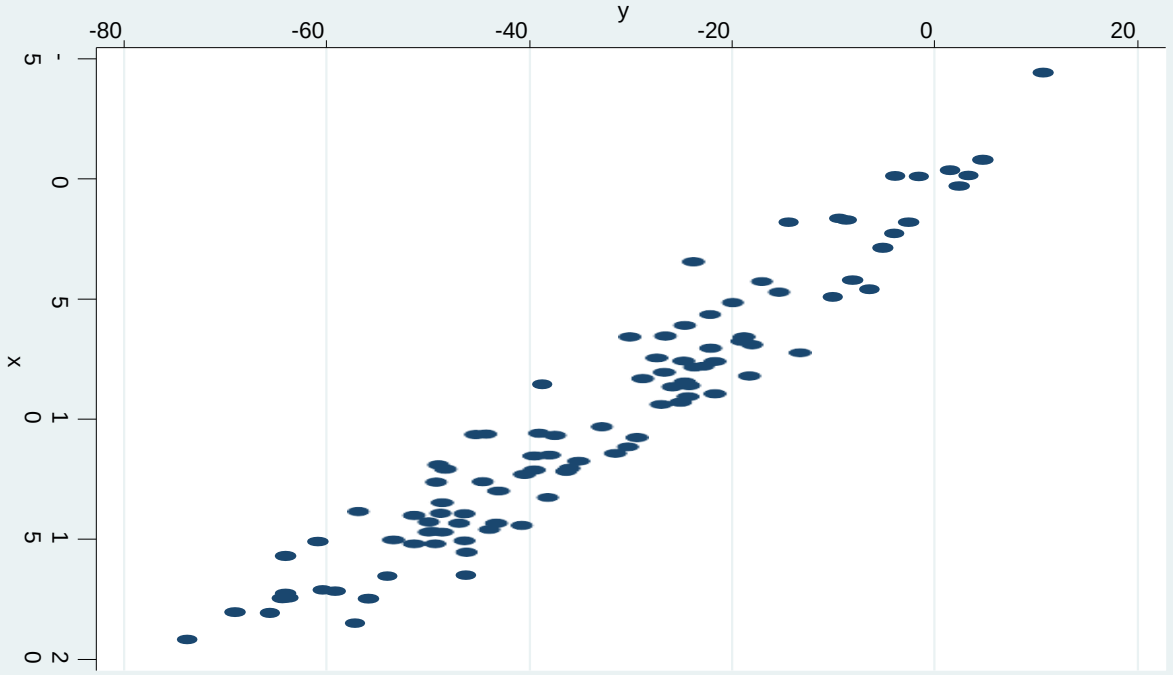
    农作物产量 $=f$ (气温，降雨量，阳光，施肥量……)

    对变量间统计依赖关系的考察主要是通过相关分析或回归分析实现的


y
-20
-10
0
10
20
x
-10
0
10
20
30


y
-20
0
20
40
60
80
x
-5
0
5
10
15
20


y
0
200
400
600
x
-10
0
10
20
30


y
-80
-60
-40
-20
0
20
x
-5
0
5
10
15
20

注意：

- ①不线性相关并不意味着不相关；
- ②有相关关系并不意味着一定有因果关系；
- ③**回归分析/相关分析**研究一个变量对另一个（些）变量的统计依赖关系，但它们并不意味着一定有因果关系。
- ④**相关分析**对称地对待任何（两个）变量，两个变量都被看作是随机的。**回归分析**对变量的处理方法存在不对称性，即区分因变量（被解释变量）和自变量（解释变量）：前者是随机变量，后者不是。

（二）回归分析的基本概念

- 回归一词最早由弗朗西斯·高尔顿（Francis Galton）引入。
- 他通过研究发现，虽然父母高，儿女也高；父母矮，儿女也矮，但给定父母的身高，儿女辈的平均身高却趋向于或者回归到全体人口的平均身高。

1078对夫妇　　$y=33.73+0.516x$

回归分析(regression analysis)是研究一个变量与另一个（些）变量的具体依赖关系的计算方法和理论。

- 其用意：在于通过后者的已知或设定值，去估计和（或）预测前者的（总体）均值。

$$y = \beta_0 + \beta_1 x + u$$

- 这里：y被称为被解释变量（Explained Variable）或因变量（Dependent Variable），x变量被称为解释变量（Explanatory Variable）或自变量（Independent Variable）。u:误差项，扰动项，未观测到的其他影响因素。

terminology for Simple regression 简单回归术语	
y	x
Dependent variable 因变量	Independent variable 自变量
Explained variable 被解释变量	Explanatory variable 解释变量
Response variable 响应变量	Control variable 控制变量
Predicted variable 被预测变量	Predictor variable 预测变量
Regressand 回归子	Regressor 回归元

- **回归分析构成计量经济学的方法论基础，其主要内容包括：**
  - －根据样本观察值对经济计量模型参数进行估计，求得回归方程；
  - －对回归方程、参数估计值进行显著性检验；
  - －利用回归方程进行分析、评价及预测。

（三）回归的函数形式

- 可以是线性或非线性的。
- 经典的一元线性回归模型：　$Y=\beta_0+\beta_1X+u$
- **线性函数**
- 其中，β_0、β_1是未知参数，称为**回归系数**（regressioncoefficients）
- 相应的回归方程：　$\hat{Y}=\hat{\beta}_1+\hat{\beta}_2X$

（三）总体回归函数

- 由于变量间关系的随机性，**回归分析关心的是根据解释变量的已知或给定值，考察被解释变量的总体均值**，即当解释变量取某个确定值时，与之统计相关的被解释变量所有可能出现的对应值的平均值。

- **例：**一个假想的社区有100户家庭组成，要研究该社区每月**家庭消费支出Y**与每月**家庭可支配收入X**的关系。

- 即如果知道了家庭的月收入，能否预测该社区家庭的平均月消费支出水平。

- 为达到此目的，将该100户家庭划分为组内收入差不多的10组，以分析每一收入组的家庭消费支出。

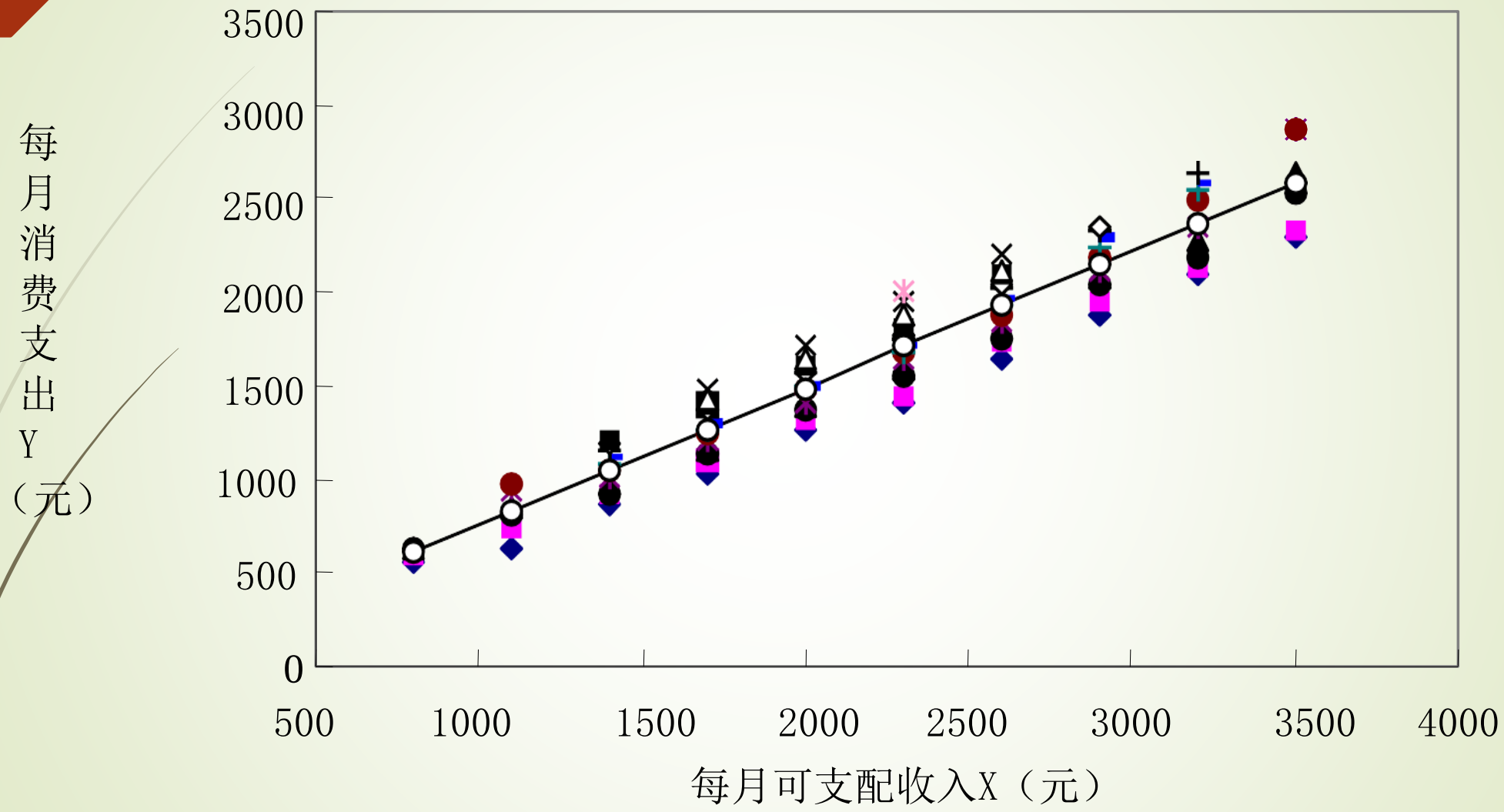
某社区家庭每月收入与消费支出统计表

	每月家庭可支配收入X（元）									
	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
每月家庭消费支出Y（元）	561	638	869	1023	1254	1408	1650	1969	2090	2299
	594	748	913	1100	1309	1452	1738	1991	2134	2321
	627	814	924	1144	1364	1551	1749	2046	2178	2530
	638	847	979	1155	1397	1595	1804	2068	2266	2629
		935	1012	1210	1408	1650	1848	2101	2354	2860
		968	1045	1243	1474	1672	1881	2189	2486	2871
			1078	1254	1496	1683	1925	2233	2552	
			1122	1298	1496	1716	1969	2244	2585	
			1155	1331	1562	1749	2013	2299	2640	
			1188	1364	1573	1771	2035	2310		
			1210	1408	1606	1804	2101			
				1430	1650	1870	2112			
				1485	1716	1947	2200			
						2002				
共计	2420	4950	11495	16445	19305	23870	25025	21450	21285	15510

- 给定收入X的值X_i，可得消费支出Y的**条件均值**（conditional mean）或**条件期望**（conditional expectation）：$E(Y|X=X_i)$

- 该例中：
- $P(Y=561|X=800)=1/4$
- $E(Y=605|X=800)=2420/4$

- **总体回归函数（population regression function, PRF）**：$E(Y|X_i)=f(X_i)$


每月消费支出Y（元）
3500
3000
2500
2000
1500
1000
500
0
500
1000
1500
2000
2500
3000
3500
4000
每月可支配收入X（元）

（四）随机扰动项

$$\mu_i = Y_i - E(Y \mid X_i)$$

- 称μi为观察值Yi围绕它的期望值E(Y | Xi)的离差（deviation）
- 是一个不可观测的随机变量
- 又称为随机干扰项（stochastic disturbance）或随机误差项（stochastic error）。

- 对于独个家庭：

$$Y_i = E(Y \mid X_i) + \mu_i = \beta_0 + \beta_1 X_i + \mu_i$$

- 称为**总体回归函数**（方程）PRF的随机设定形式。**表明被解释变量除了受解释变量的系统性影响外，还受其他因素的随机性影响。**

随机误差项主要包括以下因素的影响：

- 遗漏变量
- 观测误差
- 模型设定偏误
- 其它随机因素

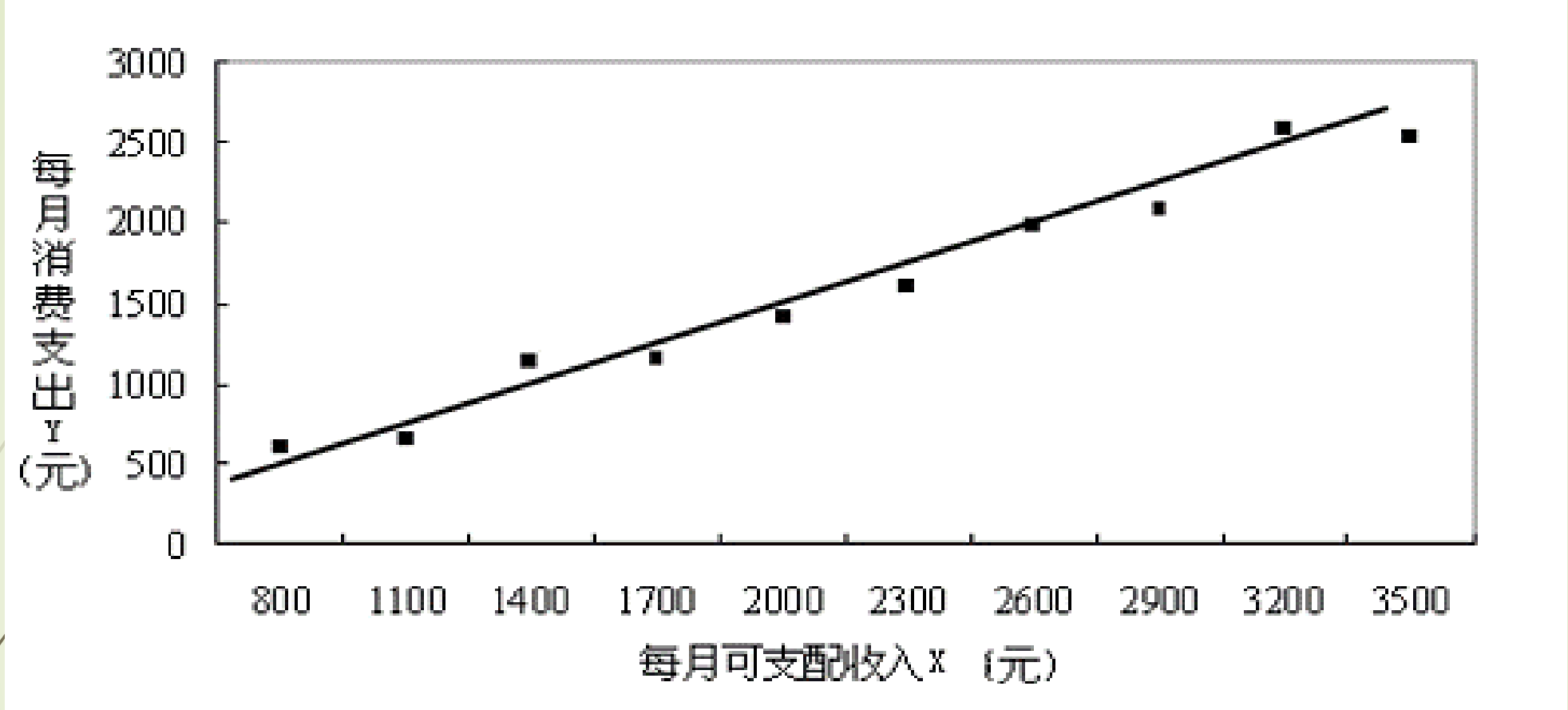
产生并设计随机误差项的主要原因：

- 理论的含糊性
- 数据的欠缺
- 节省原则

（五）样本回归函数

表　家庭消费支出与可支配收入的一个随机样本

Y	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
X	594	638	1122	1155	1408	1595	1969	2078	2585	2530



样本回归函数（sampleregressionfunction，SRF）

$$\hat{Y}_i = \int(X_i) = \widehat{\beta_0} + \widehat{\beta_1} X_i$$

思考：样本到总体？

回归分析的主要目的：根据样本回归函数SRF，估计总体回归函数PRF。

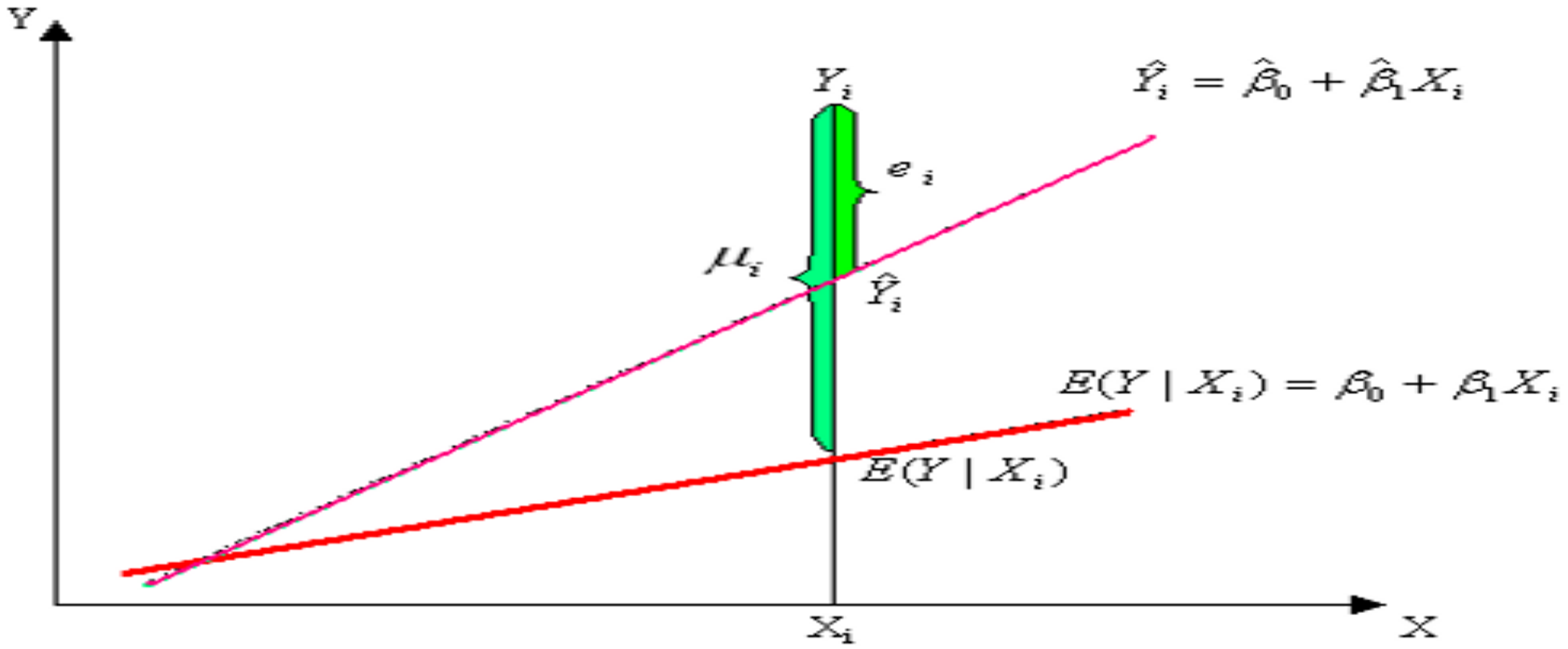
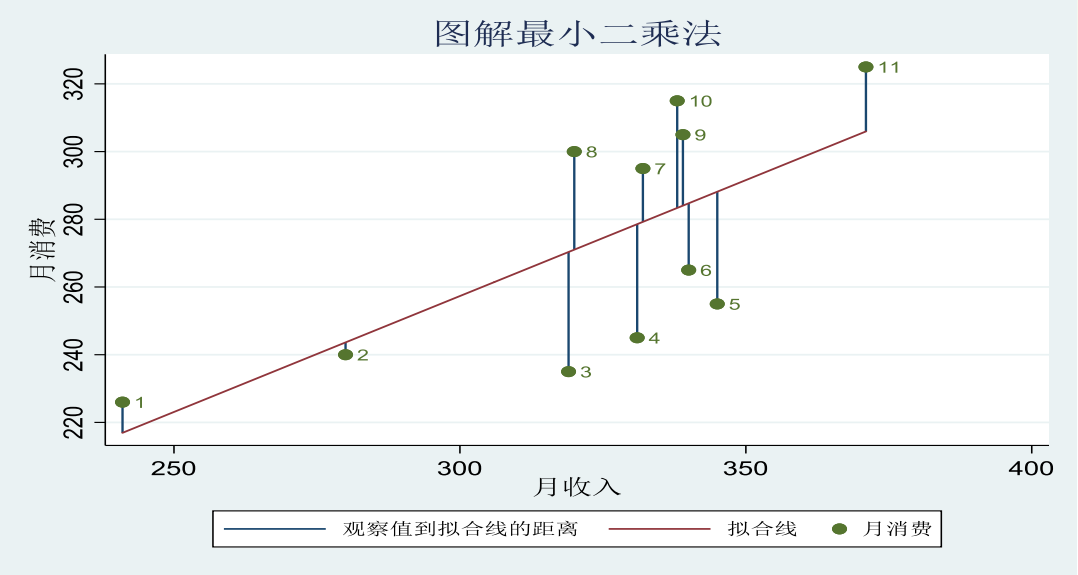


图　　总体回归线与样本回归线的基本关系

二、最小二乘法(OLS)

id	income	consume
1	241	226
2	280	240
3	319	235
4	331	245
5	345	255
6	340	265
7	332	295
8	320	300
9	339	305
10	338	315
11	371	325


图解最小二乘法
月消费
220
240
260
280
300
320
250
300
350
400
月收入
1
2
3
4
5
6
7
8
9
10
11
观察值到拟合线的距离
拟合线
月消费

- （一）经典假设（高斯假设/经典线性回归模型）：
- H1:零条件均值假定，也称强外生性，保证参数估计的无偏性

$$E(\varepsilon_i \mid X) = 0$$

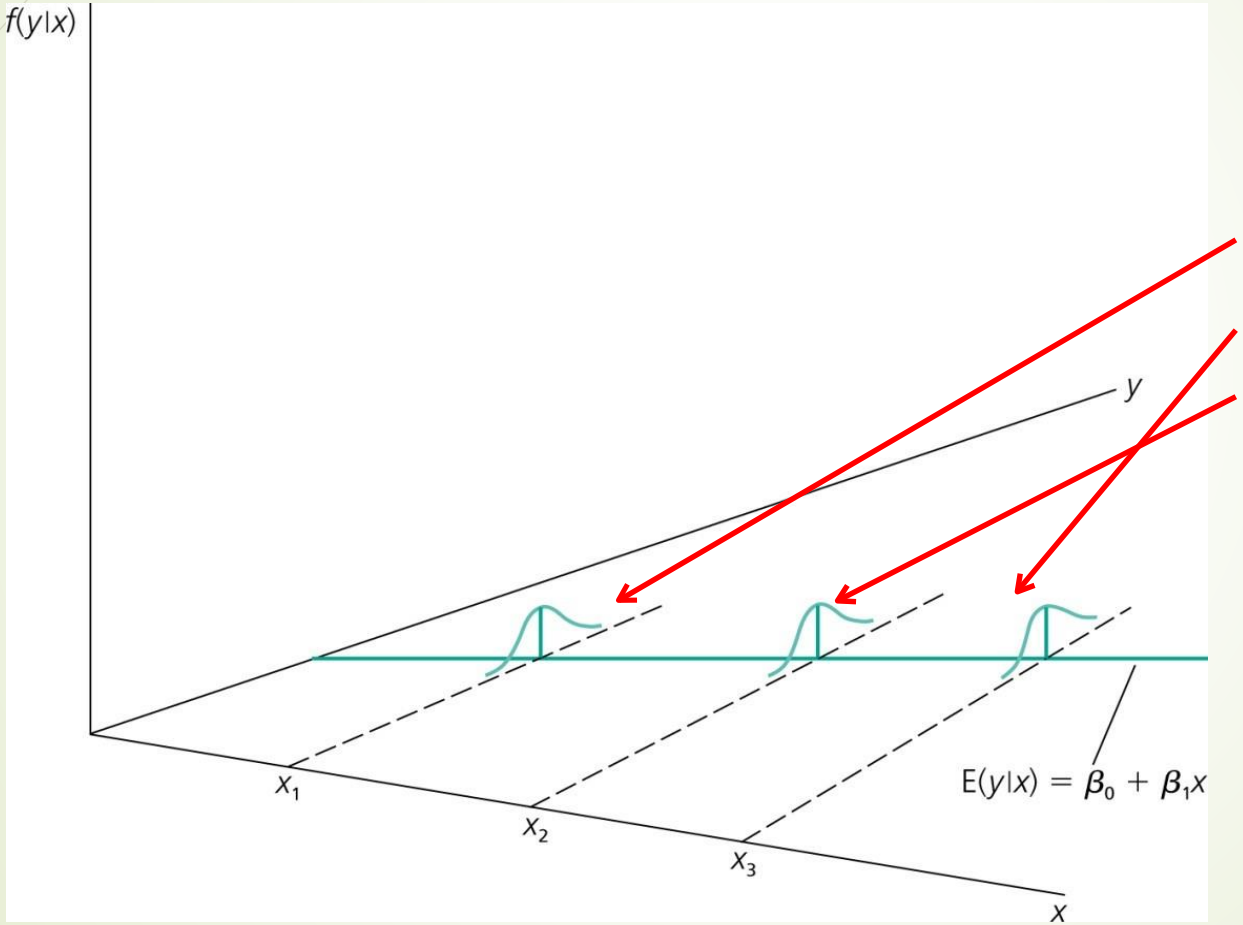
或 $$E(X\varepsilon_i) = 0$$

H2：同方差与无自相关假定（或称**球形扰动**），即随机扰动项的方差-协方差为同方差和无自相关同时成立的假定，违反将会影响到参数估计的有效性。

$$Var(\varepsilon \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

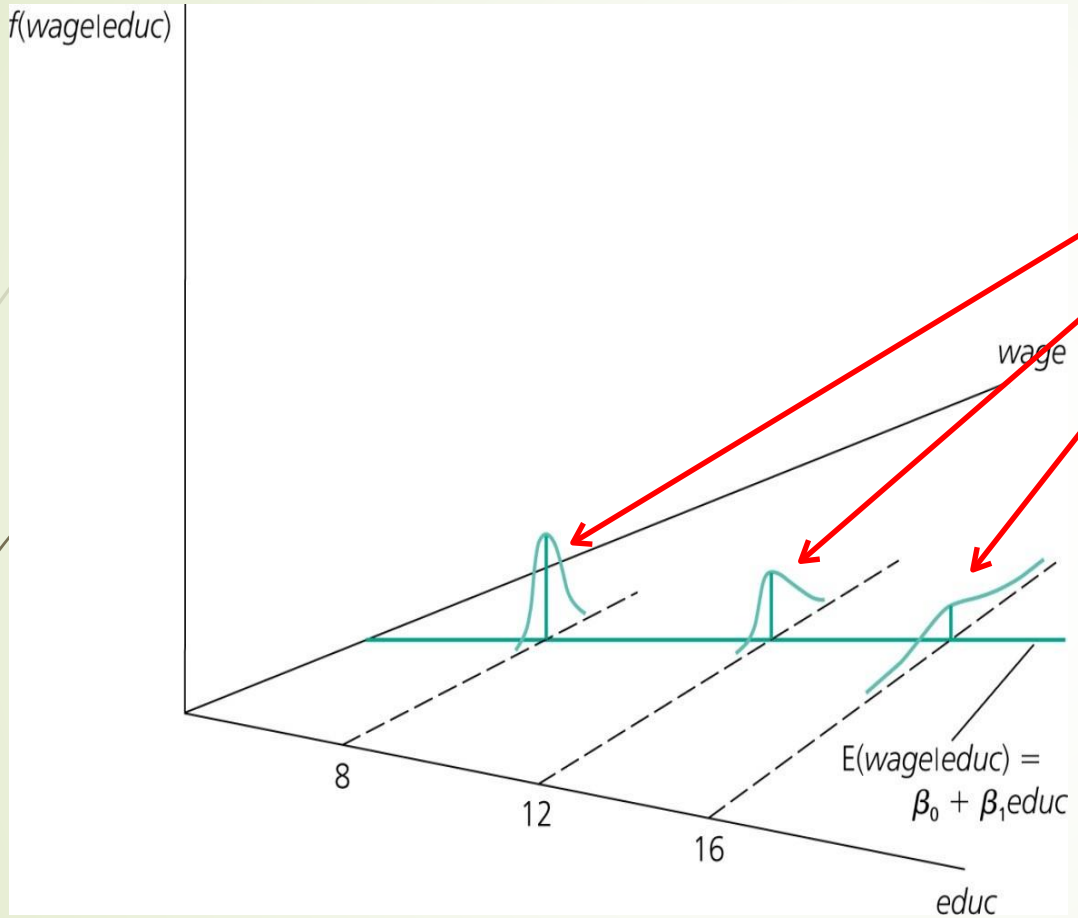
$$\sigma^2 = \frac{1}{\mathrm{n} - k} e^2$$

同方差的图形化表示

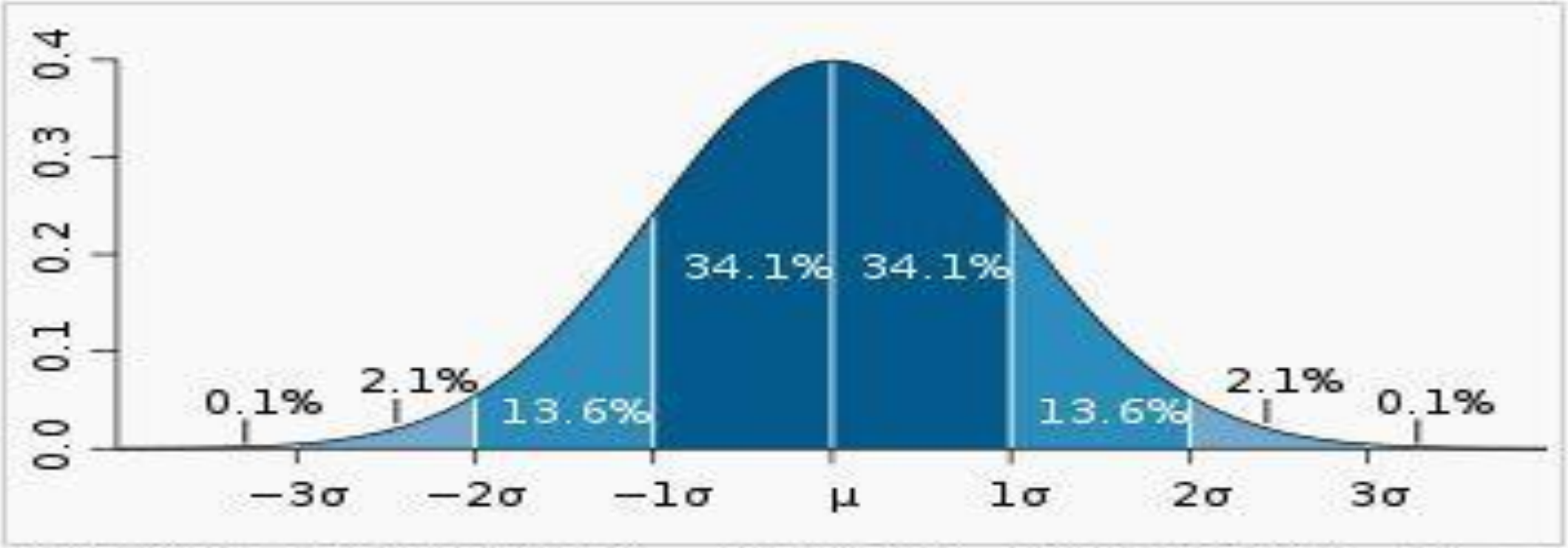


未观测到的影响的差异性
不依赖于解释变量的大小

- 异方差的图形化表示：工资和教育水平



随着教育水平的提高，未观测到的工资影响因素的方差增加



深蓝区域是距平均值小于一个标准差之内的数值范围。在正态分布中，此范围所占比率为全部数值之 68%。 根据正态分布，两个标准差之内（深蓝，蓝）的比率合起来为 95%。根据正态分布，三个标准差之内（深蓝，蓝，浅蓝）的比率合起来为 99%。

- H3:随机扰动项与解释变量不相关，也称外生性，违反将会影响到参数估计一致性

$$Cov(x_{ji}, \varepsilon_i \mid \boldsymbol{X}) = 0$$

- H4:识别条件：满秩/无多重共线性，即各解释变量之间线性无关

$$Rank(\mathbf{X}^{'}\mathrm{X}) = k \qquad （\mathrm{k<n}）$$

- H5:正态性假定:

$$\varepsilon_i \sim N(0, \sigma^2)$$

（二）最小二乘法

- 1. 高斯—马尔可夫定理(Gauss-Markov theory)
- <u>在给定经典线性回归的假定下，最小二乘估计量是具有最小方差的线性无偏估计量（BLUE）。</u>

Best: Least variance of bb_1 (most efficient)

Linear: bb_1 is linear in YY_{ii}&uu_{ii}

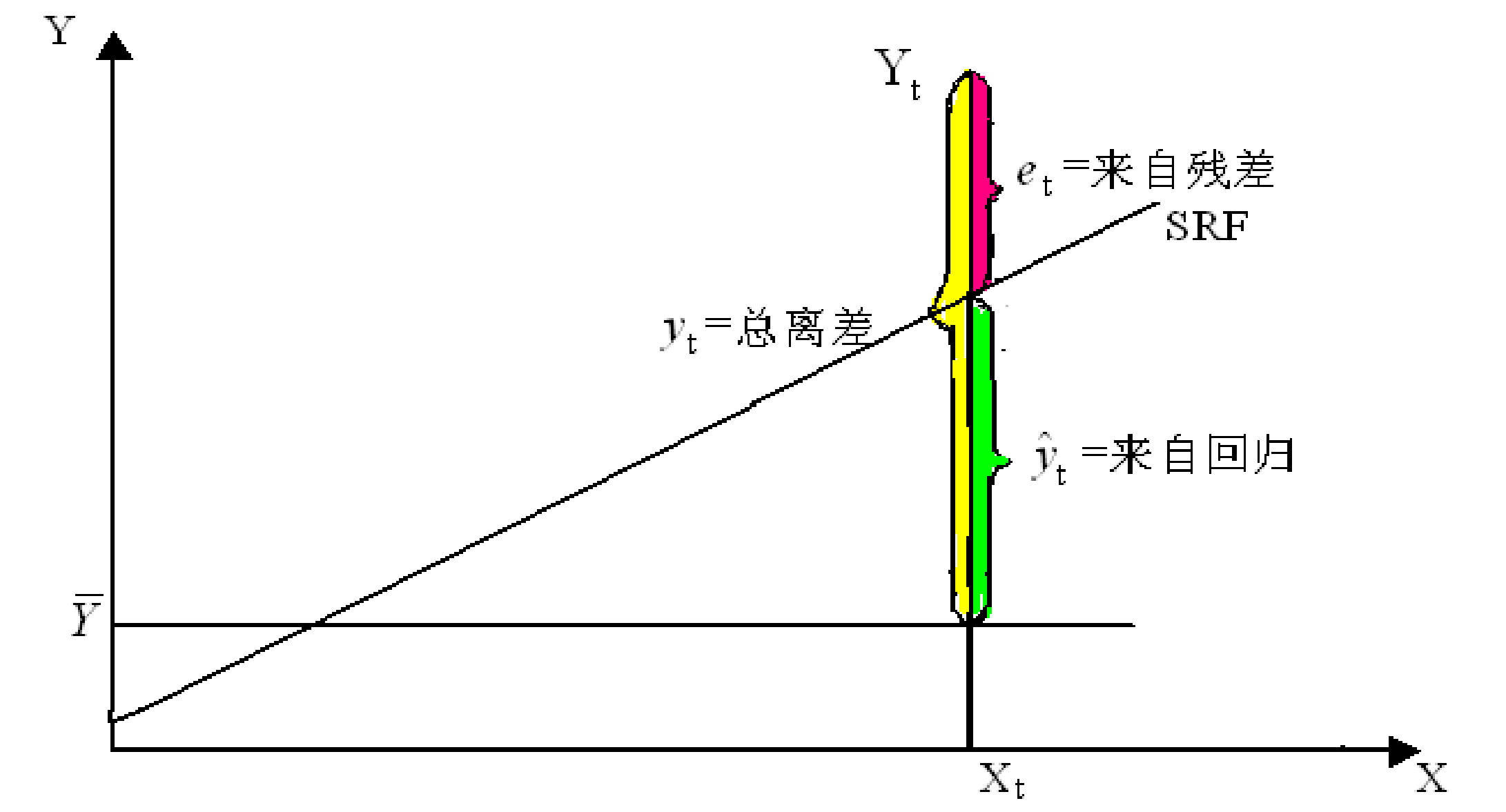
Unbiased: $E(b)=\beta$

Estimator: $\beta = (X'X)^{-1}X'Y$

Source	SS	df	MS
Model	5768.20676	1	5768.20676
Residual	6849.97506	9	761.10834
Total	12618.1818	10	1261.81818

Number of obs	=	11
F(1, 9)	=	7.58
Prob > F	=	0.0224
R-squared	=	0.4571
Adj R-squared	=	0.3968
Root MSE	=	27.588

consume	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.6848014	.2487525	2.75	0.022	.1220841	1.247519
_cons	51.89511	80.84397	0.64	0.537	-130.9867	234.7769


Y
Y_t
e_t =来自残差
SRF
y_t =总离差
$\hat{y}_t$ =来自回归
$\bar{Y}$
X_t
X

三、最小二乘法的STATA操作

- reg y x, vce(r)
- predict yhat

  predict e, res
- test