

第 9 章 模型设定与数据问题

如果模型设定(model specification)不当, 如解释变量选择不当、测量误差、函数形式不妥等, 会出现“设定误差”(specification error)。

数据本身也可能存在问题, 如多重共线性、对回归结果影响很大的极端数据等。

Produced with a Trial Version of PDF Annotator - www.PDFAnno

2. The variance of the estimator

- Under the same assumptions for unbiasedness (assumptions 1-3), plus Assumption 4, the spherical variance assumption, we can write

$$\text{Var}(b|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix} \quad (6)$$

- A simplified example can shed more light onto this expression. Consider the case of simple regression, where the regressors are only the constant term and a single explanatory variable x .
- The lower right element of $\sigma^2(\mathbf{X}'\mathbf{X}^{-1})$ is:

$$y = \beta_0 + \beta_1 x_i + \varepsilon \quad \text{Var}(b|\mathbf{X}) = \frac{\sigma^2}{n \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{n \overline{\text{Var}(x)}} \quad (7)$$

- This tells us that the sample variance of b will be low (so that β will be more precisely estimated) if:
 - σ^2 is low (the error term has low variance)
 - n is high (more observations)
 - or x has high variance

Produced with a Trial Version of PDF Annotator - www.PDFAnno

- The general case with K variables.
- Call \mathbf{x}_k the column vector in \mathbf{X} corresponding to the k th variable, and $\mathbf{X}_{(k)}$ the $n \times (K-1)$ data matrix consisting of the remaining variables.
- Then write out $\mathbf{X}'\mathbf{X}$ as the partitioned matrix

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}'_{(k)} \\ \mathbf{x}'_k \end{bmatrix} \begin{bmatrix} \mathbf{X}_{(k)} & \mathbf{x}_k \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_{(k)}\mathbf{X}_{(k)} & \mathbf{X}'_{(k)}\mathbf{x}_k \\ \mathbf{x}'_k\mathbf{X}_{(k)} & \mathbf{x}'_k\mathbf{x}_k \end{bmatrix} \quad (8)$$

$n \times (k-1)$ $n \times 1$

- Note that since we can order the variables however we like, there is no problem with putting \mathbf{x}_k into the last column.

Produced with a Trial Version of PDF Annotator - www.PDFAnno

- Then, by the formula for a partitioned inverse from linear algebra (A.5.3 in Greene's text provides more details), the lower right block of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ (which is a scalar in this example) is

$$\sigma^2 \left(\mathbf{x}'_k \mathbf{x}_k - \mathbf{x}'_k \mathbf{X}_{(k)} (\mathbf{X}'_{(k)} \mathbf{X}_{(k)})^{-1} \mathbf{X}'_{(k)} \mathbf{x}_k \right)^{-1} \quad (9)$$

$$= \sigma^2 \left(\mathbf{x}'_k \mathbf{x}_k - \mathbf{x}'_k \mathbf{P}_{(k)} \mathbf{x}_k \right)^{-1} \quad (10)$$

$$\text{where } \mathbf{P}_{(k)} \equiv \mathbf{X}_{(k)} (\mathbf{X}'_{(k)} \mathbf{X}_{(k)})^{-1} \mathbf{X}'_{(k)} \quad (11)$$

$$= \sigma^2 \left(\mathbf{x}'_k \mathbf{M}_{(k)} \mathbf{x}_k \right)^{-1} \quad (12)$$

$$= \sigma^2 \left(\mathbf{x}'_k \mathbf{M}'_{(k)} \mathbf{M}_{(k)} \mathbf{x}_k \right)^{-1} \quad (13)$$

where $\mathbf{M}_{(k)} \equiv \mathbf{I}_n - \mathbf{P}_{(k)}$.

Handwritten note: $\mathbf{x}_k \in \mathbf{X}_{(k)}$ 的残差

- Therefore

$$\text{Var}(b_k | \mathbf{X}) = \sigma^2 \left(\mathbf{x}'_k \mathbf{M}'_{(k)} \mathbf{M}_{(k)} \mathbf{x}_k \right)^{-1} \quad (14)$$

Produced with a Trial Version of PDF Annotator - www.PDFAnno

- Note that the vector $M_{(k)}\mathbf{x}_k$ is the vector of residuals obtained by regressing x_k on the other x variables.
- Therefore, in sample form,

$$\text{Var}(b_k|\mathbf{X}) = \frac{\sigma^2}{\sum_{i=1}^n \underbrace{(x_{ik} - \hat{x}_{ik})^2}_{\text{red wavy lines}}} \quad (15)$$

where \hat{x}_{ik} is the fitted value of x_{ik} from a regression of x_k on the other x variables.

- This formula reduces to the special case described earlier because when “the other variables” is a constant, the fitted value of x_{ik} is simply its sample mean.

Produced with a Trial Version of PDF Annotator - www.PDFAnno

- When “the other variables” include a constant term, we can use the R^2 decomposition to simplify the expression further
- Remember $SST = SSE + SSR$
- Or in this context,

$$\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2 = \sum_{i=1}^n (\hat{x}_{ik} - \bar{x}_k)^2 + \sum_{i=1}^n (x_{ik} - \hat{x}_{ik})^2 \quad (16)$$

- Therefore, we can rewrite the previous formula as

$$\text{Var}(b_k | \mathbf{X}) = \frac{\sigma^2}{SSR} = \frac{\sigma^2}{SST - SSE} = \frac{\sigma^2}{SST(1 - R_x^2)} \quad (17)$$

- or

$$\frac{\sigma^2}{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2 \cdot (1 - R_x^2)} \quad (18)$$

where R_x^2 is the R-squared from regressing x_k on the other x variables.

\downarrow
 x_k 在 $X_{(k)}$
 的回归中的 R^2
 $\sqrt{\text{Var}(x_k)}$

- Again, this formula simplifies to the one for the simple regression case because when there are no other x variables in the model (except for the constant term), $R_x^2 = 0$.
- What insights do we gain from this formula?
- The previous findings still apply (that is variance of b_k is low when the variance of x_k is high, when n is high, and when σ^2 is low).
- The new insight is, when x_k is highly correlated with the other x variables (that is, R_x^2 is high), the variance of b_k is also high.
- Put another way, it is difficult to estimate b_k precisely when x_k is highly correlated with the other x variables.
- In an extreme case, when x_k is an exact linear function of the other x variables (that is, when $R_x^2 = 1$), the sampling variance is infinite and b_k cannot be estimated at all.

9.1 遗漏变量

中心问题.

假设真实的模型为

$$y_i = x'_{i1}\beta_1 + x'_{i2}\beta_2 + \varepsilon_i$$

其中, x_1, x_2 可以是向量, 且与扰动项 ε 不相关。而实际估计的模型(estimated model)为

$$y_i = x'_{i1}\beta_1 + u_i \leftarrow x_2$$

遗漏变量(omitted variables) $x'_{i2}\beta_2$, 被归入新扰动项 $u_i = x'_{i2}\beta_2 + \varepsilon_i$ 。

考虑以下两种情形：

(1) $\text{Cov}(x_{i1}, x_{i2}) = 0$ 。

OLS 一致。遗漏变量 $x'_{i2}\beta_2$ 归入扰动项 u_i 中，可能增大扰动项的方差，影响估计精度。

(2) $\text{Cov}(x_{i1}, x_{i2}) \neq 0$

OLS 不一致，其偏差为“遗漏变量偏差” (omitted variable bias)。

解决遗漏变量偏差的方法主要有：*Michael Spencer*：

- (i) 加入尽可能多的控制变量(control variable);
- (ii) 使用“代理变量”(proxy variable);
- (iii) 工具变量法(第 10 章);
- (iv) 使用面板数据(第 15-17 章);
- (v) 随机实验与自然实验(第 18 章)。

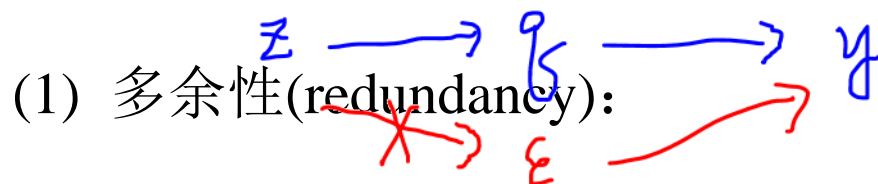
第(i)种方法：尽可能去收集数据。或从理论上说明，遗漏变量不会与解释变量相关，或相关性很弱。

例 李宏彬等(2012)通过就业调查数据，研究“官二代”大学毕业生的起薪是否高于非官二代。

由于可能存在遗漏变量，该文包括了尽可能多的控制变量，比如年龄、性别、城镇户口、父母收入、父母学历、高考成绩、大学成绩、文理科、党员、学生会干部、兼职实习经历、拥有技术等级证书等。

第(ii)种方法，即代理变量法。比如，在教育投资回归中，可用智商(IQ)来作为个人能力的代理变量。

理想的代理变量应满足以下两个条件：



即代理变量仅通过影响遗漏变量而作用于被解释变量。比如，“智商”仅通过对“能力”的作用来影响工资收入。假如有“能力”的数据，引入“智商”量就是多余的。

(2) 剩余独立性: Q 去掉 Z 后剩余部分 v , $\text{Cov}(X_k, v) = 0$

遗漏变量中不受代理变量影响的剩余部分与所有解释变量均不 $\forall k$. \checkmark 相关。

命题 如果上述两个条件满足，使用代理变量能获得一致估计。

证明： 假设真实模型为

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \underbrace{\gamma q}_{\text{代理变量}} + \varepsilon$$

其中， q 为不可观测的遗漏变量。假定 $\text{Cov}(x_k, \varepsilon) = 0, \forall k$ ，但 q 与某解释变量 x_m 相关 ($1 \leq m \leq K$)，即 $\text{Cov}(x_m, q) \neq 0$ ，故 OLS 不一致。

假设找到代理变量 z ，满足

Linear Projection

$$q = \delta_0 + \delta_1 z + v, \quad \text{Cov}(z, v) = 0$$

根据第一个条件(多余性)，代理变量 z 只通过 q 对 y 发生作用，

故在回归方程已经包含 q 的情况下, z 与 y 的扰动项 ε 不相关, 即 $\text{Cov}(z, \varepsilon) = 0$ 。

根据第二个条件, q 的扰动项 v 与所有解释变量均不相关, 即 $\text{Cov}(x_k, v) = 0, \forall k$ 。将 q 的表达式代入原模型可得

$$y = (\beta_0 + \gamma\delta_0) + \beta_1 x_1 + \dots + \beta_K x_K + \gamma\delta_1 z + (\gamma v + \varepsilon)$$

容易证明, 新扰动项 $(\gamma v + \varepsilon)$ 与所有解释变量均不相关,

$$\text{Cov}(x_k, \gamma v + \varepsilon) = \underbrace{\gamma \text{Cov}(x_k, v)}_{\text{condition 2}} + \underbrace{\text{Cov}(x_k, \varepsilon)}_{\text{assumption}} = 0 + 0 = 0 \quad (\forall k)$$

$$\text{Cov}(z, \gamma v + \varepsilon) = \underbrace{\gamma \text{Cov}(z, v)}_{\text{assumption}} + \underbrace{\text{Cov}(z, \varepsilon)}_{\text{condition 1}} = 0 + 0 = 0$$

故 OLS 一致。如果代理变量不满足这两个条件, 则不一致。

Linear Projection

$z \rightarrow q \rightarrow y$
 ε

任何实证研究中几乎总是存在遗漏变量。

论文应说明，如何在存在遗漏变量的情况下避免遗漏变量偏差。

9.2 无关变量

假设真实模型为

$$y_i = x'_{i1}\beta_1 + \varepsilon_i$$

其中， $\text{Cov}(x_{i1}, \varepsilon_i) = 0$ 。而实际估计的模型为

$$y_i = x'_{i1}\beta_1 + x'_{i2}\beta_2 + \underbrace{(\varepsilon_i - x'_{i2}\beta_2)}_{=0}$$

由于真实参数 $\beta_2 = 0$ ，故可将模型写为 $y_i = x'_{i1}\beta_1 + x'_{i2}\beta_2 + \varepsilon_i$ 。

由于 x_2 与 y 无关，故 x_2 也与 y 的扰动项 ε 无关，即 $\text{Cov}(x_{i2}, \varepsilon_i) = 0$ 。

故 OLS 一致，即 $\text{plim}_{n \rightarrow \infty} \hat{\beta}_1 = \beta_1$ ， $\text{plim}_{n \rightarrow \infty} \hat{\beta}_2 = \beta_2 = 0$ 。

但引入无关变量后，估计量 $\hat{\beta}_1$ 的方差一般会增大。

9.3 建模策略：“由小到大”还是“由大到小”

看文献。

“由小到大” (specific to general) 的建模方式，首先从最简单的小模型开始，逐渐增加解释变量。

但小模型很可能存在遗漏变量，导致估计量不一致， t 检验、 F 检验都将失效，很难确定该如何取舍变量。

“由大到小” (general to specific) 的建模方式，从尽可能大的模型开始，收集所有可能的解释变量，逐步剔除不显著的解释变量。



虽冒着包含无关变量的危险，但危害性没有遗漏变量严重。但在实际操作上，常常很难找到足够多的解释变量。

实践中，常采用这两种策略的折衷方案。↓

控制 β_1 . $P(\text{认为 } \beta_1 \neq 0 \mid \beta_1 = 0)$, False Positive, I.

检测 β_2 . $P(\text{认为 } \beta_2 = 0 \mid \beta_2 \neq 0)$, False Neg. II.

9.4 解释变量个数的选择

数据选择.

加入过多解释变量可提高模型解释力，但牺牲模型的简洁性 (parsimony)。权衡标准：

(1) 校正可决系数 \bar{R}^2 ：选择解释变量的个数 K 以最大化 \bar{R}^2 。

(2) “赤池信息准则” (Akaike Information Criterion, 简记 AIC)：选择解释变量的个数 K ，使得以下目标函数最小化：

$$\min_K \text{AIC} \equiv \ln(e'e / n) + \frac{2K}{n}$$

regularization.

右边第一项为对模型拟合度的奖励(减少残差平方和), 第二项为对解释变量过多的惩罚(解释变量个数 K 的增函数)。

当 K 上升时, 第一项下降而第二项上升。

(3)“贝叶斯信息准则”(Bayesian Information Criterion, 简记 BIC) 或“施瓦茨信息准则”(Schwarz Information Criterion, 简记 SIC 或 SBIC):

$$\min_K \text{BIC} \equiv \ln(e'e / n) + \frac{\ln n}{n} K$$

一般来说, $\ln n > 2$, 故 BIC 准则对于解释变量过多的惩罚比 AIC 严厉。BIC 准则更强调模型的简洁性。

(4) “汉南-昆信息准则” (Hannan-Quinn Information Criterion, 简记 HQIC):

$$\min_K \text{HQIC} \equiv \ln(\mathbf{e}'\mathbf{e} / n) + \frac{\ln[\ln(n)]}{n} K$$

在时间序列模型中，常用信息准则来确定滞后阶数。

比如，AR(p)模型:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t, \quad t = 1, \dots, T$$

根据 BIC 或 HQIC 计算的 \hat{p} 是 p 的一致估计，即当 $T \rightarrow \infty$ 时， $\Pr(\hat{p} < p) \rightarrow 0$ ， $\Pr(\hat{p} = p) \rightarrow 1$ ， $\Pr(\hat{p} > p) \rightarrow 0$ 。

根据 AIC 计算的 \hat{p} 不一致，在大样本中可能高估 p ，虽然 $\Pr(\hat{p} < p) \rightarrow 0$ ，但 $\Pr(\hat{p} > p) \rightarrow c > 0$ 。

在实践中，常用 AIC 与 BIC。

虽然 BIC 一致而 AIC 不一致，但现实样本有限，而 BIC 准则可能导致模型过小，故 AIC 准则依然常用。

9.5 对函数形式的检验

如果回归方程中存在非线性项，则边际效应不再是常数。

【例】

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma x_1^2 + \delta x_2 x_3 + \varepsilon$$

各变量的边际效应为：

$$\frac{E(y)}{\partial x_1} = \beta_1 + 2\gamma x_1, \quad \frac{E(y)}{\partial x_2} = \beta_2 + \delta x_3, \quad \frac{E(y)}{\partial x_3} = \beta_3 + \delta x_2$$

如怀疑边际效应非常数，应考虑中引入非线性项。

“Ramsey’s RESET 检验” (Regression Equation Specification Error Test)的基本思想是，如怀疑遗漏非线性项，则引入非线性项，并检验其系数是否显著。

假设线性回归模型为

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$$

回归拟合值 $\hat{y} = \mathbf{x}'\mathbf{b}$ 。 \hat{y} 是 \mathbf{x} 的线性组合， \hat{y}^2 包含解释变量二次项 (含平方项与交叉项) 的信息， \hat{y}^3 包含解释变量三次项的信息，等等。

考虑回归方程：

$$y = \mathbf{x}'\boldsymbol{\beta} + \delta_2 \hat{y}^2 + \delta_3 \hat{y}^3 + \delta_4 \hat{y}^4 + \varepsilon$$

对 $H_0 : \delta_2 = \delta_3 = \delta_4 = 0$ 作 F 检验。如拒绝 H_0 ，说明应有高次项；

如接受 H_0 ，可使用线性模型。

RESET 检验的缺点是，拒绝 H_0 时，不知道具体遗漏哪些高次项。

另一检验为“连接检验”(link test)。“连接”指的是，将解释变量与被解释变量连接在一起的函数形式是否正确。

进行以下回归：

$$y = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error$$

检验“ $H_0 : \delta_2 = 0$ ”。如果模型设定正确，则 \hat{y}^2 不应对 y 有解释力。如果拒绝 $H_0 : \delta_2 = 0$ ，则认为模型设定有误，可考虑加入非线性项或改变回归的函数形式(比如，取对数)。

在确定回归方程的函数形式时，最好从经济理论出发。

如缺乏理论指导，可从线性模型出发，再进行 RESET 或连接检验，看是否应加入非线性项。

9.6 多重共线性

如果数据矩阵 \mathbf{X} 不满列秩，即某一解释变量可由其他解释变量线性表出，则存在“严格多重共线性”。

近似的多重共线性表现为，将第 k 个解释变量 x_k 对其余的解释变量 $\{x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_K\}$ 进行回归，所得可决系数(记为 R_k^2)较高。

在多重共线性下，OLS 仍是 BLUE，但不表示 OLS 估计量方差在绝对意义上小。

$$\frac{\sigma^2}{n \text{Var}(x) (1 - R_x^2)}$$

由于存在多重共线性，矩阵 $(\mathbf{X}'\mathbf{X})$ 变得几乎不可逆， $(\mathbf{X}'\mathbf{X})^{-1}$ 变得很“大”，致使方差 $\text{Var}(\mathbf{b} | \mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ 增大，系数估计不准确。

\mathbf{X} 中元素轻微变化就会引起 $(\mathbf{X}'\mathbf{X})^{-1}$ 很大变化，导致 OLS 估计值 \mathbf{b} 发生很大变化。

通常的“症状”是，虽然整个回归方程的 R^2 较大、 F 检验也很显著，但单个系数的 t 检验却不显著。

另一“症状”是，增减解释变量使得系数估计值发生较大变化(比如，最后加入的解释变量与已有解释变量构成多重共线性)。

Stata²⁰: $i \cdot X_1$

如果两个(或多个)解释变量高度相关,则不易区分各自对被解释变量的影响。如一个变量是另一变量的倍数,则完全无法区分。

可以证明,协方差矩阵主对角线上的第 k 个元素为

$$\text{Var}(b_k | \mathbf{X}) = \frac{\sigma^2}{(1 - R_k^2) S_{kk}}$$

其中, $S_{kk} \equiv \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2$ 为 x_k 的离差平方和,反映 x_k 的变动幅度。如 x_k 完全不变, $S_{kk} = 0$, 则完全无法估计 b_k 。

定义第 k 个解释变量 x_k 的“方差膨胀因子”(Variance Inflation Factor, 简记 VIF)为

$$\text{VIF}_k \equiv \frac{1}{1 - R_k^2}$$

则 $\text{Var}(b_k | \mathbf{X}) = \text{VIF}_k \cdot (\sigma^2 / S_{kk})$ 。VIF越大, 多重共线性问题越严重。

经验规则: $\max\{\text{VIF}_1, \dots, \text{VIF}_K\}$ 不超过 10。

处理多重共线性的方法:

(1) 如果不关心具体的回归系数, 只关心整个方程的预测能力, 则通常可不必理会多重共线性。多重共线性的主要后果是使得对单个变量的贡献估计不准, 但所有变量的整体效应仍可准确估计。

(2) 如果关心具体的回归系数, 但多重共线性并不影响所关心变

量的显著性，也可不必理会。即使在有方差膨胀的情况下，这些系数依然显著；如果没有多重共线性，只会更加显著。

(3) 如果多重共线性影响到所关心变量的显著性，则需要增大样本容量，剔除导致严重共线性的变量，或对模型设定进行修改。

9.7 极端数据

如果样本数据中的少数观测值离大多数观测值很远，可能对OLS的回归系数产生很大影响，称为“极端观测值”(outliers or influential data)。

winorization : $\geq 99\%$, = 99%
 $\leq 1\%$, = 1%

trimming : 去掉
 y_i

OLS :

资产定价: r_{it}

Volatility

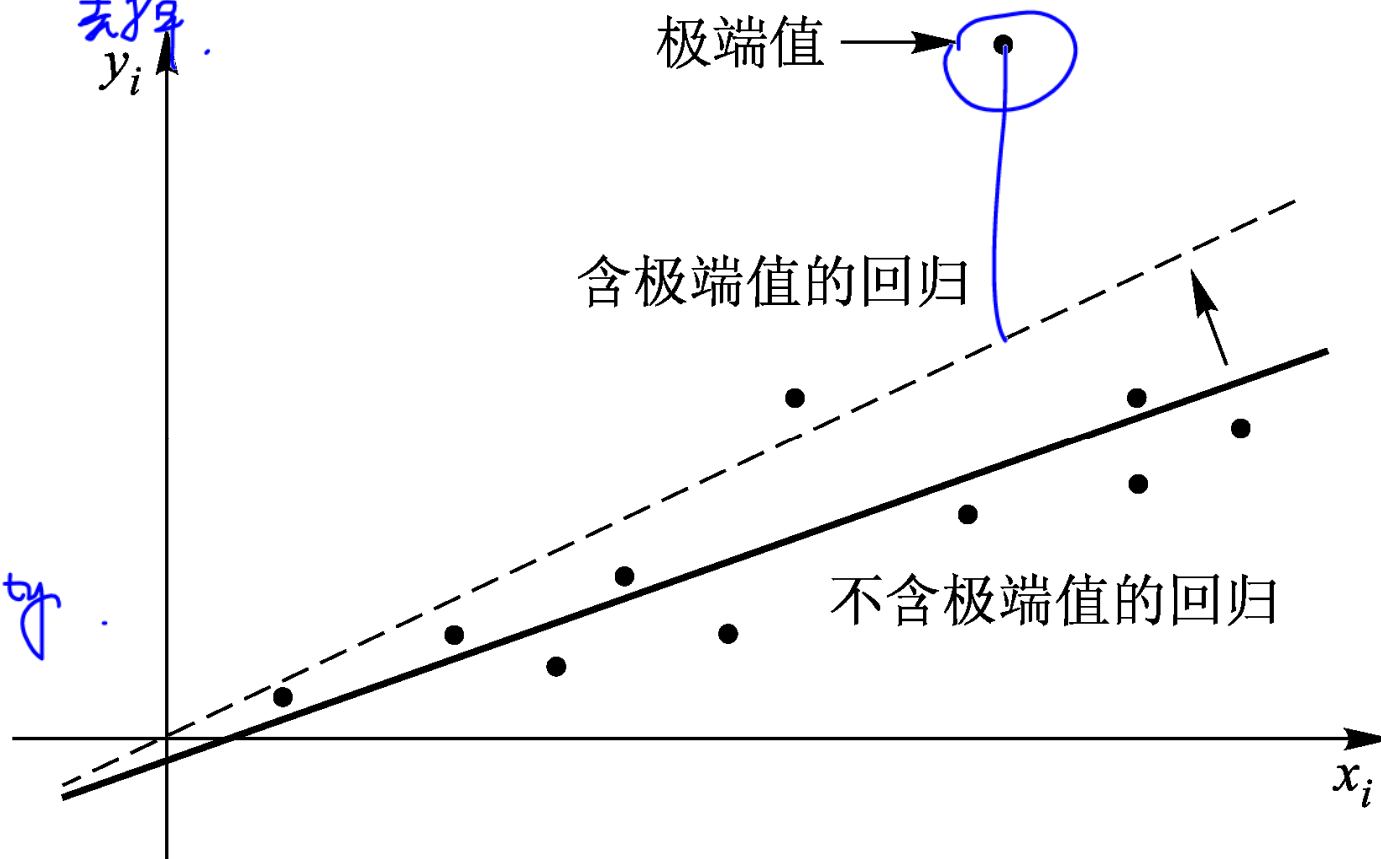


图 9.1 极端观测值对回归系数的影响

第 i 个观测数据对回归系数的“影响力”或“杠杠作用”(leverage) 可通过投影矩阵 $\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 的第 i 个主对角线元素来表示:

$$\text{lev}_i \equiv \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

所有观测数据的影响力 lev_i 满足:

(i) $0 \leq \text{lev}_i \leq 1, (i=1, \dots, n);$

$\frac{1}{n} \leq \text{lev}_i \leq 1$

(ii) $\sum_{i=1}^n \text{lev}_i = K$ (解释变量个数)。影响力 lev_i 的平均值为 (K/n) 。

记 $\mathbf{b}^{(i)}$ 为去掉第 i 个观测数据后的 OLS 估计值，可以证明：

"Greene"

$$\mathbf{b} - \mathbf{b}^{(i)} = \left(\frac{1}{1 - \text{lev}_i} \right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i e_i$$

lev_i 越大则 $(\mathbf{b} - \mathbf{b}^{(i)})$ 的变化越大。

如果 lev_i 比平均值 (K/n) 高很多，则可能对回归系数有很大影响。

如何处理极端观测值：

首先，应检查是否因数据输入有误差导致极端观测值。

其次，对出现极端观测值的个体进行背景调查，看看是否由与研究课题无关的特殊现象所致，必要时可以删除极端数据。

最后，可同时汇报“全样本” (full sample)与删除极端数据后的“子样本” (subsample)的回归结果，让读者自己做判断。

9.8 虚拟变量

对于“定性数据”(qualitative data)或“分类数据”(categorical data), 需引入“虚拟变量”, 即取值为 0 或 1 的变量。

比如, 性别分男女, 可定义 $D = \begin{cases} 1, & \text{男} \\ 0, & \text{女} \end{cases}$ 。

$i=1=f$
 $i=m,$
Sex
0
1
0
1
✓

对于全球的五大洲, 则需要四个虚拟变量, 即

$D_1 = \begin{cases} 1, & \text{Asia} \\ 0, & \text{other} \end{cases}$, $D_2 = \begin{cases} 1, & \text{America} \\ 0, & \text{other} \end{cases}$, $D_3 = \begin{cases} 1, & \text{Europe} \\ 0, & \text{other} \end{cases}$

$D_4 = \begin{cases} 1, & \text{Africa} \\ 0, & \text{other} \end{cases}$

Asia
0
0
1
0
0
0
America
1
0
0
0
0
0

$i=Asia$ 28

如果 $D_1 = D_2 = D_3 = D_4 = 0$ ，则表明为大洋洲。

在有常数项的模型中，如定性指标分 M 类，最多只能有 $(M - 1)$ 个虚拟变量。

如果引入 M 个虚拟变量，会产生严格多重共线性，因为如果将这 M 个虚拟变量对应的列向量相加，就得到常数项。

这种情况称为“虚拟变量陷阱” (dummy variable trap)。

Stata 会自动识别严格多重共线性，这种担心已不重要。

如果模型中没有常数项，可以有 M 个虚拟变量。

在模型中引入虚拟变量，会带来什么影响呢？

考虑一个有关中国经济的时间序列模型：


$$y_t = \alpha + \beta x_t + \varepsilon_t, \quad t = 1950, \dots, 2000$$

由于经济结构可能在 1978 年后有变化，引入虚拟变量：

$$D = \begin{cases} 1, & \text{若 } t \geq 1978 \\ 0, & \text{其他} \end{cases}$$

考虑以下两种情况。

(1) 仅仅引入虚拟变量本身

$$y_t = \alpha + \beta x_t + \gamma D_t + \varepsilon_t$$


该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ \underbrace{(\alpha + \gamma)} + \beta x_t + \varepsilon_t, & \text{若 } t \geq 1978 \end{cases}$$

仅引入虚拟变量，相当于在不同时期使用不同的截距项。

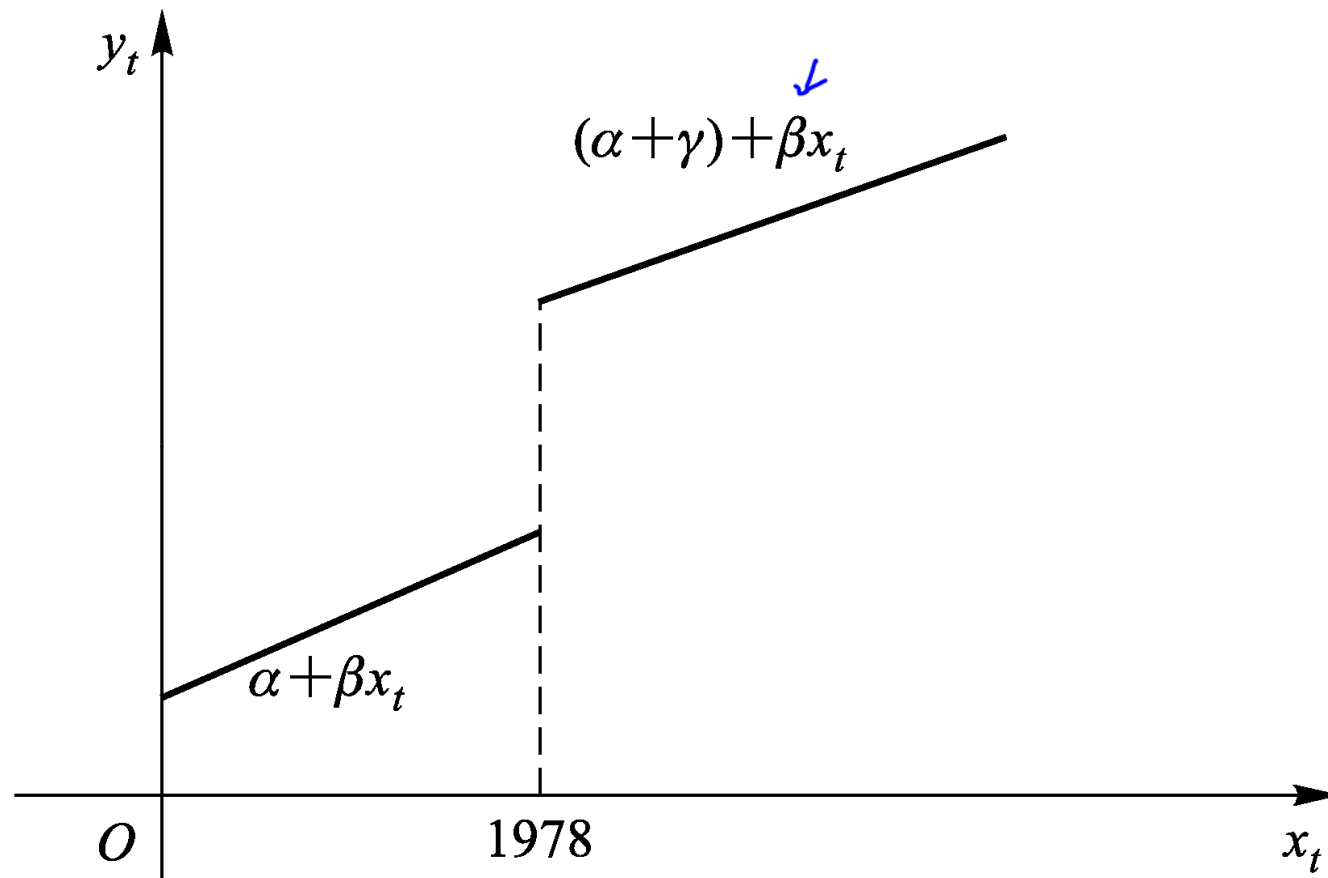


图 9.2 仅引入虚拟变量的效果

(2) 引入虚拟变量，以及虚拟变量与解释变量的“互动项” (interaction term):

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t$$

该模型等价于

Diff-in-Diff

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ (\alpha + \gamma) + \underbrace{(\beta + \delta)}_{\text{斜率}} x_t + \varepsilon_t, & \text{若 } t \geq 1978 \end{cases}$$

引入虚拟变量及其互动项相当于，在不同时期使用不同的截距项与斜率。

如果仅仅引入互动项，则仅改变斜率。

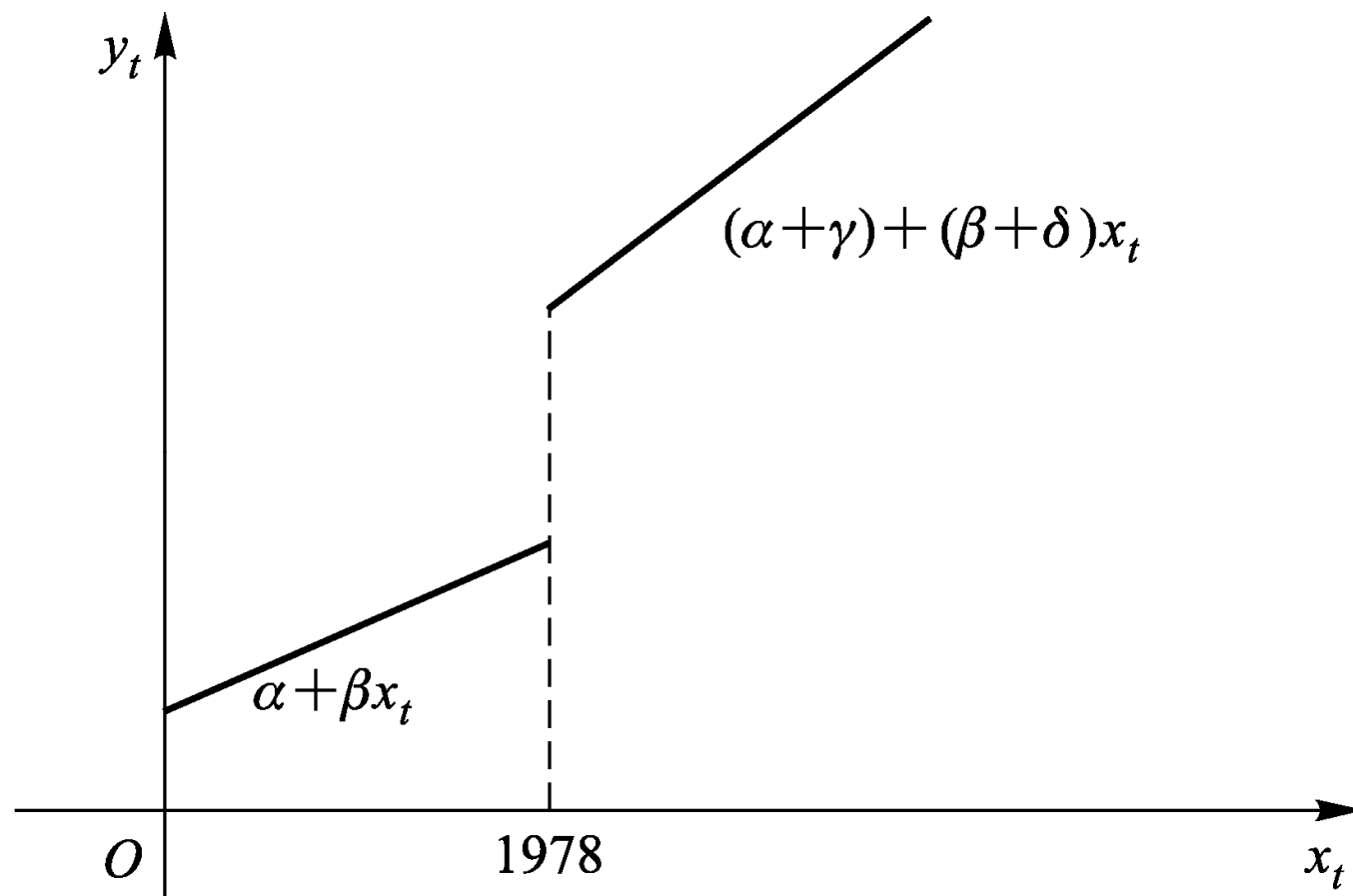


图 9.3 引入虚拟变量及其互动项的效果

Bai & Perron (1998)

9.9 经济结构变动的检验

(2003)
"strucchange"

1. 结构变动日期已知

如果存在“结构变动” (structural break), 但未加考虑, 也是一种模型设定误差。

首先考虑结构变动日期已知的情形。

假设要检验中国经济是否在 1978 年发生结构变动。

定义第 1 个时期为 $1950 \leq t < 1978$, 第 2 个时期为 $1978 \leq t \leq 2000$,

两个时期对应的回归方程可以分别记为

$$y^1 = X^1 \beta^1 + \varepsilon^1$$

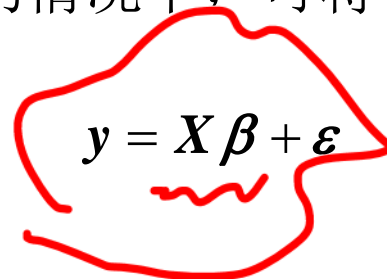
$$y^2 = X^2 \beta^2 + \varepsilon^2$$

需要检验的原假设为，经济结构没有变化，即 $H_0: \beta^1 = \beta^2$ 。

假设有 K 个解释变量，则 H_0 共有 K 个约束。

在无约束的情况下，可对两个时期分别进行回归。

在有约束(即 H_0 成立)的情况下，可将模型合并为


$$y = X \beta + \varepsilon$$

其中， $\mathbf{y} \equiv \begin{pmatrix} \mathbf{y}^1 \\ \mathbf{y}^2 \end{pmatrix}$ ， $\mathbf{X} \equiv \begin{pmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \end{pmatrix}$ ， $\boldsymbol{\varepsilon} \equiv \begin{pmatrix} \boldsymbol{\varepsilon}^1 \\ \boldsymbol{\varepsilon}^2 \end{pmatrix}$ 。可将所有样本数据合在一起回归。

传统的“邹检验” (Chow, 1960). *Gregory Chow.*

首先，回归整个样本 $1950 \leq t \leq 2000$ ，得到残差平方和 $\mathbf{e}'\mathbf{e}$ 。

其次，回归第 1 部分子样本 $1950 \leq t < 1978$ ，得到残差平方和 $\mathbf{e}'_1\mathbf{e}_1$ 。

最后，回归第 2 部分子样本 $1978 \leq t \leq 2000$ ，得到残差平方和 $\mathbf{e}'_2\mathbf{e}_2$ 。

$\mathbf{e}'\mathbf{e} \geq \mathbf{e}'_1\mathbf{e}_1 + \mathbf{e}'_2\mathbf{e}_2$ ，因为将整个样本一起回归为“有约束 OLS”，而将样本一分为二为“无约束 OLS”，故前者的拟合优度比后者更差。

如果 H_0 成立, 则 $(\mathbf{e}'\mathbf{e} - \mathbf{e}'_1\mathbf{e}_1 - \mathbf{e}'_2\mathbf{e}_2)$ 应该比较小。

如果 $(\mathbf{e}'\mathbf{e} - \mathbf{e}'_1\mathbf{e}_1 - \mathbf{e}'_2\mathbf{e}_2)$ 很大, 倾向于认为 H_0 不成立, 存在结构变动。

由于约束条件共有 K 个, 而无约束回归的解释变量个数为 $2K$, 故根据似然比检验原理的 F 统计量为

$$F = \frac{(\mathbf{e}'\mathbf{e} - \mathbf{e}'_1\mathbf{e}_1 - \mathbf{e}'_2\mathbf{e}_2)/K}{(\mathbf{e}'_1\mathbf{e}_1 + \mathbf{e}'_2\mathbf{e}_2)/(n - 2K)} \sim F(K, n - 2K)$$

其中, n 为样本容量, K 为回归方程中解释变量的个数(含截距项)。

检验结构变动的另一简便方法是引入虚拟变量，并检验所有虚拟变量以及其与解释变量交叉项的系数的联合显著性。

进行如下回归：

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t$$

然后检验“ $H_0 : \gamma = \delta = 0$ ”。此

此检验的 F 统计量与传统的邹检验完全相同，二者等价。

虚拟变量法的优点：

(1) 只需生成虚拟变量即可检验，十分方便；

(2) 邹检验在同方差假设下得到，不适用于条件异方差的情形。在条件异方差的情况下，仍可使用虚拟变量法，只要使用异方差稳健的标准误即可。

(3) 如发现存在结构变动，邹检验并不提供究竟是截距项还是斜率变动的信息，而虚拟变量法可提供这些信息。

2. 结构变动日期未知

假设不知道结构变动的具体时间。比如，也许不能肯定结构变动一定发生在 1978 年。

选择一个区间 $[\tau_0, \tau_1] \subseteq [1, T]$ (无法检验过于靠近端点的位置)，其中 T 为样本容量，而 1950 年对应于第 1 年。

trimming parameter.

0.15

计算在此区间中的每一年份 t ($\tau_0 \leq t \leq \tau_1$) 所对应的 F 统计量，然后取其最大者。此统计量称为“匡特似然比” (Quandt Likelihood Ratio, 简记 QLR), 是邹统计量的推广。

QLR 统计量不再服从 F 分布, 其分布取决于约束条件的个数, 以及 (τ_0 / T) 与 (τ_1 / T) 。

如果 τ_0 太接近于 1, 或 τ_1 太接近于 T , 则 QLR 统计量的渐近分布对有限样本分布的近似将变得不准确。

通常选择 $\tau_0 = 0.15T$, $\tau_1 = 0.85T$ (选择最接近的整数), 称为“15% 修边” (15% trimming)。

表 9.1 QLR 统计量临界值表(15%修边)

约束条件个数	10%	5%	1%
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23
11	2.40	2.62	3.09
12	2.33	2.54	2.97

13	2.27	2.46	2.87
14	2.21	2.40	2.78
15	2.16	2.34	2.71
16	2.12	2.29	2.64
17	2.08	2.25	2.58
18	2.05	2.20	2.53
19	2.01	2.17	2.48
20	1.99	2.13	2.43

资料来源：Stock and Watson (2011), p.559, Table 14.6。

如果 QLR 统计量小于临界值，则接受“无结构变动”的原假设。反之，则认为发生了结构变动，而 F 统计量取最大值的那个日期 $\hat{\tau}$ 就是对结构变动日期(break date) τ 的一致估计。

9.10 缺失数据与线性插值

在数据缺失不严重的情况下，为了保持样本容量，可采用“线性插值” (linear interpolation) 的方法来补上缺失数据。

考虑最简单的情形。已知 x_{t-1} 与 x_{t+1} ，但缺失 x_t 的数据，则 x_t 对时间 t 的线性插值为

$$\hat{x}_t = \frac{x_{t-1} + x_{t+1}}{2}$$

一般地，假设与 x (通常为时间) 对应的 y 缺失，而最临近的两个点分别为 (x_0, y_0) 与 (x_1, y_1) ，且 $x_0 < x < x_1$ ，则 y 对 x 的线性插值为

$$\hat{y} = \frac{y_1 - y_0}{x_1 - x_0} (x - x_0) + y_0$$

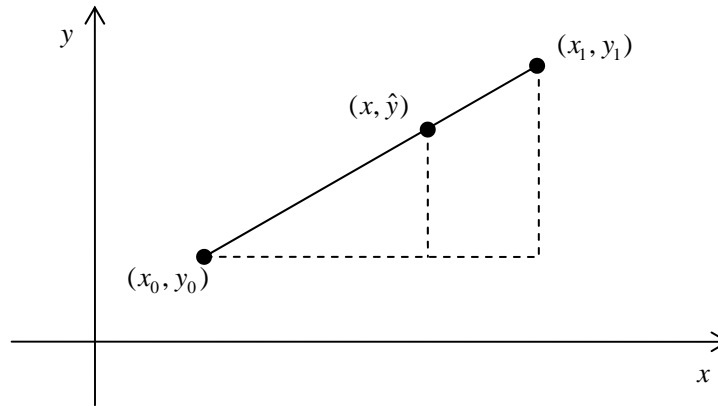


图 9.5 线性插值示意图

如果变量 y 有指数增长趋势(比如 **GDP**), 则应先取对数, 再用 $\ln y$ 进行线性插值, 以避免偏差。

如果需要以原变量 y 进行回归, 可将线性插值的对数值 $\ln \hat{y}$ 再取反对数(antilog), 即计算 $\exp(\ln \hat{y})$ 。

9.11 变量单位的选择

在选择变量单位时，应尽量避免变量间的数量级差别过于悬殊，以免出现计算机运算的较大误差。

比如，通货膨胀率通常小于 1，而如果模型中有 **GDP** 这个变量，则 **GDP** 应该使用亿或万亿作为单位。

否则，变量 **GDP** 的取值将是通货膨胀率的很多倍，即数据矩阵 \mathbf{X} 中某列的数值是另一列的很多倍，这可能使计算机在对 $(\mathbf{X}\mathbf{X})^{-1}$ 进行数值计算时出现较大误差。