

第 1 章 绪 论

1.1 什么是计量经济学

“计量经济学”(Econometrics)是运用概率统计方法对经济变量之间的(因果)关系进行定量分析的科学。

计量经济学常不足以确定经济变量间的因果关系(由于实验数据的缺乏);

多数实证分析正是要确定变量间的因果关系(X 是否导致 Y), 而非仅仅是相关关系。

计量经济学为经济理论服务。(数据建模)

统计学分支。

联立方程模型

实验

因果 \Rightarrow 预测

【例】看到街上人们带伞，可预测今天要下雨。这是相关关系；“人们带伞”并不造成“下雨”。

计量分析须建立在经济理论基础。但即使有理论，因果关系依然不好分辨。

首先，可能存在“逆向因果” (reverse causality)。

【例】FDI 促进经济增长，但 FDI 也可能被吸引到高增长地区。

Foreign Direct Investment

其次，可能是被遗漏的第三个变量(Z)对这两个变量(X, Y)同时起作用。

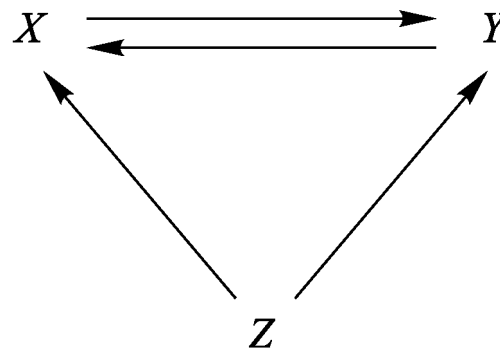


图 1.1 可能的因果关系

例：决定教育投资回报率(returns to schooling)的因素

$$\ln W_i = \alpha + \beta S_i + \varepsilon_i$$

其中， $\ln W$ (工资对数)为“被解释变量” (dependent variable),
 S (教育年限)为“解释变量” (explanatory variable, regressor), ε
为“随机扰动项”(stochastic disturbance)或“误差项”(error term);

下标 i 表示第 i 个观测值(个体 i); α 与 β 为待估参数。

用数据估计此一元回归会发现,工资与受教育年限显著正相关,而且教育投资回报率 β 还挺高。

但工资收入也与能力有关;能力无法观测,而能力高的人通常选择接受更多教育。教育的高回报率包含了对能力的回报。

影响工资收入的因素还可能包括工作经验、毕业学校、人种、性别、外貌等。

须尽可能多地引入“控制变量”(control variables),即多元回归的方法,才能准确估计“感兴趣的参数”(parameters of interest),即本例的教育投资回报率 β 。

现实中总有某些相关变量无法观测，即“遗漏变量” (omitted variables)，都被纳入随机扰动项 ε_i 中。

如果真实模型为 $\frac{\partial E(\ln W_i)}{\partial S} = \beta + \gamma S$

$$\ln W_i = \alpha + \beta S_i + \gamma S_i^2 + \varepsilon_i$$

$E(\varepsilon_i) = 0$ 围绕
 S^2
 “中心话题”

则 γS_i^2 被纳入到扰动项中了(可视为遗漏变量)。

如果变量测量得不准确，则测量误差也被放入扰动项中了。

扰动项就像是“垃圾桶”，所有不想要、无法把握的东西都往里面扔。但又希望扰动项有很好的性质，常导致自相矛盾。

“The devil is in the details.” \Rightarrow “The devil is in the error term.”

计量经济学的很多玄妙之处就在于扰动项。

1.2 经济数据的特点与类型

经济学通常无法像自然科学那样做“控制实验”(controlled experiment)，故经济数据一般不是“实验数据”(experimental data)，而是自然发生的“观测数据”(observational data)。

由于个人行为的随机性，经济变量原则上都是随机变量。

$$y = x\beta + \varepsilon$$

关系
随机
fixed
传统最小二乘法

本科教学中，有时假设解释变量是非随机的、固定的(fixed regressors)。

这只是教学法上的权宜之计。如果解释变量为非随机，则无法考虑其与扰动项的相关性。

在本研究生课程中，所有变量都是随机的（即使非随机的常数，也可视为退化的随机变量）。

经济数据按照其性质，可大致分成三种类型：

- 横截面数据(cross-sectional data, 简称截面数据): 多个经济个体的变量在同一时点上的取值。比如，2012 年中国各省的 GDP。

$N \rightarrow \infty$

- 时间序列数据(time series data): 某个经济个体的变量在不同时点上的取值。比如，在 1978—2012 年山东省每年的 GDP。

$T \rightarrow \infty$

- 面板数据(panel data): 多个经济个体的变量在不同时点上的取值。比如，在 1978—2012 年中国各省每年的 GDP。

短面板

横截面: N : 大 $\leftarrow \infty$

时间: T : 小 $\leftarrow \text{finite}$

长: $T \rightarrow \infty$
 N 有限

第 3 章 小样本 OLS

Finite 有限. N 多大没有关系, 结论都对.

大样本. $N \rightarrow \infty$ 结论才对.

3.1 古典线性回归模型的假定

“最小二乘法” (Ordinary Least Square, OLS) 是单一方程线性回归模型的基本估计方法。“古典线性回归模型” (Classical Linear Regression Model) 的假定如下。

CLR \neq OLS

假定 3.1 线性假定(linearity)。总体(population)模型为

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \quad (i = 1, \dots, n)$$

$i: 1, \dots, n (N)$

β 个数: K 个.

1

$$y_i = \vec{x}_i' \vec{\beta} + \varepsilon_i, (i = 1, \dots, n)$$

n 为样本容量，解释变量 x_{ik} 的第一个下标表示第 i 个观测值，第二个下标则表示第 k 个解释变量 ($k = 1, \dots, K$)。

如有常数项，令第一个解释变量为单位向量，即 $x_{i1} \equiv 1, \forall i$ 。

$\beta_1, \beta_2, \dots, \beta_K$ 为待估参数，称为“回归系数” (regression coefficients)。

线性假设的含义是 x_{ik} 对 y_i 的边际效应为常数，比如 $\frac{\partial \overset{\text{期望}}{\text{E}(y_i)}}{\partial x_{i1}} = \beta_1$ 。

如果边际效应可变，可加入平方项 (x_{ik}^2) 或交叉互动项 ($x_{ik}x_{im}$)。

线性: β 的线性组合.

比如, $y_i = \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \gamma x_{ik} x_{im} + \varepsilon_i$

则 x_{ik} 对 y_i 的平均边际效应为 $\frac{\partial E(y_i)}{\partial x_{ik}} = \beta_k + \gamma x_m$ 。

只要把高次项也作为解释变量来看待, 则依然满足线性假定。

总体模型也称“数据生成过程” (Data Generating Process, DGP)。^{主观}

记第 i 个观测数据为 $\mathbf{x}_i \equiv (x_{i1} \ x_{i2} \ \cdots \ x_{iK})'$, $\boldsymbol{\beta} \equiv (\beta_1 \ \beta_2 \ \cdots \ \beta_K)'$, 则

$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n)$

^{同分布}
↓ ↓

强调: 同分布

把所有个体的方程叠放可得

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix} \underset{k \times 1}{\boldsymbol{\beta}} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\vec{\mathbf{x}}_1 : k \times 1$
 \mathbf{x}'_1 (circled)

定义 $\mathbf{y} \equiv (y_1 \ y_2 \ \cdots \ y_n)'$, 数据矩阵 $\mathbf{X} \equiv (\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n)'$, $\boldsymbol{\varepsilon} \equiv (\varepsilon_1 \ \varepsilon_2 \ \cdots \ \varepsilon_n)'$, 则

$\mathbf{X} : \text{行: 观测}$
 $\text{列: 不同 } \mathbf{x}_k$
 $n \times k$

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\mathbf{y} : n \times 1$ $\mathbf{X} : n \times k$ $\boldsymbol{\beta} : k \times 1$ $\boldsymbol{\varepsilon} : n \times 1$

- The linearity assumption is on the *parameters*, i.e. as long as we can write the model in terms of linear combinations of the $\boldsymbol{\beta}$, we could use this framework. Thus it is not quite so restrictive as it might seem.

- For example, the model

$$y = Ax^{\beta}e^{\epsilon} \quad (7)$$

meets the linearity assumption because taking the log of both sides reveals

$$\ln(y) = \alpha + \beta \ln(x) + \epsilon \quad (8)$$

- The most important implication of the linearity assumption is that the marginal effects (the β parameters) are constant and do not depend on the x variables.

Linearity: examples

- Let's look at some examples.
- Consider the simple consumption function

$$\underbrace{CON}_\alpha_i = \beta_1 + \underbrace{\beta_2}_{\alpha} \underbrace{YD}_i + \epsilon_i \quad (9)$$

$0 < \beta_2 < 1$
 $X_1 = 1$
 $X_2 = YD_i$

where CON is consumption and YD is disposable income. The unit of observation i can be an individual household, or it can be an economywide aggregate for a certain year.

- In this case, the error term represents other variables besides YD that influence consumption (for instance, financial assets or the “mood” of the consumer).
- The parameter β_2 is the marginal propensity to consume out of disposable income, and it should be between zero and one.

Linearity: examples

- When the equation has only one nonconstant regressor, we call it the simple regression model.
- It can be written in matrix form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\mathbf{y} = \begin{bmatrix} CON_1 \\ CON_2 \\ \dots \\ CON_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x'_1 \\ x'_2 \\ \dots \\ x'_n \end{bmatrix} = \begin{bmatrix} 1 & YD_1 \\ 1 & YD_2 \\ \dots & \dots \\ 1 & YD_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix} \quad (10)$$

Linearity: examples

- As a second example, consider a simplified version of the wage equation routinely estimated in labor economics

$$\ln(WAGE_i) = \beta_1 + \beta_2 S_i + \beta_3 TENURE_i + \beta_4 EXPR_i + \epsilon_i, \quad (11)$$

Handwritten notes: S_i 1單位 \rightarrow wage %

where $WAGE$ is the individual's wage rate, S is education in years, $TENURE$ is years in the current job, and $EXPR$ is experience in the labor market.

- This equation is said to be in **semi-log** form because only the dependent variable is in logs.
- This is derived from the following nonlinear relationship between the level of the wage rate and the regressors:

$$WAGE_i = \exp(\beta_1) \exp(\beta_2 S_i) \exp(\beta_3 TENURE_i) \exp(\beta_4 EXPR_i) \exp(\epsilon_i) \quad (12)$$

Handwritten notes: Red arrows pointing down to each exponential term in equation (12).

- The coefficients have the interpretation of percentage changes.
- For instance, if $\beta_2 = 0.05$, an additional year of schooling raises wages by 5%.

- This is because $e^x \approx 1 + x$ when x is close to 0, so

$$\frac{WAGE_{\text{new}}}{WAGE} = e^{\beta_2} \approx 1 + \beta_2.$$

- Sometimes when β_2 is large, it is preferred to report the exact change in percentage, $e^{\beta_2} - 1$.

$$= \frac{Wage_{\text{new}}}{Wage}$$

Linearity: examples

- Certain other forms of nonlinearities can also be accommodated.
- For instance, suppose that the marginal effect of education on wages declines as the level of education gets higher.
- This can be captured by including S^2 as an additional independent variable in the wage equation.
- Then the marginal effect of education on wages is

$$\frac{\partial \ln(WAGE)}{\partial S} = \beta_2 + 2\beta_5 S \quad (13)$$

- If β_5 is negative, then the marginal effect of education declines as S increases.

Linearity: examples

- Another common specification is the loglinear or constant elasticity model:

$$\ln(y) = \beta_1 + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + \cdots + \beta_K \ln(x_K) + \epsilon \quad (14)$$

Handwritten notes: $x\% \rightarrow y\%$ and β_3 is circled with the Chinese characters "弹性" (elasticity).

- It is called the “constant elasticity” model because the elasticity of y with respect to changes in x is

$$\frac{\partial \ln(y)}{\partial \ln(x_k)} = \beta_k \frac{\frac{\partial y}{y}}{\frac{\partial x}{x}} \quad (15)$$

Handwritten notes: $\frac{\partial y}{y}$ and $\frac{\partial x}{x}$ are written above the fraction in the equation.

which is constant and does not depend on x_k .

- The point is that the linear framework is much more flexible than it might first appear.
- Flexible though this framework may be, there are still cases of genuine nonlinearity that this model cannot accommodate.
- For instance, in the wage equation, if the error term entered additively, the model could not be linearized:

$$WAGE_i = \exp(\beta_1)\exp(\beta_2 S_i)\exp(\beta_3 TENURE_i)\exp(\beta_4 EXPR_i) + \epsilon_i \quad (21)$$

- Another example:

$$y = \alpha + \frac{1}{\beta_1 + \beta_2 x} + \epsilon \quad (22)$$

x : fixed

★ 假定 3.2 严格外生性(strict exogeneity)

x : 随机.

$$E(\varepsilon_i | x) = 0.$$

$$E(\varepsilon_i) = 0$$

$$E(\varepsilon_i | \mathbf{X}) = E(\varepsilon_i | x_1, \dots, x_n) = 0 \quad (i = 1, \dots, n)$$

$$E(\varepsilon_i) = E(E(\varepsilon_i | x)) = 0$$

$n \times k$

ε_i 均值独立于所有解释变量的观测数据，而不仅仅是同一观测数据 \mathbf{x}_i 中的解释变量。
 $i = 1, \dots, n$ $i = 3$

$$E(\varepsilon_3 | x_1, x_2, x_3, \dots, x_n) \neq 0.$$

ε_i 与所有解释变量都不相关，即 $\text{Cov}(\varepsilon_i, x_{jk}) = 0, \forall j, k$ 。此假定很强，在第 5 章可放松。
时间序列中通常不对。

均值独立仅要求 $E(\varepsilon_i | \mathbf{X}) = c$ ， c 为某常数，不一定为 0。

当回归方程有常数项时，如果 $E(\varepsilon_i | \mathbf{X}) = c \neq 0$ ，总可以把 c 归入常数项要求。

$$\text{Cov}(\varepsilon_i, x_{jk}) = E(\varepsilon_i x_{jk}) - E(\varepsilon_i) E(x_{jk})$$

$$= E(\varepsilon_i x_{jk}) = E(E(\varepsilon_i x_{jk} | x_{jk})) = E(x_{jk} E(\varepsilon_i | x_{jk}))$$

$$E[\varepsilon_i | x] = 0 \Rightarrow \text{Cov}(\varepsilon_i, x_{jk}) = 0, \forall i, j, k \Rightarrow 0$$

$$E[\varepsilon_i | X] = c \neq 0, \quad y = \beta_0 + x\beta + \varepsilon = \underbrace{\beta_0 + c}_{\beta_0} + x\beta + \underbrace{(\varepsilon - c)}_{\varepsilon}$$

命题 $E(\varepsilon_i) = 0$ ，即扰动项的无条件期望为 0。

证明：根据迭代期望定律， $E(\varepsilon_i) = E_X \underbrace{E(\varepsilon_i | X)}_{=0} = E_X(0) = 0$ 。

定义 如果随机变量 X, Y 满足 $E(XY) = 0$ ，则称 X, Y 正交 (orthogonal)。

命题 解释变量与扰动项正交。

证明： $0 = \text{Cov}(x_{jk}, \varepsilon_i) = E(x_{jk}\varepsilon_i) - E(x_{jk}) \underbrace{E(\varepsilon_i)}_{=0} = E(x_{jk}\varepsilon_i)$ 。

$$\text{Saving} = \beta_0 \text{Income} + \varepsilon$$

$$\text{Income} = 0, \quad \text{Saving} = 0.$$

CAPM. α ?

$$E(r_i) = \alpha + \beta E(r_m - r_f) \quad ?$$

↑
①

假定 3.3 不存在“严格多重共线性”(strict multicollinearity),
即数据矩阵 X 满列秩, $\text{rank}(X) = K$, 其中“rank”表示矩阵的秩。
 $n \times k$, $n > k$,

如果不满足此条件, 则 β “不可识别”(unidentified), 因为 X 中
某个或多个变量为多余。

根据 OLS 估计, $b = (X'X)^{-1} X'y$ 。

$$y = X\beta + \varepsilon$$
$$b = (X'X)^{-1} X'y$$

如 X 满列秩, $X'X$ 正定, 故 $(X'X)^{-1}$ 存在; 反之, $(X'X)^{-1}$ 不存在。

实际数据不易出现严格多重共线性。

x_1, x_2, \dots, x_k

如果出现, Stata 也会自动识别。

X

虚拟变量

亚洲 1
其它 0

假定 3.4 球型扰动项(spherical disturbance), 即扰动项满足“同方差”、“无自相关”的性质,

$$\vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad X \quad n \times k$$

$$\text{Var}(\varepsilon | X) = E(\varepsilon \varepsilon' | X) = \sigma^2 I_n$$

$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$
无自相关

$$I = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$$

同方差.

$$\text{Var}(\varepsilon_i) = \sigma^2, i = 1, \dots, n.$$

I_n 为 n 阶单位矩阵。

协方差矩阵 $\text{Var}(\varepsilon | X)$ 的主对角线元素都等于 σ^2 , 即满足“条件同方差” (conditional homoskedasticity); 反之, 则存在“条件异方差” (conditional heteroskedasticity)。

协方差矩阵 $\text{Var}(\varepsilon | X)$ 的非主对角线元素都为0, 不同个体的扰动项之间无“自相关” (autocorrelation); 反之, 则存在自相关。

CLR: 理论模型 (经济理论, 近似, 脑子里想)

OLS: 估计 CLR 方法. 假设 CLR 对的. 上述假设是对的
Ordinary Least Squares.

3.2 OLS 的代数推导 最小二乘法

对于 β 的任意假想值 $\tilde{\beta}$, 记个体 i 的拟合误差 (即残差, residual) 为 $e_i = y_i - x_i' \tilde{\beta}$.

$$y_i = x_i' \beta + \varepsilon_i$$

将所有个体的残差叠放, 可得残差向量 $e = (e_1 \ e_2 \ \cdots \ e_n)' = y - X\tilde{\beta}$.

最小二乘法寻找能使残差平方和 (Sum of Squared Residuals, SSR) $\sum_{i=1}^n e_i^2$ 最小的 $\tilde{\beta}$.

几何上, 一元回归就是寻找最佳拟合的回归直线;

二元回归就是寻找最佳拟合的回归平面;

多元回归, 则寻找最佳拟合的回归超平面 (superplane)。

最小化问题:

$e: n \times 1$

$$\min_{\tilde{\beta}} \text{SSR}(\tilde{\beta}) = \sum_{i=1}^n e_i^2 = e'e \quad (\text{平方和写成向量内积})$$

$$= (y - X\tilde{\beta})'(y - X\tilde{\beta}) \quad (\text{残差向量的表达式})$$

$$= (y' - \tilde{\beta}'X')(y - X\tilde{\beta}) \quad (\text{矩阵转置性质})$$

$$= y'y - y'X\tilde{\beta} - \tilde{\beta}'X'y + \tilde{\beta}'X'X\tilde{\beta} \quad (\text{乘积展开})$$

$$= y'y - 2y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta} \quad (\text{合并同类项})$$

$(y'X\tilde{\beta})' = \tilde{\beta}'X'y$ (对称矩阵), 为 1×1 常数, 故相等, 可合为 $2y'X\tilde{\beta}$ 。

目标函数 $\text{SSR}(\tilde{\beta})$ 是 $\tilde{\beta}$ 的二次函数(二次型)。

$$\begin{matrix} X \\ n \times k \end{matrix}, \quad \begin{matrix} X' & y \\ k \times n & n \times 1 \end{matrix}, \quad k \times 1$$

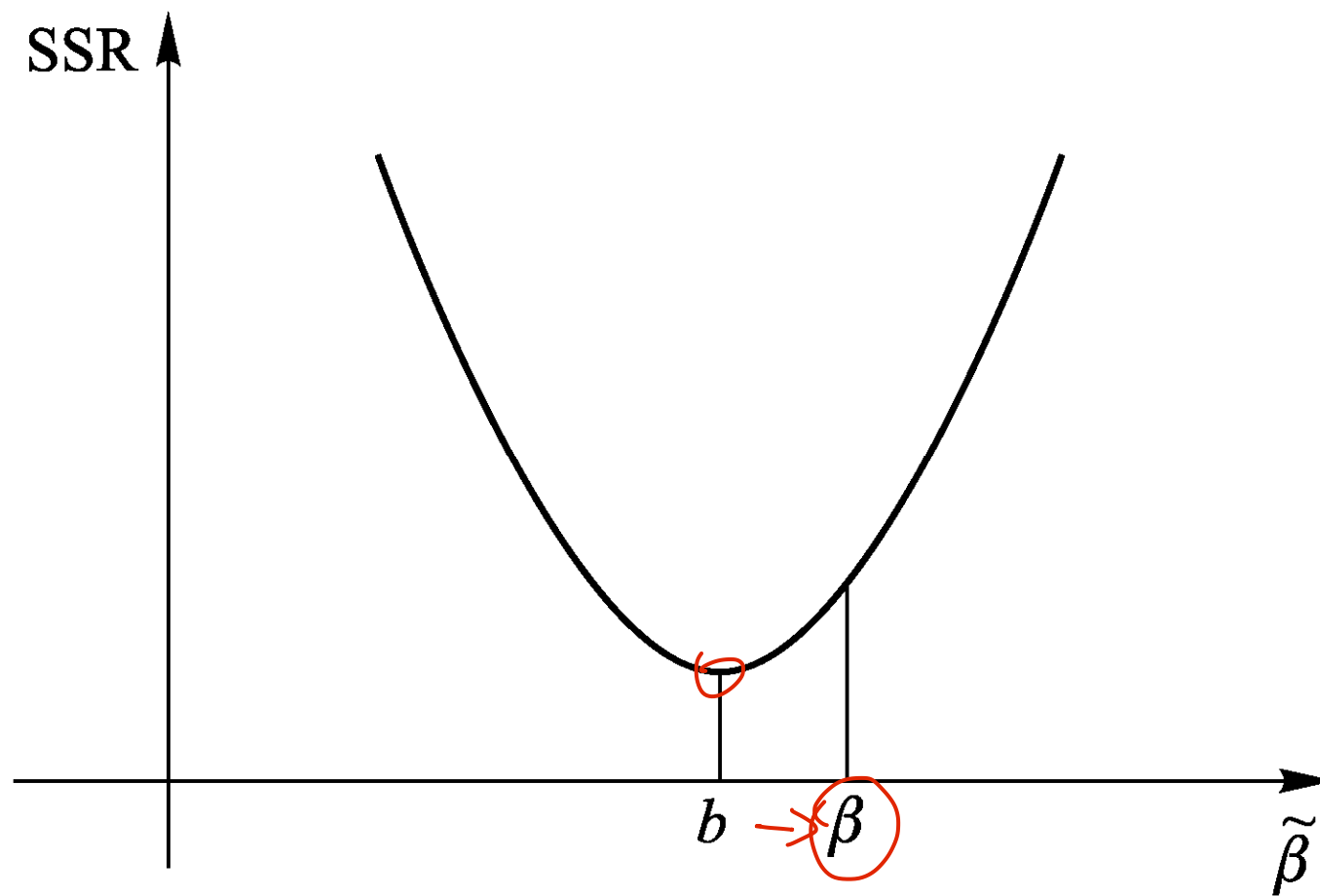


图 3.1 参数的假想值 $\tilde{\beta}$ 、真实值 β 与 OLS 估计值 b

$$\frac{\partial SSR}{\partial \tilde{\beta}}$$

向量微分的规则：

记 $\mathbf{a} = (a_1 \ a_2 \ \cdots \ a_K)'$ ，则 $\mathbf{a}'\tilde{\beta} = \sum_{i=1}^K a_i \tilde{\beta}_i$ 。

$$\frac{\partial(\mathbf{a}'\tilde{\beta})}{\partial \tilde{\beta}_{K \times 1}} \equiv \left(\frac{\partial(\mathbf{a}'\tilde{\beta})}{\partial \tilde{\beta}_1} \ \frac{\partial(\mathbf{a}'\tilde{\beta})}{\partial \tilde{\beta}_2} \ \cdots \ \frac{\partial(\mathbf{a}'\tilde{\beta})}{\partial \tilde{\beta}_K} \right)' = (a_1 \ a_2 \ \cdots \ a_K)' = \mathbf{a}_{K \times 1}$$

$a_1\beta_1 + a_2\beta_2 + \cdots + a_K\beta_K$

类似于对一次函数求导。假设 A 为 K 阶对称矩阵，可证：

$$\frac{\partial(\tilde{\beta}' A \tilde{\beta})}{\partial \tilde{\beta}} \equiv \left(\frac{\partial(\tilde{\beta}' A \tilde{\beta})}{\partial \tilde{\beta}_1} \ \frac{\partial(\tilde{\beta}' A \tilde{\beta})}{\partial \tilde{\beta}_2} \ \cdots \ \frac{\partial(\tilde{\beta}' A \tilde{\beta})}{\partial \tilde{\beta}_K} \right)' = 2A\tilde{\beta}$$

类似于对二次函数求导。

$$\tilde{\beta}' A \tilde{\beta} = \beta_1^2 a_{11} + \beta_1 \beta_2 a_{12} + \cdots + \beta_1 \beta_K a_{1K} + \beta_2 \beta_1 a_{21} + \beta_2^2 a_{22} + \cdots$$

使用向量微分规则，可得最小化的一阶条件：

$$\frac{\partial(\text{SSR})}{\partial \tilde{\beta}} = \underbrace{-2X'y}_{K \times 1} + \underbrace{2X'X\tilde{\beta}}_{K \times 1} = 0$$

最小二乘估计量 b 满足：

Normal Equation

$$(X'X)_{K \times K} b_{K \times 1} = X'_{K \times n} y_{n \times 1} \quad (\text{正规方程组, } K \text{ 个方程, } K \text{ 个未知数})$$

$$X'y - (X'X)b = 0 \quad (\text{移项})$$

$$X' \underbrace{(y - Xb)}_{=e} = 0 \quad (\text{向左提取共同的矩阵因子 } X')$$

因此， $X'e = 0$ ，其中残差向量 $e \equiv y - Xb$ 。

$X'e$ $n \times K$

$\frac{1}{n} \sum_{i=1}^n x_{ik} e_i = 0, \quad k=1, \dots, K$

$\frac{1}{n} \sum_{i=1}^n x_{ik} e_i \xrightarrow{\text{大数定律}} E(x_{ik} e_i) = 0$

假设 2 $E(x_{ik} e_i) = 0, \quad k=1, \dots, K$

$$E(XY) = 0.$$

$$CLR \leftrightarrow OLS.$$

残差向量 e 与解释变量 X 正交，是 OLS 的一大特征。

没问题, $e_i \rightarrow \varepsilon_i$

$$\frac{1}{n} \sum_{i=1}^n x_{ik} e_i \rightarrow$$

$$E(x_{ik} \varepsilon_i)$$

求解可得 OLS 估计量:

$$E(x_{ik} \varepsilon_i) \neq 0$$

再用 OLS 估计 CLR 就不对.
OLS 自动 Normal Equation

$$b \equiv (X'X)^{-1} X'y$$

二阶条件要求黑赛矩阵(Hessian)

$$\frac{\partial^2 (SSR)}{\partial \tilde{\beta} \partial \tilde{\beta}'} \equiv \frac{\partial \left(\frac{\partial SSR}{\partial \tilde{\beta}} \right)}{\partial \tilde{\beta}'} \equiv \begin{pmatrix} \frac{\partial^2 SSR}{\partial^2 \tilde{\beta}_1} & \dots & \frac{\partial^2 SSR}{\partial \tilde{\beta}_1 \partial \tilde{\beta}_K} \\ \dots & \dots & \dots \\ \frac{\partial^2 SSR}{\partial \tilde{\beta}_K \partial \tilde{\beta}_1} & \dots & \frac{\partial^2 SSR}{\partial^2 \tilde{\beta}_K} \end{pmatrix} = \underline{2X'X} \text{ 为正定矩阵}$$

$$x'x$$

14

$$a'(x'x)a = (xa)'xa \geq 0$$

因为 \mathbf{X} 满列秩，故 $\mathbf{X}'\mathbf{X}$ 正定。

被解释变量的“拟合值” (fitted values) 或 “预测值” (predicted values):

$$\hat{\mathbf{y}} \quad \hat{\boldsymbol{\beta}}$$

$$\hat{\mathbf{y}} \equiv (\hat{y}_1 \ \hat{y}_2 \ \cdots \ \hat{y}_n)' \equiv \mathbf{X}\mathbf{b}$$

可把被解释变量 \mathbf{y} 分解为两个正交的部分，即 $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$ ，而 $\hat{\mathbf{y}}$ 与 \mathbf{e} 正交，因为

$$\hat{\mathbf{y}}'\mathbf{e} = (\mathbf{X}\mathbf{b})'\mathbf{e} = \mathbf{b}'\mathbf{X}'\mathbf{e} = \mathbf{b}' \cdot \mathbf{0} = \mathbf{0}$$

对于扰动项方差 $\sigma^2 = \text{Var}(\varepsilon_i)$ 的估计：

$$s^2 \equiv \frac{1}{n-K} \sum_{i=1}^n e_i^2$$

Frequentist 频率学派.

β : 客观存在, 不变常数.

$\hat{\beta}$: 估计 β , 随机变量.

$$y = x\beta + \varepsilon.$$

$$\begin{aligned} \hat{\beta} &= \frac{\sum y}{\sum x} \text{ 来估计 } \beta \\ &= \frac{\sum (x\beta + \varepsilon)}{\sum x} = \beta + \frac{\sum \varepsilon}{\sum x} \end{aligned}$$

假设 x 固定.

ε 都大, $\hat{\beta}$ 偏大.
小, $\hat{\beta}$ 偏小.

假设知道 β , x ① 随机抽 ε , 45 个.

$$\text{计算 45 个 } y. \rightarrow \hat{\beta}_1 = \frac{\sum y_1}{\sum x}.$$

② 再随机抽 ε , 45 个.

$$\hat{\beta}_2 = \frac{\sum y_2}{\sum x}$$

$\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{1000}$ β 抽样分布.

$$\begin{aligned} &y_1, \dots, y_{45} \\ &x_1, \dots, x_{45} \end{aligned}$$

β 只有一个值.

β^*

其中， $(n-K)$ 为自由度。为什么除以 $(n-K)$ 而不除以 n ?

随机变量 $\{e_1, e_2, \dots, e_n\}$ 必须满足 K 个正规方程 $\mathbf{X}'\mathbf{e} = \mathbf{0}$ ，故只有其中 $(n-K)$ 个 e_i 是相互独立（自由）的。

经过校正后，才是无偏估计， $E(s^2) = \sigma^2$ 。

如果样本容量 n 很大($n \rightarrow \infty$)，则 $\frac{n-K}{n} \rightarrow 1$ ，是否进行“小样本校正”并没有多少差别。

称 $s = \sqrt{s^2}$ 为“回归方程的标准误差”(standard error of the regression)，简称“回归方程的标准误”。

称某统计量的标准差为该统计量的“标准误”(standard error)。

3.3 OLS 的几何解释

\hat{y} 是 y 向超平面 X 的投影(projection), 因为 e 与 X 正交。

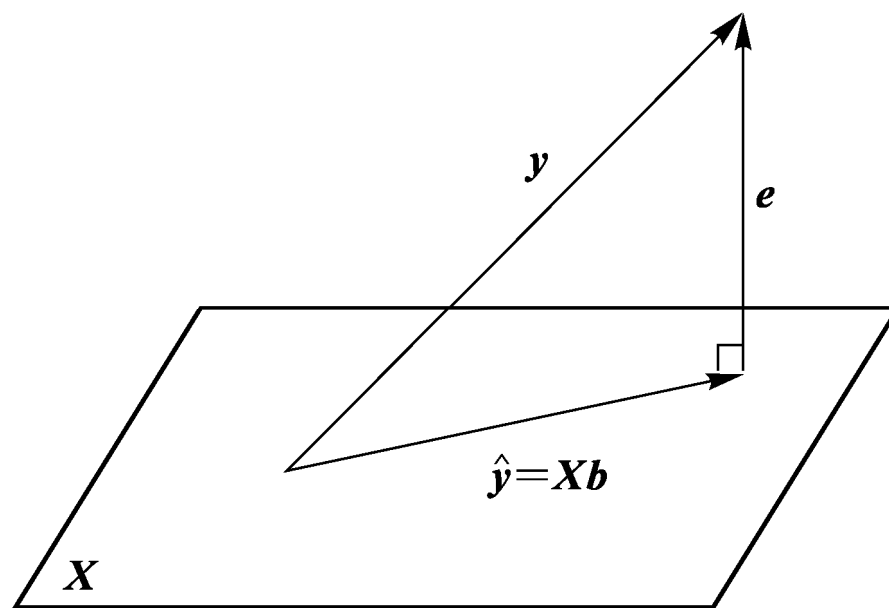


图 3.2 最小二乘法的正交性

由于 $\hat{\mathbf{y}} \equiv \mathbf{X}\mathbf{b} = \mathbf{X} \underbrace{(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}}_{=\mathbf{b}} \equiv \mathbf{P}\mathbf{y}$ ，故 $\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ 称为“投影矩阵” (projection matrix)。

用 \mathbf{P} 左乘任何向量，就得到该向量在超平面 \mathbf{X} 上的投影。

由于 $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{P}\mathbf{y} = (\mathbf{I}_n - \mathbf{P})\mathbf{y} \equiv \mathbf{M}\mathbf{y}$ ，故 $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{P}$ 称为“消灭矩阵” (annihilator matrix)。

用消灭矩阵 \mathbf{M} 左乘任何向量，就得到该向量对超平面 \mathbf{X} 投影后的残差向量。

矩阵 \mathbf{P} 与 \mathbf{M} 的性质(参见习题):

(i) $\mathbf{P}\mathbf{X} = \mathbf{X}$; (自己的投影还是自己)

(ii) $\mathbf{P}\mathbf{e} = \mathbf{0}$; (垂直于 \mathbf{X} 的向量 \mathbf{e} 投影于 \mathbf{X} 则退化为一个点)

(iii) $\mathbf{M}\mathbf{X} = \mathbf{0}$; (自己对自己投影, 其残差为 $\mathbf{0}$)

(iv) \mathbf{P} 与 \mathbf{M} 都是对称阵;

(v) $\mathbf{P}^2 = \mathbf{P}$; (再次投影的效果等于一次投影)

(vi) $\mathbf{M}^2 = \mathbf{M}$ 。 (再次消灭的效果等于一次消灭)

把残差写成 $\boldsymbol{\varepsilon}$ 的函数：

$$\boldsymbol{e} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \underbrace{\boldsymbol{M}\boldsymbol{X}}_{=\mathbf{0}}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon}$$

把残差平方和写为 $\boldsymbol{\varepsilon}$ 的函数：

$$\text{SSR} = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{M}\boldsymbol{\varepsilon})'\boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}^2\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{M}\boldsymbol{\varepsilon}$$

3.4 拟合优度

如果回归方程有常数项，则 $\sum_{i=1}^n (y_i - \bar{y})^2$ 可分解为：

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n e_i^2$$

其中, $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ 为样本均值。

被解释变量 y_i 的离差平方和可分为两部分, 即可由模型解释的部分 $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, 与无法由模型解释的残差部分 $\sum_{i=1}^n e_i^2$ 。

平方和分解公式能成立正是由于 OLS 的正交性(参见习题)。

定义 “拟合优度” (goodness of fit) R^2 为

$$0 \leq R^2 \equiv \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \leq 1$$

拟合优度 R^2 也称“可决系数” (coefficient of determination)。

可以证明(参见习题), 在有常数项的情况下, 拟合优度等于被解释变量 y_i 与拟合值 \hat{y}_i 之间相关系数的平方, 即 $R^2 = [\text{Corr}(y_i, \hat{y}_i)]^2$ 。

R^2 越高, 拟合程度越好。

缺点: 如果增加解释变量, R^2 只增不减, 因为至少可让新增解释变量的系数为 0 而保持 R^2 不变。

通过调整自由度, 对解释变量过多(模型不够简洁)进行惩罚。

定义 校正拟合优度 (adjusted R^2) \bar{R}^2 为

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^n e_i^2 / (n - K)}{\sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)}$$

缺点: \bar{R}^2 可能为负。

无论 R^2 还是 \bar{R}^2 , 只反映拟合程度好坏, 除此无太多意义。

评估回归方程是否显著, 应使用 F 检验(R^2 与 F 统计量有联系)。

如果回归模型无常数项，则平方和分解公式不成立。仍可将 $\sum_{i=1}^n y_i^2$ 分解为：

$$\sum_{i=1}^n y_i^2 = \mathbf{y}'\mathbf{y} = (\hat{\mathbf{y}} + \mathbf{e})'(\hat{\mathbf{y}} + \mathbf{e}) = \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\underbrace{\hat{\mathbf{y}}'\mathbf{e}}_{=0} + \mathbf{e}'\mathbf{e} = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n e_i^2$$

定义“非中心 R^2 ” (Uncentered R^2):

$$R_{uc}^2 \equiv \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y}}$$

如果无常数项，则 Stata 汇报 R_{uc}^2 。

3.5 OLS 的小样本性质

(1) 线性性：OLS 估计量 $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ 为 \mathbf{y} 的线性组合。

(2) 无偏性： $E(\mathbf{b} | \mathbf{X}) = \boldsymbol{\beta}$ ，即 \mathbf{b} 不会系统地高估或低估 $\boldsymbol{\beta}$ 。

证明：抽样误差(sampling error)为

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} \equiv \mathbf{A}\boldsymbol{\varepsilon}$$

其中，记 $\mathbf{A} \equiv (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ 。所以

$$E(\mathbf{b} - \boldsymbol{\beta} | \mathbf{X}) = E(\mathbf{A}\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{A} \underbrace{E(\boldsymbol{\varepsilon} | \mathbf{X})}_{=0} = \mathbf{0} \quad (\text{严格外生性})$$

移项可得， $E(\mathbf{b} | \mathbf{X}) = \boldsymbol{\beta}$ 。在此证明中，严格外生性不可或缺。

推论 无条件期望 $E(\mathbf{b}) = \boldsymbol{\beta}$ 。

证明： $E(\mathbf{b}) = E_{\mathbf{X}} E(\mathbf{b} | \mathbf{X}) = E_{\mathbf{X}}(\boldsymbol{\beta}) = \boldsymbol{\beta}$ (常数的期望仍为常数)。

(3) 估计量 \mathbf{b} 的方差为 $\text{Var}(\mathbf{b} | \mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ 。

证明： $\text{Var}(\mathbf{b} | \mathbf{X}) = \text{Var}(\mathbf{b} - \boldsymbol{\beta} | \mathbf{X})$ ($\boldsymbol{\beta}$ 是常数)
 $= \text{Var}(\mathbf{A}\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{A} \text{Var}(\boldsymbol{\varepsilon} | \mathbf{X}) \mathbf{A}' = \mathbf{A} \sigma^2 \mathbf{I}_n \mathbf{A}'$
 $= \sigma^2 \mathbf{A} \mathbf{A}' = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$

球形扰动项假定是证明的关键。

如存在条件异方差，则方差表达式不同，应使用“稳健标准误” (robust standard error)，参见第 5 章。

(4) “高斯-马尔可夫定理” (Gauss-Markov Theorem): 最小二乘法是最佳线性无偏估计 (Best Linear Unbiased Estimator, 简记 BLUE), 即在所有线性无偏估计中, 最小二乘法的方差最小。

证明: OLS 估计量 \mathbf{b} 为线性无偏估计。

假设 $\hat{\boldsymbol{\beta}}$ 为任一线性无偏估计, 需证明 $\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) \geq \text{Var}(\mathbf{b} | \mathbf{X})$, 即 $\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) - \text{Var}(\mathbf{b} | \mathbf{X})$ 为半正定矩阵。

$\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) - \text{Var}(\mathbf{b} | \mathbf{X})$ 半正定, 则 $\text{Var}(\mathbf{b} | \mathbf{X})$ 的主对角线元素 (即方差) 一定小于或等于 $\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X})$ 的主对角线上对应元素 (参见习题)。

由于 $\hat{\boldsymbol{\beta}}$ 为线性估计, 故存在常数矩阵 $\mathbf{C}_{K \times n}$, 使得 $\hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$ 。

已知 $\mathbf{b} = \mathbf{A}\mathbf{y}$ ，其中 $\mathbf{A} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 。定义 $\mathbf{D} \equiv \mathbf{C} - \mathbf{A}$ ，则

$$\hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y} = (\mathbf{D} + \mathbf{A})\mathbf{y} = \mathbf{D}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) + \mathbf{b} = \mathbf{DX}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{b}$$

利用 $\hat{\boldsymbol{\beta}}$ 的无偏性可得，

$$\boldsymbol{\beta} = \mathbf{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) = \mathbf{E}(\mathbf{DX}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{b} | \mathbf{X}) = \mathbf{DX}\boldsymbol{\beta} + \underbrace{\mathbf{D}\mathbf{E}(\boldsymbol{\varepsilon} | \mathbf{X})}_{=\mathbf{0}} + \underbrace{\mathbf{E}(\mathbf{b} | \mathbf{X})}_{=\boldsymbol{\beta}} = \mathbf{DX}\boldsymbol{\beta} + \boldsymbol{\beta}$$

对于任意 $\boldsymbol{\beta}$ ，都有 $\mathbf{DX}\boldsymbol{\beta} = \mathbf{0}$ ，故 $\mathbf{DX} = \mathbf{0}$ 。 $\hat{\boldsymbol{\beta}}$ 的表达式简化为

$$\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{DX}}_{=\mathbf{0}}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{b} = \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{b}$$

$\hat{\boldsymbol{\beta}}$ 的抽样误差为

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{b} - \boldsymbol{\beta} = \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{A}\boldsymbol{\varepsilon} = (\mathbf{D} + \mathbf{A})\boldsymbol{\varepsilon}$$

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) &= \text{Var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \mathbf{X}) = \text{Var}[(\mathbf{D} + \mathbf{A})\boldsymbol{\varepsilon} | \mathbf{X}] = (\mathbf{D} + \mathbf{A}) \text{Var}(\boldsymbol{\varepsilon} | \mathbf{X})(\mathbf{D} + \mathbf{A})' \\
&= \sigma^2 (\mathbf{D} + \mathbf{A})(\mathbf{D}' + \mathbf{A}') = \sigma^2 (\mathbf{D}\mathbf{D}' + \underbrace{\mathbf{A}\mathbf{D}'}_{=\mathbf{0}} + \underbrace{\mathbf{D}\mathbf{A}'}_{=\mathbf{0}} + \mathbf{A}\mathbf{A}') \\
&= \sigma^2 [\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}] \quad (\mathbf{D}\mathbf{A}' = \underbrace{\mathbf{D}\mathbf{X}}_{=\mathbf{0}}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0})
\end{aligned}$$

$$\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) - \text{Var}(\mathbf{b} | \mathbf{X}) = \sigma^2 [\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}] - \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \mathbf{D}\mathbf{D}'$$

由于 $\mathbf{D}\mathbf{D}'$ 为半正定矩阵，故高斯-马尔可夫定理成立。

如果没有球型扰动项的假定，则最小二乘法不是 BLUE，还存在其他更优的线性无偏估计，参见第 7 章。

(5)方差的无偏估计： $E(s^2 | \mathbf{X}) = \sigma^2$ 。

证明：因为

$$E(s^2 | \mathbf{X}) = E\left(\frac{\mathbf{e}'\mathbf{e}}{n-K} \mid \mathbf{X}\right) = E\left(\frac{\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}}{n-K} \mid \mathbf{X}\right) = \frac{1}{n-K} E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} | \mathbf{X})$$

故只要证明 $E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} | \mathbf{X}) = (n-K)\sigma^2$ ，即可。

定义 任意方阵 \mathbf{A} 的迹(trace)就是其主对角线元素之和，记为 $\text{trace}(\mathbf{A})$ 。

迹运算的性质：

$$\text{trace}(\mathbf{A} + \mathbf{B}) = \text{trace}(\mathbf{A}) + \text{trace}(\mathbf{B})$$

$\text{trace}(kA) = k \text{ trace}(A)$, k 为常数

$\text{trace}(AB) = \text{trace}(BA)$ 只要 AB 与 BA 都存在

如果 A 为 1×1 矩阵(常数), 则 $\text{trace}(A) = A$ 。

$$E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \mid X) = E[\text{trace}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \mid X)] \quad (\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \text{ 为 } 1 \times 1)$$

$$= E[\text{trace}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid X)] \quad (\boldsymbol{\varepsilon}' \text{ 与 } \mathbf{M}\boldsymbol{\varepsilon} \text{ 换次序})$$

$$= \text{trace}[E(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid X)] \quad (\text{期望算子与迹算子换次序})$$

$$= \text{trace}[\mathbf{M}\sigma^2 \mathbf{I}_n] \quad (\text{球型扰动项})$$

$$= \sigma^2 \text{trace}(\mathbf{M}) \quad (\text{迹运算的线性性})$$

$$\text{trace}(\mathbf{M}) = \text{trace}\left[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \quad (\text{消灭矩阵 } \mathbf{M} \text{ 的定义})$$

$$= \text{trace}(\mathbf{I}_n) - \text{trace}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \quad (\text{迹运算的线性性})$$

$$= n - \text{trace}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right] \quad (\mathbf{X} \text{ 与 } (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{ 互换})$$

$$= n - \text{trace}(\mathbf{I}_K) \quad (\mathbf{X}'\mathbf{X} \text{ 为 } K \times K \text{ 矩阵})$$

$$= n - K \quad (K \text{ 阶单位阵的迹为 } K)$$

对协方差阵 $\text{Var}(\mathbf{b} | \mathbf{X})$ 的无偏估计为 $s^2(\mathbf{X}'\mathbf{X})^{-1}$ ，在 Stata 中记为“VCE” (Variance-Covariance Matrix Estimated)。

3.6 对单个系数的 t 检验

假定 3.5 在给定 X 的情况下, $\boldsymbol{\varepsilon}|X$ 的条件分布为正态, 即 $\boldsymbol{\varepsilon}|X \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ 。

考虑对单个系数进行检验, “原假设” (null hypothesis, 零假设) 为 $H_0: \beta_k = \bar{\beta}_k$, 其中 $\bar{\beta}_k$ 为给定常数。

通常 $\bar{\beta}_k = 0$, 检验变量 x_{ik} 的系数是否显著地不等于 0。

假设检验是概率意义上的反证法, 即首先假设原假设成立, 然后看在原假设成立的前提下, 是否导致不太可能发生的“小概率事件”在一次抽样的样本中出现。

如果小概率事件竟在一次抽样中被观测到，则说明原假设不可信，应该拒绝原假设，接受“替代假设”(alternative hypothesis, 备择假设) $H_1: \beta_k \neq \bar{\beta}_k$ 。

如果未知参数 β_k 的估计值 b_k 离 $\bar{\beta}_k$ 较远，则倾向于拒绝原假设。这类检验称为“沃尔德检验”(Wald test)。

在衡量距离远近时，绝对距离依赖变量单位，需以标准差为基准来考虑相对距离。

由于 $\boldsymbol{\varepsilon} | \boldsymbol{X} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ ，而 $\boldsymbol{b} - \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\varepsilon}$ 为 $\boldsymbol{\varepsilon}$ 的线性函数，故 $(\boldsymbol{b} - \boldsymbol{\beta}) | \boldsymbol{X}$ 服从正态分布。

进 一 步 ， $E(\boldsymbol{b} - \boldsymbol{\beta} | \boldsymbol{X}) = \mathbf{0}$ ， $\text{Var}(\boldsymbol{b} | \boldsymbol{X}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$ ， 故

$$(\mathbf{b} - \boldsymbol{\beta}) | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}).$$

在原假设 “ $H_0: \beta_k = \bar{\beta}_k$ ” 成立的情况下，其第 k 个分量 $(b_k - \bar{\beta}_k) | \mathbf{X} \sim N(0, \sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1})$ ，其中 $(\mathbf{X}'\mathbf{X})_{kk}^{-1}$ 为矩阵 $(\mathbf{X}'\mathbf{X})^{-1}$ 的 (k, k) 元素，而 $\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}$ 为 b_k 的方差。

$$\text{如果 } \sigma^2 \text{ 已知，则统计量 } z_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}} \sim N(0, 1)。$$

通常 σ^2 未知，称为“厌恶参数” (nuisance parameter): 虽然对 σ^2 不感兴趣，但 σ^2 却出现在表达式中。

合格的“检验统计量” (test statistic) 须满足两个条件：能够根据样本数据计算；概率分布已知。

以估计值 s^2 来替代 σ^2 ，可得 t 统计量。

定理(t 统计量的分布) 在假定 3.1-3.5 均满足，且原假设“ $H_0 : \beta_k = \bar{\beta}_k$ ”也成立的情况下， t 统计量

$$t_k \equiv \frac{b_k - \bar{\beta}_k}{\text{SE}(b_k)} \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}} \sim t(n - K)$$

其中， $\text{SE}(b_k)$ 是 b_k 的“估计标准误差”(estimated standard error)，简称“标准误”。

证明：将统计量 t_k 变形：

$$t_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}} = \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}} \cdot \sqrt{\frac{\sigma^2}{s^2}} = \frac{z_k}{\sqrt{s^2/\sigma^2}} = \frac{z_k}{\sqrt{\frac{\mathbf{e}'\mathbf{e}}{(n-K)\sigma^2}}} \equiv \frac{z_k}{\sqrt{q/(n-K)}}$$

其中, $z_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}}$, $q \equiv \frac{\mathbf{e}'\mathbf{e}}{\sigma^2}$ 。

已知 $z_k \sim N(0,1)$, 下面将证明,

(1) $q|\mathbf{X} \sim \chi^2(n-K)$;

(2) $z_k|\mathbf{X}$ 与 $q|\mathbf{X}$ 相互独立,

则根据 t 分布的定义, $\frac{z_k}{\sqrt{q/(n-K)}} \sim t(n-K)$ 。

$$(1) \quad q \equiv \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'}{\sigma} \mathbf{M} \frac{\boldsymbol{\varepsilon}}{\sigma} \quad (\text{二次型}).$$

由于 $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, 故 $\frac{\boldsymbol{\varepsilon}}{\sigma}|\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n)$ 。已知消灭矩阵 \mathbf{M} 为“幂等矩阵” (idempotent matrix, 即 $\mathbf{M}^2 = \mathbf{M}$)。

根据线性代数知识, 对于幂等矩阵 \mathbf{M} , $\text{rank}(\mathbf{M}) = \text{trace}(\mathbf{M}) = n - K$ 。

根据数理统计知识, $q|\mathbf{X} \sim \chi^2(n - K)$ 。由于 \mathbf{M} 不满秩, $q|\mathbf{X}$ 的自由度降为 $(n - K)$ 。

(2) $z_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 (X'X)^{-1}_{kk}}}$ 是 \mathbf{b} 的函数, q 是 \mathbf{e} 的函数, 故只要证明 \mathbf{b} 与 \mathbf{e} 相互独立即可。

由于 $\mathbf{b} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}$ 与 $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$ 都是正态扰动项 $\boldsymbol{\varepsilon}$ 的线性函数, 故 (\mathbf{b}, \mathbf{e}) 的联合分布也是正态, 故只要证明 $\text{Cov}(\mathbf{b}, \mathbf{e}) = \mathbf{0}$ 即可。

$$\text{Cov}(\mathbf{b}, \mathbf{e} \mid \mathbf{X}) = \text{Cov}(\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}, \mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) \quad (\text{代入 } \mathbf{b} \text{ 与 } \mathbf{e} \text{ 表达式})$$

$$= \text{Cov}(\mathbf{A}\boldsymbol{\varepsilon}, \mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) \quad (\text{去掉常数 } \boldsymbol{\beta})$$

$$= \text{E}(\mathbf{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}) - \underbrace{\text{E}(\mathbf{A}\boldsymbol{\varepsilon})}_{=\mathbf{0}} \underbrace{\text{E}(\mathbf{M}\boldsymbol{\varepsilon})'}_{=\mathbf{0}'} \quad (\text{协方差矩阵公式})$$

$$= \mathbf{A}\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{M} = \sigma^2 \mathbf{A}\mathbf{M} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{M} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \underbrace{(\mathbf{M}\mathbf{X})'}_{=\mathbf{0}} = \mathbf{0}$$

(OLS 的正交性)

t 检验的步骤

第一步：计算 t_k 。如果 $|t_k|$ 很大，则 H_0 较不可信。如果 H_0 为真，则 $|t_k|$ 很大的概率将很小(为小概率事件)，不应在抽样中观测到。

第二步：计算“显著性水平”(significance level)为 α 的“临界值”(critical value) $t_{\alpha/2}(n-K)$

$$P\{t(n-K) > t_{\alpha/2}(n-K)\} = P\{t(n-K) < -t_{\alpha/2}(n-K)\} = \alpha/2$$

其中， $t(n-K)$ 服从 $t(n-K)$ 分布。 $t(n-K)$ 大于 $t_{\alpha/2}(n-K)$ ，或小于 $-t_{\alpha/2}(n-K)$ 的概率都是 $\alpha/2$ 。通常 $\alpha = 5\%$ ，则 $\alpha/2 = 2.5\%$ 。有时 $\alpha = 1\%$ 或 10% 。

第三步：如果 t_k 落入“拒绝域”(rejection region)，则拒绝 H_0 ； t_k 落入“接受域”(acceptance region)，则接受 H_0 。

拒绝域分布在 t 分布两边，称为“双边检验”(two-tailed)。

计算 p 值

定义 给定检验统计量的样本观测值，称原假设可被拒绝的最小显著性水平为此假设检验问题的 p 值(probability value，即 p -value)。

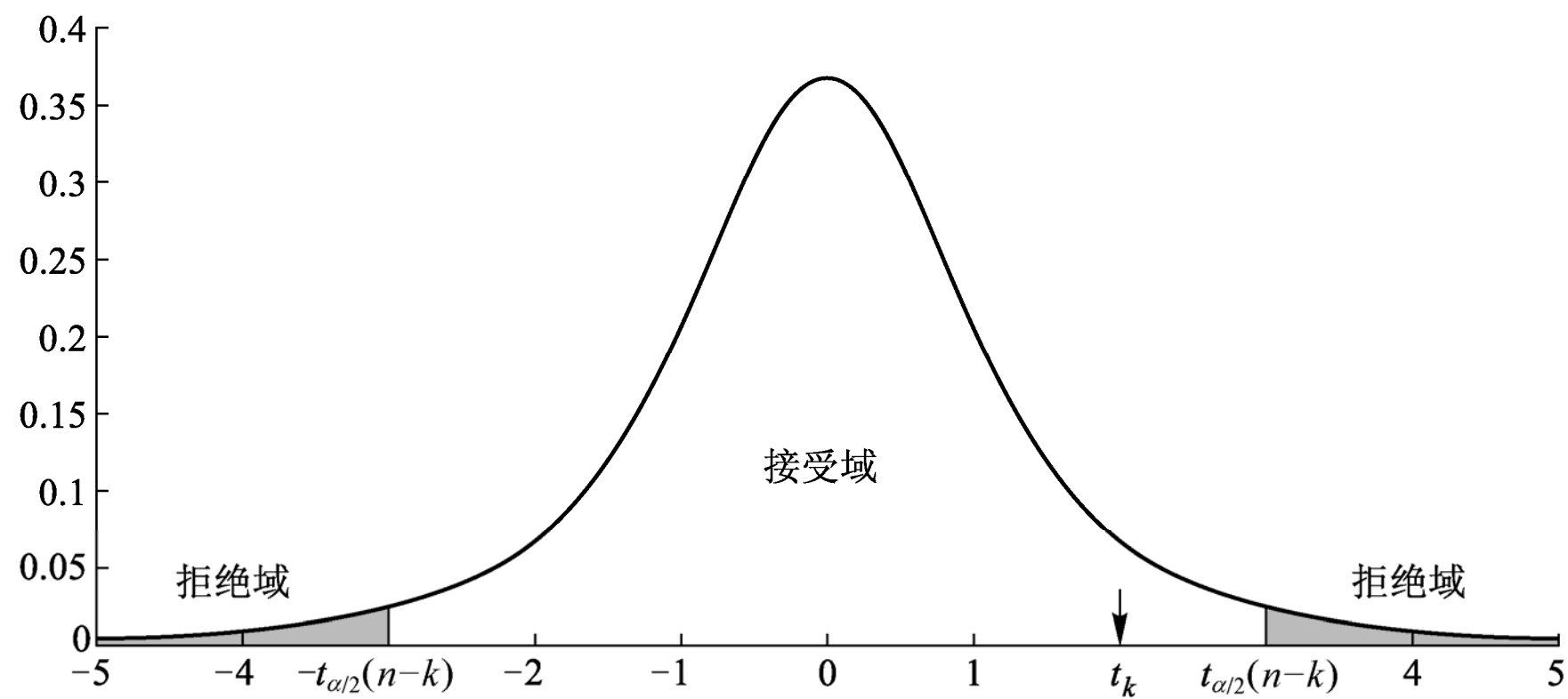


图 3.3 t 检验

在 t 检验中, p 值为 $P(t > |t_k|) \times 2$, 其中 t_k 为检验统计量的样本观测值。

p 值越小则越倾向于拒绝原假设。

【例】 p 值 = 0.03, 则可在 5% 的显著性水平上拒绝原假设。而且, “ p 值 = 0.03” 还可在 3% 的显著性水平上拒绝原假设。

使用 p 值进行假设检验一般比临界值更有信息量。

当 Stata 直接给出 p 值时, 就不需要知道临界值了。

计算置信区间

假设“置信度”(confidence level)为 $(1-\alpha)$ (比如 $\alpha=5\%$, 则 $1-\alpha=95\%$), 要找到“置信区间”(confidence interval), 使得该区间覆盖真实参数 β_k 的概率为 $(1-\alpha)$ 。

由于 $\frac{b_k - \beta_k}{\text{SE}(b_k)} \sim t(n-K)$, 故

$$P\left\{-t_{\alpha/2} < \frac{b_k - \beta_k}{\text{SE}(b_k)} < t_{\alpha/2}\right\} = 1 - \alpha \quad (t_{\alpha/2} \text{ 的定义})$$

$$P\{b_k - t_{\alpha/2} \text{SE}(b_k) < \beta_k < b_k + t_{\alpha/2} \text{SE}(b_k)\} = 1 - \alpha \quad (\text{不等式变形})$$

即置信区间为 $[b_k - t_{\alpha/2} \text{SE}(b_k), b_k + t_{\alpha/2} \text{SE}(b_k)]$ 。

置信区间是随机区间，随着样本不同而不同。

如果置信度为 95%，抽样 100 次，得到 100 个置信区间，大约 95 个置信区间能覆盖到真实参数 β_k 。

第 I 类错误与第 II 类错误

定义 “第 I 类错误” (Type I error)指的是，虽然原假设为真，但却根据观测数据做出了拒绝原假设的错误判断，即“弃真”。第 I 类错误的发生概率为 $P(\text{reject } H_0 | H_0)$ 。

定义 “第 II 类错误”(Type II error)指的是，虽然原假设为假(替代假设为真)，但却根据观测数据做出了接受原假设的错误判断，即“存伪”。第 II 类错误的发生概率为 $P(\text{accept } H_0 | H_1)$ 。

除非增加样本容量，减少第 I 类错误的发生概率，必然导致第 II 类错误的发生概率增加，反之亦然。

在进行检验时，一般先指定可接受的发生第 I 类错误的最大概率，即“显著性水平”，比如 5%；而不指定第 II 类错误的发生概率(通常更难计算)。

定义 称“1 减去第 II 类错误的发生概率”为统计检验的“功效”或“势”(power)，即“ $1 - P(\text{accept } H_0 | H_1)$ ”。换言之，功效为在原假设为错误的情况下，拒绝原假设的概率。

进行检验时，通常知道第 I 类错误的发生概率，不知道第 II 类错误的发生概率。

拒绝原假设，比较理直气壮，因为知道犯错概率（显著性水平）。

接受原假设，比较没有把握，通常不知犯错概率（可能较高）。

3.7 对线性假设的 F 检验

检验回归方程的显著性，即检验原假设 “ $H_0: \beta_2 = \cdots = \beta_K = 0$ ” (β_1 为常数项)。

更一般地，检验 m 个线性假设是否同时成立：

$$H_0: \underbrace{\mathbf{R}}_{m \times K} \underbrace{\boldsymbol{\beta}}_{K \times 1} = \underbrace{\mathbf{r}}_{m \times 1}$$

其中， \mathbf{r} 为 m 维列向量， \mathbf{R} 为 $m \times K$ 矩阵， $\text{rank}(\mathbf{R}) = m$ ，即 \mathbf{R} 满行秩，没有多余或自相矛盾的方程。

例 对于模型 $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ ，检验“ $H_0: \beta_2 = \beta_3$ 且 $\beta_4 = 0$ ”。

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{因为}$$

$$\begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_2 - \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Wald 检验原理： \mathbf{b} 是 $\boldsymbol{\beta}$ 的估计量，如果 H_0 成立，则 $(\mathbf{Rb} - \mathbf{r})$ 应比较接近 $\mathbf{0}$ 。

定理(F 统计量的分布) 在假定 3.1-3.5 均满足，且原假设“ $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ ”也成立的情况下，则 F 统计量

$$F \equiv \frac{(\mathbf{Rb} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{r}) / m}{s^2} \sim F(m, n - K)$$

证明：由于 $s^2 = \mathbf{e}'\mathbf{e}/(n - K)$ ，可将 F 写成

$$F \equiv \frac{(\mathbf{Rb} - \mathbf{r})' [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{r}) / m}{(\mathbf{e}'\mathbf{e} / \sigma^2) / (n - K)} \equiv \frac{w/m}{q/(n - K)}$$

其中, $w \equiv (\mathbf{R}\mathbf{b} - \mathbf{r})' [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{r})$ 。

下面将证明:

(1) $w|\mathbf{X} \sim \chi^2(m)$;

(2) $q|\mathbf{X} \sim \chi^2(n-K)$; (已在 t 检验定理中证明)

(3) $w|\mathbf{X}$ 与 $q|\mathbf{X}$ 相互独立,

则根据 F 分布的定义, $\frac{w/m}{q/(n-K)} \sim F(m, n-K)$ 。

(1) 定义 $\boldsymbol{v} \equiv \boldsymbol{R}\boldsymbol{b} - \boldsymbol{r}$ 。在 H_0 成立的情况下,

$$\boldsymbol{v} \equiv \boldsymbol{R}\boldsymbol{b} - \boldsymbol{r} = \boldsymbol{R}\boldsymbol{b} - \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{R}(\boldsymbol{b} - \boldsymbol{\beta})$$

由于 \boldsymbol{b} 为正态, 故 $\boldsymbol{v} | \boldsymbol{X}$ 为 m 维正态, 且 $E(\boldsymbol{v} | \boldsymbol{X}) = \mathbf{0}$, 方差为

$$\text{Var}(\boldsymbol{v} | \boldsymbol{X}) = \text{Var}[\boldsymbol{R}(\boldsymbol{b} - \boldsymbol{\beta}) | \boldsymbol{X}] = \boldsymbol{R}\text{Var}(\boldsymbol{b})\boldsymbol{R}' = \sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{R}'$$

根据数理统计知识,

$$w \equiv (\boldsymbol{R}\boldsymbol{b} - \boldsymbol{r})' [\sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{R}']^{-1} (\boldsymbol{R}\boldsymbol{b} - \boldsymbol{r}) = \boldsymbol{v}' [\text{Var}(\boldsymbol{v} | \boldsymbol{X})]^{-1} \boldsymbol{v} \sim \chi^2(m)$$

由于 \boldsymbol{R} 满行秩, 故 $[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{R}']^{-1}$ 存在。

(3) w 是 \mathbf{b} 的函数, q 是 \mathbf{e} 的函数, 由于 \mathbf{b} 与 \mathbf{e} 相互独立, 故 $w|\mathbf{X}$ 与 $q|\mathbf{X}$ 相互独立。

F 检验的步骤

第一步: 计算 F 统计量。如果 F 统计量很大, 则 H_0 较不可信。

第二步: 计算显著性水平为 α 的临界值 $F_\alpha(m, n-K)$,

$$P\{F(m, n-K) > F_\alpha(m, n-K)\} = \alpha$$

其中, $F(m, n-K)$ 服从 $F(m, n-K)$ 分布。

第三步: 如果 F 统计量落入右边拒绝域, 则拒绝 H_0 ; F 统计量

落入接受域，则接受 H_0 。拒绝域只在右侧，为“单边右侧检验”。

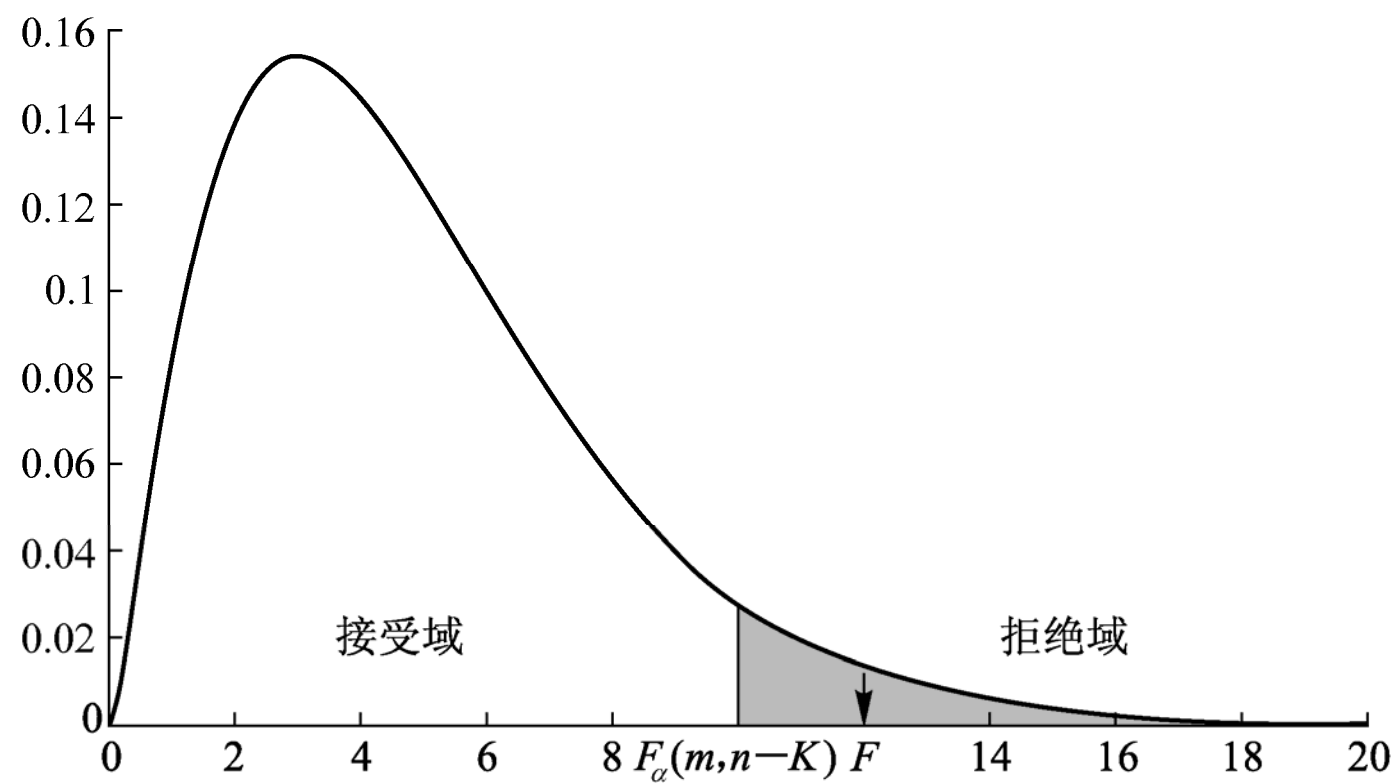


图 3.4 F 检验

3.8 F 统计量的似然比原理表达式

使用约束条件下的最小二乘法，即 “约束最小二乘法” (Restricted OLS, Constrained OLS), 可得 F 统计量的简便表达式。

考虑以下约束极值问题：

$$\begin{aligned} \min_{\tilde{\beta}} \quad & \text{SSR}(\tilde{\beta}) \\ \text{s.t.} \quad & \mathbf{R}\tilde{\beta} = \mathbf{r} \end{aligned}$$

如果 “ $H_0 : \mathbf{R}\beta = \mathbf{r}$ ” 正确，则加上此约束不应使残差平方和 $\text{SSR}(\tilde{\beta})$ 的最小值增大很多。

求解此约束极值问题，可证明：

$$F = \frac{(\mathbf{e}^{*'} \mathbf{e}^* - \mathbf{e}' \mathbf{e})/m}{\mathbf{e}' \mathbf{e}/(n-K)}$$

其中， \mathbf{e} 为无约束残差， \mathbf{e}^* 为有约束残差， m 为约束条件个数。

这种通过比较“条件极值”与“无条件极值”而进行的检验，统称为“似然比检验”(Likelihood ratio test)。

命题 对于线性回归方程“ $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$ ”，检验原假设“ $H_0: \beta_2 = \cdots = \beta_K = 0$ ”(即该方程的显著性)的 F 统计量等于 $\frac{R^2/(K-1)}{(1-R^2)/(n-K)}$ 。

证明：由于共有 $(K-1)$ 个约束，根据似然比原理的 F 统计量为

$$F = \frac{(\mathbf{e}^{*'} \mathbf{e}^* - \mathbf{e}' \mathbf{e}) / (K-1)}{\mathbf{e}' \mathbf{e} / (n-K)} = \frac{\frac{(\mathbf{e}^{*'} \mathbf{e}^* - \mathbf{e}' \mathbf{e})}{\sum_{i=1}^n (y_i - \bar{y})^2} / (K-1)}{\frac{\mathbf{e}' \mathbf{e}}{\sum_{i=1}^n (y_i - \bar{y})^2} / (n-K)}$$

其中， \mathbf{e} 为无约束残差，而 \mathbf{e}^* 为约束残差。记约束回归的拟合优度为 R_*^2 ，由于 $\frac{\mathbf{e}' \mathbf{e}}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - R^2$ ， $\frac{\mathbf{e}^{*'} \mathbf{e}^*}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - R_*^2$ ，故

$$F = \frac{[(1 - R_*^2) - (1 - R^2)] / (K-1)}{(1 - R^2) / (n-K)} = \frac{(R^2 - R_*^2) / (K-1)}{(1 - R^2) / (n-K)}$$

只需证明 $R_*^2 = 0$ 即可。

当 “ $H_0 : \beta_2 = \cdots = \beta_K = 0$ ” 成立时, $y_i = \beta_1 + \varepsilon_i$, 而 $b_1^* = \bar{y}$ (只对常数项 β_1 进行回归), 故 $\hat{y}_i^* = b_1^* = \bar{y}$ 。

由此可知 $\sum_{i=1}^n (\hat{y}_i^* - \bar{y})^2 = 0$, 故 $R_*^2 = 0$ 。

3.9 分块回归与偏回归 (选读)

Partitioned regression

- We are going to discuss **partitioned regression** now.
- Partitioned regression is useful when we have many x variables in our model but are only interested in the coefficients for some of them.
- It also permits us to simplify theoretical results in some cases (for instance, if we want a formula for only the constant term in a regression, and not the whole parameter vector).

- First partition the data matrix in the following way:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}, \quad (81)$$

where \mathbf{X} is the original $n \times K$ data matrix, and \mathbf{X}_1 and \mathbf{X}_2 are $n \times K_1$ and $n \times K_2$ data matrices, respectively.

- Correspondingly, β is partitioned as

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad (82)$$

- We are interested in the parameters of β_2 .
- The regression model can be written as

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon \quad (83)$$

- Define

$$P_1 \equiv X_1(X_1'X_1)^{-1}X_1' \quad (84)$$

$$M_1 \equiv I_n - P_1 \quad (85)$$

$$\tilde{X}_2 \equiv M_1X_2 \quad (86)$$

(the k th column of \tilde{X}_2 represents the vector of residuals when the k th column of X_2 is regressed on X_1)

$$\tilde{y} \equiv M_1y \quad (87)$$

(the residuals from a regression of y on X_1)

- The normal equations:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y} \quad (88)$$

$$\begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \end{bmatrix} \mathbf{y} \quad (89)$$

$$\begin{bmatrix} \mathbf{X}'_1\mathbf{X}_1 & \mathbf{X}'_1\mathbf{X}_2 \\ \mathbf{X}'_2\mathbf{X}_1 & \mathbf{X}'_2\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1\mathbf{y} \\ \mathbf{X}'_2\mathbf{y} \end{bmatrix} \quad (90)$$

- Equation (90) can be written as

$$\mathbf{X}'_1\mathbf{X}_1\mathbf{b}_1 + \mathbf{X}'_1\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}'_1\mathbf{y} \quad (91)$$

$$\mathbf{X}'_2\mathbf{X}_1\mathbf{b}_1 + \mathbf{X}'_2\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}'_2\mathbf{y} \quad (92)$$

- Our goal is to solve for \mathbf{b}_2 .

- Premultiply equation (91) by $\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ and solve for $\mathbf{X}_1\mathbf{b}_1$, we have

$$\mathbf{X}_1\mathbf{b}_1 = -\mathbf{P}_1\mathbf{X}_2\mathbf{b}_2 + \mathbf{P}_1\mathbf{y} \quad (93)$$

- Substitute this expression into (92):

$$\mathbf{X}_2'(-\mathbf{P}_1\mathbf{X}_2\mathbf{b}_2 + \mathbf{P}_1\mathbf{y}) + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{y} \quad (94)$$

$$\mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 - \mathbf{X}_2'\mathbf{P}_1\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{y} - \mathbf{X}_2'\mathbf{P}_1\mathbf{y} \quad (95)$$

$$\mathbf{X}_2'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{y} \quad (96)$$

$$\mathbf{X}_2'(\mathbf{M}_1)\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'(\mathbf{M}_1)\mathbf{y} \quad (97)$$

$$\mathbf{X}_2'(\mathbf{M}_1'\mathbf{M}_1)\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'(\mathbf{M}_1'\mathbf{M}_1)\mathbf{y} \quad (98)$$

(since \mathbf{M}_1 is symmetric and idempotent)

$$\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2\mathbf{b}_2 = \tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} \quad (99)$$

- Therefore,

$$\mathbf{b}_2 = (\tilde{\mathbf{X}}_2' \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2' \tilde{\mathbf{y}} \quad (100)$$

- $\tilde{\mathbf{X}}_2' \tilde{\mathbf{X}}_2 = \mathbf{X}_2' \mathbf{M}_1' \mathbf{M}_1 \mathbf{X}_2$ is invertible by the full rank assumption.
 - Essentially, if \mathbf{X}_2 (and, by extension, $\mathbf{M}_1 \mathbf{X}_2$, which is the component of \mathbf{X}_2 orthogonal to \mathbf{X}_1) does not have full column rank, \mathbf{X} cannot have full column rank either.
- This solution for \mathbf{b}_2 is an important result known as the **Frisch-Waugh Theorem**.
- It states that you can obtain the coefficients of \mathbf{b}_2 in this way:
 - 1 Regress y on the variables in \mathbf{X}_1 , and obtain the residuals.
 - 2 Regress each of the variables in \mathbf{X}_2 on all of the variables in \mathbf{X}_1 , and obtain the residuals.
 - 3 Regress the residuals in (1) on the residuals in (2).

- Conceptually, this process of regressing y and the variables in X_2 on X_1 , then taking the residuals is known as **partialling out** or **netting out** the effect of X_1 .
- For this reason, the coefficients in a multiple regression are sometimes called **partial regression coefficients**.
- As a corollary, when X_2 and X_1 are orthogonal (so that $M_1 X_2 = X_2$), no partialling out is necessary – you can just regress y on X_2 directly and the variables in X_1 will not make any difference at all.

3.10 预 测

有时也用计量模型进行预测(prediction, forecasting), 即给定解释向量 \mathbf{x}_0 的(未来)取值, 预测被解释变量 y_0 的取值。

假设计量模型对所有观测值都成立(包括外推到未来的观测值),

$$y_0 = \mathbf{x}_0' \boldsymbol{\beta} + \varepsilon_0$$

用 $\hat{y}_0 \equiv \mathbf{x}_0' \mathbf{b}$ 对 y_0 作点预测, \mathbf{b} 为 $\boldsymbol{\beta}$ 的 OLS 估计量。

“预测误差” (prediction error) $(\hat{y}_0 - y_0)$ 可写为

$$\hat{y}_0 - y_0 = \mathbf{x}_0' \mathbf{b} - \mathbf{x}_0' \boldsymbol{\beta} - \varepsilon_0 = \mathbf{x}_0' (\mathbf{b} - \boldsymbol{\beta}) - \varepsilon_0$$

由于 \mathbf{b} 是 $\boldsymbol{\beta}$ 的无偏估计，故 $E(\hat{y}_0 - y_0) = \mathbf{x}'_0 E(\mathbf{b} - \boldsymbol{\beta}) - E(\varepsilon_0) = 0$ 。

“无偏预测” (unbiased predictor): 用 \hat{y}_0 来预测 y_0 不会系统高估或低估。

预测 \hat{y}_0 的方差为:

$$\text{Var}(\hat{y}_0) = \text{Var}(\mathbf{x}'_0 \mathbf{b}) = \mathbf{x}'_0 \text{Var}(\mathbf{b}) \mathbf{x}_0 = \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0$$

此方差反映由于抽样误差 $(\mathbf{b} - \boldsymbol{\beta})$ 所带来的预测量 \hat{y}_0 的波动。如果知道 $\boldsymbol{\beta}$ ，则 $\text{Var}(\hat{y}_0) = \text{Var}(\mathbf{x}'_0 \boldsymbol{\beta}) = 0$ 。

预测误差($\hat{y}_0 - y_0$)的方差为:

$$\begin{aligned}\text{Var}(\hat{y}_0 - y_0) &= \text{Var}[\mathbf{x}'_0(\mathbf{b} - \boldsymbol{\beta}) - \varepsilon_0] \\ &= \text{Var}(\varepsilon_0) + \text{Var}[\mathbf{x}'_0(\mathbf{b} - \boldsymbol{\beta})] \\ &= \sigma^2 + \sigma^2 \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0\end{aligned}$$

其中, 假设 ε_0 与 \mathbf{b} 不相关(估计 \mathbf{b} 没用到 ε_0 的信息), 故 $\text{Cov}[\mathbf{x}'_0(\mathbf{b} - \boldsymbol{\beta}), \varepsilon_0] = 0$ 。

预测误差的方差有两个来源, 即抽样误差 $\sigma^2 \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0$ (不能精确知道参数 $\boldsymbol{\beta}$), 以及 y_0 本身的不确定性(ε_0 的方差 σ^2)。

假设扰动项为正态，则 $\hat{y}_0 - y_0 \sim N\left(0, \sigma^2 + \sigma^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)$ 。

用 s^2 估计 σ^2 ，得到 t 统计量：

$$\frac{\hat{y}_0 - y_0}{s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \sim t(n - K)$$

y_0 的置信度为 $(1 - \alpha)$ 的置信区间为

$$\left(\hat{y}_0 - t_{\alpha/2} s \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}, \hat{y}_0 + t_{\alpha/2} s \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \right)$$