计量经济学

Ender

2024 年 1 月 20 日

目录

回归模型

$Y = E(Y|X) + e$

↑ CEF　↑ 回归误差/CEF误差

$E(e|X) = 0$　$Ee = 0$

$E(Y|X)$是未知的

$E(Y|X)$为线性函数 →

线性CEF模型

$Y = X^T\beta + e$

↑ 线性CEF

$E(e|X) = 0$

线性CEF依然未知，β未知

依据MSE最小得到

$\beta^* = (E(X^TX))^{-1}E(XY)$

线性投影模型

$Y = X^T\beta^* + \varepsilon$ ← 也称为回归模型

↑ 投影误差

性质：$E(X\varepsilon) = 0$　若有常数项 $E\varepsilon = 0$

Best Line Predictor $P(Y|X) = X^T\beta^*$

总体

样本

给定 Random Sample

$\{(Y_1, \vec{X_1}), \dots, (Y_n, \vec{X_n})\}$ i.i.d

OLS估计量 $\hat{\beta}_{OLS} = \underset{\vec{\beta} \in \mathbb{R}^k}{\operatorname{argmin}}\, \hat{S}(\vec{\beta})$

$\hat{S}(\vec{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \vec{X_i}^T\vec{\beta})^2$ 为MSE的矩估计

第一章　计量经济学诸论

第二章　一元线性回归模型

回归模型源自于数理统计, 其被用于研究具有相关关系的变量, 对于计量经济学而言, 回归分析则是进一步探讨理论上具有因果关系的变量间的关系 (由因果性而具有相关性). 本章将介绍最为简单的一元线性回归模型, 之后的章节则在理解了一元回归模型后, 将进一步拓展至多元的情形.

依据概率论与数理统计的范式, 本章以及之所研究的感兴趣的变量, 均是**随机变量**, 相应地本笔记使用**斜体大写字母** (例如 X, Y, M, D) 或**首字母大小的英文** (例如 $Wage, Post, Treat$) 表示[1], 而对于多个变量 $X_1, \cdots, X_n$, 本笔记使用**非斜体粗体大写字母**记有随机向量 $\mathbf{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$. 特别地, 对于经过函数处理后的变量, 则采用函数名 + 变量的方式表示 (例如 $\ln X$ 或对数工资 $\ln Wage$). 这样表示的变量, 对应了研究问题的总体. 对于研究变量所对应的样本, 本笔记使用形如 $\boldsymbol{X}, \boldsymbol{Y}$ 的矩阵符号[2]记录全部样本的随机变量, 单个样本的随机变量记为 Y_i, X_i 的形式, 而具体的观测值则使用形如 $y_1, y_2, \cdots, y_n$ 的**斜体小写字母**来表示.

计量经济学遵循了数理统计的方法, 是利用样本去出推断总体, 因而就本笔记所涉及的样本, 不作特别说明时均假定样本是通过**概率抽样**取得的, 进一步地则是假定为简单随机抽样的样本, 故本笔记所探讨的样本个体都是**独立同分布**的. 具体地, 对于样本容量为 K 个的样本, 在仅有一个变量 X 的情况下则记第 k 个样本) 个体 ($k = 1, 2, \cdots, K$) 为随机变量 X_k, 其抽样的观测值为 x_k.

明确了符号的使用可以减少后续笔记中出现误解的情况. 由于所研究的变量均是随机变量, 本笔记也采用概率论的符号来表示随机变量的数值特征, 利用对于变量工资 $Wage$ 和受教育年限 $School$, 则

$$\mathbb{E}\left(Wage|School\right)$$

意味着给定受教育年限 $School$ 下工资 $Wage$ 的条件期望.

---

[1]在 LaTeX 的数学环境中, 通过斜体命令`\textit{}`输入.

[2]在 LaTeX 的数学环境中, 命令`\mathbf{X}`输出 $\mathbf{X}$, 而命令`\boldsymbol{X}`输出 $\boldsymbol{X}$.

2.1 回归函数

倘若在学习概率论前, 读者便已有接触过回归分析, 则很可能错误的将线性回归模型 $Y = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$ 等同于“回归函数”. 本节内容即是要纠正这一错误并介绍概率论中回归函数的概念.

随机变量是对不确定性的一种刻画方式[3], 只有试验发生了, 才能知道结果得到随机变量的观测值. 对于随机变量具体的取值, 可以用概率去进行精细的刻画, 然而在实际应用中即是知道随机变量取值的概率了, 多数情况下便是依据大数定律去理解 (例如重复掷塞子 100 次, 大约有 50 次朝上的点数是奇数), 仍旧不是一个“确切”的结果——只有试验发生了, 我们才能够观测到结果.

幸运的是. 概率论告诉了我们随机变量的数值特征, 例如: 期望、方差、高阶矩等. 只要知道了随机变量的概率分布[4], 通常而言我们是可以考虑其完全“确定”的数值特征[5], 而且数值特征亦能够帮助我们了解随机变量所具有的性质. 例如知道了随机变量的的期望 μ 和标准差 σ, 依据 Markov 不等式, 便能得到随机变量最可能出现的取值情况 (尽管这个例子还是带有不确定性, 但还是帮助我们理解了数字特征的“确定性”作用).

现考虑两个连续性随机变量 X 和 Y, 要研究二者具有的关系, 用概率论的方法则是考虑联合分布 (X, Y) 的联合密度函数 $f_{XY}(x, y)$, 这样 X 和 Y 的概率密度函数 (亦边缘密度函数) 为

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y)\,\mathrm{d}y, \quad f_Y(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y)\,\mathrm{d}x,$$

但这只包含了单独的 X 或 Y 的信息. 一种能够得知 X 和 Y 间关系的方式, 即是依据条件概率的思想, **计算条件**期望 $\mathbb{E}(Y|X)$, 用计量经济学的术语, 则称为**回归函数**.

定义 2.1.1 (条件期望函数, 总体回归函数)**.** *设随机变量 X 和 Y, 则称 $\mathbb{E}(Y|X) = f(X)$ 为 Y 关于 X 的*__条件期望函数__ *(conditional expectation function). 在计量经济学中, 则称 $\mathbb{E}(Y|X)$ 为*__总体回归函数__ *(population regression function), 简称*__回归函数__.

[3]在经济学中, 文献 (cite) 将风险 (risk) 定义为有概率分布的随机变量, 而不确定性 (uncertainty) 则是不具有概率分布的随机变量. 但进来的研究也未有遵循这个定义 (cite).

[4]一般而言, 概率论中具有概率密度函数 (probability density function,PDF) 的随机变量称为连续性随机变量, 然而更多变量不是连续性随机变量, 亦不属于离散型随机变量 (例如随机变量 $X \sim P(\lambda)$, 随机变量 Y 服从标准正态分布, 那么 $X + Y$ 既不是连续性随机变量, 又不是离散型随机变量.) 本笔记为了方便讨论, 所研究的变量均为连续性随机变量, 即其具有 PDF.

[5]具有 PDF 的连续性随机变量, 其数值特征并非一定存在. 例如服从 Cauchy 分布的随机变量不存在数学期望 (广义积分是发散的). 为了便于讨论, 计量经济学所研究的变量通常都是假定其数字特征是存在的.

对于给定的 $X=x$, 条件期望函数 $\mathbb{E}(Y|X=x)$ 利用了 $X=x$ 提供的信息, 是在此之下随机变量 Y 所具有的期望, 求期望后则消去 Y 的不确定性, 但会保留有 $X=x$, 因而条件期望函数是关于随机变量 X 的观测值 x 的函数. 假设随机变量 X,Y 均为连续性随机变量且存在联合密度函数 $f_{XY}(x,y)$, 那么 Y 关于 X 的条件期望函数为

$$\mathbb{E}(Y|X)=\int_{-\infty}^{+\infty} y f_{Y|X}(y|x)\,\mathrm{d}y=\int_{-\infty}^{+\infty} y\frac{f_{XY}(x,y)}{f_X(x)}\mathrm{d}y,$$

其中

$$f_{Y|X}(y|x)=\frac{f_{XY}(x,y)}{f_X(x)}$$

为 Y 关于 X 的条件密度函数.

在没有歧义的情形下, 本笔记将条件期望函数 $\mathbb{E}(Y|X=x)$ 记为 $m(x)$.

为什么计量经济学要以条件期望函数作为回归函数? 在解释这个问题前, 我们首先需要回顾一下条件期望的相关性质.

定理 2.1.1 (简单迭代期望定律,simple law of iterated expectations)**.** *设随机变量 X 和 Y, 若 $\mathbb{E}Y<\infty$, 则有 $\mathbb{E}(\mathbb{E}(Y|X))=\mathbb{E}Y$.*

定理 (2.1.1)[6]给出了非条件期望一种计算方法, 即是条件期望的期望等于非条件期望. 具体地, 将期望算符 $\mathbb{E}$ 所对应的随机向量标准, 则有 $\boxed{\mathbb{E}_X(\mathbb{E}_Y(Y|X))=\mathbb{E}Y}$. 可以直观理解, 首先是对 Y 求条件期望, 得到 $\mathbb{E}(Y|X)$ 是关于随机变量 X 的函数, 亦是一个随机变量, 故迭代后即是对 X 求期望. 下面给出定理 (2.1.1) 的证明.

证明. 下面的证明用到了期望的性质

$$\mathbb{E}g(X)=\int_{-\infty}^{+\infty} g(x)f_X(x)\,\mathrm{d}x,\quad \mathbb{E}g(Y)=\sum_{k=1}^{\infty} g(Y)\mathbb{P}(Y=y_k).$$

设随机变量 X 和 Y 均为连续性随机变量, 则有

$$\begin{aligned}
&\mathbb{E}(\mathbb{E}(Y|X))\\
=&\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} y\frac{f_{XY}(x,y)}{f_X(x)}\mathrm{d}y\right)f_X(x)\,\mathrm{d}x=\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} y f_{XY}(x,y)\,\mathrm{d}y\right)\frac{f_X(x)}{f_X(x)}\mathrm{d}x\\
=&\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} y f_{XY}(x,y)\,\mathrm{d}y\right)\mathrm{d}x\overset{\text{交换积分次序}}{=\!=\!=}\int_{-\infty}^{+\infty} y\left(\int_{-\infty}^{+\infty} f_{XY}(x,y)\,\mathrm{d}x\right)\mathrm{d}y\\
=&\int_{-\infty}^{+\infty} y f_Y(x)\,\mathrm{d}y=\mathbb{E}Y.
\end{aligned}$$

[6] “简单迭代期望定律” 也可称 “迭代期望定律”, 强调 “简单” 这一限定词是为了同定理 (2.1.2) 区别.

再在设随机变量 X 和 Y 为离散型随机变量, 则有

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right)=\sum_{j=1}^{\infty}\mathbb{E}\left(Y|X=x_j\right)\mathbb{P}\left(X=x_j\right)\\
=&\sum_{j=1}^{\infty}\left(\left(\sum_{k=1}^{\infty}y_k\mathbb{P}\left(Y=y_k|X=x_j\right)\right)\mathbb{P}\left(X=x_j\right)\right)\\
=&\sum_{j=1}^{\infty}\left(\left(\sum_{k=1}^{\infty}y_k\frac{\mathbb{P}\left(Y=y_k,X=x_j\right)}{\mathbb{P}\left(X=x_j\right)}\right)\mathbb{P}\left(X=x_j\right)\right)\\
=&\sum_{j=1}^{\infty}\left(\sum_{k=1}^{\infty}y_k\mathbb{P}\left(Y=y_k,X=x_j\right)\right)\overset{\text{交换求和次序}}{=\!=\!=}\sum_{k=1}^{\infty}\left(\sum_{j=1}^{\infty}y_k\mathbb{P}\left(Y=y_k,X=x_j\right)\right)\\
=&\sum_{k=1}^{\infty}y_k\mathbb{P}\left(Y=y_k\right)=\mathbb{E}Y.
\end{aligned}$$

上面的证明中无论积分还是求和, 都涉及到了交换次序, 这实际上需要严格的数学条件, 但本笔记的证明略过这些条件. □

从定理 (2.1.1) 中, 尤其是离散形式的证明中可以知道, 简单迭代迭代期望定律实际上可以理解为“先求组内均值, 再将各均值加权相加”. 考虑工资和性别, 一个例子则是

$$\begin{aligned}
&\mathbb{E}\left(\text{工资}\right)=\mathbb{E}\left(\mathbb{E}\left(\text{工资}|\text{性别}\right)\right)\\
=&\mathbb{E}\left(\text{工资}|\text{性别}=\text{男}\right)\mathbb{P}\left(\text{性别}=\text{男}\right)+\mathbb{E}\left(\text{工资}|\text{性别}=\text{女}\right)\mathbb{P}\left(\text{性别}=\text{女}\right),
\end{aligned}$$

即先计算不同性别的平均工资, 然后以性别比例作为权重计算得到平均工资.

定理 2.1.2 (迭代期望定律,law of iterated expectations)**.** *设随机变量 Y,X_1 和 X_2, 若 $\mathbb{E}Y<\infty$, 则有 $\mathbb{E}\left(\mathbb{E}\left(Y|X_1,X_2\right)|X_1\right)=\mathbb{E}\left(Y|X_1\right)$.*

显然, 定理 (2.1.1) 是定理 (2.1.2) 的特例, 这只需有 $X_2=X$ 及 X_1 为常数. 对于连续性随机变量，定理 (2.1.2) 中 $\mathbb{E}\left(Y|X_1,X_2\right)$ 的一般形式为

$$\mathbb{E}\left(Y|X_1,X_2\right)=\int_{-\infty}^{+\infty}yf_{Y|X_1,X_2}\left(y|x_1,x_2\right)\mathrm{d}y=\int_{-\infty}^{+\infty}y\frac{f_{X_1X_2Y}\left(x_1,x_2,y\right)}{f_{X_1,X_2}\left(x_1,x_2\right)}\mathrm{d}y,$$

其中 $f_{X_1X_2Y}\left(x_1,x_2,y\right)$ 为 $\left(X_1,X_2,Y\right)$ 的联合密度函数,$f_{X_1,X_2}\left(x_1,x_2\right)$ 为 $\left(X_1,X_2\right)$ 的联合密度函数. 一种对定理 (2.1.2) 的精辟概述是:*the smaller information set wins*, 即仅用 X_1 的信息不能得到拥有 $\left(X_1,X_2\right)$ 的信息的 $\mathbb{E}\left(Y|X_1,X_2\right)$, 只能得到拥有 X_1 的信息的 $\mathbb{E}\left(Y|X_1\right)$.

定理 2.1.3 (条件定理,conditioning theorem)**.** *设随机变量 X 和 Y, 若 $\mathbb{E}Y<\infty$, 则有*

$\mathbb{E}(g(X)Y|X) = g(X)\mathbb{E}(Y|X)$. 如果 $\mathbb{E}|g(X)| < \infty$, 则有

$$E(g(X)Y) = \mathbb{E}(g(X)\mathbb{E}(Y|X)).$$

定理 (2.1.3) 的前一个命题很好理解, 即是给定随机变量 X 的情况下 $g(X)$ 相当于已知的常数, 依据期望算符 $\mathbb{E}$ 的线性性质便可以提出 $g(X)$. 而后一个命题则是给出了 $E(g(X)Y)$ 的计算方法.$g(X)Y$ 的依然是随机变量, 若求 $g(X)Y$, 先求其概率密度函数并不总是容易, 但利用定理 (2.1.3) 则有

$$\mathbb{E}(g(X)Y) = \mathbb{E}_X(g(X)\mathbb{E}_Y(Y|X)) = \mathbb{E}_X(\mathbb{E}_Y(g(X)Y|X)).$$

有了定理 (2.1.1)(2.1.2)(2.1.3), 我们便可以回答为何用条件期望函数作为回归函数. 计量经济学的核心目的是进行因果推断, 以 $\mathbb{E}(Y|X)$ 作为总体回归函数, 其意味着在已知信息 X 的情况下用条件期望去解释 Y, 或者说用 $\mathbb{E}(Y|X)$ 去预测 (predict)Y. 我们将证明, 以回归方差为标准, 条件期望函数 $\mathbb{E}(Y|X)$ 即是 X 对于 Y 的**最优预测** (best predict).

定义 2.1.2 (条件期望函数误差, 回归误差). 设随机变量 X 和 Y, 则称

$$e = Y - \mathbb{E}(Y|X)$$

为**条件期望函数误差** *(CEF error)*, 又称为**回归误差** *(regression error)*.

依据定义 (2.1.2), 移项则有

$$Y = \mathbb{E}(Y|X) + e, \tag{2.1}$$

式 (2.1) 即是最一般化的计量经济学模型, 我们称之 Y 对 X[7]的为**回归方程** (regression equation). 我们常将随机变量 Y 称为**被解释变量** (explained variable)、**因变量** (dependent variable)或**回归子** (regressand), 将随机变量 X 称为**解释变量** (explanatory variable)、**自变量** (independent variable)或**回归元** (regressor), 而回归误差 e 又叫**随机扰动项** (stochastic disturbance)或**随机误差项** (stochastic error).

模型 2.1.1 (回归模型). 满足以下设定的模型称为**回归模型** *(regression model)*.

$$Y = \mathbb{E}(Y|X) + e$$

[7]这里是依据英文“Y on X”, 尽管依据中文含义, 表述“X 对 Y”可能更合理. 本笔记后续内容中“Y 对 X”均理解为“Y on X”.

回归误差 e 则是由随机向量 (X,Y) 的联合分布所决定的一个**随机变量**, 下面我们回归误差 e 的性质.

定理 2.1.4 (回归误差 e 的误差)**.** *回归误差 e 有如下性质:*

a. $\mathbb{E}(e|X)=0$;

b. $\mathbb{E}e=0$;

c. $\forall r>1$, *若* $\mathbb{E}|Y|^r<\infty$, *则* $\mathbb{E}|e|^r<\infty$;

d. $\forall h(X)$, *若* $\mathbb{E}|h(X)e|<\infty$, *则* $\mathbb{E}|h(X)e|=0$.

证明. 现仅证明性质 a 和 b, 余下证明见 Bruce(cite) 的 Section 2.33.

对于性质 a, 即有

$$\mathbb{E}(e|X)=\mathbb{E}(Y-\mathbb{E}(Y|X)|X)=\mathbb{E}(Y|X)-\mathbb{E}(\mathbb{E}(Y|X)|X)=\mathbb{E}(Y|X)-\mathbb{E}(Y|X)=0.$$

对于性质 b, 依据定理 (2.1.1) 有 $\mathbb{E}(e|X)=\mathbb{E}(\mathbb{E}(e|X))=\mathbb{E}0=0$. □

在概率论中, 如果随机变量 X 和 Y 满足有 $\mathbb{E}(Y|X=x)=\mathbb{E}Y$, 则称 Y 对于 X **均值独立** (mean independence)[8], 或 Y 均值独立于 X. 反过来,$\mathbb{E}(X|Y)=\mathbb{E}X$ 则称 X 均值独立于 Y. 容易知道,Y 均值独立于 X 的必要条件是 $\mathbb{E}(Y|X)=0$, 同理性质 b 的证明, 即得 $\mathbb{E}(Y|X)=0=\mathbb{E}$. 因此, 回归误差 e 均值独立于解释变量 X.

回归误差 e 是随机变量, 其以差值的形式刻画了条件期望函数 $\mathbb{E}(Y|X)$ 对被解释变量 Y 的解释或预测的偏离程度,$\mathbb{E}e=0$ 表明在平均意义上 $\mathbb{E}(Y|X)$ 对 Y 的偏差是不存在的, 仅给定任意的随机变量 $X=x$ 时 $\mathbb{E}(e|X=x)=0$ 依然说明了条件期望函数的预测在平均意义上是准确的. 回归误差 e 均值独立于解释变量 X, 则进一步表明了回归误差 e 与 X 是线性无关的.

如果仅从 $\mathbb{E}e=0$ 的角度说明条件期望函数具有最优的预测性质是不过的, 因为在期望的意义上有正有负的回归误差 e 会出现正负相抵的情况 (但 $\mathbb{E}e=0$ 依然是条件期望函数的重要性质). 为了避免分析中这个问题. 一种简单的方法即是研究回归误差的平方, 由此来定义最优预测.

定义 2.1.3 (最优预测量)**.** *设随机变量 X 和 Y, 记任意的 $g(X)$ 为 Y 的一个预测函数, 我们称使得 $\mathbb{E}(Y-g(X))^2$ 最小的函数为给定 X 下 Y 的***最优预测量*** (best predictor).*

这样 $\mathbb{E}(Y-g(X))^2$ 是预测误差平方的平均值, 实际上, 这里是借用数理统计中**均**

[8]可以证明, 随机变量的独立性是随机变量均值独立的必要条件, 随机变量均值独立是随机变量线性无关的必要条件, 即有 互相独立 ⇒ 均值独立 ⇒ 不相关.

方误差 (mean square error)的概念. 对于参数的估计量, 均方误差 MSE 越小则表明其估计的偏差越小[9], 数理统计中则认为 $\hat{\theta}$ 的估计更加有效. 依据定义 (2.1.3) 我们下面即可证明, 条件期望函数 $\mathbb{E}(Y|X)$ 是最优预测量.

定理 2.1.5 (最优预测). *设随机变量 X 和 Y, 记任意的 $g(X)$ 为 Y 的一个预测函数, 则有不等式*

$$\mathbb{E}(Y-g(X))^2 \geqslant \mathbb{E}(Y-\mathbb{E}(Y|X))^2 = \mathbb{E}e^2$$

成立, 当且仅当 $g(X)=\mathbb{E}(Y|X)$ 时取等.

证明. 注意到

$$\begin{aligned}\mathbb{E}(Y-g(X))^2 =& \mathbb{E}(Y-m(X)+m(X)-g(X))^2 = \mathbb{E}(e+m(X)-g(X))^2\\ =& \mathbb{E}\left(e^2+2(m(X)-g(X))e+(m(X)-g(X))^2\right)\\ =& \mathbb{E}e^2+2\mathbb{E}((m(X)-g(X))e)+\mathbb{E}(m(X)-g(X))^2,\end{aligned}$$

而依据定理 (2.1.3) 有

$$\mathbb{E}((m(X)-g(X))e) = \mathbb{E}((m(X)-g(X))\cdot\mathbb{E}(e|X)) = \mathbb{E}((m(X)-g(X))\cdot 0) = 0,$$

故有

$$\mathbb{E}(Y-g(X))^2 = \mathbb{E}e^2+\mathbb{E}(m(X)-g(X))^2 \geqslant \mathbb{E}e^2$$

成立, 当且仅当 $g(X)=m(X)=\mathbb{E}(Y|X)$ 时取等. □

上述证明, 实际上表明了对于 Y 的任意的关于 X 的预测量 $g(X)$, 其误差平方的期望 $\mathbb{E}(Y-g(X))^2$ 存在一个可达的下界 $\mathbb{E}e^2$, 在非 $g(X)=m(X)$ 的情况下 $\mathbb{E}(Y-g(X))^2$ 总是大于 $\mathbb{E}e^2$ 的. 可以说, 定理 (2.1.5) 即是计量经济学以条件期望函数 $\mathbb{E}(Y|X)$ 作为回归函数的原因.

然而, 仅从均值上考虑随机变量 e 是不够的,Bruce(cite) 在书中讲述了一个经济学家的笑话:

> An economist was standing with one foot in a bucket of boiling waterand the other foot in a bucket of ice. When asked how he felt, he replied, "*On average I feel just fine.*"

Bruce(cite) 的笑话即是忽略了变量的离散 (dispersion) 情况, 尽管均值为 0, 但观测值的取值则是两个极端. 因此, 我们有必要研究回归误差 e 的方差.

---

[9]容易证明, 参数 θ 的估计量 $\hat{\theta}$ 的均方误差 MSE 满足有 $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}(\theta-\hat{\theta})^2 = \mathrm{Var}(\hat{\theta})+(\mathbb{E}\hat{\theta}-\theta)^2$, 当估计量 $\hat{\theta}$ 为无偏估计时, 其均方误差即为 $\mathrm{Var}(\hat{\theta})$. 特别地, 若估计量 $\hat{\theta}$ 比 θ 的其他任意估计量具有更小的 MSE, 则称 $\hat{\theta}$ 为**最小方差无偏估计** (minimum variance unbiased estimate), 简称为 **MVU 估计**.

定义 2.1.4 (误差平方和, 回归方差). *对于回归误差 e, 我们记其方差为*

$$\sigma^2 = \mathrm{Var}(e) = \mathbb{E}\left(e - \mathbb{E}e\right)^2 = \mathbb{E}e^2,$$

*则称 σ^2 为***回归方差** *(regression variance),σ=$\sqrt{\mathrm{Var}\left(e\right)}$ 为***回归标准差**. *若 $\mathbb{E}e^2 < \infty$, 我们将回归误差 e 关于 X 的***条件方差** *(conditional variance)记为*

$$\sigma^2\left(x\right) = \mathrm{Var}\left(e|X = x\right) = \mathbb{E}\left(\left(e - \mathbb{E}e\right)^2|X = x\right) = \mathbb{E}\left(e^2|X = x\right),$$

*我们将 $\sigma^2\left(X\right)$ 视为随机变量, 将 $\sigma^2\left(X = x\right)$ 视为关于随机变量 $X = x$ 的函数. 同时, 我们将 $\sigma\left(X\right) = \sqrt{\mathrm{Var}\left(e|X\right)}$ 称为***条件标准差** *(conditional standard deviation).*

依据定义 (2.1.4) 以及定理 (2.1.1), 则有

$$\sigma^2 = \mathbb{E}e^2 = \mathbb{E}\left(\mathbb{E}\left(e^2|X\right)\right) = \mathbb{E}\left(\sigma^2\left(X\right)\right),$$

即回归方差 σ^2 是条件方差 $\sigma^2\left(X\right)$ 的平均值.

利用条件方差函数, 我们可以对回归误差 e 进行标准化, 令

$$u = \frac{e - 0}{\sigma\left(X\right)} = \frac{e}{\sigma\left(X\right)},$$

则有

$$\mathbb{E}u = \mathbb{E}\left(\frac{e}{\sigma\left(X\right)}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{e}{\sigma\left(X\right)}\middle| X\right)\right) = \mathbb{E}\left(\sigma\left(X\right)\mathbb{E}\left(e|X\right)\right) = \mathbb{E}\left(\sigma\left(X\right)\cdot 0\right) = 0,$$

$$\mathrm{Var}\left(u\right) = \mathbb{E}\left(u^2\right) = \mathbb{E}\left(\mathbb{E}\left(u^2|X\right)\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{e^2}{\sigma^2\left(X\right)}\middle| X\right)\right) = \mathbb{E}\left(\frac{\mathbb{E}\left(e^2|\,X\right)}{\sigma^2\left(X\right)}\right) = \mathbb{E}1 = 1,$$

这样式 (2.1) 便有

$$Y = \mathbb{E}(Y|X) + \sigma(X)u, \tag{2.2}$$

也即条件期望函数有表达式 $m(X) = Y - \sigma(X)u$.

对于条件方差函数 $\sigma^2(X)$, 我们有如下的重要定义.

定义 2.1.5 (同方差, 异方差). *对于回归误差 e 的条件方差 $\sigma^2(X) = \mathrm{Var}(e|X)$, 若 $\sigma^2(X)$ 不依赖于随机变量 X, 即满足 $\sigma^2(X) = \sigma^2$, 那么称回归误差 e 满足***同方差** *(homoskedasticity); 若 $\sigma^2(X)$ 依赖随机变量 X, 则称回归误差 e 满足***异方差** *(heteroskedasticity).*

同方差和异方差在后续讨论 OLS 估计量的性质时将发挥重要作用, 尤其是在同方

差的情形下,OLS 估计量可以成为最优无偏估计量[10]. 需要指出, 回归误差 e 满足同方差实际上是异方差的一种特例情况, 异方差是一种更加普遍的情况.

由于方差是期望运算, 本节最后补充方差的分解公式以帮助简化求解方差.

定理 2.1.6. *设随机变量 Y 和 X, 若 $\mathbb{E}Y^2 < \infty$ 则有*

$$\operatorname{Var}(Y) = \mathbb{E}(\operatorname{Var}(Y|X)) + \operatorname{Var}(\mathbb{E}(Y|X)).$$

证明. 对于条件方差 $\operatorname{Var}(Y|X)$ 有

$$\begin{aligned}\operatorname{Var}(Y|X) &= \mathbb{E}\left((Y - \mathbb{E}(Y|X))^2 |X\right) = \mathbb{E}\left((Y - m(X))^2 |X\right)\\ &= \mathbb{E}\left(Y^2 - 2Ym(X) + (m(X))^2 |X\right)\\ &= \mathbb{E}\left(Y^2|X\right) - 2(m(X))^2 + (m(X))^2\\ &= \mathbb{E}\left(Y^2|X\right) - (m(X))^2.\end{aligned}$$

将 $m(X)$ 视为随机变量, 利用公式 $\operatorname{Var}(Y|X) = \mathbb{E}(Y^2|X) - (m(X))$ 和 $\operatorname{Var}(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2$ 以及定理 (2.1.1), 对 $\operatorname{Var}(Y|X)$ 求期望有

$$\begin{aligned}\mathbb{E}(\operatorname{Var}(Y|X)) =& \mathbb{E}\left(\mathbb{E}\left(Y^2|X\right) - (m(X))^2\right) = \mathbb{E}\left(\mathbb{E}\left(Y^2|X\right)\right) - \mathbb{E}\left[(m(X))^2\right]\\ =& \mathbb{E}Y^2 - \left(\operatorname{Var}(\mathbb{E}(Y|X)) + (\mathbb{E}m(X))^2\right)\\ =& \operatorname{Var}(Y) + (\mathbb{E}Y)^2 - \operatorname{Var}(\mathbb{E}(Y|X)) - (\mathbb{E}m(X))^2\\ =& \operatorname{Var}(Y) - \operatorname{Var}(\mathbb{E}(Y|X)) + (\mathbb{E}Y)^2 - (\mathbb{E}Y)^2\\ =& \operatorname{Var}(Y) - \operatorname{Var}(\mathbb{E}(Y|X)),\end{aligned}$$

移项得 $\operatorname{Var}(Y) = \mathbb{E}(\operatorname{Var}(Y|X)) + \operatorname{Var}(\mathbb{E}(Y|X))$. □

定理 (2.1.6) 将无条件方差 $\operatorname{Var}(Y)$ 分解为**组内方差** (within group variance) $\mathbb{E}(\operatorname{Var}(Y|X))$ 和**组间方差** (across group variance)$\operatorname{Var}(\mathbb{E}(Y|X))$. 例如, 如果 X 是教育水平, 那么第一项组内方差就是教育水平条件预期的预期方差, 第二项组间方差是控制教育程度后的方差.

2.2 一元线性投影模型的总体参数

上一节我们介绍 (总体) 回归函数 $\mathbb{E}(Y|X)$, 并由此得到了最为一般的回归模型 $Y = \mathbb{E}(Y|X) + e$. 依据数理统计的方法范式, 统计推断是建立已知总体的情况下利用样本数

[10]即 Gauss-Markov 定理, 在同方差的假定下,OLS 估计量是最优线性无偏估计量 (BLUE), 而 Bruce(cite) 最新的研究证明了此条件下 OLS 估计量是最优无偏估计量 (BUE), 用数理统计的概念而言即是**最小方差无偏** (minimum variance unbiased,MVU) 估计.

据区推断. 通常而言, 在研究解释变量 X 对被解释变量 Y 的因果关系时, 就模型 (2.1.1) 而言我们只能够搜集到样本个体就 X 与 Y 的观测值, 若没有关于总体的理论则我们不能得知条件期望函数 $m(X)$ 的具体形式[11], 因而不能得到 $\mathbb{E}(Y|X)$ 的观测值, 进而回归误差 e 也是不可观测的. 一种处理方法便是设定 $m(X)$ 的具体形式, 本节我们将研究 $m(X)$ 为线性函数的情况, 并进一步研究最为广泛使用的线性投影模型.

首先我们介绍线性 CEF 模型.

定义 2.2.1 (线性条件期望函数). 对于随机变量 $Y \in \mathbb{R}$ 和随机向量 $\mathbf{X} = (X_1, \cdots X_K)^{\mathrm{T}} \in \mathbb{R}^{k\times 1}$, 若 Y 关于 X 的条件期望函数为

$$m(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X}) = \beta_1 X_1 + \cdots + \beta_K X_K, \tag{2.3}$$

则称 $m(X)$ 为**线性条件期望函数** *(linear CEF).*

利用矩阵运算, 记 $\boldsymbol{\beta} = (\beta_1 . \cdots, \beta_K)^{\mathrm{T}}$[12], 则当模型 (2.1.1) 中的 $\mathbb{E}(Y|X)$ 为式 (2.3) 时, 我们可以得到如下的模型 (2.2.1) 和模型 (2.2.2).

模型 2.2.1 (线性 CEF 模型). 满足如下设定的回归模型称为**线性 *CEF* 模型** *(linear CEF model)*或**线性回归模型** *(linear regression model).*

$$\begin{aligned} Y &= \mathbb{E}(Y|X) + e \\ \mathbb{E}(Y|X) &= \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} \\ \mathbb{E}(e|\mathbf{X}) &= 0 \end{aligned}$$

模型 2.2.2 (同方差线性 CEF 模型). 满足如下设定的回归模型称为**同方差线性 *CEF* 模型** *(homoskedastic linear CEF model)*或**同方差线性回归模型** *(homoskedastic linear regression model).*

$$\begin{aligned} Y &= \mathbb{E}(Y|X) + e \\ \mathbb{E}(Y|X) &= \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} \\ \mathbb{E}(e|\mathbf{X}) &= 0 \\ \mathbb{E}\left(e^2\middle|\mathbf{X}\right) &= \sigma^2 \end{aligned}$$

线性 CEF 模型 (2.2.1) 与回归模型 (2.1.1) 的差别仅就在于, 模型 (2.2.1) 将条件期望函数设定为了关于参数 $\boldsymbol{\beta}$ 线性函数, 因而线性 CEF 模型 (2.2.1) 只是回归模型 (2.1.1)

[11]依据样本数据, 可以使用核密度估计等方法去推断总体分布, 但这不是本章要讨论的问题.

[12]在计量经济学中, 矩阵转置更多的是使用 $\boldsymbol{A}'$ 表示, 而本笔记遵从线性代数的习惯使用 $\boldsymbol{A}^{\mathrm{T}}$ 表示矩阵转置.

的**特例**情形, 故线性 CEF 模型又被称为线性回归模型. 然而, 由于后续的模型 (2.2.3) 和模型 () 也可以被称为线性回归模型, 故为了区别, 本笔记将模型 (2.2.1) 统一称呼为“线性 CEF 模型”.

定理 (2.1.4) 依旧适用于线性 CEF 模型 (2.2.1) 的回归误差.

需要强调, 线性 CEF 模型的“线性”是指条件期望函数是参数 $\boldsymbol{\beta} = (\beta_1 \cdots, \beta_K)^{\mathrm{T}}$ 的线性函数, 这样而言,

$$\ln Y = \beta_1 X_1 + \beta_1 X_1^2 + \beta_3 \ln X_2 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$$

依然是线性回归模型. 但形如

$$\ln Y = \beta_1^2 X_1 + \beta_0 + e \quad \text{或} \quad Y = AK^{\alpha} L^{\beta} e^{\varepsilon}$$

的模型不是线性回归模型 (但可以通过变量替换转变为线性回归模型).

本节以及后续将关注最为简单的**一元线性 CEF 模型**, 其结构为

$$Y = \mathbb{E}(Y|X) = \beta_0 + \beta_1 X + e. \tag{2.4}$$

式 (2.4) 称为一元**线性回归方程** (linear regression equation). 这里的“一元”指的是只含有一个解释变量 X, 我们将 $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ 称为**回归系数** (regression coefficient), 具体地, 系数 β_1 被称为被解释变量 Y 对解释变量 X 的回归系数 (regression coefficient of Y on X), 系数 β_0 则称为**截距项** (intercept term)或**常数项** (constant term).

对于初学者而言, 式 (2.4) 其实是具有误导性的. 如果依据模型 (2.2.1), 实际上一元线性回归模型的设定应该为

$$Y = \beta_1 X + e,$$

即不含有截距项 β_0. 更一般的, 式 (2.4) 实际上应该视为特殊的二元线性回归模型, 即

$$Y = \beta_0 X_0 + \beta_1 X + e, \quad X_0 \equiv 1, \tag{2.5}$$

其中解释变量 X_0 是取值为 1 的常数 (也视为随机变量), 这时回归元被称为 **demeaned regressor**. 对于含有常数项的 demeaned regressor, 后续章节中本笔记将会详细探讨其回归模型.

我们采用模型 (2.2.1) 来研究变量 Y 和变量 X 的关系, 我们仍不能依据样本观测值来计算条件期望函数 $\mathbb{E}(Y|X)$, 因为模型 (2.1.1) 中的参数 $\boldsymbol{\beta}$ 是未知的. 但在此含义上, 我们可以对参数 $\boldsymbol{\beta}$ 进一步设定以进行统计推断, 从而使得模型更具有可操作性. 定理 (2.1.5) 证明了条件期望函数具有最小的均方误差 MSE, 则以$\boldsymbol{\beta}$ 的线性函数作为条件期望函数的 $m(X)$ 也应该满足这项要求 (必要条件), 这样, 我们从线性 CEF 模型衍生出了**线性投影模型**. 我们记一元线性投影模型为

$$Y = \mathscr{P}(Y|X) + \varepsilon = \beta_0^* + \beta_1^* X + \varepsilon, \tag{2.6}$$

其中 $\boxed{\mathscr{P}(Y|X)=\beta_0^*+\beta_1^*X}$ 称为给定 X 下 Y 的**最优线性预测量** (best linear predictor)[13], 式 (2.6) 称为一元**线性投影方程** (linear projection equation).$\mathscr{P}(Y|X)$ 是定义 (2.1.3) 的一种特殊情况, 即是假定了最优预测量为关于参数 $\boldsymbol{\beta}^*=(\beta_0^*,\beta_1^*)^{\mathrm{T}}$ 的线性函数.

需要指出,$\mathscr{P}(Y|X)=\beta_0^*+\beta_1^*X$ 中的的 β^* 是为了与式 (2.4) 的系数 β 区别, 前者我们通过后面的推导将给出具体的形式, 而后者属于假定的线性条件期望函数, 是未知的. 我们将 $\varepsilon=Y-\mathscr{P}(Y|X)=Y-\beta_0^*-\beta_1^*X$ 称为**投影误差** (projection error), 其不同于定义 (2.1.2) 的回归误差 e, 因为 β^* 并不是模型 (2.1.1) 中未知的 β. 当且仅当 $\mathbb{E}(Y|X)$ 确实为式 (2.4) 中的线性 CEF 时, 投影误差 ε 等于回归误差 e.

同理于式 (2.5), 一元线性投影方程 (2.6) 的真正形式为

$$Y=\beta_0^*X_0+\beta_1^*X+e,\quad X_0\equiv 1, \tag{2.7}$$

是具有 demeaned regressor 的二元线性投影模型. 我们之所以在一元回归模型以及一元线性投影模型中引入常数项, 一方面是对于具有常数项的线性投影模型, 其具有一些独有的性质, 另一方面在 $(X,Y)\in\mathbb{R}^2$ 中点斜式直线方程的形式为 $y=kx+b$, 一元回归下的 OLS 便是求解了一条距离样本点平均“距离”最小的直线. 不做特别说明, 本章所说的一元线性 CEF 模型或一元线性投影模型均指的是如式 (2.4) 或式 (2.6) 那样的含有截距项的模型.

对于线性投影模型, 该模型的基本假设如下.

假设 2.2.1. *对于随机变量 Y 和随机向量 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$, 假设有*

(a) $\mathbb{E}Y^2<\infty$;

(b) $\mathbb{E}\|\mathbf{X}\|^2<\infty$;

(c) $\boldsymbol{Q}_{\mathbf{XX}}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)$ *是正定矩阵.*

我们研究的模型实质上具有的是式 (2.7) 的形式, 对应地假设 (2.2.1) 中后两项是 $\mathbb{E}X^2$ 存在且有限以及对 $\mathbf{X}$ 的 Gramian 矩阵的期望

$$\boldsymbol{Q}_{\mathbf{XX}}=\mathbb{E}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)=\mathbb{E}\begin{pmatrix}1 & X\\ X & X^2\end{pmatrix}$$

是正定矩阵. 不难发现, 一元线性投影模型的假设只需有随机变量 Y 和随机变量 X 的二阶原点矩存在且有限.

线性回归模型采用了线性函数作为条件期望函数, 但参数 $\boldsymbol{\beta}=(\beta_1.\cdots,\beta_K)^{\mathrm{T}}$ 依旧是未知的. 线性投影模型则进一步限定了参数 $\boldsymbol{\beta}$ 需要满足的条件. 对于形如式 (2.4) 的

[13]符号 $\mathscr{P}$ 为手写花体字母 P, 在 LaTeX 中通过命令`\mathscr{}`输入.

一元线性 CEF 模型, 设 MSE 的期望为

$$S(\boldsymbol{\beta}) = S(\beta_0, \beta_1) = \mathbb{E}(Y - \beta_0 - \beta_1 X)^2, \tag{2.8}$$

则一元线性投影模型中 $\mathscr{P}(Y|X)$ 的参数 $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^{\mathrm{T}}$ 被定义为

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\operatorname{argmin}} \mathbb{E}(Y - \beta_0 - \beta_1 X)^2,$$

即能够使得均方误差 MSE 最小. 这样, 一元线性投影模型的参数 $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^{\mathrm{T}}$ 即是一个最优化问题的解.

现在求解使得式 (2.8) 最小的参数 $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^{\mathrm{T}}$, 本笔记将介绍两种方法, 一种是基于代数的方法, 另一种则是各教材上常见的微积分方法. 首先介绍代数求解方法. 注意到对一元线性回归方程 (2.4) 求期望有

$$\mathbb{E}Y = \beta_0 + \beta_1 \mathbb{E}X, \tag{2.9}$$

将式 (2.9) 代入式 (2.8) 消去 β_0, 即有

$$S(\beta_1) = \mathbb{E}(Y - (\mathbb{E}Y - \beta_1\mathbb{E}X + \beta_1 X))^2 = \mathbb{E}(Y - [\mathbb{E}Y + \beta_1(X - \mathbb{E}X)])^2,$$

这时 S 变成了关于单参数 β_1 的二次函数, 尝试配方得[14]

$$\begin{aligned}
S(\beta_1) =& \mathbb{E}((X - \mathbb{E}X)\beta_1 - (Y - \mathbb{E}Y))^2 \\
=& \beta_1^2 \mathbb{E}(X - \mathbb{E}X)^2 - 2\beta_1 \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) + \mathbb{E}(Y - \mathbb{E}Y)^2 \\
=& \beta_1^2 \mathrm{Var}(X) - 2\beta_1 \mathrm{Cov}(X, Y) + \mathrm{Var}(Y) \\
=& \mathrm{Var}(X)\left(\beta_1 - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\right)^2 - \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)} + \mathrm{Var}(Y),
\end{aligned}$$

则当一元线性 CEF 模型的 MSE 可以取得最小值时,β_1 满足

$$\beta_1 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

而 MSE 的最小值为

$$\begin{aligned}
S_{\min} =& \mathrm{Var}(Y) - \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)} \\
=& \mathrm{Var}(Y)\left(1 - \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)\mathrm{Var}(Y)}\right) = \mathrm{Var}(Y)\left(1 - \rho_{XY}^2\right).
\end{aligned}$$

[14]假定交换求和及交换积分次序均成立, 这里以及之后涉及到期望算符 $\mathbb{E}$ 时, 需始终牢记算符 $\mathbb{E}$ 是**线性**的.

利用式 (2.9) 则有

$$\beta_0 = \mathbb{E}Y - \beta_1 \mathbb{E}X = \mathbb{E}Y - \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}\mathbb{E}X.$$

以上代数法的推导, 即是一元线性回归方程 (2.4) 满足 MSE 最小这一必要条件时, 未知参数 $\boldsymbol{\beta}$ 所具有的性质. 则对于一元线性投影方程 (2.6), 在仅保证 MSE 最小的定义下 $\mathscr{P}(Y|X)$ 的参数 $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^{\mathrm{T}}$ 满足有

$$\beta_1^* = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}, \quad \beta_0^* = \mathbb{E}Y - \beta_1^* \mathbb{E}X = \mathbb{E}Y - \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}\mathbb{E}X.$$

下面我们再使用微积分的方法求解参数 $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^{\mathrm{T}}$. 对于式 (2.8), 求一阶偏导有

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2\mathbb{E}(Y - \beta_0 - \beta_1 X), \quad \frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2\mathbb{E}((Y - \beta_0 - \beta_1 X) X),$$

令偏导数为 0, 得到极值点的一阶条件为

$$\begin{cases} \mathbb{E}(Y - \beta_0 - \beta_1 X) = 0, \\ \mathbb{E}(X(Y - \beta_0 - \beta_1 X)) = 0, \end{cases} \Leftrightarrow \begin{cases} \mathbb{E}e = 0, \\ \mathbb{E}(Xe) = 0, \end{cases} \tag{2.10}$$

方程组 (2.10) 通常被称为一元线性 CEF 模型的**正规方程组** (normal equations)[15]或一元线性回归模型的正规方程组. 当 MSE 取最小值时, 则一元线性投影方程 (2.6) 也满足有方程组 (2.10), 这样对于式投影误差 ε 满足必要条件

$$\begin{cases} \mathbb{E}(Y - \beta_0^* - \beta_1^* X) = 0, \\ \mathbb{E}(X(Y - \beta_0^* - \beta_1^* X)) = 0, \end{cases} \Leftrightarrow \begin{cases} \mathbb{E}\varepsilon = 0, \\ \mathbb{E}(X\varepsilon) = 0, \end{cases} \tag{2.11}$$

成立.

方程组 (2.11) 表明了一元线性投影模型中投影误差 ε 满足有非条件期望为 0 以及与被解释变量 X **正交**[16].

现将正规方程组 (2.11) 展开有

$$\mathbb{E}Y - \beta_0^* - \beta_1^* \mathbb{E}X = 0, \tag{2.12}$$

$$\mathbb{E}(YX) - \beta_0^* \mathbb{E}X - \beta_1^* \mathbb{E}X^2 = 0, \tag{2.13}$$

[15]也有翻译为“常规方程组”, 见林文夫 (cite). 有问题, 正规方程组应该是特指 OLS 估计量的一阶条件?

[16]这里的正交 (orthogonality) 是概率论中的概念: 若两个随机变量 X, Y, 其积 XY 的数学期望 $\mathbb{E}XY$ 为 0, 那么则称这两个随机变量是正交的.

将式 (2.12) 代入式 (2.13) 中, 消去 β_0 得

$$\mathbb{E}(YX) - (\mathbb{E}Y - \beta_1^*\mathbb{E}X)\,\mathbb{E}X - \beta_1^*\mathbb{E}X^2 = 0,$$

化简有

$$\left(\mathbb{E}X^2 - (\mathbb{E}X)^2\right)\beta_1^* = \mathbb{E}YX - \mathbb{E}Y\mathbb{E}X,$$

解得

$$\beta_1^* = \frac{\mathbb{E}YX - \mathbb{E}Y\mathbb{E}X}{\mathbb{E}X^2 - (\mathbb{E}X)^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)},$$

将 β_1 代回式 (2.12) 可得到 β_0^* 的表达式. 可以发现, 代数法和微积分方法所解得的参数 $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^{\mathrm{T}}$ 是相同的, 由此我们得到了**一元线性投影模型**.

模型 2.2.3 (一元线性投影模型)**.** *在假设 (2.2.1) 成立时, 满足如下设定的模型称为一元***线性投影模型** *(linear projection model).*

$$\begin{aligned}
&Y = \mathscr{P}(Y|X) + \varepsilon \\
&\mathscr{P}(Y|X) = \beta_0 + \beta_1 X, \quad \begin{cases} \beta_0 = \mathbb{E}Y - \beta_1\mathbb{E}X \\ \beta_1 = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \end{cases} \\
&\mathbb{E}(X\varepsilon) = 0 \\
&\mathbb{E}(\varepsilon) = 0 \quad (\text{需要常数项}\beta_0)
\end{aligned}$$

对比一元线性投影模型 (2.2.3) 与一元线性 CEF 模型 (2.2.1), 两者的最大差别在于对于被解释变量 Y 的解释或预测是不同的. 后者使用条件期望函数 $m(X) = \mathbb{E}(Y|X)$ 作为对被解释变量 Y 的解释或预测, 只不过 $\mathbb{E}(Y|X)$ 使用了线性 CEF(但参数 $\boldsymbol{\beta}$) 未知, 而前者使用最优线性预测量 $\mathscr{P}(Y|X)$ 来对 Y 进行解释或预测, 而且 $\mathscr{P}(Y|X)$ 的参数是有具体表达式的.

一元线性投影模型 (2.2.3) 又被称为 Y 在 X 上的**线性投影** (linear projection), 且和模型 (2.1.1) 一样也被称为 Y 对 X 的**回归**. 这就导致了“回归 (regression)”一词有时是存在歧义的. 本笔记将会严格区分线性 CEF 模型 (2.2.1) 和线性投影模型 (2.2.3), 对于后者仅使用“投影”一词来表述.

另外, 之前推导 $\mathscr{P}(Y|X)$ 时为了与一元线性 CEF 方程 (2.4) 的未知参数 β 区分, 我们将 $\mathscr{P}(Y|X)$ 的参数写为 β^*. 在理解了模型 (2.2.1) 与模型 (2.2.3) 的区别后, 本笔记不再区分式 (2.4) 与式 (2.6) 中参数的符号, 都采用希腊字母 β 表示其参数. 即将一元线性投影方程 (2.6) 记为

$$Y = \mathscr{P}(Y|X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon. \tag{2.14}$$

之所以称模型 (2.2.3) 为投影模型, 这是因为在给定使得 MSE 最小的线性条件期望函数时, 回归函数 $\mathbb{E}(Y|X)$ 即是将 Y 在随机变量 X 所张成的线性空间中投影, 这一点我们将在多元线性回归模型中详细讨论.

需要指出, 模型 (2.2.3) 中 $\mathbb{E}(e)=0$ 这一性质, 实际上是因为式 (2.4) 中含有常数项 β_0才成立, 这是因为对于不含有常数项的一元线性投影模型

$$Y=\mathscr{P}(Y|X)+\varepsilon=\beta_1 X+\varepsilon$$

求解使得 MSE 最小的最优化问题时, 不存在一阶条件

$$\frac{\partial}{\partial\beta_0}S(\beta_0,\beta_1)=-2\mathbb{E}(Y-\beta_0-\beta_1X),$$

故在不含有常数项的线性投影模型中并不能够保证 $\mathbb{E}(e)=\mathbb{E}(Y-\beta_1X)=0$.

现在我们研究模型 (2.2.3) 的回归方差. 依据定义 (2.1.4), 对式 (2.14) 求方差, 则

$$\mathrm{Var}(Y)=\mathrm{Var}(\beta_0+\beta_1X+\varepsilon)=\beta_1^2\mathrm{Var}(X)+\mathrm{Var}(\varepsilon)+2\beta_1\mathrm{Cov}(X,\varepsilon),$$

而模型 (2.2.3) 含有常数项有

$$\mathrm{Cov}(X,\varepsilon)=\mathbb{E}(X\varepsilon)-\mathbb{E}X\mathbb{E}\varepsilon=0+\mathbb{E}X\cdot 0=0,$$

即X 与投影误差 ε 不相关, 于是一元线性投影模型的方差 σ^2 满足有

$$\sigma^2=\mathrm{Var}(\varepsilon)=\mathrm{Var}(Y)-\beta_1^2\mathrm{Var}(X). \tag{2.15}$$

2.3 一元线性投影模型的样本估计量

在理清了一元线性投影模型的总体参数后, 本节研究使用样本估计对应的估计量. 为了对模型 (2.2.3) 进行参数估计, 设抽取了样本数为 n 的**简单随机样本** (simple random sample), 即样本 $\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$ 满足**独立同分布** (independent and identically distributed), 具体地, 即 (X_i,Y_i) 与 (X_j,Y_j) 是相互独立的 $(i\neq j)$, 而样本个体 (X_i,Y_i) 与总体 (X,Y) 具有相同的概率分布 $(i=1,\cdots,n)$. 对于总体所具有的概率分布, 在计量经济学中又被称为**数据生成过程** (data generating process), 这意味着样本来自于一个无穷大的潜在总体或理论上的总体.

于是, 本节研究的总体与样本分别为

$$\begin{aligned}
\text{总体}:&Y=\beta_0+\beta_1X+\varepsilon\\
\text{样本}:&\boldsymbol{Y}=\beta_0\mathbf{1}+\beta_1\boldsymbol{X}+\boldsymbol{\varepsilon}\\
\text{样本个体}:&Y_i=\beta_0+\beta_1X_i+\varepsilon_i,\quad i=1,\cdots,n
\end{aligned}$$

其中列向量 $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\varepsilon}, \mathbf{1}$ 分别为

$$\boldsymbol{Y}=\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X}=\begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_n \end{pmatrix}, \quad \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathbf{1}=\begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}.$$

对于模型 (2.2.3), 其参数 $\boldsymbol{\beta}=(\beta_1,\beta_1)^{\mathrm{T}}$ 为

$$\beta_0=\mathbb{E}Y-\beta_1\mathbb{E}X, \quad \beta_1=\frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)},$$

则对于参数 $\boldsymbol{\beta}=(\beta_1,\beta_1)^{\mathrm{T}}$ 的一种估计方法是基于**矩估计** (moment estimation)的方法, 先使用样本矩[17]去估计 $\operatorname{Var}(X)$ 以及 $\operatorname{Cov}(X,Y)$, 则有参数 $\boldsymbol{\beta}$ 的估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$(分别读作 β_0 hat 和 β_1 hat, 其余估计量的读法同理) 为

$$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X},$$

和

$$\hat{\beta}_1=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\frac{1}{n^2}\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n^2}\left(\sum_{i=1}^{n}X_i\right)^2}$$

其中 $\bar{X}$ 和 $\bar{Y}$(分别读作 X bar 和 Y bar, 其余样本均值的读法同理) 为样本均值, 分别满足有

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i^2, \quad \bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i^2.$$

对于投影误差 ε, 尽管我们知道模型 (2.2.3) 中参数 $\boldsymbol{\beta}$ 的表达式, 但由于总体的概率分布未知, 我们只能使用样本去估计 $\boldsymbol{\beta}$, 因而 ε 是不可观测的. 而对投影误差 ε 的一种合理的估计量为 $\hat{\varepsilon}_i=Y_i-(\beta_0+\beta_1X_i)$. 同时, 上一节我们计算了一元线性投影模型的回归方差为式 (2.15), 则对于 $\sigma^2=\operatorname{Var}(\varepsilon)=\operatorname{Var}(Y)-\beta_1^2\operatorname{Var}(X)$ 的矩估计为

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2-\frac{\hat{\beta}_1^2}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2. \tag{2.16}$$

[17]对于随机变量 X, 其有简单随机样本 $\{X_1,\cdots,X_n\}$, 对于正整数 k. 设

$$a_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k, \quad m_k=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^k,$$

则称 a_k 为 k 阶**样本原点矩** (sample origin moment),m_k 为 k 阶**样本中心矩** (sample center moment).

上面给出了模型 (2.2.3) 的矩估计. 但对于线性投影模型而言, 为了估计参数 $\boldsymbol{\beta}$, 更加常用的方法是使用普通最小二乘法.

定义 2.3.1 (最小二乘法估计量). *对于模型 (2.2.3), 设有简单随机样本* $\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$, *则均方误差* $\mathbb{E}\left(Y-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)^2$ *的矩估计量为*

$$\hat{S}(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2=\frac{1}{n}\mathrm{SSE}(\beta_0,\beta_1),$$

其中

$$\mathrm{SSE}(\beta_0,\beta_1)=\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2$$

*被称为***误差平方和** *(sum of squared errors)函数. 对于使得* $\hat{S}(\beta_0,\beta_1)$ *最小的*

$$\hat{\boldsymbol{\beta}}=\left(\hat{\beta}_0,\hat{\beta}_1\right)^{\mathrm{T}}=\underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\operatorname{argmin}}\ \hat{S}(\boldsymbol{\beta}),$$

则称 $\left(\hat{\beta}_0,\hat{\beta}_1\right)$ *为***最小二乘估计量** *(least squares estimator)或***普通最小二乘估计量** *(ordinary least squares estimator), 简称为* ***LS* 估计量** *(LS estimator)或* ***OLS* 估计量** *(OLS estimator).*

定义 (2.3.1) 给出的 $\left(\hat{\beta}_0,\hat{\beta}_1\right)^{\mathrm{T}}$ 是对于模型 (2.2.3) 的总体参数 $\boldsymbol{\beta}=(\beta_1,\beta_1)^{\mathrm{T}}$ 的估计量. 需要强调, 参数 $\boldsymbol{\beta}=(\beta_1,\beta_1)^{\mathrm{T}}$ 是由总体确定的非随机的、固定的的数值, 而估计量 $\left(\hat{\beta}_0,\hat{\beta}_1\right)^{\mathrm{T}}$ 则是随样本变化的随机变量. 另外, 为了表示估计方法,OLS 估计量 $\hat{\boldsymbol{\beta}}$ 时常被加上角标后记为 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$; 为了表示样本容量为 n, 同理 OLS 估计量加上角标后记为 $\hat{\boldsymbol{\beta}}_n$.

现在我们求解 OLS 估计量的具体形式, 与求解模型 (2.2.3) 的总体参数类似,$\hat{\boldsymbol{\beta}}$ 同样有代数和微积分两种方法求解. 同样的, 本笔记将使用两种方法求解. 首先使用代数法, 对于

$$\hat{S}(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X\right)^2,$$

展开即有

$$\begin{aligned}
&\hat{S}(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left[Y_i^2-2Y_i\left(\beta_0+\beta_1X_i\right)+\left(\beta_0+\beta_1X_i\right)^2\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\frac{2}{n}\sum_{i=1}^{n}\left[Y_i\left(\beta_0+\beta_1X_i\right)\right]+\frac{1}{n}\sum_{i=1}^{n}\left(\beta_0+\beta_1X_i\right)^2\\
=&\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\frac{2\beta_0}{n}\sum_{i=1}^{n}Y_i-\frac{2\beta_1}{n}\sum_{i=1}^{n}Y_iX_i+\left({\beta_0}^2+\frac{2\beta_0\beta_1}{n}\sum_{i=1}^{n}X_i+\frac{\beta_1^2}{n}\sum_{i=1}^{n}X_i^2\right),
\end{aligned}$$

分别记随机变量 X,Y 和 XY 的 2 阶样本原点矩为

$$a_2^X = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad a_2^Y = \frac{1}{n}\sum_{i=1}^{n} Y_i^2, \quad a_2^{XY} = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i,$$

则 $\hat{S}(\boldsymbol{\beta})$ 有

$$\begin{aligned}
&\hat{S}(\boldsymbol{\beta}) = a_2^Y - 2\beta_0\bar{Y} - 2\beta_1 a_2^{XY} + \left(\beta_0{}^2 + 2\beta_0\beta_1\bar{X} + a_2^X\beta_1^2\right)\\
=&\beta_0{}^2 + a_2^X\beta_1^2 + 2\bar{X}\beta_0\beta_1 - 2\bar{Y}\beta_0 - 2a_2^{XY}\beta_1 + a_2^Y\\
=&\beta_0{}^2 - 2\beta_0\left(\bar{Y} - \bar{X}\beta_1\right) + a_2^X\beta_1^2 - 2a_2^{XY}\beta_1 + a_2^Y\\
=&\left[\beta_0 - \left(\bar{Y} - \bar{X}\beta_1\right)\right]^2 - \left(\bar{Y} - \bar{X}\beta_1\right)^2 + a_2^X\beta_1^2 - 2a_2^{XY}\beta_1 + a_2^Y\\
=&\left[\beta_0 - \left(\bar{Y} - \bar{X}\beta_1\right)\right]^2 - \left(\bar{Y}^2 - 2\bar{Y}\bar{X}\beta_1 + \bar{X}^2\beta_1^2\right) + a_2^X\beta_1^2 - 2a_2^{XY}\beta_1 + a_2^Y\\
=&\left[\beta_0 - \left(\bar{Y} - \bar{X}\beta_1\right)\right]^2 + \left(a_2^X - \bar{X}^2\right)\beta_1^2 - 2\left(a_2^{XY} - \bar{X}\bar{Y}\right)\beta_1 + a_2^Y - \bar{Y}^2\\
=&\left[\beta_0 - \left(\bar{Y} - \bar{X}\beta_1\right)\right]^2 + \left(a_2^X - \bar{X}^2\right)\left(\beta_1 - \frac{a_2^{XY} - \bar{X}\bar{Y}}{a_2^X - \bar{X}^2}\right)^2 - \frac{\left(a_2^{XY} - \bar{X}\bar{Y}\right)^2}{a_2^X - \bar{X}^2} + a_2^Y - \bar{Y}^2,
\end{aligned}$$

最终配方得到标准二次型[18]

$$\begin{aligned}
\hat{S}(\boldsymbol{\beta}) = &\left[\beta_0 - \left(\bar{Y} - \bar{X}\beta_1\right)\right]^2 + \left(a_2^X - \bar{X}^2\right)\left(\beta_1 - \frac{a_2^{XY} - \bar{X}\bar{Y}}{a_2^X - \bar{X}^2}\right)^2\\
&+ \left(\frac{a_2^{XY} - \bar{X}\bar{Y}}{a_2^X - \bar{X}^2}\right)^2 + a_2^Y - \bar{Y}^2.
\end{aligned} \tag{2.17}$$

对于式 (2.17), 前三项分别为平方项, 这三项的取值很不为负, 而第四项

$$a_2^Y - \bar{Y}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2 = \frac{1}{n^2}\left(n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right),$$

依据 Cauchy-Schwarz 不等式[19]有

$$\frac{1}{n^2}\left(n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right) \geqslant \frac{1}{n^2}\left(n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i^2\right)\left(\sum_{i=1}^{n} 1^2\right)\right) = 0,$$

[18]在线性代数中, 对于二次型的配方问题存在通法, 任意的二次型可以通过矩阵相合 (congruent) 化简为标准二次型.

[19]设 $a_i, b_i \in \mathbb{R}$ 且 $b_i \neq 0$, 则 Cauchy-Schwarz 不等式为

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leqslant \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right),$$

当且仅当 $\dfrac{a_1}{b_1} = \cdots = \dfrac{a_n}{b_n}$ 时取等. 在很多问题中,Cauchy-Schwarz 不等式又被称为内积不等式.

当且仅当 $Y_1 = \cdots = Y_n$ 时取等. 这样,$\hat{S}(\boldsymbol{\beta})$ 的四项均非负数, 且不难得知 $\hat{S}(\beta_0, \beta_1)$ 作为关于 $\boldsymbol{\beta} = (\beta_1, \beta_1)^{\mathrm{T}}$ 存在最小值. 对于式 (2.17), 前两项为关于 $\boldsymbol{\beta} = (\beta_1, \beta_1)^{\mathrm{T}}$ 的二次项, 当这两项取 0 时便有

$$\hat{\beta}_0 = \bar{Y} - \bar{X}\hat{\beta}_1$$

以及

$$\hat{\beta}_1 = \frac{a_2^{XY} - \bar{X}\bar{Y}}{a_2^X - \bar{X}^2} = \frac{\frac{1}{n}\sum_{i=1}^n Y_i X_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \frac{1}{n^2}\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^n X_i\right)^2},$$

这与本节开头给出的矩估计是一样的. 同时, 代数法的方便之处在于通过式 (2.17) 我们可以得知 OLS 估计量下均方误差的矩估计量为

$$\hat{S}(\beta_0, \beta_1) = \frac{\left(a_2^{XY} - \bar{X}\bar{Y}\right)^2}{a_2^X - \bar{X}^2} + a_2^Y - \bar{Y}^2 = \frac{\left(\frac{1}{n}\sum_{i=1}^n Y_i X_i - \bar{X}\bar{Y}\right)^2}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2} + \frac{1}{n}\sum_{i=1}^n Y_i^2 - \bar{Y}^2,$$

不难发现, 这正是代数法解出模型 (2.2.3) 的总体参数 $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ 后所计算的 MSE 的矩估计.

现在我们使用微积分来求解 OLS 估计量. 为了得到

$$\hat{S}(\boldsymbol{\beta}) = \frac{1}{n}\mathrm{SSE}(\beta_0, \beta_1),$$

的最小值, 只需要 $\mathrm{SSE}(\beta_0, \beta_1)$ 取最小值即可. 则对

$$\mathrm{SSE}(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

求一阶偏导有

$$\frac{\partial}{\partial \beta_0}\mathrm{SSE} = -2\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i), \quad \frac{\partial}{\partial \beta_1}\mathrm{SSE} = -2\sum_{i=1}^n [(Y_i - \beta_0 - \beta_1 X_i) X_i],$$

求二阶偏导数有

$$\frac{\partial^2}{\partial \beta_0^2}\mathrm{SSE} = 2n, \quad \frac{\partial^2}{\partial \beta_1 \partial \beta_0}\mathrm{SSE} = 2\sum_{i=1}^n X_i,$$

以及

$$\frac{\partial^2}{\partial \beta_0 \partial \beta_1}\mathrm{SSE} = 2\sum_{i=1}^n X_i, \quad \frac{\partial^2}{\partial \beta_1^2}\mathrm{SSE} = 2\sum_{i=1}^n X_i^2.$$

SSE 最优化有一阶条件为

$$\begin{cases} \dfrac{\partial}{\partial \beta_0}\mathrm{SSE} = 0, \\ \dfrac{\partial}{\partial \beta_1}\mathrm{SSE} = 0, \end{cases}$$

即有

$$\begin{cases} \sum\limits_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right) = 0, \\ \sum\limits_{i=1}^{n} \left[\left(Y_i - \beta_0 - \beta_1 X_i\right) X_i\right] = 0, \end{cases} \Rightarrow \begin{cases} \sum\limits_{i=1}^{n} \hat{\varepsilon}_i = 0, \\ \sum\limits_{i=1}^{n} \hat{\varepsilon}_i X_i = 0, \end{cases} \tag{2.18}$$

其中 $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ 称为**残差** (residual)[20], 其为回归误差 e_i 的估计量. 一阶条件得到的式 (2.18) 也称为 OLS 估计量的**正规方程组**, 其与式 (2.11) 是对应的. 此外, 正规方程组可以用矩阵语言书写为

$$\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}} \mathbf{1} = 0, \quad \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}} \boldsymbol{X} = 0, \tag{2.19}$$

这里 $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n)^{\mathrm{T}}$ 为**残差向量**.

将式 (2.18) 两边同时除以 n, 化简得

$$\begin{cases} \bar{Y} - \beta_0 - \bar{X}\beta_1 = 0, \\ \dfrac{1}{n}\sum\limits_{i=1}^{n} Y_i X_i - \beta_0 \bar{X} - \dfrac{\beta_1}{n}\sum\limits_{i=1}^{n} X_i^2 = 0, \end{cases}$$

这即是关于 β_0 和 β_1 的二元一次方程组[21], 消去 β_0 得

$$\frac{1}{n}\sum_{i=1}^{n} Y_i X_i + \bar{X}^2 \beta_1 - \bar{X}\bar{Y} - \frac{\beta_1}{n}\sum_{i=1}^{n} X_i^2 = 0,$$

解得

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum\limits_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{n\sum\limits_{i=1}^{n} X_i Y_i - \left(\sum\limits_{i=1}^{n} X_i\right)\left(\sum\limits_{i=1}^{n} Y_i\right)}{n\sum\limits_{i=1}^{n} X_i^2 - \left(\sum\limits_{i=1}^{n} X_i\right)^2},$$

而 $\hat{\beta}_0 = \bar{Y} - \bar{X}\hat{\beta}_1$.

---

[20] 上述一阶条件中并没有给 β 标上 $\hat{}$, 这是因为在求解极值过程中视 β 为变量, 而最终求得的极值点 $\hat{\beta}$ 是满足一阶条件的.

[21] 国内不少计量经济学的教材, 推导到此处后使用 Cramer 法则来求解 $\hat{\beta}_0$ 和 $\hat{\beta}_1$, 这未免有点 “高射炮打蚊子”, 令人怀疑教材的编者是否真的动手推导了 OLS 估计量. 但反过来讲, 也恰好就这种情形 “适合” 使用 Cramer 法则了...XD

现在检验极值点的二阶条件[22], 注意到 Hessen 矩阵为

$$\boldsymbol{H}=\begin{pmatrix}\dfrac{\partial^2}{\partial\beta_0^2}\mathrm{SSE} & \dfrac{\partial^2}{\partial\beta_1\partial\beta_0}\mathrm{SSE}\\ \dfrac{\partial^2}{\partial\beta_0\partial\beta_1}\mathrm{SSE} & \dfrac{\partial^2}{\partial\beta_1^2}\mathrm{SSE}\end{pmatrix}=\begin{pmatrix}2n & 2\sum\limits_{i=1}^{n}X_i\\ 2\sum\limits_{i=1}^{n}X_i & 2\sum\limits_{i=1}^{n}X_i^2\end{pmatrix},$$

则

$$|\boldsymbol{H}|=4n\sum_{i=1}^{n}X_i^2-4\left(\sum_{i=1}^{n}X_i\right)^2=4\left[n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right],$$

同理依据 Cauchy-Schwarz 不等式知 $|\boldsymbol{H}|\geqslant 0$, 但且仅当 $X_1=\cdots=X_n$ 时取等. 通常情况, 简单随机样本不会有所有观察值均相同, 故 $|\boldsymbol{H}|>0$, 可以判断 $|\boldsymbol{H}|$ 为正定矩阵, 则求解的 $\hat{\boldsymbol{\beta}}=\left(\hat\beta_0,\hat\beta_1\right)^{\mathrm{T}}$ 为极小值点, 也是 SSE 的最小值点.

最终, 无论是代数法还是微积分方法, 我们解得的的 OLS 估计量与本节开篇提出的矩估计的形式一样的.

定理 2.3.1. *模型 (2.2.3) 的参数 β_0 和 β_1 的 OLS 估计量分别为*

$$\hat\beta_0=\bar Y-\hat\beta_1\bar X, \tag{2.20}$$

$$\begin{aligned}\hat\beta_1=&\frac{\sum\limits_{i=1}^{n}\left(X_i-\bar X\right)\left(Y_i-\bar Y\right)}{\sum\limits_{i=1}^{n}\left(X_i-\bar X\right)^2}=\frac{\sum\limits_{i=1}^{n}\left(X_i-\bar X\right)Y_i}{\sum\limits_{i=1}^{n}\left(X_i-\bar X\right)^2}\\ =&\frac{\sum\limits_{i=1}^{n}Y_iX_i-n\bar X\bar Y}{\sum\limits_{i=1}^{n}X_i^2-n\bar X^2}=\frac{n\sum\limits_{i=1}^{n}X_iY_i-\left(\sum\limits_{i=1}^{n}X_i\right)\left(\sum\limits_{i=1}^{n}Y_i\right)}{n\sum\limits_{i=1}^{n}X_i^2-\left(\sum\limits_{i=1}^{n}X_i\right)^2},\end{aligned} \tag{2.21}$$

且 OLS 估计量满足有

$$\begin{cases}\sum\limits_{i=1}^{n}\left(Y_i-\hat\beta_0-\hat\beta_1X_i\right)=0,\\ \sum\limits_{i=1}^{n}\left[\left(Y_i-\hat\beta_0-\hat\beta_1X_i\right)X_i\right]=0,\end{cases}\Rightarrow\begin{cases}\sum\limits_{i=1}^{n}\hat\varepsilon_i=0,\\ \sum\limits_{i=1}^{n}\hat\varepsilon_iX_i=0.\end{cases}$$

2.4 一元 OLS 估计量的 Gauss-Markov 定理

本节我们将研究 OLS 估计量所具有的有限性质. 在概率论与数理统计中, 对于点估计的**估计准则**, 可以分为**有限样本性质** (或小样本性质) 和**大样本性质**. 前者是考虑样本

[22]依据 SSE 的结构, 显然其不存在最大值, 则求解的极值点为极小值点.

容量 n 有限的情况下, 研究估计量的无偏性、有效性, 后者则是在 $n \to \infty$ 时, 研究估计量相合性、渐进分布等问题. 本节将详细研究模型 (2.2.3) 的 OLS 估计量以及其衍生估计量的有限性质, 同时将证明 Gauss-Markov 定理.

在前面的章节中, 模型 (2.2.1) 和模型 (2.2.3) 都可以被称为线性回归模型, 为了区分我们分别称之为一元线性 CEF 模型 (2.2.1) 与一元线性投影模型 (2.2.3). 这样区别的另一个原因, 则是为了将“线性回归模型”的名称留给下面的模型.

模型 2.4.1 (一元线性回归模型)**.** *在假设 (2.2.1) 以及*

$$\mathbb{E}(\varepsilon|X) = 0 \tag{2.22}$$

*成立时, 满足如下设定的模型称为一元***线性回归模型** *(linear regression model).*

$$\begin{aligned}
&Y = \mathscr{P}(Y|X) + \varepsilon \\
&\mathscr{P}(Y|X) = \beta_0 + \beta_1 X, \quad \begin{cases} \beta_0 = \mathbb{E}Y - \beta_1 \mathbb{E}X \\ \beta_1 = \dfrac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)} \end{cases} \\
&\mathbb{E}(X\varepsilon) = 0 \\
&\mathbb{E}(\varepsilon) = 0 \quad (\underline{\text{不}}\text{需要常数项}\beta_0)
\end{aligned}$$

不难发现, 模型 (2.4.1) 即是在一元线性投影模型之上补充了假设式 (2.22), 而 $\mathbb{E}(\varepsilon|X) = 0$ 则是线性 CEF 模型 (2.2.1) 的性质. 实际上, 对于模型 (2.4.1) 的总体回归方程

$$Y = \mathscr{P}(Y|X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon \tag{2.23}$$

求给定 X 下的条件期望, 则

$$\mathbb{E}(Y|\,X) = \mathbb{E}(\beta_0 + \beta_1 X + \varepsilon|\,X) = \beta_0 + \beta_1 X + \mathbb{E}(\varepsilon|\,X),$$

而模型 (2.4.1) 中补充的假设式 (2.22), 这样模型 (2.4.1) 的条件期望函数即满足有

$$\mathbb{E}(Y|\,X) = \beta_0 + \beta_1 X,$$

即为符合定义 (2.2.1) 的线性条件期望函数函数. 这就是说, 若一元线性投影模型 (2.2.3) 满足假设 $\mathbb{E}(\varepsilon|X) = 0$, 那么即为一元线性 CEF 模型 (2.2.1).

模型 (2.4.1) 即是将模型 (2.2.3) 和模型 (2.2.1) 合二为一[23], 我们也不必再区分 CEF 模型与投影模型的区别, 故将模型 (2.4.1) 称为**一元线性回归模型**. 对于大多数的计量经

[23]实质上, 从推导的逻辑而言, 线性投影模型 (2.2.3的性质也是2.2.1) 所具有的. 但我们无法确定研究问题中 Y 与 X 的具体关系, 故分别设定了 CEF 和最优线性预测量.

济学教材, 其讨论的线性回归模型便是模型 (2.4.1). 由于模型 (2.4.1) 的条件期望函数为线性的, 这样其投影误差 $\varepsilon = Y - \beta_0 - \beta_1 X$ 即是定义 (2.1.2) 中**回归误差** (regression error), 我们仍记为 ε 以表明其源自线性投影模型.

在研究 OLS 估计量的有限样本性质前, 我们先研究一下模型 (2.4.1) 简单随机样本 $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ 下具有的性质. 对于样本个体

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots, n, \tag{2.24}$$

由 (Y_i, X_i) 独立同分布 $(i = 1, \cdots, n)$ 知投影误差 $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$ 也是独立同分布的, 因而 (ε_i, X_i) 也是独立同分布的. 这样, 对于式 (2.24) 而言, 样本个体的被解释变量 Y_i 和解释变量 X_i 以及投影残差 ε_i 的非条件期望以及非条件方差满足有

$$\begin{aligned}
&\mathbb{E}X_i = \mathbb{E}X, \quad \operatorname{Var}(X_i) = \operatorname{Var}(X), \\
&\mathbb{E}Y_i = \mathbb{E}Y, \quad \operatorname{Var}(Y_i) = \operatorname{Var}(Y), \\
&\mathbb{E}\varepsilon_i = \mathbb{E}\varepsilon, \quad \operatorname{Var}(\varepsilon_i) = \operatorname{Var}(\varepsilon),
\end{aligned}$$

这是依据同分布条件得到的, 样本的随机变量的数值特征与总体的数值特征相同. 而对于条件期望 $\mathbb{E}(Y_i|\boldsymbol{X})$ 则有,

$$\boxed{\mathbb{E}(Y_i|\boldsymbol{X})} = \mathbb{E}(Y_i|X_i) = \mathbb{E}(Y|X_i) = \boxed{m(x_i)} = \beta_0 + \beta_1 x_i,$$

第一个等号是因为不同 i 的样本是独立的, 而由于样本个体 i 是同分布的, 则 $E(Y_i|X_i) = \mathbb{E}(Y|X_i)$ 成立, 这意味着所以的样本个体 Y_i具有相同函数形式的条件期望$m(\cdot)$, 但是其确切的值依赖于样本个体 X_i 的观测值 x_i, 也就是说, 条件期望 $\mathbb{E}(Y_i|\boldsymbol{X})$ 对于所以的样本个体具有相同的函数形式, 但具体值由索引 i 决定. 用矩阵符号, 则有

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \mathbf{1}_{n\times 1}\beta_0 + \boldsymbol{X}\beta_1 = (\mathbf{1}_{n\times 1}, \boldsymbol{X}) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{1\times n} \\ \boldsymbol{X}^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}} \boldsymbol{\beta}.$$

对于回归误差 ε_i 的条件期望 $(i = 1 \cdots, n)$, 同理上述推导有

$$\begin{aligned}
&\mathbb{E}(\varepsilon_i|\boldsymbol{X}) = \mathbb{E}(\varepsilon_i|X_i) = \mathbb{E}(\varepsilon|X = x_i) \overset{\text{式 (2.22)}}{=\!=\!=} 0, \\
&\mathbb{E}(\varepsilon_i\varepsilon_j|\boldsymbol{X}) \overset{\text{样本 i.i.d.}}{=\!=\!=} \mathbb{E}(\varepsilon_i|\boldsymbol{X})\mathbb{E}(\varepsilon_j|\boldsymbol{X}) = 0, \quad i \neq j, \\
&\mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right) = \mathbb{E}\left(\varepsilon_i^2\middle|X_i\right) = \mathbb{E}\left(\varepsilon^2\middle|X = x_i\right) = \sigma^2(X = x_i),
\end{aligned}$$

因为模型 (2.4.1) 中假设了预测误差 ε 均值独立于被解释变量 X, 故 $\mathbb{E}(\varepsilon_i|\boldsymbol{X})$ 的取值为 0. 依定理 (2.1.2), 这又意味 $\mathbb{E}\varepsilon_i$ 着而 $\mathbb{E}(\varepsilon_i^2|\boldsymbol{X})$ 的形式为条件方差函数 $\sigma^2(X)$, 但实际的观测值则依赖于索引 i. 若假设

$$\mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right) = \sigma^2(X = x_i) = \sigma^2, \tag{2.25}$$

则模型 (2.4.1) 满足同方差假设.

模型 2.4.2 (同方差一元线性回归模型). 在模型 *(2.4.1)* 的假设中加入式 *(2.25)*, 则称该模型为**同方差线性回归模型** *(homoskedastic linear regression model)*.

在同方差假设下, 则被解释变量样本个体 Y_i 的的条件方差为

$$\begin{aligned}\boxed{\operatorname{Var}(Y_i|\boldsymbol{X})} &= \operatorname{Var}(Y_i|X_i) = \operatorname{Var}(\beta_0+\beta_1X_i+\varepsilon_i|X_i) = \boxed{\operatorname{Var}(\varepsilon_i|X_i)} \\ &= \mathbb{E}\left((\varepsilon_i-\mathbb{E}(\varepsilon_i|X_i))^2\middle|X_i\right) = \boxed{\mathbb{E}\left(\varepsilon_i^2\middle|X_i\right)} = \sigma^2.\end{aligned}$$

对于模型 (2.4.1) 的正交性质 $\mathbb{E}(X\varepsilon)=0$, 则对于样本有 $\mathbb{E}(X_i\varepsilon_j)=0$, 这也意味着对于一元线性回归模型, 样本矩阵 (列向量)$\boldsymbol{X}=(X_1,\cdots,X_n)^{\mathrm{T}}$ 与误差向量 $\boldsymbol{\varepsilon}=(\varepsilon_1,\cdots,\varepsilon_n)^{\mathrm{T}}$ 正交, 即 $\mathbb{E}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)=0$.

我们现考虑样本均值 $\bar{Y}$, 在数理统计中我们知道其为总体均值的无偏估计. 现在我们计算 $\bar{Y}$ 的条件均值, 则有

$$\begin{aligned}\mathbb{E}\left(\bar{Y}\middle|\boldsymbol{X}\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Y_i\middle|\boldsymbol{X}\right) = \frac{1}{n}\sum_{i=1}^n\mathbb{E}(Y_i|\boldsymbol{X}) = \boxed{\frac{1}{n}\sum_{i=1}^n\mathbb{E}(Y_i|X_i)} \\ &= \frac{1}{n}\sum_{i=1}^n\mathbb{E}(Y|X_i) = \beta_0+\frac{\beta_1}{n}\sum_{i=1}^nX_i = \beta_0+\beta_1\bar{X},\end{aligned}$$

即样本均值 $\bar{Y}$ 的条件期望是条件期望函数 $m(X)$ 的样本均值, 依据条件期望定理, 不难得知 $\mathbb{E}\left(\bar{Y}\middle|\boldsymbol{X}\right)$ 是样本均值 $\bar{Y}$ 的无偏估计. 需要注意的是, 因为样本 $\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$ 是独立同分布的, 上面计算中标红的过程才得以成立. 而对于 $\bar{Y}$ 的条件方差, 同样计算有

$$\begin{aligned}\operatorname{Var}\left(\bar{Y}\middle|\boldsymbol{X}\right) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^nY_i\middle|\boldsymbol{X}\right) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^nY_i\middle|\boldsymbol{X}\right) \\ &\overset{\text{样本 i.i.d}}{=\!=\!=} \frac{1}{n^2}\sum_{i=1}^n\operatorname{Var}(Y_i|\boldsymbol{X}) = \frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left(\varepsilon_i^2\middle|X_i\right),\end{aligned}$$

若是满足同方差假设的模型 (2.4.2), 则有 $\operatorname{Var}\left(\bar{Y}\middle|\boldsymbol{X}\right)=\dfrac{1}{n^2}\sum\limits_{i=1}^n\sigma^2=n^{-1}\sigma^2$, 样本均值 $\bar{Y}$ 的条件方差与非条件方差相同.

下面我们研究 OLS 估计量的无偏性. 我们首先证明 OLS 估计量是关于 $Y_1,\cdots,Y_n$ 的线性组合, 然后证明 $\hat{\boldsymbol{\beta}}=\left(\hat{\beta}_0,\hat{\beta}_1\right)^{\mathrm{T}}$ 是 $\boldsymbol{\beta}=(\beta_0,\beta_1)^{\mathrm{T}}$ 的无偏估计, 最后证明 OLS 估计量的 Gauss-Markov 定理.

定理 2.4.1. 模型 *(2.2.3)* 的 *OLS* 估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是 $Y_1,\cdots,Y_n$ 的线性组合.

证明. 我们首先给出使用求和符号的证明, 然后给出基于向量符号的证明.

由定理 (2.3.1) 有

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2},$$

记 $\boxed{A^{-1}(\boldsymbol{X}) = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$ 以及离差 $\boxed{B_i(\boldsymbol{X}) = X_i - \bar{X}}$, 其中 $\boldsymbol{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$, $i = 1, \cdots, n$, 易知 $\sum_{i=1}^{n} B_i = 0$, 则

$$\begin{aligned}\hat{\beta}_1 &= A\sum_{i=1}^{n} B_i\left(Y_i - \bar{Y}\right) = A\sum_{i=1}^{n} B_i Y_i - A\sum_{i=1}^{n} B_i \bar{Y} \\ &= A\sum_{i=1}^{n} B_i Y_i - A\bar{Y}\sum_{i=1}^{n} B_i = A\sum_{i=1}^{n} B_i Y_i,\end{aligned}$$

由样本均值 $\bar{X}$ 可知 B_i 不全为 0, 故 $\hat{\beta}_0$ 是 $Y_1, \cdots, Y_n$ 的线性组合.

而 $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, 其中 $\bar{Y}$ 与 $\hat{\beta}_1 \bar{X}$ 均是 $Y_1, \cdots, Y_n$ 的线性组合, 故 $\hat{\beta}_0$ 也是 $Y_1, \cdots, Y_n$ 的线性组合. 当然, 也计算得

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n}\sum_{i=1}^{n} Y_i - A\bar{X}\sum_{i=1}^{n} B_i Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - A\bar{X}B_i\right) Y_i,$$

这在之后的定理的证明中是需要的.

下面我们使用向量符号来书写, 与求和符号相比, 向量更能够体现出 OLS 估计量的几何性质, 这对于之后理解多元线性回归模型的 OLS 估计量中很有帮助.

记离差向量 $\boxed{\boldsymbol{b}(\boldsymbol{X}) = (B_1, \cdots, B_n)^{\mathrm{T}}}$ 和 $\mathbf{1} = (1, \cdots, 1)^{\mathrm{T}}$, 则有 $\boldsymbol{b}^{\mathrm{T}}\mathbf{1} = 0$, 即向量 $\boldsymbol{b}$ 同 $\mathbf{1}$ 正交. 记 $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$. 对于 OLS 估计量 $\hat{\beta}_1$, 其用向量乘法可以表示为

$$\hat{\beta}_1 = A\boldsymbol{b}^{\mathrm{T}}\left(\boldsymbol{Y} - \bar{Y}\cdot\mathbf{1}\right) = A\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y} - A\bar{Y}\left(\boldsymbol{b}^{\mathrm{T}}\mathbf{1}\right) = A\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y},$$

而 $A\boldsymbol{b}$ 的元素不全为 0, 故 $\hat{\beta}_0$ 是 $Y_1, \cdots, Y_n$ 的线性组合.$\hat{\beta}_0$ 也是 $Y_1, \cdots, Y_n$ 的线性组合的证明同上, 但我们还是可以计算 $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = \bar{Y} - A\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}\bar{X} = n^{-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y} - A\bar{X}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y} = \left(n^{-1}\mathbf{1} - A\bar{X}\boldsymbol{b}\right)^{\mathrm{T}}\boldsymbol{Y}$, 后续的定理的证明需要此. □

定理 2.4.2. *模型 (2.2.3) 的 OLS 估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是参数 β_0 和 β_1 的无偏估计量.*

证明. 同样地, 我们先使用求和符号书写证明, 然后再给出向量证明的形式.

首先使用求和符号书写. 为了证明 OLS 估计量是无偏的, 除了直接计算期望 $\mathbb{E}\hat{\beta}$, 另一种方法便是使用迭代期望定律, 通常而言使用后者更简单且能够得到条件期望. 对于

$\hat{\beta}_1$, 计算条件期望有

$$\begin{aligned}\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)&=\mathbb{E}\left(A\left(\boldsymbol{X}\right)\sum_{i=1}^{n}B_i\left(\boldsymbol{X}\right)Y_i\middle|\boldsymbol{X}\right)\\&=A\left(\boldsymbol{X}\right)\mathbb{E}\left(\sum_{i=1}^{n}B_i\left(\boldsymbol{X}\right)\left(\beta_0+\beta_1X_i+\varepsilon_i\right)\middle|\boldsymbol{X}\right)\\&=A\beta_0\sum_{i=1}^{n}B_i+A\beta_1\sum_{i=1}^{n}B_iX_i+A\mathbb{E}\left(\sum_{i=1}^{n}B_i\varepsilon_i\middle|\boldsymbol{X}\right)\\&=A\beta_1\sum_{i=1}^{n}B_iX_i+A\mathbb{E}\left(\sum_{i=1}^{n}B_i\varepsilon_i\middle|\boldsymbol{X}\right),\end{aligned}$$

对于第一项, 回顾数理统计中样本方差的计算可知

$$A\sum_{i=1}^{n}B_iX_i=A\sum_{i=1}^{n}\left(X_i-\bar{X}\right)X_i=A\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)=\frac{\sum\limits_{i=1}^{n}X_i^2-n\bar{X}^2}{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2}=1,$$

而第二项有

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{n}B_i\varepsilon_i\middle|\boldsymbol{X}\right)&=\mathbb{E}\left(\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\varepsilon_i\middle|\boldsymbol{X}\right)=\sum_{i=1}^{n}\mathbb{E}\left(\left(X_i-\bar{X}\right)\varepsilon_i\middle|\boldsymbol{X}\right)\\&=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\mathbb{E}\left(\varepsilon_i\middle|\boldsymbol{X}\right)=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\cdot0=0,\end{aligned}$$

故 $\hat{\beta}_0$ 的条件期望为 $\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)=\beta_1\cdot1+A\cdot0=\beta_1$. 依据定理 (2.1.1) 知 $\mathbb{E}\left(\hat{\beta}_1\right)=\mathbb{E}\left(\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)\right)=\beta_1$, 即 OLS 估计量 $\hat{\beta}_1$ 是 β_1 的无偏估计.

对于 $\hat{\beta}_0$, 其条件期望为

$$\mathbb{E}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\bar{Y}-\hat{\beta}_1\bar{X}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\bar{Y}\middle|\boldsymbol{X}\right)-\bar{X}\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)=\bar{Y}-\bar{X}\beta_1,$$

同理得

$$\begin{aligned}\mathbb{E}\hat{\beta}_0&=\mathbb{E}\left(\mathbb{E}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)\right)=\mathbb{E}\left(\bar{Y}-\bar{X}\beta_1\right)\\&=\mathbb{E}\bar{Y}-\beta_1\mathbb{E}\bar{X}=\mathbb{E}Y-\beta_1\mathbb{E}X=\beta_0,\end{aligned}$$

即 $\hat{\beta}_0$ 是参数 β_0 的无偏估计.

现在使用向量来证明 OLS 估计量的无偏性. 依据定理2.4.1的证明结果, 对于 $\hat{\beta}_1=A\left(\boldsymbol{X}\right)\boldsymbol{b}\left(\boldsymbol{X}\right)^{\mathrm{T}}\boldsymbol{Y}$ 求条件期望有

$$\begin{aligned}&\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)=\mathbb{E}\left(A\left(\boldsymbol{X}\right)\boldsymbol{b}\left(\boldsymbol{X}\right)^{\mathrm{T}}\boldsymbol{Y}\middle|\boldsymbol{X}\right)\xlongequal{\text{向量乘法是线性变换}}A\boldsymbol{b}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{Y}\middle|\boldsymbol{X}\right)\\=&A\boldsymbol{b}^{\mathrm{T}}\left(\mathbf{1}\beta_0+\boldsymbol{X}\beta_1\right)=A\beta_0\boldsymbol{b}^{\mathrm{T}}\mathbf{1}+A\beta_1\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}=A\beta_0\cdot0+A\beta_1\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}=A\beta_1\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X},\end{aligned}$$

而 $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)X_i=A^{-1}$, 故有 $\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)=\beta_1$. 对于同理可证. 故 $\hat{\beta}_0,\hat{\beta}_1$ 分别是 β_0,β_1 的无偏估计. □

定理 2.4.3 (一元线性回归模型的 Gauss-Markov 定理,Gauss-Markov theorem)**.** *设模型 (2.4.1) 的 OLS 估计量为 $\hat{\beta}_0$ 和 $\hat{\beta}_1$, 且满足有*同方差假设$\mathbb{E}\left(\varepsilon^2|\boldsymbol{X}\right)=\sigma^2$. *对于模型 (2.4.1) 的任意无偏估计量 $\tilde{\beta}_0$ 和 $\tilde{\beta}_1$, 若 $\tilde{\beta}_0$ 和 $\tilde{\beta}_1$ 是 $Y_1,\cdots,Y_n$ 的线性组合, 则有*

$$\operatorname{Var}\left(\tilde{\beta}_0\middle|\boldsymbol{X}\right)\geqslant\operatorname{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right),\quad\operatorname{Var}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)\geqslant\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right).$$

证明. 我们首先需要计算 OLS 估计量的条件方差. 对于 $\hat{\beta}_1$ 有

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)&=\operatorname{Var}\left(A\left(\boldsymbol{X}\right)\sum_{i=1}^{n}B_i\left(\boldsymbol{X}\right)Y_i\middle|\boldsymbol{X}\right)=A^2\operatorname{Var}\left(\sum_{i=1}^{n}B_i\left(\boldsymbol{X}\right)Y_i\middle|\boldsymbol{X}\right)\\&\xlongequal{\text{样本 i.i.d}}A^2\sum_{i=1}^{n}\operatorname{Var}\left(B_i\left(\boldsymbol{X}\right)Y_i|\boldsymbol{X}\right)=A^2\sum_{i=1}^{n}B_i^2\operatorname{Var}\left(Y_i|\boldsymbol{X}\right)\\&=A^2\sum_{i=1}^{n}B_i^2\mathbb{E}\left(\varepsilon_i^2\middle|X_i\right),\end{aligned}$$

而在同方差假设下 $\mathbb{E}\left(\varepsilon_i^2|X_i\right)=\sigma^2$, 则有

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)&=A^2\sum_{i=1}^{n}B_i^2\sigma^2=\sigma^2A^2\sum_{i=1}^{n}B_i^2=\sigma^2A^2\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\\&=\sigma^2A^2\cdot A^{-1}=\boxed{A\sigma^2}=\boxed{\frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}}.\end{aligned}$$

对于 $\hat{\beta}_0$, 由定理 (2.4.1) 的证明有

$$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}=\frac{1}{n}\sum_{i=1}^{n}Y_i-A\bar{X}\sum_{i=1}^{n}B_iY_i=\sum_{i=1}^{n}\left(\frac{1}{n}-A\bar{X}B_i\right)Y_i,$$

则 $\hat{\beta}_0$ 的条件方差为

$$\begin{aligned}&\operatorname{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)\\=&\operatorname{Var}\left(\sum_{i=1}^{n}\left(\frac{1}{n}-A\bar{X}B_i\right)Y_i\middle|\boldsymbol{X}\right)\xlongequal{\text{样本 i.i.d.}}\sum_{i=1}^{n}\operatorname{Var}\left(\left(\frac{1}{n}-A\bar{X}B_i\right)Y_i\middle|\boldsymbol{X}\right)\\=&\sum_{i=1}^{n}\left(\frac{1}{n}-A\bar{X}B_i\right)^2\operatorname{Var}\left(Y_i|\boldsymbol{X}\right)=\sum_{i=1}^{n}\left(\frac{1}{n}-A\bar{X}B_i\right)^2\mathbb{E}\left(\varepsilon_i^2\middle|X_i\right),\end{aligned}$$

而同方差假设 $\mathbb{E}\left(\varepsilon_i^2 \middle| X_i\right)=\sigma^2$ 下, 则有

$$\begin{aligned}
&\operatorname{Var}\left(\hat{\beta}_0 \middle| \boldsymbol{X}\right)=\sum_{i=1}^n\left(\frac{1}{n}-A\bar{X}B_i\right)^2\sigma^2=\sigma^2\sum_{i=1}^n\left(\frac{1}{n}-A\bar{X}B_i\right)^2\\
=&\sigma^2\sum_{i=1}^n\left(\frac{1}{n^2}-\frac{2}{n}A\bar{X}B_i+A^2\bar{X}^2B_i^2\right)=\sigma^2\left(\frac{1}{n}-\frac{2}{n}A\bar{X}\sum_{i=1}^n B_i+A^2\bar{X}^2\sum_{i=1}^n B_i^2\right)\\
=&\sigma^2\left(\frac{1}{n}-\frac{2}{n}A\bar{X}\cdot 0+A^2\bar{X}^2A^{-1}\right)=\boxed{\sigma^2\left(\frac{1}{n}+A\bar{X}^2\right)}.
\end{aligned}$$

这也可以化简得

$$\begin{aligned}
&\operatorname{Var}\left(\hat{\beta}_0 \middle| \boldsymbol{X}\right)=\sigma^2\left(\frac{1}{n}+A\bar{X}^2\right)\\
=&\sigma^2\left(\frac{1}{n}+\frac{\left(\sum_{i=1}^n X_i\right)^2}{n^2\sum_{i=1}^n\left(X_i-\bar{X}\right)}\right)=\sigma^2\cdot\frac{n\sum_{i=1}^n\left(X_i-\bar{X}\right)^2+\left(\sum_{i=1}^n X_i\right)^2}{n^2\sum_{i=1}^n\left(X_i-\bar{X}\right)^2}\\
=&\sigma^2\cdot\frac{n\left(\sum_{i=1}^n X_i^2-\frac{1}{n}\left(\sum_{i=1}^n X_i^2\right)^2\right)+\left(\sum_{i=1}^n X_i\right)^2}{n^2\sum_{i=1}^n\left(X_i-\bar{X}\right)^2}=\boxed{\sigma^2\cdot\frac{\sum_{i=1}^n X_i^2}{n\sum_{i=1}^n\left(X_i-\bar{X}\right)^2}}.
\end{aligned}$$

这样我们便求出了 OLS 估计量 $\hat{\beta}_0$ 以及 $\hat{\beta}_1$ 的条件方差分别为

$$\operatorname{Var}\left(\hat{\beta}_1 \middle| \boldsymbol{X}\right)=A\sigma^2=\frac{\sigma^2}{\sum_{i=1}^n\left(X_i-\bar{X}\right)^2}, \tag{2.26}$$

$$\operatorname{Var}\left(\hat{\beta}_0 \middle| \boldsymbol{X}\right)=\sigma^2\left(\frac{1}{n}+A\bar{X}^2\right)=\frac{\sigma^2\sum_{i=1}^n X_i^2}{n\sum_{i=1}^n\left(X_i-\bar{X}\right)^2}. \tag{2.27}$$

现考虑模型 (2.4.1) 的任意无偏估计量 $\tilde{\beta}_0$ 和 $\tilde{\beta}_1$, 其满足有

$$\tilde{\beta}_0=\bar{Y}-\tilde{\beta}_1\bar{X},\quad \tilde{\beta}_1=\sum_{i=1}^n C_i\left(\boldsymbol{X}\right)Y_i=\boldsymbol{c}^{\mathrm{T}}\boldsymbol{Y},$$

其中列向量 $\boldsymbol{c}\left(\boldsymbol{X}\right)=\left(C_1\left(\boldsymbol{X}\right),\cdots,C_n\left(\boldsymbol{X}\right)\right)^{\mathrm{T}}$ 的元素不全为 0. 无偏估计量 $\tilde{\beta}_0$ 和 $\tilde{\beta}_1$ 的

条件期望有

$$\begin{aligned}\mathbb{E}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)=&\mathbb{E}\left(\sum_{i=1}^{n}C_i\left(\boldsymbol{X}\right)Y_i\middle|\boldsymbol{X}\right)=\sum_{i=1}^{n}\mathbb{E}\left(C_i\left(\boldsymbol{X}\right)Y_i|\boldsymbol{X}\right)\\=&\sum_{i=1}^{n}C_i\mathbb{E}\left(Y_i|\boldsymbol{X}\right)=\boldsymbol{c}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{X}\right)=\boldsymbol{c}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{X}\right)\\=&\boldsymbol{c}^{\mathrm{T}}\left(\mathbf{1}\beta_0+\boldsymbol{X}\beta_1\right)=\beta_0\boldsymbol{c}^{\mathrm{T}}\mathbf{1}+\beta_1\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}\end{aligned}$$

及

$$\mathbb{E}\left(\tilde{\beta}_0\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\bar{Y}-\tilde{\beta}_1\bar{X}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\bar{Y}|\boldsymbol{X}\right)-\bar{X}\mathbb{E}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)=\bar{Y}-\bar{X}\tilde{\beta}_1,$$

则只需有 $\mathbb{E}\left(\mathbb{E}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)\right)=\beta_1$ 便可以保证 $\tilde{\beta}_0$ 和 $\tilde{\beta}_1$ 为参数 β_0 和 β_1 的无偏估计, 因而依据

$$\mathbb{E}\tilde{\beta}_1=\mathbb{E}\left(\mathbb{E}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)\right)=\mathbb{E}\left(\beta_0\boldsymbol{c}^{\mathrm{T}}\mathbf{1}+\beta_1\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}\right)=\beta_0\mathbb{E}\left(\boldsymbol{c}^{\mathrm{T}}\mathbf{1}\right)+\beta_1\mathbb{E}\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}\right)=\beta_1$$

成立待定系数得到必要条件[24]

$$\begin{cases}\mathbb{E}\left(\boldsymbol{c}^{\mathrm{T}}\mathbf{1}\right)=0,\\\mathbb{E}\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}\right)=1,\end{cases}\quad\Leftarrow\boxed{\begin{cases}\boldsymbol{c}^{\mathrm{T}}\mathbf{1}=0,\\\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}=1.\end{cases}}\tag{2.28}$$

现计算任意线性无偏估计量 $\tilde{\beta}_1$ 与 $\tilde{\beta}_0$ 的条件方差为

$$\begin{aligned}\mathrm{Var}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)=&\mathrm{Var}\left(\sum_{i=1}^{n}C_i\left(\boldsymbol{X}\right)Y_i\middle|\boldsymbol{X}\right)\overset{\text{样本 i.i.d.}}{=\!=\!=}\sum_{i=1}^{n}\mathrm{Var}\left(C_i\left(\boldsymbol{X}\right)Y_i|\boldsymbol{X}\right)\\=&\sum_{i=1}^{n}C_i^2\mathrm{Var}\left(Y_i|\boldsymbol{X}\right)=\sum_{i=1}^{n}C_i^2\mathbb{E}\left(\varepsilon_i^2\middle|X_i\right)\end{aligned}$$

和

$$\begin{aligned}&\mathrm{Var}\left(\tilde{\beta}_0\middle|\boldsymbol{X}\right)=\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i-\bar{X}\sum_{i=1}^{n}C_i\left(\boldsymbol{X}\right)Y_i\middle|\boldsymbol{X}\right)\\=&\mathrm{Var}\left(\sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}C_i\left(\boldsymbol{X}\right)\right)Y_i\middle|\boldsymbol{X}\right)\overset{\text{样本 i.i.d.}}{=\!=\!=}\sum_{i=1}^{n}\mathrm{Var}\left(\left(\frac{1}{n}-\bar{X}C_i\left(\boldsymbol{X}\right)\right)Y_i\middle|\boldsymbol{X}\right)\\=&\sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}C_i\right)^2\mathrm{Var}\left(Y_i|\boldsymbol{X}\right)=\sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}C_i\right)^2\mathrm{Var}\left(Y_i|\boldsymbol{X}\right)\\=&\sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}C_i\right)^2\mathbb{E}\left(\varepsilon_i^2\middle|X_i\right),\end{aligned}$$

[24]严格的来说, 式 (2.28) 的必要条件只有 $\begin{cases}\mathbb{E}\left(\boldsymbol{C}^{\mathrm{T}}\mathbf{1}\right)=0,\\\mathbb{E}\left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{X}\right)=1,\end{cases}$, 这时利用定理 (2.1.6) 我们可以证明非条件方差形式的 Guass-Markov 定理.

同方差假设有 $\mathbb{E}\left(\varepsilon_i^2 \middle| X_i\right)=\sigma^2$, 而

$$\sum_{i=1}^n\left(\frac{1}{n}-\bar{X}C_i\right)^2=\frac{1}{n}-\frac{2\bar{X}}{n}\sum_{i=1}^n C_i+\bar{X}^2\sum_{i=1}^n C_i^2=\frac{1}{n}+\bar{X}^2\sum_{i=1}^n C_i^2,$$

于是 $\tilde{\beta}_1$ 与 $\tilde{\beta}_0$ 的条件方差分别为

$$\operatorname{Var}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)=\sigma^2\sum_{i=1}^n C_i^2, \tag{2.29}$$

$$\operatorname{Var}\left(\tilde{\beta}_0\middle|\boldsymbol{X}\right)=\sigma^2\left(\frac{1}{n}+\bar{X}^2\sum_{i=1}^n C_i^2\right). \tag{2.30}$$

现在我们来比较 OLS 估计量与任意线性无偏估计量的非条件方差, 将式 (2.29) 减去式 (2.26)、式 (2.30) 减去式 (2.27) 得

$$\operatorname{Var}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)-\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)=\sigma^2\sum_{i=1}^n C_i^2-\sigma^2 A=\sigma^2\left(\sum_{i=1}^n C_i^2-A\right)$$

与

$$\begin{aligned}\operatorname{Var}\left(\tilde{\beta}_0\middle|\boldsymbol{X}\right)-\operatorname{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)=&\sigma^2\left(\frac{1}{n}+\bar{X}^2\sum_{i=1}^n C_i^2\right)-\sigma^2\left(\frac{1}{n}+A\bar{X}^2\right)\\=&\sigma^2\bar{X}^2\left(\sum_{i=1}^n C_i^2-A\right),\end{aligned}$$

这样, 我们只需要研究 $\sum\limits_{i=1}^n C_i^2-A$. 由 $A^{-1}=\sum\limits_{i=1}^n\left(X_i-\bar{X}\right)^2=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}$,$B_i=X_i-\bar{X}$, 不难注意到式 (2.28) 有

$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{b}=\boldsymbol{c}^{\mathrm{T}}\left(\boldsymbol{X}-\mathbf{1}\bar{X}\right)=\boldsymbol{C}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{C}^{\mathrm{T}}\mathbf{1}\cdot\bar{X}=1,$$

那么

$$\begin{aligned}\sum_{i=1}^n C_i^2-A=&\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c}-A=\left(\boldsymbol{c}-A\boldsymbol{b}+A\boldsymbol{b}\right)^{\mathrm{T}}\left(\boldsymbol{c}-A\boldsymbol{b}+A\boldsymbol{b}\right)-A\\&=\left(\boldsymbol{c}-A\boldsymbol{b}\right)^{\mathrm{T}}\left(\boldsymbol{c}-A\boldsymbol{b}\right)+2\left(\boldsymbol{c}-A\boldsymbol{b}\right)^{\mathrm{T}}A\boldsymbol{b}+\left(A\boldsymbol{b}\right)^{\mathrm{T}}\left(A\boldsymbol{b}\right)-A\\&=\left(\boldsymbol{c}-A\boldsymbol{b}\right)^{\mathrm{T}}\left(\boldsymbol{c}-A\boldsymbol{b}\right)+2A\boldsymbol{c}^{\mathrm{T}}\boldsymbol{b}-2\left(A\boldsymbol{b}\right)^{\mathrm{T}}A\boldsymbol{b}+\left(A\boldsymbol{b}\right)^{\mathrm{T}}\left(A\boldsymbol{b}\right)-A\\&=\left(\boldsymbol{c}-A\boldsymbol{b}\right)^{\mathrm{T}}\left(\boldsymbol{c}-A\boldsymbol{b}\right)+2A-A^2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}-A\\&=\left(\boldsymbol{c}-A\boldsymbol{b}\right)^{\mathrm{T}}\left(\boldsymbol{c}-A\boldsymbol{b}\right)+2A-A^2A^{-1}-A\\&=\left(\boldsymbol{c}-A\boldsymbol{b}\right)^{\mathrm{T}}\left(\boldsymbol{c}-A\boldsymbol{b}\right)\geqslant 0,\end{aligned}$$

当且仅当 $\boldsymbol{c}=A\boldsymbol{b}$ 即 $\tilde{\beta}$ 为 OLS 估计量时取等, 故有

$$\operatorname{Var}\left(\tilde{\beta}_0\middle|\boldsymbol{X}\right)\geqslant\operatorname{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right),\quad\operatorname{Var}\left(\tilde{\beta}_1\middle|\boldsymbol{X}\right)\geqslant\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right).$$

□

依据定理 (2.4.3),OLS 估计量又被称为**最优线性无偏估计量** (best linear unbiased estimator), 简称为 **BLUE**, 该性质可以说是 OLS 估计量最为重要的性质之一, 其意味着对于一元线性回归模型 (2.4.1), 从 MSE 的角度出发, 其他任何**线性**估计量均没有 OLS 估计量有效. 一个自然的问题便是是否存在模型 (2.4.1) 的非线性估计量, 其比 OLS 估计量更为有效? 长久以来, 这是一个开放问题 (open problem), 但 Bruce 在 2021 年证明了现代的 Gauss-Markov 定理:OLS 估计量是线性回归模型的 MVU 估计. 也因此,OLS 估计量又可以被称为**最优无偏估计量** (best unbiased estimator), 简称为 **BUE**.

下面我们利用 Monte Carlo 模拟, 进一步说明 OLS 估计量的有限性质, 代码见本章最后一节. 假设被解释变量 Y 与随机变量 X 具有线性条件期望函数 $E(Y|X) = 10x+25$, 数据生成过程为 $Y = E(Y|X) + e$, 其中 e 服从正态分布 $N(0, 5^2)$. 现在我们用 OLS 去回归方程 $Y_1 = \beta_0 + \beta_1 X_i + e_i$ 的参数, 我们进行重复抽样 $k = 1000$ 次, 每次抽样的样本容量为 $n = 50$, 抽样方法则是有放回的简单随机抽样. 这样, 我们每抽样一次便都能得到由该样本所决定的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$.

定理 (2.4.2) 说明了 OLS 估计量是参数 β 的无偏估计, 而大数定律则指出对于任意随机变量 X 的的样本均值 $\bar{X}$, 其依概率收敛于非条件期望 $\mathbb{E}X$, 这意味着为了估计 $\mathbb{E}X$ 我们可以通过**反复抽样求平均值**来消除概率抽样所导致的误差, 尽管不能完全消除. 而 OLS 估计量 $\hat{\beta}_{\text{OLS}}$ 也是由样本所决定的随机变量, 其无偏性便表明为了得到 β 的准确估计, 重复抽样计算均值便可以提升估计的准确性.

图 (2.1) 和图 (2.3) 绘制了第 k 次抽样后计算的 $\hat{\beta}_0$ 以及 $\hat{\beta}_1$ 的样本均值. 可以发现, 与设定的 $E(Y|X) = 10x + 25$ 的真实值相比较, 首次抽样计算的 OLS 估计量与真实值还是有较大的差距, 但随着抽样次数 k 的增大, 估计量的样本均值便趋于真实值. 图 (2.1) 中 $\hat{\beta}_0$ 的样本均值大约在 300 次抽样后便稳定在真实值 $\beta_0 = 25$ 附近, 而图 (2.3) 中 $\hat{\beta}_1$ 的样本均值大约在 150 次抽样后对于真实值 $\beta_1 = 10$ 的偏离情况非常小了.

图 (2.2) 和图 (2.4) 则是绘制了 $k = 1000$ 次重复抽样后 OLS 估计量的直方图以及拟合的概率密度函数, 图中的蓝色竖线为对应的真实值. 不难发现, 蓝线与拟合的概率密度函数的期望值是非常接近的. 为了进一步地感受这点, 图 (2.5) 和图 (2.6) 这是依次绘制了 $k = 25$ 次、$k = 100$ 次以及 $k = 1000$ 次下拟合的 OLS 估计量的概率密度函数, 可以发现, 随着 K 的增大, 拟合的 OLS 估计量的概率密度函数最终其期望趋向于真实值[25]. 这便通过 Monte Carlo 模拟验证了 OLS 估计量的无偏性.

[25]这里绘制的其实是**样本均值的分布**. 对于分布, 中心极限定理说明了样本均值的分布依概率收敛于正态分布, 同时大数定律说明了样本均值依概率收敛于期望.

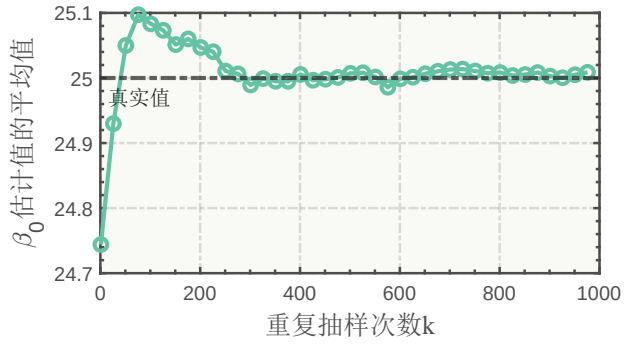

图 2.1: $\hat{\beta}_0$ 的样本均值

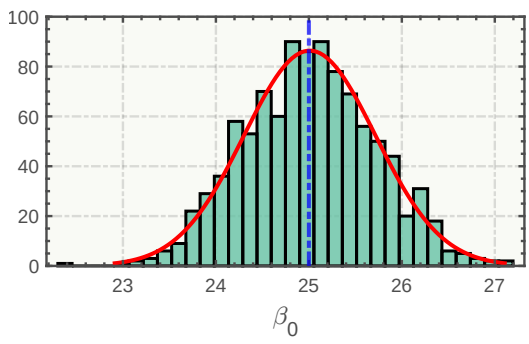

图 2.2: $\hat{\beta}_0$ 的直方图

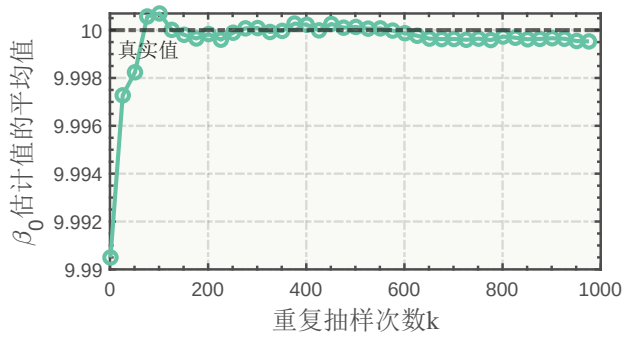

图 2.3: $\hat{\beta}_1$ 的样本均值

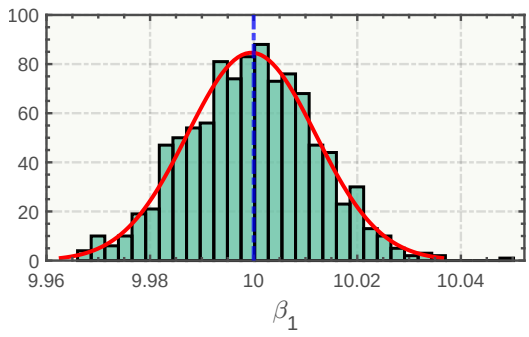

图 2.4: $\hat{\beta}_1$ 的直方图

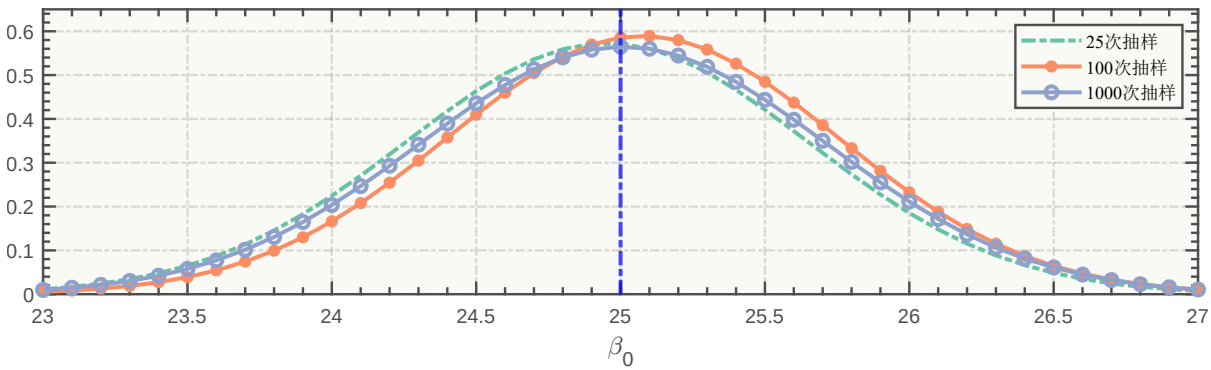

图 2.5: 重复抽样下 $\hat{\beta}_0$ 的概率密度函数

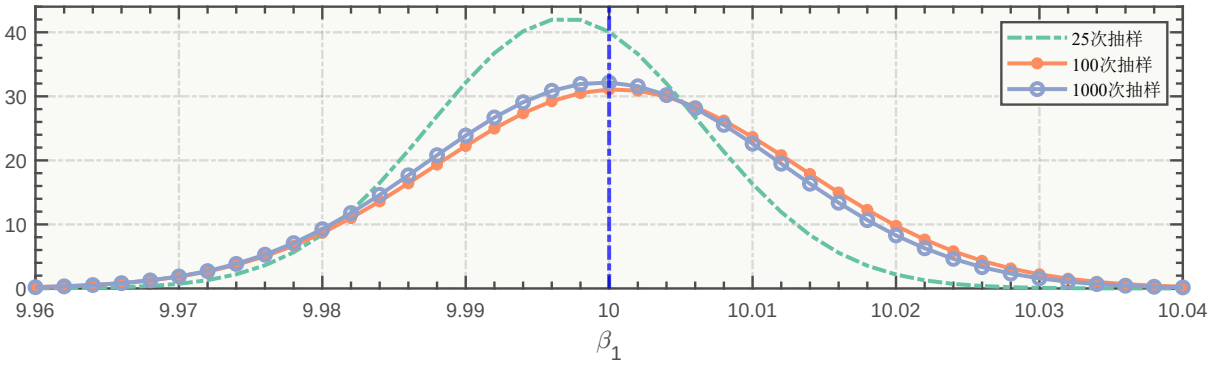

图 2.6: 重复抽样下 $\hat{\beta}_1$ 的概率密度函数

2.5 一元 OLS 估计量下的拟合值、残差

本节我们介绍由 OLS 估计量所导出的拟合值以及残差, 并介绍一元线性回归模型下的投影矩阵 (projection matrix) 与归零矩阵 (annihilator matrix), 并借此研究拟合值和残差具有有限样本性质.

依据普通最小二乘法, 我们同样可以得到模型 (2.4.1) 的 OLS 估计量, 这样我们使用 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 可以得到模型 (2.2.3) 的其他部分的估计量.

定义 2.5.1 (拟合值, 残差). *对于模型 (2.4.1), 设其参数估计为 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^{\mathrm{T}}$, 则记 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ 为 Y_i 的* **拟合值** *(fitted value), 拟合值 $\hat{Y}_i$ 为 Y_i 的估计量. 记 $\hat{\varepsilon}_i = \hat{Y} - Y_i$ 为* **残差** *(residual), 残差 $\hat{\varepsilon}_i$ 为回归误差 $\varepsilon_i = Y_i - \mathbb{E}(Y|X_i)$ 的估计量.*

我们使用向量符号, 记全部样本个体的拟合值、残差分别为拟合值向量 $\hat{\boldsymbol{Y}} = \left(\hat{Y}_1, \cdots, \hat{Y}_n\right)^{\mathrm{T}}$ 和残差向量 $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n)^{\mathrm{T}}$.

由 $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ 及 $\hat{\beta}_1 = A \sum_{i=1}^{n} B_i Y_i = A\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}$, 我们现研究 OLS 估计量下拟合值的表达式. 对于单独的样本个体, 其拟合值 $\hat{Y}_i$ 有

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i = \bar{Y} + \hat{\beta}_1 \left(X_i - \bar{X}\right) = \bar{Y} + \hat{\beta}_1 B_i \\ &= \frac{1}{n}\sum_{j=1}^{n} Y_j - AB_i \sum_{j=1}^{n} B_j Y_j = \sum_{j=1}^{n}\left(\frac{1}{n} - AB_i B_j\right) Y_j,\end{aligned}$$

其中的 B_i 是上一节用以表示离差的符号, 可见 $\hat{Y}_i$ 的形式谈不上直观. 但我们考虑拟合值向量 $\hat{\boldsymbol{Y}}$, 同样变形有

$$\begin{aligned}\hat{\boldsymbol{Y}} &= \mathbf{1}\hat{\beta}_0 + \boldsymbol{X}\hat{\beta}_1 = \mathbf{1}\bar{Y} + \left(\boldsymbol{X} - \mathbf{1}\bar{X}\right)\hat{\beta}_1 = \mathbf{1}\bar{Y} + \boldsymbol{b}\hat{\beta}_1 \\ &= \frac{1}{n}\mathbf{1}\left(\mathbf{1}^{\mathrm{T}}\boldsymbol{Y}\right) + \boldsymbol{b}A\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y} = \left(\frac{1}{n}\left(\mathbf{1}\mathbf{1}^{\mathrm{T}}\right) + A\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}\right)\boldsymbol{Y},\end{aligned}$$

注意到 $n=\mathbf{1}^{\mathrm{T}}\mathbf{1}, A^{-1}=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}$, 这里的 $\boldsymbol{b}$ 为上一节的离差向量, 于是利用矩阵的分块运算有

$$\begin{aligned}\hat{\boldsymbol{Y}}&=\left(\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1}\left(\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)+\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}\right)\boldsymbol{Y}\overset{\text{内积视为1×1矩阵}}{=\!=\!=}\left(\mathbf{1}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1}\mathbf{1}^{\mathrm{T}}+\boldsymbol{b}\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\boldsymbol{b}^{\mathrm{T}}\right)\boldsymbol{Y}\\&=\left(\mathbf{1}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1},\boldsymbol{b}\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\right)\begin{pmatrix}\mathbf{1}^{\mathrm{T}}\\\boldsymbol{b}^{\mathrm{T}}\end{pmatrix}\boldsymbol{Y}=(\mathbf{1},\boldsymbol{b})\begin{pmatrix}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1}&0\\0&\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{1}^{\mathrm{T}}\\\boldsymbol{b}^{\mathrm{T}}\end{pmatrix}\boldsymbol{Y},\end{aligned}$$

不难发现, 中间的矩阵 $\begin{pmatrix}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1}&0\\0&\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\end{pmatrix}$ 即为矩阵 $\begin{pmatrix}\mathbf{1}^{\mathrm{T}}\\\boldsymbol{b}^{\mathrm{T}}\end{pmatrix}(\mathbf{1},\boldsymbol{b})$ 的逆, 这是因为

$$\begin{aligned}\left(\begin{pmatrix}\mathbf{1}^{\mathrm{T}}\\\boldsymbol{b}^{\mathrm{T}}\end{pmatrix}(\mathbf{1},\boldsymbol{b})\right)^{-1}&=\begin{pmatrix}\mathbf{1}^{\mathrm{T}}\mathbf{1}&\mathbf{1}^{\mathrm{T}}\boldsymbol{b}\\\boldsymbol{b}^{\mathrm{T}}\mathbf{1}&\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\end{pmatrix}^{-1}\\&=\begin{pmatrix}\mathbf{1}^{\mathrm{T}}\mathbf{1}&0\\0&\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\end{pmatrix}^{-1}\overset{\text{对角矩阵的逆}}{=\!=\!=}\begin{pmatrix}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1}&0\\0&\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\end{pmatrix}.\end{aligned}$$

这样, 拟合值向量 $\hat{\boldsymbol{Y}}$ 便具有了**对称性**的形式, 记

$$\boldsymbol{X}^{*}=(\mathbf{1},\boldsymbol{b})=(\mathbf{1},\boldsymbol{X})-(\mathbf{0},\mathbf{1})\bar{X}=\begin{pmatrix}1&X_1-\bar{X}\\\vdots&\vdots\\1&X_n-\bar{X}\end{pmatrix},$$

则矩阵 $\boldsymbol{X}^{*}$ 包含了符合式 (2.7) 形式的样本个体 (也即 demeaned regressor 的样本个体), 我们记n 阶方阵

$$\boldsymbol{P}=(1,\boldsymbol{b})_{n\times 2}\begin{pmatrix}\left(1^{\mathrm{T}}1\right)^{-1}&0\\0&\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\end{pmatrix}_{2\times 2}\begin{pmatrix}1^{\mathrm{T}}\\\boldsymbol{b}^{\mathrm{T}}\end{pmatrix}_{2\times n}=\boldsymbol{X}^{*}\left(\left(\boldsymbol{X}^{*}\right)^{\mathrm{T}}\boldsymbol{X}^{*}\right)^{-1}\left(\boldsymbol{X}^{*}\right)^{\mathrm{T}},\tag{2.31}$$

则称矩阵 $\boldsymbol{P}$ 为**投影矩阵** (projection matrix).

投影矩阵 $\boldsymbol{P}$ 的含义十分清晰,$\hat{\boldsymbol{Y}}=\boldsymbol{P}\boldsymbol{Y}$ 即将被解释变量 $\hat{\boldsymbol{Y}}$ 映射为 $\hat{\boldsymbol{Y}}$, 但其 “投影” 几何含义我们将在下一章介绍.

对于投影矩阵 $\boldsymbol{P}$, 其满足有如下性质.

定理 2.5.1 (一元 OLS 估计量的投影矩阵 $\boldsymbol{P}$ 的性质). *对于模型 (2.4.1), 则 OLS 估计量下投影矩阵 $\boldsymbol{P}=\boldsymbol{X}^{*}\left(\left(\boldsymbol{X}^{*}\right)^{\mathrm{T}}\boldsymbol{X}^{*}\right)^{-1}\left(\boldsymbol{X}^{*}\right)^{\mathrm{T}}$ 具有如下性质:*

(a) $\boldsymbol{P}\boldsymbol{X}^{*}=\boldsymbol{X}^{*},\boldsymbol{P}\boldsymbol{Y}=\hat{\boldsymbol{Y}}$;

(b) 投影矩阵 $\boldsymbol{P}$ 是对称矩阵, 即 $\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{P}$;

(c) 投影矩阵 $\boldsymbol{P}$ 为幂等矩阵, 即 $\boldsymbol{P}^{2}=\boldsymbol{P}$;

(d) 投影矩阵的 $\boldsymbol{P}$ 的迹为 *2;*

(e) 投影矩阵 $\boldsymbol{P}$ 有 2 个特征值为 1, 余下的 $n-2$ 个特征值为 *0;*

(f) 投影矩阵 $\boldsymbol{P}$ 的秩为 2.

证明. 我们仅证明性质 (a)(b)(c)(d), 余下的性质见下一章定理 (3.6.1) 的证明. 依据式 (2.31), 对于性质 (a) 直接计算有 $\boldsymbol{P}\boldsymbol{X}^* = \boldsymbol{X}^* \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{-1} \left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^* = \boldsymbol{X}^*$.

性质 (b) 计算得

$$\begin{aligned}\boldsymbol{P}^{\mathrm{T}} &= \left(\boldsymbol{X}^* \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{-1} \left(\boldsymbol{X}^*\right)^{\mathrm{T}}\right)^{\mathrm{T}} = \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}}\right)^{\mathrm{T}} \left(\left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{\mathrm{T}}\right)^{-1} \left(\boldsymbol{X}^*\right)^{\mathrm{T}} \\ &= \boldsymbol{X}^* \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{-1} \left(\boldsymbol{X}^*\right)^{\mathrm{T}} = \boldsymbol{P}.\end{aligned}$$

性质 (c) 计算得

$$\begin{aligned}\boldsymbol{P}^2 &= \boldsymbol{X}^* \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{-1} \left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^* \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{-1} \left(\boldsymbol{X}^*\right)^{\mathrm{T}} \\ &= \boldsymbol{X}^* \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{-1} \left(\boldsymbol{X}^*\right)^{\mathrm{T}} = \boldsymbol{P}.\end{aligned}$$

性质 (d) 计算得

$$\begin{aligned}\mathrm{tr}\boldsymbol{P} &= \mathrm{tr}\left(\boldsymbol{X}^* \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{-1} \left(\boldsymbol{X}^*\right)^{\mathrm{T}}\right) \xlongequal{\text{迹的性质}} \mathrm{tr}\left(\left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}} \boldsymbol{X}^*\right)^{-1} \boldsymbol{X}^* \left(\boldsymbol{X}^*\right)^{\mathrm{T}}\right) \\ &= \mathrm{tr}\left(\left[(\mathbf{1}, \boldsymbol{b})^{\mathrm{T}} (\mathbf{1}, \boldsymbol{b})\right]^{-1} (\mathbf{1}, \boldsymbol{b})^{\mathrm{T}} (\mathbf{1}, \boldsymbol{b})\right) = \mathrm{tr}\left(\boldsymbol{I}_2\right) = 2.\end{aligned}$$

□

我们现在研究残差. 对于样本个体而言, 其残差 $\hat{\varepsilon}_i$ 即为

$$\hat{\varepsilon}_i = \hat{Y}_i - Y_i = \sum_{j=1}^{n} \left(\frac{1}{n} - AB_iB_j\right) Y_j - Y_i,$$

单独来看依旧缺少对称性. 考虑残差向量 $\hat{\boldsymbol{\varepsilon}}$, 则

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = \left(\boldsymbol{I}_n - \boldsymbol{P}\right) \boldsymbol{Y},$$

我们记 n 阶方阵 $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P}$, 则称矩阵 $\boldsymbol{M}$ 为**归零矩阵** (annihilator matrix), 这样残差向量也具有**对称性**的形式 $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{Y}$. 对于归零矩阵 $\boldsymbol{M}$, 其具有投影矩阵 $\boldsymbol{P}$ 类似的代数性质.

定理 2.5.2 (一元 OLS 估计量的阵归零矩阵 $\boldsymbol{M}$ 的性质)**.** *对于模型 (2.4.1), 则 OLS 估计量下归零矩阵 $\boldsymbol{M}=\boldsymbol{I}_n-\boldsymbol{P}$ 具有如下性质:*

(a) $\boldsymbol{MX}^*=\boldsymbol{O}$,$\boldsymbol{MY}=\boldsymbol{M\varepsilon}=\hat{\boldsymbol{\varepsilon}}$;

(b) 归零矩阵 $\boldsymbol{M}$ 是对称矩阵, 即 $\boldsymbol{M}^{\mathrm{T}}=\boldsymbol{M}$;

(c) 归零矩阵 $\boldsymbol{P}$ 为幂等矩阵, 即 $\boldsymbol{M}^2=\boldsymbol{M}$;

(d) 归零矩阵的 $\boldsymbol{M}$ 的迹为 $n-2$.

证明. 这些性质都是简单的代数计算便可得的. 对于性质 (a), 即有 $\boldsymbol{MX}^*=(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{X}^*=\boldsymbol{X}^*-\boldsymbol{P}\cdot\boldsymbol{X}^*=\boldsymbol{X}^*-\boldsymbol{X}^*=\boldsymbol{O}_{n\times 2}$ 及 $\boldsymbol{MY}=\boldsymbol{M}(\boldsymbol{X}^*\boldsymbol{\beta}+\boldsymbol{\varepsilon})=\boldsymbol{MX}^*\boldsymbol{\beta}+\boldsymbol{M\varepsilon}=\boldsymbol{M\varepsilon}=\hat{\boldsymbol{Y}}$.

对于性质 (b), 则 $\boldsymbol{M}^{\mathrm{T}}=(\boldsymbol{I}_n-\boldsymbol{P})^{\mathrm{T}}=(\boldsymbol{I}_n-\boldsymbol{P})=\boldsymbol{M}$.

对于性质 (c), 即 $\boldsymbol{M}^2=(\boldsymbol{I}_n-\boldsymbol{P})(\boldsymbol{I}_n-\boldsymbol{P})=\boldsymbol{I}_n-2\boldsymbol{P}+\boldsymbol{P}=\boldsymbol{I}_n-\boldsymbol{P}=\boldsymbol{M}$.

对于性质 (d), 这是因为 $\mathrm{tr}\boldsymbol{M}=\mathrm{tr}(\boldsymbol{I}_n-\boldsymbol{P})=\mathrm{tr}\boldsymbol{I}_n-\mathrm{tr}\boldsymbol{P}=n-2$. □

有了投影矩阵 $\boldsymbol{P}$ 及归零矩阵 $\boldsymbol{M}$, 我们对拟合值 $\hat{\boldsymbol{Y}}$ 和残差 $\hat{\boldsymbol{\varepsilon}}$ 相关性质的研究将容易很多. 现在, 我们对回归方差进行分解. 由于

$$\boldsymbol{Y}=\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}},$$

则考虑被解释变量的样本平方和 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 有

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\left(\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}}\right)=\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}}+2\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}},$$

注意到

$$\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=(\boldsymbol{PY})^{\mathrm{T}}(\boldsymbol{MY})=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{PMY}=\boldsymbol{Y}^{\mathrm{T}}[\boldsymbol{P}(\boldsymbol{I}-\boldsymbol{P})]\boldsymbol{Y}=0,$$

这也即意味着拟合值 $\hat{\boldsymbol{Y}}$ 与残差 $\hat{\boldsymbol{\varepsilon}}$ 正交, 故样本平方和 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 可以被分解为

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}},\quad\Leftrightarrow\quad\sum_{i=1}^{n}Y_i^2=\sum_{i=1}^{n}\hat{Y}_i^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2,\tag{2.32}$$

式 (2.32) 称为**平方和分解**公式, 这个等式其仅仅利用了投影矩阵 $\boldsymbol{P}$ 和归零矩阵 $\boldsymbol{M}$ 的性质. 由于模型 (2.4.1) 含义常数项 β_0, 另一种更常用的分解公式是考虑

$$\boldsymbol{Y}-\mathbf{1}\bar{Y}=\left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)+\hat{\boldsymbol{\varepsilon}},$$

则离差平方和 $\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)$ 满足

$$\begin{aligned}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right) &= \left[\left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)+\hat{\boldsymbol{\varepsilon}}\right]^{\mathrm{T}}\left[\left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)+\hat{\boldsymbol{\varepsilon}}\right] \\ &= \left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)+2\left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} \\ &= \left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}},\end{aligned}$$

即得到了

$$\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)=\left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}}-\mathbf{1}\bar{Y}\right)+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}, \tag{2.33}$$

或

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2, \tag{2.34}$$

式 (2.33) 或 (2.34) 被称为普通最小二乘法的**方差分析** (analysis-of-variance). 需要指出, 为了得到 (2.33), 我们使用了含有截距项时残差和为 0 的性质, 则是正规方程组 (2.18) 所得到的.OLS 的方差分析, 最为广泛的引用便是定义了统计量拟合优度 R^2.

定义 2.5.2 (一元线性回归模型的拟合优度). *对于模型 (2.4.1), 依据式 (2.34), 我们称* $\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$ *为***总体平方和** *(total sum of square)*或**离差平方和** *(sum of squared deviations),*$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$ *称为***解释平方和** *(explained sum of square)*或**回归平方和** *(regression sum of square),*$\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ *为***残差平方和** *(residual sum of square), 记*

$$\mathrm{TSS}=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2,\quad \mathrm{ESS}=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2,\quad \mathrm{RSS}=\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2,$$

则定义统计量 $\boldsymbol{R}^2$ *为*

$$R^2=\frac{\mathrm{ESS}}{\mathrm{TSS}}=1-\frac{\mathrm{RSS}}{\mathrm{TSS}}=\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}=1-\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2},$$

*我们称 R^2 为***拟合优度** *(goodness of fit)*或**可决系数** *(coefficient of determination).*

对于定义 (2.5.2) 的拟合优度, 依据式 (2.34) 不难得知 $R^2\in[0,1]$. 拟合优度 R^2 通常被描述为“由最小二乘法拟合值 $\hat{Y}$ 所解释的Y 的样本方差的百分比”[26]. 需要指出, 拟合优度仅度量了样本回归函数 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$ 对于样本的拟合程度,R^2 数值越接近于

[26]Bruce(cite) 上原文:...the fraction of the sample variance of Y which is explained by the least squares-fit. R2 is a crude measure of regression fit.

1, 说明 RSS 越小, 即拟合值 $\hat{Y}$ 与样本观察值 Y 的偏差越小——这并不表明计量模型是否更具有合理性.

拟合优度的定义依赖于回归模型有常数项, 没有常数项时则式 (2.33) 或 (2.34) 不会成立. 此时可以使用式 (2.32) 来类似地定义非中心化的 R^2_{uc}, 除此之外, 更多的有关度量拟合的内容见第3.10节.

需要强调, 定义 (2.5.2) 中 RSS 是残差平方和, 在一些计量经济学教材中有记为误差平方和 SSE, 但本笔记不使用这个说法, 因为我们已经在定义中 (2.3.1) 将 $\sum\limits_{i=1}^{n} e_i^2 = \sum\limits_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$ 称为了 SSE, 其是回归误差 e 的平方和, 是不可观察的. 由于参数 $\boldsymbol{\beta}$ 是未知的, 我们时常变化 $\text{SSE}(\boldsymbol{\beta})$ 中的 $\boldsymbol{\beta}$, 因而 SSE 与 RSS 的关系为 $\boxed{\text{SSE}\left(\hat{\boldsymbol{\beta}}\right) = \text{RSS}}$.

本节的最后我们研究一下拟合值 $\hat{Y}_i$ 与残差 $\hat{\varepsilon}_i$ 的有限样本性质. 倘若使用投影矩阵 $\boldsymbol{P}$ 以及 $\boldsymbol{M}$, 则对于拟合值向量 $\hat{\boldsymbol{Y}}$ 及残差向量 $\hat{\varepsilon}$ 其条件期望、条件方差将有非常简洁的推导和最终结果, 但笔者将这些内容放在下一章多元线性回归分析里. 在下一章, 我们会发现拟合值和残差值的条件期望以及条件方差都依赖于投影矩阵 $\boldsymbol{P}$ 的对角线元素, 这些元素在统计学中被称为**杠杆值**.

本笔记现使用求和符号来推导拟合值 $\hat{Y}_i$ 与残差 $\hat{\varepsilon}_i$ 的条件期望、条件方差.

对于样本 (Y_i, X_i), 其拟合值 $\hat{Y}_i$ 就给定 $\boldsymbol{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$ 的条件期望为

$$\begin{aligned}\mathbb{E}\left(\hat{Y}_i \middle| \boldsymbol{X}\right) &= \mathbb{E}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i \middle| \boldsymbol{X}\right) = \mathbb{E}\left(\hat{\beta}_0 \middle| \boldsymbol{X}\right) + X_i \mathbb{E}\left(\hat{\beta}_1 \middle| \boldsymbol{X}\right) \\ &= \boxed{\beta_0 + \beta_1 X_i} = \mathbb{E}\left(Y \middle| X_i\right),\end{aligned}$$

故 $\mathbb{E}\hat{Y}_i = \mathbb{E}\left(\mathbb{E}\left(\hat{Y}_i \middle| \boldsymbol{X}\right)\right) = \mathbb{E}\left(\beta_0 + \beta_1 X_i\right) = \boxed{\beta_0 + \beta_1 \mathbb{E}X}$, 也就是说, 模型 (2.4.1) 的拟合值 $\hat{Y}_i$ 并不是样本个体 $Y_i = \beta_0 + \beta_1 X_i$ 的无偏估计. 除非解释变量 X 为常数, 即满足有 $X = X_i = \mathbb{E}X$ 时, 这时一元线性回归模型的拟合值便是对于被解释变量 Y_i 的无偏估计. 我们再考虑拟合值 $\hat{Y}_i$ 的条件方差, 则依据公式

$$\text{Var}\left(aX + bY \middle| Z\right) = a^2 \text{Var}\left(X \middle| Z\right) + b^2 \text{Var}\left(Y \middle| Z\right) + 2ab\text{Cov}\left(X, Y \middle| Z\right),$$

则有

$$\begin{aligned}\text{Var}\left(\hat{Y}_i \middle| \boldsymbol{X}\right) &= \text{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i \middle| \boldsymbol{X}\right) \\ &= \text{Var}\left(\hat{\beta}_0 \middle| \boldsymbol{X}\right) + X_i^2 \text{Var}\left(\hat{\beta}_1 \middle| \boldsymbol{X}\right) + 2X_i \text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1 \middle| \boldsymbol{X}\right),\end{aligned}$$

在定理 (2.4.3) 的证明中我们已经计算了 OLS 估计量的条件方差, 现在计算条件协方差 $\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1 \middle| \boldsymbol{X}\right)$. 由方差的分配律

$$\text{Cov}\left(aX_1 + bX_2, Y \middle| Z\right) = a\text{Cov}\left(X_1, Y \middle| Z\right) + b\text{Cov}\left(X_2, Y \middle| Z\right)$$

得

$$\begin{aligned}\operatorname{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\boldsymbol{X}\right)&=\operatorname{Cov}\left(\bar{Y}-\hat{\beta}_1\bar{X},\hat{\beta}_1\middle|\boldsymbol{X}\right)=\operatorname{Cov}\left(\bar{Y},\hat{\beta}_1\middle|\boldsymbol{X}\right)-\bar{X}\operatorname{Cov}\left(\hat{\beta}_1,\hat{\beta}_1\middle|\boldsymbol{X}\right)\\&=\operatorname{Cov}\left(\bar{Y},\hat{\beta}_1\middle|\boldsymbol{X}\right)-\bar{X}\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right),\end{aligned}$$

而依据定理 (2.4.3) 的证明中计算的 $\operatorname{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)=\left(\frac{1}{n}+A\bar{X}^2\right)\sigma^2$, 这样我们由

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)&=\operatorname{Var}\left(\bar{Y}-\hat{\beta}_1\bar{X}\middle|\boldsymbol{X}\right)\\&=\operatorname{Var}\left(\bar{Y}\middle|\boldsymbol{X}\right)+\bar{X}^2\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)-2\bar{X}\operatorname{Cov}\left(\bar{Y},\hat{\beta}_1\middle|\boldsymbol{X}\right)\end{aligned}$$

可以推导出

$$\begin{aligned}-2\bar{X}\operatorname{Cov}\left(\bar{Y},\hat{\beta}_1\middle|\boldsymbol{X}\right)&=\left(\frac{1}{n}+A\bar{X}^2\right)\sigma^2-\operatorname{Var}\left(\bar{Y}\middle|\boldsymbol{X}\right)+\bar{X}^2\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)\\&=\left(\frac{1}{n}+A\bar{X}^2\right)\sigma^2-\frac{1}{n}\sigma^2-\bar{X}^2A\sigma^2=0,\end{aligned}$$

即样本均值 $\bar{Y}$ 与 OLS 估计量 $\hat{\beta}_1$ 完全不相关, 故

$$\boxed{\operatorname{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\boldsymbol{X}\right)=-\bar{X}\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)=-A\bar{X}\sigma^2}.$$

回到拟合值 $\hat{Y}_i$ 的条件方差的计算, 于是

$$\begin{aligned}\operatorname{Var}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)&=\operatorname{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)+X_i^2\operatorname{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)+2X_i\operatorname{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\boldsymbol{X}\right)\\&=\left(\frac{1}{n}+A\bar{X}^2\right)\sigma^2+X_i^2A\sigma^2-2X_iA\bar{X}\sigma^2\\&=\left(\frac{1}{n}+A\bar{X}^2+AX_i^2-2AX_i\bar{X}\right)\sigma^2=\boxed{\left[\frac{1}{n}+A\left(X_i-\bar{X}\right)^2\right]\sigma^2}.\end{aligned}$$

注意到

$$\boldsymbol{P}=\frac{1}{n}\left(\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)+A\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}=\frac{1}{n}\begin{pmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&&\vdots\\1&1&\cdots&1\end{pmatrix}+A\begin{pmatrix}B_1^2&B_1B_2&\cdots&B_1B_n\\B_2B_1&B_2^2&\cdots&B_2B_n\\\vdots&\vdots&&\vdots\\B_nB_1&B_nB_2&\cdots&B_n^2\end{pmatrix},$$

则投影矩阵对角线上第 i 个元素 h_{ii} 为

$$h_{ii}=\frac{1}{n}+A\bar{X}^2+AX_i^2-2AX_i\bar{X}=\frac{1}{n}+A\left(X_i-\bar{X}\right)^2,\quad i=1,\cdots,n,$$

在统计学中, 我们称 h_{ii} 为**杠杆值** (leverage value), 因此同方差假设下拟合值 $\hat{Y}_i$ 的条件方差为 $\boxed{\mathrm{Var}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)=h_{ii}\sigma^2}$.

现在我们来计算残差 $\hat{\varepsilon}_i=Y_i-\hat{Y}_i$ 的条件期望和条件方差, 容易知道

$$\boxed{\mathbb{E}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right)}=\mathbb{E}\left(Y_i-\hat{Y}_i\middle|\boldsymbol{X}\right)=\mathbb{E}\left(Y_i\middle|\boldsymbol{X}\right)-\mathbb{E}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)=\mathbb{E}\left(Y\middle|X_i\right)-\mathbb{E}\left(Y\middle|X_i\right)=0,$$

这样可以推知, 残差 $\hat{\varepsilon}_i$ 即是回归误差 ε_i 的无偏估计. 同时 $\boxed{\mathrm{Var}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right)}=\mathbb{E}\left(\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right)=\mathbb{E}\left[\left(Y_i-\hat{Y}_i\right)^2\middle|\boldsymbol{X}\right]$. 由于

$$\begin{aligned}\left(Y_i-\hat{Y}_i\right)^2&=\left[\beta_0+\beta_1X_i+\varepsilon_i-\left(\hat{\beta}_0+\hat{\beta}_1X_i\right)\right]^2=\left[\left(\beta_0-\hat{\beta}_0\right)+\left(\beta_1-\hat{\beta}_1\right)X_i+\varepsilon_i\right]^2\\&=\left(\beta_0-\hat{\beta}_0\right)^2+X_i^2\left(\beta_1-\hat{\beta}_1\right)^2+\varepsilon_i^2+2X_i\left(\beta_0-\hat{\beta}_0\right)\left(\beta_1-\hat{\beta}_1\right)\\&\quad+2\varepsilon_i\left(\beta_0-\hat{\beta}_0\right)+2X_i\varepsilon_i\left(\beta_1-\hat{\beta}_1\right),\end{aligned}$$

则

$$\begin{aligned}&\mathbb{E}\left[\left(Y_i-\hat{Y}_i\right)^2\middle|\boldsymbol{X}\right]\\=&\mathrm{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)+X_i^2\mathrm{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)+\mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right)+2X_i\mathrm{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\boldsymbol{X}\right)\\&+2\mathbb{E}\left(\varepsilon_i\left(\beta_0-\hat{\beta}_0\right)\middle|\boldsymbol{X}\right)+2X_i\mathbb{E}\left(\varepsilon_i\left(\beta_1-\hat{\beta}_1\right)\middle|\boldsymbol{X}\right)\\=&\mathrm{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)+X_i^2\mathrm{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)+\mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right)+2X_i\mathrm{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\boldsymbol{X}\right)\\&-2\mathbb{E}\left(\varepsilon_i\hat{\beta}_0\middle|\boldsymbol{X}\right)-2X_i\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)\end{aligned}$$

前四项我们已经计算过了, 现计算后两项. 首先计算

$$\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\varepsilon_iA\sum_{j=1}^nB_jY_j\middle|\boldsymbol{X}\right)=A\sum_{j=1}^nB_j\mathbb{E}\left(Y_j\varepsilon_i\middle|\boldsymbol{X}\right),$$

而

$$\begin{aligned}\mathbb{E}\left(Y_j\varepsilon_i\middle|\boldsymbol{X}\right)&\xlongequal{\text{样本 i.i.d.}}\mathbb{E}\left(Y_j\middle|\boldsymbol{X}\right)\mathbb{E}\left(\varepsilon_i\middle|\boldsymbol{X}\right)=\mathbb{E}\left(Y_j\middle|\boldsymbol{X}\right)\cdot0=0,\quad i\neq j,\\\boxed{\mathbb{E}\left(Y_i\varepsilon_i\middle|\boldsymbol{X}\right)}&=\mathrm{Cov}\left(Y_i,\varepsilon_i\middle|\boldsymbol{X}\right)+\mathbb{E}\left(Y_i\middle|\boldsymbol{X}\right)\mathbb{E}\left(\varepsilon_i\middle|\boldsymbol{X}\right)=\boxed{\mathrm{Cov}\left(Y_i,\varepsilon_i\middle|\boldsymbol{X}\right)}\\&=\mathrm{Cov}\left(\beta_0+\beta_1X_i+\varepsilon_i,\varepsilon_i\middle|\boldsymbol{X}\right)\\&=\underbrace{\mathrm{Cov}\left(\beta_0,\varepsilon_i\middle|\boldsymbol{X}\right)}_{\beta_0\text{为常数}}+\beta_1\underbrace{\mathrm{Cov}\left(X_i,\varepsilon_i\middle|\boldsymbol{X}\right)}_{\text{不相关}}+\mathrm{Cov}\left(\varepsilon_i,\varepsilon_i\middle|\boldsymbol{X}\right)\\&=\mathrm{Var}\left(\varepsilon_i\middle|\boldsymbol{X}\right)\xlongequal{\text{同方差假设}}\sigma^2,\end{aligned}$$

故 $\boxed{\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)=AB_i\sigma^2}$. 再者,

$$\begin{aligned}
&\mathbb{E}\left(\varepsilon_i\hat{\beta}_0\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\varepsilon_i\left(\bar{Y}-\hat{\beta}_1\bar{X}\right)\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\varepsilon_i\bar{Y}\middle|\boldsymbol{X}\right)-\bar{X}\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)\\
=&\mathrm{Cov}\left(\bar{Y},\varepsilon_i\middle|\boldsymbol{X}\right)+\mathbb{E}\left(\bar{Y}\middle|\boldsymbol{X}\right)\mathbb{E}\left(\varepsilon_i\middle|\boldsymbol{X}\right)-\bar{X}\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)\\
=&\mathrm{Cov}\left(\frac{1}{n}\sum_{j=1}^{n}Y_j,\varepsilon_i\middle|\boldsymbol{X}\right)-\bar{X}\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)=\frac{1}{n}\sum_{j=1}^{n}\underbrace{\mathrm{Cov}\left(Y_j,\varepsilon_i\middle|\boldsymbol{X}\right)}_{\text{样本 i.i.d.}}-\bar{X}\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)\\
=&\frac{1}{n}\mathrm{Cov}\left(Y_i,\varepsilon_i\middle|\boldsymbol{X}\right)-\bar{X}\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)=\frac{1}{n}\sigma^2-\bar{X}AB_i\sigma^2,
\end{aligned}$$

即有 $\boxed{\mathbb{E}\left(\varepsilon_i\hat{\beta}_0\middle|\boldsymbol{X}\right)=\left(\frac{1}{n}-\bar{X}AB_i\right)\sigma^2}$. 这样, 残差 $\hat{\varepsilon}_i$ 的条件方差为

$$\begin{aligned}
&\mathrm{Var}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right)\\
=&\mathrm{Var}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)+X_i^2\mathrm{Var}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)+\mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right)+2X_i\mathrm{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\boldsymbol{X}\right)\\
&-2\mathbb{E}\left(\varepsilon_i\hat{\beta}_0\middle|\boldsymbol{X}\right)-2X_i\mathbb{E}\left(\varepsilon_i\hat{\beta}_1\middle|\boldsymbol{X}\right)\\
=&\left(\frac{1}{n}+A\bar{X}^2\right)\sigma^2+X_i^2A\sigma^2+\sigma^2+2X_i\left(-A\bar{X}\sigma^2\right)-2\left(\frac{1}{n}-\bar{X}AB_i\right)\sigma^2-2AB_i\sigma^2\\
=&\left[1-\frac{1}{n}+A\left(\bar{X}^2-2X_i\bar{X}+X_i^2+2\bar{X}B_i-2X_iB_i\right)\right]\sigma^2\\
=&\left[1-\frac{1}{n}+A\left(\left(X_i-\bar{X}\right)^2+2B_i\left(\bar{X}-X_i\right)\right)\right]\sigma^2\\
=&\left[1-\frac{1}{n}+A\left(\left(X_i-\bar{X}\right)^2-2\left(X_i-\bar{X}\right)^2\right)\right]\sigma^2\\
=&\boxed{\left[1-\frac{1}{n}-A\left(X_i-\bar{X}\right)^2\right]\sigma^2}=\boxed{\left(1-h_{ii}\right)\sigma^2}.
\end{aligned}$$

也就是说, 同方差假设下,$\mathrm{Var}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right)=\left(1-h_{ii}\right)\sigma^2<\sigma^2=\mathrm{Var}\left(\varepsilon_i\middle|\boldsymbol{X}\right)$, 即残差的条件方差要小于回归误差的条件方差.

2.6 一元 OLS 估计量下的方差估计量

本节我们将介绍同方差假设下, 由 OLS 估计量所导出的回归方差 σ^2 的估计量.

对于一元线性回归模型 (2.4.1), 因为

$$\sigma^2=\mathrm{Var}\left(\varepsilon\right)=\mathbb{E}\varepsilon^2,$$

则理想的情况下, 回归误差 ε_i 可以被观察, 则回归方差 σ^2 的矩估计 σ^2_{ME} 为

$$\sigma^2_{\mathrm{ME}}=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon},$$

且

$$\mathbb{E}\sigma_{\mathrm{ME}}^2=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2\right)=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\varepsilon_i^2=\frac{1}{n}\sum_{i=1}^n\sigma^2=\sigma^2,$$

即 σ_{ME}^2 是回归方差的无偏估计. 但可惜的是, 回归误差 ε_i **不可被观察**, 必须选择其他的估计方法.

上一节中, 我们证明了残差 $\hat{\varepsilon}_i$ 是 ε_i 的无偏估计, 则用 $\hat{\varepsilon}_i$ 替换 ε_i, 得到估计量

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^n\hat{\varepsilon}_i^2=\frac{1}{n}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\frac{1}{n}\left(\boldsymbol{M}\boldsymbol{\varepsilon}\right)^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon},$$

但由于

$$\sigma_{\mathrm{ME}}^2-\hat{\sigma}^2=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}-\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{I}-\boldsymbol{M}\right)\boldsymbol{\varepsilon}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\varepsilon}\geqslant 0,$$

这里的投影矩阵 $\boldsymbol{P}$ 是半正定, 取期望则有 $\sigma^2=\mathbb{E}\sigma_{\mathrm{ME}}^2\geqslant\mathbb{E}\hat{\sigma}^2$, 即直接使用 $\hat{\varepsilon}_i$ 替换 ε_i 得到的方差估计量是有偏的. 这样, 为了得到回归方差的无偏估计, 我们需要计算残差平方和的期望. 我们将采用两种方法来计算 RSS 的条件期望.

第一种方法则是利用上一节计算的 $\operatorname{Var}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right)=\left(1-h_{ii}\right)\sigma^2$, 则

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^n\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right)=&\sum_{i=1}^n\mathbb{E}\left(\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right)=\sum_{i=1}^n\left(1-h_{ii}\right)\sigma^2\\=&\sigma^2\mathrm{tr}\left(\boldsymbol{I}_n-\boldsymbol{P}\right)=\sigma^2\mathrm{tr}\boldsymbol{M}=\left(n-2\right)\sigma^2,\end{aligned}$$

这里利用了杠杆值 h_{ii} 的是投影矩阵 $\boldsymbol{P}$ 的对角线元素以及定理 (2.5.2). 由此定义

$$s^2=\frac{1}{n-2}\sum_{i=1}^n\hat{\varepsilon}_i^2=\frac{\mathrm{RSS}}{n-2},$$

则有 $\mathbb{E}s^2=\mathbb{E}\left(s^2|\boldsymbol{X}\right)=\sigma^2$, 即 s^2 是回归方差 σ^2 的无偏估计, 我们称 s^2 为 σ^2 的**偏差校正估计量** (bias-corrected estimator).

我们再使用另一个方法计算 RSS 的条件期望, 利用式 (2.32) 则有

$$\begin{aligned}&\mathbb{E}\left(\sum_{i=1}^n\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\sum_{i=1}^nY_i^2\middle|\boldsymbol{X}\right)-\mathbb{E}\left(\sum_{i=1}^n\hat{Y}_i^2\middle|\boldsymbol{X}\right)\\=&\sum_{i=1}^n\mathbb{E}\left(Y_i^2\middle|\boldsymbol{X}\right)-\sum_{i=1}^n\mathbb{E}\left(\hat{Y}_i^2\middle|\boldsymbol{X}\right)\\=&\sum_{i=1}^n\left[\operatorname{Var}\left(Y_i|\boldsymbol{X}\right)+\left(\mathbb{E}\left(Y_i|\boldsymbol{X}\right)\right)^2\right]-\sum_{i=1}^n\left[\operatorname{Var}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)+\left(\mathbb{E}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)\right)^2\right]\\=&\sum_{i=1}^n\left[\operatorname{Var}\left(Y_i|\boldsymbol{X}\right)+\left(\mathbb{E}\left(Y|X_i\right)\right)^2\right]-\sum_{i=1}^n\left[\operatorname{Var}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)+\left(\mathbb{E}\left(Y|X_i\right)\right)^2\right]\\=&\sum_{i=1}^n\operatorname{Var}\left(Y_i|\boldsymbol{X}\right)-\sum_{i=1}^n\operatorname{Var}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)=n\sigma^2-\sum_{i=1}^nh_{ii}\sigma^2=\left(n-\sum_{i=1}^nh_{ii}\right)\sigma^2\\=&\left(n-\mathrm{tr}\boldsymbol{P}\right)\sigma^2=\left(n-2\right)\sigma^2.\end{aligned}$$

这样, 我们同样可以推导出校正偏差的 s^2 作为 σ^2.

s^2 是最广泛使用的回归方差的无偏估计量, 除此之外, 还有一种被广泛使用的 σ^2 的无偏估计量则是使用标准化残差还代替矩估计 σ^2_{ME} 中不可观察的 ε_i. 由 $\mathbb{E}\left(\hat{\varepsilon}_i|\boldsymbol{X}\right)=0$ 以及 $\operatorname{Var}\left(\hat{\varepsilon}_i|\,\boldsymbol{X}\right)=\left(1-h_{ii}\right)\sigma^2$, 记

$$\bar{\varepsilon}_i=\frac{\hat{\varepsilon}_i-0}{\sqrt{1-h_{ii}}},$$

则称 $\bar{\varepsilon}_i$ 为**标准化残差** (standardized residual). 显然 $\bar{\varepsilon}_i$ 满足有 $\mathbb{E}\left(\bar{\varepsilon}_i|\,\boldsymbol{X}\right)=0$ 及

$$\operatorname{Var}\left(\bar{\varepsilon}_i|\,\boldsymbol{X}\right)=\mathbb{E}\left(\bar{\varepsilon}_i^2\middle|\,\boldsymbol{X}\right)=\mathbb{E}\left(\frac{\hat{\varepsilon}_i^2}{1-h_{ii}}\middle|\,\boldsymbol{X}\right)=\frac{\mathbb{E}\left(\hat{\varepsilon}_i^2|\,\boldsymbol{X}\right)}{1-h_{ii}}=\frac{\left(1-h_{ii}\right)\sigma^2}{1-h_{ii}}=\sigma^2,$$

这样对于 $\sum\limits_{i=1}^{n}\bar{\varepsilon}_i^2=\sum\limits_{i=1}^{n}\dfrac{\hat{\varepsilon}_i^2}{1-h_{ii}}$ 便有

$$\mathbb{E}\left(\sum_{i=1}^{n}\bar{\varepsilon}_i^2\middle|\,\boldsymbol{X}\right)=\sum_{i=1}^{n}\mathbb{E}\left(\bar{\varepsilon}_i^2\middle|\,\boldsymbol{X}\right)=\sum_{i=1}^{n}\sigma^2=n\sigma^2,$$

容易知道, 使用标准化残差代替回归误差的矩估计

$$\bar{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\bar{\varepsilon}_i^2=\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varepsilon}_i^2}{1-h_{ii}}$$

是回归方差 σ^2 的无偏估计.

2.7 一元 OLS 估计量的区间估计与假设检验

本节将介绍一元正态回归模型, 其为一元线性回归模型的一种特殊情形. 在之前的模型 (2.4.1) 中, 我们尚无设定被解释变量 Y 以及回归误差 ε 的概率分布, 而在本节我们将研究具有正态分布下的线性回归模型——有了概率分布, 我们便可以引入区间估计以及假设检验.

模型 2.7.1 (一元正态回归模型). *在一元线性回归模型 (2.4.1) 的假设之上, 补充假设*

$$\varepsilon|\,X\sim\mathrm{N}\left(0,\sigma^2\right) \tag{2.35}$$

成立时, 满足如下设定的模型称为一元**正态回归模型** *(normal regression model).*

$$
\begin{aligned}
&Y=\mathscr{P}(Y|X)+\varepsilon\\
&\mathscr{P}(Y|X)=\beta_0+\beta_1X,\quad \begin{cases}\beta_0=\mathbb{E}Y-\beta_1\mathbb{E}X\\ \beta_1=\dfrac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}\end{cases}\\
&\mathbb{E}(X\varepsilon)=0\\
&\mathbb{E}(\varepsilon)=0\quad (\underline{不}需要常数项\beta_0)\\
&\operatorname{Var}(\varepsilon)=\mathbb{E}\varepsilon^2=\mathbb{E}(\varepsilon|X)=\sigma^2
\end{aligned}
$$

一元正态回归模型较模型 (2.4.1) 的差别, 仅在于假设了回归误差 ε 在给定解释变量 X 下的**条件分布**服从正态分布, 而式 (2.35) 通常被称为**正态假设** (normality assumption). 总体的回归误差 $\varepsilon|\,X\sim \mathrm{N}(0,\sigma^2)$, 这里的正态分布 $\mathrm{N}(0,\sigma^2)$ 是**不依赖于**解释变量 X, 也就是说条件随机变量 $\varepsilon|\,X$ 的各种性质均与 X 无关. 这样, 简单随机样本下样本个体的回归误差 ε_i 独立同分布, 且也满足 $\varepsilon_i|\,X_i\sim \mathrm{N}(0,\sigma^2)$, 因而对于误差向量便有 $\boldsymbol{\varepsilon}|\,\boldsymbol{X}\sim \mathrm{N}(0,\sigma^2\boldsymbol{I}_n)$,$\boldsymbol{\varepsilon}|\,\boldsymbol{X}$ 的联合分布不受到 $\boldsymbol{X}$ 的影响.

于是对于一元正态回归模型 (2.7.1), 同方差假设

$$\mathbb{E}\left(\varepsilon_i^2\middle|\,\boldsymbol{X}\right)=\mathbb{E}\varepsilon_i^2=\sigma^2$$

直接成立.

因此, 在引入条件分布服从正态分布的回归误差后, 模型 (2.7.1) 的总体和样本便是

$$
\begin{aligned}
\text{总体}:&Y=\beta_0+\beta_1X+\varepsilon,\quad \varepsilon|\,X\sim \mathrm{N}\left(0,\sigma^2\right),\\
\text{样本}:&\boldsymbol{Y}=\beta_0\mathbf{1}+\beta_1\boldsymbol{X}+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}|\,\boldsymbol{X}\sim \mathrm{N}\left(0,\sigma^2\boldsymbol{I}_n\right),\\
\text{样本个体}:&Y_i=\beta_0+\beta_1X_i+\varepsilon_i,\quad ,\varepsilon_i|\,X_i \text{ i.i.d.}\sim \mathrm{N}\left(0,\sigma^2\right),i=1,\cdots,n,
\end{aligned}
$$

一元正态回归模型至少假定了回归误差的条件分布, 而对于研究的变量 (Y,X) 并没有限定其联合分布. 如果在模型 (2.4.1) 上假定了 (Y,X) 服从联合正态分布, 那么得到的模型则称为线性回归模型的经典模型.

模型 2.7.2 (一元经典线性回归模型)**.** 在模型 *(2.4.1)* 的假设中加入 (Y,X) 服从联合正态分布, 则称该模型为一元**经典线性回归模型** *(classic linear regression model).*

由于回归误差 ε 是 (Y,X) 的线性变换, 模型 (2.7.2) 中 (Y,X) 服从联合正态分布, 则可以得到 ε 服从正态分布 $\mathrm{N}(0,\sigma^2)$. 同时在线性回归模型中,X 与 ε 不相关, 对于两个服从正态分布的随机变量而言不相关等价于独立, 因此则模型 (2.7.2) 也满足有

$\varepsilon|\,X\sim \mathrm{N}\,(0,\sigma^2)$. 故一元经典线性回归模型 (2.7.2) 只是一元正态回归模型 (2.7.1) 的特例情形.

一元正态回归模型给定了回归误差的条件分布, 现在我们也可以推导其他随机变量的条件分布. 显然, 被解释变量 Y 的总体具有条件分布 $Y|\,X\sim \mathrm{N}\,(\beta_0+\beta_1X,\sigma^2)$, 相应地 Y 的样本个体具有条件分布 $Y_i|\,X\sim \mathrm{N}\,(\beta_0+\beta_1X_i,\sigma^2)$.

对于参数的 OLS 估计量

$$\hat{\beta}_1=A\,(\boldsymbol{X})\sum_{i=1}^{n}B_i\,(\boldsymbol{X})\,Y_i,\quad \hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X},$$

给定样本 $\boldsymbol{X}=(X_1,\cdots,X_n)^{\mathrm{T}}$ 的情形下, 则 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 均是 $Y_1,\cdots,Y_n$ 的线性组合, 而正态分布的线性组合仍是线性组合, 依据此前章节计算的条件期望和条件方差, 这样 OLS 估计量便有

$$\hat{\beta}_1\Big|\,\boldsymbol{X}\sim \mathrm{N}\left(\beta_1,A\sigma^2\right),\quad \hat{\beta}_0\Big|\,\boldsymbol{X}\sim \mathrm{N}\left(\beta_0,\left(\frac{1}{n}+A\bar{X}^2\right)\sigma^2\right). \tag{2.36}$$

同理可以推知, 拟合值 $\hat{Y}_i$ 与残差 $\hat{\varepsilon}_i$ 具有条件正态分布, 即

$$\hat{Y}_i\Big|\,\boldsymbol{X}\sim \mathrm{N}\left(\beta_0+\beta_1X_i,h_{ii}\sigma^2\right),\quad \hat{\varepsilon}_i|\,\boldsymbol{X}\sim \mathrm{N}\left(0,(1-h_{ii})\,\sigma^2\right). \tag{2.37}$$

我们现在介绍回归方差估计量 $s^2=\dfrac{1}{n-2}\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2$ 的条件分布, 利用线性代数的知识可以证明

$$\frac{(n-2)\,s^2}{\sigma^2}\Big|\,\boldsymbol{X}\sim\chi^2\,(n)\,, \tag{2.38}$$

但这个证明并不算轻松, 我们将在下一章的多元线性回归模型里去详细介绍. 在此, 我们给出另一种 RSS 的分解以说明式 (2.38) 成立.

首先, 对于样本个体可以推知

$$Y_i=\beta_0+\beta_1X_i+\varepsilon_i\quad\Rightarrow\quad\bar{Y}=\beta_0+\beta_1\bar{X}+\bar{\varepsilon}\quad\Rightarrow\quad Y_i-\bar{Y}=\beta_1\left(X_i-\bar{X}\right)+(\varepsilon_i-\bar{\varepsilon})\,,$$

其中

$$\bar{\varepsilon}=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$$

为回归误差 ε_i 的样本均值 (不可观察). 同样使用前面的符号记离差 $B_i=X_i-\bar{X},i=1,\cdots,n$, 则有模型 (2.4.1) 即有消去截距 β_0 的**中心化**形式

$$Y_i-\bar{Y}=\beta_1B_i+(\varepsilon_i-\bar{\varepsilon})\,. \tag{2.39}$$

我们再注意到 OLS 估计量 $\hat{\beta}_1$ 其较参数 β_1 的差值为

$$\begin{aligned}\hat{\beta}_1-\beta_1=&\beta_1-A\sum_{i=1}^{n}B_iY_i=A\sum_{i=1}^{n}B_i\left(\beta_1B_i+(\varepsilon_i-\bar{\varepsilon})+\bar{Y}\right)-\beta_1\\=&A\beta_1\sum_{i=1}^{n}B_i^2+A\sum_{i=1}^{n}B_i\left(\varepsilon_i-\bar{\varepsilon}\right)-\beta_1=A\beta_1A^{-1}+A\sum_{i=1}^{n}B_i\left(\varepsilon_i-\bar{\varepsilon}\right)-\beta_1\\=&A\sum_{i=1}^{n}B_i\left(\varepsilon_i-\bar{\varepsilon}\right),\end{aligned}$$

即$\hat{\beta}_1-\beta_1$ 是回归误差 $\varepsilon_1,\cdots,\varepsilon_n$ 的线性组合, 而 $\varepsilon_i|\,X_i$ i.i.d. $\sim \mathrm{N}\left(0,\sigma^2\right),i=1,\cdots,n$, 于是

$$\hat{\beta}_1-\beta_1\Big|\,\boldsymbol{X}\sim\mathrm{N}\left(0,A\sigma^2\right),$$

当然,$\hat{\beta}_1-\beta_1$ 的条件分布也可以用式 (2.36) 得到, 我们这里是为了得到 $\hat{\beta}_1-\beta_1$ 的表达式. 一般地, 我们称参数的估计量与参数的差值为**抽样误差** (sampling error).

现在我们利用式 (2.39) 对 $\mathrm{RSS}=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ 进行分解, 不难算得

$$\begin{aligned}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=&\sum_{i=1}^{n}\left[\left(\beta_1-\hat{\beta}_1\right)B_i+(\varepsilon_i-\bar{\varepsilon})\right]^2\\=&\sum_{i=1}^{n}\left[\left(\beta_1-\hat{\beta}_1\right)^2B_i^2-2\left(\hat{\beta}_1-\beta_1\right)B_i\left(\varepsilon_i-\bar{\varepsilon}\right)+\left(\varepsilon_i-\bar{\varepsilon}\right)^2\right]\\=&A^{-1}\left(\beta_1-\hat{\beta}_1\right)^2+2\left(\beta_1-\hat{\beta}_1\right)\sum_{i=1}^{n}B_i\left(\varepsilon_i-\bar{\varepsilon}\right)+\sum_{i=1}^{n}\left(\varepsilon_i-\bar{\varepsilon}\right)^2\\=&A^{-1}\left(\beta_1-\hat{\beta}_1\right)^2-2A^{-1}\left(\beta_1-\hat{\beta}_1\right)^2+\sum_{i=1}^{n}\left(\varepsilon_i-\bar{\varepsilon}\right)^2\\=&\sum_{i=1}^{n}\left(\varepsilon_i-\bar{\varepsilon}\right)^2-A^{-1}\left(\beta_1-\hat{\beta}_1\right)^2,\end{aligned}$$

这里用到了 $A^{-1}=\sum_{i=1}^{n}B_i^2$, 故有

$$\sum_{i=1}^{n}\left(\varepsilon_i-\bar{\varepsilon}\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2+A^{-1}\left(\beta_1-\hat{\beta}_1\right)^2.\tag{2.40}$$

而对于样本 2 阶中心距, 我们熟知 $\sum_{i=1}^{n}\left(\varepsilon_i-\bar{\varepsilon}\right)^2=\sum_{i=1}^{n}\varepsilon_i^2-n\bar{\varepsilon}^2$, 代入式 (2.40) 中整理得

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2+A^{-1}\left(\beta_1-\hat{\beta}_1\right)^2+n\bar{\varepsilon}^2,\tag{2.41}$$

这即是对于误差平方和 SSE 的分解. 我们对于式 (2.41) 两边同除以 σ^2, 整理即得到

$$\boxed{\sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sigma}\right)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2+\left(\frac{\beta_1-\hat{\beta}_1}{\sqrt{A\sigma^2}}\right)^2+\left(\frac{\bar{\varepsilon}}{\sqrt{\sigma^2/n}}\right)^2}. \tag{2.42}$$

不难发现, 式 (2.42) 中回归误差 ε_i、抽样误差 $\hat{\beta}_1-\beta_1$ 和样本均值 $\bar{\varepsilon}$ 均进行了标准化处理. 在没有正态假设下, 式 (2.40)、(2.41) 及 (2.42) 也是成立的, 且可以证明式 (2.42) 右边三项是互相独立的. 这样, 对式 (2.42) 求条件期望, 也可以由此推导偏差校正估计量 s^2. 而加入了正态假设, 回归误差 ε_i、抽样误差 $\hat{\beta}_1-\beta_1$ 和样本均值 $\bar{\varepsilon}$ 这三者的条件分布均服从正态分布, 标准化再平方后则有

$$\sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sigma}\right)^2\Bigg|\boldsymbol{X}\sim\chi^2(n),\quad\left(\frac{\beta_1-\hat{\beta}_1}{\sqrt{A\sigma^2}}\right)^2\Bigg|\boldsymbol{X}\sim\chi^2(1),\quad\left(\frac{\bar{\varepsilon}}{\sqrt{\sigma^2/n}}\right)^2\Bigg|\boldsymbol{X}\sim\chi^2(1)$$

成立. 这样式 (2.42) 的左边是自由度为 n 的卡方分布, 右边后两项都是自由度为 1 的卡方分布成立, 利用概率论中矩生成函数 (moment generating function, 见第4.3节) 可以证明

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\Bigg|\boldsymbol{X}\sim\chi^2(n-2),$$

也即证明了式 (2.38).

现在我们便得到了正态假设下 OLS 估计量下课统计量的概率分布, 由此我们即可进行区间估计和假设检验.

定义 2.7.1 (一元正态回归模型的 t 统计量)**.** *对于模型 (2.7.1) 的 OLS 估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$, 其标准化后有*

$$\frac{\beta_1-\hat{\beta}_1}{\sqrt{A\sigma^2}}\Bigg|\boldsymbol{X}\sim\mathrm{N}(0,1),\quad\frac{\beta_0-\hat{\beta}_0}{\sqrt{\left(\frac{1}{n}+A\bar{X}^2\right)\sigma^2}}\Bigg|\boldsymbol{X}\sim\mathrm{N}(0,1),$$

*由于 σ^2 为未知的***厌恶参数** *(nuisance parameter), 使用 σ^2 的偏差校正估计量 s^2 代替, 记*

$$\mathrm{SE}\left(\hat{\beta}_1\right)=\sqrt{As^2},\quad\mathrm{SE}\left(\hat{\beta}_0\right)=\sqrt{\left(\frac{1}{n}+A\bar{X}^2\right)s^2}, \tag{2.43}$$

*我们称式 (2.43) 为 OLS 估计量的***标准误差** *(standard error)*或***标准误**, *则我们称*

$$T=\frac{\beta_i-\hat{\beta}_i}{\mathrm{SE}\left(\hat{\beta}_i\right)} \tag{2.44}$$

*为 **t 统计量** (t-statistic).*

对于 $\hat{\beta}_1$ 的 t 统计量, 由于

$$T=\frac{\beta_1-\hat{\beta}_1}{\mathrm{SE}\left(\hat{\beta}_1\right)}=\frac{\beta_1-\hat{\beta}_1}{\sqrt{A\sigma^2}}\frac{\sqrt{A\sigma^2}}{\mathrm{SE}\left(\hat{\beta}_1\right)}=\frac{\beta_i-\hat{\beta}_i}{\sqrt{A\sigma^2}}\Bigg/\sqrt{\frac{\left(\mathrm{SE}\left(\hat{\beta}_i\right)\right)^2}{A\sigma^2}},$$

而

$$\frac{\beta_1-\hat{\beta}_1}{\sqrt{A\sigma^2}}\Bigg|\boldsymbol{X}\sim\mathrm{N}\left(0,1\right),\quad\frac{\left(\mathrm{SE}\left(\hat{\beta}_i\right)\right)^2}{A\sigma^2}\Bigg|\boldsymbol{X}=\frac{\left(n-2\right)s^2/\sigma^2}{n-2}\Bigg|\boldsymbol{X}\sim\frac{\chi^2\left(n-2\right)}{n-2},$$

故依据数理统计的知识, 在给定 $\boldsymbol{X}$ 的条件下 $\hat{\beta}_1$ 的 t 统计量服从自由度为 $n-2$ 的 t 分布. 同理可以证明 $\hat{\beta}_0$ 的 t 统计量 T 也满足有 $T|\boldsymbol{X}\sim t\left(n-2\right)$.

利用 t 统计量作为**枢轴变量** (pivotal variable), 我们便可以对总体参数 β 进行区间估计. 对于参数 β, 若存在区间 $\hat{I}$ 使得

$$\mathbb{P}\left(\beta\in\hat{I}\right)=1-\alpha$$

成立, 则称区间估计量 $\hat{I}$ 是参数 β 的**置信水平** (confidence level)为 $1-\alpha$ 的**置信区间** (confidence interval). 对于概率 $\mathbb{P}\left(\beta\in\hat{I}\right)$ 的理解, 应该是将置信区间 $\hat{I}$ 视为随机的, 而总体参数 β 应该视为“已知”或固定的, 这样, 置信水平为 $1-\alpha$ 的置信区间 $\hat{I}$ 意味着**重复抽样下**区间 $\hat{I}$ 含义参数 β 的真实值的概率为 $1-\alpha$.

在计量经济学中, 对于回归系数 β 而言我们通常构造如下的置信区间:

$$\hat{I}=\left[\hat{\beta}_i-c\cdot\mathrm{SE}\left(\hat{\beta}_i\right),\hat{\beta}_i+c\cdot\mathrm{SE}\left(\hat{\beta}_i\right)\right],$$

其中 c 是某个大于 0 的常数. 如果 $\hat{I}$ 的置信水平为 $1-\alpha$, 则有

$$\begin{aligned}\mathbb{P}\left(\hat{\beta}_i-c\cdot\mathrm{SE}\left(\hat{\beta}_i\right)\leqslant\beta_i\leqslant\hat{\beta}_i+c\cdot\mathrm{SE}\left(\hat{\beta}_i\right)\right)&=1-\alpha\\\Leftrightarrow\mathbb{P}\left(-c\leqslant\frac{\beta-\hat{\beta}_i}{\mathrm{SE}\left(\hat{\beta}_i\right)}\leqslant c\right)&=1-\alpha,\end{aligned}$$

也即是 t 统计量满足

$$\mathbb{P}\left(-c\leqslant t\leqslant c\right)=F\left(c\right)-F\left(-c\right)\xlongequal{\text{对称性}}F\left(c\right)-\left(1-F\left(c\right)\right)=2F\left(c\right)-1=1-\alpha,$$

即有 $F\left(c\right)=1-\dfrac{\alpha}{2}$, 这里 F 是 t 分布的累计密度函数. 对于连续的累计密度函数函数 F, 其反函数 F^{-1} 一定是存在的, 故 $c=F^{-1}\left(1-\dfrac{\alpha}{2}\right)$. 在数理统计中我们也已经学过, 这里的 c 即是上 $\alpha/2$ 分位数 $t_{\alpha/2}$.

一条有用的经验法则是, 当 $n-K>61$(一元 OLS 即是 $k=2$) 时, 参数 β 的 95% 水平的置信区间近似为 $\left[\hat{\beta}_i-2\mathrm{SE}\left(\hat{\beta}_i\right),\hat{\beta}_i+2\mathrm{SE}\left(\hat{\beta}_i\right)\right]$.

同样, 以卡方分布作为枢轴变量, 考虑回归方差 σ^2 的置信水平为 $1-\alpha$ 的置信区间 $\hat{I}$ 有

$$\begin{aligned}\mathbb{P}\left(c_1\leqslant\frac{(n-2)s^2}{\sigma^2}\leqslant c_2\right)&=F(c_2)-F(c_1),=1-\alpha\\ \Leftrightarrow\mathbb{P}\left(\frac{(n-2)s^2}{c_2}\leqslant\sigma^2\leqslant\frac{(n-2)s^2}{c_1}\right)&=1-\alpha,\end{aligned}$$

即 σ^2 的置信区间为

$$\hat{I}=\left[\frac{(n-2)s^2}{c_2},\frac{(n-2)s^2}{c_1}\right],$$

其中 $c_1=F^{-1}(\alpha/2),c_2=F^{-1}(1-\alpha/2)$, 也即是卡方分布的上 $\alpha/2$ 分位数和上 $1-\alpha/2$ 分位数.

关于假设检验问题, 见第3.13节.

2.8 一元 OLS 估计量的渐进性质

简单介绍什么是“大样本”性质, 具体参加后面章节.

2.9 代码

```
%% OLS估计量无偏性
%设定E(Y|X)=10x+25,e~N(0,5^2)
%考虑重复抽样k次,单次样本容量为n
clear
clc
k=1000;%抽样次数
n=50;%单次抽样容量
s=RandStream('mcg16807','Seed',114514);%设定复现的随机流以及种子
Beta=zeros(2,k);%存储每次估计的系数beta
for i=1:k
X=[ones(n,1),randsample(s,-100:100,n)'];%生成样本X
e=5*randn(s,n,1);%生成扰动项
Y=X*[25;10]+e;%数据生成过程
beta=X\Y;%OLS估计量
Beta(:,i)=beta;%保存OLS估计量
end
beta_hat=cumsum(Beta,2)./(1:k);%OLS估计的平均值
```

```
18 selfGrootDefault(2)%绘图美化函数,没有请注释
19 %绘制估计值均值变化
20 figure(1)
21 set(gcf,'unit','centimeters', ...
22 'Position',[0 0 21*1.25 29.7*0.35*1.25]);%设置figure为A4纸宽度
23 tiledlayout(2,2);%创建2*2画布
24 nexttile
25 histfit(Beta(1,:))%beta0的直方图和拟合的pdf
26 xline(25,'-.b','LineWidth',2)
27 xlabel('\beta_0','FontSize',12.5)
28
29 nexttile
30 plot(1:25:k,beta_hat(1,1:25:k),'-o')
31 xlabel('\fontname{宋体}重复抽样次数\fontname{Times New ...
       Roman}k','FontSize',12.5)
32 ylabel('\beta_0\fontname{宋体}估计值的平均值','FontSize',12.5);
33 yline(25,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
       ...
34 ,'left','LabelVerticalAlignment','bottom')
35
36 nexttile
37 histfit(Beta(2,:))%beta1的直方图和拟合的pdf
38 xline(10,'-.b','LineWidth',2)
39 xlabel('\beta_1','FontSize',12.5)
40
41 nexttile
42 plot(1:25:k,beta_hat(2,1:25:k),'-o')
43 xlabel('\fontname{宋体}重复抽样次数\fontname{Times New ...
       Roman}k','FontSize',12.5)
44 ylabel('\beta_0\fontname{宋体}估计值的平均值','FontSize',12.5);
45 yline(10,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
       ...
46 ,'left','LabelVerticalAlignment','bottom')
47 %绘制重复抽样pdf
48 figure(2)
49 set(gca,'FontName','Times New Roman');
50 set(gcf,'unit','centimeters', ...
51 'Position',[0 0 21*1.25 29.7*0.4*1.25]);%设置figure为A4纸宽度
52 tiledlayout(2,1);%创建2*2画布
53 nexttile
54 pd1=fitdist(Beta(1,1:25)','Normal');%beta0拟合的pdf_25次抽样
55 pd2=fitdist(Beta(1,1:100)','Normal');%beta0拟合的pdf_100次抽样
```

```
pd3=fitdist(Beta(1,1:k)','Normal');%beta0拟合的pdf_k次抽样
x_values=20:0.1:30;
y1=pdf(pd1,x_values);%计算pdf_25次抽样
y2=pdf(pd2,x_values);%计算pdf_100次抽样
y3=pdf(pd3,x_values);%计算pdf_k次抽样
plot(x_values,y1,'-.')
hold on
plot(x_values,y2,'-*')
plot(x_values,y3,'-o')
xline(25,'-.b','LineWidth',2,'DisplayName','\beta_0 ...
    \fontname{宋体}真实值')
xlabel('\beta_0','FontSize',12.5)
legend('\fontname{Times New ...
    Roman}25\fontname{宋体}次抽样','\fontname{Times New ...
    Roman}100\fontname{宋体}次抽样','\fontname{Times New ...
    Roman}1000\fontname{宋体}次抽样')
xlim([23,27])
ylim([0,0.65])

nexttile
pd1=fitdist(Beta(2,1:25)','Normal');%beta1拟合的pdf_25次抽样
pd2=fitdist(Beta(2,1:100)','Normal');%beta1拟合的pdf_100次抽样
pd3=fitdist(Beta(2,1:k)','Normal');%beta1拟合的pdf_k次抽样
x_values=9.5:0.002:10.5;
y1=pdf(pd1,x_values);%计算pdf_25次抽样
y2=pdf(pd2,x_values);%计算pdf_100次抽样
y3=pdf(pd3,x_values);%计算pdf_k次抽样
plot(x_values,y1,'-.')
hold on
plot(x_values,y2,'-*')
plot(x_values,y3,'-o')
xline(10,'-.b','LineWidth',2,'DisplayName','\beta_1 ...
    \fontname{宋体}真实值')
xlabel('\beta_1','FontSize',12.5)
legend('\fontname{Times New ...
    Roman}25\fontname{宋体}次抽样','\fontname{Times New ...
    Roman}100\fontname{宋体}次抽样','\fontname{Times New ...
    Roman}1000\fontname{宋体}次抽样')
xlim([9.96,10.04])
ylim([0,44])
```

第三章　多元线性回归模型

本章我们将一元线性回归模型拓展至多元的情况, 即从单个解释变量 X 拓展至 K 个解释变量 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$. 本章内容结构上与上一节并无多大差异, 利用矩阵符号, 一元 OLS 的各类问题都可以自然对应至多元情形, 且有着更为更一般的视角.

3.1　多元线性回归模型的矩阵符号

本节介绍多元线性回归模型将使用的矩阵符号. 我们所研究的被解释变量依旧记为 Y, 而感兴趣的解释变量则共有 K 个, 分别是随机变量 $X_1,\cdots,X_K$, 我们用随机向量 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$ 一概表示.

本章考虑的依旧是定义 (2.2.1) 的线性 CEF, 我们使用与模型 (2.2.1) 相同的矩阵符号, 即记回归系数为 $K\times 1$ 的列向量 $\boldsymbol{\beta}=(\beta_1.\cdots,\beta_K)^{\mathrm{T}}$, 那么线性 CEF 可以用向量内积表示为

$$m(\mathbf{X})=\mathbb{E}(Y|\mathbf{X})=\beta_1X_1+\cdots+\beta_KX_K=\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}.$$

在上一章中, 我们区别了一元投影模型 (2.2.3) 以及一元线性回归模型 (2.4.1) 的区别. 相应地, 我们将这两个模型拓展为多元情形. 对于投影模型, 给定 $\mathbf{X}$ 下 Y 的最优线性预测量记为

$$\mathscr{P}(Y|\mathbf{X})=\beta_1X_1+\cdots+\beta_nX_K=\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta},$$

则投影模型方程为 $Y=\mathscr{P}(Y|\mathbf{X})+\varepsilon$, 其中 ε 为投影误差. 当回归模型 $Y=\mathbb{E}(Y|\mathbf{X})+e$ 中条件期望函数确实为线性 CEF 时, 投影模型即为线性回归模型, 则是投影误差 ε 等价于回归误差 e. 这样, 线性回归模型的总体方程为

$$Y=\beta_1X_1+\cdots+\beta_nX_K+\varepsilon=\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon. \tag{3.1}$$

现在考虑样本容量为 n 的抽样样本, 对于样本个体 $i=1,\cdots,n$, 则其线性回归方程为

$$Y_i=\beta_1X_{i1}+\cdots+\beta_nX_{iK}+\varepsilon_i=\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i, \tag{3.2}$$

其中 $\mathbf{X}_i = (X_{i1}, \cdots, X_{iK})^{\mathrm{T}}$ 为个体 i 对应的被解释变量 $\mathbf{X}$. 我们将 n 个样本回归方程以线性方程组的形式写出, 即有

$$\begin{cases} Y_1 = \beta_1 X_{11} + \cdots + \beta_n X_{K1} + \varepsilon_1, \\ \cdots \\ Y_i = \beta_1 X_{1i} + \cdots + \beta_n X_{Ki} + \varepsilon_i, \\ \cdots \\ Y_n = \beta_1 X_{1n} + \cdots + \beta_n X_{Kn} + \varepsilon_n, \end{cases}$$

依据线性代数的知识, 我们将线性方程组以使用矩阵符号书写为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.3}$$

其中

$$\boldsymbol{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X}_{n\times K} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1K} \\ \vdots & \vdots & & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{iK} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nK} \end{pmatrix}, \quad \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

通常而言,n 维列向量 $\boldsymbol{Y}$ 被称为**数据向量** (data vector), 而 $n \times k$ 维矩阵 $\boldsymbol{X}$ 被称为**数据矩阵** (data matrix)或**设计矩阵** (design matrix). 数据矩阵 $\boldsymbol{X}$ 满足有

$$\boldsymbol{X}_{n\times K} = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{X}_i^{\mathrm{T}} \\ \vdots \\ \mathbf{X}_n^{\mathrm{T}} \end{pmatrix} = (\mathbf{X}_1, \cdots, \mathbf{X}_i, \cdots \mathbf{X}_n)^{\mathrm{T}},$$

这用矩阵的分块运算

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Leftrightarrow \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta} \\ \vdots \\ \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} \\ \vdots \\ \mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} \Leftrightarrow \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{X}_i^{\mathrm{T}} \\ \vdots \\ \mathbf{X}_n^{\mathrm{T}} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

是显而易见的.

多元线性回归模型的总体方程 (3.1) 不含有截距, 但我们只需令 $X_1 \equiv 1$, 相应地样本个体满足 $\mathbf{X}_i = (1, X_{i2}, \cdots, X_{iK})^{\mathrm{T}}$, 即得到了含有截距的线性回归模型为

$$\text{总体 }: Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_n X_K + \varepsilon,$$
$$\text{样本个体 }: Y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_n X_{iK} + \varepsilon_i,$$

此时数据矩阵 $\boldsymbol{X}$ 满足有

$$\boldsymbol{X}_{n\times K} = \begin{pmatrix} 1 & X_{12} & \cdots & X_{1K} \\ 1 & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \cdots & X_{nK} \end{pmatrix},$$

因此, 式 (3.3) 便已经包含了含有截距的线性回归模型.

3.2 非简单随机样本的模型假设

在一元线性回归模型中, 我们选择了简单随机样本以推导 OLS 估计量, 但简单随机样本并非是一种普遍的情形. 就线性回归模型而言, 计量经济学对于样本的假设更为一般化.

假设 3.2.1 (线性回归模型的一般假设). 对于线性回归模型, 我们假设**样本**服从如下假设:

(a) **线性假设** *(linearity assumption)*: 对于样本个体 i, 给定样数据矩阵 $\boldsymbol{X}$ 下 Y_i 条件期望函数为定义 *(2.2.1)* 的线性函数, 即有

$$\mathbb{E}\left(Y_i \middle| \boldsymbol{X}\right) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, \quad i = 1, \cdots, n; \tag{3.4}$$

(b) **严格外生性** *(strict exogeneity)*: 样本个体的回归误差 ε_i 均值独立于数据矩阵 $\boldsymbol{X}$, 即有

$$\mathbb{E}\left(\varepsilon_i \middle| \boldsymbol{X}\right) = 0, \quad i = 1, \cdots, n; \tag{3.5}$$

(c) **无多重共线性** *(no multicollinearity)*: 数据矩阵 $\boldsymbol{X}$ 以概率为 1 列满秩, 即

$$\mathbb{P}\left(\operatorname{rank}\left(\boldsymbol{X}_{n\times K}\right) = K\right) = 1; \tag{3.6}$$

(d) **球形误差方差** *(spherical error variance)*: 样本个体的回归误差 ε_i 满足**条件同方差** *(conditional homoskedasticity)*假设

$$\mathbb{E}\left(\varepsilon_i^2 \middle| \boldsymbol{X}\right) = \sigma^2 > 0, \quad i = 1, \cdots, n \tag{3.7}$$

和不存在**自相关** *(autocorrelation)*或**序列相关** *(serial correlation)*, 即

$$\mathbb{E}(\varepsilon_i\varepsilon_j|\boldsymbol{X})=0,\quad i,j=1,\cdots,n且i\neq j. \tag{3.8}$$

我们现在将对假设 (3.2.1) 的四条假设逐一分析.

对于线性假设式 (3.4), 此假设使得线性投影模型等价于线性回归模型. 可以说假设 (3.2.1)(a) 是线性回归模型相关推导的基础, 否则这些推导都只能归类至线性投影模型 (尽管对于线性回归模型是成立的).

对于严格外生性式 (3.5), 利用简单迭代期望定律 (2.1.1) 便有

$$\mathbb{E}\varepsilon_i=\mathbb{E}(\mathbb{E}(\varepsilon_i|\boldsymbol{X}))=0,\quad i=1,\cdots,n$$

成立. 这条假设则是对应了总体的回归误差 ε 的性质 (见定理 (2.1.4)). 进一步, 我们由严格外生性可以推出回归误差 ε_i 与任意的样本观测值 X_{ij} 是完全不相关的, 这是因为对于索引 $i,j=1,\cdots,n,k=1,\cdots,K$, 依据迭代期望定律 (2.1.2) 有

$$\mathbb{E}(\varepsilon_i|X_{jk})=\mathbb{E}(\mathbb{E}(\varepsilon_i|\boldsymbol{X})|X_{jk})=\mathbb{E}(0|X_{jk})=0,$$

再依据定理 (2.1.3) 有

$$\mathbb{E}(X_{jk}\varepsilon_i)=\mathbb{E}(X_{jk}\mathbb{E}(\varepsilon_i|X_{jk}))=\mathbb{E}(X_{jk}\cdot 0)=0,$$

即回归误差 ε_i 与任意的样本观测值 X_{jk} 是正交的, 故有

$$\operatorname{Cov}(X_{jk},\varepsilon_i)=\mathbb{E}(X_{jk}\varepsilon_i)-\mathbb{E}(X_{jk})\mathbb{E}(\varepsilon_i)=0-\mathbb{E}(X_{jk})\cdot 0=0.$$

特别地, 当 $i=j$ 时 $\operatorname{Cov}(X_{ik},\varepsilon_i)$ 意味着样本个体 i 自身的各解释变量与回归误差完全不相关.

式 (3.4) 和式 (3.5) 用矩阵符号则有 $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})=\boldsymbol{X}\boldsymbol{\beta}$ 和 $\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\mathbf{0}_{n\times 1}$.

对于无多重共线性式 (3.6),$\boldsymbol{X}$ 列满秩也即以为着其列向量组是线性无关的, 而 $\boldsymbol{X}$ 的 K 个列向量代表了总体被解释变量 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$ 的各个分量, 也就是说式 (3.6) 假设了是各个解释变量 X_i 互相线性无关以概率为 1 成立. 之所以用概率考虑则是考虑到了解释变量为随机变量. 为了确保式 (3.6) 成立, 则有必要条件 $n\geqslant\operatorname{rank}(\boldsymbol{X}_{n\times K})=K$ 成立.

对于球形误差方差式 (3.7) 和式 (3.8), 这是从协方差矩阵的角度考虑. 对于任意的随机向量 $\mathbf{X}=(X_1,\cdots,X_n)^{\mathrm{T}}$, 给定条件 Z 下其**条件协方差矩阵** (conditional covariance

matrix)定义为

$$\begin{aligned}\operatorname{Var}(\mathbf{X}|Z) &= \mathbb{E}\left[(\mathbf{X}-\mathbb{E}(\mathbf{X}|Z))(\mathbf{X}-\mathbb{E}(\mathbf{X}|Z))^{\mathrm{T}}\middle|Z\right] \\ &= \begin{pmatrix} \operatorname{Var}(X_1|Z) & \operatorname{Cov}(X_1,X_2|Z) & \cdots & \operatorname{Cov}(X_1,X_n|Z) \\ \operatorname{Cov}(X_2,X_1|Z) & \operatorname{Var}(X_2|Z) & \cdots & \operatorname{Cov}(X_2,X_n|Z) \\ \vdots & \vdots & & \vdots \\ \operatorname{Cov}(X_n,X_1|Z) & \operatorname{Cov}(X_n,X_2|Z) & \cdots & \operatorname{Var}(X_n|Z) \end{pmatrix}.\end{aligned}$$

由此误差向量 $\boldsymbol{\varepsilon}=(\varepsilon_1,\cdots,\varepsilon_n)^{\mathrm{T}}$ 的条件协方差为

$$\boxed{\operatorname{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})} = \mathbb{E}\left[(\boldsymbol{\varepsilon}-\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}))(\boldsymbol{\varepsilon}-\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}))^{\mathrm{T}}\middle|\boldsymbol{X}\right] \overset{\text{式 (3.5)}}{=\!=\!=} \boxed{\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)},$$

这里已经使用了严格外生性假设, 进一步的在式 (3.7) 和式 (3.8) 下误差向量的协方差矩阵为

$$\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right) = \begin{pmatrix} \mathbb{E}(\varepsilon_1^2|\boldsymbol{X}) & \mathbb{E}(\varepsilon_1\varepsilon_2|\boldsymbol{X}) & \cdots & \mathbb{E}(\varepsilon_1\varepsilon_n|\boldsymbol{X}) \\ \mathbb{E}(\varepsilon_2\varepsilon_1|\boldsymbol{X}) & \mathbb{E}(\varepsilon_2^2|\boldsymbol{X}) & \cdots & \mathbb{E}(\varepsilon_2\varepsilon_n|\boldsymbol{X}) \\ \vdots & \vdots & & \vdots \\ \mathbb{E}(\varepsilon_n\varepsilon_1|\boldsymbol{X}) & \mathbb{E}(\varepsilon_n\varepsilon_2|\boldsymbol{X}) & \cdots & \mathbb{E}(\varepsilon_n^2|\boldsymbol{X}) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix},$$

即有 $\operatorname{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\sigma^2\boldsymbol{I}$, “球形”便是对此的具象化描述.

现在我们对比假设 (3.2.1) 和简单随机样本的下的假设, 由于简单随机样本满足有独立同分布, 正如在前一章的计算一样, 对于样本个体 $i=1,\cdots,n$, 回归误差 ε_i 满足有

$$\begin{aligned}\mathbb{E}(\varepsilon_i|\boldsymbol{X}) =& \mathbb{E}(\varepsilon_i|\mathbf{X}_i), \\ \mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right) =& \mathbb{E}\left(\varepsilon_i^2\middle|\mathbf{X}_i\right), \\ \mathbb{E}(\varepsilon_i\varepsilon_j|\boldsymbol{X}) =& \mathbb{E}(\varepsilon_i|\mathbf{X}_i)\,\mathbb{E}(\varepsilon_j|\mathbf{X}_j), \quad i\neq j,\end{aligned}$$

同分布保证了上述的期望具有相同的形式, 但具体的取值由 $\mathbf{X}_i$ 的具体观测值 $\mathbf{x}_i$ 所决定. 因此, 严格外生性以及球形误差方差假设仍然需要, 但可以简化. 严格外生性式 (3.5) 退化为 $\mathbb{E}(\varepsilon_i|\mathbf{X}_i)=0$, 这样无自相关式 (3.8) 便直接成立, 条件同方差式 (3.7) 也可以退化为 $\mathbb{E}(\varepsilon_i^2|\mathbf{X}_i)=\sigma^2$. 假设 (3.2.1) 中的其余假设仍是必须的.

对于线性回归模型, 假设 (3.2.1) 是 OLS 估计量具有良好性质的必要条件. 这在之后的章节中, 我们也会逐步放宽上述假设要求. 如果严格外生性式 (3.5) 不成立, 则计量经济学模型具有**内生性** (endogeneity)问题——这可以说是实证研究中最为常见的问题了——计量经济学家为此建立了反事实的现代因果推断方法, 包括工具变量法、双重差分法、断点回归、合成控制法等等方法, 我们将在后续的章节中介绍这些方法. 球形误

差方差具有式 (3.7) 和 (3.8) 两条假设, 倘若仅有无自相关 (3.8) 成立, 则回归误差 ε_i 存在**异方差** (heteroskedasticity)问题, 这可以写成

$$\mathbb{E}\left(\varepsilon_i^2 \middle| \boldsymbol{X}\right) = \sigma_i^2 > 0, \quad i = 1, \cdots, n, \tag{3.9}$$

这时误差向量 $\boldsymbol{\varepsilon}$ 的条件协方差矩阵满足有

$$\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} = \operatorname{diag}\left(\sigma_1^2, \cdots, \sigma_n^2\right). \tag{3.10}$$

我们将式 (3.9) 称为**条件异方差** (conditional heteroskedasticity), 同时记对角阵 $\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, \cdots, \sigma_n^2\right)$. 对角阵 $\boldsymbol{D}$ 被称为**条件异方差矩阵**. 容易知道, 条件同方差式 (3.7) 实际上是条件异方差 (3.9) 的特殊情形, 即 $\boldsymbol{D} = \sigma^2 \boldsymbol{I}$.

本节的最后, 我们介绍条件协方差矩阵的有用性质.

定理 3.2.1 (条件协方差矩阵的性质). *对于任意的随机向量 $\mathbf{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$, 给定条件 Z 下, 则其协方差矩阵 $\operatorname{Var}(\mathbf{X}|Z)$ 满足如下性质:*

(a) *协方差矩阵计算公式为 $\operatorname{Var}(\mathbf{X}|Z) = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} \middle| Z\right) - \mathbb{E}(\mathbf{X}|Z)\,\mathbb{E}(\mathbf{X}|Z)^{\mathrm{T}}$,*

(b) *条件协方差矩阵 $\operatorname{Var}(\mathbf{X}|Z)$ 为对称方阵;*

(c) *条件协方差矩阵 $\operatorname{Var}(\mathbf{X}|Z)$ 是半正定的;*

(d) *对于任意的**非随机**矩阵 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ 和 m 维常数列向量 $\boldsymbol{a}$, 则有 $\operatorname{Var}(\boldsymbol{A}\mathbf{X} + \boldsymbol{a}|Z) = \boldsymbol{A}\operatorname{Var}(\mathbf{X}|Z)\,\boldsymbol{A}^{\mathrm{T}}$.*

这些性质的证明仅涉及一些基本的矩阵运算, 留给读者自证不难.

3.3 多元线性投影模型的总体参数

本节我们介绍线性投影模型的总体参数, 尽管在线性假设式 (3.4) 成立的情况下, 线性投影模型即是线性回归模型, 但这只是一个假设 (assumption, 而非 hypothesis).

多元线性投影模型的总体方程为

$$Y = \mathscr{P}(Y|\mathbf{X}) + \varepsilon, \tag{3.11}$$

其中 $\mathscr{P}(Y|\mathbf{X}) = \beta_1 X_1 + \cdots + \beta_n X_K = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$ 为给定

$$\mathbf{X} = (X_1, \cdots, X_K)^{\mathrm{T}}$$

下 Y 的**最优线性预测量** (best linear predictor).

由于 $\mathscr{P}(Y|\mathbf{X})$ 使得投影误差 ε 具有最小的 MSE, 据此我们定义了线性投影模型的回归系数 $\boldsymbol{\beta}$ 为

$$\boldsymbol{\beta}=\underset{\boldsymbol{\beta}\in\mathbb{R}^{K\times 1}}{\operatorname{argmin}}\operatorname{MSE}(\boldsymbol{\beta})=\underset{\boldsymbol{\beta}\in\mathbb{R}^{K\times 1}}{\operatorname{argmin}}\mathbb{E}\left(Y-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)^{2}. \tag{3.12}$$

我们来求解这个最优化问题, 将 $\operatorname{MSE}(\boldsymbol{\beta})$ 展开有

$$\begin{aligned}\operatorname{MSE}(\boldsymbol{\beta})&=\mathbb{E}\left(Y-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)^{2}=\mathbb{E}\left(Y^{2}-2Y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}+\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right)\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\right)\\&=\mathbb{E}Y^{2}-2\boldsymbol{\beta}^{\mathrm{T}}\mathbb{E}(\mathbf{X}Y)+\boldsymbol{\beta}^{\mathrm{T}}\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta},\end{aligned}$$

利用矩阵微积分的知识[1], 对参数 $\boldsymbol{\beta}$ 求导得到一阶条件

$$\frac{\partial}{\partial\boldsymbol{\beta}}\operatorname{MSE}(\boldsymbol{\beta})=-2\mathbb{E}(\mathbf{X}Y)+2\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}=0,$$

即有 $\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}=\mathbb{E}(\mathbf{X}Y)$ 成立. 在假设 (2.2.1) 下便可以解出线性投影模型的总体参数为

$$\boxed{\boldsymbol{\beta}=\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y)}. \tag{3.13}$$

在式 (3.13) 中,$\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}$ 为 K 阶方阵,$\mathbb{E}(\mathbf{X}Y)$ 为 $K\times 1$ 的列向量, 因此矩阵乘法的结果即是 $K\times 1$ 的列向量.

由于 $\operatorname{MSE}(\boldsymbol{\beta})$ 是非负的, 当其取极值时只能为极小值, 由此保障了式 (3.13) 满足式 (3.12) 的最优化要求.

在一元线性投影模型中, 我们知道投影误差 ε 与解释变量 X 正交, 这对于多元的情形同样是成立的. 我们将一阶条件 $\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}=\mathbb{E}(\mathbf{X}Y)$ 中的 Y 替换有

$$\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}=\mathbb{E}\left(\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon\right)\right)\Leftrightarrow\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}+\mathbb{E}(\mathbf{X}\varepsilon),$$

即有 $\boxed{\mathbb{E}(\mathbf{X}\varepsilon)=\mathbf{0}_{K\times 1}}$ 成立, 这等价于 $\boxed{\mathbb{E}(X_i\varepsilon)=0}, i=1,\cdots,n$. 也就是说, 对于多元线性投影模型,$K$ 个解释变量均与投影误差 ε 正交.

特别地, 当最优线性预测了含有常数时, 即 $X_1\equiv 1$, 则 $\mathbb{E}(X_i\varepsilon)=\mathbb{E}(1\times\varepsilon)=0$, 即包含常数项时线性投影模型满足有 $\mathbb{E}\varepsilon=0$. 注意, 线性投影模型不需要假设 (3.2.1), 而只需要假设 (2.2.1), 因此截距项对于线性投影模型具有非凡的意义.

在求解一元线性投影模型时, 我们还使用配方法来解决问题. 实际上, 对于任意的关于列向量 $\boldsymbol{a}$ 的函数 $f(\boldsymbol{a})=c-2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{a}+\boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{a}$, 其中 c 为常数,$\boldsymbol{b}$ 为常数列向量,$\boldsymbol{S}$ 为正定

[1]参考附录A.3.

的实对称方阵, 则 $f(\boldsymbol{a})$ 可以配方[2]为

$$\begin{aligned} f(\boldsymbol{a}) &= c - 2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{a} + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{a} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{b} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{b} \\ &= \left(c - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{b}\right) + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{a} - 2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{a} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{b} \\ &= \left(c - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{b}\right) + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{a} - 2\boldsymbol{b}^{\mathrm{T}}\left(\boldsymbol{S}^{-1}\boldsymbol{S}\right)\boldsymbol{a} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{b} \\ &= \left(c - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{b}\right) + \left(\boldsymbol{a} - \boldsymbol{S}^{-1}\boldsymbol{b}\right)^{\mathrm{T}}\boldsymbol{S}\left(\boldsymbol{a} - \boldsymbol{S}^{-1}\boldsymbol{b}\right), \end{aligned}$$

这样,$\mathrm{MSE}(\boldsymbol{\beta}) = \mathbb{E}Y^2 - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbb{E}(\mathbf{X}Y) + \boldsymbol{\beta}^{\mathrm{T}}\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}$ 亦可以配方为如此的形式, 同样 MSE 取最小值时即可得到式 (3.13), 我们利用此还可以得到 MSE 取最小值为

$$\mathbb{E}Y^2 - \left(\mathbb{E}(\mathbf{X}Y)\right)^{\mathrm{T}}\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y).$$

上述推导的线性投影模型并没有直接包含有常数项, 我们可以令 $X_1 \equiv 1$ 来得到含义常数项的线性投影模型, 另一种计算含有常数项的投影模型的方法则是进行中心化处理. 设含义常数项的线性投影模型为

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_K + \varepsilon = \alpha + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \tag{3.14}$$

对式 (3.14) 求期望则有

$$\mathbb{E}Y = \alpha + \mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta} + \mathbb{E}\varepsilon,$$

当含有常数项 α 时 $\mathbb{E}\varepsilon = 0$, 故有

$$\mathbb{E}Y = \alpha + \mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}, \tag{3.15}$$

将式 (3.11) 减去式 (3.15), 即可消去常数项 α 并得到新的线性投影模型

$$Y - \mathbb{E}Y = \left(\mathbf{X} - \mathbb{E}(\mathbf{X})\right)^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon. \tag{3.16}$$

式 (3.16) 的形式与式 (3.11) 是一致的, 于是依据 MSE 最小的要求我们便得到式 (3.16) 中参数 $\boldsymbol{\beta}$ 的表达式为

$$\begin{aligned} \boldsymbol{\beta} &= \left[\mathbb{E}\left(\left(\mathbf{X} - \mathbb{E}(\mathbf{X})\right)\left(\mathbf{X} - \mathbb{E}(\mathbf{X})\right)^{\mathrm{T}}\right)\right]^{-1}\mathbb{E}\left(\mathbf{X} - \mathbb{E}(\mathbf{X})\right)(Y - \mathbb{E}Y) \\ &= \left(\mathrm{Var}(\mathbf{X})\right)^{-1}\mathrm{Cov}(\mathbf{X}, Y), \end{aligned}$$

这与我们求解的一元线性投影模型的参数 $\beta_1 = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$ 具有相同的形式, 只不过换成了协方差矩阵以及矩阵的逆. 利用式 (3.15) 还可以得到常数项 α 的表达式为 $\alpha = \mathbb{E}Y - \mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}$.

我们将本节的多元线性投影模型总结如下.

[2]这个配方过程依然能够通过矩阵相合实现.

模型 3.3.1 (多元线性投影模型). *在假设 (2.2.1) 成立时, 满足如下设定的模型称为多元*线性投影模型 *(linear projection model).*

$$
\begin{aligned}
&Y = \mathscr{P}(Y|\mathbf{X}) + \varepsilon \\
&\quad \mathscr{P}(Y|\mathbf{X}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \quad \boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y) \\
&\quad \mathbb{E}(\mathbf{X}\varepsilon) = \mathbf{0} \\
&\quad \mathbb{E}(\varepsilon) = 0 \quad (\text{需要常数项})
\end{aligned}
$$

模型 3.3.2 (多元线性投影模型 (含有截距项)). *在假设 (2.2.1) 成立时, 满足如下设定的模型称为含有截距项的多元*线性投影模型 *(linear projection model).*

$$
\begin{aligned}
&Y = \mathscr{P}(Y|\mathbf{X}) + \varepsilon \\
&\quad \mathscr{P}(Y|\mathbf{X}) = \alpha + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \quad \begin{cases} \boldsymbol{\beta} = (\mathrm{Var}(\mathbf{X}))^{-1}\mathrm{Cov}(\mathbf{X}, Y) \\ \alpha = \mathbb{E}Y - \mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\right) \end{cases} \\
&\quad \mathbb{E}(\mathbf{X}\varepsilon) = 0 \\
&\quad \mathbb{E}(\varepsilon) = 0 \quad (\text{需要常数项})
\end{aligned}
$$

3.4 多元线性投影的样本估计量

这一节我们介绍多元线性投影模型的 OLS 方法, 并由此得到 OLS 估计量.

依据本章第一节介绍的符号, 对于线性投影方程 (3.11), 设有样本容量为 n 的样本 $\{(Y_1, \mathbf{X}_1), \cdots, (Y_n, \mathbf{X}_n)\}$, 则样本个体的线性投影方程为

$$
Y_i = \mathscr{P}(Y_i|\mathbf{X}_i) + \varepsilon = \mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i,
$$

同样可以矩阵形式表示 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

在第一章中我们介绍了普通最小二乘法, 对于多元的情形我们仍旧依据定义 (2.3.1), 即依据样本对于总体的 MSE 的矩估计为

$$
\hat{S}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}\right)^2 = \frac{1}{n}\mathrm{SSE}(\boldsymbol{\beta}),
$$

而误差平方和 $\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}\right)^2$. 这样,**OLS 估计量** (OLS estimator)即被定义为

$$
\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{K\times 1}}{\operatorname{argmin}}\,\hat{S}(\boldsymbol{\beta}). \tag{3.17}
$$

为了求解式 (3.17) 的最优化问题, 同样只需要研究 $\mathrm{SSE}(\boldsymbol{\beta})$ 的极值. 由于

$$\mathrm{SSE}(\boldsymbol{\beta})=\sum_{i=1}^{n}\left(Y_i-\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}\right)^2=\sum_{i=1}^{n}\left[Y_i-\left(\beta_1X_{i1}+\cdots+\beta_nX_{iK}\right)\right]^2,$$

则对参数 $\boldsymbol{\beta}$ 的分量依次求导, 得到包含有 K 个方程的一阶条件

$$\begin{cases}\dfrac{\partial}{\partial\beta_1}\mathrm{SSE}(\boldsymbol{\beta})=-2\sum\limits_{i=1}^{n}X_{i1}\left[Y_i-\left(\beta_1X_{i1}+\cdots+\beta_nX_{iK}\right)\right]=0,\\ \cdots\\ \dfrac{\partial}{\partial\beta_k}\mathrm{SSE}(\boldsymbol{\beta})=-2\sum\limits_{i=1}^{n}X_{ik}\left[Y_i-\left(\beta_1X_{i1}+\cdots+\beta_nX_{iK}\right)\right]=0,\\ \cdots\\ \dfrac{\partial}{\partial\beta_K}\mathrm{SSE}(\boldsymbol{\beta})=-2\sum\limits_{i=1}^{n}X_{iK}\left[Y_i-\left(\beta_1X_{i1}+\cdots+\beta_nX_{iK}\right)\right]=0,\end{cases}$$

也即有

$$\begin{cases}\sum\limits_{i=1}^{n}X_{i1}\hat{\varepsilon}_i=0,\\ \cdots\\ \sum\limits_{i=1}^{n}X_{ik}\hat{\varepsilon}_i=0,\\ \cdots\\ \sum\limits_{i=1}^{n}X_{iK}\hat{\varepsilon}_i=0,\end{cases}\Leftrightarrow\begin{cases}\left(X_{11},\cdots,X_{n1}\right)\cdot\hat{\boldsymbol{\varepsilon}}=0,\\ \cdots\\ \left(X_{1k},\cdots,X_{nk}\right)\cdot\hat{\boldsymbol{\varepsilon}}=0,\\ \cdots\\ \left(X_{1K},\cdots,X_{nK}\right)\cdot\hat{\boldsymbol{\varepsilon}}=0,\end{cases}$$

这里的 $\varepsilon_i=Y_i-\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ 为样本残差, 其依赖于 OLS 估计量, 而 OLS 估计量满足一阶条件. 注意到 $\left\{\left(X_{11},\cdots,X_{n1}\right)^{\mathrm{T}},\cdots,\left(X_{1k},\cdots,X_{nk}\right)^{\mathrm{T}},\cdots\left(X_{1K},\cdots,X_{nK}\right)^{\mathrm{T}}\right\}$ 为数据矩阵 $\boldsymbol{X}$ 的列向量组, 于是一阶条件等价于

$$\left(\left(X_{11},\cdots,X_{n1}\right),\cdots,\left(X_{1k},\cdots,X_{nk}\right),\cdots\left(X_{1K},\cdots,X_{nK}\right)\right)^{\mathrm{T}}\cdot\hat{\boldsymbol{\varepsilon}}=\mathbf{0}_{K\times1}$$

即有

$$\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\mathbf{0}_{K\times1}\tag{3.18}$$

成立. 式 (3.18) 即是多元线性投影模型的 OLS 估计量的正规方程组. 我们将式 (3.18) 中的残差向量 $\hat{\varepsilon}$ 替换掉有

$$\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)=\mathbf{0}\Leftrightarrow\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

于是解得多元线性投影模型的 OLS 估计量为

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.\tag{3.19}$$

同样 $\mathrm{SSE}(\boldsymbol{\beta})$ 是非负的, 一阶条件求得的极值点 $\hat{\boldsymbol{\beta}}$ 便是极小值点, 因而式 (3.19) 符合式 (3.17) 的要求.

注意到

$$\boldsymbol{X}_{n\times K}=\begin{pmatrix}\mathbf{X}_1^{\mathrm{T}}\\ \vdots\\ \mathbf{X}_i^{\mathrm{T}}\\ \vdots\\ \mathbf{X}_n^{\mathrm{T}}\end{pmatrix},\quad \boldsymbol{X}^{\mathrm{T}}=\left(\mathbf{X}_1,\cdots,\mathbf{X}_i,\cdots\mathbf{X}_n\right),$$

则

$$\underbrace{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}_{K\text{阶方阵}}=\left(\mathbf{X}_1,\cdots,\mathbf{X}_i,\cdots\mathbf{X}_n\right)\begin{pmatrix}\mathbf{X}_1^{\mathrm{T}}\\ \vdots\\ \mathbf{X}_i^{\mathrm{T}}\\ \vdots\\ \mathbf{X}_n^{\mathrm{T}}\end{pmatrix}=\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}},$$

$$\underbrace{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}}_{K\times 1\text{列向量}}=\left(\mathbf{X}_1,\cdots,\mathbf{X}_i,\cdots\mathbf{X}_n\right)\begin{pmatrix}Y_1\\ \vdots\\ Y_i\\ \vdots\\ Y_n\end{pmatrix}=\sum_{i=1}^{n}\mathbf{X}_iY_i,$$

其中 $\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$ 为 $K\times 1$ 列向量与 $1\times K$ 行向量的积,$\mathbf{X}_iY_i$ 是 $K\times 1$ 列向量与 1×1 标量的积. 这样 OLS 估计量便可以求和符号表达式

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\mathbf{X}_iY_i. \tag{3.20}$$

由于 $\mathrm{SSE}(\boldsymbol{\beta})$ 可以用向量内积表示为

$$\mathrm{SSE}\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\left(Y_i-\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}\right)^2=\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right),$$

熟悉矩阵微积分的读者则可以直接对参数 $\boldsymbol{\beta}$ 求梯度矩阵, 得到一阶条件

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\boldsymbol{\beta}\right)=-2\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)=\mathbf{0}_{K\times 1},$$

这样亦可推导出式 (3.19) 的 OLS 估计量.

对于正规方程组 (3.18), 其等价于

$$\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\mathbf{0}\Leftrightarrow\left(\mathbf{X}_1,\cdots,\mathbf{X}_i,\cdots\mathbf{X}_n\right)\hat{\boldsymbol{\varepsilon}}=\mathbf{0}\Leftrightarrow\boxed{\sum_{i=1}^{n}\mathbf{X}_i\hat{\varepsilon}_i=\mathbf{0}}\Leftrightarrow\sum_{i=1}^{n}X_{ik}\hat{\varepsilon}_i=0,$$

依据严格外生性假设式 (3.5) 我们推知回归误差 ε_i 同任意的解释变量样本 X_{jk} 是正交的, 因而可以推知 $\mathbb{E}(\mathbf{X}\varepsilon_i)=0$, 其在 OLS 估计量的对应形式便是 $\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\mathbf{0}$. 特别地, 当线性投影模型含有常数项$X_{i1}\equiv 1$ 时, 即有残差和 $\sum\limits_{i=1}^{n}X_{i1}\hat{\varepsilon}_i=\sum\limits_{i=1}^{n}\hat{\varepsilon}_i=0$.

我们现在推导含有常数项时的 OLS 估计量, 我们依旧是利用中心化的方法. 设含有常数项 α 的样本个体有投影方程

$$Y_i=\alpha+\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i,$$

则求样本均值得 $\bar{Y}=\alpha+\bar{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}+\bar{\varepsilon}$, 其中 $\bar{\mathbf{X}}=\frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{X}_i,\bar{\varepsilon}=\frac{1}{n}\sum\limits_{i=1}^{n}\varepsilon_i$, 于是有样本个体有中心化的投影方程

$$Y_i-\bar{Y}=\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)^{\mathrm{T}}\boldsymbol{\beta}+(\varepsilon_i-\bar{\varepsilon})=\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)^{\mathrm{T}}\boldsymbol{\beta}+\mu_i,$$

这里我们使用 $\mu_i=\varepsilon_i-\bar{\varepsilon}$ 定义了新的投影误差, 于是依据式 (3.20) 得到含有常数项 α 时的 OLS 估计量为

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\left(\sum_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)\left(Y_i-\bar{Y}\right),\tag{3.21}$$

由于含有常数项时残差和为 0, 故常数项 α 的 OLS 估计量为 $\hat{\alpha}=\bar{Y}-\bar{\mathbf{X}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. 这与定理 (2.3.1) 是完全一致的.

定理 3.4.1 (OLS 估计量,OLS estimator)**.** *模型 (3.3.1) 的参数 $\boldsymbol{\beta}$ 的 OLS 估计量为*

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\quad\text{或}\quad\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\mathbf{X}_iY_i$$

且 OLS 估计量满足有

$$\hat{\boldsymbol{\varepsilon}}=\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}},\quad\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\mathbf{0}.$$

若是含有常数项 α 的模型 (3.3.2), 则 OLS 估计量为

$$\begin{aligned}\hat{\boldsymbol{\beta}}&=\left(\sum_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)\left(Y_i-\bar{Y}\right)\\\hat{\alpha}&=\bar{Y}-\bar{\mathbf{X}}^{\mathrm{T}}\hat{\boldsymbol{\beta}},\end{aligned}$$

且 OLS 估计量满足有

$$\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\mathbf{0},\quad\mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\mathbf{0}.$$

3.5 多元 OLS 估计量下 Guass-Markov 定理

本节我们将研究多元 OLS 估计量的有限样本性质, 本节以及之后, 我们假设线性投影模型 (3.3.1) 满足有假设 (3.2.1), 这时线性投影模型便等价于线性回归模型.

模型 3.5.1 (多元线性回归模型). *对于回归模型 (2.1.1), 在假设 (2.2.1) 以及*

$$\mathbb{E}(Y|\mathbf{X}) = \mathscr{P}(Y|\mathbf{X}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \quad \boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y)$$

*成立时, 我们称满足如下设定的模型称为多元***线性回归模型** *(linear regression model).*

$$\begin{aligned} Y &= \mathbb{E}(Y|\mathbf{X}) + \varepsilon \\ \mathbb{E}(\varepsilon|\mathbf{X}) &= 0 \\ \mathbb{E}(\mathbf{X}\varepsilon) &= 0 \\ \mathbb{E}(\varepsilon) &= 0 \quad (\text{不需要常数项}) \end{aligned}$$

多元线性回归模型 (3.5.1) 与多元线性投影模型 (3.3.1) 的最大区别在于, 前者的回归误差满足有 $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$, 但后者的投影误差则不然.

模型 (3.5.1) 是多元线性回归模型的统计总体的设定, 而样本的假设则遵从假设 (3.2.1). 本节以及后续, 我们并不设定样本为独立同分布的简单随机样本, 而是依据假设 (3.2.1) 这样更一般的设定, 两者的区别我们已经在 3.2 节辨析过了.

我们首先研究 OLS 估计量的代数性质. 由于 OLS 估计量为式 (3.19), 于是多元线性回归模型 (3.5.1) 的**抽样误差** (sampling error)即为

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{\beta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta} + \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = \boxed{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}}, \end{aligned}$$

这便表明抽样误差 $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ 是误差向量 ε 的线性组合.

我们再来计算 OLS 估计量的协方差矩阵, 由于 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\beta_1, \cdots, \beta_K)^{\mathrm{T}}$ 是一个 $K \times 1$ 维的**随机向量**, 则依据定理 (3.2.1) 得到

$$\begin{aligned} \operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) &= \operatorname{Var}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\middle|\boldsymbol{X}\right) \\ &= \left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right)\operatorname{Var}(\boldsymbol{Y}|\boldsymbol{X})\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\operatorname{Var}(\boldsymbol{Y}|\boldsymbol{X})\,\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}, \end{aligned}$$

而条件协方差矩阵 $\operatorname{Var}(\boldsymbol{Y}|\boldsymbol{X})$ 同样计算有

$$\boxed{\operatorname{Var}(\boldsymbol{Y}|\boldsymbol{X})} = \operatorname{Var}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}|\boldsymbol{X}) = \boxed{\operatorname{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})} = \boxed{\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)}.$$

这样, 对于模型 (3.5.1) 而言,OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的条件协方差矩阵为

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\sigma^{2}\boldsymbol{I}\right)\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\boxed{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^{2}},$$

这一结果依赖于条件同方差式 (3.7). 如果模型 (3.5.1) 具有异方差问题, 则依据式 (3.10) 得到

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},$$

由于矩阵乘法中乘以对角矩阵仅作用于对角线元素, 因此对于 $\boldsymbol{X}^{\mathrm{T}}=(\mathbf{X}_1,\cdots,\mathbf{X}_n)$ 不难得到

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}=\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\sigma_i^2,$$

故异方差下 $\hat{\boldsymbol{\beta}}$ 的条件协方差矩阵也可以写作

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)=\underbrace{\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}}_{K\text{阶方阵}}\underbrace{\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\sigma_i^2\right)}_{K\text{阶方阵}}\underbrace{\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}}_{K\text{阶方阵}}.\tag{3.22}$$

定理 3.5.1 (OLS 估计量的协方差矩阵)**.** *对于模型 (3.5.1) 的 OLS 估计量 $\hat{\boldsymbol{\beta}}$, 我们记 $\hat{\boldsymbol{\beta}}$ 的条件协方差为 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$, 则条件异方差式 (3.9) 下有*

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\tag{3.23}$$

当条件同方差式 (3.7) 成立时则有

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^{2}.\tag{3.24}$$

现在我们研究 OLS 估计量的有限性质. 对于一元的情形我们首先证明了 $\hat{\beta}$ 是 $Y_1,\cdots,Y_n$ 的线性组合, 在多元情形下 $\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 这是平凡的. 现在我们研究 OLS 估计量的有效性.

定理 3.5.2 (OLS 估计量的无偏性)**.** *模型 (3.5.1) 的 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 是参数 $\boldsymbol{\beta}$ 的无偏估计量.*

证明. 对抽样误差 $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$ 求条件期望有

$$\begin{aligned}&\mathbb{E}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)\\&=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)\xlongequal{\text{严格外生性式 (3.5)}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\cdot\mathbf{0}_{n\times1}=\mathbf{0}_{K\times1},\end{aligned}$$

则有 $\mathbb{E}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{\beta}|\boldsymbol{X}\right)=\boldsymbol{\beta}$. 当然亦可直接计算

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)=&\mathbb{E}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\middle|\boldsymbol{X}\right)\\&=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{X}\right)=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{\beta}.\end{aligned}$$

于是, 依据定理 (2.1.1) 得 $\mathbb{E}\hat{\boldsymbol{\beta}}=\mathbb{E}\left(\mathbb{E}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)\right)=\boldsymbol{\beta}$. □

我们再证明 OLS 估计量是模型 (3.5.1) 的**最优线性无偏估计量** (best linear unbiased estimator), 即 Gauss-Markov 定理.

定理 3.5.3 (Gauss-Markov 定理,Gauss-Markov theorem)**.** *设模型 (3.5.1) 的 OLS 估计量为 $\hat{\boldsymbol{\beta}}$, 且满足有*条件同方差式 *(3.7). 对于模型 (3.5.1) 的任意线性无偏估计量 $\tilde{\boldsymbol{\beta}}=\boldsymbol{A}\left(\boldsymbol{X}\right)\boldsymbol{Y}$, 则有*

$$\mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)\geqslant\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^{2},$$

这里 $\geqslant$ 表示 $\mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)-\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ 为半正定矩阵. 这也意味着 $\mathrm{Var}\left(\tilde{\beta}_k\middle|\boldsymbol{X}\right)\geqslant\mathrm{Var}\left(\hat{\beta}_k\middle|\boldsymbol{X}\right),k=1,\cdots,K$.

证明. 估计量 $\tilde{\boldsymbol{\beta}}=\boldsymbol{A}\left(\boldsymbol{X}\right)\boldsymbol{Y}$ 是参数 $\hat{\boldsymbol{\beta}}$ 的无偏估计量, 而

$$\mathbb{E}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{A}\left(\boldsymbol{X}\right)\boldsymbol{Y}|\boldsymbol{X}\right)=\boldsymbol{A}\mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{X}\right)=\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta},$$

故 $\tilde{\boldsymbol{\beta}}$ 满足有必要条件[3] $\boldsymbol{A}\boldsymbol{X}=\boldsymbol{I}_K$.

这样, 线性无偏估计量 $\tilde{\boldsymbol{\beta}}$ 的条件协方差为

$$\begin{aligned}\mathrm{Var}\left(\tilde{\beta}_k\middle|\boldsymbol{X}\right)&=\mathrm{Var}\left(\boldsymbol{A}\boldsymbol{Y}|\boldsymbol{X}\right)=\boldsymbol{A}\mathrm{Var}\left(\boldsymbol{Y}|\boldsymbol{X}\right)\boldsymbol{A}^{\mathrm{T}}\\&=\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^{\mathrm{T}}\overset{\text{条件同方差式 (3.7)}}{=\!=\!=\!=\!=\!=\!=\!=\!=\!=}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\sigma^{2}.\end{aligned}$$

于是,$\mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)-\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\sigma^{2}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^{2}=\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)\sigma^{2}$.

Guass 在 1823 年时便注意到可以将 $\boldsymbol{A}$ 分解为 $\boldsymbol{A}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$, 令 $\boldsymbol{C}=\boldsymbol{A}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$, 则有

$$\begin{aligned}&\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\left(\boldsymbol{C}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right)\left(\boldsymbol{C}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\=&\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}+\boldsymbol{C}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\=&\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}+\boldsymbol{A}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\=&\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\=&\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}},\end{aligned}$$

[3]严格的说, 依据 $\mathbb{E}\tilde{\boldsymbol{\beta}}=\mathbb{E}\left(\mathbb{E}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{A}\boldsymbol{X}\right)\boldsymbol{\beta}=\boldsymbol{\beta}$ 得到的必要条件为 $\mathbb{E}\left(\boldsymbol{A}\boldsymbol{X}\right)=\boldsymbol{I}_K$.

这样便将 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$ 转化为了 $K\times n$ 矩阵 $\boldsymbol{C}=\boldsymbol{A}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ 的 **Gram 矩阵**. 对于任意矩阵 $\boldsymbol{A}$, 其 Gram 矩阵 $\boldsymbol{G}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 是半正定的, 当且仅当 $\boldsymbol{A}$ 的列向量组线性无关时 Gram 矩阵 $\boldsymbol{G}$ 是正定的. 于是一定有

$$\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)-\boldsymbol{V}_{\hat{\beta}}=\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)\sigma^{2}\geqslant 0$$

成立. □

定理 (3.5.3) 表明了 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ 是最优线性无偏估计量, 而 Bruce(cite) 给出了现代形式的 Gauss-Markov 定理, 在条件同方差式 (3.7) 的假设下, 他证明了 OLS 估计量即是**最优无偏估计量** (best unbiased estimator), 是模型 (3.5.1) 参数 $\boldsymbol{\beta}$ 的 MVU 估计.

对于 OLS 有限样本的性质, 我们已经在一元形式讨论过了其意义, 并进行了 Monte Carlo 模拟以验证.

3.6 正交投影与 OLS 估计量

到目前为止, 本笔记所涉及的不少模型和概念, 都与“投影 (Projection)”这个词语相关——实质上, “投影”这个术语源自于线性代数. 本节将介绍线性代数中的正交投影的概念, 并介绍其与 OLS 估计量之间的关联.

在线性代数中, 对于向量空间 (或线性空间)V, 若线性变换 $\boldsymbol{f}:V\to V$ 满足有 $\boldsymbol{f}\circ\boldsymbol{f}=\boldsymbol{f}\left(\boldsymbol{f}\left(\cdot\right)\right)=\boldsymbol{f}$, 则称线性变换 $\boldsymbol{f}$ 为**投影** (projection). 如果 V 定义了内积 $\langle\cdot,\cdot\rangle$, 对于任意的 $\boldsymbol{x},\boldsymbol{y}\in V$, 投影 $\boldsymbol{f}$ 满足有 $\langle\boldsymbol{f}\left(\boldsymbol{x}\right),\boldsymbol{y}\rangle=\langle\boldsymbol{x},\boldsymbol{f}\left(\boldsymbol{y}\right)\rangle$, 则称投影 $\boldsymbol{f}$ 为**正交投影** (orthogonal projection). 显然, 这里的投影是个一般化的抽象概念, 它反映了投影在代数上的本质, 但对于我们理解线性模型中出现了投影基本上没有帮助.

为了理解上述的抽象定义, 我们回顾中学数学中的向量投影的概念, 但加以线性代数的语言来叙述. 设想一个如图 (3.1) 的情景, 已知二维平面 $\mathbb{R}^2$ 中的一条过原点的直线 $\boldsymbol{l}$ 和一点 $\boldsymbol{x}$, 由于他们都有显式的二维坐标, 故我们可以直接使用列向量 $\boldsymbol{l}$ 和 $\boldsymbol{x}$ 来表示二者. 在中学数学中, 我们知道点 $\boldsymbol{x}$ 在直线 $\boldsymbol{l}$ 上的投影, 便是 $\boldsymbol{l}$ 上的与点 $\boldsymbol{x}$ 距离最小的一点. 在二维平面中, 我们用内积定义距离, 也即是 Euclid 距离.

依据勾股定理, 不难得知距离最小意味着点 $\boldsymbol{x}$ 与其投影点的连线垂直于直线 $\boldsymbol{l}$, 否则, 总可以找到使得距离更小的投影点. 利用平面几何的知识, 我们可以证明将点 $\boldsymbol{x}$ 投影至直线 $\boldsymbol{l}$ 是一个线性变换, 而线性变换可以用矩阵乘法表示, 这样我们不妨记投影矩阵 $\boldsymbol{P}_{\boldsymbol{l}}\in\mathbb{R}^{2\times 2}$, 则点 $\boldsymbol{x}$ 在直线 $\boldsymbol{l}$ 上的投影点为 $\boldsymbol{P}_{\boldsymbol{l}}\boldsymbol{x}$.

于是我们描绘了如图 (3.1) 所示的 (正交) 投影, 这一情形是可以自然推广至更高维的情形. 现思考一个 n 维的向量空间 $\mathbb{R}^n$, 则其中的点 $\boldsymbol{x}\in\mathbb{R}^n$ 包含有 n 个分量 (视为列向量). 在高维时与二维平面中过原点直线 $\boldsymbol{l}$ 所对应的存在, 数学中称之为**超平面**

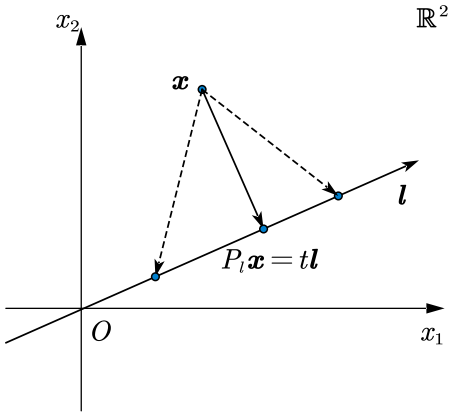

图 3.1: 二维平面中的正交投影

(hyperplane), 其是 $\mathbb{R}^n$ 的一个子空间[4], 我们仍旧可以记为列向量 $\boldsymbol{l}\in\mathbb{R}^n$.

为了确定投影矩阵 $\boldsymbol{P_l}$ 的具体形式, 我们定义

$$\begin{aligned}\boldsymbol{P_l}=\underset{\boldsymbol{P_l}\in\mathbb{R}^{n\times n}}{\operatorname{argmin}}\left\|\boldsymbol{P_l}\boldsymbol{x}-\boldsymbol{x}\right\|=\underset{\boldsymbol{P_l}\in\mathbb{R}^{n\times n}}{\operatorname{argmin}}\sqrt{\left(\boldsymbol{P_l}\boldsymbol{x}-\boldsymbol{x}\right)^{\mathrm{T}}\left(\boldsymbol{P_l}\boldsymbol{x}-\boldsymbol{x}\right)},\\ s.t.\quad \boldsymbol{P_l}\boldsymbol{x}=\boldsymbol{l}t,\quad t\in\mathbb{R},\end{aligned}\tag{3.25}$$

其中 $\boldsymbol{P_l}\boldsymbol{x}=t\boldsymbol{l}$ 为约束条件, 使得投影点在 $\boldsymbol{l}$ 之上. 不难发现, 式 (3.25) 与 OLS 估计量式 (3.17) 在形式上是相似的. 为了求解式 (3.25), 一种思路即是 Lagrange 乘数法, 但式 (3.25) 的约束条件非常简单, 直接代入优化函数得

$$s\left(t\right)=\left(\boldsymbol{l}t-\boldsymbol{x}\right)^{\mathrm{T}}\left(\boldsymbol{l}t-\boldsymbol{x}\right)=t^2\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}-2t\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x},$$

对 $s\left(t\right)$ 求导得到一阶条件

$$\frac{\mathrm{d}}{\mathrm{d}t}s\left(t\right)=2t\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}-2\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}=0\quad\Rightarrow,\quad t\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}=\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x},$$

解得 $t=\left(\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}\right)^{-1}\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}$. 故有 $\boldsymbol{P_l}\boldsymbol{x}=\boldsymbol{l}t=\boldsymbol{l}\left(\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}\right)^{-1}\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}$, 则将点 $\boldsymbol{x}$ 投影至 $\boldsymbol{l}$ 上的线性变换可以用投影矩阵

$$\boldsymbol{P_l}=\boldsymbol{l}\left(\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}\right)^{-1}\boldsymbol{l}^{\mathrm{T}}\tag{3.26}$$

表示. 不难发现, 投影点 $\boldsymbol{P_l}\boldsymbol{x}$ 至点 $\boldsymbol{x}$ 的向量 $\boldsymbol{x}-\boldsymbol{P_l}\boldsymbol{x}$ 与 $\boldsymbol{l}$ 是正交的, 即

$$\boldsymbol{l}^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{P_l}\boldsymbol{x}\right)=\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{l}^{\mathrm{T}}\boldsymbol{P_l}\boldsymbol{x}=\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}\left(\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}\right)^{-1}\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{l}^{\mathrm{T}}\boldsymbol{x}=0,$$

因此,$\boldsymbol{P_l}\boldsymbol{x}$ 称为点 $\boldsymbol{x}$ 在 $\boldsymbol{l}$ 上的正交投影.

[4] $\boldsymbol{l}$ 过原点确保了子空间是封闭的.

对于超平面 $\boldsymbol{l}$, 我们进一步考虑其上的 K 个向量 $\boldsymbol{l}_1, \cdots, \boldsymbol{l}_K$, 由于 $\boldsymbol{l}$ 是向量空间 $\mathbb{R}^n$ 的子空间, 则向量组 $\{\boldsymbol{l}_1, \cdots, \boldsymbol{l}_k\}$ 的线性组合仍属于超平面 $\boldsymbol{l}$. 这一过程也是线性变换, 记矩阵 $\boldsymbol{L} = (\boldsymbol{l}_1, \cdots, \boldsymbol{l}_k) \in \mathbb{R}^{n\times K}$, 则 $\boldsymbol{L}$ 与任意的 k 维列向量的积仍属于超平面内, 但乘积蕴含了向量组 $\{\boldsymbol{l}_1, \cdots, \boldsymbol{l}_k\}$ 所**张成** (span)的向量空间. 我们将式 (3.26) 中的 $\boldsymbol{l}$ 替换为 $\boldsymbol{L}$, 则得到 n 阶方阵

$$\boldsymbol{P}_{\boldsymbol{L}} = \boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}, \tag{3.27}$$

则式 (3.27) 是点 $\boldsymbol{x}$ 投影至 $\{\boldsymbol{l}_1, \cdots, \boldsymbol{l}_k\}$ 所张成向量空间的**投影矩阵** (projection matrix).

这样, 我们便可以理解为什么 OLS 估计量与投影存在关联了. 模型 (3.5.1) 的 OLS 估计量为 $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, 对应的拟合值为

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y},$$

依据上文分析即可得知, 拟合值 $\hat{\boldsymbol{Y}}$ 实际上是被解释变量的样本 $\boldsymbol{Y}$ 在数据矩阵 $\boldsymbol{X}$ 的列向量组所张成的线性空间上的正交投影. 实际上, 投影是一种简化 n 维信息的方法, 其将信息压缩到一个 (超) 平面上. 这在社会科学领域尤其有用, 因为我们所研究的现象非常复杂，不可能做出准确的预测. 相反, 我们往往希望构建一些模型, 由繁杂多变的数据得到更简单、更合理的解释, 投影便是实现这一目标的统计方法.

我们通过实例来理解 OLS 估计量的投影性质. 考虑一元线性回归方程 (2.23), 且有如表 (3.1) 所示的样本观测值, 则有

$$\boldsymbol{Y} = \begin{pmatrix} -0.5 \\ 1.25 \\ 2 \end{pmatrix}, \quad \boldsymbol{X}^* = \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} = (\boldsymbol{X}, \boldsymbol{1}).$$

表 3.1: 正交投影示例观测值

y	x	常数项
-0.5	-1	1
1.25	1	1
2	0	1

我们将数据向量 $\boldsymbol{Y}$ 和数据矩阵 $\boldsymbol{X}^*$ 的列向量组都视为 $\mathbb{R}^3$ 中的列向量, 则数据矩阵 $\boldsymbol{X}^*$ 可以视为一个线性变换, 其即是对应了列向量 $\boldsymbol{X}$ 和 $\boldsymbol{1}$ 所张成的过 $O(0,0,0)$ 的平面 $\boldsymbol{L}$. 由于 OLS 估计量和拟合值为

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{pmatrix} = \left(\left(\boldsymbol{X}^*\right)^{\mathrm{T}}\boldsymbol{X}^*\right)^{-1}\left(\boldsymbol{X}^*\right)^{\mathrm{T}}\boldsymbol{Y} = \begin{pmatrix} 0.8750 \\ 0.9167 \end{pmatrix}, \quad \hat{\boldsymbol{Y}} = \boldsymbol{X}^*\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{pmatrix} = \begin{pmatrix} 0.0417 \\ 1.7917 \\ 0.9167 \end{pmatrix}$$

我们在图 (3.2) 中绘制了样本点以及样本回归直线. 依据投影的视角, 拟合值向量 $\hat{\boldsymbol{Y}}$ 是平面 $\boldsymbol{L}$ 中的向量, 残差向量 $\hat{\boldsymbol{\varepsilon}}$ 即是连接了 $\hat{\boldsymbol{Y}}$ 和 $\boldsymbol{Y}$ 两者终点的向量, 且残差向量 $\hat{\boldsymbol{\varepsilon}}$ 同平面 $\boldsymbol{L}$ 是垂直的, 这样结果如图 (3.3) 所示.

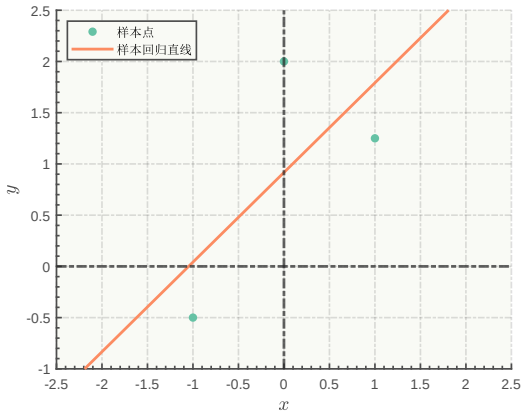

图 3.2: 一元线性回归模型的拟合直线

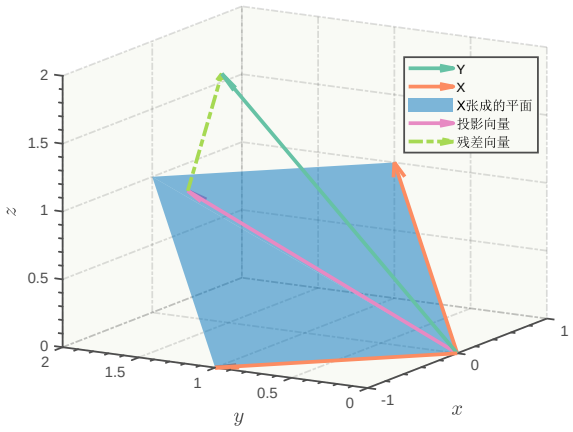

图 3.3: 一元线性回归模型的正交投影

下面我们研究式 (3.27) 的投影矩阵的性质.

定理 3.6.1 (投影矩阵的性质). *对于式 (3.27) 的投影矩阵 $\boldsymbol{P}_{\boldsymbol{L}}=\boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}},\boldsymbol{L}\in\mathbb{R}^{n\times K},n\geqslant K$, 其满足有如下性质:*

(a) 投影矩阵 $\boldsymbol{P}_{\boldsymbol{L}}$ 为幂等矩阵, 即 $\boldsymbol{P}^2=\boldsymbol{P}$, 且 $\boldsymbol{P}_{\boldsymbol{L}}\boldsymbol{L}=\boldsymbol{L}$;

(b) 投影矩阵 $\boldsymbol{P}_{\boldsymbol{L}}$ 是对称矩阵, 即 $\boldsymbol{P}_{\boldsymbol{L}}^{\mathrm{T}}=\boldsymbol{P}_{\boldsymbol{L}}$;

(c) 投影矩阵的 $\boldsymbol{P}_{\boldsymbol{L}}$ 的迹为 K;

(d) 投影矩阵 $\boldsymbol{P}_{\boldsymbol{L}}$ 有 K 个特征值为 1, 余下的 $n-K$ 个特征值为 0;

(e) 投影矩阵 $\boldsymbol{P}_{\boldsymbol{L}}$ 的秩为 K.

证明. (a) 计算有 $\boldsymbol{P}_{\boldsymbol{L}}^2=\boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}=\boldsymbol{P}_{\boldsymbol{L}}$, 这即是符合本节开头所讲的 $\boldsymbol{f}\circ\boldsymbol{f}=\boldsymbol{f}$, 复合映射两次后的结果相同. 而 $\boldsymbol{P}_{\boldsymbol{L}}\boldsymbol{L}=\boldsymbol{L}$ 也是平凡的, 其含义便是将 $\boldsymbol{L}$ 投影至 $\boldsymbol{L}$ 上, 在投影后距离最小的约束条件下则 $\boldsymbol{P}_{\boldsymbol{L}}\boldsymbol{L}$ 自然是 $\boldsymbol{L}$ 本身.

(b) 这也是平凡的, 这条性质保证了投影是正交的, 注意到

$$\boldsymbol{L}^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{P}_{\boldsymbol{L}}\boldsymbol{x}\right)=\left(\boldsymbol{P}_{\boldsymbol{L}}\boldsymbol{L}\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{P}_{\boldsymbol{L}}\boldsymbol{x}\right)=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{L}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{L}}\boldsymbol{x}=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{L}}^{\mathrm{T}}\left(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{L}}\right)\boldsymbol{x},$$

则 $\boldsymbol{P}_{\boldsymbol{L}}^{\mathrm{T}}=\boldsymbol{P}_{\boldsymbol{L}}$ 是 $\boldsymbol{L}^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{P}_{\boldsymbol{L}}\boldsymbol{x}\right)=0$ 的充分条件.

性质 (c) 至 (e) 则是线性代数中多个命题的综合, 本笔记下面证明中所使用的结论, 相应地都可以在线性代数的基础教材中找到.

首先, 对于投影矩阵 $\boldsymbol{P_L}$, 其特征值与特征向量即是关于实数 λ 和非零列向量 $\boldsymbol{x}$ 的矩阵方程 $\boldsymbol{P_L x} = \lambda \boldsymbol{x}$ 的解, 由于 $\boldsymbol{P_L}$ 是幂等矩阵, 则有

$$\boldsymbol{P_L x} = \boldsymbol{x}\lambda, \Rightarrow \boldsymbol{P_L P_L x} = \boldsymbol{P_L x}\lambda, \Rightarrow \boldsymbol{P_L x} = \boldsymbol{x}\lambda \cdot \lambda, \Rightarrow \boldsymbol{x}\lambda = \boldsymbol{x}\lambda^2, \Rightarrow \boldsymbol{x}\lambda\left(1-\lambda\right) = 0,$$

由于 $\boldsymbol{x}$ 是非零向量, 故投影矩阵 $\boldsymbol{P_L}$ 的特征值只能是 0 或 1.

这样, 我们计算投影矩阵投影矩阵 $\boldsymbol{P_L}$ 的迹为

$$\operatorname{tr}\left(\boldsymbol{P_L}\right) = \operatorname{tr}\left(\boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}\right) = \operatorname{tr}\left(\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right) = \operatorname{tr}\left(\boldsymbol{I}_K\right) = K,$$

即证性质 (c). 而方阵的特征值的和又满足有

$$\sum_{i=1}^{n}\lambda_i = \operatorname{tr}\left(\boldsymbol{P_L}\right) = k,$$

则投影矩阵 $\boldsymbol{P_L}$ 的特征值必有 K 个等于 1, 余下的 $n-K$ 个等于 0, 即证性质 (d).

最后, 由于投影矩阵 $\boldsymbol{P_L}$ 为实对称方阵, 且其特征值非负, 这等价于 $\boldsymbol{P_L}$ 为半正定矩阵, 而半正定矩阵的秩等于其非负特征值的数量, 即得 $\operatorname{rank}\boldsymbol{P_L} = K$, 性质 (e) 证毕. □

在一元线性回归模型 (2.4.1), 我们粗略地得到投影矩阵后便定义了回归矩阵, 在此我们同样定义

$$\boldsymbol{M_L} = \boldsymbol{I}_n - \boldsymbol{P_L} \tag{3.28}$$

为**归零矩阵** (annihilator matrix), 其具有如下的性质.

定理 3.6.2 (归零矩阵的性质). *对于形如式 (3.28) 的归零矩阵 $\boldsymbol{M_L} = \boldsymbol{I}_n - \boldsymbol{P_L}$, 其具有如下性质:*

(a) $\boldsymbol{M_L L} = \boldsymbol{O}_{n\times K}, \boldsymbol{M P_L} = \boldsymbol{O}_n$;

(b) 归零矩阵 $\boldsymbol{M_L}$ *为幂等矩阵, 即* $\boldsymbol{M}^2 = \boldsymbol{M}$;

(c) 归零矩阵 $\boldsymbol{M_L}$ *是对称矩阵, 即* $\boldsymbol{M}^{\mathrm{T}} = \boldsymbol{M}$;

(d) 归零矩阵的 $\boldsymbol{M_L}$ *的迹为* $n-K$;

(e) 归零矩阵 $\boldsymbol{M_L}$ *有* $n-K$ *个特征值为 1, 余下的* n *个特征值为 0;*

(f) 归零矩阵 $\boldsymbol{M_L}$ *的秩为* $n-K$.

证明. 性质 (a) 至 (c) 的是平凡的, 余下的性质与定理 (3.6.1) 后三条性质的证明同理. 相信读者自证不难. □

最后, 本节介绍线性代数中实对称方阵的谱分解以及应用, 这在后续章节中会使用.

定理 3.6.3 (实对称方阵的谱分解)**.** 设任意的实对称方阵 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, 则其一定可以通过**正交矩阵** *(orthogonal matrix)*相似至对角阵 $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$, 即有

$$\boldsymbol{A} = \boldsymbol{H}^{-1}\boldsymbol{\Lambda}\boldsymbol{H},$$

其中 $\boldsymbol{\Lambda}$ 的元素为 $\boldsymbol{A}$ 的特征值, 正交矩阵 $\boldsymbol{H}$ 满足有 $\boldsymbol{H}^{-1} = \boldsymbol{H}^{\mathrm{T}}$, 我们称此为实对称方阵 $\boldsymbol{A}$ 的或**谱分解** *(spectral decomposition)*.

定理 (3.6.3) 可以在基础的线性代数教材上找到 (cite), 其一个有用的应用便是考虑幂等的 n 阶实对称方阵 $\boldsymbol{A}$, 设 $\operatorname{rank}\boldsymbol{A} = r$, 则有

$$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) = \underbrace{\operatorname{diag}(1, \cdots, 1, 0 \cdots, 0)}_{r\text{个}1, n-r\text{个}0},$$

于是

$$\boldsymbol{A} = \boldsymbol{A}^2 = \boldsymbol{H}^{-1}\boldsymbol{\Lambda}\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{\Lambda}\boldsymbol{H} = \boldsymbol{H}^{-1}\boldsymbol{\Lambda}^2\boldsymbol{H} = \boldsymbol{H}^{-1}\boldsymbol{\Lambda}\boldsymbol{H},$$

亦即有

$$\boldsymbol{A} = \boldsymbol{H}^{-1}\begin{pmatrix} \boldsymbol{I}_r & \boldsymbol{O}_{r\times(n-r)} \\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{O}_{n-r} \end{pmatrix}\boldsymbol{H}. \tag{3.29}$$

定理 (3.6.3) 的另一个应用便是正定矩阵的平方根. 对于任意的方阵 $\boldsymbol{A}$, 若存在方阵 $\boldsymbol{B}$ 使得 $\boldsymbol{A} = \boldsymbol{B}^2 = \boldsymbol{B}\cdot\boldsymbol{B}$, 则称方阵 $\boldsymbol{B}$ 是**矩阵 $\boldsymbol{A}$ 的平方根** (square root of the matrix $\boldsymbol{A}$), 特别地, 如果方阵 $\boldsymbol{A}$ 是正定的, 则有如下命题成立.

定理 3.6.4 (正定矩阵的平方根)**.** 对于任意的正定矩阵 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, 则存在唯一的平方根 $\boldsymbol{B}$, 使得 $\boldsymbol{A} = \boldsymbol{B}^2$ 成立. 记有 $\boldsymbol{B} = \boldsymbol{A}^{1/2}$. 且有

$$\boldsymbol{A}^{1/2} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H},$$

这里的 $\boldsymbol{H}$ 为使得 $\boldsymbol{A}$ 相似至对角阵 $\boldsymbol{\Lambda}$ 的正交矩阵,$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ 为 $\boldsymbol{A}$ 的特征值构成的对角阵,$\boldsymbol{\Lambda}^{1/2} = \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_n}\right)$.

证明. 对于任意的 n 阶正定的实对称方阵 $\boldsymbol{A}$, 即有 $\operatorname{rank}A = n$, 那么其所有的特征值都是正数, 依据定理 (3.6.3) 即有 $\boldsymbol{A} = \boldsymbol{H}^{-1}\boldsymbol{\Lambda}\boldsymbol{H}, \boldsymbol{H}^{-1} = \boldsymbol{H}^{\mathrm{T}}$. 不难注意到

$$\begin{aligned}\boldsymbol{A} =& \boldsymbol{H}^{-1}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{I}_n\boldsymbol{\Lambda}^{1/2}\boldsymbol{H} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H} \\ =& \left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H}\right)\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H}\right) = \left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H}\right)^2,\end{aligned}$$

令 $\boldsymbol{B} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H}$, 则有 $\boldsymbol{A} = \boldsymbol{B}^2$ 成立, 亦即有 $\boldsymbol{A}^{1/2} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H}$. 显然, 正定矩阵 $\boldsymbol{A}$ 的平方根是实对称方阵. 我们还需要证明 $\boldsymbol{A}^{1/2}$ 的唯一性, 这个问题留给读者思考. □

平凡地, 对于任意的正定的对角矩阵 $\boldsymbol{\Lambda}=\operatorname{diag}\left(a_1,\cdots,a_n\right)$, 其所有对角元都是正数, 则 $\boldsymbol{\Lambda}$ 的平方根为 $\boldsymbol{\Lambda}^{1/2}=\operatorname{diag}\left(\sqrt{a_1},\cdots,\sqrt{a_n}\right)$.

3.7 多元 OLS 估计量下拟合值、残差

本节将利用上一节介绍的投影矩阵, 研究 OLS 估计量的**拟合值** (fitted value)向量 $\hat{\boldsymbol{Y}}$ 和**残差** (residual)向量 $\hat{\boldsymbol{\varepsilon}}$ 的性质.

正如我们在上一章研究过 $\hat{\boldsymbol{Y}}$ 和残差向量 $\hat{\boldsymbol{\varepsilon}}$, 利用投影矩阵 $\boldsymbol{P}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ 和归零矩阵 $\boldsymbol{M}=\boldsymbol{I}-\boldsymbol{P}$ 得

$$\hat{\boldsymbol{Y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\boxed{\boldsymbol{P}\boldsymbol{Y}},$$

$$\hat{\boldsymbol{\varepsilon}}=\hat{\boldsymbol{Y}}-\boldsymbol{Y}=\boldsymbol{P}\boldsymbol{Y}-\boldsymbol{Y}=\boxed{\boldsymbol{M}\boldsymbol{Y}}=\boldsymbol{M}\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right)=\mathbf{0}\cdot\boldsymbol{\beta}+\boldsymbol{M}\boldsymbol{\varepsilon}=\boxed{\boldsymbol{M}\boldsymbol{\varepsilon}},$$

且拟合值向量 $\hat{\boldsymbol{Y}}$ 和残差向量 $\hat{\boldsymbol{\varepsilon}}$ 是正交的, 这是因为

$$\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\left(\boldsymbol{P}\boldsymbol{Y}\right)^{\mathrm{T}}\left(\boldsymbol{M}\boldsymbol{Y}\right)=\boldsymbol{Y}^{\mathrm{T}}\left(\boldsymbol{P}\boldsymbol{M}\right)\boldsymbol{Y}=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{O}_n\boldsymbol{Y}=0.$$

对于投影矩阵 $\boldsymbol{P}$, 其第 i 个对角线元素我们记为 h_{ii}, 则由

$$\boldsymbol{P}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}=\begin{pmatrix}\mathbf{X}_1^{\mathrm{T}}\\ \vdots\\ \mathbf{X}_n^{\mathrm{T}}\end{pmatrix}\underbrace{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}}_{n\text{阶方阵}}\left(\mathbf{X}_1,\cdots,\mathbf{X}_n\right)$$

不难推知

$$h_{ii}=\mathbf{X}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbf{X}_i,$$

我们将 h_ii 称为**杠杆值** (leverage value).

定理 3.7.1 (杠杆值的性质). *对于投影矩阵 $\boldsymbol{P}$ 得对角线元素 h_{ii}, 杠杆值 h_{ii} 满足如下性质.*

(a) $0\leqslant h_{ii}\leqslant 1$;

(b) 当模型 (3.5.1) 含有常数项时, $h_{ii}\geqslant n^{-1}$;

(c) $\sum\limits_{i=1}^{n}h_{ii}=\operatorname{tr}\boldsymbol{P}=K$.

现在我们来研究拟合值向量 $\hat{\boldsymbol{Y}}$ 和残差向量 $\hat{\boldsymbol{\varepsilon}}$ 的有限样本性质. 首先, 对于 $\hat{\boldsymbol{Y}}$ 其条件期望、方差分别为

$$\mathbb{E}\left(\hat{\boldsymbol{Y}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{P}\boldsymbol{Y}\middle|\boldsymbol{X}\right)=\boldsymbol{P}\mathbb{E}\left(\boldsymbol{Y}\middle|\boldsymbol{X}\right),$$

$$\operatorname{Var}\left(\hat{\boldsymbol{Y}}\middle|\boldsymbol{X}\right)=\boldsymbol{P}\operatorname{Var}\left(\boldsymbol{Y}\middle|\boldsymbol{X}\right)\boldsymbol{P}^{\mathrm{T}}\overset{\text{异方差假设式 (3.9)}}{=\!=\!=\!=\!=\!=\!=\!=\!=\!=}\boldsymbol{P}\boldsymbol{D}\boldsymbol{P},$$

在条件同方差式 (3.7) 成立时, 便可以得到

$$\operatorname{Var}\left(\hat{\boldsymbol{Y}}\middle|\boldsymbol{X}\right)=\boldsymbol{P}\left(\boldsymbol{I}\sigma^2\right)\boldsymbol{P}=\boldsymbol{P}\sigma^2,$$

则有 $\boxed{\operatorname{Var}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)=h_{ii}\sigma^2}$.

对于残差向量 $\hat{\boldsymbol{\varepsilon}}$, 同样计算条件期望、方差得

$$\mathbb{E}\left(\hat{\boldsymbol{\varepsilon}}|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}\right)=\boldsymbol{M}\mathbb{E}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right)\overset{\text{严格外生性式 (3.5)}}{=\!=\!=}\boldsymbol{0},$$

$$\operatorname{Var}\left(\hat{\boldsymbol{\varepsilon}}|\boldsymbol{X}\right)=\operatorname{Var}\left(\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}\right)=\boldsymbol{M}\operatorname{Var}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right)\boldsymbol{M}^{\mathrm{T}}\overset{\text{异方差假设式 (3.9)}}{=\!=\!=}\boldsymbol{M}\boldsymbol{D}\boldsymbol{M},$$

当条件同方差式 (3.7) 成立时, 则有

$$\operatorname{Var}\left(\hat{\boldsymbol{\varepsilon}}|\boldsymbol{X}\right)=\boldsymbol{M}\left(\boldsymbol{I}\sigma^2\right)\boldsymbol{M}=\boldsymbol{M}\sigma^2,$$

此时残差 $\hat{\varepsilon}_i$ 的条件方差为 $\boxed{\operatorname{Var}\left(\hat{\varepsilon}_i|\boldsymbol{X}\right)=\left(1-h_{ii}\right)\sigma^2}$.

与第一章中计算模型 (2.4.2) 的拟合值、残差相比, 我们刚才利用矩阵工具来运算, 要便捷的多. 我们在第一章中提出了**标准化残差** (standardized residual)为 $\bar{\varepsilon}_i=\left(1-h_{ii}\right)^{-\frac{1}{2}}\hat{\varepsilon}_i$, 则标准化残差的向量形式可以表示成

$$\bar{\boldsymbol{\varepsilon}}=\operatorname{diag}\left(\frac{1}{\sqrt{1-h_{11}}},\cdots,\frac{1}{\sqrt{1-h_{11}}}\right)\boldsymbol{M}\boldsymbol{\varepsilon},$$

记 $\boldsymbol{M}^*=\operatorname{diag}\left(\frac{1}{\sqrt{1-h_{11}}},\cdots,\frac{1}{\sqrt{1-h_{11}}}\right)$, 则标准化残差为 $\bar{\boldsymbol{\varepsilon}}=\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{\varepsilon}$. 对于标准化残差, 计算条件期望、方差得

$$\mathbb{E}\left(\bar{\boldsymbol{\varepsilon}}|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}\right)=\boldsymbol{M}^*\boldsymbol{M}\mathbb{E}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right)=\boldsymbol{0}$$

与

$$\begin{aligned}\operatorname{Var}\left(\bar{\boldsymbol{\varepsilon}}|\boldsymbol{X}\right)&=\operatorname{Var}\left(\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{\varepsilon}|\boldsymbol{X}\right)=\boldsymbol{M}^*\boldsymbol{M}\operatorname{Var}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right)\boldsymbol{M}\boldsymbol{M}^*\\&\overset{\text{异方差假设式 (3.9)}}{=\!=\!=}\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*\overset{\text{同方差假设式 (3.7)}}{=\!=\!=}\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}^*\sigma^2,\end{aligned}$$

则

$$\operatorname{Var}\left(\bar{\varepsilon}_i|\boldsymbol{X}\right)=\frac{1}{\sqrt{1-h_{ii}}}\cdot\left(1-h_{ii}\right)\sigma^2\cdot\frac{1}{\sqrt{1-h_{ii}}}=\sigma^2,$$

可见标准化残差 $\bar{\varepsilon}_i$ 与回归误差 ε_i 具有相同的期望和方差.

3.8 分块回归与残差回归

本节考虑分块回归, 这一技巧或工具对于我们研究所关心的回归系数时十分有用, 并由此衍生出残差回归的方法.

在实证研究中, 我们通常仅关心一个或几个关键的感兴趣的回归系数, 而其余的回归系数则是作为协变量引入, 不是我们所重点关注的, 在这种情形下, 我们自然会考虑那少数几个回归系数的 OLS 估计量是何种的形式, 因而可以去研究其他因素对于估计的影响 (例如第6.13节中近似多重共线性对 OLS 估计量的影响、第八章中介绍的内生性问题的处理).

考虑样本的线性回归模型 $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, 置分块 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2)$, 其中 $\boldsymbol{X}_1\in\mathbb{R}^{n\times K_1}$ 包含了 K_1 个解释变量 $\mathbf{X}_1$ 的样本,$\boldsymbol{X}_2\in\mathbb{R}^{n\times K_2}$ 包含了 K_2 个解释变量 $\mathbf{X}_2$ 的样本, 而 $K_1+K_2=K$. 这样, 各分块对应的回归系数亦有分块 $\boldsymbol{\beta}=\left(\boldsymbol{\beta}_1^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}}\right)^{\mathrm{T}}$, 其中 $\boldsymbol{\beta}_1\in\mathbb{R}^{K_1\times 1},\boldsymbol{\beta}_2\in\mathbb{R}^{K_2\times 1},K_1+K_2=K$. 于是我们有**分块回归** (partitioned regression)或 **regression components**的方程

$$\boldsymbol{Y}=(\boldsymbol{X}_1,\boldsymbol{X}_2)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}+\boldsymbol{\varepsilon}=\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{X}_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}. \tag{3.30}$$

分块回归式 (3.30) 的总体与多元线性回归模型是相同的, 因而我们同样使用 OLS 估计量, 可以得到回归系数分块的估计量 $\hat{\boldsymbol{\beta}}=\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{T}},\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}\right)^{\mathrm{T}},\hat{\boldsymbol{\beta}}_1\in\mathbb{R}^{K_1\times 1},\hat{\boldsymbol{\beta}}_2\in\mathbb{R}^{K_2\times 1}$, 我们现在研究的是对于单个感兴趣的分块回归系数 $\boldsymbol{\beta}_1$ 或 $\boldsymbol{\beta}_2$ 的估计量.

既然我们使用 OLS 估计量去估计式 (3.30), 则被解释变量可以被分解为拟合值与残差之和, 即 $\boldsymbol{Y}=\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1+\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2+\hat{\boldsymbol{\varepsilon}}$, 且在误差平方和 SSE 最小的准则下, 我们有分块估计量

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{pmatrix}=\underset{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}{\operatorname{argmin}}\,\mathrm{SSE}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right)=\underset{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}{\operatorname{argmin}}\left(\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2\right), \tag{3.31}$$

假设参数 $\boldsymbol{\beta}_1$ 是我们更感兴趣的参数, 我们希望得到其单独的 OLS 估计量 $\hat{\boldsymbol{\beta}}_1$. 为此这里的关键在于, 式 (3.31) 可以等价地考虑为

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{pmatrix}=\underset{\boldsymbol{\beta}_1}{\operatorname{argmin}}\left(\min_{\boldsymbol{\beta}_2}\mathrm{SSE}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right)\right), \tag{3.32}$$

其中 $\min_{\boldsymbol{\beta}_2}\mathrm{SSE}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right)$ 表示固定 $\boldsymbol{\beta}_1$ 时关于 $\boldsymbol{\beta}_2$ 的函数 SSE 所能取得的最小值. 在此之上进一步求解关于 $\boldsymbol{\beta}_1$ 的 SSE 的最优化问题, 我们便得到了全局最小的 SSE, 因而得到了 OLS 估计量. 对于 $\min_{\boldsymbol{\beta}_2}\mathrm{SSE}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right)$, 依据式 (3.31) 中的 SSE, 固定了 $\boldsymbol{\beta}_1$ 我们便可以视之为$\underline{\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1\text{ 对 }\boldsymbol{X}_2\text{ 的回归}}$, 这样依据 OLS 估计量的一般形式 $\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, 依据式 (3.32) 我们得

$$\tilde{\boldsymbol{\beta}}_2=\underset{\boldsymbol{\beta}_2}{\operatorname{argmin}}\,\mathrm{SSE}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right)=\left(\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1\right), \tag{3.33}$$

我们使用 $\tilde{}$ 表示与真正的 OLS 估计量区分, 因为 $\tilde{\boldsymbol{\beta}}_2$ 还与 $\boldsymbol{\beta}_1$ 的取值有关. 回顾第3.7节的内容, 在代数上我们使用归零矩阵将残差向量表示为 $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{M}\boldsymbol{Y}$, 则对于式 (3.33), 记

其有分块残差向量

$$\hat{\boldsymbol{\varepsilon}}_2 = (\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1) - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{M}_2(\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1), \tag{3.34}$$

其中归零矩阵为

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{P}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2\left(\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^{\mathrm{T}}, \tag{3.35}$$

故我们得到

$$\min_{\boldsymbol{\beta}_2} \mathrm{SSE}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \hat{\boldsymbol{\varepsilon}}_2^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}_2 = (\boldsymbol{M}_2\boldsymbol{Y} - \boldsymbol{M}_2\boldsymbol{X}_1\boldsymbol{\beta}_1)^{\mathrm{T}}(\boldsymbol{M}_2\boldsymbol{Y} - \boldsymbol{M}_2\boldsymbol{X}_1\boldsymbol{\beta}_1). \tag{3.36}$$

现在我们来研究最优化问题 $\underset{\boldsymbol{\beta}_1}{\mathrm{argmin}}\left(\min_{\boldsymbol{\beta}_2}\mathrm{SSE}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\right)$. 注意到式 (3.36) 仍旧是误差平方和的形式, 因而我们将式 (3.32) 外面的最优化问题视为$\underline{\boldsymbol{M}_2\boldsymbol{Y} \text{ 对 } \boldsymbol{M}_2\boldsymbol{X}_1 \text{ 的回归}}$, 即得分块估计系数 $\boldsymbol{\beta}_1$ 的 OLS 估计量为

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= \underset{\boldsymbol{\beta}_1}{\mathrm{argmin}}\left(\min_{\boldsymbol{\beta}_2}\mathrm{SSE}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\right) \\
&= \left((\boldsymbol{M}_2\boldsymbol{X}_1)^{\mathrm{T}}(\boldsymbol{M}_2\boldsymbol{X}_1)\right)^{-1}(\boldsymbol{M}_2\boldsymbol{X}_1)^{\mathrm{T}}\boldsymbol{M}_2\boldsymbol{Y} \\
&= \left(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{M}_2\boldsymbol{Y},
\end{aligned} \tag{3.37}$$

这里用到了归零矩阵 $\boldsymbol{M}_2$ 是幂等矩阵的性质. 对偶地, 分块回归系数 $\boldsymbol{\beta}_2$ 其单独的 OLS 估计量为

$$\hat{\boldsymbol{\beta}}_2 = \underset{\boldsymbol{\beta}_2}{\mathrm{argmin}}\left(\min_{\boldsymbol{\beta}_1}\mathrm{SSE}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\right) = \left(\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{M}_1\boldsymbol{Y}, \tag{3.38}$$

其中归零矩阵

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{P}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1\left(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^{\mathrm{T}}, \tag{3.39}$$

而 $\boldsymbol{M}_1$ 对应了以 K_1 个变量 $\mathbf{X}_1$ 为解释变量的回归.

由于归零矩阵与残差紧密相连, 因而这给我们提供了一个从残差向量角度去思考的分块回归的视角. 对于 $\hat{\boldsymbol{\beta}}_2 = \left[(\boldsymbol{M}_1\boldsymbol{X}_2)^{\mathrm{T}}\boldsymbol{M}_1\boldsymbol{X}_2\right]^{-1}(\boldsymbol{M}_1\boldsymbol{X}_2)^{\mathrm{T}}(\boldsymbol{M}_1\boldsymbol{Y})$, 记

$$\tilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2, \quad \tilde{\boldsymbol{\varepsilon}}_1 = \boldsymbol{M}_1\boldsymbol{Y},$$

依据 $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\varepsilon}$, 则有 $\hat{\boldsymbol{\beta}}_2 = \left(\tilde{\boldsymbol{X}}_2^{\mathrm{T}}\tilde{\boldsymbol{X}}_2\right)^{-1}\tilde{\boldsymbol{X}}_2^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}_1$, 我们可以通过先计算两次回归的残差, 再通过**残差回归** (residual regression)得到 OLS 估计量 $\boldsymbol{\beta}_2$, 这一过程被称为 **Frisch-Waugh-Lovell 定理** (Frisch-Waugh-Lovell theorem), 简称为 **FWL 定理** (FWL theorem).

定理 3.8.1 (Frisch-Waugh-Lovell 定理). *对于分块回归 $\boldsymbol{Y}=\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{X}_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}$, 系数 $\boldsymbol{\beta}_2$ 的 OLS 估计量*

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2&=\left[\left(\boldsymbol{M}_1\boldsymbol{X}_2\right)^{\mathrm{T}}\boldsymbol{M}_1\boldsymbol{X}_2\right]^{-1}\left(\boldsymbol{M}_1\boldsymbol{X}_2\right)^{\mathrm{T}}\left(\boldsymbol{M}_1\boldsymbol{Y}\right)\\&=\left(\tilde{\boldsymbol{X}}_2^{\mathrm{T}}\tilde{\boldsymbol{X}}_2\right)^{-1}\tilde{\boldsymbol{X}}_2^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}_1,\end{aligned}$$

可以通过如下的残差回归得到:

(a) $\boldsymbol{X}_2$ 对 $\boldsymbol{X}_1$ 回归, 得到残差 $\tilde{\boldsymbol{X}}_2=\boldsymbol{M}_1\boldsymbol{X}_2$;

(b) $\boldsymbol{Y}$ 对 $\boldsymbol{X}_1$ 回归, 得到残差向量 $\tilde{\boldsymbol{\varepsilon}}_1=\boldsymbol{M}_1\boldsymbol{Y}$;

(c) 残差向量 $\tilde{\boldsymbol{\varepsilon}}_1$ 对残差 $\tilde{\boldsymbol{X}}_2$ 回答, 得到回归系数的估计量 $\hat{\boldsymbol{\beta}}_2=\left(\tilde{\boldsymbol{X}}_2^{\mathrm{T}}\tilde{\boldsymbol{X}}_2\right)^{-1}\tilde{\boldsymbol{X}}_2^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}_1$.

需要指出, 定理 (3.8.1) 中 (c) 的回归所得到残差, 即是多元线性回归方程式 (3.30) 的残差, 这是因为

$$\begin{aligned}\mathrm{RSS}\left(\hat{\boldsymbol{\beta}}_2\right)&=\left(\tilde{\boldsymbol{\varepsilon}}_1-\tilde{\boldsymbol{X}}_2\hat{\boldsymbol{\beta}}_2\right)^{\mathrm{T}}\left(\tilde{\boldsymbol{\varepsilon}}_1-\tilde{\boldsymbol{X}}_2\hat{\boldsymbol{\beta}}_2\right)=\left(\boldsymbol{M}_1\boldsymbol{Y}-\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\right)^{\mathrm{T}}\left(\boldsymbol{M}_1\boldsymbol{Y}-\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\right)\\&=\min_{\boldsymbol{\beta}_2}\left(\boldsymbol{M}_1\boldsymbol{Y}-\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2\right)^{\mathrm{T}}\left(\boldsymbol{M}_1\boldsymbol{Y}-\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2\right)=\min_{\boldsymbol{\beta}_2}\left(\min_{\boldsymbol{\beta}_1}\mathrm{SSE}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right)\right)\\&=\min_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\mathrm{SSE}\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\right).\end{aligned}$$

3.9 Leave-One-Out 回归与预测误差

如果读者在此前有学习过机器学习的相关内容, 则应该了解机器学习中不少方法都是基于**子样本** (sub-sample)来实现的, 例如自助法 Bootstrap、k-折交叉验证等方法. 本节则介绍基于**差一法** (leave one out)[5]的线性回归模型, 也即 LOO 回归.

差一法, 即是剔除某个单独的样本后得一个子样本, 然后再将某种统计方法运用于子样本之上, 这即契合了其英文名称 “leave one out”. 据此定义, 显然对于样本容量为 n 的情形 LLO 需要重复统计方法 n 次, 这是十分繁琐的. 但就线性回归模型 (3.5.1),LLO 的结果有着非常简洁的形式, 并且可以利用此去对回归结果做一些细致分析.

模型 3.9.1 (leave-one-out 回归). *在线性回归模型 (3.5.1) 的基础上, 采用 LLO 的回归模型称为 **LLO 回归** (LLO regression). 具体地, 设第 i 个样本为 Y_i 及 $\mathbf{X}_i=\left(X_{i1},\cdots,X_{ik},\cdots,X_{iK}\right)^{\mathrm{T}}$, 则将利用剔除了第 i 个样本得到的 OLS 估计量记为 $\hat{\boldsymbol{\beta}}_{(-i)}$,*

[5]国内就 leave one out 并没有统一的译法, 最常见的翻译即是 “留一法”, 但显然这个翻译是 leave one out 的相反含义. 本笔记后文将 leave one out 简称为 LLO, 而尽量不使用中文译名.

并称第 i 个样本的 **LLO 预测值** *(LLO prediction value)*为 $\tilde{Y}_i = \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(-i)}$, 同时我们记

$$\tilde{\varepsilon}_i = Y_i - \tilde{Y}_i = Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(-i)}, \tag{3.40}$$

则称 $\tilde{\varepsilon}_i$ 为第 i 个样本的**预测误差** *(prediction error)*、**预测残差** *(prediction residual)*或 ***LLO* 残差** *(LLO residual)*.

依据式 (3.20), 则 $\hat{\boldsymbol{\beta}}_{(-i)}$ 可以表示为

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\sum_{j\neq i}\mathbf{X}_j\mathbf{X}_j^{\mathrm{T}}\right)^{-1}\sum_{j\neq i}\mathbf{X}_jY_j,$$

而式 (3.20) 又有依据矩阵分块下的“内积”运算得出的, 这样我们便可将 $\hat{\boldsymbol{\beta}}_{(-i)}$ 写作

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \mathbf{X}_iY_i\right). \tag{3.41}$$

式 (3.41) 可以化简为更简洁的形式, 并与 OLS 估计量联系起来. 将式 (3.41) 左乘以 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$, 得到

$$\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)\hat{\boldsymbol{\beta}}_{(-i)} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \mathbf{X}_iY_i,$$

化简有

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{(-i)} - \mathbf{X}_i\underbrace{\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(-i)}}_{\tilde{Y}_i} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \mathbf{X}_iY_i \quad\Rightarrow\quad \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{(-i)} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \mathbf{X}_i\underbrace{\left(Y_i - \tilde{Y}_i\right)}_{\tilde{\varepsilon}_i},$$

故 $\hat{\boldsymbol{\beta}}_{(-i)} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbf{X}_i\tilde{\varepsilon}_i$, 注意到 OLS 估计量为 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, 于是有剔除了第 i 个样本的 LLO 回归结果为

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbf{X}_i\tilde{\varepsilon}_i. \tag{3.42}$$

进一步地, 依据式 (3.42) 则预测误差 $\tilde{\varepsilon}_i$ 为

$$\begin{aligned}\tilde{Y}_i = \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(-i)} &= \mathbf{X}_i^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbf{X}_i\tilde{\varepsilon}_i\right)\\ &= \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{X}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbf{X}_i\tilde{\varepsilon}_i\\ &= \hat{Y}_i - h_{ii}\tilde{\varepsilon}_i,\end{aligned}$$

亦即有

$$\left(Y_i - \hat{Y}_i\right) + h_{ii}\tilde{\varepsilon}_i = \left(Y_i - \tilde{Y}_i\right) \Leftrightarrow \hat{\varepsilon}_i + h_{ii}\tilde{\varepsilon}_i = \tilde{\varepsilon}_i,$$

最终化简得到 OLL 回归的预测误差为

$$\tilde{\varepsilon}_i = \frac{1}{1-h_{ii}}\hat{\varepsilon}_i. \tag{3.43}$$

需要指出,OLL 方法估计出的回归系数 $\hat{\boldsymbol{\beta}}_{(-i)}$ 是不含有第 i 个样本的信息, 而依据式 (3.42) 和式 (3.40), 则我们也可以从全样本的 OLS 估计量中“剔除”第 i 个样本的信息的信息, 得到 $\hat{\boldsymbol{\beta}}_{(-i)}$. 这样计算 $\hat{\boldsymbol{\beta}}_{(-i)}$ 实则又使用了第 i 个样本的信息.

依据式 (3.43), 我们将预测误差用矩阵符号表示为

$$\tilde{\boldsymbol{\varepsilon}} = \operatorname{diag}\left(\frac{1}{1-h_{11}}, \cdots, \frac{1}{1-h_{nn}}\right)\hat{\boldsymbol{\varepsilon}} = \operatorname{diag}\left(\frac{1}{1-h_{11}}, \cdots, \frac{1}{1-h_{nn}}\right)\boldsymbol{M}\boldsymbol{\varepsilon},$$

记

$$\boldsymbol{M}^{*} = \operatorname{diag}\left(\frac{1}{\sqrt{1-h_{11}}}, \cdots, \frac{1}{\sqrt{1-h_{nn}}}\right),$$

和

$$\boldsymbol{M}^{*2} = (\boldsymbol{M}^{*})^{2} = \operatorname{diag}\left(\frac{1}{1-h_{11}}, \cdots, \frac{1}{1-h_{nn}}\right),$$

则预测误差向量满足有 $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{M}^{*2}\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}^{*2}\boldsymbol{M}\boldsymbol{\varepsilon}$.

我们现在计算预测误差向量 $\tilde{\boldsymbol{\varepsilon}}$ 的条件期望、方差, 即有

$$\mathbb{E}\left(\tilde{\boldsymbol{\varepsilon}} \middle| \boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{M}^{*2}\boldsymbol{M}\boldsymbol{\varepsilon} \middle| \boldsymbol{X}\right) = \boldsymbol{M}^{*2}\boldsymbol{M}\mathbb{E}\left(\boldsymbol{\varepsilon} \middle| \boldsymbol{X}\right) = 0$$

和

$$\begin{aligned}\operatorname{Var}\left(\tilde{\boldsymbol{\varepsilon}} \middle| \boldsymbol{X}\right) &= \operatorname{Var}\left(\boldsymbol{M}^{*2}\hat{\boldsymbol{\varepsilon}} \middle| \boldsymbol{X}\right) = \boldsymbol{M}^{*2}\operatorname{Var}\left(\hat{\boldsymbol{\varepsilon}} \middle| \boldsymbol{X}\right)\boldsymbol{M}^{*2} \\ &\xlongequal{\text{异方差式 (3.9)}} \boldsymbol{M}^{*2}\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^{*2} \xlongequal{\text{同方差式 (3.7)}} \boldsymbol{M}^{*2}\boldsymbol{M}\boldsymbol{M}^{*2}\sigma^{2},\end{aligned}$$

也不难得到

$$\operatorname{Var}\left(\tilde{\varepsilon}_{i} \middle| \boldsymbol{X}\right) = \frac{\operatorname{Var}\left(\hat{\varepsilon}_{i} \middle| \boldsymbol{X}\right)}{\left(1-h_{ii}\right)^{2}} = \frac{\left(1-h_{ii}\right)\sigma^{2}}{\left(1-h_{ii}\right)^{2}} = \frac{\sigma^{2}}{1-h_{ii}}.$$

待补充的内容:

$$\boxed{\text{拟合} \neq \text{插值} \neq \text{预测}}$$

LLO 回归的一个重要用途是用于寻找具有有影响的观测值.

3.10 拟合的度量

我们在上一章介绍了一元线性回归模型 (2.4.1) 的拟合优度 (见定义 (2.5.2)), 本节我们将系统的解释有关拟合的度量的更多概念.

我们在解释拟合优度时, 是基于线性回归模型的方差分析公式对平方和进行了分解, 这在多元的情形自然是称成立的. 利用投影矩阵 $\boldsymbol{P}$ 和归零矩阵 $\boldsymbol{M}$, 则由

$$\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}, \quad \hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon}, \quad \hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} = 0, \quad \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\mathbf{1} = 0\left(\text{需要常数项}\right),$$

可得

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\left(\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}}\right)=\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}}+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}\quad\Leftrightarrow\quad\sum_{i=1}^{n}Y_i^2=\sum_{i=1}^{n}\hat{Y}_i^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \tag{3.44}$$

和

$$\begin{aligned}&\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)\\=&\left(\hat{\boldsymbol{Y}}-\bar{Y}\mathbf{1}+\hat{\boldsymbol{\varepsilon}}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}}-\bar{Y}\mathbf{1}+\hat{\boldsymbol{\varepsilon}}\right)\quad\Leftrightarrow\quad\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2\\=&\left(\hat{\boldsymbol{Y}}-\bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}}-\bar{Y}\mathbf{1}\right)+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}\end{aligned} \tag{3.45}$$

成立. 式 (3.44) 的分解用到了拟合值向量 $\hat{\boldsymbol{Y}}$ 与残差向量 $\hat{\boldsymbol{\varepsilon}}$ 正交的性质, 式 (3.44) 在此之上还依赖于模型 (3.5.1) 含义常数项时残差和为 0 的性质. 利用式 (3.44) 和 (3.44), 我们依次定义**非中心化** R^2(uncentered R^2)为

$$R_{\mathrm{uc}}^2=1-\frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}}=1-\frac{\sum\limits_{i=1}^{n}\hat{Y}_i^2}{\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2}, \tag{3.46}$$

R^2, 亦即**拟合优度** (goodness of fit)或**可决系数** (coefficient of determination)为

$$R^2=1-\frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)}=1-\frac{\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum\limits_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}. \tag{3.47}$$

对于拟合优度 R^2 或是非中心化的 R_{uc}^2, 其都是利用残差平方和 RSS 度量了模型 (3.5.1) 的可解释能力——RSS 越小, 则意味着拟合值 $\hat{\boldsymbol{V}}$ 与真实值 $\boldsymbol{Y}$ 越接近. 依据式 (3.44) 和 (3.44) 可知,R^2 或 R_{uc}^2 的取值范围都是 $[0,1]$, 取值为 0 时表明解释变量 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$ 与被解释变量 Y 线性无关 (这几乎不会出现), 或出现在被解释变量只有常数项的情形; 取值为 1 时, 则有 $\mathrm{RSS}=0$, 这意味着样本数据即是纯粹的线性关系, 不存在有误差 ε.

拟合优度 R^2 或是非中心化的 R_{uc}^2 存在的一个较为严重的问题, 则是当解释变量的总数 K 增大时,R^2 和 R_{uc}^2 也会响应的增加. 这是因为在样本容量 n 不变时残差平方和 RSS 是关于 K 的非增函数, 因而 R^2 和 R_{uc}^2 便是 K 的增函数. 这一特性意味解释变量越多, 拟合的效果自然越好. 对此 Theil(1961) 提出了使用**调整的** R^2(adjusted R^2)或 $\bar{R}^2$来代替 R^2, 经调整的 $\bar{R}^2$ 被定义为

$$\bar{R}^2=1-\frac{(n-K)^{-1}\,\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{(n-1)^{-1}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)}=1-\frac{\dfrac{1}{n-K}\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2}{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}, \tag{3.48}$$

其即是对残差平方和 RSS 和总体平方和 TSS 予以其自由度的约束 (也即是替换为了 s^2 和 Y 的样本方差), 经过了除以自由度的惩罚后, 在 n 不变的情况下,$\bar{R}^2$ 中的分子不再是关于 K 的非增函数, 这在一定程度上校正了 R^2 和 R^2_{uc} 的问题. 需要指出, 式 (3.48) 可能出现负值, 这是其缺点.

较上述的三种 R^2 而言,Bruce(cite) 给出了一种更为推荐的 $\tilde{R}^2$, 其使用预测误差 $\tilde{\varepsilon}_i$ 定义了

$$\tilde{R}^2 = 1 - \frac{\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}}{\left(\boldsymbol{Y} - \bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \bar{Y}\mathbf{1}\right)} = 1 - \frac{\sum\limits_{i=1}^{n}\left(1 - h_{ii}\right)^{-2}\hat{\varepsilon}_i^2}{\sum\limits_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}, \tag{3.49}$$

依据第节的内容, 可以推知 $\tilde{R}^2$ 可以作为预测方差中被能回归模型所够预测的百分比的估计量.$\tilde{R}^2$ 能够完全避免由于解释变量导致的数值提升问题, 其取值也可能出现负值的情形, 但这仅在模型的预测能力比仅有常数项的模型还差时出现.

需要指出, 本节介绍的四种R^2 仅仅度量了模型对于样本数据的拟合情况, 其不能说明回归模型的经济学含义. 对于同一个经济学模型, 将之做适当变换后得到的计量模型的形式往往不止一个, 这些形式都可以被用于正确地估计出经济学模型中的参数, 但相应的回归结果往往具有不同的 R^2. 此外, 本节介绍的四种 R^2 在机器学习中也被广泛使用, 且在处理高维数据和非参数问题时 $\tilde{R}^2$ 被广泛采用.

最后, 我们给出 R^2 与 $\bar{R}^2$ 和 R^2_{uc} 直接的换算关系. 对于调整的 R^2. 由于

$$\begin{aligned}
\bar{R}^2 =& 1 - \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/\left(n-K\right)}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)/\left(n-1\right)} = 1 - \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)} \cdot \frac{n-1}{n-K} \\
=& \frac{n-1}{n-K}\left(\frac{n-K}{n-1} - \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)}\right) \\
=& \frac{n-1}{n-K}\left(\frac{n-K}{n-1} - 1 + 1 - \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)}\right) \\
=& \frac{n-1}{n-K}\left(\frac{n-K}{n-1} - 1 + R^2\right) = \frac{n-1}{n-K}\left(\frac{1-K}{n-1} + R^2\right),
\end{aligned}$$

故化简有

$$1 - R^2 = \frac{n-K}{n-1}\left(1 - \bar{R}^2\right). \tag{3.50}$$

对于 R^2 与非中心化 R^2, 则有

$$\begin{aligned}
1 - R^2 =& \frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)} = \frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)} \cdot \frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}} \\
=& \frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)} \cdot \left[1 - \left(1 - \frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}}\right)\right] = \frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)} \cdot \left(1 - R^2_{\text{uc}}\right) \\
=& \frac{\left(\boldsymbol{Y} - \bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \bar{Y}\mathbf{1}\right) + n\bar{Y}^2}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)} \cdot \left(1 - R^2_{\text{uc}}\right) = \left[1 + \frac{n\bar{Y}^2}{\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \mathbf{1}\bar{Y}\right)}\right]\left(1 - R^2_{\text{uc}}\right)
\end{aligned}$$

即有

$$1-R^2=\left[1+\frac{n\bar{Y}^2}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right]\left(1-R_{\mathrm{uc}}^2\right). \tag{3.51}$$

3.11 多元 OLS 估计量下的方差估计量

本节我们将研究模型 (3.5.1) 的各种方差估计量, 除了在上一章已经有涉及的回归方差 σ^2 的估计量外, 本节我们还将介绍 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的方差估计量.

现在我们研究总体参数回归方差 $\mathbb{E}\varepsilon^2=\sigma^2$ 的估计量, 同样地, 结合条件同方差 $\mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right)=\sigma^2$ 依据矩估计的思想有

$$\hat{\sigma}_{\mathrm{ideal}}^2=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon},$$

且

$$\hat{\sigma}_{\mathrm{ideal}}^2=\frac{1}{n}\mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right)\overset{\text{条件同方差式 (3.7)}}{=\!=\!=}\frac{1}{n}\sum_{i=1}^{n}\sigma^2=\sigma^2,$$

即可以推知 $\hat{\sigma}_{\mathrm{ideal}}^2$ 是 σ^2 的无偏估计. 但回归误差 ε_i 是不可观察的, 使用残差 $\hat{\varepsilon}_i$ 替换有

$$\hat{\sigma}^2=\frac{1}{n}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{M}\boldsymbol{\varepsilon}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{\varepsilon}\leqslant\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}=\hat{\sigma}_{\mathrm{ideal}}^2,$$

可以知悉直接替换的得到的 $\frac{1}{n}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}$ 是有偏的. 为此, 计算条件期望

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)&=\mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\mathrm{tr}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}\right)\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right)\middle|\boldsymbol{X}\right)\\&\overset{\text{迹是线性变换}}{=\!=\!=}\mathrm{tr}\left(\mathbb{E}\left(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\right)=\mathrm{tr}\left(\boldsymbol{M}\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\right)\overset{\text{异方差式 (3.9)}}{=\!=\!=}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{D}\right),\end{aligned}$$

则在同方差假设式 (3.7) 有

$$\mathbb{E}\left(\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)=\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{I}\sigma^2\right)=\left(n-K\right)\sigma^2,$$

同样地, 我们得到 σ^2 的**偏差校正估计量** (bias-corrected estimator)为

$$s^2=\frac{1}{n-K}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\frac{1}{n-K}\sum_{i=1}^{n}\varepsilon_i^2. \tag{3.52}$$

另一种回归方差方式则是使用具体相同方差的标准化残差 $\bar{\varepsilon}_i$ 替换回归误差 ε_i, 即有方差估计量

$$\bar{\sigma}^2=\frac{1}{n}\bar{\boldsymbol{\varepsilon}}^{\mathrm{T}}\bar{\boldsymbol{\varepsilon}}=\frac{1}{n}\sum_{i=1}^{n}\bar{\varepsilon}_i^2=\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varepsilon}_i^2}{\sqrt{1-h_{ii}}}, \tag{3.53}$$

不难计算

$$\begin{aligned}
\mathbb{E}\left(\bar{\boldsymbol{\varepsilon}}^{\mathrm{T}}\bar{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right) &= \mathbb{E}\left(\left(\boldsymbol{M}^{*}\boldsymbol{M}\boldsymbol{\varepsilon}\right)^{\mathrm{T}}\left(\boldsymbol{M}^{*}\boldsymbol{M}\boldsymbol{\varepsilon}\right)\middle|\boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{M}\boldsymbol{M}^{*}\right)\left(\boldsymbol{M}^{*}\boldsymbol{M}\right)\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right) \\
&= \mathbb{E}\left(\mathrm{tr}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\left(\boldsymbol{M}^{*}\right)^{2}\boldsymbol{M}\boldsymbol{\varepsilon}\right)\middle|\boldsymbol{X}\right) = \mathbb{E}\left(\mathrm{tr}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{M}^{*}\right)^{2}\boldsymbol{M}\boldsymbol{\varepsilon}\right)\middle|\boldsymbol{X}\right) \\
&= \mathbb{E}\left(\mathrm{tr}\left(\left(\boldsymbol{M}^{*}\right)^{2}\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right)\middle|\boldsymbol{X}\right) = \mathrm{tr}\left(\mathbb{E}\left(\left(\boldsymbol{M}^{*}\right)^{2}\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\right) \\
&= \mathrm{tr}\left(\left(\boldsymbol{M}^{*}\right)^{2}\boldsymbol{M}\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\right) \xlongequal{\text{异方差式 (3.9)}} \mathrm{tr}\left(\left(\boldsymbol{M}^{*}\right)^{2}\boldsymbol{M}\boldsymbol{D}\right) \\
&= \mathrm{tr}\left(\mathrm{diag}\left(1-h_{11},\cdots,1-h_{nn}\right)\boldsymbol{M}\boldsymbol{D}\right) \\
&\xlongequal{\text{同方差式 (3.7)}} \sigma^{2}\mathrm{tr}\left(\mathrm{diag}\left(1-h_{11},\cdots,1-h_{nn}\right)\boldsymbol{M}\right) \\
&= \sigma^{2}\sum_{i=1}^{n}\frac{1-h_{ii}}{1-h_{ii}} = n\sigma^{2},
\end{aligned}$$

由此可以推知, 式 (3.53) 的 $\bar{\sigma}^2$ 是 σ^2 的无偏估计, 因而也是 σ^2 的偏差校正估计量.

线性回归模型 (3.3.1) 的方差中除了 σ^2 之外, 我们还关注 OLS 估计量的协方差矩阵, 因为由其可以衍生出的各个 $\hat{\beta}_i$ 的方差. 依据定理 (3.5.1), 我们已经计算出 OLS 估计量的协方差矩阵为

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},$$

在同方差式 (3.7) 的情况下则协方差矩阵化简为了 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2$. 由于 σ^2 未知, 则需要使用偏差校正估计量去代替. 对于同方差的情形, 最为简单的协方差矩阵估计量即为

$$\hat{\boldsymbol{V}}^{0}_{\hat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^{2}, \tag{3.54}$$

对协方差矩阵 $\hat{\boldsymbol{V}}^{0}_{\hat{\boldsymbol{\beta}}}$ 计算条件期望得

$$\mathbb{E}\left(\hat{\boldsymbol{V}}^{0}_{\hat{\boldsymbol{\beta}}}\middle|\boldsymbol{X}\right) = \mathbb{E}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^{2}\middle|\boldsymbol{X}\right) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbb{E}\left(s^{2}\middle|\boldsymbol{X}\right) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^{2} = \boldsymbol{V}_{\hat{\boldsymbol{\beta}}},$$

则可以进一步推知在**同方差**的条件下 $\hat{\boldsymbol{V}}^{0}_{\hat{\boldsymbol{\beta}}}$ 是 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ 的无偏估计. 实际上, 通过这里的计算也不难知悉, 用任意的偏差校正估计量替换 σ^2, 得到的协方差矩阵估计量都是无偏的.$\hat{\boldsymbol{V}}^{0}_{\hat{\boldsymbol{\beta}}}$ 是多数计量经济学软件 (包括 Stata、eViews) 的默认的协方差矩阵估计量.

倘若我们的模型确实满足假设 (3.2.1) 而不存在异方差问题, 利用式 (3.54) 去计算 OLS 估计量的方差, 通常而言不会有较大的偏差. 但一旦模型 (3.5.1) 是存在异方差式 (3.9), 则式 (3.2.1) 将与 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$ 存在较大的偏离, 在 Bruce(cite) 中这种偏离可以高达 30 倍! 研究异方差下如何计算 OLS 估计量的协方差矩阵的估计值, 这是一个非常重要的问题.

记对角矩阵 $\boldsymbol{D}_{\boldsymbol{\varepsilon}} = \mathrm{diag}\left(\varepsilon_1^2,\cdots,\varepsilon_n^2\right)$, 则有

$$\mathbb{E}\left(\boldsymbol{D}_{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right) = \mathrm{diag}\left(\sigma_1^2,\cdots,\sigma_n^2\right) = \boldsymbol{D},$$

故 $\boldsymbol{D}_{\boldsymbol{\varepsilon}}$ 是条件异方差矩阵 $\boldsymbol{D}$ 的条件无偏估计.

异方差的情况下,OLS 估计量的协方差矩阵 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$, 由于

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}=\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\sigma_i^2,$$

则我们需要对 $\sigma_1^2,\cdots,\sigma_n^2$ 进行估计. 显然, 理想的情况即是使用回归误差的平方 $\varepsilon_1^2,\cdots,\varepsilon_n^2$ 去估计, 也即是使用 $\boldsymbol{D}_{\boldsymbol{\varepsilon}}$ 去估计 $\boldsymbol{D}$. 记

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{ideal}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}_{\boldsymbol{\varepsilon}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\varepsilon_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\tag{3.55}$$

则

$$\begin{aligned}&\mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{ideal}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}_{\boldsymbol{\varepsilon}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\middle|\boldsymbol{X}\right)\\&=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{D}_{\boldsymbol{\varepsilon}}|\boldsymbol{X}\right)\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\boldsymbol{V}_{\hat{\boldsymbol{\beta}}},\end{aligned}$$

即 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{ideal}}$ 是协方差矩阵 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ 的无偏估计量. 可惜, 回归误差 $\boldsymbol{\varepsilon}$ 不可观察, 因此我们需要其他的估计方法.

最基本的方法即是即是使用残差 $\hat{\varepsilon}_i$ 替代回归误差 ε_i, 即有

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\tag{3.56}$$

这里的上标 HC 是 heteroskedasticity-consistent 的缩写, 而 HC0 则是代表 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}$ 是异方差下协方差矩阵估计量的基准.$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}$ 是由 Eicker(1963,cite) 提出的, 最早由 White(1980,cite) 引入计量经济学中, 因此 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}$ 又被称为 **Eicker-White 协方差矩阵估计量** (Eicker-White covariance matrix estimator)或 **White 协方差矩阵估计量** (White covariance matrix estimator).

依据我们多次计算的经验, 显然, 由于 $\hat{\varepsilon}_i$ 的条件方差较 ε_i 偏小, 则式 (3.56) 会是有偏的, 仿照校正直接替换 $\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$ 的方法, 则 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}$ 的校正估计量为

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC1}}=\frac{n}{n-K}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}=\frac{n}{n-K}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\tag{3.57}$$

式 (3.57) 比式 (3.56) 更为推荐, 其予以了 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}$ 的 $n-K$ 的自由度的惩罚, 这种形式的协方差矩阵估计量由 Hinkley(1977,cite) 提出, 在 Stata 软件中输入可选命令 “ `,r`” 时,Stata 便使用 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC1}}$ 作为协方差矩阵的估计量.

另几种常见的方法便是使用标准化残差 $\bar{\varepsilon}_i$ 或预测误差 $\tilde{\varepsilon}_i$ 去替换残差, 即有

$$\begin{aligned}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\bar{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}(1-h_{ii})^{-1}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\end{aligned} \tag{3.58}$$

$$\begin{aligned}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\tilde{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}(1-h_{ii})^{-2}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\end{aligned} \tag{3.59}$$

我们将式 (3.56)、(3.57)、(3.58) 和 (3.59) 统称为**稳健的协方差矩阵估计量** (robust covariance matrix estimator)、**异方差一致的协方差矩阵估计量** (heteroskedasticity-consistent covariance matrix estimator)或**异方差稳健的协方差矩阵估计量** (heteroskedasticity-robust covariance matrix estimator). 使用使用标准化残差 $\bar{\varepsilon}_i$ 替换得到的 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$ 是 Horn 和 Duncan(1975,cite) 提出的, 在 Stata 中可以使用可选项 “ `,vce(2)`” 来汇报依据 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$ 计算的稳健的回归结果.$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}$ 则是最晚提出的协方差矩阵估计量, 其依据 jackknife 原则由 MacKinnonand(1985,cite) 提出, 在 Stata 中通过可选命令 “ `,vce(3)`” 来汇报相关结果.

由于 $1<\dfrac{1}{1-h_{ii}}<\dfrac{1}{(1-h_{ii})^2}, 0<h_{ii}<1$, 则可以推知

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} < \hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} < \hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}, \tag{3.60}$$

这里 $<$ 是针对二次型而言的, 即 $\boldsymbol{A}<\boldsymbol{B}$ 等价于 $\boldsymbol{B}-\boldsymbol{A}$ 是正定的二次型.

对于 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$、$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$ 和 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}$, 我们可以尝试计算器条件期望以判断其无偏性. 以 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$ 为例, 即有

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}\middle|\boldsymbol{X}\right) &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbb{E}\left(\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\middle|\boldsymbol{X}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\mathbb{E}\left(\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right)\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\end{aligned}$$

对于 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$ 和 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}$ 的计算是同理的, 也就是说,$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$、$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$ 和 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}$ 的条件期望依次取决于条件二阶原点矩 $\mathbb{E}\left(\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right)$、$\mathbb{E}\left(\bar{\varepsilon}_i^2\middle|\boldsymbol{X}\right)$ 和 $\mathbb{E}\left(\tilde{\varepsilon}_i^2\middle|\boldsymbol{X}\right)$. 由于残差、标准化残差和预测误差的条件期望为 0, 则其条件二阶矩即等于其期望, 依据此前的计算有

$$\mathrm{Var}\left(\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right) \overset{\text{异方差}}{=\!=} \boldsymbol{MDM} \Rightarrow \mathbb{E}\left(\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right) = \mathrm{Var}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right) = \left(\boldsymbol{MDM}\right)_{ii} \overset{\text{同方差}}{=\!=} (1-h_{ii})\,\sigma^2,$$

以及

$$\mathbb{E}\left(\bar{\varepsilon}_i^2 \middle| \boldsymbol{X}\right) = \operatorname{Var}\left(\bar{\varepsilon}_i \middle| \boldsymbol{X}\right) = \operatorname{Var}\left(\frac{\hat{\varepsilon}_i}{\sqrt{1-h_{ii}}} \middle| \boldsymbol{X}\right) = \frac{\operatorname{Var}\left(\hat{\varepsilon}_i \middle| \boldsymbol{X}\right)}{1-h_{ii}} \overset{\text{同方差}}{=\!=} \sigma^2,$$
$$\mathbb{E}\left(\tilde{\varepsilon}_i^2 \middle| \boldsymbol{X}\right) = \operatorname{Var}\left(\tilde{\varepsilon}_i \middle| \boldsymbol{X}\right) = \operatorname{Var}\left(\frac{\hat{\varepsilon}_i}{1-h_{ii}} \middle| \boldsymbol{X}\right) = \frac{\operatorname{Var}\left(\hat{\varepsilon}_i \middle| \boldsymbol{X}\right)}{(1-h_{ii})^2} \overset{\text{同方差}}{=\!=} \frac{\sigma^2}{1-h_{ii}},$$

异方差的情形下需要计算 $\boldsymbol{MDM}$, 其形式较为复杂, 在此暂不展开讨论. 但在同方差式 (3.7) 的条件下, 对于 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$、$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$ 和 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}$ 这三个异方差稳健的估计量, 则容易得到

$$\mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} \middle| \boldsymbol{X}\right) < \boldsymbol{V}_{\hat{\boldsymbol{\beta}}} < \mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}} \middle| \boldsymbol{X}\right), \quad \boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} \middle| \boldsymbol{X}\right), \tag{3.61}$$

这里的 $<$ 依然是针对二次型的正定性而言的. 式 (3.61) 与式 (3.60) 的结果是相同的. 尽管这三个估计量是针对异方差情形的, 但在同方差的情况下我们还是能够借此对三者进行直观的比较. 对于 OLS 估计量的条件协方差矩阵的估计量, 可以见第5.2节的表5.1的总结.

在具体的回归分析中, 无论我们选择了哪一个估计量去估计 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$, 我们总是关心协方差矩阵的对角线元素, 其是 OLS 估计量的方差估计值.

定义 3.11.1 (标准误差, 标准误). *对于 OLS 估计量协方差矩阵 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ 的估计量 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, 我们称 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ 的对角线元素的平方根*

$$\operatorname{SE}\left(\hat{\beta}_k\right) = \sqrt{\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\right)_{kk}} \tag{3.62}$$

*为 OLS 估计量 $\hat{\beta}_k$ 的***标准误差** *(standard error)*或***标准误**.

依据定义 (3.11.1), 标准误$\operatorname{SE}\left(\hat{\beta}_k\right)$ 实际上是 OLS 估计量 $\hat{\beta}_k$ 的标准差的**估计量**. 当模型 (3.5.1) 满足有同方差假设式 (3.7) 时, 则依据式 (3.54) 则式 (3.62) 等于

$$\operatorname{SE}\left(\hat{\beta}_k\right) = s\sqrt{\left((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right)_{kk}}. \tag{3.63}$$

下面我们通过数据实例来说明 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{0}$、$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$、$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}}$、$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$ 和 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}$ 这五个协方差矩阵估计量的区别. 考虑一个如下的存在异方差的一元线性回归模型: 假设解释变量 X 仅有取值 $X = 0, 5, 10, 15$, 数据生成过程为 $Y_i = 10 + 25X_i + \varepsilon_i$ 且有条件异方差

$$\mathbb{E}\left(\varepsilon_i \middle| X = 0\right) = 25^2,\ \mathbb{E}\left(\varepsilon_i \middle| X = 5\right) = 30^2,\ \mathbb{E}\left(\varepsilon_i \middle| X = 15\right) = 15^2,\ \mathbb{E}\left(\varepsilon_i \middle| X = 20\right) = 5^2,$$

成立. 设 X 取值 $0, 5, 10, 15$ 的样本数分别为 $n_1 = 25, n_2 = 30, n_3 = 15, n_4 = 20$, 则样本

容量 n 为 $n = n_1 + n_2 + n_3 + n_4 = 90$, 这样, 我们便可以得到该模型的协方差矩阵为

$$\boldsymbol{D} = \begin{pmatrix} 25^2\boldsymbol{I}_{25} & & & \\ & 30^2\boldsymbol{I}_{30} & & \\ & & 15^2\boldsymbol{I}_{15} & \\ & & & 5^2\boldsymbol{I}_{20} \end{pmatrix}.$$

我们对这个异方差的回归模型使用 Monte Carlo 模拟以求解 OLS 估计量. 依据式 (3.19) 则有

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \begin{pmatrix} 21.3049 \\ 10.2696 \end{pmatrix},$$

即有样本回归直线为 $Y = 10.2696 + 21.3049X$, 可见异方差下 $\hat{\beta}_1$ 与真实值 $\beta_1 = 25$ 有较大的偏差. 现我们依据本节的推导, 计算 OLS 估计量的标准误, 结果如表 (3.2) 所示. 不难发现, 与真实的 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ 所计算的 OLS 估计量的标准差相比, 同方差假设下 $\hat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$ 计算的标准误出现了非常大的偏差, 对于 $\hat{\beta}_0$ 的标准差估计的误差为 11.53%, 而 $\hat{\beta}_0$ 的标准差估计的误差更是高达 -29.54%! 而余下的四种考虑了异方差的协方差矩阵估计量, 其计算的标准误较真实的标准差而言, 则误差控制在了 5% 以下, 且不难验证式 (3.60) 的二次型关系 (对角元的大小).

表 3.2: 协方差矩阵估计量

	$\hat{\beta}_0$ 的标准误	$\mathrm{SE}(\hat{\beta}_0)$ 与真实值的误差	$\hat{\beta}_1$ 的标准误	$\mathrm{SE}(\hat{\beta}_1)$ 与真实值的误差
真实值	4.2300	0.00%	0.3336	0.00%
$\hat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$	3.7425	11.53%	0.4321	-29.54%
$\hat{\boldsymbol{V}}^{\mathrm{HC0}}_{\hat{\boldsymbol{\beta}}}$	4.3036	-1.74%	0.3394	-1.74%
$\hat{\boldsymbol{V}}^{\mathrm{HC1}}_{\hat{\boldsymbol{\beta}}}$	4.3522	-2.89%	0.3433	-2.89%
$\hat{\boldsymbol{V}}^{\mathrm{HC2}}_{\hat{\boldsymbol{\beta}}}$	4.3526	-2.90%	0.3439	-3.10%
$\hat{\boldsymbol{V}}^{\mathrm{HC3}}_{\hat{\boldsymbol{\beta}}}$	4.4024	-4.08%	0.3485	-4.47%

3.12 多元正态回归模型及其区间估计

本节我们将介绍多元正态线性回归模型, 其是引入了正态假设的模型 (3.5.1), 是一种特例的线性回归模型. 在此之上, 我们便可以给出 OLS 估计量的区间估计与假设检验.

模型 3.12.1 (多元正态回归模型). 在多元线性回归模型 *(3.5.1)* 的假设之上, 补充假设

$$\text{总体}:\varepsilon|\,X\sim \mathrm{N}\left(0,\sigma^2\right), \tag{3.64}$$

$$\text{样本}:\boldsymbol{\varepsilon}|\,\boldsymbol{X}\sim \mathrm{N}\left(\mathbf{0},\boldsymbol{I}_n\sigma^2\right), \tag{3.65}$$

成立时, 满足如下设定的模型称为多元**正态回归模型** *(normal regression model).*

$$\begin{aligned}
&Y=\mathscr{P}\left(Y|\mathbf{X}\right)+\varepsilon\\
&\quad\mathscr{P}\left(Y|\mathbf{X}\right)=\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta},\quad \boldsymbol{\beta}=\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}Y\right)\\
&\quad\mathbb{E}\left(\mathbf{X}\varepsilon\right)=0\\
&\quad\mathbb{E}\left(\varepsilon\right)=0\quad(\text{不需要常数项})\\
&\quad\mathrm{Var}\left(\varepsilon\right)=\mathrm{E}\varepsilon^2=\mathrm{E}\left(\varepsilon|\mathbf{X}\right)=\sigma^2
\end{aligned}$$

与在上一章的分析一样, 当总体满足正态假设 (3.64) 时, 则 $Y=\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon$ 满足有条件分布 $Y|\,\mathbf{X}\sim \mathrm{N}\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta},\sigma^2\right)$. 相应地, 正态假设 (3.65) 使得样本 $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 有条件分布 $\boldsymbol{Y}|\,\boldsymbol{X}\sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{I}_n\sigma^2\right)$ 成立, 亦即有 $Y_i|\,\mathbf{X}_i\sim \mathrm{N}\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta},\sigma^2\right)$. 注意, 正态假设的式 (3.64) 和 (3.65) 是给定了回归误差的条件分布为正态分布——正态分布仅仅依赖于期望和方差——此时的回归误差便不受到解释变量的影响, 即模型 (3.12.1) 自然有同方差假设成立.

在此之上, 我们现在研究 OLS 估计量所具有的概率分布. 由于抽样误差为 $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$, 则其是服从多余正态分布的回归误差 $\boldsymbol{\varepsilon}$ 向量的**线性变换**, 故抽样误差依然服从正态分布, 且满足有

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\Big|\,\boldsymbol{X}\sim \mathrm{N}\left(\mathbf{0},\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2\right),$$

由此推出 OLS 估计量的条件分布为 $\boxed{\hat{\boldsymbol{\beta}}\Big|\,\boldsymbol{X}\sim \mathrm{N}\left(\boldsymbol{\beta},\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2\right)}$, 据此则可以得到

$$\hat{\beta}_k\Big|\,\boldsymbol{X}\sim \mathrm{N}\left(\beta_k,\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2\right)_{kk}\right). \tag{3.66}$$

对于拟合值向量 $\hat{\boldsymbol{Y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$ 和残差向量 $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{M}\boldsymbol{\varepsilon}$, 同理可以推知

$$\hat{\boldsymbol{Y}}\Big|\,\boldsymbol{X}\sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{P}\sigma^2\right),\quad \hat{\boldsymbol{\varepsilon}}|\,\boldsymbol{X}\sim \mathrm{N}\left(\mathbf{0},\boldsymbol{M}\sigma^2\right). \tag{3.67}$$

特别地, 由于抽样误差 $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ 和残差向量 $\hat{\boldsymbol{\varepsilon}}$ 均是误差向量 $\boldsymbol{\varepsilon}$ 的线性变换, 故我们考虑矩阵分块有

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\boldsymbol{\varepsilon}}\end{pmatrix}=\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\\ \boldsymbol{M}\boldsymbol{\varepsilon}\end{pmatrix}=\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\\ \boldsymbol{M}\end{pmatrix}\boldsymbol{\varepsilon},$$

这样我们计算 $\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}},\boldsymbol{M}\right)^{\mathrm{T}}\boldsymbol{\varepsilon}$ 的条件期望、方差为

$$\begin{aligned}\mathbb{E}\left(\left.\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\\ \boldsymbol{M}\end{pmatrix}\boldsymbol{\varepsilon}\right|\boldsymbol{X}\right)&=\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\\ \boldsymbol{M}\end{pmatrix}\mathbb{E}\left(\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)\\&=\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\\ \boldsymbol{M}\end{pmatrix}\cdot\mathbf{0}=\mathbf{0}_{(K+n)\times 1}\end{aligned}$$

及

$$\begin{aligned}&\mathrm{Var}\left(\left.\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\\ \boldsymbol{M}\end{pmatrix}\boldsymbol{\varepsilon}\right|\boldsymbol{X}\right)\\=&\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\\ \boldsymbol{M}\end{pmatrix}\mathrm{Var}\left(\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)\left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\boldsymbol{M}\right)\\=&\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\\ \boldsymbol{M}\end{pmatrix}\boldsymbol{I}_n\sigma^2\left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\boldsymbol{M}\right)=\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2 & \boldsymbol{O}_{K\times n}\\ \boldsymbol{O}_{n\times K} & \boldsymbol{M}\sigma^2\end{pmatrix},\end{aligned}$$

故有

$$\left.\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\boldsymbol{\varepsilon}}\end{pmatrix}\right|\boldsymbol{X}\sim\mathrm{N}\left(\mathbf{0}_{(K+n)\times 1},\begin{pmatrix}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2 & \boldsymbol{O}_{K\times n}\\ \boldsymbol{O}_{n\times K} & \boldsymbol{M}\sigma^2\end{pmatrix}\right).\tag{3.68}$$

式 (3.68) 表明, 抽样误差 $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ 和残差向量 $\hat{\boldsymbol{\varepsilon}}$ 的联合分布是一个多元正态分布, 且依据协方差矩阵可知 $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ 与 $\hat{\boldsymbol{\varepsilon}}$ 是完全不相关的, 而对于服从正态分布的随机向量而言, 完全不相关与相互独立是等价的, 即式 (3.68) 表明了在多元正态回归模型 (3.12.1) 下, 抽样误差与残差向量独立. 这还可以进一步的推出,OLS 估计量 $\hat{\boldsymbol{\beta}}$ 与残差向量 $\hat{\boldsymbol{\varepsilon}}$ 是相互独立的.

现在我们在来研究方差估计量 s^2 的条件分布, 在一元正态回归模型 (2.7.1) 中我们说明了其经过变形后服从于卡方分布. 现在我们利用线性代数的知识来证明这一点. 对于 RRS 有 $\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\left(\boldsymbol{M}\boldsymbol{\varepsilon}\right)^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}$, 而由于投影矩阵 $\boldsymbol{M}$ 是幂等的实对称方阵, 则依据式 (3.29) 得到

$$\boldsymbol{M}=\boldsymbol{H}^{-1}\begin{pmatrix}\boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n}\\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{O}_K\end{pmatrix}\boldsymbol{H},$$

其中 $\boldsymbol{H}$ 是依赖于 $\boldsymbol{X}$ 的正交方阵, 满足有 $\boldsymbol{H}^{-1}=\boldsymbol{H}^{\mathrm{T}}$. 故 RSS 有

$$\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{H}^{-1}\begin{pmatrix}\boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n}\\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{O}_K\end{pmatrix}\boldsymbol{H}\boldsymbol{\varepsilon}=\left(\boldsymbol{H}\boldsymbol{\varepsilon}\right)^{\mathrm{T}}\begin{pmatrix}\boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n}\\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{O}_K\end{pmatrix}\boldsymbol{H}\boldsymbol{\varepsilon}$$

成立. 令 $\boldsymbol{\mu}=\boldsymbol{H}\boldsymbol{\varepsilon}=\left(\mu_1,\cdots,\mu_n\right)^{\mathrm{T}}$, 即是对回归误差向量 $\boldsymbol{\varepsilon}$ 做正交变换, 这自然也是线性变换, 则 $\boldsymbol{\mu}$ 服从多元正态分布. 不难计算有

$$\begin{aligned}&\mathbb{E}\left(\boldsymbol{\mu}|\,\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{H}\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)=\boldsymbol{H}\mathbb{E}\left(\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)=\boldsymbol{H}\mathbf{0}=\mathbf{0},\\&\mathrm{Var}\left(\boldsymbol{\mu}|\,\boldsymbol{X}\right)=\mathrm{Var}\left(\boldsymbol{H}\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)=\boldsymbol{H}\mathrm{Var}\left(\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{H}\left(\boldsymbol{I}_n\sigma^2\right)\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{I}_n\sigma^2,\end{aligned}$$

故 $\boldsymbol{\mu}|\,\boldsymbol{X}\sim \mathrm{N}\left(\boldsymbol{0},\boldsymbol{I}_n\sigma^2\right)$, 且 $\boldsymbol{\mu}$ 的分量满足有 $\mu_i|\,\boldsymbol{X}\sim \mathrm{N}\left(0,\sigma^2\right)$.

我们将 $\boldsymbol{\mu}$ 进行分块处理, 设 $\boldsymbol{\mu}=\left(\boldsymbol{\mu}_1^{\mathrm{T}},\boldsymbol{\mu}_2^{\mathrm{T}}\right)$, 则不难得知 $\boldsymbol{\mu}_1|\,\boldsymbol{X}\sim \mathrm{N}\left(\boldsymbol{0},\boldsymbol{I}_{n-K}\sigma^2\right)$. 再计算 RSS 得

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} =&\boldsymbol{\mu}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n}\\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{O}_K\end{pmatrix}\boldsymbol{\mu}\\ =&\left(\boldsymbol{\mu}_1^{\mathrm{T}},\boldsymbol{\mu}_2^{\mathrm{T}}\right)\begin{pmatrix}\boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n}\\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{O}_K\end{pmatrix}\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}\\ =&\left(\boldsymbol{\mu}_1^{\mathrm{T}},\boldsymbol{0}_{1\times K}\right)\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}=\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1,\end{aligned}$$

故可以推知方差估计量 s^2 满足有

$$\left.\frac{(n-K)\,s^2}{\sigma^2}\right|\boldsymbol{X}=\left.\frac{\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1}{\sigma^2}\right|\boldsymbol{X}=\left.\sum_{i=1}^{n-K}\left(\frac{\mu_i-0}{\sigma}\right)^2\right|\boldsymbol{X}\sim\chi^2\left(n-K\right). \tag{3.69}$$

由于正态假设 (3.64) 和 (3.65) 保证了模型 (3.12.1) 是满足条件同方差式 (3.7), 我们在本章无需考虑异方差问题, 对于需要 σ^2 的情形便使用偏差校正估计量 s^2 去代替.

现在我们便可以对 OLS 估计量进行区间估计. 与定义 (2.7.1) 完全一致, 我们使用 t 统计量进行区间估计.

定义 3.12.1 (多元正态回归模型的 t 统计量)**.** *对于模型 (3.12.1) 的 OLS 估计量 $\hat{\boldsymbol{\beta}}$, 其标准化后有*

$$\left.\frac{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}}{\sqrt{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2}}\right|\boldsymbol{X}\sim\mathrm{N}\left(\boldsymbol{0},\boldsymbol{I}\right),$$

*由于 σ^2 为未知的***厌恶参数** *(nuisance parameter), 使用 σ^2 的偏差校正估计量 s^2 代替, 则对于 OLS 估计量的第 k 个参数估计量, 我们称*

$$T=\frac{\hat{\beta}_k-\beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)}=\frac{\hat{\beta}_k-\beta_k}{\sqrt{s^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}} \tag{3.70}$$

为 ***t* 统计量** *(t-statistic).*

显然,$\hat{\beta}_k$ 的 t 统计量 T 有

$$T=\frac{\beta_k-\hat{\beta}_k}{\sqrt{\sigma^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}\Big/\frac{\sqrt{s^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}{\sqrt{\sigma^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}=\frac{\beta_k-\hat{\beta}_k}{\sqrt{\sigma^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}\Big/\sqrt{\frac{(n-K)\,s^2/\sigma^2}{n-K}},$$

而

$$\left.\frac{\beta_k-\hat{\beta}_k}{\sqrt{\sigma^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}\right|\boldsymbol{X}\sim\mathrm{N}\left(0,1\right),\quad\left.\frac{\left(n-K\right)s^2}{\sigma^2}\right|\boldsymbol{X}\sim\chi^2\left(n-K\right)$$

故 t 统计量 T 的条件分布服从自由度为 $n-K$ 的 t 分布.

在上一章第2.7节中, 我们已经详细讲解了以 t 统计量作为枢轴变量去构造 OLS 估计量的置信区间, 本笔记不再重复叙述. 在此, 我们列出如下定理.

定理 3.12.1 (OLS 估计量的置信区间). *在正态假设 (3.64) 和 (3.65) 下, 模型 (3.12.1) 总体参数 $\boldsymbol{\beta}$ 的第 k 个回归系数 β_k 的置信水平为 $1-\alpha$ 的置信区间为*

$$\hat{I}=\left[\hat{\beta}_k-c\cdot\mathrm{SE}\left(\hat{\beta}_k\right),\hat{\beta}_k+c\cdot\mathrm{SE}\left(\hat{\beta}_k\right)\right],\tag{3.71}$$

其中 $c=F^{-1}\left(1-\dfrac{\alpha}{2}\right)$,$F$ 是自由度为 $n-K$ 的 t 分布的累计密度函数,F^{-1} 为 F 的反函数. 常数 c 亦即是自由度为 $n-K$ 的 t 分布的上 $\alpha/2$ 分位数 $t_{\alpha/2}$.
对于回归方差 σ^2, 其置信水平为 $1-\alpha$ 的置信区间为

$$\hat{I}=\left[\frac{\left(n-2\right)s^2}{c_2},\frac{\left(n-2\right)s^2}{c_1}\right],\tag{3.72}$$

其中 $c_1=F^{-1}\left(\alpha/2\right),c_2=F^{-1}\left(1-\alpha/2\right)$,$F$ 是自由度为 $n-K$ 的卡方分布的累计密度函数,F^{-1} 为 F 的反函数. 常数 c_1 和 c_2 亦即是自由度为 $n-K$ 的卡方分布的上 $\alpha/2$ 分位数和上 $1-\alpha/2$ 分位数.

特别地, 对于置信区间式 (3.71), 一条有用的经验公式为当 $n-K\geqslant 61$ 时,β_k 的置信水平为 95% 的置信区间近似为

$$\hat{I}=\left[\hat{\beta}_k-2\cdot\mathrm{SE}\left(\hat{\beta}_k\right),\hat{\beta}_k+2\cdot\mathrm{SE}\left(\hat{\beta}_k\right)\right].\tag{3.73}$$

3.13 多元正态回归模型的假设检验

作为区间估计的置信区间同点估计一样, 其都是被视为由抽取的样本所决定的**随机变量**, 因此其具有的概率性质, 必须从重复抽样的前提下去理解——这即是大数定律赋予随机变量的现实意义. 然而, 在多数情况下, 计量经济学所要研究的问题是不能重复调查的, 相应地问题所需的样本仅有唯一一次的抽样结果. 就单独的一次样本而言,OLS 估计量的无偏性、置信区间的 $1-\alpha$ 的置信水平似乎都没有意义——就我们为了确保估计结果的准确性而言.

但假设检验可以就单次样本的估计结果予以可信度的说明. 就模型 (3.12.1) 的回归结果而言, 其结果是否正确我们可以给出自己的判断, 这种判断可以源自于经济学的理

论, 由此我们便对于回归结果提出了**假设** (hypothesis). 通常而言, 对于某个总体参数 θ, 我们认为其具有一定的限制, 我们用集合 B_0 表示该限制, 则我们称

$$\mathrm{H}_0 : \theta \in B_0 \tag{3.74}$$

为**原假设** (null hypothesis), 相应地, 我们称原假设不成立的对应情况

$$\mathrm{H}_1 : \theta \notin B_0 \tag{3.75}$$

为**备择假设** (alternative hypothesis). 对于原假设 H_0 为真, 我们记为 $\theta \in \mathrm{H}_0$, 反之则记为 $\theta \in \mathrm{H}_1$.

注意, 原假设 H_0 与备择假设 H_1 的选取取决于我们对于所谓“限制”的判断. 例如, 我们要研究一个命题 $\theta \neq \theta_0$, 则原假设 H_0 与备择假设 H_1 的选取通常有如下两种抉择:

(i) 如果我们要证明命题 $\theta \neq \theta_0$ 为真, 依据反证法的思想, 令其逆命题为原假设 $\mathrm{H}_0 : \theta_0 = 0$, 而备择假设即是 $\mathrm{H}_1 : \theta \neq \theta_0$;

(ii) 依据实际问题考虑, 选取犯第一类错误 (见后文) 后危害更小的情形, 将其作为原假设 H_0.

通常而言, 原假设 H_0 与备择假设 H_1 的选择一般都是遵从 (i) 的方法, 这种选择为假设检验的讨论提供了一个公认的范式.

假设检验 (hypothesis test)即是通过样本数据构造**检验统计量** (test statistic), 据此判断原假设 H_0 是否正确. 对于依据检验统计量的实现值来论断, 原假设 H_0 是否成立的方法, 我们称为**检验** (test), 且常记为 ϕ. 例如, 我们有针对回归系数 β_k 原假设 $\mathrm{H}_0 : \beta_k > c$, 则以 OLS 估计量为检验统计量的

$$\phi : \hat{\beta}_k > C_0$$

便是原假设 H_0 的一个检验, 其意味当 $\hat{\beta}_0 > C_0$ 时我们接受 (不拒绝) 原假设 $\mathrm{H}_0 : \beta_k > c$. 我们将检验 ϕ 中使得接受 (不拒绝) 原假设 H_0的检验统计量的集合, 称为检验 ϕ 的**接受域** (acceptance region), 反之使得拒绝原假设 H_0 的集合称为检验 ϕ 的**拒绝域** (rejection region). 接受域与拒绝域的边界通常称为**临界值** (critical value). 例如上例检验 ϕ 中的接受域是集合 $\left\{ \hat{\beta}_k \middle| \hat{\beta}_k > C_0 \right\}$, 拒绝域是 $\left\{ \hat{\beta}_k \middle| \hat{\beta}_k \leqslant C_0 \right\}$, 而 C_0 即是临界值.

对于一个原假设 H_0, 其检验 ϕ 是多种的, 而 ϕ 所使用的检验统计量是依赖于抽取的样本, 因而检验 ϕ 实际上是具有随机性的. 为了避免源自抽样的误差的影响, 我们需要对于检验 ϕ 提出一个优劣判断准则.

定义 3.13.1 (功效函数). 设关于总体参数 θ 的原假设 H_0, 其有基于样本构造检验统计量的检验 ϕ, 我们称

$$\beta_\phi(\theta) = \mathbb{P}(\text{检验}\phi\text{下}\mathrm{H}_0\text{被拒绝}) \tag{3.76}$$

为检验 ϕ 的**功效函数** *(power function)*.

不难知晓, 当 $\theta \in \mathrm{H}_0$ 时功效函数 $\beta_\phi(\theta)$ 应该尽可能小; 当 $\theta \in \mathrm{H}_1$ 时功效函数应该尽可能大 (体现了“功效 (power)”一词). 故我们可以以 $\beta_\phi(\theta)$ 作为检验 ϕ 的判断准则.

在假设检验中, 由于检验 ϕ 具有随机性, 对于真实情形 $\theta \in \mathrm{H}_0$ 或 $\theta \in \mathrm{H}_1$ 便可能出现判断错误. 假设检验存在有两类错误:

- **第一类错误** (type I error):H_0 正确但被拒绝 (**弃真**);
- **第二类错误** (type II error):H_0 错误但被接受 (**取伪**).

这两类错误可以用表 (3.3) 直观展示.

表 3.3: 假设检验的判断结果

	$\theta \in \mathrm{H}_0$	$\theta \in \mathrm{H}_1$
接受 (不拒绝)	正确	第二类错误
拒绝	第一类错误	正确

由于 $\theta \in \mathrm{H}_0$ 和 $\theta \in \mathrm{H}_1$ 两者只能出现一种情形, 不难计算, 对于检验 ϕ, 记其出现第一、二类错误的概率分布为 $\alpha_1(\theta|\phi), \alpha_2(\theta|\phi)$, 则有

$$\alpha_1(\theta|\phi) = \begin{cases} \beta_\phi(\theta), \theta \in \mathrm{H}_0, \\ 0, \theta \in \mathrm{H}_1, \end{cases} \qquad \alpha_2(\theta|\phi) = \begin{cases} 0, \theta \in \mathrm{H}_0, \\ 1 - \beta_\phi(\theta), \theta \in \mathrm{H}_1 \end{cases}$$

成立. 为了保障假设检验的准确性, 被广泛接受的一种方法即是控制犯第一类错误的概率$\alpha_1(\theta|\phi)$ 在某一较小的水平下. 在前文的按照 (i) 去选择原假设 H_0 时, 即依据反证法的思想将要证明的命题的逆命题设定为原假设, 控制了 $\alpha_1(\theta|\phi)$ 较小, 则意味着 $1 - \beta_\phi(\theta)$ 有较大的值, 这就是说检验 ϕ 下接受 (不拒绝) 原假设的概率较大. 这也意味着, 如果检验 ϕ 能够拒绝原假设 H_0, 其有检验统计量确定的概率是较小的, 故需要有充分的证据才能拒绝原假设, 也由此避免了检验 ϕ 的随机误差.

定义 3.13.2 (显著性水平). 设 ϕ 是原假设 H_0 的一个检验, 对于常数 $0 \leqslant \alpha \leqslant 1$, 若 $\forall \theta \in \mathrm{H}_0$ 总有 $\beta_\phi(\theta) \leqslant \alpha$, 我们称检验 ϕ 是原假设 H_0 的一个**显著性水平** *(significance level)*为 α 的检验.

显然, 一个显著性水平为 α 的检验 ϕ 表保障了犯第一类错误的概率小于等于常数 α, 通常而言, 尝试 α 的取值为 10%、5% 和 1%.

经过了上述的理论分析, 现在我们就 OLS 估计量进行假设检验的实际操作. 对于单个回归系数 β_k, 最为常见的假设检验问题便是

$$\mathrm{H}_0 : \beta_k = c \quad v.s. \quad \mathrm{H}_1 : \beta_k \neq c,$$

由于定理 (3.5.3) 指出 OLS 估计量 $\hat{\beta}_k$ 是参数 β_k 的最优线性无偏估计量, 我们便可使用 $\hat{\beta}_k$ 作为 β_k 的近似值. 一种检验 ϕ 即是认为, 如果 OLS 估计量 $\hat{\beta}_k$ 与原假设的 c 偏离较大, 则应该拒绝 H_0, 这种检验方法被称为 **Wald 检验** (Wald test). 由于我们还没有进行检验, 则不能拒绝原假设 $\mathrm{H}_0 : \beta_k = c$ 为假, 则我们便可以依据抽样误差的绝对值 $\left|\hat{\beta}_k - \beta_k\right| = |\beta_k - c|$ 来设立检验 ϕ, 但为了具有可比性, 还需要考虑系数自身大小影响, 故我们考虑对抽样误差做标准化处理——也就是说, 我们使用 $\hat{\beta}_k$ 的 t 统计量

$$T = \frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)} \overset{\mathrm{H}_0}{=\!=} \frac{\beta_k - c}{\mathrm{SE}\left(\hat{\beta}_k\right)}$$

作为检验 ϕ 的检验统计量, 且检验 ϕ 为

$$\phi : \text{若 } |T| > C \text{则拒绝 } \mathrm{H}_0, \text{其中} C > 0.$$

现在我们来研究检验 ϕ 可以具有的置信水平. 考虑其功效函数 $\beta_\phi(\beta_k)$ 为

$$\begin{aligned}
\beta_\phi(\beta_k) =& \mathbb{P}\left(\text{检验}\phi\text{下 } \mathrm{H}_0\text{被拒绝}\right) = \mathbb{P}(|T| > C) \\
=& \mathbb{P}(T < -C) + \mathbb{P}(T > C) = F(-C) + 1 - F(C) \\
=& 1 - F(C) + 1 - F(C) = 2(1 - F(C)),
\end{aligned}$$

故检验 ϕ 犯第一类错误的概率为 $\alpha_1(\theta|\,\phi) = \begin{cases} 2(1 - F(C)), \beta_k = c, \\ 0, \beta_k \neq c. \end{cases}$ 这样, 为了保障检验 ϕ 具有 α 的显著性水平, 只需有

$$\alpha_1(\theta|\,\phi) \leqslant \alpha \Rightarrow 2(1 - F(C)) \leqslant \alpha \Rightarrow F(C) \geqslant 1 - \frac{1}{2}\alpha \Rightarrow C = F^{-1}\left(1 - \frac{1}{2}\alpha\right),$$

这里的 F 是自由度为 $n - K$ 的 t 分布的累计密度函数,F^{-1} 为 F 的函数. 上面的推导过程用到了 t 分布为对称分布以及 F 为增函数这两条性质.$C = F^{-1}\left(1 - \frac{1}{2}\alpha\right)$ 便是 t 分布的上 $\alpha/2$ 分位数.

定理 3.13.1 (回归模型单个系数的 t 检验). *对于模型 (3.12.1) 的回归系数 β_k, 就假设*

$$\mathrm{H}_0 : \beta_k = c \quad v.s. \quad \mathrm{H}_1 : \beta_k \neq c,$$

而言, 以 $\hat{\beta}_k$ 的 t 统计量

$$T = \frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)} \overset{\mathrm{H}_0}{=\!=} \frac{\beta_k - c}{\mathrm{SE}\left(\hat{\beta}_k\right)} \tag{3.77}$$

作为检验统计量, 则有显著性水平为 α的检验

$$\phi : 若\, |T| > F^{-1}\left(1 - \frac{1}{2}\alpha\right) 则拒绝\ H_0,$$

其中 F 是自由度为 $n-K$ 的 t 分布的累计密度函数,F^{-1} 为 F 的函数. 我们称检验 ϕ 为回归系数 β_k 的 ***t* 检验** *(t test).*

特别地, 当定理 (3.13.1) 中假设常数 c 取 0 时, 则此时的 t 检验称为回归系数 β_k 的**显著性检验** (significance test). 如果检验不能通过, 则意味着我们接受 (不能拒绝)$\mathrm{H}_0 : \beta_k = 0$, 这意味着在 α 的显著性水平下,β_k 对应的解释变量 X_k 对于被解释变量 Y 没有解释力 (因为系数为 0).

除了计算检验统计量 T 外, 我们在概率论与数理统计中学过另一种判断假设是否成立的方法即是计算 p 值.

定义 3.13.3 (p 值). *对于一个假设检验, 设检验统计量为 S, 若检验 ϕ 为*

$$\phi : 若 S > C 则拒绝 \mathrm{H}_0, C \in \mathbb{R},$$

则检验 ϕ 的 ***p* 值** *(p-value)为 $p = 1 - G(S)$, 其中 G 是 H_0 下检验统计量的累计密度函数.*

p 值是原假设 H_0 为真时依据检验统计量的累计密度函数所计算的, 进一步计算有

$$p = 1 - G(S) = 1 - \mathbb{P}\left(s \leqslant S_{实现值} \middle| \mathrm{H}_0\right) = \mathbb{P}\left(s \geqslant S_{实现值} \middle| \mathrm{H}_0\right)$$

即 p 值是原假设成立的情形下, 检验统计量 S 的取值不小于由样本计算得到的实现值的概率. 显然, 对于检验 ϕ 是 S 的实现值越大则越有可能拒绝 H_0, 则计算的 p 值越小, 则出现大于 S 的实现值的情况越少. 若检验 ϕ 具有 α 的显著性水平, 则有

$$\beta_\phi(\beta_k) = \mathbb{P}\left(检验\phi下\ \mathrm{H}_0被拒绝\right) = \mathbb{P}(S > C) \leqslant \alpha$$

成立, 这实际上也就是说检验统计量 S 的取值出现大于 C 的概率要小于 α. 若我们计算出 p 值满足有 $p \leqslant \alpha$, 即检验统计量 S 出现了由样本决定的实现值的概率小于等于 α,

那么必然有 S 的现实值大于 C, 这论断可以通过如图 (3.4) 所示的累计密度函数函数形象说明.

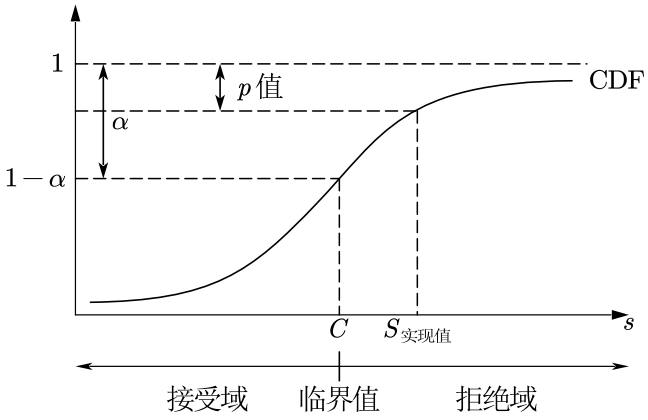

图 3.4: 检验 ϕ 的 p 值

一种常见的误解是认为 p 值是原假设成立的概率, 这是错误的.p 值仅仅是一个由样本计算的检验统计量的现实值所决定的**随机变量**,p 值越小, 则表明样本观测值越不支持原假设 H_0, 且 p 值与显著性水平 α 是相关联的.

对于定理 (3.13.1), 则其 p 值为

$$\begin{aligned} p =& \mathbb{P}\left(s \geqslant S_{实现值} \middle| \mathrm{H}_0\right) = \mathbb{P}\left(|t| \geqslant |T|\right) = \mathbb{P}\left(t \geqslant |T|\right) + \mathbb{P}\left(t \leqslant -|T|\right) \\ =& 1 - F\left(|T|\right) + F\left(-|T|\right) = 1 - F\left(|T|\right) + 1 - F\left(|T|\right) \\ =& 2\left(1 - F\left(|T|\right)\right), \end{aligned}$$

其中 F 是自由度服从 $n-K$ 的 t 分布的累计密度函数.

既然我们能够对某个单个系数 β_k 进行假设检验, 自然对于模型 (3.12.1) 我们需要进行 K 次 t 检验才能得到

$$\begin{cases} \mathrm{H}_{01}: \beta_1 = c_1 \quad v.s. \quad \mathrm{H}_{11}: \beta_1 \neq c_1, \\ \cdots \\ \mathrm{H}_{0k}: \beta_k = c_k \quad v.s. \quad \mathrm{H}_{1k}: \beta_k \neq c_k, \\ \cdots \\ \mathrm{H}_{0K}: \beta_K = c_K \quad v.s. \quad \mathrm{H}_{1K}: \beta_K \neq c_K \end{cases}$$

这 K 个原假设 $\mathrm{H}_{01}, \cdots, \mathrm{H}_{0K}$ 的检验结果. 我们是否能够据此得到合并这 K 个原假设, 依据 K 次 t 检验来给出

$$\mathrm{H}_0: \boldsymbol{\beta} = \boldsymbol{c} = (c_1, \cdots, c_K)^{\mathrm{T}}$$

的结果? 答案是否定的, 因为单个系数的 t 检验都是独立进行的, 没有考虑参数 $\boldsymbol{\beta}$ 的联合分布. 如果把 K 次 t 检验的结果作为 $\mathrm{H}_0:\boldsymbol{\beta}=\boldsymbol{c}$ 的结果, 还会积累这 K 次 t 检验的第一类错误. 倘若每次 t 检验都在 α 的显著性水平上拒绝了 t 检验的原假设, 以此结果认为能够拒绝 $\mathrm{H}_0:\boldsymbol{\beta}=\boldsymbol{c}$, 那么累计的犯第一类错误概率便有上界 α^K. 故此, 我们需要对于 $\mathrm{H}_0:\boldsymbol{\beta}=\boldsymbol{c}$ 需要新的检验方法——F 检验.

假设模型 (3.12.1) 的回归线性 $\boldsymbol{\beta}$ 有假设

$$\mathrm{H}_0:\boldsymbol{R\beta}=\boldsymbol{r}\quad v.s.\quad \mathrm{H}_1:\boldsymbol{R\beta}\neq\boldsymbol{r}$$

需要检验, 其中矩阵 $\boldsymbol{R}\in\mathbb{R}^{r\times K}$ 满足有行满秩 $\mathrm{rank}\boldsymbol{R}=r$, 而列向量 $\boldsymbol{r}\in\mathbb{R}^{r\times 1}$, 则原假设即为

$$\begin{aligned}\boldsymbol{R\beta}=\boldsymbol{r}\Leftrightarrow&\begin{pmatrix}R_{11}&R_{12}&\cdots&R_{1K}\\R_{21}&R_{22}&\cdots&R_{2K}\\\vdots&\vdots&&\vdots\\R_{r1}&R_{r2}&\cdots&R_{rK}\end{pmatrix}\begin{pmatrix}\beta_1\\\beta_2\\\vdots\\\beta_K\end{pmatrix}=\begin{pmatrix}r_1\\r_2\\\vdots\\r_r\end{pmatrix}\\\Leftrightarrow&\begin{cases}R_{11}\beta_1+R_{12}\beta_2+\cdots+R_{1K}\beta_K=r_1,\\R_{21}\beta_1+R_{22}\beta_2+\cdots+R_{2K}\beta_K=r_2,\\\cdots\\R_{r1}\beta_1+R_{m2}\beta_2+\cdots+R_{rK}\beta_K=r_r,\end{cases}\end{aligned}$$

这便是 r 个关于参数 $\beta_1,\cdots,\beta_K$ 的线性方程, 其构成对于参数 $\boldsymbol{\beta}$ 的一个线性约束条件. 由于要检验的假设等价于

$$\mathrm{H}_0:\boldsymbol{R\beta}-\boldsymbol{r}=\boldsymbol{0}\quad v.s.\quad \mathrm{H}_0:\boldsymbol{R\beta}-\boldsymbol{r}\neq\boldsymbol{0},$$

我们同样使用 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 去代替总体参数 $\boldsymbol{\beta}$, 依据 Wald 检验的思想, 如果原假设成立则向量模长 $\left\|\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right\|$ 应该有较小的取值, 反之若 $\left\|\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right\|$ 的取值较大, 则应该拒绝原假设 H_0.

令 $\boldsymbol{u}=\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\in\mathbb{R}^{r\times 1}$, 则 $\boldsymbol{u}$ 为 OLS 估计量的一个线性变换, 其条件分布依然是多元正态分布. 在 $\mathrm{H}_0:\boldsymbol{R\beta}-\boldsymbol{r}=\boldsymbol{0}$ 成立的情况下, 计算 $\boldsymbol{u}$ 的条件期望、方差为

$$\begin{aligned}&\mathbb{E}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\middle|\boldsymbol{X}\right)=\boldsymbol{R}\mathbb{E}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)-\boldsymbol{r}=\boldsymbol{R\beta}-\boldsymbol{r}\overset{\mathrm{H}_0}{=\!=}\boldsymbol{0},\\&\mathrm{Var}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)=\mathrm{Var}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\middle|\boldsymbol{X}\right)=\boldsymbol{R}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)\boldsymbol{R}^{\mathrm{T}}=\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2,\end{aligned}$$

故 r 维列向量 $\boldsymbol{u}|\boldsymbol{X}\sim\mathrm{N}\left(\boldsymbol{0},\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)$.

为了度量 $\left\|\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right\|=\|\boldsymbol{u}\|$ 偏离原假设的程度, 我们还需要对 $\boldsymbol{u}$ 进行标准化, 将其转化为标准的多元正态分布, 显然这是可以通过线性变换来实现的. 由于假设式 (3.6) 保

证了 $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$ 的存在性, 则可以推知协方差矩阵 $\operatorname{Var}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)=\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2$ 是正定的. 依据定理 (3.6.4), 则可将协方差矩阵写作 $\operatorname{Var}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)=\boldsymbol{B}^2$, 其中 r 阶方阵 $\boldsymbol{B}$ 为 $\operatorname{Var}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)$ 的平方根. 不难注意到

$$\begin{aligned}\boldsymbol{u}^{\mathrm{T}}\left(\operatorname{Var}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)\right)^{-1}\boldsymbol{u}=&\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{B}^2\right)^{-1}\boldsymbol{u}=\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{B}\boldsymbol{B}\right)^{-1}\boldsymbol{u}\\=&\boldsymbol{u}^{\mathrm{T}}\boldsymbol{B}^{-\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{u}=\left(\boldsymbol{B}^{-1}\boldsymbol{u}\right)^{\mathrm{T}}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\right),\end{aligned}$$

在原假设下, 对于 $\boldsymbol{B}^{-1}\boldsymbol{u}$ 有

$$\mathbb{E}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\middle|\boldsymbol{X}\right)=\boldsymbol{B}^{-1}\mathbb{E}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)=\boldsymbol{0},$$

$$\operatorname{Var}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\middle|\boldsymbol{X}\right)=\operatorname{Var}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\middle|\boldsymbol{X}\right)=\boldsymbol{B}^{-1}\operatorname{Var}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)\boldsymbol{B}^{-1}=\boldsymbol{B}^{-1}\left(\boldsymbol{B}^2\right)\boldsymbol{B}^{-1}=\boldsymbol{I}_r,$$

故有 $\boldsymbol{B}^{-1}\boldsymbol{u}\middle|\boldsymbol{X}\sim\mathrm{N}\left(\boldsymbol{0},\boldsymbol{I}_r\right)$ 成立. 我们以 r 维列向量 $\boldsymbol{B}^{-1}\boldsymbol{u}$ 的模长 $\left\|\boldsymbol{B}^{-1}\boldsymbol{u}\right\|$ 来度量原假设的偏离情况, 也通过列向量 $\boldsymbol{B}^{-1}\boldsymbol{u}$ 的内积来衡量. 注意到 $\boldsymbol{B}^{-1}\boldsymbol{u}=\left(u_1,\cdots,u_r\right)^{\mathrm{T}}$ 的每个分量都是独立同分布的服从标准正态分布, 那么其内积便满足有

$$\boldsymbol{u}^{\mathrm{T}}\left(\operatorname{Var}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)\right)^{-1}\boldsymbol{u}=\left(\boldsymbol{B}^{-1}\boldsymbol{u}\right)^{\mathrm{T}}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\right)=\sum_{i=1}^{r}u_i^2,$$

故有 $\boldsymbol{u}^{\mathrm{T}}\left(\operatorname{Var}\left(\boldsymbol{u}\middle|\boldsymbol{X}\right)\right)^{-1}\boldsymbol{u}\middle|\boldsymbol{X}\sim\chi^2\left(r\right)$, 也即有

$$\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)\Bigg|\boldsymbol{X}\sim\chi^2\left(r\right).\tag{3.78}$$

对于形如式 (3.78) 的具有二次型的统计量

$$W=\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right),\tag{3.79}$$

我们将式 (3.79) 称为 **Wald 统计量** (Wald statistic). 式 (3.78) 即是回归系数在线性约束条件下的 Wald 统计量. 对于 Wald 统计量, 在学习了渐进理论后我们将对其进行详细讨论.

我们现在继续研究 $\mathrm{H}_0:\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{r}=\boldsymbol{0}$ 的检验问题, 需要指出式 (3.78) 较式 (3.79) 仍是有区别的, 因为式 (3.78) 中有未知的冗余参数 σ^2, 但真正的 Wald 统计量是使用了协方差矩阵的估计量 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}$. 由于正态回归模型 (3.12.1) 是自然满足同方差式 (3.7), 于是我们使用方差估计量 s^2 来替代 σ^2, 则有

$$\begin{aligned}W=&\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}s^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)\\=&\frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)}{s^2}\\=&\frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)}{s^2/\sigma^2}\\=&\frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)}{\dfrac{\left(n-K\right)s^2}{\sigma^2}/\left(n-K\right)},\end{aligned}$$

记

$$W^0=\frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\right)}{\frac{(n-K)\,s^2}{\sigma^2}/\,(n-K)},\tag{3.80}$$

我们称式 (3.80) 为**同方差的 Wald 统计量** (homoskedastic Wald statistic). 注意到 s^2 是残差向量 $\hat{\boldsymbol{\varepsilon}}$ 的函数, 式 (3.78) 是 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的函数, 而 $\hat{\varepsilon}$ 和 $\hat{\boldsymbol{\beta}}$ 是独立的, 故其两者函数也是独立的. 依据式 (3.69), 不难得知, 令 $F=W^0/r$, 则有

$$F=\frac{W^0}{r}=\frac{\chi^2\,(r)\,/r}{\chi^2\,(r)\,/\,(n-K)}\sim F\,(r,n-K)\,,\tag{3.81}$$

于是我们便得到了一个不依赖冗余参数 σ^2 的检验估计量 F, 其条件分布服从自由度为 $(r,n-K)$ 的 F 分布.

由于 $F=W^0/r$,r 为线性约束条件的方程个数, 其是固定的, 则我们依旧遵循有 F 越大, 相应地 W^0 越大, 则应该拒绝原假设 $\mathrm{H}_0:\boldsymbol{R\beta}-\boldsymbol{r}=\boldsymbol{0}$ 的检验方法. 这从 F 分布的角度而言也是如此. 设检验

$$\phi:\text{若}F>C\text{则拒绝 }\mathrm{H}_0,$$

则在 α 的显著性水平下功效函数满足有

$$\beta_\phi\,(\boldsymbol{\beta})=\mathbb{P}\left(\text{检验}\phi\text{下 }\mathrm{H}_0\text{被拒绝}\right)=\mathbb{P}\,(F>C)=1-F\,(C)\leqslant\alpha,$$

即可推知

$$F\,(C)\geqslant 1-\alpha\Rightarrow C\geqslant F^{-1}\,(1-\alpha)\,,$$

故有

$$\phi:\text{若}F\ \ >F^{-1}\,(1-\alpha)\,\text{则拒绝 }\mathrm{H}_0,$$

其中第一个 F 是检验统计量式 (3.81) 的实现值, 第二个 F 为自由度服从 $(r,n-K)$ 的 F 分布的累计密度函数,F^{-1} 为 F 的反函数,$F^{-1}\,(1-\alpha)$ 也便是 F 分布的上 α 分位数.

定理 3.13.2 (回归系数线性约束的 F 检验)**.** *对于模型 (3.12.1) 的回归系数 $\boldsymbol{\beta}$, 就假设*

$$\mathrm{H}_0:\boldsymbol{R\beta}=\boldsymbol{r}\quad v.s.\quad \mathrm{H}_1:\boldsymbol{R\beta}\neq\boldsymbol{r}$$

而言, 以同方差的 Wald 统计量式 (3.80) 得到

$$F=\frac{W^0}{r}$$

我们称 F 为 ***F* 统计量** *(F statistic). 一个显著性水平为 α的检验为*

$$\phi:\text{若}F\ \ >F^{-1}\,(1-\alpha)\,\text{则拒绝 }H_0,$$

其中第一个 F 是检验统计量式 *(3.81)* 的实现值, 第二个 $F(\cdot)$ 为自由度服从 $(r, n-K)$ 的 F 分布的累计密度函数. 我们称检验 ϕ 为回归系数 $\boldsymbol{\beta}$ 线性约束的 **F 检验** *(F test)*.

定理 (3.13.2) 的 p 值为

$$p = \mathbb{P}(f \geqslant F_{\text{实现值}} | \mathrm{H}_0) = \mathbb{P}(f \geqslant F_{\text{实现值}}) = 1 - F(F_{\text{实现值}}),$$

其中 $F(\cdot)$ 为自由度服从 $(r, n-K)$ 的 F 分布的累计密度函数.

对于原假设 $\mathrm{H}_0 : \boldsymbol{R\beta} = \boldsymbol{r}$, 定理 (3.13.2) 中统计检验量 F 是依赖于样本所确定的 OLS 估计量 $\hat{\boldsymbol{\beta}}$, 而 $\hat{\boldsymbol{\beta}}$ 则是定义于式 (3.17) 的最优化问题——这是一个无约束条件的优化问题. 我们在推导统计检验量 F 中是用到了原假设 $\mathrm{H}_0 : \boldsymbol{R\beta} = \boldsymbol{r}$ 成立的条件, 但是 OLS 估计量的推导没有用到, 故我们还可以从受约束的最优化问题

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{CLS}} &= \underset{\boldsymbol{\beta} \in \mathbb{R}^{K \times 1}}{\operatorname{argmin}} \hat{S}(\boldsymbol{\beta}), \\ s.t.\ & \boldsymbol{R\beta} = \boldsymbol{r}, \end{aligned}$$

来推导总体参数 $\boldsymbol{\beta}$ 的估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{CLS}}$, 此时我们得到的估计量称为**受约束的最小二乘估计量** (constrained least squares estimator). 我们可以以 $\hat{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 来研究 $\mathrm{H}_0 : \boldsymbol{R\beta} = \boldsymbol{r}$ 的检验问题, 但这我们留在本笔记的第6.7节. 我们仅在此介绍依据 $\hat{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 得到的 F 统计量的似然比检验, 且可以证明, 似然比检验与定理 (3.13.2) 是等价的.

设 **F 比率** (F-ratio)为

$$F = \frac{\left(\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}} - \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}\right)/r}{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}/(n-K)} = \frac{(\mathrm{RSS}^* - \mathrm{RSS})/r}{\mathrm{RSS}/(n-K)}, \tag{3.82}$$

其中 $\tilde{\boldsymbol{\varepsilon}}$ 是 CLS 估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 计算的残差向量,RSS^* 是 CLS 估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 计算的残差平方和, 而 $\hat{\boldsymbol{\varepsilon}}$ 和 RSS 则是 OLS 估计量的残差向量和残差平方和. 在第6.7节中, 可以证明, 式 (3.82) 与式 (3.81) 是相等的. 我们以 F 比率对假设 $\mathrm{H}_0 : \boldsymbol{R\beta} = \boldsymbol{r}$ 的检验, 称为 F 比率的**似然比检验** (likelihood ratio test), 其与定理 (3.13.2) 的相同的.

回到我们引出 F 检验的问题. 特别地, 当针对总体参数 $\boldsymbol{\beta}$ 的线性约束假设为 $\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0}$ 时, 即 $\boldsymbol{R}$ 为 K 阶单位阵,$\boldsymbol{r}$ 为 K 维零向量, 此时我们称 F 检验或似然比检验为模型 (3.12.1) 或方程 (3.2) 的**显著性检验** (significance test). 这时, 通过 F 检验拒绝原假设 $\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0}$, 即是意味着在给定的显著性水平 α 下, 至少有一个回归系数 β_k 不为 0.

方程 (3.2) 的显著性检验可以与拟合优度 R^2 联系起来. 当模型 (3.12.1) 含有常数项

时, 在 $\mathrm{H}_0:\boldsymbol{\beta}_{K\times 1}=(1,0,\cdots,0)^{\mathrm{T}}$ 下注意到式 (3.82) 有

$$\begin{aligned}F=&\frac{\left(\left(\boldsymbol{\varepsilon}^*\right)^{\mathrm{T}}\boldsymbol{\varepsilon}^*-\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)/r}{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n-K)}=\frac{\left(\frac{\left(\boldsymbol{\varepsilon}^*\right)^{\mathrm{T}}\boldsymbol{\varepsilon}^*}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}-\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right)(K-1)^{-1}}{\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\cdot(n-K)^{-1}}\\=&\frac{\left[1-\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}-\left(1-\frac{\left(\boldsymbol{\varepsilon}^*\right)^{\mathrm{T}}\boldsymbol{\varepsilon}^*}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right)\right](K-1)^{-1}}{\left[1-\left(1-\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right)\right]\cdot(n-K)^{-1}}\\=&\frac{\left(R^2-\left(R^*\right)^2\right)(K-1)^{-1}}{\left(1-R^2\right)\cdot(n-K)^{-1}}\end{aligned}$$

其中 $\left(R^*\right)^2$ 为 CLS 估计量下的拟合优度. 而 CLS 估计量在 $\mathrm{H}_0:\boldsymbol{\beta}_{K\times 1}=(1,0,\cdots,0)^{\mathrm{T}}$ 下回归的个体方程即为 $Y_i=\beta_1+\varepsilon_i$, 即是仅含有常数项的 OLS 估计量, 这时的 OLS 估计量的拟合值即是样本均值 $\bar{Y}$, 依据拟合优度的定义式 (3.47) 便可推知 $\left(R^*\right)^2=0$. 故在 $\mathrm{H}_0:\boldsymbol{\beta}_{K\times 1}=(1,0,\cdots,0)^{\mathrm{T}}$ 下,F 比率即为

$$F=\frac{R^2/(K-1)}{\left(1-R^2\right)/(n-K)},\tag{3.83}$$

也就是说, 对于含有常数项的模型 (3.12.1), 其方程的显著性检验的检验统计量 F 比率是关于拟合优度 R^2 的函数. 不难得知, 在 n 和 K 不变的情形下,F 是关于 R^2 的增函数.

对于式 (3.83), 其分子除以的自由度 $n-K$ 不难让我们联想至调整的拟合优度 $\bar{R}^2$, 则利用式 (3.50) 对式 (3.83) 代换得

$$\begin{aligned}F=&\frac{R^2/(K-1)}{\left(1-R^2\right)/(n-K)}=\left(\frac{1}{1-R^2}-1\right)\cdot\frac{n-K}{K-1}\\=&\left(\frac{n-1}{n-K}\left(1-\bar{R}^2\right)^{-1}-1\right)\cdot\frac{n-K}{K-1}=\frac{n-1}{K-1}\left(1-\bar{R}^2\right)^{-1}-\frac{n-K}{K-1},\end{aligned}$$

也即有

$$\bar{R}^2=1-\frac{n-1}{(K-1)F+(n-K)}.\tag{3.84}$$

这样, 针对在 $\mathrm{H}_0:\boldsymbol{\beta}_{K\times 1}=(1,0,\cdots,0)^{\mathrm{T}}$ 的 F 检验便通过式 (3.84) 将调整的 R^2 与 F 比率联系在一起, 同样地, 在 n 与 K 不变的情形下,$\bar{R}^2$ 与 F 是同向变化的. 式 (3.84) 的意义在于, 由于如果检验统计量取值较大 F 比率是显著的, 则关于 F 递增的 $\bar{R}^2$ 也是“显著的”. 这对于拟合优度 R^2 和 F 也是一样的. 在此意义而言, 对于含有常数项的模型 (3.12.1), 其方程的显著性检验亦即是针对各类拟合优度的显著性检验.

上述的分析似乎表明, 各类拟合优度越大则 F 统计量越大, 因而模型越显著——这是错误的, 因为没有考虑到样本容量 n 与解释变量的个数 K. 实际上, 只要样本容量 n

足够大, 方程的显著性总能够取得非常显著的结果.Bruce(cite) 在他的计量经济学教材中谈到:

> This was a popular statistic in the early days of econometric reporting when sample sizes were very small and researchers wanted to know if there was "any explanatory power" to their regression. This is rarely an issue today as sample sizes are typically sufficiently large that this F statistic is nearly always highly significant. While there are special cases where this F statistic is useful these cases are not typical. As a general rule there is no reason to report this F statistic.

也就是说, 针对单个方程显著性的 F 检验源自于早期计量经济学发展所面临的样本有限性, 为了确保他们的回归具有解释能力, 需要汇报 F 比率, 这一方法便被传承了下来. 我们可以分析, 尽管 R^2 非常小, 但足够大的样本容量便能使得方程的显著性检验在给定 α 的显著性水平下通过. 依据定理 (3.13.2), 如果方程的显著性检验在 α 的显著性水平下拒绝 $\mathrm{H}_0 : \boldsymbol{\beta}_{K\times 1} = (1, 0, \cdots, 0)^{\mathrm{T}}$, 则由式 (3.83) 可以推出样本容量 n 满足有

$$\left(\frac{1}{1-R^2}-1\right)\cdot\frac{n-K}{K-1} \geqslant F^{-1}(1-\alpha) \Rightarrow n-K \geqslant \frac{1-R^2}{R^2}\cdot(K-1)\cdot F^{-1}(1-\alpha).$$

我们通过 Monte Carlo 模拟来说明这一结构. 我们给定显著性水平 $\alpha = 1\%$ 以及一个非常小的 $R^2 = 0.01$, 且解释变量共有 $K = 8$ 个, 我们编写如下的 MATLAB 代码来求解所需的样本容量 n.

```
clear
clc
alpha=0.01;%显著性水平
K=8;%变量数
R2=0.01;%假定的拟合优度
n=K+1;%样本数初始值
while 1
    n=n+1;%样本数循环+1
    F=R2/(1-R2)*(n-K)/(K-1);%F统计量
    cv=finv(1-alpha,K-1,n-K);%计算临界值
    if F≥cv %显著时退出
        disp(['所需样本数:',num2str(n)])
        break
    end
end
```

代码输出的结果即是需要 $n = 1844$ 大小的样本容量. 这个数值似乎看上去很大, 但对于当代的计量经济学而言其实谈不上大样本. 中国现行的地级市共有 293 个, 如果以之作为我们研究问题的总体, 考虑 2010-2019 年的跨度为 10 年的市级年度数据, 即使是数据存在缺失, 我们也是可以轻松获得一个包含有 200 个中国地级市的跨度 10 年的面板面板数据——这便有了 $n = 2000$ 的样本容量. 如果是研究上市公司等问题, 则可获得的样本数据将更多. 在就机器学习的角度而言,$n = 1844$ 的样本容量可以说是有些苛刻了. 故对于当代的实证研究而言, 汇报方程的显著性检验其实不具有充分的理由. 而从这一角度来看, 即使是 $R^2 = 0.01$ 足够大的样本容量也使得回归结果也是显著的,R^2、$\bar{R}^2$ 和 R_{uc^2} 的意义也仅在于度量拟合上. 需要强调, 这样不是说我们应该一味地追求“大数据”的回归分析, 好的抽样比样本总量更为重要.

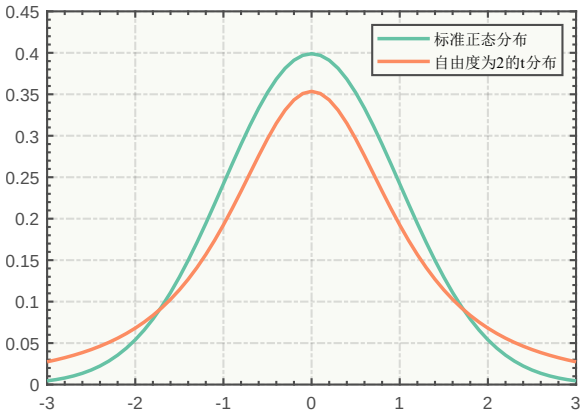

图 3.5: 正态分布与 t 分布的概率密度函数

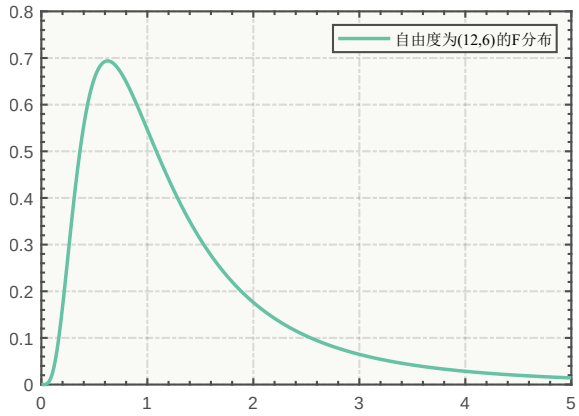

图 3.6: F 分布的概率密度函数

3.14 多元线性回归模型的预测

3.15 虚拟变量的 OLS 估计量

虚拟变量 (dummy variable), 又被译为“**哑变量**”, 即取值仅有 1 和 0 的变量, 其中 1 表示属于某个标准,0 则是不属于该标准, 因而将总体区分为了不同的**子总体** (subpopulation). 设 X 是虚拟变量, 则有

$$X = \begin{cases} 1, \text{是/属于}, \\ 0, \text{否/不属于}. \end{cases} \tag{3.85}$$

具体地, 如果虚拟变量 X 表示是否为大学生, 那么取值为 1 即为大学生,0 为不是大学生. 虚拟变量还可以被称为**零一变量** (binary variable)、**指示变量** (indicator variable).

考虑使用虚拟变量 X 对 Y 进行预测, 则有条件期望函数为

$$\mathbb{E}(Y|X)=\begin{cases}\mu_1, X=1,\\ \mu_0, X=0,\end{cases} \tag{3.86}$$

这可以等价的写为 $\mathbb{E}(Y|X)=(\mu_1-\mu_0)X+\mu_0=\beta_1X+\beta_2$, 即将 $\mathbb{E}(Y|X)$ 表示成了关于参数 $\boldsymbol{\beta}$ 线性函数. 其中参数 β_2 表示不属于标准时被解释变量的条件期望值, 参数 β_1 表示是属于标准与不属于标准的条件期望的差值. 如果 Y 表示工资收入, 虚拟变量 X 表示是否上大学, 那么参数 β_2 则是不上大学的平均工资收入, 参数 β_1 表示上大学的平均工资收入与不上大学的平均工资收入的差值.

现在假设有两个虚拟变量 X_1 和 X_2, 那么条件期望函数

$$\mathbb{E}(Y|X_1,X_2)=\begin{cases}\mu_{11}, X_1=1, X_2=1,\\ \mu_{01}, X_1=0, X_2=1,\\ \mu_{10}, X_1=1, X_2=0,\\ \mu_{00}, X_1=0, X_2=0,\end{cases}$$

可以等价的写成 $\mathbb{E}(Y|X_1,X_2)=\beta_1X_1+\beta_2X_2+\beta_3X_1X_2+\beta_4$, 变动 X_1,X_2 的取值, 则参数 $\boldsymbol{\beta}$ 的含义可以通过线性方程组

$$\begin{pmatrix}1&1&1&1\\0&1&0&1\\1&0&0&1\\0&0&0&1\end{pmatrix}\begin{pmatrix}\beta_1\\\beta_2\\\beta_3\\\beta_4\end{pmatrix}=\begin{pmatrix}\mu_{11}\\\mu_{01}\\\mu_{10}\\\mu_{00}\end{pmatrix}$$

解出为

$$\begin{pmatrix}\beta_1\\\beta_2\\\beta_3\\\beta_4\end{pmatrix}=\begin{pmatrix}0&0&1&-1\\0&1&0&-1\\1&-1&-1&1\\0&0&0&1\end{pmatrix}\begin{pmatrix}\mu_{11}\\\mu_{01}\\\mu_{10}\\\mu_{00}\end{pmatrix}=\begin{pmatrix}\mu_{10}-\mu_{00}\\\mu_{01}-\mu_{00}\\\mu_{11}-\mu_{01}-\mu_{10}+\mu_{00}\\\mu_{00}\end{pmatrix}.$$

特别地, 如果 X_1 表示是否为大学生,X_2 表示是否结婚 ($X_2=1$ 表示结婚,$X_2=0$ 表示不结婚),β_4 是非大学生的未婚者的平均工资收入, 以其为基准, 那么参数 β_1 是上大学对于不上大学的未婚者的**平均**工资收入的影响,β_2 是结婚对不上大学的未婚者的**平均**工资收入的影响, 而 β_3 则可以解释为结婚对于大学生与非大学生的平均工资收入的影响的差异, 也可以理解为上大学对于已婚者和未婚者的平均工资收入的影响的差异. 系数 β_3 对应的变量 X_2X_3 通常被称为解释变量 X_1 和 X_2 的**交互项** (interaction term)或**交叉项** (cross term), 而回归系数 β_3 则被称为**交互效应** (interaction effect).

一般地, 如果使用 p 个虚拟变量 $X_1,\cdots,X_p$ 作为解释变量, 则有条件期望函数

$$\mathbb{E}\left(Y|\,X_1,\cdots,X_p\right)=\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta},\quad \boldsymbol{\beta}\in\mathbb{R}^{2^p\times 1}, \tag{3.87}$$

其中 $\mathbf{X}$ 的分量为 $\{X_1,\cdots,X_p\}$ 的所有子集, 涵盖了所有的虚拟变量的一次项、各种交互项和一个常数项 (空集), 依据集合论的知识可知共有 2^p 个分量. 而回归系数 $\boldsymbol{\beta}$ 的含义也可以按照上述的实例去解释.

在实践中, 一个属性可以被分为多个标准, 因而出现了**分类变量** (categorical variable)或**名义变量** (nominal variable). 假设分类变量 Z 对应了 m 个类别 (不同的子总体), 即有

$$Z=\begin{cases}1,\text{第}1\text{类},\\ \vdots\\ m,\text{第}m\text{类},\end{cases} \tag{3.88}$$

那么可以定义 m 个虚拟变量

$$X_i=\begin{cases}1,\text{第}i\text{类},\\ 0,\text{其他},\end{cases},\quad i=1,\cdots,m,$$

则在给定 Z 下的 Y 的条件期望函数可以等价的表示为

$$\mathbb{E}\left(Y|\,Z\right)=\begin{cases}\mu_1,Z=1,\\ \vdots\\ \mu_m,Z=m,\end{cases}=\mathbb{E}\left(Y|\,X_1,\cdots,X_m\right)=X_1\beta_1+\cdots+X_m\beta_m, \tag{3.89}$$

此时依据 X_i 的取值不同, 则参数 $\boldsymbol{\beta}$ 满足有

$$\begin{pmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&1\end{pmatrix}\begin{pmatrix}\beta_1\\\beta_2\\\vdots\\\beta_m\end{pmatrix}=\begin{pmatrix}\mu_1\\\mu_2\\\vdots\\\mu_m\end{pmatrix}\Leftrightarrow\begin{pmatrix}\beta_1\\\beta_2\\\vdots\\\beta_m\end{pmatrix}=\begin{pmatrix}\mu_1\\\mu_2\\\vdots\\\mu_m\end{pmatrix},$$

因而 β_i 即表示属于第 i 类时被解释变量 Y 的条件期望. 需要指出, 式 (3.89) 与式 (3.87) 的不同之处在于, 含有 m 个虚拟变量的 $(Y|\,X_1,\cdots,X_m)$ 是不含有常数项的.

在通常设定中, 回归模型总是含有常数项的, 而式 (3.89) 则不含有常数项, 我们能将其转化为含有常数项的形式? 答案是肯定的. 因为在给定前 $m-1$ 个虚拟变量时,

$$\mathbb{E}\left(Y|\,Z\right)=\begin{cases}\mathbb{E}\left(Y|\,X_1,\cdots,X_{m-1},1\right)\\ \mathbb{E}\left(Y|\,X_1,\cdots,X_{m-1},0\right)\end{cases}=\begin{cases}X_1\beta_1+\cdots+X_{m-1}\beta_{m-1}+\beta_m,X_m=1,\\ X_1\beta_1+\cdots+X_{m-1}\beta_{m-1},X_m=0,\end{cases}$$

则 $\mathbb{E}(Y|Z)$ 相当于一个由 X_m 所决定的条件期望函数, 则仿照式 (3.86) 的等价变形, 代入 μ_1 和 μ_0 得到

$$\mathbb{E}(Y|Z) = \mathbb{E}(Y|X_1,\cdots,X_{m-1}) = \beta_m + X_1\beta_1 + \cdots + X_{m-1}\beta_{m-1}, \tag{3.90}$$

对于含有截距的给定分类变量 Z 的条件期望函数式 (3.90), 其回归系数 $\beta_1.\cdots,\beta_{m-1}$ 不同于式 (3.89). 式 (3.90) 的 β_i 的含义为第 i 类子总体与第 m 类子总体的被解释变量 Y 的条件期望的差值 $(i=1,\cdots,m-1)$, 而式 (3.89) 的 β_i 的含义就是第 i 类子总体的被解释变量 Y 的条件期望.

在实证中, 对于含有 m 个类别的分类变量 Z, 可以依据含有 m 个虚拟变量的式 (3.89) 构建回归模型, 此时的模型没有截距项, 也可以使用含有 $m-1$ 个虚拟变量的式 (3.90) 构建回归模型, 此时模型含有截距项.

特别地, 如果使用了含有 m 个虚拟变量的式 (3.89) 构造回归模型, 但又引入了截距项 β_0, 此时样本回归模型

$$Y_i = X_{i1}\beta_1 + \cdots + X_{im}\beta_m + \beta_0 + \varepsilon_i$$

便会存在严格多重共线性, 我们称为**虚拟变量陷阱** (dummy variable trap). 这是因为若设解释变量的第 i 个样本个体为 $\mathbf{X}_i = \left(\boldsymbol{\alpha}_{1\,i}^{\mathrm{T}}, 1\right)^{\mathrm{T}}$, 其中 m 维列向量 $\boldsymbol{\alpha}_i$ 仅有一个元素为 1, 其余元素都为 0, 那么在构造设计矩阵时

$$\boldsymbol{X}^{\mathrm{T}} = (\mathbf{X}_1,\cdots,\mathbf{X}_n) = \begin{pmatrix} \boldsymbol{\alpha}_1 & \cdots & \boldsymbol{\alpha}_n \\ 1 & \cdots & 1 \end{pmatrix}$$

则有 $\sum_{i=1}^{n}\boldsymbol{\alpha}_i = (1,\cdots,1)^{\mathrm{T}}$, 即引入的常数项 β_0 的解释变量被 m 个虚拟变量线性表出, 存在了严格多重共线性.

3.16 实证分析中 OLS 估计量的呈现方式

3.17 代码

```
%% 正交投影与OLS估计量
clear
clc
Y=[-0.5;1.25;2];
X=[-1  1
    1  1
```

```
7       0   1 ];
8  Ze=[0,0];%设定箭头位置
9  % a=0:1;%线性变换系数
10 a=[0  0  1  1;
11      0  1  1  0];%线性变换矩阵
12 Span=X*a;%张成的空间
13 Yhat=X*(X\Y);%拟合值_投影
14 e=Y-Yhat;%残差
15 selfGrootDefault(2)%绘图美化函数,没有请注释
16 figure(1)
17 quiver3(0,0,0,Y(1),Y(2),Y(3),"off","LineWidth",2.5)%绘制Y
18 hold on
19 quiver3(Ze,Ze,Ze,X(1,:),X(2,:),X(3,:),"off","LineWidth",2.5)%绘制X
20 fill3(Span(1,:),Span(2,:),Span(3,:),[0  0.4470  0.7410], ...
21      "LineStyle","none")%X张成的平面
22 quiver3(0,0,0,Yhat(1,:),Yhat(2,:),Yhat(3,:), ...
23      "off","LineWidth",2.5)%绘制投影
24 quiver3(Yhat(1,:),Yhat(2,:),Yhat(3,:),e(1,:),e(2,:),e(3,:), ...
25      "off",'-.',"LineWidth",2.5)%绘制残差
26 % quiver3(-2,0,0,1,0,0,4,'k','filled','LineWidth',2);%绘制x轴
27 % quiver3(0,0,0,0,1,0,2.5,'k','filled','LineWidth',2);%绘制y轴
28 % quiver3(0,0,0,0,0,1,2.5,'k','filled','LineWidth',2);%绘制z轴
29 % xlim([-1.5,1.5])
30 % ylim([0,2.5])
31 % zlim([0  3])
32 xlabel('$x$','FontSize',14,'Interpreter','latex')
33 ylabel('$y$','FontSize',14,'Interpreter','latex')
34 zlabel('$z$','FontSize',14,'Interpreter','latex')
35 legend('Y','X','X张成的平面','投影向量','残差向量','Location','north')
36
37 figure(2)%绘制拟合直线和散点图
38 hold on
39 xline(0,'-.','LineWidth',2)
40 yline(0,'-.','LineWidth',2)
41 scatter(X(:,1),Y,"filled","LineWidth",2)
42 y=[-5:1:5;ones(size(-5:1:5))]'*(X\Y);
43 plot(-5:1:5,y,'LineWidth',2)
44 xlim([-2.5  2.5])
45 ylim([-1  2.5])
46 xlabel('$x$','FontSize',14,'Interpreter','latex')
47 ylabel('$y$','FontSize',14,'Interpreter','latex')
48 legend('','','样本点','样本回归直线','Location','northwest')
```

第四章　概率极限及渐进理论

本章的内容旨在介绍和补充一般概统课程中未能详细论述的**概率极限**知识. 概率极限对于发展计量经济学的大样本理论极为重要, 然而国内鲜有一本专门介绍这些知识的教材, 或者需要基于测度论去建立这些内容——笔者即是提供了一个较为初等的视角, 所需的数学工具不会超出之前的章节.

4.1　依概率收敛

在微积分中, 我们已经学习过关于数列极限的知识. 对于一个数列 $\{x_n\}$, 我们使用 $\varepsilon-\delta$ 语言定义了

$$\lim_{n\to\infty} x_n = a \quad \Leftrightarrow \quad \forall \varepsilon>0, \exists N\in\mathbb{N}, \forall n>N \ s.t. \ |x_n-a|<\varepsilon,$$

则称数列 $\{x_n\}$ 收敛于常数 $a\in\mathbb{R}$. 可以说, 极限是整个微积分的基石, 微积分的所有分析都离不开极限的影子. 而本节我们将介绍概率极限的知识, 其便是数理统计的**渐进理论** (asymptotic theorem)的基石.

首先考虑一个序列 $\{X_n\}$, 其中的每一个 X_n 都是随机变量, 则我们称之为**随机序列** (random sequence). 我们现研究随机序列 $\{X_n\}$ 的性质, 自然, 我们可以按照数列极限的定义, 假设有 $\{X_n\}$ 收敛于常数 a, 但这样而言对于随机序列而言则是一个过强的假设, 且不能与随机变量的性质联系在一起. 故我们就随机变量提出依概率收敛.

定义 4.1.1 (随机变量依概率收敛). 对于一个随机序列 $\{X_n\}$, 对于任意的 $\varepsilon,\delta>0$, 若存在常数 $a\in\mathbb{R}$, 使得

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n-a|>\delta\right)=0 \quad \Leftrightarrow \quad \forall \varepsilon,\delta>0, \exists N\in\mathbb{N}, \forall n>N \ s.t. \ \mathbb{P}\left(|X_n-a|>\delta\right)<\varepsilon,$$

则称随机序列 $\{X_n\}$ **依概率收敛** *(converge in probability)*于常数 a, 同时记为

$$X_n \xrightarrow{\mathrm{P}} a\,(n\to\infty) \quad 或 \quad \operatorname*{plim}_{n\to\infty} X_n = a.$$

常数 a 也称为随机序列 $\{X_n\}$ 的**概率极限** *(probability limit)*.

对于事件 $(|X_n - a| > \delta)$, 其对立事件为 $(|X_n - a| \leqslant \delta)$, 则随机序列 $\{X_n\}$ 依概率收敛也等价于

$$\lim_{n\to\infty} \mathbb{P}(|X_n - a| \leqslant \delta) = 1$$

成立. 这样, 对比概率极限与数列极限的定义, 不能发现, 随机序列 $\{X_n\}$ 依概率收敛其实是概率 $\mathbb{P}(|X_n - a| \leqslant \delta)$ 的极限. 我们以此来认识何为概率极限.

我们知道, 如果数列 $\{x_n\}$ 收敛于 a, 则对于充分大的 n, 数列 $\{x_n\}$ 将全部落在点 a 的任意小的邻域之中——即无限接近于 a, 但不意味着等于 a. 这对于极限 $\lim\limits_{n\to\infty} \mathbb{P}(|X_n - a| \leqslant \delta) = 1, \forall \delta > 0$ 而言, 即是意味着对于充分大的 n, 随机序列 $\{X_n\}$ 落在点 a 的任意小的邻域中的**概率**, 将可以无限的接近于 1, 但是不意味为 1. 这也就是说, 随机序列 $\{X_n\}$ 收敛于常数 a, 则无论是多么大的 n, 总有随机变量 X_n 会出现在常数 a 的给定邻域之外.

我们下面研究几个依概率收敛的具体例子.

例 4.1.1. *设随机序列 $\{X_n\}$ 满足有两点分布 $\mathbb{P}(X_n = a_n) = p_n, \mathbb{P}(X_n = 0) = 1 - p_n$, 其中 $0 \leqslant p_n \leqslant 1$. 若 p_n 收敛于 0 或 a_n 收敛于 0, 则随机序列 $\{X_n\}$ 依概率收敛于 0.*

证明. 当 p_n 收敛于 0 时, 则有 $n \to \infty$ 时 $\mathbb{P}(X_n = 0) = 1 - p_n \to 1. \forall \delta > 0$, 对于事件 $|X_n - 0| = |X_n| \leqslant \delta$, 其包含有事件 $X_n = 0$, 故有

$$\mathbb{P}(X_n = 0) \leqslant \mathbb{P}(|X_n| \leqslant \delta) \leqslant 1,$$

夹逼即得 $\lim\limits_{n\to\infty} \mathbb{P}(|X_n| \leqslant \delta) = 1$.

当 a_n 收敛于 0 时,$\forall \delta > 0$ 则存在充分大的 N 使得 $\forall n > N, |a_n| \leqslant \delta$, 而 X_n 的取值仅有 0 和 a_n, 故 $\forall n > N, |X_n| \leqslant \delta$, 此时

$$\begin{aligned}\mathbb{P}(|X_n| \leqslant \delta) =& \mathbb{P}(|a_n| \leqslant \delta) + \mathbb{P}(X_n = 0)\\ =& \mathbb{P}(X_n = a_n) + \mathbb{P}(X_n = 0) = p_n + 1 - p_n = 1,\end{aligned}$$

即知 $\lim\limits_{n\to\infty} \mathbb{P}(|X_n| \leqslant \delta) = 1$. □

例 4.1.2. *设随机变量 X, 定义随机序列 $\{X_n\}$ 为 $X_n = b_n X$, 其中 $\{b_n\}$ 是实数列. 若数列 $\{b_n\}$ 收敛于 0, 则随机序列 $\{X_n\}$ 依概率收敛于 0.*

证明. 对于任意的 $\forall \delta > 0$, 则 $\mathbb{P}(|X_n| \leqslant \delta) = \mathbb{P}(-\delta \leqslant b_n X \leqslant \delta)$. 随机变量 X 可能是无界的, 但无论 X 实现值 x 如何大, 由于 b_n 收敛于 0，我们总能找到充分大的 n 使得 $-\delta \leqslant b_n X \leqslant \delta$ 成立 (序列的生成共有同一个随机变量 X), 故有 $\lim\limits_{n\to\infty} \mathbb{P}(|X_n| \leqslant \delta) = 1$. □

随机序列 $\{X_n\}$ 能依概率收敛为常数, 自然我们可以考虑其能否“收敛”至随机变量. 就此而言, 一种常用的定义是, 对于随机序列 $\{X_n\}$ 以及随机变量 X, 如果 $n \to 0$ 时有 $X_n - X \xrightarrow{\mathrm{P}} 0$, 则称随机序列 $\{X_n\}$依概率收敛于随机变量 X.

依概率收敛亦可拓展至多元的情形. 设随机向量 $\mathbf{X}_n = (X_{n1}, \cdots X_{nK})^{\mathrm{T}} \in \mathbb{R}^{K\times 1}$, 其每个分量都是随机变量, 则我们亦可称

$$\mathbf{X}_1 = (X_{11}, \cdots X_{1K})^{\mathrm{T}}, \mathbf{X}_2 = (X_{21}, \cdots X_{2K})^{\mathrm{T}}, \cdots \mathbf{X}_n = (X_{n1}, \cdots X_{nK})^{\mathrm{T}}$$

构成了随机序列 $\{\mathrm{X}_n\}$.

定义 4.1.2 (随机向量依概率收敛). 对于一个 $K \times 1$ 维随机序列 $\{\mathbf{X}_n\}$, 对于任意的 $\delta > 0$, 若存在非随机的向量 $\boldsymbol{a} \in \mathbb{R}^{K\times 1}$, 使得

$$\lim_{n\to\infty} \mathbb{P}\left(\|\mathbf{X}_n - \boldsymbol{a}\| > \delta\right) = 0$$

则称随机序列 $\{\mathrm{X}_n\}$ **依概率收敛** *(converge in probability)*于非随机的向量 $\boldsymbol{a}$, 同时记

$$\mathbf{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{a}\,(n \to \infty) \quad \text{或} \quad \operatorname*{plim}_{n\to\infty} \mathbf{X}_n = \boldsymbol{a}.$$

对于随机向量 $\mathbf{X} = (X_1, \cdots X_K)^{\mathrm{T}} \in \mathbb{R}^{K\times 1}$, 我们称

$$\mathbf{X}_n - \mathbf{X} \xrightarrow{\mathrm{P}} \mathbf{0}\,(n \to \infty)$$

为随机序列 $\{\mathbf{X}_n\}$ 依概率收敛于随机向量 $\mathbf{x}$.

可以证明, 随机序列 $\{\mathrm{X}_n\}$ 依概率收敛于非随机的 $\boldsymbol{a}$, 等价于 $\mathbf{X}_n = (X_{n1}, \cdots X_{nK})^{\mathrm{T}}$ 的每个分量收敛于 $\boldsymbol{a} = (a_1, \cdots, a_K)^{\mathrm{T}}$ 的分量.

依概率收敛是通过概率的极限来定义的, 而另一种思路则是考虑极限的概率, 由此我们得到了几乎处处收敛.

定义 4.1.3 (几乎处处收敛). 对于一个 $K \times 1$ 维随机序列 $\{\mathbf{X}_n\}$, 若存在非随机的向量 $\boldsymbol{a} \in \mathbb{R}^{K\times 1}$ 使得

$$\mathbb{P}\left(\lim_{n\to\infty} \mathbf{X}_n = \boldsymbol{a}\right) = 1$$

则称随机序列 $\{\mathrm{X}_n\}$ **几乎处处收敛** *(converge almost surely)*于非随机的向量 $\boldsymbol{a}$, 同时记 $\mathbf{X}_n \xrightarrow{a.s.} \boldsymbol{a}\,(n \to \infty)$.
对于随机向量 $\mathbf{X} = (X_1, \cdots X_K)^{\mathrm{T}} \in \mathbb{R}^{K\times 1}$, 我们称 $\mathbf{X}_n - \mathbf{X} \xrightarrow{a.s.} \mathbf{0}\,(n \to \infty)$ 为随机序列 $\{\mathbf{X}_n\}$ 几乎处处收敛于随机向量 $\mathbf{x}$.
特别地, 当 $K = 1$ 时, 即有随机序列 $\{X_n\}$ 几乎处处收敛于常数 $a(X_n \xrightarrow{a.s.} a\,(n \to \infty))$ 和随机序列 $\{X_n\}$ 几乎处处收敛于随机变量 $X(X_n - X \xrightarrow{a.s.} 0\,(n \to \infty))$.

在概率论中, **几乎处处** (almost surely)指的是概率为 1 的事件——这样的事件并不是一定发生的. 例如考虑几何概型问题, 记在单位圆中选得一给定点为事件 A, 则事件

A 是概率为 0 的事件, 但可以发生; 事件的 A 的对立事件概率为 1, 但可以不发生. 几乎处处收敛便是一种将数列收敛的概率化的方法, 其比定义 (4.1.2) 的收敛性要更强, 即 $\mathbf{X}_n \xrightarrow{a.s.} \boldsymbol{a}\,(n\to\infty) \Rightarrow \mathbf{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{a}\,(n\to\infty)$, 这是因为

$$\mathbb{P}\left(\lim_{n\to\infty}\mathbf{X}_n=\boldsymbol{a}\right)=1 \quad \Leftrightarrow \quad \forall\delta>0, \mathbb{P}\left(\|\mathbf{X}_n-\boldsymbol{a}\|\leqslant\delta\right)=1,$$

而常数 $\mathbb{P}\left(\|\mathbf{X}_n-\boldsymbol{a}\|\leqslant\delta\right)=1$ 的极限仍是常数.

回顾本节的例 (4.1.1) 的两点分布的随机序列, 由于 X_n 的取值为 0 或 a_n, 则可以得知随机序列 $\{X_n\}$ 满足数列收敛于 0, 故,$\{X_n\}$ 是几乎处处收敛于 0, 但 p_n 收敛于 0 时则不能保证几乎处处收敛. 几乎处处收敛的详细讨论需要涉及测度论的知识, 在此不过多介绍.

对于参数 θ, 如果估计量 $\hat{\theta}$ 满足有 $\hat{\theta}\xrightarrow{P}\theta$, 我们便称 $\hat{\theta}$ 是 θ 的**弱相合估计** (weak consistent estimation); 如果估计量 $\hat{\theta}$ 满足有 $\hat{\theta}\xrightarrow{a.s.}\theta$, 我们便称 $\hat{\theta}$ 是 θ 的**强相合估计** (strong consistent estimation). 通常而言, 我们将弱相合估计简称为**相合估计** (consistent estimation).

需要指出, “consistent” 一词在计量经济学中更多被翻译为 “一致的”, 故 “consistent estimation” 在绝大多数中文计量经济学教材中被称为**一致估计**. 然而, 在数学分析中 “一致性” 一词更多的是指 “uniformity”, 例如一致收敛 (uniform convergence)、一致连续 (uniform continuity)、一致可积 (uniform integrability) 等. 本笔记遵从数学分析和数理统计的范式, 将 “consistent estimation” 统称为相合估计.

4.2 大数定律与连续映射定理

本节我们介绍大数定律与连续映射定理, 将两者结合我们便可以推知样本估计量所具有的相合性.

大数定律 (law of large numbers, 简称 LLN), 可以说是概率论与数理统计最早研究的问题之一, 其的发现伴随了数学家对于何为概率的认识, 并直接导致了概率论的频率学派的诞生. 在介绍大数定律前, 本笔记首先阐述其证明工具——Markov 不等式与 Chebyshev 不等式.

定理 4.2.1 (Markov 不等式)**.** *对于非负的随机变量 X, 设其 $\mathbb{E}X$ 存在, 则对于 $\forall\varepsilon>0$ 有*

$$\mathbb{P}\left(X\geqslant\varepsilon\right)\leqslant\frac{\mathbb{E}X}{\varepsilon}, \tag{4.1}$$

我们称式 (4.1) 为 ***Markov 不等式*** *(Markov's inequality).*

证明. 假设非负的随机变量 X 具有概率密度函数 f, 注意到

$$\begin{aligned}\mathbb{P}(X \geqslant \varepsilon) &= \int_{\varepsilon}^{+\infty} f(x)\,\mathrm{d}x \leqslant \int_{0}^{+\infty} f(x)\,\mathrm{d}x \\ &\leqslant \int_{0}^{+\infty} \frac{x}{\varepsilon} f(x)\,\mathrm{d}x = \frac{1}{\varepsilon}\int_{0}^{+\infty} x f(x)\,\mathrm{d}x = \frac{\mathbb{E}X}{\varepsilon},\end{aligned}$$

这里第二个“$\leqslant$”用到了事件 $X \geqslant \varepsilon$.

再设非负的随机变量 X 是离散的, 则同理可得

$$\begin{aligned}\mathbb{P}(X \geqslant \varepsilon) &= \sum_{x_i \geqslant \varepsilon}^{\infty} \mathbb{P}(X = x_i) \leqslant \sum_{i=1}^{\infty} \mathbb{P}(X = x_i) \\ &\leqslant \sum_{i=1}^{\infty} \frac{x_i}{\varepsilon}\mathbb{P}(X = x_i) = \frac{1}{\varepsilon}\sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i) = \frac{\mathbb{E}X}{\varepsilon}.\end{aligned}$$

这样, 无论非负随机变量是连续的还是离散的, 我们都证明了式 (4.1) 是成立的. □

对于非负的随机变量 X,Markkov 不等式为事件 $X \leqslant \varepsilon$ 的发生概率提供了一个与随机变量 X 的期望有关的上界.

定理 4.2.2 (Chebyshev 不等式)**.** *对于随机变量 X, 设 $\mathbb{E}X$ 和 $\mathrm{Var}(X)$ 存在, 则对于 $\forall \varepsilon > 0$ 有*

$$\mathbb{P}(|X - \mathbb{E}X| \geqslant \varepsilon) \leqslant \frac{\mathrm{Var}(X)}{\varepsilon^2} \tag{4.2}$$

我们称式 (4.2) 为 ***Chebyshev* 不等式** *(Chebyshev's inequality).*

证明. 依据定理 (4.2.1), 对于非负的随机变量 $|X - \mathbb{E}X|$ 则有

$$\mathbb{P}(|X - \mathbb{E}X| \geqslant \varepsilon) = \mathbb{P}\left((X - \mathbb{E}X)^2 \geqslant \varepsilon^2\right) \leqslant \frac{\mathbb{E}(X - \mathbb{E}X)^2}{\varepsilon^2} = \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

□

由定理 (4.2.2) 的证明不难看出,Chebyshev 不等式 (4.2) 即是 Markov 不等式 (4.1) 的一个特例.Chebyshev 不等式的意义在于, 其对于随机变量 X 出现偏离于 $\mathbb{E}X$ 的情况给出了一个概率上界, 这个上界与随机变量的方差有关.Chebyshev 不等式亦可以用于理解为何方差刻画了随机变量的离散程度.

我们下面通过计算来详细说明 Monte Carlo 不等式.

(切比雪夫不等式的意义; 蒙特卡洛模拟阐述)

例 4.2.1. *对于随机序列 $\{X_n\}$, 设 $\mathrm{Var}(X_n) = \sigma_n^2$ 是有限的. 如果数列 $\{\sigma_n^2\}$ 收敛于 0, 则随机序列 $\{X_n\}$ 依概率收敛于 0.*

证明. 依据 Chebyshev 不等式 (4.2) 则有

$$0 \leqslant \mathbb{P}\left(|X_n - \mathbb{E}X_n| \geqslant \varepsilon\right) \leqslant \frac{\sigma_n^2}{\varepsilon^2},$$

由于 σ_n^2 收敛于 0, 故依据夹逼定理有 $\lim\limits_{n\to\infty} \mathbb{P}\left(|X_n - \mathbb{E}X_n| \geqslant \varepsilon\right) = 0$, 这样我们即知道随机序列 $\{X_n\}$ 是依概率收敛的, 但收敛于常数或是随机变量则依赖于 $\mathbb{E}X_n$. 在一些计量经济学教材中, 随机序列的方差的序列 $\{\sigma_n^2\}$ 收敛于 0 被称为**依均方收敛** (converge in mean square). □

有了 Chebyshev 不等式 (4.2) 和 Markov 不等式 (4.1), 我们现在便可以刻画随机变量的大数定律.

定理 4.2.3 (弱大数定律,weak law of large number)**.** *设 $X_i, i = 1\cdots, n$ 是独立同分布的随机变量, 且共有的总体 X 的满足 $\mathbb{E}|X| < \infty$, 则有*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mathbb{E}X, \quad (n \to \infty). \tag{4.3}$$

证明. 假设共有的总体具有有限的方差 $\mathrm{Var}(X)$, 则依据 Chebyshev 不等式 (4.2) 可得

$$0 \leqslant \mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}X\right| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}\left(\bar{X}_n\right)}{\varepsilon^2} = \frac{\mathrm{Var}(X)}{n\varepsilon^2},$$

当 $n \to \infty$ 时, 夹逼定理即得 $\bar{X}_n \xrightarrow{\mathrm{P}} \mathbb{E}X$.

上面的证明为了使用 Chebyshev 不等式, 我们假定了总体具有有限的方差 $\mathrm{Var}(X)$, 但对于定理 (4.2.3) 并不需要此条件. 由于 $\operatorname{plim}\limits_{n\to\infty} \bar{X}_n = \mathbb{E}X \Leftrightarrow \operatorname{plim}\limits_{n\to\infty} \left(\bar{X}_n - \mathbb{E}X\right) = 0$, 则令 $Y_i = X_i - \mathbb{E}X$, 定理 (4.2.3) 便等价于

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{\mathrm{P}} 0, \quad (n \to \infty),$$

不需要方差假定, 我们现在证明这是成立的. 依据 Markov 不等式 (4.1), 即得

$$\mathbb{P}\left(\left|\bar{Y}_n\right| \geqslant \varepsilon\right) \leqslant \frac{\mathbb{E}\left|\bar{Y}_n\right|}{\varepsilon},$$

可以证明, 当 $n \to \infty$ 时,$\mathbb{E}\left|\bar{Y}_n\right|$ 是收敛于 0 的[1], 这样依据夹逼定理即证当 $n \to \infty$ 时 $\bar{Y}_n \xrightarrow{\mathrm{P}} 0$. □

[1]一种方法是依据 Bahr-Esseen 不等式: 对于独立同分布的随机变量 $X_i, i = 1, \cdots, n$, 若 $\mathbb{E}X_i = 0$, 则对于任意的 $0 < r \leqslant 2$ 有

$$\mathbb{E}\left|\sum_{i=1}^{n} X_i\right|^r \leqslant 2\sum_{i=1}^{n} \mathbb{E}\left|X_i\right|^r$$

这样, 我们便证明了定理 (4.2.3), 其意义在于, 对于**独立同分布**的随机样本, 其样本均值将依概率收敛于总体均值. 一个非常直接且广泛的应用便是考虑两点分布, 设事件 A 发生的概率为 p, 记事件 A 发生时随机变量 $X=1$, 不发生则 $X=0$, 于是 X 有分布律 $\mathbb{P}(X=1)=p, \mathbb{P}(X=0)=1-p$. 考虑 n 次独立重复试验, 每次实验的结果用独立同分布的 $X_i, i=1,\cdots,n$ 表示, 则定理 (4.2.3) 意味着

$$\bar{X}_n=\frac{\text{事件}A\text{发生的次数}}{n}\xrightarrow{\mathrm{P}}p,\quad (n\to\infty),$$

换句话说,n 次独立重复试验中事件 A 发生的频率将随 n 的增大依概率收敛于概率 p, 则即是概率论的频率学派的基石.

弱大数定理——加以“弱 (weak)”字修饰, 则是因为还有强大数定律.

定理 4.2.4 (强大数定律,strong law of large number)**.** *设 $X_i, i=1\cdots,n$ 是独立同分布的随机变量, 且共有的总体 X 的满足 $\mathbb{E}|X|<\infty$, 则有*

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i\xrightarrow{a.s.}\mathbb{E}X,\quad (n\to\infty). \tag{4.4}$$

定理 (4.2.4) 指出样本均值 $\bar{X}_n$ 是几乎处处收敛于总体期望 $\mathbb{E}X$, 这个结论是强于定理 (4.2.3) 的依概率收敛. 自然, 强大数定律可推出弱大数定律. 对于定理 (4.2.4), 其证明需要使用测度论, 这超出本笔记所涉及的范围.

大数定理自然可以拓展至高维的情形. 设独立同分布的 k 维随机列向量 $\mathbf{X}_i=(X_{i1},\cdots,X_{iK})^{\mathrm{T}}, i=1,\cdots,n$, 且共有的总体 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$ 有期望向量 $\mathbb{E}\mathbf{X}=(\mathbb{E}X_1,\cdots,\mathbb{E}X_K)^{\mathrm{T}}$ 和协方差矩阵

$$\begin{aligned}\mathrm{Var}(\mathbf{X})=&\mathbb{E}\left((\mathbf{X}-\mathbb{E}\mathbf{X})(\mathbf{X}-\mathbb{E}\mathbf{X})^{\mathrm{T}}\right)\\ =&\begin{pmatrix}\mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_K)\\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2,X_K)\\ \vdots & \vdots & & \vdots\\ \mathrm{Cov}(X_K,X_1) & \mathrm{Cov}(X_K,X_2) & \cdots & \mathrm{Var}(X_K)\end{pmatrix}.\end{aligned}$$

协方差矩阵 $\mathrm{Var}(\mathbf{X})$ 我们也常记有 $\boldsymbol{\Sigma}$ 或 $\boldsymbol{\Sigma}_{\mathbf{X}}$.

成立. 对于 $\bar{Y}_n$ 即有

$$\mathbb{E}\left|\bar{Y}_n\right|=\mathbb{E}\left|\sum_{i=1}^{n}\frac{Y_i}{n}\right|\leqslant 2\sum_{i=1}^{n}\mathbb{E}\left|\frac{Y_i}{n}\right|=\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}|Y_i|=\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}|Y|,$$

由于 $\mathbb{E}|X|<\infty$, 则 $\mathbb{E}|Y|<\infty$, 故可以得知 $\mathbb{E}\left|\bar{Y}_n\right|$ 是收敛于 0. 对此之外, 还有多种方法可以证明 $\mathbb{E}\left|\bar{Y}_n\right|$ 的收敛性.

定理 4.2.5 (大数定律,law of large number)**.** 设有 $k \times 1$ 维随机向量 $\mathbf{X}_i = (X_{i1}, \cdots, X_{iK})^{\mathrm{T}}, i = 1, \cdots, n$ 是独立同分布的随机变量, 且共有的总体 $\mathbf{X}$ 的满足 $\mathbb{E}\|\mathbf{X}\| < \infty$, 则有

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \xrightarrow{P} \mathbb{E}\mathbf{X}, \quad (n \to \infty) \tag{4.5}$$

和

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \xrightarrow{a.s.} \mathbb{E}\mathbf{X}, \quad (n \to \infty), \tag{4.6}$$

其中式 *(4.5)* 是**弱大数定律** *(weak law of large number)*, 式 *(4.6)* 是**强大数定律** *(strong law of large number)*.

定理 (4.2.5) 的成立条件有三, 随机向量 $\mathbf{X}_i$ 互相独立、同分布、总体期望 $\mathbb{E}\|\mathbf{X}\|$ 有限缺一不可. 任意条件的缺失都不能保证定理 (4.2.5) 成立. 当然, 大数定理的成立条件也可以进一步的调整, 我们将在第5.1节中介绍.

依据定理 (4.2.5), 我们不难得到如下的结果.

定理 4.2.6 (弱大数定律的应用)**.** 利用式 *(4.5)*, 则有

1. 如果 $\mathbb{E}\|\mathbf{X}\| < \infty$, 则有 $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \xrightarrow{P} \mathbb{E}\mathbf{X}, n \to \infty$;
2. 如果 $\mathbb{E}\|\mathbf{X}\|^m < \infty$, 则有 $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{X}_i\|^m \xrightarrow{P} \mathbb{E}\|\mathbf{X}\|^m, n \to \infty$;
3. 如果 $\mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right) < \infty$, 则有 $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i \xrightarrow{P} \mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right), n \to \infty$;
4. 如果 $\mathbb{E}\left\|e^{\mathbf{X}}\right\| < \infty$, 则有 $\frac{1}{n}\sum_{i=1}^{n}e^{\mathbf{X}_i} \xrightarrow{P} \mathbb{E}e^{\mathbf{X}}, n \to \infty$;
5. 如果 $\mathbb{E}\|\mathbf{X}\| < \infty$, 则有 $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \xrightarrow{P} \mathbb{E}\mathbf{X}, n \to \infty$;
6. 如果 $\mathbb{E}\|\ln\mathbf{X}\| < \infty$, 则有 $\frac{1}{n}\sum_{i=1}^{n}\ln\mathbf{X}_i \xrightarrow{P} \mathbb{E}\ln\mathbf{X}, n \to \infty$;

上述的函数均是作用于分量. 一般地, 设映射 $\boldsymbol{h}: \mathbb{R}^{K\times 1} \to \mathbb{R}^{m\times 1}$, 如果 $\mathbb{E}\|\boldsymbol{h}(\mathbf{X})\| < \infty$, 则有

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{h}(\mathbf{X}_i) \xrightarrow{P} \mathbb{E}\boldsymbol{h}(\mathbf{X}), n \to \infty. \tag{4.7}$$

定理 (4.2.5) 只是给出了针对**随机向量**的收敛结果, 一个自然的问题是如果 $\mathbb{E}\left\|\mathbf{X}\mathbf{X}^{\mathrm{T}}\right\| < \infty$, 那么是否有 $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} \xrightarrow{P} \mathbb{E}\left\|\mathbf{X}\mathbf{X}^{\mathrm{T}}\right\|, n \to \infty$ 成立? 熟知矩阵微积分的读者知道, 对于矩阵我们并不能直接的从多元微积分迁移过去, 还需要考虑不少细节问题. 在此, 答案是肯定的.$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$ 是一个 K 阶方阵, 定理 (4.2.5) 可以保证其收敛性——

我们将其的证明留在的第节.

我们通过 Monte Carlo 模拟来说明定理 (4.2.6).

大数定律的结论还可以进一步被加强, 现在我们介绍概率极限的连续映射定理.

定理 4.2.7 (连续映射定理,continuous mapping theorem)**.** 对于随机向量的序列 $\{\mathbf{X}_n\}$, 设 $\mathbf{X}_n \in \mathbb{R}^{K\times 1}$, 映射 $\boldsymbol{h}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{m\times 1}$ 是一个连续映射, 则当 $n\to\infty$ 时,

1. 若 $\mathbf{X}\xrightarrow{P}\mathbf{X}$, 则有 $\boldsymbol{h}(\mathbf{X}_n)\xrightarrow{P}\boldsymbol{h}(\mathbf{X})$;
2. 若 $\mathbf{X}\xrightarrow{a.s.}\mathbf{X}$, 则有 $\boldsymbol{h}(\mathbf{X}_n)\xrightarrow{a.s.}\boldsymbol{h}(\mathbf{X})$.

定理 (4.2.7) 实际上对应了一元微积分中连续函数所具有的极限性质, 只不过这里是对应的概率的收敛. 相似地, 我们对于连续映射 h 的要求可以进一步缩小, 我们只需收敛的随机变量 $\mathbf{X}\in\mathbb{R}^{K\times 1}$ 是概率有界的, 即 $\exists C\subset\mathbb{R}^{K\times 1}$ 使得 $\mathbb{P}(\mathbf{X}\in C)=1$, 而映射 $\boldsymbol{h}$ 在 C 中连续, 这就可以确保定理 (4.2.7) 成立. 定理 (4.2.7) 的直接应用如下.

定理 4.2.8. 设 $K\times 1$ 维随机向量的序列 $\{\mathbf{X}_n\}$ 依概率收敛于 $\mathbf{X}$,$\boldsymbol{a}\in\mathbb{R}^{K\times 1}$ 为非随机的向量, 则有

1. 当 $n\to\infty$, 则 $\mathbf{X}_n+\boldsymbol{a}\xrightarrow{P}\mathbf{X}+\boldsymbol{a}$;
2. 当 $n\to\infty$, 则 $\boldsymbol{a}^{\mathrm{T}}\mathbf{X}_n\xrightarrow{P}\boldsymbol{a}^{\mathrm{T}}\mathbf{X}$;
3. 当 $n\to\infty$, 则 $\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n\xrightarrow{P}\mathbf{X}^{\mathrm{T}}\mathbf{X}$;
4. 若 $\|\mathbf{X}_n\|\neq 0$ 且 $\|\mathbf{X}\|\neq 0$, 当 $n\to\infty$, 则 $\dfrac{\|\boldsymbol{a}\|}{\|\mathbf{X}_n\|}\xrightarrow{P}\dfrac{\|\boldsymbol{a}\|}{\|\mathbf{X}\|}$.

结合大数定律与连续映射定理, 我们可以得到更为便捷形式的大数定律.

定理 4.2.9 (弱大数定律与连续映射定理的结合)**.** 设 $\mathbf{X}_i, i=1,\cdots,n$ 是独立同分布的且有共有的总体 $\mathbf{X}\in\mathbb{R}^{K\times 1}$, 记 $\bar{\mathbf{X}}=\frac{1}{n}\sum_{i=1}^n\mathbf{X}_i$, 则有

1. 如果 $\mathbb{E}\|\mathbf{X}\|<\infty$, 则有 $\bar{\mathbf{X}}^2\xrightarrow{P}(\mathbb{E}\mathbf{X})^2, n\to\infty$;
2. 如果 $\mathbb{E}\|\mathbf{X}\|<\infty$, 则有 $\bar{\mathbf{X}}^m\xrightarrow{P}(\mathbb{E}\mathbf{X})^m, n\to\infty$;
3. 如果 $\mathbb{E}\|\mathbf{X}\|<\infty$, 则有 $e^{\bar{\mathbf{X}}}\xrightarrow{P}e^{\mathbb{E}\mathbf{X}}, n\to\infty$;
4. 如果 $\mathbb{E}\|\mathbf{X}\|<\infty$, 则有 $\ln\bar{\mathbf{X}}\xrightarrow{P}\ln\mathbb{E}\mathbf{X}, n\to\infty$;
5. 如果 $\mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)<\infty$, 则有 $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^n\mathbf{X}_i^2-\bar{\mathbf{X}}^2\xrightarrow{P}\sigma^2=\mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)-(\mathbb{E}\mathbf{X})^2, n\to\infty$;

6. 如果 $\mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)<\infty$, 则有 $\hat{\sigma}=\sqrt{\hat{\sigma}^2}\xrightarrow{P}\sigma=\sqrt{\sigma^2}, n\to\infty$.

上述的函数均是作用于分量. 一般地, 设映射 $\boldsymbol{g}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{q\times 1}$ 满足有 $\mathbb{E}\|\boldsymbol{g}(\mathbf{X})\|<\infty$, 则依据式 *(4.7)* 可知

$$\hat{\boldsymbol{\theta}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\mathbf{X}_i)\xrightarrow{P}\boldsymbol{\theta}=\mathbb{E}\boldsymbol{g}(\mathbf{X}), n\to\infty.$$

定义参数 $\boldsymbol{\beta}=\boldsymbol{h}(\boldsymbol{\theta})$, 其中 $\boldsymbol{h}$ 为连续映射 $\boldsymbol{h}:\mathbb{R}^{q\times 1}\to\mathbb{R}^{m\times 1}$, 记参数 $\boldsymbol{\beta}$ 的**插入估计量** *(plug-in estimator)*为

$$\hat{\boldsymbol{\beta}}=\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right), \tag{4.8}$$

则依定理 *(4.2.7)* 知当 $n\to\infty$, 则 $\hat{\boldsymbol{\beta}}=\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)\xrightarrow{P}\boldsymbol{\beta}=\boldsymbol{h}(\boldsymbol{\theta})$.

4.3 矩生成函数

按照通常的教材安排, 一般是先介绍大数定律, 然后介绍中心极限定理. 本笔记也遵从这一顺序, 但在介绍中心极限定理前, 我们首先介绍一个概率论中有用的工具——矩生成函数.

定义 4.3.1 (矩生成函数)**.** 对于随机变量 X, 我们称函数 $M(t)=\mathbb{E}e^{tX}$ 为随机变量 X **矩生成函数** *(moment generating function).*

依据定义 (4.3.1), 如果随机变量 X 是连续型随机变量, 则其矩生成函数为

$$M(t)=\int_{-\infty}^{+\infty}e^{tx}f(x)\,\mathrm{d}x, \tag{4.9}$$

相应地, 如果随机变量 X 是离散型随机变量, 则有

$$M(t)=\sum_{i=1}^{\infty}e^{tx_i}\mathbb{P}(X=x_i). \tag{4.10}$$

我们知道, 并不是所有的随机变量其分布是存在数学期望的, 例如 Cauchy 分布的概率密度函数不能使得期望的定义积分是收敛的. 同理, 也并不是所有的随机变量都具有矩生成函数 $M(t)$[2]. 矩生成函数 $M(t)=\mathbb{E}e^{tX}$ 使用了期望, 其存在性需要广义积分或

[2]考虑复数域 $\mathbb{C}$, 若定义函数 $C(t)=\mathbb{E}e^{i\cdot tX}$, 其中 i 是虚数单位, 则函数 $C(t)$ 其是一定存在的, 我们称之为随机变量 X 的**特征函数** (characteristic function). 如果读者有学习过复分析, 不难发现, 实际上随机变量 X 的矩生成函数 $M(t)$ 是一个 Laplace 变换, 特征函数 $C(t)$ 是一个 Fourier 变换.

无穷级数收敛, 这所有的随机变量都可以保证的. 我们本节讨论的矩生成函数都是默认其存在的.

正如其名, 矩生成函数 $M(t)$ 与随机变量的矩是紧密相关的, 我们可以使用 $M(t)$ 去计算随机变量 X 的矩.

定理 4.3.1 (矩与矩生成函数). *设随机变量 X 的矩生成函数 $M(t)$ 在 $t=0$ 的邻域内存在, 如果 X 具有有限的任意矩, 则有*

$$M^{(n)}(0)=\frac{\mathrm{d}^n}{\mathrm{d}t^n}M(t)\bigg|_{t=0}=\mathbb{E}X^n$$

证明. 我们仅考虑连续型随机变量 X 显然 $M(0)=\mathbb{E}e^{0\cdot X}=\mathbb{E}1=1$. 当 $n=1$ 时有

$$M'(t)=\frac{\mathrm{d}}{\mathrm{d}t}M(t)=\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{+\infty}e^{tx}f(x)\,\mathrm{d}x=\int_{-\infty}^{+\infty}xe^{tx}f(x)\,\mathrm{d}x,$$

这里用到了变限积分的求导公式[3]

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{+\infty}F(x,t)\,\mathrm{d}x=\int_{-\infty}^{+\infty}\frac{\partial}{\partial t}F(x,t)\,\mathrm{d}x,$$

故有 $t=0$ 时有 $M'(0)=\mathbb{E}X$ 成立. 容易用数学归纳法证明

$$M^{(n)}(t)=\frac{\mathrm{d}^n}{\mathrm{d}t^n}M(t)=\int_{-\infty}^{+\infty}x^ne^{tx}f(x)\,\mathrm{d}x.$$

则有

$$M^{(n)}(0)=\int_{-\infty}^{+\infty}x^nf(x)\,\mathrm{d}x=\mathbb{E}X^n.$$

□

例 4.3.1. *使用矩生成函数计算正态分布 $N(\mu,\sigma^2)$ 的原点矩.*

解. 设 $X\sim N(\mu,\sigma^2)$, 则

$$M(t)=\int_{-\infty}^{+\infty}e^{tx}f(x)\,\mathrm{d}x=\int_{-\infty}^{+\infty}e^{tx}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{(x-\mu)^2}{2\sigma^2}+tx}\mathrm{d}x,$$

[3]一般地, 完整的变限积分的求导公式

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{g(t)}^{h(t)}F(x,t)\,\mathrm{d}x=F(h,t)\,h'-F(g,t)\,g'+\int_{g(t)}^{h(t)}\frac{\partial}{\partial t}F(x,t)\,\mathrm{d}x$$

又被称为 Leibniz 律.

注意到

$$\begin{aligned}-\frac{(x-\mu)^2}{2\sigma^2}+tx &= -\frac{1}{2\sigma^2}\left[x-\left(\mu+\sigma^2t\right)\right]^2+\frac{1}{2\sigma^2}\left[\left(\mu+\sigma^2t\right)^2-\mu^2\right]\\ &= -\frac{1}{2\sigma^2}\left[\left(x-\left(\mu+\sigma^2t\right)\right)^2+\mu^2-\left(\mu+\sigma^2t\right)^2\right],\end{aligned}$$

令 $u=x-\left(\mu+\sigma^2t\right), x=u+\left(\mu+\sigma^2t\right)$, 则有

$$\begin{aligned}M(t) &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}\left[u^2+\mu^2-\left(\mu+\sigma^2t\right)^2\right]}\mathrm{d}x\\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}u^2}\cdot e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2t\right)^2\right]}\mathrm{d}x\\ &= \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2t\right)^2\right]}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}u^2}\mathrm{d}x\\ &= \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2t\right)^2\right]}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}u^2}\mathrm{d}u\\ &\xlongequal{\text{正态分布的 pdf}} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2t\right)^2\right]}\cdot\sigma\sqrt{2\pi}\\ &= e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2t\right)^2\right]} = e^{-\frac{1}{2\sigma^2}\left(\mu^2-\mu^2-2\mu\sigma^2t-\sigma^4t^2\right)} = e^{\frac{1}{2}\sigma^2t^2+\mu t},\end{aligned}$$

即正态分布 $N(\mu,\sigma^2)$ 的矩生成函数为 $\boxed{M(t)=e^{\frac{1}{2}\sigma^2t^2+\mu t}}$. 下面我们来求 $M(t)$ 的高阶导数. 注意到 e^x 在 $x=0$ 处的幂级数为

$$e^x=\sum_{n=0}^{\infty}\frac{1}{n!}x^n=1+x+\frac{1}{2!}x^2+\frac{1}{3!}x^3+\cdots,$$

则可以得到 $M(t)$ 的幂级数为

$$M(t)=\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{1}{2}\sigma^2t^2+\mu t\right)^n=1+\mu t+\frac{1}{2!}\left(\mu^2+\sigma^2\right)t^2+\frac{1}{3!}\left(\mu^3+3\mu\sigma^2\right)t^3+\cdots, \quad (4.11)$$

由于 $M(t)$ 在 $t=0$ 处的幂级数

$$M(t)=\sum_{n=0}^{\infty}\frac{1}{n!}M^{(n)}(0)\,t^n$$

是唯一的, 对比系数便可得到

$$\mathbb{E}X=\mu,\quad \mathbb{E}X^2=\mu^2+\sigma^2,\quad \mathbb{E}X^3=\mu^3+3\mu\sigma^2,\quad \cdots$$

特别地, 由于 $Z=X-\mathbb{E}X=X-\mu\sim \mathrm{N}(0,\sigma^2)$, 则随机变量 X 的中心距 $\mathbb{E}(X-\mathbb{E}X)^n=\mathbb{E}Z^n$ 可以通过计算 $\mu=0$ 的矩生成函数来得到. 令 $\mu=0$, 式 (4.11) 即有

$$M(t)=\sum_{n=0}^{\infty}\frac{\sigma^{2n}t^{2n}}{2^n n!}=\begin{cases}0, k\in \text{奇数},\\ \frac{1}{k!}\frac{k!\sigma^k}{2^{k/2}\left(\frac{k}{2}\right)!}t^k, k\in \text{偶数},\end{cases}$$

不难得知, 正态分布 $N(\mu,\sigma^2)$ 的奇数阶的中心矩为 0, 偶数阶的中心矩可以依据上述公式计算. □

我们将式 (4.11) 计算的正态分布 $N(\mu,\sigma^2)$ 的高阶矩整理如表 (4.1).

表 4.1: 正态分布 $N(\mu,\sigma^2)$ 的高阶矩

阶数 n	原点矩 $\mathbb{E}X^n$	中心矩 $\mathbb{E}\left(X-\mathbb{E}X\right)^n$
1	μ	0
2	$\mu^2+\sigma^2$	σ^2
3	$\mu^2+3\mu\sigma^2$	0
4	$\mu^4+6\mu^2\sigma^2+3\sigma^4$	$3\sigma^4$
5	$\mu^5+10\mu^3\sigma^2+15\mu\sigma^4$	0
6	$\mu^6+15\mu^4\sigma^2+45\mu^2\sigma^4+15\sigma^6$	$15\sigma^6$
7	$\mu^7+21\mu^5\sigma^2+105\mu^3\sigma^4+105\mu\sigma^6$	0
8	$\mu^8+28\mu^6\sigma^2+210\mu^4\sigma^4+420\mu^2\sigma^6+105\sigma^8$	$105\sigma^8$

利用矩生成函数计算矩并不是其最重要的性质, 矩生成函数最宝贵的性质在于其唯一性.

定理 4.3.2 (矩生成函数的唯一性)**.** *设随机变量 X 的矩生成函数 $M(t)$ 存在, 则 $M(t)$ 的形式是唯一的.*

定理 (4.3.2) 的证明涉及了 Laplace 变换的反演问题, 其超出了本笔记的范围. 定理 (4.3.2) 的一个直接应用便是, 如果随机变量 X 和 Y 具有相同的矩生成函数, 那么 X 和 Y 的概率分布是相同的. 当然, 由于不是所有的随机变量 X 都存在矩生成函数[4], 定理 (4.3.2) 在使用上便具有一定局限性.

与矩生成函数相关的另一种技巧便是累积量与累积量生成函数.

定义 4.3.2 (累积量与累积量生成函数)**.** *设随机变量 X 存在有矩生成函数 $M(t)=\mathbb{E}e^{tX}$, 我们称函数 $K(t)=\ln M(t)$ 为随机变量 X 的***累积量生成函数*** (cumulant generating function). 我们称累积量生成函数 $K(t)$ 的在 $t=0$ 处的 n 阶导数 $\kappa_n=K^{(n)}(0)$ 为随机变量 X 的 n 阶***累积量*** (cumulant).*

累积量 κ_n 使用希腊字母 κ 表示, 英语单词写作 Kappa. 由于累积量生成函数

[4]任意的随机变量 X 都具有特征函数, 与矩生成函数类似, 随机变量 X 的特征函数也是唯一的.

$K(t)=\ln M(t)$, 显然 $K(0)=\ln 1=0$. 利用复合函数求导公式不难得到

$$K'(t)=\frac{M'(t)}{M(t)},\quad K''(t)=\frac{M''(t)}{M(t)}-\left(\frac{M'(t)}{M(t)}\right)^2,$$

对于正态分布 $X\sim N(\mu,\sigma^2)$, 则对应的累积量 κ_n 为

$$\kappa_1=\frac{M'(0)}{M(0)}=\mu,\quad \kappa_2=\frac{M''(0)}{M(0)}-\left(\frac{M'(0)}{M(0)}\right)^2=\mu^2+\sigma^2-\mu^2=\sigma^2,$$

即我们 X 的 1 阶累积量、2 阶累积量分别是 X 的期望与方差. 对于更为高阶的累积量, 我们可以利用复合函数的高阶导数公式 (又称为 Faà di Bruno 公式) 去计算. 可以证明, 前三个累积量 κ_n 等于随机变量的对应阶的中心矩, 但更高阶的累积量 κ_n 是则是随机变量中心矩的多项式函数. 依据定义 (4.3.2), 我们可以得到累积量生成函数在 $t=0$ 处的幂级数为

$$K(t)=\sum_{n=0}^{\infty}\frac{\kappa_n}{n!}t^n=0+\kappa_1 t+\frac{\kappa_2}{2!}t^2+\frac{\kappa_3}{3!}t^3+\cdots \tag{4.12}$$

我们将矩、累积量与常用的随机变量的数值特征总结为表 (4.2).

表 4.2: 矩、累积量与随机变量的数值特征

阶数	矩			累积量	
	原点矩	中心矩	标准化	自身	归一化
1	均值	0	0	均值	\
2	\	方差	1	方差	1
3	\	\	偏度	\	\
4	\	\	峰度	\	超峰度

本节的最后我们研究一下样本均值 $\bar{X}_n$ 的生成函数. 设随机变量 $X_i, i=1,\cdots,n$ 是独立同分布的且具有共有的总体 X,X 的期望和方差记有 $\mu=\mathbb{E}X$ 和 $\sigma^2=\mathbb{E}(X-\mathbb{E}X)^2$. 我们知道, 样本均值 $\bar{X}_n=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$ 的期望和方差分别有 $\mathbb{E}\bar{X}_n=\mu$ 和 $\mathrm{Var}\left(\bar{X}_n\right)=n^{-1}\sigma^2$. 令

$$Z_n=\sqrt{n}\left(\bar{X}_n-\mu\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu), \tag{4.13}$$

这样随机变量 Z_n 便具有零均值和 σ^2 的方差. 我们现在计算 Z_n 的矩生成函数 $M_{Z_n}(t)$ 以及累积量生成函数 $K_{Z_n}(t)$. 记总体 X 的矩生成函数 $M_X(t)=\mathbb{E}e^{tX}$ 和累积量生成函数 $K_X(t)=\ln M_X(t)$, 则 Z_n 的矩生成函数为

$$M_{Z_n}(t)=\mathbb{E}e^{tZ_n}=\mathbb{E}e^{t\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}(X_i-\mu)}=\mathbb{E}\left(\prod_{i=1}^{n}e^{t\frac{X_i-\mu}{\sqrt{n}}}\right)\overset{\text{i.i.d.}}{=\!=\!=}\prod_{i=1}^{n}\mathbb{E}\left(e^{t\frac{X_i-\mu}{\sqrt{n}}}\right)$$

$$=\prod_{i=1}^{n}\mathbb{E}\left(e^{t\frac{X_i}{\sqrt{n}}}\cdot e^{t\frac{-\mu}{\sqrt{n}}}\right)=\prod_{i=1}^{n}e^{t\frac{-\mu}{\sqrt{n}}}\mathbb{E}e^{t\frac{X_i}{\sqrt{n}}}=\prod_{i=1}^{n}e^{t\frac{-\mu}{\sqrt{n}}}M_X\left(\frac{t}{\sqrt{n}}\right)=e^{-\mu\sqrt{n}t}\left(M_X\left(\frac{t}{\sqrt{n}}\right)\right)^n,$$

相应地,Z_n 的累积量生成函数为

$$\begin{aligned} K_{Z_n}(t) &= \ln M_{Z_n}(t) = \ln\left[e^{-\mu\sqrt{n}t}\left(M_X\left(\frac{t}{\sqrt{n}}\right)\right)^n\right] \\ &= n\ln M_X\left(\frac{t}{\sqrt{n}}\right) - \mu\sqrt{n}t = nK_X\left(\frac{t}{\sqrt{n}}\right) - \mu\sqrt{n}t. \end{aligned}$$

对于总体 X, 利用式 (4.12) 以及 $\kappa_1=\mu, \kappa_2=\sigma^2$ 便得到 Z_n 的累积量生成函数有

$$\begin{aligned} K_{Z_n}(t) &= n\sum_{m=0}^{\infty}\frac{\kappa_m}{m!}n^{-\frac{m}{2}}t^m - \mu\sqrt{n}t = \sum_{m=0}^{\infty}\frac{\kappa_m}{m!}n^{1-\frac{m}{2}}t^m - \mu\sqrt{n}t \\ &= -\mu\sqrt{n}t + \left(\kappa_1 n^{\frac{1}{2}}t + \frac{\kappa_2}{2!}t^2 + \sum_{m=3}^{\infty}\frac{\kappa_m}{m!}n^{1-\frac{m}{2}}t^m\right) \\ &\xlongequal{\kappa_1=\mu,\kappa_2=\sigma^2} -\mu\sqrt{n}t + \mu n^{\frac{1}{2}}t + \frac{\sigma^2}{2!}t^2 + \sum_{m=3}^{\infty}\frac{\kappa_m}{m!}n^{1-\frac{m}{2}}t^m \\ &= \boxed{\frac{\sigma^2}{2!}t^2 + \sum_{m=3}^{\infty}\frac{\kappa_m}{m!}n^{1-\frac{m}{2}}t^m}, \end{aligned}$$

亦即有

$$\boxed{M_{Z_n}(t) = \exp\left(K_{Z_n}(t)\right) = \exp\left(\frac{\sigma^2}{2!}t^2 + \sum_{m=3}^{\infty}\frac{\kappa_m}{m!}n^{1-\frac{m}{2}}t^m\right)}.$$

4.4 渐进分布与中心极限定理

本节我们将介绍渐进分布的内容, 与概率极限相比, 其具有更为广泛的应用场景.

定义 4.4.1 (依分布收敛). 设 $K\times 1$ 维的随机向量的序列 $\{\mathrm{X}_n\}$,$\mathbf{X}_n=(X_{n1},\cdots,X_{nK})^{\mathrm{T}}$ 的累计密度函数为 $F_n(\boldsymbol{x})=\mathbb{P}(X_{n1}<x_1,\cdots,X_{nK}<x_1)$. 若存在 $K\times 1$ 维随机变量 $\mathbf{X}$, 其具有连续的累计密度函数为 $F\{\boldsymbol{x}\}$, 若极限

$$\lim_{n\to\infty}F_n(\boldsymbol{x}) = F(\boldsymbol{x})$$

成立, 则我们称随机序列 $\{\mathrm{X}_n\}$ **依分布收敛** *(converges in distribution)*于 $\mathbf{X}$, 记作 $\mathbf{X}_n\xrightarrow{d}\boldsymbol{X}$. 我们也称 F_n 弱收敛,*converge weakly*于 F. 对于随机序列 $\mathbf{X}$ 和其分布 F, 我们称之为随机向量的序列 $\{\mathrm{X}_n\}$ 的的**渐进分布** *(asymptotic distribution)*、**大样本分布** *(large sample distribution)*或**极限分布** *(limit distribution)*.

依据定义 (4.4.1), 如果 F 对应是一个非随机的向量 $\boldsymbol{a}$ 的累计密度函数函数, 即有 $\mathbb{P}(\mathbf{X}=\boldsymbol{a})=1$, 这时我们称 F 是退化的 (degenerate). 此时当 $n\to\infty$ 时,$\mathbf{X}_n\xrightarrow{d}\boldsymbol{a}$ 便也有 $\mathbb{P}(\mathbf{X}_n=\boldsymbol{a})=1$, 则可以推出 $\mathbf{X}_n\xrightarrow{P}\boldsymbol{a}$.

如果随机序列 $\{\mathbf{X}_n\}$ 依概率收敛于随机向量 X, 则一定意味着依分布收敛于 $\mathbf{X}$. 这是因为, 对于 $\varepsilon > 0$, 我们有不等式[5]

$$\begin{aligned}&\mathbb{P}\left(X_1 < x_1 - \varepsilon, \cdots, X_K < x_1 - \varepsilon\right) - \mathbb{P}\left(\|\mathbf{X}_n - \boldsymbol{X}\| > \varepsilon\right)\\ \leqslant &\mathbb{P}\left(X_{n1} < x_1, \cdots, X_{nK} < x_1\right)\\ \leqslant &\mathbb{P}\left(X_1 < x_1 + \varepsilon, \cdots, X_K < x_1 + \varepsilon\right) + \mathbb{P}\left(\|\mathbf{X}_n + \boldsymbol{X}\| > \varepsilon\right)\end{aligned}$$

成立, 亦即

$$F\left(\boldsymbol{x} - \mathbf{1}\varepsilon\right) - \mathbb{P}\left(\|\mathbf{X}_n - \boldsymbol{X}\| > \varepsilon\right) \leqslant F_n\left(\boldsymbol{x}\right) \leqslant F\left(\boldsymbol{x} + \mathbf{1}\varepsilon\right) + \mathbb{P}\left(\|\mathbf{X}_n - \boldsymbol{X}\| > \varepsilon\right). \quad (4.14)$$

由于 $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$, 则 $n \to \infty$ 时得式 (4.14) 便有

$$F\left(\boldsymbol{x} - \mathbf{1}\varepsilon\right) \leqslant \lim_{n\to\infty} F_n\left(\boldsymbol{x}\right) \leqslant F\left(\boldsymbol{x} + \mathbf{1}\varepsilon\right),$$

由于 ε 具有任意性, 当 $\varepsilon \to 0$ 时, 即有 F_n 收敛于 F. 我们将依概率收敛与依分布收敛收敛的关系总结如下.

定理 4.4.1. *对于的随机向量的序列 $\{\mathrm{X}_n\}$,*

1. *当 $n \to \infty$ 时, 如果 $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$, 则有 $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$;*
2. *当 $n \to \infty$ 时, 如果 $\mathbf{X}_n \xrightarrow{d} \boldsymbol{a}$, 其中 $\boldsymbol{a}$ 为非随机的向量, 则有 $\mathbf{X}_n \xrightarrow{P} \boldsymbol{a}$.*

通过定理 (4.4.1), 我们可以知道依概率收敛是比依分布收敛更为强的收敛. 而大数定律告诉我们样本均值 $\bar{X}_n$ 依概率收敛于总体期望 $\mathbb{E}X$, 自然的结论便是样本均值 $\bar{X}_n$ 依分布收敛于 $\mathbb{E}X$. 但 $\mathbb{E}X$ 是非随机的, 我们希望能够得到 $\bar{X}_n$ 更为精确的渐进性质. 我们在微积分中曾学习过估计极限的阶 (order) 的方法, 例如极限

$$\lim_{n\to 0} \frac{\sin n - n}{n} = 0$$

即意味着 $\sin -n$ 收敛于 0 与 n 收敛于 0 是同阶的, 这也被认为 $\sin -n$ 与 n 有相同的“速度”收敛至 0, 我们使用小 o 表示法记有当 $n \to 0$ 时 $\sin n - n = o\left(n\right)$. 在概率极限中, 我们也有类似概率估阶以及概率小 o 表示法, 但我们不打算展开介绍这些内容.

回顾式 (4.13) 标准化后期望为 0、方差为 σ^2 的 Z_n, 则当 $n \to \infty$ 时 Z_n 与 $1/\sqrt{n}$ 是同阶. 我们现在研究 Z_n 的概率分布, 首先我们指出如下定理.

定理 4.4.2 (Lévy 连续性定理)**.** *对于随机变量的序列 $\{X_n\}$, 若存在矩生成函数*

[5]这个不等式也不显然啊 XD...

$M_{X_n}(t)=\mathbb{E}e^{tX_n}$ 满足有

$$\lim_{n\to\infty}M_{X_n}(t)=\lim_{n\to\infty}M_X(t)=\mathbb{E}e^{tX},\forall t\in\mathbb{R},$$

则有 $n\to\infty$ 时 $X_n\xrightarrow{d}X$.

对于经过标准化后的 $Z_n=\sqrt{n}\left(\bar{X}_n-\mu\right)$, 我们计算了其累积量生成函数为

$$K_{Z_n}(t)=\frac{\sigma^2}{2!}t^2+\sum_{m=3}^{\infty}\frac{\kappa_m}{m!}n^{1-\frac{m}{2}}t^m,$$

由于 $m\geqslant 3$ 时 $1-\frac{m}{2}$ 小于 0, 故有

$$\lim_{n\to\infty}K_{Z_n}(t)=\lim_{n\to\infty}\left(\frac{\sigma^2}{2!}t^2+\sum_{m=3}^{\infty}\frac{\kappa_m}{m!}n^{1-\frac{m}{2}}t^m\right)=\frac{\sigma^2}{2!}t^2,$$

因而 Z_n 的矩生成函数满足有

$$\lim_{n\to\infty}M_{Z_n}(t)=\lim_{n\to\infty}e^{K_{Z_n}(t)}=e^{\frac{\sigma^2}{2!}t^2},$$

依据定理 (4.4.2) 则可得当 $n\to\infty$ 时 Z_n 是依分布收敛于某个随机变量. 注意到在例 (4.3.1) 中我们计算的正态分布的矩生成函数为 $M(t)=e^{\frac{1}{2}\sigma^2t^2+\mu t}$, 当 $\mu=0$ 时即为 Z_n 矩生成函数的极限. 也就是说, 对于独立同分布计算出的标准化的样本均值 Z_n, 当样本容量 $n\to\infty$ 时,Z_n 的渐进分布为 $\mathrm{N}(0,\sigma^2)$. 我们便由此得到了中心极限定理.

定理 4.4.3 (Lindeberg-Lévy 中心极限定理,Lindeberg-Lévy central limit theorem)**.** 对于独立同分布的样本 $X_i,i=1,\cdots,n$, 其共有总体满足有 $\mathbb{E}X^2<\infty$, 当 $n\to\infty$ 则

$$Z_n=\sqrt{n}\left(\bar{X}_n-\mu\right)\xrightarrow{d}\mathrm{N}\left(0,\sigma^2\right),\tag{4.15}$$

其中 $\mu=\mathbb{E}X$,$\sigma^2=\mathrm{Var}(X)=\mathbb{E}(X-\mu)^2$.

式 (4.15) 意味着当 n 充分大时候,Z_n 的累计密度函数函数与 $\mathrm{N}(0,\sigma^2)$ 的累计密度函数的差异可以任意的小, 因而我们常常使用 $\mathrm{N}(0,\sigma^2)$ 来近似大样本情况下 Z_n 的概率分布. 具有这样性质的 Z_n 我们又称其具有**渐近正态性** (asymptotic normality). 式 (4.15) 的一种常见变形便是

$$\bar{X}_n\overset{a}{\sim}\mathrm{N}\left(\mu,\frac{\sigma^2}{n}\right),\tag{4.16}$$

这里 $\sim$ 上的 a 表示 “近似 (approximation)”, 即式 (4.16) 表示当 n 足够大时, 我们可以认为其分布是近似于正态分布的.

需要指出, 我们在利用定理 (4.4.2) 去描绘中心极限定理为何会出现时, 假设了随机变量 X 的矩生成函数是存在的, 这不是所有的随机变量都具有的. 更一般地, 我们可以利用所有随机变量都一定存在特征函数 $\mathbb{E}e^{itX}$ 去证明定理 (4.15).

我们现在将中心极限定理定理推广到多元. 对于随机向量, 其是否依分布收敛与一元情形有如下的关联.

定理 4.4.4 (Cramér-Wold 方式,Cramér-Wold device)**.** *对于随机向量 $\mathbf{X}_n \in \mathbb{R}^{K\times 1}$, 设 $\forall \boldsymbol{\lambda} \in \mathbb{R}^{K\times 1}$ 且 $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\lambda} = 1$, 则有 $\mathbf{X}_n \xrightarrow{d} \mathbf{X} \Leftrightarrow \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{X}_n \xrightarrow{d} \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{X}$.*

我们利用定理 (4.4.4) 来推导随机变量的中心极限定理定理.

定理 4.4.5 (Lindeberg-Lévy 中心极限定理,Lindeberg-Lévy central limit theorem)**.** *对于独立同分布的样本 $\mathbf{X}_i \in \mathbb{R}^{K\times 1}, i = 1, \cdots, n$, 其共有总体满足有 $\mathbb{E}\left\|\mathbf{X}\right\|^2 < \infty$, 则*

$$\mathbf{Z}_n = \sqrt{n}\left(\bar{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \tag{4.17}$$

其中 $\boldsymbol{\mu} = \mathbb{E}\mathbf{X}, \boldsymbol{\Sigma} = \mathrm{Var}\left(\mathbf{X}\right) = \mathbb{E}\left(\left(\mathbf{X} - \boldsymbol{\mu}\right)\left(\mathbf{X} - \boldsymbol{\mu}\right)^{\mathrm{T}}\right)$.

证明. 设

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i = \left(\bar{X}_1, \cdots, \bar{X}_K\right)^{\mathrm{T}} = \left(\frac{1}{n}\sum_{i=1}^{n}X_{i1}, \cdots, \frac{1}{n}\sum_{i=1}^{n}X_{iK}\right)^{\mathrm{T}}$$

及 $\boldsymbol{\mu} = \mathbb{E}\mathbf{X} = \left(\mathbb{E}X_1, \cdots, \mathbb{E}X_K\right)^{\mathrm{T}}$, 对于 $\forall \boldsymbol{\lambda} \in \mathbb{R}^{K\times 1}$ 且 $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\lambda} = 1$, 则

$$\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}_n = \sqrt{n}\boldsymbol{\lambda}^{\mathrm{T}}\left(\bar{\mathbf{X}}_n - \boldsymbol{\mu}\right) = \sum_{k=1}^{K}\lambda_k\sqrt{n}\left(\bar{X}_k - \mu_k\right)$$

而依据定理 (4.4.3) 知当 $n \to \infty$ 时 $\lambda_k\sqrt{n}\left(\bar{X}_k - \mu_k\right) \xrightarrow{d} \mathrm{N}\left(0, \lambda_k^2\mathrm{Var}\left(X_k\right)\right)$, 这样 $\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}_n$ 在 $n \to$ 时收敛于 K 个正态分布的和, 则其渐进分布仍是服从正态分布的. 显然, 这个正态分布的均值为 0, 而其方差的计算需要考虑到 K 个变量 X_i 间的相关关系, 其结果可以使用二次型表示为 $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\lambda}$, 故有

$$\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}_n = \sum_{k=1}^{K}\lambda_k\sqrt{n}\left(\bar{X}_k - \mu_k\right) \xrightarrow{d} \mathrm{N}\left(0, \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\lambda}\right).$$

设 $\mathbf{Z} \in \mathbb{R}^{K\times 1} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$, 注意到 $\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\lambda}\right)$, 即得

$$\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}_n \xrightarrow{d} \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}, \quad n \to \infty,$$

故依据定理 (4.4.4) 得到当 $n \to \infty$ 时有式 (4.17) 成立. □

尽管依分布收敛要弱于依概率收敛, 但中心极限定理为我们提供了关于样本均值的具体的渐进分布. 同样地, 依分布收敛已有连续映射定理, 以此我们可以推知不少统计量的渐进分布.

定理 4.4.6 (连续映射定理,continuous mapping theorem)**.** *对于随机向量的序列* $\{\mathbf{X}_n\}$*, 设* $\mathbf{X}_n \in \mathbb{R}^{K\times 1}$*, 映射* $\boldsymbol{h}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{m\times 1}$ *是一个连续映射, 则当* $n\to\infty$ *时, 若* $\mathbf{X}\xrightarrow{d}\mathbf{X}$*, 则有* $\boldsymbol{h}(\mathbf{X}_n)\xrightarrow{d}\boldsymbol{h}(\mathbf{X})$.

同样地, 定理 (4.4.6) 的使用条件只需保障局部是连续的:$\exists C\subset\mathbb{R}^{K\times 1}$ 使得 $\mathbb{P}(\mathbf{X}\in C)=1$, 而映射 $\boldsymbol{h}$ 在 C 中连续. 定理 (4.4.6) 最早由 Mann 和 Wald (1943) 证明, 故又称为 **Mann-Wald 定理** (Mann-Wald Theorem). 定理 (4.4.6) 的一个特例是 Slutsky 定理[6].

定理 4.4.7 (Slutsky 定理,Slutsky' s Theorem)**.** *当* $n\to\infty$ *时, 若* $\mathbf{X}_n$ *和* $\mathbf{Y}_n$ *满足有* $\mathbf{X}_n\xrightarrow{d}\mathbf{X},\mathbf{Y}_n\xrightarrow{P}\mathbf{Y}$*, 则*

1. *当* $n\to\infty$ *时,*$\mathbf{X}_n+\mathbf{Y}_n\xrightarrow{d}\mathbf{X}+\mathbf{Y}$*;*
2. *当* $n\to\infty$ *时,*$\mathbf{X}_n^{\mathrm{T}}\mathbf{Y}_n\xrightarrow{d}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$*;*
3. *当* $n\to\infty$ *且* $\|\mathbf{Y}\|\neq 0$ *时,*$\dfrac{\mathbf{X}_n}{\|\mathbf{Y}_n\|}\xrightarrow{d}\dfrac{\mathbf{X}}{\|\mathbf{Y}\|}$.

定理 (4.4.7) 中涉及的计算都是连续映射, 故其是定理 (4.4.6) 的特例.

连续映射定理并没有告诉我们映射后的渐进分布, 其只是保证了收敛性是成立的. 为了得到确切的渐进分布, 我们考虑一个连续可微的映射 $\boldsymbol{h}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{m\times 1}$, 对于我们感兴趣的参数 $\boldsymbol{\theta}\in\mathbb{R}^{K\times 1}$, 则 $\boldsymbol{h}$ 在 $\boldsymbol{\theta}$ 处的 Jacobi 矩阵为

$$\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})=\begin{pmatrix}\dfrac{\partial h_1(\boldsymbol{\theta})}{\partial x_1} & \cdots & \dfrac{\partial h_1(\boldsymbol{\theta})}{\partial x_K}\\ \vdots & & \vdots\\ \dfrac{\partial h_m(\boldsymbol{\theta})}{\partial x_1} & \cdots & \dfrac{\partial h_m(\boldsymbol{\theta})}{\partial x_K}\end{pmatrix}_{m\times K}, \tag{4.18}$$

这里的 x 表示求偏导的分量, 因而 $\boldsymbol{h}(\boldsymbol{u})$ 在 $\boldsymbol{u}=\boldsymbol{\theta}$ 处的一阶 Taylor 展开为

$$\boldsymbol{h}(\boldsymbol{u})=\boldsymbol{h}(\boldsymbol{\theta})+\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\cdot\boldsymbol{u}+R(\boldsymbol{\zeta}),$$

其中

$$R(\boldsymbol{\zeta})=\frac{1}{2}\left(\sum_{j=1}^{K}u_j\frac{\partial}{\partial x_j}\right)^2\boldsymbol{h}(\boldsymbol{\zeta})$$

[6]Slutsky(1880-1948), 在微观经济学中我们学过他的综合了收入效应和替代效应的 Slutsky 方程.

为 Lagrange 余项, 而 $\boldsymbol{\zeta}$ 介于 $\boldsymbol{u}$ 与 $\boldsymbol{\theta}$ 之间. 余项 $R(\boldsymbol{\zeta})$ 满足有 $R(\boldsymbol{\xi}) = o(\|\boldsymbol{u}\|), \|\boldsymbol{u}-\boldsymbol{\theta}\| \to 0$.

设我们有参数 $\boldsymbol{\theta}$ 的相合估计量 $\hat{\boldsymbol{\theta}}$, 那么映射 $\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)$ 满足有

$$\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{h}(\boldsymbol{\theta}) + \boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}} + R(\boldsymbol{\zeta}), \tag{4.19}$$

其中余项 $R(\boldsymbol{\zeta})$ 的 $\boldsymbol{\zeta}$ 介于估计量 $\hat{\boldsymbol{\theta}}$ 与参数 $\boldsymbol{\theta}$ 之间. 倘若我们已经知道估计量 $\hat{\boldsymbol{\theta}}$ 满足有渐进分布 $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \xrightarrow{d} \boldsymbol{\xi}$, 则随机向量 $\boldsymbol{\xi}$ 的期望为 $\mathbf{0}$. 我们记随机向量 $\boldsymbol{\xi}$ 的协方差矩阵为 $\mathrm{Var}(\boldsymbol{\xi}) = \boldsymbol{\Sigma}$, 对于估计量 $\hat{\boldsymbol{\theta}}$ 依分布收敛意味着

$$\mathrm{Var}\left[\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right] = \boxed{n\mathrm{Var}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{\Sigma}}, \quad n \to \infty.$$

考虑用插入估计量 $\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)$ 来估计 $\boldsymbol{h}(\boldsymbol{\theta})$, 利用式 (4.19), 考虑对插入估计量 $\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)$ 的标准化变化 $\sqrt{n}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{h}(\boldsymbol{\theta})\right)$, 其协方差矩阵有

$$\begin{aligned}
&\mathrm{Var}\left[\sqrt{n}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{h}(\boldsymbol{\theta})\right)\right] = \mathrm{Var}\left[\sqrt{n}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}} + R(\boldsymbol{\zeta})\right)\right] \\
=&n\mathrm{Var}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}} + R(\boldsymbol{\zeta})\right) = n\mathrm{Var}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}}\right) = n \cdot \boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\,\mathrm{Var}\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\right)^{\mathrm{T}},
\end{aligned}$$

这里的计算中我们将余项 $R(\boldsymbol{\zeta})$ 视为非随机的列向量. 当 $n \to \infty$ 时, 利用 $\hat{\boldsymbol{\theta}}$ 渐进分布的即有

$$\mathrm{Var}\left[\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right] = n \cdot \boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\frac{\boldsymbol{\Sigma}}{n}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\right)^{\mathrm{T}} = \boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\,\boldsymbol{\Sigma}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\right)^{\mathrm{T}} = \mathrm{Var}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\,\boldsymbol{\xi}\right)$$

成立. 也就是说, 插入估计量 $\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)$ 的渐进分布是与 $\boldsymbol{\xi}$ 有关的线性变换. 由此, 我们介绍 Delta 方法.

定理 4.4.8 (Delta 方法,Delta method)**.** *设映射 $\boldsymbol{h}: \mathbb{R}^{K\times 1} \to \mathbb{R}^{m\times 1}$ 在参数 $\boldsymbol{\theta} \in \mathbb{R}^{K\times 1}$ 的邻域内连续可微. 若参数 $\boldsymbol{\theta}$ 的估计量为 $\hat{\boldsymbol{\theta}}$, 且满足有渐进分布*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \xrightarrow{d} \boldsymbol{\xi},$$

则嵌入估计量 $\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)$ 具有渐进分布

$$\sqrt{n}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{h}(\boldsymbol{\theta})\right) \xrightarrow{d} \boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\,\boldsymbol{\xi},$$

其中 $m \times K$ 矩阵 $\boldsymbol{J}\boldsymbol{h}(\cdot)$ 为形如式 (4.18) 的映射 $\boldsymbol{h}$ 的 Jacobi 矩阵.

我们将定理 (4.4.5)(4.4.6)(4.4.8) 综合起来为如下的应用.

定理 4.4.9 (渐进分布的应用). 结合定理 *(4.4.5)(4.4.6)*，我们将 *Delat* 方法从一元推广至二元，再推广至多元:

(a) 对于独立同分布的随机变量 $X_i, i=1,\cdots,n$, 共有总体 X 满足 $\mathbb{E}X^2<\infty$, 记 $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i, \mu=\mathbb{E}X, \sigma^2=\mathrm{Var}(X)$, 则

1. 当 $n\to\infty$ 时, $\sqrt{n}\left(\bar{X}_n-\mu\right)\xrightarrow{d}\mathrm{N}\left(0,\sigma^2\right)$;
2. 当 $n\to\infty$ 时, $\sqrt{n}\left(\bar{X}_n^2-\mu^2\right)\xrightarrow{d}\mathrm{N}\left(0,(2\mu)^2\sigma^2\right)$, 其中 $h'(\mu)=2\mu$;
3. 当 $n\to\infty$ 时, $\sqrt{n}\left(e^{\bar{X}_n}-e^{\mu}\right)\xrightarrow{d}\mathrm{N}\left(0,e^{2\mu}\sigma^2\right)$, 其中 $h'(\mu)=e^{\mu}$;
4. 当 $n\to\infty$ 时, $\sqrt{n}\left(\ln\bar{X}_n-\ln\mu\right)\xrightarrow{d}\mathrm{N}\left(0,\frac{\sigma^2}{\mu^2}\right)$, 其中 $h'(\mu)=\mu^{-1}$.

(b) 对于独立同分布的随机变量 $(X_i,Y_i)^{\mathrm{T}}, i=1,\cdots,n$, 共有总体 $(X,Y)^{\mathrm{T}}$ 满足 $\mathbb{E}X^2<\infty$ 和 $\mathbb{E}Y^2<\infty$, 记 $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i, \bar{Y}_n=\frac{1}{n}\sum_{i=1}^n Y_i, \mu_X=\mathbb{E}X, \mu_Y=\mathbb{E}Y, (X,Y)^{\mathrm{T}}$ 的协方差矩阵为 $\boldsymbol{\Sigma}$, 则

1. 当 $n\to\infty$ 时, $\sqrt{n}\left(\bar{X}_n-\mu_X,\bar{Y}_n-\mu_Y\right)^{\mathrm{T}}\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Sigma}\right)$;
2. 当 $n\to\infty$ 时, $\sqrt{n}\left(\bar{X}_n\bar{Y}_n-\mu_X\mu_Y\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{l}\boldsymbol{\Sigma}\boldsymbol{l}^{\mathrm{T}}\right)$, 其中 $\boldsymbol{h}(\mu_X,\mu_Y)=\mu_X\mu_Y, \boldsymbol{l}=\boldsymbol{J}\boldsymbol{h}(\mu_X,\mu_Y)=(\mu_Y,\mu_X)_{1\times 2}$;
3. 当 $n\to\infty$ 时且 $\mu_Y\neq 0$, $\sqrt{n}\left(\frac{\bar{X}_n}{\bar{Y}_n}-\frac{\mu_X}{\mu_Y}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{l}\boldsymbol{\Sigma}\boldsymbol{l}^{\mathrm{T}}\right)$, 其中 $\boldsymbol{h}(\mu_X,\mu_Y)=\frac{\mu_X}{\mu_Y}, \boldsymbol{l}=\boldsymbol{J}\boldsymbol{h}(\mu_X,\mu_Y)=\left(\frac{1}{\mu_Y},-\frac{\mu_X}{\mu_Y^2}\right)_{1\times 2}$;
4. 当 $n\to\infty$ 时, $\sqrt{n}\left(\bar{X}_n^2+\bar{X}_n\bar{Y}_n-(\mu_X^2+\mu_X\mu_Y)\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{l}\boldsymbol{\Sigma}\boldsymbol{l}^{\mathrm{T}}\right)$, 其中 $\boldsymbol{h}(\mu_X,\mu_Y)=\mu_X^2+\mu_X\mu_Y, \boldsymbol{l}=\boldsymbol{J}\boldsymbol{h}(\mu_X,\mu_Y)=(2\mu_X+\mu_Y,\mu_X)_{1\times 2}$.

(c) 设 $\mathbf{X}_i, i=1,\cdots,n$ 是独立同分布的且有共有的总体 $\mathbf{X}\in\mathbb{R}^{K\times 1}$, 且总体 $\mathbf{X}$ 满足有 $\mathbb{E}\|\mathbf{X}\|^2<\infty$, 则有定理 *(4.4.8)* 有

1. 如果 $\boldsymbol{\xi}\sim\mathrm{N}(0,\boldsymbol{V})$, 当 $n\to\infty$ 时,
$$\sqrt{n}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{h}(\boldsymbol{\theta})\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\boldsymbol{V}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\right)^{\mathrm{T}}\right).$$
特别地, 当 $\boldsymbol{\theta}$ 和 $\boldsymbol{h}$ 均为标量时, $\sqrt{n}\left(h\left(\hat{\theta}\right)-h(\theta)\right)\xrightarrow{d}\mathrm{N}\left(0,\left(h\prime(\theta)\right)^2V\right)$;
2. 如果参数 $\boldsymbol{\theta}=\mathbb{E}(g(\mathbf{X}))$, 定义参数 $\boldsymbol{\beta}=\boldsymbol{h}(\boldsymbol{\theta})$, 若 $\mathbb{E}\|g(\mathbf{X})\|<\infty$, 则参数 $\boldsymbol{\theta}$ 的相合估计为 $\hat{\boldsymbol{\theta}}=\frac{1}{n}\sum_{i=1}^n g(\mathbf{X}_i)$, 参数 $\boldsymbol{\beta}=\boldsymbol{h}(\boldsymbol{\theta})$ 的插入估计量为 $\hat{\boldsymbol{\beta}}=\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)$, 当 $n\to\infty$ 时,
$$\boldsymbol{\beta}=\boldsymbol{h}(\boldsymbol{\theta})\xrightarrow{P}\hat{\boldsymbol{\beta}}=\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right),\quad\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\beta}\right),$$
其中协方差矩阵 $\boldsymbol{V}_{\beta}=\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\boldsymbol{V}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\theta})\right)^{\mathrm{T}}$, 而 $\boldsymbol{V}$ 是参数 $\boldsymbol{\theta}=\mathbb{E}(g(\mathbf{X}))$ 的协方差矩阵, 其满足有 $\boldsymbol{V}=\mathbb{E}\left[(g(\mathbf{X})-\boldsymbol{\theta})(g(\mathbf{X})-\boldsymbol{\theta})^{\mathrm{T}}\right]$.

对于定理 (4.4.9) 的 (c), 知道渐进分布后我们自然会研究协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\beta}}$ 和 $\boldsymbol{V}$ 的估计量. 显然, 依据弱大数定律 (4.2.5) 知当 $n \to \infty$ 时插入估计量

$$\hat{\boldsymbol{V}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(g\left(\mathbf{X}_i\right) - \hat{\boldsymbol{\theta}} \right) \left(g\left(\mathbf{X}_i\right) - \hat{\boldsymbol{\theta}} \right)^{\mathrm{T}} \xrightarrow{P} \boldsymbol{V}, \tag{4.20}$$

进而依据连续映射定理 (4.2.7) 得到嵌入估计量

$$\boldsymbol{Jh}\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{P} \boldsymbol{Jh}\left(\boldsymbol{\theta}\right), \quad \hat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \boldsymbol{Jh}\left(\hat{\boldsymbol{\theta}}\right) \hat{\boldsymbol{V}} \left(\boldsymbol{Jh}\left(\hat{\boldsymbol{\theta}}\right) \right)^{\mathrm{T}} \xrightarrow{P} \boldsymbol{V}_{\boldsymbol{\beta}}, \quad n \to \infty. \tag{4.21}$$

4.5 随机过程的渐进性质

第五章　OLS 估计量的渐进性质

本节我们介绍 OLS 估计量在大样本下的渐进性质: 我们将讨论简单随机样本与平稳随机序列不同两种假设下的结果. 同时, 我们将从大样本的视角去考察线性投影模型与线性回归模型的差异. 利用渐进分布, 我们可以得到不依赖于正态假设的 OLS 估计量的渐进区间估计.

5.1　简单随机样本下 OLS 估计量的渐进性质

本节介绍简单随机样本下 OLS 估计量的渐进性质, 无论是线性回归模型 (3.5.1) 还是线性投影模型 (3.3.1), 本节都遵循有简单随机样本的设定和假设 (2.2.1).

假设 5.1.1. 对于随机变量 Y 和随机向量 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$, 假设有

(a) 样本 $\left(Y_i,\mathbf{X}_i^{\mathrm{T}}\right),i=1\cdots,n$ 是独立同分布的;

(b) 总体 $\mathbb{E}Y^2<\infty$;

(c) 总体 $\mathbb{E}\left\|\mathbf{X}\right\|^2<\infty$;

(d) 总体 $\boldsymbol{Q}_{\mathbf{XX}}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)$ 是正定矩阵.

对于线性投影模型的总体参数 $\boldsymbol{\beta}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbb{E}\left(\mathbf{X}Y\right)$, 我们有 OLS 估计量

$$\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_iY_i\right).$$

我们现在来研究 $\hat{\boldsymbol{\beta}}$ 的相合性. 我们首先将 OLS 写为样本均值的组合, 即

$$\hat{\boldsymbol{\beta}}=\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\right),\tag{5.1}$$

由于函数 $\boldsymbol{A}^{-1}\boldsymbol{B}$ 是连续的, 若 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 都具有概率极限, 则由上一章的连续映射定理不难推知 $\boldsymbol{A}^{-1}\boldsymbol{B}$ 也是存在概率极限的. 注意到总体参数 $\boldsymbol{\beta}$ 由 $\boldsymbol{Q}_{\mathbf{XX}}=\mathrm{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)$ 和

$\boldsymbol{Q}_{\mathbf{X}Y}=\mathrm{E}(\mathbf{X}Y)$ 组成, 显然依据弱大数定律 (4.2.5) 知 OLS 估计量式 (5.1) 后一部分有

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\xrightarrow{P}\boldsymbol{Q}_{\mathbf{X}Y}=\mathrm{E}(\mathbf{X}Y),\quad n\to\infty, \tag{5.2}$$

但前一部分不能直接使用大数定律, 因为其为矩阵. 对于 K 个解释变量的总体列向量 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$ 而言,

$$\begin{aligned}\boldsymbol{Q}_{\mathbf{XX}}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)=&\mathbb{E}\left[\begin{pmatrix}X_1\\\vdots\\X_K\end{pmatrix}(X_1,\cdots,X_K)\right]=\mathbb{E}\begin{pmatrix}X_1^2&X_1X_2&\cdots&X_1X_K\\X_2X_1&X_2^2&\cdots&X_2X_K\\\vdots&\vdots&&\vdots\\X_KX_1&X_KX_2&\cdots&X_K^2\end{pmatrix}\\=&\begin{pmatrix}\mathbb{E}X_1^2&\mathbb{E}(X_1X_2)&\cdots&\mathbb{E}(X_1X_K)\\\mathbb{E}(X_2X_1)&\mathbb{E}X_2^2&\cdots&\mathbb{E}(X_2X_K)\\\vdots&\vdots&&\vdots\\\mathbb{E}(X_KX_1)&\mathbb{E}(X_KX_2)&\cdots&\mathbb{E}X_K^2\end{pmatrix},\end{aligned}$$

则设计矩阵 $\boldsymbol{X}^{\mathrm{T}}=(\mathbf{X}_1,\cdots,\mathbf{X}_n)$ 对应的有

$$\begin{aligned}\boldsymbol{G}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}=&\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}=\sum_{i=1}^{n}\begin{pmatrix}X_{i1}^2&X_{i1}X_{i2}&\cdots&X_{i1}X_{iK}\\X_{i2}X_{i1}&X_{i2}^2&\cdots&X_{i2}X_{iK}\\\vdots&\vdots&&\vdots\\X_{iK}X_{i1}&X_{iK}X_{i2}&\cdots&X_{iK}^2\end{pmatrix}\\=&\begin{pmatrix}\sum\limits_{i=1}^{n}X_{i1}^2&\sum\limits_{i=1}^{n}X_{i1}X_{i2}&\cdots&\sum\limits_{i=1}^{n}X_{i1}X_{iK}\\\sum\limits_{i=1}^{n}X_{i2}X_{i1}&\sum\limits_{i=1}^{n}X_{i2}^2&\cdots&\sum\limits_{i=1}^{n}X_{i2}X_{iK}\\\vdots&\vdots&&\vdots\\\sum\limits_{i=1}^{n}X_{iK}X_{i1}&\sum\limits_{i=1}^{n}X_{iK}X_{i2}&\cdots&\sum\limits_{i=1}^{n}X_{iK}^2\end{pmatrix},\end{aligned}$$

则矩阵 $\boldsymbol{G}$ 的对角元素 $\boldsymbol{G}_{ii}$ 与非对角元素 $\boldsymbol{G}_{ij}$ 依弱大数定律 (4.2.5) 有

$$\frac{\boldsymbol{G}_{jj}}{n}=\frac{1}{n}\sum_{i=1}^{n}X_{ij}^2\xrightarrow{P}\mathbb{E}X_j^2,\qquad\frac{\boldsymbol{G}_{jt}}{n}=\frac{1}{n}\sum_{i=1}^{n}X_{ij}X_{it}\xrightarrow{P}\mathbb{E}(X_jX_t),\quad n\to\infty.$$

对于一个随机矩阵而言, 如果其所以元素都是依概率收敛, 那么我们便认为该矩阵

是依概率收敛的, 这也即是说当 $n \to \infty$ 时,

$$\begin{aligned}\frac{\boldsymbol{G}}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} = \frac{1}{n}\begin{pmatrix} \sum\limits_{i=1}^{n}X_{i1}^2 & \sum\limits_{i=1}^{n}X_{i1}X_{i2} & \cdots & \sum\limits_{i=1}^{n}X_{i1}X_{iK} \\ \sum\limits_{i=1}^{n}X_{i2}X_{i1} & \sum\limits_{i=1}^{n}X_{i2}^2 & \cdots & \sum\limits_{i=1}^{n}X_{i2}X_{iK} \\ \vdots & \vdots & & \vdots \\ \sum\limits_{i=1}^{n}X_{iK}X_{i1} & \sum\limits_{i=1}^{n}X_{iK}X_{i2} & \cdots & \sum\limits_{i=1}^{n}X_{iK}^2 \end{pmatrix} \\ \xrightarrow{P} \boldsymbol{Q}_{\mathbf{XX}} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right) = \begin{pmatrix} \mathbb{E}X_1^2 & \mathbb{E}\left(X_1X_2\right) & \cdots & \mathbb{E}\left(X_1X_K\right) \\ \mathbb{E}\left(X_2X_1\right) & \mathbb{E}X_2^2 & \cdots & \mathbb{E}\left(X_2X_K\right) \\ \vdots & \vdots & & \vdots \\ \mathbb{E}\left(X_KX_1\right) & \mathbb{E}\left(X_KX_2\right) & \cdots & \mathbb{E}X_K^2 \end{pmatrix},\end{aligned} \tag{5.3}$$

在由定理 (4.2.7) 得

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1} \xrightarrow{P} \boldsymbol{Q}_{\mathbf{XX}} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right), \quad n \to \infty. \tag{5.4}$$

这样结合式 (5.4) 和式 (5.2), 依据定理定理 (4.2.7) 得

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\right) \xrightarrow{P} \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathrm{E}\left(\mathbf{X}Y\right) = \boldsymbol{\beta}, \quad n \to \infty, \tag{5.5}$$

即 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 是线性投影模型 (3.5.1) 总体参数 $\boldsymbol{\beta}$ 的 (弱) 相合估计.

我们还可以通过抽样误差

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = \left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\varepsilon_i\right)$$

来证明 OLS 估计量的无偏性. 记

$$\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\varepsilon_i = \begin{pmatrix} \frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{X}_{i1}\varepsilon_i \\ \vdots \\ \frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{X}_{iK}\varepsilon_i \end{pmatrix}, \quad \boldsymbol{Q}_{\mathbf{X}\varepsilon} = \mathbb{E}\left(\mathbf{X}\varepsilon\right) = \begin{pmatrix} \mathbb{E}X_1\varepsilon \\ \vdots \\ \mathbb{E}X_K\varepsilon \end{pmatrix} = \mathbf{0},$$

则依据定理 (4.2.5) 可得当 $n \to \infty$ 时 $\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon} \xrightarrow{P} \boldsymbol{Q}_{\mathbf{X}\varepsilon} = \mathbb{E}\left(\mathbf{X}\varepsilon\right) = \mathbf{0}$, 于是依据连续映射定理 (4.2.7) 得抽样误差

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\varepsilon_i\right) = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon} \xrightarrow{P} \boldsymbol{Q}_{\mathbf{XX}}^{-1}\mathbf{0} = \mathbf{0}, \quad n \to \infty,$$

故亦可以推知 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 是总体参数 $\boldsymbol{\beta}$ 的相合估计.

下面我们通过 Monte Carlo 模拟来说明定理 OLS 估计量的相合性. 考虑数据生成过程 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, 其中 ε_i 独立同分布且具有共有的 $\mathrm{N}(0, 50^2)$. 我们使用 Monte Carlo 模拟生成一万个样本个体, 并逐个计算 OLS 估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$, 绘制了如图 (5.1) 和图 (5.2) 所示的半对数坐标的 OLS 估计量的折线图.

图 (5.1) 和图 (5.2) 的横轴是 10 为底数的对数, 这样对于 $n = 10000$ 个回归系数的估计值, 图 (5.1) 和图 (5.2) 可以较为直观地展示不同数量级下 *OLS* 估计量与真实值的差异. 首先, 对于 10^0 的数量级, 不能发现在最初样本容量很小时偏差很大. 这是因为,$n = 1$ 时起始点的 $\hat{\beta}_1 = 0$, 这是因为只有一个样本点的一元 OLS 即等价于直线 $Y = \bar{Y}$, 当 $n = 2$ 时则是一元 OLS 便是平面中两点确定的一条直线, 信息有限 OLS 估计量不能剔除回归误差 ε_i 的影响——尽管我们这次模拟中 β_0 的估计非常准确. 随着 n 的增加, 在 10^1 的数量级的样本容量内 β_1 的估计值没有最初那样的大的偏离, 但依旧与真实值差距不小. 对于 10^1 的数量级, 图 (5.2) 中可以直观感受 n 的增加缩小了 $\hat{\beta}_1$ 与 β_1 差距, 尤其是较 10^0 的数量级时. 当样本容量 n 破百后, 在 10^2 的数量级下 $\hat{\beta}_1$ 的偏差经历了一个小幅的增长过程, 则也说明了依概率收敛、几乎处处收敛都不能保证有充分大的 n 使得估计量与真实值有任意小的偏差. 当 n 的数量级达到 10^3 的数量级, 不难发现 $\hat{\beta}_1$ 与真实值 $\beta_1 = 25$ 的偏差在图中很难体现, 这样便是说 10^3 的数量级下 OLS 估计量要较更小的数量级更为精确.

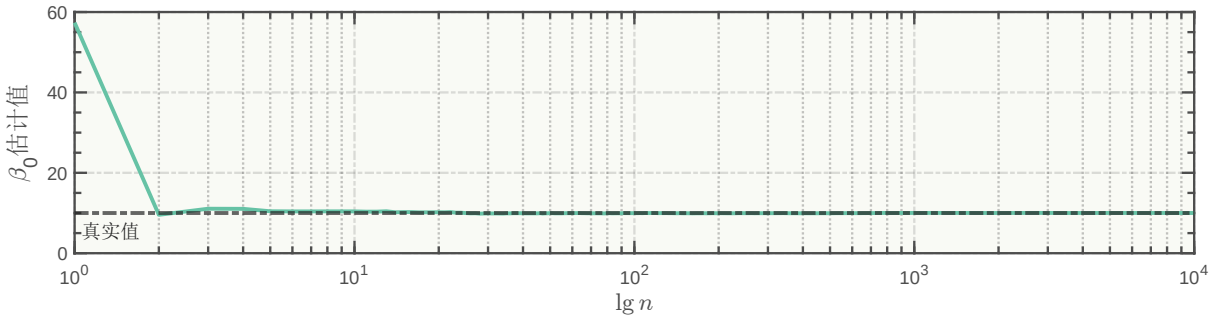

图 5.1: 样本容量增大时的 $\hat{\beta}_0$

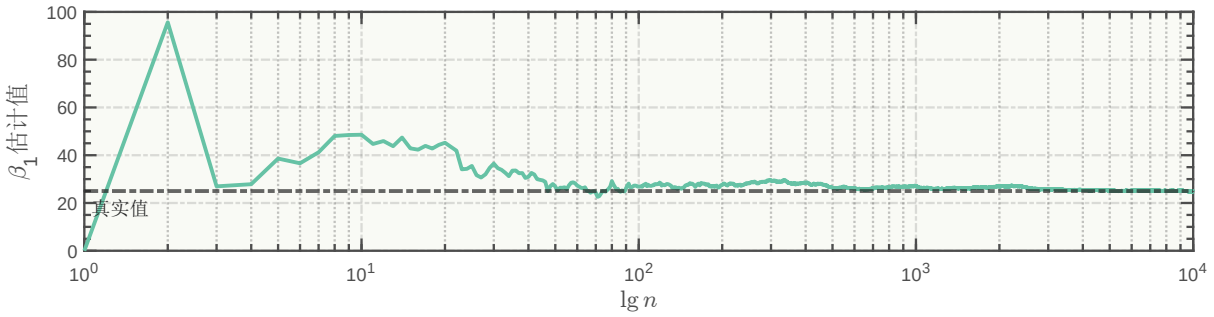

图 5.2: 样本容量增大时的 $\hat{\beta}_1$

现在我们通过 OLS 相合性来进一步说明线性投影模型 (3.3.1) 与线性回归模型 (3.5.1) 的区别. 我们在最初已经证明了, 在 MSE 最小的视角下条件期望函数函数是解释变量 $\mathbf{X}$ 对被解释变量 Y 的最优预测量, 尽管通常我们不知道 $\mathbb{E}(Y|\mathbf{X})$ 具体形式. 为了使得我们的模型具有可行性 (feasibility), 我们假设了条件期望函数 $\mathbb{E}(Y|\mathbf{X})$ 具有形如定义 (2.2.1) 的关于参数 $\boldsymbol{\beta}$ 线性表达式

$$m(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X}) = \beta_1 X_1 + \cdots + \beta_K X_K,$$

这时, 我们便可以得到线性回归模型——尽管线性 CEF 只是一个假设. 于是我们另辟蹊径, 无论 CEF 是何种形式, 我们直接考虑使用线性函数

$$\mathscr{P}(Y|\mathbf{X}) = \beta_1 X_1 + \cdots + \beta_n X_K = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$$

去解释 Y, 且 $\mathscr{P}(Y|\mathbf{X})$ 同样遵循 MSE 最小的原则, 于是我们便得到了最优线性预测量以及线性投影模型.

在 MSE 最小的原则下最优预测量必然是条件期望 $\mathbb{E}(Y|\mathbf{X})$, 对比线性投影模型 (3.3.1) 与线性回归模型 (3.5.1) 的区别, 前者即是假设了 $\mathbb{E}(Y|\mathbf{X})$ 是关于参数线性的, 后者则是考虑使得 MES 最小的最优线性预测量, 而不是从最优预测量入手. 这样而言, 线性投影模型 (3.3.1) 是更为广泛的设定, 当且仅当条件期望 $\mathbb{E}(Y|\mathbf{X})$ 是线性 CEF 时线性投影模型即等价于线性回归模型.

我们现在通过一个 Monte Carlo 模拟再来辨析线性投影模型 (3.3.1) 与线性回归模型 (3.5.1). 我们设一元回归模型和一元线性投影模型分别为

$$Y = \mathbb{E}(Y|\mathbf{X}) + e = X^2 + e, \tag{5.6}$$

$$Y = \mathscr{P}(Y|X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon, \tag{5.7}$$

其中式 (5.6) 也是真实的数据生成过程. 这样, 在给定解释变量 X 下被解释变量 Y 下的条件期望与最优线性预测量是不相同的. 不妨假设解释变量 X 服从正态分布 $\mathrm{N}(\mu, \sigma^2)$, 依据第4.3节的表4.1, 不难计算有

$$\begin{aligned}
\mathbb{E}Y &= \mathbb{E}\left(X^2 + e\right) = \mathbb{E}X^2 = \mu^2 + \sigma^2, \\
\mathbb{E}(XY) &= \mathbb{E}\left[X\left(X^2 + e\right)\right] = \mathbb{E}X^3 = \mu^2 + 3\mu\sigma^2,
\end{aligned}$$

以及

$$\begin{aligned}
\mathrm{Cov}(X, Y) =& \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y = \left(\mu^2 + 3\mu\sigma^2\right) - \mu\left(\mu^2 + \sigma^2\right) \\
=& \mu^2 + 2\mu\sigma^2 - \mu^3 = \mu\left(\mu + 2\sigma^2 - \mu^2\right).
\end{aligned}$$

对于一元线性投影模型, 式 (5.7) 的总体参数即为

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \frac{\mu\left(\mu + 2\sigma^2 - \mu^2\right)}{\sigma^2}$$

和

$$\begin{aligned}\beta_0 =& \mathbb{E}Y - \beta_1 \mathbb{E}X = \mu^2 + \sigma^2 - \frac{\mu\left(\mu + 2\sigma^2 - \mu^2\right)}{\sigma^2} \cdot \mu \\ =& \frac{\left(\mu^2 + \sigma^2\right)\sigma^2 - \mu^2\left(\mu + 2\sigma^2 - \mu^2\right)}{\sigma^2} \\ =& \frac{\mu^2\sigma^2 + \sigma^4 - \mu^3 - 2\mu^2\sigma^2 + \mu^4}{\sigma^2} = \frac{\sigma^4 + \mu^4 - \mu^2\sigma^2 - \mu^3}{\sigma^2}.\end{aligned}$$

特别地, 当 $X \sim \mathrm{N}\left(1, \sigma^2\right)$, 则线性投影模型的参数可以化简为

$$\beta_1 = 2, \quad \beta_0 = \frac{\sigma^4 - \sigma^2}{\sigma^2} = \sigma^2 - 1,$$

我们现对 $\sigma = 2$ 的情形进行 Monte Carlo 模拟. 与说明 OLS 估计量相合性的模拟一样, 我们随机生成 $n = 10^4$ 个样本个体, 逐个增加样本来计算 OLS 估计量, 结果如图 (5.3) 所示. 当样本的数量级较小时,OLS 估计的 $\hat{\beta}_1$ 与 $\hat{\beta}_0$ 有真实值有较大的偏离. 当样本容量的数量级达到 10^3 后,OLS 估计量与真实值的差距可见是越来越小, 最终是依概率收敛于了线性投影模型的总体参数 β_0 和 β_1.

图 (5.4) 绘制了 $n = 10^4$ 个样本点、条件期望函数 $\mathbb{E}\left(Y|X\right) = X^2$、最优线性预测量 $\mathscr{P}\left(Y|X\right) = \beta_0 + \beta_1 X$ 和 OLS 估计量 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$. 当样本的数据生成过程是二次函数形式的 $Y = m\left(X\right) + e = X^2 + e$ 时, 显然使用线性投影模型 $Y = \mathscr{P}\left(Y|X\right) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$ 在模型形式上便存在了设定错误的问题, 对此 OLS 估计量得到的样本回归直线, 随着样本容量 n 的提升最终会依概率收敛于最优线性预测量 $\mathscr{P}\left(Y|X\right) = \beta_0 + \beta_1 X$.

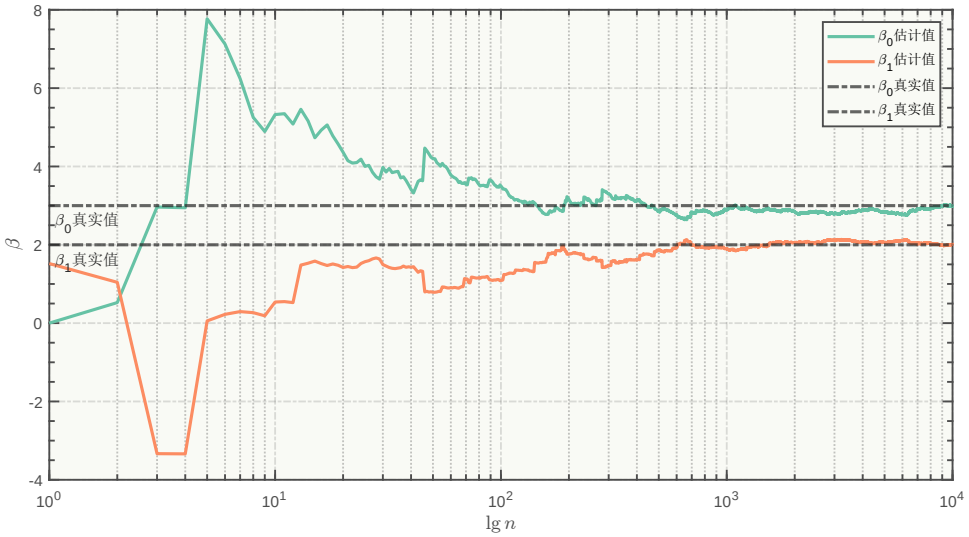

图 5.3: OLS 估计量依概率收敛于最优线性预测 $\mathscr{P}\left(Y|X\right)$

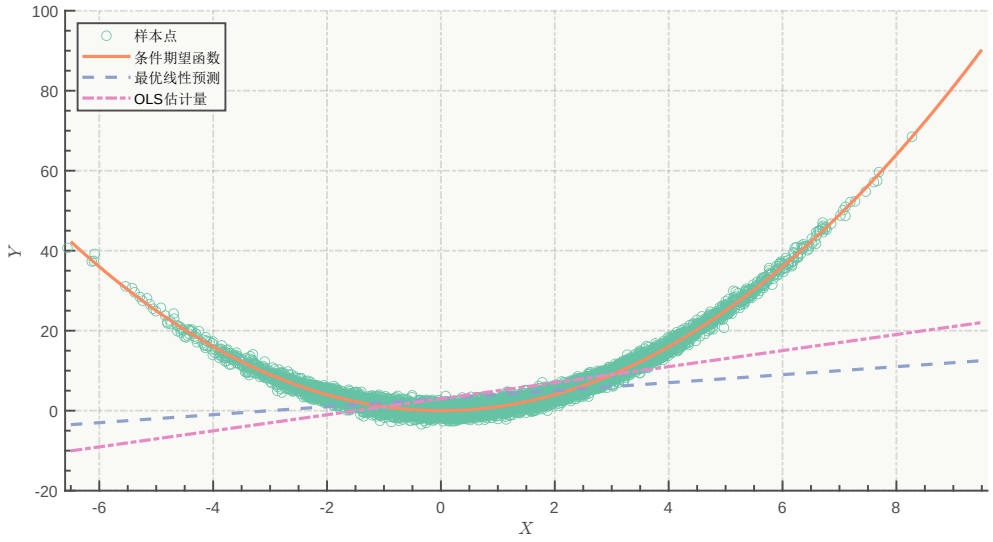

图 5.4: 非线性 CEF 下的 OLS 估计量

我们将 OLS 估计量的相合性总结如下.

定理 5.1.1 (OLS 估计量的相合性). *对于 OLS 估计量, 在假设 (5.1.1) 下, 则当 $n \to \infty$ 有*

$$\hat{\boldsymbol{Q}}_{\mathbf{XX}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} \xrightarrow{P} \boldsymbol{Q}_{\mathbf{XX}} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right),$$

$$\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1} \xrightarrow{P} \boldsymbol{Q}_{\mathbf{XX}}^{-1} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1},$$

$$\hat{\boldsymbol{Q}}_{\mathbf{X}Y} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i \xrightarrow{P} \boldsymbol{Q}_{\mathbf{X}Y} = \mathrm{E}\left(\mathbf{X}Y\right),$$

$$\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\varepsilon_i \xrightarrow{P} \boldsymbol{Q}_{\mathbf{XX}} = \mathbb{E}\left(\mathbf{X}\varepsilon\right) = \mathbf{0},$$

以及

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\right) = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\hat{\boldsymbol{Q}}_{\mathbf{X}Y}$$

$$\xrightarrow{P}\boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathrm{E}\left(\mathbf{X}Y\right) = \boldsymbol{Q}_{\mathbf{XX}}^{-1}\boldsymbol{Q}_{\mathbf{X}Y}.$$

如果仅仅得知 OLS 估计量是相合估计, 那么大样本理论并不比有限样本下更有优势. 下面我们研究 OLS 估计量具有的**渐近正态性**. 对于抽样误差 $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ 则有

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i\varepsilon_i\right) = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\cdot\sqrt{n}\left(\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon} - \mathbf{0}\right)$$

我们首先来研究 $\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon}$ 的渐进分布.$\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon}$ 是 $\boldsymbol{X}\varepsilon$ 估计量, 且其即是 $\boldsymbol{X}\varepsilon$ 的样本均值, 不难

得知

$$\mathbb{E}\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon}=\mathbb{E}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i1}\varepsilon_i\\ \vdots\\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{iK}\varepsilon_i\end{pmatrix}=\begin{pmatrix}\mathbb{E}X_1\varepsilon\\ \vdots\\ \mathbb{E}X_K\varepsilon\end{pmatrix}=\mathbb{E}\left(\mathbf{X}\varepsilon\right)=\mathbf{0}.$$

我们设 $\boldsymbol{X}\varepsilon$ 的协方差矩阵为 $\boldsymbol{\Omega}$, 则

$$\boldsymbol{\Omega}=\operatorname{Var}\left(\mathbf{X}\varepsilon\right)=\mathbb{E}\left[\left(\mathbf{X}\varepsilon-\mathbf{0}\right)\left(\mathbf{X}\varepsilon-\mathbf{0}\right)^{\mathrm{T}}\right]=\mathbb{E}\left[\left(\mathbf{X}\varepsilon\right)\left(\varepsilon\mathbf{X}^{\mathrm{T}}\right)\right]=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right),$$

可以证明[1], 对于独立同分布的样本, 总体解释变量 $\mathbf{X}$ 与被解释变量 Y 若满足有 $\mathbb{E}Y^4<\infty,\mathbb{E}\left\|\mathbf{X}\right\|^4<\infty$, 则有 $\boxed{\left\|\boldsymbol{\Omega}\right\|\leqslant\mathbb{E}\left\|\mathbf{X}^{\mathrm{T}}\mathbf{X}\varepsilon^2\right\|<\infty}$, 回顾上一章介绍的中心极限定理 (4.4.5), 我们便得到

$$\sqrt{n}\left(\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon}-\mathbf{0}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right),\quad n\to\infty.\tag{5.8}$$

再依据连续映射定理 (4.4.6), 则得到 OLS 估计量满足有渐进分布

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\cdot\sqrt{n}\left(\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon}-\mathbf{0}\right)\xrightarrow{d}\boldsymbol{Q}_{\mathbf{XX}}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right)=\mathrm{N}\left(\mathbf{0},\boldsymbol{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\mathbf{XX}}^{-1}\right),\tag{5.9}$$

这里的 $\boldsymbol{Q}_{\mathbf{XX}}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)$ 是实对称方阵. 我们将 OLS 估计量的渐进分布总结如下.

定理 5.1.2 (OLS 估计量的渐近正态性)**.** *在假设 (5.1.1) 以及 $\mathbb{E}Y^4<\infty,\mathbb{E}\left\|\mathbf{X}\right\|^4<\infty$ 下, 当 $n\to\infty$ 时,OLS 估计量具有***渐近正态性*** (asymptotic normality):*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right),$$

其中协方差矩阵

$$\boldsymbol{V}_{\boldsymbol{\beta}}=\boldsymbol{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\mathbf{XX}}^{-1},\tag{5.10}$$

*而 $\boldsymbol{\Omega}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right),\boldsymbol{Q}_{\mathbf{XX}}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)$. 我们将 $\boldsymbol{V}_{\boldsymbol{\beta}}$ 称为 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的***渐进协方差矩阵*** (asymptotic covariance matrix), 我们将 $\boldsymbol{V}_{\boldsymbol{\beta}}$ 又记为 $\operatorname{Avar}\left(\hat{\boldsymbol{\beta}}\right)$.*

我们此前计算了 OLS 估计量的条件协方差矩阵, 并总结为了定理 (3.5.1). 我们对比两者,

$$\text{条件协方差矩阵}:\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},$$

$$\text{渐进协方差矩阵}:\boldsymbol{V}_{\boldsymbol{\beta}}=\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right)\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1},$$

则不难发现, 前者是由样本所决定的随机矩阵 (但含有未知的总体参数 $\boldsymbol{D}$), 是 OLS 估计量这一统计量的数值特征, 而后者是 OLS 估计量渐进分布的总体的数值特征, 是一个

[1] 见 Bruce(2021),p.166

参数矩阵. 由于

$$n\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\right)\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1},$$

可以证明, 当 $n\to\infty$ 时,$n\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\xrightarrow{d}\boldsymbol{V}_{\boldsymbol{\beta}}$. 特别地, 类似于同方差式 (3.7), 假设当总体满足有 $\operatorname{Cov}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}},\varepsilon^2\right)=\mathbf{0}$ 时, 则

$$\boldsymbol{\Omega}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right)=\operatorname{Cov}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}},\varepsilon^2\right)+\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\mathbb{E}\varepsilon^2=\boldsymbol{Q}_{\mathbf{XX}}\sigma^2,$$

此时渐进协方差矩阵有

$$\boldsymbol{V}_{\boldsymbol{\beta}}^0=\boldsymbol{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\mathbf{XX}}^{-1}=\boldsymbol{Q}_{\mathbf{XX}}^{-1}\boldsymbol{Q}_{\mathbf{XX}}\sigma^2\boldsymbol{Q}_{\mathbf{XX}}^{-1}=\boldsymbol{Q}_{\mathbf{XX}}^{-1}\sigma^2,\tag{5.11}$$

我们称 $\boldsymbol{V}_{\boldsymbol{\beta}}^0$ 为 $\hat{\boldsymbol{\beta}}$ 的**同方差渐进协方差矩阵** (homoskedastic asymptotic covariance matrix). 需要指出,$\boldsymbol{V}_{\boldsymbol{\beta}}^0$ 被视为是一个总体参数矩阵, 无论总体是否满足 $\operatorname{Cov}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}},\varepsilon^2\right)=\mathbf{0}$ 这一条件.

5.2 简单随机样本下方差估计量的渐进性质

这一节我们介绍简单随机样本下 OLS 估计量的各类方差估计量. 我们在上一节中介绍了 OLS 估计量所具有的渐进分布, 且推导了渐进分布所具有的总体参数渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\beta}}$. 我们在第三章中介绍了有限样本下 OLS 估计量对于回归方差 σ^2 和协方差矩阵 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ 的估计量, 并进一步的提出了标准误 SE 的概念. 本节将在大样本理论下发展这些内容, 并讨论其具有的渐进性质.

我们知道, 对于总体参数回归误差 $\mathbb{E}\varepsilon^2=\sigma^2$, 使用残差 $\hat{\varepsilon}_i$ 替换回归误差 ε_i 得到的矩估计

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{2}\hat{\varepsilon}_i^2,\tag{5.12}$$

是有偏的, 因而我们使用了形如式 (3.52) 的偏差校正估计量 s^2. 我们现在研究两者的渐进性质. 对于式 (5.12), 我们使用投影矩阵 $\boldsymbol{P}$ 和归零矩阵 $\boldsymbol{M}$, 由

$$\boldsymbol{P}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}},\quad \boldsymbol{M}=\boldsymbol{I}-\boldsymbol{P}$$

可以得到抽样误差 $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$ 满足 $\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=\boldsymbol{P}\boldsymbol{\varepsilon}$, 故使用矩阵符号, 式 (5.12) 可以分解为

$$\begin{aligned}\hat{\sigma}^2&=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{\varepsilon}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{\varepsilon}-\boldsymbol{P}\boldsymbol{\varepsilon}\right)\\&=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\left[\boldsymbol{\varepsilon}-\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right]=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}-\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right),\end{aligned}$$

依据大数定律 (4.2.5) 可以得到当 $n \to \infty$ 时

$$\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{2}\varepsilon_i^2 \xrightarrow{P} \mathbb{E}\varepsilon^2 = \sigma^2,$$
$$\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{2}\varepsilon_i\mathbf{X}_i^{\mathrm{T}} = \left(\frac{1}{n}\sum_{i=1}^{2}\varepsilon_i X_{i1}, \cdots, \frac{1}{n}\sum_{i=1}^{2}\varepsilon_i X_{iK}\right) \xrightarrow{P} \mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\varepsilon\right) = \mathbf{0},$$

且抽样误差依概率收敛于 $\mathbf{0}$, 故依据连续映射定理 (4.2.7) 则的方差估计量

$$\hat{\sigma}^2 = \frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} - \frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{P} \sigma^2 - \mathbf{0}_{1\times K}\cdot\mathbf{0}_{K\times 1} = \sigma^2, \quad n \to \infty,$$

即有偏的 $\hat{\sigma}^2$ 是回归方差 σ^2 的相合估计. 进一步地, 对于偏差校正估计量 s^2, 可以立即推知

$$s^2 = \frac{n}{n-K}\hat{\sigma}^2 \xrightarrow{P} \sigma^2, \quad n \to \infty$$

成立, 即 s^2 也是回归方差 σ^2 的相合估计.

定理 5.2.1 (方差估计量的相合性)**.** *对于 OLS 估计量导出的方差估计量 $\hat{\sigma}^2$ 和 s^2, 当 $n \to \infty$ 时,*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{2}\hat{\varepsilon}_i^2 \xrightarrow{P} \sigma^2, \quad s^2 = \frac{1}{n-K}\sum_{i=1}^{2}\hat{\varepsilon}_i^2 \xrightarrow{P} \sigma^2.$$

我们现在来介绍 OLS 估计量的各类协方差矩阵估计量所具有的渐进性质. 对于形如式 (3.56)、(3.57)、(3.58) 和 (3.59) 的异方差稳健的协方差矩阵估计量, 我们尝试使用矩阵符号将其转变为由 $\hat{\boldsymbol{Q}}_{\mathbf{XX}}$ 所构成的形式. 由于

$$\hat{\boldsymbol{Q}}_{\mathbf{XX}} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}, \quad \hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2,$$

则 White 协方差矩阵估计量可以写为

$$\begin{aligned}\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} =& \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\ =& \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\ =& \frac{1}{n}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1},\end{aligned}$$

令

$$\boldsymbol{V}_{\boldsymbol{\beta}}^{\mathrm{HC0}} = n\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}, \tag{5.13}$$

则协方差矩阵 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}}$ 是渐进协方差矩阵式 (5.10) 的一个估计量. 类似地, 将 HC1、HC2 和 HC3 协方差矩阵乘以 n, 同样可以得到 OLS 估计量的渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\beta}}$ 的估计量

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC1}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}} = n \cdot \frac{n}{n-K}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} = \frac{n}{n-K}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}}, \tag{5.14}$$

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC2}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\bar{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}, \quad \bar{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\bar{\varepsilon}_i^2, \tag{5.15}$$

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC3}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}} = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\tilde{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}, \quad \tilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\tilde{\varepsilon}_i^2, \tag{5.16}$$

其中 $\bar{\varepsilon}_i = (1-h_{ii})^{-1}\hat{\varepsilon}_i$ 为标准化残差,$\tilde{\varepsilon}_i = (1-h_{ii})^{-2}\hat{\varepsilon}_i$ 为 LLO 回归的预测误差 (见第3.9节). 可以证明[2], 对于协方差矩阵 $\hat{\boldsymbol{\Omega}}$、$\bar{\boldsymbol{\Omega}}$ 和 $\tilde{\boldsymbol{\Omega}}$, 当 $n \to \infty$ 时

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2 \xrightarrow{p} \boldsymbol{\Omega} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right)$$

$$\bar{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\bar{\varepsilon}_i^2 \xrightarrow{p} \boldsymbol{\Omega} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right)$$

$$\tilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\tilde{\varepsilon}_i^2 \xrightarrow{p} \boldsymbol{\Omega} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right)$$

再依据连续映射定理 (4.2.7) 不难得知 $\boldsymbol{V}_{\boldsymbol{\beta}}^{\mathrm{HC0}}$、$\boldsymbol{V}_{\boldsymbol{\beta}}^{\mathrm{HC1}}$、$\boldsymbol{V}_{\boldsymbol{\beta}}^{\mathrm{HC2}}$ 和 $\boldsymbol{V}_{\boldsymbol{\beta}}^{\mathrm{HC3}}$ 都依概率收敛于渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\beta}}$. 我们将有限样本和大样本下 OLS 估计量的协方差矩阵总结为表5.1.

在后续的内容中, 我们使用 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ 表示渐进协方差矩阵的估计量, 其所指的是 HC0 至 HC3 或同方差的协方差矩阵.

[2]对于随机矩阵, 不能直接使用大数定理. 证明可参见 Bruce(2021)p.172,174.

表 5.1: OLS 估计量的各类协方差矩阵

	有限样本 $\hat{\boldsymbol{\beta}}$	大样本 $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$	渐进性质 $n\to\infty$
总体	条件协方差矩阵 $\boldsymbol{V}_{\hat{\beta}}=$ $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$	渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\beta}}=\boldsymbol{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\mathbf{XX}}^{-1}=$ $\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^{2}\right)\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}$	$n\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\xrightarrow{P}\boldsymbol{V}_{\boldsymbol{\beta}}$
	同方差时 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{0}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^{2}$	同方差时 $\boldsymbol{V}_{\boldsymbol{\beta}}^{0}=\boldsymbol{Q}_{\mathbf{XX}}^{-1}\sigma^{2}$	$n\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{0}\xrightarrow{P}\boldsymbol{V}_{\boldsymbol{\beta}}^{0}$
估计量	$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_{i}\mathbf{X}_{i}^{\mathrm{T}}\hat{\varepsilon}_{i}^{2}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}}=n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}}\xrightarrow{P}\boldsymbol{V}_{\boldsymbol{\beta}}$
	$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}}=\frac{n}{n-K}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC1}}=n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC1}}\xrightarrow{P}\boldsymbol{V}_{\boldsymbol{\beta}}$
	$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_{i}\mathbf{X}_{i}^{\mathrm{T}}\bar{\varepsilon}_{i}^{2}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC2}}=n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC2}}\xrightarrow{P}\boldsymbol{V}_{\boldsymbol{\beta}}$
	$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_{i}\mathbf{X}_{i}^{\mathrm{T}}\tilde{\varepsilon}_{i}^{2}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC3}}=n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC3}}\xrightarrow{P}\boldsymbol{V}_{\boldsymbol{\beta}}$
	同方差时 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{0}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^{2}$	同方差时 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}=\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}s^{2}=n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{0}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}\xrightarrow{P}\boldsymbol{V}_{\boldsymbol{\beta}}^{0}$

有时我们会研究由总体参数 $\boldsymbol{\beta}$ 定义的总体参数 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})$, 其中 $\boldsymbol{h}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{q\times 1}$, 这时我们便可以使用上一章中的 Delta 方法来推导参数 $\boldsymbol{\theta}$ 的渐进分布. 我们使用插入估计量 $\hat{\boldsymbol{\theta}}=\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)$ 来估计参数 $\boldsymbol{\theta}$, 则依据定理 (4.4.9) 中 (c) 知, 估计量 $\hat{\boldsymbol{\theta}}$ 具有渐进分布

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=\sqrt{n}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{h}(\boldsymbol{\beta})\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\theta}}\right), \tag{5.17}$$

其中 $\boldsymbol{V}_{\boldsymbol{\theta}}=\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\beta})\boldsymbol{V}_{\boldsymbol{\beta}}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\beta})\right)^{\mathrm{T}}$ 是 $\hat{\boldsymbol{\theta}}$ 的的渐进协方差矩阵, 而

$$\boldsymbol{J}\boldsymbol{h}(\cdot)=\begin{pmatrix}\dfrac{\partial h_1(\cdot)}{\partial x_1} & \cdots & \dfrac{\partial h_1(\cdot)}{\partial x_K}\\ \vdots & & \vdots\\ \dfrac{\partial h_q(\cdot)}{\partial x_1} & \cdots & \dfrac{\partial h_q(\cdot)}{\partial x_K}\end{pmatrix}_{q\times K}$$

是映射 $\boldsymbol{h}$ 的 Jacobi 矩阵. 显然, 渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\theta}}$ 的一个估计量可以是

$$\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}=\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}=n\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}. \tag{5.18}$$

正如定义 (3.11.1) 一样, 我们将参数标准差的估计量称为标准误, 则参数 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_q)^{\mathrm{T}}$ 的第 j 个分量的标准误为

$$\begin{aligned}\mathrm{SE}\left(\hat{\theta}_j\right)=&\sqrt{\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}\right)_{jj}}=\sqrt{\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right)_{jj}}\\ =&\sqrt{\left(\frac{1}{n}\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right)_{jj}},\end{aligned} \tag{5.19}$$

这里我们使用的不是渐进协方差矩阵的对角线元素. 在大样本下, 我们将标准误 $\mathrm{SE}\left(\hat{\theta}_j\right)=\sqrt{\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}\right)_{jj}}$ 称为参数 θ_j 的**渐进标准误** (asymptotic standard error). 特别地, 当映射 $\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\beta}$, 即参数 $\boldsymbol{\theta}=\boldsymbol{\beta}$, 上述的渐进标准误也是适用于 OLS 估计量的.

特别地, 当回归模型满足有同方差假设时, 则回归系数 $\boldsymbol{\beta}$ 有 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}=\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}s^2$, $\boldsymbol{V}_{\boldsymbol{\beta}}^{0}=\boldsymbol{Q}_{\mathbf{XX}}^{-1}\sigma^2$, 于是式 (5.17) 和式 (5.18) 中的协方差矩阵可以化简得

$$\boldsymbol{V}_{\boldsymbol{\theta}}^{0}=\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\beta})\boldsymbol{Q}_{\mathbf{XX}}^{-1}\left(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\beta})\right)^{\mathrm{T}}\sigma^2,\quad \hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{0}=\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}s^2. \tag{5.20}$$

5.3 回归区间与预测区间

在有限样本下, 我们通过正态假设得以对 OLS 估计量进行区间估计和假设检验——尽管正态假设不能总是保障. 在大样本理论下, 由于 OLS 估计量是渐进正态的, 当 n 充分大时, 我们便可以直接将相关内容迁移与此, 讨论渐近正态下 OLS 估计量的区间估计. 本节我们将介绍回归区间与预测区间.

我们同样考虑上一节的参数 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})$ 以及其插入估计量 $\hat{\boldsymbol{\theta}}=\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)$, 当 $n\to\infty$ 时, 由于 $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\theta}}\right)$, 则参数 $\boldsymbol{\theta}$ 的分量满足有

$$\sqrt{n}\left(\hat{\theta}_j-\theta_j\right)\xrightarrow{d}\mathrm{N}\left(0,\left(\boldsymbol{V}_{\boldsymbol{\theta}}\right)_{jj}\right),$$

故对于充分大的 n, 依据连续映射定理 (4.4.6) 知 **t 统计量** (t-statistic)或 **t 比率** (t-ratio)有

$$T=\frac{\hat{\theta}_j-\theta_j}{\mathrm{SE}\left(\hat{\theta}_j\right)}=\frac{\sqrt{n}\left(\hat{\theta}_j-\theta_j\right)}{\sqrt{n}\mathrm{SE}\left(\hat{\theta}_j\right)}=\frac{\sqrt{n}\left(\hat{\theta}_j-\theta_j\right)}{\sqrt{n\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}\right)_{jj}}}=\frac{\sqrt{n}\left(\hat{\theta}_j-\theta_j\right)}{\sqrt{\left(\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}\right)_{jj}}}\xrightarrow{d}\mathrm{N}(0,1). \tag{5.21}$$

由于式 (5.21) 的渐进分布为标准正态分布, 其不依赖于未知的参数, 故我们称统计量是**渐进枢轴** (asymptotically pivotal)的统计量. 我们现在直接迁移定理 (3.12.1) 的内容, 当 n 充分大时, 依据 t 统计量的渐近正态性可得**渐进置信水平** (asymptotic confidence level)为 $1-\alpha$ 的置信区间为

$$\hat{I}=\left[\hat{\theta}_j-c\cdot\mathrm{SE}\left(\hat{\theta}_j\right),\hat{\theta}_j+c\cdot\mathrm{SE}\left(\hat{\theta}_j\right)\right], \tag{5.22}$$

其中常数 c 满足有当 $\boxed{n\to\infty}$ 时

$$\mathbb{P}(-c\leqslant T\leqslant c)=\Phi(c)-\Phi(-c)=2\Phi(c)-1=1-\alpha\Rightarrow c=\Phi^{-1}\left(1-\frac{\alpha}{2}\right),$$

这里的 Φ 为标准正态分布的累计密度函数,c 亦被称为标准正态分布的上 $\alpha/2$ 分位数. 我们将区间 $\hat{I}$ 称为参数 θ_j 的**渐进置信区间** (asymptotic confidence interval).

需要指出, 正态回归模型 (3.12.1) 的置信水平为 $1-\alpha$ 的置信区间式 (3.71) 是使用 t 分布构造, 但式 (5.22) 则是使用的 $n\to\infty$ 下的正态分布. 两者有如下辨析:

a. 置信区间式 (3.71) 仅针对总体参数 β 成立, 而渐进置信区间式 (5.22) 则是对于任意的由 $\boldsymbol{\beta}$ 得到的参数 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})$. 后者的适用范围更广, 特别地,$\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\beta}$ 时式 (5.22) 便是回归系数 β_j 的渐进置信区间;

b. 相同的 $1-\alpha$ 的置信水平下, 由于 t 分布较标准正态分布更“矮胖”(这可以从图 (3.5) 中观察到),t 分布的上 $\alpha/2$ 分位数 c 要大于标准正态分布的, 因而置信区间式 (3.71) 的长度较渐进置信区间式 (5.22) 的更长, 估计的结果也更为保守. 当 n 充分大时, 相同的 $1-\alpha$ 的置信水平下两者的常数 c 区别不大;

c. 置信区间式 (3.71) 和渐进置信区间式 (5.22) 都准确度都依赖于标准误 SE. 标准误可以是同方差的, 也可以是异方差稳健的.SE 越小, 则区间估计越为精确, 参数估计的不确定性越少.

当 $\alpha = 0.05$ 时, 标准正态分布的常数 $c \approx 1.96$, 这样我们依然有参数 θ_j 的 95% 的置信水平的渐进置信区间近似于

$$\hat{I} = \left[\hat{\theta}_j - 2 \cdot \mathrm{SE}\left(\hat{\theta}_j\right), \hat{\theta}_j + 2 \cdot \mathrm{SE}\left(\hat{\theta}_j\right)\right] \tag{5.23}$$

这一经验法则成立[3].

下面我们来研究回归函数的置信区间. 对于线性回归模型, 对于给定的解释变量 $\mathbf{X} = \boldsymbol{x}$(或解释变量的实现值为 $\boldsymbol{x}$), 则最优预测量为

$$m(\boldsymbol{x}) = \mathbb{E}(Y | \mathbf{X} = \boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta},$$

显然 $m(\boldsymbol{x})$ 可以视为参数 $\boldsymbol{\beta}$ 的一个映射 $m : \mathbb{R}^{K\times 1} \to \mathbb{R}$, 则 $m(\boldsymbol{x})$ 有估计量 $\hat{m}(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, 其亦为 OLS 估计量的拟合值. 不难计算, 映射 $m : \mathbb{R}^{K\times 1} \to \mathbb{R}$ 的 Jacobi 矩阵为

$$\boldsymbol{J}m(\boldsymbol{x}) = (x_1, \cdots, x_K)_{1\times K}$$

故条件期望函数 $m(\boldsymbol{x})$ 的渐进协方差矩阵退化为了结果为二次型的方差

$$\boldsymbol{V} = \mathrm{Var}(m(\boldsymbol{x})) = \boldsymbol{J}m(\boldsymbol{x})\,\boldsymbol{V}_{\boldsymbol{\beta}}\,(\boldsymbol{J}m(\boldsymbol{x}))^{\mathrm{T}} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{x}$$

即有 $\boxed{\mathrm{SE}(\hat{m}(\boldsymbol{x})) = \sqrt{\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{x}}}$. 于是依据式 (5.22), 线性回归模型的条件期望函数 $m(\mathbf{X} = \boldsymbol{x})$ 的 95% 置信水平的渐进置信区间近似为

$$\hat{I} = \left[\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - 1.96 \cdot \sqrt{\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{x}}, \boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}} + 1.96 \cdot \sqrt{\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{x}}\right], \tag{5.24}$$

我们称式 (5.24) 为线性回归模型的 95% 置信水平的**回归区间** (Regression Interval).

我们通过 Monte Carlo 模拟来说明回归区间. 依旧考虑一元线性回归模型 (5.6), 但这次我们正确设定回归方程为 $Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$, 则使用相同的数据, 计算得到 OLS 估计量为

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} -0.0174 \\ 1.0012 \end{pmatrix}$$

且 $\hat{\boldsymbol{\beta}}$ 的条件协方差矩阵的估计量为

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}} = \begin{pmatrix} 0.0067 & -0.0038 \\ -0.0038 & 0.0061 \end{pmatrix},$$

于是依据式 (5.24) 有

$$\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}} = (1, x^2)\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \hat{\beta}_0 + \hat{\beta}_1 X^2 = -0.0174 + 1.0012x^2,$$

$$\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{x} = (1, x^2)\begin{pmatrix} 0.0067 & -0.0038 \\ -0.0038 & 0.0061 \end{pmatrix}\begin{pmatrix} 1 \\ x^2 \end{pmatrix} = 0.0067 + 0.0061x^4 - 0.0076x^2,$$

[3]事实上, 可以计算式 (5.23) 的渐进置信水平约为 95.4%.

故条件期望函数 $m(X=x)=x^2$ 的 95% 的置信区间的上下边界为

$$-0.0174+1.0012x^2 \pm 1.96 \cdot \sqrt{0.0067+0.0061x^4-0.0076x^2}.$$

图 (5.5) 绘制了回归函数/条件期望函数 $m(X=x)=x^2$、正确设定下 OLS 估计量计算的样本回归直线和依据 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}}$ 计算的置信水平为 95% 的回归区间. 需要指出, 在理解图 (5.5) 中的回归区间时, 应当考虑的是平行与 Y 轴的直线与阴影部分所截得的竖直方向的区间, 这才是符合式 (5.24) 的含义. 不难发现, 当 X 的实现值 x 越接近于 0, 回归区间的长度越小, 但当 x 偏离 0 时, 回归区间的长度是增大的, 即意味着对于条件期望函数 $m(X=x)=x^2$ 的估计越不准确.

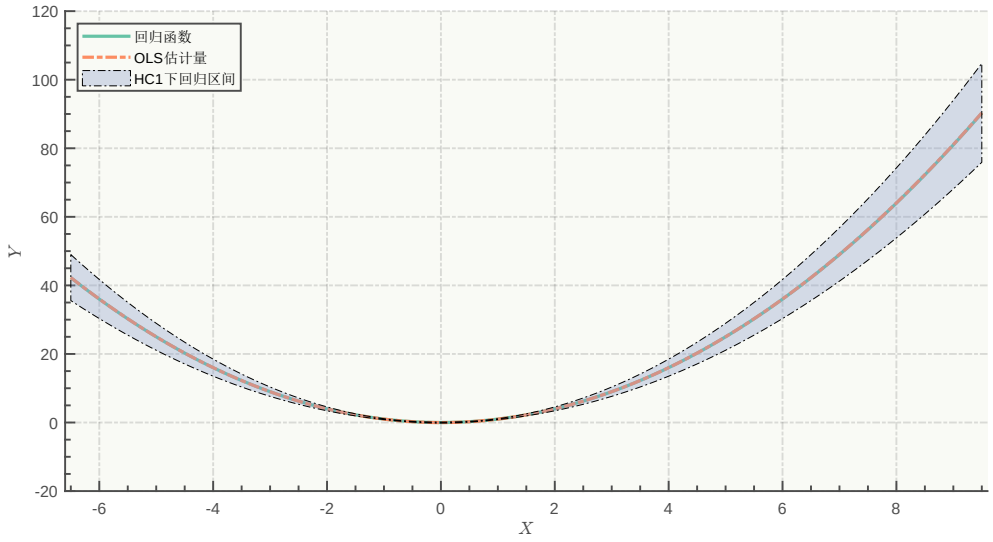

图 5.5: $m(X=x)=X^2$ 的回归区间

预测区间, 结合第3.14节, 需要结合 LLO.

5.4 Wald 统计量与 Wald 检验

在第 (3.13) 节中, 我们指出了 F 检验的统计量式 (3.78) 是 Wald 统计量. 现在我们将在大样本理论下介绍 Wald 统计量.

定义 5.4.1 (Wald 统计量). 对于线性投影模型的总体参数 $\boldsymbol{\beta}$, 设感兴趣的参数 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta}):\mathbb{R}^{K\times 1}\to\mathbb{R}^{q\times 1}$, 则 $\boldsymbol{\theta}$ 有插入估计量 $\hat{\boldsymbol{\theta}}=\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)$. 我们称二次型

$$W(\boldsymbol{\theta})=\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=n\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{\mathrm{T}}\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \tag{5.25}$$

为 **Wald 统计量** (*Wald statistic*), 其中 $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}=n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}$ 是参数 $\boldsymbol{\theta}$ 的渐进协方差矩阵的估计量,$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}$ 是 $\boldsymbol{\theta}$ 的协方差矩阵的估计量.

Wald 统计量以协方差矩阵为权重, 度量了估计量 $\hat{\boldsymbol{\theta}}$ 与参数 $\boldsymbol{\theta}$ 的加权距离. 特别地, 当映射 $\boldsymbol{h}$ 为 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})=\beta_k, k=1,\cdots,K$, 即选取了回归系数第 k 个分量的估计量, 此时 Wald 统计量为

$$W(\boldsymbol{\beta})=\left(\hat{\beta}_k-\beta_k\right)^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\beta}_k}^{-1}\left(\hat{\beta}_k-\beta_k\right)=\frac{\left(\hat{\beta}_k-\beta_k\right)^2}{\operatorname{Var}\left(\hat{\beta}_k\middle|\boldsymbol{X}\right)}=T^2,$$

即退化为了 t 统计量的平方. 因此,t 统计量可以视为 Wald 统计量的一种特殊情形.

在第3.13节中, 我们证明了当 $\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{R}\boldsymbol{\beta}$ 和 $\mathrm{H}_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{r}$ 时由正态假设可以推出 Wald 统计量服从卡方分布, 进而用于假设检验之中. 我们现在证明, 式 (5.25) 的渐进分布即是卡方分布. 首先, 依据 Delta 方法的应用 (见定理 (4.4.9)), 由 OLS 估计量 $\boldsymbol{\beta}$ 的渐近正态性可得参数 $\boldsymbol{\theta}$ 也是渐近正态的, 即但 $n\to\infty$ 时 $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow{d}\boxed{\boldsymbol{Z}\sim\mathrm{N}(\boldsymbol{0},\boldsymbol{V}_{\boldsymbol{\theta}})}$, 而 $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ 依概率收敛于渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}$, 于是由连续映射定理 (4.4.6) 得

$$W(\boldsymbol{\theta})=\left[\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right]^{\mathrm{T}}\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow{d}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Z}$$

即 Wald 统计量具有渐进分布 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Z}$, 形如第3.13节中对式 (3.78) 的证明一样, 可以证明 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Z}$ 服从自由度为 q 的卡方分布. 依据定理 (3.6.4), 若渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\theta}}$ 是正定的[4], 则存在唯一的 q 阶实对称方阵 $\boldsymbol{B}$ 使得 $\boldsymbol{V}_{\boldsymbol{\theta}}=\boldsymbol{B}^2$. 于是渐进分布可以分解为

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\right)^{-1}\boldsymbol{Z}=\left(\boldsymbol{B}^{-1}\boldsymbol{Z}\right)^{\mathrm{T}}\left(\boldsymbol{B}^{-1}\boldsymbol{Z}\right),$$

而随机向量 $\boldsymbol{B}^{-1}\boldsymbol{Z}$ 是多元正态分布 $\boldsymbol{Z}$ 的线性组合, 依然服从多元正态分布, 且有

$$\operatorname{Var}\left(\boldsymbol{B}^{-1}\boldsymbol{Z}\middle|\boldsymbol{X}\right)=\boldsymbol{B}^{-1}\boldsymbol{V}_{\boldsymbol{\theta}}\boldsymbol{B}^{-1}=\boldsymbol{B}^{-1}\boldsymbol{B}^2\boldsymbol{B}^{-1}=\boldsymbol{I}_q,$$

故 $\boldsymbol{B}^{-1}\boldsymbol{Z}\sim\mathrm{N}(\boldsymbol{0},\boldsymbol{I}_q)$, 这也保证了 $\boldsymbol{B}^{-1}\boldsymbol{Z}$ 的各分量是独立的. 不难推知当渐进分布

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{Z}=\left(\boldsymbol{B}^{-1}\boldsymbol{Z}\right)^{\mathrm{T}}\left(\boldsymbol{B}^{-1}\boldsymbol{Z}\right)\sim\chi^2(q),$$

即有当 $n\to\infty$ 时,$W(\boldsymbol{\theta})\xrightarrow{d}\chi^2(q)$.

定理 5.4.1. *在假设 (5.1.1) 以及 $\mathbb{E}Y^4<\infty,\mathbb{E}\|\mathbf{X}\|^4<\infty$ 和 $\operatorname{rank}\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\beta})=q$ 下, 形如式 (5.25) 的 Wald 统计量满足有当 $n\to\infty$ 时 $W(\boldsymbol{\theta})\xrightarrow{d}\chi^2(q)$.*

特别地, 当线性回归模型满足有同方差假设式 (3.7) 时, 则 Wald 统计量可以简化为

$$W^0(\boldsymbol{\theta})=\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}^0\right)^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=n\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}^0\right)^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right),$$

[4] $\boldsymbol{V}_{\boldsymbol{\theta}}=\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\beta})\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\beta}))^{\mathrm{T}}$, 其中渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\beta}}$ 的正定性由假设 (5.1.1) 中 (d) 来确保, 故只需有 $\boxed{\operatorname{rank}\boldsymbol{J}\boldsymbol{h}(\boldsymbol{\beta})=q}$ 即可保证渐进协方差矩阵 $\boldsymbol{V}_{\boldsymbol{\theta}}$ 的正定性?.

这里的 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}^{0}$ 和 $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{0}$ 都是同方差的协方差矩阵估计量, 我们称 $W^{0}(\boldsymbol{\theta})$ 为**同方差的 Wald 统计量** (homoskedastic Wald statistic). 依据式 (5.20), 即得同方差的 Wald 统计量有

$$\begin{aligned}
W^{0}(\boldsymbol{\theta}) =& n\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{\mathrm{T}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}s^{2}\right)^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \\
=& n\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{\mathrm{T}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right)^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)s^{-2} \\
=& \left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{\mathrm{T}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right)^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)s^{-2}.
\end{aligned}$$

由于同方差是异方差的特例, 因而当 $n\to\infty$ 时 $W^{0}(\boldsymbol{\theta})\xrightarrow{d}\chi^{2}(q)$.

Wald 统计量被广泛用于假设检验中, 第3.13节中我们便指出了 F 统计量是同方差的 Wald 统计量, 现在我们介绍一般化的 Wald 检验. 考虑关于参数 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})$ 原假设

$$\mathrm{H}_{0}:\boldsymbol{\theta}=\boldsymbol{\theta}_{0}\quad v.s.\quad \mathrm{H}_{1}:\boldsymbol{\theta}\neq\boldsymbol{\theta}_{0},$$

这也等价于

$$\mathrm{H}_{0}:\boldsymbol{\beta}\in B=\{\boldsymbol{\beta}:\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\theta}_{0}\}\quad v.s.\quad \mathrm{H}_{1}:\boldsymbol{\beta}\notin B=\{\boldsymbol{\beta}:\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\theta}_{0}\}.$$

我们使用 Wald 统计量来定义关于 H_{0} 的检验. 若 H_{0} 成立, 则依据式 (5.18) 有

$$\begin{aligned}
W =& n\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{0}\right) \\
=& n\left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(n\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{\theta}_{0}\right) \\
=& \left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{\theta}_{0}\right),
\end{aligned}$$

由于 Wald 统计量是度量了估计量 $\hat{\boldsymbol{\beta}}$ 与参数 $\boldsymbol{\theta}=\boldsymbol{\theta}_{0}$ 的距离, 则原假设为真时,Wald 统计量的数值应该尽可能的小, 这样我们便可以利用 Wald 统计量的渐进分布进行检验.

定理 5.4.2 (Wald 检验)**.** *在假设 (5.1.1) 下, 对于模型 (3.5.1) 的回归系数 $\boldsymbol{\beta}$, 设参数 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})$, 其中 $\boldsymbol{h}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{q\times 1}$, 就假设*

$$\mathrm{H}_{0}:\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\theta}_{0}\quad v.s.\quad \mathrm{H}_{1}:\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})\neq\boldsymbol{\theta}_{0}$$

而言, 以 Wald 统计量式 (5.25) 得到

$$W=\left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{\theta}_{0}\right)\xrightarrow{d}\chi^{2}(q),\quad n\to\infty \tag{5.26}$$

作为检验统计量, 则有渐进显著性水平为 α的检验

$$\phi: 若W > F^{-1}(1-\alpha) 则拒绝\ H_0,$$

其中 $F(\cdot)$ 是自由度为 q 的卡方分布的累计密度函数,$F^{-1}(\cdot)$ 为 $F(\cdot)$ 的逆函数. 我们称检验 ϕ 为原假设 $\mathrm{H}_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ 的 **Wald 检验** *(Wald test).Wald* 检验的渐进 p 值为 $p = 1 - F(W)$.

特别地, 当样本满足条件同方差式 (3.7) 时, 则有式 (5.26) 即有

$$W^0 = \frac{\left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right) - \boldsymbol{\theta}_0\right)^{\mathrm{T}} \left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{J}\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right)^{-1} \left(\boldsymbol{h}\left(\hat{\boldsymbol{\beta}}\right) - \boldsymbol{\theta}_0\right)}{s^2} \xrightarrow{d} \chi^2(q),$$

这时的 Wald 检验与定理 (5.4.2) 并无差异.

Wald 检验的一个实例是考虑原假设为线性约束 $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{\theta}_0$.

定理 5.4.3 (线性约束的 Wald 检验)**.** 在假设 *(5.1.1)* 下, 对于模型 *(3.12.1)* 的回归系数 $\boldsymbol{\beta}$, 就假设

$$\mathrm{H}_0: \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{\theta}_0 \quad v.s. \quad \mathrm{H}_1: \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{\theta}_0,$$

其中 $\boldsymbol{R} \in \mathbb{R}^{q\times K}$,rank$\boldsymbol{R} = q$, 以形如式 *(5.25)* 的 *Wald* 统计量作为检验统计量, 则有

$$W = \left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)^{\mathrm{T}} \left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1} \left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \chi^2(q), \quad n \to \infty, \tag{5.27}$$

$$W^0 = \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)^{\mathrm{T}} \left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1} \left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)}{s^2} \xrightarrow{d} \chi^2(q), \quad n \to \infty. \tag{5.28}$$

我们称渐进显著性水平为 α的检验

$$\phi: 若W > F^{-1}(1-\alpha) 则拒绝\ H_0$$

为 H_0 的 *Wald* 检验, 其中 $F(\cdot)$ 是自由度为 q 的卡方分布的累计密度函数,$F^{-1}(\cdot)$ 为 $F(\cdot)$ 的逆函数.*Wald* 检验的渐进 p 值为 $p = 1 - F(\lambda_{\mathrm{LR}})$.

在第3.13节中, 我们介绍了有限样本下正态回归模型的假设检验, 而我们指出了 t 统计量和 F 统计量是 Wald 统计量的特例, 因而我们现在将定理 (3.13.1) 和定理 (3.13.2) 拓展至大样本下.

对于 t 统计量式 (3.77), 我们知道 T^2 即是 Wald 统计量, 因而有 T^2 具有渐进分布 $\chi^2(1)$, 这样对于单个系数 β_k 的原假设 $\mathrm{H}_0: \beta_k = \theta_0$ 即可使用 T^2 去进行假设检验. 当然, 依据 OLS 估计量的渐进分布式 (5.9) 我们也可以得到 t 统计量是渐进正态的.

定理 5.4.4 (大样本下单个系数的 t 检验). *在假设 (5.1.1) 下, 对于模型 (3.5.1) 的回归系数 β_k, 就假设*

$$\mathrm{H}_0 : \beta_k = \theta_0 \quad v.s. \quad \mathrm{H}_1 : \beta_k \neq \theta_0,$$

而言, 以 $\hat{\beta}_k$ 的 t 统计量

$$T = \frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)} \overset{\mathrm{H}_0}{=\!=} \frac{\beta_k - \theta_0}{\mathrm{SE}\left(\hat{\beta}_k\right)} \xrightarrow{d} \mathrm{N}(0,1), \quad n \to \infty, \tag{5.29}$$

作为检验统计量, 则有渐进显著性水平为 α的检验

$$\phi : \text{若 } |T| > \Phi^{-1}\left(1 - \frac{1}{2}\alpha\right) \text{ 则拒绝 } H_0,$$

我们称检验 ϕ 为回归系数 β_k 的 ***t 检验*** *(t test). 其中 Φ 的标准正态分布累计密度函数,Φ^{-1} 为 Φ 的逆函数.t 检验的 p 值为 $p = 2\left(1 - \Phi\left(|T|\right)\right) = 2\Phi\left(-|T|\right)$.*

在第3.13节中, 我们知道定理 (3.13.2) 所使用的 F 统计量或 F 比率是同方差的 Wald 统计量 W^0 除以约束条件的个数, 则依据定理 (5.4.3) 可知当 $n \to \infty$ 时 $F = W^0/q \xrightarrow{d} \chi^2(q)/q$, 因而 F 检验的大样本版本即是定理 (5.4.3). 特别地, 对于方程的显著性检验, 这时 F 统计量的渐进分布为 $\chi^2\left(K-1\right)/\left(K-1\right)$, 这里的$K$ 是包含了一个截距项的解释变量的个数. 依据式 (3.83) 我们知道其等价于对拟合优度 R^2 的检验, 且有

$$R^2 = \frac{\left(K-1\right)F}{\left(K-1\right)F + n - K} = 1 - \frac{n-K}{\left(K-1\right)F + n - K}, \tag{5.30}$$

我们现在尝试推导 R^2 的渐进分布. 对于 Wald 统计量而言, 注意到当 $n \to \infty$ 时

$$\frac{W\left(\boldsymbol{\theta}\right)}{n} = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^{\mathrm{T}} \hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{P} \mathbf{0}^{\mathrm{T}} \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \mathbf{0} = 0,$$

因而有 $F/n \xrightarrow{P} 0$, 则依据连续映射定理 (4.4.6) 知

$$nR^2 = \frac{\left(K-1\right)}{\left(K-1\right)\frac{F}{n} + 1 - \frac{K}{n}} F \xrightarrow{d} \left(K-1\right)F = \chi^2\left(K-1\right), \quad n \to \infty, \tag{5.31}$$

则在大样本下, 对于方程显著性的检验可以直接使用拟合优度 R^2 去作为检验统计量. 这一点我们将在后续中经常使用.

定理 5.4.5 (大样本下对方程显著性的检验). *在假设 (5.1.1) 下, 对于含有截距项的模型 (3.12.1) 进行方程的显著性检验, 则原假设为*

$$\mathrm{H}_0 : \beta_2 = \cdots = \beta_K = 0.$$

使用 F 统计量进行 *Wald* 检验, 等价于使用拟合优度 R^2 得到

$$nR^2 \xrightarrow{d} \chi^2(K-1), \quad n \to \infty, \tag{5.32}$$

作为检验统计量进行检验, 则有渐进显著性水平为 α的检验

$$\phi: 若nR^2 > F^{-1}(1-\alpha) 则拒绝\ H_0,$$

其中 F 是自由度为 $K-1$ 的卡方分布的累计密度函数,F^{-1} 为 F 的逆函数. 检验 ϕ 的 p 值为 $p = 1 - F(nR^2)$.

最后我们介绍Stata软件中 Wald 检验. 在Stata软件中我们test命令对回归系数进行假设检验, 其方法便是 F 统计量形式的 Wald 检验. 例如对于单个系数所进行的就是以 $F = W^0/q = T^2/1$ 作为统计量的 F 检验, 并使用 F 分布区计算 p 值. 这意味着Stata实际上使用的是有限样本下的定理 (3.13.2) 进行假设检验.

5.5 遍历平稳下 OLS 估计量的渐进性质

5.6 代码

```
%% OLS估计量的相合性
clear
clc
s=RandStream('mcg16807','Seed',114514);%设定复现的随机流以及种子
%设定E(Y|X)=10x+25,e~N(0,50^2)
n=10000;%样本容量
X=[ones(n,1),randsample(s,-100:100,n,true)'];%生成样本X
e=50*randn(s,n,1);%生成扰动项
Y=X*[25;10]+e;%数据生成过程
Beta=zeros(n,2);%初始化存储每次回归结果的Beta
for i=1:n
    beta=X(1:i,:)\Y(1:i,:);%OLS估计量
    Beta(i,:)=beta;%保存OLS估计量
end
selfGrootDefault(2)%绘图美化函数,没有请注释
%绘制估计值均值变化
figure(1)%普通坐标
set(gcf,'unit','centimeters', ...
```

```
19      'Position',[0 0 21*1.25 29.7*0.35*1.25]);%设置figure为A4纸宽度
20  tiledlayout(2,1);%创建2*1画布
21  nexttile
22  plot(Beta(:,1),'LineWidth',2)
23  xlabel('\fontname{宋体}样本容量\fontname{Times New ...
        Roman}n','FontSize',12.5)
24  ylabel('\beta_1\fontname{宋体}估计值','FontSize',12.5);
25  yline(25,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
        ...
26      ,'left','LabelVerticalAlignment','bottom')
27
28  nexttile
29  plot(Beta(:,2),'LineWidth',2)
30  xlabel('\fontname{宋体}样本容量\fontname{Times New ...
        Roman}n','FontSize',12.5)
31  ylabel('\beta_0\fontname{宋体}估计值','FontSize',12.5);
32  yline(10,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
        ...
33      ,'left','LabelVerticalAlignment','bottom')
34
35  figure(2)%对数坐标
36  set(gcf,'unit','centimeters', ...
37      'Position',[0 0 21*1.25 29.7*0.35*1.25]);%设置figure为A4纸宽度
38  tiledlayout(2,1);%创建2*1画布
39  nexttile
40  semilogx(Beta(:,1),'LineWidth',2)
41  xlabel('$\ln n$','FontSize',12.5,'Interpreter','latex')
42  ylabel('\beta_1\fontname{宋体}估计值','FontSize',12.5);
43  yline(25,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
        ...
44      ,'left','LabelVerticalAlignment','bottom')
45
46  nexttile
47  semilogx(Beta(:,2),'LineWidth',2)
48  xlabel('$\ln n$','FontSize',12.5,'Interpreter','latex')
49  ylabel('\beta_0\fontname{宋体}估计值','FontSize',12.5);
50  yline(10,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
        ...
51      ,'left','LabelVerticalAlignment','bottom')
```

第六章　线性回归模型的拓展估计问题

本章介绍对于线性回归模型的拓展问题, 包含予以约束条件、放松假设 (3.2.1), 并介绍针对这些问题的解决方法.

6.1　受限回归与受约束最小二乘法 (CLS)

在第3.13节中, 我们介绍了关于线性约束的假设检验问题, 并指出 F 检验可以通过受约束的 LS 去构造, 这便是我们本节要介绍的内容. 对于线性投影模型 (3.3.1) 的总体

$$Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon = \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon,$$

我们设回归系数 $\boldsymbol{\beta}$ 被施加有**约束** (constraint)条件或**限制** (restriction)条件, 例如一个最为常见的**线性约束** (linear constraint)条件即为

$$\boldsymbol{R}_{q\times K}\boldsymbol{\beta} = \boldsymbol{c}_{q\times 1} \Leftrightarrow \begin{cases} R_{11}\beta_1 + \cdots + R_{1K}\beta_K = c_1, \\ \cdots \\ R_{q1}\beta_1 + \cdots + R_{qK}\beta_K = c_q, \end{cases} \tag{6.1}$$

其中矩阵 $\boldsymbol{R} \in \mathbb{R}^{q\times K}$,rank$\boldsymbol{R} = q$,$\boldsymbol{c} \in \mathbb{R}^{q\times 1}$, 即是对于参数 $\boldsymbol{\beta}$ 施加了 q 个线性无关的线性方程. 我们将关于 $\boldsymbol{\beta}$ 的线性方程组 (6.1) 的解空间 $B = \{\boldsymbol{\beta} : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{c}\}$ 称为模型 (3.5.1) 的受限的参数空间.

需要指出, 我们是对线性投影模型施加了关于参数 $\boldsymbol{\beta}$ 约束条件, 但是这并没有参数 $\boldsymbol{\beta}$ 的解析式 $\boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}Y\right)$. 但在**取值**上, 参数 $\boldsymbol{\beta}$ 满足有约束条件. 之所以施加例如式 (6.1) 的约束条件, 通常是因为有经济理论支撑了这些约束条件是正确的.

对应本节研究的模型, 我们总结如下.

模型 6.1.1 (受限回归模型)**.** *对于模型 (3.3.1)、模型 (3.5.1) 或模型 (3.12.1), 若增加线性约束式 (6.1), 则我们称总体模型为***受限回归模型** *(restricted regression model).*

需要指出, 模型 (6.1.1) 的约束条件可以是非线性约束, 亦或是不等式约束, 本节仅介绍线性约束下的回归模型.

下面我们来研究在线性约束式 (6.1) 下模型 (3.5.1) 的参数估计. 同样地, 我们依旧以使得 MSE 的估计量 n^{-1}SSE 最小的最优化问题的解来进行参数估计.

定义 6.1.1 (受约束的最小二乘估计量). *对于模型 (3.3.1), 设参数 $\boldsymbol{\beta}$ 满足有线性约束式 (6.1), 我们称*

$$\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} = \underset{\boldsymbol{R\beta}=\boldsymbol{c}}{\operatorname{argmin}} \operatorname{SSE}(\boldsymbol{\beta}) = \underset{\boldsymbol{R\beta}=\boldsymbol{c}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 \tag{6.2}$$

*为总体参数 $\boldsymbol{\beta}$ 的***受约束的最小二乘估计量** *(constrained least squares estimator), 简称为 **CLS 估计量**. 在一些计量经济学教材中,CLS 估计量有时也被称为**受限制的最小二乘估计量** (restricted least square estimator), 简称 **RLS 估计量**.*

定义 (6.1.1) 中我们使用 "~" 来表示 CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$, 这是区别于 OLS 估计量 $\hat{\boldsymbol{\beta}}$. 下面我们求解 CLS 估计量的表达式. 于是式 (6.2), 最为常规的求解方法即是采用 **Lagrange 乘数法** (Lagrange multipliers). 由于

$$\begin{aligned}\operatorname{SSE}(\boldsymbol{\beta}) =& \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ =& \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - 2\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta},\end{aligned}$$

不妨设

$$\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2}\operatorname{SSE}(\boldsymbol{\beta}) + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{R\beta} - \boldsymbol{c}), \tag{6.3}$$

其中 q 为列向量 $\boldsymbol{\lambda}$ 是 Lagrange 乘数. 为了使得 SSE 最小, 则有一阶条件

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathscr{L}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}, \tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}\right) = -\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} + \boldsymbol{R}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}} = \mathbf{0}_{K\times 1}, \tag{6.4}$$

$$\frac{\partial}{\partial \boldsymbol{\lambda}} \mathscr{L}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}, \tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}\right) = \boldsymbol{R}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} - \boldsymbol{c} = \mathbf{0}_{K\times 1} \tag{6.5}$$

成立. 注意到 OLS 估计量为 $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, 则由一阶条件式 (6.4) 可以得到 CLS 估计量为

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} =& \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{R}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}\right] \\ =& \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}} \\ =& \boxed{\hat{\boldsymbol{\beta}} - \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}},\end{aligned}$$

则我们只需解出 Lagrange 乘数 $\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}$. 由一阶条件式 (6.5) 有 $\boxed{\boldsymbol{R}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} = \boldsymbol{c}}$, 将一阶条件

式 (6.4) 左乘以 $\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$, 即有

$$\begin{aligned}
&\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}+\boldsymbol{R}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}\right)=\mathbf{0},\\
\Rightarrow&-\boldsymbol{R}\hat{\boldsymbol{\beta}}+\boldsymbol{R}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}+\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}=\mathbf{0},\\
\Rightarrow&-\boldsymbol{R}\hat{\boldsymbol{\beta}}+\boldsymbol{c}+\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}=\mathbf{0},
\end{aligned}$$

故解得 Lagrange 乘数为

$$\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}}=\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\mathbf{c}\right), \tag{6.6}$$

这里 $\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}$ 的可逆性依赖于 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的正定性和 $\boldsymbol{R}$ 满秩. 这样我们便有式 (6.6) 得到 CLS 估计量为

$$\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}=\hat{\boldsymbol{\beta}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\mathbf{c}\right), \tag{6.7}$$

若使用 $\hat{\boldsymbol{Q}}_{\mathbf{XX}}=n^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)$, 则式 (6.7) 可以记为

$$\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}=\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\mathbf{c}\right). \tag{6.8}$$

下面我们介绍 CLS 估计量一些有用的代数性质. 记 CLS 估计量的残差为 $\tilde{\varepsilon}_i=Y_i-\mathbf{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$(区别于 LLO 回归的预测残差式 (3.40)), 使用矩阵符号有 $\tilde{\boldsymbol{\varepsilon}}=\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$. 记 K 阶实对称矩阵

$$\boldsymbol{A}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},$$

则有 CLS 估计量有如下代数性质.

定理 6.1.1 (CLS 估计量的代数性质)**.** *在予以线性约束式 (6.1) 下, 线性投影模型 (3.3.1) 有 CLS 估计量式 (6.7), 其与 OLS 估计量 $\hat{\boldsymbol{\beta}}$、回归误差 ε 等有如下的性质:*

(a) 总体参数 $\boldsymbol{\beta}$ 和 CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 满足有约束条件:

$$\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{R}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}=\boldsymbol{c}. \tag{6.9}$$

(b) OLS 估计量与约束条件的关系:

$$\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{c}=\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}. \tag{6.10}$$

(c) CLS 估计量的抽样误差:

$$\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}. \tag{6.11}$$

(d) CLS 估计量的残差向量:

$$\tilde{\boldsymbol{\varepsilon}} = \left(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}. \tag{6.12}$$

(e) K 阶实对称方阵 $\boldsymbol{A}$ 是半正定的.n 阶方阵 $\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}$ 是实对称的幂等矩阵, 且 $\operatorname{tr}\left(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right) = n - K + q$.

证明. (a) 是平凡的, 对于 (b) 则代入式 (6.9) 有

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c} = \boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{R}\underbrace{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}_{\text{抽样误差}} = \boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}.$$

对于 (c), 注意到

$$\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} - \boldsymbol{\beta} = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right), \tag{6.13}$$

则我们将 CLS 估计量的抽样误差分解了 OLS 估计量的抽样误差与 $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 的差值. 对于后者, 利用式 (6.7) 与式 (6.10) 即有

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right) \\
&= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} \\
&= \boxed{\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}},
\end{aligned} \tag{6.14}$$

将 OLS 估计量的抽样误差与 $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 代入式 (6.13) 即得

$$\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} - \boldsymbol{\beta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} - \boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = \left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}.$$

对于 (d), 这是类似于 OLS 估计量的残差向量 $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{\varepsilon} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon}$, 其中 $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ 和 $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}$ 分别是投影矩阵和归零矩阵. 对于 CLS 的残差向量, 计算有

$$\begin{aligned}
\tilde{\boldsymbol{\varepsilon}} =& \boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} = \boldsymbol{\varepsilon} - \boldsymbol{X}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} - \boldsymbol{\beta}\right) \\
=& \boldsymbol{\varepsilon} - \boldsymbol{X}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon} \\
=& \left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon} = \left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}.
\end{aligned}$$

对于 (e), 由于 $\boldsymbol{A}$ 是实对称的, 故方阵 $\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}$ 也是实对称的. 而

$$\begin{aligned}
&\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)^2 = \left(\boldsymbol{M} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)^2 \\
=& \boldsymbol{M}^2 + \boldsymbol{M}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}} \\
=& \boldsymbol{M} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}},
\end{aligned}$$

又

$$\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}=\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}},$$

故 $\left(\boldsymbol{I}_n-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)^2=\boldsymbol{M}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}$, 即 $\boldsymbol{I}_n-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}$ 是幂等矩阵. 对于 $\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}$, 由于 $\boldsymbol{R}$ 是行满秩的, 且 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 可逆, 故可以推知幂等方阵 $\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}$ 是秩为 q, 即有 $\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)=q$, 于是由迹运算的线性性知 $\operatorname{tr}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)=\operatorname{tr}\left(\boldsymbol{I}-\boldsymbol{P}\right)+\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)=n-K+q$. □

6.2 CLS 估计量的有限样本性质

本节我们介绍 CLS 估计量的统计样本性质, 包括有限样本性质以及大样本性质. 我们依旧遵循介绍 OLS 估计量性质的顺序, 对于有限样本性质, 首先介绍 CLS 估计量的无偏性, 然后讨论 CLS 估计量的方差估计量, 再讨论引入正态假设后估计量的统计分布; 对于大样本性质, 则先讨论 CLS 估计量的相合性, 再分析其具有的渐进分布.

我们现在考虑施加了线性约束式 (6.1) 下线性回归模型 (3.5.1) 的 CLS 估计量. 我们首先研究有限样本性质.

定理 6.2.1 (CLS 估计量的无偏性). *在予以线性约束式 (6.1) 下, 线性回归模型 (3.5.1) 有 CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$, 则有 $\mathbb{E}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\boldsymbol{\beta}$.*

证明. 对于式 (6.7) 求条件期望得

$$\begin{aligned}\mathbb{E}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=&\mathbb{E}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)-\mathbb{E}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\mathbf{c}\right)\middle|\boldsymbol{X}\right)\\=&\boldsymbol{\beta}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\mathbb{E}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\mathbf{c}\middle|\boldsymbol{X}\right)\\=&\boldsymbol{\beta}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\left(\boldsymbol{R}\boldsymbol{\beta}-\mathbf{c}\right)\\=&\boldsymbol{\beta}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\cdot\mathbf{0}=\boldsymbol{\beta}.\end{aligned}$$

亦可对于 CLS 估计量的抽样误差式 (6.11) 求条件期望

$$\begin{aligned}\mathbb{E}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\middle|\boldsymbol{X}\right)=&\mathbb{E}\left(\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)\\=&\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\mathbb{E}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\mathbf{0},\end{aligned}$$

即有 $\mathbb{E}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\boldsymbol{\beta}$. □

定理 6.2.2 (CLS 估计量的条件协方差矩阵). *在予以线性约束式 (6.1) 下, 线性回归模型 (3.5.1) 有 CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$, 则 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 的条件协方差矩阵为*

$$\boldsymbol{V}_{\tilde{\beta}}=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{D}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{6.15}$$

当模型满足有同方差假设式 (3.7) 时, 则 CLS 估计量有同方差的条件协方差矩阵

$$\boldsymbol{V}_{\tilde{\beta}}^{0}=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)\sigma^{2}. \tag{6.16}$$

证明. 当样本存在条件异方差时, 利用式 (3.10),CLS 估计量的条件协方差矩阵为

$$\begin{aligned}
\boldsymbol{V}_{\tilde{\beta}}=&\mathbb{E}\left(\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\right)\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\middle|\boldsymbol{X}\right)\\
=&\mathbb{E}\left(\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}}\middle|\boldsymbol{X}\right)\\
=&\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}}\\
=&\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{D}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}},
\end{aligned}$$

即证式 (6.15).

特别地, 当模型满足有同方差假设式 (3.7) 时,$\boldsymbol{D}=\boldsymbol{I}_n\sigma^2$, 则式 (6.15) 可化简有

$$\begin{aligned}
\boldsymbol{V}_{\tilde{\beta}}^{0}&=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}}\sigma^{2}\\
&=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}\right)\sigma^{2}\\
&=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{A}+\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}\right)\sigma^{2}\\
&=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-2\boldsymbol{A}+\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\right)\sigma^{2},
\end{aligned}$$

这里用到了方阵 $\boldsymbol{A}$ 是实对称的. 注意到

$$\begin{aligned}
&\boxed{\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\\
=&\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\
=&\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\boxed{\boldsymbol{A}},
\end{aligned}$$

故 $\boldsymbol{V}_{\tilde{\beta}}^{0}=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-2\boldsymbol{A}+\boldsymbol{A}\right)\sigma^{2}=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)\sigma^{2}$. □

由于回归方差 σ^2 未知, 我们现在研究 CLS 估计量下的方差估计量. 由式 (6.12) 计

算 CLS 估计量的残差平方和, 即有

$$\begin{aligned}
&\mathbb{E}\left(\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\operatorname{tr}\left(\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}\right)\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\operatorname{tr}\left(\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}\right)\middle|\boldsymbol{X}\right)\\
=&\mathbb{E}\left(\operatorname{tr}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}\right)\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\operatorname{tr}\left(\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)\middle|\boldsymbol{X}\right)\\
=&\operatorname{tr}\left(\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)\right)\overset{\text{同方差}}{=\!=\!=}\operatorname{tr}\left(\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\cdot\sigma^{2}\right)\\
=&\operatorname{tr}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\cdot\sigma^{2}=(n-K+q)\,\sigma^{2},
\end{aligned}$$

故有 CLS 估计量可以定义偏差校正估计量

$$s_{\mathrm{CLS}}^{2}=\frac{\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}}{n-K+q}=\frac{1}{n-K+q}\sum_{i=1}^{n}\tilde{\varepsilon}_{i}^{2},\tag{6.17}$$

于是同方差的 $\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}^{0}$ 有协方差矩阵估计量

$$\hat{\boldsymbol{V}}_{\tilde{\boldsymbol{\beta}}}^{0}=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)s_{\mathrm{CLS}}^{2}.\tag{6.18}$$

定理 6.2.3 (CLS 估计量的方差估计量). *在予以线性约束式 (6.1) 下, 线性回归模型 (3.5.1) 有 CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$, 则在同方差假设式 (3.7) 下 $\mathbb{E}(s_{\mathrm{CLS}}^{2}|\boldsymbol{X})=\sigma^{2},\mathbb{E}\left(\hat{\boldsymbol{V}}_{\tilde{\boldsymbol{\beta}}}^{0}\middle|\boldsymbol{X}\right)=\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}^{0}$.*

证明. 在同方差假设式 (3.7) 下, 则有

$$\mathbb{E}\left(s_{\mathrm{CLS}}^{2}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\frac{\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}}{n-K+q}\middle|\boldsymbol{X}\right)=\frac{\mathbb{E}\left(\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)}{n-K+q}=\frac{(n-K+q)\,\sigma^{2}}{n-K+q}=\sigma^{2}$$

和

$$\begin{aligned}
&\mathbb{E}\left(\hat{\boldsymbol{V}}_{\tilde{\boldsymbol{\beta}}}^{0}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)s_{\mathrm{CLS}}^{2}\middle|\boldsymbol{X}\right)\\
=&\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)\mathbb{E}\left(s_{\mathrm{CLS}}^{2}\middle|\boldsymbol{X}\right)=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)\sigma^{2}=\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}^{0}.
\end{aligned}$$

□

我们由式 (6.18) 得到 CLS 估计量第 k 个分量的标准误为

$$\mathrm{SE}\left(\tilde{\beta}_{k}\right)=\sqrt{\left(\hat{\boldsymbol{V}}_{\tilde{\boldsymbol{\beta}}}^{0}\right)_{kk}}=\sqrt{s_{\mathrm{CLS}}^{2}\left[\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)\right]_{kk}},\quad k=1,\cdots,K.\tag{6.19}$$

我们在引入正态假设式 (3.64) 和式 (3.65), 研究正态回归模型 (3.12.1) 下的 CLS 估计量.$\boldsymbol{\varepsilon}|\boldsymbol{X}\sim\mathrm{N}\left(\boldsymbol{0},\boldsymbol{I}_{n}\sigma^{2}\right)$ 下同方差是假设是自然成立的, 依据式 (6.11) 和式 (6.12) 是 $\boldsymbol{\varepsilon}$ 的线性组合, 则可推知

$$\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\middle|\boldsymbol{X}\sim\mathrm{N}\left(\boldsymbol{0},\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}^{0}\right),\quad\frac{(n-K+q)\,s_{\mathrm{CLS}}^{2}}{\sigma^{2}}\middle|\boldsymbol{X}\sim\chi^{2}\left(n-K+q\right),$$

这里用到了 $\operatorname{rank}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)=n-K+q$, 详细的证明同理第3.12节.

与无约束条件一样,CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 与其残差向量 $\tilde{\boldsymbol{\varepsilon}}$ 的独立的. 计算两者的条件协方差矩阵有

$$\begin{aligned}
&\operatorname{Cov}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}, \tilde{\boldsymbol{\varepsilon}} \middle| \boldsymbol{X}\right)=\mathbb{E}\left(\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\right) \tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}} \middle| \boldsymbol{X}\right)+\mathbb{E}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta} \middle| \boldsymbol{X}\right)\left(\mathbb{E}\left(\tilde{\boldsymbol{\varepsilon}} \middle| \boldsymbol{X}\right)\right)^{\mathrm{T}} \\
=&\mathbb{E}\left(\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\right) \tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}} \middle| \boldsymbol{X}\right)+\mathbf{0}_{K\times 1} \cdot \mathbf{0}_{1\times K} \\
=&\mathbb{E}\left(\left(\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right) \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{I}_n-\boldsymbol{P}+\boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right)^{\mathrm{T}} \middle| \boldsymbol{X}\right) \\
=&\left(\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right) \mathbb{E}\left(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right)\left(\boldsymbol{I}_n-\boldsymbol{P}+\boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right) \\
=&\left(\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right)\left(\boldsymbol{I}_n-\boldsymbol{P}+\boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right) \sigma^2,
\end{aligned}$$

而

$$\begin{aligned}
&\left(\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right)\left(\boldsymbol{I}_n-\boldsymbol{P}+\boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right) \\
=&\left(\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right)\left(\boldsymbol{M}+\boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right) \\
=&\left(\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}\right) \boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}=\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^{\mathrm{T}} \\
=&\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}-\boldsymbol{A} \boldsymbol{X}^{\mathrm{T}}=\boldsymbol{O}_{K\times n},
\end{aligned}$$

这里用到了 $\boldsymbol{M}\boldsymbol{X}=\boldsymbol{O}$ 和 $\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}=\boldsymbol{A}$. 故有 $\operatorname{Cov}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}, \tilde{\boldsymbol{\varepsilon}} \middle| \boldsymbol{X}\right)=\boldsymbol{O}_{K\times n}$, 即 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 与 $\tilde{\boldsymbol{\varepsilon}}$ 是不相关的, 在正态假设下即意味着两者是独立的, 这可以进一步推出抽样误差 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}$ 与方差估计量 s^2_{CLS} 是独立的. 于是 t 统计量

$$T=\frac{\tilde{\beta}_k-\beta_k}{\operatorname{SE}\left(\tilde{\beta}_k\right)}=\frac{\left(\tilde{\beta}_k-\beta_k\right) / \sqrt{\sigma^2\left[\left(\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)\right]_{kk}}}{\sqrt{\frac{\left(n-K+q\right) s^2_{\mathrm{CLS}} / \sigma^2}{n-K+q}}} \sim t\left(n-K+q\right).$$

定理 6.2.4 (正态假设下的 CLS 估计量). *在予以线性约束式 (6.1) 下, 线性回归模型 (3.12.1) 有 CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$, 则有*

$$\begin{aligned}
&\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta} \middle| \boldsymbol{X} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}^0_{\tilde{\boldsymbol{\beta}}}\right), \\
&\frac{\left(n-K+q\right) s^2_{\mathrm{CLS}}}{\sigma^2} \middle| \boldsymbol{X} \sim \chi^2\left(n-K+q\right), \\
&T \middle| \boldsymbol{X}=\frac{\tilde{\beta}_k-\beta_k}{\operatorname{SE}\left(\tilde{\beta}_k\right)} \middle| \boldsymbol{X} \sim t\left(n-K+q\right),
\end{aligned}$$

且 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 和 $\tilde{\boldsymbol{\varepsilon}}$ 是独立的.

我们知道, 在同方差假设式 (3.7) 下 Guass-Markov 指出 OLS 估计量是线性估计量中最有效, 即任意线性估计量的协方差矩阵与 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} \sigma^2$ 的差都是半正定的.

然而, 当我们对于总体参数 $\boldsymbol{\beta}$ 施加线性约束式 (6.1) 后, 同方差下 CLS 估计量有协方差矩阵式 (6.18), 直觉上对比,

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{0}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^{2},\quad \boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}^{0}=\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{A}\right)\sigma^{2},$$

似乎 OLS 估计量的条件协方差矩阵 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{0}$ 要“大于”CLS 估计量的条件协方差矩阵 $\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}^{0}$. 事实上,

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{0}-\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}^{0}=\boldsymbol{A}\sigma^{2}\geqslant 0,$$

由实对称方阵 $\boldsymbol{A}$ 是半正定矩阵可以推知 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{0}-\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}^{0}$ 是半正定的, 也就是说, 当参数 $\boldsymbol{\beta}$ 满足线性约束式 (6.1) 时,CLS 估计量比 OLS 估计量更有效.

有趣的是, 我们可以进一步的计算 OLS 估计量与 CLS 估计量的条件协方差矩阵, 由于 $\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}=\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$, 则为了计算简便, 我们先计算

$$\begin{aligned}
&\operatorname{Cov}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)\\
=&\mathbb{E}\left(\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\mathbb{E}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)\right)\cdot\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\mathbb{E}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)\right)^{\mathrm{T}}\middle|\boldsymbol{X}\right)\\
=&\mathbb{E}\left(\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)\cdot\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{X}\boldsymbol{A}\right)\middle|\boldsymbol{X}\right)\\
=&\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{X}\boldsymbol{A}\right)=\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}-\boldsymbol{X}\boldsymbol{A}\right)\sigma^{2}\\
=&\left(\boldsymbol{A}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\right)\sigma^{2}=\boldsymbol{O}_{K},
\end{aligned}$$

即两估计量之差 $\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 与 CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 是完全不相关的, 这在正态假设下即意味着两者是独立的. 而依据协方差矩阵的性质, 则有

$$\begin{aligned}
&\operatorname{Cov}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\operatorname{Cov}\left(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)-\operatorname{Cov}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)\\
=&\operatorname{Cov}\left(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)-\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\operatorname{Cov}\left(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)-\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}=\boldsymbol{O}_{K},
\end{aligned}$$

即得 OLS 估计量与 CLS 估计量的条件协方差矩阵为

$$\operatorname{Cov}\left(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}}.\tag{6.20}$$

式 (6.20) 的结果非常有趣, 两个估计量的条件协方差矩阵等于了更为有效的 CLS 估计量的条件协方差矩阵. 据此, 我们计算两估计量之差 $\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$ 的条件协方差矩阵为

$$\begin{aligned}
&\operatorname{Var}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)+\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)-2\operatorname{Cov}\left(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)\\
=&\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)+\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)-2\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)-\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right),
\end{aligned}$$

又同方差下

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\operatorname{Var}\left(\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\operatorname{Var}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)\boldsymbol{X}\boldsymbol{A}=\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\sigma^{2}=\boldsymbol{A}\sigma^{2},$$

故有

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)-\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\middle|\boldsymbol{X}\right)=\boldsymbol{A}\sigma^2\geqslant 0 \tag{6.21}$$

成立. 式 (6.21) 被称为 **Hausman 等式** (Hausman equality), 其表明估计量之差的方差等于估计量的方差之差, 这主要出现在比较有效估计量与非有效估计量时.

6.3 受限回归下最小距离估计量 (MDE) 及其渐进有效性

上一节我们发现如果回归模型确实具有线性约束, 且在满足同方差假设下,CLS 估计量比 OLS 估计量是更为有效的. 一个自然的问题便是还有比 CLS 估计量更为有效的估计量吗? 本节我们研究最小距离估计量, 并通过其渐进性质来回答这个问题.

定义 6.3.1 (最小距离估计量)**.** *对于线性投影模型 (2.2.3), 设其 OLS 估计量为 $\hat{\boldsymbol{\beta}}$. 对于模型施加线性约束条件式 (6.1) 后, 记*

$$\tilde{\boldsymbol{\beta}}_{\mathrm{md}}=\underset{\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}}\, J\left(\boldsymbol{\beta}\right)=\underset{\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}}\left[n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\hat{\boldsymbol{W}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right], \tag{6.22}$$

*我们称 $\tilde{\boldsymbol{\beta}}_{\mathrm{md}}$ 为***最小距离估计量** *(minimum distance estimator), 简称为 **MD 估计量**. 式 (6.22) 中*

$$J\left(\boldsymbol{\beta}\right)=n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\hat{\boldsymbol{W}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \tag{6.23}$$

*是有由 K 阶正定权重矩阵 $\hat{\boldsymbol{W}}$ 所定义的***准则函数** *(criterion function).*

准则函数式 (6.23) 是一个二次型, 其可以解释为加权的 Euclidean 距离的平方, 即是度量了 $\boldsymbol{\beta}$ 与 OLS 估计量的距离. 这样,MD 估计量式 (6.22) 便是使得距离非约束的 OLS 估计量数值最近的估计量, 且 $\tilde{\boldsymbol{\beta}}_{\mathrm{md}}$ 满足有线性约束式 (6.1). 由于权重矩阵 $\hat{\boldsymbol{W}}$ 是预先给定的, 故 MD 估计量可以具有不同形式, 其依赖于权重矩阵 $\tilde{\boldsymbol{\beta}}_{\mathrm{md}}$.

特别地, 当权重矩阵满足有

$$\hat{\boldsymbol{W}}=\hat{\boldsymbol{Q}}_{\mathbf{XX}}=\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}},$$

我们记

$$J^0\left(\boldsymbol{\beta}\right)=n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right), \tag{6.24}$$

则有 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}=\underset{\boldsymbol{R\beta}=\boldsymbol{c}}{\operatorname{argmin}} J^0(\boldsymbol{\beta})$. 这是因为 CLS 估计量对应的 SSE 满足有

$$
\begin{aligned}
&\mathrm{SSE}(\boldsymbol{\beta})=\sum_{i=1}^{n}\left(Y_i-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2=\sum_{i=1}^{n}\left(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}+\hat{\varepsilon}_i-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2=\sum_{i=1}^{n}\left(\mathbf{X}_i^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\hat{\varepsilon}_i\right)^2\\
=&\sum_{i=1}^{n}\left(\mathbf{X}_i^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\hat{\varepsilon}_i\right)^2=\left[\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\hat{\boldsymbol{\varepsilon}}\right]^{\mathrm{T}}\left[\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\hat{\boldsymbol{\varepsilon}}\right]\\
=&\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\underbrace{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)}_{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\boldsymbol{X}=\mathbf{0}}+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}\\
=&n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\mathrm{RSS}\\
=&J^0(\boldsymbol{\beta})+\mathrm{RSS},
\end{aligned}
$$

即

$$
\mathrm{SSE}(\boldsymbol{\beta})=J^0(\boldsymbol{\beta})+\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}\right)=J^0(\boldsymbol{\beta})+\mathrm{RSS}. \tag{6.25}
$$

式 (6.25) 表明, 误差平方和 SSE 可以分解为准则函数 $J^0(\boldsymbol{\beta})$ 与 OLS 估计量的残差平方和 RSS 的和.OLS 估计量是无约束条件下给定的结果, 于是在线性约束式 (6.1) 下最优化问题定义的 CLS 估计量式 (6.2) 与 $\underset{\boldsymbol{R\beta}=\boldsymbol{c}}{\operatorname{argmin}} J^0(\boldsymbol{\beta})$ 是等价的, 故 CLS 估计量是最小距离估计量的一种特例情形.

我们现在来求解具有权重矩阵 $\hat{\boldsymbol{W}}$ 的 MD 估计量的具体形式, 我们同样适用 Lagrange 乘数法. 设 Lagrange 函数为

$$
(\boldsymbol{\beta},\boldsymbol{\lambda})=\frac{1}{2}J(\boldsymbol{\beta})+\boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{R\beta}-\boldsymbol{c}), \tag{6.26}
$$

注意到求解 CLS 估计量时式 (6.3) 中 SSE 为

$$
\mathrm{SSE}(\boldsymbol{\beta})=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-2\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}+(\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta},
$$

由于

$$
J(\boldsymbol{\beta})=n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\mathrm{T}}\hat{\boldsymbol{W}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=n\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\beta}}-2\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{\beta}+\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{\beta}\right),
$$

则对比最优化问题的系数有

$$
\begin{cases}
n\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{W}}\hat{\boldsymbol{\beta}}=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y},\\
n\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{W}}=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X},\\
n\hat{\boldsymbol{W}}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X},
\end{cases}
$$

于是我们可以直接由 CLS 估计量的求解过程得到 MD 估计量, 只需将式 (6.6) 与式 (6.7) 中相同系数替换, 得到

$$\tilde{\boldsymbol{\lambda}}_{\mathrm{md}} = n\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right), \tag{6.27}$$

$$\tilde{\boldsymbol{\beta}}_{\mathrm{md}} = \hat{\boldsymbol{\beta}} - n^{-1}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}_{\mathrm{CLS}} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right). \tag{6.28}$$

由于 MD 估计量依赖于权重矩阵 $\hat{\boldsymbol{W}}$, 一个自然的问题便是该选择什么作为 $\hat{\boldsymbol{W}}$? 我们将通过渐进理论来阐述该问题. 首先我们需要如下假设.

假设 6.3.1 (MD 估计量渐进性质的假设). *对于形如式 (6.1) 的线性约束和式 (6.28) 的 MD 估计量, 假设其满足如下假设:*

(a) 线性约束 $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{c}$ 满足 $\mathrm{rank}\boldsymbol{R}_{q\times K} = q$;

(b) K 阶正定的权重矩阵 $\hat{\boldsymbol{W}}$ 满足有 $\hat{\boldsymbol{W}} \xrightarrow{P} \boldsymbol{W} > 0$.

现在我们来研究 MD 估计量的大样本性质. 首先 MD 估计量是相合估计, 这是因为依据连续映射定理 (4.2.7) 当 $n \to \infty$ 时有

$$\tilde{\boldsymbol{\beta}}_{\mathrm{md}} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right) \xrightarrow{P} \boldsymbol{\beta} - \boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\boldsymbol{\beta} - \mathbf{c}\right),$$

而参数 $\boldsymbol{\beta}$ 满足约束条件式 (6.9), 故有 $\tilde{\boldsymbol{\beta}}_{\mathrm{md}} \xrightarrow{P} \boldsymbol{\beta}$. 由于 CLS 估计量式 $\hat{\boldsymbol{W}} = \hat{\boldsymbol{Q}_{\mathbf{XX}}}$ 的特殊情形, 故当 $n \to \infty$ 时有 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}} \xrightarrow{P} \boldsymbol{\beta}$.

现在我们来研究 MD 估计量的渐进分布. 同样地, 我们先计算 MD 估计量的抽样误差, 仿照证明式 (6.11) 的过程, 将 MD 估计量的抽样误差分解为 $\tilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta} = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{md}}\right)$, 而由式 (6.28) 和式 (6.9) 有

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{md}} =& \hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)\\ =& \hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right),\end{aligned} \tag{6.29}$$

故有

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta} &= \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\\ &= \left[\boldsymbol{I} - \hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\right]\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \boxed{\hat{\boldsymbol{B}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)},\end{aligned} \tag{6.30}$$

其中 $\hat{\boldsymbol{B}} = \boldsymbol{I} - \hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}$. 式 (6.30) 即是说明,MD 估计量的抽样误差 $\tilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta}$ 是 OLS 估计量的抽样误差 $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ 的线性组合. 由假设 (6.3.1) 的 (b), 根据连续

映射定理 (4.2.7) 有[1]

$$\hat{\boldsymbol{B}}=\boldsymbol{I}-\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\xrightarrow{P}\boldsymbol{B}=\boldsymbol{I}-\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}.$$

注意到当 $n\to\infty$ 时,OLS 估计量具有渐进分布

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{P}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right),\quad \boldsymbol{V}_{\boldsymbol{\beta}}=\boldsymbol{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\mathbf{XX}}^{-1},$$

故依据连续映射定理 (4.4.6) 得 MD 估计量有

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{md}}-\boldsymbol{\beta}\right)=\hat{\boldsymbol{B}}\cdot\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}\boldsymbol{B}\cdot\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right)=\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\left(\boldsymbol{W}\right)\right),$$

其中

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{\beta}}\left(\boldsymbol{W}\right)=\boldsymbol{B}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{B}^{\mathrm{T}}\\
=&\left[\boldsymbol{I}-\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\right]\boldsymbol{V}_{\boldsymbol{\beta}}\left[\boldsymbol{I}-\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\right]^{\mathrm{T}}\\
=&\left[\boldsymbol{V}_{\boldsymbol{\beta}}-\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\right]\left[\boldsymbol{I}-\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{W}^{-1}\right]\\
=&\boldsymbol{V}_{\boldsymbol{\beta}}-\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{W}^{-1}-\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\\
&+\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{W}^{-1}.
\end{aligned}$$

特别地,CLS 估计量满足有 $\hat{\boldsymbol{W}}=\hat{\boldsymbol{Q}}_{\mathbf{XX}}=n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, 故 $n\to\infty$ 时 CLS 的渐进分布为 $\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\mathrm{CLS}}\right)$, 其中

$$\begin{aligned}
&\boldsymbol{B}=\boldsymbol{I}-\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\\
=&\boldsymbol{I}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\left[\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right]^{-1}\boldsymbol{R}=\boldsymbol{I}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}
\end{aligned}$$

及

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{\beta}}\left(\boldsymbol{Q}_{\mathbf{XX}}\right)=\boldsymbol{V}_{\mathrm{CLS}}=\left(\boldsymbol{I}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\boldsymbol{V}_{\boldsymbol{\beta}}\left(\boldsymbol{I}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{\mathrm{T}}\\
=&\left(\boldsymbol{V}_{\boldsymbol{\beta}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V}_{\boldsymbol{\beta}}\right)\left(\boldsymbol{I}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}\right)\\
=&\boldsymbol{V}_{\boldsymbol{\beta}}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V}_{\boldsymbol{\beta}}-\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}.
\end{aligned}$$

定理 6.3.1 (MD 估计量的渐进分布)**.** *在假设 (5.1.1) 和假设 (6.3.1) 下, 对于 MD 估计量式 (6.28) 当 $n\to\infty$ 时, 其具有渐进分布*

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{md}}-\boldsymbol{\beta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\left(\boldsymbol{W}\right)\right),\tag{6.31}$$

[1]这里是随机矩阵, 似乎不应该直接使用... 后续也是..XD

其中 $\boldsymbol{B}=\boldsymbol{I}-\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}$, 而 *MD* 估计量的渐进协方差矩阵为

$$\begin{aligned}&\boldsymbol{V}_{\beta}(\boldsymbol{W})=\boldsymbol{B}\boldsymbol{V}_{\beta}\boldsymbol{B}^{\mathrm{T}}\\=&\boldsymbol{V}_{\beta}-\boldsymbol{V}_{\beta}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{W}^{-1}-\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\beta}\\&+\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\beta}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{W}^{-1}.\end{aligned}\tag{6.32}$$

特别地, 当 $\hat{\boldsymbol{W}}=\hat{\boldsymbol{Q}}_{\mathbf{XX}}=n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 时 *MD* 估计量即为 *CLS* 估计量. 当 $n\to\infty$ 时 *CLS* 估计量具有渐进分布

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}-\boldsymbol{\beta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\mathrm{CLS}}\right),\tag{6.33}$$

其中 $\boldsymbol{B}=\boldsymbol{I}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$,*CLS* 估计量的渐进协方差矩阵为

$$\boldsymbol{V}_{\mathrm{CLS}}=\boldsymbol{V}_{\beta}-\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V}_{\beta}-\boldsymbol{V}_{\beta}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V}_{\beta}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}.\tag{6.34}$$

上述的 $\boldsymbol{V}_{\beta}$ 为 *OLS* 估计量的渐进协方差矩阵.

有了 MD 估计量的渐进分布, 我们现在便可以回归在线性约束式 (6.1) 下对于回归系数 $\boldsymbol{\beta}$ 的估计何时是最有效的. 直觉而言, 如果依赖于权重矩阵 $\hat{\boldsymbol{W}}$ 的 MD 估计量的渐进协方差矩阵具有是“最小”的, 那么这应该是 MD 估计量是最有效的充分条件. 对于 $\boldsymbol{V}_{\beta}(\boldsymbol{W})=\boldsymbol{B}\boldsymbol{V}_{\beta}\boldsymbol{B}^{\mathrm{T}}$, 可以证明[2], 当 $\hat{\boldsymbol{W}}=\hat{\boldsymbol{V}}_{\beta}^{-1},\boldsymbol{W}=\boldsymbol{V}_{\beta}^{-1}$ 时, 即有最小渐进协方差矩阵的 MD 估计量.

定义 6.3.2 (有效最小距离估计量)**.** 对于 *MD* 估计量式 *(6.28)*, 当权重矩阵满足有当 $n\to\infty$ 时,$\hat{\boldsymbol{W}}=\hat{\boldsymbol{V}}_{\beta}^{-1}\xrightarrow{P}\boldsymbol{W}=\boldsymbol{V}_{\beta}^{-1}$, 我们称此时的 *MD* 估计量为**有效最小距离估计量** *(efficient minimum distance estimator)*, 简称为 ***EMD* 估计量**. 我们记 *EMD* 估计量为

$$\tilde{\boldsymbol{\beta}}_{\mathrm{emd}}=\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{V}}_{\beta}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\beta}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\mathbf{c}\right),\tag{6.35}$$

其渐进协方差矩阵为

$$\boldsymbol{V}_{\boldsymbol{\beta},\mathrm{emd}}=\boldsymbol{V}_{\beta}-\boldsymbol{V}_{\beta}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\beta}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\beta},\tag{6.36}$$

其中 $\boldsymbol{V}_{\beta}$ 为 *OLS* 估计量的渐进协方差矩阵.

如果回归模型满足有线性约束式 (6.1), 那么无论模型是否满足同方差假设,OLS 估

[2] 见 Bruce(2021)p.217.

计量或 MD 估计量的渐进协方差矩阵总满足有

$$\boldsymbol{V}_{\boldsymbol{\beta},\mathrm{emd}} \leqslant \boldsymbol{V}_{\boldsymbol{\beta}}, \quad \boldsymbol{V}_{\boldsymbol{\beta},\mathrm{emd}} \leqslant \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}),$$

这里的 $\leqslant$ 是表示二次型之差为半正定的. 也就是说, 对于线性约束式 (6.1) 的受限回归,EMD 是最有效的. 一种特殊的情况是, 当模型满足同方差时, 依据式 (5.11) 有

$$\boldsymbol{W} = \left(\boldsymbol{V}_{\beta}^{0}\right)^{-1} = \left(\boldsymbol{Q}_{\mathbf{XX}}^{-1}\sigma^{2}\right)^{-1} = \boldsymbol{Q}_{\mathbf{XX}}\sigma^{-2}, \quad \hat{\boldsymbol{W}} = \left(\hat{\boldsymbol{V}}_{\beta}^{0}\right)^{-1} = \left(\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}s^{2}\right)^{-1} = \hat{\boldsymbol{Q}}_{\mathbf{XX}}s^{-2},$$

则 EMD 估计量有

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_{\mathrm{emd}} &= \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}s^{-2}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}s^{-2}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)\\ &= \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right) = \tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\end{aligned}$$

即形如式 (6.8) 的 CLS 估计量是模型总体同方差时的有效估计量.

在实证中,EMD 估计量的渐进协方差矩阵的一种常用估计量为

$$\tilde{\boldsymbol{V}}_{\boldsymbol{\beta},\mathrm{emd}} = \tilde{\boldsymbol{V}}_{\boldsymbol{\beta}} - \tilde{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\tilde{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\tilde{\boldsymbol{V}}_{\boldsymbol{\beta}},$$

其中

$$\tilde{\boldsymbol{V}}_{\boldsymbol{\beta}} = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\tilde{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}, \quad \tilde{\boldsymbol{\Omega}} = \frac{1}{n-K+q}\sum_{i=1}^{n}\mathbf{X}_{i}\mathbf{X}_{i}^{\mathrm{T}}\tilde{\varepsilon}_{i}^{2}, \quad \tilde{\varepsilon}_{i} = Y_{i} - \mathbf{X}_{i}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{\mathrm{emd}}$$

考虑了调整的自由度为 $n-(K-q)$.

最后我们介绍渐进理论下的 Hausman 等式. 我们由式 (6.29) 可以得到

$$\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{emd}} = \hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = \hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right),$$

而依据参数满足约束条件式 (6.9), 则有

$$\sqrt{n}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right) = \sqrt{n}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c} - \boldsymbol{R}\boldsymbol{\beta} - \mathbf{c}\right) \xrightarrow{\mathrm{P}} \mathrm{N}\left(\mathbf{0}, \boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right),$$

故依据连续映射定理 (4.4.6) 得

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{emd}}\right) =& \hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\cdot\sqrt{n}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)\\ \xrightarrow{\mathrm{P}}& \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\cdot\mathrm{N}\left(\mathbf{0}, \boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right),\end{aligned}$$

即 $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{emd}}$ 是渐近正态的, 且有渐进协方差矩阵

$$\begin{aligned}\mathrm{Avar}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{emd}}\right) =& \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\\ =& \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}.\end{aligned}$$

这样, 我们便可发现 $\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\text{emd}}$、OLS 估计量 $\hat{\boldsymbol{\beta}}$ 和 EMD 估计量 $\tilde{\boldsymbol{\beta}}_{\text{emd}}$ 都满足有渐近正态性, 当 $n\to\infty$ 时

$$\begin{aligned}&\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\text{emd}}\right)\xrightarrow{\text{P}}\text{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\text{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\text{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\right),\\&\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{\text{P}}\text{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right),\quad\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_{\text{emd}}-\boldsymbol{\beta}\right)\xrightarrow{\text{P}}\text{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta},\text{emd}}\right),\end{aligned}$$

且有

$$\begin{aligned}&\text{Avar}\left(\hat{\boldsymbol{\beta}}\right)-\text{Avar}\left(\tilde{\boldsymbol{\beta}}_{\text{emd}}\right)=\boldsymbol{V}_{\boldsymbol{\beta}}-\boldsymbol{V}_{\boldsymbol{\beta},\text{emd}}\\=&\boldsymbol{V}_{\boldsymbol{\beta}}-\left[\boldsymbol{V}_{\boldsymbol{\beta}}-\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\text{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\text{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\right]=\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\text{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\text{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}},\end{aligned}$$

即有

$$\text{Avar}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\text{emd}}\right)=\text{Avar}\left(\hat{\boldsymbol{\beta}}\right)-\text{Avar}\left(\tilde{\boldsymbol{\beta}}_{\text{emd}}\right),\tag{6.37}$$

我们称式 (6.37) 为 **Hausman 等式** (Hausman equality). 式 (6.37) 的含义是一个估计量与一个有效估计量之差的渐近协方差矩阵等于两者的渐近协方差矩阵之差.

6.4 非线性约束的受限回归模型

6.5 极大似然估计

本节将介绍使用极大似然估计估计取解决回归系数 $\boldsymbol{\beta}$ 的估计问题, 并介绍信息矩阵的相关内容.

极大似然估计 (maximum likelihood estimation)是一种已知总体的概率分布的参数估计方法, 其依据样本, 去推断已知形式的概率分布的参数. 对于计量经济学而言, 我们研究正态回归模型 (3.12.1) 的极大似然估计. 由总体的正态扰动项式 (3.64) 可以推知被解释变量 Y 的总体满足条件分布

$$\text{Y}|\,\mathbf{X}\sim\text{N}\left(\mathbf{X}^{\text{T}}\boldsymbol{\beta},\sigma^2\right)$$

我们便是使用极大似然估计去估计 $\text{N}\left(\mathbf{X}^{\text{T}}\boldsymbol{\beta},\sigma^2\right)$ 中的参数 $\boldsymbol{\beta}$ 和 σ^2.

对于第 i 个样本个体而言 $(i=1,\cdots,n)$, 则有其满足有 $Y_i|\,\mathbf{X}_i\sim\text{N}\left(\mathbf{X}_i^{\text{T}}\boldsymbol{\beta},\sigma^2\right)$, 于是 Y_i 的条件密度函数为

$$f\left(y|\,\mathbf{X}\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(y-\mathbf{X}^{\text{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}},$$

故**似然函数** (likelihood function)为

$$\begin{aligned}L\left(\boldsymbol{\beta},\sigma^2\right)=&\prod_{i=1}^{n}f\left(y_i|\,\mathbf{X}_i\right)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(y_i-\mathbf{X}_i^{\text{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}}\\=&\left(2\pi\sigma^2\right)^{-\frac{n}{2}}\exp\left(\sum_{i=1}^{n}-\frac{\left(y_i-\mathbf{X}_i^{\text{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right),\end{aligned}\tag{6.38}$$

则**对数似然函数** (log-likelihood function)为

$$\ln L\left(\boldsymbol{\beta},\sigma^2\right)=-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2=-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\mathrm{SSE}\left(\boldsymbol{\beta}\right), \tag{6.39}$$

这里我们使用误差平方和 SSE, 这会帮助我们简化后续的推导.

由于正态假设下总体 Y 具有连续分布, 依据式 (6.39) 即定义

$$\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)=\underset{\boldsymbol{\beta}\in\mathbb{R}^{K\times1},\sigma^2>0}{\operatorname{argmin}}\ln L\left(\boldsymbol{\beta},\sigma^2\right) \tag{6.40}$$

为**极大似然估计量** (maximum likelihood estimator), 简称为 **ML 估计量**. ME 估计量是使得式 (6.39) 和式 (6.38) 取得最大值的估计量, 则其有是极大值点的必要条件. 我们对对数似然函数式 (6.39) 求导[3], 得到一阶条件

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ln L\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)=-\frac{1}{2\hat{\sigma}^2_{\mathrm{ML}}}\cdot\frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)=0, \tag{6.41}$$

$$\frac{\partial}{\partial\sigma^2}\ln L\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)=-\frac{n}{2\hat{\sigma}^2_{\mathrm{ML}}}+\frac{\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)}{2\left(\hat{\sigma}^2_{\mathrm{ML}}\right)^2}=0. \tag{6.42}$$

依据式 (6.41) 即可解出 $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, 继而再由式 (6.42) 解出 $\hat{\sigma}^2_{\mathrm{ML}}$. 不难发现, 式 (6.41) 即是 OLS 估计量推导过程中的一阶条件 (使得 SSE 最小的极值点), 则可以立刻得到

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}}=\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{Y}_i\right)=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}, \tag{6.43}$$

在将式 (6.43) 代入一阶条件式 (6.42), 解出

$$\hat{\sigma}^2_{\mathrm{ML}}=\frac{\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)}{n}=\frac{\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)}{n}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right)^2=\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{n}. \tag{6.44}$$

这里的残差 $\hat{\varepsilon}_i$ 即是 OLS 估计量的残差, 也是 ML 估计量的残差. 现在我们要确保式 (6.43) 和式 (6.44) 的极大值点, 我们计算对数似然函数的 Hessian 矩阵. 计算二阶偏导数有

$$\begin{aligned}
\frac{\partial^2}{\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\boldsymbol{\beta}}\ln L\left(\boldsymbol{\beta},\sigma^2\right)&=-\frac{1}{2\sigma^2}\cdot\frac{\partial}{\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\boldsymbol{\beta}\right),\\
\frac{\partial^2}{\partial\sigma^2\partial\boldsymbol{\beta}}\ln L\left(\boldsymbol{\beta},\sigma^2\right)&=\frac{1}{2\left(\sigma^2\right)^2}\cdot\frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\boldsymbol{\beta}\right),\\
\frac{\partial^2}{\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\sigma^2}\ln L\left(\boldsymbol{\beta},\sigma^2\right)&=\frac{1}{2\left(\sigma^2\right)^2}\cdot\left(\frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\boldsymbol{\beta}\right)\right)^{\mathrm{T}},\\
\frac{\partial^2}{\partial\left(\sigma^2\right)^2}\ln L\left(\boldsymbol{\beta},\sigma^2\right)&=\frac{n}{2\left(\sigma^2\right)^2}-\frac{\mathrm{SSE}\left(\boldsymbol{\beta}\right)}{\left(\sigma^2\right)^3},
\end{aligned}$$

[3]这里的矩阵微积分参见附录 ()

其中

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathrm{SSE}\left(\boldsymbol{\beta}\right)=-2\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right),\quad \frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}\partial \boldsymbol{\beta}}\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)=2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X},$$

故有

$$\begin{aligned}\boldsymbol{H}\left(\boldsymbol{\beta},\sigma^2\right)&=\begin{pmatrix}\dfrac{\partial^2}{\partial \boldsymbol{\beta}^{\mathrm{T}}\partial \boldsymbol{\beta}}\ln L\left(\boldsymbol{\beta},\sigma^2\right) & \dfrac{\partial^2}{\partial \sigma^2\partial \boldsymbol{\beta}}\ln L\left(\boldsymbol{\beta},\sigma^2\right)\\ \dfrac{\partial^2}{\partial \boldsymbol{\beta}\partial \sigma^2}\ln L\left(\boldsymbol{\beta},\sigma^2\right) & \dfrac{\partial^2}{\partial \sigma^2\partial \sigma^2}\ln L\left(\boldsymbol{\beta},\sigma^2\right)\end{pmatrix}\\ &=\begin{pmatrix}\dfrac{-1}{\sigma^2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & -\dfrac{1}{\left(\sigma^2\right)^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\boldsymbol{X}\\ -\dfrac{1}{\left(\sigma^2\right)^2}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right) & \dfrac{n}{2\left(\sigma^2\right)^2}-\dfrac{\mathrm{SSE}\left(\boldsymbol{\beta}\right)}{\left(\sigma^2\right)^3}\end{pmatrix},\end{aligned}\tag{6.45}$$

将 ML 估计量代入式 (6.45) 即有

$$\boldsymbol{H}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)=\begin{pmatrix}\dfrac{-1}{\hat{\sigma}^2_{\mathrm{ML}}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{0}_{K\times 1}\\ \boldsymbol{0}_{1\times K} & \dfrac{-n}{2\left(\sigma^2\right)^2}\end{pmatrix},$$

依据假设式 (3.6) 实对称方阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 是可逆的, 故是正定的, 于是 $-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, 由此推知 Hessian 矩阵 $\boldsymbol{H}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)$ 是负定的, 故 $\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)$ 是式 (6.39) 的极大值点. 又 $\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)$ 是唯一的极大值点, 故 $\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)$ 是便是对数似然函数的最大值点, 式 (6.43) 和式 (6.44) 便是极值问题式 (6.40) 的解.

对回归系数 $\boldsymbol{\beta}$ 的 ML 估计量与 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 相同, 这并不意外. 注意到对数似然函数式 (6.39) 包含有负号的 $\ln\sigma$ 与 $\mathrm{SSE}\left(\boldsymbol{\beta}\right)$, 则为了使之最大便有充分条件 $\mathrm{SSE}\left(\boldsymbol{\beta}\right)$ 最小, 这与 OLS 估计量的思路是一致的. 需要强调, 似然函数依赖于已知的总体分布——即正态假设式 (3.64)——只有在正态回归模型 (3.12.1) 下 ML 估计量等于 OLS 估计量. 不满足正态假设便没有已知的总体分布,ML 估计量也无从谈起.

我们将 ML 估计量代入对数似然函数式 (6.39) 中, 即有式 (6.39) 的最小值

$$\ln L\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)=-\frac{n}{2}\ln\left(2\pi\hat{\sigma}^2_{\mathrm{ML}}\right)-\frac{n}{2},\tag{6.46}$$

式 (6.46) 通常被用于汇报拟合的度量. 越小越好?

回归模型的极大似然估计需要给定正态假设, 这较 OLS 估计量而言显然要求更为严格, 但由于利用了已知分布的信息, 相应地 ML 估计量具有较 OLS 估计量更为优秀的渐进性质. 我们现在研究 $\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)$ 的统计性质.

由于形式上 ML 估计量与 OLS 估计量是一致的, 其有限样本性质可以之间参见第三章, 大样本性质亦满足第五章的推导. 我们下面介绍极大似然估计独有的渐进性质.

对于连续性随机变量的 LS 估计量, 我们研究对数似然函数式 (6.39) 的梯度向量 (也即偏导数、Jacobi 矩阵的转置), 对于正态回归模型 (3.12.1) 即有

$$\begin{aligned}\boldsymbol{S} =&\nabla \ln L\left(\boldsymbol{\beta}, \sigma^2\right)=\begin{pmatrix} \dfrac{\partial}{\partial \boldsymbol{\beta}} \ln L\left(\boldsymbol{\beta}, \sigma^2\right) \\ \dfrac{\partial}{\partial \sigma^2} \ln L\left(\boldsymbol{\beta}, \sigma^2\right) \end{pmatrix} \\ =&\begin{pmatrix} -\dfrac{1}{2\sigma^2} \cdot \dfrac{\partial}{\partial \boldsymbol{\beta}} \mathrm{SSE}(\boldsymbol{\beta}) \\ -\dfrac{n}{2\sigma^2}+\dfrac{\mathrm{SSE}(\boldsymbol{\beta})}{2\left(\sigma^2\right)^2} \end{pmatrix}=\begin{pmatrix} \dfrac{1}{\sigma^2} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\varepsilon} \\ -\dfrac{n}{2\sigma^2}+\dfrac{\boldsymbol{\varepsilon}^{\mathrm{T}} \boldsymbol{\varepsilon}}{2\left(\sigma^2\right)^2} \end{pmatrix},\end{aligned} \tag{6.47}$$

我们称 $\boldsymbol{S}$ 为**有效得分** (efficient score), 其在与 ML 估计量有关的渐进分布和假设检验中有重要的用途[4]. 有效得分 $\boldsymbol{S}$ 是依赖于样本的随机变量, 一个重要性质便是其期望值为零向量, 我们对式 (6.47) 取给定 $\boldsymbol{X}$ 的条件期望, 即有

$$\begin{aligned}\mathbb{E}(\boldsymbol{S} | \boldsymbol{X})=&\begin{pmatrix} \mathbb{E}\left(\dfrac{1}{\sigma^2} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\varepsilon} \middle| \boldsymbol{X}\right) \\ \mathbb{E}\left(-\dfrac{n}{2\sigma^2}-\dfrac{\boldsymbol{\varepsilon}^{\mathrm{T}} \boldsymbol{\varepsilon}}{2\left(\sigma^2\right)^2} \middle| \boldsymbol{X}\right) \end{pmatrix}=\begin{pmatrix} \dfrac{1}{\sigma^2} \boldsymbol{X}^{\mathrm{T}} \mathbb{E}(\boldsymbol{\varepsilon} | \boldsymbol{X}) \\ -\dfrac{n}{2\sigma^2}+\dfrac{\mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}} \boldsymbol{\varepsilon} \middle| \boldsymbol{X}\right)}{2\left(\sigma^2\right)^2} \end{pmatrix} \\ =&\begin{pmatrix} \dfrac{1}{\sigma^2} \boldsymbol{X}^{\mathrm{T}} \mathbf{0} \\ -\dfrac{n}{2\sigma^2}+\dfrac{n\sigma^2}{2\left(\sigma^2\right)^2} \end{pmatrix}=\mathbf{0}_{(K+1)\times 1},\end{aligned}$$

则依据定理 (2.1.1) 知 $\mathbb{E}\boldsymbol{S}=\mathbf{0}$.

在数理统计中, 利用有效得分 $\boldsymbol{S}$, 我们定义

$$\mathscr{I}=\mathbb{E}\left(\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}\right), \tag{6.48}$$

式 (6.48) 被称为 **Fisher 信息矩阵** (Fisher information matrix), 简称**信息矩阵** (information matrix), 这里的符号 $\mathscr{I}$ 是花体的大写字母 I. 再定义

$$\mathscr{H}=-\mathbb{E}\left(\boldsymbol{H}\left(\boldsymbol{\beta}, \sigma^2\right)\right) \tag{6.49}$$

为**期望 Hessian 矩阵** (expected Hessian matrix), 其中符号 $\mathscr{H}$ 为花体大写字母 H. 信息矩阵 $\mathscr{I}$ 即是有效得分 $\boldsymbol{S}$ 的条件协方差, 而期望 Hessian 矩阵 $\mathscr{H}$ 可以视为对数似然函数的二阶导数的平均变化情形. 在模型是设定正确[5]以及样本 $\boldsymbol{Y}$ 不依赖于未知的总

[4]严格的说, 有效得分是对数似然函数在未知参数的真实值处的梯度. 本笔记中涉及参数的真实值时使用的即是参数符号本身, 在此特别声明.

[5]设定正确 (correctly specified) 是指对于独立同分布的 Y_i, 其共有的条件密度函数 $f(y|\mathbf{X}) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{\left(y-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}\right]$ 只取决于**唯**一的未知的总体参数.

体参数 $(\boldsymbol{\beta},\sigma^2)$ 时, 则有 $\mathscr{I}=\mathscr{H}$ 成立. 这样, 对于 ML 估计量, 依据式 (6.45) 我们计算给定条件 $\boldsymbol{X}$ 下的条件信息矩阵 $\mathscr{I}_{\boldsymbol{X}}$ 为

$$\begin{aligned}
\mathscr{I}_{\boldsymbol{X}} &= \mathscr{H}_{\boldsymbol{X}} = -\mathbb{E}\left(\boldsymbol{H}\left(\boldsymbol{\beta},\sigma^2\right)\middle|\boldsymbol{X}\right) \\
&= \begin{pmatrix} \mathbb{E}\left(\frac{1}{\sigma^2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\middle|\boldsymbol{X}\right) & \mathbb{E}\left(\frac{1}{\left(\sigma^2\right)^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\boldsymbol{X}\middle|\boldsymbol{X}\right) \\ \mathbb{E}\left(\frac{1}{\left(\sigma^2\right)^2}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\middle|\boldsymbol{X}\right) & \mathbb{E}\left(\frac{\mathrm{SSE}\left(\boldsymbol{\beta}\right)}{\left(\sigma^2\right)^3}-\frac{n}{2\left(\sigma^2\right)^2}\middle|\boldsymbol{X}\right) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \frac{1}{\left(\sigma^2\right)^2}\mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\boldsymbol{X} \\ \frac{1}{\left(\sigma^2\right)^2}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right) & \frac{\mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)}{\left(\sigma^2\right)^3}-\frac{n}{2\left(\sigma^2\right)^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{0}_{K\times 1} \\ \boldsymbol{0}_{1\times K} & \frac{n}{2\sigma^4} \end{pmatrix},
\end{aligned}$$

即有条件信息矩阵为

$$\mathscr{I}_{\boldsymbol{X}} = \mathscr{H}_{\boldsymbol{X}} = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{0}_{K\times 1} \\ \boldsymbol{0}_{1\times K} & \frac{n}{2\sigma^4} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}n\hat{\boldsymbol{Q}}_{\mathbf{XX}} & \boldsymbol{0}_{K\times 1} \\ \boldsymbol{0}_{1\times K} & \frac{n}{2\sigma^4} \end{pmatrix} \tag{6.50}$$

信息矩阵 $\mathscr{I}$ 的重要应用在于确定估计量的有效性.

定理 6.5.1 (信息矩阵与 Cramér-Rao 下界). *对于未知的总体参数 $\boldsymbol{\theta}$, 在模型设定正确时, 对于其任意的有无偏估计量 $\hat{\boldsymbol{\theta}}$, 则有*

$$\mathrm{Var}\left(\hat{\boldsymbol{\theta}}\middle|\boldsymbol{X}\right) \geqslant \mathscr{I}^{-1} \tag{6.51}$$

成立, 其中 $\mathscr{I}$ 为信息矩阵, 这里的 $\geqslant$ 表示矩阵之差的正定性. 我们称 $\mathscr{I}^{-1}$ 为总体参数 $\boldsymbol{\theta}$ 的 Cramér-Rao 下界,Cramér-Rao lower bound, 简称为 ***C-R 下界****.*

定理 (6.5.1) 给出了确定无偏参数是否最为有效的一种判断方法, 即是对比估计量的方差与 C-R 下界. 如果估计量的方差为 C-R 下界, 那么其便是 MVU 估计量. 极大似然估计最为有重要的性质, 便是 ML 估计量的渐进分布为 C-R 下界.

定理 6.5.2 (极大似然估计的渐进性质). *在正则条件 (regularity conditions) 下, 对于未知的总体参数 $\boldsymbol{\theta}$, 其极大似然估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$(如果可行) 满足有当 $n\to\infty$ 时,*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}-\boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0},\mathscr{I}^{-1}\right)$$

其中 $\mathscr{I}$ 是参数 $\boldsymbol{\theta}$ 的信息矩阵.

定理 (6.5.2) 中所需的正则条件较为复杂, 但对于通常的回归模型是可以满足的. 定理 (6.5.2) 意味着, 如果 ML 估计量是参数的无偏估计, 那么在大样本下,ML 估计量便是

最有效的. 这对于正态回归模型 (3.12.1) 而言, 在有限样本下, 依据式 (6.50) 其 C-R 下界为

$$\mathscr{I}_{\boldsymbol{X}}^{-1}=\mathscr{H}_{\boldsymbol{X}}^{-1}=\begin{pmatrix}\dfrac{1}{\sigma^2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{0}_{K\times 1}\\ \boldsymbol{0}_{1\times K} & \dfrac{n}{2\sigma^4}\end{pmatrix}^{-1}=\begin{pmatrix}\sigma^2\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} & \boldsymbol{0}_{K\times 1}\\ \boldsymbol{0}_{1\times K} & \dfrac{2\sigma^4}{n}\end{pmatrix}. \tag{6.52}$$

而在前面的章节, 我们计算了同方差下 OLS 估计量具有形如式 (3.24) 的条件协方差矩阵 $\boldsymbol{V}_{\hat{\beta}}=\sigma^2\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$, 则其于式 (6.52) 是的第一个对角块是相等的. 这也便是说, 在给定正态假设下,OLS 估计量 (或 ML 估计量)$\hat{\boldsymbol{\beta}}$ 达到了 C-R 下界, 即具有 “最小” 的条件协方差矩阵, 故 OLS 估计量是 MVU 估计, 且为 BUE——这一结论强于经典的 Gauss-Markov 定理的 BLUE.

对于 ML 估计量 $\hat{\sigma}_{\mathrm{ML}}^2$, 在第3.11节中我们分析了其是有偏的, 因而 C-R 下界对其不具有意义. 我们现在来考察偏差校正的方差估计量 s^2 的有效性. 在第3.12节, 依据正态假设, 我们证明了残差 $\hat{\boldsymbol{\varepsilon}}$ 满足有

$$\hat{\boldsymbol{\varepsilon}}|\,\boldsymbol{X}\sim\mathrm{N}\left(\boldsymbol{0},\boldsymbol{M}\sigma^2\right),\quad \left.\frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\sigma^2}\right|\boldsymbol{X}\sim\chi^2\left(n-K\right).$$

对于自由度为 q 的卡方分布, 其期望和方差分别为 q 和 $2q$, 故有

$$\mathbb{E}\left(\left.\frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\sigma^2}\right|\boldsymbol{X}\right)=n-K,\quad \mathrm{Var}\left(\left.\frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\sigma^2}\right|\boldsymbol{X}\right)=2\left(n-K\right),$$

于是在正态假设下,$\hat{\sigma}_{\mathrm{ML}}^2$ 和 s^2 的条件方差分别为

$$\mathrm{Var}\left(\hat{\sigma}_{\mathrm{ML}}^2\middle|\,\boldsymbol{X}\right)=\mathrm{Var}\left(\frac{\sigma^2}{n}\cdot\left.\frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\sigma^2}\right|\boldsymbol{X}\right)=\frac{\sigma^4}{n^2}\cdot 2\left(n-K\right)=\frac{2\left(n-K\right)\sigma^4}{n^2},$$

$$\mathrm{Var}\left(s^2\middle|\,\boldsymbol{X}\right)=\mathrm{Var}\left(\left.\frac{n}{n-K}\hat{\sigma}_{\mathrm{ML}}^2\right|\boldsymbol{X}\right)=\frac{n^2}{\left(n-K\right)^2}\cdot\frac{2\left(n-K\right)\sigma^4}{n^2}=\frac{2\sigma^4}{n-K},$$

这样便有 $\mathrm{Var}\left(s^2|\,\boldsymbol{X}\right)\geqslant 2n^{-1}\sigma^4$, 即无偏估计量 s^2 的方差没有达到 C-R 下界. 但这并不意味着 s^2 不是好的估计量, 林文夫 (cite) 指出, 不存在比 s^2 具有更小方差的方差估计量.

我们现在从大样本的视角来考察 ML 估计量与 C-R 边界. 类似于条件协方差矩阵估计量与渐进协方差矩阵估计量的关系, 对于正态回归模型 (3.12.1),ML 估计量的渐进协方差矩阵为

$$\begin{aligned}\mathscr{I}^{-1}=&\lim_{n\to\infty}n\mathscr{I}_{\boldsymbol{X}}^{-1}=\lim_{n\to\infty}n\begin{pmatrix}\dfrac{\sigma^2}{n}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1} & \boldsymbol{0}_{K\times 1}\\ \boldsymbol{0}_{1\times K} & \dfrac{2\sigma^4}{n}\end{pmatrix}\\ =&\lim_{n\to\infty}\begin{pmatrix}\sigma^2\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1} & \boldsymbol{0}_{K\times 1}\\ \boldsymbol{0}_{1\times K} & 2\sigma^4\end{pmatrix}=\begin{pmatrix}\sigma^2\boldsymbol{Q}_{\mathbf{XX}}^{-1} & \boldsymbol{0}_{K\times 1}\\ \boldsymbol{0}_{1\times K} & 2\sigma^4\end{pmatrix},\end{aligned}$$

依据定理 (6.5.2), 则 ML 估计量渐进协方差矩阵是 C-R 下界. 注意到 $\sigma^2\boldsymbol{Q}_{\mathbf{XX}}^{-1}$ 即是 OLS 估计量的同方差的渐进协方差矩阵 $\boldsymbol{V}_\beta^0$, 这样便是说在正态假设下,OLS 估计量的渐进分布的协方差矩阵也达到了 C-R 下界. 另一方面,$\hat{\sigma}^2_{\mathrm{ML}}$ 和 s^2 的具有相同的渐进分布, 其渐进方差也是 C-R 下界. 这一结果既是极大似然估计的优秀性质之一, 又是说明了了 OLS 估计量的良好性质.

6.6 广义最小二乘法 (GLS) 与加权最小二乘法 (WLS)

本节我们介绍放松假设 (3.2.1) 中 (d) 球形扰动项假设后, 更为一般地针对线性回归模型 (3.12.1) 的估计方法.

对于线性回归模型 (3.5.1), 其误差 $\boldsymbol{\varepsilon}$ 条件协方差矩阵可以更一般地写作

$$\operatorname{Var}(\boldsymbol{Y}|\boldsymbol{X}) = \operatorname{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right) = \boldsymbol{\Sigma}\sigma^2, \tag{6.53}$$

其中 $\boldsymbol{\Sigma}$ 是正定矩阵, 且可以是关于设计矩阵 $\boldsymbol{X}$ 的函数. 这样, 条件同方差式 (3.7) 和无自相关式 (3.8), 则可以写成 $\boldsymbol{\Sigma} = \boldsymbol{I}_n$ 的特例情形.

对于回归模型 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, 我们考虑式 (6.53) 时的 OLS 估计量 $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, 此时 $\hat{\boldsymbol{\beta}}$ 依然是无偏的、相合的, 这是因为定理 (3.5.2) 和定理 (5.1.1) 的证明只用到了严格外生性式 (3.5), 而没有涉及到 $\boldsymbol{\varepsilon}$ 的协方差矩阵的形式. 但在此时,Guass-Markov 定理不再成立,OLS 估计量不再是最有效的线性估计量.

具体而言, 我们可以计算 OLS 估计量的条件协方差矩阵矩阵为

$$\begin{aligned}\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\operatorname{Var}(\boldsymbol{Y}|\boldsymbol{X})\,\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2,\end{aligned} \tag{6.54}$$

这是类似于式 (3.23) 的. 一个自然的问题是, 在式 (6.14) 下我们能否找到一个关于回归系数 $\boldsymbol{\beta}$ 的有效估计量? 在此, 我们介绍更为一般的 Gauss-Markov 定理.

定理 6.6.1 (GLS 下的 Gauss-Markov 定理)**.** *将放松条件同方差式 (3.7) 和无自相关式 (3.8) 为式 (6.53) 时, 对于模型 (3.5.1) 的任意线性无偏估计量 $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}(\boldsymbol{X})\boldsymbol{Y}$, 则有*

$$\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) \geqslant \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\sigma^2$$

这里 $\geqslant$ 表示 $\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) - \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\sigma^2$ 为半正定矩阵. 这也意味着 $\operatorname{Var}\left(\tilde{\beta}_k\middle|\boldsymbol{X}\right) \geqslant \left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\sigma^2\right)_{kk}, k = 1, \cdots, K.$

证明. 由于 $\boldsymbol{\Sigma}$ 是正定矩阵, 依据定理 (3.6.4) 可知 $\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}^{1/2}\right)^2, \boldsymbol{\Sigma}^{-1} = \left(\boldsymbol{\Sigma}^{-1/2}\right)^2$, 这里 $\boldsymbol{\Sigma}^{-1/2}$ 是 $\boldsymbol{\Sigma}$ 的平方根, 其为实对称正定矩阵. 类似于定理 (3.5.3) 的证明, 无偏性意味着

$\boldsymbol{A}$ 满足有 $\boldsymbol{AX}=\boldsymbol{I}$, 且有

$$\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)=\boldsymbol{A}\operatorname{Var}\left(\boldsymbol{Y}|\boldsymbol{X}\right)\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}\sigma^2.$$

我们将 $\boldsymbol{A}$ 分解为 $\boldsymbol{A}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$, 令 $\boldsymbol{C}=\boldsymbol{A}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$, 则

$$\begin{aligned}
&\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\\
=&\left(\boldsymbol{C}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\Sigma}\left(\boldsymbol{C}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\\
=&\left(\boldsymbol{C}\boldsymbol{\Sigma}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right)\left(\boldsymbol{C}^{\mathrm{T}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\right)-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\\
=&\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{\mathrm{T}}+\boldsymbol{C}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\\
=&\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{\mathrm{T}}+\boldsymbol{C}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}},
\end{aligned}$$

而依据 $\boldsymbol{AX}=\boldsymbol{I}$ 则有

$$\begin{aligned}
&\boldsymbol{C}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)=\left(\boldsymbol{A}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)\\
=&\boldsymbol{A}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)\\
=&\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X},
\end{aligned}$$

故有

$$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}=\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{\mathrm{T}}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}+\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{\mathrm{T}}=\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{\mathrm{T}}\geqslant 0,$$

这里的 $\geqslant$ 由 $\boldsymbol{\Sigma}$ 的正定性保证, 故有 $\operatorname{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)\geqslant\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\sigma^2$ 成立. □

定理 (6.6.1) 是一般化的 Gauss-Markov 定理, 在式 (6.53) 的情形下, 其提供了对于任意线性无偏估计量的协方差矩阵的下界. 特别地, 当 $\Sigma=\boldsymbol{I}_n$ 时, 即得到了定理 (3.5.3), 此时 OLS 估计量的条件协方差矩阵便可达到此下界.

我们现在利用定理 (6.6.1) 回答此前的问题, 当误差向量 $\boldsymbol{\varepsilon}$ 满足更为一般的式 (6.53) 时, 在知道 $\boldsymbol{\Sigma}$ 的理想情况下, 存在最佳线性无偏估计量.

考虑将 $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 左乘以 $\boldsymbol{\Sigma}$ 的平方根的逆 $\boldsymbol{\Sigma}^{-1/2}$, 得

$$\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon},$$

令

$$\tilde{\boldsymbol{Y}}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y},\quad \tilde{\boldsymbol{X}}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X},\quad \tilde{\boldsymbol{\varepsilon}}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon},$$

则有新的回归模型

$$\tilde{\boldsymbol{Y}}=\tilde{\boldsymbol{X}}\boldsymbol{\beta}+\tilde{\boldsymbol{\varepsilon}},\tag{6.55}$$

此时新的误差向量满足有

$$\mathbb{E}(\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = \mathbb{E}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right) = \boldsymbol{\Sigma}^{-1/2}\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbf{0},$$
$$\mathrm{Var}(\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = \mathrm{Var}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right) = \boldsymbol{\Sigma}^{-1/2}\mathrm{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}\sigma^2 = \boldsymbol{I}_n\sigma^2,$$

即回归模型式 (6.55) 满足了假设 (3.2.1), 且仍有待估计的回归系数 $\boldsymbol{\beta}$. 我们对方程式 (6.55) 使用 OLS 去估计 $\boldsymbol{\beta}$, 则可以得到此前分析的 OLS 估计量的各种性质.

定义 6.6.1 (广义最小二乘估计量). *对于回归模型式 (6.55), 我们将其 OLS 估计量记为*

$$\tilde{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{Y}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}, \tag{6.56}$$

*我们称 $\tilde{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 为***广义最小二乘估计量*** (Generalized least squares estimator), 简称为 ***GLS* 估计量** (GLS estimator).*

由于回归模型式 (6.55) 符合模型 (3.5.1), 因而不难推知, 由式 (6.55) 的 OLS 估计量所衍生出的 GLS 估计量是最优线性无偏估计量. 实际上, 我们亦可计算 $\tilde{\boldsymbol{\beta}}_{\mathrm{GLS}}$ 的条件期望和条件协方差矩阵分别为

$$\begin{aligned}\mathbb{E}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{GLS}}\middle|\boldsymbol{X}\right) &= \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\mathbb{E}\left(\tilde{\boldsymbol{Y}}\middle|\boldsymbol{X}\right) = \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\mathbb{E}\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta}+\tilde{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)\\ &= \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta}+\mathbb{E}(\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X})\right) = \boldsymbol{\beta}\end{aligned}$$

和

$$\begin{aligned}&\mathrm{Var}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{GLS}}\middle|\boldsymbol{X}\right) = \mathrm{Var}\left(\left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{Y}}\middle|\boldsymbol{X}\right)\\ &= \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\mathrm{Var}(\tilde{\boldsymbol{\varepsilon}}|\boldsymbol{X})\,\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\\ &= \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{I}_n\sigma^2\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1} = \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\left(\boldsymbol{I}_n\sigma^2\right)\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\\ &= \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\sigma^2 = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X}\right)^{-1}\sigma^2,\end{aligned}$$

因而 GLS 估计量不仅是无偏的, 而且其协方差矩阵达到了定理 (6.6.1) 中的下界.

显然, 当线性回归模型 (3.5.1) 存在条件异方差式 (3.9) 问题时, 即有

$$\boldsymbol{\Sigma} = \boldsymbol{D} = \begin{pmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \sigma_n^2\end{pmatrix} = \mathrm{diag}\left(\sigma_1^2,\cdots,\sigma_n^2\right),$$

在已知 $\boldsymbol{\Sigma}$ 的理想情况下我们可以使用 GLS 估计量去估计回归系数 $\boldsymbol{\beta}$. 此时有

$$\boldsymbol{\Sigma}^{1/2} = \mathrm{diag}\left(\sigma_1,\cdots,\sigma_n\right), \quad \boldsymbol{\Sigma}^{-1/2} = \mathrm{diag}\left(\sigma_1^{-1},\cdots,\sigma_n^{-1}\right), \quad \boldsymbol{\Sigma}^{-1} = \mathrm{diag}\left(\sigma_1^{-2},\cdots,\sigma_n^{-2}\right),$$

故回归模型 (6.55) 的各变量为

$$\tilde{\boldsymbol{Y}}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y}=\left(\frac{Y_1}{\sigma_1},\cdots,\frac{Y_n}{\sigma_n}\right)^{\mathrm{T}},\quad \tilde{\boldsymbol{X}}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}=\left(\frac{\mathbf{X}_1}{\sigma_1},\cdots,\frac{\mathbf{X}_n}{\sigma_n}\right)^{\mathrm{T}}$$

和

$$\tilde{\boldsymbol{\varepsilon}}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}=\left(\frac{\varepsilon_1}{\sigma_1},\cdots,\frac{\varepsilon_n}{\sigma_n}\right)^{\mathrm{T}}.$$

我们称此时的 GLS 估计量为**加权最小二乘估计量** (weighted least squares estimator), 简称为 **WLS 估计量** (WLS estimator).

尽管 GLS 估计量具有如此优异的性质, 但这一切前提都是建立在已知协方差矩阵 $\boldsymbol{\Sigma}$ 的“理想”情况下——显示是, 通常我们不知道矩阵 $\boldsymbol{\Sigma}$, 因而上述推导的 GLS 估计量是不可行的 (infeasible). 一个自然的想法是使用 $\boldsymbol{\Sigma}$ 的估计量 $\hat{\boldsymbol{\Sigma}}$ 去代替, 然而, 实对称的 n 阶矩阵 $\boldsymbol{\Sigma}$ 即使剔除掉了重复元素, 仍有 $1+2+\cdots+n=n(n+1)/2$ 个参数需要估计, 而我们的样本容量仅有 n 个——这直接宣判了得到一个一般化的协方差矩阵估计量 $\hat{\boldsymbol{\Sigma}}$ 是不可行的. 除非有关于 $\boldsymbol{\Sigma}$ 的特定的模型设定, 我们不能得到 $\hat{\boldsymbol{\Sigma}}$.

下面我们介绍可行的 GLS 估计量以及其渐进性质.

6.7 三类渐进假设检验

本节介绍计量经济学中常见的三种渐进假设检验:Wald 检验、似然比检验和 Lagrange 乘数法检验. 在大样本下, 这三种检验方法的检验统计量的渐进分布都是卡方分布, 但在实践的使用中又有一定的差异. 本节设计的假设检验的基本原理, 参见第3.13节.

本节我们研究的研究假设检验问题是

$$\mathrm{H}_0:\boldsymbol{\theta}=\boldsymbol{\theta}_0\quad v.s.\quad \mathrm{H}_1:\boldsymbol{\theta}\neq\boldsymbol{\theta}_0, \tag{6.57}$$

其中映射 $\boldsymbol{h}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{q\times 1}$,$\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})$ 是有回归系数 $\boldsymbol{\beta}$ 所定义的我们感兴趣的参数, 则原假设式 (6.57) 等价于

$$\mathrm{H}_0:\boldsymbol{\beta}\in B=\{\boldsymbol{\beta}:\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\theta}_0\}\quad v.s.\quad \mathrm{H}_1:\boldsymbol{\beta}\notin B=\{\boldsymbol{\beta}:\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\theta}_0\}. \tag{6.58}$$

我们将参数 $\boldsymbol{\theta}$ 的值域记为 $\boldsymbol{\Theta}$, 称 $\boldsymbol{\Theta}$ 为参数空间. 特别地, 本节将重点研究关于线性约束的原假设 $\mathrm{H}_0:\boldsymbol{R\beta}=\boldsymbol{\theta}_0$ 的假设检验问题, 其中 $\boldsymbol{R}\in\mathbb{R}^{q\times K}$,$\mathrm{rank}\boldsymbol{R}=q$. 这时参数 $\boldsymbol{\theta}$ 为线性变换 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{R\beta}$.

Wald 检验我们在第5.4节已经介绍了, 我们现在先介绍基准统计量.Wald 统计量是通过欧氏距离度量了估计量与原假设的差异, 更为一般的则却采用某种准则函数来度量与原假设的差异.

定义 6.7.1 (基准统计量). 对于原假设式 *(6.57)*, 设关于回归系数 $\boldsymbol{\beta}$ 的**准则函数** *(criterion function)*为

$$J(\boldsymbol{\beta}): \mathbb{R}^{q\times 1} \to \mathbb{R}, \tag{6.59}$$

记不受限与受限的估计量分别为

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} J(\boldsymbol{\beta}), \quad \tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{\operatorname{argmin}} J(\boldsymbol{\beta})$$

其中参数空间 $\boldsymbol{\Theta}$ 满足 $\{\boldsymbol{\theta}: \boldsymbol{\theta} = \boldsymbol{\theta}_0\} \subseteq \boldsymbol{\Theta}$. 我们称

$$J = \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} J\left(\tilde{\boldsymbol{\beta}}\right) - \min_{\boldsymbol{\theta}\in\Theta} J\left(\tilde{\boldsymbol{\beta}}\right) \geqslant 0 \tag{6.60}$$

为关于原假设式 *(6.57)* 的**基准统计量** *(criterion-based statistic)*.

原假设式 (6.57) 是等价于关于回归系数的原假设式 (6.58), 则我们可以依据回归系数 $\boldsymbol{\beta}$ 的估计量来检验式 (6.57). 由于无限制的参数空间 $\boldsymbol{\Theta}$ 包含了原假设式 (6.57), 因而以原假设式 (6.57) 作为约束条件时准则函数 $J(\boldsymbol{\beta})$ 能取得的最小值也被限制了, 这样基准统计量式 (6.60) 的取值是非负的.

非约束条件下准则函数的最小值包含了约束条件, 如果原假设式 (6.57) 成立, 则依据抽取的样本所计算的估计量会使得基准统计量式 (6.60) 的值较小, 反之若原假设式 (6.57) 不成立有更小的取值. 基于此便可定义一种检验方法, 我们称为**基准检验** (criterion-based test).

显然, 基准检验依赖于准则函数 $J(\boldsymbol{\beta})$ 的选择, 我们将介绍基准检验的两种具体的检验方法: 最小距离检验和似然比检验. 前者以 MD 估计量的准则函数作为 $J(\boldsymbol{\beta})$, 后者则是使用的对数似然函数.

考虑原假设为线性约束 $\mathrm{H}_0: \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{\theta}_0$, 我们使用 MD 估计量的准则函数式 (6.23) 作为

$$J(\boldsymbol{\beta}) = n\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}} \hat{\boldsymbol{W}} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right),$$

则依据第6.3节的内容知

$$\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} J(\boldsymbol{\theta}) = \hat{\boldsymbol{\beta}}, \quad \underset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{\operatorname{argmin}} J(\boldsymbol{\theta}) = \tilde{\boldsymbol{\beta}}_{\mathrm{md}},$$

且有

$$J\left(\hat{\boldsymbol{\beta}}\right) = 0, \quad J\left(\tilde{\boldsymbol{\beta}}_{\mathrm{md}}\right) = n\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{md}}\right)^{\mathrm{T}} \hat{\boldsymbol{W}} \left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{md}}\right),$$

故基准统计量为

$$J = J\left(\tilde{\boldsymbol{\beta}}_{\mathrm{md}}\right) - J\left(\hat{\boldsymbol{\beta}}\right) = J\left(\tilde{\boldsymbol{\beta}}_{\mathrm{md}}\right), \tag{6.61}$$

我们称式 (6.61) 为**最小距离统计量** (minimum distance statistic), 简称为 **MD 统计量**.

MD 估计量并不唯一, 因而 MD 统计量也不唯一. 我们取 $\hat{\boldsymbol{W}}=\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}$, 即使用最有效率的 EMD 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{emd}}$, 则有基准统计量为

$$J^{*}=J\left(\tilde{\boldsymbol{\beta}}_{\mathrm{emd}}\right)=n\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{emd}}\right)^{\mathrm{T}}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{emd}}\right), \tag{6.62}$$

此时我们称 J^{*} 为**有效最小距离统计量** (efficient minimum distance statistic), 简称为 **EMD 统计量**. 依据第6.3节的式6.35, 则式 (6.62) 可以等价于

$$\begin{aligned}J^{*}=&n\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)\\ =&n\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right),\end{aligned}$$

由于 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}=n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, 不难注意到 J^{*} 与第5.4节的 Wald 统计量式 (5.27) 是相同的. 这就是说, 对于原假设线性约束 $\mathrm{H}_{0}:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{\theta}_{0}$,Wald 统计量与 EMD 统计量是相同的, 因而两者就有相同的检验方法. 这也便意味着

$$J^{*}=n\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)=W\xrightarrow{d}\chi^{2}(q)\quad n\to\infty.$$

特别地, 当回归模型满足同方差时,$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}=\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}s^{2}=n\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^{2}$,EMD 估计量即为 $\tilde{\boldsymbol{\beta}}_{\mathrm{emd}}$ 即为 CLS 估计量 $\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$, 此时的基准统计量则有

$$\begin{aligned}J^{0}=&J\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)=n\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)^{\mathrm{T}}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)\\ =&\frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)}{s^{2}}=W^{0}\xrightarrow{d}\chi^{2}(q),\quad n\to\infty.\end{aligned}$$

EMD 统计量的渐进分布还可以通过 Hausman 等式式 (6.37) 来说明, 由于

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{emd}}\right)\xrightarrow{\mathrm{P}}\mathrm{N}\left(0,\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{V}_{\boldsymbol{\beta}}\right),$$

则仿照第3.12的证明可知 $J^{*}\xrightarrow{d}\chi^{2}(q)$.

定理 6.7.1 (EMD 统计量的基准检验). *在假设 (5.1.1) 下, 对于原假设线性约束 $\mathrm{H}_{0}:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{\theta}_{0}$, 形如式 (6.62) 的 EMD 统计量等于 Wald 统计量, 因而*

$$J^{*}=n\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(\boldsymbol{R}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)=W\xrightarrow{d}\chi^{2}(q),\quad n\to\infty, \tag{6.63}$$

$$J^{0}=\frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{\theta}_{0}\right)}{s^{2}}=W^{0}\xrightarrow{d}\chi^{2}(q),\quad n\to\infty, \tag{6.64}$$

此时基准检验等同于定理 (5.4.2).

现在我们考虑使用对数似然函数作为准则函数. 同样考虑原假设 $\mathrm{H}_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{\theta}$, 由式 (6.39) 得到准则函数

$$J\left(\boldsymbol{\beta},\sigma^2\right)=2\ln L\left(\boldsymbol{\beta},\sigma^2\right)=-n\ln\left(2\pi\sigma^2\right)-\frac{1}{\sigma^2}\mathrm{SSE}\left(\boldsymbol{\beta}\right),$$

这里乘以 2 是为了计算的方便, 并不影响估计量的确定. 这里与定义 (6.7.1) 有所差异由于对数似然函数是取值越大越“优”, 因而实际遵循定义 (6.7.1) 来确定参数估计量时应当使用的是 $-2\ln L$. 参考第6.1节和第6.5节中最优化问题的一阶条件, 可以得到

$$\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}-J\left(\boldsymbol{\beta},\sigma^2\right)=\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right),\quad \underset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{\operatorname{argmin}}-J\left(\boldsymbol{\beta},\sigma^2\right)=\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\tilde{\sigma}^2_{\mathrm{CLS}}\right),$$

且有方差估计量 $\hat{\sigma}^2$ 与 SSE 满足有

$$\tilde{\sigma}^2_{\mathrm{CLS}}=\frac{\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}}{n}=\frac{\mathrm{SSE}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)}{n},\quad \hat{\sigma}^2_{\mathrm{ML}}=\frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{n}=\frac{\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)}{n},\tag{6.65}$$

故有

$$\begin{aligned}J\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)&=\ln L\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)=-n\ln\left(2\pi\hat{\sigma}^2_{\mathrm{ML}}\right)-n,\\ J\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\tilde{\sigma}^2_{\mathrm{CLS}}\right)&=\ln L\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\tilde{\sigma}^2_{\mathrm{CLS}}\right)=-n\ln\left(2\pi\tilde{\sigma}^2_{\mathrm{CLS}}\right)-n,\end{aligned}$$

于是基准统计量为

$$\begin{aligned}J=&-J\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\tilde{\sigma}^2_{\mathrm{CLS}}\right)-\left(-J\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}},\hat{\sigma}^2_{\mathrm{ML}}\right)\right)\\ =&n\ln\left(2\pi\tilde{\sigma}^2_{\mathrm{CLS}}\right)-n\ln\left(2\pi\hat{\sigma}^2_{\mathrm{ML}}\right)=n\ln\frac{\tilde{\sigma}^2_{\mathrm{CLS}}}{\hat{\sigma}^2_{\mathrm{ML}}},\end{aligned}\tag{6.66}$$

我们将式 (6.66) 称为**似然比统计量** (likelihood ratio statistic), 并常记有

$$\lambda_{\mathrm{LR}}=n\ln\frac{\tilde{\sigma}^2_{\mathrm{CLS}}}{\hat{\sigma}^2_{\mathrm{ML}}}.\tag{6.67}$$

由于似然函数需要给定分布, 故似然比统计量依赖于正态假设. 此时,ML 估计量与 OLS 估计量是相同的, 依据第3.12节与第6.2节, 我们有

$$\left.\frac{n\tilde{\sigma}^2_{\mathrm{CLS}}}{\sigma^2}\right|\boldsymbol{X}\sim\chi^2\left(n-K+q\right),\quad \left.\frac{n\hat{\sigma}^2_{\mathrm{ML}}}{\sigma^2}\right|\boldsymbol{X}\sim\chi^2\left(n-K\right),$$

这样我们可以进一步推导似然比统计量式 (6.67) 的有限样本分布, 尽管其形式并不简单. 更常见地, 我们使用似然比统计量的渐进分布去进行假设检验.Wilks 证明了, 形如式 (6.67) 的似然比的渐进分布为卡方分布, 其自由度为非限制的参数空间与限制的参数空间的维度之差. 对于正态回归模型 (3.12.1) 而言, 即有当 $n\to\infty$ 时,$\lambda_{\mathrm{LR}}\xrightarrow{d}\chi^2\left(q\right)$ 成立. 我们由此得到似然比检验.

定理 6.7.2 (似然比检验). *在假设 (5.1.1) 下, 对于模型 (3.12.1) 的回归系数 $\boldsymbol{\beta}$, 设参数 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})$, 其中 $\boldsymbol{h}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{q\times 1}$, 就假设*

$$\mathrm{H}_0:\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\theta}_0 \quad v.s. \quad \mathrm{H}_1:\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})\neq\boldsymbol{\theta}_0$$

而言, 以对数似然函数作为准则函数, 得到的基准统计量为形如式 (6.66) 的似然比统计量, 则有

$$\lambda_{\mathrm{LR}}=n\ln\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\xrightarrow{d}\chi^2(m), \quad n\to\infty, \tag{6.68}$$

其中 $\hat{\sigma}^2$ 是无限制的方差估计量,$\tilde{\sigma}^2$ 是以 H_0 为限制的方差估计量,m 为非限制的参数空间与限制的参数空间的维度之差.. 我们称渐进显著性水平为 α的检验

$$\phi:\text{若}\lambda_{\mathrm{LR}}>F^{-1}(1-\alpha)\text{则拒绝 }H_0$$

*为**似然比检验** (likelihood ratio test), 简称为 **LR 检验**, 其中 $F(\cdot)$ 是自由度为 q 的卡方分布的累计密度函数,$F^{-1}(\cdot)$ 为 $F(\cdot)$ 的逆函数. 似然比检验的渐进 p 值为 $p=1-F(\lambda_{\mathrm{LR}})$.*

定理 6.7.3 (线性约束的似然比检验). *在假设 (5.1.1) 下, 对于模型 (3.12.1) 的回归系数 $\boldsymbol{\beta}$, 就假设*

$$\mathrm{H}_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{\theta}_0 \quad v.s. \quad \mathrm{H}_1:\boldsymbol{R}\boldsymbol{\beta}\neq\boldsymbol{\theta}_0,$$

其中 $\boldsymbol{R}\in\mathbb{R}^{q\times K}$,rank$\boldsymbol{R}=q$, 以对数似然函数作为准则函数, 得到的基准统计量为形如式 (6.66) 的似然比统计量, 则有

$$J=\lambda_{\mathrm{LR}}=n\ln\frac{\hat{\sigma}^2_{\mathrm{CLS}}}{\tilde{\sigma}^2_{\mathrm{ML}}}\xrightarrow{d}\chi^2(q), \quad n\to\infty. \tag{6.69}$$

我们称渐进显著性水平为 α的检验

$$\phi:\text{若}\lambda_{\mathrm{LR}}>F^{-1}(1-\alpha)\text{则拒绝 }H_0$$

为 H_0 的似然比检验, 其中 $F(\cdot)$ 是自由度为 q 的卡方分布的累计密度函数,$F^{-1}(\cdot)$ 为 $F(\cdot)$ 的逆函数. 似然比检验的渐进 p 值为 $p=1-F(\lambda_{\mathrm{LR}})$.

在第3.13节中, 我们指出了 F 统计量或 F 比率的 F 检验与似然比检验是相同的. 对于形如式 (3.82), 我们有更一般的形式为

$$F=\frac{\left(\tilde{\sigma}^2_{\mathrm{CLS}}-\hat{\sigma}^2_{\mathrm{OLS}}\right)/q}{\hat{\sigma}^2_{\mathrm{OLS}}/(n-K)}, \tag{6.70}$$

我们称式 (6.70) 为 **F 比率** (F-ratio). 依据式 (6.65), 则式 (6.70) 可以化简为

$$\begin{aligned} F =& \frac{\left(\tilde{\sigma}_{\mathrm{CLS}}^2-\hat{\sigma}_{\mathrm{OLS}}^2\right)/q}{\hat{\sigma}_{\mathrm{OLS}}^2/\left(n-K\right)}=\frac{n\left(\tilde{\sigma}_{\mathrm{CLS}}^2-\hat{\sigma}_{\mathrm{OLS}}^2\right)/q}{n\hat{\sigma}_{\mathrm{OLS}}^2/\left(n-K\right)}=\frac{\left(\mathrm{SSE}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)-\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)\right)/q}{\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)/\left(n-K\right)} \\ =& \frac{\mathrm{SSE}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)-\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)}{q}\Big/\frac{\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)}{n-K}=\frac{\mathrm{SSE}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)-\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)}{qs_{\mathrm{OLS}}^2}. \end{aligned} \tag{6.71}$$

正态回归模型 (3.12.1) 是同方差的, 因而依据第6.3节式 (6.25), 则有

$$\mathrm{SSE}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)=J^0\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)+\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right),$$

其中

$$J^0\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)=n\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)^{\mathrm{T}}\hat{\boldsymbol{Q}}_{\mathbf{xx}}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right),$$

故 F 比率进一步等价变形有

$$F=\frac{J^0\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)}{qs_{\mathrm{OLS}}^2}=\frac{n\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)^{\mathrm{T}}\hat{\boldsymbol{Q}}_{\mathbf{xx}}\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)}{qs_{\mathrm{OLS}}^2}=\frac{J^0}{q}=\frac{W^0}{q}\xrightarrow{d}\frac{\chi^2\left(q\right)}{q},\quad n\to\infty,$$

这里的 J^0 和 W^0 分别是同方差下的 EMD 统计量式 (6.64) 和 Wald 统计量式 (5.28). 这也便证明了第3.13节中式 (3.82) 与式 (3.81) 是相等的, 即 F 比率是同方差的 Wald 统计量的一种等价形式. 在此意义上,F 比率的假设检验与 Wald 检验是完全一样的, 定理 (3.13.2) 是同方差的 Wald 检验的一种具体应用.

F 比率的一大优势的在于其的可计算性, 依据式 (6.70) 和式 (6.71), 计算出 F 只需知道方差估计量的结果或者残差平方和 $\mathrm{RSS}=\mathrm{SSE}\left(\boldsymbol{\beta}_{\text{估计量}}\right)$, 这在通常的回归分析结果中是会被汇报的.F 比率与同方差的 Wald 统计量相差了一个除以线性约束的个数 q 的差距. 在给定了正态假设时, 我们知道 F 比率式 (3.81) 服从有 $F(q,n-K)$, 因而在小样本的情形下使用了 F 分布确定检验的临界值和 p 值. 而在大样本下,F 比率则具有渐进分布 $\chi^2\left(q\right)/q$, 以此来计算临界值和 p 值. 就 $F(q,n-K)$ 与 $\chi^2\left(q\right)/q$ 而言, 在相同的显著性水平下前者的临界值要越大, 因而 p 值也更大, 检验的结果更为保守. 对于 F 比率, 其渐进分布的一条经验法则是, 对于充分大的 n, 渐进显著性水平为 5% 的临界值与 q 是反向变动的,q 越大, 临界值越小. 当 $q\geqslant 7$ 时,F 比率的临界值小于 2.

我们第3.13节中还指出,F 比率的 F 检验与似然比检验的等价的. 这是因为, 对比式 (6.67) 和式 (6.70) 即有

$$F=\frac{n-K}{q}\cdot\frac{\tilde{\sigma}_{\mathrm{CLS}}^2-\hat{\sigma}_{\mathrm{OLS}}^2}{\hat{\sigma}_{\mathrm{OLS}}^2}=\frac{n-K}{q}\cdot\left(\frac{\tilde{\sigma}_{\mathrm{CLS}}^2}{\hat{\sigma}_{\mathrm{OLS}}^2}-1\right)\Rightarrow\frac{\tilde{\sigma}_{\mathrm{CLS}}^2}{\hat{\sigma}_{\mathrm{OLS}}^2}=\frac{qF}{n-K}+1, \tag{6.72}$$

因而

$$\lambda_{\mathrm{LR}}=n\ln\frac{\hat{\sigma}_{\mathrm{ML}}^2}{\tilde{\sigma}_{\mathrm{CLS}}^2}=n\ln\left(\frac{qF}{n-K}+1\right)=n\ln\left(\frac{qF}{n-K}+1\right), \tag{6.73}$$

即 F 比率与似然比 λ_{LR} 是一一对应的单调变换. 这意味着, 在给定的显著性水平下, 当统计量 F 或 λ_{LR} 的值较大时拒绝原假设是等价. 设一个显著性水平为 α 的似然比检验为

$$\phi: 若\lambda_{\mathrm{LR}} > c_1 则拒绝\ \mathrm{H}_0,$$

则等价的 F 检验为

$$\phi: 若F > c_2 则拒绝\ \mathrm{H}_0,$$

其中依据式 (6.73) 得

$$\lambda_{\mathrm{LR}} > c_1 \Rightarrow n \ln\left(\frac{qF}{n-K}+1\right) > c_1 \Rightarrow F > c_2 = \frac{n-K}{q}\left(e^{\frac{c_1}{n}}-1\right).$$

我们现在介绍得分检验——在计量经济学中更多的被称为 Lagrange 乘数检验. 在此前分析的统计量中,Wald 统计量是仅基于非限制的估计量 $\hat{\boldsymbol{\beta}}$, 而基准统计量 (包括 MD 统计量、似然比统计量) 则包含了受限统计量 $\tilde{\boldsymbol{\beta}}$ 与非限制统计量 $\hat{\boldsymbol{\beta}}$, 而下面我们将介绍的得分检验则仅依赖于受限统计量 $\tilde{\boldsymbol{\beta}}$.

定义 6.7.2 (得分统计量). *对于原假设式 (6.57), 我们以对数似然函数*

$$\ln L\left(\boldsymbol{\beta}, \sigma^2\right) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\mathrm{SSE}\left(\boldsymbol{\beta}\right)$$

*来定义***得分统计量** *(score statistic)为*

$$S = \left(\frac{\partial}{\partial\boldsymbol{\beta}}\ln L\left(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\right)\right)^{\mathrm{T}}\left(-\frac{\partial}{\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\boldsymbol{\beta}}\ln L\left(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\right)\right)^{-1}\left(\frac{\partial}{\partial\boldsymbol{\beta}}\ln L\left(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\right)\right), \tag{6.74}$$

其中 $\left(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\right)$ *依据* H_0 *作为约束条件计算的受限估计量.*

得分检验最早由 Rao 在 1948 年提出, 之后 1959 年 Silvey 证明了得分检验等价于对 Lagrange 乘数的检验, 故得分检验又称为 Lagrange 检验.1980 年因 Breusch 和 Pagan 对异方差的检验使用的是 Lagrange 乘数检验, 此后 Lagrange 乘数检验一名称被计量经济学所广泛接受. 我们现在来揭示式 (6.74) 与 Lagrange 乘数的关系. 考虑以原假设式 (6.57) 作为约束条件的 ML 估计量, 则有 Lagrange 函数

$$\mathscr{L}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda}\right) = \ln L\left(\boldsymbol{\beta}, \sigma^2\right) + \boldsymbol{\lambda}^{\mathrm{T}}\left(\boldsymbol{h}\left(\boldsymbol{\beta}\right) - \boldsymbol{\theta}_0\right),$$

其中 Lagrange 乘数 $\boldsymbol{\lambda} \in \mathbb{R}^{K\times 1}$. 一阶条件有

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathscr{L} = \frac{\partial}{\partial\boldsymbol{\beta}}\ln L\left(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\right) + \boldsymbol{\lambda}^{\mathrm{T}}\frac{\partial\boldsymbol{h}\left(\tilde{\boldsymbol{\beta}}\right)}{\partial\boldsymbol{\beta}} = 0 \Rightarrow \frac{\partial}{\partial\boldsymbol{\beta}}\ln L\left(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\right) = -\left(\boldsymbol{J}\boldsymbol{h}\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{\lambda}\right)^{\mathrm{T}}$$

成立, 即知对似然函数的梯度是关于 Lagrange 乘数的线性变换, 这是单调的, 因而得分统计量 S 是依赖于受限估计量的 Lagrange 乘数, 对得分统计量 S 的检验也等价于对 Lagrange 乘数 $\boldsymbol{\lambda}$ 的检验. 也是因此, 得分统计量 S 又被称为 **Lagrange 乘数统计量** (Lagrange multiplier statistic), 简称 **LM 统计量**.

如果得分统计量 S 的值越小, 则意味着对似然函数的梯度接近于零向量, 因而受限估计量 $\tilde{\boldsymbol{\beta}}$ 越接近非受限的极大似然估计, 说明约束条件是非严格的, 此时抽样误差的影响不容易排除, 因而不应该轻易拒绝原假设 H_0. 反之, 得分统计量 S 越大, 则说明约束条件越严格, 更有可能拒绝原假设.

具体地, 依据正态回归模型3.12.1的对数似然函数, 利用附录 () 的矩阵微积分可以计算

$$\frac{\partial^2}{\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\boldsymbol{\beta}}\ln L\left(\boldsymbol{\beta},\sigma^2\right)=-\frac{1}{2\sigma^2}\cdot\frac{\partial}{\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\boldsymbol{\beta}\right),\quad \frac{\partial}{\partial\boldsymbol{\beta}}\ln L\left(\boldsymbol{\beta},\sigma^2\right)=-\frac{1}{2\sigma^2}\cdot\frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\boldsymbol{\beta}\right),$$

以及

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\boldsymbol{\beta}\right)=-2\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right),\quad \frac{\partial}{\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\boldsymbol{\beta}}\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)=2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X},$$

故得分统计量 S 化简为

$$\begin{aligned}S=&\left(\frac{1}{\tilde{\sigma}^2}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\left(\frac{1}{\tilde{\sigma}^2}\cdot\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\frac{1}{\tilde{\sigma}^2}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)\right)\\ =&\frac{1}{\tilde{\sigma}^2}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right)=\frac{1}{\tilde{\sigma}^2}\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\boldsymbol{P}\tilde{\boldsymbol{\varepsilon}},\end{aligned}\tag{6.75}$$

其中 $\boldsymbol{P}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ 是 OLS 估计量的投影矩阵,$\tilde{\boldsymbol{\varepsilon}}$ 是受限估计量 $\tilde{\boldsymbol{\beta}}$ 的残差向量.

式 (6.75) 的形式在正态假设下, 可以用于推导得分统计量 S 的有限样本性质. 特别地, 考虑原假设为线性约束 $\mathrm{H}_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{\theta}_0$, 则有受限估计量为 $\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}},\tilde{\sigma}^2_{\mathrm{CLS}}\right)$, 依据第6.1节中定理 (6.1.1) 得

$$\tilde{\sigma}^2=\frac{1}{n}\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}=\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon},$$

因而

$$\begin{aligned}S=&\frac{1}{\tilde{\sigma}^2}\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\boldsymbol{P}\tilde{\boldsymbol{\varepsilon}}=\frac{1}{\tilde{\sigma}^2}\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{P}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}\\ =&\frac{1}{\tilde{\sigma}^2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{\varepsilon}=\frac{1}{\tilde{\sigma}^2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\\ =&\frac{1}{\tilde{\sigma}^2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}=\frac{1}{\tilde{\sigma}^2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\\ =&\frac{1}{\tilde{\sigma}^2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon},\end{aligned}$$

又

$$\begin{aligned}&\mathrm{rank}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}=\mathrm{tr}\left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)=q,\\ &\mathrm{rank}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)=\mathrm{tr}\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\right)=n-K+q,\end{aligned}$$

故有 $n\tilde{\sigma}^2 \big| \boldsymbol{X} \sim \chi^2(n-K+q)$ 和 $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} \big| \boldsymbol{X} \sim \chi^2(q)$. 但这并不意味着有

$$\frac{(n-K+q)S}{qn} = \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}/q}{n\tilde{\sigma}^2/(n-K+q)} \sim F(q, n-K+q)$$

成立, 因为分子与分母的卡方分布都由误差 $\boldsymbol{\varepsilon}$ 导出, 两者不是独立的. 在有限样本下 S 的分布较为复杂[6], 不便于用于进行假设检验.

我们现在来推导得分统计量的渐进分布. 将受限估计量的残差 $\tilde{\boldsymbol{\beta}}$ 分解为

$$\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right) + \hat{\boldsymbol{\varepsilon}},$$

则由 OLS 估计量的正规方程组式 (3.18) 得

$$\boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{X}^{\mathrm{T}}\left[\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right) + \hat{\boldsymbol{\varepsilon}}\right] = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right),$$

故式 (6.75) 可以等价变换为

$$\begin{aligned}
S =& \frac{1}{\tilde{\sigma}^2}\left(\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right) + \hat{\boldsymbol{\varepsilon}}\right)^{\mathrm{T}} \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right) + \hat{\boldsymbol{\varepsilon}}\right) \\
=& \frac{1}{\tilde{\sigma}^2}\left(\boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}\right)^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}} = \frac{1}{\tilde{\sigma}^2}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right) \\
=& \frac{1}{\tilde{\sigma}^2}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right) \\
=& \frac{1}{\tilde{\sigma}^2}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right) = \frac{J^0\left(\tilde{\boldsymbol{\beta}}\right)}{\tilde{\sigma}^2},
\end{aligned}$$

即有

$$S = \frac{1}{\tilde{\sigma}^2}\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right) = \frac{J^0\left(\tilde{\boldsymbol{\beta}}\right)}{\tilde{\sigma}^2}, \tag{6.76}$$

这里的 $J^0\left(\tilde{\boldsymbol{\beta}}\right)$ 是 MD 估计量的准则函数式 (6.24).

当受限估计量为 $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}$, 则得分统计量为 $S = J^0\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)/\tilde{\sigma}^2_{\mathrm{CLS}}$, 其是同方差的 Wald 统计量式 (5.28) 将 OLS 估计量的 s^2 替换为了 CLS 估计量的方差估计量. 利用式 (6.7) 可以得到

$$S = \frac{\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)}{\tilde{\sigma}^2_{\mathrm{CLS}}} = \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}\right)}{\tilde{\sigma}^2_{\mathrm{CLS}}}.$$

我们进一步证明 CLS 估计量的得分统计量 S 是形如式 (6.70) 的 F 比率的单调变

[6]两个相关的卡方分布之比的分布, 可以参见资料https://stats.stackexchange.com/a/608566以及https://doi.org/10.1007/s00362-007-0105-0.

换, 利用式 (6.65) 和式 (6.72), 即有

$$\begin{aligned} S =& \frac{J^0\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)}{\tilde{\sigma}_{\mathrm{CLS}}^2}=\frac{\mathrm{SSE}\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)-\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)}{\tilde{\sigma}_{\mathrm{CLS}}^2} \\ =& n \frac{\tilde{\sigma}_{\mathrm{CLS}}^2-\hat{\sigma}_{\mathrm{OLS}}^2}{\tilde{\sigma}_{\mathrm{CLS}}^2}=n\left(1-\frac{\hat{\sigma}_{\mathrm{OLS}}^2}{\tilde{\sigma}_{\mathrm{CLS}}^2}\right)=n\left(1-\frac{1}{1+\frac{q}{n-K} F}\right), \end{aligned}$$

故 CLS 估计量的得分统计量 S 是 F 比率的单增变换, 因而使用 S 作为检验统计量时有与 F 比率相同的方法. 在此意义上, 进而使用得分统计量 S 进行假设检验与似然比检验是等价的. 进一步的, 依据式 (6.67) 有

$$S=n\left(1-\frac{\hat{\sigma}_{\mathrm{OLS}}^2}{\tilde{\sigma}_{\mathrm{CLS}}^2}\right)=n\left(1-e^{-\frac{\lambda_{\mathrm{LR}}}{n}}\right),$$

而极限

$$\lim_{n\to\infty} n\left(1-e^{-\frac{A}{n}}\right)=\lim_{n\to\infty}\frac{1-e^{-\frac{A}{n}}}{\frac{1}{n}}=\lim_{n\to\infty}\frac{-\left(-\frac{A}{n}\right)}{\frac{1}{n}}=A,$$

故依据连续映射定理 (4.4.6) 我们推出得分统计量 S 的渐进分布为

$$S=n\left(1-e^{-\frac{\lambda_{\mathrm{LR}}}{n}}\right) \xrightarrow{d} \lambda_{\mathrm{LR}}=\chi^2(q), \quad n\to\infty,$$

这一结论可以推广至更为一般的得分统计量式 (6.76). 我们由此得到了得分检验.

定理 6.7.4 (得分检验/Lagrange 乘数检验)**.** *在假设 (5.1.1) 下, 对于模型 (3.12.1) 的回归系数 $\boldsymbol{\beta}$, 设参数 $\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})$, 其中 $\boldsymbol{h}:\mathbb{R}^{K\times 1}\to\mathbb{R}^{q\times 1}$, 就假设*

$$\mathrm{H}_0:\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{\theta}_0 \quad v.s. \quad \mathrm{H}_1:\boldsymbol{\theta}=\boldsymbol{h}(\boldsymbol{\beta})\neq\boldsymbol{\theta}_0$$

而言, 形如式 (6.76) 的得分统计量作为检验统计量, 则有

$$S=\frac{J^0\left(\tilde{\boldsymbol{\beta}}\right)}{\tilde{\sigma}^2}=\frac{\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\left(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\right)}{\tilde{\sigma}^2}\xrightarrow{d}\chi^2(m), \quad n\to\infty, \tag{6.77}$$

其中 $\tilde{\boldsymbol{\beta}}$ 是以 H_0 为限制的回归系数的估计量,$\tilde{\sigma}^2$ 相应的方差估计量,m 为非限制的参数空间与限制的参数空间的维度之差.. 我们称渐进显著性水平为 α的检验

$$\phi: 若 S>F^{-1}(1-\alpha) 则拒绝\ H_0$$

*为**得分检验** (score test), 其中 $F(\cdot)$ 是自由度为 q 的卡方分布的累计密度函数,$F^{-1}(\cdot)$ 为 $F(\cdot)$ 的逆函数. 得分检验的渐进 p 值为 $p=1-F(S)$. 得分检验又称为 **Lagrange 乘数检验** (Lagrange multiplier test), 简称为 **LM 检验**.*

定理 6.7.5 (线性约束的得分检验/Lagrange 乘数检验). 在假设 *(5.1.1)* 下, 对于模型 *(3.12.1)* 的回归系数 $\boldsymbol{\beta}$, 就假设

$$\mathrm{H}_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{\theta}_0 \quad v.s. \quad \mathrm{H}_1 : \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{\theta}_0,$$

其中 $\boldsymbol{R} \in \mathbb{R}^{q\times K}$,$\mathrm{rank}\boldsymbol{R} = q$, 以得分统计量为检验统计量, 则有

$$S = \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}\right)^{\mathrm{T}} \left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1} \left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}\right)}{\tilde{\sigma}^2_{\mathrm{CLS}}} \xrightarrow{d} \chi^2(q), \quad n \to \infty. \tag{6.78}$$

我们称渐进显著性水平为 α的检验

$$\phi : 若 S > F^{-1}(1-\alpha) 则拒绝\ H_0$$

为 H_0 的得分检验或 *Lagrange* 乘数检验, 其中 $F(\cdot)$ 是自由度为 q 的卡方分布的累计密度函数,$F^{-1}(\cdot)$ 为 $F(\cdot)$ 的逆函数. 似然比检验的渐进 p 值为 $p = 1 - F(S)$.

介绍三种假设检验的异同. 我们将目前介绍的假设检验总结为如下表6.1.

表 6.1: 线性回归模型的假设检验

	原假设 H_0	检验统计量		p 值	
		有限样本	大样本	有限样本	大样本
t 检验单系数	$\beta_k = \theta_0$	$T \sim t(n-K)$	$T \xrightarrow{d} \mathrm{N}(0,1)$	$p = 2(1 - F(\lvert T\rvert))$	$p = 2(1 - \Phi(\lvert T\rvert))$
F 检验	线性约束 $\boldsymbol{R\beta} = \boldsymbol{\theta}_0$	$F \sim F(q, n-K)$	$F \xrightarrow{d} \dfrac{\chi^2(q)}{q}$	$p = 1 - F\left(F_{\text{实现值}}\right)$	$p = 1 - F\left(F_{\text{实现值}}\right)$
方程的显著性检验	$\beta_2 = \cdots = \beta_K = 0$ (β_1 为截距项)	$F \sim F(K-1, n-K)$	$F \xrightarrow{d} \dfrac{\chi^2(K-1)}{K-1}$ 或 $nR^2 \xrightarrow{d} \chi^2(K-1)$	$p = 1 - F\left(F_{\text{实现值}}\right)$	$p = 1 - F\left(F_{\text{实现值}}\right)$ 或 $p = 1 - F\left(nR^2\right)$
Wald 检验	$\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ 线性约束 $\boldsymbol{R\beta} = \boldsymbol{\theta}_0$	线性约束下 $\dfrac{W^0}{q} = F \sim F(q, n-K)$	$W \xrightarrow{d} \chi^2(q)$,$W^0 \xrightarrow{d} \chi^2(q)$	$p = 1 - F\left(F_{\text{实现值}}\right)$	$p = 1 - F(W)$ $p = 1 - F\left(W^0\right)$
EMD 检验	线性约束 $\boldsymbol{R\beta} = \boldsymbol{\theta}_0$	$J* = W, J^0 = W^0$	$J^* = W \xrightarrow{d} \chi^2(q)$ $J^0 = W^0 \xrightarrow{d} \chi^2(q)$	同 Wald 检验	$p = 1 - F(J^*)$ $p = 1 - F\left(J^0\right)$
似然比检验	$\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ 线性约束 $\boldsymbol{R\beta} = \boldsymbol{\theta}_0$	$\lambda_{\mathrm{LR}} = n \ln\left(\hat{\sigma}^2/\tilde{\sigma}^2\right)$ $\lambda_{\mathrm{LR}} = n \ln\left(\hat{\sigma}^2_{\mathrm{CLS}}/\tilde{\sigma}^2_{\mathrm{ML}}\right)$	$\lambda_{\mathrm{LR}} \xrightarrow{d} \chi^2(m)$ $\lambda_{\mathrm{LR}} \xrightarrow{d} \chi^2(q)$	\	$p = 1 - F(\lambda_{\mathrm{LR}})$ $p = 1 - F(\lambda_{\mathrm{LR}})$
得分检验 (LM 检验)	$\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ 线性约束 $\boldsymbol{R\beta} = \boldsymbol{\theta}_0$	线性约束下 $S = J^0\left(\tilde{\boldsymbol{\beta}}_{\mathrm{CLS}}\right)/\tilde{\sigma}^2_{\mathrm{CLS}}$	$S \xrightarrow{d} \chi^2(m)$ $S \xrightarrow{d} \chi^2(q)$	\	$p = 1 - F(S)$ $p = 1 - F(S)$

6.8 Hausman 检验

6.9 异方差的检验与处理

在本章我们已经介绍了不少拓展的方法, 现在我们利用这些方法来研究经典 (3.2.1) 设定不成立时的应该如何处理. 本节研究的重点问题是条件同方差式 (3.7) 不成立——存在条件异方差式 (3.9) 下对于多元线性回归模型 (3.5.1) 应该如何估计.

在无自相关式 (3.8) 下, 异方差意味着 $\mathbb{E}\left(\varepsilon_i^2 \middle| \boldsymbol{X}\right) = \sigma_i^2 > 0$ 误差向量满足有

$$\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} = \operatorname{diag}\left(\sigma_1^2, \cdots, \sigma_n^2\right) = \boldsymbol{D},$$

这是我们在第三章中所熟悉的. 对于 OLS 估计量, 我们在第三章中已经推知

$$\begin{aligned}
&\operatorname{Var}\left(\boldsymbol{Y} \middle| \boldsymbol{X}\right) = \operatorname{Var}\left(\boldsymbol{\varepsilon} \middle| \boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right) = \boldsymbol{D}, \\
&\operatorname{Var}\left(Y_i \middle| \boldsymbol{X}\right) = \operatorname{Var}\left(\varepsilon_i \middle| \boldsymbol{X}\right) = \mathbb{E}\left(\varepsilon_i^2 \middle| \boldsymbol{X}\right) = \sigma_i^2 > 0, \\
&\operatorname{Var}\left(\hat{\boldsymbol{Y}} \middle| \boldsymbol{X}\right) = \boldsymbol{P}\operatorname{Var}\left(\boldsymbol{Y} \middle| \boldsymbol{X}\right)\boldsymbol{P} \overset{\text{异方差}}{=\!=\!=} \boldsymbol{P}\boldsymbol{D}\boldsymbol{P} \overset{\text{同方差}}{=\!=\!=} \boldsymbol{P}\sigma^2, \\
&\operatorname{Var}\left(\hat{Y}_i \middle| \boldsymbol{X}\right) \overset{\text{异方差}}{=\!=\!=} \left(\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}\right)_{ii} \overset{\text{同方差}}{=\!=\!=} h_{ii}\sigma^2, \\
&\operatorname{Var}\left(\hat{\boldsymbol{\varepsilon}} \middle| \boldsymbol{X}\right) = \operatorname{Var}\left(\boldsymbol{M}\boldsymbol{\varepsilon} \middle| \boldsymbol{X}\right) = \boldsymbol{M}\operatorname{Var}\left(\boldsymbol{\varepsilon} \middle| \boldsymbol{X}\right)\boldsymbol{M} \overset{\text{异方差}}{=\!=\!=} \boldsymbol{M}\boldsymbol{D}\boldsymbol{M} \overset{\text{同方差}}{=\!=\!=} \boldsymbol{M}\sigma^2, \\
&\operatorname{Var}\left(\hat{\varepsilon}_i \middle| \boldsymbol{X}\right) \overset{\text{异方差}}{=\!=\!=} \left(\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\right)_{ii} \overset{\text{同方差}}{=\!=\!=} \left(1 - h_{ii}\right)\sigma^2, \\
&\operatorname{Var}\left(\hat{\boldsymbol{\beta}} \middle| \boldsymbol{X}\right) = \boldsymbol{V}_{\hat{\boldsymbol{\beta}}} \overset{\text{异方差}}{=\!=\!=} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \overset{\text{同方差}}{=\!=\!=} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2,
\end{aligned}$$

可见在同方差下,OLS 估计量的不少问题都得到了简化. 异方差式 (3.9) 的存在, 并不影响 OLS 估计量的无偏性和相合性, 但是形如定理 (3.5.3) 的 Gauss-Markov 定理不成立, 因而 OLS 估计量不是最佳线性无偏估计量.

而对于 OLS 估计量的标准误, 我们在第3.11节中已经介绍了异方差稳健的方差估计量式 (3.56) 至式 (3.59), 且在第5.2节中给出了其相应的渐进协方差矩阵估计量式 (5.13) 至式 (5.16), 并总结为了表5.1. 只要选取正确的方差估计量, 由标准误所构建的区间估计、假设检验都没有问题. 而同方差是异方差的特例, 因而在现在的实证研究中我们更多的使用是 HC 这类的稳健的协方差矩阵估计量.

下面我们介绍对样本数据是否存在异方差式 (3.9) 的假设检验.

我们首先介绍 Breusch–Pagan 检验. 我们在此前介绍了线性投影模型 (3.3.1) 与线性回归模型 (3.5.1) 的区别, 其核心的问题在于最优预测量 $\mathbb{E}\left(Y|\boldsymbol{X}\right)$ 是否为线性 CEF.

我们依据定义 (2.1.4), 类似地假设条件方差函数 $\sigma^2(\mathbf{X})$ 为线性函数

$$\sigma^2(\mathbf{X}) = \mathbb{E}\left(\varepsilon^2 \middle| \mathbf{X}\right) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\gamma} = \gamma_0 + X_1\gamma_1 + \cdots + X_K\gamma_K, \tag{6.79}$$

其中 $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_K)^{\mathrm{T}}$ 是未知的参数, 依据定义 (2.1.4) 有 $\sigma^2 = \mathbb{E}\sigma^2(\mathbf{X}) = (\mathbb{E}\mathbf{X})^{\mathrm{T}}\boldsymbol{\gamma}$.Breusch-Pagantest 检验便是假设了形如式 (6.79) 的条件方差函数, 显然其符合式 (3.9). 就是否存在异方差, 则可以同方差式 (3.7) 作为原假设, 提出假设检验的问题为

$$\mathrm{H}_0 : \gamma_1 = \cdots = \gamma_K = 0 \quad v.s. \quad \mathrm{H}_1 : \gamma_1, \cdots, \gamma_K\text{不全为0}.$$

在理想的情况下, 如果误差 ε_i 是已知的, 那么我们便可以依据式 (6.79) 建立回归模型

$$\varepsilon_i^2 = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} + u_i = \gamma_0 + X_{i1}\gamma_1 + \cdots + X_{iK}\gamma_K + u_i, \tag{6.80}$$

其中 u_i 为误差项. 通过对回归模型 (6.80) 进行依据方程的显著性检验, 我们便可诊断模型 (3.5.1) 是否存在异方差的问题. 可惜的是现实中 ε_i 不可观察, 因而使用 OLS 估计量的残差 $\hat{\varepsilon}_i$ 代替误差 ε_i, 则得到回归模型

$$\hat{\varepsilon}_i^2 = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} + u_i = \gamma_0 + X_{i1}\gamma_1 + \cdots + X_{iK}\gamma_K + u_i, \tag{6.81}$$

依据定理 (5.4.5) 则可以以 nR^2 作为检验统计量, 由此我们得到了异方差的 Breusch-Pagan 检验.

定理 6.9.1 (Breusch-Pagan 检验)**.** *在假设 (5.1.1) 下, 对于模型 (3.5.1) 是否存在异方差问题, 假设有形如式 (6.79) 的线性条件方差函数, 则同方差有原假设*

$$\mathrm{H}_0 : \gamma_1 = \cdots = \gamma_K = 0,$$

依据定理 (5.4.5), 以回归模型式 (6.81) 的拟合优度 R^2 作为检验统计量, 则有

$$nR^2 \xrightarrow{d} \chi^2(K), \quad n \to \infty.$$

我们称渐进显著性水平为 α的检验

$$\phi : \text{若} nR^2 > F^{-1}(1-\alpha)\text{则拒绝 } H_0,$$

为 ***Breusch–Pagan*** *检验 (Breusch-Pagan test), 简称为* ***BP*** *检验 (BP test), 其中 F 是自由度为 K 的卡方分布的累计密度函数,F^{-1} 为 F 的逆函数. 检验 ϕ 的 p 值为 $p = 1 - F(nR^2)$.*

Breusch-Pagan 检验的核心在于假设了条件方差函数是线性的, 因而其检验的异方差的形式只能是一次的. 在实践中, 式 (6.81) 不要求一定包含有所有的解释变量, 而是根据需要去选择以确定可能的异方差的形式.

异方差的另一种常用检验方法是 White 检验, 其基于 OLS 估计量的渐进性质. 在第5.2节中, 我们知道异方差下渐进协方差矩阵式 (5.10) 是夹心形式, 类似于式 (5.3), 其中间的矩阵 $\boldsymbol{\Omega}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right)$ 有估计量

$$\hat{\boldsymbol{\Omega}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2=\frac{1}{n}\begin{pmatrix}\sum\limits_{i=1}^{n}X_{i1}^2\hat{\varepsilon}_i^2 & \sum\limits_{i=1}^{n}X_{i1}X_{i2}\hat{\varepsilon}_i^2 & \cdots & \sum\limits_{i=1}^{n}X_{i1}X_{iK}\hat{\varepsilon}_i^2\\ \sum\limits_{i=1}^{n}X_{i2}X_{i1}\hat{\varepsilon}_i^2 & \sum\limits_{i=1}^{n}X_{i2}^2\hat{\varepsilon}_i^2 & \cdots & \sum\limits_{i=1}^{n}X_{i2}X_{iK}\hat{\varepsilon}_i^2\\ \vdots & \vdots & & \vdots\\ \sum\limits_{i=1}^{n}X_{iK}X_{i1}\hat{\varepsilon}_i^2 & \sum\limits_{i=1}^{n}X_{iK}X_{i2}\hat{\varepsilon}_i^2 & \cdots & \sum\limits_{i=1}^{n}X_{iK}^2\hat{\varepsilon}_i^2\end{pmatrix},$$

White(1980) 指出, 在条件同方差的情形下以及总体误差 ε 独立于 $\mathbf{X}$ 时, 当 $n\to\infty$ 时, 则有

$$\hat{\boldsymbol{\Omega}}-\hat{\boldsymbol{Q}}_{\mathbf{XX}}s^2=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}s^2=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i^2-s^2\right)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\xrightarrow{P}\boldsymbol{O}_{K\times K},$$

则可以利用 $\hat{\boldsymbol{\Omega}}-\hat{\boldsymbol{Q}}_{\mathbf{XX}}s^2$ 的差异程度来检验是否存在异方差. 由于 $\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$ 是实对称方阵, 为了简便计算, 对于解释变量的总体 $\mathbf{X}$, 取矩阵

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}=\begin{pmatrix}X_1^2 & X_1X_2 & \cdots & X_1X_K\\ X_2X_1 & X_2^2 & \cdots & X_2X_K\\ \vdots & \vdots & & \vdots\\ X_KX_1 & X_KX_2 & \cdots & X_K^2\end{pmatrix}$$

的非重复元素, 记列向量

$$\boldsymbol{\psi}=\left(X_1^2,X_1X_2,X_2^2,\cdots,X_1X_K,X_2X_K,\cdots,X_K^2\right)^{\mathrm{T}}\in\mathbb{R}^{K(K+1)/2\times1},$$

则有样本个体 $\boldsymbol{\psi}_i=\left(X_{i1}^2,X_{i1}X_{i2},X_{i2}^2,\cdots,X_{i1}X_{iK},X_{i2}X_{iK},\cdots,X_{iK}^2\right)^{\mathrm{T}}$.

在一定的条件下, 依据大数定律和中心极限定理,White(1980) 证明了

$$\boldsymbol{c}=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i^2-s^2\right)\boldsymbol{\psi}_i\xrightarrow{P}\mathbf{0}_{K(K+1)/2\times1},\quad\sqrt{n}\boldsymbol{c}\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{B}\right),\quad n\to\infty,$$

其中 $\boldsymbol{B}$ 是列向量 $\boldsymbol{c}$ 的渐进协方差矩阵矩阵, 使用 $\boldsymbol{B}$ 的估计量 $\hat{\boldsymbol{B}}$ 可以得到

$$n\boldsymbol{c}\hat{\boldsymbol{B}}\boldsymbol{c}\xrightarrow{d}\chi^2\left(\frac{K\left(K+1\right)}{2}\right),$$

这即意味着 $n\boldsymbol{c}\hat{\boldsymbol{B}}\boldsymbol{c}$ 可以作为检验异方差的检验统计量. 如果仅仅是因为提出了 $n\boldsymbol{c}\hat{\boldsymbol{B}}\boldsymbol{c}$, 那么 White 检验并不会广泛使用. 进一步的,White(1980) 指出了对于 $\boldsymbol{\psi}_i$, 考虑回归模型

$$\hat{\varepsilon}_i^2=\gamma_0+\boldsymbol{\psi}_i^{\mathrm{T}}\boldsymbol{\gamma}+u_i,\tag{6.82}$$

则模型式 (6.82) 的显著性检验的检验统计量等价地有

$$nR^2 \xrightarrow{d} \chi^2\left(\frac{K\left(K+1\right)}{2}\right), \quad n \to \infty,$$

于是我们得到了 White 检验.

定理 6.9.2 (White 检验). *在假设 (5.1.1) 下, 对于模型 (3.5.1) 是否存在异方差问题, 同方差的原假设*

$$\mathrm{H}_0 : \mathbb{E}\left(\varepsilon_i^2 \middle| \boldsymbol{X}\right) = \sigma^2$$

等价于对回归模型式 (6.82) 的方程的显著性检验的原假设

$$\mathrm{H}_0 : \gamma_1 = \cdots = \gamma_{K(K+1)/2} = 0.$$

依据定理 (5.4.5), 以回归模型式 (6.82) 的拟合优度 R^2 作为检验统计量, 则有

$$nR^2 \xrightarrow{d} \chi^2\left(\frac{K\left(K+1\right)}{2}\right), \quad n \to \infty.$$

我们称渐进显著性水平为 α的检验

$$\phi : \text{若} nR^2 > F^{-1}\left(1-\alpha\right) \text{则拒绝 } H_0,$$

*为 **White** 检验 (White test), 其中 F 是自由度为 K 的卡方分布的累计密度函数,F^{-1} 为 F 的逆函数. 检验 ϕ 的 p 值为 $p = 1 - F\left(nR^2\right)$.*

不难发现,White 检验即是 Breusch–Pagan 检验的拓展形式, 其检验了异方差的来源不仅含有解释变量的一次项, 还包含有二次项与交互项. 但无论是 White 检验还是 Breusch–Pagan 检验, 其依旧可以归类到假设条件期望函数式 (6.79) 是关于参数 $\boldsymbol{\gamma}$ 线性的.

对于 Breusch–Pagan 检验和 White 检验, 拒绝原假设我们便有接受模型存在异方差的问题, 但不能拒绝原假设意味着什么?Breusch–Pagan 检验仅涉及了解释变量的一次项, 不能拒绝原假设只能说明这次检验所设定的异方差形式是得不到支持的. 在 Breusch–Pagan 检验之上 White 检验则进一步的涵盖了二次型和交互项, 但更高次形式的异方差问题仍不能够排除. 但就 White 检验的原理而言, 其实质上通过 $\hat{\boldsymbol{\Omega}} - \hat{\boldsymbol{Q}}_{\mathbf{XX}} s^2$ 检验了原假设 $\mathrm{H}_0 : \boldsymbol{\Omega} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right) = \boldsymbol{Q}_{\mathbf{XX}}\sigma^2$, 如果不能拒绝原假设, 那么便也意味着不能拒绝 $\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right) = \boldsymbol{Q}_{\mathbf{XX}}\sigma^2$, 在这种情况下依据式 (5.10) 即就退化了式 (5.11), 因而使用同方差的协方差矩阵估计量 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0 = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} s^2$ 是合理的.

我们知道,GLS 估计量能够用于处理具有任意形式的协方差矩阵的误差向量 $\boldsymbol{\varepsilon}$, 并得到最为有效的线性估计量——前提是条件协方差矩阵 $\boldsymbol{\Sigma}$ 是已知的. 当条件协方差矩阵

$\boldsymbol{\Sigma}$ 不存在自相关时, 即有条件协方差矩阵 $\boldsymbol{\Sigma}$ 为对角阵 $\boldsymbol{\Sigma}=\boldsymbol{D}=\operatorname{diag}\left(\sigma_1^2,\cdots,\sigma_n^2\right)$, 这时的 GLS 估计量即为 WLS 估计量. 现在我们来介绍可行的 WLS 估计量. 同样依据 MSE 最小的思想, 条件方差函数 $\sigma^2(\mathbf{X})$ 的最优预测是条件期望函数, 故可以使用 Breusch–Pagan 检验的回归模型式 (6.81) 得到 $\boldsymbol{\Sigma}$ 的估计量

$$\hat{\boldsymbol{\Sigma}}_1=\hat{\boldsymbol{D}}_1=\operatorname{diag}\left(\mathbf{X}_1^{\mathrm{T}}\hat{\boldsymbol{\gamma}},\cdots,\mathbf{X}_n^{\mathrm{T}}\hat{\boldsymbol{\gamma}}\right),$$

但这样估计一个问题便是 $\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\gamma}}$ 可能会出现负数的情形. 在实践中, 通常利用指数函数假设条件期望函数为

$$\sigma_i^2=\exp\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\delta}\right)=\exp\left(\delta_0+X_{i1}\delta_1+\cdots+X_{iK}\delta_K\right),$$

取对数后通过回归模型

$$\ln\sigma_i^2=\delta_0+X_{i1}\delta_1+\cdots+X_{iK}\delta_K+\mu_i$$

到了参数 $\boldsymbol{\delta}$ 的 OLS 估计量 $\hat{\boldsymbol{\delta}}$, 则有每个元素非负的协方差矩阵估计量

$$\hat{\boldsymbol{\Sigma}}_2=\hat{\boldsymbol{D}}_2=\operatorname{diag}\left(\exp\left(\mathbf{X}_1^{\mathrm{T}}\hat{\boldsymbol{\delta}}\right),\cdots,\exp\left(\mathbf{X}_n^{\mathrm{T}}\hat{\boldsymbol{\delta}}\right)\right).$$

6.10 自相关的检验与处理

当球形扰动项不满足无自相关式 (3.8), 则有

$$\mathbb{E}\left(\varepsilon_i\varepsilon_j\middle|\boldsymbol{X}\right)=\rho_{ij},\quad i,j=1,\cdots,n且i\neq j, \tag{6.83}$$

其中 ρ_{ij} 不全部为 0. 在严格外生性式 (3.5) 和同方差式 (3.7) 下, 这意味着误差向量 $\boldsymbol{\varepsilon}$ 的条件协方差矩阵为

$$\operatorname{Var}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)=\begin{pmatrix}\sigma^2 & \rho_{12} & \cdots & \rho_{1n}\\ \rho_{21} & \sigma^2 & \cdots & \rho_{2n}\\ \vdots & \vdots & & \vdots\\ \rho_{n1} & \rho_{n2} & \cdots & \sigma^2\end{pmatrix},$$

此时, 对多元线性回归模型 (3.5.1) 的 OLS 估计量仍是无偏的、相合的 (一致的) 渐近正态的, 但依赖于球形扰动项的 Gauss-Markov 定理不再成立, 因而 OLS 估计量不再是最优无偏线性估计量. 同时球形扰动项下的协方差矩阵估计量 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^2$ 也存在较大偏差, 由其得到标准误 SE 也不是一个良好的估计量.

自相关问题多存在于随机序列分析中, 我们在此问题的中使用下标 t 代替 i, 即表示第 t 期的观测值.

对于自相关, 通常有 4 种检验的方法: 残差散点图、Breusch-Godfrey 检验、Q 检验和 DW 检验.

由于残差 $\hat{\varepsilon}_t$ 是误差 ε_t 的估计量, 通过绘制 $\hat{\varepsilon}_t - \hat{\varepsilon}_{t-p}$ 即可观察误差 ε_i 与 ε_{t-p} 的相关关系. 另一种绘图方法则是计算扰动项相关系数, 考虑 i 阶自相关系数

$$\rho_i = \frac{\operatorname{Cov}(\varepsilon_t, \varepsilon_{t-i})}{\sqrt{\operatorname{Var}(\varepsilon_t)\operatorname{Var}(\varepsilon_{t-i})}} = \frac{\mathbb{E}(\varepsilon_t \varepsilon_{t-i})}{\operatorname{Var}(\varepsilon_t)}, \tag{6.84}$$

则有样本估计量

$$\hat{\rho}_i = \frac{\sum_{t=i+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-i}}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}, \tag{6.85}$$

则绘制了 $(i, \hat{\rho}_i)$ 的置信带图称为**自相关图** (coorelogram).

Breusch-Godfrey 检验是针对 p 阶自回归的扰动项的检验方法. 假设误差项 ε_i 满足有 AR(p) 过程

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \cdots + \rho_p \varepsilon_{t-p} + u_t, \tag{6.86}$$

其中 u_t 为白噪声, 且有 $\operatorname{Cov}(\varepsilon_{t-1}, u_t) = 0$. 则以无自相关式 (3.8) 为原假设, 即等价地有

$$\mathrm{H}_0 : \rho_1 = \cdots = \rho_p = 0 \quad v.s. \quad \mathrm{H}_1 : \rho_1, \cdots, \rho_p \text{不全为0},$$

使用可获得的残差 $\hat{\varepsilon}_i$ 替代不可观察的误差 ε_i, 则可使用方程的显著性检验无自相关的原假设. 进一步的, 考虑到残差 $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{X}$ 是关于解释变量 $\boldsymbol{X}$ 的线性函数, 为了确保不存在遗漏变量偏差 (见第6.11节), 则有回归模型

$$\hat{\varepsilon}_t = \alpha_0 + \mathbf{X}_t^{\mathrm{T}} \boldsymbol{\alpha} + \rho_1 \hat{\varepsilon}_{t-1} + \cdots + \rho_p \hat{\varepsilon}_{t-p} + u_t, \quad t > p, \tag{6.87}$$

对模型式 (6.87) 检验原假设 $\mathrm{H}_0 : \rho_1 = \cdots = \rho_p = 0$ 即可.

定理 6.10.1 (Breusch-Godfrey 检验)**.** *在假设 (5.1.1) 下, 对于模型 (3.5.1) 是否存在自相关问题, 假设误差项 ε_i 有形如式 (6.86) 的 p 阶自回归过程, 则无自相关有原假设*

$$\mathrm{H}_0 : \rho_1 = \cdots = \rho_p = 0,$$

以回归模型式 (6.87) 的拟合优度 R^2 作为检验统计量, 则有

$$(n-p) R^2 \xrightarrow{d} \chi^2(p), \quad n \to \infty.$$

我们称渐进显著性水平为 α的检验

$$\phi : \text{若 } (n-p) R^2 > F^{-1}(1-\alpha) \text{ 则拒绝 } H_0,$$

为 **Breusch-Godfrey 检验** *(Breusch-Godfrey test)*, 简称为 **BG 检验** *(BG test)*, 其中 F 是自由度为 p 的卡方分布的累计密度函数,F^{-1} 为 F 的逆函数. 检验 ϕ 的 p 值为 $p=1-F\left((n-p)R^2\right)$.

由于模型式 (6.87) 中包含了 $\hat{\varepsilon}_{t-p}$, 则需要有 $t>p$, 因而模型式 (6.87) 的 OLS 估计量和 BG 检验损失了 p 个样本个体.Davidson 和 MacKinnon 提出使用残差的期望值 $\mathbb{E}\hat{\varepsilon}_i=\mathbb{E}\varepsilon_i=0$ 来弥补损失值, 即有

$$\hat{\varepsilon}_{t-p}=\begin{cases}0, t\leqslant p,\\ \hat{\varepsilon}_{t-p}, t>p,\end{cases}$$

此时则为模型式 (6.87) 和 BG 检验补充了 p 个观测值, 这也是Stata所采用的方法.BG 检验又称为自相关的 Lagrange 乘数检验.

Q 检验则是检验了自相关系数是否为零, 考虑形如式 (6.84) 的 p 自相关系数, 则有原假设

$$\mathrm{H}_0:\rho_1=\cdots=\rho_p=0,$$

在 H_0 下由残差 $\hat{\boldsymbol{\varepsilon}}$ 构建的自回归系数的估计量 $\hat{\boldsymbol{\rho}}=(\hat{\rho}_1,\hat{\rho}_1\cdots,\hat{\rho}_p)^{\mathrm{T}}$ 是渐近正态性的, 由此可以进行假设检验.Box 和 Pierce 提出了 Q 统计量, 当 $n\to\infty$ 时

$$Q_{\mathrm{BP}}=n\sum_{i=1}^{p}\hat{\rho}_i^2\xrightarrow{d}\chi^2(p),$$

在此之上 Ljung 和 Box 改进了 Q 统计量, 同样在 $n\to\infty$ 时

$$Q_{\mathrm{LB}}=n(n+2)\sum_{i=1}^{p}\frac{\hat{\rho}_i^2}{n-i}\xrightarrow{d}\chi^2(p),$$

则得到了自相关的 Q 检验.

定理 6.10.2 (自相关的 Q 检验). *在假设 (5.1.1) 下, 对于模型 (3.5.1) 是否存在自相关问题, 考虑形如式 (6.84) 的自回归系数 $\rho_i, i=1\cdots,p$, 则无自相关有原假设*

$$\mathrm{H}_0:\rho_1=\cdots=\rho_p=0,$$

以自回归系数的估计量 $\hat{\rho}_i$ 构造检验统计量

$$Q_{\mathrm{BP}}=n\sum_{i=1}^{p}\hat{\rho}_i^2\xrightarrow{d}\chi^2(p),\quad n\to\infty, \tag{6.88}$$

$$Q_{\mathrm{LB}}=n(n+2)\sum_{i=1}^{p}\frac{\hat{\rho}_i^2}{n-i}\xrightarrow{d}\chi^2(p),\quad n\to\infty, \tag{6.89}$$

我们由渐进显著性水平为 α的检验

$$\phi_{\mathrm{BP}}: 若Q_{\mathrm{BP}} > F^{-1}(1-\alpha) 则拒绝\ H_0,$$
$$\phi_{\mathrm{LB}}: 若Q_{\mathrm{LB}} > F^{-1}(1-\alpha) 则拒绝\ H_0,$$

其中 F 是自由度为 p 的卡方分布的累计密度函数,F^{-1} 为 F 的逆函数. 我们称检验 ϕ_{BP} 为 ***Box-Pierce Q* 检验** *(Box-Pierce Q test)*, 简称为 ***BP Q* 检验** *(BP Q test)*, 其 p 值为 $p = 1 - F(Q_{\mathrm{BP}})$; 检验 ϕ_{LB} 为 ***Ljung-Box Q* 检验** *(Ljung-Box Q test)*, 简称为 ***LB Q* 检验** *(LB Q test)*, 其 p 值为 $p = 1 - F(Q_{\mathrm{LB}})$

在大样本下,Q_{BP} 和 Q_{LB} 具有相同的渐进分布, 这是因为当 $n \to \infty$ 时

$$\begin{aligned} Q_{\mathrm{LB}} - Q_{\mathrm{BP}} =& n(n+2)\sum_{i=1}^{p}\frac{\hat{\rho}_i^2}{n-i} - n\sum_{i=1}^{p}\hat{\rho}_i^2 = n\sum_{i=1}^{p}\left[\frac{(n+2)}{n-i} - 1\right]\hat{\rho}_i^2 \\ =& n\sum_{i=1}^{p}\frac{2+i}{n-i}\hat{\rho}_i^2 = \sum_{i=1}^{p}\frac{n(2+i)}{n-i}\hat{\rho}_i^2 \xrightarrow{P} \sum_{i=1}^{p}(2+i)\cdot 0 = 0. \end{aligned}$$

而在实践中我们不可能具有无限样本, 而是近似的去使用渐进分布.Q_{BP} 在有限样本下与渐进分布的差距较大, 使得假设检验的推断更可能犯错, 因而 Ljung 和 Box 提出了有限样本下分布更好的 Q_{LB}.

无论是 Breusch-Godfrey 检验或是自相关的 Q 检验, 都面临着自相关阶数 p 的选择问题. 一种方法即是基于第6.12节的信息准则, 而`Stata`的采用的是 $p = \min\{\lfloor n/2 \rfloor - 2, 40\}$, 其中 $\lfloor\cdot\rfloor$ 表示向下取整函数,$\lfloor n/2 \rfloor$ 即为不超过 $n/2$ 的最大整数.

除了上述的检验方法外, 还有一个传统的检验方法称为 **Durbin–Watson 检验** (Durbin–Watson test), 简称为 **DW 检验** (DW test). 在严格外生性下,Durbin–Watson 检验构造统计量

$$d = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} = \frac{\sum_{t=2}^{n}\hat{\varepsilon}_t^2 + \sum_{t=2}^{n}\hat{\varepsilon}_{t-1}^2 - 2\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} \xrightarrow{P} 2(1-\rho_1), \quad n \to \infty, \tag{6.90}$$

其中 ρ_1 为形如式 (6.84) 的一阶自回归系数. 我们称式 (6.90) 为 **Durbin–Watson 统计量** (Durbin–Watson statistic), 简称 **DW 统计量** (DW statistic). 统计量 d 的渐进分布为 $2(1-\rho_1)$, 而相关系数的取值范围为 $[-1, 1]$, 因而 d 的取值范围为 $[0, 4]$.

DW 统计量 d 即是对一阶自回归系数的参数的估计量, 因而可以使用其假设检验, 然而不幸的是,d 渐进分布的推导十分复杂, 在给定了渐进显著性水平为 α 的情形下, 则图6.1展示了 DW 检验的检验结果.

DW 检验有两个临界值 $d_{\alpha,L}$ 和 $d_{\alpha,U}$, 依据统计量 d 所落在区间则有:

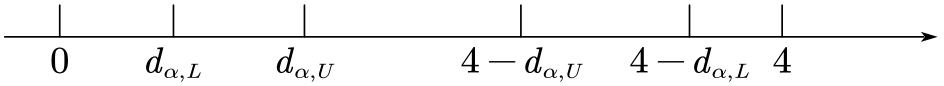

图 6.1: DW 统计量的检验结果

(a) 若 $d \leqslant d_{\alpha,L}$, 则拒绝原假设 $\mathrm{H}_0 : \rho_1 \leqslant 0$, 认为存在一阶正自相关;

(b) 若 $d \geqslant 4 - d_{\alpha,L}$, 则拒绝原假设 $\mathrm{H}_0 : \rho_1 \geqslant 0$, 认为存在一阶负自相关;

(c) 若 $d_{\alpha,U} < d < 4 - d_{\alpha,U}$, 则不能拒绝原假设 $\mathrm{H}_0 : \rho_1 = 0$, 没有证据拒绝无自相关;

(d) 若 $d \in (d_{\alpha,L}, d_{\alpha,U}] \cup [4 - d_{\alpha,U}, 4 - d_{\alpha,L})$, 则检验没有结果.

较其他检验而言,DW 检验不仅方法复杂, 而且还存在无法得出结论的情形, 现在已经较少使用.

下面我们介绍自相关的处理方法, 通常而言有两种方法: 使用 HAC 标准误和准差分法.

正如面临异方差时, 我们使用了异方差一致的标准误, 针对自相关则有**异方差自相关一致的协方差矩阵估计量** (heteroskedasticity and autocorrelation consistent covariance matrix estimators). 依据式 (6.54), 则 OLS 估计量的关键在于对存在自相关的协方差矩阵 $\boldsymbol{\Sigma}$ 的估计, 由于 $\boldsymbol{\Sigma}$ 为实对称方阵, 我们将其分解为对角矩阵 $\boldsymbol{D}$ 和对角元素为 0 实对称方阵 $\boldsymbol{T}$, 即 $\boldsymbol{\Sigma} = \boldsymbol{D} + \boldsymbol{T}$, 那么则有

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{X} = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{D} + \boldsymbol{T})\boldsymbol{X} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{X},$$

而前者 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}$ 即是异方差一致的协方差矩阵中夹心部分所要估计的矩阵, 我们只需考虑对于 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{X}$ 的估计量,Newey 和 West 给出了一个估计量为

$$\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{T}}\boldsymbol{X} = \frac{1}{n}\sum_{j=1}^{p}\sum_{t=p+1}^{n}\left(1 - \frac{j}{p+1}\right)\hat{\varepsilon}_t\hat{\varepsilon}_{t-j}\left(\mathbf{X}_t\mathbf{X}_{t-j}^{\mathrm{T}} + \mathbf{X}_{t-j}\mathbf{X}_t^{\mathrm{T}}\right),$$

则有 HAC 协方差矩阵估计量为

$$\begin{aligned}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HAC}} =& \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{X}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^2 \\ =& \hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC}} + \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{X}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^2,\end{aligned} \tag{6.91}$$

使用 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HAC}}$ 去计算标准误, 便可缓解自相关和异方差带来的估计的偏差.

处理异方差的另一种方法是准差分法, 其假设扰动项符合一阶自回归 AR(1) : $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, 其中 u_t 为白噪声且与 ε_{t-1} 完全不无关,$|\rho| < 1$. 对于回归模型

$$Y_t = \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_t = \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\beta} + \rho\varepsilon_{t-1} + u_t, \tag{6.92}$$

只要 $\rho\varepsilon_{t-1}+u_t$ 能去掉 $\rho\varepsilon_{t-1}$, 那么白噪声 u_t 即是符合球形扰动项的. 为此考虑 $t-1$ 期个体的回归方程, 将其乘以 ρ 得

$$\rho Y_{t-1}=\rho\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}+\rho\varepsilon_{t-1},\quad t>1, \tag{6.93}$$

将式 (6.92) 减去式 (6.93), 即有

$$(1-\rho)\left(Y_t-Y_{t-1}\right)=(1-\rho)\left(\mathbf{X}_t^{\mathrm{T}}-\mathbf{X}_{t-1}^{\mathrm{T}}\right)\boldsymbol{\beta}+u_t,$$

令

$$\tilde{Y}_t=(1-\rho)\left(Y_t-Y_{t-1}\right),\quad \tilde{\mathbf{X}}_t^{\mathrm{T}}=(1-\rho)\left(\mathbf{X}_t^{\mathrm{T}}-\mathbf{X}_{t-1}^{\mathrm{T}}\right),\quad \tilde{\varepsilon}_t=u_t$$

则回归模型

$$\tilde{Y}_t=\tilde{\mathbf{X}}_t^{\mathrm{T}}\boldsymbol{\beta}+\tilde{\varepsilon}_t,\quad t>1, \tag{6.94}$$

是满足球形扰动项的, 因而可以排除满足 AR (1) 的自相关对于 OLS 估计量的影响, 这种估计方法被称为 **Cochrane–Orcutt 估计量** (Cochrane–Orcutt estimation), 简称为 **CO 估计** (CO estimation). 需要指出, 式 (6.94) 只有 $n-1$ 个样本, 因而没能充分利用样本信息. 一个自然的想法是将 $t=1$ 的式 (6.92) 去补充, 但这样回归模型的扰动项的协方差矩阵为

$$\operatorname{Var}\left(\tilde{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\middle|\boldsymbol{X}\right)=\begin{pmatrix}\operatorname{Var}\left(\varepsilon_1\middle|\boldsymbol{X}\right) & 0 & \cdots & 0\\ 0 & \sigma_u^2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \sigma_u^2\end{pmatrix},$$

从而又导致了异方差的问题. 对一阶自回归取条件方差 $\operatorname{Var}\left(\varepsilon_t^2\middle|\boldsymbol{X}\right)=\rho^2\operatorname{Var}\left(\varepsilon_{t-1}^2\middle|\boldsymbol{X}\right)+\sigma_u^2$, 考虑 $t\to\infty$ 时, 满足有

$$\sigma_\varepsilon^2=\rho^2\sigma_\varepsilon^2+\sigma_u^2\Rightarrow\sigma_\varepsilon^2=\frac{\sigma_u^2}{1-\rho^2},$$

用 σ_ε^2 代替 $\operatorname{Var}\left(\varepsilon_1\middle|\boldsymbol{X}\right)$, 则协方差矩阵满足有

$$\boldsymbol{\Sigma}\sigma_u^2=\begin{pmatrix}\dfrac{1}{1-\rho^2} & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix}\sigma_u^2,\quad \boldsymbol{\Sigma}^{1/2}=\begin{pmatrix}\dfrac{1}{\sqrt{1-\rho^2}} & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix},$$

这样依据 GLS 估计量的思想我们则有满足经典假设的回归模型

$$\begin{aligned}&t=1\text{期}:\sqrt{1-\rho^2}\tilde{Y}_1=\sqrt{1-\rho^2}\tilde{\mathbf{X}}_1^{\mathrm{T}}\boldsymbol{\beta}+\sqrt{1-\rho^2}\tilde{\varepsilon}_1,\\ &t>1\text{期}:\tilde{Y}_t=\tilde{\mathbf{X}}_t^{\mathrm{T}}\boldsymbol{\beta}+\tilde{\varepsilon}_t,\end{aligned} \tag{6.95}$$

采用回归模型 (6.95) 的估计方法称为 **Prais–Winsten 估计量** (Prais–Winsten estimation), 简称为 **PW 估计** (PW estimation).

需要指出, 无论是 CO 估计还是 PW 估计, 其需要知道 AR (1) 的自回归系数 ρ, 因而在实践中通常使用式 (6.90) 或式 (6.85) 去先计算估计量 $\hat{\rho}$. 为了提供估计的精度, 通常采用迭代法以计算 $\hat{\rho}$.

6.11 遗漏变量与无关变量

假设总体回归模型被设定正确为

$$Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \quad \mathbb{E}(\mathbf{X}\varepsilon) = \mathbf{0}, \tag{6.96}$$

我们将解释变量分块为

$$Y = \mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{X}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon = \left(\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}}\right)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \varepsilon,$$

其中

$$\mathbf{X} = \begin{pmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix}.$$

对于总体回归模型式 (6.96), 则总体参数回归系数 $\boldsymbol{\beta}$ 满足有 $\boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y)$. 现考虑遗漏了解释变量 $\mathbf{X}_2$ 的设定错误的总体模型

$$Y = \mathbf{X}_1^{\mathrm{T}}\boldsymbol{\gamma}_1 + u, \quad \mathbb{E}(\mathbf{X}u) = \mathbf{0}, \tag{6.97}$$

则参数 $\boldsymbol{\gamma}$ 满足有

$$\begin{aligned}
\boldsymbol{\gamma}_1 =& \left(\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}_1 Y) = \left(\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{X}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon\right)\right) \\
=& \left(\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{X}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon\right)\right) \\
=& \left(\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}}\right)\right)^{-1}\left[\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}}\right)\boldsymbol{\beta}_1 + \mathbb{E}\left(\mathbf{X}_1\mathbf{X}_2^{\mathrm{T}}\right)\boldsymbol{\beta}_2 + \mathbb{E}(\mathbf{X}_1\varepsilon)\right] \\
=& \boldsymbol{\beta}_1 + \left(\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_2^{\mathrm{T}}\right)\boldsymbol{\beta}_2 \\
=& \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2,
\end{aligned}$$

其中 $\boldsymbol{\Gamma}_{12} = \left(\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}_1\mathbf{X}_2^{\mathrm{T}}\right)$. 类比与回归系数 $\boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y)$, 则可知 $\boldsymbol{\Gamma}_{12}$ 的含义为 $\mathrm{X}_2\mathrm{z}$ 中各分量对 $\mathbf{X}_1$ 的回归系数的矩阵. 一般地, 我们将设定正确的总体回归模型式 (6.96) 称为**长回归** (long regression), 设定错误的总体回归模型式 (6.97) 称为**短回归** (short regression), 而

$$\boldsymbol{\gamma}_1 - \boldsymbol{\beta}_1 = \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2 \tag{6.98}$$

则被称为**遗漏变量偏差** (omitted variable bias).

在实证中, 我们是使用样本去估计总体参数, 对于设定错误的模型式 (6.97) 而言, 无论估计的有多准确, 所得到仍是参数 $\boldsymbol{\gamma}_1$ 的估计量而非真正的参数 $\boldsymbol{\beta}_1$ 的估计. 不难得知, 当且仅当遗漏变量偏差 $\boldsymbol{\gamma}_1 - \boldsymbol{\beta}_1$ 为 $\mathbf{0}$ 时设定错误的模型才能得到正确的结果, 这又对应了两种情形:

(a) $\boldsymbol{\Gamma}_{12} = \mathbf{0}$, 这即以为着分块解释变量 $\mathbf{X}_2$ 与解释变量 $\mathbf{X}_1$ 是线性无关的;

(b) 回归系数 $\boldsymbol{\beta}_2 = \mathbf{0}$, 即分块解释变量 $\mathbf{X}_2$ 对于被解释变量 Y 完全没有解释能力.

对于情形 (a), 即便如此就实证而言, 分块解释变量 $\mathbf{X}_2$ 被纳入进了误差项 u 中, 因为不相关则设定错误的回归模型式 (6.97) 仍就是严格外生的, 但 OLS 估计量的方差可能会增大. 而情形 (b) 实际上即意味着回归模型 (6.96) 退化为回归模型 (6.97), 不存在“设定错误”问题.

对于遗漏变量问题, 通常有如下的解决方法:

(a) 加入尽可能多的控制变量 (control variable), 即将可能遗漏的解释变量纳入回归模型中以提高估计的准确性, 具体的控制变量的选择需要依赖于理论建模. 这是最直接的处理方法, 但通常而言由于数据的可获得性等原因, 在实证中可能无法实现;

(b) 构造随机实验或寻找 (准) 自然实验, 这是现代因果推断方法, 前者通过随机分配处理组 (又称实验组) 和对照组并以控制除了实验因素以外的其他变量, 后者则是选择 (准) 在自然环境中发生的具有随机实验性质的过程, 从而减小遗漏变量偏差;

(c) 与解释变量相关的遗漏变量会导致回归模型存在内生性问题, 因而使用工具变量法进行参数估计量. 工具变量是满足误差项 (遗漏变量) 不相关、与内生变量 (与遗漏变量相关的解释变量) 相关, 可以通过 2SLS 将遗漏变量的影响区分出来, 从而得到更为准确的估计;

(d) 使用面板数据 (Panel Data) 建立回归模型, 其包括多个样本个体在不同时间点的观测数据, 通过控制个体固定效应和时间固定效应可以减小遗漏变量偏差.

无关变量则是与遗漏变量的相反的一种情形, 这时设定正确模型为式 (6.97), 而设定错误的模型为式 (6.96), 且有 $\boldsymbol{\beta}_2 = \mathbf{0}$, 因而依据式 (6.98) 有 $\boldsymbol{\beta}_1 = \boldsymbol{\gamma}_1$. 这既是说, 对于加入无关变量的设定错误的回归模型而言, 所估计量的总体参数 $\boldsymbol{\beta}_1$ 即为设定正确的真实参数 $\boldsymbol{\gamma}_1$.

但就 OLS 估计量而言, 由于加入的额外的解释变量 $\mathbf{X}_2$, 如果样本有限则增大 $\operatorname{Var}\left(\hat{\boldsymbol{\beta}} \middle| \boldsymbol{X}\right)$, 降低估计量的准确性.

6.12 信息准则

解释变量的选择有 3 种常用方法: 基于调整的 R^2 抉择、基于信息准则 (information criterion) 和“由大到小的序贯 t 规则”(general-to-specific sequential t rule).

(a) 基于调整的 R^2 抉择, 即所选择的解释变量的个数 K 要使得调整的 R^2 最大, 这是因为

$$\bar{R}^2 = 1 - \frac{(n-K)^{-1}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{(n-1)^{-1}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)} = 1 - \frac{\dfrac{1}{n-K}\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2}{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}$$

一方面反映了模型的解释能力 (RSS 越小), 另一方面同时惩罚了解释变量个数 K 增加带来的拟合效果的提升 ($-(n-K)^{-1}$), 使得 $\bar{R}^2$ 最大的 K 既有不错的拟合能力, 由控制了变量的个数;

(b) 信息准则实则是一个极小型指标, 其数值越小则表明模型越优异. 常用的信息准则包括**赤池信息准则** (Akaike information criterion)和**贝叶斯信息准则** (Bayesian information criterion), 在陈强的教材中其公式为[7]

$$\min_K \mathrm{AIC} = \ln\frac{\mathrm{RSS}}{n} + \frac{2}{n}K, \tag{6.101}$$

$$\min_K \mathrm{BIC} = \ln\frac{\mathrm{RSS}}{n} + \frac{\ln n}{n}K, \tag{6.102}$$

其含义是是使得使得 AIC 和 BIC 最小的 K 即为所需的变量个数.AIC 和 BIC 仅相差了后一项 K 的系数, 其前一项即是要求 K 使得 RSS 尽可能的小 (模型的拟合尽可能的好), 后一项则是要求解释变量的个数 K 尽可能的小. 由此而言,AIC 和 BIC 都是一个综合性的极小值指标.

[7]严格的说, 依据定义的 AIC 和 BIC 的公式为

$$\mathrm{AIC} = n\ln\left(2\pi\hat{\sigma}^2\right) + n + 2K, \tag{6.99}$$

$$\mathrm{BIC} = n\ln\left(2\pi\hat{\sigma}^2\right) + n + K\ln n, \tag{6.100}$$

其中方差估计量为

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\mathrm{RSS} = \frac{1}{n}\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)^2 = \frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\boldsymbol{\varepsilon}}{n}.$$

但就实际应用而言, 其实上述的公式都是等价的. 信息准则旨在帮助确定选择最合适的变量个数 K, 而 AIC 和 BIC 即是一个极小型指标, 其数值越小表示对应的 K 越好. 在给定样本容量为 n 的情况下, 陈强教材使用的式 (6.101) 和式 (6.102), 较依据定义的式 (6.99) 和式 (6.100) 而言其去掉了与 K 无关的常数项, 在此之上并除以了 n, 这是一个单调变换, 也不会改变 K 的选择.`Stata` 软件使用的便是式 (6.99) 和式 (6.100).

(c) 由大到小的序贯 t 规则, 常用于时间序列模型中确认所需的滞后阶数 p. 其思路是先指定一个最大滞后期数 $p_{\max}$, 记 $p^{(1)}=p_{\max}$, 随后对估计量 $\widehat{p^{(1)}}$ 进行原假设为 $\mathrm{H}_0: p^{(1)}=p_{\max}=0$ 的显著性检验, 若拒绝原假设, 则选择 $p=p_{\max}$; 若不能拒绝原假设, 则设新的 $p^{(2)}=p_{\max}-1$, 随后对估计量 $\widehat{p^{(2)}}$ 进行原假设为 $\mathrm{H}_0: p^{(2)}=p_{\max}-1=0$ 的显著性检验. 重复上述过程, 直至检验结果显著为止.

6.13 多重共线性的影响

多重共线性 (multicollinearity)包含两种情形, 严格 (或完全) 多重共线性和近似 (或弱) 多重共线性. 前者即是经典假设 (3.6) 不成立, 此时矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 不可逆, 因而 OLS 估计量是不存在的, 故回归系数 $\boldsymbol{\beta}$ 是不可识别的, 此时需要修改回归模型; 后者这是指解释变量 $\mathbf{X}=(X_1,\cdots,X_K)^{\mathrm{T}}$ 的各分量之间存在高度的相关关系 (但非完全相关).

具体而言, 考虑将回归模型 $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 写作分块形式

$$\boldsymbol{Y}=(\boldsymbol{X}_{-k},\boldsymbol{X}_k)\begin{pmatrix}\boldsymbol{\beta}_{-k}\\ \beta_k\end{pmatrix}+\boldsymbol{\varepsilon}=\boldsymbol{X}_{-k}\boldsymbol{\beta}_{-k}+\boldsymbol{X}_k\beta_k+\boldsymbol{\varepsilon}, \tag{6.103}$$

其中 $\boldsymbol{X}_k\in\mathbb{R}^{n\times 1}$ 为第 k 的解释变量 X_k 的样本 $(k=1,\cdots,K)$, 而 $\boldsymbol{X}_{-k}\in\mathbb{R}^{n\times(K-1)}$ 为设计矩阵 $\boldsymbol{X}$ 剔除了 $\boldsymbol{X}_k$ 的矩阵. 这样分块回归的系数满足有 $\beta_k\in\mathbb{R},\boldsymbol{\beta}_{-k}\in\mathbb{R}^{(K-1)\times 1}$. 依据线性代数的知识, 可以求出分块回归, 则第 k 个回归系数有估计量

$$\hat{\beta}_k=\left(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{M}_{-k}\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{M}_{-k}\boldsymbol{Y},$$

其中

$$\boldsymbol{M}_k=\boldsymbol{I}_n-\boldsymbol{X}_k\left(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k^{\mathrm{T}},\quad \boldsymbol{M}_{-k}=\boldsymbol{I}_n-\boldsymbol{X}_{-k}\left(\boldsymbol{X}_{-k}^{\mathrm{T}}\boldsymbol{X}_{-k}\right)^{-1}\boldsymbol{X}_{-k}^{\mathrm{T}}$$

分别为分块解释变量 $\boldsymbol{X}_k$ 和 $\boldsymbol{X}_{-k}$ 所对应的归零矩阵[8]. 由于归零矩阵是幂等矩阵[9], 则回归系数 β_k 可以变形为

$$\begin{aligned}\hat{\beta}_k&=\left(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{M}_{-k}\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{M}_{-k}\boldsymbol{Y}\\&=\left(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{M}_{-k}\boldsymbol{M}_{-k}\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{M}_{-k}\boldsymbol{M}_{-k}\boldsymbol{Y}\\&=\left(\tilde{\boldsymbol{X}}_k^{\mathrm{T}}\tilde{\boldsymbol{X}}_k\right)^{-1}\tilde{\boldsymbol{X}}_k^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}_{-k},\end{aligned} \tag{6.104}$$

其中 $\tilde{\boldsymbol{X}}_k=\boldsymbol{M}_{-k}\boldsymbol{X}_k,\tilde{\boldsymbol{\varepsilon}}_{-k}=\boldsymbol{M}_{-k}\boldsymbol{Y}$. 这样回归系数 β_k 即是回归模型 $\tilde{\boldsymbol{\varepsilon}}_{-k}=\tilde{\boldsymbol{X}}_k\beta_k+\boldsymbol{u}_1$ 的 OLS 估计量.

[8]对于回归模型 $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 而言, 其归零矩阵 $\boldsymbol{M}=\boldsymbol{I}_n-\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\boldsymbol{X}$ 满足有 $\boldsymbol{M}\boldsymbol{X}=\boldsymbol{O}$, 因而得名归零矩阵 (annihilator matrix). 进一步地,OLS 估计量的残差可以使用归零矩阵 $\boldsymbol{M}$ 表示为 $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{M}\boldsymbol{Y}$.

[9]如果方阵 $\boldsymbol{A}$ 满足有 $\boldsymbol{A}^2=\boldsymbol{A}$, 则称方阵 $\boldsymbol{A}$ 是幂等矩阵.

依据归零矩阵 $\boldsymbol{M}$ 的代数性质,$\tilde{\boldsymbol{X}}_k$ 即是回归模型 $\boldsymbol{X}_k = \boldsymbol{X}_{-k}\boldsymbol{\gamma}_2 + \boldsymbol{u}_3$ 的残差向量,$\tilde{\boldsymbol{\varepsilon}}_{-k}$ 是 $\boldsymbol{Y} = \boldsymbol{X}_{-k}\boldsymbol{\gamma}_1 + \boldsymbol{u}_2$ 的残差, 由此我们便可以通过三个回归方程得到 β_k 的 OLS 估计量[10]. 在条件同方差的情形下, 即有

$$\operatorname{Var}\left(\hat{\beta}_k \middle| \boldsymbol{X}\right) = \left(\tilde{\boldsymbol{X}}_k^{\mathrm{T}}\tilde{\boldsymbol{X}}_k\right)^{-1}\sigma^2,$$

而 $\tilde{\boldsymbol{X}}_k^{\mathrm{T}}\tilde{\boldsymbol{X}}_k$ 是回归模型 $\boldsymbol{X}_k = \boldsymbol{X}_{-k}\boldsymbol{\gamma}_2 + \boldsymbol{u}_3$ 的残差平方和, 这个回归的含义是将第 k 个解释变量 X_k 回归到余下的 $K-1$ 个解释变量上, 我们记该回归模型的拟合优度为 R_k^2, 则有

$$R_k^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\tilde{\boldsymbol{X}}_k^{\mathrm{T}}\tilde{\boldsymbol{X}}_k}{\left(\boldsymbol{X}_k - \bar{X}_k\cdot\boldsymbol{1}\right)^{\mathrm{T}}\left(\boldsymbol{X}_k - \bar{X}_k\cdot\boldsymbol{1}\right)} = 1 - \frac{\tilde{\boldsymbol{X}}_k^{\mathrm{T}}\tilde{\boldsymbol{X}}_k}{\sum\limits_{i=1}^{n}\left(X_{ik} - \bar{X}_{ik}\right)},$$

即有

$$\left(\tilde{\boldsymbol{X}}_k^{\mathrm{T}}\tilde{\boldsymbol{X}}_k\right)^{-1} = \frac{1}{(1-R_k^2)\sum\limits_{i=1}^{n}\left(X_{ik} - \bar{X}_{ik}\right)},$$

而注意到随机变量 X 的无偏方差估计量为

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right),$$

故 OLS 估计量的 $\hat{\beta}_k$ 的条件协方差为

$$\operatorname{Var}\left(\hat{\beta}_k \middle| \boldsymbol{X}\right) = \frac{\sigma^2}{(1-R_k^2)\sum\limits_{i=1}^{n}\left(X_{ik} - \bar{X}_{ik}\right)} = \frac{\sigma^2}{(1-R_k^2)(n-1)s_{X_k}^2}. \tag{6.105}$$

式 (6.105) 将估计量 $\hat{\beta}_k$ 的条件方差分解为四个构成的部分: 回归方差 σ^2、解释变量 X_k 的方差估计量 $s_{X_k}^2$、样本容量 n 和拟合优度 R_k^2. 而式 (6.105) 的平方根即为标准误 $\mathrm{SE}\left(\hat{\beta}_k\right)$, 这四部分对于 OLS 估计量的影响如下:

(a) 回归方差 σ^2 是误差项 ε_i 的非条件方差, 依据式 (6.105),σ^2 越大,$\mathrm{SE}\left(\hat{\beta}_k\right)$ 越大, 即说明误差波动的增大降低了 OLS 估计量 β_k 的估计的准确性;

(b) 方差估计量 $s_{X_k}^2$ 度量了第 k 个解释变量 X_k 的样本数据的波动情况, 其与式 (6.105) 是负相关的, 这意味着如果选择的解释变量 X_k 的数据波动不大, 那么 OLS 估计量 $\hat{\beta}_k$ 识别出 β_k 的准确性会降低, 增大了估计量的不确定性;

(c) 样本容量 n 与标准误 $\mathrm{SE}\left(\hat{\beta}_k\right)$ 是反向变动的, 即样本个数越多, 估计的不确定性越少;

[10]这种方法又被称为 Frisch-Waugh-Lovell 定理.

(d) 拟合优度 R_k^2 度量了其余的 $K-1$ 个解释变量对解释变量 X_k 的解释能力 (线性条件期望函数), 也即反映了 X_k 能否表示为其余解释变量的线性组合. 依据式 (6.105), 则 R_k^2 越趋于 1, 即解释变量 X_k 与其余解释变量的线性关系越强, 那么 OLS 估计量 $\hat{\beta}_k$ 的标准误越大, 因而单个系数 β_k 估计的不确定性越大.

依据上述分析, 我们便可以得到近似多重共线性的影响. 依据 (d) 可知, 如果存在了近似多重共线性, 在其余影响因素不变的前提下, 单个系数的 OLS 估计量 $\hat{\beta}_k$ 的标准误会增大, 进而使得单个系数的估计的不确定性增大. 我们将

$$\mathrm{VIF}_k = \frac{1}{1-R_k^2} \tag{6.106}$$

称为第 k 个解释变量**方差膨胀因子** (variance inflation factor), 其可以作为近似多重共线性的一个独立指标. 恰如其名, 当其余条件不变时 VIF_k 的增大即使得 OLS 估计量 $\hat{\beta}_k$ 的方差增大了相应的倍数. 由于一个回归模型有 K 个解释变量, 因而可以计算 K 个方差膨胀因子, 一个经验法则是对于任意的 VIF_k 都应该满足有 $\mathrm{VIF}_k < 10$, 亦即任意的拟合优度 $R_k^2 < 0.9$, 这样近似多重共线性对回归结果的影响是有限的.

对于近似多重共线性, 通常采取的方法是“无为而治”, 这样因为:

(1) 如果是研究回归模型的预测能力, 这时考虑的是条件期望函数 $m(\boldsymbol{x}) = \mathbb{E}(Y|\mathbf{X}=\boldsymbol{x})$, 可以计算条件期望函数的标准误为 $\mathrm{SE}(\hat{m}(\boldsymbol{x})) = \sqrt{\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{x}}$, 近似多重共线性对协方差矩阵估计量 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ 的影响不严重;

(2) 实证研究中通常仅关心我们所感兴趣变量的回归系数 β_k, 一种理想情况是即是存在了近似多重共线性但 β_k 仍然通过了单个系数的显著性检验, 那么则是安然无事, 即使处理了近似多重共线性那也便会得到更显著的结果. 在存在近似多重共线性的情形下, 如果我们感兴趣的回归系数 β_k 不显著, 这时可以采用修改模型设定、增加样本容量 n、将解释变量的高次项标准化等方法.

6.14 代码

第七章　多变量回归

7.1　从多元 (multiple) 到多变量 (multivariate)

7.2　多变量回归的 OLS 估计量

7.3　多变量回归的 OLS 估计量的有限样本性质

7.4　多变量回归的 OLS 估计量的渐进性质

第八章　内生性与工具变量

本节我们将介绍内生性——这是计量经济学区别于数理统计的核心问题之一. 内生性即是违背了线性回归模型假设 (3.2.1) 中的严格外生性, 其导致了 OLS 估计量的良好性质不再成立. 解决内生性是实证分析中的核心问题, 最为经典的方法即是使用工具变量.

8.1　内生变量

在此前的章节中, 我们是先介绍了线性投影模型模型 (3.3.1) 的总体参数, 依据 MSE 最小的原则, 其核心设定在于

$$Y = \mathscr{P}(Y|\mathbf{X}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \quad \mathbb{E}(\mathbf{X}\varepsilon) = \mathbf{0}, \tag{8.1}$$

即投影误差 ε 与解释变量 $\mathbf{X}$ 是概率论意义上正交的. 我们还得到了线性投影模型的总体参数为 $\boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y)$.

对于回归模型 (2.1.1), 其核心设定为

$$Y = \mathbb{E}(Y|\mathbf{X}) + e, \quad \mathbb{E}(e|\mathbf{X}) = 0, \tag{8.2}$$

即回归误差 e 的在给定解释变量 $\mathbf{X}$ 时的条件期望为 0, 亦或回归误差 e 均值独立于解释变量 $\mathbf{X}$. 利用定理 (2.1.3),$\mathbb{E}(e|\mathbf{X}) = 0$ 可以推出 $\mathbb{E}(\mathbf{X}\varepsilon) = \mathbf{0}$, 因而回归误差 e 与解释变量 $\mathbf{X}$ 是正交的.

之所以引入线性投影模型, 是因为我们不能确定回归函数 $\mathbb{E}(Y|\mathbf{X})$ 的具体形式, 因而使用了最优线性预测量 $\mathscr{P}(Y|\mathbf{X})$ 去预测或解释 Y. 当回归函数是线性条件期望函数时, 即有 $\mathbb{E}(Y|\mathbf{X}) = \mathscr{P}(Y|\mathbf{X})$ 时, 这时线性投影模型便是线性回归模型, 因而投影误差 ε 即为回归误差 e, 我们统一将线性回归模型 (3.5.1) 的误差记为 ε. 在此意义上, 线性回归模型 (3.5.1) 的回归系数 $\boldsymbol{\beta}$ 便是线性投影模型的总体参数, 因而我们便可使用 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 去估计 $\boldsymbol{\beta}$.

无论是线性回归模型还是线性投影模型, 在 MSE 最小的原则下, 其误差项都满足于解释变量 $\mathbf{X}$ 是正交的. 本章我们考虑一个更一般的线性模型

$$Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \tag{8.3}$$

我们称式 (8.3) 为**结构方程** (structural equation), 其总体参数 $\boldsymbol{\beta}$ 称为**结构参数** (structural parameter). 为了与线性投影模型 (3.3.1) 区别, 本章在此之后将线性投影模型 (3.3.1) 的总体参数记为 $\boldsymbol{\beta}^*$.

我们现在引入内生变量的概念.

定义 8.1.1 (外生变量与内生变量)**.** *对于结构方程式 (8.3), 若随机变量 X_i 与误差项 ε 正交, 即*

$$\mathbb{E}(X_i\varepsilon) = 0,$$

*我们称 X_i 是***外生变量** *(exogenous variable), 或者 X_i 是外生的; 若随机变量 X_i 与误差项 ε 不正交, 即*

$$\mathbb{E}(X_i\varepsilon) \neq 0,$$

*我们称 X_i 是***内生变量** *(endogenous variable), 或者 X_i 是外生的.*

显然, 依据式 (8.1) 和式 (8.2), 线性投影模型 (3.3.1) 和线性回归模型 (3.5.1) 中的解释变量都是外生变量, 我们称模型满足了**外生性** (exogeneity). 相反地, 如果结构方程式 (8.3) 存在内生的解释变量, 那么我们称式 (8.3) 存在**内生性** (endogeneity).

为什么要提出内生性? 这是因为, 如果结构方程式 (8.3) 存在内生性, 即有 $\mathbb{E}(\mathbf{X}\varepsilon) \neq \mathbf{0}$, 那么式 (8.3) 的 OLS 估计量的总体参数 (即线性投影模型的总体参数) 为

$$\begin{aligned}\boldsymbol{\beta}^* =& \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y) = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon\right)\right)\\ =& \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta}+\mathbb{E}(\mathbf{X}\varepsilon)\right)\\ =& \boldsymbol{\beta}+\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}\varepsilon),\end{aligned} \tag{8.4}$$

因而使用 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 有 $\hat{\boldsymbol{\beta}} \xrightarrow{P} \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y) = \boldsymbol{\beta}^* \neq \boldsymbol{\beta}$, 即 OLS 估计量既不是结构参数 $\boldsymbol{\beta}$ 的相合估计, 亦不是无偏估计.

一个有意思的结论是, 根据定义 (8.1.1), 被解释变量 Y 是一个内生变量, 这是因为

$$\mathbb{E}(Y\varepsilon) = \mathbb{E}\left[\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon\right)\varepsilon\right] = \mathbb{E}\left(\mathbf{X}^{\mathrm{T}}\varepsilon\right)\boldsymbol{\beta} + \mathbb{E}\varepsilon^2 \geqslant \mathbb{E}\varepsilon^2 = \sigma^2 > 0,$$

这一结果总是成立的, 无论结构方程式 (8.3) 是否存在内生性.

在实证中, 内生性有三种常见的来源: 测量误差、联立方程和选择偏差. 为了分析内生性, 我们需要复习一下概率论中独立性、均值独立和正交性的关系: 在误差项 ε 满足 $\mathbb{E}\varepsilon = 0$ 的条件下, 则随机变量 X 有单向推出的命题:

$$X\text{与}\varepsilon\text{独立} \Rightarrow \underbrace{\mathbb{E}(\varepsilon|\,X) = 0}_{\varepsilon\text{均值独立于}X} \Rightarrow \underbrace{\mathbb{E}(X\varepsilon) = 0}_{\varepsilon\text{正交于}X} \Leftrightarrow \underbrace{\mathrm{Cov}(X,\varepsilon) = 0}_{\varepsilon\text{与}X\text{不相关}},$$

第二个 $\Rightarrow$ 利用迭起期望定理是易证的.

例 8.1.1 (解释变量的测量误差). 对于回归模型 $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$, 设随机变量满足 $\mathbf{X}^* = \mathbf{X} + \boldsymbol{u}$, 其中 $\mathbf{X}^*$ 是可以观察的结果,$\mathbf{X}$ 是真实的结果.$\boldsymbol{u}$ 为解释变量 $\mathbf{X}$ 的**测量误差** *(measurement error)*, 其满足有 $\mathbb{E}\boldsymbol{u} = \mathbf{0}$, 且与 $\mathbf{X}^*$ 和 $\mathbf{X}$ 不相关.

对于例 (8.1.1), 依据测量误差与 $\mathbf{X}^*$ 和 $\mathbf{X}$ 不相关, 不难推知 $\mathbb{E}\left(\mathbf{X}\boldsymbol{u}^{\mathrm{T}}\right) = \mathbb{E}\left(\mathbf{X}^*\boldsymbol{u}^{\mathrm{T}}\right) = \mathbf{0}$, 且有

$$\mathbb{E}\left(\mathbf{X}^*\boldsymbol{u}^{\mathrm{T}}\right) = \mathbb{E}\left[\left(\mathbf{X} + \boldsymbol{u}\right)\boldsymbol{u}^{\mathrm{T}}\right] = \mathbb{E}\left(\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\right).$$

假设真实的回归模型 $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$ 满足有 $\mathbb{E}\left(\mathbf{X}\varepsilon\right) = \mathbf{0}$, 即不存在内生性问题. 但在存在测量误差的情形下, 我们只能得到 $\mathbf{X}^*$, 因而实际的回归模型为

$$Y = \left(\mathbf{X}^* - \boldsymbol{u}\right)^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon = \left(\mathbf{X}^*\right)^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon = \left(\mathbf{X}^*\right)^{\mathrm{T}}\boldsymbol{\beta} + v,$$

其中实际的误差项 $v = \varepsilon - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\beta}$. 不难计算,

$$\begin{aligned}\mathbb{E}\left(\mathbf{X}^*v\right) =& \mathbb{E}\left[\left(\mathbf{X} + \boldsymbol{u}\right)\left(\varepsilon - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\beta}\right)\right] \\ =& \mathbb{E}\left(\mathbf{X}\varepsilon\right) + \mathbb{E}\left(\boldsymbol{u}\varepsilon\right) - \mathbb{E}\left(\mathbf{X}\boldsymbol{u}^{\mathrm{T}}\right)\boldsymbol{\beta} - \mathbb{E}\left(\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\right)\boldsymbol{\beta} \\ =& -\mathbb{E}\left(\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\right)\boldsymbol{\beta} \neq \mathbf{0},\end{aligned}$$

即解释变量的测量误差导致了回归模型存在内生性. 特别地, 当仅有一个解释变量 $\mathbf{X} = X$ 时, 此时测量误差有 $X^* = X + u, v = \varepsilon - u^{\mathrm{T}}\beta$, 依据式 (8.4) 我们可以得到 OLS 估计量的总体参数为

$$\begin{aligned}\beta^* =& \beta + \frac{\mathbb{E}\left(X^*v\right)}{\mathbb{E}\left(X^*\right)^2} = \beta + \frac{\mathbb{E}\left[X^*\left(\varepsilon - u^{\mathrm{T}}\beta\right)\right]}{\mathbb{E}\left(X^*\right)^2} \\ =& \beta - \frac{\mathbb{E}u^2}{\mathbb{E}\left(X^*\right)^2}\beta = \beta\left(1 - \frac{\mathbb{E}u^2}{\mathbb{E}\left(X^*\right)^2}\right) < \beta,\end{aligned} \tag{8.5}$$

即 OLS 估计量的收敛结果将小于真实参数 β, 我们将这种偏差成为**测量误差偏差** (measurement error bias). 另一方面, 依据式 (8.5) 不难得如果测量误差 u 的波动幅度较 X^* 越大, 则 β^* 越趋于 0, 因而测量误差偏差又被称为**衰减偏差** (attenuation bias).

例 (8.1.1) 考虑的是解释变量的测量误差, 我们现在考虑被解释变量 Y 的测量误差. 设随机变量 $Y^* = Y + u$, 其中 Y^* 是解释变量 Y 可观察的结果, 而 u 为 Y 的测量误差. 那么实际的回归模型为

$$Y^* - u = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon \Rightarrow Y^* = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + u + \varepsilon = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + v,$$

其中误差项 $v = u + \varepsilon$. 如果被解释变量 Y 的测量误差 u 与解释变量 $\mathbf{X}$ 不相关 (这通常都是成立的), 那么误差项 v 与解释变量 $\mathbf{X}$ 不相关, 因而 v 与解释变量 $\mathbf{X}$ 仍是正交的, 不会导致内生性问题.

例 8.1.2 (联立方程偏差). 考虑微观经济学中的供给曲线与需求曲线

$$\begin{aligned} &\text{需求曲线}: Q=-\beta_1 P+\alpha_1+\varepsilon_1, \\ &\text{供给曲线}: Q=\beta_2 P+\alpha_2+\varepsilon_2, \end{aligned} \tag{8.6}$$

其中 Q 的商品的数量,P 为价格, 参数 β_1,β_2 非负.$\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2)^{\mathrm{T}}$ 为随机向量,$\mathbb{E}\boldsymbol{\varepsilon}=\mathbf{0},\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right)=\boldsymbol{I}_n$, 即有 ε_1 和 ε_2 不相关.

考虑线性投影模型 $Q=\beta^* P+\alpha^*+\varepsilon$, 依据模型 (2.2.3) 可以得到参数 β^* 和 α^* 为

$$\beta^*=\frac{\operatorname{Cov}(Q,P)}{\operatorname{Var}(P)},\quad \alpha^*=\mathbb{E}Q-\beta^*\mathbb{E}P.$$

一个自然的问题是, 我们通过 OLS 估计量去估计 β^* 和 α^*, 那么 β^* 和 α^* 与式 (8.6) 是什么关系? 为此我们需要将商品数量 Q 和价格 P 视为随机变量, 这可以通过联立方程求解式 (8.6), 将 Q 和 P 解出为关于随机变量 $\boldsymbol{\varepsilon}$ 的函数——亦即市场取得均衡点. 不难解得, 式 (8.6) 的均衡点为

$$Q^*=\frac{\beta_1\alpha_2+\beta_2\alpha_1}{\beta_1+\beta_2}+\frac{\beta_1\varepsilon_2+\beta_2\varepsilon_1}{\beta_1+\beta_2},\quad P^*=\frac{\alpha_1-\alpha_2}{\beta_1+\beta_2}+\frac{\varepsilon_1-\varepsilon_2}{\beta_1+\beta_2},$$

因而有

$$\begin{aligned} &\mathbb{E}P=\frac{\alpha_1-\alpha_2}{\beta_1+\beta_2}+\mathbb{E}\frac{\varepsilon_1-\varepsilon_2}{\beta_1+\beta_2}=\frac{\alpha_1-\alpha_2}{\beta_1+\beta_2},\quad \mathbb{E}Q=\frac{\beta_1\alpha_2+\beta_2\alpha_1}{\beta_1+\beta_2} \\ &\operatorname{Var}(P)=\mathbb{E}(P-\mathbb{E}P)^2=\mathbb{E}\left(\frac{\varepsilon_1-\varepsilon_2}{\beta_1+\beta_2}\right)^2=\frac{2}{(\beta_1+\beta_2)^2}, \\ &\operatorname{Cov}(Q,P)=\operatorname{Cov}\left(\frac{\beta_1\varepsilon_2+\beta_2\varepsilon_1}{\beta_1+\beta_2},\frac{\varepsilon_1-\varepsilon_2}{\beta_1+\beta_2}\right)=\frac{1}{(\beta_1+\beta_2)^2}\operatorname{Cov}(\beta_1\varepsilon_2+\beta_2\varepsilon_1,\varepsilon_1-\varepsilon_2) \\ &\quad=\frac{1}{(\beta_1+\beta_2)^2}\operatorname{Cov}(\beta_1\varepsilon_2+\beta_2\varepsilon_1,\varepsilon_1-\varepsilon_2)=\frac{\beta_2-\beta_1}{(\beta_1+\beta_2)^2}, \end{aligned}$$

故总体参数为

$$\beta^*=\frac{\beta_2-\beta_1}{(\beta_1+\beta_2)^2}\cdot\frac{(\beta_1+\beta_2)^2}{2}=\frac{\beta_2-\beta_1}{2},$$

和

$$\begin{aligned} \alpha^*=&\frac{\beta_1\alpha_2+\beta_2\alpha_1}{\beta_1+\beta_2}-\frac{\beta_2-\beta_1}{2}\cdot\frac{\alpha_1-\alpha_2}{\beta_1+\beta_2} \\ =&\frac{1}{2(\beta_1+\beta_2)}\left[2\beta_1\alpha_2+2\beta_2\alpha_1-(\beta_2-\beta_1)(\alpha_1-\alpha_2)\right] \\ =&\frac{1}{2(\beta_1+\beta_2)}\left(2\beta_1\alpha_2+2\beta_2\alpha_1-\beta_2\alpha_1+\beta_2\alpha_2+\beta_1\alpha_1-\beta_1\alpha_2\right) \\ =&\frac{1}{2(\beta_1+\beta_2)}\left(\beta_1\alpha_2+\beta_2\alpha_1+\beta_2\alpha_2+\beta_1\alpha_1\right) \\ =&\frac{1}{2(\beta_1+\beta_2)}(\beta_2+\beta_1)(\alpha_1+\alpha_2)=\frac{\alpha_1+\alpha_2}{2}. \end{aligned}$$

由此, 我们得到 OLS 估计量最终将收敛于

$$\hat{Q} = \hat{\beta}P + \hat{\alpha} \xrightarrow{P} Q = \frac{\beta_2 - \beta_1}{2}P + \frac{\alpha_1 + \alpha_2}{2}, \quad n \to \infty,$$

即为将需求曲线和供给曲线相加后计算的平均值, 其会经过市场均衡点. 像例 (8.1.2) 这样, 如果被解释变量 Y 和解释变量 $\mathbf{X}$ 通过方程组被联立决定 (simultaneously determined), 则 OLS 估计量收敛于线性投影模型的总体参数而偏离了联立方程的参数, 我们将此称为**联立方程偏差** (simultaneous equations bias).

代码蒙特卡洛模型绘图解释.

例 8.1.3 (教育回报的研究). 内容...

8.2 工具变量

在上一节中, 我们介绍了结构方程式 (8.3) 存在内生性时 OLS 估计量将无法估计出结构参数 $\boldsymbol{\beta}$. 对于一个参数, 如果我们其可以写成关于可观察变量 (the observable) 的概率分布的唯一函数, 那么我们称这个参数是**可识别的** (identified). 本节我们介绍工具变量的概念, 通过工具变量我们可以识别结构参数 $\boldsymbol{\beta}$ 并得到相应的估计量.

定义 8.2.1 (工具变量). 设随机向量 $\mathbf{Z} \in \mathbb{R}^{l\times 1}$ 满足有

$$\text{外生性}: \mathbb{E}(\mathbf{Z}\varepsilon) = \mathbf{0}, \tag{8.7}$$

$$\text{正定性}: \mathbb{E}\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right) > 0, \tag{8.8}$$

$$\text{相关条件}: \operatorname{rank}\left(\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right)\right) = K, \tag{8.9}$$

则我们称 $\mathbf{Z}$ 是结构方程式 *(8.3)* 的**工具变量** *(instrumental variable)*, 或者称 $\mathbf{Z}$ 是解释变量 $\mathbf{X}$ 的工具变量.

我们将结构方程式 (8.3) 写成分块形式

$$Y = \mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{X}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon, \tag{8.10}$$

其中

$$\mathbf{X} = \left(\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}}\right)^{\mathrm{T}} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}\right)^{\mathrm{T}} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}.$$

我们设解释变量 $\mathbf{X}_1 \in \mathbb{R}^{K_1\times 1}$ 包含了 K_1 个外生变量, 解释变量 $\mathbf{X}_2 \in \mathbb{R}^{K_2\times 1}$ 包含了 K_2 个内生变量, 即有

$$\mathbb{E}(\mathbf{X}_1\varepsilon) = 0, \quad \mathbb{E}(\mathbf{X}_2\varepsilon) \neq 0, \quad K_1 + K_2 = K.$$

依据定义 (8.10), 在假设5.1.1的条件下, 显然外生变量 $\mathbf{X}_1$ 是自身的工具变量 (包括了截距项的情形). 这样对于分块的结构方程式 (8.10), 我们设有包含了 $\mathbf{X}_1$ 的工具变量 $\mathbf{Z} \in \mathbb{R}^{l\times 1}$, 写成分块形式即有

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{Z}_2 \end{pmatrix},$$

其中第一个工具变量 $\mathbf{Z}_1 = \mathbf{X}_1 \in \mathbb{R}^{K_1\times 1}$, 另一个工具变量 $\mathbf{Z}_2 \in \mathbb{R}^{l_2\times 1}$ 假设是已知的, 且有 $\boxed{K_1 + l_2 = l}$. 通常而言, 在实证研究中工具变量 $\mathbf{Z}_2 \in \mathbb{R}^{l_2\times 1}$ 并不是基准模型的解释变量, 是需要额外去寻找的.

线性投影模型的核心设定是 $\mathbb{E}(\mathbf{X}\varepsilon) = \mathbf{0}$, 依据此正交条件我们推导了参数 $\boldsymbol{\beta}^*$ 的表达式. 但存在内生性时 $\mathbb{E}(\mathbf{X}\varepsilon) \neq \mathbf{0}$, 我们需要额外的信息来推导. 相似地, 利用工具变量 $\mathbf{Z}$ 的外生性式 (8.7), 对于结构参数 $\boldsymbol{\beta}$ 的一个可能的**识别** (identification)是, 将结构方程式 (8.3) 代入式 (8.7), 即有 $\mathbb{E}\left[\mathbf{Z}\left(Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\right] = \mathbf{0}$, 进而我们得到

$$\underbrace{\mathbb{E}(\mathbf{Z}Y)}_{l\times 1} = \underbrace{\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right)}_{l\times K}\boldsymbol{\beta}, \tag{8.11}$$

式 (8.11) 是一个关于结构参数 $\boldsymbol{\beta}$ 的线性方程组, 具体而言, 其包含了 K 个未知元和 l 个方程, 显然我们希望方程组式 (8.11) 能够有唯一的解 $\boldsymbol{\beta}$, 只有这样才能回答我们所研究的问题. 回顾在线性代数的知识, **在方程组式 (8.11) 有解的情形下**[1], 存在唯一解的充要条件是齐次线性方程组 $\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right)\boldsymbol{\beta} = \mathbf{0}$ 有唯一的解 (也即平凡的零解), 即意味着系数矩阵的秩等于未知元的个数:$\operatorname{rank}\left(\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right)\right) = K$. 这便是是定义 (8.2.1) 中需要**相关条件** (relevance condition)式 (8.9) 的原因, 只有满足了使得 $\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right)$ 列满秩的的外生变量 $\boldsymbol{Z}$ 才能帮助我们识别出结构参数 $\boldsymbol{\beta}$.

线性代数还告诉我们矩阵的秩不会大于其维度, 因而工具变量的相关条件式 (8.9) 便意味着 $l \geqslant \operatorname{rank}\left(\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right)\right) = K$, 又内生变量的工具变量便是其自身, 因而我们将工具变量的识别又分为如下两种情形:

(i) 若 $l = K$, 即 $l_2 = K_2$, 工具变量 $\mathbf{Z}_2$ 的变量数与内生变量 $\mathbf{X}_2$ 的变量数相同, 这时我们称结构方程式 (8.3) 是**恰好识别的** (just-identified);

(ii) 若 $l > K$, 即 $l_2 > K_2$, 工具变量 $\mathbf{Z}_2$ 的变量数大于与内生变量 $\mathbf{X}_2$ 的变量数, 这时我们称结构方程式 (8.3) 是**过度识别的** (over-identified).

如果工具变量 $\mathbf{Z}_2$ 的变量数的个数少于内生变量的变量数, 则方程组 (8.11) 没有唯一解, 那么这时的结构参数是**不可识别的** (unidentified). 显然这也违反了定义 (8.2.1).l_2

[1]非齐次线性方程组 $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{b}$ 有解的充分必要条件是 $\operatorname{rank}\boldsymbol{A} = \operatorname{rank}(\boldsymbol{A}, \boldsymbol{b})$, 但系数矩阵 $\boldsymbol{A}$ 是列满秩的并不能保证 $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{b}$ 有解.

与 K_2 数量关系并不能决定识别 (有解时还需要相关条件), 但我们通常将其**阶条件** (order condition).

我们现在考虑恰好识别的情形, 即 $\mathbf{Z}_2$ 的变量数等于内生变量 $\mathbf{X}_2$ 的变量数, 那么 $l = K_1 + l_2 = K_1 + K_2 = K$, 而 $\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right)$ 为满秩的 K 阶方阵. 此时方程组式 (8.11) 有唯一解

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{Z}Y\right), \tag{8.12}$$

类似于设计矩阵 $\boldsymbol{X}$ 的矩阵符号, 记工具变量 $\mathbf{Z} = (Z_1, \cdots, Z_l)^{\mathrm{T}}$ 的第 i 个样本个体为随机变量 $\mathbf{Z}_i = (Z_{il}, \cdots, Z_{il})^{\mathrm{T}}$,$i = 1 \cdots, n$, 则 $\boldsymbol{Z}^{\mathrm{T}} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_l)$. 依据式 (8.12), 我们对结构参数 $\boldsymbol{\beta}$ 有矩估计

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{IV}} &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_iY_i\right)\\ &= \left(\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_iY_i\right) = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y},\end{aligned} \tag{8.13}$$

我们称式 (8.13) 便是结构方程式 (8.3) 的**工具变量估计量** (instrumental variables estimator), 简称为 **IV 估计量** (IV estimator). 有时候式 (8.13) 又被称为**间接最小二乘估计量** (indirect least squares estimator), 简称为 **ILS 估计量** (ILS estimator), 记为 $\hat{\boldsymbol{\beta}}_{\mathrm{ILS}}$.

我们定义 IV 估计量的残差为 $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$, 用矩阵的符号表示则有 $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$, 不难计算

$$\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y} = \mathbf{0}, \tag{8.14}$$

即 IV 估计量的残差向量与工具变量 $\mathbf{Z}$ 的“设计矩阵”是正交的.

和 OLS 估计量一样, 我们将含截距项的结构方程单独写出, 研究其恰好识别时的 IV 估计量. 设含有截距项的结构方程为

$$Y = \alpha + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \tag{8.15}$$

其中截距项 α 对应了 demeaned regressor. 那么此时工具变量 $\mathbf{Z}$ 中也有 demeaned regressor, 即 $\boldsymbol{Z}^{\mathrm{T}} = (\mathbf{1}, \mathbf{Z}_1, \cdots, \mathbf{Z}_l)$. 依据正交性式 (8.14) 即得

$$\boldsymbol{Z}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} = \mathbf{0} \Rightarrow \sum_i \mathbf{Z}_i\hat{\varepsilon}_i = 0 \Rightarrow \sum_i Z_{ik}\hat{\varepsilon}_i = 0,$$

这时 IV 估计量的残差为 $\hat{\varepsilon}_i = Y_i - \hat{\alpha}_{\mathrm{IV}} - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$, 因而有关于 IV 估计量的线性方程组

$$\begin{cases}\displaystyle\sum_{i=1}^{n}\left(Y_i - \hat{\alpha}_{\mathrm{IV}} - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}\right) = 0,\\ \displaystyle\sum_{i=1}^{n}\mathbf{Z}_i\left(Y_i - \hat{\alpha}_{\mathrm{IV}} - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}\right) = 0,\end{cases}$$

可以解得含有截距项时的 IV 估计量为

$$\begin{cases} \hat{\alpha}_{\mathrm{IV}} = \bar{Y} - \bar{\mathbf{X}}^T \hat{\boldsymbol{\beta}}_{\mathrm{IV}}, \\ \hat{\boldsymbol{\beta}}_{\mathrm{IV}} = \left(\sum\limits_{i=1}^{n} \left(\mathbf{Z}_i - \bar{\mathbf{Z}} \right) \left(\mathbf{X}_i - \bar{\mathbf{X}} \right)^{\mathrm{T}} \right)^{-1} \sum\limits_{i=1}^{n} \left(\mathbf{Z}_i - \bar{\mathbf{Z}} \right) \left(Y_i - \bar{Y} \right). \end{cases} \tag{8.16}$$

此外, 由于截距项的 demeaned regressor 使得一个工具变量 $Z_{ik} \equiv 1$, 因而根据 $\sum\limits_i Z_{ik} \hat{\varepsilon}_i = 0$ 便可推知残差和 $\sum\limits_{i=1}^{n} \hat{\varepsilon}_i = 0$.

我们现在以最简单的一元线性模型来说明工具变量的恰好识别. 设结构方程为

$$Y = \alpha + \beta X + \varepsilon, \tag{8.17}$$

其中假设解释变量 X 是内生的. 我们考虑已知工具变量 $\mathbf{Z} = (Z_1, Z_2)^{\mathrm{T}} = (1, Z)^{\mathrm{T}}$, 那么依据式 (8.11) 即有关于结构参数的线性方程组

$$\mathbb{E} \begin{pmatrix} Y \\ ZY \end{pmatrix} = \mathbb{E} \begin{pmatrix} 1 & X \\ Z & ZX \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

不难解得

$$\begin{aligned} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} =& \frac{1}{\mathbb{E}(ZX) - \mathbb{E}Z\mathbb{E}X} \begin{pmatrix} \mathbb{E}(ZX) & -\mathbb{E}Z \\ -\mathbb{E}X & 1 \end{pmatrix} \begin{pmatrix} \mathbb{E}Y \\ \mathbb{E}(ZY) \end{pmatrix} \\ =& \frac{1}{\mathbb{E}(ZX) - \mathbb{E}Z\mathbb{E}X} \begin{pmatrix} \mathbb{E}(ZX)\mathbb{E}Y - \mathbb{E}Z\mathbb{E}(ZY) \\ \mathbb{E}(ZY) - \mathbb{E}X\mathbb{E}Y \end{pmatrix}, \end{aligned} \tag{8.18}$$

这时的相关条件式 (8.9) 等价于 $\mathrm{Cov}(Z, X) = \mathbb{E}(ZX) - \mathbb{E}Z\mathbb{E}X \neq 0$, 即意味着工具变量 Z 与内生变量 X 是相关的. 特别地, 如果解释变量 X 是外生的, 那么式 (8.18) 便退化为了一元线性回归模型的总体参数, 相应地 IV 估计量便退化为了 OLS 估计量.

8.3 内生变量的简约式

上一节中我们介绍了在恰好识别的情形下, 如何使用工具变量得到结构参数 $\boldsymbol{\beta}$ 的估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$. 本节我们研究过度识别下 ($l > K$ 时) 内生变量与工具变量的简约式, 并由此得到结果参数的估计量.

考虑分块的结构方程式 (8.10), 我们在前面已经指出依据定义 (8.1.1), 被解释变量 Y 也是内生的, 为了区分我们将内生变量 $\mathbf{X}_2$ 又称为**右侧内生变量** (endogenous right-hand-side variable). 我们将结构方程式 (8.10) 中的所有内生变量记为列向量 $\overrightarrow{Y} = \left(Y_1, \mathbf{Y}_2^{\mathrm{T}}\right)^{\mathrm{T}}$, 其中 $Y_1 = Y$ 为被解释变量,$\mathbf{Y}_2 = \mathbf{X}_2$ 为右侧内生变量, 那么结构方程式 (8.10) 可以写作

$$Y_1 = \mathbf{Z}_1^{\mathrm{T}} \boldsymbol{\beta}_1 + \mathbf{Y}_2^{\mathrm{T}} \boldsymbol{\beta}_2 + \varepsilon. \tag{8.19}$$

我们现在介绍**简约式** (reduced form)的概念: 我们将内生变量与工具变量的关系 (relationship) 称为简约式. 对于内生变量 $\mathbf{Y}_2$, 我们称其与工具变量 $\mathbf{Z}$ 的一个线性的简约式为

$$\mathbf{Y}_2 = \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z} + \boldsymbol{u}_2 = \boldsymbol{\Gamma}_{12}^{\mathrm{T}}\mathbf{Z}_1 + \boldsymbol{\Gamma}_{22}^{\mathrm{T}}\mathbf{Z}_2 + \boldsymbol{u}_2, \tag{8.20}$$

式 (8.20) 是第七章所介绍的多变量回归的一个总体方程, 其中 $\mathbf{Y}_2 \in \mathbb{R}^{K_2\times 1}$ 为数据向量,$\boldsymbol{u}_2 \in \mathbb{R}^{K_2\times 1}$ 为误差项[2],

$$\boldsymbol{\Gamma} = \left(\boldsymbol{\Gamma}_{12}^{\mathrm{T}}, \boldsymbol{\Gamma}_{22}^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{l\times K_2}, \quad \boldsymbol{\Gamma}_{12} \in \mathbb{R}^{K_1\times K_2}, \quad \boldsymbol{\Gamma}_{22} \in \mathbb{R}^{l_2\times K_2}$$

为对应变量的系数矩阵. 依据 (), 则式 (8.20) 的总体参数满足

$$\boldsymbol{\Gamma} = \left(\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{Z}\mathbf{Y}_2^{\mathrm{T}}\right), \quad \mathbb{E}\left(\mathbf{Z}\boldsymbol{u}_2^{\mathrm{T}}\right) = \mathbf{0}. \tag{8.21}$$

我们将式 (8.20) 代入结构方程式 (8.19) 中, 则有

$$\begin{aligned} Y_1 =& \mathbf{Z}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \left(\boldsymbol{\Gamma}_{12}^{\mathrm{T}}\mathbf{Z}_1 + \boldsymbol{\Gamma}_{22}^{\mathrm{T}}\mathbf{Z}_2 + \boldsymbol{u}_2\right)^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon \\ =& \mathbf{Z}_1^{\mathrm{T}}\left(\boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2\right) + \mathbf{Z}_2^{\mathrm{T}}\boldsymbol{\Gamma}_{22}\boldsymbol{\beta}_2 + \boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon, \end{aligned}$$

令

$$\boldsymbol{\lambda}_1 = \underbrace{\boldsymbol{\beta}_1}_{K_1\times 1} + \underbrace{\boldsymbol{\Gamma}_{12}}_{K_1\times K_2}\underbrace{\boldsymbol{\beta}_2}_{K_2\times 1} \in \mathbb{R}^{K_1\times 1}, \quad \boldsymbol{\lambda}_2 = \underbrace{\boldsymbol{\Gamma}_{22}}_{l_2\times K_2}\underbrace{\boldsymbol{\beta}_2}_{K_2\times 1} \in \mathbb{R}^{l_2\times 1}, \quad u_1 = \boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon,$$

则我们得到了 Y_1 的简约式为

$$Y_1 = \mathbf{Z}_1^{\mathrm{T}}\boldsymbol{\lambda}_1 + \mathbf{Z}_2^{\mathrm{T}}\boldsymbol{\lambda}_2 + u_1 = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\lambda} + u_1, \tag{8.22}$$

其中

$$\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2 \\ \boldsymbol{\Gamma}_{22}\boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{K_1} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{O}_{l_2\times K_1} & \boldsymbol{\Gamma}_{22} \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \overline{\boldsymbol{\Gamma}}\boldsymbol{\beta} \in \mathbb{R}^{l\times 1}, \tag{8.23}$$

$$\overline{\boldsymbol{\Gamma}} = \begin{pmatrix} \boldsymbol{I}_{K_1} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{O}_{l_2\times K_1} & \boldsymbol{\Gamma}_{22} \end{pmatrix} \in \mathbb{R}^{l\times K}, . \tag{8.24}$$

联立式 (8.22) 和式 (8.20), 我们便可得到 $\overrightarrow{Y} = \left(Y_1, \mathbf{Y}_2^{\mathrm{T}}\right)^{\mathrm{T}}$ 的简约式为

$$\begin{pmatrix} Y_1 \\ \mathbf{Y}_2 \end{pmatrix} = \mathbf{Z}^{\mathrm{T}}\begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\Gamma} \end{pmatrix} + \begin{pmatrix} u_1 \\ \boldsymbol{u}_2 \end{pmatrix} \Leftrightarrow \overrightarrow{Y} = \begin{pmatrix} \boldsymbol{\lambda}_1^{\mathrm{T}} & \boldsymbol{\lambda}_2^{\mathrm{T}} \\ \boldsymbol{\Gamma}_{12}^{\mathrm{T}} & \boldsymbol{\Gamma}_{22}^{\mathrm{T}} \end{pmatrix}\mathbf{Z} + \boldsymbol{u}. \tag{8.25}$$

[2]这里的误差项 $\boldsymbol{\mu}_2$ 与工具变量 $\mathbf{Z}$ 应是正交的, 即 $\mathbb{E}\left(\mathbf{Z}\boldsymbol{u}_2^{\mathrm{T}}\right) = \boldsymbol{O}$.

对于式 (8.22) 和式 (8.20), 则投影模型的总体参数为

$$\boldsymbol{\lambda} = \left(\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{Z}Y_1\right), \tag{8.26}$$

$$\boldsymbol{\Gamma} = \left(\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{Z}\mathbf{Y}_2^{\mathrm{T}}\right), \tag{8.27}$$

可见正定性式 (8.8) 保证了参数 $\boldsymbol{\lambda}$ 和 $\boldsymbol{\Gamma}$ 的存在性, 这样, 我们有矩估计量

$$\hat{\boldsymbol{\lambda}} = \left(\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_iY_{1i}\right) = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}_1, \tag{8.28}$$

$$\hat{\boldsymbol{\Gamma}} = \left(\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Y}_{2i}^{\mathrm{T}}\right) = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}_2. \tag{8.29}$$

简约式的意义在于, 其为我们提供了估计结构参数 $\boldsymbol{\beta}$ 的可行方法. 由于工具变量是外生的, 因而对于 Y_1 的简约式式 (8.22) 而言,

$$Y_1 = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\lambda} + u_1 = \boxed{\mathbf{Z}^{\mathrm{T}}\overline{\boldsymbol{\Gamma}}}\boldsymbol{\beta} + u_1$$

扰动项 u_1 与工具变量 $\mathbf{Z}$ 是正交的, 故 OLS 估计量不会存在偏差. 进一步地, 利用式 (8.23) 我们将式 (8.22) 可以等价变换为关于结构参数 $\boldsymbol{\beta}$ 的线性回归模型, 其中解释变量为 $\overline{\boldsymbol{\Gamma}}^{\mathrm{T}}\mathbf{Z}$, 即是工具变量 $\mathbf{Z}$ 的线性组合.

我们自然想知道线性组合 $\overline{\boldsymbol{\Gamma}}^{\mathrm{T}}\mathbf{Z}$ 的含义. 对于矩阵 $\overline{\boldsymbol{\Gamma}}$, 依据式 (8.24), 考虑关于结构方程式 (8.3) 的解释变量 $\mathbf{X} = \left(\mathbf{Z}_1^{\mathrm{T}}, \mathbf{Y}_2^{\mathrm{T}}\right)^{\mathrm{T}}$, 则有

$$\mathbf{X} = \begin{pmatrix}\mathbf{Z}_1\\ \mathbf{Y}_2\end{pmatrix} = \begin{pmatrix}\mathbf{Z}_1\\ \boldsymbol{\Gamma}_{12}^{\mathrm{T}}\mathbf{Z}_1 + \boldsymbol{\Gamma}_{22}^{\mathrm{T}}\mathbf{Z}_2 + \boldsymbol{u}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{I}_{K_1} & \boldsymbol{O}_{l_2\times K_1}\\ \boldsymbol{\Gamma}_{12}^{\mathrm{T}} & \boldsymbol{\Gamma}_{22}^{\mathrm{T}}\end{pmatrix}\begin{pmatrix}\mathbf{Z}_1\\ \mathbf{Z}_2\end{pmatrix} + \begin{pmatrix}\mathbf{0}\\ \boldsymbol{u}_2\end{pmatrix},$$

因而 $\overline{\boldsymbol{\Gamma}}$ 是简约式

$$\mathbf{X} = \overline{\boldsymbol{\Gamma}}^{\mathrm{T}}\mathbf{Z} + \boldsymbol{\mu}, \quad \boldsymbol{\mu} = \left(\mathbf{0}, \boldsymbol{u}_2^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{8.30}$$

的系数矩阵. 由此我们有 $\overline{\boldsymbol{\Gamma}}$ 的总体参数以及对应的矩估计[3]

$$\overline{\boldsymbol{\Gamma}} = \left(\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{Z}\mathbf{X}^{\mathrm{T}}\right), \tag{8.31}$$

$$\widehat{\overline{\boldsymbol{\Gamma}}} = \left(\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{X}_i^{\mathrm{T}}\right) = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}. \tag{8.32}$$

8.4 两阶段最小二乘法

上一节中我们已经介绍了如何通过简约式将工具变量 $\mathbf{Z}$ 和结构参数 $\boldsymbol{\beta}$ 联系在一起, 本节我们介绍在实证研究中更为常见的两阶段最小二乘估计量.

[3]这里 OLS 估计量不存在内生性问题, 由于式 (8.20) 中误差项 $\boldsymbol{\mu}_2$ 与工具变量 $\mathbf{Z}$ 正交, 则可知 $\mathbb{E}\left(\mathbf{Z}\boldsymbol{\mu}^{\mathrm{T}}\right) = \mathbb{E}\left[\mathbf{Z}\left(\mathbf{0}, \boldsymbol{u}_2^{\mathrm{T}}\right)\right] = \left(\mathbf{0}, \mathbb{E}\left(\mathbf{Z}\boldsymbol{u}_2^{\mathrm{T}}\right)\right) = \boldsymbol{O}$.

8.5 弱工具变量

8.6 识别的假设检验

8.7 代码

第九章　广义矩估计

第十章　重抽样方法

第十一章　时间序列

第十二章　面板数据

附录 A　数学工具

A.1　线性代数

A.2　概率论

A.3　矩阵微积分

A.4　最优化方法

概念索引

符号

B

C

D

E

I

J

K

L

M

N

O

P

Q

R

S

T

W

X

Y

Z

模型索引