计量经济学

Ender

2023 年 12 月 5 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

给定 Random Sample

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

样本

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

$$\mathbb{E}\,(Wage|School)$$

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 $\mathbb{E}\left(Y|X=x\right)$ 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 $f_{XY}\left(x,y\right)$, 那么 Y 关于 X 的条件期望函数为

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

其中

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

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 $\mathbb{E}\left(Y|X=x\right)$ 记为 $m\left(x\right)$.

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 $\mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right)=\mathbb{E}Y$.*

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 $\boxed{\mathbb{E}_X\left(\mathbb{E}_Y\left(Y|X\right)\right)=\mathbb{E}Y}$. 可以直观理解, 首先是对 Y 求条件期望, 得到 $\mathbb{E}\left(Y|X\right)$ 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

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

设随机变量 X 和 Y 均为连续性随机变量, 则有

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

[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

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

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<u>工资</u>和<u>性别</u>, 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. $\mathbb{E}\left(e|X\right) = 0$*;*

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

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

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

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

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

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

在概率论中, 如果随机变量 X 和 Y 满足有 $\mathbb{E}\left(Y|X = x\right) = \mathbb{E}Y$, 则称 Y 对于 X **均值独立** (mean independence)[8], 或 Y 均值独立于 X. 反过来,$\mathbb{E}\left(X|Y\right) = \mathbb{E}X$ 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 $\mathbb{E}\left(Y|X\right) = 0$, 同理性质 b 的证明, 即得 $\mathbb{E}\left(Y|X\right) = 0 = \mathbb{E}$. 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<u>条件期望函数 $\mathbb{E}\left(Y|X\right)$ 对被解释变量 Y 的解释或预测</u>的偏离程度,$\mathbb{E}e = 0$ 表明在平均意义上 $\mathbb{E}\left(Y|X\right)$ 对 Y 的偏差是不存在的, 仅给定任意的随机变量 $X = x$ 时 $\mathbb{E}\left(e|X = x\right) = 0$ 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 $\mathbb{E}\left(Y - g(X)\right)^2$ 最小的函数为给定 X 下 Y 的***最优预测量** *(best predictor).*

这样 $\mathbb{E}\left(Y - g(X)\right)^2$ 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, <u>随机变量的独立性</u>是<u>随机变量均值独立</u>的必要条件, <u>随机变量均值独立</u>是<u>随机变量线性无关</u>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 $\mathbb{E}\left(Y|X\right)$ 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*

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

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

证明. 注意到

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

而依据定理 (2.1.3) 有

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

故有

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

成立, 当且仅当 $g(X)=m(X)=\mathbb{E}\left(Y|X\right)$ 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 $\mathbb{E}\left(Y-g(X)\right)^2$ 存在一个可达的下界 $\mathbb{E}e^2$, 在非 $g(X)=m(X)$ 的情况下 $\mathbb{E}\left(Y-g(X)\right)^2$ 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 $\mathbb{E}\left(Y|X\right)$ 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),$\sigma = \sqrt{\mathrm{Var}\left(e\right)}$ 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<u>线性 CEF 模型</u>又被称为<u>线性回归模型</u>. 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

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

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 为取值为 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<u>$\boldsymbol{\beta}$ 的线性函数</u>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

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

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) $\boldsymbol{Q}_{\mathbf{XX}}=\mathbb{E}\left(\mathbf{XX}^{\mathrm{T}}\right)$ *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

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

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

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

令偏导数为 0, 得到极值点的一阶条件为

$$\begin{cases}\mathbb{E}(Y-\beta_0-\beta_1X)=0,\\ \mathbb{E}(X(Y-\beta_0-\beta_1X))=0,\end{cases}\Leftrightarrow\begin{cases}\mathbb{E}e=0,\\ \mathbb{E}(Xe)=0,\end{cases}\tag{2.10}$$

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

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

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

将式 (2.12) 代入式 (2.13) 中, 消去 β_0 得

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

[15]也有翻译为“常规方程组”, 见林文夫 (cite).

[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.

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

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

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).*

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

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

$$Y=\mathscr{P}(Y|X)+\varepsilon=\beta_0+\beta_1X+\varepsilon. \tag{2.14}$$

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<u>含有常数项 β_0</u>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

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

故在不含有常数项的线性投影模型中并不能够保证 $\mathbb{E}(e)=\mathbb{E}(Y-\beta_1 X)=0$.

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

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

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<u>X 与投影误差 ε 不相关</u>, 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

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

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

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

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 $\hat{\varepsilon}_i = Y_i - (\beta_0 + \beta_1 X_i)$. 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

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

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

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

[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).

其中

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

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

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

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 $\boldsymbol{\beta}=\left(\beta_1, \beta_1\right)^{\mathrm{T}}$ 的估计量. 需要强调, 参数 $\boldsymbol{\beta}=\left(\beta_1, \beta_1\right)^{\mathrm{T}}$ 是由总体确定的<u>非随机的</u>、<u>固定的</u>的数值, 而估计量 $\left(\hat{\beta}_0, \hat{\beta}_1\right)^{\mathrm{T}}$ 则是<u>随样本变化</u>的<u>随机变量</u>. 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

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

展开即有

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

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

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

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

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

---

[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.

以及

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

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

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

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

$$\begin{cases}\sum\limits_{i=1}^{n}(Y_i-\beta_0-\beta_1X_i)=0, \\ \sum\limits_{i=1}^{n}[(Y_i-\beta_0-\beta_1X_i)X_i]=0,\end{cases} \Rightarrow \begin{cases}\sum\limits_{i=1}^{n}\hat{\varepsilon}_i=0, \\ \sum\limits_{i=1}^{n}\hat{\varepsilon}_iX_i=0,\end{cases} \tag{2.18}$$

其中 $\hat{\varepsilon}_i=Y_i-\hat{\beta}_0-\hat{\beta}_1X_i$ 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

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

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

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

解得

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

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

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

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

[20]上述一阶条件中并没有给 β 标上 ^, 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.

[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.

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*

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{2.20}$$

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right) Y_i}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \\ &= \frac{\sum_{i=1}^{n} Y_i X_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2},\end{aligned} \tag{2.21}$$

且 OLS 估计量满足有

$$\begin{cases}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0, \\ \sum_{i=1}^{n}\left[\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) X_i\right] = 0,\end{cases} \Rightarrow \begin{cases}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0, \\ \sum_{i=1}^{n} \hat{\varepsilon}_i X_i = 0.\end{cases}$$

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 估计量以及其衍生变量的有限性质, 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*

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

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, <u>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).</u>

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**. <u>对于大多数的计量经济学教材, 其讨论的线性回归模型便是模型 (2.4.1).</u> 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

[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.

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

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

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 $\mathbb{E}\left(Y_i|\boldsymbol{X}\right)$ 则有,

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

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 $E\left(Y_i|\, X_i\right) = \mathbb{E}\left(Y|\, X_i\right)$ 成立, 这意味意味着所以的样本个体 Y_i<u>具有相同函数形式的条件期望</u>$m\left(\cdot\right)$, 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 $\mathbb{E}\left(Y_i|\, \boldsymbol{X}\right)$ 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

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

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

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

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 $\mathbb{E}\left(\varepsilon_i|\, \boldsymbol{X}\right)$ 的取值为 0. 依定理 (2.1.2), 这又意味 $\mathbb{E}\varepsilon_i$ 着而 $\mathbb{E}\left(\varepsilon_i^2|\, \boldsymbol{X}\right)$ 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

$$\begin{aligned}
\boxed{\mathrm{Var}\left(Y_i|\, \boldsymbol{X}\right)} &= \mathrm{Var}\left(Y_i|\, X_i\right) = \mathrm{Var}\left(\beta_0 + \beta_1 X_i + \varepsilon_i|\, X_i\right) = \boxed{\mathrm{Var}\left(\varepsilon_i|\, X_i\right)} \\
&= \mathbb{E}\left(\left(\varepsilon_i - \mathbb{E}\left(\varepsilon_i|\, X_i\right)\right)^2\middle|\, X_i\right) = \boxed{\mathbb{E}\left(\varepsilon_i^2\middle|\, X_i\right)} = \sigma^2.
\end{aligned}$$

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

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

即样本均值 $\bar{Y}$ 的条件期望是<u>条件期望函数 $m(X)$ 的样本均值</u>, 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

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

若是满足同方差假设的模型 (2.4.2), 则有 $\operatorname{Var}\left(\bar{Y}\middle|\boldsymbol{X}\right)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=n^{-1}\sigma^2$, 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

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

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

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

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

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

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

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

故 $\hat{\beta}_0$ 的条件期望为 $\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)=\beta_1\cdot 1+A\cdot 0=\beta_1$. 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

$$\begin{aligned}&\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)=\mathbb{E}\left(A\left(\boldsymbol{X}\right)\boldsymbol{b}\left(\boldsymbol{X}\right)^{\mathrm{T}}\boldsymbol{Y}\middle|\boldsymbol{X}\right)\xlongequal{\text{向量乘法是线性变换}}A\boldsymbol{b}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{Y}\middle|\boldsymbol{X}\right)\\=&A\boldsymbol{b}^{\mathrm{T}}\left(\boldsymbol{1}\beta_0+\boldsymbol{X}\beta_1\right)=A\beta_0\boldsymbol{b}^{\mathrm{T}}\boldsymbol{1}+A\beta_1\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}=A\beta_0\cdot 0+A\beta_1\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}=A\beta_1\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X},\end{aligned}$$

而 $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}=\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)X_i=A^{-1}$, 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)**.** *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

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

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

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

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

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

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

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

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

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

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

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

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

及

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

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]

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

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

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

和

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

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

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

---

[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.

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 $\sum_{i=1}^{n}C_i^2-A$. 由 $A^{-1}=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}$,$B_i=X_i-\bar{X}$, 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

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

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 $Y_1=\beta_0+\beta_1X_i+e_i$ 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.

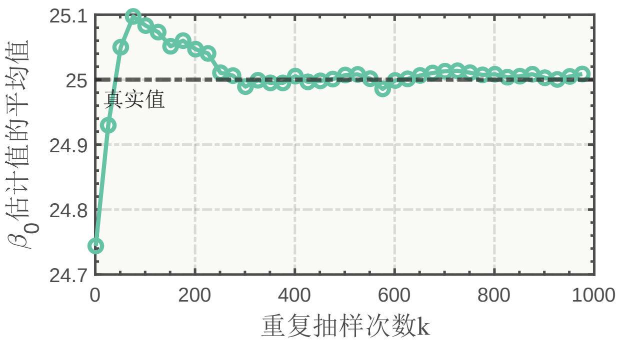

图 2.1: $\hat{\beta}_0$ 的样本均值

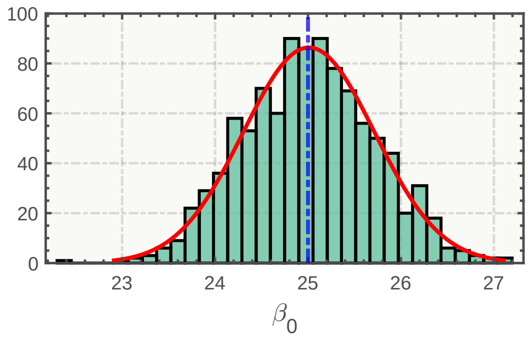

图 2.2: $\hat{\beta}_0$ 的直方图

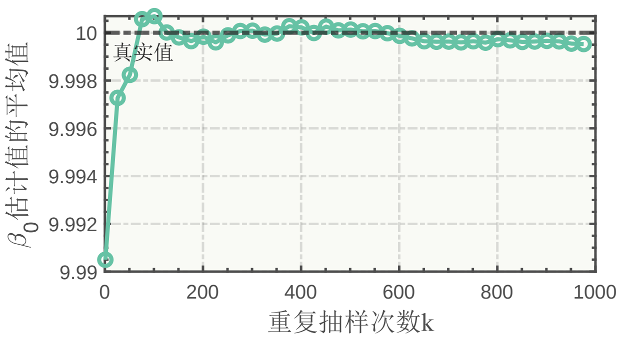

图 2.3: $\hat{\beta}_1$ 的样本均值

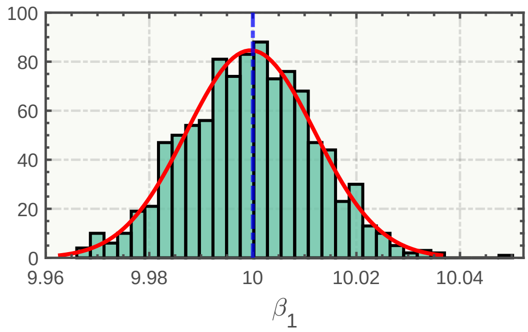

图 2.4: $\hat{\beta}_1$ 的直方图

图 (2.1) 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.

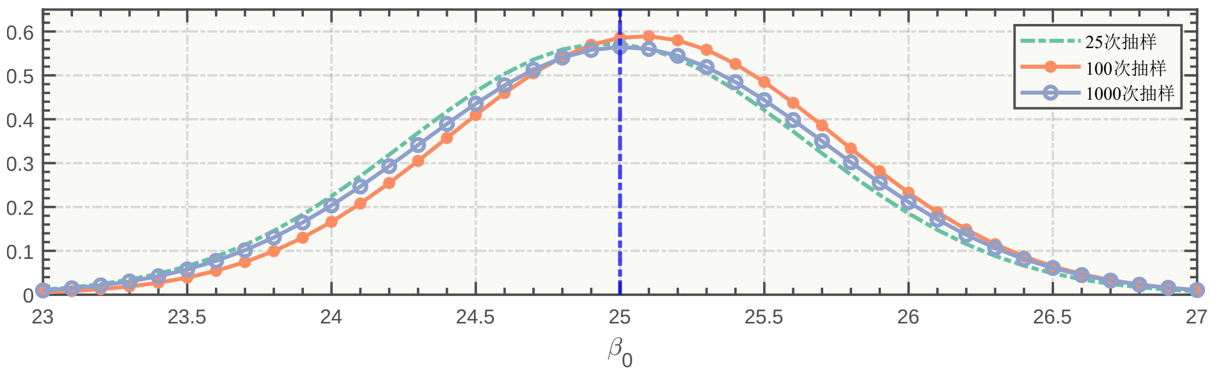

图 2.5: 重复抽样下 $\hat{\beta}_0$ 的概率密度函数

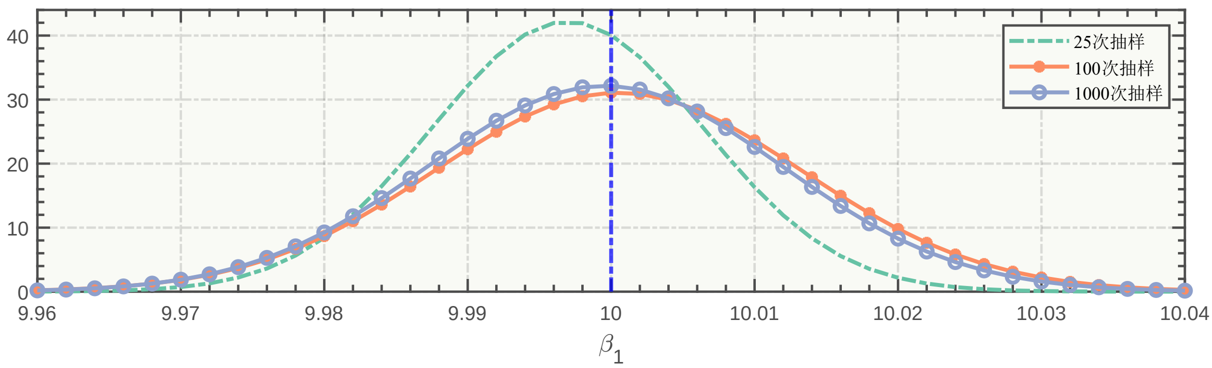

图 2.6: 重复抽样下 $\hat{\beta}_1$ 的概率密度函数

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* $\hat{\varepsilon}_i = \hat{Y} - Y_i$

[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.

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 $\hat{\beta}_1 = A \sum\limits_{i=1}^{n} B_i Y_i = A\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}$, 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

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

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

$$\begin{aligned}
\hat{\boldsymbol{Y}} &= \left(\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1}\left(\mathbf{1}\mathbf{1}^{\mathrm{T}}\right) + \left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}\right)\boldsymbol{Y} \xlongequal{\text{内积视为}1\times1\text{矩阵}} \left(\mathbf{1}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1}\mathbf{1}^{\mathrm{T}} + \boldsymbol{b}\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\boldsymbol{b}^{\mathrm{T}}\right)\boldsymbol{Y} \\
&= \left(\mathbf{1}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1}, \boldsymbol{b}\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\right)\begin{pmatrix}\mathbf{1}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}}\end{pmatrix}\boldsymbol{Y} = (\mathbf{1}, \boldsymbol{b})\begin{pmatrix}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)^{-1} & 0 \\ 0 & \left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}\right)^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{1}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}}\end{pmatrix}\boldsymbol{Y},
\end{aligned}$$

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

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

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

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

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 $\boldsymbol{P} = \boldsymbol{X}^* \left((\boldsymbol{X}^*)^T \boldsymbol{X}^* \right)^{-1} (\boldsymbol{X}^*)^T$ 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 (3.7.1) 的证明. 依据式 (2.31), 对于性质 (a) 直接计算有 $\boldsymbol{P}\boldsymbol{X}^* = \boldsymbol{X}^* \left((\boldsymbol{X}^*)^T \boldsymbol{X}^* \right)^{-1} (\boldsymbol{X}^*)^T \boldsymbol{X}^* = \boldsymbol{X}^*$.

性质 (b) 计算得

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

性质 (c) 计算得

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

性质 (d) 计算得

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

□

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

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

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) $\boldsymbol{M}\boldsymbol{X}^* = \boldsymbol{O}, \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\varepsilon} = \hat{\boldsymbol{\varepsilon}}$*;*

(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 $\boldsymbol{M}\boldsymbol{X}^* = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X}^* = \boldsymbol{X}^* - \boldsymbol{P}\cdot\boldsymbol{X}^* = \boldsymbol{X}^* - \boldsymbol{X}^* = \boldsymbol{O}_{n\times 2}$ 及 $\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}(\boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{M}\boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon} = \hat{\boldsymbol{Y}}$.

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 $\operatorname{tr}\boldsymbol{M} = \operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{P}) = \operatorname{tr}\boldsymbol{I}_n - \operatorname{tr}\boldsymbol{P} = n - 2$. □

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

$$\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=(\boldsymbol{P}\boldsymbol{Y})^{\mathrm{T}}(\boldsymbol{M}\boldsymbol{Y})=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{M}\boldsymbol{Y}=\boldsymbol{Y}^{\mathrm{T}}\left[\boldsymbol{P}\left(\boldsymbol{I}-\boldsymbol{P}\right)\right]\boldsymbol{Y}=0,$$

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** (re-*

gression sum of square), $\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2$ 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

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

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 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 $\sum\limits_{i=1}^{n}e_i^2=\sum\limits_{i=1}^{n}(Y_i-\beta_0-\beta_1X_i)^2$ 称为了 SSE, 其是回归误差 e 的平方和, 是不可观察的.

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

$$\begin{aligned}\mathbb{E}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)&=\mathbb{E}\left(\hat{\beta}_0+\hat{\beta}_1X_i\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\hat{\beta}_0\middle|\boldsymbol{X}\right)+X_i\mathbb{E}\left(\hat{\beta}_1\middle|\boldsymbol{X}\right)\\&=\boxed{\beta_0+\beta_1X_i}=\mathbb{E}\left(Y|X_i\right),\end{aligned}$$

[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.

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

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

则有

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

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 $\mathrm{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\middle|\,\boldsymbol{X}\right)$. 由方差的分配律

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

得

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

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

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

可以推导出

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

即<u>样本均值 $\bar{Y}$ 与 OLS 估计量 $\hat{\beta}_1$ 完全不相关</u>, 故

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

回到拟合值 $\hat{Y}_i$ 的条件方差的计算, 于是

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

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 $\boxed{\operatorname{Var}\left(\hat{Y}_i\middle|\boldsymbol{X}\right)=h_{ii}\sigma^2}$.

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 $\boxed{\operatorname{Var}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right)}=\mathbb{E}\left(\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right)=\mathbb{E}\left[\left(Y_i-\hat{Y}_i\right)^2\middle|\boldsymbol{X}\right]$. 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

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

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

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

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

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

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

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

也就是说, 同方差假设下,$\operatorname{Var}\left(\hat{\varepsilon}_i \mid \boldsymbol{X}\right) = \left(1 - h_{ii}\right)\sigma^2 < \sigma^2 = \operatorname{Var}\left(\varepsilon_i \mid \boldsymbol{X}\right)$, 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

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

即 σ^2_{ME} 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

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

这里的投影矩阵 $\boldsymbol{P}$ 是半正定, 取期望则有 $\sigma^2 = \mathbb{E}\sigma^2_{\mathrm{ME}} \geqslant \mathbb{E}\hat{\sigma}^2$, 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 $\mathrm{Var}\left(\hat{\varepsilon}_i|\,\boldsymbol{X}\right) = \mathbb{E}\left(\hat{\varepsilon}_i^2|\,\boldsymbol{X}\right) = \left(1 - h_{ii}\right)\sigma^2$, 则

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

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

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

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

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

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).*

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

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 $\varepsilon|\, X \sim \mathrm{N}\left(0, \sigma^2\right)$, 这里的正态分布 $\mathrm{N}\left(0, \sigma^2\right)$ 是**不依赖**于解释变量 X, 也就是说<u>条件随机变量 $\varepsilon|\, X$ 的各种性质均与 X 无关</u>. 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 $\varepsilon_i|\, X_i \sim \mathrm{N}\left(0, \sigma^2\right)$, 因而对于误差向量便有 $\boldsymbol{\varepsilon}|\, \boldsymbol{X} \sim \mathrm{N}\left(0, \sigma^2 \boldsymbol{I}_n\right), \boldsymbol{\varepsilon}|\, \boldsymbol{X}$ 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

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

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 $Y|\, X \sim \mathrm{N}(\beta_0+\beta_1 X,\sigma^2)$, 相应地 Y 的样本个体具有条件分布 $Y_i|\, X \sim \mathrm{N}(\beta_0+\beta_1 X_i,\sigma^2)$.

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

$$\hat{Y}_i\Big|\,\boldsymbol{X} \sim \mathrm{N}\left(\beta_0+\beta_1 X_i, h_{ii}\sigma^2\right), \quad \hat{\varepsilon}_i|\,\boldsymbol{X} \sim \mathrm{N}\left(0,(1-h_{ii})\,\sigma^2\right). \tag{2.37}$$

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

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

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

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

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

$$Y_i - \bar{Y} = \beta_1 B_i + (\varepsilon_i - \bar{\varepsilon}). \tag{2.39}$$

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

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

即<u>$\hat{\beta}_1 - \beta_1$ 是回归误差 $\varepsilon_1, \cdots, \varepsilon_n$ 的线性组合</u>, 而 $\varepsilon_i | X_i$ i.i.d. $\sim \mathrm{N}(0, \sigma^2), i = 1, \cdots, n$, 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 $\mathrm{RSS} = \sum\limits_{i=1}^{n} \hat{\varepsilon}_i^2$ 进行分解, 不难算得

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

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

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

而对于样本 2 阶中心距, 我们熟知 $\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2=\sum_{i=1}^{n}\varepsilon_i^2-n\bar{\varepsilon}^2$, 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<u>互相独立</u>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*

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

*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

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

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 $F(c)=1-\frac{\alpha}{2}$, 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 $c=F^{-1}\left(1-\frac{\alpha}{2}\right)$. 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, 请见第节.

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
selfGrootDefault(2)%绘图美化函数,没有请注释
%绘制估计值均值变化
figure(1)
set(gcf,'unit','centimeters', ...
'Position',[0 0 21*1.25 29.7*0.35*1.25]);%设置figure为A4纸宽度
tiledlayout(2,2);%创建2*2画布
nexttile
histfit(Beta(1,:))%beta0的直方图和拟合的pdf
xline(25,'-.b','LineWidth',2)
xlabel('\beta_0','FontSize',12.5)

nexttile
plot(1:25:k,beta_hat(1,1:25:k),'-o')
xlabel('\fontname{宋体}重复抽样次数\fontname{Times New ...
    Roman}k','FontSize',12.5)
ylabel('\beta_0\fontname{宋体}估计值的平均值','FontSize',12.5);
yline(25,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
    ...
,'left','LabelVerticalAlignment','bottom')

nexttile
histfit(Beta(2,:))%beta1的直方图和拟合的pdf
xline(10,'-.b','LineWidth',2)
xlabel('\beta_1','FontSize',12.5)

nexttile
plot(1:25:k,beta_hat(2,1:25:k),'-o')
xlabel('\fontname{宋体}重复抽样次数\fontname{Times New ...
    Roman}k','FontSize',12.5)
ylabel('\beta_0\fontname{宋体}估计值的平均值','FontSize',12.5);
yline(10,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
    ...
,'left','LabelVerticalAlignment','bottom')
%绘制重复抽样pdf
figure(2)
set(gca,'FontName','Times New Roman');
set(gcf,'unit','centimeters', ...
'Position',[0 0 21*1.25 29.7*0.4*1.25]);%设置figure为A4纸宽度
```

```
tiledlayout(2,1);%创建2*2画布
nexttile
pd1=fitdist(Beta(1,1:25)','Normal');%beta0拟合的pdf_25次抽样
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

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

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

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

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

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

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

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

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). 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

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

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

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

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

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

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

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

再依据定理 (2.1.3) 有

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

即回归误差 ε_i 与任意的样本观测值 X_{jk} 是正交的, 故有

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

特别地, 当 $i=j$ 时 $\mathrm{Cov}\left(X_{ik},\varepsilon_i\right)$ 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 $\mathbb{E}\left(\boldsymbol{Y}|\,\boldsymbol{X}\right)=\boldsymbol{X}\boldsymbol{\beta}$ 和 $\mathbb{E}\left(\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)=\mathbf{0}_{n\times 1}$.

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 $\mathbf{X}=\left(X_1,\cdots,X_K\right)^{\mathrm{T}}$ 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 $n\geqslant\mathrm{rank}\left(\boldsymbol{X}_{n\times K}\right)=K$ 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 $\mathbf{X}=\left(X_1,\cdots,X_n\right)^{\mathrm{T}}$, 给定条件 Z 下其**条件协方差矩阵** (conditional covariance matrix)定义为

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

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

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

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

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

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

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

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

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

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

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

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

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

即有 $\mathbb{E}\left(\mathbf{X}\mathbf{X}^T\right)\boldsymbol{\beta} = \mathbb{E}(\mathbf{X}Y)$ 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

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

在式 (3.13) 中,$\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^T\right)\right)^{-1}$ 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 $\mathbb{E}\left(\mathbf{X}\mathbf{X}^T\right)\boldsymbol{\beta} = \mathbb{E}(\mathbf{X}Y)$ 中的 Y 替换有

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

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]详见参考资料 (cite)

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

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

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

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

对式 (3.14) 求期望则有

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

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

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

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

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

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

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

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 $\alpha = \mathbb{E}Y - \mathbb{E}\left(\mathbf{X}^T\right)\boldsymbol{\beta}$.

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).*

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

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).*

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

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

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

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

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

而误差平方和 $\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(Y_i - {\mathbf{X}_i}^T\boldsymbol{\beta}\right)^2$. 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 SSE$(\boldsymbol{\beta})$ 的极值. 由于

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

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

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

也即有

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

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 $\left\{(X_{11},\cdots,X_{n1})^{\mathrm{T}},\cdots,(X_{1k},\cdots,X_{nk})^{\mathrm{T}},\cdots(X_{1K},\cdots,X_{nK})^{\mathrm{T}}\right\}$ 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 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

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

则

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

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

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

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

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

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

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

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

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 $\mathbb{E}\left(\mathbf{X}\varepsilon_i\right)=0$, 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

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

则求样本均值得 $\bar{Y}=\alpha+\bar{\mathbf{X}}^T\boldsymbol{\beta}+\bar{\varepsilon}$, 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

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

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 $\hat{\alpha}=\bar{Y}-\bar{\mathbf{X}}^T\hat{\boldsymbol{\beta}}$. 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*

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

且 OLS 估计量满足有

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

3.5 多元线性投影模型的分块回归

3.6 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 3.6.1 (多元线性回归模型). *在假设 (2.2.1) 以及*

$$\mathbb{E}\left(Y|\,\mathbf{X}\right)=\mathscr{P}\left(Y|\mathbf{X}\right)\quad \text{或}\quad \mathbb{E}\left(\varepsilon|\,\mathbf{X}\right)=0$$

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

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

模型 (3.6.1) 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 (3.6.1) 的**抽样误差** (sampling error)即为

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

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 $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\left(\beta_1,\cdots,\beta_K\right)^{\mathrm{T}}$ 是一个 $K\times 1$ 维的**随机向量**, 则依据定理 (3.2.1) 得到

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

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

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

这样, 对于模型 (3.6.1) 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 (3.6.1) 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 3.6.1 (OLS 估计量的协方差矩阵). *对于模型 (3.6.1) 的 OLS 估计量 $\hat{\boldsymbol{\beta}}$, 我们记 $\hat{\boldsymbol{\beta}}$ 的条件协方差为 $\boldsymbol{V}_{\hat{\beta}}$, 则条件异方差式 (3.9) 下有*

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

当条件同方差式 (3.7) 成立时则有 $\boldsymbol{V}_{\hat{\beta}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2$.

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 3.6.2 (OLS 估计量的无偏性). *模型 (3.6.1) 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 $\mathbb{E}\left(\hat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{\beta}\middle|\boldsymbol{X}\right)=\boldsymbol{\beta}$. 当然亦可直接计算

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

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 (3.6.1) 的**最优线性无偏估计量** (best linear unbiased estimator), 即 Gauss-Markov 定理.

定理 3.6.3 (Gauss-Markov 定理). *设模型 (3.6.1) 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 (3.6.1) 的任意线性无偏估计量 $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}(\boldsymbol{X})\boldsymbol{Y}$, 则有*

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

这里 $\geqslant$ 表示 $\mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right) - \boldsymbol{V}_{\hat{\beta}}$ 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 $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}(\boldsymbol{X})\boldsymbol{Y}$ 是参数 $\hat{\boldsymbol{\beta}}$ 的无偏估计量, 而

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

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

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

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

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

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

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

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

[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 $\mathbb{E}\left(\boldsymbol{A}\boldsymbol{X}\right) = \boldsymbol{I}_K$.

成立. $\square$

定理 (3.6.3) 表明了 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 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 (3.6.1) 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.

3.7 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 $\boldsymbol{f} \circ \boldsymbol{f} = \boldsymbol{f}(\boldsymbol{f}(\cdot)) = \boldsymbol{f}$, 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 $\langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{f}(\boldsymbol{y}) \rangle$, 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 $\boldsymbol{P_l} \in \mathbb{R}^{2\times 2}$, 则点 $\boldsymbol{x}$ 在直线 $\boldsymbol{l}$ 上的投影点为 $\boldsymbol{P_l x}$.

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** (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

$$
\begin{aligned}
\boldsymbol{P_l} = \underset{\boldsymbol{P_l} \in \mathbb{R}^{n\times n}}{\operatorname{argmin}} \|\boldsymbol{P_l x} - \boldsymbol{x}\| &= \underset{\boldsymbol{P_l} \in \mathbb{R}^{n\times n}}{\operatorname{argmin}} \sqrt{(\boldsymbol{P_l x} - \boldsymbol{x})^{\mathrm{T}} (\boldsymbol{P_l x} - \boldsymbol{x})}, \\
s.t. \quad \boldsymbol{P_l x} = \boldsymbol{l} t, &\quad t \in \mathbb{R},
\end{aligned}
\tag{3.23}
$$

[4] $\boldsymbol{l}$ 过原点确保了子空间是封闭的.

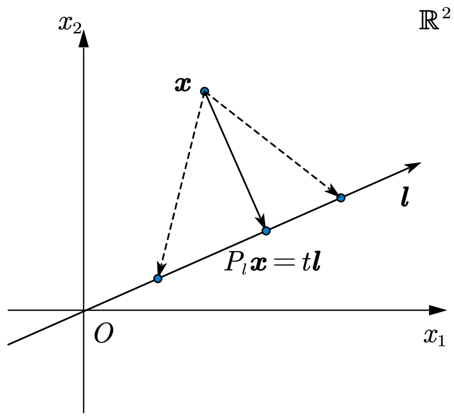

图 3.1: 二维平面中的正交投影

其中 $\boldsymbol{P_l x} = t\boldsymbol{l}$ 为约束条件, 使得投影点在 $\boldsymbol{l}$ 之上. 不难发现, 式 (3.23) 与 OLS 估计量式 (3.17) 在形式上是相似的. 为了求解式 (3.23), 一种思路即是 Lagrange 乘数法, 但式 (3.23) 的约束条件非常简单, 直接代入优化函数得

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

对 $s(t)$ 求导得到一阶条件

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

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

$$\boldsymbol{P_l} = \boldsymbol{l}\left(\boldsymbol{l}^{\mathrm{T}}\boldsymbol{l}\right)^{-1}\boldsymbol{l}^{\mathrm{T}} \tag{3.24}$$

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

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

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

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 (3.24) 中的 $\boldsymbol{l}$ 替换为 $\boldsymbol{L}$, 则得到 n 阶方阵

$$\boldsymbol{P_L} = \boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}, \tag{3.25}$$

则式 (3.25) 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 (3.6.1) 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<u>数据矩阵 $\boldsymbol{X}$ 的列向量组所张成的线性空间</u>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

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

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 $\hat{\varepsilon}$ 即是连接了 $\hat{\boldsymbol{Y}}$ 和 $\boldsymbol{Y}$ 两者终点的向量, 且残差向量 $\hat{\varepsilon}$ 同平面 $\boldsymbol{L}$ 是垂直的, 这样结果如图 (3.3) 所示.

下面我们研究式 (3.25) 的投影矩阵的性质.

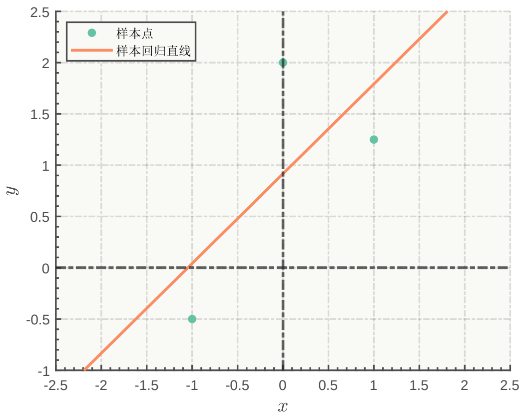

图 3.2: 一元线性回归模型的拟合直线

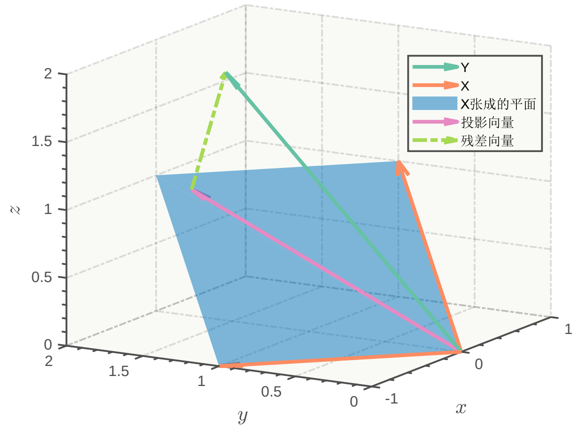

图 3.3: 一元线性回归模型的正交投影

定理 3.7.1 (投影矩阵的性质). *对于式 (3.25) 的投影矩阵 $\boldsymbol{P_L}=\boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}$,$\boldsymbol{L}\in\mathbb{R}^{n\times K}$,$n\geqslant K$, 其满足有如下性质:*

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

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

(c) 投影矩阵的 $\boldsymbol{P_L}$ 的迹为 K;

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

(e) 投影矩阵 $\boldsymbol{P_L}$ 的秩为 K.

证明. (a) 计算有 $\boldsymbol{P_L}^2=\boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^{\mathrm{T}}=\boldsymbol{P_L}$, 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 $\boldsymbol{P_L L}=\boldsymbol{L}$ 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 $\boldsymbol{P_L L}$ 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

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

则 $\boldsymbol{P_L}^{\mathrm{T}}=\boldsymbol{P_L}$ 是 $\boldsymbol{L}^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{P_L x}\right)=0$ 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

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

即证性质 (c). 而方阵的特征值的和又满足有

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

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

$$\boldsymbol{M_L} = \boldsymbol{I}_n - \boldsymbol{P_L} \tag{3.26}$$

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

定理 3.7.2 (归零矩阵的性质)**.** *对于形如式 (3.26) 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 (3.7.1) 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 3.7.3 (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 (3.7.3) 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

$$\boldsymbol{A}=\boldsymbol{H}^{-1}\begin{pmatrix}\boldsymbol{I}_r & \boldsymbol{O}_{r\times(n-r)}\\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{O}_{n-r}\end{pmatrix}\boldsymbol{H}. \tag{3.27}$$

定理 (3.7.3) 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 3.7.4 (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 (3.7.3) 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 $\boldsymbol{\Lambda}=\operatorname{diag}(a_1,\cdots,a_n)$, 其所有对角元都是正数, 则 $\boldsymbol{\Lambda}$ 的平方根为 $\boldsymbol{\Lambda}^{1/2}=\operatorname{diag}\left(\sqrt{a_1},\cdots,\sqrt{a_n}\right)$.

3.8 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

$$\begin{aligned}\hat{\boldsymbol{Y}}&=\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=\boxed{\boldsymbol{PY}},\\ \hat{\boldsymbol{\varepsilon}}&=\hat{\boldsymbol{Y}}-\boldsymbol{Y}=\boldsymbol{PY}-\boldsymbol{Y}=\boxed{\boldsymbol{MY}}=\boldsymbol{M}\left(\boldsymbol{X\beta}+\boldsymbol{\varepsilon}\right)=\boldsymbol{0}\cdot\boldsymbol{\beta}+\boldsymbol{M\varepsilon}=\boxed{\boldsymbol{M\varepsilon}},\end{aligned}$$

且<u>拟合值向量 $\hat{\boldsymbol{Y}}$ 和残差向量 $\hat{\boldsymbol{\varepsilon}}$ 是正交的</u>, 这是因为

$$\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\left(\boldsymbol{PY}\right)^{\mathrm{T}}\left(\boldsymbol{MY}\right)=\boldsymbol{Y}^{\mathrm{T}}\left(\boldsymbol{PM}\right)\boldsymbol{Y}=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{O}_n\boldsymbol{Y}=0.$$

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 3.8.1 (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 *(3.6.1)* 含有常数项时,$h_{ii}\geqslant n^{-1}$;

(c) $\sum\limits_{i=1}^{n}h_{ii}=\mathrm{tr}\boldsymbol{P}=K.$

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

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{Y}}\middle|\boldsymbol{X}\right)&=\mathbb{E}\left(\boldsymbol{PY}|\boldsymbol{X}\right)=\boldsymbol{P}\mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{X}\right),\\ \mathrm{Var}\left(\hat{\boldsymbol{Y}}\middle|\boldsymbol{X}\right)&=\boldsymbol{P}\mathrm{Var}\left(\boldsymbol{Y}|\boldsymbol{X}\right)\boldsymbol{P}^{\mathrm{T}}\xlongequal{\text{异方差假设式 (3.9)}}\boldsymbol{PDP},\end{aligned}$$

在条件同方差式 (3.7) 成立时, 便可以得到

$$\mathrm{Var}\left(\hat{\boldsymbol{Y}}\middle|\boldsymbol{X}\right)=\boldsymbol{P}\left(\boldsymbol{I}\sigma^2\right)\boldsymbol{P}=\boldsymbol{P}\sigma^2,$$

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

$$
\begin{aligned}
&\mathbb{E}\left(\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{M}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\boldsymbol{M}\mathbb{E}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)\xlongequal{\text{严格外生性式 (3.5)}}\mathbf{0},\\
&\operatorname{Var}\left(\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)=\operatorname{Var}\left(\boldsymbol{M}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\boldsymbol{M}\operatorname{Var}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)\boldsymbol{M}^{\mathrm{T}}\xlongequal{\text{异方差假设式 (3.9)}}\boldsymbol{M}\boldsymbol{D}\boldsymbol{M},
\end{aligned}
$$

当条件同方差式 (3.7) 成立时, 则有

$$
\operatorname{Var}\left(\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)=\boldsymbol{M}\left(\boldsymbol{I}\sigma^2\right)\boldsymbol{M}=\boldsymbol{M}\sigma^2,
$$

此时残差 $\hat{\varepsilon}_i$ 的条件方差为 $\boxed{\operatorname{Var}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right)=\left(1-h_{ii}\right)\sigma^2}$.

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 $\bar{\varepsilon}_i=(1-h_{ii})^{-\frac{1}{2}}\hat{\varepsilon}_i$, 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

$$
\mathbb{E}\left(\bar{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\boldsymbol{M}^*\boldsymbol{M}\mathbb{E}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\mathbf{0}
$$

与

$$
\begin{aligned}
\operatorname{Var}\left(\bar{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)&=\operatorname{Var}\left(\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\boldsymbol{M}^*\boldsymbol{M}\operatorname{Var}\left(\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)\boldsymbol{M}\boldsymbol{M}^*\\
&\xlongequal{\text{异方差假设式 (3.9)}}\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*\xlongequal{\text{同方差假设式 (3.7)}}\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}^*\sigma^2,
\end{aligned}
$$

则

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

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.

3.9 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.

[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.

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 (3.6.1),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 (3.6.1) 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 $\mathbf{X}_i = (X_{i1}, \cdots, X_{ik}, \cdots, X_{iK})^{\mathrm{T}}$, 则将利用剔除了第 i 个样本得到的 OLS 估计量记为 $\hat{\boldsymbol{\beta}}_{(-i)}$, 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*

$$\tilde{\varepsilon}_i = Y_i - \tilde{Y}_i = Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(-i)}, \tag{3.28}$$

*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

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \mathbf{X}_iY_i\right). \tag{3.29}$$

式 (3.29) 可以化简为更简洁的形式, 并与 OLS 估计量联系起来. 将式 (3.29) 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

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbf{X}_i\tilde{\varepsilon}_i. \tag{3.30}$$

进一步地, 依据式 (3.30) 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

$$\tilde{\varepsilon}_i=\frac{1}{1-h_{ii}}\hat{\varepsilon}_i. \tag{3.31}$$

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 (3.30) 和式 (3.28), 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 (3.31), 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

$$\hat{\boldsymbol{Y}} = \boldsymbol{PY}, \quad \hat{\boldsymbol{\varepsilon}} = \boldsymbol{M\varepsilon}, \quad \hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} = 0, \quad \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\mathbf{1} = 0\,(\text{需要常数项})\,,$$

可得

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \left(\hat{\boldsymbol{Y}} + \hat{\boldsymbol{\varepsilon}}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}} + \hat{\boldsymbol{\varepsilon}}\right) = \hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}} + \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} \quad \Leftrightarrow \quad \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \tag{3.32}$$

和

$$\begin{aligned}
&\left(\boldsymbol{Y} - \bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \bar{Y}\mathbf{1}\right) \\
=&\left(\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1} + \hat{\boldsymbol{\varepsilon}}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1} + \hat{\boldsymbol{\varepsilon}}\right) \quad \Leftrightarrow \quad \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \\
=&\left(\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}\right) + \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}
\end{aligned} \tag{3.33}$$

成立. 式 (3.32) 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 (3.32) 在此之上还依赖于模型 (3.6.1) 含义常数项时残差和为 0 的性质. 利用式 (3.32) 和 (3.32), 我们依次定义**非中心化** R^2(uncentered R^2)为

$$R_{\mathrm{uc}}^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}} = 1 - \frac{\sum\limits_{i=1}^{n}\hat{Y}_i^2}{\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2}, \tag{3.34}$$

R^2, 亦即**拟合优度** (goodness of fit)或**可决系数** (coefficient of determination)为

$$R^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{\left(\boldsymbol{Y} - \bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \bar{Y}\mathbf{1}\right)} = 1 - \frac{\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum\limits_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}. \tag{3.35}$$

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 (3.6.1) 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 (3.32) 和 (3.32) 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 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^2_{uc} 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^2_{uc} 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^2_{uc} 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

$$\bar{R}^2 = 1 - \frac{(n-K)^{-1}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{(n-1)^{-1}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)} = 1 - \frac{\dfrac{1}{n-K}\sum\limits_{i=1}^{n}\hat{\varepsilon}_i^2}{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}, \tag{3.36}$$

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 (3.36) 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

$$\tilde{R}^2 = 1 - \frac{\tilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}}{\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\bar{Y}\mathbf{1}\right)} = 1 - \frac{\sum\limits_{i=1}^{n}\left(1-h_{ii}\right)^{-2}\hat{\varepsilon}_i^2}{\sum\limits_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}, \tag{3.37}$$

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

$$\begin{aligned}
\bar{R}^2 =& 1 - \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n-K)}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)/(n-1)} = 1 - \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)} \cdot \frac{n-1}{n-K} \\
=& \frac{n-1}{n-K}\left(\frac{n-K}{n-1} - \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right) \\
=& \frac{n-1}{n-K}\left(\frac{n-K}{n-1} - 1 + 1 - \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right) \\
=& \frac{n-1}{n-K}\left(\frac{n-K}{n-1} - 1 + R^2\right) = \frac{n-1}{n-K}\left(\frac{1-K}{n-1} + R^2\right),
\end{aligned}$$

故化简有

$$1-R^2=\frac{n-K}{n-1}\left(1-\bar{R}^2\right). \tag{3.38}$$

对于 R^2 与非中心化 R^2, 则有

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

即有

$$1-R^2=\left[1+\frac{n\bar{Y}^2}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right]\left(1-R_{\mathrm{uc}}^2\right). \tag{3.39}$$

3.11 多元 OLS 估计量下的方差估计量

本节我们将研究模型 (3.6.1) 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

$$\hat{\sigma}_{\mathrm{ideal}}^2=\frac{1}{n}\mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left(\varepsilon_i^2\middle|\boldsymbol{X}\right)\xlongequal{\text{条件同方差式 (3.7)}}\frac{1}{n}\sum_{i=1}^n\sigma^2=\sigma^2,$$

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 $\dfrac{1}{n}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}$ 是有偏的. 为此, 计算条件期望

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)&=\mathbb{E}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\mathrm{tr}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}\right)\middle|\boldsymbol{X}\right)=\mathbb{E}\left(\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right)\middle|\boldsymbol{X}\right)\\
&\xlongequal{\text{迹是线性变换}}\mathrm{tr}\left(\mathbb{E}\left(\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\right)=\mathrm{tr}\left(\boldsymbol{M}\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right)\right)\xlongequal{\text{异方差式 (3.9)}}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{D}\right),\end{aligned}$$

则在同方差假设式 (3.7) 有

$$\mathbb{E}\left(\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right)=\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{I}\sigma^2\right)=(n-K)\,\sigma^2,$$

同样地, 我们得到 σ^2 的**偏差校正估计量** (bias-corrected estimator)为

$$s^2 = \frac{1}{n-K}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} = \frac{1}{n-K}\sum_{i=1}^{n}\varepsilon_i^2. \tag{3.40}$$

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

$$\bar{\sigma}^2 = \frac{1}{n}\bar{\boldsymbol{\varepsilon}}^{\mathrm{T}}\bar{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\bar{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varepsilon}_i^2}{\sqrt{1-h_{ii}}}, \tag{3.41}$$

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 (3.41) 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 (3.6.1), 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

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0 = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^2, \tag{3.42}$$

对协方差矩阵 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0$ 计算条件期望得

$$\mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0\middle|\boldsymbol{X}\right) = \mathbb{E}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}s^2\middle|\boldsymbol{X}\right) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\mathbb{E}\left(s^2\middle|\boldsymbol{X}\right) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2 = \boldsymbol{V}_{\hat{\boldsymbol{\beta}}},$$

则可以进一步推知在**同方差**的条件下 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0$ 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.$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0$ 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 (3.42) 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 (3.6.1) 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 $\boldsymbol{D}_{\varepsilon}=\operatorname{diag}\left(\varepsilon_1^2,\cdots,\varepsilon_n^2\right)$, 则有

$$\mathbb{E}\left(\boldsymbol{D}_{\varepsilon}\middle|\,\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\,\boldsymbol{X}\right)=\operatorname{diag}\left(\sigma_1^2,\cdots,\sigma_n^2\right)=\boldsymbol{D},$$

故 $\boldsymbol{D}_{\varepsilon}$ 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 $\boldsymbol{D}_{\varepsilon}$ 去估计 $\boldsymbol{D}$. 记

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{ideal}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}_{\varepsilon}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\varepsilon_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}, \tag{3.43}$$

则

$$\begin{aligned}&\mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{ideal}}\middle|\,\boldsymbol{X}\right)=\mathbb{E}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}_{\varepsilon}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\middle|\,\boldsymbol{X}\right)\\ =&\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{D}_{\varepsilon}\middle|\,\boldsymbol{X}\right)\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\boldsymbol{V}_{\hat{\boldsymbol{\beta}}},\end{aligned}$$

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

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}, \tag{3.44}$$

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 (3.44) 会是有偏的, 仿照校正直接替换 $\dfrac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$ 的方法, 则 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}$ 的校正估计量为

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC1}}=\frac{n}{n-K}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}=\frac{n}{n-K}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}, \tag{3.45}$$

式 (3.45) 比式 (3.44) 更为推荐, 其予以了 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$ 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 $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}}$ 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

$$
\begin{aligned}
\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\bar{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}(1-h_{ii})^{-1}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},
\end{aligned}
\tag{3.46}
$$

$$
\begin{aligned}
\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\tilde{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}(1-h_{ii})^{-2}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},
\end{aligned}
\tag{3.47}
$$

我们将式 (3.44)、(3.45)、(3.46) 和 (3.47) 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

$$
\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} < \hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} < \hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}, \tag{3.48}
$$

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

$$
\operatorname{Var}\left(\hat{\boldsymbol{\varepsilon}}\middle|\boldsymbol{X}\right) \xlongequal{\text{异方差}} \boldsymbol{MDM} \Rightarrow \mathbb{E}\left(\hat{\varepsilon}_i^2\middle|\boldsymbol{X}\right) = \operatorname{Var}\left(\hat{\varepsilon}_i\middle|\boldsymbol{X}\right) = \left(\boldsymbol{MDM}\right)_{ii} \xlongequal{\text{同方差}} \left(1-h_{ii}\right)\sigma^2,
$$

以及

$$\mathbb{E}\left(\bar{\varepsilon}_i^2 \middle| \boldsymbol{X}\right) = \mathrm{Var}\left(\bar{\varepsilon}_i \middle| \boldsymbol{X}\right) = \mathrm{Var}\left(\frac{\hat{\varepsilon}_i}{\sqrt{1-h_{ii}}} \middle| \boldsymbol{X}\right) = \frac{\mathrm{Var}\left(\hat{\varepsilon}_i \middle| \boldsymbol{X}\right)}{1-h_{ii}} \xlongequal{\text{同方差}} \sigma^2,$$
$$\mathbb{E}\left(\tilde{\varepsilon}_i^2 \middle| \boldsymbol{X}\right) = \mathrm{Var}\left(\tilde{\varepsilon}_i \middle| \boldsymbol{X}\right) = \mathrm{Var}\left(\frac{\hat{\varepsilon}_i}{1-h_{ii}} \middle| \boldsymbol{X}\right) = \frac{\mathrm{Var}\left(\hat{\varepsilon}_i \middle| \boldsymbol{X}\right)}{(1-h_{ii})^2} \xlongequal{\text{同方差}} \frac{\sigma^2}{1-h_{ii}},$$

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

$$\mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} \middle| \boldsymbol{X}\right) < \boldsymbol{V}_{\hat{\boldsymbol{\beta}}} < \mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}} \middle| \boldsymbol{X}\right), \quad \boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \mathbb{E}\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} \middle| \boldsymbol{X}\right), \tag{3.49}$$

这里的 $<$ 依然是针对二次型的正定性而言的. 式 (3.49) 与式 (3.48) 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.

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 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$, 我们总是关系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, 标准差)**.** *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*

$$\mathrm{SE}\left(\hat{\beta}_k\right) = \sqrt{\left(\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\right)_{kk}} \tag{3.50}$$

*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$\mathrm{SE}\left(\hat{\beta}_k\right)$ 实际上是 OLS 估计量 $\hat{\beta}_k$ 的标准差的**估计量**. 当模型 (3.6.1) 满足有同方差假设式 (3.7) 时, 则依据式 (3.42) 则式 (3.50) 等于

$$\mathrm{SE}\left(\hat{\beta}_k\right) = s\sqrt{\left((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right)_{kk}}. \tag{3.51}$$

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

$$\mathbb{E}\left(\varepsilon_i | X = 0\right) = 25^2,\ \mathbb{E}\left(\varepsilon_i | X = 5\right) = 30^2,\ \mathbb{E}\left(\varepsilon_i | X = 15\right) = 15^2,\ \mathbb{E}\left(\varepsilon_i | X = 20\right) = 5^2,$$

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 $\boldsymbol{V}_{\hat{\beta}}$ 所计算的 OLS 估计量的标准差相比, 同方差假设下 $\hat{\boldsymbol{V}}_{\hat{\beta}}^{0}$ 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 (3.48) 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	SE($\hat{\beta}_0$) 与真实值的误差	$\hat{\beta}_1$ 的标准误	SE($\hat{\beta}_1$) 与真实值的误差
真实值	4.2300	0.00%	0.3336	0.00%
$\hat{\boldsymbol{V}}_{\hat{\beta}}^{0}$	3.7425	11.53%	0.4321	-29.54%
$\hat{\boldsymbol{V}}_{\hat{\beta}}^{\mathrm{HC0}}$	4.3036	-1.74%	0.3394	-1.74%
$\hat{\boldsymbol{V}}_{\hat{\beta}}^{\mathrm{HC1}}$	4.3522	-2.89%	0.3433	-2.89%
$\hat{\boldsymbol{V}}_{\hat{\beta}}^{\mathrm{HC2}}$	4.3526	-2.90%	0.3439	-3.10%
$\hat{\boldsymbol{V}}_{\hat{\beta}}^{\mathrm{HC3}}$	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 (3.6.1), 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 (3.6.1) 的假设之上, 补充假设*

$$\text{总体}: \varepsilon|\,X\sim\mathrm{N}\left(0,\sigma^2\right), \tag{3.52}$$

$$\text{样本}: \boldsymbol{\varepsilon}|\,\boldsymbol{X}\sim\mathrm{N}\left(\boldsymbol{0},\boldsymbol{I}_n\sigma^2\right), \tag{3.53}$$

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)*.

$$\begin{aligned}
Y &= \mathscr{P}(Y|\mathbf{X}) + \varepsilon \\
\mathscr{P}(Y|\mathbf{X}) &= \mathbf{X}^T\boldsymbol{\beta}, \quad \boldsymbol{\beta} = \left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^T\right)\right)^{-1}\mathbb{E}(\mathbf{X}Y) \\
\mathbb{E}(\mathbf{X}\varepsilon) &= 0 \\
\mathbb{E}(\varepsilon) &= 0 \quad (\text{不需要常数项}) \\
\operatorname{Var}(\varepsilon) &= \mathrm{E}\varepsilon^2 = \mathrm{E}(\varepsilon|\mathbf{X}) = \sigma^2
\end{aligned}$$

与在上一章的分析一样, 当总体满足正态假设 (3.52) 时, 则 $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$ 满足有条件分布 $Y|\,\mathbf{X} \sim \mathrm{N}\left(\mathbf{X}^T\boldsymbol{\beta}, \sigma^2\right)$. 相应地, 正态假设 (3.53) 使得样本 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 有条件分布 $\boldsymbol{Y}|\,\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n\sigma^2)$ 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 (3.52) 和 (3.53) 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

$$\hat{\beta}_k\Big|\,\boldsymbol{X} \sim \mathrm{N}\left(\beta_k, \left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2\right)_{kk}\right). \tag{3.54}$$

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

$$\hat{\boldsymbol{Y}}\Big|\,\boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{P}\sigma^2\right), \quad \hat{\boldsymbol{\varepsilon}}|\,\boldsymbol{X} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{M}\sigma^2\right). \tag{3.55}$$

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

$$\begin{aligned}
\mathbb{E}\left(\begin{pmatrix} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \\ \boldsymbol{M} \end{pmatrix}\boldsymbol{\varepsilon}\Bigg|\,\boldsymbol{X}\right) &= \begin{pmatrix} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \\ \boldsymbol{M} \end{pmatrix}\mathbb{E}(\boldsymbol{\varepsilon}|\,\boldsymbol{X}) \\
&= \begin{pmatrix} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \\ \boldsymbol{M} \end{pmatrix}\cdot\mathbf{0} = \mathbf{0}_{(K+n)\times 1}
\end{aligned}$$

及

$$
\begin{aligned}
&\operatorname{Var}\left(\left.\begin{pmatrix} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \\ \boldsymbol{M} \end{pmatrix}\boldsymbol{\varepsilon}\right|\boldsymbol{X}\right) \\
=&\begin{pmatrix} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \\ \boldsymbol{M} \end{pmatrix}\operatorname{Var}\left(\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)\left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\boldsymbol{M}\right) \\
=&\begin{pmatrix} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} \\ \boldsymbol{M} \end{pmatrix}\boldsymbol{I}_n\sigma^2\left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\boldsymbol{M}\right)=\begin{pmatrix} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2 & \boldsymbol{O}_{K\times n} \\ \boldsymbol{O}_{n\times K} & \boldsymbol{M}\sigma^2 \end{pmatrix},
\end{aligned}
$$

故有

$$
\left.\begin{pmatrix} \hat{\boldsymbol{\beta}}-\boldsymbol{\beta} \\ \hat{\boldsymbol{\varepsilon}} \end{pmatrix}\right|\boldsymbol{X}\sim\mathrm{N}\left(\boldsymbol{0}_{(K+n)\times 1},\begin{pmatrix} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2 & \boldsymbol{O}_{K\times n} \\ \boldsymbol{O}_{n\times K} & \boldsymbol{M}\sigma^2 \end{pmatrix}\right). \tag{3.56}
$$

式 (3.56) 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 (3.56) 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 $\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=(\boldsymbol{M}\boldsymbol{\varepsilon})^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\varepsilon}$, 而由于投影矩阵 $\boldsymbol{M}$ 是幂等的实对称方阵, 则依据式 (3.27) 得到

$$
\boldsymbol{M}=\boldsymbol{H}^{-1}\begin{pmatrix} \boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n} \\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{I}_K \end{pmatrix}\boldsymbol{H},
$$

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

$$
\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}=\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{H}^{-1}\begin{pmatrix} \boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n} \\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{I}_K \end{pmatrix}\boldsymbol{H}\boldsymbol{\varepsilon}=(\boldsymbol{H}\boldsymbol{\varepsilon})^{\mathrm{T}}\begin{pmatrix} \boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n} \\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{O}_K \end{pmatrix}\boldsymbol{H}\boldsymbol{\varepsilon}
$$

成立. 令 $\boldsymbol{\mu}=\boldsymbol{H}\boldsymbol{\varepsilon}=(\mu_1,\cdots,\mu_n)^{\mathrm{T}}$, 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

$$
\begin{aligned}
&\mathbb{E}\left(\boldsymbol{\mu}|\,\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{H}\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)=\boldsymbol{H}\mathbb{E}\left(\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)=\boldsymbol{H}\boldsymbol{0}=\boldsymbol{0}, \\
&\operatorname{Var}\left(\boldsymbol{\mu}|\,\boldsymbol{X}\right)=\operatorname{Var}\left(\boldsymbol{H}\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)=\boldsymbol{H}\operatorname{Var}\left(\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right)\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{H}\left(\boldsymbol{I}_n\sigma^2\right)\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{I}_n\sigma^2,
\end{aligned}
$$

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

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} =&\boldsymbol{\mu}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n}\\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{O}_K\end{pmatrix}\boldsymbol{\mu}\\ &=\left(\boldsymbol{\mu}_1^{\mathrm{T}},\boldsymbol{\mu}_2^{\mathrm{T}}\right)\begin{pmatrix}\boldsymbol{I}_{n-K} & \boldsymbol{O}_{(n-K)\times n}\\ \boldsymbol{O}_{K\times(n-K)} & \boldsymbol{O}_K\end{pmatrix}\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}\\ &=\left(\boldsymbol{\mu}_1^{\mathrm{T}},\mathbf{0}_{1\times K}\right)\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}=\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1,\end{aligned}$$

故可以推知方差估计量 s^2 满足有

$$\left.\frac{(n-K)\,s^2}{\sigma^2}\right|\boldsymbol{X}=\left.\frac{\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1}{\sigma^2}\right|\boldsymbol{X}=\left.\sum_{i=1}^{n-K}\left(\frac{\mu_i-0}{\sigma}\right)^2\right|\boldsymbol{X}\sim\chi^2(n-K)\,. \tag{3.57}$$

由于正态假设 (3.52) 和 (3.53) 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

$$\left.\frac{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}}{\sqrt{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\sigma^2}}\right|\boldsymbol{X}\sim\mathrm{N}\left(\mathbf{0},\boldsymbol{I}\right),$$

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

$$T=\frac{\hat{\beta}_k-\beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)}=\frac{\hat{\beta}_k-\beta_k}{\sqrt{s^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}} \tag{3.58}$$

为 ***t* 统计量** *(t-statistic)*.

显然,$\hat{\beta}_k$ 的 t 统计量 T 有

$$T=\frac{\beta_k-\hat{\beta}_k}{\sqrt{\sigma^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}\Bigg/\frac{\sqrt{s^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}{\sqrt{\sigma^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}=\frac{\beta_k-\hat{\beta}_k}{\sqrt{\sigma^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}\Bigg/\sqrt{\frac{(n-K)\,s^2/\sigma^2}{n-K}},$$

而

$$\left.\frac{\beta_k-\hat{\beta}_k}{\sqrt{\sigma^2\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right)_{kk}}}\right|\boldsymbol{X}\sim\mathrm{N}\left(0,1\right),\qquad\left.\frac{(n-K)\,s^2}{\sigma^2}\right|\boldsymbol{X}\sim\chi^2(n-K)$$

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 (3.52) 和 (3.53) 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*

$$\hat{I}=\left[\hat{\beta}_k-c\cdot\mathrm{SE}\left(\hat{\beta}_k\right),\hat{\beta}_k+c\cdot\mathrm{SE}\left(\hat{\beta}_k\right)\right],\tag{3.59}$$

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

$$\hat{I}=\left[\frac{(n-2)\,s^2}{c_2},\frac{(n-2)\,s^2}{c_1}\right],\tag{3.60}$$

其中 $c_1=F^{-1}(\alpha/2)\,,c_2=F^{-1}(1-\alpha/2)$,$F$ 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 (3.59), 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

$$\hat{I}=\left[\hat{\beta}_k-2\cdot\mathrm{SE}\left(\hat{\beta}_k\right),\hat{\beta}_k+2\cdot\mathrm{SE}\left(\hat{\beta}_k\right)\right].\tag{3.61}$$

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

$$\mathrm{H}_0:\theta\in B_0\tag{3.62}$$

为**原假设** (null hypothesis), 相应地, 我们称原假设不成立的对应情况

$$\mathrm{H}_1 : \theta \notin B_0 \tag{3.63}$$

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*

$$\beta_\phi(\theta) = \mathbb{P}\left(\text{检验}\phi\text{下}\mathrm{H}_0\text{被拒绝}\right) \tag{3.64}$$

*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<u>控制犯第一类错误的概率</u>$\alpha_1(\theta|\phi)$ 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<u>检验 ϕ 下接受 (不拒绝) 原假设的概率较大</u>. 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 (3.6.3) 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

$$T = \frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)} \xlongequal{\mathrm{H}_0} \frac{\beta_k - c}{\mathrm{SE}\left(\hat{\beta}_k\right)}$$

作为检验 ϕ 的检验统计量, 且检验 ϕ 为

$$\phi: 若\ |T| > C 则拒绝\ \mathrm{H}_0, 其中 C > 0.$$

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

$$\begin{aligned}\beta_\phi(\beta_k) =& \mathbb{P}\left(检验\phi下\ \mathrm{H}_0 被拒绝\right) = \mathbb{P}(|T| > C) \\ =& \mathbb{P}(T < -C) + \mathbb{P}(T > C) = F(-C) + 1 - F(C) \\ =& 1 - F(C) + 1 - F(C) = 2(1 - F(C)),\end{aligned}$$

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

$$T = \frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)} \xlongequal{\mathrm{H}_0} \frac{\beta_k - c}{\mathrm{SE}\left(\hat{\beta}_k\right)}$$

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

$$p = 1 - G(S) = 1 - \mathbb{P}(s \leqslant S_{实现值} | \mathrm{H}_0) = \mathbb{P}(s \geqslant S_{实现值} | \mathrm{H}_0)$$

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.

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

$$\begin{aligned}
p =& \mathbb{P}(s \geqslant S_{实现值} | \mathrm{H}_0) = \mathbb{P}(|t| \geqslant |T|) = \mathbb{P}(t \geqslant |T|) + \mathbb{P}(t \leqslant -|T|) \\
=& 1 - F(|T|) + F(-|T|) = 1 - F(|T|) + 1 - F(|T|) \\
=& 2(1 - F(|T|)),
\end{aligned}$$

其中 F 是自由度服从 $n - K$ 的 t 分布的累计密度函数.

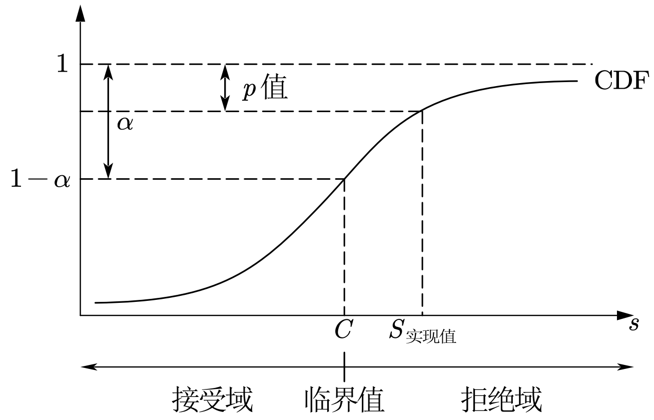

图 3.4: 检验 ϕ 的 p 值

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

$$\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{r}\Leftrightarrow\begin{pmatrix}R_{11}&R_{12}&\cdots&R_{1K}\\R_{21}&R_{22}&\cdots&R_{2K}\\\vdots&\vdots&&\vdots\\R_{r1}&R_{r2}&\cdots&R_{rK}\end{pmatrix}\begin{pmatrix}\beta_1\\\beta_2\\\vdots\\\beta_K\end{pmatrix}=\begin{pmatrix}r_1\\r_2\\\vdots\\r_r\end{pmatrix}$$

$$\Leftrightarrow\begin{cases}R_{11}\beta_1+R_{12}\beta_2+\cdots+R_{1K}\beta_K=r_1,\\R_{21}\beta_1+R_{22}\beta_2+\cdots+R_{2K}\beta_K=r_2,\\\cdots\\R_{r1}\beta_1+R_{m2}\beta_2+\cdots+R_{rK}\beta_K=r_r,\end{cases}$$

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

$$\mathrm{H}_0:\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{r}=\mathbf{0}\quad v.s.\quad\mathrm{H}_0:\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{r}\neq\mathbf{0},$$

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 $\boldsymbol{u}=\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\in\mathbb{R}^{r\times1}$, 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 $\mathrm{H}_0:\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{r}=\mathbf{0}$ 成立的情况下, 计算 $\boldsymbol{u}$ 的条件期望、方差为

$$\begin{aligned}&\mathbb{E}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)=\mathbb{E}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\middle|\,\boldsymbol{X}\right)=\boldsymbol{R}\mathbb{E}\left(\hat{\boldsymbol{\beta}}\middle|\,\boldsymbol{X}\right)-\boldsymbol{r}=\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{r}\xlongequal{\mathrm{H}_0}\mathbf{0},\\&\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)=\mathrm{Var}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\middle|\,\boldsymbol{X}\right)=\boldsymbol{R}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\middle|\,\boldsymbol{X}\right)\boldsymbol{R}^{\mathrm{T}}=\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2,\end{aligned}$$

故 r 维列向量 $\boldsymbol{u}|\,\boldsymbol{X}\sim\mathrm{N}\left(\mathbf{0},\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)$.

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 $\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)=\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2$ 是正定的. 依据定理 (3.7.4), 则可将协方差矩阵写作 $\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)=\boldsymbol{B}^2$, 其中 r 阶方阵 $\boldsymbol{B}$ 为 $\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)$ 的平方根. 不难注意到

$$\begin{aligned}\boldsymbol{u}^{\mathrm{T}}\left(\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)\right)^{-1}\boldsymbol{u}=&\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{B}^2\right)^{-1}\boldsymbol{u}=\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{B}\boldsymbol{B}\right)^{-1}\boldsymbol{u}\\=&\boldsymbol{u}^{\mathrm{T}}\boldsymbol{B}^{-\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{u}=\left(\boldsymbol{B}^{-1}\boldsymbol{u}\right)^{\mathrm{T}}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\right),\end{aligned}$$

在原假设下, 对于 $\boldsymbol{B}^{-1}\boldsymbol{u}$ 有

$$\begin{aligned}&\mathbb{E}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\middle|\,\boldsymbol{X}\right)=\boldsymbol{B}^{-1}\mathbb{E}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)=\mathbf{0},\\&\mathrm{Var}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\middle|\,\boldsymbol{X}\right)=\mathrm{Var}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\middle|\,\boldsymbol{X}\right)=\boldsymbol{B}^{-1}\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)\boldsymbol{B}^{-1}=\boldsymbol{B}^{-1}\left(\boldsymbol{B}^2\right)\boldsymbol{B}^{-1}=\boldsymbol{I}_r,\end{aligned}$$

故有 $\boldsymbol{B}^{-1}\boldsymbol{u}|\,\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_r)$ 成立. 我们以 r 维列向量 $\boldsymbol{B}^{-1}\boldsymbol{u}$ 的模长 $\|\boldsymbol{B}^{-1}\boldsymbol{u}\|$ 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 $\boldsymbol{B}^{-1}\boldsymbol{u} = (u_1, \cdots, u_r)^{\mathrm{T}}$ 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

$$\boldsymbol{u}^{\mathrm{T}}\left(\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)\right)^{-1}\boldsymbol{u} = \left(\boldsymbol{B}^{-1}\boldsymbol{u}\right)^{\mathrm{T}}\left(\boldsymbol{B}^{-1}\boldsymbol{u}\right) = \sum_{i=1}^{r} u_i^2,$$

故有 $\boldsymbol{u}^{\mathrm{T}}\left(\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)\right)^{-1}\boldsymbol{u}\big|\,\boldsymbol{X} \sim \chi^2(r)$, 也即有

$$\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)\Big|\,\boldsymbol{X} \sim \chi^2(r)\,. \tag{3.65}$$

对于形如式 (3.65) 的具有二次型的统计量

$$W = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^{\mathrm{T}}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right), \tag{3.66}$$

我们将式 (3.66) 称为 **Wald 统计量** (Wald statistic). 式 (3.65) 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 $\mathrm{H}_0 : \boldsymbol{R\beta} - \boldsymbol{r} = \boldsymbol{0}$ 的检验问题, 需要指出式 (3.65) 较式 (3.66) 仍是有区别的, 因为式 (3.65) 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

$$\begin{aligned} W &= \left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}s^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right) \\ &= \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)}{s^2} \\ &= \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)}{s^2/\sigma^2} \\ &= \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)}{\dfrac{(n-K)\,s^2}{\sigma^2}\Big/(n-K)}, \end{aligned}$$

记

$$W^0 = \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)^{\mathrm{T}}\left(\boldsymbol{R}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{R}^{\mathrm{T}}\sigma^2\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right)}{\dfrac{(n-K)\,s^2}{\sigma^2}\Big/(n-K)}, \tag{3.67}$$

我们称式 (3.67) 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 (3.65) 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 (3.57), 不难得知, 令 $F=W^0/r$, 则有

$$F=\frac{W^0}{r}=\frac{\chi^2(r)/r}{\chi^2(r)/(n-K)}\sim F(r,n-K), \tag{3.68}$$

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 $\mathrm{H}_0:\boldsymbol{R\beta}-\boldsymbol{r}=\mathbf{0}$ 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

$$\phi:\text{若}F\quad>F^{-1}(1-\alpha)\,\text{则拒绝 }\mathrm{H}_0,$$

其中第一个 F 是检验统计量式 (3.68) 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 (3.67) 得到

$$F=\frac{W^0}{r}$$

作为检验统计量, 则有显著性水平为 α的检验

$$\phi:\text{若}F\quad>F^{-1}(1-\alpha)\,\text{则拒绝 }H_0,$$

其中第一个 F 是检验统计量式 (3.68) 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

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathrm{CLS}} = & \underset{\boldsymbol{\beta}\in\mathbb{R}^{K\times 1}}{\arg\min}\, \hat{S}(\boldsymbol{\beta}), \\
& s.t.\ \boldsymbol{R\beta}=\boldsymbol{r},
\end{aligned}
$$

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 () 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

$$
F=\frac{\left(\left(\boldsymbol{\varepsilon}^{*}\right)^{\mathrm{T}}\boldsymbol{\varepsilon}^{*}-\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)/r}{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n-K)}=\frac{(\mathrm{RSS}^{*}-\mathrm{RSS})/r}{\mathrm{RSS}/(n-K)}, \tag{3.69}
$$

其中 * 表示 CLS 的回归结果, 即 $\boldsymbol{\varepsilon}^{*}$ 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 $\boldsymbol{\varepsilon}$ 和 RSS 则是 OLS 估计量的残差向量和残差平方和. 可以证明, 式 (3.69) 与式 (3.68) 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 (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 (3.69) 有

$$
\begin{aligned}
F= & \frac{\left(\left(\boldsymbol{\varepsilon}^{*}\right)^{\mathrm{T}}\boldsymbol{\varepsilon}^{*}-\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\right)/r}{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n-K)}=\frac{\left(\frac{(\boldsymbol{\varepsilon}^{*})^{\mathrm{T}}\boldsymbol{\varepsilon}^{*}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}-\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right)(K-1)^{-1}}{\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\cdot(n-K)^{-1}} \\
= & \frac{\left[1-\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}-\left(1-\frac{(\boldsymbol{\varepsilon}^{*})^{\mathrm{T}}\boldsymbol{\varepsilon}^{*}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right)\right](K-1)^{-1}}{\left[1-\left(1-\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\mathbf{1}\bar{Y}\right)}\right)\right]\cdot(n-K)^{-1}} \\
= & \frac{\left(R^2-(R^{*})^2\right)(K-1)^{-1}}{(1-R^2)\cdot(n-K)^{-1}}
\end{aligned}
$$

其中 $(R^*)^2$ 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 (3.35) 便可推知 $(R^*)^2 = 0$. 故在 $\mathrm{H}_0 : \boldsymbol{\beta}_{K\times 1} = (1, 0, \cdots, 0)^{\mathrm{T}}$ 下,F 比率即为

$$F = \frac{R^2 / (K-1)}{(1-R^2) / (n-K)}, \tag{3.70}$$

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 (3.70), 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 (3.38) 对式 (3.70) 代换得

$$\begin{aligned} F =& \frac{R^2 / (K-1)}{(1-R^2) / (n-K)} = \left(\frac{1}{1-R^2} - 1\right) \cdot \frac{n-K}{K-1} \\ =& \left(\frac{n-1}{n-K}\left(1-\bar{R}^2\right)^{-1} - 1\right) \cdot \frac{n-K}{K-1} = \frac{n-1}{K-1}\left(1-\bar{R}^2\right)^{-1} - \frac{n-K}{K-1}, \end{aligned}$$

也即有

$$\bar{R}^2 = 1 - \frac{n-1}{(K-1)F + (n-K)}. \tag{3.71}$$

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 (3.71) 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 (3.71) 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 (3.70) 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.

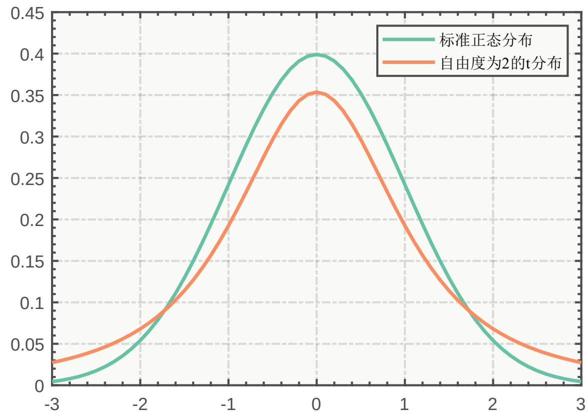

图 3.5: 正态分布与 t 分布的概率密度函数

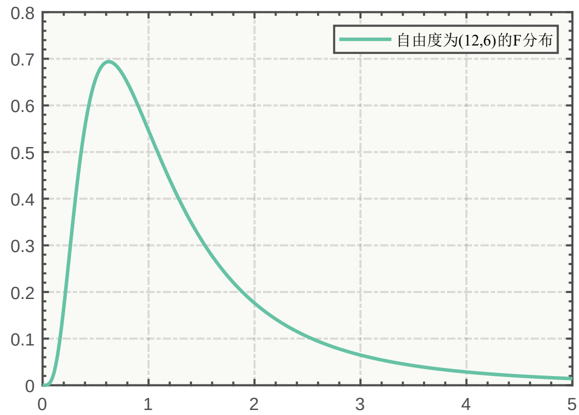

图 3.6: F 分布的概率密度函数

3.14 多元线性回归模型的预测

3.15 虚拟变量的 OLS 估计量

3.16 代码

```
%% 正交投影与OLS估计量
clear
clc
Y=[-0.5;1.25;2];
X=[-1 1
    1  1
    0  1 ];
Ze=[0,0];%设定箭头位置
% a=0:1;%线性变换系数
a=[0 0 1 1;
   0 1 1 0];%线性变换矩阵
Span=X*a;%张成的空间
Yhat=X*(X\Y);%拟合值_投影
e=Y-Yhat;%残差
selfGrootDefault(2)%绘图美化函数,没有请注释
figure(1)
quiver3(0,0,0,Y(1),Y(2),Y(3),"off","LineWidth",2.5)%绘制Y
hold on
quiver3(Ze,Ze,Ze,X(1,:),X(2,:),X(3,:),"off","LineWidth",2.5)%绘制X
fill3(Span(1,:),Span(2,:),Span(3,:),[0 0.4470 0.7410], ...
```

```
    "LineStyle","none")%X张成的平面
quiver3(0,0,0,Yhat(1,:),Yhat(2,:),Yhat(3,:), ...
    "off","LineWidth",2.5)%绘制投影
quiver3(Yhat(1,:),Yhat(2,:),Yhat(3,:),e(1,:),e(2,:),e(3,:), ...
    "off",'-.',"LineWidth",2.5)%绘制残差
% quiver3(-2,0,0,1,0,0,4,'k','filled','LineWidth',2);%绘制x轴
% quiver3(0,0,0,0,1,0,2.5,'k','filled','LineWidth',2);%绘制y轴
% quiver3(0,0,0,0,0,1,2.5,'k','filled','LineWidth',2);%绘制z轴
% xlim([-1.5,1.5])
% ylim([0,2.5])
% zlim([0 3])
xlabel('$x$','FontSize',14,'Interpreter','latex')
ylabel('$y$','FontSize',14,'Interpreter','latex')
zlabel('$z$','FontSize',14,'Interpreter','latex')
legend('Y','X','X张成的平面','投影向量','残差向量','Location','north')

figure(2)%绘制拟合直线和散点图
hold on
xline(0,'-.','LineWidth',2)
yline(0,'-.','LineWidth',2)
scatter(X(:,1),Y,"filled","LineWidth",2)
y=[-5:1:5;ones(size(-5:1:5))]'*(X\Y);
plot(-5:1:5,y,'LineWidth',2)
xlim([-2.5 2.5])
ylim([-1 2.5])
xlabel('$x$','FontSize',14,'Interpreter','latex')
ylabel('$y$','FontSize',14,'Interpreter','latex')
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*

$$\lim_{n\to\infty}\mathbb{P}(|X_n-a|>\delta)=0 \quad \Leftrightarrow \quad \forall\varepsilon,\delta>0, \exists N\in\mathbb{N}, \forall n>N \ s.t. \ \mathbb{P}(|X_n-a|>\delta)<\varepsilon,$$

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 $X_n=b_nX$, 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 $\mathbb{P}(|X_n|\leqslant\delta)=\mathbb{P}(-\delta\leqslant b_nX\leqslant\delta)$. 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 $-\delta\leqslant b_nX\leqslant\delta$ 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<u>非随机的向量 $\boldsymbol{a}$</u>, 同时记

$$\mathbf{X}_n \xrightarrow{\mathrm{P}} \boldsymbol{a}\,(n \to \infty) \quad 或 \quad \operatorname*{plim}_{n\to\infty} \mathbf{X}_n = \boldsymbol{a}.$$

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<u>随机向量 $\mathbf{x}$</u>.

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<u>概率的极限</u>来定义的, 而另一种思路则是考虑<u>极限的概率</u>, 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<u>非随机的向量 $\boldsymbol{a}$</u>, 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<u>随机向量 $\mathbf{x}$</u>.
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*

$$\mathbb{P}(X\geqslant\varepsilon)\leqslant\frac{\mathbb{E}X}{\varepsilon}, \tag{4.1}$$

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

$$0 \leqslant \mathbb{P}\left(\left|X_n - \mathbb{E}X_n\right| \geqslant \varepsilon\right) \leqslant \frac{\sigma_n^2}{\varepsilon^2},$$

由于 σ_n^2 收敛于 0, 故依据夹逼定理有 $\lim\limits_{n\to\infty} \mathbb{P}\left(\left|X_n - \mathbb{E}X_n\right| \geqslant \varepsilon\right) = 0$, 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 $\mathbb{E}\left|X\right| < \infty$, 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 $\mathrm{Var}\left(X\right)$, 则依据 Chebyshev 不等式 (4.2) 可得

$$0 \leqslant \mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}X\right| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}\left(\bar{X}_n\right)}{\varepsilon^2} = \frac{\mathrm{Var}\left(X\right)}{n\varepsilon^2},$$

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 $\mathrm{Var}\left(X\right)$, 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

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

协方差矩阵 $\operatorname{Var}(\mathbf{X})$ 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* $\mathbb{E}\left\|\boldsymbol{g}\left(\mathbf{X}\right)\right\| < \infty$*, 则依据式 (4.7) 可知*

$$\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}\left(\mathbf{X}_i\right) \xrightarrow{P} \boldsymbol{\theta} = \mathbb{E}\boldsymbol{g}\left(\mathbf{X}\right), n \to \infty.$$

定义参数 $\boldsymbol{\beta} = \boldsymbol{h}\left(\boldsymbol{\theta}\right)$*, 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* $\hat{\boldsymbol{\beta}} = \boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{P} \boldsymbol{\beta} = \boldsymbol{h}\left(\boldsymbol{\theta}\right)$*.*

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* $M\left(t\right) = \mathbb{E}e^{tX}$ *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

$$M\left(t\right) = \int_{-\infty}^{+\infty} e^{tx} f\left(x\right) \mathrm{d}x, \tag{4.9}$$

相应地, 如果随机变量 X 是离散型随机变量, 则有

$$M\left(t\right) = \sum_{i=1}^{\infty} e^{tx_i}\mathbb{P}\left(X = x_i\right). \tag{4.10}$$

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 $M\left(t\right) = \mathbb{E}e^{tX}$ 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 $C\left(t\right) = \mathbb{E}e^{i\cdot tX}$, 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

$$
\begin{aligned}
-\frac{(x-\mu)^2}{2\sigma^2}+tx &= -\frac{1}{2\sigma^2}\left[x-\left(\mu+\sigma^2 t\right)\right]^2+\frac{1}{2\sigma^2}\left[\left(\mu+\sigma^2 t\right)^2-\mu^2\right] \\
&= -\frac{1}{2\sigma^2}\left[\left(x-\left(\mu+\sigma^2 t\right)\right)^2+\mu^2-\left(\mu+\sigma^2 t\right)^2\right],
\end{aligned}
$$

令 $u=x-\left(\mu+\sigma^2 t\right), x=u+\left(\mu+\sigma^2 t\right)$, 则有

$$
\begin{aligned}
M(t) &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}\left[u^2+\mu^2-\left(\mu+\sigma^2 t\right)^2\right]}\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}u^2}\cdot e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2 t\right)^2\right]}\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2 t\right)^2\right]}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}u^2}\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2 t\right)^2\right]}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}u^2}\mathrm{d}u \\
&\xlongequal{\text{正态分布的 pdf}} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2 t\right)^2\right]}\cdot\sigma\sqrt{2\pi} \\
&= e^{-\frac{1}{2\sigma^2}\left[\mu^2-\left(\mu+\sigma^2 t\right)^2\right]} = e^{-\frac{1}{2\sigma^2}\left(\mu^2-\mu^2-2\mu\sigma^2 t-\sigma^4 t^2\right)} = e^{\frac{1}{2}\sigma^2 t^2+\mu t},
\end{aligned}
$$

即正态分布 $N(\mu,\sigma^2)$ 的矩生成函数为 $\boxed{M(t)=e^{\frac{1}{2}\sigma^2 t^2+\mu t}}$. 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

$$
M(t)=\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{1}{2}\sigma^2 t^2+\mu t\right)^n=1+\mu t+\frac{1}{2!}\left(\mu^2+\sigma^2\right)t^2+\frac{1}{3!}\left(\mu^3+3\mu\sigma^2\right)t^3+\cdots, \quad (4.11)
$$

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 $\mathbb{E}(X-\mathbb{E}X)^n$
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

$$\begin{aligned}M_{Z_n}(t)&=\mathbb{E}e^{tZ_n}=\mathbb{E}e^{t\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}(X_i-\mu)}=\mathbb{E}\left(\prod_{i=1}^{n}e^{t\frac{X_i-\mu}{\sqrt{n}}}\right)\xlongequal{\text{i.i.d.}}\prod_{i=1}^{n}\mathbb{E}\left(e^{t\frac{X_i-\mu}{\sqrt{n}}}\right)\\&=\prod_{i=1}^{n}\mathbb{E}\left(e^{t\frac{X_i}{\sqrt{n}}}\cdot e^{t\frac{-\mu}{\sqrt{n}}}\right)=\prod_{i=1}^{n}e^{t\frac{-\mu}{\sqrt{n}}}\mathbb{E}e^{t\frac{X_i}{\sqrt{n}}}=\prod_{i=1}^{n}e^{t\frac{-\mu}{\sqrt{n}}}M_X\left(\frac{t}{\sqrt{n}}\right)=e^{-\mu\sqrt{n}t}\left(M_X\left(\frac{t}{\sqrt{n}}\right)\right)^n,\end{aligned}$$

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 $\sin n - n = o(n)$. 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 $M(t) = e^{\frac{1}{2}\sigma^2 t^2 + \mu t}$, 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

$$\boldsymbol{Jh}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial h_1(\boldsymbol{\theta})}{\partial x_1} & \cdots & \dfrac{\partial h_1(\boldsymbol{\theta})}{\partial x_K} \\ \vdots & & \vdots \\ \dfrac{\partial h_m(\boldsymbol{\theta})}{\partial x_1} & \cdots & \dfrac{\partial h_m(\boldsymbol{\theta})}{\partial x_K} \end{pmatrix}_{m\times K}, \tag{4.18}$$

因而 $\boldsymbol{h}(\boldsymbol{u})$ 在 $\boldsymbol{u}=\boldsymbol{\theta}$ 处的一阶 Taylor 展开为

$$\boldsymbol{h}(\boldsymbol{u}) = \boldsymbol{h}(\boldsymbol{\theta}) + \boldsymbol{Jh}(\boldsymbol{\theta})\cdot\boldsymbol{u} + R(\boldsymbol{\zeta}),$$

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

$$\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{h}(\boldsymbol{\theta}) + \boldsymbol{Jh}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}} + R(\boldsymbol{\zeta}), \tag{4.19}$$

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

$$\begin{aligned}&\mathrm{Var}\left[\sqrt{n}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{h}(\boldsymbol{\theta})\right)\right] = \mathrm{Var}\left[\sqrt{n}\left(\boldsymbol{Jh}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}} + R(\boldsymbol{\zeta})\right)\right]\\ =&n\mathrm{Var}\left(\boldsymbol{Jh}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}} + R(\boldsymbol{\zeta})\right) = n\mathrm{Var}\left(\boldsymbol{Jh}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}}\right) = n \cdot \boldsymbol{Jh}(\boldsymbol{\theta})\,\mathrm{Var}\left(\hat{\boldsymbol{\theta}}\right)(\boldsymbol{Jh}(\boldsymbol{\theta}))^{\mathrm{T}},\end{aligned}$$

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

$$\mathrm{Var}\left[\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\right] = n \cdot \boldsymbol{Jh}(\boldsymbol{\theta}) \frac{\boldsymbol{\Sigma}}{n} (\boldsymbol{Jh}(\boldsymbol{\theta}))^{\mathrm{T}} = \boldsymbol{Jh}(\boldsymbol{\theta})\,\boldsymbol{\Sigma}\,(\boldsymbol{Jh}(\boldsymbol{\theta}))^{\mathrm{T}} = \mathrm{Var}(\boldsymbol{Jh}(\boldsymbol{\theta})\,\boldsymbol{\xi})$$

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 $\boldsymbol{\theta} \in \mathbb{R}^{K\times 1}$ 的邻域内联系可微. 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

$$\sqrt{n}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{h}(\boldsymbol{\theta})\right) \xrightarrow{d} \boldsymbol{Jh}(\boldsymbol{\theta})\,\boldsymbol{\xi},$$

其中 $m \times K$ 矩阵 $\boldsymbol{Jh}(\cdot)$ 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,$\sqrt{n}\left(\bar{X}_n\bar{Y}_n - \mu_X\mu_Y\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{l\Sigma l}^{\mathrm{T}}\right)$, 其中 $\boldsymbol{h}(\mu_X, \mu_Y) = \mu_X\mu_Y$,$\boldsymbol{l} = \boldsymbol{Jh}(\mu_X, \mu_Y) = (\mu_Y, \mu_X)_{1\times 2}$;
3. 当 $n \to \infty$ 时且 $\mu_Y \neq 0$,$\sqrt{n}\left(\frac{\bar{X}_n}{\bar{Y}_n} - \frac{\mu_X}{\mu_Y}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{l\Sigma l}^{\mathrm{T}}\right)$, 其中 $\boldsymbol{h}(\mu_X, \mu_Y) = \frac{\mu_X}{\mu_Y}$,$\boldsymbol{l} = \boldsymbol{Jh}(\mu_X, \mu_Y) = \left(\frac{1}{\mu_Y}, -\frac{\mu_X}{\mu_Y^2}\right)_{1\times 2}$;
4. 当 $n \to \infty$ 时,$\sqrt{n}\left(\bar{X}_n^2 + \bar{X}_n\bar{Y}_n - (\mu_X^2 + \mu_X\mu_Y)\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{l\Sigma l}^{\mathrm{T}}\right)$, 其中 $\boldsymbol{h}(\mu_X, \mu_Y) = \mu_X^2 + \mu_X\mu_Y$,$\boldsymbol{l} = \boldsymbol{Jh}(\mu_X, \mu_Y) = (2\mu_X + \mu_Y, \mu_X)_{1\times 2}$.

(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 $\mathbb{E}\left\|\mathbf{X}\right\|^2 < \infty$, 则有定理 (4.4.8) 有

1. 如果 $\boldsymbol{\xi} \sim \mathrm{N}(0, \boldsymbol{V})$, 当 $n \to \infty$ 时,
$$\sqrt{n}\left(\boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{h}(\boldsymbol{\theta})\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{Jh}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{Jh}(\boldsymbol{\theta}))^{\mathrm{T}}\right).$$
特别地, 当 $\boldsymbol{\theta}$ 和 $\boldsymbol{h}$ 均为标量时,$\sqrt{n}\left(h\left(\hat{\theta}\right) - h(\theta)\right) \xrightarrow{d} \mathrm{N}\left(0, (h\prime(\theta))^2 V\right)$;
2. 如果参数 $\boldsymbol{\theta} = \mathbb{E}(g(\mathbf{X}))$, 定义参数 $\boldsymbol{\beta} = \boldsymbol{h}(\boldsymbol{\theta})$, 若 $\mathbb{E}\left\|g(\mathbf{X})\right\| < \infty$, 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,
$$\boldsymbol{\beta} = \boldsymbol{h}(\boldsymbol{\theta}), \hat{\boldsymbol{\beta}} = \boldsymbol{h}\left(\hat{\boldsymbol{\theta}}\right), \sqrt{n}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\beta}\right),$$
其中协方差矩阵 $\boldsymbol{V}_{\beta} = \boldsymbol{Jh}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{Jh}(\boldsymbol{\theta}))^{\mathrm{T}}$, 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 $\boldsymbol{V}_\beta$ 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

$$\boldsymbol{Jh}\left(\hat{\boldsymbol{\theta}}\right)\xrightarrow{P}\boldsymbol{Jh}\left(\boldsymbol{\theta}\right),\quad \hat{\boldsymbol{V}}_\beta=\boldsymbol{Jh}\left(\hat{\boldsymbol{\theta}}\right)\hat{\boldsymbol{V}}\left(\boldsymbol{Jh}\left(\hat{\boldsymbol{\theta}}\right)\right)^{\mathrm{T}}\xrightarrow{P}\boldsymbol{V}_\beta,\quad n\to\infty. \tag{4.21}$$

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 (3.6.1) 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 估计量后一部分

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i Y_i \xrightarrow{P} \boldsymbol{Q}_{\mathbf{X}Y} = \mathrm{E}(\mathbf{X}Y), \quad n\to\infty, \tag{5.2}$$

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,

$$\boldsymbol{Q}_{\mathbf{XX}} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right) = \mathbb{E}\left[\begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix}(X_1, \cdots, X_K)\right] = \mathbb{E}\begin{pmatrix} X_1^2 & X_1X_2 & \cdots & X_1X_K \\ X_2X_1 & X_2^2 & \cdots & X_2X_K \\ \vdots & \vdots & & \vdots \\ X_KX_1 & X_KX_2 & \cdots & X_K^2 \end{pmatrix}$$
$$= \begin{pmatrix} \mathbb{E}X_1^2 & \mathbb{E}(X_1X_2) & \cdots & \mathbb{E}(X_1X_K) \\ \mathbb{E}(X_2X_1) & \mathbb{E}X_2^2 & \cdots & \mathbb{E}(X_2X_K) \\ \vdots & \vdots & & \vdots \\ \mathbb{E}(X_KX_1) & \mathbb{E}(X_KX_2) & \cdots & \mathbb{E}X_K^2 \end{pmatrix},$$

则设计矩阵 $\boldsymbol{X}^{\mathrm{T}} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$ 对应的有

$$\boldsymbol{G} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} = \sum_{i=1}^{n}\begin{pmatrix} X_{i1}^2 & X_{i1}X_{i2} & \cdots & X_{i1}X_{iK} \\ X_{i2}X_{i1} & X_{i2}^2 & \cdots & X_{i2}X_{iK} \\ \vdots & \vdots & & \vdots \\ X_{iK}X_{i1} & X_{iK}X_{i2} & \cdots & X_{iK}^2 \end{pmatrix}$$
$$= \begin{pmatrix} \sum\limits_{i=1}^{n}X_{i1}^2 & \sum\limits_{i=1}^{n}X_{i1}X_{i2} & \cdots & \sum\limits_{i=1}^{n}X_{i1}X_{iK} \\ \sum\limits_{i=1}^{n}X_{i2}X_{i1} & \sum\limits_{i=1}^{n}X_{i2}^2 & \cdots & \sum\limits_{i=1}^{n}X_{i2}X_{iK} \\ \vdots & \vdots & & \vdots \\ \sum\limits_{i=1}^{n}X_{iK}X_{i1} & \sum\limits_{i=1}^{n}X_{iK}X_{i2} & \cdots & \sum\limits_{i=1}^{n}X_{iK}^2 \end{pmatrix},$$

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,

$$\frac{\boldsymbol{G}}{n}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}=\begin{pmatrix}\sum\limits_{i=1}^{n}X_{i1}^2 & \sum\limits_{i=1}^{n}X_{i1}X_{i2} & \cdots & \sum\limits_{i=1}^{n}X_{i1}X_{iK}\\ \sum\limits_{i=1}^{n}X_{i2}X_{i1} & \sum\limits_{i=1}^{n}X_{i2}^2 & \cdots & \sum\limits_{i=1}^{n}X_{i2}X_{iK}\\ \vdots & \vdots & & \vdots\\ \sum\limits_{i=1}^{n}X_{iK}X_{i1} & \sum\limits_{i=1}^{n}X_{iK}X_{i2} & \cdots & \sum\limits_{i=1}^{n}X_{iK}^2\end{pmatrix}$$
$$\xrightarrow{P}\boldsymbol{Q}_{\mathbf{XX}}=\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)=\begin{pmatrix}\mathbb{E}X_1^2 & \mathbb{E}\left(X_1X_2\right) & \cdots & \mathbb{E}\left(X_1X_K\right)\\ \mathbb{E}\left(X_2X_1\right) & \mathbb{E}X_2^2 & \cdots & \mathbb{E}\left(X_2X_K\right)\\ \vdots & \vdots & & \vdots\\ \mathbb{E}\left(X_KX_1\right) & \mathbb{E}\left(X_KX_2\right) & \cdots & \mathbb{E}X_K^2\end{pmatrix}, \tag{5.3}$$

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 (5.3), 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 (3.6.1) 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 $Y_i=\beta_0+\beta_1X_i+\varepsilon_i$, 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.

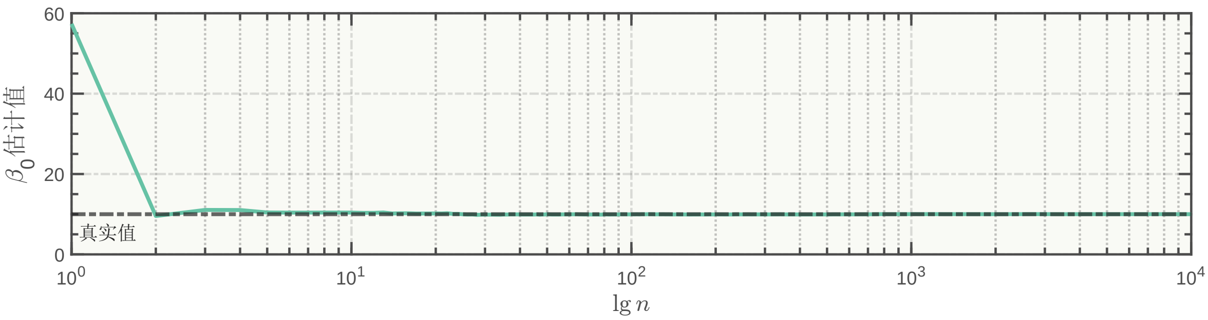

图 5.1: 样本容量增大时的 $\hat{\beta}_0$

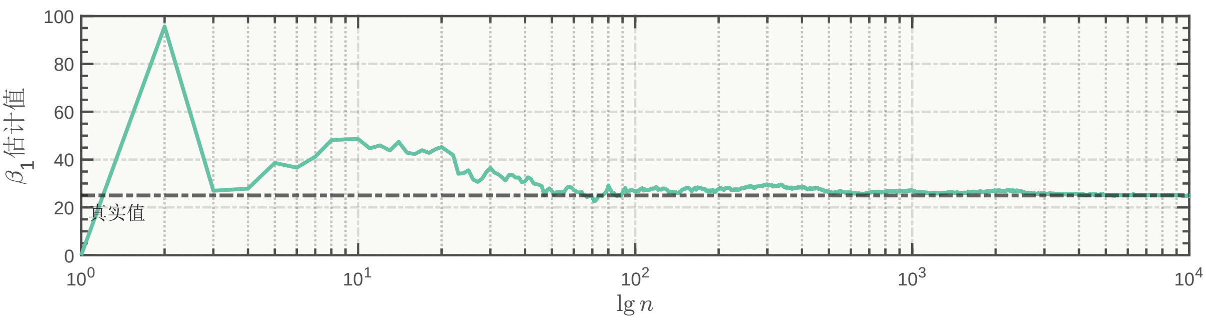

图 5.2: 样本容量增大时的 $\hat{\beta}_1$

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 (3.6.1) 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<u>最优预测量</u>, 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 <u>只是一个假设</u>. 于是我们另辟蹊径, 直接考虑使用线性函数

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

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<u>最优线性预测量</u>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 (3.6.1) 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 (3.6.1). 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. 特别地, 假设解释变量 X 服从正态分布 $\mathrm{N}(\mu, \sigma^2)$, 依据表 (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 $\mathscr{P}\left(Y|X\right)=\beta_0+\beta_1X$ 和 OLS 估计量 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$. 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 $Y=\mathscr{P}\left(Y|X\right)+\varepsilon=\beta_0+\beta_1X+\varepsilon$ 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 $\mathscr{P}\left(Y|X\right)=\beta_0+\beta_1X$.

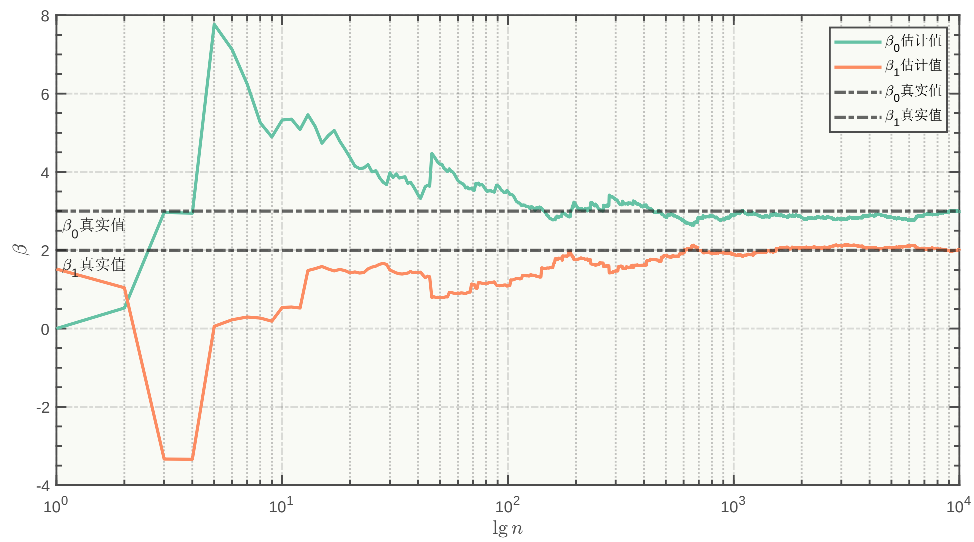

图 5.3: OLS 估计量依概率收敛于最优线性预测 $\mathscr{P}\left(Y|X\right)$

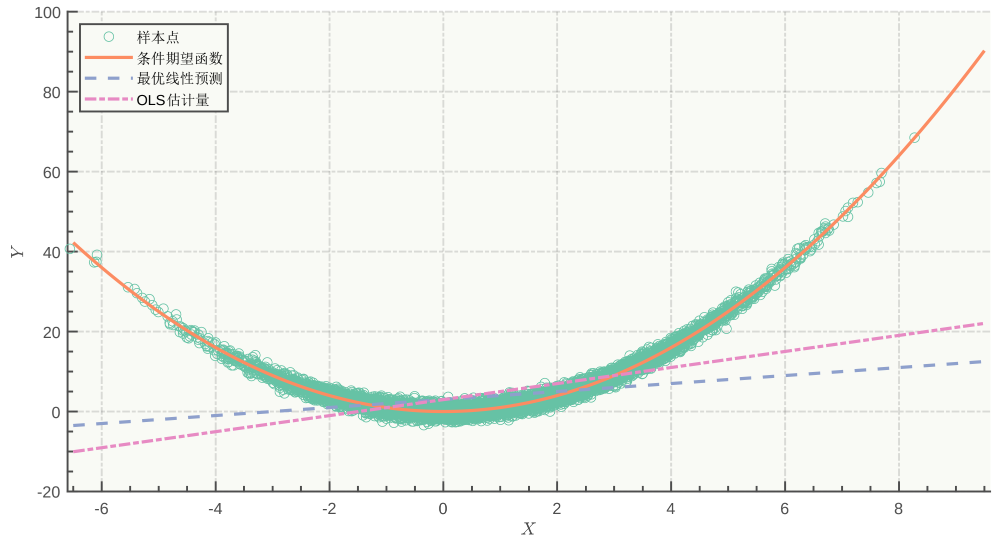

图 5.4: 非线性 CEF 下的 OLS 估计量

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

$$\mathbb{E}\hat{\boldsymbol{Q}}_{\mathbf{X}\varepsilon}=\mathbb{E}\begin{pmatrix}\dfrac{1}{n}\sum\limits_{i=1}^{n}\mathbf{X}_{i1}\varepsilon_i\\ \vdots\\ \dfrac{1}{n}\sum\limits_{i=1}^{n}\mathbf{X}_{iK}\varepsilon_i\end{pmatrix}=\begin{pmatrix}\mathbb{E}X_1\varepsilon\\ \vdots\\ \mathbb{E}X_K\varepsilon\end{pmatrix}=\mathbb{E}\left(\mathbf{X}\varepsilon\right)=\mathbf{0}.$$

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 (3.6.1). 我们对比两者,

$$\begin{aligned}&\text{条件协方差矩阵}:\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1},\\&\text{渐进协方差矩阵}:\boldsymbol{V}_{\boldsymbol{\beta}}=\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right)\left(\mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)\right)^{-1},\end{aligned}$$

则不难发现, 前者是有样本所决定的随机矩阵 (但含有未知的总体参数 $\boldsymbol{D}$), 而后者是渐进分布的总体的一个参数矩阵. 由于

$$n\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}=\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}=\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\right)\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1},$$

[1] 见 Bruce(2021),p.166

可以证明, 当 $n\to\infty$ 时,$n\boldsymbol{V}_{\hat{\beta}}\xrightarrow{d}\boldsymbol{V}_{\boldsymbol{\beta}}$. 特别地, 类似于同方差式 (3.7), 假设当总体满足有 $\operatorname{Cov}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}},\varepsilon^2\right)$ 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).

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 $\boldsymbol{V}_{\hat{\beta}}$ 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 (3.40) 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

$$\begin{aligned}&\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}=\frac{1}{n}\sum_{i=1}^{2}\varepsilon_i^2\xrightarrow{P}\mathbb{E}\varepsilon^2=\sigma^2,\\&\frac{1}{n}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}=\frac{1}{n}\sum_{i=1}^{2}\varepsilon_i\mathbf{X}_i^{\mathrm{T}}=\left(\frac{1}{n}\sum_{i=1}^{2}\varepsilon_iX_{i1},\cdots,\frac{1}{n}\sum_{i=1}^{2}\varepsilon_iX_{iK}\right)\xrightarrow{P}\mathbb{E}\left(\boldsymbol{X}^{\mathrm{T}}\varepsilon\right)=\mathbf{0},\end{aligned}$$

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 (3.44)、(3.45)、(3.46) 和 (3.47) 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其他估计量的形式. 由于

$$\hat{\boldsymbol{Q}}_{\mathbf{XX}} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}, \quad \hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2,$$

则 White 协方差矩阵估计量可以写为

$$\begin{aligned}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\ &=\frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\\ &=\frac{1}{n}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1},\end{aligned}$$

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

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC1}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}} = n\cdot\frac{n}{n-K}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} = \frac{n}{n-K}\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}}, \tag{5.14}$$

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC2}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\bar{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}, \quad \bar{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\bar{\varepsilon}_i^2, \tag{5.15}$$

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC3}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}} = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}\tilde{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}, \quad \tilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\tilde{\varepsilon}_i^2, \tag{5.16}$$

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

$$
\begin{aligned}
\hat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2 \xrightarrow{p} \boldsymbol{\Omega} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right),\\
\bar{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\bar{\varepsilon}_i^2 \xrightarrow{p} \boldsymbol{\Omega} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right),\\
\tilde{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\tilde{\varepsilon}_i^2 \xrightarrow{p} \boldsymbol{\Omega} = \mathbb{E}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\varepsilon^2\right),
\end{aligned}
$$

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.

[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 $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$	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	$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\hat{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}} \xrightarrow{P} \boldsymbol{V}_{\boldsymbol{\beta}}$
	$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}} = \frac{n}{n-K}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC1}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC1}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC1}} \xrightarrow{P} \boldsymbol{V}_{\boldsymbol{\beta}}$
	$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\bar{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC2}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC2}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC2}} \xrightarrow{P} \boldsymbol{V}_{\boldsymbol{\beta}}$
	$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\tilde{\varepsilon}_i^2\right)\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC3}} = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{\mathrm{HC3}}$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC3}} \xrightarrow{P} \boldsymbol{V}_{\boldsymbol{\beta}}$
	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 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 = \hat{\boldsymbol{Q}}_{\mathbf{XX}}^{-1}s^2 = n\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0$	$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 \xrightarrow{P} \boldsymbol{V}_{\boldsymbol{\beta}}^0$

5.3 回归区间与预测区间

5.4 Wald 统计量

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', ...
    'Position',[0 0 21*1.25 29.7*0.35*1.25]);%设置figure为A4纸宽度
tiledlayout(2,1);%创建2*1画布
nexttile
plot(Beta(:,1),'LineWidth',2)
xlabel('\fontname{宋体}样本容量\fontname{Times New ...
  Roman}n','FontSize',12.5)
ylabel('\beta_1\fontname{宋体}估计值','FontSize',12.5);
yline(25,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
    ...
    ,'left','LabelVerticalAlignment','bottom')

nexttile
plot(Beta(:,2),'LineWidth',2)
```

```
xlabel('\fontname{宋体}样本容量\fontname{Times New ...
    Roman}n','FontSize',12.5)
ylabel('\beta_0\fontname{宋体}估计值','FontSize',12.5);
yline(10,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
    ...
    ,'left','LabelVerticalAlignment','bottom')

figure(2)%对数坐标
set(gcf,'unit','centimeters', ...
    'Position',[0 0 21*1.25 29.7*0.35*1.25]);%设置figure为A4纸宽度
tiledlayout(2,1);%创建2*1画布
nexttile
semilogx(Beta(:,1),'LineWidth',2)
xlabel('$\ln n$','FontSize',12.5,'Interpreter','latex')
ylabel('\beta_1\fontname{宋体}估计值','FontSize',12.5);
yline(25,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
    ...
    ,'left','LabelVerticalAlignment','bottom')

nexttile
semilogx(Beta(:,2),'LineWidth',2)
xlabel('$\ln n$','FontSize',12.5,'Interpreter','latex')
ylabel('\beta_0\fontname{宋体}估计值','FontSize',12.5);
yline(10,'-.','\fontname{宋体}真实值','LineWidth',2,'LabelHorizontalAlignment' ...
    ...
    ,'left','LabelVerticalAlignment','bottom')
```

概念索引

符号

B

C

D

E

F

G

H

J

K

L

S

T

W

X

Y

模型索引