
计量经济学 (ECONOMETRICS)

知识总结

PRODUCED BY: 蒋岳祥

JIANGYUEXIANG@ZJU.EDU.CN

EDITOR: RANKFAN MONICARONG

FANXIAOLONG98@GMAIL.COM XMU2020WYR@GMAIL.COM

公众号：经济知识综合

VERSION 1.0

2021 年 2 月 15 日

IF YOU COME ACROSS ANY PROBLEMS, READ OTHER BOOKS FOR POSSIBLE
SOLUTIONS OR CONTACT ME AT FANXIAOLONG98@GMAIL.COM & XMU2020WYR@GMAIL.COM
HAPPY READ-ING!

各位读者好，我现在是一名经管学科的在读研究生，在大三时看过浙江大学蒋岳祥老师的中级计量经济学，后来又看了部分高级计量经济学。推荐资料：[Econometrics Analysis](#)；推荐课程：蒋岳祥老师的[中高级计量经济学](#)，厦门大学洪永淼老师的[高级计量经济学](#)和[概率论与统计学](#)

我被蒋岳祥老师对于计量经济学所吸引，蒋岳祥老师课程曾将“小概率”事件与佛学的“因果循环”联系在一起，顿时让我感受到计量经济学的“哲学”性质。

后来我在网上搜集到蒋岳祥老师的讲稿，由于计量经济学课程公式很多， \LaTeX 在这方面又有很大的优势，历时半年多，我终于将蒋岳祥老师讲稿的主要内容整理成这个总结。

这个总结主要分为四个部分，第一部分：“基础”；第二部分：“线性”；第三部分：“应用”；第四部分：“附录”。

非常感谢我的同学给予的支持，由于我的能力有限，在整理蒋岳祥老师讲稿时未免不出现一些错误，请各位读者见谅。

RankFan
2021 年 2 月 15 日

Preface	i
目录	iii
I 基础	1
1 导言	7
1.1 什么是计量经济学	7
1.2 随机现象的描述	8
1.3 数据本质与统计规律	9
1.3.1 数据的一般属性	9
1.3.2 数据的本质属性	9
1.4 数据的价值与《易经》	10
1.4.1 回归的本质	12
1.5 计量经济学实例	14
1.5.1 消费	14
1.5.2 奥肯定律	15
1.5.3 Solow 模型	16
2 数学基础	17
2.1 矩阵及其二次型 (Matrix and its Quadratic Forms)	17
2.1.1 矩阵的基本概念与运算	17
2.1.2 特殊矩阵	18
2.1.3 矩阵的迹 (trace)	18
2.1.4 矩阵的秩 (rank)	19
2.1.5 统计量的矩阵表示	19

2.1.6	幂等矩阵与二次型	25
2.1.7	微分及其矩阵的微分表示	25
2.2	分布函数 (Distribution function), 数学期望 (Expectation) 与方差 (Variance)	30
2.2.1	概率 (Probability)	30
2.2.2	随机变量 (Random Variable) 和概率分布函数 (Probability Distribution Function)	32
2.2.3	数字期望及方差	35
2.3	条件分布、条件数学期望及其条件方差	38
2.3.1	极限分布理论 (Limit Distribution Theory)	40
2.3.2	实例	43
2.4	数理统计 (Mathematical Statistics)	44
2.4.1	数理统计的基本概	44
2.4.2	正态母体子样的线性函数的分布	46
2.4.3	几种与正态分布 $N(0,1)$ 有关的常用分布	47
2.4.4	统计量的分布与独立性	48
2.4.5	线性变换及二次型的独立性	49
2.4.6	参数估计的常用方法	49
2.4.7	估计的有效性	51
2.4.8	假设检验	53
3	协整理论与误差修正模型	59
3.1	单整 (Integration)	59
3.1.1	稳定序列	59
3.1.2	单整	60
3.1.3	单整的检验	60
3.2	协整 (Cointegration)	61
3.2.1	定义及意义	61
3.2.2	协整的检验	62
3.3	误差修正模型 (ECM)	63
3.3.1	误差修正模型	63
3.3.2	ECM 与协整的关系	64
3.4	单方程计量经济学模型的贝叶斯估计	64
3.4.1	概念	64
3.4.2	单方程计量经济学模型贝叶斯估计的过程	65
3.4.3	正态线性单方程计量经济学模型的贝叶斯估计	66
II	线性	71
4	古典线性回归模型	73
4.1	线性回归模型及其假定	73

4.1.1	构成古典线性回归模型的一组基本假设	74
4.1.2	模型假定的几点说明	74
4.1.3	随机干扰项	75
4.2	最小二乘法	75
4.2.1	最小二乘系数	75
4.2.2	回归拟合的评价	77
4.2.3	方差分析表 (ANOVA)	78
4.3	最小二乘法估计量的统计特征	80
4.3.1	b 是 β 的无偏估计	80
4.3.2	a 是 α 的无偏估计	81
4.3.3	a, b 估计量的协方差矩阵	81
4.3.4	b 是 β 的最小线性无偏估计	82
4.4	最小二乘估计量的统计推断	82
4.4.1	σ^2 的无偏估计量的推导	82
4.4.2	抽样分布	84
4.4.3	β 假设检验	84
4.5	预测	85
4.5.1	个体预测 (Individual Prediction)	85
4.5.2	均值预测 (Mean Prediction)	86
5	多元线性回归模型	87
5.1	多元线性模型及其假定	87
5.2	最小二乘回归	88
5.2.1	最小二乘向量系数	88
5.3	最小二乘估计量的统计特性	89
5.3.1	X 非随机回归量	89
5.3.2	X 随机回归量	90
5.4	最小二乘估计量的统计推断	91
5.4.1	回归系数的假设检验	92
5.4.2	线性约束检验	94
5.4.3	随机 X 及正态 ε 下的检验统计量	95
5.4.4	拟合优度和方差分析	95
5.4.5	回归的显著性检验	98
5.5	预测	98
5.6	分块回归和偏回归	99
5.6.1	偏回归和偏相关系数	101
5.6.2	对均值的离差——对常数回归	101
5.7	偏离正态性的检测 (正态性的哈尔克-贝拉 (Jarque-Bera) BJ 检验)	102
5.7.1	样本偏度与样本峰度	103

5.8 思考题	104
6 正态线性统计模型的最大似然估计	107
6.1 最大似然估计	107
6.2 最优化问题	108
7 带有线性约束的多元线性回归模型及其假设检验	111
7.1 线性约束的检验	111
7.2 参数带有约束的最小二乘估计	113
7.2.1 带有约束的最小二乘函数	113
7.2.2 对约束的检验的另一个方法	115
7.2.3 实例	116
7.3 结构变化与邹至庄检验 (Structure Change and Chou-Test)	117
7.3.1 问题提出	117
7.3.2 问题的模型表述	117
III 应用	121
8 古典线性回归的大样本理论	123
8.1 最小二乘法的有限样本特性	123
8.2 古典回归模型的渐近分布理论	124
8.2.1 依概率分布	125
8.2.2 最小二乘估计量的渐近正态性	127
8.2.3 标准检验统计量的渐近行为	128
8.3 s^2 的一致性和 $\text{Var}[b]$ 的估计量	130
8.3.1 B 的函数的渐近分布——得尔塔方法	131
9 非线性回归模型	133
9.1 线性化回归	133
9.2 非线性最小二乘估计	134
9.2.1 非线性最小二乘准则函数	135
10 非球形扰动项与广义最小二乘 (GLS)	137
10.1 问题的提出	137
10.1.1 异方差性	137
10.1.2 自相关	138
10.1.3 普通最小二乘法的结果	138
10.1.4 有限样本特性	138
10.1.5 最小二乘法的渐近特性	139
10.2 广义最小二乘 (GLS)	140

10.2.1 GLS 大样本的特性	141
10.3 最大似然估计和似然比统计量	142
10.3.1 二元正态分布	142
10.3.2 多元正态分布	142
10.3.3 最大似然估计	143
10.3.4 最大似然估计量的渐进特性	144
10.3.5 似然比统计量	145
10.3.6 线性约束检验的另一个统计量	147
10.4 可行的最小二乘估计 (FGLS)	147
10.4.1 Ω 知的条件下	147
10.5 广义矩法 (GMM) 估计量	149
10.5.1 广义矩法 (GMM) 估计量	149
10.5.2 一致估计——矩方法	150
10.5.3 随机抽样和分布的参数估计	150
10.6 异方差的检验异方差的多数检验均基于下述策略	154
10.6.1 怀特的一般检验 (White's General Test)	154
10.6.2 戈德菲尔德-匡特检验 (The Goldfeld-Quandt Test)	155
10.7 练习	156

IV Appendix 157

A 关于生产函数的计量经济模型 159

A.1 几个重要基本概念	159
A.1.1 生产函数	159
A.1.2 要素替代弹性	161
A.1.3 要素的产出弹性	162
A.1.4 技术进步	163
A.2 以要素之间替代性质的描述为线索的生产函数模型的发展	164
A.2.1 线性生产函数模型	164
A.2.2 投入产出生产函数模型	164
A.2.3 C-D 生产函数模型	164
A.2.4 不变替代弹性 (CES) 生产函数模型	166
A.2.5 变替代弹性 (VES) 生产函数模型	168
A.2.6 多要素生产函数模型	169
A.2.7 超越对数生产函数模型	171
A.3 以技术要素的描述为线索的生产函数模型的发展	171
A.3.1 将技术要素作为一个不变参数的生产函数模型	171
A.3.2 改进的 C-D、CES 生产函数模型	172
A.3.3 含体现型技术进步的生产函数模型	173

A.3.4	边界生产函数模型	175
A.4	几个重要生产函数模型的参数估计方法	176
A.4.1	线性生产函数模型的估计	176
A.4.2	C-D 生产函数模型及其改进型的估计	176
A.4.3	CES 生产函数模型及其改进型的估计	176
A.4.4	VES 生产函数的估计	177
A.4.5	二级 CES 生产函数模型的估计	178
A.4.6	含体现型技术进步的生产函数模型的估计	179
A.4.7	确定性统计边界生产函数模型的修正的普通最小二乘估计 (COLS)	179
A.5	生产函数模型在技术进步分析中的应用	180
A.5.1	技术进步速度的测定	181
A.5.2	技术进步对增长的贡献	181
A.5.3	部门之间、企业之间技术进步水平的比较分析	182
A.6	建立生产函数模型中的数据质量问题	183
A.6.1	样本数据的一致性问题	183
A.6.2	样本数据的准确性问题	183
A.6.3	样本数据的可比性问题	184
B	计量经济模型的类别	185
B.1	模型的类别	185
B.2	事例：社会保障水平与国内生产总值	186
C	其他	189
C.1	经济学家预测奥运奖牌	189
C.1.1	经济学家无需分析运动员实力	189
C.1.2	美俄中占据前三位	190
C.1.3	奖牌分布趋势	190
C.1.4	最后结果	190
C.2	学生邮件	191
C.3	2005 年计量经济学作业参考答案	193
C.3.1	2005 年计量经济学作业参考答案一	193
C.3.2	2005 年计量经济学作业参考答案二	194
C.3.3	2005 年计量经济学作业参考答案三	195
C.3.4	2005 年计量经济学作业参考答案四	196
C.3.5	总结	197

Part I

基础

课程简介

课程号：

课程名称 (含英文主)：研究生《中级计量经济学》(Intermediate Econometrics)

学分：2

周学时：4

预修课程：经济学、高等数学、概率论和数理统计

内容简介：

首先介绍计量经济学中必须具备的数学知识如高等代数中矩阵、概率论与数理统计中点估计、有效估计、一致估计、区间估计、假设检验、大样本与极限理论等。而后介绍古典线性回归模型、多元线性回归模型、带有线性约束的多元线性回归模型及其假设检验、正态线性统计模型的最大似然估计、古典线性的大样本理论、非球形扰动与广义最小二乘、异方差性、非线性回归模型等。

选用教材或参考书：

教材：William H. Greene, **Econometrics Analysis**, fourth edition。

参考书：

1. William H. Greene, 经济计量分析, Econometrics Analysis 的翻译, 中国社会科学出版社。
2. 课件。

教学大纲

课程号：

课程名称 (含英文主)：研究生《中级计量经济学》(Intermediate Econometrics)

学分：2

周学时：4

预修课程：经济学、高等数学、概率论和数理统计

一、课程的教学目的和基本要求：

本课程为已具备经济学，概率论和数理统计以及初级计量经济学的研究生开设的《中级计量经济学》。目的是为他们今后在经济和金融领域能够独立开展科学研究和调查提供坚实的统计与计量经济学的方法与技巧。本课程的重点是使学生充分掌握和理解以下三个方面的知识 with 技能：

- 1. 计量经济学的理论与原理；
- 2. 计量经济学中广泛使用的统计推断知识，方法与技巧；
- 3. 掌握各种模型需要的条件与模型的局限性和适用性。

二、课程内容与学时分配：

第一章 引言 1 学时

- 计量经济学概念
- 为什么学计量经济学
- 计量经济学模型

第二章 矩阵的基础知识 4 学时

- 矩阵的概念与运算
- 矩阵的特征根与特征向量
- 矩阵的二次型与二项式
- 矩阵的微分

第三章 概率论与数理统计 4 学时

- 随机变量与概率
- 分布函数与中心极限定理

- 二元态分布与多元正态分布
- 样本与样本的分布函数
- 统计量及其分布
- 点估计
- 有效估计
- 一致估计
- 区间估计
- 假设检验

第四章 古典线性回归模型 4 学时

- 古典线性回归模型与其假设条件
- 最小二乘回归
- 方差分析
- 最小二乘统计量的有限样本性质
- 预测

第五章 多元线性回归模型 4 学时

- 多元线性回归模型与其假设条件
- 最小二乘回归
- 方差分析
- 最小二乘统计量的有限样本性质
- 预测

第六章 带有线性约束的多元线性回归模型及其假设检验 4 学时

- 带有线性约束的多元线性回归模型与其假设条件
- 线性约束的检验
- 参数带有约束的最小二乘回归
- Wald 检验
- 实例

第七章 正态线性统计模型的最大似然估计 4 学时

- 模型及其假设条件
- 模型求解
- 与最小二乘估计量的比较

第八章 古典线性回归的大样本理论 4 学时

- 最小二乘统计量的有限样本性质
- 古典回归模型的渐近分布理论
- 最小二乘估计量的渐近正态性
- 标准检验统计量的渐近行为

第九章 非线性回归模型 2 学时

- 非线性回归模型
- 可供选择的几个统计量
- 假设检验与参数约束
- Box-Cox 变换

第十章 异方差性 2 学时

- OLS 估计的探讨
- 异方差性的检验
- GLS 估计
- 二阶段估计

CHAPTER 1

导言

学习计量经济学应该准备什么？怎么去做？首先推荐浙江大学蒋岳祥老师的[中高级计量经济学](#)，厦门大学洪永淼老师的[高级计量经济学](#)和[概率论与统计学](#)。

§ 1.1 什么是计量经济学

计量经济学是由挪威经济学家 R.Fisher 在三十年代首先创立的一门学科，是关于运用统计方法测量经济关系的艺术与科学，已经成为现代经济学的重要组成部分之一。

如果要给**计量经济学 (Econometrics)** 下一个较为确切的定义，我们可以这样界定：

计量经济学是这样一门学科，它根据以往历史的经济资料与数据，从经济理论出发，运用数理统计的分析方法对经济关系建立经济计量模型，并依据所建立的模型对经济系统进行结构分析，经济预测和政策评价。所以计量经济学涉及数学学科中的统计学领域和经济学领域，统计学与经济理论是计量经济学的两块基石。

经济现象包罗万象，影响经济的因素有很多，如果我们企图将所有的因素作为研究的对象，我们可能什么结论也得不到，研究经济问题的一般方法是：我们总是选用最重要的因素变量而屏弃一些非本质的因素（变量），还需要了解哪些经济现象是有待解释的，哪些重要因素是有助于解释这些经济现象的，如何度量量化那些因素，并努力寻求它们之间存在的数量关系，并用统计推断来检验这些关系，如图 1.1.1 on the following page。

计量经济学的研究内容：随机现象，客观事物所反映的规律。

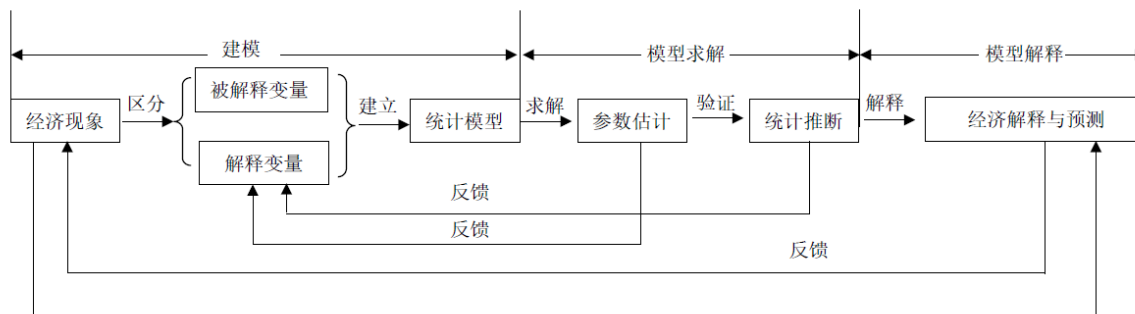


图 1.1.1: 计量经济模型建立, 求解, 解释过程图

§ 1.2 随机现象的描述

工学和商学的思维模式与分析方法存在较大差异。一个是在确定下环境下少数可控变量关系的研究。另一个是研究在不确定性环境下无限且不可控变量关系的问题。

实质上金融经济理论是研究不确定性的情况下, 未来资源的如何最佳配置的问题。

银行为什么按揭贷款给年轻人? 根据不同人来确定吗? 未来的资产拿到今天来用。研究随机现象是概率论与数理统计的知识。随机现象的处理, 增加一个维度, 例如基金经理投资成功的概率。

$$\lim_{n \rightarrow \infty} p^n = 0 \quad \text{where } 0 < p < 1$$

研究对象的数据化, 随机现象的描述——数据化。我们所看到的任何随机事件从理论上讲都可以通过将随机事件与数值联系。这样试验的结果就能有一个数来表示, 这个数是随着实验的结果的不同而变化, 也即它是样本点一个函数, 这种量被称为随机变量。

对于随机现象, 最重要的是两件事, 其一随机现象, 其二出现结果的可能性。从而用来描述随机变量的大小变化和变化的可能性都可以量化。

随机现象 A_1, A_2, \dots, A_N , 为 \mathbf{A} 事件的集合体。 I_{A_i} 表示事件是否发生。

$$\begin{aligned} \mathbf{A} &= A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n \\ &= \mathbf{A}\Omega = \mathbf{A} \sum_{i=1}^n A_i = \sum_{i=1}^n \mathbf{A}A_i \\ P(\mathbf{A}) &= P\left(\sum_{i=1}^n \mathbf{A}A_i\right) = \sum_{i=1}^n P(\mathbf{A}A_i) = \sum_{i=1}^n P(\mathbf{A})P(A_i) \end{aligned}$$

§ 1.3 数据本质与统计规律

概率论研究的分布函数，反映了事物变化和概率结合在一起。研究变量分为连续性变量，离散性变量。计量经济学的三个定义，一是研究变量的分布函数，即古人所说的易，变化和概率（可能性）。

$$\begin{aligned} F(X) &= P(X \leq x) = \int_{-\infty}^x f(x)dx \\ F(x_2) - F(x_1) &= P(x_1 < X < x_2) \\ Y &= X\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \implies Y \sim N(X\beta, \sigma^2) \end{aligned} \quad (1.3.1)$$

计量模型估计 β , $\varepsilon^2 \rightarrow$ 求 Y 的分布函数。在均值附近变化一点，概率变化会很大。我国“古代思想”的概率论数据是基于随机现象的产生和发展得到的记录了事物发展的轨迹得到的数据，不一定能够反映事物的本质不利用数据是否可以分析事物就是我们所说的玄学。

§ 1.3.1 数据的一般属性

$X_1, X_2, X_3, \dots, X_n \sim F_\theta(x)$ 表明 $X_1, X_2, X_3, \dots, X_n$ 独立同分布，

研究一个事物的要求分布函数，回归即是求分布函数，数据的分布（ $X \sim F_\theta(x)$ ）获得全样本是小概率事件，样本数据 \implies 母本数据。

$$\lim_{n \rightarrow \infty} \left\{ P \left(X = \frac{1}{n} \sum X_i \leq x \right) \right\} = G_\theta(x) \sim N(\mu, \sigma^2)$$

$n \rightarrow \infty$ 即为大样本， $\mu = \mathbb{E}(X)\sigma^2 = \text{var}(X)$ 平均属性 \rightarrow 平均的分布函数，查表时需进行标准化。哲学思想：天下乌鸦一般黑，事物随机现象的平均属性都服从正态分布。特例黑天鹅事件，

$$P \left(\frac{X - \mu}{\sigma} \right) \sim N(0, 1)$$

§ 1.3.2 数据的本质属性

Fisher & Tippett 1928 年已经完成，其极限分布只有三种类型。既所谓的极值分布是下列三种类型中的一种。最大值分布的特征事物的本质特征，事物最基本的本质。

$$\begin{aligned} P[\max(X_1, X_2, X_3, \dots, X_n)] &< x \\ \lim_{n \rightarrow \infty} P[\max(X_1, X_2, X_3, \dots, X_n)] &=? \end{aligned}$$

1. Frechet 分布 $\Leftrightarrow \rho > 0$

2. Gumbel 分布 $\Leftrightarrow \rho < 0$

3. Weibull 分布 $\Leftrightarrow \rho = 0$

$$\min(X_1, X_2, X_3, \dots, X_n) = \max(-X_1, -X_2, -X_3, \dots, -X_n)$$

总结: 假设研究对象是随机的, 增加一个维度, 用概率论的分布函数描述描述事物。

Definition 1.1 研究对象的分布函数;

$$\mathbf{Y} = \beta' \mathbf{X} + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \implies \mathbf{Y} \sim N(\beta' \mathbf{X}, \sigma^2)$$

Definition 1.2 分部函数怎么确定? 随机现象的平均属性基本相同, 2008 年金融危机后研究点, 侧重于本质属性。MLE 方法、OLS 方法和矩估计等都可以估计参数 $\mathbf{X} \sim F_\theta(\mathbf{x})$

$$\begin{cases} P(\bar{X} < x) & \implies \text{Law of Large Numbers} \\ P(\max(X_1, X_2, \dots, X_n) < x) & \implies \text{Law of Large Numbers} \end{cases}$$

$G_X(X)$ 可以用一个等式表示:

$$H_p(X) = \begin{cases} e^{\{-(1+\rho x)^{-1/\rho}\}} & \implies \text{if } \rho \neq 0 \\ e^{\{-e^{-x}\}} & \implies \text{if } \rho = 0 \end{cases}$$

$\rho = 0$ 状态就为太极 (极限的极限), 道的最高境界 (无或双有)。数据是错的怎么办, 利用古代人的做法, 还原本质。

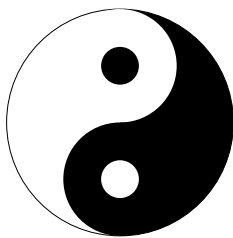


图 1.3.2: 太极图

§ 1.4 数据的价值与《易经》

回归包括线性回归和非线性回归, 数据决定模型。计量经济学一般会求均值 (一阶矩) 和方差 (二阶矩)。一般来说 $\bar{X} \sim N(\mu, \sigma^2)$ 。用样本去估计母体, t 检验使用 t 统计量。为什么会使用 t 统计量。检验优先样本可以获得 \bar{x} , S^2

单个变量之间所表现出来的规律，但世界之间是有联系的，X 对 Y 有无影响， $H_0: \beta = 0$ ，用 t 统计量。拒绝原假设时， $\beta \neq 0$ 。 $\beta = 0$ 是不是小概率事件？

$$Y = \beta' X + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \implies Y \sim N(\beta' X, \sigma^2)$$

$$(X_1, Y_1) \quad (X_2, Y_2) \quad (X_3, Y_3) \cdots \cdots (X_N, Y_N) \implies \text{Regression Coefficient}$$

表 1.1 两类错误

检验策略	H_0 为真	H_0 非真
拒绝 H_0	犯 I 类错误 (α)	正确
接受 H_0	正确	犯 II 类错误 (β)

一个事情虽然说用很小，但未必不发生作用，发生时就会为黑天鹅事件，计量经济学的背景为不存在黑天鹅事件。现实社会中一个政策把事物的性质可以发生改变。很多因素都存在着非回归的关系。

数据的两大属性：数据的一般属性和本质属性对应的反映事物的一般属性和本质属性，

任何真正反映事物发展轨迹的数据是和事物本质一样，有“性”，并同样存在同性恋的情况。对于任何研究，首先要做定性分析，而非定量分析分析，任何事物都有“性”，“性”是事物的本质，而且可以做定量分析。《易经》与统计学的关系，《易经》研究的是风险，规避统计学建立的模型分析和预测也是风险规避。“易”是事物变化的统计规律，“易”是事物的变化和可能性。“易”变化的复杂和简单相统一，极限是科学的语言极限，是中心极限定理、小概率事件的基础。

Example 1.1 传统文化为什么阻碍中国人的科学技术进步？中国偏向于小概率事件。

Example 1.2 就业愿望的衡量。劳动参与率 (*Labor Force Participation Rate*)。L 为收益率， ε 为冲击。

$$LFPR = \alpha + \beta L + \varepsilon$$

可以有如下原假设： $H_0: \beta = 0$ ， $H_0: \beta < 0$ 。将工资水平考虑到估计中，分析工资的挤出效应。

$$LFPR = \alpha + \beta_1 L + \beta_2 FH + \varepsilon$$

计量经济学常用的表达式：

1) ε 假设

2) $\hat{\alpha} \quad \hat{\beta}_1 \quad \hat{\beta}_2$ 线性部分， $\hat{\sigma}^2$ 随机部分

3) $H_0 : \beta_1 = 0, H_0 : \beta_2 = 0$

4) $H_0 : \beta_1 = \beta_2 = 0$

5) R^2 拟合优度, 判定系数

6) 预测, 随机现象。

§ 1.4.1 回归的本质

设随机变量 $\beta' \mathbf{X}, \mathbf{Y}, \mathbf{X}^T = (X_1, \dots, X_m)$ 是 m 维随机向量, 它是可以预先测量的, 希望通过 \mathbf{X} 预测 \mathbf{Y} , 也就是说要寻找一个函数 $y = M(x_1, \dots, x_m)$ 当 \mathbf{X} 的观察值为 \mathbf{x} 时, 就把 $M(\mathbf{x})$ 作为对 \mathbf{Y} 的预测值。当然一般总希望一个好的预测, 其均方预测误差应达到最小。

均值处理随机性, $M(\mathbf{x})$ 与 \mathbf{Y} 最大相关, 此时相关系数最大。

$$\begin{aligned} \mathbb{E}[Y - L(X)]^2 &= \mathbb{E}[Y - M(X)]^2 + \mathbb{E}[M(X) - L(X)]^2 \\ \rho[Y, M(x)] &= \max_L \rho[Y, L(X)] \end{aligned}$$

设 (\mathbf{X}, \mathbf{Y}) 的分布密度函数是 $f(x, y)$, 设 \mathbf{X} 的边际分布密度函数是 $f_1(x)$, \mathbf{Y} 关于 \mathbf{X} 的条件分布密度是:

$$Y = \begin{cases} f(x, y)/f_1(x) & f_1(x) \neq 0 \\ 0 & f_1(x) = 0 \end{cases}$$

$$\begin{aligned} M(x) &\triangleq E(Y/x) = \int y f(y/x) dy \\ \mathbb{E}[Y - L(X)]^2 &= \iint \dots \int [y - L(x)]^2 f(x, y) dx dy \\ &= \iint \dots \int [y - L(x) + M(x) - M(x)]^2 f(x, y) dx dy \\ &= \iint \dots \int [y - M(x)]^2 f(x, y) dx dy \\ &\quad + 2 \iint \dots \int [y - M(x)][M(x) - L(x)] f(x, y) dx dy \\ &\quad + \iint \dots \int [M(x) - L(x)]^2 f(x, y) dx dy \\ \Rightarrow &\iint \dots \int [y - M(x)][M(x) - L(x)] f(x, y) dx dy \\ &= \iint \dots \int [y - M(x)][M(x) - L(x)] f(y/x) f_1(x) dy dx \\ &= \int \dots \int [M(x) - L(x)] f_1(x) \left[\int y f(y/x) dy - M(x) \right] dx = 0 \end{aligned}$$

$$\mathbb{E}[Y - L(X)]^2 = \mathbb{E}[Y - M(X)]^2 + \mathbb{E}[M(X) - L(X)]^2 \quad (1.4.1)$$

右边第一项与 $L(X)$ 无关, 第二项大于等于零, 它等于零的充要条件是 $M(X) = L(X)$, 它表示当 $M(X) = L(X)$ 时, $\mathbb{E}[Y - L(X)]^2$ 达到最小值 $\mathbb{E}[Y - M(X)]^2$ 。在统计学上, 我们称 $Y = M(X) = \mathbb{E}(Y|X)$ 为 Y 关于 X 的回归曲线。

$$\min [Y - M(x)]^2 = \min_L [Y - L(x)]^2$$

Theorem 1.4.1 (条件均值)

$$M(x) = \mathbb{E}(Y|X)$$

$$Y = \mathbb{E}(Y|x) + [Y - \mathbb{E}(Y|x)] = \mathbb{E}(Y|x) + \varepsilon \quad \mathbb{E}(\varepsilon) = 0$$

$\mathbb{E}(\varepsilon) \neq 0$ 将会如何? 当 X, Y 满足二元正态分布时, 可以得到一般线性回归, 求 $Y = \mathbb{E}(Y|x) + \varepsilon$, 即指导联合分布。

$$Y = F(\beta, X) + \varepsilon, \mathbb{E}(\varepsilon) = 0, \mathbb{E}(\varepsilon\varepsilon') = \sigma^2\Omega$$

$\mathbb{E}(\varepsilon\varepsilon')$ 为协方差矩阵, $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ 为样本点, OLS 回归会假设服从正态性分布。随机变量之间的联系是通过协方差系数。

$$\begin{aligned} \rho_{XY} &= \frac{\mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y))}{\sqrt{DX}\sqrt{DY}} \\ \text{cov}(\varepsilon_i, X_j) &= 0 \\ \rho_{Y, M(X)}^2 &= \frac{\text{cov}^2[Y, E(Y|X)]}{\text{var}(Y)\text{var}[E(Y|X)]} \\ &= \frac{\text{cov}^2[E(Y|X), E(Y|X)]}{\text{var}(Y)\text{var}[E(Y|X)]} \\ &= \frac{\text{var}[E(Y|X)]}{\text{var}(Y)} = R^2 \\ \Rightarrow \widehat{R_{Y, \hat{Y}}^2} &= \frac{SSR}{SST} = \widehat{\rho_{Y, \hat{Y}}^2} \end{aligned} \quad (1.4.2)$$

统计学中求均值, 方差, 相关系数, ρ 为相关系数阵, 是变量之间的关系网。古典线性回归是对现实问题的简化再简化。样本数据一般会存在异方差问题, $\sigma_1^2 \neq \sigma_2^2$, 如组间数据异方差。

自相关问题，一阶自回归模型。

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & \sigma^2 & \rho & & \rho^{n-2} \\ \rho^2 & \rho & \sigma^2 & & \vdots \\ \vdots & & & \ddots & \rho \\ \rho^{n-1} & \cdots & & \rho & \sigma^2 \end{pmatrix}$$

ARCH (条件异方差), GARCH (广义条件异方差)。对 $L(\alpha, \beta)$ 求偏导, 得到 $\hat{\alpha}, \hat{\beta} \Rightarrow Y_i = \hat{\alpha} + \hat{\beta}X_i + e_i$ 。
得到 e_1, e_2, \dots, e_n 检验 $\varepsilon \sim N(0, \sigma^2)$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix}$$

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad L(\alpha, \beta) = \sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2$$

1. $X_1, X_2, \dots \Rightarrow Y$, 利用联合分布 $f(X_1, X_2, \dots, X_m) \Rightarrow L(\cdot)$
2. $Y = \beta X + \varepsilon \quad H_0: \beta_1 = 0 / H_0: \beta_2 = 0$
3. $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, 对于 PRF, $\mathbb{E}(\varepsilon) = 0 \quad \mathbb{E}(\varepsilon \varepsilon') = \sigma^2 \Omega$ 是否满足。
4. OLS。 $L(\alpha, \beta) = \sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2 \Rightarrow Y = \hat{\alpha} + \hat{\beta}X + e$ (Sample Regression Function) 理解为条件回归 + 波动, 一般有几个假定, 就会有几个检验。

§ 1.5 计量经济学实例

§ 1.5.1 消费

学过经济学中凯恩斯经济理论的人都知道, 理论上说消费和收入存在着密切的联系, 如果 C 表示消费, Y 表示收入。则 C 与 Y 的关系, 可用消费函数表示:

$$C = f(Y) \tag{1.5.1}$$

这样的函数满足:

- 1) 边际消费倾向 (MPC) 位于 0 和 1 之间;
- 2) 平均消费倾向 (APC) 是随着收入的增加而减少。

$$\begin{aligned}
 \frac{dC}{dY} &= \frac{d(C \cdot \frac{1}{Y})}{dY} \\
 &= \frac{dC}{dY} \cdot \frac{1}{Y} - \frac{1}{Y^2} C \\
 &= \frac{1}{Y} \left(\frac{dC}{dY} - \frac{C}{Y} \right) = \frac{1}{Y} (MPC - APC) < 0
 \end{aligned}$$

在现实经济社会中，消费与收入之间的关系很难确切地用方程（1）表示收入，我们所能采集到的数据往往受到这样那样的影响，我们可用随机扰动来表示这些影响，所以，我们要对方程（1）要作适当调整，于是消费和收入之间的关系可以写成如下形式：

$$C = f(Y, \varepsilon) \quad (1.5.2)$$

其中 ε 是随机扰动。满足凯恩斯条件的 ε 很多，无法枚举穷尽，但我们可以大致将它们分为线性模型与非线性模型两类。

1. 线性模型 (Linear Model)

方程(1.5.2)的一个最简单的情况，是 C 与 Y 的线性关系，即：

$$C = \alpha + \beta Y + \varepsilon \quad (1.5.3)$$

如果我们现在从历史记录中或观察到 N 个样本，即 (Y_t, C_t) , $t = 1, 2, \dots, N$ ，于是我们有如下一组方程：

$$\begin{aligned}
 C_1 &= \alpha + \beta Y_1 + \varepsilon_1 \\
 C_2 &= \alpha + \beta Y_2 + \varepsilon_2 \\
 &\dots\dots \\
 C_N &= \alpha + \beta Y_N + \varepsilon_N
 \end{aligned}$$

2. 非线性模型 (Nonlinear Model)

一般情况下，方程(1.5.2)都是非线性的情况。现在我们假设 $0 < \nu < 1$, $MPC > 0$ ，即该模型满足凯恩斯的两个条件，这就是一个典型非线性模型。例如：

$$C = \alpha + \beta Y^\nu + \varepsilon$$

§ 1.5.2 奥肯定律

奥肯定律：失业率每上升 1%，实际 GDP 下降 2%，这是线性回归的结果。

§ 1.5.3 Solow 模型

带有 μ 的 Solow 模型, 西方经济理论在中国未必行得通, 因为具体分布不同。希克斯提出中性技术条件。超越对数的生产函数, α 劳动的产出弹性, β 资本的产出弹性。

$$\begin{aligned} Y &= A(t)f(L, K) = A_0 e^{\mu t} f(L, K) \\ \ln Y &= \ln A_0 + \mu t + \ln f(L, K) \end{aligned} \quad (1.5.4)$$

$$\begin{aligned} \frac{d \ln Y}{dt} &= \mu + \frac{d \ln f(L, K)}{dt} \\ \frac{d \ln Y}{Y dt} &= \mu + \frac{\partial f}{f \partial L} \frac{dL}{dt} + \frac{\partial f}{f \partial K} \frac{dK}{dt} \\ &= \mu + \frac{L}{f} \frac{\partial f}{\partial L} \frac{1}{L} \frac{dL}{dt} + \frac{L}{f} \frac{\partial f}{\partial K} \frac{1}{K} \frac{dK}{dt} \end{aligned}$$

$$\frac{d \ln Y}{Y dt} = \mu + \frac{L}{A_0 e^{\mu t} f} \frac{\partial f}{\partial L} \frac{1}{L} \frac{dL}{dt} + \frac{L}{A_0 e^{\mu t} f} \frac{\partial f}{\partial K} \frac{1}{K} \frac{dK}{dt} \quad (1.5.5)$$

$$\begin{aligned} \frac{\Delta Y}{Y} &\stackrel{dt=1}{=} \mu + \alpha \frac{\Delta L}{L} + \beta \frac{\Delta K}{K} \\ 1 &= \frac{\mu}{\Delta Y/Y} + \alpha \frac{\Delta L/L}{\Delta Y/Y} + \beta \frac{\Delta K/K}{\Delta Y/Y} \\ \frac{\mu}{\Delta Y/Y} &\text{ is the Technological innovation} \end{aligned}$$

CD 函数, $\alpha + \beta = 1$, 现对规模效应做检验, 则 $\beta = 1 - \alpha$

$$\frac{\Delta Y}{Y} = \mu + \alpha \frac{\Delta L}{L} + (1 - \alpha) \frac{\Delta K}{K}$$

§ 2.1 矩阵及其二次型 (Matrix and its Quadratic Forms)

§ 2.1.1 矩阵的基本概念与运算

一个 $m \times n$ 矩阵可表示为:

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

矩阵的加法较为简单, 若 $\mathbf{C} = \mathbf{A} + \mathbf{B}$, $c_{i,j} = a_{i,j} + b_{i,j}$ 。但矩阵的乘法的定义比较特殊, 若 \mathbf{A} 是一个 $m \times n_1$ 的矩阵, \mathbf{B} 是一个 $n_1 \times n$ 的矩阵, 则 $\mathbf{C} = \mathbf{AB}$ 是一个 $m \times n$ 的矩阵, 而且, 一般来讲, $\mathbf{AB} \neq \mathbf{BA}$, 但如下运算是成立的:

a. 结合律 (Associative Law) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$

b. 分配律 (Distributive Law) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$

向量 (Vector) 是一个有序的数组, 既可以按行, 也可以按列排列。行向量 (row vector) 是只有一行的向量, 列向量 (column vector) 只有一列的向量。如果 α 是一个标量, 则 $\alpha\mathbf{A} = [\alpha a_{ij}]$ 。

矩阵 \mathbf{A} 的转置矩阵 (transpose matrix) 记为 \mathbf{A}' , 是通过把 \mathbf{A} 的行向量变成相应的列向量而得到。显然 $(\mathbf{A}')' = \mathbf{A}$, 而且 $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$ 。

乘积的转置 (Transpose of production) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$

可逆矩阵 (inverse matrix), 如果 n 级方阵 (square matrix) \mathbf{A} 和 \mathbf{B} , 满足 $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$ 。则称 \mathbf{B} 、 \mathbf{B} 是可逆矩阵, 显然 $\mathbf{A} = \mathbf{B}^{-1}$, $\mathbf{B} = \mathbf{A}^{-1}$ 。如下结果是成立的:

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}, \quad (\mathbf{A}')^{-1} = (\mathbf{A}^{-1})', \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

§ 2.1.2 特殊矩阵

- 1) 恒等矩阵 (identity matrix): 对角线上元素全为 1, 其余全为 0, 可记为 \mathbf{I} ;
- 2) 标量矩阵 (scalar matrix): 即形如 $\alpha\mathbf{I}$ 的矩阵, 其中 α 是标量;
- 3) 幂等矩阵 (idempotent matrix): 如果矩阵 \mathbf{A} 具有性质 $\mathbf{A} \cdot \mathbf{A} = \mathbf{A}^2 = \mathbf{A}$, 这样的矩阵称为幂等矩阵。

Theorem 2.1.1 幂等矩阵的特征根要么是 1, 要么是零。

- 4) 正定矩阵 (positive definite) 和负定矩阵 (negative definite), 非负定矩阵 (nonnegative) 或半正定矩阵 (positive semi-definite), 非正定矩阵 (nonpositive definite) 或半负定矩阵 (negative semi-definite);

对于任意的非零向量, 如有 $\vec{x}'\mathbf{A}\vec{x} > 0$ (< 0), 则称 \mathbf{A} 是正 (负) 定矩阵; 如有 $\vec{x}'\mathbf{A}\vec{x} \geq 0$ (≤ 0), 非负 (非正) 定矩阵。如果 \mathbf{A} 是非负定的, 则记为 $\mathbf{A} \geq 0$; 如果是正定的, 则记为 $\mathbf{A} > 0$ 。协方差矩阵是半正定矩阵, 几个结论:

- a) 恒等矩阵或单位矩阵是正定的;
 - b) 如果 \mathbf{A} 是正定的, 则 \mathbf{A}^{-1} 也是正定的;
 - c) 如果 \mathbf{A} 是正定的, \mathbf{B} 是可逆矩阵, 则 $\mathbf{B}'\mathbf{A}\mathbf{B}$ 是正定的;
 - d) 如果 \mathbf{A} 是一个 $n \times m$ 矩阵, 且 $n > m$, $r(\mathbf{A}) = m$, 则 $\mathbf{A}'\mathbf{A}$ 是正定的, $\mathbf{A}\mathbf{A}'$ 是非负定矩阵。
- 5) 对称矩阵 (symmetric matrix): 如果 $\mathbf{A} = \mathbf{A}'$, 则 \mathbf{A} 称为对称矩阵。

§ 2.1.3 矩阵的迹 (trace)

一个 $n \times n$ 矩阵的迹被定义为它的对角线上的元素之和, 记为 $tr(\mathbf{A})$, 则 $tr(\mathbf{A}) = \sum_i^n a_{ii}$ 。如下结论是显然的。

- 1) $tr(\alpha\mathbf{A}) = \alpha tr(\mathbf{A}) \cdots (\alpha \text{ 是标量}) \cdots$ Special case : $tr(\mathbf{I}) = n$;
- 2) $tr(\mathbf{A}') = tr(\mathbf{A})$;
- 3) $tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$;

- 4) $tr(\mathbf{AB}) = tr(\mathbf{BA})$, 特例 $tr(\mathbf{A}'\mathbf{A}) = \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2$
- 5) 循环排列原则: $tr(\mathbf{ABCD}) = tr(\mathbf{BCDA}) = tr(\mathbf{CDAB}) = tr(\mathbf{DABC})$

Theorem 2.1.2 实对称矩阵 \mathbf{A} 的迹等于它的特征根之和。

因为 \mathbf{A} 是实对称矩阵, 故有在矩阵 \mathbf{C} , 使得 $\mathbf{C}'\mathbf{A}\mathbf{C} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$, 其中 $\mathbf{C}'\mathbf{C} = \mathbf{I}$, 因此 $\sum_{i=1}^n \lambda_i = tr(\mathbf{A}) = tr(\mathbf{C}'\mathbf{A}\mathbf{C}) = tr(\mathbf{A}\mathbf{C}'\mathbf{C}) = tr(\mathbf{A}\mathbf{I}) = tr(\mathbf{A})$

§ 2.1.4 矩阵的秩 (rank)

一个矩阵 \mathbf{A} 的行秩和列秩一定相等, 一个矩阵的秩就可以定义为它的行秩或列秩, 记为 $r(\mathbf{A})$, 不加证明, 我们给出如下结果:

- 1) $r(\mathbf{A}) = r(\mathbf{A}') \leq \min(\text{row}, \text{col})$
- 2) $r(\mathbf{A}) + r(\mathbf{A}) - n_1 \leq r(\mathbf{AB}) \leq \min(r(\mathbf{A}), r(\mathbf{B}))$, 其中 \mathbf{A} 、 \mathbf{B} 分别为 $m \times n_1$ 、 $n_1 \times n$ 矩阵, 特例: 如果 \mathbf{A} 、 \mathbf{B} 为 $n \times n$ 矩阵, 而且 $\mathbf{AB} = \mathbf{0}$, 则 $r(\mathbf{A}) + r(\mathbf{A}) \leq n$
- 3) $r(\mathbf{A}) = r(\mathbf{AA}') = r(\mathbf{A}'\mathbf{A})$, 其中是 $n \times n$ 的方阵
- 4) $r(\mathbf{AB}) \leq r(\mathbf{A}) + r(\mathbf{B})$
- 5) \mathbf{A} 设是 $n \times n$ 矩阵, 且 $\mathbf{A}^2 = \mathbf{I}$, 则 $r(\mathbf{A} + \mathbf{I}) + r(\mathbf{A} - \mathbf{I}) = n$
- 6) \mathbf{A} 设是 $n \times n$ 矩阵, 且 $\mathbf{A}^2 = \mathbf{A}$, 则 $r(\mathbf{A}) + r(\mathbf{A} - \mathbf{I}) = n$

§ 2.1.5 统计量的矩阵表示

向量可理解为特殊的矩阵。 $\vec{\mathbf{i}}$ 是一个其元素都为 1 的 n 维列向量, 即 $\vec{\mathbf{i}} = (1, 1, \dots, 1)$, 如果我们再假定 $\vec{\mathbf{x}}' = (X_1, X_2, \dots, X_n)$, 计量经济模型中的许多统计量就可以用矩阵的形式表示出来, 很方便进行数学推导。

显而易见, $\sum_{i=1}^n x_i = \vec{\mathbf{i}}' \cdot \vec{\mathbf{x}}$, $\sum_{i=1}^n x_i^2 = \vec{\mathbf{x}}' \cdot \vec{\mathbf{x}}$, 样本的均值与方差的矩阵表示如下:

- 1) 样本均值矩阵表示:

事实上 $\vec{i}'\vec{i} = n$ 即 $\frac{\vec{i}'\vec{i}}{n} = 1$, 而 $\vec{i}\vec{i}' = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \vdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$, $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\vec{i}' \cdot \vec{x}}{n}$

2) 样本方差矩阵表示:

易知: $\begin{pmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{pmatrix} = \vec{i}\bar{x} = \vec{i} \cdot \frac{\vec{i}' \cdot \vec{x}}{n} = \frac{\vec{i}\vec{i}'\vec{x}}{n}$ 。其中矩阵 $\frac{1}{n}\vec{i}\vec{i}'$ 是一个每个元素都为 $\frac{1}{n}$ 的 n 阶阵, 从而 $\begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} = (\vec{x} - \vec{i}\bar{x}) = \left(\vec{x} - \frac{\vec{i}\vec{i}'\vec{x}}{n} \right) = \left(I - \frac{\vec{i} \cdot \vec{i}'}{n} \right) \vec{x} \triangleq \mathbf{M}^0 \vec{x}$

矩阵 \mathbf{M}^0 的对角线上的元素为 $1 - \frac{1}{n}$, 非对角线的元素为 $-\frac{1}{n}$, 是一个对称矩阵。故样本方差:

$$\begin{aligned} S^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} (\vec{x} - \bar{x})' (\vec{x} - \bar{x}) \\ &= \frac{1}{n} \vec{x} \cdot \mathbf{M}^{0'} \mathbf{M}^0 \vec{x} = \frac{1}{n} \vec{x} \mathbf{M}^{02} \vec{x} = \frac{1}{n} \vec{x}' \mathbf{M}^0 \vec{x} \end{aligned}$$

Theorem 2.1.3 矩阵 \mathbf{M}^0 是幂等矩阵。

§ 2.1.5.1 矩阵的二次型与多元正态分布

1) 矩阵的二次型 (Quadratic Forms) 和线性变换 (linear transferring)

设 \mathbb{P} 是一数域, 一个系数在数域 \mathbb{P} 中的 $f(x_1, x_2, \cdots, x_n)$ 的二次齐次多项式

$$\begin{aligned} f(x_1, x_2, \cdots, x_n) &= a_{11}x_1^2 + 2a_{12}x_1x_2 + \cdots + 2a_{1n}x_1x_n \\ &\quad + a_{22}x_2^2 + \cdots + 2a_{2n}x_2x_n \\ &\quad \cdots \cdots \cdots \\ &\quad + a_{nn}x_n^2 \end{aligned} \tag{2.1.1}$$

称为数域 \mathbb{P} 上的一个 n 元二次型, 或者, 在不致引起混淆时简称二次型。例如

$$x_1^2 + x_1x_2 + 3x_1x_3 + 2x_2^2 + 4x_2x_3 + 3x_3^2$$

就是有理数域上的一个三元二次型, 为了以后讨论上的方便, 在(2.1.1)中, $x_ix_j (i < j)$ 的系数写在 $2a_ia_j$ 。而不简单地写成 a_ia_j 。

和在几何中一样, 在处理许多其它问题时也常常希望通过变量的线性替换简化有关的二次型, 为此, 我们引入

Definition 2.1 设 $x_1, \dots, x_n; y_1, \dots, y_n$ 是两组文字, 系数在数域 \mathbb{P} 中的一级关系式。

$$\begin{cases} x_1 = c_{11}y_1 + c_{12}y_2 + \dots + c_{1n}y_n \\ x_2 = c_{21}y_1 + c_{22}y_2 + \dots + c_{2n}y_n \\ \dots\dots\dots \\ x_n = c_{n1}y_1 + c_{n2}y_2 + \dots + c_{nn}y_n \end{cases} \quad (2.1.2)$$

称为由 x_1, \dots, x_n 到 y_1, \dots, y_n 的一个线性替换, 或简称线性替换, 如果系数行列式

$$|c_{ij}| \neq 0$$

那么线性替换 (2.1.2) 就称为非退化的。在讨论二次型时, 矩阵是一个有力的工具, 因此我们先把二次型与线性替换用矩阵来表示。

令 $a_j a_i = a_i a_j, i < j$ 。由于 $x_j x_i = x_i x_j$ 所以二次型 (2.1.2) 可以写成

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= a_{11}x_1^2 + a_{12}x_1x_2 + \dots + a_{1n}x_1x_n \\ &\quad + a_{21}x_2x_1 + a_{22}x_2^2 + \dots + a_{2n}x_2x_n \\ &\quad \dots\dots\dots \\ &\quad + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \dots + a_{nn}x_n^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j \end{aligned} \quad (2.1.3)$$

把 (2.1.3) 的系数排成一个 $n \times n$ 矩阵

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \quad (2.1.4)$$

(2.1.4) 就称为二次型 (2.1.3) 的矩阵, 因为 $a_{ij} = a_{ji}, i, j = 1, \dots, n$ 。所以 $\mathbf{A} = \mathbf{A}'$ 。我们把这样的矩阵称为对称矩阵, 因此二次型的矩阵都是对称的。

$$\text{令 } \mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ 于是, 二次型可以用矩阵的乘积表示出来,}$$

$$\begin{aligned}
f(x_1, x_2, \dots, x_n) &= \mathbf{X}' \mathbf{A} \mathbf{X} \\
&= (x_1, x_2, \dots, x_n) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\
&= (x_1, x_2, \dots, x_n) \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{pmatrix} \\
&= \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j
\end{aligned}$$

应该看到, 二次型 (2.1.1) 的矩阵 \mathbf{A} 的元素 $a_i a_j = a_j a_i$, 正是它的 $x_i x_j = x_j x_i$ 项的系数的一半, 因此二次型和它的矩阵是相互唯一决定的, 由此还能得到, 若二次型

$$f(x_1, x_2, \dots, x_n) = \mathbf{X}' \mathbf{A} \mathbf{X} = \mathbf{X}' \mathbf{B} \mathbf{X}$$

且 $\mathbf{A} = \mathbf{A}', \mathbf{B} = \mathbf{B}'$ 则 $\mathbf{A} = \mathbf{B}$ 。令:

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

于是线性替换 (2.1.2) 可以写成

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{or} \quad \mathbf{X} = \mathbf{C} \mathbf{Y}$$

我们知道, 经过一个非退化的线性替换, 二次型还是变成二次型, 现在就来看一下, 替换后的二次型与原来的二次型之间有什么关系, 也就是说, 找出替换后的二次的矩阵与原二次型的矩阵之间的关系。

$$\text{Suppose } f(x_1, x_2, \dots, x_n) = \mathbf{X}' \mathbf{A} \mathbf{X}, \quad \mathbf{A} = \mathbf{A}' \quad (2.1.5)$$

是一个二次型, 作非退化线性替换

$$\mathbf{X} = \mathbf{C} \mathbf{Y} \quad (2.1.6)$$

我们得到一个 y_1, y_2, \dots, y_n 的二次型 $\Rightarrow Y'BY$

现在来看矩阵 B 与 A 的关系。把(2.1.6)代入(2.1.5), 有

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= X'AX = (CY)'A(CY) = Y'C'ACY \\ &= Y'(C'AC)Y = Y'BY \end{aligned}$$

容易看出, 矩阵 $C'AC$ 也是对称的, 事实上,

$$(C'AC)' = C'A'C'' = C'AC$$

由此, 即得 $B = C'AC$ 这就是前后两个二次型的矩阵的关系, 与之相应, 我们引入

Definition 2.2 数域 \mathbb{P} 上 $n \times n$ 短阵 A, B 称为合同的, 如果有数域 \mathbb{P} 上可逆的 $n \times n$ 矩阵 C , 使 $B = C'AC$ 。

合同是矩阵之间的一个关系, 不难看出, 合同关系具有

- 1) 反身性: $A = E'AE$
- 2) 对称性: 由 $B = C'AC$ 即得 $B = (C^{-1})'AC^{-1}$
- 3) 传递性: 由 $A_1 = C_1'AC_1, A_2 = C_2'AC_2$ 即得 $A_2 = (C_1C_2)'A(C_1C_2)$

因之, 经过非退化的线性替换, 新二次型的矩阵与原二次型的矩阵是合同的。这样, 我们就把二次型的变换通过矩阵表示出来, 为以下的探讨提供了有力的工具。

最后指出, 在变换二次型时, 我们总是要求所作的线性替换是非退化的。从几何上看, 这一点是自然的, 因为坐标变换一定是非退化的, 一般地, 当线性替换。

$$X = CY$$

是非退化时, 由上面的关系即得

$$Y = C^{-1}X$$

这也是一个线性替换, 它把所得的二次型还原。这样就使我们从所得二次型的性质可以推知原来二次型的一些性质。

Theorem 2.1.4 定理: 若 A 是实对称矩阵, 则存在可逆矩阵 C , 满足: $C'AC = \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$

2) 多元正态分布

- a) 二元正态分布 直观上, 二元正态分布是两个正态随机变量的联合分布。如果两个随机变量 X_1 和 X_2 的联合密度函数为

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{\Sigma}{2}\right\}^{-1}$$

where $-\infty < x_1, x_2 < \infty, \quad \sigma_1 > 0, \quad \sigma_2 > 0, \quad -1 < \rho < 1$

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right]$$

我们称 X_1 和 X_2 服从二元正态分布。通过计算可得 X_1 和 X_2 的边缘分布分别为 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ 。上式中的参数 ρ 是 X_1 和 X_2 的相关系数。

如果 X_1 和 X_2 服从二元正态分布, 那么在给定 $X = x$ 的条件下 X_2 的条件分布也是正态的。它的条件密度函数为

$$f(x_2 | x_1) \sim N(b, \sigma_2^2(1-\rho^2)), \quad \text{where } b = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

条件均值 $b = \mathbb{E}(X_2 | X_1)$ 是 x_1 的线性函数。并且, 二元正态分布具有一个独特的性质, 那就是如果 $\rho = 0$, 那么 X_1 和 X_2 是相互独立的。这是由于当 $\rho = 0$ 时, 我们有 $f(x_2 | x_1)$ 。这对于一般的两个随机变量是不对的。

有时如果把联合概率密度函数写成矩阵的形式, 则从形式上来看就简单多了。记 $X' = (X_1, X_2)$, 那么二元正态概率密度函数可以写成如下的简单形式

$$f(x) = (2\pi)^{-1} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}$$

where $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}$

b) 多元正态分布

$g(x) = \frac{1}{2\pi^{\frac{n}{2}}\Sigma^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^n$ 这就是均值为 μ 协方差矩阵为 Σ 的多元正态分布, 记为 $\mathbf{X} \sim N(\mu, \Sigma)$

c) 多元正态分布的二次型的分布

如果 $\mathbf{X} \sim N(\mu, \Sigma)$, 那么

$$\mathbf{Y} = (\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu) \sim \chi_{(n)}^2$$

这里 n 是 \mathbf{X} 的维数。我们可以简单地证明这个结果。由于 Σ 是对称可逆矩阵，那么存在一个可逆的矩阵 \mathbf{A} ，使得 $\mathbf{A}\Sigma\mathbf{A}' = \mathbf{I}$ 。我们有 $\mathbf{AX} \sim N(\mathbf{A}\mu, \mathbf{I})$, $\mathbf{Z} = \mathbf{A}(\mathbf{X} - \mu) \sim N(0, \mathbf{I})$ 所以 $\mathbf{Y} = \mathbf{Z}'\mathbf{Z} = (\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu) \sim \chi^2_{(n)}$ 。

§ 2.1.6 幂等矩阵与二次型

幂等矩阵满足 $\mathbf{A}^2 = \mathbf{A}$ 的矩阵称为幂等矩阵。

幂等矩阵可以是对称的，也可以是非对称的，但在我们计量统计学中，所研究的幂等矩阵都是对称的。与幂等矩阵的有关的结论有：

1) 幂等矩阵的特征根要么是 1，要么是零。

Proof 2.1 设 λ 是 \mathbf{A} 的特征根，则 $\mathbf{AE} = \lambda\mathbf{E}$ ，同时 $\lambda\mathbf{E} = \mathbf{A} = \mathbf{A}^2 = \lambda^2\mathbf{E}$ ，故 $\lambda^2 = \lambda$ ，从而 $\lambda = 1$ 或 $\lambda = 0$

2) 唯一满秩的对称幂等矩阵是单位矩阵。

Proof 2.2 $\because \mathbf{A}^2 = \mathbf{A} \Rightarrow \mathbf{A}(\mathbf{I} - \mathbf{A}) = 0 \Rightarrow \mathbf{A}^{-1}\mathbf{A}(\mathbf{I} - \mathbf{A}) = 0 \Rightarrow \mathbf{I} = \mathbf{A}$

3) \mathbf{A} 是幂等矩阵，则 $\mathbf{I} - \mathbf{A}$ 也是幂等矩阵，且秩 $(\mathbf{A}) + \text{秩}(\mathbf{I} - \mathbf{A}) = n$ 。

4) 对称幂等矩阵的秩等于它的迹。

从而我们很容易知道 \mathbf{M}^0 的秩。因 \mathbf{M}^0 的每个对角元素都是 $1 - \frac{1}{n}$ 。因此 $\text{tr}(\mathbf{M}^0) = n \cdot \left(1 - \frac{1}{n}\right) = n - 1 = r(\mathbf{M}^0)$

5) nS_n^2 的服从 $\chi^2(n-1)$ 分布 (如果 $\mathbf{X}_i \sim N(0, \mathbf{I}), i = 1, \dots, n$)

这是因为: $nS_n^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \vec{x}\mathbf{M}^0\vec{x}$ 和 $r(\mathbf{M}^0) = n - 1$

6) $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ \mathbf{X} 是一个 $n \times m$ 的矩阵，秩 $(\mathbf{X}) = m$ ，则 \mathbf{M} 是幂等矩阵。

§ 2.1.7 微分及其矩阵的微分表示

1、 矩阵的微分 如果 $y = f(x_1, x_2, \dots, x_n)$ 或写成 $y = f(\mathbf{x})$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \partial y / \partial x_1 \\ \partial y / \partial x_2 \\ \vdots \\ \partial y / \partial x_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$$

二阶偏导数矩阵为

$$\frac{\partial^2 f^2(x)}{\partial x \partial x'} = \begin{bmatrix} \partial^2 y / \partial x_1 \partial x_1 & \partial^2 y / \partial x_1 \partial x_2 & \cdots & \partial^2 y / \partial x_1 \partial x_n \\ \partial^2 y / \partial x_2 \partial x_1 & \partial^2 y / \partial x_2 \partial x_2 & \cdots & \partial^2 y / \partial x_2 \partial x_n \\ \cdots & \cdots & \cdots & \cdots \\ \partial^2 y / \partial x_n \partial x_1 & \partial^2 y / \partial x_n \partial x_2 & \cdots & \partial^2 y / \partial x_n \partial x_n \end{bmatrix}$$

特别地, 如果 $y = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a} = \sum_{i=1}^n a_i x_i$, 那么

$$\frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}$$

同样地可得

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}'$$

如果 \mathbf{A} 是对称矩阵, 那么

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$$

思考题:

- 1、证明: $tr(\mathbf{A}\mathbf{A}') = \sum_{j=1}^n \sum_{i=1}^n a_{ij}^2$
- 2、证明矩阵 \mathbf{M}^0 是幂等矩阵。
- 3、如果 $L_1, L_2 \cdots L_n$ 的百分比变动较小 $\Delta L_1, \cdots \Delta L_n$

如果 $Y_1, Y_2 \cdots Y_m$ 中百分比变动较小 $\Delta Y_1, \cdots \Delta Y_m$ 则如下计算公式是否可行?

- a) $\Delta(L_1, L_2 \cdots L_n) \approx \sum_{i=1}^n \Delta L_i$
- b) $\Delta\left(\frac{Y_1 \cdots Y_m}{L_1 \cdots L_n}\right) \approx \sum_{i=1}^m \Delta Y_i - \sum_{i=1}^n \Delta L_i$

2、矩阵的分块 (partitioned matrix)

在表述一个矩阵的元素时——如构造一个方程组——将一些元素以子矩阵的形式进行分组有时是有用的, 例如, 我们可以写

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 5 \\ 2 & 9 & 3 \\ 8 & 9 & 6 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

\mathbf{A} 称为一个分块矩阵, 子矩阵的下标和矩阵中的元素的下标按同样方式定义, 一个普通的特殊情形是分块对角矩阵。其中 \mathbf{A}_{11} 和 \mathbf{A}_{22} 都是方阵。

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}$$

分块矩阵的加法和乘法

加法和乘法可以推广到分块矩阵，对一致的分块矩阵 A 和 B 有：

$$A + B = \begin{bmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \end{bmatrix} \quad (2.1.7)$$

$$\begin{aligned} AB &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix} \end{aligned} \quad (2.1.8)$$

其中所有矩阵必须适于所用运算，对于加法， A_{ij} 和 B_{ij} 的阶数必须相同；在乘法中，对所有的数对 i 和 j ， A_{ij} 的列数必须等于 B_{ij} 的行数，即矩阵相乘所必需的条件都要得到满足。两个经常遇到的情况是如下的形式：

$$\begin{aligned} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}' \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} &= \begin{bmatrix} A_1' & A_2' \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \\ &= [A_1' A_1 + A_2' A_2] \end{aligned} \quad (2.1.9)$$

$$\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}' \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11}' A_{11} & 0 \\ 0 & A_{22}' A_{22} \end{bmatrix} \quad (2.1.10)$$

分块矩阵的行列式

类似于对角矩阵的行列式，分块对角矩阵的行列式可以得到

$$\begin{vmatrix} A_{11} & 0 \\ 0 & A_{22} \end{vmatrix} = |A_{11}| \cdot |A_{22}| \quad (2.1.11)$$

一个一般的 2×2 分块矩阵的结果为：

大于 2×2 分块矩阵的结果极其繁琐，且在我们的工作中也不必要。

分块矩阵的逆

分块对角矩阵的逆是：

$$\begin{aligned} \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} &= |A_{22}| \cdot |A_{11} - A_{12} A_{22}^{-1} A_{21}| \\ &= |A_{11}| \cdot |A_{22} - A_{21} A_{11}^{-1} A_{12}| \end{aligned} \quad (2.1.12)$$

这可以最简单地用逆去乘 A 来证实。由于计算的对称性，左上块可以写作：

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix} \quad (2.1.13)$$

这可由直接相乘证实。对一般的 2×2 分块矩阵，分块逆的一个形式是：

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} \left(\mathbf{I} + \mathbf{A}_{12} \mathbf{F}_2 \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \right) & -\mathbf{A}_{11}^{-1} \mathbf{A}_{12} \mathbf{F}_2 \\ -\mathbf{F}_2 \mathbf{A}_{21} \mathbf{A}_{11}^{-1} & \mathbf{F}_2 \end{bmatrix} \quad (2.1.14)$$

where $\mathbf{F}_2 = (\mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12})^{-1}$

这可以最简单地用逆去乘 \mathbf{A} 来证实。由于计算的对称性，左上块可以写作：

$$\mathbf{F}_1 = (\mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21})^{-1}$$

对均值的偏差

上述内容的一个有用的应用是如下的计算：假设我们从一个 n 个元素的列向量 \mathbf{x} 开始。且令

$$\mathbf{A} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} = \begin{bmatrix} \mathbf{i}'\mathbf{i} & \mathbf{i}'\mathbf{x} \\ \mathbf{x}'\mathbf{i} & \mathbf{x}'\mathbf{x} \end{bmatrix}$$

我们关心的是 \mathbf{A}^{-1} 中的右下角元素，根据(2.1.14)中 \mathbf{F}_2 的定义，这将是

$$\begin{aligned} \mathbf{F}_2 &= \left[\mathbf{x}'\mathbf{x} - (\mathbf{x}'\mathbf{i})(\mathbf{i}'\mathbf{i})^{-1}(\mathbf{i}'\mathbf{x}) \right]^{-1} \\ &= \left\{ \mathbf{x}' \left[\mathbf{I}\mathbf{x} - \mathbf{i} \left(\frac{1}{n} \right) \mathbf{i}'\mathbf{x} \right] \right\}^{-1} \\ &= \left\{ \mathbf{x}' \left[\mathbf{I} - \left(\frac{1}{n} \right) \mathbf{i}'\mathbf{i} \right] \mathbf{x} \right\}^{-1} \\ &= [\mathbf{x}'\mathbf{M}^0\mathbf{x}]^{-1} \end{aligned}$$

所以，逆矩阵中的右下角值是

$$(\mathbf{x}'\mathbf{M}^0\mathbf{x})^{-1} = \frac{1}{\sum_i (x_i - \bar{x})^2} = a_{22}$$

现在，假设以含有若干列的矩阵 \mathbf{X} 代替只有一列的 \mathbf{x} ，我们要求 $[\mathbf{Z}'\mathbf{Z}]^{-1}$ 中的右下块，这里 $\mathbf{Z} = [\mathbf{i}, \mathbf{X}]$ ，类似的结果是

$$\begin{aligned} (\mathbf{Z}'\mathbf{Z})^{-1} &= \left[\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'\mathbf{X} \right]^{-1} \\ &= [\mathbf{X}'\mathbf{M}^0\mathbf{X}]^{-1} \end{aligned} \quad (2.1.15)$$

这暗示着 $[\mathbf{Z}'\mathbf{Z}]^{-1}$ 的右下块, $K \times K$ 矩阵是第 jk 元素为 $\sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ 的 $K \times K$ 矩阵的逆, 这样, 当一个数据矩阵含有一列 1 时, 平方和及交叉积矩阵的逆的元素将用原始数据以对其相对应列均值的离差的形式计算得出。

§ 2.2 分布函数 (Distribution function), 数学期望 (Expectation) 与方差 (Variance)

本节主要介绍概率及其分布函数, 数学期望, 方差等方面的基础知识。

§ 2.2.1 概率 (Probability)

1、 概率定义 (Definition of Probability)

在自然界和人类社会中有两类不同的现象, 一类是决定性现象, 其特征是在一定条件必然会发生的现象; 另一类是随机现象, 其特征是在基本条件不变的情况下, 观察到或试验的结果会不同。换句话说, 就个别的试验或观察而言, 它会时而出现这种结果, 时而出现那样结果, 呈现出一种偶然情况, 这种现象称为随机现象。

随机现象有其偶然性的一面, 也有其必然性的一面, 这种必然性表现为大量试验中随机事件出现的频率的稳定性, 即一个随机事件出现的频率常在某固定的常数附近变动, 这种规律性我们称之为统计规律性。频率的稳定性说明随机事件发生可能性大小是随机事件本身固定的, 不随人们意志而改变的一种客观属性, 因此可以对它进行度量。

对于一个随机事件 A , 用一个数 $P(A)$ 来表示该事件发生的可能性大小, 这个数 $P(A)$ 就称为随机事件 A 的概率, 因此, 概率度量了随机事件发生的可能性的的大小。

对于随机现象, 光知道它可能出现什么结果, 价值不大, 而指出各种结果出现的可能性的的大小则具有很大的意义。有了概率的概念, 就使我们能对随机现象进行定量研究, 由此建立了一个新的数学分支——概率论。

概率的定义:

定义在事件域 F 上的一个集合函数 P 称为概率, 如果它满足如下三个条件:

- i、 $P(A) \geq 0$, 对一切 $A \in \mathcal{F}$
- ii、 $P(\Omega) = 1$
- iii、 若 $A_i \in \mathcal{F}$, $i = 1, 2, \dots$, 且两两互不相容, 则 $P(\sum_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

性质 (iii) 称为可列可加性 (conformable addition) 或完全可加性。

推论 1: 对任何事件 A 有 $P(\bar{A}) = 1 - P(A)$

推论 2: 不可能事件的概率为 0, 即 $P(\phi) = 0$

推论 3: $P(A \cup B) = P(A) + P(B) - P(AB)$

2、条件概率 (Conditional Probability) 如果 $P(B) > 0$, 记 $P(A | B) = \frac{P(AB)}{P(B)}$ 称 $P(A | B)$ 为在事件 B 发生的条件下事件 A 发生的条件概率。

转化后有: $P(AB) = P(A) \cdot P(B | A) = P(B) \cdot P(A | B)$ 如果 $P(A) > 0$, 称为概率的乘法原理。

推广后的乘法原理: $P(A_1 A_2 \cdots A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 A_2) \cdots P(A_n | A_1 A_2 \cdots A_{n-1})$

3、全概率公式与贝叶斯 (Bayes) 公式

设事件 $A_1, A_2, \cdots, A_n, \cdots$ 是样本空间的一个分割, 即 $A_i A_j = \phi, i \neq j$, 而且: $\sum_{i=1}^{\infty} A_i = \Omega$ 从而 $B = \sum_{i=1}^{\infty} A_i B$ 这里 $A_i B$ 也两两互不相容。

则 $P(B) = \sum_{i=1}^{\infty} P(A_i B) = \sum_{i=1}^{\infty} P(A_i) \cdot P(B | A_i)$ 这个公式称为全概率公式。

由于 $P(A_i B) = P(B)P(A_i | B) = P(A_i)P(B | A_i)$ 。故 $P(A_i | B) = \frac{P(A_i)P(B | A_i)}{P(B)}$

再利用全概率公式即得:

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{\sum_{i=1}^{\infty} P(A_i)P(B | A_i)}$$

这个公式称为贝叶斯公式。贝叶斯公式在概率论和数理统计中有着多方面的应用, 假定 A_1, A_2, \cdots , 是导致试验结果的“原因”, $P(A_i)$ 称为先验概率, 它反映了各种“原因”发生的可能性大小, 一般是以往经验的总结, 在这次试验前已经知道, 现在若试验产生了事件 B, 这个信息将有助于探讨事件发生的“原因”, 条件概率 $P(A | B)$ 称为后验概率, 它反映了试验之后对各种“原因”发生的可能性大小的新知识。

4、事件 (Random event) 独立性 (Independence)

1)、两个事件的独立性

Definition 2.3 对事件 A 及 B, 若 $P(AB) = P(A)P(B)$ 则称它们是统计独立的, 简称独立的。

推论 1: 若事件独立, 且 $P(B) > 0$ 则 $P(A | B) = P(A)$

Proof 2.3 由条件概率定义 $P(A | B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$

因此, 若事件 A, B 相互独立, 由 A 关于 B 的条件概率等于无条件概率 $P(A)$, 这表示 B 的发生对于事件 A 是否发生没有提供任何消息, 独立性就是把这种关系从数学上加以严格定义。

推论 2: 若事件 A 与 B 独立, 则下列各对事件也相互独立: $\{\bar{A}, B\}, \{A, \bar{B}\}, \{\bar{A}, \bar{B}\}$

Proof 2.4

$$\begin{aligned}
 P(\bar{A}B) &= P(B - AB) = P(B) - P(AB) \\
 &= P(B) - P(A)P(B) = P(B)[1 - P(A)] \\
 &= P(\bar{A})P(B)
 \end{aligned} \tag{2.2.1}$$

所以 \bar{A} 与 B 相互独立, 由它立刻推出 \bar{A} 与 \bar{B} 相互独立, 由 $\bar{\bar{A}} = A$ 又推出 A 与 \bar{B} 相互独立。

2)、多个事件的独立性

Definition 2.4 对 n 个事件 A_1, A_2, \dots, A_n , 若对于所有可能的组合 $1 \leq i < j < \dots \leq n$ 成立着

$$\left. \begin{aligned}
 P(A_i A_j) &= P(A_i) P(A_j) \\
 P(A_i A_j A_k) &= P(A_i) P(A_j) P(A_k) \\
 &\dots\dots \\
 P(A_1 A_2 \dots A_n) &= P(A_1) P(A_2) \dots P(A_n)
 \end{aligned} \right\}$$

则称 A_1, A_2, \dots, A_n , 相互独立。这里第一行有 $\binom{n}{2}$ 这里第二行有 $\binom{n}{3}$ 等等, 因此共应满足 $\binom{n}{2} + \binom{n}{3} + \dots + \binom{n}{n} = 2^n - n - 1$ 个等式。

§ 2.2.2 随机变量 (Random Variable) 和概率分布函数 (Probability Distribution Function)

1、随机变量 (Random Variable)

如果 A 为某个随机事件, 则一定可以通过如下示性函数使它与数值发生联系:

$$I_A = \begin{cases} 1, & \text{如果 } A \text{ 发生} \\ 0, & \text{如果 } A \text{ 不发生} \end{cases}$$

这样试验的结果就能有一个数 X 来表示, 这个数是随着试验的结果的不同而变化, 也即它是样本点的一个函数, 这种量以后称为随机变量, 随机变量可分为离散型随机变量和连续型随机变量。

2、概率分布函数 (P.d.f = probability density function)

称 $F(x) = P\{X < x\}$, $-\infty < x < \infty$ 为随机变量 X 的分布函数 CDF, 对于连续型随机变量, 存在可能函数 $f(x)$, 使

$$F(x) = \int_{-\infty}^x f(x) dX, \quad f(x) \text{ 称为随机变量的 (分布) 密度函数 (density function)}.$$

3、随机向量 (Random Vector) 及其分布

在有些随机现象中, 每次试验的结果不能只用一个数来描述, 而要同时用几个数来描述。试验的结果将是一个向量 (X_1, X_2, \dots, X_n) , 称 n 维随机向量。

随机向量的联合分布函数也有离散型与连续型的分别, 在离散型场合, 概率分布集中在有限或可列个点上, 多项分布, 就是一个例子; 在连续型场合, 存在着非负函数 $f(x_1, x_2, \dots, x_n)$ 使

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, \dots, y_n) dy_1 \cdots dy_n$$

这里的 $f(x_1, x_2, \dots, x_n)$ 称为密度函数, 满足如下两个条件:

$$\begin{aligned} f(x_1, \dots, x_n) &\geq 0 \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_1 \cdots dx_n &= 1 \end{aligned}$$

一般地, 若 (ξ, η) 是二维随机向量, 其分布函数为 $F(x, y)$, 我们能由 $F(x, y)$ 得出 ξ 或 η 的分布函数, 事实上,

$$F_1(x) = P\{\xi < x\} = P\{\xi < x, \eta < \infty\} = F(x, +\infty)$$

同理 $F_2(y) = P\{\eta < y\} = F(+\infty, y)$, $F_1(x)$ 及 $F_2(y)$ 称为 $F(x, y)$ 的边际分布函数 (Marginal Distribution Function)。

Example 2.1 若 $F(x, y)$ 是连续型分布函数, 有密度函数 $f(x, y)$, 那么 $f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$ 。因此 $F_1(x)$ 是连续型分布函数, 其密度函数为 $f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$ 同理 $F_2(x)$ 是连续型分布函数, 其密度函数为 $f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx$ 。

[二元正态分布] 函数

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} \cdot \left[\frac{(x-a)^2}{\sigma_1^2} - \frac{2r(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2} \right] \right\}$$

这里 $a, b, \sigma_1, \sigma_2, r$ 为常数, $\sigma_1 > 0$, $\sigma_2 > 0$, $|r| < 1$, 称为二元正态分布密度函数。

Theorem 2.2.1 (二元正态分布的边际分布仍为正态分布)

$$N\left(\mu_2 + \rho \frac{\sigma_2^2}{\sigma_1^2} (x - \mu_1), \sigma_2^2 (1 - \rho^2)\right)$$

其均值 $\mu = \mu_2 + \rho \frac{\sigma_2^2}{\sigma_1^2} (x - \mu_1)$ 是 x 的线性函数, 这个结论在一些统计问题中很重要。

条件分布 (Conditional Distribution)

离散型：若已知 $\xi = x_i$, ($p_1(x_i) > 0$) 则事件 $\{\eta = y_j\}$ 的条件概率为

$$P\{\eta = y_j \mid \xi = x_i\} = \frac{P\{\xi = x_i, \eta = y_j\}}{P\{\xi = x_i\}} = \frac{P(x_i, y_j)}{p_1(x_i)}$$

这式子定义了随机变量 η 关于随机变量 ξ 的条件分布。

连续型：在给定 $\xi = x$ 条件下, η 的分布密度函数为 $f(y \mid x) = \frac{f(x, y)}{f_1(x)}$ 。同理可行在给定 $\eta = y$ 的条件下, ξ 的分布密度函数为 $f(y \mid x) = \frac{f(x, y)}{f_2(y)}$ 。这里当然也要求 $f_2(y) \neq 0$ 。

4、随机变量的独立性

Definition 2.5 设 ξ_1, \dots, ξ_n 为 n 个随机变量, 若对于任意的 x_1, \dots, x_n 成立

$$P\{\xi_1 < x_1, \dots, \xi_n < x_n\} = P\{\xi_1 < x_1\} \cdots P\{\xi_n < x_n\} \quad (2.2.2)$$

则称 ξ_1, \dots, ξ_n 是相互独立的。若 ξ_i 的分布函数为 $F(\xi_1, \dots, \xi_n)$ 则(2.2.2)等价于对一切 x_1, \dots, x_n 成立 $F(x_1, \dots, x_n) = F_1(x_1) \cdots F_n(x_n)$ 。在这种场合, 由每个随机变量的(边际)分布函数可以唯一地确定联合分布函数 (Joint Distribution Function)。

对于离散型随机变量, (2.2.2) 等价于任何一组可能取的值 (x_1, \dots, x_n) 成立

$$P\{\xi_1 = x_1, \dots, \xi_n = x_n\} = P\{\xi_1 = x_1\} \cdots P\{\xi_n = x_n\}$$

对于连续型随机变量, (2.2.2) 等价于任何一组可能取的值 (x_1, \dots, x_n) 成立

$$f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$$

这里 $f(x_1, \dots, x_n)$ 是联合分布密度函数 (Joint density function), 而 $f_i(x_i)$ 是各随机变量的密度函数。

此外, 注意到若 $\xi_1, \xi_2, \dots, \xi_n$ 相互独立, 则其中的任意 r ($2 \leq r < n$) 个随机变量也相互独立, 例如, 我们证明 $\xi_1, \xi_2, \dots, \xi_{n-1}$ 相互独立。

$$\begin{aligned} P\{\xi_1 < x_1, \dots, \xi_{n-1} < x_{n-1}\} &= P\{\xi_1 < x_1, \dots, \xi_{n-1} < x_{n-1}, \xi_n < \infty\} \\ &= P\{\xi_1 < x_1\} \cdots P\{\xi_{n-1} < x_{n-1}\} P\{\xi_n < \infty\} \\ &= P\{\xi_1 < x_1\} \cdots P\{\xi_{n-1} < x_{n-1}\} \end{aligned}$$

随机变量的独立性概念是概率论中最基本的概念之一, 也是最重要的概念之一。

5、随机向量变换 (Transformation) 及其分布

若 (ξ_1, \dots, ξ_n) 的密度函数为 $f(x_1, \dots, x_n)$, 求 $\eta_1 = f_1(\xi_1, \dots, \xi_n), \dots, \eta_n = f_n(\xi_1, \dots, \xi_n)$ 的分布, 这时有

$$\begin{aligned} G(y_1, \dots, y_n) &= P\{\eta_1 < y_1, \dots, \eta_n < y_n\} \\ &= \int \cdots \int f(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &\quad \text{where } f_1(x_1, \dots, x_n) < y_1 \\ &\quad \quad \quad f_m(x_1, \dots, x_n) < y_m \end{aligned} \quad (2.2.3)$$

若对 $f(x_1, \dots, x_n)$ 存在唯一的反函数 $x_i(y_1, \dots, y_n) = x_i, (i = 1, \dots, n)$, 且 (η_1, \dots, η_n) 的密度函数为 $q(y_1, \dots, y_n)$ 那么

$$\begin{aligned} G(y_1, \dots, y_n) &= \int \cdots \int q(u_1, \dots, u_n) du_1 \cdots du_n \\ &\quad \text{where } u_1 < y_1, \quad u_n < y_n \end{aligned} \quad (2.2.4)$$

比较(2.2.3)与(2.2.4)可知

$$q(y_1, \dots, y_n) = \begin{cases} f(x_1, \dots, x_n) |J|, & \text{若 } (y_1, \dots, y_n) \text{ 属于 } f_1, \dots, f_n \text{ 的值域} \\ 0, & \text{其它} \end{cases}$$

其中 J 为坐标变换的雅可比行列式 (Jacobian Determinant), 这里, 我们假定上述偏导数存在而且连续。

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial y_n} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

随机变量的函数的独立性

Theorem 2.2.2 若 ξ_1, \dots, ξ_n 是相互独立的随机变量, 则 $f_1(\xi_1), \dots, f_n(\xi_n)$ 也是相互独立的, 这里 $f_i (i = 1, \dots, n)$ 是任意的一元函数。

§ 2.2.3 数学期望及方差

1、数学期望 一般地, 如果 X 是随机变量, 它的概率密度函数为 $f(x)$, 那么它的期望值为

$$\mathbb{E}[X] = \begin{cases} \sum_x x f(x) & \text{当 } X \text{ 是离散型随机变量时} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{当 } X \text{ 是连续型} \end{cases}$$

在许多问题中我们不仅要知道 $E[X]$, 而且还想知道 X 的某个函数 $g(X)$ 的数学期望。

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x)f(x) & \text{当 } X \text{ 是离散型时} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{当 } X \text{ 是连续时} \end{cases}$$

我们可以用同样的方法定义多元随机变量的函数的数学期望。假设随机变量 X_1, X_2, \dots, X_n 的联合概率密度函数为 $f(x_1, x_2, \dots, x_n)$, $Y = g(X_1, X_2, \dots, X_n)$

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n$$

如果随机变量是离散的, 那么上面公式里的积分号用和号代替。

利用这个定义我们可以得到下列结果

如果 a_0, a_1, \dots, a_n 是常数, 那么

$$\mathbb{E}[a_0 + a_1 X_1 + \cdots + a_n X_n] = a_0 + a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n]$$

如果 X_1, X_2, \dots, X_n 是相互独立的随机变量, 那么

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mathbb{E}[X_1] \mathbb{E}[X_2] \cdots \mathbb{E}[X_n]$$

2、 方差 (Variance) 与协方差 (Covariance)

一个随机变量 X 的 r 阶中心矩被定义为 $\mathbb{E}[(X - \mu)^r]$ 记为 μ_r 。如果 $r = 2$, $\mathbb{E}[(X - \mu)^2]$ 被称为 X 的分布的方差或 X 的方差, 常常记为 σ^2 或 $\text{Var}(X)$ 。 σ^2 的正平方根 σ 被称为 X 的标准差。关于方差, 我们有一个有用的公式

$$\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

X 和 Y 之间的协方差, 记为 σ_{XY} 或 $\text{cov}(X, Y)$

$$\sigma_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

X 和 Y 之间的协方差是对它们之间的相关性的一个测度。如果 X 和 Y 是相互独立的, 那么 $\text{cov}(X, Y) = 0$ 。这导致下面的相关系数的定义, X 和 Y 之间的相关系数记为 ρ_{XY} 被定义为

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$$

由这个定义, ρ_{XY} 的取值一定在 -1 和 1 之间。如果 X 和 Y 是相互独立的, 那么 $\rho_{XY} = 0$ 。如果 $Y = aX + b$, 这里 a, b 是不等于 0 的常数, 那么 $\|\rho_{XY}\| = 1$, 此时, 我们说 X 和 Y 是完全相关的。 X 和 Y 的值越接近线性关系, $\|\rho_{XY}\|$ 值接近 1。

利用这些定义, 我们可以得到下面的结果: 如果 a_0, a_1, \dots, a_n 是常数, X_1, X_2, \dots, X_n 是相互独立的随机变量, 那么

$$\text{var}[a_0 + a_1 X_1 + \dots + a_n X_n] = \sum a_i^2 \text{var}(X_i) + 2 \sum_{i < j} a_i a_j \text{cov}(X_i, X_j)$$

特别地, 有

$$\text{var}(a_0 + a_1 X_1) = a_1^2 \text{var}(X_1)$$

$$\text{var}(X_1 \pm X_2) = \text{var}(X_1) + \text{var}(X_2) \pm 2 \text{cov}(X_1, X_2)$$

3、随机向量的协方差矩阵

对于随机向量而言, 我们可以相似地定义它的期望和协方差矩阵。用 \mathbf{X} 表示随机变量组成的向量, 即

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

假设 $\mathbb{E}(X_i) = \mu_i, \text{var}(X_i) = \sigma_i^2, \text{cov}(X_i, X_j) = \sigma_{ij}$ 那么 \mathbf{X} 的期望值为

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \boldsymbol{\mu}$$

也即是一个随机向量的期望值等于它的各个分量的期望值组成的向量。我们定义一个随机向量 \mathbf{X} 的协方差矩阵 (Covariance Matrix) 如下:

$$\begin{aligned} \text{cov}(\mathbf{X}) &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])'] \\ &= \mathbb{E} \begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_n - \mu_n) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_n - \mu_n) \\ \cdots & \cdots & \cdots & \cdots \\ (X_n - \mu_n)(X_1 - \mu_1) & (X_n - \mu_n)(X_2 - \mu_2) & \cdots & (X_n - \mu_n)^2 \end{bmatrix} \\ &= \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{var}(X_n) \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} \end{aligned}$$

\mathbf{X} 的协方差矩阵常常记为 Σ_x , 它是一个正定矩阵, 如下是证明: 对于任意的不为零的向量 $\mathbf{a}' = (a_1, a_2, \dots, a_n)$ 。我们构造一个变量 $\mathbf{Y} = \mathbf{a}'\mathbf{X}$ 那么 \mathbf{Y} 的方差 $\text{Var}(\mathbf{Y}) = \text{Var}(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\Sigma_x\mathbf{a} \geq 0$ 即证明了 Σ_x 是非负定的。

§ 2.3 条件分布 (Conditional Distribution)、条件数学期望 (Conditional Expectation) 及其条件方差 (Conditional Variance)

条件均值 (Conditional Mean) 是条件分布的均值, 其定义为:

$$\mathbb{E}[y | x] = \begin{cases} \int_y y f(y | x) dy & \text{若 } y \text{ 是连续的,} \\ \sum_y y f(y | x) & \text{若 } y \text{ 是离散的} \end{cases}$$

条件均值函数 $\mathbb{E}[y | x]$ 的回归对称称为 y 对 x 的回归。

条件方差 (Conditional Variance) 是条件分布的方差:

$$\begin{aligned} \text{Var}[y | x] &= \mathbb{E}[(y - \mathbb{E}[y | x])^2 | x] \\ &= \int_y (y - \mathbb{E}[y | x])^2 f(y | x) dy \\ \text{或} \quad &\sum_y (y - \mathbb{E}[y | x])^2 f(y | x) \quad (\text{离散时}) \end{aligned}$$

利用下式可以简化计算

$$\text{Var}[y | x] = \mathbb{E}[y^2 | x] - (\mathbb{E}[y | x])^2$$

并且有: $\mathbb{E}[y] = \mathbb{E}_x[\mathbb{E}[y | x]]$ 。记号 $\mathbb{E}_x[\cdot]$ 表示对 \mathbf{X} 的值的期望。

几个重要的公式:

$$1. \mathbb{E}(XY | X) = X\mathbb{E}(Y | X)$$

思考: $\mathbb{E}[g(X)Y | X] = g(X)\mathbb{E}(Y | X)$ 是否成立? ?

$$2. \mathbb{E}(XY) = \mathbb{E}[X\mathbb{E}(Y | X)]$$

$$3. \text{方差分解公式 (Decomposition of Variance)}$$

推导: 分两步, 先证明

$$i) \mathbb{E}(Y | X) \text{ 和 } Z = Y - \mathbb{E}(Y | X) \text{ 是不相关的, 即 } \text{cov}(\mathbb{E}(Y | X), Z) = 0$$

这是因为：

$$\begin{aligned}\mathbb{E}(Z | X) &= \mathbb{E}\{[Y - \mathbb{E}(Y | X)] | X\} \\ &= \mathbb{E}(Y | X) - \mathbb{E}(Y | X) = 0, \\ \text{从而 } \mathbb{E}Z &= \mathbb{E}[\mathbb{E}(Z | X)] = 0\end{aligned}$$

进而有 $\text{cov}[\mathbb{E}(Y | X), Z] = \mathbb{E}[Z\mathbb{E}(Y | X)]$

我们考察 $\mathbb{E}\{[Z\mathbb{E}(Y | X)] | X\} = \mathbb{E}(Y | X)\mathbb{E}(Z | X) = 0$

$$\therefore \mathbb{E}[Z\mathbb{E}(Y | X)] = \mathbb{E}\{\mathbb{E}[Z\mathbb{E}(Y | X) | X]\} = 0$$

ii) 对于任意 Y 有： $Y = Y - \mathbb{E}(Y | X) + \mathbb{E}(Y | X) = Z + \mathbb{E}(Y | X)$ ，因为 X 与 $\mathbb{E}(Y | X)$ 是不相关，故

$$\text{Var}(Y) = \text{Var}[Y - \mathbb{E}(Y | X)] + \text{Var}[\mathbb{E}(Y | X)]$$

$$\begin{aligned}\text{Var}(Y - \mathbb{E}(Y | X)) &= \mathbb{E}(Y - \mathbb{E}(Y | X))^2 \\ &= \mathbb{E}\{\mathbb{E}[Y - \mathbb{E}(Y | X)]^2 | X\} \\ &= \mathbb{E}[\text{Var}_x(Y | X)]\end{aligned}$$

我们得到方差分解公式：

$$\text{Var}(Y) = \text{Var}[\mathbb{E}(Y | X)] + \mathbb{E}[\text{Var}_x(Y | X)]$$

方差分解结果表明，在双变量分布中，y 的变差出自两个来源：

1. 由于 $\mathbb{E}[y | x]$ 随 x 变化的事实所产生的变差为回归方差 (Regression Variance)：

$$\text{回归方差} = \text{Var}_x[\mathbb{E}[y | x]]$$

2. 由于在每一条件分布中，y 都围绕条件均值变化而产生的变差为残差方差 (Residual Variance)：

$$\text{残差方差} = \text{Var}_x[\text{var}[y | x]]$$

这样 $\text{Var}[y] = \text{回归方差} + \text{残差方差}$ 。

由方差分解公式，我们得到 $\text{Var}[\mathbb{E}(y | x)] \leq \text{Var}[y]$ ，这个是非常重要的公式，它常被应用到寻求最小方差估计量的方法中。我们可以看一个实际的例子。

Example 2.2 设 X 和 Y 服从二元正态分布联合分布, 我们已经知道, 在给定 X 的条件下, 其条件分布仍然是正态分布, 并且 $\mathbb{E}(Y | X) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (X - \mu_1)$, 并且 $\mathbb{E}[\mathbb{E}(Y | X)] = \mu_2 = \mathbb{E}Y$, 然而

$$\begin{aligned} \text{Var}(\mathbb{E}(Y | X)) &= \mathbb{E}[E(Y | X) - \mu_2]^2 = \mathbb{E}\left(\rho \frac{\sigma_2}{\sigma_1} (X - \mu_1)\right)^2 \\ &= \rho^2 \frac{\sigma_2^2}{\sigma_1^2} \mathbb{E}(X - \mu_1)^2 = \rho^2 \sigma_2^2 \end{aligned}$$

在 $-1 < \rho < 1$ 条件下, $\sigma_2^2 > \rho^2 \sigma_2^2$ 。满足方差分解公式, 并且我们很容易知道, $\mathbb{E}(\text{Var}_x(Y | X)) = \sigma_2^2 - \rho^2 \sigma_2^2 = \sigma_2^2 (1 - \rho^2)$ 。

§ 2.3.1 极限分布理论 (Limit Distribution Theory)

1 几个极限的定义

1) 分布函数的弱收敛 (Weak Convergence of the Distribution Function)

Definition 2.6 对于分布函数列 $\{F_n(x)\}$, 如果存在一个非降函数 $F(x)$ 使 $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ 。在 $F(x)$ 的每一连续点上都成立, 则称 $F_n(x)$ 弱收敛于 $F(x)$, 并记为 $F_n(x) \xrightarrow{w} F(x)$, 中心极限定理就是一个分布函数弱收敛的例子。

2) 随机变量的收敛性 (Convergence of the Random Variable)

概率论中的极限定理研究的是随机变量序列的某种收敛性, 对随机变量收敛性的不同定义将导致不同的极限定理, 而随机变量的收敛性的确可以有各种不同的定义, 理解这些不同的极限定义, 对于我们分析线性回归的大样本结果很重要。现在就来讨论这个问题。

a) 依分布收敛 (Convergence in Distribution)

分布函数弱收敛的讨论启发我们引进如下定义。

Definition 2.7 (依分布收敛) 设随机变量 ξ_n, ξ 的分布函数分别为 $F_n(x)$ 及 $F(x)$ 如果 $F_n(x) \xrightarrow{w} F(x)$ 则称 ξ_n 依分布收敛于 ξ , 并记为 $\xi_n \xrightarrow{L} \xi$

b) 依概率收敛 (Convergence in Probability)

Definition 2.8 (依概率收敛) 如果 $\lim_{n \rightarrow \infty} P\{|\xi_n - \xi| \geq \varepsilon\} = 0$ 对任意的 $\varepsilon > 0$ 成立, 则称 ξ_n 依概率收敛于 ξ , 并记为 $\xi_n \xrightarrow{P} \xi$

c) r -阶收敛

Definition 2.9 (r -阶收敛) 设对随机变量 ξ_n 及 ξ 有 $\mathbb{E}|\xi_n|^r < \infty$, $\mathbb{E}|\xi|^r < \infty$, 其中 $r > 0$ 为常数, 如果 $\lim_{n \rightarrow \infty} \mathbb{E}|\xi_n - \xi|^r = 0$, 则称 $\{\xi_n\}$ r -阶收敛于 ξ , 并记为 $\xi_n \xrightarrow{r} \xi$

下面定理揭示了 r -阶收敛与依概率收敛的关系。

Theorem 2.3.1 $\xi_n \xrightarrow{r} \xi \implies \xi_n \xrightarrow{P} \xi$

3) 极限的应用——贝努里分布与泊松分布

a) 近似计算在 n 次贝努里试验中正好出现 k 次成功的概率 $b(k; n, p)$:

$$b(k; n, p) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \dots, n$$

其中 $q = 1 - p$ 。 $b(k; n, p)$, $k = 0, 1, 2, \dots, n$ 称为二项分布。

在很多应用问题中, 我们常常遇到这样的贝努里试验, 其中, 相对地说 n 大, p 小, 而乘积 $\lambda = np$ 大小适中, 在这种情况下, 有一个便于使用的近似公式。

Theorem 2.3.2 (泊松) 在贝努里试验中, 以 pn 代表事件 A 在试验中出现的概率, 它与试验总数 n 有关, 如果 $np_n \rightarrow \lambda$, 则当 $n \rightarrow \infty$ 时, $b(k; n, P_n) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}$

若 $X_1, X_2, \dots, X_n, \dots$ 是一串相互独立相同分布的随机变量序列, 且 $\mathbb{E}X_k = \mu, \text{Var}(X_k) = \sigma^2$

我们来讨论标准化随机变量和 $\xi_n = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^n (X_k - \mu)$ 的极限分布。林德贝格与勒维 (Lindeberg and Levy) 建立了下列中心极限定理。

Theorem 2.3.3 (林德贝格-勒维) 若 $0 < \sigma^2 < \infty$, 则 $\lim_{n \rightarrow \infty} P\{\xi_n < x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$

2 契比雪夫 (Chebyshevs Inequality) 不等式

对于任何具有有限方差的随机变量 X , 都有

$$P\{|X - \mathbb{E}X| \geq \varepsilon\} \leq \frac{\text{Var}(X)}{\varepsilon^2}, \quad \text{其中 } \varepsilon \text{ 是任一正数} \quad (2.3.1)$$

Proof 2.5 若 $F(x)$ 是 X 的分布函数, 则显然有

$$\begin{aligned} P\{|X - \mathbb{E}X| \geq \varepsilon\} &= \int_{|x - \mathbb{E}X| \geq \varepsilon} dF(x) \\ &\leq \int_{|x - \mathbb{E}X| \geq \varepsilon} \frac{(x - \mathbb{E}X)^2}{\varepsilon^2} dF(x) \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{\infty} (x - \mathbb{E}X)^2 dF(x) = \frac{\text{Var}(X)}{\varepsilon^2} \end{aligned} \quad (2.3.2)$$

这就证得了不等式(2.3.1), 有时把(2.3.1)改写成

$$P\{|X - \mathbb{E}X| < \varepsilon\} \geq 1 - \frac{\text{Var}(X)}{\varepsilon^2}, \quad \text{or } P\left\{\left|\frac{X - \mathbb{E}X}{\sqrt{\text{Var}(X)}}\right| \geq \delta\right\} \leq \frac{1}{\delta^2} \quad (2.3.3)$$

契比雪夫不等式利用随机变量 X 的数学期望 $\mathbb{E}(X)$ 及方差 $\text{Var}(X) = \sigma^2$ 对 X 的概率分布进行估计。例如(2.3.3)断言不管 X 的分布是什么, X 落在 $(\mathbb{E}X - \sigma\delta, \mathbb{E}X + \sigma\delta)$ 中的概率不小于 $1 - \frac{1}{\delta^2}$, 因为契比雪夫不等式只利用数学期望及方差就描述了随机变量的变化情况, 因此它在理论研究及实际应用中很有价值。

3 大数定律

Definition 2.10 若 $\xi_1, \xi_2, \dots, \xi_n, \dots$ 是随机变量序列, 令 $\eta_n = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}$ 。如果存在这样的一个常数序列 $a_1, a_2, \dots, a_n, \dots$, 对任意的 $\varepsilon > 0$ 恒有

$$\lim_{n \rightarrow \infty} P\{|\eta_n - a_n| < \varepsilon\} = 1$$

Definition 2.11 (契比雪夫大数定律) 设 $X_1, X_2, \dots, X_n, \dots$ 是由两两不相关的随机变量所构成的序列, 每一随机变量都有有限的方差, 并且它们有公共上界 C , 即

$$\text{Var}(X_1) \leq c, \text{Var}(X_2) \leq c, \dots, \text{Var}(X_n) \leq c, \dots$$

则对任意的 $\varepsilon > 0$ 恒有

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{1}{n} \sum_{k=1}^n X_k - \frac{1}{n} \sum_{k=1}^n \mathbb{E}X_k\right| < \varepsilon\right\} = 1 \quad (2.3.4)$$

Proof 2.6 因为 $\{\xi_k\}$ 两两不相关, 故

$$\text{Var}\left(\frac{1}{n} \sum_{k=1}^n X_k\right) = \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) \leq \frac{C}{n}$$

再由契比雪夫不等式得到

$$\begin{aligned} P\left\{\left|\frac{1}{n} \sum_{k=1}^n X_k - \frac{1}{n} \sum_{k=1}^n \mathbb{E}X_k\right| < \varepsilon\right\} &\geq 1 - \frac{\text{Var}\left(\frac{1}{n} \sum_{k=1}^n X_k\right)}{\varepsilon^2} \geq 1 - \frac{C}{n\varepsilon^2} \\ \therefore 1 &\geq P\left\{\left|\frac{1}{n} \sum_{k=1}^n X_k - \frac{1}{n} \sum_{k=1}^n \mathbb{E}X_k\right| < \varepsilon\right\} \geq 1 - \frac{C}{n\varepsilon^2} \end{aligned}$$

于是, 当 $n \rightarrow \infty$ 时有(2.3.4)。That's it.

Definition 2.12 (贝努里大数定律) 设 μ_n 是 n 次贝努里试验中事件 A 出现的次数, 而 p 是事件 A 在每次试验中出现的概率, 则对任意 $\varepsilon > 0$, 都有

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right\} = 1$$

Proof 2.7 定义随机变量 $X_i = \begin{cases} 1, & \text{第 } i \text{ 次试验出现 } A \\ 0, & \text{第 } i \text{ 次试验不出现 } A \end{cases}$

$$\mathbb{E}X_i = p, \quad \text{Var}(X_i) = pq \leq \frac{1}{4}$$

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^n X_k - \frac{1}{n} \sum_{k=1}^n \mathbb{E}X_k &= \frac{\mu_n}{n} - p \\ P\left\{\left|\frac{\mu_n}{n} - p\right| \geq \varepsilon\right\} &\leq \frac{1}{\varepsilon^2} \text{Var}\left(\frac{\mu_n}{n}\right) = \frac{1}{n\varepsilon^2} \text{Var}(X_i) \leq \frac{1}{4n\varepsilon^2} \end{aligned}$$

贝努里大数定律建立了在大量重复独立试验中事件出现频率的稳定性，正因为这种稳定性，概率的概念才有客观意义，贝努里大数定律还提供了通过试验来确定事件概率的方法，既然频率 $\frac{\mu_n}{n}$ 与概率 p 有较大偏差的可能性很小，那么我们便可以通过做试验确定某事件发生的频率并把它作为相应概率的估计，这种方法称为参数估计，它是数理统计中的主要研究课题之一，参数估计的重要理论基础就是大数定律。

§ 2.3.2 实例

在一次全民选举中，总共有 5 个候选人 A、B、C、D、E 竞选总统，经全民投票后，结果如下：

A	c
ABCDE	33%
BDCEA	16%
CDBAE	3%
CEBDA	8%
DECBA	18%
ECBDA	22%

问谁是总统？请制定一些合理的选举规则，使 5 个候选人都有可能当选。

§ 2.4 数理统计 (Mathematical Statistics)

数理统计的方法及考虑的问题不同于一般的资料统计, 它更侧重于应用随机现象本身的规律性来考虑资料的收集、整理和分析, 从而找出相应的随机变量的分布律或它的数字特征。由于大量的随机试验必能呈现出它的规律性, 因而从理论上讲, 只要对随机现象进行足够多次观察, 被研究的随机现象的规律性一定能清楚地呈现出来, 但是实际上所允许的观察永远只能是有限的, 有时甚至是少量的。因此我们所关心的是怎样有效地利用有限的资料, 便能去掉那些由于资料不足所引起的随机干扰, 而把那些实质性的东西找出来, 一个好的统计方法就在于能有效地利用所获得的资料, 尽可能作出精确而可靠的结论。

§ 2.4.1 数理统计的基本概

1) 母体和子样

我们把所研究的全部元素组成的集合称为母体或总体, 而把组成母体的每个元素称为个体。

为了对母体的分布律进行各种研究, 就必需对母体进行抽样观察。一般来说, 我们还不止进行一次抽样观察, 而要进行几次观察。设 X_1, X_2, \dots, X_n 是所观察到的结果, 显然它是随机变量, 称它为容量是 n 的子样。把 X_1, X_2, \dots, X_n 所取值的全体称为子样空间。

我们抽取子样的目的是为了对母体的分布律进行各种分析推断, 因而要求抽取的子样能很好地反映母体的特性, 这就必须对随机抽样的方法提出一定的要求。通常提出下面两点:

(i) 代表性: 要求子样的每个分量 X_i 与所考察的母体 \mathbf{X} 具有相同的分布 $F(\mathbf{X})$

(ii) 独立性: X_1, X_2, \dots, X_n 为相互独立的随机变量, 也就是说, 每个观察结果即不影响其它观察结果, 也不受其它观察结果的影响。

满足上述两点性质的子样称为简单随机子样, 获得简单随机子样的抽样方法称为简单随机抽样。

对于简单随机子样 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 其分布可以由母体的分布函数 $F(\mathbf{X})$ 完全决定, \mathbf{X} 的分布函数

$$\prod_{i=1}^n F(x_i)$$

2) 统计量

一般来说, 子样的某种不含任何未知参数的函数, 在统计学中都可以称为统计量。

统计量: $\frac{1}{3}(X_1 + X_2 + X_3), X_1^2 + X_2^2, \frac{1}{3}(2X_1 + X_2)$; 非统计量: $\frac{1}{3}(Z_1 + Z_2 + Z_3) - \mu, \frac{Z_1 - Z_2}{\sigma}$

3) 常用的统计量—子样矩

r 阶矩 (或 r 阶原点矩) $A_r = \frac{1}{n} \sum_{i=1}^n X_i^r$ 特别地, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 为子样均值。

r 阶中心矩: $B_r = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^r$ 特别地, $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 为子样方差。

总结: 对于母体, 我们有母体均值 μ , 母体方差 σ^2 , 母体的 k 阶原点矩 $k\mu$ 和 k 阶中心矩 $k\sigma$;

我们可以得到如下结论:

Theorem 2.4.1 设母体服从分布 $F(\mathbf{X})$, $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 是从该母体中抽得的一个简单随机子样, 如果的 $F(\mathbf{X})$ 二阶矩阵存在, 则对子样均值 \bar{X} , 有

$$E\bar{X} = \mu \text{ 和 } \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

$$\begin{aligned} \text{Proof: } E(\bar{X}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu \\ \text{Var}(\bar{X}) &= E(\bar{X} - \mu)^2 = E\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu\right)^2 \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^n (X_i - \mu)\right]^2 = \frac{1}{n^2} \sum_{i=1}^n E(X_i - \mu)^2 = \frac{\sigma^2}{n} \end{aligned}$$

思考: 是否存在更简单的证明方法?

Theorem 2.4.2 对于子样方差 S_n^2 , 其均值 $E S_n^2 = \frac{n-1}{n} \sigma^2$

$$\begin{aligned} \text{Proof: } \because S_n^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \frac{1}{n^2} \left(\sum_{i=1}^n X_i\right)^2 \\ \therefore E S_n^2 &= E \left\{ \frac{1}{n} \sum_{i=1}^n X_i^2 - \frac{1}{n^2} \left(\sum_{i=1}^n X_i\right)^2 \right\} \\ &= \alpha_2 - \frac{1}{n^2} E \left\{ \sum_{i=1}^n X_i^2 + \sum_{i \neq j} \sum_{j=1}^n X_i X_j \right\}, \text{ where } \alpha_2 = E \frac{1}{n} \sum_{i=1}^n X_i^2 \\ &= \alpha_2 - \frac{1}{n^2} (n\alpha_2 + n(n-1)\mu^2) = \frac{n-1}{n} (\alpha_2 - \mu^2) = \frac{n-1}{n} \sigma^2 \end{aligned}$$

4) 顺序统计量、经验分布函数与子样矩

设是从母体 (X_1, X_2, \dots, X_n) 中抽取的一个子样, 记 (x_1, x_2, \dots, x_n) 是子样的一个观察值, 将观察值的各分量按大小递增次序排列, 得到

$$x_1^* \leq x_2^* \leq \dots \leq x_n^*$$

当 (X_1, X_2, \dots, X_n) 取值为 (x_1, x_2, \dots, x_n) 。我们定义 $X_k^{(n)}$ 取值为 x_k^* 称由此得到的 $X_1^{(n)}, X_2^{(n)}, \dots, X_n^{(n)}$ 为 (X_1, X_2, \dots, X_n) 的一组顺序统计量。显然 $X_1^{(n)} \leq X_2^{(n)} \leq \dots \leq X_n^{(n)}$, $X_1^{(n)} = \min_{1 \leq i \leq n} X_i$ 即 $X_1^{(n)}$ 的观察值是子样观察值中最小的一个, 而 $X_n^{(n)} = \max_{1 \leq i \leq n} X_i$, $X_n^{(n)}$ 的观察值是子样观察值中最大的一个。记:

$$F_n^*(x) \begin{cases} 0, & \text{when } x \leq x_1^* \\ \frac{k}{n}, & \text{when } x < x_{k+1}^*, k=1, 2, \dots, n-1 \\ 1, & \text{when } x > x_n^* \end{cases}$$

$0 \leq F_n^*(x) \leq 1$, 且作为 x 的函数是一非减左连续函数, 把 $F_n^*(x)$ 看作为 x 的函数, 它具备分布函数所要求的性质, 故称为经验分布函数 (或子样分布函数)。

经验分布函数也是子样的函数, 它与子样矩之间具有下列关系: 设 (x_1, x_2, \dots, x_n) 是子样观察值, $F_n^*(x)$ 是对应的经验分布函数, 则有:

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i = \int x dF_n^*(x) \\ s_n^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \int (x - \bar{x})^2 dF_n^*(x) \\ A_v &= \frac{1}{n} \sum_{i=1}^n x_i^v = \int x^v dF_n^*(x), \quad v = 2, 3, \dots \\ B_v &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^v = \int (x - \bar{x})^v dF_n^*(x), \quad v = 3, 4, \dots\end{aligned}$$

§ 2.4.2 正态母体子样的线性函数的分布

Theorem 2.4.3 X_1, X_2, \dots, X_n 是抽自正态母体 $N(\mu, \sigma^2)$ 的一个子样, 统计量 U 是子样的任一确定的线性函数

$$U = a_1 X_1 + a_2 X_2 + \dots + a_n X_n \quad (2.4.1)$$

则 U 也是正态随机变量, 均值、方差分别为

$$\mathbb{E}(U) = \mu \sum_{k=1}^n a_k, \quad \text{Var}(U) = \sigma^2 \sum_{k=1}^n a_k^2 \quad (2.4.2)$$

在 (2.4.1) 式中, 特别地取 $a_k = \frac{1}{n}, k = 1, \dots, n$ 此时行到的 U 是子样均值 \bar{X} 。 $E(\bar{X}) = \mu$, $\text{Var}(\bar{X}) = \frac{1}{n} \sigma^2$

由此可见, \bar{X} 具有与 X 相同的均值, 但是它更向数学期望集中, 集中程度与子样容量 n 的大小有关。

Theorem 2.4.4 (1) X_1, X_2, \dots, X_n 是独立同分布随机变量, 同服从于正态分布 $N(\mu, \sigma^2)$

(2) $A = (a_{ij})$ 是 $p \times n$ 矩阵, 记

$$Y \triangleq \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots \\ a_{p1} & \cdots & a_{pn} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \triangleq AX \triangleq (Y_1, Y_2, \dots, Y_p)^T$$

则 Y_1, Y_2, \dots, Y_p 也是正态随机变量, 均值、方差、协方差分别为:

$$\begin{aligned}\mathbb{E}Y_i &= \mu \sum_{k=1}^n a_{ik}, i = 1, \dots, p \\ \text{Var}(Y_i) &= \sigma^2 \sum_{k=1}^n a_{ik}^2, i = 1, \dots, p \\ \text{cov}(Y_i, Y_j) &= \sigma^2 \sum_{k=1}^n a_{ik} a_{jk}, i, j = 1, \dots, p\end{aligned}$$

特别地, 当 $\mu = 0$, 且 \mathbf{A} 是一 $n \times n$ 正交矩阵时, Y_1, Y_2, \dots, Y_p 也是相互独立且同服从于 $N(\mu, \sigma^2)$ 分布的随机变量。

§ 2.4.3 几种与正态分布 $N(0,1)$ 有关的常用分布

1) χ^2 分布

Definition 2.13 设 X_1, X_2, \dots, X_n 是相互独立, 且同服从于 $N(0,1)$ 分布的随机变量,

$$x_n^2 = \sum_{i=1}^n X_i^2$$

所服从的分布为 χ^2 分布, χ_n^2 称为自由度为 n 的 χ^2

Theorem 2.4.5 设 $X_1 \sim \chi^2(n_1)$ 和 $X_2 \sim \chi^2(n_2)$, 且 X_1, X_2 相互独立, 则 $X_1 + X_2 \sim \chi^2(n_1 + n_2)$

2) t 分布

设 $X \sim N(0,1)$ 和 $Y \sim \chi^2(n)$, 且 X 和 Y 相互独立, 则称随机变量 $T = \frac{X}{\sqrt{Y/n}}$

所服从的分布为 t 分布。n 称为它的自由度, 且记 $T \sim t(n)$ 。

3) F 分布

Definition 2.14 设 X 和 Y 是相互独立的 χ^2 分布随机变量, 自由度分别为 m 和 n , 则称随机变量

$$F = \frac{X/m}{Y/n} = \frac{X}{Y} \cdot \frac{n}{m}$$

所服从的分布为 F 分布, (m, n) 称为它的自由度, 且通常写为 $F \sim F(m, n)$ 。

推论 如果 $X/\sigma^2 \sim \chi^2(m), Y/\sigma^2 \sim \chi^2(n)$ 且相互独立, 则 $F = \frac{X}{Y} \cdot \frac{n}{m} \sim F(m, n)$ 分布。

推论 如果 $X \sim F(m, n)$ 分布, 则 $1/X \sim F(n, m)$

结论 设 X_1, \dots, X_m 和 Y_1, \dots, Y_n 分别是来自正态母体 $N(\mu_1, \sigma^2)$ 和 $N(\mu_2, \sigma^2)$ 中所抽取的独立子样。则

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{mS_{1m}^2 + nS_{2n}^2}} \cdot \sqrt{\frac{mn(m+n-2)}{m+n}}$$

服从于 $t(m+n-2)$ 分布。

Example 2.3 (练习) 设 X_1, X_2, \dots, X_n 是从正态 $N(\mu_1, \sigma^2)$ 分布的母体中抽取的简单子样, \bar{X}, S_n^2 分别表示它的子样均值和子样方差。又设 $X_{n+1} \sim N(\mu_1, \sigma^2)$ 且与 X_1, X_2, \dots, X_n 独立。试求统计量

$$\frac{X_{n+1} - \bar{X}}{S_n} \sqrt{\frac{n-1}{n+1}} \quad (\text{提示: 服从 } t(n-1) \text{ 分布})$$

§ 2.4.4 统计量的分布与独立性

Theorem 2.4.6 若 $x \sim N[0, I]$ 且 $x'Ax$ 和 $x'Bx$ 是 x 的两个幂等二次型, 则 $x'Ax$ 和 $x'Bx$ 在 $AB = 0$ 时是独立的。

Proof 2.8 由于 A 和 B 都是对称的和幂等的, $A = A'A$ 和 $B = B'B$, 所以二次型是:

$$\begin{cases} x'Ax = x'A'A x = x'_1 x_1 & \text{其中 } x_1 = Ax \\ x'Bx = x'B'B x = x'_2 x_2 & \text{其中 } x_2 = Bx \end{cases}$$

两个向量都有零均值向量, 所以 X_1 和 X_2 协方差矩阵是 $\mathbb{E}(X_1 X_2') = AIB' = AB = 0$

由于 AX 和 BX 都是一个正态分布随机向量的线性函数, 因而它们也都服从正态分布, 零协方差矩阵暗示它们是统计上独立的。所以, 它们的函数形式 $X'AX$ 和 $X'BX$ 是独立的, 这就证明了两个二次型统计量的独立性。

Example 2.4 易知 $X'X = \sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2$

$$\begin{aligned} X'X &= \sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2 \\ &\begin{cases} \sum_{i=1}^n (x_i - \bar{x})^2 = X'M^0 X, \\ n\bar{x}^2 = X'(I - M^0)X \end{cases} \\ \therefore M^0(I - M^0) &= M^0 - M^{0^2} = 0 \\ \therefore n\bar{x}^2 &\text{ 与 } \sum_{i=1}^n (x_i - \bar{x})^2 \text{ 是相互独立的} \end{aligned}$$

§ 2.4.5 线性变换及二次型的独立性

Theorem 2.4.7 标准正态向量的一个线性函数 Lx 和一个幂等二次型 $x'Ax$, 当 $LA = 0$ 时两个统计量是独立的。证明遵循与对两个二次型的证明同样的逻辑, 将 $x'Ax$ 写作 $x'A'x = (AX)'AX$, 变量 Lx 和 Ax 的协方差矩阵是 $LA = 0$, 这证实了这两个随机向量的独立性, 线性函数和二次型的独立性就可以立即推导。

Example 2.5 $\sqrt{n}\bar{x} = \frac{1}{\sqrt{n}}\vec{i}'X$, $S^2 = \frac{X'M^0X}{n-1} \Rightarrow M^0\vec{i} = \left[I - \frac{1}{n}\vec{i}\vec{i}'\right] \cdot \vec{i} = 0$ 所以上面两个统计量是相互独立的。从而

$$\frac{\sqrt{n}\bar{X}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \frac{1}{n-1}}} = \frac{\sqrt{n}\bar{X}}{S} \sim t(n-1)$$

设 X_1, X_2, \dots, X_n 是从正态母体 $N(\mu_1, \sigma^2)$ 中抽取的一个简单子样。记

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k, \quad S_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2$$

则有 (1) \bar{X} 和 S_n 独立;

$$(2) \quad \bar{X} \sim N\left(\mu, \frac{1}{n}\sigma^2\right);$$

$$(3) \quad nS_n^2/\sigma^2 \sim \chi^2(n-1) \quad T = \frac{\bar{X} - \mu}{S_n/\sigma} \sqrt{n-1} \sim t(n-1)$$

Proof 2.9 $\therefore \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1), nS_n^2/\sigma^2 \sim \chi^2(n-1)$

$$\therefore \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1), \quad nS_n^2/\sigma^2 \sim \chi^2(n-1)$$

$$\therefore T = \frac{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}}{\sqrt{\frac{nS_n^2}{\sigma^2(n-1)}}} = \frac{(\bar{X} - \mu)}{S_n} \sqrt{n-1} \quad \text{服从自由度为 } (n-1) \text{ 的 } t \text{ 分布。}$$

§ 2.4.6 参数估计的常用方法

在参数估计问题中, 我们总是首先假设母体 X 具有一族可能的分布 F , 且 F 的函数形式是已知的, 仅包含有几个未知参数, 记 θ 是支配这分布的未知参数 (可以是向量), 在统计学上, 我们把分布 F 的未知参数 θ 的全部可容许值组成的集合称为参数空间, 记为 Θ 。

我们用 $F(\cdot; \theta)$ 表示 X 的分布, 又称集合 $\{F(\cdot, \theta), \theta \in \Theta\}$ 为 X 的分布函数族。类似地, 如果 X 是连续型随机变量, 我们有概率密度函数族, 如果 X 是离散型随机变量, 我们有概率分布族。

一个参数估计问题就是要求通过子样估计母体分布所包含的未知参数 θ 。

一般地, 设母体具有分布族 $\{F(\cdot, \theta), \theta \in \Theta\}$, X_1, X_2, \dots, X_n 是它的一个子样。点估计问题就是要求构造一个统计量 $T(X_1, X_2, \dots, X_n)$ 作为参数 θ 的估计 (T 的维数与 θ 的维数相同)。在统计学上, 我们称 T 为 θ 的估计量。

1) 矩方法

设 $\{F(\cdot, \theta), \theta \in \Theta\}$ 是母体 X 的可能分布族, $\theta = (\theta_1, \dots, \theta_k)$ 是待估计的未知参数, 假定母体分布的 k 阶矩存在, 则母体分布的 v 阶矩

$$a_v(\theta_1, \dots, \theta_k) = \int_{-\infty}^{\infty} x^v dF(x; \theta), \quad 1 \leq v \leq k \quad \text{是 } \theta = (\theta_1, \dots, \theta_k) \text{ 的函数}$$

对于子样 $X = (X_1, X_2, \dots, X_n)$, 其 v 阶子样矩是

$$A_v = \frac{1}{n} \sum_{i=1}^n X_i^v = \int_{-\infty}^{\infty} x^v dF_n^*(x), \quad 1 \leq v \leq k$$

现在用子样矩作为母体矩的估计, 即令

$$a_v(\theta_1, \dots, \theta_k) = A_v = \frac{1}{n} \sum_{i=1}^n X_i^v, \quad v = 1, 2, \dots, k \quad (2.4.3)$$

这样, (2.4.3)式确定了包含 k 个未知参数 $\theta = (\theta_1, \dots, \theta_k)$ 的 k 个方程式。

Example 2.6 母体均值和方差的矩估计。

设 X_1, \dots, X_n 是一子样, 设母体的二阶矩存在, 则有 $\alpha_2 = \sigma^2 + \mu^2$ 用矩方法得方程组

$$\begin{cases} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \\ \hat{\mu}^2 + \hat{\sigma}^2 &= \hat{\alpha}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases} \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = S_n^2$$

所以母体均值 μ 和方差 σ^2 的矩估计分别是子样均值 \bar{x} 和子样方差 S_n^2 。运用以前的有关定理有

$$\begin{aligned} \mathbb{E}\hat{\mu} &= \mathbb{E}\bar{X} = \mu \\ \text{Var}(\hat{\mu}) &= \frac{1}{n}\sigma^2 \\ \mathbb{E}(\hat{\sigma}^2) &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

由此可见, $\hat{\mu}$ 作为 μ 的估计它是在 μ 的真值的周围波动, 且其平均值恰好是真值 μ , 这一性质在统计学上称为无偏性。

2) 最大似然估计方法

一般地, 设母体具有分布密度族 $\{F(\cdot, \theta), \theta \in \Theta\}$, 其中 $\theta = (\theta_1, \dots, \theta_k)$ 是一个未知的 k 维参数向量, 需待估计, 又设 (x_1, x_2, \dots, x_n) 是子样 (X_1, X_2, \dots, X_n) 的一个观察值, 那么子样 (X_1, X_2, \dots, X_n) 落在点 (x_1, x_2, \dots, x_n) 的邻域里的概率 $\prod_{i=1}^n f(x_i; \theta) dx_i$

为方便起见, 记 $L(x; \theta) \triangleq \prod_{i=1}^n f(x_i; \theta)$, (θ 可以是向量) 它看作为 θ 的函数称为 θ 的似然函数。如果选取使下式

$$L(x_1, \dots, x_n; \hat{\theta}) = \sup_{\theta \in \Theta} L(x_1, \dots, x_n; \theta) \quad (2.4.4)$$

成立的 $\hat{\theta}(X) = [\hat{\theta}_1(X), \dots, \hat{\theta}_k(X)]$ 作为 θ 的估计, 则称 $\hat{\theta}(X)$ 是 θ 的最大似然估计。由于 $\log x$ 是 x 的单调函数, 所以(2.4.4)式可等价地写为:

$$\log L(x_1, \dots, x_n; \hat{\theta}) = \sup_{\theta \in \Theta} \log L(x_1, \dots, x_n; \theta) \quad (2.4.5)$$

如果 Θ 是开集, 且 $f(x; \theta)$ 关于 θ 可微, 则满足 (4) 式的解 $\hat{\theta}$ 也一定满足下列似然方程

$$\left. \frac{\partial \log L(x_1, \dots, x_n; \theta)}{\partial \theta_j} \right|_{\theta=\hat{\theta}} = 0, \quad j = 1, \dots, k$$

设 $X = (X_1, X_2, \dots, X_n)$ 是取自均匀分布的子样, 试求 θ 的最大似然估计。

$$f(x; \theta) = \begin{cases} \frac{1}{\theta}, & 0 < x \leq \theta \\ 0, & \text{其它} \end{cases} \quad (\theta > 0)$$

$$\text{此时 } L(x; \theta) = \prod_{i=1}^n f(x_i; \theta) = \begin{cases} \theta^{-n}, & \text{当 } 0 < \max_{1 \leq i \leq n} x_i \leq \theta \\ 0, & \text{其它} \end{cases} \quad (\text{注意: 条件 } 0 < x_i \leq \theta, \quad i = 1, \dots, n)$$

和条件 $0 < \max_{1 \leq i \leq n} x_i \leq \theta$ 是等价的。

显然当 $\theta = \max_{1 \leq i \leq n} x_i$ 时, $L(x; \theta)$ 取到最大值, 所以 $\hat{\theta}_L = \max_{1 \leq i \leq n} X_i = X_n^{(n)}$ 是 θ 的最大似然估计。可以计算出 $E_{\theta}(\hat{\theta}_L) = \frac{n}{n+1}\theta$

§ 2.4.7 估计的有效性

1) 无偏估计

Definition 2.15 一般地, 如果 $T(X)$ 是未知参数 θ 的一个估计量, 且满足下面的关系式, $E_{\theta}T(X) = \theta$ 对一切 $\theta \in \Theta$ 则称 $T(X)$ 是 θ 的无偏估计。

2) 有效估计

Definition 2.16 对两个无偏估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$, 若 $\hat{\theta}_1$ 的方差小于 $\hat{\theta}_2$ 的方差, 即 $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$ 。

判别方式: 在多数情形中, 比较基于两个估计量的协方差矩阵, 若 $\text{Var}(\hat{\theta}_1) - \text{Var}(\hat{\theta}_2)$ 是非负定矩阵, 则 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效。

3) 渐近无偏估计

如果有一列 θ 的估计 $\mathbf{T}_n \triangleq T_n(X_1, \dots, X_n)$ 满足下面的关系式, 则称 \mathbf{T}_n 是 θ 的渐近无偏估计。

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\theta}(\mathbf{T}_n) = \theta, \text{ 对一切 } \theta \in \Theta$$

4) 一致估计

设 X_1, X_2, \dots, X_n 是取自分布族 $\{F(\cdot, \theta), \theta \in \Theta\}$ 的子样, $\mathbf{T}_n = T_n(X_1, X_2, \dots, X_n)$ 是 θ 的一个估计。如果序列 \mathbf{T}_n 随机收敛到真参数值 θ , 即对任意 $\varepsilon > 0$, $\lim_{n \rightarrow \infty} P_{\theta}\{|\mathbf{T}_n - \theta| > \varepsilon\} = 0$, 对一切 $\theta \in \Theta$ 。则称 \mathbf{T}_n 是 θ 的一致估计。

5) 最小方差无偏估计

一般地若 \mathbf{T}_1 是 θ 的一个无偏估计, 关于 θ 的任一无偏估计 \mathbf{T}_2 成立下式 $\text{Var}_{\theta}(\mathbf{T}_1) \leq \text{Var}_{\theta}(\mathbf{T}_2)$, 对一切 $\theta \in \Theta$ 。则称 \mathbf{T}_1 是 θ 的最小方差无偏估计。

6) 线性估计

如果估计 T 是子样的线性函数, 即 T 可以表示为 $T = \sum_{i=1}^n a_i X_i$, 其中 a_1, \dots, a_n 是固定常数, 则称 T 为线性估计。类似地可以定义, 如果 T 是线性估计, 且满足无偏性条件, 则 T 称为线性无偏估计; 如果 U_L 表示 θ 的具有有限方差的线性无偏估计的全体所组成的集合, 而对 $T_0 \in U_L$, 有 $\text{Var}_{\theta}(T_0) \leq \text{Var}_{\theta}(T)$, 对一切 $\theta \in \Theta$ 和 $T \in U_L$ 则称 T_0 为 θ 的最小方差线性无偏估计。

高斯—马尔科夫定理 在线性无偏估计量中, 最小二乘估计量具有最小方差。

7) 克拉美—劳 (Cramer-Rao) 下界

Definition 2.17 (克拉美—劳 (Cramer-Rao) 下界) 假定 x 的密度满足一定的正则条件, 参数 θ 的一个无偏估计量的方差将大于等于:

$$\begin{aligned} \text{Var}(\theta) \geq [I(\theta)]^{-1} &= \left(-E \left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] \right)^{-1} \\ &= \left(E \left[\left(\frac{\partial \ln L(\theta)}{\partial \theta} \right) \left(\frac{\partial \ln L(\theta)}{\partial \theta'} \right) \right] \right)^{-1} \end{aligned}$$

$I(\theta)$ 是样本的信息数。再考虑一个多变量情形。若 θ 是一个参数向量, $I(\theta)$ 是信息矩阵。克拉美—劳定理, 任何无偏估计量的方差矩阵与信息矩阵的逆 $[I(\theta)]^{-1}$ 的差将是一个非负定矩阵, 其中

$$\begin{aligned}
 [I(\theta)]^{-1} &= \left\{ -\mathbb{E} \left[\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] \right\}^{-1} \\
 &= \left\{ \mathbb{E} \left[\left(\frac{\partial \ln L(\theta)}{\partial \theta} \right) \left(\frac{\partial \ln L(\theta)}{\partial \theta'} \right) \right] \right\}^{-1} \\
 \text{即} \quad I(\theta) &= \begin{bmatrix} -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_1^2} \right) & -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} \right) & \cdots & -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_1 \partial \theta_k} \right) \\ -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} \right) & -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_2^2} \right) & \cdots & -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_2 \partial \theta_k} \right) \\ \cdots & \cdots & \cdots & \cdots \\ -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_k \partial \theta_1} \right) & -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_k \partial \theta_2} \right) & \cdots & -\mathbb{E} \left(\frac{\partial^2 L}{\partial \theta_k^2} \right) \end{bmatrix}
 \end{aligned}$$

这个矩阵的逆矩阵 $[I(\theta)]^{-1}$ 称为 C-R 下界或 CRLB。

§ 2.4.8 假设检验

1) 正态母样参数检验

前面我们介绍了两种常用的参数估计方法。实践中还提出了统计推断问题。先看一个例子

Example 2.7 某厂有一批产品, 共一万件, 须经检验后方可出厂。按规定标准, 次品率不得超过 5%, 今在其中任意选取 50 件产品进行检查, 发现有次品 4 件, 问这批产品能否出厂?

在这个例子中, 我们事先对这批产品次品率的情况一无所知, 当然, 从频率稳定性来说, 我们可以用被检查的 50 件产品的次品率 $4/50$ 来估计这整批产品的次品率, 但是我们目前所关心的问题是: 如何根据抽样的次品率 v/n ($=4/50$) 推断这批产品的次品率是否超过了 5%, 也就是说, 首先我们可以对整批产品作一种假设: 次品率低于 5%, 然后利用子样的次品率 v/n 来检验我们所作这一假设的正确性。

我们把任何一个在母体的未知分布上所作的假设称为统计假设。并记为 H_0 。对上面所举的例子中, 统计假设分别是: $H_0: p$ (次品率) ≤ 0.05 。

由于母体的真分布完全被几个未知参数所决定。因此任何一个关于母体未知分布的假设总可以等价地给出在它的未知参数上。这种仅涉及到母体分布中所包含的几个未知参数的统计假设称为参数假设。

对于一个假设检验问题, 首先是根据实际问题的要求提出统计假设 H_0 , 但这仅是第一步, 提出统计假设的目的是要求进一步推断所提出的统计假设 H_0 是否正确。这就要求建立推断统计假设 H_0 的方法。在统计学上, 称判断给定统计假设 H_0 的方法为统计假设检验, 或简称为统计检验。

如果一个统计问题中仅提出一个统计假设, 而且我们的目的也仅仅是判断这一个统计假设是否成立, 并不同时研究其它统计假设。这类检验问题称为显著性检验。显著性检验问题的处理一般步骤是:

- (1) 建立统计假设 H_0 ;
- (2) 构造一个合适的统计量 U 和从子样观察值计算出统计量 U 的观察值 u ;
- (3) 规定一个显著水平 α (一般取 0.05 或 0.01), 求出在 H_0 成立条件下能使 $P_{H_0}\{|U| \geq u_0\} \leq \alpha$ 满足的值 u_0 ;
- (4) 比较观察值 u 和 u_0 , 如果 $|u| \geq u_0$, 则拒绝设 H_0 。

显然, 寻找检验统计量 U 的分布, 至少对于给定的 α 要找出满足 $P_{H_0}\{|U| \geq u_0\} \leq \alpha$ 的临界值 u_0 是很重要的。按进行检验时所取的子样容量的大、小, 分为小样和大样两类问题, 对于小样的显著性检验, 需要给出检验统计量 U 的精确分布, 而对于大样问题可利用 U 的极限分布作为近似。

正态母体参数的显著性检验可总结如下表 2.1。

表 2.1 正态母体参数的显著性检验

检验参数	假设 H_0	统计量	分布
μ	$\mu = \mu_0 (\sigma = \sigma_0)$	$U = \frac{\bar{X} - \mu_0}{\sigma_0} \sqrt{n}$	$N(0, 1)$
μ	$\mu_1 = \mu_2 (\sigma_1, \sigma_2 \text{ 已知})$	$U = (\bar{X} - \bar{Y}) / \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$	$N(0, 1)$
μ	$\mu_1 = \mu_2, \sigma^2 > 0$	$T = \frac{\bar{X} - \mu_0}{S_n} \sqrt{n-1}$	$t(n-1)$
μ	$\mu_1 = \mu_2, \sigma_1 = \sigma_2$	$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{mS_{1m}^2 + nS_{2n}^2}} \cdot \sqrt{\frac{mn(m+n-2)}{m+n}}$	$t(m+n-2)$
σ^2	$\sigma = \sigma_0$	$x^2 = \frac{nS_n^2}{\sigma_0^2}$	$\chi^2(n-1)$
σ^2	$\sigma_1^2 = \sigma_2^2$	$F = \frac{mS_{1m}^2}{nS_{2n}^2} \cdot \frac{n-1}{m-1}$	$F(m-1, n-1)$

上例的解: 为简单起见, 我们可将此问题归结为希望利用次品率 v/n 来检验母体次品率 p 是否满足假设 $H_0: p = p_0 (= 0.05)$ 用 Y 记母体元素的指标, 有 $Y = \begin{cases} 0, & \text{好品} \\ 1, & \text{次品} \end{cases}$ 则在假设 H_0 成立时。 $P\{Y = 0\} =$

表 2.2 正态母体参数的置信区间

待估参数	条件	置信区间下限	置信区间上限	对应的检验统计量
μ	$\sigma = \sigma_0$	$\bar{X} - u_{\alpha/2} \cdot \sigma_0 / \sqrt{n}$	$\bar{X} + u_{\alpha/2} \cdot \sigma_0 / \sqrt{n}$	$U = \frac{\bar{X} - \mu}{\sigma_0} \sqrt{n}$
μ	σ 未知	$\bar{X} - t_{\alpha/2} \cdot \sigma_0 / \sqrt{n-1}$	$\bar{X} + t_{\alpha/2} \cdot \sigma_0 / \sqrt{n-1}$	$U = \frac{\bar{X} - \mu}{\sigma_0} \sqrt{n-1}$
$\mu_1 - \mu_2$	$\sigma_1 = \sigma_2 = ?$	$-\bar{t}_{\alpha/2} \frac{(\bar{X} - \bar{Y}) \sqrt{mS_{1m}^2 + nS_{2n}^2} \cdot \sqrt{m+n}}{\sqrt{(m+n-2)mn}}$	$+\bar{t}_{\alpha/2} \frac{(\bar{X} - \bar{Y}) \sqrt{mS_{1m}^2 + nS_{2n}^2} \cdot \sqrt{m+n}}{\sqrt{(m+n-2)mn}}$	$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{mS_{1m}^2 + nS_{2n}^2}} \times \sqrt{\frac{(m+n-2)m \cdot n}{m+n}}$
σ^2	单子样	$\frac{1}{C_{2a}} \cdot nS_n^2$	$\frac{1}{C_{1a}} \cdot nS_n^2$	$\chi^2 = \frac{nS_n^2}{\sigma_0^2}$
$\frac{\sigma_1^2}{\sigma_2^2}$	双子样	$\frac{1}{f_{2a}} \frac{mS_{1m}^2}{nS_{2n}^2} \cdot \frac{n-1}{m-1}$	$\frac{1}{f_{1a}} \frac{mS_{1m}^2}{nS_{2n}^2} \cdot \frac{n-1}{m-1}$	$F = \frac{mS_{1m}^2/\sigma_1^2}{nS_{2n}^2/\sigma_2^2} \cdot \frac{n-1}{m-1}$

$1 - p_o, P\{Y = 1\} = p_o; \mathbb{E}Y = p_o, \text{Var}(Y) = p_o(1 - p_o)$, 设 x_1, x_2, \dots, x_n , 则 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{v}{n}$, 其中 v 表示子样中的次品数。

由中心极限定理知道, 在 $H_0(p = p_o)$ 成立的条件下,

$$U = \frac{(\bar{X} - \mathbb{E}X)}{\sqrt{\text{Var}(X)}} \sqrt{n} = \frac{\left(\frac{v}{n} - p_o\right)}{\sqrt{p_o(1-p_o)}} \sqrt{n} = \frac{(v - np_o)}{\sqrt{np_o(1-p_o)}} \quad (2.4.6)$$

渐近于 $N(0, 1)$ 分布, 因此当 n 较大时 (一般在 30 以上), 可把(2.4.6)式决定的 U 近似地作为正态变量来处理。

2) 正态母体参数的置信区间

在许多实际问题中, 我们往往希望通过子样的观察给出一个范围, 使得这个范围能按足够大的概率 (给定的) 包含我们所感兴趣的参数, 在统计学上, 我们称这个范围叫置信区间 (或置信域), 这类问题称为区间估计问题。

参数的置信区间与参数的假设检验之间有着密切的联系。可以直接正从态母体参数的各种检验法构造正态母体参数的各种置信区间。正态母体参数的各科置信区间的情况可总结如下表2.2。

3) 联合置信域

下面我们讨论正态分布均值和方差的联合置信域。

(μ, σ^2) 的联合置信域可以运用 \bar{X} 和 S_n^2 的联合分布来构造。因为是独立的, 因此, 如果我们希望寻找置信水平为 0.95 的置信域, 我们可以找到数 α , 和 c_1, c_2 , 使得

$$\begin{cases} P \left\{ -\alpha < \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} < \alpha \right\} = \sqrt{0.95} \approx 0.975 \\ P \left\{ c_1 < \frac{nS_n^2}{\sigma_0^2} < c_2 \right\} = \sqrt{0.95} \approx 0.975 \end{cases} \implies P \left\{ -\alpha < \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} < \alpha, c_1 < \frac{nS_n^2}{\sigma_0^2} < c_2 \right\} = 0.95$$

$$\therefore P \left\{ (\mu_0 - \bar{X})^2 < a^2 \sigma_0^2 / n, \frac{nS_n^2}{c_2} < \sigma_0^2 < \frac{nS_n^2}{c_1} \right\} = 0.95 \quad (2.4.7)$$

由此可见, (μ, σ^2) 的置信度为 0.95 的联合置信域是(2.4.7)式大括号内不等式 μ, σ^2 对所给出的范围。

4) 广义似然比检验

设 $\mathbf{X} = (X_1, \dots, X_n)$ 是从母体中抽取的子样, 其可能分布族 $\{f(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, 其中 $\boldsymbol{\theta}$ (可以是向量) 是未知参数 (当母体是连续型变量时 f 表示分布密度, 当母体是离散型变量时 f 表示概率分布)。要求检验假设 $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ 。这里应指出, $\boldsymbol{\theta}_0$ 有时是表示一个集合, 如在运用 t 检验法检验假设 $H_0: \mu = \mu_0$ 时, 那里

$$\begin{aligned} \boldsymbol{\theta} &\triangleq (\mu, \sigma^2) \\ \Theta &= \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\} \\ \boldsymbol{\theta}_0 &\triangleq \{(\mu_0, \sigma^2) > 0\} \end{aligned}$$

它是一个未知参数的集合而不是一个单点。现在我们引进一个统计量:

$$\lambda(x) \triangleq \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}_0} \prod_{i=1}^n f(x_i; \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \prod_{i=1}^n f(x_i; \boldsymbol{\theta})}$$

习惯上称 $\lambda(x)$ 为广义似然比, 显然它是子样的函数, 不依赖于未知参数 $\boldsymbol{\theta}$ 。由于 $\boldsymbol{\theta} \in \Theta$, 所以 $0 \leq \lambda(x) \leq 1$

类似于最大似然原理, 如果 $\lambda(x)$ 取值较小, 这说明当 H_0 为真时观察到样点 x 的概率比 H_0 不真时观察到样点 x 的概率要小得多, 此时我们有理由怀疑假设 H_0 不真。所以从广义似然比出发, 该检验问题是当下式 ($\lambda(x) \leq \lambda_0$) 成立时拒绝 H_0 , 其中 λ_0 的选取是使得下式成立。

$$P_{\boldsymbol{\theta}} \{ \lambda(X) \leq \lambda_0 \} \leq \alpha, \text{ 对一切 } \boldsymbol{\theta} \in \boldsymbol{\theta}$$

给出的检验法称为水平为 α 的广义似然比检验。当 $\boldsymbol{\theta}_0$ 是一个单点时可写为 $P_{\boldsymbol{\theta}_0} \{ \lambda(X) \leq \lambda_0 \} \leq \alpha$

进一步分析这样一个参数假设的显著性检验过程，就会发现有一系列问题有待解决。如由于采取接受或拒绝假设 H_0 的判断是根据子样观察值作出的，而子样是随机变量。子样观察值的出现带有随机性，因此判断有可能发生错误。则能发生那些类型的错误和发生各类错误的概率有多大？

可能犯下面两种类型的错误：当原假设 H_0 为真的时候，即 θ 的真实值落在 Θ_0 中时，作出拒绝 H_0 的决策 a_1 ——它称为第一类错误；另一种错误是当备选假设为真时，即 θ 的真实值落在 Θ 之中时，作出接受原假设 H_0 的决策 a_0 ——它称为第二类错误（见表2.3）。这两类错误所造成的影响常常很不一样。例如我们要求检验病人是否患有某种疾病。若我们取原假设是该人患此种疾病，则第二类错误（无病当作有病）造成由于使用不必要的药品而引起病人的痛苦和经济上的浪费，但第一类错误（有病当作无病）就有可能导致死亡。

表 2.3 两类错误

	H_0 为真	H_1 为真
接受 H_0	正 确	第II类错误
拒绝 H_0	第I类错误	正 确

当然，我们希望所作出的检验能使得犯这两种类型错误的概率同时尽可能地小，最好全为零，但实际上这是不可能的，当子样的容量（即观察个数）给定后，犯这两种类型错误的概率就不能同时被控制。

当许多传统的计量经济学模型在 20 世纪 70 年代的经济动荡面前预测失灵时，误差修正模型却显示了它的稳定性和可靠性。对其原因进行深入分析之后发现，误差修正模型的非稳定的单整变量之间存在一种长期稳定关系。C.J.Granger 把这种长期稳定关系称为“协整关系”，于是，一种新的理论——协整理论诞生了。虽然协整理论诞生于误差修正模型之后，但在本节中，为了便于理解，我们首先介绍协整理论，然后引出误差修正模型。

将协整理论误差修正模型作为单方程计量经济学模型的扩展的理由在于，传统的计量经济学模型是以某种经济理论或对经济行为的认识来确立模型的理论关系形式，而在这里，则是从经济变量的数据中所显示的关系出发，确定模型包含的变量和变量之间的理论关系。这是 20 世纪 80 年代以来计量经济学模型建模理论的一个重大发展。

§ 3.1 单整 (Integration)

§ 3.1.1 稳定序列

如果一个时间序列 x_t 是稳定的，则

- (1) 其均值 $\mathbb{E}(x_t)$ 与时间 t 无关；
- (2) 其方差 $\text{Var}(x_t)$ 是有限的，并不随着 t 的推移产生系统的变化。

于是，时间序列 x_t 将趋于返回它的均值，以一种相对不变的振幅围绕均值波动。如果一个时间序列 x_t 是非稳定的，则其均值、方差将随 t 而改变。例如，随机游动序列 $x_t = x_{t-1} + \varepsilon_t$ $\varepsilon_t \sim N(0, \delta^2)$

若 $x_0 = 0$, 则 $x_t = \sum_{i=1}^t \varepsilon_i$ $\text{Var}(x_t) = t\delta^2$ 。当 $t \rightarrow \infty$ 时, $\text{Var}(x_t) \rightarrow \infty$ 均值就无意义了, 实际上序列 x_t 返回曾经达到过的某一点的期望时间是无穷大。

一个稳定序列一般以用一个自回归移动平均表达式 $ARMA(p, q)$ 表示:

$$x_t = \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + \theta_1 \xi_{t-1} + \cdots + \theta_q \xi_{t-q}$$

§ 3.1.2 单整

如果一个序列在成为稳定序列之前必须经过 d 次差分, 则该序列被称为 d 单整。记为 $I(d)$ 。换句话说, 如果序列 x_t 是非稳定序列, $\Delta^d x_t$ 是稳定序列, 则是 x_t 序列 $I(d)$ 。其中

$$\Delta x_t = x_t - x_{t-1}, \Delta^2 x_t = \Delta(\Delta x_t), \quad \Delta^d x_t = \Delta(\Delta^{d-1} x_t)$$

如果有两个序列分别为 d 阶单整和 e 阶单整, 即

$$x_t \sim I(d), y_t \sim I(e), e > d$$

则二序列的线性组合是 e 阶单整序列, 即

$$z_t = \alpha x_t + \beta y_t \sim I[\max(d, e)]$$

§ 3.1.3 单整的检验

对于时间序列 x_t , 建立下列方程:

$$x_t = \rho x_{t-1} + \varepsilon_t \quad \text{or} \quad \Delta x_t = (\rho - 1)x_{t-1} + \varepsilon_t$$

如果 ρ 不显著为 0, 则序列 x_t 至少为 1 阶单整 $I(1)$ 问题在于如何判断 ρ 是否是显著为 0。

构造统计量, 但这时统计量服从由 Dickey 和 Fuller 于 1979 年提出的 Dickey-Fuller 分布, 即 DF 分布。像第二章中介绍的变量显著性检验统计量的计算一样, 计算得到统计量的值; 从 DF 分布表中查出给定显著性水平下的临界值; 如果统计量的绝对值大于临界值的绝对值, 则拒绝假设, 序列至少为 1 阶单整 $I(1)$ 。这就是 Dickey-Fuller 检验, 也称单位根检验。

通过了 1 阶单整检验后, 再建立如下方程:

$$\Delta^2 x_t = (\rho - 1)\Delta x_{t-1} + \varepsilon_t$$

进行同样过程的检验, 如果通过检验, 则序列 x_t 至少为 2 阶单整 $I(2)$ ……直到不能通过检验为止。通过该检验, 同时也就确定了序列 x_t 的单整的阶数。

在 DF 检验中, 由于不能保证方程中的一般项, 在经济数据中, 表示流量的序列, 例如以不变价格表示的消费额、收入等经常表现为 1 阶单整; 表示存量的序列, 例如以下不变价格表示的资产总值、储蓄余额等经常表现为 2 阶单整; 由于价格指数的作用, 也经常表现为 2 阶单整; 而像利率等序列, 经常表现为 0 阶单整。了解这些, 对于选择什么变量进入模型是十分重要的。

在 DF 检验中, 由于不能保证方程中的 ε_t 所以得到的 $(\rho - 1)$ 的估计值不是无偏的。于是 Dickey 和 Fuller 于 1979、1980 年对 DF 检验进行了扩充, 形成了 ADF (Augment Dickey-Fuller) 检验。这是目前普遍应用的单整检验方法。由于其过程较为复杂, 这里不作介绍。

§ 3.2 协整 (Cointegration)

§ 3.2.1 定义及意义

如果序列 $X_{1t}, X_{2t}, \dots, X_{kt}$ 都是 d 阶单整, 存在一个向量 $a = (a_1, a_2, \dots, a_k)$, 使得 $Z_t = aX_t' \sim I(d-b)$, 其中, $b > 0$, $X_t' = (X_{1t}, X_{2t}, \dots, X_{kt})'$, 则认为序列 $X_{1t}, X_{2t}, \dots, X_{kt}$ 是 $(d-b)$ 阶协整, 记为 $X_t \sim CI(d, b)$, a 为协整向量

例如, 居民收入时间序列 Y_t 为 1 阶单整序列, 居民消费时间序列 C_t 也为 1 阶单整序列, 如果二者的线性组合 $a_1 Y_t + a_2 C_t$ 构成的新序列为 0 阶单整序列, 于是认为序列 Y_t 与 C_t 是 $(1, 1)$ 阶协整。

由此可见, 如果两个变量都是单整变量, 只有当它们的单整阶相同时, 才可能协整, 例如上面的居民收入 Y_t 和居民消费 C_t ; 如果它们的单整阶不相同, 就不可能协整, 例如居民消费 C_t 和居民储蓄余额 S_t (一般讲作为存量的居民储蓄余额 S_t 为 2 阶单整)。

三个以上的变量, 如果具有不同的单整阶数, 有可能经过线性组合构成低阶单整变量。例如, 如果存在

$$W_t \sim I(1), V_t \sim I(2), U_t \sim I(2)$$

并且

$$P_t = aV_t + bU_t \sim I(1)$$

$$Q_t = cW_t + eP_t \sim I(0)$$

那么认为

$$V_t, U_t \sim CI(2, 1)$$

$$W_t, P_t \sim CI(1, 1)$$

从协整的定义可以看出协整的经济意义在于: 两个变量, 虽然它们具有各自的长波动规律, 但是如果它们是协整的, 则它们之间存在着一个长期稳定的比例关系。例如居民收入 Y_t 和居民消费 C_t , 如果它们各自都是 1 阶单整, 并且它们是 $(1, 1)$ 阶协整, 则说明它们之间存在着一个长期稳定的比例关系, 而这个比例关系就

是消费倾向,也就是说消费倾向是不变的。从计量经济学模型的意义讲,建立如下消费函数模型:

$$C_t = a_0 + a_1 Y_t + \mu_t$$

变量选择是合理的,随机误差项一定是“白噪声”(即均值为0、方差不变的稳定随机序列),模型参数有合理的经济解释。

反过来,如果两个变量,具有各自的长期波动规律,但是它们不是协整的,则它们之间就不存在着一个长期稳定的比例关系。例如居民消费 C_t 和居民储蓄余额 S_t ,由于它们单整阶数不同,所以它们不是协整的,则说明它们之间不存在着一个长期稳定的比例关系。从计量经济学模型的意义讲,建立如下消费函数模型:

$$C_t = a_0 + a_1 S_t + \mu_t$$

或者

$$C_t = a_0 + a_1 Y_t + a_2 S_t + \mu_t$$

变量选择是不合理的,随机误差项一定不是“白噪声”,模型参数没有合理的经济解释。

从这里,我们已经初步认识,检验变量之间的协整关系,在建立计量经济学模型中的重要性。而且,从变量之间是否具有协整关系出发选择模型的变量,其数据基是牢固的,其统计性质是优良的。从协整理论出发,在建立消费函数模型时,就不会选择居民储蓄余额作为居民消费的解释变量;但是,按照传统的计量经济学建模理论,从已经认识的经济理论出发选择模型的变量,那么选择居民储蓄余额和居民收入共同作为居民消费的解释变量,不仅不感到奇怪,而且被认为是完全合理的,在本书第五章中将会看到,按照“生命周期消费理论”建立的消费函数模型正是这样的。

§ 3.2.2 协整的检验

(1) 两变量的 Engle-Granger 检验。为了检验两变量 Y_t, X_t 是否为协整,Engle 和 Granger 于 1987 年提出两步检验法。第一步,用 OLS 方法估计下列方程:

$$Y_t = aX_t + \varepsilon_t$$

得到

$$\begin{aligned}\hat{Y}_t &= \hat{a}X_t \\ \hat{\varepsilon}_t &= Y_t - \hat{Y}_t\end{aligned}$$

称为协整回归。

第二步,检验 $\hat{\varepsilon}_t$ 的单整性。如果 $\hat{\varepsilon}_t$ 为稳定序列,则认为变量 Y_t, X_t 为 (1, 1) 阶协整;如果 Y_t 为 1 阶单整,则认为变量 Y_t, X_t 为 (2, 1) 阶协整……检验 $\hat{\varepsilon}_t$ 的单整性的方法即是上述的 DF 检验。

(2) 多变量协整关系的检验。上述 Engle-Granger 检验通常用于检验两变量之间的协整关系, 对于多变量之间的协整关系, Johansen 于 1988 年, 以及与 Juselius 于 1990 年提出了一种用极大似然法进行检验的方法, 通常称为 Jo-hansen 检验。

§ 3.3 误差修正模型 (ECM)

误差修正模型 (Error Correction Model) 是一种具有特定形式的计量经济学模型, 它的主要形式是由 Davidson、Hendry、Srba 和 Yeo 于 1978 年提出的, 称为 DHSY 模型。为了便于理解, 我们通过一个具体的模型来介绍它的结构。

§ 3.3.1 误差修正模型

对于 (1, 1) 阶自回归分布滞后模型

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 y_t + \beta_3 z_{t-1} + \varepsilon_t \quad (3.3.1)$$

移项后得到

$$\begin{aligned} \Delta y_t &= \beta_0 + \beta_1 \Delta z_t + (\beta_2 - 1) y_{t-1} + \beta_3 z_{t-1} + \beta_1 z_{t-1} + \varepsilon_t \\ &= \beta_0 + \beta_1 \Delta z_t + (\beta_2 - 1) \left(y - \frac{\beta_1 + \beta_3}{1 - \beta_2} z \right)_{t-1} + \varepsilon_t \end{aligned} \quad (3.3.2)$$

方程 3.3.2 即为误差修正模型。其中 $y = \frac{\beta_1 + \beta_3}{1 - \beta_2} z$ 为误差修正项。

显然, 3.3.2 实际上是一个短期模型, 反映了 y_t 的短期波动 δy_t 是如何被决定的。如果变量 y 和 z 之间存在长期均衡关系, 即存在 $y = az$ 例如在 3.3.1 中, 若 $z = \bar{z}$, 那么 y 的均衡值与 \bar{z} 有下列均衡关系

$$\bar{y} = \frac{\beta_1 + \beta_3}{1 - \beta_2} \bar{z}$$

3.3.2 中的误差修正项正是与它相一致的。所以它反映长期均衡对短期波动的影响; 3.3.2 中的差分项反映变量短期波动的影响。于是, 被解释变量的波动被分成两部分: 一部分为短期波动, 一部分为长期均衡。

模型 3.3.2 可以写成

$$\Delta y_t = \beta_0 + \beta_1 \Delta z_t + \gamma ecm + \varepsilon_t \quad (3.3.3)$$

其中 ecm 表示误差修正项。由 3.3.1 可知, 一般情况下 $|\beta_2| < 1$, 所以有 $\gamma = \beta_2 - 1 < 0$ 我们可以据此分析 ecm 的修正作用: 若 $(t-1)$ 时刻大于其长期均衡解 $\frac{\beta_1 + \beta_3}{1 - \beta_2} z$, ecm 为正, $\gamma \times ecm$ 为负, 使得 δy_t 减少; 若 $(t-1)$ 时刻 y 小于其长期均衡解 $\frac{\beta_1 + \beta_3}{1 - \beta_2} z$, ecm 为负, $\gamma \times ecm$ 为正, 使得 δy_t 增大。体现了长期均衡误差对 y_t 的控制。

§ 3.3.2 ECM 与协整的关系

对于上述 (1, 1) 阶自回归分布滞后模型, 如果

$$y_t \sim I(1), z_t \sim I(1)$$

对么, 3.3.2式左边

$$\Delta y_t \sim I(0)$$

右边的 $\Delta z_t \sim I(0)$ 只有 y 与 z 协整, 才能保证右边也是 $I(0)$ 。此时, $\frac{\beta_1 + \beta_3}{1 - \beta_2}$ 为协整系数, $y_t - \frac{\beta_1 + \beta_3}{1 - \beta_2} z_t$ 即为均衡误差。

§ 3.3.2.1 从协整理论到误差修正模型

前面提到, 实际上是先有误差修正模型, 然后用协整理论去解释误差修正模型。那么在今天, 我们就可以首先对变量进行协整分析, 以发现变量之间的协整关系, 即长期均衡关系, 求出协整系数, 并以这种关系构成误差修正项。然后建立短期模型, 将误差修正项看作一个解释变量, 加同其他反映短期波动的解释变量一起, 建立短期模型, 即误差修正模型。

§ 3.4 单方程计量经济学模型的贝叶斯估计

贝叶斯 (Bayes) 统计是由 T.R.Bayes 于 19 世纪创立的数理统计的一个重要分支, 20 世纪 50 年代, 以 H.Robbins 为代表提出了在计量经济学模型估计中将经验贝叶斯方法与经典方法相结合, 引起了广泛的重视, 得到了广泛的应用。贝叶斯方法的数学描述较为复杂, 内容也很广泛, 在本节中主要介绍其基本概念, 并尽可能用简单的语言介绍其用于计量经济学模型估计的过程, 只涉及一些简单的内容, 使读者对这种方法有一个概念性了解, 为深入学习与应用建立一个基础。

§ 3.4.1 概念

1. 贝叶斯方法的基本思路

贝叶斯方法是与传统 (也称经典的) 计量经济学模型的估计方法相对的一种统计学方法。它的基本思路是: 认为要估计的模型参数是服从一定分布的随机变量, 根据经验给出待估参数的先验分布 (也称为主观分布), 关于这些先验分布的信息被称为先验信息; 然后根据这些先验信息, 并与样本信息相结合, 应用贝叶斯定理, 求出待估参数的后验分布; 再应用损失函数, 得出后验分布的一些特征值, 并把它们作为待估参数的估计量。

贝叶斯方法与经典估计方法的主要不同之处是：

(1) 关于参数的解释不同。经典估计方法认为待估参数具有确定值，它的估计量才是随机的，如果估计量是无偏的，该估计量的期望等于那个确定的参数；而贝叶斯方法认为待估参数是一个服从某种分布的随机变量。

(2) 所利用的信息不同。经典方法只利用样本信息；贝叶斯方法要求事先提供一个参数的先验分布，即人们有关参数的主观认识，被称为先验信息，是非样本信息，在参数估计过程中，这些非样本信息与样本信息一起被利用。

(3) 对随机误差项的要求不同。经典方法，除了最大或然法，在参数估计过程中并不要求知道随机误差项的具体分布形式，但是在假设检验与艾是估计时是需要的；贝叶斯方法需要知道随相误差的具体分布形式。

(4) 选择参数估计量的准则不同。经典方法或者以最小二乘，或者以最大或然为准则，求解参数估计量；贝叶斯方法则需要构造一个损失函数，并以损失函数最小化为准则求得参数估计量。

2. 贝叶斯定理

贝叶斯定理是贝叶斯估计方法的理论基础。贝叶斯定理表达如下：

$$g(\theta/Y) = \frac{f(Y/\theta)g(\theta)}{f(Y)} \quad (3.4.1)$$

其中 θ 为待估参数； Y 为样本观测值信息，即样本信息； $g(\theta)$ 是待估参数 θ 的先验分布密度函数； $g(\theta | Y)$ 为 θ 的后验分布密度函数； $f(Y)$ 和是 $f(Y | \theta)$ 的密度函数。因为对 θ 而言， $f(Y)$ 可以认为是常数（样本观测值独立于待估参数）， $f(Y | \theta)$ 在形式上又同的或然函数 $L(\theta | Y)$ 一致（在最大或然法中已经了解），于是 3.4.1 可以改写为

$$g(\theta | Y) = L(\theta | Y) \cdot g(\theta) \quad (3.4.2)$$

即后验信息正比于样本信息与先验信息的乘积。3.4.2 表明，可以通过样本信息对先验信息的修正来得到更准确的后验信息。得到后验分布的密度函数后，. 就可以此为基础进行参数的点估计、区间估计与假设检验。

3. 损失函数

常用的损失函数有线性函数和二次函数。不同的损失函数，得到的参数估计值是不同的。在下面的估计过程中再作进一步的说明。

§ 3.4.2 单方程计量经济学模型贝叶斯估计的过程

采用贝叶斯方法估计单方程计量经济学模型的过程可概括如下：

1. 确定模型的形式, 指出待估参数 B
2. 给出待估参数 B 的先验分布

通常待估参数 B 是一个多维向量, 即需要给出多参数的联合信息先验。如果对参数一无所知, 可以认为 B 是均匀分布。实际上常用的也是均匀分布和共轭分布 (尤以自然共轭分布为多)。

3. 利用样本信息, 修正先验分布

利用贝叶斯定理, 导出 B 的后验密度函数 (密度函数简称为 p.d.f.)。到此为止, 实际上已经得到 B 的所有信息。

4. 利用 B 的后验密度函数, 进一步推断出 B 的点估计值, 或进行区间估计与假设检验。

一般利用损失函数推断出 B 的点估计值, 利用最高后验密度区间进行 B 的区间估计, 利用最高后验密度区间或者利用后验优抛比进和假设检验。

5. 预测

根据求得的参数估计值进行预测。这并不是贝叶斯估计过程所必须, 但对于检验估计结果是重要的。

§ 3.4.3 正态线性单方程计量经济学模型的贝叶斯估计

下面以正态线性单方程计量经济学模型为例分析贝叶斯估计方法。选择正态线性单方程计量经济学模型, 主要因为: (1) 在第一章已经介绍, 多元线性单方程计量经济学模型具有普遍性意义; (2) 随计误差项是大量随机扰动之总和, 根据中心极限定理, 可以认为它是渐近正态分布; (3) 计算简单, 使用方便, 并能完整地体现贝叶斯估计方法的主要内容。

正态线性单方程计量经济学模型又分为随计误差项方差已知和方差未知两种情况。作为方法的演示分析, 我们只讨论方差已知的情况。

1. 有先验信息的后验分布

对于正态线性单方程计量经济学模型

$$Y = B'X + \mu$$

其中 $\mu \sim N(0, \sigma^2 I)$, 其他变量及参数的含义同前。

为方便起见, 在下面的讨论中将只涉及正态分布的“核”, 即其指数部分, 而忽略其常数部分, 不影响讨论结果。选择 B 的先验分布为自然共轭分布, 即 B 的密度函数和它的或然函数, 以及二者结合后产生的函数服从同一分布。 B 的自然共轭先验密度函数为正态密度函数

$$g(B) = e^{-\frac{1}{2}(B-\bar{B})' \Sigma_b^{-1}(B-\bar{B})} \quad (3.4.3)$$

其中 \bar{B} 为待估参数先验均值, $\bar{\Sigma}_B$ 为待估参数先验协方差矩阵。 B 的或然函数 $L(B | Y)$ 等同于它的联合密度函数, 即

$$L(B | Y) = e^{-\frac{1}{2\sigma^2}(Y - XB)'(Y - XB)} \quad (3.4.4)$$

利用贝叶斯定理, 得到 B 的后验密度函数为

$$\begin{aligned} g(B | Y) &= g(B) \cdot L(B | Y) \\ &= \exp \left\{ -\frac{1}{2\sigma^2} \left(\left(A^{\frac{1}{2}} B - A^{\frac{1}{2}} \bar{B} \right)' \left(A^{\frac{1}{2}} B - A^{\frac{1}{2}} \bar{B} \right) + (Y - XB)'(Y - XB) \right) \right\} \end{aligned} \quad (3.4.5)$$

其中 $A = \sigma^2 \bar{\Sigma}_B^{-1}$, 令 $W = \begin{pmatrix} A^{\frac{1}{2}} \bar{B} \\ Y \end{pmatrix}_{(k+n) \times n}$ $G = \begin{pmatrix} A^{\frac{1}{2}} \\ X \end{pmatrix}_{(k+n) \times n}$, 于是有:

$$g(B | Y) = \exp \left\{ -\frac{1}{2\sigma^2} (W - GB)'(W - GB) \right\} \quad (3.4.6)$$

用 \bar{B} 表示待估参数后验均值, $\bar{\Sigma}_B$ 示待估参数后验协方差矩阵。并且应用下结论:

$$(W - GB)'(W - GB) = (B - \bar{B})' G' G (B - \bar{B}) + (W - G\bar{B})'(W - G\bar{B}) \quad (3.4.7)$$

其中

$$\begin{aligned} \bar{B} &= (G'G)^{-1} G'W = (A + X'X)^{-1} (A\bar{B} + X'X\beta) \\ \beta &= (X'X)^{-1} X'Y \end{aligned}$$

β 是 B 的 OLS 估计值。将(3.4.7)代入(3.4.6), 代入时舍去(3.4.7)的右边第二项, 因为其不含 B , 可以作为常数项处理, 只考虑核。得到

$$\begin{aligned} g(B | Y) &= \exp \left\{ -\frac{1}{2\sigma^2} (B - \bar{B})' G' G (B - \bar{B}) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (B - \bar{B})' (A + X'X) (B - \bar{B}) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (B - \bar{B})' \left(\bar{\Sigma}_B^{-1} + X'X/\sigma^2 \right) (B - \bar{B}) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (B - \bar{B})' \left(\bar{\Sigma}_B^{-1} \right) (B - \bar{B}) \right\} \end{aligned} \quad (3.4.8)$$

$$\bar{\Sigma}_B^{-1} = \bar{\Sigma}_B^{-1} + X'X/\sigma^2 \quad (3.4.9)$$

$$\begin{aligned} \bar{B} &= (A + X'X)^{-1} (A\bar{B} + X'X\beta) \\ &= \left(\bar{\Sigma}_B^{-1} + X'X/\sigma^2 \right)^{-1} \left(\bar{\Sigma}_B^{-1} \bar{B} + (X'X/\sigma^2) \beta \right) \\ &= \bar{\Sigma}_B^{-1} \left(\bar{\Sigma}_B^{-1} \bar{B} + (X'X/\sigma^2) \beta \right) \end{aligned} \quad (3.4.10)$$

于是, 3.4.8 正好是均值为 $\bar{\bar{B}}$ 、方差为 $\bar{\bar{\Sigma}}_B$ 的多元正态分布的核, 即 B 的后验密度函数为

$$(B | Y) \sim N(\bar{\bar{B}}, \bar{\bar{\Sigma}}_B) \quad (3.4.11)$$

将协方差矩阵的逆 $\bar{\bar{\Sigma}}_B^{-1}$ 定义为精确度矩阵, 那么可以看出: (1) 后验精确度矩阵 $\bar{\bar{\Sigma}}_B^{-1}$ 是先验精确度矩阵 $\bar{\Sigma}_B^{-1}$ 与样本信息精确度矩阵 $X'X/\sigma^2$ 之和, 故后验精确度总是高于先验精确度; (2) 后验均值 $\bar{\bar{B}}$ 是先验均值 \bar{B} 与样本信息 OLS 估计值 β 的加权平均和, 权数为各自的精确度。

2. 无先验信息的后验分布

在对待估参数一无所知的情况下, 仍然可以用贝叶斯方法求得待估参数的后验分布。这时可以认为待估参数的所有元素服从上的均匀分布, 且互不相关。于是, B 的先验密度函数为

$$g(B) = g(\beta_1) \cdot g(\beta_2) \cdots g(\beta_k) = c$$

B 的或然函数仍然与 (3.4.4) 相同, 则后验密度函数与或然函数形式相同:

$$\begin{aligned} g(B | Y) &= g(B) \cdot L(B | Y) \approx L(B | Y) \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (Y - XB)'(Y - XB) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} ((B - \beta)'X'X(B - \beta) + (Y - XB)'(Y - XB)) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (B - \beta)'X'X(B - \beta) \right\} \end{aligned} \quad (3.4.12)$$

得到的 B 的后验密度函数仍然是正态的, 均值为 β , 协方差矩阵为 $\sigma^2 (X'X)^{-1}$

$$(B | Y) \sim N(\beta, \sigma^2 (X'X)^{-1}) \quad (3.4.13)$$

从形式上看, 无信息先验得到的后验分布均值与样本信息的 OLS 估计相同, 但二者有不同的含义。贝叶斯结论 3.4.13 中的 B 是随机的, 而均值 β 在样本确定后是固定的; 样本信息的结论中, B 作为期望值, 而 β 是随机变量, 即有 $\beta \sim N(\beta, \sigma^2 (X'X)^{-1})$ 。

事实上, 也可以直接由 3.4.8、3.4.9、3.4.10 得到无信息先验下的后验密度, 只要把无信息先验作为有信息先验的一种特殊情况, 即无信息先验的精确度为 0。即

令 $\bar{\Sigma}_B^{-1} = 0$, 代入 3.4.9、3.4.10 可得到

$$\bar{\bar{\Sigma}}_B^{-1} = X'X/\sigma^2$$

从而 $\bar{\bar{\Sigma}}_B = \sigma^2 X'X$

$$\bar{\bar{B}} = (X'X/\sigma^2)^{-1} \cdot (X'X/\sigma^2) \cdot \beta = \beta$$

所以

$$(B | Y) \sim N(\beta, \sigma^2 (X'X)^{-1})$$

3. 点估计

在得到贝叶斯估计的后验密度函数后, 即可以此为出发点, 进行点估计。其思路是利用损失函数并使平均损失最小。为此需要确定一个损失函数。

先讨论一般情况。假设 \hat{B} 为 B 的估计量, 损失函数为 $Loss(B, \hat{B})$ 表示参数为 B 时采用 \hat{B} 为估计量所造成的损失, 故总存在 $Loss(B, \hat{B}) \geq 0$ 于是满足 $\min[Loss(B, \hat{B})]$ 的 \hat{B} 即是需要的估计量。由于损失函数依赖样本, 故取加权平均损失, 权数为后验密度函数, 表示为 $g(B | Y)$

$$\mathbb{E}(Loss(B, \hat{B}) | Y) = \int Loss(B, \hat{B}) \cdot g(B | Y) dB \quad (3.4.14)$$

能使3.4.14所表示的加权平均后验损失最小的 \hat{B} 的值, 即为 B 的贝叶斯点估计值。

下面以二次损失函数为例说明点估计过程。损失函数的二次型为

$$Loss = (B - \hat{B})' M (B - \hat{B}) \quad (3.4.15)$$

M 为一个正定矩阵, \hat{B} 是 B 的一个估计量。为使 $\mathbb{E}(Loss)$ 最小,

$$\begin{aligned} \mathbb{E}(Loss) &= \mathbb{E}(B - \hat{B})' M (B - \hat{B}) \\ &= \mathbb{E}\{(B - \mathbb{E}(B)) - (\hat{B} - \mathbb{E}(B))\}' M \{(B - \mathbb{E}(B)) - (\hat{B} - \mathbb{E}(B))\} \\ &= \mathbb{E}\{(B - \mathbb{E}(B))' M (B - \mathbb{E}(B)) + (\hat{B} - \mathbb{E}(B))' M (\hat{B} - \mathbb{E}(B))\} \end{aligned} \quad (3.4.16)$$

上述推导过程中出现的交叉项为 0, 即

$$\begin{aligned} \mathbb{E}\{(B - \mathbb{E}(B))' M (\hat{B} - \mathbb{E}(B))\}' &= \mathbb{E}\{(B - \mathbb{E}(B))' M (\hat{B} - \mathbb{E}(B))\} \\ &= 0 \cdot M \cdot (\hat{B} - \mathbb{E}(B)) = 0 \end{aligned}$$

考察3.4.16式, 第1项不含 \hat{B} , 又因为 M 为一个正定矩阵, 故第2项

$$(\hat{B} - \mathbb{E}(B))' M (\hat{B} - \mathbb{E}(B)) \geq 0$$

所以使3.4.16式最小的 \hat{B} 为 $\hat{B} = \mathbb{E}(B)$ 。即二次损失函数的点估计值为后验均值。

4. 区间估计

类似于点估计, 可以根据 B 的后验密度函数进行区间估计。这里需要引入最高后验密度区间的概念: 区间内每点的后验密度函数值大于区间外任何一点的后验密度函数值, 这样的区间称为最高后验密度区间 (HPD 区间)。

对于单参数 ($k = 1$) 模型的区间估计, 可参照下式:

$$\int_a^b g(\boldsymbol{\beta} | \mathbf{Y}) d\boldsymbol{\beta} = 1 - \alpha \quad (3.4.17)$$

其中 $(1 - \alpha)$ 为置信水平。这种区间估计与经典样本信息理论中的区间估计是一致的。特别在正态分布情况下, 只要稍作变换, 查正态分布表便可很容易地得到结论。

对于多参数 ($k > 1$) 模型的区间估计, 可根据 \mathbf{B} 的后验密度函数, 求出 \mathbf{B} 的每一元素的边际后验密度函数, 再按照单参数情况求出每一参数的最高后验密度区间。

需要说明的是, 参数的最高后验密度区间在形式上与经典样本信息理论中的置信区间是一致的, 但解释并不相同。在贝叶斯估计中参数的最高后验密度区间的含义是参数以 $(1 - \alpha)$ 的概率位于该区间内; 经典样本信息理论中的置信区间的含义是该区间以 $(1 - \alpha)$ 的概率包含参数。含义上的区别源于是否把待估参数作为随机变量。

5. 假设检验

可以用上述的最高后验密度区间进行假设检验, 但常用的方法是利用后验优势比检验。

Part II

线性

在引论中，我们推出了满足凯恩斯条件的消费函数与收入有关的一个最普通模型： $C = \alpha + \beta X + \varepsilon$ 其中 $\alpha > 0, 0 < \beta < 1$ ， ε 是一个随机扰动。这是一个标准的古典线性回归模型。假如我们得到如下表 4.1 的数据

表 4.1 可支配个人收入和个人消费支出

年份	可支配收入	个人消费
1971	779.2	696.8
1972	810.3	737.1
1973	864.7	767.9
1974	857.5	762.8
1975	847.9	779.4
1976	906.8	823.1
1977	942.9	964.3
1978	988.8	903.2
1979	1015.7	927.6

§ 4.1 线性回归模型及其假定

一般地，被估计模型具有如下形式：

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n$$

其中 y 是因变量或称为被解释变量， x 是自变量或称为解释变量， i 标志 n 个样本观测值中的一个。这个

形式一般被称作 y 对 x 的总体线性回归模型。在此背景下, y 称为被回归量, x 称为回归量

§ 4.1.1 构成古典线性回归模型的一组基本假设

- 1) 函数形式: $y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n$ 。
- 2) 干扰项的零均值: 对所有 i , 有: $\mathbb{E}[\varepsilon_i] = 0$ 。
- 3) 同方差性: 对所有 i , 有: $\text{Var}[\varepsilon_i] = \sigma^2$, 且 σ^2 是一个常数。
- 4) 无自相关: 对所有 $i \neq j$, 则 $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0$ 。
- 5) 回归量和干扰项的非相关: 对所有 i 和 j 有 $\text{Cov}[x_i, \varepsilon_j] = 0$ 。
- 6) 正态性: 对所有 $\varepsilon_i, \varepsilon_j$ 满足正态分布 $N[0, \sigma]$ 。

§ 4.1.2 模型假定的几点说明

1. 函数形式及其线性模型的转换

具有一般形式:

$$f(y_i) = \alpha + \beta g(x_i) + \varepsilon_i$$

对任何形式的 $g(x)$ 都符合我们关于线性模型的定义。

Example 4.1 (一个常用的函数形式是对数线性模型)

$$y = Ax^\beta$$

取对数得: $\ln y = \alpha + \beta \ln x$ ($\alpha = \ln A$) 这被称作不变弹性形式。在这个方程中, y 对于 x 的变化的弹性是:

$$\eta = \frac{dy/y}{dx/x} = \frac{d \ln y}{d \ln x} = \beta$$

它不随 x 而变化。与之相反, 线性模型的弹性是:

$$\eta = \left(\frac{dy}{dx} / \frac{y}{x} \right) = \left(\frac{x}{\alpha + \beta x} \right) \left(\frac{dy}{dx} \right) = \frac{\beta x}{\alpha + \beta x}$$

对数线性模型通常用来估计需求函数和生产函数, 尽管线性模型具有巨大的灵活性, 但在实际中存在着大量的非线性模型的形式。任何变换也不能将转化为线性回归模型。

$$y = \alpha + \frac{1}{\beta + x} \quad \& \quad y = \alpha + \beta x^v \quad (0 < v < 1)$$

2. 回归量

对于回归量即解释变量我们有两种处理方法，第一种将 X 设定为非随机变量，第二种方法将 X 设定为随机变量。

1) 当 X 为非随机变量

x_i 的值在 y_i 的概率分布中是已知的常数。这条假定暗示 y_i 的每一个值都是一个概率分布的观察值，这个概率分布具有均值和方差。

$$\mathbb{E}[y_i | x_i] = E[\alpha + \beta x_i + \varepsilon_i] = \alpha + \beta x_i + \mathbb{E}[\varepsilon_i] = \alpha + \beta x_i$$

$$\text{Var}[y_i | x_i] = \text{Var}[\alpha + \beta x_i + \varepsilon_i] = \text{Var}[\varepsilon_i] = \sigma^2$$

此外，有必要假定，对 $n \leq 1$

$$\left(\frac{1}{n}\right) S_{xx} = \left(\frac{1}{n}\right) \sum_i (x_i - \bar{x})^2$$

是一个有限正数，这个假定被称作识别条件，若 x_i 没有任何变化，我们所有的观测值将落在一条垂直线上，我们的观测数据将不允许我们作出关于回归 $\alpha + \beta \mathbf{X}$ 的任何推断。这个识别条件等同于子样的极差 $\max(X_1, \dots, X_n) - \min(X_1, \dots, X_n) \neq 0$

2) 当 X 为随机变量

若 \mathbf{x} 被当作一个随机变量，则假定 1 成为一个对 \mathbf{y} 和 \mathbf{x} 的联合分布的陈述。我们就用条件期望和方差来处理。

§ 4.1.3 随机干扰项

如果干扰项不是零均值，即 $\mathbb{E}[\varepsilon_i] = \mu$ ，对所有的 i ，则 $\alpha + \beta x + \varepsilon_i$ 等同于 $(\alpha + \mu) + \beta x + (\varepsilon_i - \mu)$ ，令 $\alpha' = \alpha + \mu$ 及 $\varepsilon'_i = \varepsilon_i - \mu$ 可得到模型，此模型满足我们原始模型的要求。

观测值中的随机部分假定是不相关的： $\mathbb{E}[\varepsilon_i, \varepsilon_j] = 0$ ，对所有 i 不等于 j ，这被称为非自相关。

§ 4.2 最小二乘法

§ 4.2.1 最小二乘系数

总体回归是 $\mathbb{E}[y_i | x_i] = \alpha + \beta x_i$ ，而我们对 $\mathbb{E}[y_i | x_i]$ 的估计记作

$$\hat{y}_i = a + bx_i$$

和第 i 的数据点相联系的干扰项是:

$$\varepsilon_i = y_i - \alpha - \beta x_i$$

对 a 和 b 的任何值, 我们用残差:

$$e_i = y_i - a - bx_i$$

来估计 ε_i , 从这些定义可知:

$$y_i = \alpha + \beta x_i + \varepsilon_i = a + bx_i + e_i$$

对任何一对值 a 和 b , 残差平方和是:

$$\sum_i e_i^2 = \sum_i (y_i - a - bx_i)^2$$

最小二乘法系数就是使这个拟合标准达到最小的 a 和 b 的值。最小化的一阶条件是

$$\begin{aligned} \frac{\partial (\sum_i e_i^2)}{\partial a} &= \sum_i 2(y_i - a - bx_i)(-1) \\ &= -2 \sum_i (y_i - a - bx_i) = 0 \\ \frac{\partial (\sum_i e_i^2)}{\partial b} &= \sum_i 2(y_i - a - bx_i)(-x_i) = 0 \\ &= -2 \sum_i x_i (y_i - a - bx_i) = 0 \end{aligned}$$

将上两式展开合并同类项后得到正规方程组:

$$\sum_i y_i = na + \left(\sum_i x_i \right) b \quad (4.2.1)$$

$$\sum_i x_i y_i = \left(\sum_i x_i \right) a + \left(\sum_i x_i^2 \right) b \quad (4.2.2)$$

(4.2.1) 暗示 $\sum_{i=1}^n e_i = 0$, 而 (4.2.2) 暗示 $\sum_i X_i e_i = 0$

为了得到解, 我们首先用 n 除 (4.2.1), 结果是 $\bar{y} = a + b\bar{x}$ 。最小二乘回归线通过均值点。现在分离 a :

$$a = \bar{y} - b\bar{x} \quad (4.2.3)$$

有了 a 后, 我们可以求解(4.2.1)得到 b 。首先, $\sum_i x_i = n\bar{x}$ 。将此和(4.2.3)代入(4.2.2)并重新安排各项。

$$\sum_i x_i y_i - n\bar{x}\bar{y} = b \left(\sum_i x_i^2 - n\bar{x}^2 \right)$$

或

$$b = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

最小的残差平方和，对 a 和 b 的二阶微分矩阵是：

$$\begin{bmatrix} \partial^2 (\sum_i e_i^2) / \partial a^2 & \partial^2 (\sum_i e_i^2) / \partial a \partial b \\ \partial^2 (\sum_i e_i^2) / \partial b \partial a & \partial^2 (\sum_i e_i^2) / \partial b^2 \end{bmatrix} = \begin{bmatrix} 2n & 2\sum_i x_i \\ 2\sum_i x_i & 2\sum_i x_i^2 \end{bmatrix}$$

我们必须表明这是一个正定矩阵，两个对角元素永远为正，所以仅需证明行列式为正， $(4n)\sum_i x_i^2 - 4(\sum_i x_i)^2$ ，但 $\sum_i x_i = n\bar{x}$ ，所以行列式为

$$4n \left(\sum_i x_i^2 - n\bar{x}^2 \right) = 4n \left(\sum_i (x_i - \bar{x})^2 \right)$$

由识别条件得知这是一个正值。这样 a 和 b 是平方和的最小化因子。

§ 4.2.2 回归拟合的评价

1) 回归量 x 是非随机变量

总变差是离差的平方和：

$$\begin{aligned} SST &= \sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i^2 + 2 \sum_i e_i \hat{y}_i \\ &= b^2 \sum_i (x_i - \bar{x})^2 + \sum_i e_i^2 \end{aligned}$$

第二个等式成立是因为： $\sum_{i=1}^n e_i \hat{y}_i = \sum_{i=1}^n e_i (a + bX_i) = a \sum e_i + b \sum_{i=1}^n e_i X_i = 0$

我们将其写作：

$$SST = SSR + SSE$$

我们利用下式得到一个关于回归直线对数据拟合程度的度量：

$$R^2 = \frac{SSR}{SST}$$

为了方便计算与分析，约定

$$\begin{aligned} S_{xx} &= \sum (x_i - \bar{x})^2, & S_x &= \sqrt{S_{xx}} \\ S_{yy} &= \sum (y_i - \bar{y})^2, & S_y &= \sqrt{S_{yy}} \\ S_{xy} &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

x 和 y 间的样本相关系数是： $r_{xy} = S_{xy} / (S_x S_y)$ ，利用 $b = S_{xy} / S_{xx}$ 我们得到 $r_{xy} = b / (S_y / S_x)$ 这表明回归的斜率和 x 、 y 间的相关系数具有相同的符号，而且

$$R^2 = \frac{SSR}{SST} = \frac{b^2 S_{xx}}{S_{yy}} = r_{xy}^2$$

这进一步证明了我们利用 R^2 作为回归模型拟合优劣指标的正确性。

§ 4.2.3 方差分析表 (ANOVA)

进一步研究回归平方和 SSR 与残差平方和 SSE，我们可以得到下面三个结论：

- a) 在 $\beta = 0$ 的假设条件下，回归平方和 $\frac{SSR}{\sigma^2}$ 服从自由度为 1 的卡方分布 $\chi^2(1)$
- b) 残差平方和 $\frac{SSE}{\sigma^2}$ 服从自由度为 $n-2$ 的卡方分布 $\chi^2(n-2)$
- c) 在 $\beta = 0$ 的假设条件下， $\frac{SSR/1}{SSE/(n-2)}$ 服从 $F(1, n-2)$ 分布。现在我们来证明这三个结论

Proof 4.1

$$\begin{aligned}
 \text{Proof(a): } c_i &= \frac{x_i - \bar{x}}{S_{xx}}, \quad b = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \bar{x}) y_i}{S_{xx}} = \sum_i c_i y_i \\
 &\Rightarrow \sum_i c_i^2 = \frac{1}{S_{xx}}, \quad \text{Suppose } \mathbf{C} = (c_1, c_2, \dots, c_n)' \\
 &\Rightarrow b = \mathbf{C}'\mathbf{Y}, \quad b^2 = \mathbf{Y}'\mathbf{C}\mathbf{C}'\mathbf{Y}, \quad SSR = b^2 \sum (x_i - \bar{x})^2 = S_{xx} b^2 = \mathbf{Y}' S_{xx} \mathbf{C}\mathbf{C}' \mathbf{Y}
 \end{aligned}$$

可以验证 $S_{xx} \mathbf{C}\mathbf{C}'$ 是幂等矩阵。

$$\begin{aligned}
 S_{xx} \mathbf{C}\mathbf{C}' \cdot S_{xx} \mathbf{C}\mathbf{C}' &= S_{xx}^2 \mathbf{C} (\mathbf{C}'\mathbf{C}) \mathbf{C} = S_{xx} \mathbf{C}\mathbf{C}' \\
 r(S_{xx} \mathbf{C}\mathbf{C}') &= \text{tr}(S_{xx} \mathbf{C}\mathbf{C}') = S_{xx} \sum_i c_i^2 = 1
 \end{aligned}$$

在 $\beta = 0$ 的假设条件下， $\frac{SSR}{\sigma^2}$ 才服从自由度为 1 的卡方分布 $\chi^2(1)$

Proof (b): 因为 $SST = \mathbf{Y}'\mathbf{M}_0\mathbf{Y}$ 及 $SST = SSR + SSE \Rightarrow SSE = \mathbf{Y}'(\mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}')\mathbf{Y}$

已验证 $\mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}'$ 也是幂等矩阵, $\mathbf{C}'\mathbf{i} = \mathbf{i}'\mathbf{C} = \sum_i c_i = 0$ 。

$$\begin{aligned}
 (\mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}')^2 &= \mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}'\mathbf{M}_0 - S_{xx} \mathbf{M}_0 \mathbf{C}\mathbf{C}' + S_{xx} \mathbf{C}\mathbf{C}' \\
 &= \mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}' + \frac{1}{n} S_{xx} \mathbf{C}\mathbf{C}' \mathbf{i}\mathbf{i}' + \frac{1}{n} S_{xx} \mathbf{i}\mathbf{i}' \mathbf{C}\mathbf{C}' = \mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}' \\
 &\Rightarrow r(\mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}') = \text{tr}(\mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}') = n - 1 - S_{xx} \cdot \frac{1}{S_{xx}} = n - 2
 \end{aligned}$$

因此 $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$ 此结论成立不需要 $\beta = 0$ 的假设条件下，为什么？

所以 SSR 与 SSE 是相互独立的统计量。从而，在 $\beta = 0$ 的假设条件下， $\frac{SSR/1}{SSE/(n-2)}$ 服从 $F(1, n-2)$ 分布，

所以，可以用来作模型的整体检验的统计量。

$$\begin{aligned}
 \text{Proof(c): } S_{xx} \mathbf{C}\mathbf{C}' \cdot (\mathbf{M}_0 - S_{xx} \mathbf{C}\mathbf{C}') &= S_{xx} \mathbf{C}\mathbf{C}' \cdot \left(\mathbf{I} - \frac{1}{n} \mathbf{i}\mathbf{i}' - S_{xx} \mathbf{C}\mathbf{C}' \right) \\
 &= S_{xx} \mathbf{C}\mathbf{C}' - \frac{1}{n} S_{xx} \mathbf{C}\mathbf{C}' \mathbf{i}\mathbf{i}' - S_{xx} \mathbf{C}\mathbf{C}' = 0
 \end{aligned}$$

表 4.2 方差分析表

变差来源	变差	自由度	均方
回归	$SSR = b^2 S_{xx}$	1	$\frac{SSR}{1}$
残差	$SSE = \sum_i e_i^2$	$n - 2$	$\frac{SSE}{n-2}$
总	$SST = S_{yy}$	$n - 1$	$\frac{S_{yy}}{n-1}$
$F[1, n - 2] = \frac{SSR/1}{SSE/(n - 2)}$			

概括这些计算的一个方便的途径是方差分析表，可总结在方差分析表4.2中。

2) 回归量 X 是随机变量

我们要利用方差分解公式:

$$\begin{aligned}
 \text{Var}(Y) &= \text{Var}(\mathbb{E}(Y | X)) + \mathbb{E}(\text{Var}_x(Y | X)) \\
 &= \text{Var}(\alpha + \beta X) + \mathbb{E}[\mathbb{E}(Y - \mathbb{E}(Y | X))^2 | X] \\
 &= \beta^2 \text{Var}(X) + \mathbb{E}[\mathbb{E}(Y - \mathbb{E}(Y | X))^2 | X]
 \end{aligned}$$

我们将它应用到子样空间里来，即

$$\frac{1}{n} \sum_i (y_i - \bar{y})^2 = b^2 \frac{1}{n} \sum_i (x_i - \bar{x})^2 + \frac{1}{n} \sum_i e_i^2$$

所以，两边去掉 $1/n$ 后得到:

$$\sum_i (y_i - \bar{y})^2 = b^2 \sum_i (x_i - \bar{x})^2 + \sum_i e_i^2$$

对消费和收入数据，方差分析表如下所示:

表 4.3 方差分析表

变差来源	变差	自由度	均方
回归	64,435.15	1	64,435.13
残差	537.00	8	67.124
总	64,972.13	9	7,219.12
$F[1, 8] = \frac{64,435.13}{67.124} = 959.94$			

我们得到了和把 X 当成非随机变量时同样的结果，因此，方差分析表也是一样的。考虑消费函数的例子，这里 C 是消费而 X 是收入，我们得到: $\bar{C} = 793.43$, $\bar{X} = 879.24$, $S_{CC} = 64,972.12$, $S_{XX} = 67,192.44$,

$S_{XC} = 65,799.34$ 。 $SST = 64,972.12$, $SSR = 64,435.13$, $SSE = 537.00 \implies R^2 = \frac{64,435.13}{64,972.12} = 0.99173$ 。显然, 此回归提供了一个很好的拟合。

另一个计算和通常 R^2 相类似公式是:

$$R^2 = 1 - \frac{\sum_i e_i^2}{S_{yy}}$$

任何一个模型的残差都可用 $y_i - \hat{y}$ 来计算。

§ 4.3 最小二乘法估计量的统计特征

我们利用了最小二乘法, 从纯粹的代数方法, 求得所拟合的最小二乘系数 a 和 b , 从统计意义上来说, 这个结果可以看作是对参数 α 和 β 的一个估计 (因为还存在着利用其他估计方法得到的估计)。我们现在对 a 、 b 的无偏性, 有效性和精确度等统计特性作分析。

我们所考虑的计量模型是:

$$y_i = \alpha + \beta x + \varepsilon_i$$

β 的最小二乘估计是:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \bar{x}) y_i}{S_{xx}} = \sum_i c_i y_i \quad (4.3.1)$$

其中权数 c_i 仅仅是 x_1, \dots, x_n 的一个函数。

$$c_i = \frac{x_i - \bar{x}}{S_{xx}} \quad (4.3.2)$$

§ 4.3.1 b 是 β 的无偏估计

将 $y_i = \alpha + \beta x + \varepsilon_i$ 代入 4.3.1, 我们得到

$$\begin{aligned} b &= \frac{\sum_i (x_i - \bar{x}) (\alpha + \beta x_i + \varepsilon_i)}{S_{xx}} \\ &= \frac{\alpha \sum_i (x_i - \bar{x})}{S_{xx}} + \frac{\beta \sum_i (x_i - \bar{x}) x_i}{S_{xx}} + \frac{\sum_i (x_i - \bar{x}) \varepsilon_i}{S_{xx}} \\ &= \beta + \sum_i c_i \varepsilon_i \end{aligned} \quad (4.3.3)$$

所以

$$\mathbb{E}[b] = \beta + \mathbb{E} \left[\sum_i c_i \varepsilon_i \right] = \beta \quad (4.3.4)$$

这是因为 $\mathbb{E}[\varepsilon_i] = 0$ 。不论 ε 的分布如何, 在我们其他假定下, b 是 β 的一个无偏估计量, 利用 4.3.3 得到 b 的样本方差

$$\text{Var}[b] = \text{Var}[b - \beta] = \text{Var} \left[\sum_i c_i \varepsilon_i \right]$$

特别要注意 b 的方差中的分母。 x 的变差越大（也就是 x 的采样范围越广），则这个方差越小。

§ 4.3.2 a 是 α 的无偏估计

对于最小二乘截距 a ，我们有：

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ &= \frac{1}{n} \sum_i y_i - b\bar{x} = \frac{1}{n} \sum_i (\alpha + \beta x_i + \varepsilon_i) - b\bar{x} \end{aligned}$$

利用4.3.3式并加以整理，我们有

$$a - \alpha = \sum_i d_i \varepsilon_i$$

其中

$$d_i = \left(\frac{1}{n} - \bar{x}c_i \right)$$

由于求和中每一项的期望都为 0，所以 a 也是 α 的估计量无偏估计量。 a 的样本方差就是 $\sum_i d_i \varepsilon_i$ 的方差，根据独立性有

$$\text{Var}[a] = \sum \sigma^2 d_i^2 = \sigma^2 \left(\frac{1}{n} + \bar{x}^2 \sum_i c_i^2 \right) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

（通过对括号中的项进行平方并利用 $\sum_i c_i = 0$ 的结果，可以得到上式中后一结果）。

§ 4.3.3 a, b 估计量的协方差矩阵

两个估计的协方差是：

$$\begin{aligned} \text{Cov}[a, b] &= \mathbb{E}[(a - \alpha)(b - \beta)] = \mathbb{E} \left[\left(\sum_i d_i \varepsilon_i \right) \left(\sum_i c_i \varepsilon_i \right) \right] \\ &= \sigma^2 \sum_i c_i d_i = \frac{-\bar{x}\sigma^2}{S_{xx}} \end{aligned}$$

a 和 b 两者都有 $\sum_i w_i y_i$ 的形式，因此它们都是线性估计量，前边给出了它们的样本均值和方差并证实了它们是无偏的。正如已指出的，还存在利用数据估计 α 和 β 的其他方法。然而，从线性无偏估计量的角度，没有任何估计量比最小二乘估计量具有更小的样本方差，这就是高斯—马尔科夫定理。

当把正态分布干扰项的假定加入上面的过程时，我们得到估计量的分布的一个完备的结果。由于 a 和 b 两者都是正态分布变量的线性函数，因而它们也都是正态分布的。其均值和方差已导出，概括起来，在正态性假设下，有：

$$\begin{bmatrix} a \\ b \end{bmatrix} \sim N \left[\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \sigma^2 \begin{bmatrix} 1/n + \bar{x}^2/S_{xx} & -\bar{x}/S_{xx} \\ -\bar{x}/S_{xx} & 1/S_{xx} \end{bmatrix} \right]$$

§ 4.3.4 b 是 β 的最小线性无偏估计

思考：证明 $b = \sum_i c_i y_i$ 是线性无偏估计量中，方差最小的一个估计量。

Proof 4.2 令另一个估计量是

$$b' = \sum_i q_i y_i = \alpha \sum_i q_i + \beta \sum_i q_i x_i + \sum_i q_i \varepsilon_i$$

在等式两边取期望，我们可以看到，若使 b' 是无偏的，必须有 $\sum_i q_i = 0$ 及 $\sum_i q_i x_i = 1$ 。这样， $b' = \beta + \sum_i q_i \varepsilon_i$ 。 b' 的方差是：

$$\text{Var}[b'] = \sigma^2 \sum_i q_i^2$$

令 $v_i = q_i - c_i$ ，则 $q_i = c_i + v_i$ 且

$$\text{Var}[b'] = \sigma^2 \sum_i (c_i + v_i)^2 = \sigma^2 (\sum_i c_i^2 + \sum_i v_i^2 + 2 \sum_i c_i v_i)$$

利用 $\sum_i q_i = 0$ 和 $\sum_i q_i x_i = 1$ ，易得到 $\sum_i c_i v_i = 0$ ，这就是在 b' 的方差中只留下两个平方项，这意味着 $\text{Var}[b']$ 一定大于 $\text{Var}[b]$ 。

Why $\sum_i c_i v_i = 0$?

$$\begin{aligned} \Rightarrow \sum_i c_i v_i &= \sum_i (q_i - c_i) c_i = \sum_i c_i q_i - \sum_i c_i^2 \\ &= \frac{\sum_i x_i q_i - \sum_i q_i \bar{x}}{S_{xx}} - \frac{1}{S_{xx}} = \frac{1}{S_{xx}} - \frac{1}{S_{xx}} = 0 \end{aligned}$$

§ 4.4 最小二乘估计量的统计推断

在前面的内容里，我们在假定干扰项是正态分布和样本 X_1, X_n 是非随机的条件下，给出了最小二乘估计量的确切的样本分布。但通常的参数估计过程包括构造置信区间和对 α 和 β 值的假设检验。为了做到这一点，我们需要参数的真正样本方差的估计，这将需要对未知参数 σ^2 的一个估计，并构造假设检验方法。

§ 4.4.1 σ^2 的无偏估计量的推导

由于 σ^2 是 ε_i^2 的期望值，而 e_i 是 ε_i 的一个估计，

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i e_i^2$$

似乎是一个自然的估计量，通过写出 $e_i = y_i - a - bx_i$ ，并把 $y_i = \alpha + \beta x_i + \varepsilon_i$ ， $\alpha = \bar{y} - \beta \bar{x} - \bar{\varepsilon}$ 和 $a = \bar{y} - b\bar{x}$ 代入，我们得到

$$\begin{aligned} e_i &= \varepsilon_i - \bar{\varepsilon} - (x_i - \bar{x})(b - \beta) \\ &= \varepsilon_i - \bar{\varepsilon} - (x_i - \bar{x}) \left(\sum_j c_j \varepsilon_j \right) \end{aligned} \quad (4.4.1)$$

我们对某一个别干扰项 ε_i 的估计受两种因素的扭曲：所有干扰项的样本平均和我们可以归于 β 并非完美估计这一事实所造成的影响。回忆所有干扰项是独立的，所以 $\mathbb{E}(\varepsilon_i \varepsilon_j) = 0$ 若 $i \neq j$ 。现在我们平方的两边并取期望值，可得到

$$\begin{aligned} \mathbb{E}[e_i^2] &= \sigma^2 + \frac{\sigma^2}{n} + \sigma^2 (x_i - \bar{x})^2 \left(\sum_j c_j^2 \right) - \frac{2\sigma^2}{n} \\ &\quad - 2\sigma^2 (x_i - \bar{x}) c_i + \frac{2\sigma^2}{n} (x_i - \bar{x}) \left(\sum_j c_j \right) \end{aligned}$$

在对这些项求和时，我们利用 $\sum_i c_i = 0$, $\sum_i (x_i - \bar{x}) c_i = 1$ 和 $\sum_i c_i^2 = \frac{1}{S_{xx}}$ 。整理后，我们有

$$E \left[\sum_i e_i^2 \right] = (n-2)\sigma^2$$

这表明 σ^2 的一个无偏估计量是：

$$s^2 = \frac{\sum_i e_i^2}{n-2}$$

这样，我们可以得到 b 的抽样方差的一个估计为：

$$\text{Est.Var}[b] = \frac{s^2}{S_{xx}}$$

以后，我们将用记号 $\text{Est.Var}[\cdot]$ 表示一个估计量的抽样方差的一个样本估计。

t 分布统计量的构造

$$z = \frac{b - \beta}{\sqrt{\sigma^2 / S_{xx}}} \quad (4.4.2)$$

的分布是标准正态。由 $\frac{SSE}{\sigma^2}$ 服从 $\chi^2(n-2)$

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2) \quad (4.4.3)$$

并且和 b 是独立的。

根据 4.4.2 和 4.4.3，我们得到：

$$t = \frac{b - \beta}{\sqrt{s^2 / S_{xx}}}$$

是一个标准正态变量和一个除以其自由度的卡方量的平方根之比，它服从自由度为 $(n-2)$ 的 t 分布。这样，记 $s_b = \frac{S}{\sqrt{S_{xx}}} = \frac{S}{S_x}$

$$\frac{b - \beta}{s_b} \sim t[n-2] \quad (4.4.4)$$

可以形成统计推断的基础。

§ 4.4.2 抽样分布

β 的置信区间将以4.4.4为基础。特别的，我们可以有

$$P\left(b - t_{\frac{\lambda}{2}} s_b \leq \beta \leq b + t_{\frac{\lambda}{2}} s_b\right) = 1 - \lambda$$

其中 $1-\lambda$ 是要求的置信水平， $t_{\lambda/2}$ 是来自于自由度为 $(n-2)$ 的 t 分布的适当的临界值。利用 a 及其估计方差，可以同样地构造 α 的置信区间。

§ 4.4.3 β 假设检验

我们也可以构造干扰项方差 σ^2 的置信区间，利用 4.4.3 和前边的同样推理，我们得到 σ^2 的 95% 置信区间是：

$$\left[\frac{(n-2)s^2}{\chi_{0.975}^2}, \frac{(n-2)s^2}{\chi_{0.025}^2} \right]$$

一个相关的过程是检验参数是否取一给定值，为了检验假设： $H_0: \beta = \beta^0; H_1: \beta \neq \beta^0$

最简单的过程是利用我们的置信区间，置信区间给出了在给定样本数据情况下， β 的一个似乎可能的值的集合，如果这个集合不包含 β^0 ，则原假设应该被拒绝。在原假设下，比率：

$$t = \frac{b - \beta^0}{s_b}$$

服从自由度为 $(n-2)$ 的 t 分布，其均值为 0。这个比率在任何尾部的极端值都将使假设值得怀疑。这样，一般地，若

$$\frac{|b - \beta^0|}{s_b} \geq t_{\frac{\lambda}{2}}$$

我们将拒绝 H_0 。这里， $t_{\frac{\lambda}{2}}$ 是来自于自由度为 $(n-2)$ 的 t 分布的 $100(1-\lambda/2)\%$ 临界值。

Example 4.2 在前边的回归中，我们得到 $a = -67.5806$ 和 $b = 0.9793$ 。为了计算标准误差，我们需要

$$s^2 = \frac{537.00}{8} = 67.125, \quad S_{xx} = 67,192.45$$

$$\bar{x} = 679.24, \quad s_a = 27.91, \quad s_b = 27.91$$

对一个自由度为 $n-2=8$ 的分布，95% 临界值是 2.306。所以， α 和 β 的 95% 置信区间分别是

$$-67.5806 \pm 2.306(27.91); \quad 0.9793 \pm 2.306(0.03161)$$

我们得到基于自由度为 $(10-2) = 8$ 的 χ^2 分布的 σ^2 的置信区间，相应的临界值是 2.18 和 15.5，所以置信区间是

$$(10-2)\frac{67.125}{17.54} < \sigma^2 < (10-2)\frac{67.125}{2.18} \text{ or } 30.62 < \sigma^2 < 246.33$$

这可能显得太宽了。然而，我们通常对 ε 的标准差比对其方差更感兴趣。基于同样这些结果的 σ 的 95% 置信区间是 5.89 至 15.69。

§ 4.5 预测

除了参数的估计外，回归的最常见的作用是进行预测。假定 x^0 是回归量的已知值，且我们对预测与 x^0 相应的 y 的取值 y^0 感兴趣。我们将试图对真值 y^0 进行预测：

§ 4.5.1 个体预测 (Individual Prediction)

$$y^0 = \alpha + \beta x^0 + \varepsilon^0$$

预测值将是 $\hat{y}^0 = a + bx^0$ ， $(\varepsilon^0 \sim N(0, \sigma^2) \text{ 且 } \mathbb{E}[\varepsilon^0 \varepsilon_i] = 0, \quad i = 1, \dots, n$

预测误差是：

$$\begin{aligned} e^0 &= y^0 - \hat{y}^0 \\ &= \alpha + \beta x^0 + \varepsilon^0 - a - bx^0 \\ &= (\alpha - a) + (\beta - b)x^0 + \varepsilon^0 \end{aligned}$$

在两边取期望有 $\mathbb{E}[e^0] = 0$ 。所以，在预测误差均值为 0 这个意义上最小二乘预测是无偏的。预测误差的方差是：

$$\begin{aligned} \text{Var}[e^0] &= \text{Var}[a] + (x^0)^2 \text{Var}[b] + 2x^0 \text{Cov}[a, b] + \text{Var}[\varepsilon^0] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{(x^0)^2}{S_{xx}} - \frac{2\bar{x}x^0}{S_{xx}} \right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^0 - \bar{x})^2}{S_{xx}} \right] \end{aligned}$$

所以 $\frac{y^0 - a - bX^0}{\sigma \sqrt{\left(1 + \frac{1}{n} + \frac{(X^0 - \bar{X})^2}{S_{xx}}\right)}} \sim N(0, 1)$ ，又因为 $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$ 。

所以 $\frac{y^0 - a - bX^0}{s \sqrt{\left(1 + \frac{1}{n} + \frac{(X^0 - \bar{X})^2}{S_{xx}}\right)}} \sim t(n-2)$

我们能够为 y^0 构造一个预测区间，它具有和个别参数置信区间相同的形式，特别地，我们的预测区间将是：

$$(a + bx^0) \pm t_{\lambda/2} \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x^0 - \bar{x})^2}{S_{xx}} \right]} \quad (4.5.1)$$

§ 4.5.2 均值预测 (Mean Prediction)

均值预测是预测值是 $y^0 = \alpha + \beta x^0$ 而不考虑随机干扰项

预测误差是：

$$\begin{aligned} e^0 &= y^0 - \hat{y}^0 \\ &= \alpha + \beta x^0 - a - bx^0 \\ &= (\alpha - a) + (\beta - b)x^0 \end{aligned}$$

在两边取期望有 $E[e^0] = 0$ 。所以，在预测误差均值为 0 这个意义上最小二乘预测是无偏的。预测误差的方差是：

$$\begin{aligned} \text{Var}[e^0] &= \text{Var}[a] + (x^0)^2 \text{Var}[b] + 2x^0 \text{Cov}[a, b] \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{(x^0)^2}{S_{xx}} - \frac{2\bar{x}x^0}{S_{xx}} \right] \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x^0 - \bar{x})^2}{S_{xx}} \right] \end{aligned}$$

所以 $\frac{y^0 - a - bX^0}{\sigma \sqrt{\left(\frac{1}{n} + \frac{(X^0 - \bar{X})^2}{S_{xx}} \right)}} \sim N(0, 1)$ ，又因为 $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$

所以 $\frac{y^0 - a - bX^0}{s \sqrt{\left(\frac{1}{n} + \frac{(X^0 - \bar{X})^2}{S_{xx}} \right)}} \sim t(n-2)$ 我们能够为 y^0 构造一个预测区间，它具有和个别参数置信区间相同

的形式，特别地，我们的预测区间将是

$$(a + bx^0) \pm t_{\lambda/2} \sqrt{s^2 \left[\frac{1}{n} + \frac{(x^0 - \bar{x})^2}{S_{xx}} \right]} \quad (4.5.2)$$

Example 4.3 如果 1980 年的可支配收入预测是 1030 美元（十亿），为了计算一个预测区间，我们需要 $a = -67.5806$, $b = 0.9793$, $s^2 = 67.125$, $\bar{x} = 879.24$, $S_{xx} = 67,192.44$, $n = 10$

t 分布的临界值是 2.306，将这些代入 3 得到一个预测区间是： $-67.5806 \pm 0.9793(1030) \pm 2.306(9.8256)$
i.e. 941.1 ± 22.658 。

CHAPTER 5

多元线性回归模型

在第四章中, 我们讨论只有一个解释变量影响被解释变量的情况, 但在实际生活中, 往往是多个解释变量同时影响着被解释变量。需要我们建立多元线性回归模型。

§ 5.1 多元线性模型及其假定

多元线性回归模型的一般形式是:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$$

令列向量 \mathbf{x} 是变量 \mathbf{x}_k , $k = 1, 2, \dots$ 的 n 个观测值, 并用这些数据组成一个 $n \times K$ 数据矩阵 \mathbf{X} , 在多数情况下, \mathbf{X} 的第一列假定为一列 1, 则 β_1 就是模型中的常数项。最后, 令 \mathbf{y} 是 n 个观测值 y_1, y_2, \dots, y_n 组成的列向量, 现在可将模型写为:

$$\mathbf{y} = \mathbf{x}_1 \beta_1 + \cdots \mathbf{x}_K \beta_K + \boldsymbol{\varepsilon}$$

构成多元线性回归模型的一组基本假设为:

假定 1 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 我们主要兴趣在于对参数向量 $\boldsymbol{\beta}$ 进行估计和推断。

$$\text{假定 2 } \mathbb{E}[\boldsymbol{\varepsilon}] = \begin{bmatrix} E[\varepsilon_1] \\ \mathbb{E}[\varepsilon_2] \\ \vdots \\ \mathbb{E}[\varepsilon_n] \end{bmatrix} = \mathbf{0}$$

$$\text{假定 3 } \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2 \mathbf{I}_n$$

假定 4 $\mathbb{E}[\varepsilon | \mathbf{X}] = 0$ 我们假定 \mathbf{X} 中不包含 ε 的任何信息, 由于

$$\text{Cov}[\mathbf{X}, \varepsilon] = \text{Cov}[\mathbf{X}, \mathbb{E}(\varepsilon | \mathbf{X})] \quad (5.1.1)$$

所以假定 4 暗示着 $\text{Cov}[\mathbf{X}, \varepsilon] = 0$

(5.1.1) 式成立是因为, 对于任何的双变量 \mathbf{X}, Y , 有 $\mathbb{E}(\mathbf{X}Y) = \mathbb{E}(\mathbf{X}\mathbb{E}(Y | \mathbf{X}))$, 而且

$$\text{Cov}(\mathbf{X}, Y) = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(Y - \mathbb{E}Y)'] = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbb{E}(Y | \mathbf{X}) - \mathbb{E}Y)'] = \text{Cov}(\mathbf{X}, \mathbb{E}(Y | \mathbf{X}))$$

这也暗示 $\mathbb{E}[\mathbf{y} | \mathbf{X}] = \mathbf{X}\beta$

假定 5 \mathbf{X} 是秩为 K 的 $n \times K$ 随机矩阵这意味着 \mathbf{X} 列满秩, \mathbf{X} 的各列是线性无关的。在需要作假设检验和统计推断时, 我们总是假定: $\varepsilon \sim N[0, \sigma^2 \mathbf{I}]$

§ 5.2 最小二乘回归

§ 5.2.1 最小二乘向量系数

采用最小二乘法寻找未知参数 β 的估计量 $\hat{\beta}$, 它要求 β 的估计 $\hat{\beta}$ 满足下面的条件

$$S(\hat{\beta}) \triangleq \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 = \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 \quad (5.2.1)$$

其中 $\|\mathbf{y} - \mathbf{X}\beta\|^2 \triangleq \sum_{j=1}^n \left(y_i - \sum_{j=1}^K x_{ij}\beta_j \right)^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$ 。min 是对所有的 m 维向量 β 取极小值。也即

$$\begin{aligned} S(\hat{\beta}) &= \sum_{i=1}^n \left(y_i - \sum_{j=1}^m X_{ij}\hat{\beta}_j \right)^2 \\ &= \min_{\beta_1, \dots, \beta_m} \sum_{i=1}^n \left(y_i - \sum_{j=1}^m X_{ij}\beta_j \right)^2 \end{aligned} \quad (5.2.2)$$

满足 5.2.1 式或 5.2.2 式的估计量 $\hat{\beta}_L = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{pmatrix}$ 称为 β 的最小二乘估计, 这种求估计量的方法称为最小二乘法 (OLS)。

展开上式得:

$$\begin{aligned} S(\beta) &= \mathbf{y}'\mathbf{y} - \beta'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ &= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta \end{aligned}$$

最小值的必要条件是:

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta = 0$$

设 \mathbf{b} 是解, 则 \mathbf{b} 满足正则方程组这正是我们曾分析的最小二乘正则方程组。因为 \mathbf{X} 是满秩的, 所以 $\mathbf{X}'\mathbf{X}$ 的逆存在, 从而得到解是:

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

为了证实这确实是最小值, 我们需要二阶偏分矩阵:

$$\left. \frac{\partial^2 S(\beta)}{\partial \beta \partial \beta'} \right|_{\beta=\mathbf{b}} = 2\mathbf{X}'\mathbf{X}$$

是一个正定矩阵。

我们现在来证明这个结果。对任意一非零向量 \mathbf{c} , 令 $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$, 则

$$q = \mathbf{v}'\mathbf{v} = \sum_i v_i^2; \quad \text{where } \mathbf{v} = \mathbf{X}\mathbf{c}$$

除非 \mathbf{v} 的每一元素都为 0, 否则 q 是正的。但若 \mathbf{v} 为零的话, 则 \mathbf{X} 的各列的一个线性组合等于 0, 这与 \mathbf{X} 满秩的假定相矛盾。

§ 5.3 最小二乘估计量的统计特性

在本节中, 我们对回归量的两种情况, 即非随机回归量和随机回归量下分别作讨论。

§ 5.3.1 \mathbf{X} 非随机回归量

若回归量当作非随机来进行处理时, 则将 \mathbf{X} 当作常数矩阵处理就可导出最小二乘估计量的各种特性。可得

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\beta + \epsilon) = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\epsilon \quad (5.3.1)$$

若 \mathbf{X} 是非随机的, 或 $\mathbb{E}(\mathbf{X}'\epsilon) = 0$, 则 5.3.1 中第二项的期望值是 0。所以, 最小二乘估计量是无偏的, 它的协方差矩阵是:

$$\begin{aligned} \text{Var}[\mathbf{b}] &= \mathbb{E}[(\mathbf{b} - \beta)(\mathbf{b} - \beta)'] \\ &= \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\epsilon\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbb{E}[\epsilon\epsilon'] \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\sigma^2 \mathbf{I}) \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

在前面的内容中, 对 $K = 2$ 的特殊 \mathbf{b} 是 β 的最小方差的线性无偏估计量。现在我们给出这个基本结果的一个更一般的证明, 令 $\tilde{\mathbf{b}} = \mathbf{C}\mathbf{y}$ 是 β 的另一个不同于 \mathbf{b} 的线性无偏估计量, 其中 \mathbf{C} 是一个 $K \times n$ 矩阵。若 $\tilde{\mathbf{b}}$ 是无偏的,

$$\mathbb{E}[\mathbf{C}\mathbf{y}] = \mathbb{E}[\mathbf{C}\mathbf{X}\beta + \mathbf{C}\epsilon] = \beta$$

这暗示着 $\mathbf{C}\mathbf{X} = \mathbf{I}$, 并且 $\tilde{\mathbf{b}} = \beta + \mathbf{C}\epsilon$ 。所以可以得到 $\tilde{\mathbf{b}}$ 的协方差矩阵是:

$$\text{Var}[\tilde{\mathbf{b}}] = \sigma^2 \mathbf{C}\mathbf{C}'$$

现在令 $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, 由假设知 $\mathbf{D} \neq 0$ 。那么, $\mathbf{b}^* = \tilde{\mathbf{b}} - \mathbf{b} = \mathbf{D}\mathbf{y}$, $\text{Var}(\mathbf{b}^*) = \mathbf{D}\Sigma_Y\mathbf{D}' = \sigma^2\mathbf{D}\mathbf{D}'$ 于是 $\mathbf{D}\mathbf{D}'$ 是非负定矩阵。

$$\begin{aligned} \Rightarrow \text{Var}[\tilde{\mathbf{b}}] &= \sigma^2 \left[\left(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \right) \left(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \right)' \right] \\ &= \sigma^2 \left[\left(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \right) \left(\mathbf{D}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \right) \right] \\ &= \sigma^2 \left(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1} \right) \end{aligned}$$

在展开这个四项和式之前, 我们注意到:

$$\mathbf{I} = \mathbf{C}\mathbf{X} = \mathbf{D}\mathbf{X} + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})$$

由于上面最后一项是 \mathbf{I} , 有 $\mathbf{D}\mathbf{X} = 0$, 所以

$$\begin{aligned} \text{Var}[\tilde{\mathbf{b}}] &= \sigma^2\mathbf{D}\mathbf{D}' + \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \text{Var}[\mathbf{b}] + \sigma^2\mathbf{D}\mathbf{D}' \end{aligned}$$

$\tilde{\mathbf{b}}$ 的方差矩阵等于 \mathbf{b} 的方差矩阵加上一个非负定矩阵。所以, $\text{Var}[\tilde{\mathbf{b}}]$ 的每个二次型都大于 $\text{Var}[\mathbf{b}]$ 的相应二次型。利用这个结果可以证明高斯-马尔科夫定理。

Theorem 5.3.1 (高斯—马尔科夫定理:) 对任意常向量 \mathbf{w} , 古典线性模型中 $\mathbf{w}'\beta$ 的最小方差线性无偏估计量是 $\mathbf{w}'\mathbf{b}$, 其中 \mathbf{b} 是最小二乘估计量。

§ 5.3.2 X 随机回归量

在这样的情况下, 为了得到最小二乘估计量特性更多的一般性, 有必要将上面的结果推广解释变量 \mathbf{X} 是来自某种概率分布的情况中去。获得 \mathbf{b} 的统计特性的一个方便的方法是, 首先, 第一步求得对 \mathbf{X} 的条件期望结果, 这等同于非随机回归量的情况, 第二步, 通过假定 5 分布得到无条件结果。此论点的关键是, 如果我们对任意 \mathbf{X} 都可能得到条件无偏性, 我们就可以得到一个无条件结果。

$$\therefore \mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$$

所以，以观测到的 \mathbf{X} 为条件我们得到

$$\mathbb{E}[\mathbf{b} \mid \mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{0} = \boldsymbol{\beta}$$

一个有用的方法是利用重期望定律:

$$\begin{aligned} \mathbb{E}[\mathbf{b}] &= \mathbb{E}_x[\mathbb{E}[\mathbf{b} \mid \mathbf{X}]] \\ &= \boldsymbol{\beta} + \mathbb{E}_x[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}]] \end{aligned}$$

因为由假定 4 有 $\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$ ，所以， \mathbf{b} 也是无条件无偏的，这样，

$$\mathbb{E}[\mathbf{b}] = \mathbb{E}_x[\mathbb{E}[\mathbf{b} \mid \mathbf{X}]] = \mathbb{E}_x[\boldsymbol{\beta}] = \boldsymbol{\beta}$$

同样，以 \mathbf{X} 为条件的 \mathbf{b} 的方差是:

$$\text{Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

为了求得确切的方差，我们使用方差分解公式:

$$\text{Var}[\mathbf{b}] = \mathbb{E}_x[\text{Var}[\mathbf{b} \mid \mathbf{X}]] + \text{Var}_x[\mathbb{E}[\mathbf{b} \mid \mathbf{X}]]$$

由于对所有 \mathbf{X} ， $\mathbb{E}[\mathbf{b} \mid \mathbf{X}] = \boldsymbol{\beta}$ ，所以第二项为零，因此:

$$\text{Var}[\mathbf{b}] = \mathbb{E}[\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2 \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}]$$

我们原来的结论要稍作改变，我们必须用其期望值 $\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}]$ 来代替原来 $(\mathbf{X}'\mathbf{X})^{-1}$ 以得到适当的协方差矩阵。

从上一段的结果可以合乎逻辑地建立高斯—马尔科夫定理，即对任何 $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$ ，在 \mathbf{X} 给定的条件下有

$$\text{Var}[\mathbf{b} \mid \mathbf{X}] \leq \text{Var}[\tilde{\boldsymbol{\beta}} \mid \mathbf{X}]$$

但若这一不等式对一特定 \mathbf{X} 成立，则必须成立:

$$\text{Var}[\mathbf{b}] = \mathbb{E}_x[\text{Var}[\mathbf{b} \mid \mathbf{X}]]$$

即，若它对每一特定 \mathbf{X} 成立，则它一定对 \mathbf{X} 的平均值也成立。这暗示， $\text{Var}(\mathbf{b}) \leq \text{Var}(\tilde{\boldsymbol{\beta}})$ 所以，不论我们是否将 \mathbf{X} 看作是随机的，即无偏性和高斯—马尔科夫定理都成立。

§ 5.4 最小二乘估计量的统计推断

迄今为止，在我们任一结果还未用到 $\boldsymbol{\varepsilon}$ 的正态性的假定 6，但这一假定对构造假设检验的统计量是有用的和必须的。

§ 5.4.1 回归系数的假设检验

我们先讨论 \mathbf{X} 非随机变量时的情况。

在5.3.1中, \mathbf{b} 是干扰向量 ϵ 的一个线性函数, 如果我们假定 ϵ 服从多重正态分布。利用前面结果及前边推导的均值向量和协方差矩阵来表示即

$$\mathbf{b} \sim N \left[\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \right]$$

这是一个多重正态分布, 所以 \mathbf{b} 的每一元素的边际分布都是正态分布的:

$$b_k \sim N \left[\beta_k, \sigma^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1} \right]$$

令 S_{kk} 是 $(\mathbf{X}'\mathbf{X})^{-1}$ 的第 k 个对角元素, 则

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S_{kk}}} \quad (5.4.1)$$

服从标准正态分布。若 σ^2 已知, 关于 β_k 的统计推断可以基于 z_k 。然而 σ^2 仍要估计, 所以5.4.1式中 Z_k 不是统计量。我们要得到 σ^2 的无偏估计量, 才能作进一步的推断。

按定义最小二乘残差向量是:

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \right) \mathbf{y} \\ &= \mathbf{M}\mathbf{y} \end{aligned}$$

\mathbf{M} 是回归分析中一个基本的 $n \times n$ 矩阵, 你可以容易地验证 \mathbf{M} 既是对称的 ($\mathbf{M} = \mathbf{M}'$) 又是幂等的 ($\mathbf{M} = \mathbf{M}^2$)。

Property 5.1 $\mathbf{X}'\mathbf{e} = 0$ 和 $\mathbf{i}'\mathbf{e} = 0$

Proof 5.1 由正则方程组, 我们得到:

$$\begin{aligned} \mathbf{X}'\mathbf{e} &= \mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{X}' \left(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \right) = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{Y} = 0 \end{aligned}$$

所以, $\mathbf{i}'\mathbf{e} = 0$

Corollary 5.1 $\mathbf{X}'\mathbf{M} = 0$ 和 $\mathbf{M}\mathbf{X} = 0$

Corollary 5.2 $\mathbf{i}'\mathbf{M} = 0$ 和 $\mathbf{M}\mathbf{i} = 0$ 。Corollary 5.2 成立是因为 \mathbf{X}' 的第一行是 $(1, 1, \dots, 1)$ 。

Property 5.2 e 和 b 互不相关。

$$\begin{aligned}\text{cov}(e, b) &= [I_n - X(X'X)^{-1}X'] \text{cov}(Y, Y) [(X'X)^{-1}X']' \\ &= \sigma^2 [I_n - X(X'X)^{-1}X'] [X(X'X)^{-1}] = 0\end{aligned}$$

从几何解释来看这一性质是显然的, e 表示 Y 到子样空间的垂线估计量, \hat{Y} 和 e 互相垂直。

Property 5.3 残差 e 的均值向量和协方差阵分别是 $\mathbb{E}(e) = 0$ 和 $\text{Var}(e) = \sigma^2 M$

Proof 5.2

$$\begin{aligned}\mathbb{E}(e) &= \mathbb{E}(Y - Xb) = \mathbb{E}Y - \mathbb{E}(Xb) = X\beta - \mathbb{E}\left(X[X'X]^{-1}X'\varepsilon + X\beta\right) = 0 \\ \text{Var}(e) &= \mathbb{E}(My(My)') = M\mathbb{E}(yy')M = \sigma^2 M^2 = \sigma^2 M \\ &= \sigma^2 \left\{ I_n - X(X'X)^{-1}X' \right\}\end{aligned}$$

$\mathbb{E}(e) = 0$, 暗示 $\hat{y} = Xb$ 是 y 的无偏估计量。

Property 5.4 $E[e'e] = (n - K)\sigma^2$

Proof 5.3 最小二乘残差是:

$$e = My = M[X\beta + \varepsilon] = M\varepsilon$$

这是由于 $MX = 0$, σ^2 的一个估计量将基于残差平方和:

$$e'e = \varepsilon'M'M\varepsilon = \varepsilon'M^2\varepsilon = \varepsilon'M\varepsilon$$

这个二次型的期望值是: $\mathbb{E}[e'e] = \mathbb{E}[\varepsilon'M\varepsilon]$

我们有 $\mathbb{E}(e'e) = \mathbb{E}(\text{tr}(e'e)) = \mathbb{E}[\text{tr}(\varepsilon'M\varepsilon)] = \mathbb{E}[\text{tr}(M\varepsilon\varepsilon')]$

由于 M 是固定的, 这就是

$$\text{tr}(M\mathbb{E}[\varepsilon\varepsilon']) = \text{tr}(M\sigma^2 I) = \sigma^2 \text{tr}(M)$$

M 的迹是:

$$\begin{aligned}\text{tr}[I_n - X(X'X)^{-1}X'] &= \text{tr}(I_n) - \text{tr}((X'X)^{-1}X'X) \\ &= \text{tr}(I_n) - \text{tr}(I_K) = n - K\end{aligned}$$

所以,

$$\mathbb{E}[e'e] = (n - K)\sigma^2$$

σ^2 的一个无偏估计量是:

$$s^2 = \frac{e'e}{n - K} \quad (5.4.2)$$

回归的标准误差是 s^2 ，其平方根为 s 。利用 s^2 ，我们可以计算估计量 b 的估计协方差矩阵：

$$\text{Est. Var } [b] = s^2 (\mathbf{X}'\mathbf{X})^{-1}$$

通过利用 s^2 替代 σ^2 ，我们导出替代 5.4.1 中 z_k 的一个统计量。此量

$$\frac{(n-K)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$$

是一个标准正态向量 $(\boldsymbol{\varepsilon}/\sigma)$ 的幂等二次型，所以，它服从自由度为 $\text{Rank}(\mathbf{M}) = \text{tr}(\mathbf{M}) = n - K$ 的 χ^2 分布。

5.4.2 中的 χ^2 分布变量独立于 5.3.1 中的标准正态变量，为了证明这一点，只要证明

$$\frac{\mathbf{b} - \boldsymbol{\beta}}{\sigma} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \quad (5.4.3)$$

独立于 $(n-K)s^2/\sigma$ 就足够了。我们知道标准正态向量 x 的一个线性式 L_x 和一个幂等二次型 $\mathbf{x}'\mathbf{A}\mathbf{x}$ 独立的充分条件是 $L_{\mathbf{A}} = 0$ ， $\boldsymbol{\varepsilon}/\sigma$ 等 x ，我们发现这里所需求的是 $(\mathbf{X}')\mathbf{X}^{-1}\mathbf{X}'\mathbf{M} = \mathbf{0}$ 。这确实成立，因为 $\mathbf{X}'\mathbf{M} = \mathbf{0}$ 。

在推导回归分析中许多检验统计量中起中心作用的一般性结果是：

若 $\boldsymbol{\varepsilon}$ 服从正态分布，最小二乘系数估计量 b 统计独立于残差向量 e 及包括 s^2 在内的 e 的所有函数。

所以，比率：

$$\begin{aligned} t_k &= \frac{(b_k - \beta_k) / \sqrt{\sigma^2 S_{kk}}}{\{[(n-K)s^2/\sigma^2] / (n-K)\}^{1/2}} \\ &= \frac{b_k - \beta_k}{\sqrt{s^2 S_{kk}}} \end{aligned} \quad (5.4.4)$$

服从自由度为 $(n-K)$ 的 t 分布。这是我们作统计推断的基础。

§ 5.4.2 线性约束检验

我们通常对含有不只一个系数的假设检验感兴趣，我们可以利用一个类似于 5.4.4 中的检验统计量。假定我们的假设是：

$$H_0 : r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = \mathbf{r}'\boldsymbol{\beta} = q$$

(通常某些 r 将为零) 左边的样本估计是：

$$r_1b_1 + r_2b_2 + \cdots + r_Kb_K = \mathbf{r}'\mathbf{b} = \hat{q}$$

若 \hat{q} 显著异于 q ，则我们推断样本数据与假设不一致。与 5.4.4 一样，将假设基于下式是很自然的。

$$t = \frac{\hat{q} - q}{se(\hat{q})} \quad (5.4.5)$$

我们需要 \hat{q} 的标准误差的一个估计。由于 \hat{q} 是 b 的一个线性函数，且我们已估计出了 b 的方差矩阵 $s^2(\mathbf{X}'\mathbf{X})^{-1}$ ，我们可用下式估计 \hat{q} 的方差。

$$\text{Est.Var} [\hat{q}] = \mathbf{r}' \left[s^2 (\mathbf{X}'\mathbf{X})^{-1} \right] \mathbf{r}$$

5.4.4 中的分母是这个量的平方根。若假设是正确的，我们的估计应该反映这一事实，至少在抽样变化性的范围内如此。这样，若前边的 t 比率的绝对值大于适当的监界值，则应对假设产生怀疑。

§ 5.4.3 随机 \mathbf{X} 及正态 ε 下的检验统计量

现在，我们考虑当 \mathbf{X} 是随机的，样本检验统计量和推断方法考虑 5.4.5 中检验 $H_0: \beta_k = \beta_k^0$ 的 t 统计量：

$$t | \mathbf{X} = \frac{(b_k - \beta_k^0)}{\left[s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1} \right]^{1/2}} \quad (5.4.6)$$

以 \mathbf{X} 为条件， $t | \mathbf{X}$ 服从自由度为 $(n-K)$ 的 t 分布。然而，我们感兴趣的是 t 的边际（即无条件）分布。正如我们所见，5.4.5 仅仅在以 \mathbf{X} 为条件时 b 才是正态分布的，我们还没有证明它的边际分布是正态分布的。类似地，当 \mathbf{X} 是随机的情况下，在给定 \mathbf{X} 的条件下，我们得到了 5.4.6 式的 t 统计量，我们还没有证明 t 边际分布也是以 $(n-K)$ 为自由度的 t 分布。事实上， t 的边际分布仍是以 $(n-K)$ 为自由度的 t 分布，不论 \mathbf{X} 的分布是什么，甚至不论 \mathbf{X} 是随机的还是非随机的或者是混合的。

这个令人迷惑的结果来自 $f(t | \mathbf{X})$ 不是 \mathbf{X} 的函数这一事实，同样的原因可以用来推演不论 \mathbf{X} 是不是随机的，通常用以检验线性约束的 F 比率都是有效的。

结论：若干扰项是正态分布的，我们可以在我们的过程中不加变化地进行检验和构造参数的置信区间，而不去考虑回归量是随机的、非随机的，还是它们的混合。

§ 5.4.4 拟合优度和方差分析

由方差分解公式，我们有： $\text{Var}(\mathbf{Y}) = \text{Var}(\mathbb{E}(\mathbf{Y} | \mathbf{X})) + \mathbb{E}_x(\text{Var}(\mathbf{Y} | \mathbf{X}))$ 。我们用幂等矩阵 \mathbf{M}^0 来表示：

$$\mathbf{Y}'\mathbf{M}^0\mathbf{Y} = \mathbf{E}'(\mathbf{Y} | \mathbf{X})\mathbf{M}_0\mathbf{E}(\mathbf{Y} | \mathbf{X}) + \mathbf{e}'\mathbf{e}$$

$$\mathbf{Y}'\mathbf{M}^0\mathbf{Y} = \mathbf{b}'\mathbf{X}'\mathbf{M}_0\mathbf{X}\mathbf{b} + \mathbf{Y}'\mathbf{M}\mathbf{Y}$$

$$SST = SSR + SSE$$

所以， $SSE = \mathbf{Y}'\mathbf{M}\mathbf{Y}$ 和 $SSR = SST - SSE = \mathbf{Y}'(\mathbf{M}^0 - \mathbf{M})\mathbf{Y}$

进一步研究回归平方和 SSR 与残差平方和 SSE，我们可以得到下面三个结论：

a) 在 $\beta = 0$ 的假设条件下，回归平方和 $\frac{SSR}{\sigma^2}$ 服从自由度为 $(K-1)$ 的卡方分布 $\chi^2(K-1)$ ；

b) 残差平方和 $\frac{SSR}{\sigma^2}$ 服从自由度为 $n - K$ 的卡方分布 $\chi^2(n - k)$;

c) 在 $\beta = 0$ 的假设条件下, $\frac{SSR/(K-1)}{SSE/(n-K)}$ 服从 $F(k - 1, n - k)$ 分布

Proof 5.4 (a) $M^0 - M$ 是幂等矩阵。先证明 $M^0 M + M M^0 = 2M$ 。

$$\begin{aligned} M^0 M + M M^0 &= \left(I - \frac{1}{n} \mathbf{i} \mathbf{i}' \right) M + M \left(I - \frac{1}{n} \mathbf{i} \mathbf{i}' \right) \\ &= M - \frac{1}{n} \mathbf{i} \mathbf{i}' M + M - \frac{1}{n} M \mathbf{i} \mathbf{i}' \\ &= 2M \end{aligned}$$

$$\begin{aligned} \Rightarrow (M^0 - M)(M^0 - M) &= M^0 - (M^0 M + M M^0) + M \\ &= M^0 - 2M + M = M^0 - M \Rightarrow \\ \therefore r(M^0 - M) &= \text{tr}(M^0 - M) = \text{tr}(M^0) - \text{tr}(M) = n - 1 - (n - K) = K - 1 \end{aligned}$$

在 $\beta = 0$ 的假设条件下, 回归平方和 $\frac{SSR}{\sigma^2}$ 服从自由度为 $(K - 1)$ 的卡方分布 $\chi^2(K - 1)$; (WHY?)

Proof 5.5 (b) 因为 M 是幂等矩阵而且 $r(M) = \text{tr}(M) = n - K$

Proof 5.6 (c) 只要验证 $M(M^0 - M) = 0$ 即可。

$$\begin{aligned} M(M^0 - M) &= M \left(I - \frac{1}{n} \mathbf{i} \mathbf{i}' - M \right) \\ &= M - \frac{1}{n} M \mathbf{i} \mathbf{i}' - M^2 = -\frac{1}{n} M \mathbf{i} \mathbf{i}' = 0 \end{aligned}$$

和前一章的情况一样, 我们要对回归模型的好坏, 作出评价, 决定系数 $R^2 = \frac{SSR}{SST}$ 就是对模型拟合的一个度量, 计算 R^2 有两个等价的方法。

$$R^2 = \frac{SSR}{SST} = \frac{\mathbf{b}' \mathbf{X}' M^0 \mathbf{X} \mathbf{b}}{\mathbf{Y}' M^0 \mathbf{Y}} = 1 - \frac{\mathbf{e}' \mathbf{e}}{\mathbf{Y}' M^0 \mathbf{Y}}$$

进一步推导和化解, 我们可以得到 R^2 另一个公式。 $\mathbf{b}' \mathbf{X}' M^0 \mathbf{X} \mathbf{b} = \hat{\mathbf{Y}}' M^0 \hat{\mathbf{Y}}$, $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}$ 以及 $M^0 \mathbf{e} = \mathbf{e}$ (表示残差已经具有零均值) 和 $\mathbf{X}' \mathbf{e} = 0$ 。

$$\hat{\mathbf{Y}}' M^0 \hat{\mathbf{Y}} = \hat{\mathbf{Y}}' M^0 \mathbf{Y} - \hat{\mathbf{Y}}' M^0 \mathbf{e} = \hat{\mathbf{Y}}' M^0 \mathbf{Y} - \mathbf{b}' \mathbf{X}' \mathbf{e} = \hat{\mathbf{Y}}' M^0 \mathbf{Y}$$

$$\begin{aligned}
R^2 &= \frac{\hat{\mathbf{Y}}' \mathbf{M}^0 \hat{\mathbf{Y}}}{\mathbf{Y}' \mathbf{M}^0 \mathbf{Y}} = \frac{(\hat{\mathbf{Y}}' \mathbf{M}_0 \hat{\mathbf{Y}})^2}{\mathbf{Y}' \mathbf{M}^0 \mathbf{Y} \cdot \hat{\mathbf{Y}}' \mathbf{M}_0 \hat{\mathbf{Y}}} \\
&= \frac{(\hat{\mathbf{Y}}' \mathbf{M}_0 \mathbf{M}_0 \mathbf{Y})^2}{\mathbf{Y}' \mathbf{M}^0 \mathbf{Y} \cdot \hat{\mathbf{Y}}' \mathbf{M}^0 \hat{\mathbf{Y}}} \\
&= \frac{[\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y})]^2}{[\sum_i (y_i - \bar{y})^2][\sum_i (\hat{y}_i - \bar{y})^2]} = r_{Y\hat{Y}}^2
\end{aligned}$$

第一个方法度量了 \mathbf{y} 的总变差中由回归变差所解释的部分，第二个是 \mathbf{y} 的观测值和由估计的回归方程所产生的预测值间的相关系数的平方。

当利用 R^2 来比较不同的线性统计模型的拟合度时，存在一个严重的缺点，就是它的值随着解释变量的增多而增大。为了克服这个缺点，我们可以用调整的 R^2 来测度一个模型的解释能力，这个调整的 R^2 被记 \bar{R}^2 ，它的表达式为：

$$\begin{aligned}
\bar{R}^2 &= 1 - \frac{\text{SSE}/(n-K)}{\text{SST}/(n-1)} = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-K)}{\text{SST}/(n-1)} \\
&= 1 - \left(\frac{n-1}{n-K} \right) (1 - R^2)
\end{aligned}$$

这里 $\frac{\mathbf{e}'\mathbf{e}}{n-K}$ 是 σ 的无偏估计量，(思考：当 \mathbf{y} 服从正态分布时， $\frac{\mathbf{y}'\mathbf{y} - n\bar{y}^2}{n-1}$ 是 σ 的一个无偏估计量)。

R^2 与 \bar{R}^2 不同的是，随着解释变量的增多，它的值可能变小，甚至要能取负值

因为 $\mathbf{M}^0 \mathbf{Y} = \mathbf{Y} - \bar{\mathbf{Y}}$

$$\begin{aligned}
\therefore \text{SSR} &= \mathbf{b}' \mathbf{X} \mathbf{M}^0 \mathbf{X} \mathbf{b} = \mathbf{b}' \mathbf{X}' \mathbf{M}^0 \mathbf{Y} \\
&= \mathbf{b}' \mathbf{X}' \mathbf{Y} - \mathbf{b}' \mathbf{X}' \bar{\mathbf{Y}} \\
&= \mathbf{b}' \mathbf{X}' \mathbf{Y} - \hat{\mathbf{Y}}' \bar{\mathbf{Y}} = \mathbf{b}' \mathbf{X}' \mathbf{Y} - n\bar{\mathbf{Y}}^2
\end{aligned}$$

我们得到了回归方差的另一个表达式，请见多元线性回归模型方差分析表。

表 5.1 多元线性回归模型方差分析

	来源	自由度	均方
回归	$\mathbf{b}' \mathbf{X}' \mathbf{y} - n\bar{y}^2$	$K - 1$	
残差	$\mathbf{e}' \mathbf{e}$	$n - K$	s^2
总	$\mathbf{y}' \mathbf{y} - n\bar{y}^2$	$n - 1$	$S_{yy(n-1)}$
$F[K-1, n-K] = \frac{\text{SSR}/(K-1)}{\text{SSE}/(n-K)} = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$			

§ 5.4.5 回归的显著性检验

一个通常要检验的假定是回归方程作为整体的显著性, 这是对除了常数项外所有常数都为 0 的假设的联合检验。若所有系数为 0, 则多重相关系数为 0, 所以我们可以将这一假定的一个检验基于 R^2 值上。统计量:

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

服从自由度为 $K-1$ 和 $n-K$ 的 F 分布, 检验的逻辑是, F 统计量是对我们强加所有斜率都是 0 的这一约束时的拟合损失的一个度量 (R^2 的全部), 若 F 大, 假设被拒绝。

§ 5.5 预测

多元回归环境下的预测结果与前一章中讨论的那些本质是一样的。假定我们希望预测与回归向量 x^0 相应的 y^0 值。它将是:

$$y^0 = \beta' x^0 + \varepsilon^0 \quad \varepsilon^0 \sim N(0, \sigma^2) \quad \& \quad \mathbb{E}[\varepsilon^0 \varepsilon_i] = 0, \quad i = 1, \dots, n$$

由高斯—马尔科夫定理知:

$$\hat{y}^0 = b' x^0$$

是 y^0 的最小方差线性无偏估计量。

个体预测 (Individual Prediction) 误差是:

$$e^0 = y^0 - \hat{y}^0 = (\beta - b)' x^0 + \varepsilon^0 \quad \varepsilon^0 \sim N(0, \sigma^2) \quad \& \quad \mathbb{E}[\varepsilon^0 \varepsilon_i] = 0, \quad i = 1, \dots, n$$

这个估计的预测方差是:

$$\begin{aligned} \text{Var}[e^0] &= \sigma^2 + \text{Var}[(\beta - b)' x^0] \\ &= \sigma^2 + x^{0'} [\sigma^2 (X'X)^{-1}] x^0 \end{aligned}$$

若回归含有一个常数项, 一个等价的表达式是:

$$\text{Var}[e^0] = \sigma^2 + \frac{\sigma^2}{n} + \sigma^2 \left\{ \sum_{j=2}^K \sum_{k=2}^K (x_j^0 - \bar{x}_j) (x_k^0 - \bar{x}_k) (\underline{X}' M^0 \underline{X})_{jk} \right\}$$

其中 \underline{X} 是 X 的不包含全为 1 的列的最后 $K-1$ 列。这表明, 和以前一样, 区间的宽度依赖于 x^0 的元素与数据中心的距离。

$$\text{因此 } \frac{y_0 - \hat{y}_0}{\sqrt{\sigma^2 (1 + x^{0'} (X'X)^{-1} x^0)}} \sim N(0, 1), \quad \text{又因为 } \frac{(n-K)S^2}{\sigma^2} \sim \chi^2(n-K)$$

由此得到 $\frac{y_0 - \hat{y}_0}{\sqrt{s^2 (1 + x'^0 (X'X)^{-1} x^0)}} \sim t(n-K)$

即 y^0 的一个置信区间将用下式形成:

$$\text{预测区间} = \hat{y}^0 \pm t_{\frac{\lambda}{2}} \text{se}(\hat{y}^0)$$

均值预测 (Mean Prediction) 均值预测是预测值是 $y^0 = \beta'x^0$ 而不考虑随机干扰项 ε^0 。

误差是:

$$e^0 = y^0 - \hat{y}^0 = (\beta - b)'x^0$$

这个估计的预测方差是:

$$\begin{aligned} \text{Var}[e^0] &= \text{Var}[(\beta - b)'x^0] \\ &= x^{0'} [\sigma^2 (X'X)^{-1}] x^0 \end{aligned}$$

因此 $\frac{y_0 - \hat{y}_0}{\sqrt{\sigma^2 (x^{0'} (X'X)^{-1} x^0)}} \sim N(0, 1)$, 又因为 $\frac{(n-K)S^2}{\sigma^2} \sim \chi^2(n-K)$

由此得到 $\frac{y_0 - \hat{y}_0}{\sqrt{s^2 (x^{0'} (X'X)^{-1} x^0)}} \sim t(n-K)$

即 y^0 的一个置信区间将用下式形成:

$$\text{预测区间} = \hat{y}^0 \pm t_{\frac{\lambda}{2}} \text{se}(\hat{y}^0)$$

§ 5.6 分块回归和偏回归

当兴趣实际上只集中于一个变量或变量全集的一个子集时, 设定一个多元回归模型是很普遍的, 但往往这个变量或变量全集的子集并不能很好地解释被解释变量, 需要我们在原有的模型中添加新的解释变量, 才能进一步完善模型例如考虑收入方程, 虽然我们的主要兴趣在于收入和教育的联系上, 将年龄包括进模型是必要的。我们已经证实从方程忽略年龄将是错误的, 这里我们考虑的问题是, 从一个多元回归模型中单独地获取一个子集变量的系数涉及什么样的计算, 例如获取前边及回归中教育的系数。

以一般术语, 假定原有回归模型是 $y = \beta_2 X_2 + \varepsilon$, 现在在原有的模型中添加新的解释变量集 X_1 , 那么现在的回归方程包括两组变量 X_1 和 X_2 , 转换为:

$$y = \beta'X + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

的代数解 b_2^* 是什么? 与原有的估计量 b_2 有何关系?

新的模型的正则方程组是:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} \quad (5.6.1)$$

利用分块逆矩阵可以得到:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} = \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix}$$

另外一个方法是可以直接处理 (1a) 和 5.6.1 以求解 b_2^* 。我们首先从 (1a) 求得解 b_1

$$\begin{aligned} b_1^* &= (X_1'X_1)^{-1} X_1'y - (X_1'X_1)^{-1} X_1'X_2 b_2^* \\ &= (X_1'X_1)^{-1} X_1'(y - X_2 b_2^*) \end{aligned} \quad (5.6.2)$$

注意此解表明 b_1^* 是 y 对 X_1 回归的系数减去一个修正向量。然后, 将其代入 5.6.1 得到:

$$X_2'X_1 (X_1'X_1)^{-1} X_1'y - X_2'X_1 (X_1'X_1)^{-1} X_1'X_2 b_2^* + X_2'X_2 b_2^* = X_2'y$$

整理各项后:

$$X_2' \left(I - X_1 (X_1'X_1)^{-1} X_1' \right) X_2 b_2^* = X_2' \left(I - X_1 (X_1'X_1)^{-1} X_1' \right) y$$

解是:

$$\begin{aligned} b_2^* &= \left[X_2' \left(I - X_1 (X_1'X_1)^{-1} X_1' \right) X_2 \right]^{-1} \left[X_2' \left(I - X_1 (X_1'X_1)^{-1} X_1' \right) y \right] \\ &= (X_2' M_1 X_2)^{-1} (X_2' M_1 y) \end{aligned} \quad (5.6.3)$$

注意出现在每个中括号中的小括号里的矩阵都是讨论过的“残差制造者”, 这里是相应于对 X_1 各列回归的。这样, $M_1 X_2$ 是一个残差矩阵, 其中每一列都是 X_2 中相应列对 X_1 中各变量回归的残差向量。利用 M_1 和 M 一样是幂等的这一事实, 我们可将 (5.6.3) 重写为:

$$b_2^* = (X_2^* X_2^*)^{-1} X_2^* y^* \quad \text{where } X_2^* = M_1 X_2; y^* = M_1 y \quad (5.6.4)$$

所以, b_2^* 是为来自一个回归的系数集合, 这个回归的被解释变量是 y 单独对 X_1 回归的残差, 解释变量是 X_2 的每一列分别对 X_1 回归所得残差的集合。这个过程通常被称作排除或筛掉 X_1 的影响。正是部分地由于这个原因, 一个多元回归中的系数通常被称作偏回归系数。

我们可以用一个例子来说, 通过首先用收入和教育对年龄 (或年龄及年龄中平方) 回归, 然后在一个简单回归中使用这两个残差, 我们能够得到教育在最小二乘回归中的系数。这一方法的一个经典的应用中, 费雪和沃 (1933) 注意到, 在时间序列环境下, 像刚才提到的那样首先通过筛掉时间的影响而消除数据趋势, 然后用消除趋势的数据简单回归和直接带有一个时间趋势变量似合所得结果是一样的。

§ 5.6.1 偏回归和偏相关系数

使用多元回归包含一个在实际中可能不能实施的概念性试验，即类似于经济学中的“假设其余情况均同”。继续考虑简介中的例子，将收入和年龄及教育相联系的回归方程使我们能够对两个同龄但教育程度不同的人的收入进行比较，即使样本中没有这样一对个人。术语偏回归系数所暗示的正是回归的这一特性。我们已经看到，获取这个结果的方法是首先用收入和教育对年龄进行回归，然后从回归方程中计算出残差，按其构造，年龄对解释这些残差没有任何能力。所以，在这种“净化”（或筛掉年龄的影响后）后的收入和教育间的任何相关都与年龄无关。

同一原理可应用于两个变量间的相关系数上。继续我们的例子，当我们在样本中得到收入和教育间的相关数为 0.7 时，那么，在何种程度上我们可以假定这一相关是由于某种直接关系，而非由于当人们变老时，收入和教育平均来说都趋于增长这一事实？为了找出答案，我们将使用偏相关系数，这与偏回归系数的计算方式一样，在我们的例子中，抑制年龄的影响，收入和教育间的偏相关系数可如下获取：

1. Y^* = 收入对年龄的回归中的残差
2. E^* = 教育对年龄的回归中的残差
3. 偏相关系数 r_{YE}^* 就是 Y^* 和 E^* 间的简单相关系数。

这似乎是一个可怕的计算量，然而存在一个方便的简捷算法，一旦计算了一个多元回归，5.4.4 中用于检验系数等于 0 的 t 比率，可用于计算

$$r_{y^k}^{*2} = \frac{t_k^2}{t_k^2 + \text{自由度}} \quad (5.6.5)$$

§ 5.6.2 对均值的离差——对常数回归

作为上一节结果的一个应用，考虑 X_1 仅为 X 中由 1 组成的第一列的这种情况，此时 b_2 的解将是带有常数项的回归中斜率。令 i 为由 1 构成的列，任何变量 z 对 i 的回归的系数是： $[i'i]^{-1} i'z = \bar{z}$ 拟合值是 $i\bar{z}$ ，残差是 $z_i - \bar{z}$ 。所以，当我们将其应用于先前结果时，会发现：将数据转换成对其均值的离差，然后用离差形式的变量 Y 对同样的离差形式的解释变量回归，可以得到含有常数项的多元回归中的斜率。

Example 5.1 (练习) 若在计算斜率前忽略了将 y 转换为对 \bar{y} 的离差，在前边的回归中将会发生什么情况？

得到了 X_2 的系数后，怎么才能取得 X_1 的系数？当然，一个方法是转换 X_1 和 X_2 的角色重复上一节中的练习，但有一个更容易的方法，对一般情形，两个正则方程组中的第一个是：

$$X_1' X_1 b_1 + X' X_2 b_2 = X' y$$

我们已经解出了 b_2 ，所以，在求解 b_1 时可以使用它：

$$b_1 = (X_1' X_1)^{-1} X_1' y - X_1' X_2 b_2 = (X_1' X_1)^{-1} X_1' (y - X_2 b_2) \quad (5.6.6)$$

若 X_1 仅为一列，(5.6.6) 中第一个将产生如下结果：

$$b_1 = y - x_2 b_2 - \cdots - x_k b_k \quad (5.6.7)$$

这我们以前已经见到过。

§ 5.7 偏离正态性的检测（正态性的哈尔克-贝拉（Jarque-Bera）BJ 检验）

本节考察的是利用最小二乘残差的矩来推断真正扰动项的分布的一般问题。

$$\mu_r = \mathbb{E}[\varepsilon^r]$$

的直观估计量是：

$$m_r = \left(\frac{1}{n}\right) \sum_i e_i^r$$

然而，最小二乘残差只是真实扰动项的不完全估计：

$$e_i = \varepsilon_i - X_i'(b - \beta)$$

由于 $p \lim b = \beta$ ，样本越大，这个估计就越好。这有时被称为逐点一致性。可以看出最小乘残差的样本收敛于真正扰动项的样本。这意味着：

$$\hat{\mu}_r = \frac{1}{n} \sum_i \varepsilon_i^r$$

是 μ_r 的一致估计量，

$$m_r = \frac{1}{n} \sum_i e_i^r$$

也是 μ_r 的一致估计量，

通常运用下列公式计算偏度（Skewness）：

$$\begin{aligned} S &= \frac{\left[\mathbb{E}(\mathbf{X} - u_x)^3\right]^2}{\left[\mathbb{E}(\mathbf{X} - u_x)^2\right]^3} \\ &= \frac{\text{三阶矩的平方}}{\text{两阶矩的立方}} \end{aligned} \quad (5.7.1)$$

因为，对于对称的概率密度函数，其三阶矩为零，因为这样的一个概率密度函数，其偏度 S 为零。一个最重要的例子就是正态分布。如果偏度 S 的值为正，则其概率密度为正偏或右偏；如果 S 的值为负，则其概率密度为负偏或左偏。

通常运用下列公式计算峰态（Kurtosis）：

$$K = \frac{\mathbb{E}(\mathbf{X} - u_x)^4}{\left[\mathbb{E}(\mathbf{X} - u_x)^2\right]^2} = \frac{\text{三阶矩的平方}}{\text{两阶矩的立方}} \quad (5.7.2)$$

概率密度的峰度 K 小于 3 时，成为低峰态的（胖的或短尾的），峰度 K 大于 3 时，称为尖峰态的（瘦的或长尾的），见图 5.7.1。正态分布的峰度 K 为 3，这样的概率密度函数称为常峰态的。

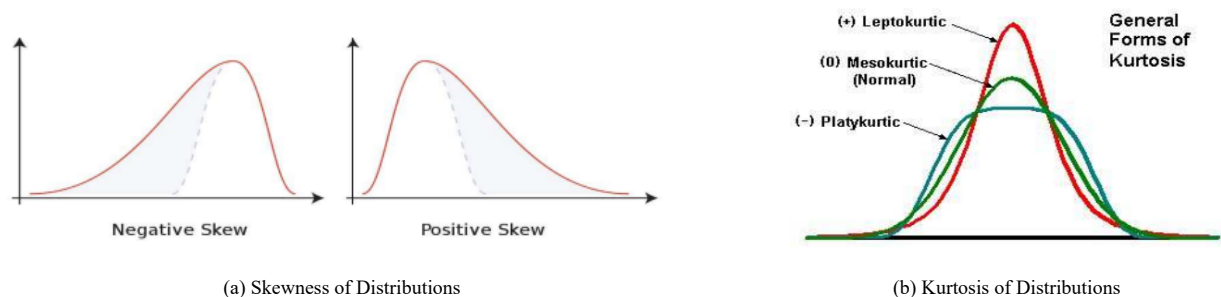


图 5.7.1: 偏度与峰态

§ 5.7.1 样本偏度与样本峰度

根据式 5.7.1 和式 5.7.2，用样本三阶矩和四阶矩来计算样本偏度与峰度。样本三阶矩（与样本方差的计算公式相对照）为：

$$\frac{\sum(\mathbf{X} - \bar{\mathbf{X}})^3}{n - 1} \quad (5.7.3)$$

样本四阶矩为：

$$\frac{\sum(\mathbf{X} - \bar{\mathbf{X}})^4}{n - 1} \quad (5.7.4)$$

前述内容可用于设计正态性的检验。正态分布是对称和常峰态的。对称意味着三阶矩 $\mathbb{E}[\epsilon^3]$ 为 0。分布对称性的标准量是偏态（Skewness）

$$\sqrt{\beta_1} = \frac{\mathbb{E}[\epsilon^3]}{(\sigma^2)^{3/2}}$$

峰态 (Kurtosis) 是分布尾部厚度的度量。此度量是:

$$\beta_2 = \frac{\mathbb{E}[\epsilon^4]}{(\sigma^2)^2}$$

正态分布对于这个度量通常是评价标准; 常峰态值是正态分布的峰度, 等于 3。因此, 我们可以通过比较偏度是否为 0 和峰度是否为 3 来判断该分布是否为正态分布。在实际中, 通常的度量是过量程度 (degree of excess) $(\beta_2 - 3)$ 。我们将使用的工具是一个沃尔德统计量。在正态性的假设下, 此检验统计量是:

$$W = n \left[\frac{b_1}{6} + \frac{(b_2 - 3)^2}{24} \right] \sim \chi^2(2)$$

称为正态性的哈尔克-贝拉 (Jarque-Bera) BJ 检验。

这渐近地服从自由度为 2 的 χ^2 分布。这些参数的可行的估计量是利用最小二乘残差计算而得到的。统计量可以参考标准 χ^2 表。

由贝拉和哈尔克 (1980 a, 1980b) 推导的这个检验统计量的皮尔逊分布的内容中是作为拉格朗日乘数检验。应该注意这个检验本质上是无建设性的。非正态性的发现不一定给出下一步如何做的建议。同样, 注意不能拒绝正态性并没有确认了正态性。这只是一个对称性和常峰态的检验。

§ 5.8 思考题

1. 对于线性统计模型 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 。假设 $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$, $n = 13, K = 3$, 最小化误差平方和 $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ 得到如下线性方程组:

$$\begin{cases} b_1 + 2b_2 + b_3 = 3 \\ 2b_1 + 5b_2 + b_3 = 9 \\ b_1 + b_2 + 6b_3 = -8 \end{cases}$$

(1) 把这个方程组写成矩阵的形式, 并利用矩阵方法求最小二乘估计量 b 的值。

(2) 如果 $\mathbf{y}'\mathbf{y} = 53$, 求 σ^2 的无偏估计量 s^2 的值

(3) 求 b 的协方差矩阵。

(4) 分别写出能够检验 $H_0: \beta_k = \beta_k^0$ 的 t 统计量 ($k=1, 2, 3$)。

(5) 写出能够检验 $H_0: \beta_1 + \beta_2 - 2\beta_3 = q$ 的 t 统计量和 F 统计量。

2. 假设 b 是 \mathbf{y} 关于 \mathbf{X} 的回归的最小二乘估计量, c 是另一 $K \times 1$ 向量, 证明两个残差平方和之差是:

$$(\mathbf{y} - \mathbf{X}c)'(\mathbf{y} - \mathbf{X}c) - (\mathbf{y} - \mathbf{X}b)'(\mathbf{y} - \mathbf{X}b) = (c - b)' \mathbf{X} \mathbf{X}' (c - b)$$

并证明这个差值是正的。

3. 假设对于同一个参数 θ , 你有两个相互独立的无偏估计量 θ^1 和 θ^2 , 它们的方差分别为 v_1 和 v_2 且 $v_1 \neq v_2$, 那么什么样的线性组合 $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ 是 θ 的最小方差无偏估计量?

4. 假设对于同一个参数 θ , 你有 n 个相互独立的无偏估计量 $\hat{\theta}_1, \dots, \hat{\theta}_n$, 它们的方差分别 v_1, \dots, v_n 。那么什么样的线性组合 $\hat{\theta} = c_1\hat{\theta}_1 + \dots + c_n\hat{\theta}_n$ 是 θ 的最小方差无偏估计量?

§ 6.1 最大似然估计

我们假定第五章的 6 个假设条件全部满足，我们就知道了 \mathbf{Y} 的分布函数，我们也就可以用其他方法如最大似然估计和矩估计等来求解出参数 β 和 σ^2 的估计量。在本章中我们用最大似然估计求出参数 β 和 σ^2 的估计量。

利用模型的假设和样本信息，我们首先求出似然函数，它是关于未知数 β 和 σ^2 的函数。由于 $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$ ，因此有 $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ 或者边际分布 $y_t \sim N(\mathbf{x}'_t \beta, \sigma^2)$ ，这里 $\mathbf{x}'_t = (x_{t1}, x_{t2}, \dots, x_{tK})$, $t = 1, 2, \dots, n$, $\{\mathbf{y}_t\}$ 是相互独立的 (Why?)。 y_t 的密度函数为

$$f(y_t | \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left\{ -\frac{(y_t - \mathbf{x}'_t \beta)^2}{2\sigma^2} \right\}$$

所以 y_1, y_2, \dots, y_n 的联合概率密度函数为

$$\begin{aligned} f(\mathbf{y} | \beta, \sigma^2) &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\sum_{t=1}^n \frac{(y_t - \mathbf{x}'_t \beta)^2}{2\sigma^2} \right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \right\} \end{aligned} \quad (6.1.1)$$

如果 y_1, y_2, \dots, y_n 的联合概率密度函数看做是未知参数 β 和 σ^2 的函数，我们称它为似然函数，记为

$$l(\beta, \sigma^2 | \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \right\}$$

§ 6.2 最优化问题

据最大似然法则, 估计未知参数 β 和 σ^2 的问题变成了选择 β 和 σ^2 的值使得对数似然函数的值达到最大, 也即是如下的最优化问题:

$$\max_{\beta, \sigma^2} L(\beta, \sigma^2 | \mathbf{y})$$

这个问题的一阶条件为

$$\begin{aligned} \frac{\partial L}{\partial \beta} &= -\frac{1}{2\sigma^2} \left[\frac{\partial(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{\partial \beta} \right] = -\frac{1}{2\sigma^2} (-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta) = 0 \\ \frac{\partial L}{\partial \sigma^2} &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^4} = 0 \end{aligned} \quad (6.2.1)$$

如果把这两个方程的解分别记为 $\hat{\beta}$ 和 $\hat{\sigma}^2$, 那么它们满足

$$\begin{aligned} \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'\mathbf{y} \\ \hat{\sigma}^2 &= \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{n} \end{aligned}$$

可以解得

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \mathbf{b} \quad (\text{与最小二乘估计量一样}) \quad (6.2.2)$$

$$\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{n} \quad (6.2.3)$$

这里 $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 。所以 β 的最大似然估计和最小二乘估计是一样的。这是由于选择 β 的值最大化对数似然函数和最小化误差平方和 $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$ 是等价的。如果记

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{M}\mathbf{y}$$

并称它为最大似然残差, 那么它和最小二乘残差是相等的, 即 $\hat{\mathbf{e}} = \mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$ 。这样 σ^2 可表示为 $\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{n} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}$

对于 $\tilde{\beta}$, 我们已经知道 $\mathbb{E}[\tilde{\beta}] = \beta$, $\text{cor}(\tilde{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ 。由于 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b}$ 是 \mathbf{y} 的线性函数, 而 \mathbf{y} 服从正态分布, 所以 $\tilde{\beta}$ 也服从正态分布, 均值为 β , 协方差矩阵为 $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, 也即 $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ 。这个结果在进行区间估计和假设检验时是非常有用的。

关于 $\hat{\sigma}^2$, 由于 $\mathbb{E}[\hat{\mathbf{e}}'\hat{\mathbf{e}}] = \mathbb{E}[(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})] = \mathbb{E}[\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}] = \sigma^2(n - K)$, 所以其期望值为

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n - K}{n} \sigma^2$$

它是 σ^2 的一个有偏估计量，其偏度为 $-K\sigma_2/n$ 。为了得到一个无偏估计量，定义

$$s^2 = \frac{\hat{e}'\hat{e}}{n-K}$$

那么它是 σ^2 的一个无偏估计量。它和我们在前一章里得到的关于 σ^2 的无偏估计量是一样的。

CHAPTER 7

带有线性约束的多元线性回归模型及其假设检验

在本章中，继续讨论第五章的模型，但新的模型中，参数 β 满足 J 个线性约束集， $R\beta = q$ ，矩阵 R 有和 β 相一致的 K 列和总共 J 个约束的 J 行，且 R 是行满秩的，我们考虑不是过度约束的情况，因此， $J < K$ 。

带有线性约束的参数的假设检验，我们可以用两种方法来处理：

第一个方法是我们按照无约束条件求出一组参数估计后，然后我们对求出的这组参数是否满足假设所暗示的约束，进行检验，我们在本章的第一节中讨论。

第二个方法是我们把参数所满足的线性约束和模型一起考虑，求出参数的最小二乘解，尔后再作检验，后者就是参数带有约束的最小二乘估计方法，我们在本章的第二节中讨论。

§ 7.1 线性约束的检验

从线性回归模型开始，

$$y = X\beta + \varepsilon \quad (7.1.1)$$

我们考虑具有如下形式的一组线性约束，

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1K}\beta_K &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2K}\beta_K &= q_2 \\ &\vdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{JK}\beta_K &= q_J \end{aligned}$$

这些可以用矩阵改写成一个方程

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q} \quad (7.1.2)$$

作为我们的假设条件 H_0 。

\mathbf{R} 中每一行都是一个约束中的系数。矩阵 \mathbf{R} 有和 $\boldsymbol{\beta}$ 相一致的 K 列和总共 J 个约束的 J 行, 且 \mathbf{R} 是行满秩的。因此, J 一定要小于或等于 K 。 \mathbf{R} 的各行必须是线性无关的, 虽然 $J = K$ 的情况并不违反条件, 但其唯一决定了 $\boldsymbol{\beta}$, 这样的约束没有意义, 我们不考虑这种情况。

给定最小二乘估计量 \mathbf{b} , 我们的兴趣集中于“差异”向量 $\mathbf{d} = \mathbf{R}\mathbf{b} - \mathbf{q}$ 。 \mathbf{d} 精确等于 $\mathbf{0}$ 是不可能的事件 (因为其概率是 0), 统计问题是 \mathbf{d} 对 $\mathbf{0}$ 的离差是否可归因于抽样误差或它是否是显著的。

由于 \mathbf{b} 是多元正态分布的, 且 \mathbf{d} 是 \mathbf{b} 的一个线性函数, 所以 \mathbf{d} 也是多元正态分布的, 若原假设为真, \mathbf{d} 的均值为 $\mathbf{0}$, 方差为

$$\text{var}[\mathbf{d}] = \text{var}[\mathbf{R}\mathbf{b} - \mathbf{q}] = \mathbf{R}(\text{var}[\mathbf{b}])\mathbf{R}' = \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \quad (7.1.3)$$

对 H_0 的检验我们可以将其基于沃尔德 (Wald) 准则:

$$\begin{aligned} W &= \chi^2(J) \\ &= \mathbf{d}'(\text{Var}[\mathbf{d}])^{-1}\mathbf{d} \\ &= (\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \right]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}) \end{aligned} \quad (7.1.4)$$

在假设正确时将服从自由度为 J 的 χ^2 分布。

直觉上, \mathbf{d} 越大, 即最小二乘满足约束的错误越大, χ^2 统计量越大, 所以, 一个大的 χ^2 值将加重对假设的怀疑。

$$\frac{(n-K)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \quad (7.1.5)$$

由于 σ 未知, 7.1.4 中的统计量是不可用的, 用 s^2 替代 σ^2 , 我们可以导出一个 $F[J, (n-K)]$ 样本统计量, 令

$$F = \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \right]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})/J}{[(n-K)s^2/\sigma^2]/(n-K)} \quad (7.1.6)$$

分子是 $(1/J)$ 乘 7.1.4 中的 W , 分母是 $1/(n-K)$ 乘 7.1.5 中的幂等二次型。所以, F 是两个除以其自由度的卡方变量的比率。如果它们是独立的, 则 F 的分布是 $F[J, (n-K)]$, 我们前边发现 \mathbf{b} 是独立于 s^2 分布的, 所以条件是满足的。

我们也可以直接推导。利用 7.1.5 及 M 是幂等的这一事实，我们可以把 F 写为

$$F = \frac{\{\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})/\sigma\}' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} \{\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})/\sigma\}/J}{[\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)]' [\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)]/(n-K)} \quad (7.1.7)$$

由于

$$\frac{\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})}{\sigma} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \left(\frac{\boldsymbol{\varepsilon}}{\sigma} \right) = \mathbf{T} \left(\frac{\boldsymbol{\varepsilon}}{\sigma} \right)$$

F 统计量是 $(\boldsymbol{\varepsilon}/\sigma)$ 的两个二次型的比率，由于 $\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)$ 和 $\mathbf{T}(\boldsymbol{\varepsilon}/\sigma)$ 都服从正态分布且它们的协方差 $\mathbf{T}\mathbf{M}$ 为 $\mathbf{0}$ ，所以二次型的向量都是独立的。 F 的分子和分母都是独立随机向量的函数，因而它们也是独立的。这就完成了证明。

消掉 7.1.6 中的两个 σ^2 ，剩下的是检验一个线性假设的 F 统计量，

$$\begin{aligned} F &= \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})/J}{\mathbf{e}'\mathbf{e}/(n-K)} \\ &= \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})' [\mathbf{s}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})}{J} \end{aligned} \quad (7.1.8)$$

我们将检验统计量

$$F[J, n-K] = \frac{(\mathbf{R}\mathbf{b} - \mathbf{q})' \left\{ \mathbf{R} [\mathbf{s}^2 (\mathbf{X}'\mathbf{X})^{-1}] \mathbf{R}' \right\}^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})}{J}$$

和 F 分布表中的临界值相比较，一个大的 F 值是反对假设的证据。

注意：将 wald 统计量中的 σ^2 用 s^2 去替代，相应的就将 J 维的卡方分布转换为维度为 $(J, n-K)$ 的 F 分布。

§ 7.2 参数带有约束的最小二乘估计

§ 7.2.1 带有约束的最小二乘函数

在许多问题中，要求其中的未知参数 $\boldsymbol{\beta}$ 满足某特定的线性约束条件： $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ ，这里 \mathbf{R} 是 $J \times K$ 矩阵 ($J < K$)，并假定它的秩为 J 维向量，常常希望求 $\boldsymbol{\beta}$ 的估计 $\hat{\boldsymbol{\beta}}$ ，使得

$$\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = \min_{\{\boldsymbol{\beta}: \mathbf{R}\boldsymbol{\beta} = \mathbf{q}\}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad (7.2.1)$$

满足条件 7.1.5 的称为 $\boldsymbol{\beta}$ 的具有线性约束 $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ 的最小二乘估计。

解 $\hat{\boldsymbol{\beta}}$ 的问题实际上是在约束条件

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$$

下求

$$f = \|Y - X\beta\|^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^m x_{ij}\beta_j \right)^2$$

的限制极值点问题。

这个问题的一个拉格朗日解可写作

$$S^* = (y - X\beta)'(y - X\beta) + 2\lambda'(R\beta - q)$$

解 b_* 和 λ 将满足必要条件

$$\begin{aligned} \frac{\partial S^*}{\partial \beta} &= -2X'(y - Xb_*) + 2R'\lambda = 0 \\ \frac{\partial S^*}{\partial \lambda} &= 2(Rb_* - q) = 0 \end{aligned}$$

展开可以得到分块矩阵方程

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} b_* \\ \lambda \end{bmatrix} = \begin{bmatrix} X'y \\ q \end{bmatrix} \quad (7.2.2)$$

或

$$Wd_* = v$$

假定括号中的分块矩阵是非奇异的, 约束最小二乘估计量

$$d_* = W^{-1}v = \begin{bmatrix} b_* \\ \lambda \end{bmatrix} \quad (7.2.3)$$

其中

$$W^{-1} = \begin{bmatrix} (X'X)^{-1} - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1} & (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1} \\ (R(X'X)^{-1}R')^{-1}R(X'X)^{-1} & - (R(X'X)^{-1}R')^{-1} \end{bmatrix}$$

的解。此外, 若 $X'X$ 是非奇异的, 则用分块逆公式可以得到 b_* 和 λ 的显示解

$$\begin{aligned} b_* &= (X'X)^{-1}X'y - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1}X'y + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}q \\ &= (X'X)^{-1}X'y - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1}X'(Xb + e) + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}q \\ &= (X'X)^{-1}X'y - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}Rb + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}q \\ &= b - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(Rb - q) \end{aligned}$$

和

$$\lambda = [R(X'X)^{-1}R']^{-1}(Rb - q)$$

格林和西克斯 (1991) 表明 \mathbf{b}_* 的协方差矩阵简单地就是 σ^2 乘以 W^{-1} 的左上块, 在 $\mathbf{X}'\mathbf{X}$ 非奇异的通常情况下, 再一次可以得到一个显性公式

$$\text{Var}[\mathbf{b}_*] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} - \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} \mathbf{R} (\mathbf{X}'\mathbf{X})^{-1}$$

这样,

$$\text{Var}[\mathbf{b}_*] = \text{Var}[\mathbf{b}] - (\text{一个非负定矩阵})$$

$\text{Var}[\mathbf{b}_*]$ 的方差比 $\text{Var}[\mathbf{b}]$ 小的一个解释是约束条件提供了更多的信息价值。

§ 7.2.2 对约束的检验的另一个方法

令 $\mathbf{e}_* = \mathbf{y} - \mathbf{X}\mathbf{b}_*$, 我们来计算新的离差平方和 $\mathbf{e}_*' \mathbf{e}_*$ 。

$$\mathbf{e}_* = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{X}(\mathbf{b}_* - \mathbf{b}) = \mathbf{e} - \mathbf{X}(\mathbf{b}_* - \mathbf{b})$$

则新的离差平方和是

$$\mathbf{e}_*' \mathbf{e}_* = \mathbf{e}' \mathbf{e} + (\mathbf{b}_* - \mathbf{b})' \mathbf{X}' \mathbf{X} (\mathbf{b}_* - \mathbf{b}) \geq \mathbf{e}' \mathbf{e}$$

其中,

$$\frac{\mathbf{e}' \mathbf{e}}{\sigma^2} \sim \chi_{n-k}^2 \quad \frac{\mathbf{e}_*' \mathbf{e}_*}{\sigma^2} \sim \chi_{n-(k-J)}^2$$

因为新的模型中参数的个数为 $k - J$ 个, J 个榆树条件是原模型中的 J 个参数可以被其他 $k - J$ 个表示。(此表达式中的中间项含有 $\mathbf{X}'\mathbf{e}$, 它是 $\mathbf{0}$)。这说明我们可以将一个约束检验基于拟合的损失。这个损失是,

$$\mathbf{e}_*' \mathbf{e}_* - \mathbf{e}' \mathbf{e} = (\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})$$

这出现在前边推导的 F 统计量的分子上, 我们得到统计量的另一个可选形式。可选形式是

$$F[J, n - K] = \frac{(\mathbf{e}_*' \mathbf{e}_* - \mathbf{e}' \mathbf{e}) / J}{\mathbf{e}' \mathbf{e} / (n - K)}$$

最后, 以 $\text{SST} = \Sigma(y - \bar{y})^2$ 除 F 的分子和分母, 我们得到第三种形式,

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

由于两个模型的拟合之差直接体现在检验统计量中, 这个形式具有一些直观吸引力。

§ 7.2.3 实例

所有柯布一道格拉斯模型的一般化是如下的对数变换模型，

$$\log Y = \beta_1 + \beta_2 \log L + \beta_3 \log K + \beta_4 \frac{\log^2 L}{2} + \beta_5 \frac{\log^2 K}{2} + \beta_6 \frac{\log L \log K}{2} + \varepsilon \tag{7.2.4}$$

无约束回归的结果在表7.1中给出。

表 7.1 无约束回归的结果

回归标准误差	0.17994
残差平方和	0.67993
R 平方	0.95486
调整 R 平方	0.94411

表 7.2 无约束回归的系数结果

变量	系数	标准误差	t 值
常数项	0.944216	2.911	0.324
$\log L$	3.61363	1.548	2.334
$\log K$	- 1.89311	1.016	- 1.863
$\frac{1}{2} \log^2 L$	- 0.96406	0.7074	- 1.363
$\frac{1}{2} \log^2 K$	0.08529	0.2926	0.291
$\log L \times \log K$	0.31239	0.4389	0.71

表 7.3 系数估计量的估计协方差矩阵

	常数项	$\log L$	$\log K$	$\frac{1}{2} \log^2 L$	$\frac{1}{2} \log^2 K$	$\log L \times \log K$
常数项	8.472					
$\log L$	- 2.388	2.397				
$\log K$	- 0.3313	- 1.231	1.033			
$\frac{1}{2} \log^2 L$	- 0.08760	- 0.6658	0.5231	0.5004		
$\frac{1}{2} \log^2 K$	0.2332	0.03477	0.02637	0.1467	0.08562	
$\log L \times \log K$	0.3635	0.1831	- 0.2255	- 0.2880	- 0.1160	0.1927

考虑了约束条件 $\beta_4 = \beta_5 = \beta_6 = 0$ 的模型就可以得到柯布一道格拉斯模型：

$$\log Y = \beta_1 + \beta_2 \log L + \beta_3 \log K + \varepsilon$$

这是一个条件约束下的无条件的多元线性回归模型。就可以用一般线性回归的方法求解模型。假如我们通过有约束条件下的无条件的多元线性回归模型得到： $\mathbf{e}_*'\mathbf{e}_* = 0.85163$ ，而且 $n - K = 21$ ，则柯布一道格拉斯

模型假设的 F 统计量是

$$F[3, 21] = \frac{(0.85163 - 0.67993)/3}{0.67993/21} = 1.768$$

查自 F 分布表的 5% 临界值是 3.07, 所以我们不能拒绝柯布一道格拉斯模型是适当的这一假设。

考虑了约束条件 $\beta_4 = \beta_5 = \beta_6 = 0$ 和条件 $\beta_2 + \beta_3 = 1$ 的模型就是满足规模效应的柯布一道格拉斯生产函数。这个模型可以推导如下:

$$\begin{aligned}\log Y &= \beta_1 + \beta_2 \log L + \beta_3 \log K + \varepsilon \\ &= \beta_1 + \beta_2 \log L + (1 - \beta_2) \log K + \varepsilon \\ \implies \log Y - \log L &= \beta_1 + \beta_2 (\log L - \log K) + \varepsilon\end{aligned}\tag{7.2.5}$$

假如我们通过有约束条件下的无条件的多元线性回归模型得到: $e_*'e_* = 0.89172$, 而且 $n - K = 21$, 则柯布一道格拉斯模型假设的 F 统计量是

$$F[4, 21] = \frac{(0.89172 - 0.67993)/4}{0.67993/21} = 1.635$$

查自 F 分布表的 5% 临界值是 2.85, 所以我们不能拒绝柯布一道格拉斯模型是规模效应的生产函数的这一假设。

§ 7.3 结构变化与邹至庄检验 (Structure Change and Chou-Test)

§ 7.3.1 问题提出

我们经常碰到这样的问题。某项政策的出台及实施, 其效果如何? 不同地区或不同时期内, 我们分别可以得到这两个地区或时期的观测值, 我们的问题是: 这两个地区或时期的情况是否不同, 经济结构有无差异。

这类问题, 被华人经济学家邹至庄用构造的 F 检验解决了 (1960 年)。这样的 F 检验的统计量, 就称为邹至庄检验 (Chou-Test)。

§ 7.3.2 问题的模型表述

设 (Z_1, Y_1) , (Z_2, Y_2) 分别表示这两个时期的观测值, 允许两个时期中系数不同的无约束回归是

$$\begin{cases} Y_1 = Z_1 \beta_1 + \varepsilon_1 \\ Y_2 = Z_2 \beta_2 + \varepsilon_2 \end{cases}$$

我们可以将其改写成一個回归方程:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}\tag{7.3.1}$$

即 $Y = Z\beta + \varepsilon$ 模型, 其中

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, Z = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

上述问题就转换成检验

$$H_0: \beta_1 = \beta_2$$

$$H_1: \beta_1 \neq \beta_2$$

的问题。我们可以用两种方式来处理问题:

1. 用约束条件 $\beta_1 = \beta_2$ 来检验。 $\beta_1 = \beta_2$ 是更一般约束条件 $R\beta = q$ 的一个特殊形式, 其中 $R = (I, -I)$ 和 $q = 0$ 。这个直接可以从基于 **Wald 统计量** 的带约束条件的 F 检验得到。(请自己推导)。

Example 7.1 用约束条件下的 F 检验推导出邹至庄检验的表达式。

Proof 7.1 在约束条件 $R\beta = q$ 下, F 检验

$$F(J, n - k) = \frac{(Rb - q)' [S^2 R (Z'Z)^{-1} R']^{-1} (Rb - q)}{J}$$

而邹至庄检验时约束条件 $R\beta = q$ 的一种特殊形式, 即 $R = (I, -I)$ 和 $q = 0$, 也即等同于条件 $\beta_1 = \beta_2$ (有 $2k$ 个参数, 并且是有 k 个约束)。故

$$\begin{aligned} F(k, n_1 + n_2 - 2k) &= \frac{(Rb - q)' [S^2 R (Z'Z)^{-1} R']^{-1} (Rb - q)}{k} \\ &= \frac{(b_1 - b_2)' \left[S^2 (I, -I) \begin{pmatrix} (Z_1'Z_1)^{-1} & 0 \\ 0 & (Z_2'Z_2)^{-1} \end{pmatrix} \begin{pmatrix} I \\ -I \end{pmatrix} \right]^{-1} (b_1 - b_2)}{k} \\ &= \frac{(b_1 - b_2)' \left[S^2 \left((Z_1'Z_1)^{-1} + (Z_2'Z_2)^{-1} \right) \right]^{-1} (b_1 - b_2)}{k} \end{aligned}$$

服从 $F(k, n_1 + n_2 - 2k)$ 的分布。

另外, 在考虑了约束条件 $\beta_1 = \beta_2$ 后, 我们可以将模型 7.3.1 改写成一个无约束的新的回归方程

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

即

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \beta_1 + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \quad (7.3.2)$$

即无约束的线性模型即 $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 模型, 其中

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix}, \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2, \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}$$

假如模型 7.3.2 的残差平方和是 $\mathbf{e}'_*\mathbf{e}_*$, 在假设条件 $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ 下, 我们可以得到 F 统计量可更简单地表示为:

$$F[k, n_1 + n_2 - 2k] = \frac{(\mathbf{e}'_*\mathbf{e}_* - \mathbf{e}'\mathbf{e})/k}{(\mathbf{e}'\mathbf{e})/(n_1 + n_2 - 2k)}$$

2. 更直接、更容易的一个处理是将约束直接构造进模型中。若两个系数向量相同, 则模型 7.3.1 就转换为 7.3.3 :

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix} \quad (7.3.3)$$

由此我们推导出可以检验的邹至庄统计量 Chou-Test。

从模型 7.3.1 中, 我们可以得到无约束最小二乘估计量是

$$\begin{aligned} \mathbf{b} &= (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{Y} \\ &= \begin{pmatrix} \mathbf{Z}'_1\mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}'_2\mathbf{Z}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Z}'_1\mathbf{Y}_1 \\ \mathbf{Z}'_2\mathbf{Y}_2 \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{Z}'_1\mathbf{Z}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{Z}'_2\mathbf{Z}_2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{Z}'_1\mathbf{Y}_1 \\ \mathbf{Z}'_2\mathbf{Y}_2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \end{aligned}$$

故

$$\begin{aligned} \mathbf{e} = \mathbf{Y} - \mathbf{Zb} &= \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} - \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{pmatrix} \begin{pmatrix} (\mathbf{Z}'_1\mathbf{Z}_1)^{-1} \mathbf{Z}'_1\mathbf{Y}_1 \\ (\mathbf{Z}'_2\mathbf{Z}_2)^{-1} \mathbf{Z}'_2\mathbf{Y}_2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} - \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{pmatrix} \begin{pmatrix} (\mathbf{Z}'_1\mathbf{Z}_1)^{-1} \mathbf{Z}'_1 & \mathbf{0} \\ \mathbf{0} & (\mathbf{Z}'_2\mathbf{Z}_2)^{-1} \mathbf{Z}'_2 \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \\ &= \left[\mathbf{I} - \begin{pmatrix} \mathbf{Z}_1 (\mathbf{Z}'_1\mathbf{Z}_1)^{-1} \mathbf{Z}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 (\mathbf{Z}'_2\mathbf{Z}_2)^{-1} \mathbf{Z}'_2 \end{pmatrix} \right] \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} \\ &= \mathbf{M}_1 \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \end{aligned}$$

因此

$$\mathbf{e}'\mathbf{e} = (\mathbf{Y}'_1\mathbf{Y}'_2) \mathbf{M}'_1\mathbf{M}_1 \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = (\mathbf{Y}'_1\mathbf{Y}'_2) \mathbf{M}_1 \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$$

则 $\mathbf{e}'\mathbf{e}/\sigma^2 \sim \chi^2(n_1 + n_2 - 2k)$ 。

对于有约束条件 $\beta_1 = \beta_2$ 限制的模型 7.3.3

$$\begin{aligned} e^* &= \left(I - \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \left((Z_1' Z_2') \cdot \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \right)^{-1} (Z_1' Z_2') \right) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &= \left(I - \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} (Z_1' Z_1 + Z_2' Z_2)^{-1} (Z_1' Z_2') \right) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &= M_2 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \end{aligned}$$

因此

$$e^{*'} e^* = (Y_1' Y_2') M_2' M_2 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = (Y_1' Y_2') M_2 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

则 $e^{*'} e^* / \sigma^2 \sim \chi^2(n_1 + n_2 - k)$ 。

$$e^{*'} e^* - e' e = (Y_1' Y_2') (M_2 - M_1) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

那么, $(e^{*'} e^* - e' e) / \sigma^2$ 服从何分布?

Proof 7.2 首先证明: $M_3 M_1 = 0$

$$\begin{aligned} (M_2 - M_1) M_1 &= M_2 M_1 - M_1^2 \\ &= M_2 M_1 - M_1 \\ &= \left[I - \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} (Z_1' Z_1 + Z_2' Z_2)^{-1} (Z_1' Z_2') \right] \cdot M_1 - M_1 \\ &= - \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} (Z_1' Z_1 + Z_2' Z_2)^{-1} (Z_1' Z_2') \left[I - \begin{pmatrix} Z_1 (Z_1' Z_1)^{-1} Z_1' & 0 \\ 0 & Z_2 (Z_2' Z_2)^{-1} Z_2' \end{pmatrix} \right] \\ &= - \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} (Z_1' Z_1 + Z_2' Z_2)^{-1} (Z_1' Z_2') + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} (Z_1' Z_1 + Z_2' Z_2)^{-1} (Z_1' Z_2') = 0 \end{aligned}$$

故 $M_2 - M_1 + M_1 = M_2$ 而且 $(M_2 - M_1) M_1 = 0$

故 $r(M_2 - M_1) = r(M_2) - r(M_1) = n_1 + n_2 - k - (n_1 + n_2 - 2k) = k$

同样 $(M_2 - M_1)$ 是幂等矩阵

故 $(e^{*'} e^* - e' e) / \sigma^2 \sim \chi^2(k)$ 且与 $e' e / \sigma^2 \sim \chi^2(n_1 + n_2 - 2k)$ 是独立的, 所以

$$F[k, n_1 + n_2 - 2k] = \frac{(e^{*'} e^* - e' e) / k}{(e' e) / (n_1 + n_2 - 2k)}$$

这个就是邹至庄检验统计量 (**Chou-Test**)。

Part III

应用

CHAPTER 8

古典线性回归的大样本理论

迄今为止的讨论涉及了最小二乘估计量的有限样本性质。根据非随机回归量和扰动项正态分布这两个假设，我们知道了最小二乘估计量的精确分布和一些检验统计量。

在本章中，我们去总结前一章关于最小二乘法的有限样本特性，然后我们重点讨论古典回归模型的大样本结果。

§ 8.1 最小二乘法的有限样本特性

古典回归模型的基本假设是：

I $y = \beta'X + \epsilon$

II X 是秩为 K 的 $n \times K$ 非随机矩阵

III $E[\epsilon] = 0$

IV $E[\epsilon\epsilon'] = \sigma^2 I$

未知参数 β 和 σ 的最小二乘估计量是：

$$b = (X'X)^{-1} X'y \quad s^2 = \frac{e'e}{(n-K)}$$

通过分析

$$b = \beta + (X'X)^{-1} X'\epsilon \quad s^2 = \frac{\epsilon'M\epsilon}{n-K}$$

我们可得下列精确的有限样本结果：

1. $\mathbb{E}[\mathbf{b}] = \boldsymbol{\beta}$, (最小二乘估计是无偏的)
2. $\text{Var}[\mathbf{b}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$
3. 任意函数 $\mathbf{r}'\boldsymbol{\beta}$ 的最小方差线性无偏估计量是 $\mathbf{r}'\mathbf{b}$ (这就是高斯—马尔科夫定理)
4. $\mathbb{E}[s^2] = \sigma^2$
5. $\text{Cov}[\mathbf{b}, \mathbf{e}] = 0$

为了构造置信区间和检验假设, 我们根据正态分布的假设 $V.\boldsymbol{\varepsilon} \sim N[0, \sigma^2 \mathbf{I}]$ 推导额外的结果, 即

6. \mathbf{b} 和 \mathbf{e} 在统计上是相互独立的。相应的, \mathbf{b} 和 s^2 无关并在统计上相互独立。
7. \mathbf{b} 的精确分布依赖于 X , $N[\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}]$
8. $(n-K)s^2/\sigma^2$ 的分布是 $\chi^2[n-K]$ 。 s^2 的均值是 σ^2 , 方差是 $2\sigma^4/(n-K)$
9. 根据 6 至 8 结果, 统计量 $t[n-K] = \frac{b_k - \beta_k}{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}$ 服从自由度为 $n-K$ 的 t 分布。
10. 用于检验一组 J 个线性约束 $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ 的检验统计量

$$\frac{(\mathbf{Rb} - \mathbf{q})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{q})/J}{\mathbf{e}'\mathbf{e}/(n-K)} = \frac{(\mathbf{Rb} - \mathbf{q})' [\mathbf{R}s^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{q})}{J}$$

服从自由度为 J 和 $n-K$ 的 F 分布。

注意, 利用 I 至 IV 建立起来的 \mathbf{b} 的各种性质和根据扰动项更进一步的正态分布假设而得到的额外推断结果之间的区别。第一组中最重要的结果是高斯—马尔科夫定理, 它与扰动项的分布无关。根据正态分布假设得到的重要的附加结果是 7、8、9、10。正态性没有产生任何额外的有限样本的最优性结果。(没有得出额外的有关统计量好坏的结论)

§ 8.2 古典回归模型的渐近分布理论

为什么要用大样本理论?

在 OLS 的方法中, 我们如果用数据得到的 wald 统计量: $W = n \left[\frac{b_1}{6} + \frac{(b_2-3)^2}{24} \right] \sim \chi^2(2)$

通不过检验, 即假设 $V.\boldsymbol{\varepsilon} \sim N[0, \sigma^2 \mathbf{I}]$ 不满足, 这样的话我们就不能用 OLS 完成相关的假设检验问题, 所以我们要用到中心极限定理: 在 n 足够大的情况下, \mathbf{Y} 和 $\boldsymbol{\varepsilon}$ 都服从正态分布。这样, 相应的判别估计量好坏的方法和标准要作相应的调整, 其中重要的概念是一致估计量。虽然估计量有可能相同, 但我们关心的是他们的一致性, 而不是无偏性。

所以我们要区分那些结论是可以在没有正态性的假设下仍然成立的，利用这些条件来推断最小二乘系数估计量的一致性。

对于满足 I 到 IV 假设的模型，可以直接推导大样本最小二乘估计量的特性

§ 8.2.1 依概率分布

Theorem 8.2.1 (依概率分布) 从具有有限均值 μ 和有限方差 σ^2 的任何总体中抽取的随机样本的均值都是 μ 的一个一致估计量

Proof 8.1 $\mathbb{E}[\bar{x}] = \mu$ 及 $\text{Var}[\bar{x}] = \sigma^2/n$ ，所以， \bar{x} 依均方收敛于 μ ，或 $p \lim \bar{x} = \mu$

Theorem 8.2.2 (斯拉茨基定理 (Slutsky)) 对一个不是 n 的函数的连续函数 $g(x_n)$ 有

$$p \lim g(x_n) = g(p \lim x_n)$$

假设

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}' \mathbf{X} = \mathbf{Q} \quad \text{是正定矩阵} \quad (8.2.1)$$

这个假设在大多数时候是不过份的，考虑一元的情况： $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$

$$\frac{1}{n} \mathbf{X}' \mathbf{X} = \frac{1}{n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \frac{n-1}{n} s^2 + \bar{x} \end{pmatrix}$$

我们知道:

$$\begin{aligned} p \lim \bar{x} &= \mu \\ p \lim \frac{\sum x_i^2}{n} &= \lim \left[\frac{n-1}{n} s^2 + \bar{x}^2 \right] = \sigma^2 + \mu^2 \\ \therefore \lim \frac{1}{n} \mathbf{X}' \mathbf{X} &= \begin{pmatrix} 1 & \mu \\ \mu & \sigma^2 + \mu^2 \end{pmatrix} \end{aligned}$$

which is positive definite as its principal submatrices all have positive determinants.

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \left(\frac{1}{n} \mathbf{X}' \boldsymbol{\epsilon} \right) \quad (8.2.2)$$

假设 Q^{-1} 存在, 因为逆矩阵是原矩阵的连续函数, 我们得到:

$$p \lim \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} p \lim \left(\frac{1}{n} \mathbf{X}' \boldsymbol{\varepsilon} \right)$$

现在我们需要最后一项的概率极限。令

$$\bar{w} = \frac{1}{n} \mathbf{X}' \boldsymbol{\varepsilon} = \frac{1}{n} \sum_i x_i \varepsilon_i = \frac{1}{n} \sum_i w_i, \quad \text{其中 } x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}, \text{ 为 } \mathbf{X}' \text{ 的列向量}$$

那么

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \bar{w}$$

且

$$p \lim \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} p \lim \bar{w}$$

因为, \mathbf{X} 是非随机矩阵, 所以

$$\mathbb{E}[\bar{w}] = \frac{1}{n} \mathbf{X}' \mathbb{E}[\boldsymbol{\varepsilon}] = 0$$

且

$$\text{Var}[\bar{w}] = \mathbb{E}[\bar{w} \bar{w}] = \frac{1}{n} \mathbf{X}' \mathbb{E}[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}'] \mathbf{X} \frac{1}{n} = \frac{\sigma^2}{n} \left(\frac{\mathbf{X}' \mathbf{X}}{n} \right)$$

于是可得:

$$\lim_{n \rightarrow \infty} \text{Var}[\bar{w}] = 0 \cdot Q = 0$$

由于 w 的均值是 0, 并且它的方差收敛于 0, 所以 \bar{w} 按均方收敛于 0, 且 $p \lim \bar{w} = 0$ 。(下面定理揭示了 r -阶收敛与依概率收敛的关系)

Theorem 8.2.3

$$\xi_n \xrightarrow{r} \xi \Rightarrow \xi_n \xrightarrow{P} \xi$$

因此

$$p \lim \left(\frac{1}{n} \mathbf{X}' \boldsymbol{\varepsilon} = 0 \right) \tag{8.2.3}$$

所以

$$p \lim \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot 0 = \boldsymbol{\beta} \tag{8.2.4}$$

这表明了在古典回归模型中, 在假设 (1) 条件下 \mathbf{b} 是 $\boldsymbol{\beta}$ 的一致估计量。

§ 8.2.2 最小二乘估计量的渐近正态性

为了导出最小二乘估计量的渐近分布，利用以前结果可得：

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left(\frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon}$$

由于逆矩阵是原矩阵的连续函数， $\lim_{n \rightarrow \infty} (\mathbf{X}'\mathbf{X}/n)^{-1} = \mathbf{Q}^{-1}$ 因此，如果极限分布存在，则统计量的极限分布与下式相同：

$$\left[\lim_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \right] \left(\frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1} \left(\frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} \quad (8.2.5)$$

因此，我们必须建立下式的极限分布：

$$\frac{1}{\sqrt{n}} \mathbf{X}'\boldsymbol{\varepsilon} = \sqrt{n}(\bar{w} - \mathbb{E}[\bar{w}])$$

其中 $\mathbb{E}[\bar{w}] = 0$ 我们可以利用林德伯格-费勒形式的中心极限定理得到 $\sqrt{n}\bar{w}$ 利用定理中的表达式，

$$\bar{w} = \frac{1}{n} \sum_i x_i \varepsilon_i$$

是 n 个互不相关的随机向量 $\mathbf{x}_i' \boldsymbol{\varepsilon}_i$ 的平均值，其中 $\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}$

ε_i 的均值为 0，方差为：

$$\text{Var}[x_i \varepsilon_i] = \sigma^2 \mathbf{x}_i \mathbf{x}_i' = \sigma^2 \mathbf{Q}_i$$

$\sqrt{n}\bar{w}$ 的方差；

$$\begin{aligned} \sigma^2 \bar{Q}_n &= \sigma^2 \left(\frac{1}{n} \right) [Q_1 + Q_2 + \cdots + Q_n] \\ &= \sigma^2 \left(\frac{1}{n} \right) \sum_i \mathbf{x}_i' \mathbf{x}_i = \sigma^2 \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right) \end{aligned}$$

下列结果的正式证明是根据林德伯格-费勒形式的中心极限定理，由施密特（1976）和怀特（1984）给出。
如果

1. 扰动项都服从具有零均值和有限方差 σ^2 的同样的分布。
2. \mathbf{X} 的元素受到限制使得 $|x_{tk}|$ 有限并且 $\lim (\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$ 是一个有限正定矩阵。则

$$\left(\frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[0, \sigma^2 \mathbf{Q}] \quad (8.2.6)$$

(这也是为什么我们要假设 \mathbf{Q} 是正定的, 因为正态的协方差都是正定的)

我们利用这一结果可得, 即作一个变换:

$$\mathbf{Q}^{-1} \left(\frac{1}{\sqrt{n}} \right) \mathbf{X}' \boldsymbol{\varepsilon} \xrightarrow{\text{stackreld}} N[\mathbf{Q}^{-1} \mathbf{0}, \mathbf{Q}^{-1} (\sigma^2 \mathbf{Q}) \mathbf{Q}^{-1}]$$

根据 8.2.5 式

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[0, \sigma^2 \mathbf{Q}^{-1}]$$

我们可以得到 \mathbf{b} 的渐近分布 (不加证明):

$$\mathbf{b} \xrightarrow{a} N \left[\boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}^{-1} \right]$$

§ 8.2.3 标准检验统计量的渐近行为

如果没有 $\boldsymbol{\varepsilon}$ 的正态性, 前面给出的 t , F 和 χ^2 统计量则不会服从相应的这些分布。因为

$$b \xrightarrow{a} N \left[\beta, \frac{\sigma^2}{n} \mathbf{Q}^{-1} \right]$$

由此得出

$$\theta_k = \frac{b_k - \beta_k}{[(\sigma^2/n) \mathbf{Q}_{kk}^{-1}]^{1/2}}$$

的渐近分布是标准正态分布。

由于 $p \lim s^2 (\mathbf{X}' \mathbf{X} / n)^{-1} = \sigma^2 \mathbf{Q}^{-1}$ (在下一节中将证明 $p \lim s^2 = \sigma^2$ 这个结果)

$$t_k = \frac{b_k - \beta_k}{[s^2 (\mathbf{X}' \mathbf{X})_{kk}^{-1}]^{1/2}}$$

将与 θ_k 有同样的渐近分布。因此, 我们可以认为, 关于 $\boldsymbol{\beta}$ 的一个元素的假设的通常统计量服从标准正态分布, 而不是 t 分布。(也就是大样本情况下, 没有 t 分布了, 相应的 t 分布是正态分布。)

用于检验一组线性约束的 F 统计量:

$$F = \frac{(\mathbf{e}'_* \mathbf{e}_* - \mathbf{e}' \mathbf{e}) / J}{\mathbf{e}' \mathbf{e} / (n - K)} = \frac{(\mathbf{Rb} - \mathbf{q})' \left[\mathbf{R} \left(s^2 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{R}' \right)^{-1} (\mathbf{Rb} - \mathbf{q}) \right]}{J}$$

不再是 F 分布, 因为分子和分母都不是要求的 χ^2 分布。不过, 沃尔德统计量 $JF[J, n - K]$ 渐近地服从 χ^2 分布并可以用来替代使用。这与扰动项正态分布情况的结果相同。在通常的假设下, 无论扰动项是否服从正态分布, 在处理古典模型的大样本时, 沃尔德统计量都可使用。

Theorem 8.2.4 (沃尔德统计量的极限分布定理) 如果 $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[0, \sigma^2 \mathbf{Q}^{-1}]$ 以及 $H_0 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = 0$ 是正确的, 那么

$$W = (\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\mathbf{R} \left(s^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right)^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}) \right] = JF$$

依分布收敛于自由度为 J 的 χ^2 统计量。(我们不要求正式严格证明)。

特别提醒与注意: 模型的整体检验统计量 这个沃尔德统计量就是可以用来作为我们模型的整体检验, 只不过检验时, 这里的 $\mathbf{R} = \mathbf{I}$, 而 $\mathbf{q} = 0$ 而已。但注意沃尔德统计量 W 是自由度为 J 的 χ^2 统计量, 而不再是用 F 分布来检验了。但 $W = JF$ 。

Proof 8.2 (定理的证明) 由于 \mathbf{R} 是常数矩阵,

$$\sqrt{n}\mathbf{R}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[0, \mathbf{R}(\sigma^2 \mathbf{Q}^{-1}) \mathbf{R}'] \quad (8.2.7)$$

又 $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, 因此

$$\sqrt{n}(\mathbf{R}\mathbf{b} - \mathbf{q}) \xrightarrow{d} N[0, \mathbf{R}(\sigma^2 \mathbf{Q}^{-1}) \mathbf{R}'] \quad (8.2.8)$$

为方便起见, 将此写成:

$$\mathbf{z} \xrightarrow{d} N[0, \mathbf{P}] \quad (8.2.9)$$

令 \mathbf{T} 满足 $\mathbf{T}^2 = \mathbf{P}^{-1}$, 并把 \mathbf{T} 记为 $\mathbf{P}^{-\frac{1}{2}}$, 即 \mathbf{T} 是 \mathbf{P} 的逆平方根。

如果 $\mathbf{z} \xrightarrow{d} N[0, \mathbf{P}]$, 那么 $\mathbf{P}^{-1/2}\mathbf{z} \xrightarrow{d} N[0, \mathbf{P}^{-1/2}\mathbf{P}\mathbf{P}^{-1/2}] = N[0, \mathbf{I}]$ 现在, 我们对随机变量函数的极限分布利用斯拉茨基 (Slutsky) 定理, 无关的 (即, 相互独立) 标准正态分布变量的平方和服从 χ^2 分布。因此, 有下面的极限分布

$$(\mathbf{P}^{-1/2}\mathbf{z})' (\mathbf{P}^{-1/2}\mathbf{z}) = \mathbf{z}' \mathbf{P}^{-1} \mathbf{z} \xrightarrow{d} \chi^2(J) \quad (8.2.10)$$

再结合前面的各部分, 不难证明:

$$\mathbf{z}' \mathbf{P}^{-1} \mathbf{z} = n(\mathbf{R}\mathbf{b} - \mathbf{q})' [\mathbf{R}(\sigma^2 \mathbf{Q}^{-1}) \mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}) \xrightarrow{d} \chi^2(J) \quad (8.2.11)$$

即我们已经证明了其极限分布是自由度为 J 的 χ^2 分布。

由于 $\mathbf{P} \lim s^2 (\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2 \mathbf{Q}^{-1}$ (在下一节中将证明这个结果), 这样:

$n(\mathbf{R}\mathbf{b} - \mathbf{q})' [\mathbf{R} (s^2 (\mathbf{X}'\mathbf{X}/n)^{-1}) \mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})$ 的极限分布式与 8.2.11 式的极限分布是一样的。约去 n , 对左边进行整理就得到沃尔德统计量 W 。证明完毕。

注意: 沃尔德统计量 W 可以用 J 乘以通常的 F 的统计量而得到。 F 仍然是 OLS 得到的 F 统计量。

§ 8.3 s^2 的一致性和 $\text{Var}[\mathbf{b}]$ 的估计量

本节证明上节用到的结果 $p\lim s^2 (\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2 \mathbf{Q}^{-1}$ 的假设, 即证明 s^2 对 σ^2 的一致性, 也就是证明 $p\lim s^2 = \sigma^2$ 。展开

$$s^2 = \frac{1}{n-K} \boldsymbol{\varepsilon}' \mathbf{M} \boldsymbol{\varepsilon}$$

可得

$$\begin{aligned} s^2 &= \frac{1}{n-K} \left[\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\varepsilon} \right] \\ &= \frac{n}{n-K} \left[\frac{\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}}{n} - \left(\frac{\boldsymbol{\varepsilon}' \mathbf{X}}{n} \right) \left(\frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \left(\frac{\mathbf{X}' \boldsymbol{\varepsilon}}{n} \right) \right] \end{aligned}$$

最前面的常数显然收敛于 1, 括号中第一项依概率收敛于 σ^2 , 因为: $\frac{\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}}{n} = \frac{1}{n} \sum \varepsilon_i^2$ 而且: $\mathbb{E}(\varepsilon_i^2) = \text{Var}(\varepsilon_i) = \sigma^2$

$$\text{Var}(\varepsilon_i^2) = \mathbb{E}(\varepsilon_i^4) - [\mathbb{E}(\varepsilon_i^2)]^2 = \mathbb{E}(\varepsilon_i^4) - \sigma^4$$

因为有: (定理从具有有限均值 μ 和有限方差 σ^2 的任何总体中抽取的随机样本的均值都是 μ 的一个一致估计量。P357(大 Green))

所以只有在 $\mathbb{E}(\varepsilon_i^4)$ 为有限的情况下, $\frac{\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}}{n}$ 是 σ^2 的一致估计量。

所以我们要假设 $\mathbb{E}(\varepsilon_i^4)$ 是有限的。这意味着

$$p\lim s^2 = \sigma^2 - p\lim \left(\frac{\boldsymbol{\varepsilon}' \mathbf{X}}{n} \right) \left(\frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \left(\frac{\mathbf{X}' \boldsymbol{\varepsilon}}{n} \right)$$

单独看 $p\lim s^2$ 的第二项, 略微整理之后, 我们有

$$\left(\frac{\boldsymbol{\varepsilon}' \mathbf{X}}{n} \right) \left(\frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \left(\frac{\mathbf{X}' \boldsymbol{\varepsilon}}{n} \right) = \left(\frac{1}{n} \right) \left(\frac{\boldsymbol{\varepsilon}' \mathbf{X}}{\sqrt{n}} \right) \left(\frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \left(\frac{\mathbf{X}' \boldsymbol{\varepsilon}}{\sqrt{n}} \right)$$

这个统计量的大样本特性与

$$q = \left(\frac{1}{n} \right) \left(\frac{\boldsymbol{\varepsilon}' \mathbf{X}}{\sqrt{n}} \right) \mathbf{Q}^{-1} \left(\frac{\mathbf{X}' \boldsymbol{\varepsilon}}{\sqrt{n}} \right)$$

的相同。注意 q 等于 $\frac{1}{n}$ 乘以正态分布向量的二次型, 该向量 $\left(\frac{\mathbf{X}' \boldsymbol{\varepsilon}}{\sqrt{n}} \right)$ 渐近方差矩阵是 $\sigma^2 \mathbf{Q}$

因此, 利用沃尔德统计量极限分布证明的结果, 我们发现 q 可以写成

$$\frac{q}{\sigma^2} = \left(\frac{1}{n} \right) \mathbf{z}' \mathbf{z}, \quad \text{其中, } \mathbf{z} = \mathbf{Q}^{-1/2} \left(\frac{\mathbf{X}' \boldsymbol{\varepsilon}}{\sigma \sqrt{n}} \right), \text{ 所以, } \mathbf{z} \xrightarrow{d} N[0, \mathbf{I}]$$

这样

$$nq/\sigma^2 \xrightarrow{d} \chi^2[K]$$

$E(nq/\sigma^2) \rightarrow \text{常数}$, 即 $\mathbb{E}(q) \rightarrow \sigma^2 \times \text{常数} / n \rightarrow 0$

而且 $\text{Var}(nq/\sigma^2) \rightarrow \text{常数}$, 即 $\text{Var}(q) \rightarrow \frac{\sigma^4 \times \text{常数}}{n^2} \rightarrow 0$, q 是二阶收敛的, 所以保证了概率收敛, 即 $p \lim_{n \rightarrow \infty} q \rightarrow Eq \rightarrow 0$ 由此可得 q 本身依均方收敛于 0。这表明了 s^2 对 σ^2 的一致性。由此 b 的渐近协方差的适当的估计量是:

$$\text{Est.Asy. Var}[b] = \left(\frac{1}{n}\right) s^2 \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} = s^2 (\mathbf{X}'\mathbf{X})^{-1}$$

§ 8.3.1 B 的函数的渐近分布——得尔塔方法

利用泰勒展开, 把 $f(\mathbf{x})$ 线性化。

令 $f(\mathbf{b})$ 是一组关于最小二乘估计量 J 个连续的线性或非线性的函数并令

$$G = \frac{\partial f(\mathbf{b})}{\partial \mathbf{b}'}$$

G 是 $J \times K$ 矩阵, 其中第 j 行是第 j 个函数关于 \mathbf{b} 的导数。利用斯拉茨基 (Slutsky) 定理,

$$p \lim f(\mathbf{b}) = f(\beta)$$

并且

$$p \lim G = \frac{\partial f(\beta)}{\partial \beta'} = \Gamma$$

于是

$$f(\mathbf{b}) \xrightarrow{a} N \left[f(\beta), \Gamma \left(\frac{\sigma^2}{n} \mathbf{Q}^{-1} \right) \Gamma' \right] \quad (8.3.1)$$

实际上, 渐近协方差矩阵的估计量是:

$$\text{Est.Asy.Var} [f(\mathbf{b})] = G \left[s^2 (\mathbf{X}'\mathbf{X})^{-1} \right] G'$$

如果某个函数是非线性的, 则 \mathbf{b} 的无偏的性质不会传给 $f(\mathbf{b})$ 。不过从 8.3.1 中可得 $f(\mathbf{b})$ 是 $f(\beta)$ 的一致估计量, 而且渐近协方差矩阵很容易获得。

Example 8.1 (P324 (小 Green))

有限样本在条件 $V.\varepsilon \sim N[0, \sigma^2 I]$ 下的结果, 用于检验一组 J 个线性约束 $R\beta = q$ 的检验统计量:

$$\frac{(R\mathbf{b} - \mathbf{q})' \left[R (\mathbf{X}'\mathbf{X})^{-1} R' \right]^{-1} (R\mathbf{b} - \mathbf{q})/J}{\mathbf{e}'\mathbf{e}/(n-K)} = \frac{(R\mathbf{b} - \mathbf{q})' \left[R s^2 (\mathbf{X}'\mathbf{X})^{-1} R' \right]^{-1} (R\mathbf{b} - \mathbf{q})}{J}$$

表 8.1 小结：有限样本和大样本的结果比较

有限样本在条件 $V.\varepsilon \sim N[0, \sigma^2 \mathbf{I}]$ 下的结果	大样本在不满足条件 $V.\varepsilon \sim N[0, \sigma^2 \mathbf{I}]$ 下的结果
1. $\mathbb{E}[\mathbf{b}] = \boldsymbol{\beta}$ 最小二乘估计是无偏的	1. $p \lim \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot 0 = \boldsymbol{\beta}$
2. $\mathbb{E}[s^2] = \sigma^2$	2. s^2 是方差 σ^2 的一致估计量
3. Est. $\text{Var}[\mathbf{b}] = s^2 (\mathbf{X}'\mathbf{X})^{-1}$	3. Est. $\text{Var}[\mathbf{b}] = \left(\frac{1}{n}\right) s^2 \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} = s^2 (\mathbf{X}'\mathbf{X})^{-1}$
4. $N\left[\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right]$	4. $\mathbf{b} \xrightarrow{a} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}^{-1}\right]$
5. $t[n - K] = \frac{b_k - \beta_k}{s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}}$	5. $t_k = \frac{b_k - \beta_k}{\left[s^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}\right]^{1/2}}$ 服从标准正态分布，而不是 t 分布
6. $\frac{(\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}) / J}{\frac{\mathbf{e}'\mathbf{e} / (n - K)}{J}} =$	6. $W = (\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\mathbf{R} \left(s^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right)^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})\right] = JF$

大样本在不满足条件 $V.\varepsilon \sim N[0, \sigma^2 \mathbf{I}]$ 下的结果，

$$W = (\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\mathbf{R} \left(s^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right)^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})\right] = JF$$

依分布收敛于自由度为 J 的 χ^2 统计量。非线性问题的处理：(利用泰勒展开，转换为线性)

§ 9.1 线性化回归

回归模型的一般形式是：

$$y_i = h(x_i, \beta_i) + \varepsilon_i \quad (9.1.1)$$

很明显，线性模型只是一种特殊情况，我们应该讨论更一般的模型 9.1.1，例如：

$$\mathbf{y} = \beta_1 + \beta_2 e^{\beta_3 \mathbf{x}} + \varepsilon \quad (9.1.2)$$

不能变换到线性形式。

非线性回归模型是：

$$\mathbf{y} = h(\mathbf{x}, \beta) + \varepsilon$$

（为简化记号，我们去掉了观测值的下标）非线性回归模型的许多结果是基于在参数向量的一个特定值 β^0 处（如由经验得到的数据时）对 $h(\mathbf{x}, \beta)$ 的一个线性泰勒级数来近似：

$$h(\mathbf{x}, \beta) \cong h(\mathbf{x}, \beta^0) + \sum_k \left. \frac{\partial h(\mathbf{x}, \beta)}{\partial \beta_k} \right|_{\beta=\beta^0} (\beta_k - \beta_k^0) \quad (9.1.3)$$

这被称为线性化回归模型。整理各项可得：

$$h(\mathbf{x}, \beta) \cong h(\mathbf{x}, \beta^0) - \sum_k \beta_k^0 \left. \frac{\partial h(\mathbf{x}, \beta)}{\partial \beta_k} \right|_{\beta=\beta^0} + \sum_k \beta_k \left. \frac{\partial h(\mathbf{x}, \beta)}{\partial \beta_k} \right|_{\beta=\beta^0}$$

令 \tilde{x}_k^0 等于第 k 个偏微分, 对于 β^0 的一个给定值, 这 \tilde{x}_k^0 是数据而不是含未知参数的函数。于是:

$$\begin{aligned} h(\mathbf{x}, \beta) &\cong \left[h^0 - \sum_k \tilde{x}_k^0 \beta_k^0 \right] + \sum \tilde{x}_k^0 \beta_k \\ &= h^0 - \tilde{\mathbf{x}}^{0'} \beta^0 + \tilde{\mathbf{x}}^0 \beta \end{aligned}$$

或

$$\mathbf{y} \cong h^0 - \tilde{\mathbf{x}}^{0'} \beta^0 + \tilde{\mathbf{x}}^0 \beta + \varepsilon$$

把已知项移到方程左边, 可得回归模型:

$$\tilde{\mathbf{y}} = \mathbf{y} - h^0 + \tilde{\mathbf{x}}^{0'} \beta^0 = \tilde{\mathbf{x}}^0 \beta + \varepsilon \quad (9.1.4)$$

有了 β^0 值, 我们就可以计算 $\tilde{\mathbf{y}}^0$ 和 $\tilde{\mathbf{x}}^0$ 并通过线性最小二乘法估计 9.1.4 中的参数。

然后, 进行再次的迭代和回归, 直至收敛和满足我们的精度要求。

Example 9.1 对于 9.1.2 所给的非线性回归模型, 线性化方程中的回归量是:

$$\tilde{x}_1^0 = \frac{\partial h(\cdot)}{\partial \beta_1} = 1, \quad \tilde{x}_2^0 = \frac{\partial h(\cdot)}{\partial \beta_2} = e^{\beta_3^0 \mathbf{x}}, \quad \tilde{x}_3^0 = \frac{\partial h(\cdot)}{\partial \beta_3} = \beta_2^0 \mathbf{x} e^{\beta_3^0 \mathbf{x}}$$

有了一组参数 β^0

$$\tilde{\mathbf{y}}^0 = \mathbf{y} - h(\mathbf{x}, \beta_1^0, \beta_2^0, \beta_3^0) + \beta_1^0 \tilde{x}_1^0 + \beta_2^0 \tilde{x}_2^0 + \beta_3^0 \tilde{x}_3^0$$

可以对前面为估计 β^1, β^2 和 β^3 而定义三个变量进行回归。

§ 9.2 非线性最小二乘估计

最小二乘法仍然是一种比较具有吸引力的估计参数的方法。对于这种估计量已经得到许多分析结果, 例如, 一致性和渐近正态性。然而, 除了在扰动项是正态分布的情况下, 我们不能肯定非线性最小二乘估计量是最有效的估计量。(这和我们对于线性模型所得的结论是一样的) 下面的一些例子将说明这一点。

在继续之前, 有必要关于回归量做些假设。贾奇等人 (1985) 和雨宫 (1985) 曾详细地讨论过精确的要求。在古典回归模型中, 为了得到渐近结果, 我们假设样本矩矩阵 $(1/n) \mathbf{X}' \mathbf{X}$ 收敛于一个正定矩阵 \mathbf{Q} 。类似地, 当线性化模型中的“回归量”在真实参数值处被计算时, 我们在其上附加相同的条件。因此, 对于非线性回归模型, 与以前类似的是:

$$p \lim \left(\frac{1}{n} \right) \tilde{\mathbf{X}}' \tilde{\mathbf{X}} = p \lim \left(\frac{1}{n} \right) \sum_i \left[\frac{\partial h(x_i, \beta^0)}{\partial \beta^0} \right] \left[\frac{\partial h(x_i, \beta^0)}{\partial \beta^0} \right]' = \tilde{\mathbf{Q}} \quad (9.2.1)$$

其中 \tilde{Q} 是正定矩阵。根据这个公式, 非线性最小二乘估计量的渐近性质可以导出。实际上, 除了在这种情况下我们把 \tilde{X} 中的导数也作为回归量之外, 它们与我们已见到的线性模型的渐近性质十分相似。

(9.2.1) 中的矩阵收敛于正定矩阵的要求还附带回归量矩阵 \tilde{X} 的各列是线性无关的条件。这是一个识别条件, 类似于线性模型中的解释变量是线性无关的要求。

§ 9.2.1 非线性最小二乘准则函数

$$S(b) = \sum_i [y_i - h(x_i, b)]^2 = \sum_i e_i^2$$

其中我们已经代入即将是解的 b 。最小化的一阶必要条件是:

$$g(b) = -2 \sum_i [y_i - h(x_i, b)] \frac{\partial h(x_i, b)}{\partial b} = 0$$

注意:

$$g(b) = -2\tilde{X}'e$$

这与线性模型的情况相同。这是非线性最优化中的一个标准问题, 可以用许多方法来求解。高斯—牛顿方法在这种情况下经常使用。回想我们关于线性回归模型的讨论, 如果 β^0 的值是可以获得的, 那里所显示的线性回归模型可以用普通最小二乘法来估计。一旦回到一个参数向量, 它就可以作为一个新的 β^0 , 计算可以继续。迭代可以一直进行到相邻两个参数向量的差是足够小可以认为已经收敛为止。这个方法的主要优点之一是在最后一步迭代, \tilde{Q}^{-1} 的估计, 除了 $\hat{\sigma}^2$, 给出了参数估计渐近方差矩阵的正确估计。 σ^2 的一致估计可以利用残差来计算:

$$\hat{\sigma}^2 = \left(\frac{1}{n}\right) \sum_i [y_i - h(x_i, b)]^2 \quad (9.2.2)$$

(自由度校正 $1/(n - K)$) 在这里没有价值, 因为所有结果在任何情况下都是渐近的和 $\frac{n-K}{n} \rightarrow 1$ 的事实。

$$b \xrightarrow{a} N \left[\beta, \frac{\sigma^2}{n} Q^{-1} \right]$$

其中

$$Q = p \lim \left(\frac{\tilde{X}'\tilde{X}}{n} \right)$$

渐近协方差矩阵的样本估计是:

$$\text{Est. Asy. Var}[b] = \hat{\sigma}^2 \left(\tilde{X}'\tilde{X} \right)^{-1}$$

只要得到这些结果，推论和假设检验就可以按前几章中所描述的同样方式进行。在评价回归拟合值中产生一个小问题，因为

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

不再保证在零到 1 的范围内,(因为估计的误差 $\sum e_i^2$ 有可能足够的大)。然而，它仍给出了一个有用的描述性度量。

CHAPTER 10

非球形扰动项与广义最小二乘 (GLS)

§ 10.1 问题的提出

多元化回归模型扰动项违背古典假设的更一般的模型是广义回归模型，即假设

$$\mathbf{y} = \boldsymbol{\beta}'\mathbf{X} + \boldsymbol{\varepsilon}, \quad \mathbb{E}[\boldsymbol{\varepsilon}] = 0, \quad \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\boldsymbol{\Omega} \quad (10.1.1)$$

其中 $\boldsymbol{\Omega}$ 是一般的正定矩阵，而不是在古典假设的情况下的单位矩阵。古典假设条件情况只是这种模型的一个特例。

我们将考察的正定矩阵 $\boldsymbol{\Omega}$ 两种特殊的情况是**异方差性**和**自相关**。

§ 10.1.1 异方差性

当扰动项有不同的方差时，它们就是异方差的，异方差性经常产生于横截面数据，其中因变量的尺度 (scales) 和模型解释能力在不同的观察值之间倾向于变动。我们仍然假设不同观测值之间扰动无关。因此 $\sigma^2\boldsymbol{\Omega}$ 是：

$$\sigma^2\boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

§ 10.1.2 自相关

自相关经常出现在时间序列数据中，经济现象中的时间序列经常表现出一种“记忆”，因为变化在不同时期之间不是独立的。时间序列数据通常是同方差的，因此 $\sigma^2 \Omega$ 可能是

$$\sigma^2 \Omega = \begin{bmatrix} 1 & \sigma_1 & \cdots & \sigma_{n-1} \\ \sigma_1 & 1 & \cdots & \sigma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n-1} & \sigma_{n-2} & \cdots & 1 \end{bmatrix}$$

非对角线上的值依赖于扰动项的模式。

§ 10.1.3 普通最小二乘法的结果

具有球形干扰项:

$$\mathbb{E}[\varepsilon] = 0 \quad \mathbb{E}[\varepsilon \varepsilon'] = \sigma^2 I \quad (10.1.2)$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\varepsilon \quad (10.1.3)$$

是最佳线性无偏的、一致的和渐近正态分布的 (CAN=Consistent and asymptotically normally distributed)，并且如果干扰项服从正态分布，在所有 CAN 估计量中它是渐近有效的。现在我们考察哪些特性在 10.1.1 模型中仍然成立。

重申前面的内容，普通最小二乘估计量，

§ 10.1.4 有限样本特性

对 10.1.3 两边取期望，如果 $\mathbb{E}[\varepsilon | \mathbf{X}] = 0$ ，则

$$\mathbb{E}[\mathbf{b}] = \mathbb{E}_{\mathbf{X}} [\mathbb{E}[\mathbf{b} | \mathbf{X}]] = \beta \quad (10.1.4)$$

如果回归量和扰动项是无关的，则最小二乘法的无偏性不受 10.1.2 假设变化的影响。

最小二乘法估计量的样本方差是:

$$\begin{aligned} \text{Var}[\mathbf{b} - \beta] &= \mathbb{E}[(\mathbf{b} - \beta)(\mathbf{b} - \beta)'] \\ &= \mathbb{E} \left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\varepsilon \varepsilon' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \right] \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (\sigma^2 \Omega) \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{\sigma^2}{n} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}\Omega\mathbf{X}}{n} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \end{aligned} \quad (10.1.5)$$

在 (10.1.3) 中, \mathbf{b} 是 ε 的线性函数。因此, 如果 ε 服从正态分布, 则

$$\mathbf{b} \sim N \left[\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}\boldsymbol{\Omega}\mathbf{X}) (\mathbf{X}'\mathbf{X})^{-1} \right]$$

由于最小二乘估计量的方差不再是 $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$, 任何基于 $s^2 (\mathbf{X}'\mathbf{X})^{-1}$ 的推断都可能导致错误。不仅使用的矩阵是错误的, 而且 s^2 也可能是 σ^2 的有偏估计量。通常无法知道 $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ 是比 \mathbf{b} 的真正方差大还是小, 因此即使有 σ^2 的一个好的估计, $\text{Var}[\mathbf{b}]$ 的传统估计量也不会有用。

§ 10.1.5 最小二乘法的渐近特性

如果 $\text{Var}[\mathbf{b}]$ 收敛于 0, 则 \mathbf{b} 是一致的。使用表现良好的回归量, $(\mathbf{X}'\mathbf{X}/n)^{-1}$ 将收敛到一个常数矩阵 (可能是 0), 并且最前面的乘子 σ^2/n 将收敛于 0。但 $\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n$ 不一定收敛, 如果它收敛, 则从 (10.1.5) 式可推断普通最小二乘是一致的和无偏的。

Theorem 10.1.1 如果 $p \lim (\mathbf{X}'\mathbf{X}/n)$ 和 $p \lim (\mathbf{X}\boldsymbol{\Omega}\mathbf{X}/n)$ 都是有限正定矩阵, 则 \mathbf{b} 是 $\boldsymbol{\beta}$ 的一致估计量。

上述结论成立的条件依赖于 \mathbf{X} 和 $\boldsymbol{\Omega}$ 。

另一种分离这两个组成部分的处理办法是:

1. $\mathbf{X}'\mathbf{X}$ 最小的特征根当 $n \rightarrow \infty$ 时无限制地增加, 这意味着 $p \lim (\mathbf{X}'\mathbf{X})^{-1} = 0$

2. $\boldsymbol{\Omega}$ 最大的特征根对于所有 n 都是有限的。对于异方差模型, 方差就是特征根。因此, 要求它们是有限的。对于有自相关的模型, 这要求 $\boldsymbol{\Omega}$ 的元素有限并且非对角线元素与对角线元素相比不是特别大。

那么, 普通最小二乘法在广义回归模型中是一致的。

说明普通最小二乘法是不一致的模型

假定回归模型是 $\mathbf{y} = \boldsymbol{\mu} + \varepsilon$, 其中 ε 的均值为 0, 方差为常数并且在不同观测值之间具有相同的相关系数 ρ 。于是

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ & & & \ddots & \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

矩阵 \mathbf{X} 是一列 1。 $\boldsymbol{\mu}$ 的普通最小二乘估计量是 \bar{y} 。把 $\boldsymbol{\Omega}$ 代入 (10.1.5), 得

$$\text{Var}[\bar{y}] = \frac{\sigma^2}{n} (1 - \rho + n\rho) \quad (10.1.6)$$

这个表达式的极限是 $\rho\sigma^2$ 而不是 0。尽管 OLS 是无偏的, 但它不是一致的。对于这个模型, $\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}/n = 1 + \rho(n-1)/n$ 不收敛。由于 \mathbf{X} 是一列 1, 因此 $\mathbf{X}\mathbf{X}' = n$ 是一个标量, 满足条件 1; 但是, $\boldsymbol{\Omega}$ 的特征根是 $1-\rho$

(重数是 $n-1$) 和 $(1-\rho+n\rho)$, 不满足条件 2; 这个例子中模型的困难是不同观测值间有太多的相关。在时间序列情况下, 我们一般要求观测值之间关于时间的相关系数随它们之间距离增加而减小。这里条件没有被满足。关于在简介中曾讨论的自相关扰动项的协方差矩阵上需要附加什么种类的要求, 这给出一些很有意义的信息。

如果

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{X}'\boldsymbol{\varepsilon} \quad (10.1.7)$$

的极限分布是正态的, 则 OLS 估计量渐近地服从正态分布。如果 $p \lim(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, 那么右边项的极限分布与

$$v = \mathbf{Q}^{-1} \frac{1}{\sqrt{n}} \mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1} \frac{1}{\sqrt{n}} \sum_i x_i \varepsilon_i \quad (10.1.8)$$

的分布相同, 其中 x'_i 是 \mathbf{X} 的一行 (当然假定极限分布确实存在)。现在, 问题是中心极限定理是否可以直接应用于 v 。如果扰动项只是异方差的而且仍是无关的, 答案通常是肯定的。在这种情况下, 很容易看到只要 \mathbf{X} 表现良好, 而且 Ω 对角元素是有限的, 最小二乘估计量是渐近正态分布的, 方差矩阵由 10.1.5 给出。

但对于大多数一般的情况, 答案是否定的, 因为 (10.1.8) 中的和不一定是相互独立或是甚至无关的随机变量的和。不过, 雨宫 (1985) 和安德森 (1971) 曾指出, 自相关扰动项的模型中 \mathbf{b} 的渐近正态性是足够普遍的, 以致于包括了我们在实际中可能遇到的大多数情况。我们可以得到结论, 除了在特别不利的情况下,

\mathbf{b} 渐近地服从均值为 $\boldsymbol{\beta}$, 方差矩阵由 (10.1.5) 给出的正态分布。

总之, OLS 在这个模型中只保留了它的一些可取性质, 它是无偏的、一致的和渐近正态分布的。不过, 它不是有效。我们需要寻求 \mathbf{b} 的有效估计。

§ 10.2 广义最小二乘 (GLS)

在广义回归模型中, $\boldsymbol{\beta}$ 的有效估计需要关于 Ω 的知识。我们只考察 Ω 是已知的、对称正定矩阵的情况, 这种情况偶尔会发生, 但在大多数的模型中 Ω 包含必须估计的未知参数。

由于 $\boldsymbol{\beta}$ 是正定对称矩阵, 它可以分解为:

$$\Omega = \mathbf{C} \boldsymbol{\Lambda} \mathbf{C}' \quad (10.2.1)$$

其中 \mathbf{C} 的各列是 Ω 的特征向量经过正交化而得到, 即 $\mathbf{C}\mathbf{C}' = \mathbf{I}$, 而且 Ω 的特征根被放在对角矩阵 $\boldsymbol{\Lambda}$ 中。令 $\boldsymbol{\Lambda}^{\frac{1}{2}}$ 是对角元素为 $\sqrt{\lambda_i}$ 的对角矩阵。

如果令 $\mathbf{P} = \mathbf{C}\boldsymbol{\Lambda}^{\frac{1}{2}}$, 则

$$\Omega^{-1} = \mathbf{P}\mathbf{P}'$$

用 P' 前乘 10.1.1 中的模型可得

$$P'y = P'X\beta + P'\varepsilon \quad \text{or} \quad y_* = X_*\beta + \varepsilon_* \quad (10.2.2)$$

ε_* 的方差是:

$$\mathbb{E}[\varepsilon_*\varepsilon_*'] = P'\sigma^2\Omega P\sigma^2I$$

因此, 这个变换后的模型就是一个我们熟悉的古典回归模型。由于 Ω 已知, 所以, y_* 和 X_* 是可观测量。在古典回归模型中, OLS 是有效的。

$$\therefore \beta = (X_*'X_*)^{-1}X_*'y_* = (X'PP'X)^{-1}X'PP'y = (X\Omega^{-1}X)^{-1}X\Omega^{-1}y \quad (10.2.3)$$

是 β 的有效估计量。这是 β 的广义最小二乘 (GLS) 估计量。按照古典回归模型, 我们有以下结论:

如果 $\mathbb{E}[\varepsilon_*|X_*] = 0$, GLS 估计量 β 是无偏的。这等价于 $\mathbb{E}[P\varepsilon|PX] = 0$, 但由于 P 是已知常数的矩阵, 即要求 $\mathbb{E}[\varepsilon|X] = 0$, 也即要求回归量与扰动项是无关的, 是我们模型的基本假设。

§ 10.2.1 GLS 大样本的特性

如果

$$p\lim\left(\frac{X_*'X_*}{n}\right) = Q_* \quad (10.2.4)$$

GLS 估计量是一致的, 其中 Q_* 是有限正定矩阵。进行替换可得

$$p\lim\left(\frac{X\Omega^{-1}X}{n}\right)^{-1} = Q_*^{-1} \quad (10.2.5)$$

我们需要的是变换后的数据 $X_* = P'X$ 而不是原始数据 X 的数据。

根据 (10.2.5) 的假设, GLS 估计量是渐近正态分布的, 均值为 β , 样本方差为:

$$\text{Var}[\hat{\beta}] = \sigma^2 (X_*'X_*)^{-1} = \sigma^2 (X\Omega^{-1}X)^{-1} \quad (10.2.6)$$

通过对 (10.2.2) 中的模型应用高斯—马尔科夫定理可得如下的艾特肯 (1935) 定理: GLS 估计量 $\hat{\beta}$ 是广义回归模型中的最小方差线性无偏估计量。(估计量 $\hat{\beta}$ 的无偏性很容易从式 (10.2.3) 中得到)。

$\hat{\beta}$ 有时被称为艾特肯估计量。这是一个一般性结果, 当 $\Omega = I$ 时高斯—马尔科夫定理是它的一个特例。

对于假设检验, 我们可以把所有结果应用到变换后的模型 (10.2.2) 中。为了检验 J 个线性约束 $R\beta = q$, 相应的统计量是

$$\begin{aligned} F[J, n-K] &= \frac{(R\hat{\beta} - q)' \left[R \left(\hat{\sigma}^2 (X_*'X_*)^{-1} R' \right)^{-1} (R\hat{\beta} - q) \right]}{J} \\ &= \frac{(\hat{\varepsilon}_c' \hat{\varepsilon}_c - \hat{\varepsilon}' \hat{\varepsilon}) / J}{\hat{\sigma}^2} \end{aligned}$$

其中残差向量是: $\hat{\varepsilon} = \mathbf{y}_* - \mathbf{X}_* \hat{\beta}$

而

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}' \hat{\varepsilon}}{n - K} = \frac{(\mathbf{y} - \mathbf{X} \hat{\beta})' \Omega^{-1} (\mathbf{y} - \mathbf{X} \hat{\beta})}{n - K}$$

有约束的 GLS 残差 $\hat{\varepsilon}_c = \mathbf{y}_* - \mathbf{X}_* \beta_c$, 基于

$$\hat{\beta}_c = \hat{\beta} - [\mathbf{X} \Omega^{-1} \mathbf{X}]^{-1} \mathbf{R}' [\mathbf{R} (\mathbf{X} \Omega^{-1} \mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R} \hat{\beta} - \mathbf{q}) \quad (\text{作业3}) \quad (10.2.7)$$

总之, 对于古典模型的所有结果, 包括通常的推断过程, 都适用于 (10.2.2) 中的模型。应该注意的是: 在广义回归模型中没有 R^2 的准确对等物。不同的统计量有不同的意义, 但使用它们时一定要谨慎。类似 R^2 的任何形式的度量纯粹是描述性的, 不象古典回归模型中一样具有确切的意义。

§ 10.3 最大似然估计和似然比统计量

§ 10.3.1 二元正态分布

直观上, 二元正态分布是两个正态随机变量的联合分布。如果两个随机变量 X_1 和 X_2 的联合密度函数为:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{\Sigma^{-1}}{2}\right\}$$

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right]$$

这里 $-\infty < x_1, x_2 < \infty$, $\sigma_1 > 0$, $\sigma_2 > 0$, $-1 < \rho < 1$ 。

我们称 X_1 和 X_2 服从二元正态分布。通过计算可得 X_1 和 X_2 的边缘分布分别为 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 。上式中的参数 ρ 是 X_1 和 X_2 的相关系数。

有时如果把联合概率密度函数写成矩阵的形式, 则从形式上来看就简单多了。记 $\mathbf{X}' = (X_1, X_2)$, 那么二元正态概率密度函数可以写成如下的简单形式:

$$f(\mathbf{x}) = (2\pi)^{-1} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

这里

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}$$

§ 10.3.2 多元正态分布

$g(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \Sigma^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$, $\mathbf{x} \in \mathbb{R}^n$ 这就是均值为 $\boldsymbol{\mu}$ 协方差矩阵为 Σ 的多元正态分布, 记为 $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$

多元正态分布的指数分布是卡方分布。

§ 10.3.3 最大似然估计

如果扰动项服从多元正态分布，则样本的对数似然函数是

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\sigma^2 \Omega| - \frac{1}{2} \epsilon' (\sigma^2 \Omega)^{-1} \epsilon$$

进行变量变换 $y - X\beta = \epsilon$ 可得

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)' \Omega^{-1} (y - X\beta) - \frac{1}{2} \ln |\Omega| \quad (10.3.1)$$

由于 Ω 是已知常数的矩阵， β 的最大似然估计量是使得广义平方和

$$S_*(\beta) = (y - X\beta)' \Omega^{-1} (y - X\beta) \quad (10.3.2)$$

最小化的向量。（因此取名广义最小二乘）使得 L 最大化的必要条件是：

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} X \Omega^{-1} (y - X\beta) = \frac{1}{\sigma^2} X_*' (y_* - X_* \beta) = 0 \quad (10.3.3)$$

和

$$\begin{aligned} \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)' \Omega^{-1} (y - X\beta) \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (y_* - X_* \beta)' (y_* - X_* \beta) = 0 \end{aligned} \quad (10.3.4)$$

利用变换后的数据，解就是 MLE 估计量：

$$\hat{\beta}_{MLE} = (X_*' X_*)^{-1} X_*' y_* = (X \Omega^{-1} X)^{-1} X \Omega^{-1} y \quad (10.3.5)$$

$$\begin{aligned} \hat{\sigma}_{MLE}^2 &= \frac{(y_* - X_* \hat{\beta})' (y_* - X_* \hat{\beta})}{n} \\ &= \frac{(y - X \hat{\beta})' \Omega^{-1} (y - X \hat{\beta})}{n} \end{aligned} \quad (10.3.6)$$

这意味着在扰动项服从正态分布的情况下，GLS 也是 MLE。正如在古典回归模型中， σ^2 的 MLE 估计量也是有偏的。

一个无偏估计量是：

$$s^2 = \frac{(y - X \hat{\beta})' \Omega^{-1} (y - X \hat{\beta})}{n - K}$$

为了得到 MLE 估计量的渐近分布，我们可以把以前的结果应用于变换后的模型中。因此， $[\hat{\beta}_{MLE}, \hat{\sigma}_{MLE}^2]$ 的渐近分布是：

$$\begin{bmatrix} \hat{\beta}_{MLE} \\ \hat{\sigma}_{MLE}^2 \end{bmatrix} \xrightarrow{a} N \left[\begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \sigma^2 (X \Omega^{-1} X)^{-1} & 0 \\ 0' & 2\sigma^4/n \end{bmatrix} \right]$$

§ 10.3.4 最大似然估计量的渐进特性

由于其大样本特性或渐近特性, 最大似然估计量最具吸引力, 若我们假定密度函数 $f(\mathbf{x}, \boldsymbol{\theta})$ 满足一定的条件, 则最大似然估计量将有如下渐近特性 (不加证明):

1. 它是一致的:

$$p \lim \hat{\boldsymbol{\theta}}_{MLE} = \boldsymbol{\theta}$$

2. 它是渐近正态的:

$$\hat{\boldsymbol{\theta}} \xrightarrow{a} N[\boldsymbol{\theta}, \{\mathbf{I}(\boldsymbol{\theta})\}^{-1}]$$

3. 它是渐近有效的, 且达到一致估计量的克拉美—劳下界:

$$\begin{aligned} \text{Asy. Var} [\hat{\boldsymbol{\theta}}_{MLE}] &= \left\{ -\mathbb{E} \left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \right\}^{-1} \\ &= \left\{ \mathbb{E} \left[\left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right) \right] \right\}^{-1} \end{aligned}$$

这三个特性解释了经济计量学中最大似然技术盛行的原因; 第二个是大大地促进了假设检验和区间估计的构造; 第三个是一个特别强有力的结果, MLE 具有一个一致估计量所能达到的最少方差。

不过, 注意到它们都是渐近特性是很重要的, MLE 在小样本情况下并不是最佳估计量。

最后, 最大似然估计量有一个非常有用的特征。

Theorem 10.3.1 (不变性) 若 $\boldsymbol{\theta}_{MLE}$ 是 $\boldsymbol{\theta}$ 的最大似然估计且 $g(\boldsymbol{\theta})$ 是一个连续函数, 则 $g(\boldsymbol{\theta})$ 的最大似然估计是 $g(\boldsymbol{\theta}_{MLE})$

不变性特征有两个有用的实际含义。**首先**, 若对一组参数已经得到了估计, 且要求它们的一个函数的估计, 则不需要重新估计模型。例如, 为了估计一正态分布的 σ , 我们可以简单地利用 $\hat{\sigma}^2$ 的平方根; **第二**, 不变性原理暗示我们可以按我们喜欢的方式自由地重参数化 (reparameterize) 一个似然函数, 这将简化估计。

克拉美—劳 (Cramer-Rao) 下界。假定 x 的密度函数满足一定的条件, 参数 $\boldsymbol{\theta}$ 的一个无偏估计量的方差将大于等于:

$$\begin{aligned} [\mathbf{I}(\boldsymbol{\theta})]^{-1} &= \left(-\mathbb{E} \left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right] \right)^{-1} \\ &= \left(-\mathbb{E} \left[\left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^2 \right] \right)^{-1} \end{aligned}$$

量 $\mathbf{I}(\boldsymbol{\theta})$ 是样本的信息数。

考虑一个多变量情形。若 θ 是一个参数向量, $I(\theta)$ 是信息矩阵。克拉美—劳定理指出, 任何无偏估计量的方差矩阵与信息矩阵的逆

$$\begin{aligned} [I(\theta)]^{-1} &= \left\{ -\mathbb{E} \left[\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] \right\}^{-1} \\ &= \left\{ \mathbb{E} \left[\left(\frac{\partial \ln L(\theta)}{\partial \theta} \right) \left(\frac{\partial \ln L(\theta)}{\partial \theta'} \right) \right] \right\}^{-1} \end{aligned}$$

的差将是一个非负定矩阵。

最大似然估计量的最重要的性质是有效性和不变性。

§ 10.3.5 似然比统计量

广义似然比检验

设 $\mathbf{X} = (X_1, \dots, X_n)$ 是从母体中抽取的子样, 其可能分布族 $\{f(\mathbf{x}; \theta), \theta \in \Theta\}$, 其中 θ (可以是向量) 是未知参数 (当母体是连续型变量时 f 表示分布密度, 当母体是离散型变量时 f 表示概率分布)。要求检验假设 $H: \theta = \theta_0$ 。这里应指出, θ_0 有时是表示一个集合, 如在运用 t-检验法检验假设 $H_0: \mu = \mu_0$ 时, 那里

$$\theta \triangleq (\mu, \sigma^2)$$

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

$$\theta_0 \triangleq \{(\mu_0, \sigma^2) : \sigma^2 > 0\}$$

它是一个未知参数的集合而不是一个单点。

现在我们引进一个统计量:

$$\lambda(\mathbf{x}) \triangleq \frac{\sup_{\theta \in \theta_0} \prod_{i=1}^n f(x_i; \theta)}{\sup_{\theta \in \Theta} \prod_{i=1}^n f(x_i; \theta)}$$

习惯上称 $\lambda(\mathbf{x})$ 为广义似然比, 显然它是子样的函数, 不依赖于未知参数 θ 。由于 $\theta_0 \subset \Theta$, 所以

$$0 \leq \lambda(\mathbf{x}) \leq 1$$

类似于最大似然原理, 如果 $\lambda(\mathbf{x})$ 取值较小, 这说明当 H_0 为真时观察到样点 \mathbf{x} 的概率比 H_0 不真时观察到样点的概率要小得多, 此时我们有理由怀疑假设 H_0 不真。所以从广义似然比出发, 该检验问题是当下式成立时拒绝 H_0 ,

$$\lambda(\mathbf{x}) \leq \lambda_0 \tag{10.3.7}$$

其中 λ_0 的选取是使得下式成立,

$$P_{\theta} \{\lambda(\mathbf{X}) \leq \lambda_0\} \leq \alpha, \text{ 对一切 } \theta \in \Theta \tag{10.3.8}$$

给出的检验法称为水平为 α 的广义似然比检验。当 θ_0 是一个单点时可写为:

$$P_{\theta_0} \{ \lambda(\mathbf{X}) \leq \lambda_0 \} \leq \alpha$$

Example 10.1 设母体 X 服从正态 $N(\mu, \sigma^2)$ 分布, 其中 μ 和 σ^2 均是未知参数, 欲检验假设 $H_0: \mu = \mu_0$

设 $\mathbf{X} = (X_1, \dots, X_n)$ 是从该母体中抽得的一个简单随机子样, 它的联合分布密度是

$$f(\mathbf{x}; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

在这里, 参数空间是 $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$

假设 H_0 所对应的子集是 $\theta_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}$, 运用微分法对 $f(\mathbf{x}; \mu, \sigma^2)$ 求最大值得到

$$\begin{aligned} \sup_{(\mu, \sigma^2) \in \theta_0} f(\mathbf{x}; \mu, \sigma^2) &= \left[2\pi \frac{1}{n} \sum (x_i - \mu_0)^2 \right]^{-n/2} e^{-n/2} \\ \sup_{(\mu, \sigma^2) \in \Theta} f(\mathbf{x}; \mu, \sigma^2) &= \left[2\pi \frac{1}{n} \sum (x_i - \bar{x})^2 \right]^{-n/2} e^{-n/2} \end{aligned}$$

所以广义似然比是:

$$\begin{aligned} \lambda(x) &= \frac{\sup_{(\mu, \sigma^2) \in \theta_0} f(\mathbf{x}; \mu, \sigma^2)}{\sup_{(\mu, \sigma^2) \in \Theta} f(\mathbf{x}; \mu, \sigma^2)} = \left\{ \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \mu_0)^2} \right\}^{n/2} \\ &= \left\{ \frac{1}{1 + n(x - \mu_0)^2 / \sum (x_i - \bar{x})^2} \right\}^{n/2} = \left\{ \frac{1}{1 + t^2/n - 1} \right\}^{n/2} \end{aligned} \quad (10.3.9)$$

其中:

$$t(x) = \frac{\sqrt{n(n-1)}(x - \mu_0)}{\sqrt{\sum (x_i - \bar{x})^2}} \quad (10.3.10)$$

所以, 广义似然比检验的是当下式成立时拒绝 H_0 ,

$$\lambda(\mathbf{x}) \leq \lambda_0 \quad (10.3.11)$$

由于 $\lambda(x)$ 是关于 $t(x)$ 的单调下降函数, 所以 10.3.11 又可等价地表示为

$$|t(\mathbf{x})| > c_0 \quad (10.3.12)$$

因为 $t(\mathbf{X})$ 当 H_0 成立时服从 $t(n-1)$ 分布, 所以 c_0 由下式决定 $\int_{c_0}^{\infty} t(\mathbf{x}; n-1) d\mathbf{x} = \alpha/2$

它就是前面我们所指出的 t -检验法

§ 10.3.6 线性约束检验的另一个统计量

为了检验 J 个线性约束 $R\beta = q$ ，相应的统计量是：

$$F[J, n-K] = \frac{(R\hat{\beta} - q)' \left[R \left(\hat{\sigma}^2 (X_*' X_*)^{-1} R' \right)^{-1} (R\hat{\beta} - q) \right]}{J} \\ = \frac{(\hat{\varepsilon}'_c \hat{\varepsilon}_c - \hat{\varepsilon}' \hat{\varepsilon}) / J}{\hat{\sigma}^2}$$

其中残差向量是 $\hat{\varepsilon} = y_* - X_* \hat{\beta}$

而

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}' \hat{\varepsilon}}{n-K} = \frac{(y - X\hat{\beta})' \Omega^{-1} (y - X\hat{\beta})}{n-K}$$

不同于这个检验的另一种用于关于 β 的假设检验的统计量是似然比统计量，同样可以检验带有 J 个线性约束 $R\beta = q$ 。

如果我们把上面的 (16) 和 (17) 代入 (12)，可以发现，在 LM 估计量处，

$$\ln \hat{L} = -\frac{n}{2} (1 + \ln(2\pi) + \ln \hat{\sigma}^2) - \frac{1}{2} \ln |\Omega|$$

因此，似然比检验基于

可以证明，此统计量渐近服从自由度为 J 的 χ^2 分布，其中 J 是约束条件的个数。

§ 10.4 可行的最小二乘估计 (FGLS)

上一节的结果是基于 Ω 必须是已知的条件基础上的。如果 Ω 含有必须估计的未知参数，则 GLS 是不可行的。但在无约束的情况下， $\sigma^2 \Omega$ 中有 $n(n+1)/2$ 个附加参数。这对于用 n 个观测值来估计这么多的参数是不现实的。只有当模型中需要估计的参数较少时，即模型中 Ω 某种结构要简化，才可以找到求解的方法

§ 10.4.1 Ω 知的条件下

具有代表性的问题涉及到一小组参数 θ ，满足 $\Omega = \Omega(\theta)$ ，例如， Ω 只有一个未知数 ρ ，其常见的表达式是：

$$\Omega = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2} \\ & & & & \ddots & \\ \rho^{n-1} & \rho^{n-2} & \cdots & & & 1 \end{bmatrix}$$

其中，也只有一个附加的未知参数 ρ 。

另外, 只包含一个新参数的异方差模型是:

$$\sigma_i^2 = \sigma^2 z_i^\alpha$$

接下来, 假定 $\hat{\theta}$ 是 θ 的一致估计量 (如果我们知道如何求得这样的估计量) 为了使 GLS 估计可行, 我们将使用

替代真正的 Ω 。我们所考虑的问题是利用 $\Omega(\theta)$ 是否要求我们改变上节的某些结果。

如果利用 $p \lim \hat{\theta} = \theta$, $\hat{\Omega}$ 似乎渐近等价于利用真正的 Ω (有时会不成立, 施密特 (1976) 构造了一个反例)。

令可行广义最小二乘 (或 FGLS) 估计量记为:

$$\hat{\beta} = (X\hat{\Omega}^{-1}X')^{-1} X\hat{\Omega}^{-1}y$$

那么, $\hat{\beta}$ 渐近等价于 β 的条件是:

$$p \lim \frac{X'\hat{\Omega}^{-1}X}{n} = p \lim \frac{X'\Omega^{-1}X}{n} \quad (10.4.1)$$

和

$$p \lim \frac{1}{\sqrt{n}} X'\hat{\Omega}^{-1}\varepsilon = p \lim \frac{1}{\sqrt{n}} X'\Omega^{-1}\varepsilon \quad (10.4.2)$$

如果 10.2.2 中变换后的回归量表现良好, 则 10.4.2 右边服从极限正态分布。这正是我们求最小二乘估计量的渐近分布时所利用的条件。因此, 当 $\hat{\Omega}$ 替 Ω 时 10.4.2 要求同样的条件成立。

这些是必须逐个情况进行核实的条件, 在大多数情况下, 它们的确成立。如果我们假设它们成立, 基于 $\hat{\theta}$ 的 FGLS 估计量与 GLS 估计量具有同样的渐近性质。这是一个相当有用的结果。特别地, 注意以下结论:

1. 一个渐近有效的 FGLS 估计量不要求我们有 θ 的有效估计量, 只需要一个一致估计量。

2. 除了最简单的情况, FGLS 估计量的有限样本性质和精确分布是未知的。FGLS 估计量的渐近有效性在小样本的情况下可能不再成立, 这是因为由估计的 Ω 引入的易变性。

对于异方差情况的一些分析由泰勒 (1977 年) 给出。自相关的模型由格涅里切斯和拉奥 (1969 年) 做了分析。在这两项研究中, 他们发现对于许多类型的参数, FGLS 比最小二乘更为有效。但是, 如果偏离古典假设不太严重, 在小样本情况下最小二乘可能比 FGLS 更有效。

§ 10.5 广义矩法 (GMM) 估计量

§ 10.5.1 广义矩法 (GMM) 估计量

在经济计量学中，尤其是在宏观经济学和金融学中，最近大部分的实证性工作都利用了所谓的 GMM 估计量。正如我们将看到的，这是一大类估计量，实际上包括我们所学过的所有估计量。这个技巧的近期推广大大丰富了应用型经济计量学家可使用的工具。

GMM 估计方法是一般矩方法技巧的直接推广。

1. OLS 估计量是一个矩法估计量

现在，考虑古典线性回归模型 (CLRM) 中参数的 OLS 估计量的明显不同情况。模型假设之一是：

$$\mathbb{E}[x_i \varepsilon_i] = 0$$

样本类似物为：

$$\frac{1}{n} \sum_i x_i e_i = \frac{1}{n} \sum_i x_i (y_i - x_i \hat{\beta}_i) = 0$$

β 的估计量满足这些矩方程。当然，这便是最小二乘估计量的正则方程组，因此，我们看到 OLS 估计量是一个矩法估计量。

2. GLS 估计量是一个矩法估计量

GLS 估计量根据

$$\mathbb{E}[\mathbf{X}'\Omega^{-1}\varepsilon] = 0$$

来定义，我们可以把上式重写成：

$$\mathbb{E}[\mathbf{X}'_*\varepsilon] = 0 \quad \text{or} \quad \mathbb{E}[\mathbf{x}'_*\varepsilon] = 0$$

3. MLE 估计量是一个矩法估计量

所有迄今为止我们曾考察过的以及后面将遇到的 MLE 都是通过让对数似然的导数等于 0 得到的。操作原理是

$$\ln L = \sum_i \ln f(y_i, x_i | \theta)$$

其中 $f(\cdot)$ 是密度函数， θ 是参数向量。构成 MLE 理论的基础是下列结果，即对于正常问题，

$$\mathbb{E}\left[\frac{\partial \ln f(y, x | \theta)}{\partial \theta}\right] = 0$$

MLE 是通过令样本类似物等于 0 得到:

$$\frac{1}{n} \frac{\partial \ln L}{\partial \theta} = \frac{1}{n} \sum_i \left[\frac{\partial \ln f(y_i, x_i | \theta)}{\partial \theta} \right] = 0$$

(除以 n 是为了使这个结果与以前的结果具有可比性, 并不改变解) 结果是, 几乎所有我们曾讨论过的和后面将要遇到的估计量都可以解释为矩法估计量 MME。

以上 OLS、GLS 和 MLE 估计量的三个例子都有一个共同方面, 在所列的每种情况中, 矩方程组的数目与要估计的参数的个数完全一样。因此, 每种情况都是恰好可识别的。矩方程组将有惟一解, 并在这个解处, 方程组将恰好被满足。在有共线性的回归模型中, 有 K 个参数但独立的矩方程个数少于 K ; 另外, 在有些情况中矩方程组超过参数个数 (如实例 3 的情况), 因此系统是过度决定的。

§ 10.5.2 一致估计——矩方法

用于许多这种情形中的技术是矩方法, 这个方法的基础如下:

Theorem 10.5.1 (最小二乘定理) 从具有有限均值 μ 和有限方差 σ^2 的任何总体中抽取的随机样本的均值都是 μ 的一个一致估计量。

Proof 10.1 $\mathbb{E}[\bar{x}] = \mu$ 及 $\text{Var}[\bar{x}] = \sigma^2/n$ 所以, \bar{x} 依均方收敛于 μ , 或 $p \lim \bar{x} = \mu$ 。

Theorem 10.5.2 (斯拉茨基定理 (Slutsky)) 对一个连续函数 $g(x_n)$, $p \lim g(x_n) = g(p \lim x_n)$

§ 10.5.3 随机抽样和分布的参数估计

考虑来自具有有限矩 $\mathbb{E}[x^k]$ 的分布随机抽样, 样本包含 n 个观测值 x_1, \dots, x_K 。第 k 阶“原始”或未中心化矩是:

$$m'_k = \frac{1}{n} \sum_i x_i^k$$

通过代换 $z_i = x_i^k$ 我们得到如下结果:

$$\mathbb{E}[m'_k] = \mu'_k = \mathbb{E}[x_i^k]$$

和

$$\text{Var}[m'_k] = \frac{1}{n} \text{Var}[x_i^k] = \frac{1}{n} (\mu'_{2k} - \mu_k'^2)$$

由最小二乘定理

$$p \lim m'_k = \mu'_k = \mathbb{E}[x_i^k]$$

最后, 由中心极限定理, 有:

$$\sqrt{n}(m'_k - \mu'_k) \xrightarrow{d} N[0, \mu'_{2k} - \mu'^2_k]$$

按惯例,

$$\mu'_1 = \mathbb{E}[x_i] = \mu$$

一般地, μ'_k 将是母体参数的一个函数, 通过计算 K 个原始矩并使它们等于这些函数, 我们得到 K 个方程, 解这些方程可以提供 K 个未知参数的估计。

在随机抽样中, 样本统计量将依概率收敛于某个常数。例如, $(1/n)\Sigma_i x_i^2$ 将依均方收敛到 x_i 的分布的方差加均值的平方。这个常数又是分布的未知参数的一个函数。为了估计 K 个参数 $\theta_1, \dots, \theta_K$, 我们计算 K 个统计量 m_1, \dots, m_K 它们的概率极限都是参数的已知函数, 这个 K 个矩等于 K 个函数, 且这些函数反过来又把参数表达为矩的函数。估计量由于斯拉茨基定理和最小二乘定理而都是一致的

虽然基于 x 的幂的各个矩提供了关于参数的信息的一个自然的来源, 数据的其他函数也可能是有用的, 令 $g_k(\cdot)$ 是不涉及样本容量 n 的任何连续函数, 并令:

$$\bar{g}_k = \frac{1}{n} \sum_i g_k(x_i) \quad k = 1, 2, \dots, K$$

这些也是数据的“矩”, 由最小二乘定理可推得:

$$p \lim \bar{g}_k = \mathbb{E}[g_k(x)] = \gamma_k(\theta_1, \dots, \theta_K)$$

我们假定 $\gamma_k(\cdot)$ 包含分布的一些或全部参数, 由于要估计 K 个参数, K 个矩方程:

$$\begin{aligned} \bar{g}_1 - \gamma_1(\theta_1, \dots, \theta_K) &= 0 \\ &\vdots \\ \bar{g}_K - \gamma_K(\theta_1, \dots, \theta_K) &= 0 \end{aligned}$$

提供 K 个未知数 $\theta_1, \dots, \theta_K$ 的 K 个方程, 则通过求解方程组获得矩法估计量:

$$\hat{\theta}_k = \hat{\theta}_k[\bar{g}_1, \dots, \bar{g}_K]$$

Example 10.2 (实例 2) 在来自伽玛分布的抽样中, 有:

$$f(x) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda x} x^{P-1}, x > 0, \quad P > 0, \quad \lambda > 0$$

$$\begin{aligned}
p \lim \frac{1}{n} \sum_i x_i &= \frac{P}{\lambda} \\
p \lim \frac{1}{n} \sum_i x_i^2 &= \frac{P(P+1)}{\lambda^2} \\
p \lim \sum_i \ln x_i &= \frac{d \ln \Gamma(P)}{dP} - \ln \lambda = \Psi(P) - \ln \lambda \\
p \lim \frac{1}{n} \sum_i \frac{1}{x_i} &= \frac{\lambda}{P-1}
\end{aligned}$$

其中函数 $\Gamma(P)$ 和 $\Psi(P)$ 是关于 P 的函数, 这些方程中的任何两个都能用来估计 λ 和 P 。

另外, 这个方程组还可能扩充, 因为对 $k = 1, 2, \dots$, 可以证明 $p \lim m'_r = \lambda^{-k} \Gamma(P+k) / \Gamma(P)$ 。

在多数情形中, 矩方法估计量都不是有效的, 例外是来自于指数族分布的随机抽样。

Definition 10.1 一个指数 (参数) 族分布是对数似然有如下形式的分布:

$$\ln L(\boldsymbol{\theta} \mid \mathbf{X}) = a(\mathbf{X}) + b(\boldsymbol{\theta}) + \sum_j c_j(\mathbf{X}) s_j(\boldsymbol{\theta})$$

其中 $a(\cdot)$ 、 $b(\cdot)$ 、 $c(\cdot)$ 和 $s(\cdot)$ 都是函数。(“族”中的成员由不同参数值来区分)。若对数似然函数具有这个形式, 则函数 $c_j(\cdot)$ 称作充分统计量。

当充分统计量存在时, 矩方法估计量将是它们的函数, 而且可以证明, 矩方法估计量将是最大似然估计量, 所以, 它们当然是有效的, 并且是渐近有效的。

Example 10.3 (实例 3 (续前例)) 伽马分布的对数似然函数是:

$$\ln L = n \left(P \ln \lambda - \ln \Gamma(P) - \lambda \sum_i x_i + (P-1) \sum_i \ln x_i \right)$$

这是有 $a(\mathbf{X}) = 0$, $b(\boldsymbol{\theta}) = n(P \ln \lambda - \ln \Gamma(P))$ 的一个指数族, 且两个充分统计量是 $\sum_i x_i$ 和 $\sum_i \ln x_i$ 。 $\sum_i x_i$ 的矩方法估计量将是最大似然估计量。

对于原版书中图 4.1 (P.98) 中的收入数据, 下面列出的 4 个矩是:

$$\begin{aligned}
\frac{1}{n} \left[\sum_i x_i, \sum_i x_i^2, \sum_i \ln x_i, \sum_i \frac{1}{x_i} \right] &= [31.278, 1453.957, 3.221387, 0.0500141] \\
&= (m'_1, m'_2, m'_*, m'_{-1})
\end{aligned}$$

将估计的参数表示为 $\theta = (P, \lambda)$, 基于这些矩中一对的 P 和 λ 的矩方法估计量如下:

$$\begin{aligned}
\hat{\theta}(m'_1, m'_2) &= (2.05682, 0.065759), & \hat{\theta}(m'_1, m'_*) &= (2.4106, 0.077707) \\
\hat{\theta}(m'_1, m'_{-1}) &= (2.7714, 0.088605), & \hat{\theta}(m'_2, m'_*) &= (2.2644, 0.071302) \\
\hat{\theta}(m'_{-1}, m'_2) &= (2.3952, 0.074797), & \hat{\theta}(m'_{-1}, m'_*) &= (3.0358, 0.101782)
\end{aligned}$$

最大似然估计是:

$$\hat{\theta}(m'_1, m'_*) = (2.4206, 0.77707)$$

Example 10.4 ([实例 4]) (例 3 继续) J 在例 3 中, 我们计算 4 个样本矩

$$\bar{g} = \frac{1}{n} \sum \left[x_i, x_i^2, \frac{1}{x_i}, \ln x_i \right]$$

分别具有概率极限 $P/\lambda, P(P+1)/\lambda^2, \lambda/(P-1)$, 任意一对可以用于估计两个参数, 但正如前面例子所显示的, 6 对产生 (P, λ) 的 6 对不同的估计。

在这样的例子中, 为利用样本的所有信息, 有必要想出一个办法来调和将出现在过度决定系统中的互相冲突的估计。

更为一般地, 假设模型包括 K 个参数 $\theta = (\theta_1, \theta_2, \dots, \theta_K)$, 以及包含一组 $J > K$ 个矩条件,

$$\mathbb{E}[m_j(y, x, z, \theta)] = 0, j = 1, \dots, J$$

其中 y_i, x_i 和 z_i 是模型中出现的变量的样本点。把相应的样本和记作

$$\bar{m}_j(y, x, Z, \theta) = \frac{1}{n} \sum_i m_j(y, x_i, z_i, \theta)$$

除非方程组是函数相关的 (*functionally dependent*), K 个未知参数 J 个方程的系统,

$$\bar{m}_j = \frac{1}{n} \sum_i m_j(y, x_i, z_i, \theta) = 0, j = 1, \dots, J$$

将不会有惟一解。必须调和可能产生的 $\binom{J}{K}$ 种不同的估计。一种可能性是最小化准则函数, 例如平方和

$$q = \sum \bar{m}_j^2 = \bar{m}(\theta)' \bar{m}(\theta) \quad (10.5.1)$$

以上是一种 *OLS* 方法。由于矩是观测值的和, 实际上它们是方差可以估计的随机变量。因此, 根据使 *GLS* 比 *OLS* 更可取的同样逻辑, 使用一种加权方法应该有所帮助, 其中权数与矩的方差成反比。令 W 是对角线元素为:

$$w_{jj} = \text{Asy. Var}[\bar{m}_j]$$

的对角矩阵。于是, 加权最小二乘方法将最小化:

$$q = \bar{m}' W^{-1} \bar{m}$$

然而, \bar{m} 的 J 个元素是自由相关的, 上式中我们使用的对角 W 阵忽视了这种相关性。为了使用 GLS 的思路与方法, 我们应定义:

$$W = \text{Asy} \cdot \text{Var}[\bar{m}] \quad (10.5.2)$$

通过选择 θ 来最小化

$$q = \bar{m}(\theta)' W^{-1} \bar{m}(\theta)$$

而定义的估计量是最小距离估计量。由于 10.5.1 中的 OLS 准则使用单位矩阵 I , 这产生了一个一致估计量, 与加权最小二乘估计量和完全 GLS 估计量一样, 剩下要决定的是使用最佳的 W 。再一次根据促使 GLS 的逻辑, 直觉告诉我们 10.5.2 中定义的是最优的。这是现在著名的汉森 (1982) 结果。这个广义矩法估计量 (GMME) 的渐近协方差矩阵是

$$\Sigma = [G' W^{-1} G]^{-1}$$

其中 G 是一个导数矩阵, 它们的第 j 行是

$$G^j = \frac{\partial \bar{m}_j}{\partial \theta'}$$

最后, 把中心极限定理应用于样本矩并把斯拉茨基定理应用于这个过程, 我们可以断定:

$$\hat{\theta} \xrightarrow{a} N[\vartheta, \Sigma]$$

§ 10.6 异方差的检验异方差的多数检验均基于下述策略

即便存在异方差性, 普通最小二乘也是 β 的一致估计量。所以, 尽管由于抽样变化而不是十分完美, 普通最小二乘残差仍将非常近似于真实扰动的异方差。因此, 在大多数情况下, 为判定异方差性是否存在而设计的检验均采用普通最小二乘残差。

§ 10.6.1 怀特的一般检验 (White' s General Test)

能对下述一般假设进行检验是乎是合理的检验

$$H_0 : \sigma_i^2 = \sigma^2 \text{ 对所有 } i$$

$$H_1 : \sigma_i^2 \neq \sigma^2$$

用 n 个样本对 n 个参数的模型进行的估计, 是一件十分困难的事, 因此, 对这种检验是极具挑战性的。但这种检验已经被怀特于 1980 年设计出来。

异方差条件下的最小二乘估计量（OLS）的协方差矩阵是：

$$\text{Var}[\mathbf{b}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} [\mathbf{X}\boldsymbol{\Omega}\mathbf{X}] [\mathbf{X}'\mathbf{X}]^{-1}$$

我们可用如下式对它加以估计：

$$\text{Est. Var}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1} \left[\sum_{i=1}^n e_i^2 (\mathbf{X}_i \mathbf{X}_i') \right] (\mathbf{X}'\mathbf{X})^{-1}$$

如果不存在异方差性，最小二乘估计量（OLS）的协方差矩阵是：

$$\text{Var}[\mathbf{b}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

可得到 $\text{Var}[\mathbf{b}]$ 的一个估计量，

$$\text{Est. Var}[\mathbf{b}] = \left(\frac{1}{n} \right) s^2 \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} = s^2 (\mathbf{X}'\mathbf{X})^{-1}$$

方法：将 e_i^2 对一个常数 X 中所有的单一变量的组合组成的变量进行回归，得到 nR^2 。

这个统计量渐近地服从 $P-1$ 个自由度的卡方分布， $nR^2 \sim \chi^2(P-1)$, (Why?) 其中 $R^2 = SSR/SST$, P 为回归量的数量，但不包含常数

怀特检验极为一般。为进行此检验，我们不需对异方差的性质作任何特定的假设。尽管这是优点，但同时也是极为严重缺点。怀特检验可揭示异方差性，但也可能导致简单地识别某些其他的设定误差（如从一个简单回归中省略 x^2 ）。此外，不同于我们要讨论的其他检验，怀特检验是非建设性，如果我们拒绝同方差假设，检验的结果对我们下一步应当做什么没有任何启示。

§ 10.6.2 戈德菲尔德—匡特检验 (The Goldfeld-Quandt Test)

另外两个相对一般性的检验是戈尔德—匡特检验（1965）和布罗施—帕甘（1979）拉格朗日乘数检验。

对于戈德菲尔德—匡特检验，我们假设观测值的扰动方差相同，而在备择假设情况下，扰动方差可存在系统性差别。此检验最理想的情形是组间异方差模型或者对某变量 x 满足 $\sigma_i^2 = \sigma^2 x_i^2$ 的这类模型。以该 x 为基础对观测值进行排列，可将观测值分成高方差和抵方差两部分。通过将样本分成具有 n_1 和 n_2 个观测值的两组来进行此检验。为获得统计上独立的方差估计量，回归是采用两组观测值分别进行估计的。该检验统计量为：

其中我们假定第一个样本中的扰动方差大于第二组（若非如此，可变换下标）。在同方差的零假设情况下，此统计量为自由度是 n_1-K 和 n_2-K 的 F 分布。可将样本值对照标准 F 表，若样本值较大，则可拒绝零假设，这样检验就完成了。

提请注意的是：如果扰动项是正态分布的，戈德菲尔德-匡特统计量在零假设下严格服从 F 分布，且该检验的名义值是合适的；但如果扰动项不是正态分布，则 F 分布是不适当的，需要具有已知大样本性质的某些备择方法。

§ 10.7 练习

1. $X \sim N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $\sigma^2 > 0$ 均是未知参数, X_1, \dots, X_n 是一子样, 试求检验问题 $H_0: \sigma^2 = \sigma_0^2$ 的广义似然比检验。
2. $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, (X_1, \dots, X_m) 是分别从抽取的独立子样, 试求检验问题的 $H_0: \frac{\sigma_1^2}{\sigma_2^2} = \Delta_0$ 广义似然比检验。
3. GLS 估计量和它与 OLS 估计量的差的协方差矩阵是什么?

$$\text{Cov}[\hat{\beta}, \hat{\beta} - b]$$

其中 $\hat{\beta} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}$ 并且 $b = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 。该结果豪斯曼 (1978) 的设定检验的发展中起到关键作用。

Part IV

Appendix

附录 A

关于生产函数的计量经济模型

在西方经济学中，生产理论是最重要内容之一；同样，在西方的计量经济学中，生产函数模型的研究与发展始终是一个重要的、最活跃的领域。本章主要内容：

§ A.1 几个重要基本概念

§ A.1.1 生产函数

1) 定义

生产函数是描述生产过程中投入的生产要素的某种组合同它可能的最大产出量之间的依存关系的数学表达。即

$$Y = f(A, K, L, \dots) \quad (\text{A.1.1})$$

其中 Y 为产出量， A 、 K 、 L 分别为技术、资本、劳动等投入要素。这里“投入的生产要素”是生产过程中发挥作用、对产出量产生贡献的生产要素；“可能的最大产出量”指这种要素组合应该形成的产出量，而不一定是实际产出量。生产要素对产出量的作用与影响，主要是由一定的技术条件决定的，所以，从本质上讲，生产函数反映了生产过程中投入要素与产出量之间的技术关系。

2) 生产函数模型的发展

从 20 世纪 20 年代末，美国数学家 Charles Cobb 和经济学家 Paul Duaglas 提出了生产函数这一名词，并用 1899—1922 年的数据资料，导出了著名的 Cobb-Dauglas 生产函数以来，不断有新的研究成果出现，使生产函数的研究与应用呈现长盛不衰的局面。下面列出的是这期间出现的主要成果。

表 A.1 生产函数模型的发展

年份	生产函数类型
1928 年 Cobb, Dauglas	C-D 生产函数
1937 年 Dauglas, Durand	C-D 生产函数的改进型
1957 年 Solow	C-D 生产函数的改进型
1960 年 Solow	含体现型技术进步生产函数
1961 年 Arrow 等	两要素 CES 生产函数
1967 年 Sato	二级 CES 生产函数
1968 年 Sato, Hoffman	VES 生产函数
1968 年 Aigner, Chu	边界生产函数
1971 年 Revanker	VES 生产函数
1973 年 Christensen, Jorgenson	超越对数生产函数
1980 年	三级 CES 生产函数

在这期间，关于生产函数估计方法的研究成果也很多，在本节中将结合生产函数的估计加以介绍。

3) 生产函数是经验的产物

上面列举的生产函数，都是在西方国家发展起来的，当然，作为西方经济理论体系的一部分，它是与特定的生产理论紧密联系的。例如，作为生产函数理论基础的要素价值论和由此产生的分配理论，都体现在生产函数模型中。那么，这些生产函数模型在中国有无应用价值？事实已经表明，西方国家发展的生产函数模型在中国应用具有一定的合理性。

首先，正如前面已经提及的，生产函数所描述的是投入要素与产出量之间的技术关系。无论何种社会制度，任何生产过程都必须是劳动、技术与生产资料的结合，也就是必须具备一定的投入要素、在一定的技术条件下才能进行生产。生产函数实际上是用数学公式对现实发生的生产过程中的投入要素与产出量之间的技术关系进行拟合，是对生产过程中量的关系的描述。

其次，西方的生产函数并不是西方生产理论的直接推导结果，而是经验的产物，是以数据为样本，反复拟合、检验、修正后得到的。换句话说，如果抛开已有的生产函数模型，用我国的有关数据，也能得到相同或相似的生产函数模型。

4) 生产函数的一阶齐次性

如果生产函数 1) 中资本、劳动等非技术要素的投入量同时增长 λ 倍，根据生产理论中规模报酬不变法则，产出量也应该增长 λ 倍。即

$$f(\lambda K, \lambda L, \cdots) = \lambda f(K, L, \cdots)$$

称为生产函数的一阶齐次性。但是在实际生活中,存在着规模报酬递增或者规模报酬递减的现象,所以并非所有生产函数模型都具有一阶齐次性。

§ A.1.2 要素替代弹性

在生产函数模型的研究与发展中,要素替代弹性是一个十分重要的概念。所谓要素替代弹性,是描述投入要素之间替代性质的一个量,主要用于描述要素之间替代能力的大小。要素替代弹性是与研究对象、样本区间甚至样本点联系在一起的。所以,在建立生产函数模型之前,需要对要素替代弹性作出假设,不同的假设,会导致差异甚大的生产函数模型。

在引入要素替代弹性的定义之前,需要首先引入如下概念:

1) 要素的边际产量

边际产量是指其他条件不变时,某一种投入要素增加一个单位时导致的产出量的增加量。用于描述投入要素对产出量的影响程度。边际产量可以表示为

$$MP_K = \partial f / \partial K$$

$$MP_L = \partial f / \partial L$$

$$\vdots$$

在一般情况下,边际产量满足

$$MP_K \geq 0, MP_L \geq 0, \dots$$

即边际产量不为负。在大多数情况下,边际产量还满足

$$\frac{\partial (MP_K)}{\partial K} = \frac{\partial^2 f}{\partial K^2} \leq 0$$

$$\frac{\partial (MP_L)}{\partial L} = \frac{\partial^2 f}{\partial L^2} \leq 0$$

$$\vdots$$

即边际产量递减规律。

2) 要素的边际替代率

当两种要素可以互相替代时,就可采用不同的要素组合生产相同数量的产出量。要素的边际替代率指的是在产量一定的情况下,某一种要素的增加与另一种要素的减少之间的比例。

用 $MRS_{K \rightarrow L}$ 表示 K 对 L 的边际替代率,即在保持产量不变的情况下,替代 1 单位 L 所需要增加的 K 的数量。于是有

$$MRS_{K \rightarrow L} = \Delta K / \Delta L \quad (Y \text{ 保持不变})$$

因为边际产量也可以表示为

$$MP_K = \Delta Y / \Delta K$$

$$MP_L = \Delta Y / \Delta L$$

所以

$$\frac{MP_L}{MP_K} = \frac{\Delta Y}{\Delta L} \bigg/ \frac{\Delta Y}{\Delta K} = \frac{\Delta K}{\Delta L}$$

于是要素的边际替代率可以表示为要素的边际产量之比, 即

$$MRS_{K \rightarrow L} = MP_L / MP_K$$

$$MRS_{L \rightarrow K} = MP_K / MP_L$$

3) 要素替代弹性

将要素替代弹性定义为两种要素的比例的变化率与边际替代率的变化率之比, 一般用 σ 表示。则有

$$\sigma = \frac{d(K/L)}{(K/L)} \bigg/ \frac{d(MP_L/MP_K)}{(MP_L/MP_K)} \quad (\text{A.1.2})$$

一般情况下, 要素替代弹性 σ 为一个正数。

如果用 K 替代 L , 则(A.1.2)式分子大于 0; 由于 L 减少, 其边际产量 MP_L 增大, 而由于 K 增加, 其边际产量 MP_K 减小, 于是(A.1.2)式分母也大于 0。所以要素替代弹性 σ 大于 0, 表明要素之间具有有限可替代性;

在特殊情况下, 要素之间不可以替代, 此时 K/L 不变, 则(A.1.2)式分子等于 0, 所以替代弹性 σ 等于 0;

另一种极端情况是, 无论要素的数量增加或者减少, 其边际产量不变, 此时(A.1.2)式分母等于 0, 替代弹性 σ 为 ∞ , 表明要素之间具有无限可替代性。

§ A.1.3 要素的产出弹性

某投入要素的产出弹性被定义为, 当其他投入要素不变时, 该要素增加 1% 所引起的产出量的变化率。这是从动态变化的角度衡量生产要素对产出量的影响的指标。如果用 E_K 表示资本的产出弹性, 用 E_L 表示劳动的产出弹性, 则有

$$\begin{aligned} E_K &= \frac{\Delta Y}{Y} \bigg/ \frac{\Delta K}{K} = \frac{\partial f}{\partial K} \cdot \frac{K}{Y} \\ E_L &= \frac{\Delta Y}{Y} \bigg/ \frac{\Delta L}{L} = \frac{\partial f}{\partial L} \cdot \frac{L}{Y} \end{aligned} \quad (\text{A.1.3})$$

一般情况下, 要素的产出弹性大于 0 小于 1。

§ A.1.4 技术进步

从本质上讲，生产函数所描述的是投入要素与产出量之间的技术关系。即是说，同样的投入要素组合，在不同的技术条件下，产出量是不同的。所以在生产函数模型中必须引入技术进步因素。技术进步是一个广泛的研究领域，这里仅就生产函数模型中涉及到的有关技术进步的一些概念略作说明。

1) 广义技术进步与狭义技术进步

所谓狭义技术进步，仅指要素质量的提高。例如，由于性能的改进，同样数量的资本在生产过程中的贡献是不一样的；由于文化水平的提高，同样数量的劳动在生产过程中的贡献是不一样的。狭义的技术进步是体现在要素上的，可以通过要素的“等价数量”来表示。例如，如果一个具有大学文化水平的劳动者对产出量的贡献是一个具有中学文化水平劳动者的3倍，那么就可以将一个具有大学文化水平的劳动者等价于3个具有中学文化水平劳动者，求得“等价劳动数量”，作为生产函数模型的样本观测值，以这样的方法来引入技术进步因素。

所谓广义技术进步，除了要素质量的提高外，还包括管理水平的提高等对产出量具有重要影响的因素，这些因素是独立于要素之外的。在生产函数模型中需要特别处理。

2) 中性技术进步

假设在生产活动中除技术以外，只有资本与劳动两种要素，定义两要素的产出弹性之比为相对资本密集度，用 ω 表示。即

$$\omega = E_L/E_K$$

a) 如果技术进步使得 ω 越来越大，即劳动的产出弹性比资本的产出弹性增长得快，则称之为节约劳动型技术进步；

b) 如果技术进步使得 ω 越来越小，即劳动的产出弹性比资本的产出弹性增长得慢，则称之为节约资本型技术进步；

c) 如果技术进步前后 ω 不变，即劳动的产出弹性与资本的产出弹性同步增长，则称之为中性技术进步。

在中性技术进步中：

如果要素之比 K/L 不随时间变化，则称为希克斯中性技术进步；

如果劳动产出率 Y/L 不随时间变化，则称为索洛中性技术进步；

如果资本产出率 Y/K 不随时间变化，则称为哈罗德中性技术进步。

不同的技术进步类型是建立生产函数模型时必须要考虑的重要因素，对生产函数模型将产生重要影响。

§ A.2 以要素之间替代性质的描述为线索的生产函数模型的发展

模型是对现实的模拟，生产函数模型是对生产活动中产出量与投入要素组合之间关系的模拟。模型总是建立在一定的假设的基础上的，没有假设，就没有模型。而假设与现实之间是有差距的，差距越小，模型对现实的描述越准确。假设向现实的逼近，导致了模型的不断发展。

生产函数模型的一个基本假设是关于要素之间替代性质的假设，由于该假设不同，导致生产函数的发展，出现了各种不同的生产函数模型。在下面的讨论中我们先只考虑两种要素的情况，最后将问题推广到多要素的情况。同时为了书写简便，我们在讨论模型的发展时，只写出它们的数理形态（即不写出随机误差项）。

§ A.2.1 线性生产函数模型

如果假设资本 K 与劳动 L 之间是无限可以替代的，则产出量 Y 与投入要素组合之间的关系可以用如下形式的模型描述：

$$Y = a_0 + a_1K + a_2L \quad (\text{A.2.1})$$

对于该模型，要素的边际产量 $MP_K = a_1$ ， $MP_L = a_2$ ，边际产量之比 $MP_K/MP_L = a_1/a_2$ 。于是有

$$d(MP_K/MP_L) = 0$$

代入(A.1.2)得到，即要素替代弹性为 $\sigma = \infty$ 。从(A.2.1)也可以直观地看出，一种要素可以被另一种要素替代直至减少为 0，产出量仍然不变。

§ A.2.2 投入产出生产函数模型

另一种极端的情况是假设资本 K 与劳动 L 之间是完全不可以替代的，则产出量 Y 与投入要素组合之间的关系可以用如下形式的模型描述——Harrod-Domar 模型：

$$Y = \min \left(\frac{K}{a}, \frac{L}{b} \right) \quad (\text{A.2.2})$$

称为投入产出型生产函数。其中 a ， b 为生产 1 单位的产出量所必须投入的资本、劳动的数量。由于 a ， b 为常数，所以产出量 Y 所必须的资本投入量 $K = aY$ ，劳动投入量 $L = bY$ ，二者之比 $K/L = a/b$ 为常数， $d(K/L) = 0$ 。代入(A.1.2)得到，即要素替代弹性为 0，资本 K 与劳动 L 之间完全不可以替代。

§ A.2.3 C-D 生产函数模型

1) 模型形式与参数的含义

1928 年美国数学家 Charles Cobb 和经济学家 Paul Douglas 提出的生产函数的数学形式为:

$$Y = AK^\alpha L^\beta \quad (\text{A.2.3})$$

根据要素的产出弹性的定义, 很容易推出

$$\begin{aligned} E_K &= \frac{\partial Y}{\partial K} / \frac{Y}{K} = A\alpha K^{\alpha-1} L^\beta \frac{Y}{K} = \alpha \\ E_L &= \frac{\partial Y}{\partial L} / \frac{Y}{L} = AK^\alpha \beta L^{\beta-1} \frac{Y}{L} = \beta \end{aligned}$$

即参数 α 、 β 分别是资本与劳动的产出弹性。那么由产出弹性的经济意义, 应该有

$$0 \leq \alpha \leq 1, 0 \leq \beta \leq 1$$

在最初提出的 C-D 生产函数中, 假定参数满足 $\alpha + \beta = 1$, 即生产函数的一阶齐次性, 也就是假定研究对象满足规模报酬不变。因为

$$A(\lambda K)^\alpha (\lambda L)^\beta = \lambda^{\alpha+\beta} AK^\alpha L^\beta = \lambda AK^\alpha L^\beta$$

即当资本与劳动的数量同时增长 λ 倍时, 产出量也增长 λ 倍。

1937 年, Durand 提出了 C-D 生产函数的改进型, 即取消了 $\alpha + \beta = 1$ 的假定, 允许要素的产出弹性之和大于 1 或小于 1, 即承认研究对象可以是规模报酬递增的, 也可以是规模报酬递减的, 取决于参数的估计结果。

模型(A.2.3)中的待估参数 A 为效率系数, 是广义技术进步水平的反映, 以后将对它进行专门讨论。显然, 应该有 $A > 0$ 。

由上可见, C-D 生产函数模型的参数具有明确的经济意义, 这是它的一个显著特点, 是它被广泛应用的一个重要原因。

Cobb 和 Douglas 利用美国 1899—1922 年的数据资料为样本, 估计模型的参数, 得到

$$Y = 1.01K^{0.25}L^{0.75}$$

2) 要素替代弹性

现在来看看模型(A.2.3)对要素替代弹性的假设。根据(A.1.2)式, 可以得到

$$\begin{aligned} \sigma &= \frac{d(K/L)}{(K/L)} / \frac{d(MP_L/MP_K)}{(MP_L/MP_K)} \\ &= d\left(\ln\left(\frac{K}{L}\right)\right) / d\left(\ln\left(\frac{MP_L}{MP_K}\right)\right) \\ &= d\left(\ln\left(\frac{K}{L}\right)\right) / d\left(\ln\left(\frac{\beta K}{\alpha L}\right)\right) \\ &= d\left(\ln\left(\frac{K}{L}\right)\right) / d\left(\ln\left(\frac{\beta}{\alpha}\right) + \ln\left(\frac{K}{L}\right)\right) \\ &= 1 \end{aligned}$$

这是一个重要的结论，它表明 **C-D 生产函数模型假设要素替代弹性为 1**。

显然，与上述要素之间可以无限替代的线性生产函数模型和要素之间完全不可以替代的投入产出生产函数模型相比较，C-D 生产函数模型假设要素替代弹性为 1，是更加逼近于生产活动的实际，是一个很大的进步。正因为此，加之 C-D 生产函数模型的参数具有明确的经济意义，使得它一经提出，就得到广泛的应用。直到今天，它仍然是应用最广泛的一种生产函数模型。

但是，C-D 生产函数模型关于要素替代弹性为 1 的假设仍然具有缺陷。根据这一假设，不管研究对象是什么，不管样本区间是什么，不管样本观测值是什么，要素替代弹性都为 1，这是与实际不符的。例如，劳动密集型的农业与资本密集型的现代工业，资本与劳动之间的替代性质是明显不同的；再例如，对于同一个研究对象，如果样本区间不同，即考察的区间不同，要素之间的替代性质也应该是不同的；即使研究对象相同、样本区间相同，对于不同的样本点，由于要素的比例不同，相互之间的替代性质也应该是不同的。所有这些，都需要人们发展新的生产函数模型。

§ A.2.4 不变替代弹性 (CES) 生产函数模型

1) 模型形式与参数的含义

在 1961 年，由 Arrow、Chenery、Mihás 和 Solow 四位学者提出了两要素不变替代弹性 (Constant Elasticity of Substitution) 生产函数模型，简称 **CES 生产函数模型**，其基本形式如下：

$$Y = A (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{1}{\rho}} \quad (\text{A.2.4})$$

其中，待估参数 A 为效率系数，是广义技术进步水平的反映，显然，应该有 $A > 0$ ； δ_1 和 δ_2 为分配系数， $0 < \delta_1 < 1$ ， $0 < \delta_2 < 1$ ，并且满足 $\delta_1 + \delta_2 = 1$ ； ρ 为替代参数，下面将专门讨论。(A.2.4) 假定研究对象具有不变规模报酬，因为

$$A (\delta_1 (\lambda K)^{-\rho} + \delta_2 (\lambda L)^{-\rho})^{-\frac{1}{\rho}} = \lambda \left(A (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{1}{\rho}} \right)$$

即当资本与劳动的数量同时增长 λ 倍时，产出量也增长 λ 倍。

后来，在应用中取消了这一假定，将(A.2.4)式改写为

$$Y = A (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{m}{\rho}} \quad (\text{A.2.5})$$

对 (8)，有

$$A (\delta_1 (\lambda K)^{-\rho} + \delta_2 (\lambda L)^{-\rho})^{-\frac{m}{\rho}} = \lambda^m \left(A (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{m}{\rho}} \right)$$

即承认研究对象可以是规模报酬递增的,也可以是规模报酬递减的,取决于参数 m 的估计结果。于是参数 m 为规模报酬参数,当 $m = 1 (< 1, > 1)$ 时,表明研究对象是规模报酬不变(递减、递增)的。(A.2.5)为实际应用的 CES 生产函数模型的理论形式。

2) 要素替代弹性

现在来看看模型(A.2.4)对要素替代弹性的假设。根据(A.1.2)式,要素替代弹性为

$$\begin{aligned}\sigma &= \frac{d(K/L)}{(K/L)} \bigg/ \frac{d(MP_L/MP_K)}{(MP_L/MP_K)} \\ &= d\left(\ln\left(\frac{K}{L}\right)\right) \bigg/ d\left(\ln\left(\frac{MP_L}{MP_K}\right)\right)\end{aligned}$$

因为

$$\begin{aligned}MP_K &= \frac{\partial Y}{\partial K} \\ &= A \left(-\frac{1}{\rho}\right) (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{1}{\rho}-1} \delta_1 (-\rho) K^{-\rho-1} \\ &= AK^{-1-\rho} (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{1}{\rho}-1} \delta_1 \\ MP_L &= AL^{-1-\rho} (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{1}{\rho}-1} \delta_2 \\ \Rightarrow \frac{MP_L}{MP_K} &= \frac{\delta_2}{\delta_1} \left(\frac{K}{L}\right)^{1+\rho}\end{aligned}$$

所以

$$\begin{aligned}\sigma &= d\left(\ln\left(\frac{K}{L}\right)\right) \bigg/ d\left(\ln\left(\frac{\delta_2}{\delta_1} \left(\frac{K}{L}\right)^{1+\rho}\right)\right) \\ &= d\left(\ln\left(\frac{K}{L}\right)\right) \bigg/ d\left(\ln\left(\frac{\delta_2}{\delta_1}\right) + (1+\rho)\ln\left(\frac{K}{L}\right)\right) \\ &= \frac{1}{1+\rho}\end{aligned}\tag{A.2.6}$$

由于要素替代弹性 σ 为一正数,所以参数 ρ 的数值范围为

$$-1 < \rho < \infty$$

由(A.2.6)可以看出,一旦研究对象确定、样本观测值给定,可以得到参数 ρ 的估计值,并计算得到要素替代弹性 σ 的估计值。对于不同的研究对象,或者同一研究对象的不同的样本区间,由于样本观测值不同,要素替代弹性是不同的。这使得 CES 生产函数比 C-D 生产函数更接近现实。但是,在 CES 生产函数中,仍然假定要素替代弹性与样本点无关,这就是不变替代弹性生产函数模型的“不变”的含义。而这一点,仍然是与实际不符的。对于不同的样本点,由于要素的比例不同,相互之间的替代性质也应该是不同的。所以,不变替代弹性生产函数模型还需要发展。

值得注意的是:在不变替代弹性生产函数模型中,如果参数 ρ 的估计值等于 0,则要素替代弹性 σ 的估计值为 1,此时 CES 生产函数退化为 C-D 生产函数。

§ A.2.5 变替代弹性 (VES) 生产函数模型

变替代弹性 (Variable Elasticity of Substitution) 生产函数模型, 简称 VES 生产函数模型, 有许多理论和方法方面的研究成果, 是生产函数研究的一个前沿领域。较著名的是 Revankar 于 1971 年提出的模型和 Sato 与 Hoffman 于 1968 年提出的模型。

前者假定要素替代弹性 σ 为要素比例的线性函数, 即

$$\sigma = a + b \cdot \frac{K}{L}$$

容易理解, 要素比例不同, 要素之间的替代性能是不同的。当 K/L 较大时, 资本替代过去就比较困难; 当 K/L 较小时, 资本替代过去就比较容易。生产函数的一般形式为 (不加证明给出)

$$Z = A \exp \int \frac{dk}{k + c \left(\frac{k}{a+bk} \right)^{1/a}} \quad (\text{A.2.7})$$

其中, $Z = Y/L, k = K/L$ 。

后者假定要素替代弹性 σ 为时间的线性函数, 即

$$t\sigma = \sigma(t) = a + b \cdot t$$

随着时间的推移, 技术的进步将使得要素之间的替代变得容易。生产函数的一般形式为

$$Y = B \left(\lambda L^{\frac{\sigma(t)-1}{\sigma(t)}} + (1-\lambda) K^{\frac{\sigma(t)-1}{\sigma(t)}} \right)^{\frac{\sigma(t)}{\sigma(t)-1}} \quad (\text{A.2.8})$$

在实际应用中, 前者可以与样本观测值相联系, 因而实用价值更大。下面将着重讨论(A.2.7)式所表示的 VES 生产函数模型。

1) 当 $b = 0$ 时, (A.2.7)变为

$$\begin{aligned} \frac{Y}{L} &= A \exp \int \frac{dk}{k + c \left(\frac{k}{a} \right)^{1/a}} \\ &= A \exp \left(\frac{a}{1-a} \ln \frac{k^{\frac{1-a}{a}}}{1 + \frac{c}{a^{1/a}} k^{\frac{1-a}{a}}} + \mu \right) \end{aligned}$$

令 $\frac{1-a}{a} = \rho$, $Ae^\mu = A'$, 则有

$$\begin{aligned} \frac{Y}{L} &= A' \left(\frac{a^{1/a} + ck^\rho}{a^{1/a} k^\rho} \right)^{-\frac{1}{\rho}} \\ &= A'' (a^{1/a} k^{-\rho} + c)^{-\frac{1}{\rho}} \\ \Rightarrow Y &= A'' \left(a^{1/a} \left(\frac{K}{L} \right)^{-\rho} + c \right)^{-\frac{1}{\rho}} \cdot L \\ &= A'' (a^{1/a} K^{-\rho} + cL^{-\rho})^{-\frac{1}{\rho}} \quad (\text{A.2.9}) \end{aligned}$$

此时, VES 生产函数模型退化为(A.2.9)所表示的 CES 生产函数模型。

2) 当 $b = 0, a = 1$ 时, (A.2.7)变为

$$\begin{aligned}\frac{Y}{L} &= A \exp \int \frac{dk}{k(1+c)} \\ &= A' \exp \left(\frac{\ln k}{1+c} \right) = A' k^{\frac{1}{1+c}} \\ \Rightarrow Y &= A' K^{\frac{1}{1+c}} \cdot L^{-\frac{1}{1+c}} \cdot L = A' K^{\frac{1}{1+c}} \cdot L^{\frac{c}{1+c}}\end{aligned}\quad (\text{A.2.10})$$

此时, VES 生产函数模型退化为(A.2.10)所表示的 C-D 生产函数模型。

3) 当 $a = 1$ 时, $\sigma = 1 + bk$, (A.2.7)可写成

$$Y = AK^{\frac{1}{1+c}} \left(L + \left(\frac{b}{1+c} \right) K \right)^{\frac{c}{1+c}} \quad (\text{A.2.11})$$

即为一般常用的 VES 生产函数模型, 其中 A 、 b 、 c 是待估参数。(A.2.11)为规模报酬不变的情况, 如果将规模报酬系数 m 作为一个待估参数, 则 VES 生产函数模型的理论形式为

$$Y = AK^{(\frac{1}{1+c})^m} \left(L + \left(\frac{b}{1+c} \right) K \right)^{(\frac{1}{1+c})^m m} \quad (\text{A.2.12})$$

§ A.2.6 多要素生产函数模型

如果作为产出量的解释变量的投入要素多于 2 个, 可以有不同的处理方法, 关键在于对要素之间替代性质的认识。下面以三要素 (资本 K 、劳动 L 和能源 E) 为例介绍几种多要素生产函数模型。

1) 多要素线性生产函数模型

如果资本 K 、劳动 L 和能源 E 互相之间都是无限可以替代的, 则产出量 Y 与投入要素组合之间的关系可以用如下形式的模型描述:

$$Y = \alpha_0 + \alpha_1 K + \alpha_2 L + \alpha_3 E \quad (\text{A.2.13})$$

2) 多要素投入产出生产函数模型

假设资本 K 、劳动 L 和能源 E 互相之间都是完全不可以替代的, 则产出量 Y 与投入要素组合之间的关系可以用如下形式的模型描述:

$$Y = \min \left(\frac{K}{a}, \frac{L}{b}, \frac{E}{c} \right) \quad (\text{A.2.14})$$

3) 多要素 C-D 生产函数模型

假设资本 K 、劳动 L 和能源 E 互相之间的替代弹性都为 1，则产出量 Y 与投入要素组合之间的关系可以用如下形式的模型描述：

$$Y = AK^\alpha L^\beta E^\gamma \quad (\text{A.2.15})$$

4) 多要素一级 CES 生产函数模型

假设资本 K 、劳动 L 和能源 E 互相之间的替代弹性相同，为同一个待估参数，则产出量 Y 与投入要素组合之间的关系可以用如下形式的模型描述：

$$Y = A (\delta_1 K^{-\rho} \delta_2 L^{-\rho} + \delta_3 E^{-\rho})^{-\frac{m}{\rho}} \quad (\text{A.2.16})$$

其中 δ_1 、 δ_2 和 δ_3 为分配系数， $0 < \delta_1 < 1$ ， $0 < \delta_2 < 1$ ， $0 < \delta_3 < 1$ ，并且满足 $\delta_1 + \delta_2 + \delta_3 = 1$ 。要素之间的替代弹性为

$$\sigma = \frac{1}{1 + \rho}$$

5) 多要素二级 CES 生产函数模型

假设资本 K 、劳动 L 和能源 E 互相之间的替代弹性不相同，例如资本与能源之间的替代弹性不同于它们与劳动之间的替代弹性，这是比较符合实际的，那么一级 CES 生产函数模型就不能描述要素之间的替代性质。许多人在探索如何既保持 CES 生产函数的性质，又能解决多要素之间不同替代弹性的问题。1967 年 Sato 提出的多要素二级 CES 生产函数模型，是一个比较成功的具有实用价值的成果。以三要素为例，二级 CES 生产函数模型表达如下：

$$\begin{aligned} Y_{KE} &= (a_1 K^{-\rho_1} + a_2 E^{-\rho_1})^{-\frac{1}{\rho_1}} \\ Y &= A (b_1 Y_{KE}^{-\rho} + b_2 L^{-\rho})^{-\frac{m}{\rho}} \end{aligned} \quad (\text{A.2.17})$$

其中 Y_{KE} 为第一级 CES 生产函数，在第二级 CES 生产函数中，将它作为一个组合要素。

李子奈教授于 1988 年曾经以 1963—1984 年的时间序列数据为样本，采用(A.2.16)和(A.2.17)分别估计我国工业生产函数模型，结果如下：采用(A.2.16)得到资本 K 、劳动 L 和能源 E 互相之间的替代弹性为 0.84；采用(A.2.17)得到资本 K 和能源 E 之间的替代弹性为 0.97，组合要素 Y_{KE} 与劳动 L 之间的替代弹性为 0.71。结果表明，采用模型(A.2.17)，资本 K 和能源 E 之间有较大的替代弹性，而它们二者的组合与劳动之间的替代弹性较小，显然比采用模型(A.2.16)更符合实际。

6) 多要素三级 CES 生产函数模型

当投入要素多于 3 个时，还可以根据要素之间的替代性质，构造三级 CES 生产函数模型，其原理与二级 CES 生产函数模型相同，不再赘述。第一个三级 CES 生产函数模型出现在 1980 年美国的一篇博士论文中。

§ A.2.7 超越对数生产函数模型

一个更具有一般性的变替代弹性生产函数模型是由 L.Christensen、D.Jor-genson 和 Lau 于 1973 年提出的超越对数生产函数模型。其形式为（为何可以采用如下形式？）

$$\ln Y = \beta_0 + \beta_K \ln K + \beta_L \ln L + \beta_{KK}(\ln K)^2 + \beta_{LL}(\ln L)^2 + \beta_{KL} \ln K \cdot \ln L \quad (\text{A.2.18})$$

该生产函数模型的显著特点是它的易估计和包容性。它是一个简单线性模型，可以直接采用单方程线性模型的估计方法进行估计。所谓包容性，是它可以被认为是任何形式的生产函数的近似。例如，如果 $\beta_{KK} = \beta_{LL} = -1/2\beta_{KL}$ ，则表现为 CES 生产函数。所以可以根据该生产函数的估计结果判断要素的替代性质；另外，可以比较容易设计一些假设检验条件，来验证有关经济理论是否正确。

以上是以要素之间的替代性质为线索发展的一系列生产函数模型。从中可以看出，如果没有前人的研究成果，只要掌握模型发展的思路，我们也有可能发展生产函数模型。正是从这个意义说，掌握研究思路比了解几种模型具有更大的意义。

§ A.3 以技术要素的描述为线索的生产函数模型的发展

技术是一种重要的生产要素，在现代生产活动中尤其重要，所以在生产函数模型中不能不考虑技术要素。如何将技术要素引入生产函数模型，如何使得模型对技术要素的描述更逼近于现实，是生产函数研究中一个重要领域，也是迄今没有很好解决的一个难题。

§ A.3.1 将技术要素作为一个不变参数的生产函数模型

在模型(A.2.3)和(A.2.5)中，已经引入了技术要素，但是仅仅将它作为独立于其他投入要素之外的一个不变的参数。其基本假设是：技术进步是广义的；技术进步是中性的；技术进步改变了由其他投入要素的数量决定的生产活动的效率；技术进步的作用在所有样本点上都是相同的。

显然，这些假设是不符合实际的。例如，技术进步的作用在所有样本点上是不相同的。在生产函数研究中，经常以时间序列数据为样本，不同的样本点表示不同的时间，而技术的发展恰恰是与时间紧密相关的。又例如，技术进步的一部分是以其他要素质量的提高为体现的，而各自要素质量提高的速度是不同的，所以技术进步不可能等地改变所有要素的效率。这就需要发展新的生产函数模型。

§ A.3.2 改进的 C-D、CES 生产函数模型

早在 1942 年, Tinbergen 就提出在生产函数中加入时间指数趋势项以测定技术进步, 1957 年 Solow 提出如下改进的 C-D 生产函数模型:

关于 A_t 的形式, 通常有两种设定:

$$A_t = Y = A_0 (1 + \gamma)^t K^\alpha L^\beta$$

$$A_t = A_0 e^{\lambda t}$$

前一种表达式中, γ 具有明确的经济含义, 即表示技术的年进步速度; 在后一种表达式中, λ 的经济含义不明确。但是, 当技术进步速度很低时, 由于

$$\ln(1 + \gamma) \approx \gamma$$

于是有

$$\ln(A_0 (1 + \gamma)^t) = \ln A_0 + \ln(1 + \gamma)^t = \ln A_0 + t\gamma$$

$$\ln(A_0 e^{\lambda t}) = \ln A_0 + t\lambda$$

所以也可以将后一种表达式中的 λ 看作为技术进步速度。改进的 C-D 生产函数模型的表达式为

$$Y = A_0 (1 + \gamma)^t K^\alpha L^\beta \quad (\text{A.3.1})$$

$$Y = A_0 e^{\lambda t} K^\alpha L^\beta \quad (\text{A.3.2})$$

同样的思路, 改进的 CES 生产函数模型的表达式为

$$Y = A_0 (1 + \gamma)^t (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{m}{\rho}} \quad (\text{A.3.3})$$

$$Y = A_0 e^{\lambda t} (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{m}{\rho}} \quad (\text{A.3.4})$$

需要特别注意的是, 上述改进的 C-D、CES 生产函数模型是在关于技术进步的特定假设下成立的, 离开了特定的假设, 这些模型表达式就不正确了。

在本节关于技术进步的概念中曾经提到 3 类技术进步和 3 类中性技术进步。在改进的 C-D、CES 生产函数模型中, 作为资本和劳动产出弹性的参数不随样本点变化, 这就是说技术进步不是节约资本和节约劳动型, 而是中性的。

而在中性技术进步中，希克斯中性技术进步假设要素之比 K/L 不随时间变化，由此可以证明（证明过程需要求解偏微分方程，我们不加证明）考虑技术进步的生产函数形式为

$$Y = A(t) f(K, L)$$

即技术进步的作用相当于在要素投入不变的情况下，使产出增加 $A(t)$ 倍。

索洛中性技术进步假设劳动产出率 Y/L 不随时间变化，由此可以证明，考虑技术进步的生产函数形式为

$$Y = f(A(t) K, L)$$

即技术进步的作用相当于使劳动要素投入增加 $A(t)$ 倍。那么我们稍加推导，就可以发现，对于改进的 C-D 生产函数模型(A.3.1)和(A.3.2)，3 类中性技术进步假设都是适宜的；而对于改进的 CES 生产函数模型(A.3.3)和(A.3.4)，则只有希克斯中性技术进步假设是适宜，因为索洛中性或哈罗德中性技术进步假设下是无法得到(A.3.3)和(A.3.4)形式的生产函数模型表达式的。

从这里我们得到一个启示，任何经济计量模型都是建立在一定假设基础上的，我们在学习与应用已有的模型时，必须搞清楚模型的假设条件，否则就会犯错误。

§ A.3.3 含体现型技术进步的生产函数模型

技术进步要素中有一部分是体现为资本、劳动等要素质量的提高，而资本、劳动等要素质量的提高使得相同数量的要素投入量具有不同的产出效果。所以，如果能将体现为资本、劳动等要素质量提高的技术进步因素从广义技术进步中分离出来，无论是对技术进步的作用机制描述，还是对技术进步作用的数量描述都是十分重要的。由 Solow 于 1964 年首先提出并由 Nelson 于 1964 年补充应用的含体现型技术进步的生产函数模型（也称为 Solow-Nelson 同期模型），就是在这个思路下发展起来的，是生产函数模型的一个重大进展。

1) 总量增长方程

在 1957 年由 Solow 提出了用总量生产函数度量技术进步的**总量增长方程**，认为产出量的增长是由资本数量的增长、劳动数量的增长和技术的进步共同贡献的结果。用数学表达式表示为

$$\frac{\Delta Y}{Y} = \frac{\Delta A}{A} + \alpha \frac{\Delta K}{K} + \beta \frac{\Delta L}{L} \quad (\text{A.3.5})$$

其中 α 和 β 分别为资本和劳动的产出弹性，那么式中后两项分别表示资本数量的增长和劳动数量的增长对产出增长的贡献； $\Delta A/A$ 被用来度量技术进步对产出增长的贡献。但是，在实际上 $\Delta A/A$ 是一个余项，是产出增长中不能被要素数量增长所解释的那一部分，是一个大杂烩，甚至被称为“垃圾箱”。如何从 $\Delta A/A$ 中将不同类型的技术进步因素分离出来，显然是有意义的。

2) 分离资本质量的含体现型技术进步的生产函数模型

将 C-D 生产函数模型改变为

$$Y_t = A'_t J_t^\alpha L_t^\beta \quad (\text{A.3.6})$$

其中 J_t 是以质量加权的资本数量, 也称为有效资本; A'_t 是除了体现为资本质量提高以外的技术效率系数。 J_t 的计算式为

$$J_t = \sum_{m=0}^t K_{mt} (1 + \lambda)^m \quad (\text{A.3.7})$$

其中, K_{mt} 为在第 m 年形成的第 t 年仍然使用的资本数量, λ 为由于资本质量提高带来的资本效率年提高速度。即认为新资本具有更高的质量, 因而具有更高的效率, 相当于资本数量增加了。(A.3.6)即为含体现型技术进步的生产函数模型, 在实际应用时, 给定 λ , 计算 J_t 的样本观测值, 即可以估计该生产函数模型。

人们经常不直接应用(A.3.6), 而是采用另外一种近似形式。引入第 t 年资本的平均寿命 \bar{a}_t , 则有效资本的增长率可以近似写成

$$\frac{\Delta Y}{Y} = \frac{\Delta K}{K} + \lambda - \lambda \cdot \Delta \bar{a} \quad (\text{A.3.8})$$

其中 $\Delta \bar{a}$ 为资本平均年龄的变化, 当资本平均年龄降低时, $\Delta \bar{a}$ 为负值; $\Delta K/K$ 为实际资本数量的变化率; 引入调整量 $\lambda \cdot \Delta \bar{a}$, 反映资本平均年龄变化的作用。于是总量增长方程(A.3.5)变为

$$\frac{\Delta Y}{Y} = \left(\frac{\Delta A'}{A'} + \alpha \lambda - \alpha \lambda \Delta \bar{a} \right) + \alpha \frac{\Delta K}{K} + \beta \frac{\Delta L}{L} \quad (\text{A.3.9})$$

其中括号中部分相当于原方程中的 $\Delta A/A$, 现在从这个“垃圾箱”中将体现资本质量提高的部分 $\alpha \lambda$ 和反映资本平均年龄变化的部分 $\alpha \lambda \Delta \bar{a}$ 分离出来了。(A.3.9)可以改写为

$$\frac{\Delta Y}{Y} = \frac{\Delta A'}{A'} + \alpha \left(\lambda - \lambda \Delta \bar{a} + \frac{\Delta K}{K} \right) + \beta \frac{\Delta L}{L} \quad (\text{A.3.10})$$

这是常用的分离资本质量的含体现型技术进步的生产函数模型。

3) 分离劳动质量的含体现型技术进步的生产函数模型

按照同样的思路, 可以从上述 $\Delta A'/A'$ 中将体现为劳动质量提高的技术进步因素分离出来。例如, 用 δ 表示由于劳动者平均受教育水平的提高带来的劳动效率年提高速度, 用 $\Delta \bar{b}$ 表示劳动者平均年龄的变化。式(A.3.9)可以改写为:

$$\frac{\Delta Y}{Y} = \frac{\Delta A''}{A''} + \alpha \left(\lambda - \lambda \Delta \bar{a} + \frac{\Delta K}{K} \right) + \beta \left(\delta - \delta \Delta \bar{b} + \frac{\Delta L}{L} \right) \quad (\text{A.3.11})$$

式中 $\Delta A''/A''$ 仅表示由于管理水平的提高等技术进步因素对产出增长的贡献。当然, 劳动质量比资本质量更为复杂, (A.3.11)只是一个示意性模型, 有待于我们进一步发展。

§ A.3.4 边界生产函数模型

前面曾经提及,从理论上讲,生产函数是描述一定的投入要素组合与最大产出量之间的关系。但是在实际应用中,人们无法得到最大产出量的样本观测值,只能用实际产出量作为样本观测值估计生产函数模型,因而得到的生产函数仅描述一定的投入要素组合与平均产出量之间的关系。人们已经习惯将后者称为生产函数,为了区别起见,我们把前者称为边界生产函数。

对于平均生产函数,实际产出量可以在它的上方,也可以在它的下方;而对于边界生产函数,实际产出量只能在它的下方。边界生产函数正是根据这一点设置的,它实质上是平均生产函数向上的平移。正是由于这一点,边界生产函数在比较不同样本点的技术效率方面具有重要的应用价值。

边界生产函数按照边界的性质分为确定性边界生产函数和随机边界生产函数两大类。

1) 确定性边界生产函数

确定性边界生产函数把影响产出量的不可控因素(例如观测误差、方程设定误差等)和可控因素(例如生产非效率因素)不加区别,统统归入一个单侧的误差项中,作为对非效率的反映。其模型可以写成

$$Y = f(K, L, \dots) e^{-u} \quad (u \geq 0) \quad (\text{A.3.12})$$

其中 Y 为实际产出量, $f(K, L, \dots)$ 为边界生产函数, $0 \leq e^{-u} \leq 1$ 反映生产非效率,实际产出量总是在 $f(K, L, \dots)$ 的下方。可见边界生产函数 $f(K, L, \dots)$ 为确定性的。

确定性边界生产函数又可以分为确定性非参数边界、确定性参数边界和确定性统计边界生产函数。区别在于估计方法的不同。

2) 随机边界生产函数

随机边界生产函数把影响产出量的不可控因素和可控因素加以区别。其模型可以写成

$$Y = f(K, L, \dots) e^{v-u} = (f(K, L, \dots) e^v) e^{-u} \quad (\text{A.3.13})$$

其中 Y 为实际产出量, $f(K, L, \dots) e^v$ 为边界生产函数, $0 \leq e^{-u} \leq 1$ 反映相对于随机边界的生产非效率,实际产出量总是在 $f(K, L, \dots) e^v$ 的下方,但可以由于随机因素的作用而处于 $f(K, L, \dots)$ 的上方,边界生产函数 $f(K, L, \dots) e^v$ 为随机性的。

§ A.4 几个重要生产函数模型的参数估计方法

§ A.4.1 线性生产函数模型的估计

对于线性生产函数模型(A.2.1), 其计量经济学形态为

$$Y = \alpha_0 + \alpha_1 K + \alpha_2 L + \mu$$

采用单方程线性计量经济学模型的估计方法, 可以很方便地估计其参数。

§ A.4.2 C-D 生产函数模型及其改进型的估计

对于 C-D 生产函数模型(A.2.3)及其改进型(A.3.1)和(A.3.2), 两边取对数, 即可化成线性模型, 然后采用单方程线性计量经济学模型的估计方法估计其参数。但是其假设条件是随机误差项可以作为方程的一个因子与理论模型相乘, 即模型的计量经济学形态为

$$Y = AK^\alpha L^\beta \mu$$

注意这里的随即干扰项并不能作 0 均值处理。

如果随机误差项作为方程的一个因子与理论模型相加, 即

$$Y = AK^\alpha L^\beta \mu + \mu$$

则要采用非线性模型的估计方法估计其参数, 但在实际应用中, 都假设为前一种情况。

§ A.4.3 CES 生产函数模型及其改进型的估计

对于(A.2.5)所表示的 CES 生产函数模型

$$Y = A (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{\rho}{1-\rho}}$$

为一个关于参数的非线性模型, 采用简单的方法难以化为线性模型。自 1961 年以来, 关于它的估计问题有许多研究, 主要有两类方法, 即利用边际生产力条件的估计方法和直接估计方法。

所谓边际生产力条件, 即当生产活动处于均衡的情况下, 存在

$$\frac{\partial Y}{\partial K} = \frac{r}{p} \quad \frac{\partial Y}{\partial L} = \frac{w}{p}$$

其中 r, w, p 分别表示资本的利率、劳动的工资率和产出的价格。将该条件应用于(A.2.5), 经过适当的变换, 可以得到线性计量经济方程。由于边际生产力条件与实际生产活动有较大距离, 在实际上我们基本不

采用这类估计方法。顺便指出,对其他形式的生产函数模型,从理论上讲,也可以利用边际生产力条件进行估计,所以我们称其为“一类”估计方法。

直接估计方法。将 C-D 生产函数模型的计量形态假设为

$$Y = A (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})^{-\frac{m}{\rho}} \mu$$

两边取对数,得到

$$\ln Y = \ln A - \frac{m}{\rho} \ln (\delta_1 K^{-\rho} + \delta_2 L^{-\rho}) + \varepsilon \quad (\text{A.4.1})$$

将其中的 $\ln (\delta_1 K^{-\rho} + \delta_2 L^{-\rho})$ 在 $\rho = 0$ 处展开泰勒级数,取 0 阶、1 阶和 2 阶项,代入(A.4.1),得到

$$\ln Y = \ln A - \frac{m}{\rho} \ln (\delta_1 K^{-\rho} + \delta_2 L^{-\rho}) + \varepsilon \quad (\text{A.4.2})$$

(A.4.2)为一个简单线性模型,通过变量置换,可以表示成

$$Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon$$

采用单方程模型的估计方法,得到 $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ 的估计值,利用对应关系和 $\delta_1 + \delta_2 = 1$,可以计算得到关于参数 $A, \rho, m, \delta_1, \delta_2$ 的估计值。

选择在 $\rho = 0$ 处展开台劳级数,是因为当 $\rho = 0$ 时,要素替代弹性等于 1,即模型退化为 C-D 生产函数,由于 C-D 生产函数的普遍适用性,所以可以假定 ρ 为接近于 0 的数。当参数估计完成后,可以根据 ρ 的估计值是否接近于 0 来检验这种估计方法的可用性。(完成从样本到母体的检验!)

从(A.4.2)可以看出,当 $\rho = 0$ 时,方程为

$$\ln Y = \ln A + \delta_1 m \ln K + \delta_2 m \ln L + \varepsilon$$

即为 C-D 生产函数模型。所以可以认为 CES 生产函数模型是对 C-D 生产函数模型的修正。

对于改进的 CES 生产函数模型(A.3.3)和(A.3.4),估计方法是相同的。

§ A.4.4 VES 生产函数的估计

这里仅讨论(A.2.12)式

$$Y = AK^{\left(\frac{1}{1+c}\right)m} \left(L + \left(\frac{b}{1+c} \right) K \right)^{\left(\frac{c}{1+c}\right)m}$$

的直接估计方法。将它的计量形态假设为

$$Y = AK^{\left(\frac{1}{1+c}\right)m} \left(L + \left(\frac{b}{1+c} \right) K \right)^{\left(\frac{c}{1+c}\right)m} \cdot \mu \quad (\text{A.4.3})$$

其对数形式为

$$\ln Y = \ln A + \frac{m}{1+c} \ln K + \frac{cm}{1+c} \ln \left(L + \frac{b}{1+c} K \right) + \varepsilon \quad (\text{A.4.4})$$

令

$$\ln \left(L + \frac{b}{1+c} K \right) = \ln(L + \lambda \cdot K) = Z(\lambda)$$

在 $\lambda = 0$, 即 $b = 0$ 处展开泰勒级数

$$Z(\lambda) = \ln L + \frac{K}{L} \cdot \lambda + 0(\lambda)$$

代入(A.4.4)得到

$$\ln Y = \ln A + \frac{m}{1+c} \ln K + \frac{cm}{1+c} \ln L + \frac{cmb}{(1+c)^2} \frac{K}{L} + \varepsilon \quad (\text{A.4.5})$$

对(A.4.5)进行变量置换, 得到

$$Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon$$

采用单方程模型的估计方法, 得到 $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ 的估计值, 利用对应关系可以计算得到关于参数 A, c, m, b 的估计值。当参数估计完成后, 可以根据 b 的估计值是否接近于 0 来检验这种估计方法的可用性。

§ A.4.5 二级 CES 生产函数模型的估计

对于二级 CES 生产函数模型(A.2.17)

$$Y_{KE} = (a_1 K^{-\rho_1} + a_2 E^{-\rho_1})^{-\frac{1}{\rho_1}}$$

$$Y = A (b_1 Y_{KE}^{-\rho} + b_2 L^{-\rho})^{-\frac{m}{\rho}}$$

首先将第二级 CES 生产函数取对数, 在 $\rho = 0$ 处展开台劳级数, 得到如下近似式:

$$\ln Y = \ln A + b_1 m \ln Y_{KE} + b_2 m \ln L - \frac{1}{2} \rho m b_1 b_2 \left(\ln \left(\frac{Y_{KE}}{L} \right) \right)^2 + \varepsilon$$

式中包含 Y_{KE} ; 再将第一级 CES 生产函数在 $\rho_1 = 0$ 处展开台劳级数, 得到关于 Y_{KE} 的近似式

$$\ln Y_{KE} = a_1 \ln K + a_2 \ln E - \frac{1}{2} \rho_1 a_1 a_2 \left(\ln \left(\frac{K}{E} \right) \right)^2$$

将 Y_{KE} 的近似式代入第一级 CES 生产函数的展开近似式中, 考虑到可能引起共线性和计算复杂性等因素, 用逐步回归筛选出如下线性方程:

$$\ln Y = \ln A + b_1 m a_1 \ln K + b_1 m a_2 \ln E + b_2 m \ln L - \frac{1}{2} m b_1 \rho_1 a_1 a_2 \left(\ln \left(\frac{K}{E} \right) \right)^2$$

$$- \frac{1}{2} \rho m b_1 b_2 \left(\ln \left(\frac{K}{L} \right) \right)^2 + \varepsilon \quad (\text{A.4.6})$$

通过变量置换, 可以表示成

$$Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \varepsilon$$

采用单方程模型的估计方法, 得到 $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ 的估计值, 利用对应关系和 $a_1 + a_2 = 1$, $b_1 + b_2 = 1$, 可以计算得到关于参数 $A, \rho, \rho_1, m, a_1, a_2, b_1, b_2$ 的估计值。对于这些估计值, 除了级数展开时展开点的选择误差和逼近误差外, 在代入后的筛选也带来了误差。

§ A.4.6 含体现型技术进步的生产函数模型的估计

以分离资本质量的含体现型技术进步的生产函数模型(A.3.10)

$$\frac{\Delta Y}{Y} = \frac{\Delta A'}{A'} + \alpha \left(\lambda - \lambda \Delta \bar{a} + \frac{\Delta K}{K} \right) + \beta \frac{\Delta L}{L} + \varepsilon$$

为例。将模型进行变量置换, 得到

$$Z_t = \alpha_0 + a X_{1t} + \beta X_{2t} + \varepsilon_t$$

其中 Z, X_2 的样本观测值可以直接取得, 在 X_1 中, $\Delta \bar{a}, \Delta K/K$ 的样本观测值也可以直接取得, 需要给定不同的值, 进行反复估计, 以拟合效果最好者作为最后估计结果。

§ A.4.7 确定性统计边界生产函数模型的修正的普通最小二乘估计 (COLS)

修正的普通最小二乘估计 (COLS) 是 Richmand 于 1974 年首先提出的在普通最小二乘估计结果的基础上对常数项进行修正的一种估计方法, 得到了广泛的应用。

对于确定性统计边界生产函数(A.3.12)

$$Y = f(K, L, \dots) e^{-u} \quad (u \geq 0)$$

如果用 C-D 生产函数的形式表示, 则写成

$$Y = AK^\alpha L^\beta e^{-u} \quad (u \geq 0)$$

其对数形式为

$$\ln Y = \ln A + \alpha \ln K + \beta \ln L - u \quad (\text{A.4.7})$$

其中实质上的边界生产函数为

$$\ln Y' = \ln A + \alpha \ln K + \beta \ln L$$

Y' 为理论上的最大产出量。设 $\mathbb{E}(u) = \mu$, $\ln A = a$, 将(A.4.7)写成

$$\ln Y = (a - \mu) + \alpha \ln K + \beta \ln L - (u - \mu) \quad (\text{A.4.8})$$

式中 $\mathbb{E}(u - \mu) = 0$, 可以用普通最小二乘法估计模型(A.4.8), 得到

$$\ln \hat{Y} = (a - \hat{\mu}) + \hat{\alpha} \ln K + \hat{\beta} \ln L \quad (\text{A.4.9})$$

这就是我们所说的平均生产函数, 它与我们所要求的边界生产函数的差别在于常数项。如何才能求得边界生产函数的常数项 α 的估计值? 显然, 应该有

$$\hat{a} = (a - \hat{\mu}) + \hat{\mu} \quad (\text{A.4.10})$$

根据边界生产函数应该使得所有实际产出量都在它的下面的特点, 可以用

$$\text{Max} (\ln Y_i - \ln \hat{Y}_i) = \text{Max} (\ln Y_i - ((a - \hat{\mu}) + \hat{\alpha} \ln K_i + \hat{\beta} \ln L_i)) \quad (\text{A.4.11})$$

作为 $\hat{\mu}$ 的值, 代入(A.4.10)得到 $\hat{\alpha}$ 。于是所要求的边界生产函数为

$$\hat{Y}' = e^{\hat{a}} K^{\hat{\alpha}} L^{\hat{\beta}}$$

该边界生产函数即是平均生产函数(A.4.9)向上平移了 $\hat{\mu}$ 。

概括起来, 用修正的普通最小二乘法估计确定性统计边界生产函数模型, 即是首先用最小二乘法估计平均生产函数, 然后计算所有样本点的产出量的观测值与平均生产函数估计值之差, 取其最大者加到平均生产函数的常数项上, 即得到边界生产函数的常数项, 进而得到边界生产函数。

§ A.5 生产函数模型在技术进步分析中的应用

生产函数模型是对生产活动进行数量分析的有效工具, 有其广泛的应用。首先, 生产函数模型的参数具有特定的经济含义, 可以直接用于生产活动的结构分析; 生产函数模型揭示了投入要素与产出量之间的技术关系, 可以用于生产预测。生产函数模型在技术进步分析中的应用, 更是其显著的功能。下面以此为例。说明生产函数模型在技术进步分析中的应用。

技术进步的定量分析, 主要包括两方面内容: 一是测算技术进步速度及其对经济增长的贡献, 显然这是从纵向研究技术进步; 一是关于部门之间、企业之间技术进步水平的比较研究, 这是从横向研究技术进步。对于这两方面研究, 生产函数模型都是有效的方法。

§ A.5.1 技术进步速度的测定

年技术进步速度，是一项反映在一定时期内技术进步快慢的综合指标。通常用常用下式定义：

$$\gamma = y - \alpha \cdot k - \beta \cdot l \quad (\text{A.5.1})$$

其中 γ 为技术进步速度； α ， β 为资本与劳动的产出弹性； y ， k ， l 分别为产出、资本和劳动数量的增长速度。显然，在(A.5.1)中是将资本与劳动数量增长之外的所有因素全部归入“技术进步”之中。

α ， β 可以通过生产函数模型估计得到， y ， k ， l 则由样本观测值计算得到，根据(A.5.1)式就可以计算得到技术进步速度 γ 的值。李子奈曾经利用 1963—1984 年我国全民所有制工业的数据为样本，估计修正的 C-D 生产函数模型，得到如下估计结果：

$$Y = 0.6479e^{0.0128t} K^{0.3608} L^{0.6756}$$

由样本数据计算得到 1963—1984 年间全民所有制工业不变价总产值、不变价固定资产原值和劳动者人数的平均增长速度为

$$y = 10.5\%$$

$$k = 8.74\%$$

$$l = 5.71\%$$

并且由 $\alpha = 0.360$ ， $\beta = 0.6765$ ，根据(A.5.1)式计算得到 1963—1984 年间我国全民所有制工业平均技术进步速度为

$$\gamma = 3.49\%$$

§ A.5.2 技术进步对增长的贡献

技术进步对增长的贡献，是一项直接反映技术进步对增长影响的综合指标。它的定义由下式给出：

$$E_A = \frac{\gamma}{y} \times 100\% \quad (\text{A.5.2})$$

它是由(A.5.1)式的两边同除以 y 后得到的：

$$\frac{\gamma}{y} = 1 - \frac{\alpha \cdot k}{y} - \frac{\beta \cdot l}{y}$$

对于 1963—1984 年间我国全民所有制工业，计算得到技术进步对增长的贡献为

$$E_A = \frac{3.49\%}{10.5\%} \times 10\% = 33.2\%$$

当然，这是除了资本与劳动数量增长的贡献外所有因素对增长的贡献。

根据最新的计算，我国的经济增长中，技术进步的贡献一直维持在 30%~40% 之间，而一些发达国家，该项指标达到 60% 以上。

上述用于测算技术进步速度和技术进步对增长的贡献的方法存在不少问题，主要是在“技术进步”中包含的因素过于复杂。所以，积极创造条件，主要是基础数据条件，使得含体现型技术进步生产函数模型进入实际应用，将会使技术进步的定量分析水平提高一大步。

§ A.5.3 部门之间、企业之间技术进步水平的比较分析

对部门之间、企业之间的技术进步水平进行比较分析，无论是正确地评价部门或企业，或是确定技术进步的方向，都是十分重要的。边界生产函数模型是一种有效的方法。

例如，建立某行业的企业确定性统计边界生产函数模型

$$Y = AK^\alpha L^\beta e^{-u} \quad (u \geq 0)$$

进行变换后用普通最小二乘法估计得到

$$\ln \hat{Y} = (a - \hat{\mu}) + \hat{\alpha} \ln K + \hat{\beta} \ln L$$

设

$$\hat{\mu}_j = \ln Y_j - \ln \hat{Y}_j$$

定义 $(a - \hat{\mu}) + \hat{\mu}_j$ 为第 j 个样本企业的技术水平。如果第 m 个样本企业处在边界上，根据修正的最小二乘法的原理，应该有

$$\hat{\mu}_m = \text{Max} (\ln Y_i - \ln \hat{Y}_i) = \text{Max} (\ln Y_i - ((a - \hat{\mu}) + \hat{\alpha} \ln K_i + \hat{\beta} \ln L_i))$$

那么我们将该企业的技术效率定义为 1，即认为它处于最佳技术状态。这样第 j 个样本企业的技术效率为

$$E_j = e^{(a - \hat{\mu}) + \hat{\mu}_j} / e^{(a - \hat{\mu}_i) + \hat{\mu}_m} = e^{\hat{\mu}_j - \hat{\mu}_n} \quad (\text{A.5.3})$$

根据(A.5.3)式可以对同一时间截面上不同样本企业的技术水平进行定量比较。

李子奈曾经以 1985 年煤炭行业 63 个大型企业的截面数据为样本分别估计煤炭企业平均生产函数和边界生产函数模型：

$$\ln \hat{Y} = -4.3279 + 0.4074 \ln K + 0.2766 \ln L + 0.2885 \ln E \quad (\text{A.5.4})$$

$$\ln \hat{Y}' = -3.6477 + 0.4074 \ln K + 0.2766 \ln L + 0.2885 \ln E \quad (\text{A.5.5})$$

其中 \hat{Y}' 为应该达到的最大产出量的估计量, E 为钢材消耗量。可见, 平均生产函数(A.5.4)和边界生产函数(A.5.5)的差别仅在于常数项。根据(A.5.3)计算每个企业的技术效率, 结果如下:

$E_j^{1985} \geq 80\%$	2
$60\% \leq E_j^{1985} < 80\%$	9
$40\% \leq E_j^{1985} < 60\%$	43
$E_j^{1985} < 40\%$	9

即技术效率化较高的只有 2 个企业, 有 9 个企业技术效率相当低。这样就将 63 个大型企业按技术水平排了队, 无论对行业主管部门, 还是对企业, 都是有重要意义的。

§ A.6 建立生产函数模型中的数据质量问题

在建立与应用生产函数模型过程中, 有许多实际问题需要认真处理, 其中较为突出的是数据质量问题。

§ A.6.1 样本数据的一致性问题

可以作为生产函数模型样本数据的有两类: 时间序列数据和截面数据。在选择哪类数据作样本时, 需要特别注意一致性问题。正如在绪论中提及的, 计量经济学模型是通过样本估计母体的参数, 那么样本必须是从母体中随机抽取的。例如, 同行业的企业截面数据只能用于该行业企业生产函数的估计, 不可以将由此得到的生产函数用于整个行业的总量分析; 行业生产函数一般选取该行业的时间序列数据为样本, 如果一定要采用截面数据的话, 也只能采用不同国家的同行业数据, 而不能采用企业截面数据; 同样, 如果采用某一企业的时间序列数据, 只能估计该企业的生产函数, 而不能作为行业的企业生产函数应用。

这个概念是重要的。曾经有人采用全国大中型煤炭企业的截面数据, 估计生产函数模型, 然后用该模型预测未来煤炭行业的产出量。这就违反了一致性原则。也曾经有人采用某钢铁企业的时间序列数据建立了一个生产函数模型, 然后将该模型作为钢铁行业的一般企业生产函数应用。这也违反了一致性原则。

§ A.6.2 样本数据的准确性问题

在生产函数模型估计中, 经常遇到样本数据口径不一致的问题。例如, 估计我国工业生产函数, 作为被解释变量的总产值是全国口径的, 作为解释变量的固定资产原值只有独立核算工业的数据; 估计企业生产函数, 作为被解释变量的产值是生产性口径的, 作为解释变量的劳动力却包含非生产性人员; 等等。这类问题几乎普遍存在。这就违反了数据的准确性原则。处理的方法, 一是按照最小口径建立模型, 然后在应用中对全口径进行估算; 二是利用其他信息对样本数据首先进行调整, 然后再估计模型。在特殊情况下, 当生产函数模型是线

性的，或者是 C-D 生产函数，在假设不同样本点上不同口径的数据之间存在着固定比例时，采用不同口径的样本数据不影响结构参数的估计结果，只影响常数项。

§ A.6.3 样本数据的可比性问题

在生产函数模型估计中，更严重的问题是样本数据的可比性问题，而这个问题经常被忽视。主要表现在不同的样本点上，实际相同的产出量或要素投入量出现不同的观测值数据。例如，产出量用当年价格计算时，采用时间序列数据为样本，由于价格的变化，会使不同样本点上实际相同的产出量表现出相差甚大的观测值。再如，固定资产原值按固定资产形成时的价格计算，对于同行业的两个规模相同、生产工艺相同、设备技术水平相同的企业，只是因为投产时间不同，账面上的固定资产原值差别会很大，作为样本数据时，尽管在不同样本点上实际投入的固定资产数量相同，却出现了不同的观测值。诸如此类的样本数据不可比问题，会给生产函数的结构参数估计值造成很大的“失真”。

曾经以 1963—1984 年全民所有制工业总产值（1970 年不变价计算）、全民所有制工业职工人数和全民所有制工业固定资产原值（分别以形成当年价格计算和以 1970 年不变价计算）作为样本，对 C-D 生产函数的参数进行估计，得到

$$Y = 0.9181e^{0.0273t} K^{-0.1789} L^{0.7386}$$

和

$$Y = 0.6479e^{0.0128t} K^{-0.3608} L^{0.6756}$$

二者的结构参数相差甚大。其中前者是以固定资产形成当年价格计算的固定资产原值为样本观测值，后者是以 1970 年不变价计算的固定资产原值作为样本观测值。这个例子足以说明样本数据的可比性对模型质量的影响。

生产函数模型离不开固定资产，而统计中的固定资产原值是按照资产形成年的当年价格计算的，这是建立生产函数模型中的一个突出的困难。简单的调整方法是：将固定资产原值按照形成的年份分解，每一年形成的固定资产又可以分解为设备与建筑安装两部分，将每一部分按照各自的价格指数向某一基准年调整，然后再汇总为各年的固定资产原值。分解得越细，调整越准确。

人们有时对计量经济学模型提出疑问：同样的研究对象、同样的模型形式、同样的样本区间，不同的研究者得到的结果不同，那么计量经济学模型还有没有科学性？上面的例子就是一个明证。所以，说样本数据的质量关系到模型的成败，是恰如其分的。这个总是不仅在生产函数模型研究中是一个突出问题，在所有的应用模型研究中都是一个突出的问题。

附录 B

计量经济模型的类别

§ B.1 模型的类别

一般的模型是广义回归模型，即假设

$$y = F(\beta, \mathbf{X}) + \varepsilon, \quad \mathbb{E}[\varepsilon] = \mathbf{0}, \quad \mathbb{E}[\varepsilon\varepsilon'] = \sigma^2\mathbf{\Omega}$$

其中 $\mathbf{\Omega}$ 是一般的正定矩阵， $\sigma^2\mathbf{\Omega}$ 是样本的协方差矩阵。

假设 $\text{Cov}(\varepsilon_i, \varepsilon_j) = \rho_{ij}$ ，样本的协方差矩阵 $\mathbf{\Sigma}$ (the covariance matrix) 是：

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \rho_{n-1,n} \\ \rho_{n1} & \cdots & \rho_{n,n-1} & \sigma_n^2 \end{pmatrix}$$

$\sigma^2\mathbf{\Omega}$ 中应该有 $1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$ 未知的参数，再加上未知参数 β 的个数，是一个只有 n 个样本点难以完成任务的，即使完成，效率和准确性是不高的。即不简化模型我们将一事无成。

模型 1 Large-sample Model

$$\mathbf{\Sigma} = \mathbb{E}(\varepsilon\varepsilon') = \sigma^2\mathbf{I}$$

模型 2 异方差 (Heteroscedasticity)

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix}$$

即使这样，也有超过 n 个未知的参数要估计，所以，进一步假设组间异方差 (group-wise)

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_1^2 & & & \\ & & & \sigma_2^2 & & \\ & & & & \ddots & \\ & & & & & \sigma_2^2 \\ & & & & & & \ddots & \\ & & & & & & & \sigma_g^2 \\ & & & & & & & & \ddots & \\ & & & & & & & & & \sigma_g^2 \end{pmatrix}$$

模型 3 自相关 (Autocorrelation)

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & \sigma^2 & \rho & & \rho^{n-2} \\ \rho^2 & \rho & \sigma^2 & & \vdots \\ \vdots & & & \ddots & \rho \\ \rho^{n-1} & \cdots & & \rho & \sigma^2 \end{pmatrix}$$

此时我们需要估计两个参数 (σ^2, ρ) 。

模型 4 ARCH(条件异方差) 或 GARCH(广义条件异方差)

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix}$$

对于不同的观测值 σ_i^2 是不同的，但是它们彼此间也存在一定关系，如：

$$\text{ARCH: } \sigma_k^2 = a + b\sigma_{k-1}^2 \text{ 在条件 } \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ 下}$$

$$\text{GARCH: } \sigma_k^2 = a + b\sigma_{k-1}^2 + c\sigma_{k-2}^2 + \cdots$$

§ B.2 事例：社会保障水平与国内生产总值

直观上看，社会保障水平的相关因素中，最主要的因素是人均国内生产总值。只有人均国内生产总值的增长，才会有资金支撑社会保障的各项支出，我们可以建立相应的线性回归模型： $Y = a + bX + \varepsilon$

利用有关国家的数据，算出常数项 a 和系数 b ，如下：

社会保障水平与人均 GDP 增长之间的相关函数和回归方程：

表 B.1 社会保障水平与人均 GDP 增长之间的相关函数和回归方程

国家	相关系数	回归方程 $Y = a + bX$	样本年份
英国	0.956	$Y = 14.1 + 0.0034X$	1960—1995
瑞典	0.964	$Y = 10.68 + 0.0064X$	
丹麦	0.940	$Y = 10.14 + 0.0056X$	
美国	0.903	$Y = 10.46 + 0.00034X$	
日本	0.988	$Y = 7.62 + 0.00078X$	
德国	0.947	$Y = 16.37 + 0.00081X$	

¹ 世界银行，世界发展报告 (1982—1998) 北京：中国财政经济出版社
² 联合国，人类发展报告 (1982—1999) 伦敦：天津大学出版社

从统计分析结果证明了 2 点：

1、社会保障水平与人均 GDP 队长之间存在着高度相关。（相关系数在 0.94 至 0.98 之间）

2、回归方程中的自变量系数 b 值，福利型国家明显都高于自保公助型国家，上述关系表明，人均 GDP 每增长一亿本币，社会保障支出相应增长，福利型国家为 0.003% ~ 0.006%，自保公助型国家为 0.0003% ~ 0.0008%，二者相差一个小数点，从而说明，在相同人均国内生产总值增长速度下，福利型国家社会保障水平的上升速度快于自保公助型国家。

问题：从样本到母体的的检验还没有完成！

§ C.1 经济学家预测奥运奖牌

距离 2004 年雅典奥运会开幕还有两个多星期。与往常一样，这次体育盛会的奖牌争夺战成为关注焦点。奥运奖牌如何分配不仅是体育界人士争论不休的问题，经济学家也不甘落后。美国两位经济学家通过分析各国人口，经济总产值以及人均收入等经济因素，预测出本届奥运会的奖牌分配情况，引起广泛兴趣。

§ C.1.1 经济学家无需分析运动员实力

体育界人士在预测奥运会奖牌分配情况时，往往侧重于各国运动员的身体和心理素质，主要强调运动员的实力。美国达特茅斯学院商学院教授伯纳德和加利福尼亚大学伯克利分校商学院客座教授布斯则表示，他们无需了解各国运动员的实力，就能预测出本届奥运会的奖牌分配情况。

美国之音报道，这两位经济学家所依靠的主要是经济数据。他们认为，每个国家在奥运会上的表现与其经济实力密切相关，人口和人均收入状况在奥运会奖牌争夺中发挥着极其重要的作用。

两位经济学家在联名发表的论文中说，人口是判断一个国家奥运实力的重要指标，因为一般来说，一个国家的人口越多，它的综合体育实力就应该越强。

国民经济总产值和人均收入状况也很重要，富裕国家可以动员更多的社会资源来发展体育运动，提高在奥运会的竞争水平。当然，除了考察人口和资源等经济因素之外，伯纳德和布斯的研究模式还注重每个国家在以往奥运会上的表现。

历史记录弥补了经济数据分析方面的不足，说明了为什么古巴和俄罗斯这样的国家在人口和经济实力上并不占优势，但是却能经常在奥运会上取得好成绩。

伯纳德和布斯的经济分析模式准确地预测出韩国在 2000 年悉尼奥运会获得奖牌的数量。对其他各主要国家的奖牌预测误差平均在四枚以内。

§ C.1.2 美俄中占据前三位

两位经济学家对这次雅典奥运会的奖牌分布情况做出如下的预测：美国仍保持奖牌总数第一名，总共将获得九十三枚；其次是俄罗斯，八十三枚；中国五十七枚；德国五十五枚；澳大利亚五十四枚；法国三十七枚；意大利三十三枚。

§ C.1.3 奖牌分布趋势

伯纳德和布斯还预测，自上世纪 80 年代开始出现的富国获得奥运奖牌的比例不断缩小、穷国获得奖牌总数的比例不断扩大的趋势还会在本届奥运会上继续下去。

伯纳德说，他这一发现和全球经济的发展趋势相符合，一些发展中国家的生活水平得到改善，从而也扩大了这些国家分享奥运光荣的机会。中国的情况最能说明两位经济学家的经济分析模式。随着中国经济起飞，中国在奥运会上获得的奖牌数量从 1988 年的二十八枚猛增到 2000 年的五十九枚。

伯纳德和布斯预测中国将在本届奥运会上获得五十七枚奖牌，但是伯纳德表示，中国的奖牌总数很可能超出他们的预测，因为中国将举办 2008 年奥运会，未来东道主的效应可能会使中国运动员超常发挥。

§ C.1.4 最后结果

表 C.1 2004 年雅典奥运会奖牌榜

国家	(金)	(银)	(铜)	最后结果	经济学家的预测结果
1 美国	35	39	29	103	93
2 中国	32	17	14	63	57
3 俄罗斯	27	27	38	92	83
4 澳大利亚	17	16	16	49	54
5 日本	16	9	12	37	
6 德国	14	16	18	48	55
7 法国	11	9	13	33	33
8 意大利	10	11	11	32	37
9 韩国	9	12	9	30	
10 英国	9	9	12	30	
11 古巴	9	7	11	27	
12 乌克兰	9	5	9	23	

13 匈牙利	8	6	3	17
14 罗马尼亚	8	5	6	19
15 希腊	6	6	4	16

§ C.2 学生邮件

尊敬的蒋岳祥老师：

您好! 我叫是 2003 级博士生，曾上过您的计量课。我现在遇到一个计量方面的问题，希望能得到您的指点。

最近我在研究城市化与农业产值的关系时，通过构建模型得到如下的公式：

$$x_1 = B_1 (\rho - z_u)^\alpha z_u^{1-\alpha} \tag{C.2.1}$$

其中 x_1 为人均农业产值，而 z_u 代表城市化率， B_1 ， ρ ， α 均为参数。

为简化起见，令 $\alpha = 0.5$ ，且我们用的是 1978 年-2003 年 26 年的时间序列数据，拟用扣除时间趋势的 OLS 方法进估计（Wooldridge，中文版，2003，pp.318-325），因此得到公式如下：

$$x_1^2 = \beta_0 + \beta_1 z_u + \beta_2 z_u^2 + \beta_3 t + \beta_4 t^2 + \mu_t \tag{C.2.2}$$

数据¹如下：

表 C.2 城市化与农业产值

年份	人均农业产值（元） x_1	城市化率 (%) z_u	年份	人均农业产值（元） x_1	人均农业产值（元） z_u
1978	116.09	0.18	1991	207.92	0.27
1979	133.21	0.19	1992	211.77	0.27
1980	136.28	0.19	1993	218.64	0.28
1981	147.67	0.20	1994	246.63	0.29
1982	162.67	0.21	1995	275.55	0.29
1983	175.89	0.22	1996	292.82	0.30
1984	193.78	0.23	1997	294.25	0.32
1985	184.84	0.24	1998	307.77	0.33
1986	189.86	0.25	1999	311.69	0.35
1987	198.46	0.25	2000	308.87	0.36
1988	191.24	0.26	2001	322.30	0.38
1989	178.88	0.26	2002	334.99	0.39
1990	208.63	0.26	2003	331.90	0.41

经运算后，得到结果如下：

表 C.3 人均农业产值与城市化的关系

	人均农业产值平方 x_1^2	
	(1)	(2)
城市化率 z_u	-641046 (-1.33)	2685430 (3.63)
城市化率平方 z_u^2	1162719 (2.05)	-4241057 (-3.71)
一次时间趋势 t	3834.87 (2.47)	-9771.19 (-3.36)
二次时间趋势 t^2		429.47 (5.04)
常数项 Constant	87623.44 (1.24)	-316236.2 (-3.37)
R^2	0.9466	0.9758
\bar{R}^2	0.9393	0.9712
F	130.03	211.87
样本数 Observations	26	26

由上可知，在扣除了二次时间趋势后，我们得到了比较满意的结果。

请问，这种估计方法用在这里合适吗？是否要用到非线性计量分析的方法？如何进一步改进？谢谢您的指导。

此致

敬礼！

于浙大玉泉
2004 年 11 月 22 日

¹数据来源：中国统计年鉴 2001 年、2004 年，其中人均农业产值为实际值

§ C.3 2005 年计量经济学作业参考答案

§ C.3.1 2005 年计量经济学作业参考答案一

1. 用 X 解释 Y , 其中 $X, Y \sim$ 二元正态分布函数, 求出回归方程

$$\because (X, Y) \sim N(\mu_1, \mu_2, \rho, \sigma_1^2, \sigma_2^2)$$

$$\therefore f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \cdot \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}$$

$$\text{则 } f_{Y|X=x}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{1}{2\pi\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2\sigma_2^2(1-\rho^2)} \cdot \left[y - \mu_2 - \rho\frac{\sigma_2}{\sigma_1} \cdot (x - \mu_1) \right]^2 \right\}$$

$$\sim N\left(\mu_2 - \rho\frac{\sigma_2}{\sigma_1} \cdot (x - \mu_1), \sigma_2^2(1-\rho^2)\right)$$

$$\text{故 } \mathbb{E}(Y | X = x) = \mu_2 - \rho\frac{\sigma_2}{\sigma_1} \cdot (x - \mu_1)$$

$$\text{即 } \mathbb{E}(Y | X) = \mu_2 - \rho\frac{\sigma_2}{\sigma_1} \cdot (x - \mu_1) \text{ 为其回归方程}$$

2. In the country of Wrknam, the velocity of money is constant. Real GDP grows by 5% per year, the money stock grows by 14% per year, and the nominal interest rate is 11 percent. What is the real interest rate?

Solution:

$$V = \frac{\text{Nominal GDP}}{M} \text{ by definition}$$

$$\because \frac{\Delta M}{M} = 14\%, \frac{\Delta V}{V} \text{ is constant}$$

$$\therefore \frac{\Delta \text{Nominal GDP}}{\text{Nominal GDP}} = \frac{\Delta M}{M} = 14\%$$

$$\text{Thus, inflation rate} = \text{Nominal GDP} - \text{Real GDP} = 14\% - 5\% = 9\%$$

$$\text{Then, real interest rate} = 11\% - 9\% = 2\%$$

§ C.3.2 2005 年计量经济学作业参考答案二

1. 构造 $\mu_1 = \rho\mu_2$ 的估计量进行检验。其中 $X_1, X_2 \cdots X_m$ 来自正态总体 $X \sim N(\mu_1, \sigma^2)$, $Y_1, Y_2 \cdots Y_n$ 来自正态总体 $Y \sim N(\mu_2, \sigma^2)$, X 、 Y 独立, 均值 \bar{X} , \bar{Y} 。

解: $X \sim N(\mu_1, \sigma^2)$, $Y \sim N(\mu_2, \sigma^2)$

$$\bar{X} - \rho\bar{Y} \sim N\left(\mu_1 - \rho\mu_2, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\rho^2\right)$$

$$\therefore \frac{(\bar{X} - \rho\bar{Y}) - (\mu_1 - \rho\mu_2)}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}\rho^2}} \sim N(0, 1)$$

$$\text{同时, 我们有 } \frac{ms_1^2}{\sigma^2} \sim \chi^2(m-1), \frac{ns_2^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\text{令 } M = \frac{ms_1^2}{\sigma^2} + \frac{ns_2^2}{\sigma^2} \sim \chi^2(m+n-2)$$

$$\text{整理有 } \frac{(\bar{X} - \rho\bar{Y}) - (\mu_1 - \rho\mu_2)}{\sqrt{ms_1^2 + ns_2^2}} \sqrt{(m+n-2) \left(\frac{mn}{m+n\rho^2} \right)} \sim t(m+n-2)$$

2. 假设对于同一个参数 θ , 你有 n 个相互独立的无偏估计量 $\hat{\theta}_1, \cdots, \hat{\theta}_n$, 它们的方差分别为 v_1, \cdots, v_n 。那么什么样的线性组合 $\hat{\theta} = c_1\hat{\theta}_1 + \cdots + c_n\hat{\theta}_n$ 是 θ 的最小方差无偏估计量?

$$\text{解: } \mathbb{E}(\hat{\theta}) = \left(\sum_{i=1}^n c_i \right) \theta = \theta$$

$$\therefore \sum_{i=1}^n c_i = 1 \cdots \cdots (1)$$

$$\text{同时 } \text{Var}(\hat{\theta}) = \sum_{i=1}^n c_i^2 v_i \cdots \cdots (2)$$

问题转化为在条件 (1) 下, 求 (2) 的最小值

利用拉格朗日函数, 求解

$$L(c_1 \cdots c_n, \lambda) = \sum_{i=1}^n c_i^2 v_i - 2\lambda \sum_{i=1}^n (c_i - 1)$$

$$\frac{\partial L(c_1 \cdots c_n, \lambda)}{\partial c_i} = 2v_i c_i - 2\lambda = 0 \cdots \cdots (3)$$

$$\frac{\partial L(c_1 \cdots c_n, \lambda)}{\partial \lambda} = \sum_{i=1}^n c_i - 1 = 0 \cdots \cdots (4)$$

$$\text{由 (3) 得: } c_i = \frac{\lambda}{v_i}, \text{ 代入 (4) } \rightarrow \lambda = \frac{1}{\sum_{i=1}^n v_i^{-1}}, c_i = \frac{1}{v_i \sum_{i=1}^n v_i^{-1}}$$

$$\text{从而 } \text{Var}(\hat{\theta}) = \sum_{i=1}^n c_i^2 v_i = \frac{1}{\left(\sum_{i=1}^n v_i^{-1}\right)^2} \sum_{i=1}^n \frac{1}{v_i^2} v_i = \frac{1}{\sum_{i=1}^n v_i^{-1}} \leq \frac{1}{\frac{1}{v_i}} = v_i$$

即说明 $\hat{\theta}$ 的方差比任何一个估计量 $\hat{\theta}_i$ 的方差小

§ C.3.3 2005 年计量经济学作业参考答案三

1. 验证 $SSE/\sigma^2 \sim \chi^2(n-2)$ 不需要 $\beta = 0$ 的假设条件即成立。

Proof C.1

$$\frac{SSE}{\sigma^2} = \left(\frac{Y}{\sigma}\right)' (M_0 - S_{xx}CC') \left(\frac{Y}{\sigma}\right)$$

可以验证 $S_{xx}CC'$ 是幂等矩阵:

$$S_{xx}CC'S_{xx}CC' = S_{xx}^2C(C'C)C' = S_{xx}CC'$$

且 $r(S_{xx}CC') = \text{tr}(S_{xx}CC') = S_{xx} \sum_i C_i^2 = 1$ 易验证 $M_0 - S_{xx}CC'$ 也是幂等矩阵

$$\begin{aligned} (M_0 - S_{xx}CC')^2 &= M_0 - S_{xx}CC'M_0 - S_{xx}M_0CC' + S_{xx}CC' \\ &= M_0 - S_{xx}CC' + \frac{1}{n}S_{xx}CC'ii' + \frac{1}{n}S_{xx}ii'CC' = M_0 - S_{xx}CC' \end{aligned}$$

且 $r(M_0 - S_{xx}CC') = \text{tr}(M_0 - S_{xx}CC') = n - 2$, 所以, $\frac{SSE}{\sigma^2} = \left(\frac{\alpha i + \beta x + \varepsilon i}{\sigma}\right)' (M_0 - S_{xx}CC') \left(\frac{\alpha i + \beta x + \varepsilon i}{\sigma}\right)$

又

$$\begin{aligned} (M_0 - S_{xx}CC')(\alpha i + \beta x) &= \alpha M_0 i - \alpha S_{xx}CC' i + \beta M_0 x - \beta S_{xx}CC' x \\ &= \alpha \left(I - \frac{1}{n}ii'\right) i - \alpha S_{xx}CC' i + \beta \left(I - \frac{1}{n}ii'\right) x - \beta S_{xx}CC' x \end{aligned}$$

$$\because \frac{1}{n}ii' i = i, C' i = 0, \therefore \alpha \left(I - \frac{1}{n}ii'\right) i = \alpha S_{xx}CC' i = 0$$

$$\beta \left(I - \frac{1}{n}ii'\right) x = \beta \left(x - \frac{1}{n}ii' x\right) = \beta \left(x - \frac{1}{n}in\bar{x}\right) = \beta(x - i\bar{x})$$

$$C' x = \frac{\sum (x_i - \bar{x}) x_i}{S_{xx}} = \frac{\sum (x_i - \bar{x}) (x_i - \bar{x})}{S_{xx}} = 1$$

$$\therefore \beta S_{xx}CC' x = \beta S_{xx}C = \beta(x - i\bar{x})$$

$$\therefore (M_0 - S_{xx}CC')(\alpha i + \beta x) = 0$$

$$\left(\frac{Y}{\sigma}\right)' (M_0 - S_{xx}CC') \left(\frac{Y}{\sigma}\right) = \left(\frac{\varepsilon}{\sigma}\right)' (M_0 - S_{xx}CC') \left(\frac{\varepsilon}{\sigma}\right)$$

故 $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$ 且与 $\beta = 0$ 无关

§ C.3.4 2005 年计量经济学作业参考答案四

1. 对于线性统计模型 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 假设 $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, $n = 13, k = 3$, 最小化误差平方和 $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ 得到如下线性方程组:

$$\begin{cases} b_1 + 2b_2 + b_3 = 3 \\ 2b_1 + 5b_2 + b_3 = 9 \\ b_1 + b_2 + 6b_3 = -8 \end{cases}$$

(1) 把这个方程组写成矩阵的形式, 并利用矩阵方法求最小二乘估计量 \mathbf{b} 的值。

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 1 \\ 1 & 1 & 6 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 9 \\ -8 \end{pmatrix} \Rightarrow \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ -2 \end{pmatrix}$$

(2) 如果 $\mathbf{y}'\mathbf{y} = 53$, 求 σ^2 的无偏估计量 s^2 的值。

$$\sigma^2 \text{ 的无偏估计量 } s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} = \frac{\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{n-k} = \frac{19}{10} = 1.9$$

(3) 求 \mathbf{b} 的协方差矩阵。

$$[\mathbf{b}] = s^2 (\mathbf{X}'\mathbf{X})^{-1} = \frac{19}{10} \begin{bmatrix} \frac{29}{4} & \frac{-11}{4} & \frac{-3}{4} \\ \frac{-11}{4} & \frac{5}{4} & \frac{1}{4} \\ \frac{-3}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

(4) 分别写出能够检验 $H_0: \beta_k = \beta_k^0$ 的 t 统计量 ($k = 1, 2, 3$)。

$$\begin{aligned} t_1 &= \frac{b_1 - \beta_1^0}{\sqrt{s^2 s_{11}}} = \frac{3 - \beta_1^0}{\sqrt{1.9 \times \frac{29}{4}}} \sim t(10) \\ t_2 &= \frac{b_2 - \beta_2^0}{\sqrt{s^2 s_{22}}} = \frac{1 - \beta_2^0}{\sqrt{1.9 \times \frac{5}{4}}} \sim t(10) \\ t_3 &= \frac{b_3 - \beta_3^0}{\sqrt{s^2 s_{33}}} = \frac{-2 - \beta_3^0}{\sqrt{1.9 \times \frac{1}{4}}} \sim t(10) \end{aligned}$$

(5) 写出能够检验 $H_0: \beta_1 + \beta_2 - 2\beta_3 = q$ 的 t 统计量和 F 统计量。记 $\mathbf{Rb} = q$, $\mathbf{R} = (1 \quad 1 \quad -2)$

$$\begin{aligned} F &= \frac{(\mathbf{Rb} - q)' [\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - q)}{s^2 J}, J = 1 \\ &= \frac{(8 - q)' \frac{1}{6} (8 - q)}{1.9} = \frac{(8 - q)^2}{11.4} \sim F(1, 10) \\ t &= \sqrt{F} = \frac{(8 - q)}{\sqrt{11.4}} \sim t(10) \end{aligned}$$

§ C.3.5 总结

$$\text{模型} = \begin{cases} \text{非线性回归模型} \\ \text{线性回归模型} \begin{cases} \text{普通最小二乘估计 (OLS)} \\ \text{最大似然估计 (MLE)} \end{cases} \end{cases}$$

最小二乘法的有限样本特性

古典回归模型的基本假设是：

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

II. \mathbf{X} 是秩为 K 的 $n \times K$ 非随机矩阵。

III. $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ 。

IV. $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2 \mathbf{I}$ 。

V. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$

未知参数 $\boldsymbol{\beta}$ 和 σ^2 的最小二乘估计量是

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}, \quad s^2 = \frac{\mathbf{e}'\mathbf{e}}{(n-K)}$$

通过分析

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

并且

$$s^2 = \frac{\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}}{n-K}$$

我们可得下列精确的有限样本结果：

1. $\mathbb{E}[\mathbf{b}] = \boldsymbol{\beta}$ (最小二乘估计是无偏的)

2. $\text{Var}[\mathbf{b}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$

3. 任意函数 $\mathbf{r}'\boldsymbol{\beta}$ 的最小方差线性无偏估计量是 $\mathbf{r}'\mathbf{b}$ 。(这就是高斯—马尔科夫定理)

4. $\mathbb{E}[s^2] = \sigma^2$

5. $\text{Cov}[\mathbf{b}, \mathbf{e}] = \mathbf{0}$

为了构造置信区间和检验假设，我们根据正态分布的假设 V 推导额外的结果，即

6. \mathbf{b} 和 \mathbf{e} 在统计上是相互独立的。相应的， \mathbf{b} 和 s^2 无关并在统计上相互独立。

7. \mathbf{b} 的精确分布依赖于 \mathbf{X} ，是 $N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$ 。

8. $(n-K)s^2/\sigma^2$ 的分布是 $\chi^2(n-K)$ 。 s^2 的均值是 σ^2 ，方差是 $2\sigma^4/(n-K)$ 。

9. 根据 6 至 8 结果, 统计量 $t(n-K) = \frac{b_k - \beta_k}{s^2(X'X)^{-1}_{kk}}^{-1}$ 服从自由度为 $n-K$ 的 t 分布。

10. 用于检验一组 J 个线性约束 $R\beta = q$ 的检验统计量

$$\frac{(Rb - q)' \left[R(X'X)^{-1} R' \right]^{-1} (Rb - q) / J}{e'e / (n - K)} = \frac{(Rb - q)' \left[Rs^2 (X'X)^{-1} R' \right]^{-1} (Rb - q)}{J}$$

服从自由度为 J 和 $n-K$ 的 F 分布。

注意, 利用 I 至 V 建立起来的 b 的各种性质和根据扰动项更进一步的正态分布假设而得到的额外推断结果之间的区别。第一组中最重要的结果是高斯—马尔科夫定理, 它与扰动项的分布无关。根据正态分布假设得到的重要的附加结果是 7、8、9、10。正态性没有产生任何额外的有限样本的最优性结果。