《中级计量经济学》

蒋岳祥 教授

目录

第一章 引言 3

第二章 矩阵 及其二次型 15

第三章 分布函数、数学期望与方差 31

第四章 数理统计 48

第五章 古典线性回归模型 65

第六章 多元线性回归模型 83

第七章 带线性约束的多元线性回归模型及其假设检验 104

第八章 正态线性模型的最大似然估计 114

第九章 非线性回归模型 117

第十章 古典线性回归的大样本理论 121

第十一章 非球形扰动项与广义最小二乘 132

第一章　引言

1.1 什么是计量经济学？

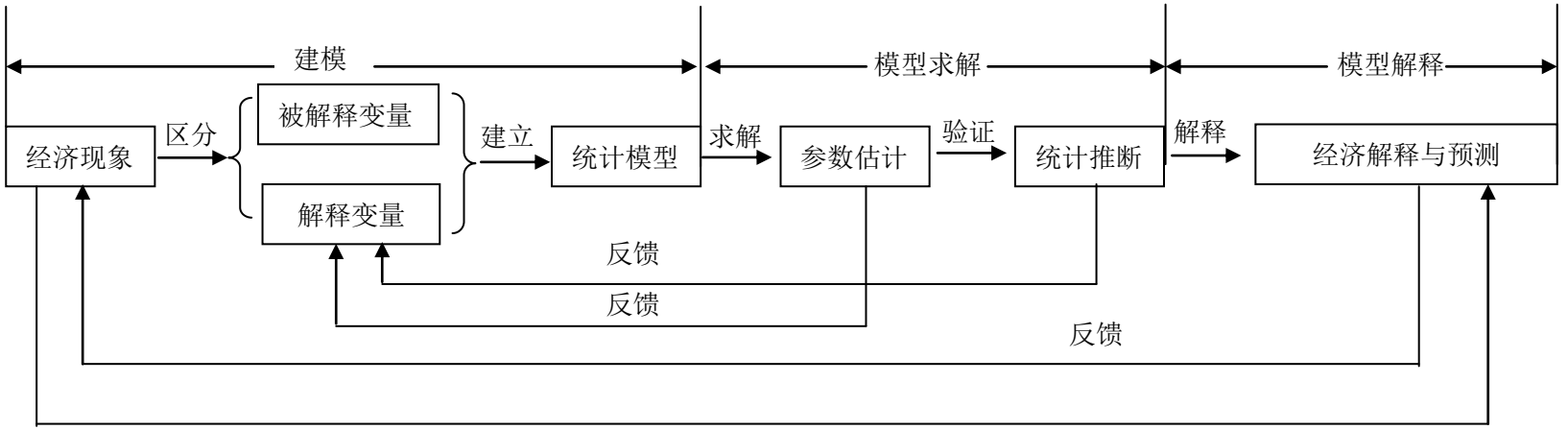
计量经济学是由挪威经济学家 R.Fisher 在三十年代首先创立的一门学科，是关于运用统计方法测量经济关系的艺术与科学，已经成为现代经济学的重要组成部分之一。

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：

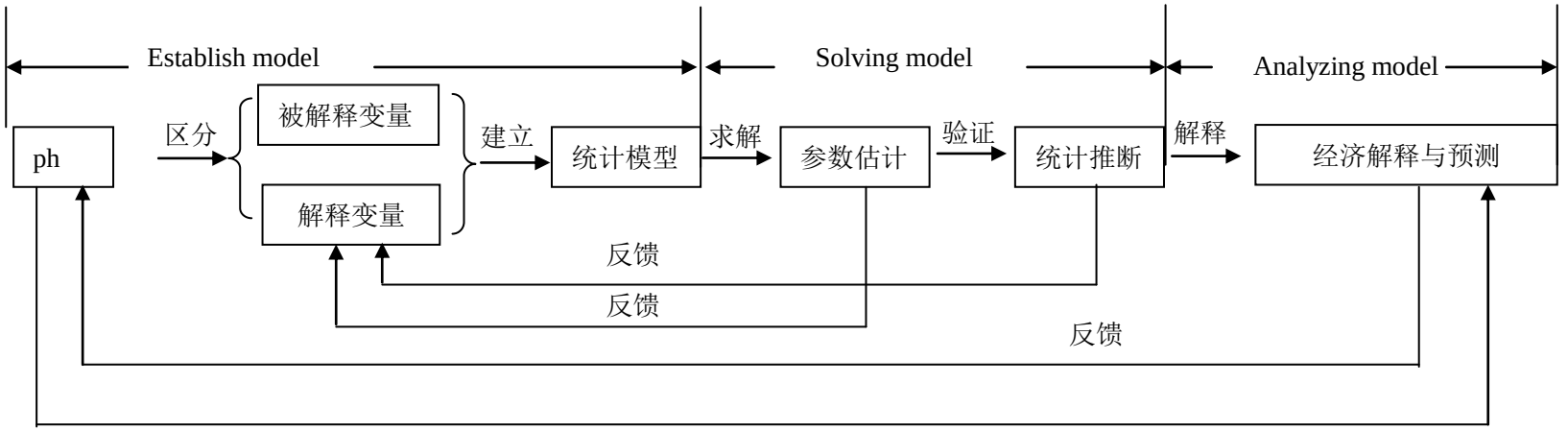
计量经济学是这样一门学科，它根据以往历史的经济资料与数据，从经济理论出发，运用数理统计的分析方法对经济关系建立经济计量模型，并依据所建立的模型对经济系统进行结构分析，经济预测和政策评价。所以计量经济学涉及数学学科中的统计学领域和经济学领域，统计学与经济理论是计量经济学的两块基石。

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，故一般建立计量经济模型的过程与方法是：

计量经济模型建立，求解，解释过程图



The flow chart of establishing, solving and analyzing econometrics model



1.2 计量经济模型(Econometric Modeling)

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：

Every one who has studied The **Keynes'** Economic theory should know, in theory, there's a close relationship between consumption and income. If we denote consumption as C, income as Y, Then a consumption fn. can be used to show the relationship between C & Y.

C=f（Y） （1）

这样的函数满足：

Fn. like this satisfies:

1）边际消费倾向（MPC）$\frac{dC}{dY}$ 位于 0 和 1 之间，即 $0<\frac{dC}{dY}<1$；

2）平均消费倾向（APC）$\frac{C}{Y}$ 是随着收入的增加而减少。

我们不妨将第二个条件作些化解，这个条件用数学语言表示是：$\frac{d\left(\frac{C}{Y}\right)}{dY}<0$，

而 $\frac{d\frac{C}{Y}}{dY}=\frac{d\left(C\cdot\frac{1}{Y}\right)}{dY}=\frac{dC}{dY}\cdot\frac{1}{Y}-\frac{1}{Y^2}C$

$$=\frac{1}{Y}\cdot\left(\frac{dC}{dY}-\frac{C}{Y}\right)=\frac{1}{Y}(MPC-APC)<0$$

即 MPC＜APC。

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 ε 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：

$$C=f(Y,\varepsilon) \qquad (2)$$

其中 ε 是随机扰动。

满足凯恩斯条件的 $f(Y\cdot\varepsilon)$ 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。

[例 1]线性模型(Linear Model)

方程（2）的一个最简单的情况，是 C 与 Y 的线性关系，即

$C=\alpha+\beta Y+\varepsilon$ （3）

其中 $0<\beta<1$，$\alpha>0$

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，……N，于是我们有如下一组方程：

$C_1=\alpha+\beta Y_1+\varepsilon_1$

$C_2=\alpha+\beta Y_2+\varepsilon_2$

……

$C_N=\alpha+\beta Y_N+\varepsilon_N$

这便是典型的一元线性回归模型。

[例 2]非线性模型(Nonlinear Model)

一般情况下，方程（2）都是非线性的情况。例如：

$C=\alpha+\beta Y^{\nu}+\varepsilon$，　其中 $0<\beta<1$，$\alpha>0$

显然，当ν=1 时，它就是例 1 的情况。$MPC=\frac{dC}{dY}=\beta\nu Y^{\nu-1}$ 而 $APC=\frac{C}{Y}=\frac{\alpha}{Y}+\beta Y^{\nu-1}$，$APC-MPC=\frac{\alpha}{Y}+(\beta-\beta\nu)Y^{\nu-1}=\frac{\alpha}{Y}+\beta(1-\nu)Y^{\nu-1}$，现在我们假设 $0<\nu<1$ 则，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。

其他实例

1、社会保障水平与国内生产总值

直现上看，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：

国家	相关系数 Y	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	1960—1995
日本	0.988	Y=7.62+0.00078X	
德国	0.947	Y=16.37+0.00081X	

资料来源：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联合同，人类发展报告，（1982—1999）伦墩：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。

2、失业、国内生产总值 GDP 与奥肯定理（Okun’s Law）

失业与实际 GDP 之间的负相关关系，首先被奥肯发现，称之为奥肯定理。

利用美国 1951 年至 1997 年的经济数据，发现：

实际 GDP 变动的百分比=3%—2 x 失业率的变动。

如果失业率保持不变，实际的 GDP 增长 3%左右，这种正常的增长是由于人口增长、资本积累和技术进步引起的。此外，失业率每上升一个百分点，实际 GDP 一般减少两个百分点。因此，如果失业率从 6%上升到 8%，那么，实际 GDP 的增长将是：

实际 GDP 变动的百分比=3%—2（8%—6%）=—1%。奥肯定理说明了，在这种情况下，GDP 将在原有的基础上下降 1%，表明经济处于衰退中。

3、带技术进步 μ 的 Solow 模型

假定生产函数为希克斯（Hicks）中性技术进步条件下的产出增长型函数，其一般形式 Solow 模型为：

$$Y = A(t) f(L, K) \qquad (1)$$

对 A（t）作进一步假定，令 $A(t)=A_0e^{\mu t}$，这里 A_0 为基本的技术水平，μ 表示由于技术进步而使产出增长的部分，称为技术进步增长率。于是（1）式变为：

$$Y=A_0e^{\mu t}f(L,K) \qquad (2)$$

对（2）式两边取对数并求导得到：

$$\frac{1}{Y}\frac{dY}{dt}=\mu+\frac{L}{Y}\frac{\partial Y}{\partial L}\frac{1}{L}\frac{dL}{dt}+\frac{K}{Y}\frac{\partial Y}{\partial K}\frac{1}{K}\frac{dK}{dt} \qquad (3)$$

由于 Y、L、K 的实际数据都是离散的，故对（3）进行离散化，并令 $\Delta t=1$ 年，于是有：

$$\frac{\Delta Y}{Y}=\mu+\alpha\cdot\frac{\Delta L}{L}+\beta\cdot\frac{\Delta K}{K} \qquad (4)$$

α 表示产出的劳动力弹性，β 表示产出的资本弹性。于是（4）式实际上就是我们的科技进步贡献率的测算模型，注意到：

$$1=\frac{\mu}{\Delta Y/Y}+\alpha\frac{\Delta L/L}{\Delta Y/Y}+\beta\frac{\Delta K/K}{\Delta Y/Y}$$

这里 $\frac{\mu}{\Delta Y/Y}$ 表示科技进步对产出增长的贡献率，$\alpha\frac{\Delta L/L}{\Delta Y/Y}$ 表示劳动力增长对产出增长的贡献率，$\beta\frac{\Delta K/K}{\Delta Y/Y}$ 表示资本增长对产出增长的贡献率。从而有：

$$\frac{\mu}{\Delta Y/Y}=1-\alpha\frac{\Delta L/L}{\Delta Y/Y}-\beta\frac{\Delta K/K}{\Delta Y/Y} \qquad (5)$$

（5）式就给出了技术进步贡献率的测算公式。

通过假定一定规模报酬不变，即 $\alpha+\beta=1$ 这一条件，比较合理有效地预防或克服了变量间可能出现的共线性。由（4）式，根据 $\beta=1-\alpha$，有：

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

设 $D_1=\frac{\Delta Y}{Y}-\frac{\Delta L}{L}, D_2=\frac{\Delta K}{K}-\frac{\Delta L}{L}$，则有：

$$D_1=\mu+\beta\cdot D_2 \qquad (6)$$

一般来讲，只要 D_1 序列不存在异方差性，（6）式就是测算科技进步增长率 μ 所用的最终模型。

1．4 回归的本质

设随机变量 $X,Y,X^{\tau}=(X_1,...,X_m)$ 是 m 维随机向量，它是可以预先测量的，希望通过 X 预测 Y，也就是说要寻找一个函数 $y=M(x_1,...,x_m)$ 当 X 的观察值为 x 时，就把 $M(x)$

作为对 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,即

$$E[Y-M(X)]^2=\min_L E[Y-L(X)]^2 \tag{1}$$

某中 min 是对一切 x 的(可测)函数 L(x)取极小,对此有

定理 1 当$M(X)$取作为条件数学期望

$$M(X)=E[Y/X] \tag{2}$$

时,使得(1)式成立,即

$$\sigma^2_{Y,X}\triangleq E[Y-E[Y/X]]^2=\min_L E[Y-L(X)]^2 \tag{3}$$

且$M(X)$与Y具有最大相关,即

$$\rho(Y,M(X))=\max_L \rho(Y,L(X)) \tag{4}$$

[证明](仅对连续型情形给出)

设(X,Y)的分布密度是$f(x,y),X$的边际分布密度是$f_1(x),Y$关于X的条件分布密度是

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

则Y关于X的条件期望是

$$M(x)\triangleq E(Y/x)=\int yf(y/x)dy$$

$$E[Y-L(X)]^2=\iint\ldots\int[y-L(x)]^2 f(x,y)dxdy$$

$$=\iint\ldots\int[y-M(x)]^2 f(x,y)dxdy$$

$$+2\iint\ldots\int[y-M(x)][M(x)-L(x)]f(x,y)dxdy$$

$$+\iint\ldots\int[M(x)-L(x)]^2 f(x,y)dxdy$$

由于

$$\iint\ldots\int[y-M(x)][M(x)-L(x)]f(x,y)dxdy$$

$$=\iint\ldots\int[y-M(x)][M(x)-L(x)]f(y/x)\,f_1(x)dydx$$

$$=\int\ldots\int[M(x)-L(x)]f_1(x)\left[\int yf(y/x)dy-M(x)\right]dx$$

$$=0 \tag{5}$$

因而

$$E[Y-L(X)]^2=E[Y-M(X)]^2+E[M(X)-L(X)]^2 \tag{6}$$

(6)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

$$M(X) = L(X) \qquad a.s.$$

它表示当 $L(X) \overset{a.s.}{=} M(X)$ 时, $E[Y - L(X)]^2$ 达到最小值 $E\left[Y - M(X)\right]^2$。

在统计学上，我们称 Y= $M(X) = E\left[Y / X\right]$ **为 Y 关于 X 的回归曲线**。

Step 2:

To prove $\rho(Y, M(X)) = \max_L \rho(Y, L(X))$, which is, the estimator M(X)=E(Y/X) has the highest correlation coefficient among all possible fn.s(L(X)).

Proof:

1.) Let L(X)=EY, as we have: $E(Y - L(X))^2 = E(Y - M(X))^2 + E(M(X) - L(X))^2$

$$E(Y - EY)^2 = Var(Y) = E(Y - M(X))^2 + E(M(X) - E(M(X)))^2$$
$$(\because E(M(X)) = E(E(Y / X)) = EY)$$

Then: $Var(Y) = E(Y - M(X))^2 + Var(M(X))$
$$= E[(Y - M(X)) - (EY - E(M(X)))]^2 + Var(M(X))$$
$$= Var(Y - M(X)) + Var(M(X))$$

2.) for all L(X)

$$Cov(Y, L(X)) = E[(Y - EY)(L(X) - E(L(X))]$$
$$= E\{[(Y - M(X)) + (M(X) - EY)](L(X) - E(L(X))\}$$
$$= E[(Y - M(X))(L(X) - E(L(X))] + E[(M(X) - E(M(X))(L(X) - E(L(X))]$$

$\because E[(Y-M(X))(L(X)-E(L(X))]$

$=\int_y\int_x\ldots\int(y-M(x))(L(x)-E(L(x)))f(x,y)dxdy$

$=\int\ldots\int(L(x)-E(L(x))[\int_y(y-M(x))f(y/x)dy]\,f_X(x)dx$

$\because\int_y(y-M(x))f(y/x)dy=\int y.f(y/x)dy-\int M(x).f(y/x)dy$

$=E(Y/X)-M(X)\int f(y/x)dy=M(X)-M(X).1=0$

$\therefore\int\ldots\int(L(x)-E(L(x))[\int_y(y-M(x))f(y/x)dy]\,f_X(x)dx$

$=\int\ldots\int(L(x)-E(L(x)).0.f(x)dx=0$

$\therefore Cov(Y,L(X))=E[(M(X)-E(M(X)))(L(X)-E(L(X)))]$

$=Cov(M(X),L(X))$(有何意义?)

当$L(x)=M(x)$时,有

$Cov(Y,M(X))=Cov(M(X),M(X))=Var(M(x)).$

从而有:

$\rho^2(Y,L(X))=Cov^2(Y,L(X))/Var(Y)Var(L(X))$

$=Cov^2(M(X),L(X))/Var(Y)Var(L(X))$

$=Cov^2(M(X),L(X))/Var(M(X))Var(L(X))*Var(M(X))Var(M(X))/Var(Y)Var(M(X))$

$\le\rho^2(L(x),M(X))\rho^2(Y,M(X))$

$\le\rho^2(Y,M(X))$

得证.

思考问题：$C=f(Y,\varepsilon)$与 C=f（Y）+ ε 两者间的区别？

计量经济学数学基础知识

调查

1、本科所学专业：_____属 1）理科　2）文科 。

2、请在你学过的课程中打“ √”：

1）高等数学　　2）概率论　　3）数理统计　　4）数学分析

5）线性回归分析　6）中级计量经济学　7）随机过程　　8）常微分方程

3、若将二次型 $Y = x_{11}^2 + 4x_1x_2 + 10x_1x_3 + 3x_{22}^2 + 2x_2x_3 + 13x_{33}^2$ 转化成 $Y = X'AX$ ，

则 $A = \begin{bmatrix} & \\ & \\ & \end{bmatrix}$ 。

4、若矩阵 $A = \begin{pmatrix} 1 & 2 & 5 \\ 1 & 3 & 7 \\ 2 & 5 & 13 \end{pmatrix}$ 求 A^{-1}。

5、若矩阵 $A = \begin{pmatrix} 2 & 4 & 3 \\ 4 & 8 & 6 \\ 3 & 6 & 5 \end{pmatrix}$，求 A 的秩、特征根及特征向量。

6、假设连续随机变量 Z，它的概率密度函数为

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{如果 a<x<b} \\ 0, & \text{其它} \end{cases}$$ ，求 E（Z），和 Var（Z）。

7、设 Z，Y 的联合概率密度函数为

$$f(x,y) = \begin{cases} (x+y) & 0<x<1，0<y<1 \\ 0 & \text{其它} \end{cases}$$

证明 Z 与 Y 的相关系数 $P_{ZY}=-\frac{1}{11}$。

8、如果 $Z_1,Z_2\cdots Z_n$ 是相互独立的标准正态分布，那么 $Y=Z_1^2+Z_2^2\cdots+Z_n^2$ 服从何分币？ $\frac{Z_1}{\sqrt{\frac{1}{n}\cdot Y}}$ 又服从何分币？

9、$Y=f(x)=\alpha x+\beta e^x$，请写出 $f(x)$ 在点 $x=0$ 处泰勒展开式。

10、设 $Z_1,Z_2\cdots Z_N$ 是一个随机样本，其总体分布为

$f(x)=(\theta+1)x^\theta$，0＜x＜1

（1）利用矩方法求参数 θ 的估计量；

（2）求参数 θ 的极大似然 ML 估计量。

11、对教师如何上好《高级计量经济学》的建议。

第二章　矩阵及其二次型(Matrix and its Quadratic Forms)

2.1 矩阵的基本概念与运算

一个 m×n 矩阵可表示为：

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

矩阵的加法较为简单，若 C=*A*+*B*，$c_{ij}=a_{ij}+b_{ij}$

但矩阵的乘法的定义比较特殊，若 *A* 是一个 m×n_1 的矩阵，*B* 是一个 n_1×n 的矩阵，则 *C*=*AB* 是一个 m×n 的矩阵，而且$c_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj}$，一般来讲，*AB*≠*BA*，但如下运算是成立的：

- 结合律（Associative Law）　(*AB*)*C*=*A*（*BC*）
- 分配律（Distributive Law）　*A*(*B*+*C*)=*AB*+*AC*

问题：$(A+B)^2=A^2+2AB+B^2$ 是否成立？

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α *A*=[α a_{ij}]。

矩阵 A 的转置矩阵(transpose matrix)记为 A'，是通过把 A 的行向量变成相应的列向量而得到。

显然$(A')'=A$，而且$(A+B)'=A'+B'$，

- 乘积的转置（Transpose of a production）　$(AB)'=B'A'$，$(ABC)'=C'B'A'$。
- 可逆矩阵（inverse matrix），如果 n 级方阵(square matrix)A 和 B，满足 AB=BA=I。

  则称 A、B 是可逆矩阵，显然 $A=B^{-1}$，$B=A^{-1}$。如下结果是成立的：

$(A^{-1})^{-1}=A$　　$(A^{-1})'=(A')^{-1}$　　$(AB)^{-1}=B^{-1}A^{-1}$。

2.2 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 I；

2）标量矩阵(scalar matrix)

即形如α I 的矩阵，其中α 是标量；

3）幂等矩阵(idempotent matrix)

如果矩阵 A 具有性质 $A\cdot A=A^2=A$，这样的矩阵称为幂等矩阵。

定理：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 $\vec{x}$，如有 $\vec{x}'A\vec{x}>0$（<0），则称 A 是正（负）定矩阵；如有 $\vec{x}'A\vec{x}$ ≥0（≤0），非负（非正）定矩阵。如果 A 是非负定的，则记为 A≥0；如果是正定的，则记为 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 A 是正定的，则 A^{-1} 也是正定的；

c）如果 A 是正定的， B 是可逆矩阵，则 $B'AB$ 是正定的；

d）如果 A 是一个 n×m 矩阵，且 n>m， $r(A)=m$，则 $A'A$ 是正定的， AA' 是非负定矩阵。

5）对称矩阵（symmetric matrix);

如果 $A=A'$ ，则 A 称为对称矩阵。

2.3 矩阵的迹（trace）

一个 n×n 矩阵的迹被定义为它的对角线上的元素之和，记为 $tr(A)$，则 $tr(A)=\sum_{i=1}^{n}a_{ii}$，如下结论是显然的。

1） $tr(\alpha A)=\alpha tr(A)$ （ α 是标量） 特例 $tr(I)=n$

2） $tr(A')=tr(A)$

3） $tr(A+B)=tr(A)+tr(B)$

4）$tr(AB)=tr(BA)$，特例$tr(A'A)=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}^2$

5）循环排列原则　*tr(ABCD)=tr(BCDA)=tr(CDAB)=tr(DABC)*

定理：实对称矩阵 A 的迹等于它的特征根之和。

因为 A 是实对称矩阵，故有在矩阵 C，使得$C'AC=\Lambda=\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}$，其中$CC'=I$，

所以，$\sum_{i=1}^{n}\lambda_i=tr(\Lambda)=tr(C'AC)=tr(AC'C)=tr(AI)=tr(A)$。

2.4 矩阵的秩(rank)

一个矩阵 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(A)，不加证明，我们给出如下结果：

1）$r(A)=r(A')\leqslant\min$（行数、列数）

2）$r(A)+r(B)-n_1\leqslant r(AB)\leqslant\min\ (r(A),r(B))$，其中 A、B 分别为 m×$n_1$、$n_1$×n 矩阵，特例：如果 A、B 为 n×n 矩阵，而且 AB=0，则$r(A)+r(B)\leqslant n$

3）$r(A)=r(AA')=r(A'A)$，其中A是 n×n 的方阵

4）$r(A+B)\leqslant r(A)+r(B)$

5）设A是 n×n 矩阵，且$A^2=I$，则$r(A+I)+r(A-I)=n$

6）设A是 n×n 矩阵，且$A^2=A$，则$r(A)+r(A-I)=n$

2.5 统计量的矩阵表示

向量可理解为特殊的矩阵。$\vec{i}$是一个其元素都为 1 的 n 维列向量，即$\vec{i}'$=（1，1，…，1），如果我们再假定$\vec{x}'=(x_1,x_2,\cdots,x_n)$，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，$\sum_{i=1}^{n}x_i=\vec{i}'\cdot\vec{x}$，$\sum_{i=1}^{n}x_i^2=\vec{x}'\cdot\vec{x}$，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$\frac{1}{n}\vec{i}'\vec{i}=1$，而$\vec{i}\,\vec{i}'=\begin{pmatrix}1 & 1\cdots 1\\ 1 & 1\cdots 1\\ \cdots & \cdots\ \cdots\\ 1 & 1\cdots 1\end{pmatrix}$，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{n}\vec{i}'\cdot\vec{x}$；

2）样本方差矩阵表示

易知：$\begin{pmatrix}\bar{x}\\ \vdots\\ \bar{x}\end{pmatrix}=\vec{i}\,\bar{x}=\vec{i}\cdot\frac{1}{n}\cdot\vec{i}'\cdot\vec{x}=\frac{1}{n}\vec{i}\,\vec{i}'\vec{x}$。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

方阵，从而$\begin{pmatrix}x_1-\bar{x}\\ x_2-\bar{x}\\ \vdots\\ x_n-\bar{x}\end{pmatrix}=(\vec{x}-i\bar{x})=(\vec{x}-\frac{1}{n}\vec{i}\,\vec{i}'\vec{x})=(I-\frac{1}{n}\vec{i}\cdot\vec{i}')\vec{x}\quad\triangleq\quad M^0\vec{x}$。

矩阵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：$S_n{}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}(\vec{x}-\bar{x})'(\vec{x}-\bar{x})$

$$=\frac{1}{n}\vec{x}\cdot M'^0M^0\vec{x}=\frac{1}{n}\vec{x}M_0^2\vec{x}=\frac{1}{n}\vec{x}'M_0\vec{x}。$$

定理：矩阵M^0是幂等矩阵。

2.6 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P是一数域，一个系数在数域P中的$x_1,x_2,\cdots,x_n$的二次齐次多项式

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

称为数域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（1）中，$x_ix_j(i<j)$

的系数写在$2a_{ij}$。而不简单地写成a_{ij}。

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

定义 1　设$x_1,\cdots,x_n$；$y_1,\cdots,y_n$是两组文字，系数在数域 P 中的一级关系式

$$\begin{cases} x_1 = c_{11}y_1 + c_{12}y_2 + \cdots + c_{1n}y_n \\ x_2 = c_{21}y_1 + c_{22}y_2 + \cdots + c_{2n}y_n \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n = c_{n1}y_1 + c_{n2}y_2 + \cdots + c_{nn}y_n \end{cases} \tag{2}$$

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

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

那么线性替换（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

$$a_{ji} = a_{ij}, \qquad i < j$$

由于

$$x_i x_j = x_j x_i$$

所以二次型（1）可以写成

$$\begin{aligned} f(x_1,x_2,\cdots,x_n) &= a_{11}x_1^2 + a_{12}x_1x_2 + \cdots + a_{1n}x_1x_n \\ &\quad + a_{21}x_2x_1 + a_{22}x_2^2 + \cdots + a_{2n}x_2x_n \\ &\quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\quad + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \cdots + a_{nn}x_n^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \end{aligned} \tag{3}$$

把（3）的系数排成一个$n\times n$矩阵

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \qquad (4)$$

它就称为二次型（3）的矩阵，因为$a_{ij}=a_{ji}$，i，$j=1,\cdots,n,$所以

$$A=A'$$

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。

令

$$X=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

于是，二次型可以用矩阵的乘积表示出来，

$$X'AX$$

$$=(x_1,x_2,\cdots,x_n)\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$=(x_1,x_2,\cdots,x_n)\begin{pmatrix} a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n \\ a_{21}x_1+a_{22}x_2+\cdots+x_{2n}x_n \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1+a_{n2}x_2+\cdots+a_{nn}x_n \end{pmatrix}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_ix_j$$

故 $$f(x_1,x_2,\cdots,x_n)=X'AX$$

应该看到，二次型（1）的矩阵A的元素$a_{ij}=a_{ji}$正是它的x_ix_j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

$$f(x_1,x_2,\cdots,x_n)=X'AX=X'BX$$

且$A'=A$，$B'=B$，则$A=B$。

令

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

于是线性替换（2）可以写成

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

或者

$$X=CY$$

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。

设

$$f(x_1,x_2,\cdots,x_n)=X'AX,\qquad A=A' \tag{5}$$

是一个二次型，作非退化线性替换

$$X=CY \tag{6}$$

我们得到一个$y_1,y_2,\cdots,y_n$的二次型

$$Y'BY$$

现在来看矩阵 B 与 A 的关系。

把（6）代入（5），有

$$f(x_1,x_2,\cdots,x_n)=X'AX=(CY)'A(CY)=Y'C'ACY$$

$$=Y'(C'AC)Y=Y'BY$$

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

定义 2　数域 P 上 $n\times n$ 矩阵 A，B 称为合同的，如果有数域 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 $A=(C^{-1})'BC^{-1}$；

3）传递性：由 $A_1=C_1'AC_1$ 和 $A_2=C_2'A_1C_2$ 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。

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

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

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

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

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的边际分布分别为$N(\mu_1,\sigma_1^2)$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。

如果X_1和X_2服从二元正态分布，那么在给定$X_1=x_1$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

$$f(x_2|x_1)\sim N(b,\sigma_2^2(1-\rho^2))$$

这里

$$b=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\text{（作业 1）}$$

条件均值$b=E[X_1|X_2]$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$f(x_2|x_1)=f(x_2)$。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

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

这里

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

b）多元正态分布

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

c）多元正态分布的二次型的分布

如果$X\sim N(\mu,\sum)$，那么

$$Y=(X-\mu)'\sum^{-1}(X-\mu)\sim x^2_{(n)}$$

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

2.7 幂等矩阵与二次型

1、幂等矩阵满足 $A^2=A$ 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果有：

1）幂等矩阵的特征根要么是 1，要么是零。

证明：设λ是 A 的特征根，则 $AE=\lambda E$，同时 $\lambda E=AE=A^2E=\lambda^2E$，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。

证明：$\because A^2=A\Rightarrow A(I-A)=0\Rightarrow A^{-1}A(I-A)=0\Rightarrow I=A$

即除了单位矩阵外，所有幂等矩阵是奇异的。

3）对称幂等矩阵的秩等于它的迹。

从而我们很容易知道 M^0 的秩。

因 M^0 的每个对角元素都是$1-\frac{1}{n}$，因此$tr(M^0)=n\cdot(1-\frac{1}{n})=n-1=r(M^0)$。

4）A 是幂等矩阵，则 $I-A$ 也是幂等矩阵，且秩（A）+秩（$I-A$）=n。

5）nS_n^2的服从$x^2(n-1)$分布（如果$X_i\sim N(0,I),i=1,n)$

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

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

则 M 是幂等矩阵。

2.8 微分及其矩阵的微分表示

1）微分的应用

微分的应用在经济学领域中被广泛地用来作近似计算。为了说明这种技巧如何运作，考虑一个例子。设 P 代表 GDP 平减指数，Y 代表实际 GDP，则名义 GDP 为 P×Y，于是有：

（P×Y）变动的百分比的≈（P 变动的百分比）+（Y 变动的百分比）；

同样一个比率变动的百分比近似地是分子变动的百分比减去分母变动的百分比。例如：设 Y 代表 GDP，而 L 代表人口数，则人均 GDP 为$\frac{Y}{L}$，则：

（Y/L）变动的百分比≈（Y 变动的百分比）－（L 变动的百分比）

问题 1：1）上述 2 个近似公式在什么条件下成立？

2）推导上述两个公式

3）宏观经济中，GDP 的确定由 4 个组成部分，即：GDP=C+I+G+NX。能否按如下公式计算 GDP 变动百分比：

GDP 变动的百分比≈（消费 C 变动的百分比）+（投资 I 变为的百分比）+（政府购买 G 变动的百分比）+（净出口 NX 变动百分比）。

如果不能，哪边的值较大？为什么？

问题 2:

In the country of Wiknam, the velocity of money is constant. Real GDP grows by 5 percent per year, the money stock grows by 14 percent per year, and the nominal interest rate is 11 percent . What is the real interest rate? （作业 2）

2）计量模型的推导

带技术进步 μ 的 Solow 模型

假定生产函数为希克斯（Hicks）中性技术进步条件下的产出增长型函数，其一般形式 Solow 模型为：

$$Y = A(t) f(L, K) \qquad (1)$$

对 $A（t）$作进一步假定，令 $A(t) = A_0 e^{\mu t}$，这里 A_0 为基本的技术水平，μ 表示由于技术进步而使产出增长的部分，称为技术进步增长率。于是（1）式变为：

$$Y = A_0 e^{\mu t} f(L, K) \qquad (2)$$

对（2）式两边取对数并求导得到：

$$\frac{1}{Y}\frac{dY}{dt}=\mu+\frac{L}{Y}\frac{\partial Y}{\partial L}\frac{1}{L}\frac{dL}{dt}+\frac{K}{Y}\frac{\partial Y}{\partial K}\frac{1}{K}\frac{dK}{dt} \qquad (3)$$

由于 Y、L、K 的实际数据都是离散的，故对（3）进行离散化，并令 $\Delta t=1$ 年，于是有：

$$\frac{\Delta Y}{Y}=\mu+\alpha\cdot\frac{\Delta L}{L}+\beta\cdot\frac{\Delta K}{K} \qquad (4)$$

α 表示产出的劳动力弹性，β 表示产出的资本弹性。于是（4）式实际上就是我们的科技进步贡献率的测算模型，注意到：

$$1=\frac{\mu}{\Delta Y/Y}+\alpha\frac{\Delta L/L}{\Delta Y/Y}+\beta\frac{\Delta K/K}{\Delta Y/Y}$$

这里 $\frac{\mu}{\Delta Y/Y}$ 表示科技进步对产出增长的贡献率，$\alpha\frac{\Delta L/L}{\Delta Y/Y}$ 表示劳动力增长对产出增长的贡献率，$\beta\frac{\Delta K/K}{\Delta Y/Y}$ 表示资本增长对产出增长的贡献率。从而有：

$$\frac{\mu}{\Delta Y/Y}=1-\alpha\frac{\Delta L/L}{\Delta Y/Y}-\beta\frac{\Delta K/K}{\Delta Y/Y} \qquad (5)$$

（5）式就给出了技术进步贡献率的测算公式。

通过假定一定规模报酬不变，即 $\alpha+\beta=1$ 这一条件，比较合理有效地预防或克服了变量间可能出现的共线性。由（4）式，根据 $\beta=1-\alpha$，有：

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

设 $D_1=\frac{\Delta Y}{Y}-\frac{\Delta L}{L}, D_2=\frac{\Delta K}{K}-\frac{\Delta L}{L}$，则有：

$$D_1=\mu+\beta\cdot D_2 \qquad (6)$$

一般来讲，只要 D_1 序列不存在异方差性，（6）式就是测算科技进步增长率 μ 所用的最终模型。

3、矩阵的微分

如果 $y=f(x_1,x_2,\cdots,x_n)$ 或写成 $y=f(x)$，那么梯度向量为

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

二阶偏导数矩阵为

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

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

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

同样地可得

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

如果 A 是对称矩阵，那么

$$\frac{\partial x'Ax}{\partial x} = 2Ax$$

一般地，有

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

思考题：

1、证明：$tr(AA') = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2$

2、证明矩阵 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、$Y_2 \cdots Y_m$ 的百分比变动较小 $\Delta Y_1, \cdots \Delta Y_m$

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

4. **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

$$A=\begin{bmatrix}1 & 4 & 5\\ 2 & 9 & 3\\ 8 & 9 & 6\end{bmatrix}$$

$$=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}$$

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**。

$$A=\begin{bmatrix}A_{11} & 0\\ 0 & A_{22}\end{bmatrix}$$

其中 A_{11} 和 A_{22} 都是方阵。

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：

$$A+B=\begin{bmatrix}A_{11}+B_{11} & A_{12}+B_{12}\\ A_{21}+B_{21} & A_{22}+B_{22}\end{bmatrix} \qquad (1)$$

和

$$AB=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix}$$

$$=\begin{bmatrix}A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22}\\ A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22}\end{bmatrix} \qquad (2)$$

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：

$$\begin{bmatrix}A_1\\ A_2\end{bmatrix}'\begin{bmatrix}A_1\\ A_2\end{bmatrix}=\begin{bmatrix}A_1' & A_2'\end{bmatrix}\begin{bmatrix}A_1\\ A_2\end{bmatrix}$$

$$=\begin{bmatrix}A_1'A_1+A_2'A_2\end{bmatrix} \qquad (3)$$

和

$$\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}' \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11}'A_{11} & 0 \\ 0 & A_{22}'A_{22} \end{bmatrix} \tag{4}$$

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

$$\begin{vmatrix} A_{11} & 0 \\ 0 & A_{22} \end{vmatrix} = |A_{11}| \cdot |A_{22}| \tag{5}$$

一个一般的 2×2 分块矩阵的结果为：

$$\begin{aligned} \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} &= |A_{22}| \cdot |A_{11} - A_{12}A_{22}^{-1}A_{21}| \\ &= |A_{11}| \cdot |A_{22} - A_{21}A_{11}^{-1}A_{12}| \end{aligned} \tag{6}$$

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：

$$\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix} \tag{7}$$

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：

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1}(I + A_{12}F_2A_{21}A_{11}^{-1}) & -A_{11}^{-1}A_{12}F_2 \\ -F_2A_{21}A_{11}^{-1} & F_2 \end{bmatrix} \tag{8}$$

其中

$$F_2 = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$$

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：

$$F_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$$

问题：请推倒上面的公式（5）、（6）、（7）和（8）。

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 x 开始。且令

$$A = \begin{bmatrix} n & \Sigma_i x_i \\ \Sigma_i x_i & \Sigma_i x_i^2 \end{bmatrix}$$

$$=\begin{bmatrix} i'i & i'x \\ x'i & x'x \end{bmatrix}$$

我们关心的是 A^{-1} 中的右下角元素，根据（8）中 F_2 的定义，这将是

$$F2=[x'x-(x'i)(i'i)^{-1}(i'x)]^{-1}$$

$$=\left\{x'\left[Ix-i\left(\frac{1}{n}\right)i'x\right]\right\}^{-1}$$

$$=\left\{x'\left[I-\left(\frac{1}{n}\right)i'i\right]x\right\}^{-1}$$

$$=[x'M^0x]^{-1}$$

所以，逆矩阵中的右下角值是

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

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

$$(Z'Z)^{22}=[X'X-X'i(i'i)^{-1}i'X]^{-1}$$

$$=[X'M^0X]^{-1}$$

这暗示着[Z′ Z]$^{-1}$ 的右下块，$K\times K$ 矩阵是第 jk 元素为 $\Sigma_i(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$ 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。

第三章 分布函数、数学期望与方差

本章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。

一、概率(Probability)

1、概率定义(Definition of Probability)

在自然界和人类社会中有着两类不同的现象，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了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）≥0，对一切 $A \in F$

（ii）P（Ω）=1；

（iii）若 $A_i \in$，i=1,2…，且两两互不相容（p(AB)=0），则

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

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 $P(\overline{A}) = 1 - P(A)$；

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 $P(A/B)=\dfrac{P(AB)}{P(B)}$，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：

$$P(A_1A_2\cdots A_n)=P(A_1)\cdot P(A_2/A_1)\cdot P(A_3\mid A_1A_2)\cdots P(A_n\mid A_1A_2\cdots A_{n-1})$$

其中 $P(A_1A_2\cdots A_{n-1})>0$。

3、全概率公式与贝叶斯（Bayes）公式

设事件 $A_1,A_2,\cdots,A_n\cdots\cdots$是样本空间Ω 的一个分割，即 $A_iA_j=\varphi$，$i\neq j$，而且：$\sum_{i=1}^{\infty}A_i=\Omega$。

从而 $B=\sum_{i=1}^{\infty}A_iB$，这里 A_iB 也两两互不相容。

则 $P(B)=\sum_{i=1}^{\infty}P(A_iB)=\sum_{i=1}^{\infty}P(A_i)\cdot P(B\mid A_i)$。

这个公式称为**全概率公式**。

由于

$$P(A_iB)=P(B)P(A_i\mid B)=P(A_i)P(B\mid A_i)$$

故

$$P(A_i\mid B)=\frac{P(A_i)P(B\mid A_i)}{P(B)}$$

再利用全概率公式即得

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

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，…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 $P(A_i|B)$ 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

定义　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

$P(A|B)=P(A)$

[证明]由条件概率定义

$$P(A\mid B)=\frac{P(AB)}{P(B)}=\frac{P(A)P(B)}{P(B)}=P(A)$$

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：

$\{\overline{A},B\},\{A,\overline{B}\},\{\overline{A},\overline{B}\}$

[证明]　由于

$$P(\overline{A}B)=P(B-AB)=P(B)-P(AB)$$

$$=P(B)-P(A)P(B)=P(B)[1-P(A)]$$

$$=P(\overline{A})P(B)$$

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

2）多个事件的独立性

定义　对 n 个事件 A_1，A_2，…，A_n，若对于所有可能的组合 $1\leqslant i<j<\cdots\leqslant n$ 成立着

$$\left.\begin{aligned}&P(A_iA_j)=P(A_i)P(A_j)\\&P(A_iA_jA_k)=P(A_i)P(A_j)P(A_k)\\&\cdots\\&P(A_1A_2\cdots A_n)=P(A_1)P(A_2)\cdots P(A_n)\end{aligned}\right\}$$

则称 A_1，A_2，…A_n 相互独立。

这里第一行有 $\binom{n}{2}$ 个式子，第二行有 $\binom{n}{3}$ 个式子，等等，因此共应满足

$$\binom{n}{2}+\binom{n}{3}+\cdots+\binom{n}{n}=2^n-n-1$$

个等式。

二、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

$F(x)=\int_{\infty}^{x}f(x)dX$，$f$（x）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，…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 $f(x_1, x_2,\cdots x_n)$，使

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

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

$$f(x_1,\cdots,x_n) \geqslant 0$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\cdots,x_n)dx_1\cdots dx_n = 1$$

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)。

［例］若 $F(x, y)$ 是连续型分布函数，有密度函数 $f(x, y)$，那么

$$F_1(x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(u,y)dudy$$

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

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

$f_1(x)$ 及 $f_2(y)$ 的边际分布密度函数。

［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, σ_1, σ_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。

定理：二元正态分布的边际分布仍为正态分布。

条件分布(Conditional Distribution)

离散型：若已知ξ $=x_i$，（$p1(x_i) > 0$）则事件{η $=y_i$}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ξ $=x$ 的条件下，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η =y 的条件下，ξ 的分布密度函数为

$$f(x \mid y)=\frac{f(x,y)}{f_2(y)}$$

这里当然也要求 $f_2(y)\neq 0$

定理：二元正态分布的条件分布仍然是正态分布

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

其均值 $\mu=\mu_2+\rho\frac{\sigma^2}{\sigma^1}(x-\mu_1)$ 是 x 的线性函数，这个结论在一些统计问题中很重要。

4、随机变量的独立性

定义 设ξ $_1$，…，ξ $_n$ 为 n 个随机变量，若对于任意的 x_1，…，x_n 成立

$$P\{\xi_1<x_1,\cdots,\xi_n<x_n\}=P\{\xi_1<x_1\}\cdots P\{\xi_n<x_n\} \qquad (1)$$

则称 $\xi_1,\cdots,\xi_n$ 是相互独立的。

若 ξ_i 的分布函数为 $F_i(x)$，它们的联合分布函数为 $F(x_1,\cdots,x_n)$，则（1）等价于对一切 $x_1,\cdots,x_n$ 成立

$$F(x_1,\cdots,x_n)=F_1(x_1)\cdots F_n(x_n)$$

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，（1）等价于任何一组可能取的值（$x_1,\cdots,x_n$）成立

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

对于连续型随机变量，条件（1）的等价形式是对一切 $x_1,\cdots,x_n$ 成立

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

这里 $f(x_1,\cdots,x_n)$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 $\xi_1,\xi_2,\cdots,\xi_n$ 相互独立，则其中的任意 r(2≤r＜n）个随机变量也相互

独立，例如，我们证明$\xi_1,\xi_2,\cdots,\xi_{n-1}$相互独立。

$$P\{\xi_1<x_1,\cdots,\xi_{n-1}<x_{n-1}\}=P\{\xi_1<x_1,\cdots,\xi_{n-1}<x_{n-1},\xi_n<\infty\}$$

$$=P\{\xi_1<x_1\}\cdots P\{\xi_{n-1}<x_{n-1}\}P\{\xi_1<\infty\}$$

$$=P\{\xi_1<x_1\}\cdots P\{\xi_{n-1}<x_{n-1}\}$$

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

$$G(y_1,\cdots,y_n)=P\{\eta_1<y_1,\cdots,\eta_n<y_n\}$$

$$=\int_{\substack{f_1(x_1,\cdots,x_n)<y_1\\ \cdots\\ f_m(x_1,\cdots,x_n)<y_m}}\cdots\int f(x_1,\cdots,x_n)dx_1\cdots dx_n \qquad (1)$$

若对$y_i=f_i(x_1,\cdots,x_n)$存在唯一的反函数$x_i(y_1,\cdots,y_n)=x_i,(i=1,\cdots,n)$，且$(\eta_1,\cdots,\eta_n)$的密度函数为$q(y_1,\cdots,y_n)$，那么

$$G(y_1,\cdots,y_n)=\int_{\substack{u_1<y_1\\ u_n<y_n}}\cdots\int q(u_1,\cdots,u_n)du_1\cdots du_n \qquad (2)$$

比较（1）与（2）可知

$$q(y_1,\cdots,y_n)$$

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

其中J为坐标变换的雅可比行列式(Jacobian Determinant)

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

这里，我们假定上述偏导数存在而且连续。

随机变量的函数的独立性

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

三、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

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

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

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

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，… X_n 的联合概率密度函数为 $f(x_1,x_2,\cdots,x_n)$，$Y=g(X_1,X_2,\cdots,X_n)$，那么

$$E[Y]=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}g(x_1,x_2,\cdots,x_n)f(x_1,x_2,\cdots,x_n)dx_1dx_2\cdots dx_n$$

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

（1）如果 $a_0,a_1\cdots,a_n$ 是常数，那么

$$E[a_0+a_1X_1+\cdots+a_nX_n]=a_0+a_1E[X_1]+\cdots+a_nE[X_n]$$

（2）如果 X_1，$X_2\cdots$，X_n 是相互独立的随机变量，那么

$$E[X_1X_2\cdots X_n]=E[X_1]E[X_2]\cdots E[X_n]$$

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 $E[(X-\mu)^r]$ 记为 μ_r。如果 $r=2,E[(X-\mu)^2]$ 被称为 X 的分布的方差或 X 的方差，常常记为 σ^2 或 $\mathrm{var}(X)$。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

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

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

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

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 $\mathrm{cov}(X,Y)$=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

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

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 ρ_{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 $a_0, a_1\cdots, a_n$ 是常数，X_1，$X_2\cdots$，X_n 是随机变量，那么

$$\mathrm{var}[a_0 + a_1X_1 + \cdots + a_nX_n] = \sum a_i^2\ \mathrm{var}(X_i) + 2\sum\sum_{i<j} a_ia_j\ \mathrm{cov}(X_i, X_j)$$

特别地，有

$$\mathrm{var}(a_0 + a_1X_1) = a_1^2\ \mathrm{var}(X_1)$$

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

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 X 表示随机变量组成的向量，即

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

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

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

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 X 的**协方差矩阵(Covariance Matrix)**如下

$$\mathrm{cov}(X)=E[(X-E[X])(X-E[X])']$$

$$=E\begin{bmatrix}(X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_n-\mu_n)\\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_n-\mu_n)\\ \cdots & \cdots & \cdots & \cdots\\ (X_n-\mu_n)(X_1-\mu_1) & (X_n-\mu_n(X_2-\mu_2) & \cdots & (X_n-\mu_n)^2\end{bmatrix}$$

$$=\begin{bmatrix}\mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \cdots & \mathrm{cov}(X_1,X_n)\\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2,X_n)\\ \cdots & \cdots & \cdots & \cdots\\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_1) & \cdots & \mathrm{var}(X_n)\end{bmatrix}$$

$$=\begin{bmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n}\\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2\end{bmatrix}$$

X 的协方差矩阵常常记为$\sum_x$，它是一个正定矩阵，如下是证明：

对于任意的不为零的向量$a'=(a_1,a_2,\cdots,a_n)$， 我们构造一个变量$Y=a'X$

那么 Y 的方差

$Var(Y)=Var(a'X)=a'\sum_x a\geq 0$（solved，how clever I am! hahaha），即证明了$\sum_x$是非负定的。

线性变换后的向量的均值与协方差

如果 P 是一个 $m\times n$ 常数矩阵，$m\leqslant n$，那么 $Z=PX$ 是一个 m 维随机向量，可以得到

a） $E[Z]=E[PX]=PE[X]=P\mu$

b） $\mathrm{cov}(Z)=\mathrm{cov}(PX)=P\sum_x P'$

四、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

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

条件均值函数 $E[y\mid x]$称为y对x的回归。

条件方差（Conditional Variance）

条件方差是条件分布的方差：

$$Var[y \mid x] = E[(y - E[y \mid x])^2 \mid x]$$

$$= \int_y (y - E[y \mid x])^2 f(y \mid x) dy$$

或

$$\sum_y (y - E[y \mid x])^2 f(y \mid x)$$（离散时）

利用下式可以简化计算

$$Var[y \mid x] = E[y^2 \mid x] - (E[y \mid x])^2$$

并且有：$$E[y] = E_x[E[y \mid x]]$$

记号 $E_x[\bullet]$表示对 X的值的期望。

几个重要的公式

1)、$E(XY \mid X) = XE(Y \mid X)$

思考：$E(g(X)Y \mid X) = g(X)E(Y \mid X)$是否成立?

2)、$E(XY) = E(XE(Y \mid X))$

3)、方差分解公式(Decomposition of Variance)

推导：分两步，先证明

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

这是因为：$E(Z \mid X) = E((Y - E(Y \mid X)) \mid X)$

$$= E(Y \mid X) - E(Y \mid X) = 0, \text{从而} EZ = E(E(Z \mid X)) = 0$$

进而有 $$\text{cov}(E(Y \mid X), Z)) = E(ZE(Y \mid X))$$

我们考察 $E[(ZE(Y \mid X)) \mid X] = E(Y \mid X)E(Z \mid X) = 0$

$\therefore$ $E(ZE(Y \mid X)) = E\{E\{ZE(Y \mid X) \mid X)\}\} = 0$

ii）对于任意 Y有：$Y = Y - E(Y \mid X) + E(Y \mid X) = Z + E(Y \mid X)$

因为 Z与 $E(Y|X)$ 是不相关，故

$$Var(Y) = Var(Y - E(E(Y \mid X)) + Var(E(Y \mid X))$$

而 $Var(Y-E(Y\mid X))=E(Y-E(Y\mid X))^2$

$=E[E(Y-E(Y\mid X)^2\mid X)]$

$=E(Var_x(Y\mid X))$

我们得到方差分解公式：

$$Var(Y)=Var(E(Y\mid X))+E(Var_x(Y\mid X))$$

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

回归方差=$Var_x[E[y|x]]$

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)：

残差方差=$E_x[Var[y|x]]$

这样， Var[y]=回归方差 + 残差方差。

由方差分解公式，我们得到$Var(E(Y\mid X))\le Var(Y)$，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。

[例子] 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

$$E(Y\mid X)=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(X-\mu_1)$$

则 $E(E(Y\mid X))=\mu_2=EY$，然而

$$Var(E(Y\mid X))=E[E(Y\mid X)-\mu_2]^2=E\left(\rho\frac{\sigma_2}{\sigma_1}(X-\mu_1)\right)^2$$

$$=\rho^2\frac{\sigma_2^2}{\sigma_1^2}E(X-\mu_1)^2=\rho^2\sigma_2^2$$

在－1＜ρ ＜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，$E(Var_x(Y\mid X))=\sigma_2^2-\rho^2\sigma_2^2=\sigma_2^2(1-\rho^2)$。

六、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)

定义 1　对于分布函数列$\{F_n(x)\}$，如果存在一个非降函数 $F(x)$使

$$\lim_{n\to\infty} F_n(x) = F(x)$$

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。

定义 2（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$\{\xi_n\}$依分布收敛于$\xi$，并记为$\xi_n \xrightarrow{L} \xi$。

b）依概率收敛(Convergence in Probability)

定义 3（依概率收敛）　如果

$$\lim_{n\to\infty} P\{|\xi_n - \xi| \geq \varepsilon\} = 0$$

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

定义 4（r-阶收敛）　设对随机变量ξ_n及ξ有$E|\xi_n|^r < \infty$，$E|\xi|^r < \infty$，其中 $r>0$为常数，如果 $\lim_{n\to\infty} E|\xi_n - \xi|^r = 0$，则称$\{\xi_n\}r$-阶收敛于$\xi$，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。

定理 8　$\xi_n \xrightarrow{r} \xi \Rightarrow \xi_n \xrightarrow{P} \xi$。

2）极限的应用

贝努里分布与普松分布

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）：

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

其中 $q=1-p$。$b(k;n,p)$，k=0, 1, 2, …，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。

定理（普松） 在贝努里试验中，以 p_n 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 $np_n \to \lambda$，则当 $n \to \infty$ 时，$b(k;n,P_n) \to \frac{\lambda^k}{k!}e^{-\lambda}$

b) **中心极限定理**(Central Limit Theorem)

若 X_1，X_2，…X_n，…是一串相互独立相同分布的随机变量序列，且

$$EX_k = \mu, Var(X_k) = \sigma^2$$

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。

定理 2（林德贝格-勒维） 若 $0 < \sigma^2 < \infty$，则

$$\lim_{n\to\infty} P\{\xi_n < x\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt$$

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

$$P\{|X - EX| \geqslant \varepsilon\} \leqslant \frac{Var(X)}{\varepsilon^2} \qquad (1)$$

其中 ε 是任一正数。

[证明] 若 F(x)是 X 的分布函数，则显然有

$$P\{|X - EX| \geqslant \varepsilon\} = \int_{|x-EX|\geq\varepsilon} dF(x)$$

$$\leqslant \int_{|x-EX|\geq\varepsilon} \frac{(x-EX)^2}{\varepsilon^2} dF(x)$$

$$\leqslant \frac{1}{\varepsilon^2}\int_{-\infty}^{\infty}(x-EX)^2 dF(x) = \frac{Var(X)}{\varepsilon^2} \qquad (2)$$

这就证得了不等式（1），有时把（1）改写成

$$P\{|X - EX| < \varepsilon\} \geqslant 1 - \frac{Var(X)}{\varepsilon^2}$$

或

$$P\left\{\left|\frac{X-EX}{\sqrt{Var(X)}}\right|\geq\delta\right\}\leqslant\frac{1}{\delta^2} \qquad (3)$$

契比雪夫不等式利用随机变量 X 的数学期望 EX 及方差 $Var(X)=\sigma^2$ 对 X 的概率分布进行估计。例如（3）断言不管 X 的分布是什么，X 落在 $(EX-\sigma\delta, EX+\sigma\delta)$ 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

定义 若 $\xi_1, \xi_2, \cdots, \xi_n, \cdots$ 是随机变量序列，令

$$\eta_n=\frac{\xi_1+\xi_2+\cdots+\xi_n}{n}$$

如果存在这样的一个常数序列 $a_1, a_2, \cdots, a_n, \cdots$，对任意的 $\varepsilon>0$，恒有

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

则称序列 $\{\xi_n\}$ 服从**大数定律**（或**大数法则**）。

契比雪夫大数定律 设 $X_1, X_2, \cdots, X_n, \cdots$ 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

$$Var(X_1)\leqslant C,\ Var(X_2)\leqslant C,\ \cdots,\ Var(X_n)\leqslant C,\ \cdots$$

则对任意的 $\varepsilon>0$，皆有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\frac{1}{n}\sum_{k=1}^{n}EX_k\right|<\varepsilon\right\}=1 \qquad (4)$$

［证明］ 因为 $\{\xi_k\}$ 两两不相关，故

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

再由契比雪夫不等式得到

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\frac{1}{n}\sum_{k=1}^{n}EX_k\right|<\varepsilon\right\}\geqslant 1-\frac{Var\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)}{\varepsilon^2}\geqslant 1-\frac{C}{n\varepsilon^2}$$

所以

$$1 \geqslant P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} EX_k\right| < \varepsilon\right\} \geqslant 1 - \frac{C}{n\varepsilon^2}$$

于是，当$n \to \infty$时有（4），因此定理得证。

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

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

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

$$EX_i = p, \qquad Var(X_i) = pq \leqslant \frac{1}{4}$$

而

$$\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} EX_k = \frac{\mu_n}{n} - p$$

$$P\left\{\left|\frac{\mu_n}{n} - p\right| \geq \varepsilon\right\} \leqslant \frac{1}{\varepsilon^2} Var\left(\frac{\mu_n}{n}\right) = \frac{1}{n\varepsilon^2} Var(X_i) \leqslant \frac{1}{4n\varepsilon^2}$$

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。

七、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：

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。

第四章　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。因此我们所关心的问题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。

1、数理统计的基本概念

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，…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，…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 X 具有相同的分布 $F(x)$；

（ⅱ）**独立性**：X_1，X_2，…，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 $X=(X_1, X_2, \cdots, X_n)$，其分布可以由母体的分布函数 $F(x)$ 完全决定，X 的分布函数是 $\prod_{i=1}^{n} F(x_i)$。

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：$\frac{1}{3}(Z_1+Z_2+Z_3)-\mu,\quad \frac{Z_1-Z_2}{\sigma}$

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,$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 为子样均值。

r 阶中心矩：$B_r=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^r$；　特别地,$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 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 μ_k 和 k 阶中心矩σ_k；

对于子样，我们有子样均值 $\overline{X}$ ，子样方差S_n^2，子样的 r 阶矩 A_r 和 r 阶中心矩 B_r。

我们可以得到如下结论：

定理 1　设母体服从分布 $F(x)$，X=（X_1，…，X_n）是从该母体中抽得的一个简单随机子样，如果 $F(x)$的二阶矩阵存在，则对子样均值 $\overline{X}$ ，有

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

[证明]　$E(\overline{X})=E(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$

$$Var(\overline{X})=E(\overline{X}-\mu)^2=E(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu)^2$$

$$=\frac{1}{n^2}E\left[\sum_{i=1}^{n}(X_i-\mu)\right]^2=\frac{1}{n^2}\sum_{i=1}^{n}E(X_i-\mu)^2=\frac{\sigma^2}{n}$$

思考：是否存在更简单的证明方法?

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

证明：因为$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n^2}(\sum_{i=1}^{n}X_i)^2$，所以

$$ES_n^2=E\left\{\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n^2}(\sum_{i=1}^{n}X_i)^2\right\}$$

$$= \alpha_2 - \frac{1}{n^2} E\{\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} \sum X_i X_j\}$$

（其中 $\alpha_2 = E\frac{1}{n}\sum X_i^2$ ）

$$= \alpha_2 - \frac{1}{n^2}(n\alpha_2 + n(n-1)\mu^2)$$

$$= \frac{n-1}{n}(\alpha_2 - \mu^2) = \frac{n-1}{n}\sigma^2$$

4）顺序统计量、经验分布函数与子样矩

设（X_1，…，X_n）是从母体　中抽取的一个子样，记（x_1,x_2…,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

$$x_1^* \leqslant x_2^* \leqslant \cdots \leqslant x_n^*$$

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

记

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

显然 0$\leqslant F_n^*(x) \leqslant$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, …, X_n）是子样观察值，$F_n^*(x)$ 是对应的经验分布函数，则有：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \int x dF_n^*(x)$$

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \int (X - \bar{X})^2 \, dF_n^*(x)$$

$$A_v = \frac{1}{n}\sum_{i=1}^{n} x_i^v = \int x^v dF_n^*(x), \quad v = 2,3,\cdots$$

$$B_v = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^v = \int (x-\bar{x})^v dF_n^*(x), \quad v = 3,4,\cdots$$

2、正态母体子样的线性函数的分布

定理 1 设 X_1，…，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

$$U = a_1X_1 + a_2X_2 + \cdots + a_nX_n \qquad (1)$$

则 U 也是正态随机变量，均值、方差分别为

$$E(U) = \mu\sum_{k=1}^{n} a_k \qquad (2)$$

$$Var(U) = \sigma^2\sum_{k=1}^{n} a_k^2 \qquad (3)$$

在（1）式中，特别地取 $a_k = \frac{1}{n}, k = 1,\cdots,n$，此时行到的 U 是子样均值 $\bar{X}$。

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{1}{n}\sigma^2$$

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。

定理 2 设

（1）X_1，X_2…，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

$$Y \underline{\Delta} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} a_{11}\cdots a_{1n} \\ a_{21}\cdots a_{2n} \\ \cdots\cdots\cdots \\ a_{p1}\cdots a_{pn} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \underline{\Delta} AX$$

$$X \underline{\Delta} (X_1X_2\cdots X_n)^\tau$$

则 Y_1，…，Y_p 也是正态随机变量，均值、方差、协方差分别为：

$EY_i = \mu\sum_{k=1}^{n} a_{ik}, i = 1,\cdots, p$。

$Var(Y_i) = \sigma^2\sum_{k=1}^{n} a_{ik}^2, i = 1,\cdots, p$

$\mathrm{cov}(Y_i, Y_j) = \sigma^2\sum_{k=1}^{n} a_{ik}a_{jk}, i, j = 1,\cdots, p$。

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

3、几种与正态分布 $N(0,1)$ 有关的常用分布

1）x^2-分布

定义　设 X_1，X_2，…，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 x^2-分布，x_n^2 称为自由度为 n 的 x^2-变量。

定理　设 $X_1 \sim x^2(n_1)$ 和 $X_2 \sim x^2(n_2)$，且 X_1, X_2 相互独立，则 $X_1 + X_2 \sim x^2(n_1 + n_2)$。

2）t-分布

设 $X \sim N(0,1)$和$Y \sim x^2(n)$，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～$t(n)$。

3）F-分布

定义　设 X 和 Y 是相互独立的 x^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～F（m,n）。

推论　如果 $X/\sigma^2 \sim x^2(m), Y/\sigma^2 \sim x^2(n)$，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～$F(m,n)$分布，则 $1/X$～$F(n,m)$分布。

结论 设 X_1，…，X_m 和 Y_1，…，Y_n 分别是从正态母体 $N(\mu_1,\sigma^2)$ 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。

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

$$\frac{X_{x+1}-\bar{X}}{S_n}\sqrt{\frac{n-1}{n+1}}$$

（提示：服从 t(n-1)分布）

4、统计量的分布与独立性

定理 若 x～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。

[证明] 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：

$$x'Ax=x'A'Ax=x_1'x_1 \qquad \text{其中} x_1=Ax$$

和 $x'Bx=x_2'x_2$ 其中 $x_2=Bx,$

两个向量都有零均值向量，所以 X_1 和 X_2 协方差矩阵是

$$E(X_1X_2')=AIB'=AB=0$$

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。

[例] 易知

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

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=X'M^0X,\text{而}n\bar{x}^2=X'(I-M^0)X$$

因为　$M^0(I-M^0)=M^0-M^{0^2}=0$

故　$n\bar{x}^2$与$\sum_{i=1}^{n}(x_i-\bar{x})^2$ 是相互独立的。

5、线性变换及二次型的独立性

定理　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。

[例]　$\sqrt{n}\bar{x}=\frac{1}{\sqrt{n}}\bar{i}'X,\qquad S^2=\frac{X'M^0X}{n-1}$

$\because$　$M^0\bar{i}=[I-\frac{1}{n}\bar{i}\bar{i}']\cdot\bar{i}=0$

所以上面两个统计量是相互独立的。

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

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

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

则有　（1）$\bar{X}$和S_n^2独立；

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

（3）$nS_n^2/\sigma^2\sim x^2(n-1)$

$$T=\frac{\bar{X}-\mu}{S_n}\sqrt{n-1}\sim t(n-1)$$

[证明]　因为$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\sim N(0,1),nS_n^2/\sigma^2\sim x^2(n-1),$

所以

$$T=\frac{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}}{\sqrt{\frac{nS_n^2}{\sigma^2(n-1)}}}=\frac{(\overline{X}-\mu)}{S_n}\sqrt{n-1}$$

服从自由度为 n－1 的 t-分布。

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$\{F(\cdot;\ \theta),\theta\in\Theta\}$，$X_1$，$X_2\cdots$，$X_n$ 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，$\cdots X_n$）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，$\theta$ =（θ_1，…，θ_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

$$a_v(\theta_1,\cdots,\theta_k)=\int_{-\infty}^{\infty}x^v dF(x;\theta),\quad 1\leqslant v\leqslant k$$

是θ =（θ_1，…，θ_k）的函数。

对于子样 X=（X_1，…，X_n），其 v 阶子样矩是

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

现在用子样矩作为母体矩的估计，即令

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

这样，（1）式确定了包含 k 个未知参数θ =（θ_1，…，θ_k）的 k 个方程式。

[例]　母体均值和方差的矩估计。

设 X_1，…，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

$$\begin{cases}\hat{\mu}=\dfrac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}\\ \hat{\mu}^2+\hat{\sigma}^2=\hat{\alpha}_2=\dfrac{1}{n}\sum_{i=1}^{n}X_i^2\end{cases}$$

解之得　　$\hat{\sigma}^2=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=S_n^2$

所以母体均值 μ 和方差 σ^2 的矩估计分别是子样均值 $\overline{X}$ 和子样方差 S_n^2。

运用以前的有关定理有

$$E\hat{\mu}=E\overline{X}=\mu$$

$$Var(\hat{\mu})=\frac{1}{n}\sigma^2$$

和

$$E(\hat{\sigma}^2)=\frac{n-1}{n}\sigma^2$$

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$\{F(x;\theta),\theta\in\Theta\}$，其中$\theta=(\theta_1,\theta_2\cdots,\theta_k)$是一个未知的 k 维参数向量，需待估计，又设（$x_1$，…，$x_n$）是子样（$X_1$，…，$X_n$）的一个观察值，那么子样（$X_1$，…，$X_n$）落在点（$x_1$，…，$x_n$）的邻域里的概率是$\prod_{i=1}^{n}f(x_i;\theta)dx_i$。

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

$$L(x_1,\cdots,x_n;\hat{\theta})=\sup_{\theta\in\Theta}L(x_1,\cdots x_n;\theta)\qquad(2)$$

成立的$\hat{\theta}(X)=(\hat{\theta}_1(X),\cdots,\hat{\theta}_k(X))$作为$\theta$ 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）式可等价地写为：

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

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

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

[例] 设X=（X_1，…，X_n）是取自均匀分布

$$f(x;\theta)=\begin{cases}\dfrac{1}{\theta}, & 0<x\leqslant\theta\\ 0, & \text{其它}\end{cases}\qquad(\theta>0)$$

的子样，试求θ的最大似然估计。

此时 $L(x;\theta)=\prod_{i=1}^{n}f(x_i;\theta)=\begin{cases}\theta^{-n},\text{当 } 0<\max\limits_{1\leq i\leq n}x_i\leq\theta\\ 0, \quad \text{其它}\end{cases}$

（注意：条件$0<x_i\leqslant\theta$，i=1,…,n和条件$0<\max\limits_{1\leq i\leq n}x_i\leq\theta$是等价的。

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

7、估计的有效性

1）无偏估计

定义 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

定义 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$Var(\hat{\theta}_1)<Var(\hat{\theta}_2)$，则称$\hat{\theta}_1$比$\hat{\theta}_2$更有效。

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$Var(\hat{\theta}_2)-Var(\hat{\theta}_1)$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$T_n\underline{\Delta}T_n(X_1,\cdots,X_n)$满足下面的关系式

$$\lim_{n\to\infty} E_\theta(T_n)=\theta, \text{对一切}\theta\in\Theta$$

则称 T_n 是θ 的**渐近无偏估计**。

4）一致估计

设 X_1，…，X_n 是取自分布族$\{F(x;\theta),\theta\in\Theta\}$的子样，$T_n$=$T_n$($X_1$，…，$X_n$)是$\theta$ 的一个估计。如果序列{T_n}随机收敛到真参数值θ ，即对任意$\varepsilon>0$，

$$\lim_{n\to\infty} P_\theta\{|T_n-\theta|>\varepsilon\}=0, \text{对一切}\theta\in\Theta$$

则称 T_n 是θ 的一致估计。

5）最小方差无偏估计

一般地若 T_1 是θ 的一个无偏估计，关于θ 的任一无偏估计 T_2 成立下式

$$Var_\theta(T_1)\leqslant Var_\theta(T_2), \text{对一切}\theta\in\Theta$$

则称 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,…, 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

$$Var_\theta(T_0)\leqslant Var_\theta(T)\text{，对一切}\theta\in\Theta\text{和}T\in U_L$$

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

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：

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

$$=\left(E\left[\left(\frac{\partial\ln L(\theta)}{\partial\theta}\right)\left(\frac{\partial\ln L(\theta)}{\partial\theta'}\right)\right]\right)^{-1}$$

量 $I(\theta)$ 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

$$[I(\theta)]^{-1}=\left\{-E\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'}\right]\right\}^{-1}$$

$$=\left\{E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)\left(\frac{\partial \ln L(\theta)}{\partial\theta'}\right)\right]\right\}^{-1}$$

即 $$I(\theta)=\begin{bmatrix} -E\left(\frac{\partial^2 L}{\partial\theta_1^2}\right) & -E\left(\frac{\partial^2 L}{\partial\theta_1\partial\theta_2}\right) & \cdots & -E\left(\frac{\partial^2 L}{\partial\theta_1\partial\theta_k}\right) \\ -E\left(\frac{\partial^2 L}{\partial\theta_2\partial\theta_1}\right) & -E\left(\frac{\partial^2 L}{\partial\theta_2^2}\right) & \cdots & -E\left(\frac{\partial^2 L}{\partial\theta_2\partial\theta_k}\right) \\ \cdots & \cdots & \cdots & \cdots \\ -E\left(\frac{\partial^2 L}{\partial\theta_k\partial\theta_1}\right) & -E\left(\frac{\partial^2 L}{\partial\theta_k\partial\theta_2}\right)\cdots & & -E\left(\frac{\partial^2 L}{\partial\theta_k^2}\right) \end{bmatrix}$$

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

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

［例］ 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ν/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ν/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 $P_{H0}\{|U| \geq u_0\} \leq \alpha$ 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 $P_{H0}\{|U| \geq u_0\}=\alpha$ 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 1。

表 1　正态母体参数的显著性检验

检验参数	假设 H0	统　计　量	分　布
μ	$\mu=\mu_0$（$\sigma=\sigma_0$）	$U=\dfrac{\overline{X}-\mu_0}{\sigma_0}\sqrt{n}$	$N(0,1)$
	$\mu_1=\mu_2$（σ_1,σ_2 已知）	$U=(\overline{X}-\overline{Y})/\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}$	
	$\mu=\mu_0\ \sigma^2>0$	$T=\dfrac{\overline{X}-\mu_0}{S_n}\sqrt{n-1}$	$t(n-1)$
	$\mu_1=\mu_2$，$\sigma_1=\sigma_2$	$T=\dfrac{(\overline{X}-\overline{Y})}{\sqrt{mS_{1m}^2+nS_{2n}^2}}\cdot\sqrt{\dfrac{mn(m+n-2)}{m+n}}$	$t(m+n-2)$
σ^2	$\sigma=\sigma_0$	$x^2=\dfrac{nS_n^2}{\sigma_0^2}$	$x^2(n-1)$
	$\sigma_1^2=\sigma_2^2$	$F=\dfrac{mS_{1m}^2}{nS_{2n}^2}\cdot\dfrac{n-1}{n-1}$	$F(m-1,n-1)$

例 1 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 $P\{Y=0\}=1-p_0$，$P\{Y=1\}=p_0$；$EY=p_0$，Var（Y）$=p_0(1-p_0)$，设 X_1，…，X_n 是一子样，则

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{v}{n}$$

其中 v 表示子样中的次品数。

由中心极限定理知道，在 $H_0(p=p_0)$ 成立的条件下，

$$U=\frac{(\overline{X}-EX)}{\sqrt{Var(X)}}\sqrt{n}=\frac{\left(\frac{v}{n}-p_0\right)}{\sqrt{p_0(1-p_0)}}\sqrt{n}$$

$$=\frac{(v-np_0)}{\sqrt{np_0(1-p_0)}} \qquad (1)$$

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

现在 p_0=0.05，n=50，v=4，代入（1）式得

$$u=\frac{\left(\frac{4}{50}-0.05\right)}{\sqrt{0.05\times0.95}}\sqrt{50}\approx0.96$$

对 α =0.01，查表得 $u_{\alpha/2}$=2.58，这时因

|u|=0.96＜2.58=$u_{\alpha/2}$

所以不能拒绝假设 H_0（p=0.05）。

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，便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。

表 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	单子样	$\sigma=\sigma_0$	$\bar{X}-u_{a/2}\cdot\sigma_0/\sqrt{n}$	$\bar{X}+u_{a/2}\cdot\sigma_0/\sqrt{n}$	$U=\dfrac{\bar{X}-\mu}{\sigma_0}\sqrt{n}$
		σ 未知	$\bar{X}-t_{a/2}\cdot S_n/\sqrt{n-1}$	$\bar{X}-t_{a/2}\cdot\dfrac{S_n}{\sqrt{n-1}}$	$T=\dfrac{\bar{X}-\mu}{S_n}\sqrt{n-1}$
$\mu_1-\mu_2$	双子样	已知$\sigma_1=\sigma_2$ 但数值未知	$-t_{a/2}\dfrac{(\bar{X}-\bar{Y})\sqrt{mS_{1m}^2+nS_{2n}^2}\cdot\sqrt{m+n}}{\sqrt{(m+n-2)mn}}$	$+t_{a/2}\dfrac{(\bar{X}-\bar{Y})\sqrt{mS_{1m}^2+nS_{2n}^2}\cdot\sqrt{m+n}}{\sqrt{(m+n-2)mn}}$	$T=\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{mS_{1m}^2+nS_{2n}^2}}\times\sqrt{\dfrac{(m+n-2)m\cdot n}{m+n}}$
σ^2	单子样		$\dfrac{1}{C_{2a}}\cdot nS_n^2$	$\dfrac{1}{C_{1a}}\cdot nS_n^2$	$x^2=\dfrac{nS_n^2}{\sigma_0^2}$
$\dfrac{\sigma_1^2}{\sigma_2^2}$	双子样		$\dfrac{1}{f_{2a}}\dfrac{mS_{1m}^2}{nS_{2n}^2}\cdot\dfrac{n-1}{m-1}$	$\dfrac{1}{f_{1a}}\dfrac{mS_{1m}^2}{nS_{2n}^2}\cdot\dfrac{n-1}{m-1}$	$F=\dfrac{mS_{1m}^2/\sigma_1^2}{nS_{2n}^2/\sigma_2^2}\cdot\dfrac{n-1}{m-1}$

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 $\bar{X}$和S_n^2 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 a, 和 c_1, c_2，使得

$$P\{-a<\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}}<a\}\sqrt{0.95}\approx 0.975$$

和

$$P\{c_1<\frac{nS_n^2}{\sigma_0^2}<c_2\}=\sqrt{0.95}\approx 0.975$$

联合概率是

$$P\{-a<\frac{\overline{X}-\mu_0}{\sigma_0/\sqrt{n}}<a, c_1<\frac{nS_n^2}{\sigma_0^2}<c_2\}=0.95$$

解得：

$$P\{(\mu_0-\overline{X})^2<a^2\sigma_0^2/n, \frac{nS_n^2}{c_2}<\sigma_0^2<\frac{nS_n^2}{c_1}\}=0.95 \qquad (1)$$

由此可见，(μ,σ^2)的置信度为0.95的联合置信域是（1）式大括号内不等式对μ,σ_0^2所给出的范围。

4）广义似然比检验

设$X=(X_1,\cdots,X_n)$是从母体中抽取的子样，其可能分布族$\{f(x;\theta),\theta\in\Theta\}$，其中$\theta$（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$H_0$：$\theta=\theta_0$。这里应指出，$\theta_0$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

$$\theta\triangleq(\mu,\sigma^2)$$

$$\Theta=\{(\mu,\sigma^2):-\infty<\mu<\infty,\sigma^2>0\}$$

$$\theta_0\triangleq\{(\mu_0,\sigma^2)>0\}$$

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：

$$\lambda(x)\triangleq\frac{\sup\limits_{\theta\in\theta_0}\prod\limits_{i=1}^{n}f(x_i;\theta)}{\sup\limits_{\theta\in\Theta}\prod\limits_{i=1}^{n}f(x_i;\theta)}$$

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θ。由于$\theta_0\subset\Theta$，所以

$$0\leqslant\lambda(x)\leqslant1$$

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成立时拒绝H_0，

$$\lambda(x)\leqslant\lambda_0 \qquad (1)$$

其中λ_0的选取是使得下式成立，

$P_\theta\{\lambda(X)\le\lambda_0\}\le\alpha$，对一切$\theta\in\theta$。　　（2）

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

$P_{\theta 0}\{\lambda(X)\le\lambda_0\}\le\alpha$

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$\Theta-\Theta_0$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（见图 1）。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。

	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

图 1

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。

第五章　古典线性回归模型

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$$\varepsilon$ 是一个随机扰动。这是一个标准的古典线性回归模型。假如我们得到如下例 1 的数据

例 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
1970	751.6	672.1
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	864.3
1978	988.8	903.2
1979	1015.7	927.6

来源：数据来自总统经济报告，美国政府印刷局，华盛顿特区，1984。

（收入和支出全为 1972 年的十亿美元）

一、线性回归模型及其假定

一般地，被估计模型具有如下形式：

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

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**。

构成古典线性回归模型的**一组基本假设**为：

1. 函数形式：$y_i=\alpha+\beta x_i+\varepsilon_i$，$i=1,\cdots,n,$

2. 干扰项的零均值：对所有 i，有：$E[\varepsilon_i]=0$。

3. 同方差性：对所有 i，有：$Var[\varepsilon_i]=\sigma^2$，且$\sigma^2$是一个常数。

4. 无自相关：对所有 $i \neq j$，则 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 Cov[x_i, ε_j]=0。

6. 正态性：对所有 i，ε_i 满足正态分布 N（0，σ^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。

[例] 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：

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

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。

例如，任何变换也不能将

$$y = \alpha + \frac{1}{\beta + x} 和 y = \alpha + \beta x^{\nu} （0 < \nu < 1）$$

转化为线性回归模型。

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

$$E[y_i \mid x_i] = E[\alpha + \beta x_i + \varepsilon_i] = \alpha + \beta x_i + E[\varepsilon_i] = \alpha + \beta x_i$$

和方差

$$Var[y_i \mid x_i] = Var[\alpha + \beta x_i + \varepsilon_i] = Var[\varepsilon_i] = \sigma^2 。$$

此外，有必要假定，对 $n \geqslant 1$

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

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 $\alpha + \beta x$ 的任何推断。这个识别条件等同于子样的极差 $\max(X_1,\cdots,X_n) - \min(X_1,\cdots,X_n) \neq 0$。

2）当 X 为随机变量

若 x 被当作一个随机变量，则假定 1 成为一个对 y 和 x 的联合分布的陈述。

我们就用条件期望和方差来处理。

3、随机干扰项

1）如果干扰项不是零均值，即 $E[\varepsilon_i] = \mu$ ，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，$y = \alpha' + \beta x + \varepsilon'$，此模型满足我们原始模型的要求。

2）观测值中的随机部分假定是不相关的：

$E[\varepsilon_i \varepsilon_j] = 0$　　对所有 i 不等于 j。

这被称为**非自相关**。

二、最小二乘法

1　最小二乘系数

总体回归是 $E[y_i|x_i] = \alpha + \beta x_i$，而我们对 $E[y_i|x_i]$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

$$\frac{\partial(\Sigma_i e_i^2)}{\partial a} = \sum_i 2(y_i - a - bx_i)(-1)$$

$$= -2\sum_i (y_i - a - bx_i) = 0$$

和

$$\frac{\partial(\Sigma_i e_i^2)}{\partial b} = \sum_i 2(y_i - a - bx_i)(-x_1) = 0$$

$$= -2\sum_i x_i (y_i - a - bx_i) = 0$$

将上两式展开合并同类项后得到正规方程组

$$\sum_i y_i = na + \left(\sum_i x_i\right) b, \qquad (1)$$

$$\sum x_i y_i = \left(\sum_i x_i\right) a + \left(\sum_i x_i^2\right) b \qquad (2)$$

（1）式暗示 $\sum_{i=1}^{n} e_i = 0$，而（2）式暗示 $\sum_i X_i e_i = 0$

为了得到解，我们首先用 n 除（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：

$$a = \bar{y} - b\bar{x} \qquad (3)$$

有了 a 后，我们可以求解（2）得到 b。首先，$\Sigma_i x_i = n\bar{x}$。将此和（3）代入（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

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

最小的残差平方和，对 *a* 和 *b* 的二阶微商矩阵是

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

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，行列式为$(4n)\Sigma_i x_i^2 - 4(\Sigma_i x_i)^2$但$\Sigma_i x_i = n\bar{x}$，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。

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：

$$SST = \sum_i (y_i - \bar{y})^2$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i^2 + 2\sum e_i \hat{y}_i$$

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

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

总平方和=回归平方和+残差平方和

或

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

$$\text{决定系数} R^2 = \frac{SSR}{SST}$$

为了方便计算与分析，约定

$$S_{xx}=\sum(x_i-\bar{x})^2,\qquad S_x=\sqrt{S_{xx}},$$

$$S_{yy}=\sum(x_i-\bar{x})^2,\qquad S_y=\sqrt{S_{yy}},$$

和

$$S_{xy}=\sum_i(x_i-\bar{x})(y_i-\bar{y})$$

x 和 y 间的样本相关系数是 $r_{xy}=S_{xy}/(S_xS_y)$。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

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

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。

3 方差分析表

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β =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 $x^2(1)$(**为什么？**)；

b）残差平方和 $\frac{SSE}{\sigma^2}$ 服从自由度为 *n-2* 的卡方分布 $x^2(n-2)$;

c）**在β =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。

证明：

a ） $b=\frac{S_{xy}}{S_{xx}}=\frac{\sum_i(x_i-\bar{x})y_i}{S_{xx}}=\sum_i c_iy_i$ ， 其 中 $c_i=\frac{x_i-\bar{x}}{S_{xx}}$ ， 易 知 $\sum_i c_i^2=\frac{1}{S_{xx}}$, 令$C=(c_1,c_2\cdots c_n)'$, 则$b=C'Y, b^2=Y'CC'Y$ ，

$SSR=b^2\sum_i(x_i-\bar{x})^2=S_{xx}b^2=Y'S_{xx}CC'Y$ 。

可以验证 $S_{xx}CC'$ 是幂等矩阵。

$$S_{xx}CC'\cdot S_{xx}CC'=S_{xx}^2C(C'C)C=S_{xx}CC'$$

$$r(S_{xx}CC')=tr(S_{xx}CC')=S_{xx}\sum_i c_i^2=1$$

在β =0 的假设条件下， $\frac{SSR}{\sigma^2}$ 才服从自由度为 1 的卡方分布 $x^2(1)$（**为什么？**）

b）因为 $SST = YM_0Y$ 及 $SST = SSR + SSE$

所以 $SSE = Y'(M_0 - S_{xx}CC')Y$

易验证 $M_0 - S_{xx}CC'$ 也是幂等矩阵

$$(M_0 - S_{xx}CC')^2 = M_0 - S_{xx}CC'M_0 - S_{xx}M_0CC' + S_{xx}CC'$$

$$= M_0 - S_{xx}CC' + \frac{1}{n}S_{xx}CC'ii' + \frac{1}{n}S_{xx}ii'CC' = M_0 - S_{xx}CC'$$

最后一个等式成立是因为 $C'i = i'C = \sum_i c_i = 0$。

所以 $r(M_0 - S_{xx}CC') = tr(M_0 - S_{xx}CC') = n - 1 - S_{xx} \cdot \frac{1}{S_{xx}} = n - 2$，从而 $\frac{SSE}{\sigma^2} \sim x^2(n-2)$。此结论成立不需要**$\beta$ =0 的假设条件下，为什么？**

c）因为 $S_{xx}CC' \cdot (M_0 - S_{xx}CC') = S_{xx}CC' \cdot (I - \frac{1}{n}ii' - S_{xx}CC')$

$$= S_{xx}CC' - \frac{1}{n}S_{xx}CC'ii' - S_{xx}CC' = 0$$

所以 SSR 与 SSE 是相互独立的统计量。从而，**在β =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。

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 1 中。

表 1　方差分析表

变差来源	变差	自由度	均方
回归	$SSR=b^2Sxx$	1	$\frac{SSR}{1}$
残差	$SST=\Sigma_i e_i^2$	$n-2$	$\frac{SSE}{n-2}$
总	$SST=S_{yy}$	$n-1$	$\frac{S_{yy}}{n-1}$

$$F[1,n-2]=\frac{SSR/1}{SSE/(n-2)}$$

2）回归量 X 是随机变量

我们要利用方差分解公式

$$Var(Y)=Var(E(Y\mid X))+E(Var_x(Y\mid X))$$

$$=Var(\alpha+\beta X)+E[E(Y-E(Y\mid X))^2\mid X]$$

$$=\beta^2Var(X)+E[E(Y-E(Y\mid X))^2\mid X]$$

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：

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

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

$$\overline{C}=793.43,\qquad \overline{X}=879.24,$$

$$S_{CC}=64{,}972.12,\qquad S_{XX}=67{,}192.44,$$

$$S_{XC}=65{,}799.34.$$

总平方和的各个部分为

总平方和=64,972.12

回归平方和=64,435.13

残差平方和=537.00

$$R^2 = \frac{64,435.13}{64,972.12} = 0.99173$$

显然，此回归提供了一个很好的拟合。

对消费和收入数据，方差分析表如下所示

例 1 数据的方差分析表

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

$$F[1,8] = \frac{64,435.13}{67.124} = 959.94$$

另一个计算和通常 R^2 相类似公式是：

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

任何一个模型的残差都可用 $y_i - \hat{y}$ 来计算。

三、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

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

$$= \sum_i c_i y_i \qquad (1)$$

其中权数，

$$c_i = \frac{x_i - \bar{x}}{S_{xx}} \qquad (2)$$

仅仅是 $x_1,\cdots,x_n$ 的一个函数。

1、b 是 β 的无偏估计

将 $y_i=\alpha+\beta x_i+\varepsilon_i$ 代入（1），我们得到

$$b=\frac{\Sigma_i(x_i-\bar{x})(\alpha+\beta x_i+\varepsilon_i)}{S_{xx}}$$

$$=\frac{\alpha\Sigma_i(x_i-\bar{x})}{S_{xx}}+\frac{\beta\Sigma_i(x_i-\bar{x})x_i}{S_{xx}}+\frac{\Sigma_i(x_i-\bar{x})\varepsilon_i}{S_{xx}}$$

$$=\beta+\sum_i c_i\varepsilon_i \qquad (3)$$

所以

$$E[b]=\beta+E\left[\sum_i c_i\varepsilon_i\right]=\beta \qquad (4)$$

这是因为 $E[\varepsilon_i]=0$。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（3）得到 b 的样本方差

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

线性回归模型的假定 4 暗示这个和的方差中的协方差项是零，所以有

$$Var[b]=\sum_i Var[c_i\varepsilon_i]=\sum_i \sigma^2 c_i^2=\frac{\sigma^2}{S_{xx}}$$

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。

2、a 是 α 的无偏估计

对于最小二乘截距 a，我们有：

$$a=\bar{y}-b\bar{x}$$

$$=\frac{1}{n}\sum_i y_i-b\bar{x}$$

$$=\frac{1}{n}\sum_i(\alpha+\beta x_i+\varepsilon_i)-b\bar{x}$$

利用（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 $\Sigma_i d_i\varepsilon_i$ 的方差，根据独立性有

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

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

3、a、b 估计量的协方差矩阵

两个估计的协方差是

$$Cov[a,b]=E[(a-\alpha)(b-\beta)]=E\left[\left(\sum_i d_i\varepsilon_i\right)\left(\sum_i c_i\varepsilon_i\right)\right]$$

$$=\sigma^2\sum_i c_i d_i=\frac{-\bar{x}\sigma^2}{S_{xx}}$$

a 和 *b* 两者都有$\Sigma_i w_i y_i$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

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

4、*b* 是 β 的最小线性无偏估计。

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

[证明]　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$\Sigma_i q_i = 0$及$\Sigma_i q_i x_i = 1$。这样，$b' = \beta + \Sigma_i q_i \varepsilon_i$。$b'$的方差是

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

令$\upsilon_i = q_i - c_i$,则$q_i = c_i + \upsilon_i$且

$$Var[b'] = \sigma^2 \sum_i (c_i + \upsilon_i)^2$$

$$= \sigma^2 \left(\sum_i c_i^2 + \sum_i \upsilon_i^2 + 2\sum_i c_i \upsilon_i \right)$$

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

推导$\sum_i c_i \upsilon_i = 0$

$$\sum_i c_i \upsilon_i = \sum_i (q_i - c_i) c_i = \sum_i c_i q_i - \sum_i c_i^2$$

$$= \frac{\sum_i x_i q_i - \sum_i q_i \bar{x}}{S_{xx}} - \frac{1}{S_{xx}} = \frac{1}{S_{xx}} - \frac{1}{S_{xx}} = 0$$

四、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。

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

$$e_i = \varepsilon_i - \bar{\varepsilon} - (x_i - \bar{x})(b - \beta)$$

$$= \varepsilon_i - \bar{\varepsilon} - (x_i - \bar{x})\left(\sum_j c_j \varepsilon_j\right) \qquad (1)$$

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$E(\varepsilon_i \varepsilon_j) = 0若i \neq j$。现在我们平方的两边并取期望值，可得到

$$E[e_i^2] = \sigma^2 + \frac{\sigma^2}{n} + \sigma^2 (x_i - \bar{x})^2 \left(\sum_j c_j^2\right) - \frac{2\sigma^2}{n}$$

$$- 2\sigma^2 (x_i - \bar{x}) c_i + \frac{2\sigma^2}{n}(x_i - \bar{x})\left(\sum_j c_j\right)$$

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

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

这样，我们可以得到 b 的抽样方差的一个估计为

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

以后，我们将用记号$Est.Var[\cdot]$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

$$z = \frac{b - \beta}{\sqrt{\sigma^2 / S_{xx}}} \qquad (1)$$

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

$$\frac{(n-2)s^2}{\sigma^2} \sim x^2(n-2) \qquad (2)$$

并且和 b 是独立的。

根据（1）和（2），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 $s_b=\frac{S}{\sqrt{S_{xx}}}=\frac{S}{S_x}$，则比率

$$\frac{b-\beta}{s_b}\sim t[n-2] \qquad (3)$$

可以形成统计推断的基础。

2、抽样分布的

β 的置信区间将以（3）为基础。特别的，我们可以有

$$P(b-t_{\lambda/2}s_b\leqslant\beta\leqslant b+t_{\lambda/2}s_b)=1-\lambda,$$

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。

3、 β 的假设检验

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

$$\frac{(n-2)s^2}{\chi^2_{0.975}} \quad 至 \quad \frac{(n-2)s^2}{\chi^2_{0.025}}$$

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

$$H_0:\beta=\beta^0 \quad 对 \quad H_1:\beta\neq\beta^0,$$

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

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

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

例子

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

$$s^2=\frac{537.00}{8}=67.125$$

$$S_{xx}=67{,}192.45$$

和

$$\bar{x}=679.24$$

$$s_a=27.91$$

$$s_b=27.91$$

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±2.306（27.91）　　或　　－131.94　　至　　－3.22

和

0.9793±2.306（0.03161）　　或　　－0.90641　　至　　－1.0522

我们得到基于自由度为（10—2）=8 的 x^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

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

或

$30.62<\sigma^2<246.33$

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。

五、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：

1．个体预测(Individual Prediction）

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

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

预测误差是

$$e^0 = y^0 - \hat{y}^0$$

$$= \alpha + \beta x^0 + \varepsilon^0 - a - bx^0$$

$$= (\alpha - a) + (\beta - b)x^0 + \varepsilon^0$$

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

$$Var[e^0] = Var[a] + (x^0)^2 Var[b] + 2x^0 Cov[a,b] + Var[\varepsilon^0]$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{(x^0)^2}{S_{xx}} - \frac{2\bar{x}x^0}{S_{xx}}\right]$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x^0 - \bar{x})^2}{S_{xx}}\right]$$

所以 $\dfrac{y^0 - a - bX^0}{\sigma\sqrt{(1 + \frac{1}{n} + \frac{(X^0 - \bar{X})^2}{S_{xx}})}} \sim N(0,1)$

又因为 $\dfrac{(n-2)s^2}{\sigma^2} \sim x^2(n-2)$

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

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

$$(a + bx^0) \pm t_{\lambda/2}\sqrt{s^2\left[1 + \frac{1}{n} + \frac{(x^0 - \bar{x})^2}{S_{xx}}\right]} \qquad (3)$$

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 ε^0。

预测误差是

$$e^0=y^0-\hat{y}^0$$

$$=\alpha+\beta x^0-a-bx^0$$

$$=(\alpha-a)+(\beta-b)x^0$$

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

$$Var[e^0]=Var[a]+(x^0)^2Var[b]+2x^0Cov[a,b]$$

$$=\sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}+\frac{(x^0)^2}{S_{xx}}-\frac{2\bar{x}x^0}{S_{xx}}\right]$$

$$=\sigma^2\left[\frac{1}{n}+\frac{(x^0-\bar{x})^2}{S_{xx}}\right]$$

所以 $\dfrac{y^0-a-bX^0}{\sigma\sqrt{(\dfrac{1}{n}+\dfrac{(X^0-\bar{X})^2}{S_{xx}}}}\sim N(0,1)$

又因为 $\dfrac{(n-2)s^2}{\sigma^2}\sim x^2(n-2)$

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

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

$$(a+bx^0)\pm t_{\lambda/2}\sqrt{s^2\left[\frac{1}{n}+\frac{(x^0-\bar{x})^2}{S_{xx}}\right]}\qquad(4)$$

例子

利用例 1 中的消费数据，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：

$$-67.5806 \pm 0.9793(1030) \pm 2.306(9.8256)$$

即

$$941.1 \pm 22.658.$$

第六章　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。

一、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 x 是变量 x_k，k=1，2，的 n 个观测值，并用这些数据组成一个 $n \times K$ 数据矩阵 X，在多数情况下，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 y 是 n 个观测值 $y_1, y_2, \cdots, y_n$ 组成的列向量，现在可将模型写为：

$$y = x_1\beta_1 + \cdots x_K\beta_K + \varepsilon$$

构成多元线性回归模型的一组基本假设为

假定 1.　$y = X\beta + \varepsilon$

我们主要兴趣在于对参数向量β 进行估计和推断。

假定 2.　$E[\varepsilon] = \begin{bmatrix} E[\varepsilon_1] \\ E[\varepsilon_2] \\ \vdots \\ E[\varepsilon_n] \end{bmatrix} = 0,$

假定 3.　$E[\varepsilon\varepsilon'] = \sigma^2 I_n$

假定 4.　$E[\varepsilon \mid X] = 0$

我们假定 X 中不包含ε 的任何信息，由于

$$Cov[X, \varepsilon] = Cov[X, E(\varepsilon \mid X)], \qquad (1)$$

所以假定 4 暗示着$Cov[X, \varepsilon] = 0$。

（1）式成立是因为，对于任何的双变量 X，Y，有 E(XY)=E(XE(Y|X))，而且

$$Cov(X, Y) = E(X - EX)(Y - EY)'] = E[(X - EX)(E(Y \mid X) - EY)']$$

$$= Cov(X, E(Y \mid X))$$

这也暗示　　$E[y \mid X] = X\beta$

假定 5　X **是秩为** K **的** $n \times K$ **随机矩阵**

这意味着 X 列满秩，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：

假定 6　$\varepsilon \sim N[0, \sigma^2 I]$

二、最小二乘回归

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

$$S(\hat{\beta}) \underline{\Delta} \left\| y - X\hat{\beta} \right\|^2 = \min_{\beta} \left\| y - X\beta \right\|^2 \qquad (2)$$

其中 $\left\| y - X\beta \right\|^2 \underline{\Delta} \sum_{j=1}^{n} \left(y_i - \sum_{j=1}^{K} x_{ij} \beta_j \right)^2 = (y - X\beta)'(y - X\beta)$，min 是对所有的 m 维向量β 取极小值。

也即　　$S(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} X_{ij} \hat{\beta}_j)^2$

$$= \min_{\beta_1, \cdots \beta_m} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} X_{ij} \beta_i)^2 \qquad (3)$$

满足（2）式或（3）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

$$S(\beta) = y'y - \beta' X'y - y' X\beta + \beta' X' X\beta$$

或

$$S(\beta) = y'y - 2\beta' X'y + \beta' X' X\beta$$

最小值的必要条件是

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0$$

设 b 是解，则 b 满足正则方程组

$$X'Xb = X'y$$

这正是我们曾分析的最小二乘正则方程组。因为 X 是满秩的，所以 $X'X$ 的逆存在，从而得到解是

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

为了证实这确实是最小值，我们需要二阶编分矩阵

$$\frac{\partial^2 S(\beta)}{\partial\beta\partial\beta'}\Big|_{\beta=b} = 2X'X$$

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 c，令 $q = c'X'Xc$，则

$$q = v'v = \sum_i \upsilon_i^2, \quad \text{其中} \quad v = Xc$$

除非 v 的每一元素都为 0，否则 q 是正的。但若 υ 为零的话，则 X 的各列的一个线性组合等于 0，这与 X 满秩的假定相矛盾。

三、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。

1、X 非随机回归量

若回归量当作非随机来进行处理时，则将 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

$$b = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \qquad (4)$$

若 X 是非随机的，或 $E(X'\varepsilon) = 0$，则（4）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

$$Var[b] = E[(b-\beta)(b-\beta)']$$

$$= E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}]$$

$$= (X'X)^{-1}X'E[\varepsilon\varepsilon']X(X'X)^{-1}$$

$$= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}$$

在前面的内容中，对 *K*=2 的特殊 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$\tilde{b} = Cy$是β 的另一个不同于 b 的线性无偏估计量，其中 C 是一个 K×n 矩阵。若$\tilde{b}$ 是无偏的，

$$E[Cy] = E[CX\beta + C\varepsilon] = \beta,$$

这暗示着 CX=I，并且$\tilde{b} = \beta + C\varepsilon$ 。所以可以得到$\tilde{b}$ 的协方差矩阵是

$$Var[\tilde{b}] = \sigma^2 CC'$$

现在令 $D = C - (X'X)^{-1}X'$，由假设知 D ≠ 0。那么，$b^* = \tilde{b} - b = Dy,$ $Var(b^*) = D\sum_Y D' = \sigma^2 DD'$，于是 DD' 是非负定矩阵。

则

$$Var[\tilde{b}] = \sigma^2[(D + (X'X)^{-1}X')(D + (X'X)^{-1}X')']$$

$$= \sigma^2[(D + (X'X)^{-1}X')(D' + X(X'X)^{-1})]$$

$$= \sigma^2(DD' + (X'X)^{-1})$$

在展开这个四项和式之前，我们注意到

$$I = CX = DX + (X'X)^{-1}(X'X)$$

由于上面最后一项是 I，有 DX=0，所以

$$Var[\tilde{b}] = \sigma^2 DD' + \sigma^2(X'X)^{-1}$$

$$= Var[b] + \sigma^2 DD'$$

$\tilde{b}$ 的方差矩阵等于 b 的方差矩阵加上一个非负定矩阵。所以，$Var[\tilde{b}]$ 的每个二次型都大于$Var[b]$的相应二次型。

利用这个结果可以证明高斯-马尔科夫定理：

高斯—马尔科夫定理：

对任意常向量 w，古典线性模型中 $w'\beta$ 的最小方差线性无偏估计量是 $w'b$，其中 b 是最小二乘估计量。

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 X 是来自某种概率分布的情况中去。获得 b 的统计特性的一个方便的方法是，首先，第一步求得对 X 的条件期望结果，这等同于非随机回归量的情况，第二步，通过条件分布得到无条件结果。此论点的关键是，如果我们对任意 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。

因为　　$b=\beta+(X'X)^{-1}X'\varepsilon$

所以，以观测到的 X 为条件我们得到

$$E[b\mid X]=\beta+(X'X)^{-1}X'E[\varepsilon\mid X]=\beta+(X'X)^{-1}X'0=\beta$$

一个有用的方法是利用重期望定律

$$E[b]=E_x[E[b\mid X]]$$

$$=\beta+E_x[(X'X)^{-1}X'E[\varepsilon\mid X]]$$

因为由假定 4 有 $E[\varepsilon\mid X]=0$，所以，b 也是无条件无偏的，这样，

$$E[b]=E_x[E[b\mid X]]=E_x[\beta]=\beta。$$

同样，以 X 为条件的 b 的方差是

$$Var[b\mid X]=\sigma^2(X'X)^{-1}$$

为了求得确切的方差，我们使用方差分解公式：

$$Var[b]=E_x[Var[b\mid X]]+Var_x[E[b\mid X]]$$

由于对所有 X，$E[b\mid X]=\beta$，所以第二项为零，因此，

$$Var[b]=E[\sigma^2(X'X)^{-1}]=\sigma^2E[(X'X)^{-1}]$$

我们原来的结论要稍作改变，我们必须用其期望值 $E[(X'X)^{-1}]$来代替原来$(X'X)^{-1}$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 $\tilde{\beta}\neq b$，在 X 给定的条件下有

$$Var[b\mid X]\le Var[\tilde{\beta}\mid X]$$

但若这一不等式对一特定 X 成立，则必须成立：

$$Var[b] = E_x[Var[b \mid X]]$$

即，若它对每一特定 X 成立，则它一定对 X 的平均值也成立。这暗示，$Var(b) \leqslant Var(\tilde{\beta})$。

所以，不论我们是否将 X 看作是随机的，即无偏性和高斯—马尔科夫定理都成立。

四、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ε 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。

1、回归系数的假设检验

我们先讨论 X 非随机变量时的情况。

在（4）中，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

$$b \sim N[\beta, \sigma^2 (X'X)^{-1}]$$

这是一个多重正态分布，所以 b 的每一元素的边际分布都是正态分布的：

$$b_k \sim N[\beta_k, \sigma^2 (X'X)^{-1}_{kk}]$$

令 S_{kk} 是 $(X'X)^{-1}$ 的第 k 个对角元素，则

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S_{kk}}} \qquad (5)$$

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）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

$$e = y - Xb$$

$$= y - X(X'X)^{-1} X'y$$

$$= (I_n - X(X'X)^{-1} X')y$$

$$= My$$

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 M 既是对称的(M=M′)又

是幂等的（M=M^2）。

性质 1：X′ *e*=0 和 *i*′ *e*=0

证明：由正则方程组，我们得到：

$$X'e = X'(Y - Xb)$$

$$= X'(Y - X(X'X)^{-1}X'Y) = X'Y - X'X(X'X)^{-1}X'Y = X'Y - X'Y = 0$$

所以， *i*′ *e*=0

由性质 1 及证明过程我们得到两个推论：

推论 1： $X'M = 0$ 和 *MX*=0。

推论 2：$i'M = 0$ 和 *Mi*=0。

推论 2 成立是因为 X′ 的第一行是（1,1,…，1）。

性质 2：e 和 b 互不相关。

$$\mathrm{cov}(e,b) = [I_n - X(X'X)^{-1}X']\mathrm{cov}(Y,Y)[(X'X)^{-1}X']'$$

$$= \sigma^2[I_n - X(X'X)^{-1}X'][X(X'X)^{-1}] = 0$$

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。

性质 3：残差 e 的均值向量和协方差阵分别是 $E(e) = 0$和$Var(e) = \sigma^2 M$

证明： $E(e) = E(Y - Xb) = EY - E(Xb) = X\beta - E(X(X'X)^{-1}X'\varepsilon + X\beta = 0$

$$Var(e) = E(My(My)')$$

$$= ME(yy')M = \sigma^2 M^2 = \sigma^2 M$$

$$= \sigma^2\{I_n - X(X'X)^{-1}X'\}$$

E(e)=0，暗示 $\hat{y} = Xb$ 是 *y* 的无偏估计量。

性质 4： $E[e'e] = (n-K)\sigma^2$

证明：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：

$$e'e = \varepsilon' M' M e = \varepsilon' M^2 \varepsilon = \varepsilon' M \varepsilon$$

这个二次型的期望值是

$$E[e'e] = E[\varepsilon' M \varepsilon]$$

我们有 $E(e'e) = E(tr(e'e)) = E[tr(\varepsilon' M \varepsilon)] = E[tr(M \varepsilon \varepsilon')]$

由于 M 是固定的，这就是

$$tr(ME[\varepsilon\varepsilon']) = tr(M\sigma^2 I) = \sigma^2 tr(M)$$

M 的迹是

$$tr[I_n - X(X'X)^{-1}X'] = tr(I_n) - tr((X'X)^{-1}X'X)$$

$$= tr(I_n) - tr(I_K) = n - K$$

所以，

$$E[e'e] = (n-K)\sigma^2,$$

σ^2 的一个无偏估计量是

$$s^2 = \frac{e'e}{n-K} \qquad (6)$$

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：

$$Est.Var[b] = s^2 (X'X)^{-1}$$

通过利用 s^2 替代 σ^2，我们导出替代（5）中 z_k 的一个统计量。此量

$$\frac{(n-K)s^2}{\sigma^2} = \frac{e'e}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right)$$

是一个标准正态向量 (ε/σ) 的幂等二次型，所以，它服从自由度为秩（M）=迹（M）=n—K 的 x^2 分布。（6）中的 x^2 分布变量独立于（4）中的标准正态变量，为了证明这一点，只要证明

$$\frac{b-\beta}{\sigma} = (X'X)^{-1}X'\left(\frac{\varepsilon}{\sigma}\right) \qquad (7a)$$

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 Lx 和一个幂等二次型 x′ Ax 独立的充分条件是 LA=0，令 ε/σ 等 x，我们发现这里所需求的是

$(X'X)^{-1}X'M=0$。这确实成立，因为 $X'M=0$。

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 ε 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

$$t_k=\frac{(b_k-\beta_k)/\sqrt{\sigma^2 S_{kk}}}{\{[(n-K)s^2/\sigma^2]/(n-K)\}^{1/2}}$$

$$=\frac{b_k-\beta_k}{\sqrt{s^2 S_{kk}}} \qquad (7)$$

服从自由度为（n—K）的 t 分布。这是我们作统计推断的基础。

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（7）中的检验统计量。假定我们的假设是

$$H_0: r_1\beta_1+r_2\beta_2+\cdots+r_K\beta_K=r'\beta=q,$$

（通常某些 r 将为零）左边的样本估计是

$$r_1b_1+r_2b_2+\cdots+r_Kb_K=r'b=\hat{q}$$

若 $\hat{q}$ 显著异于 q，则我们推断样本数据与假设不一致。与（7）一样，将假设基于下式是很自然的。

$$t=\frac{\hat{q}-q}{se(\hat{q})} \qquad (7a)$$

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 $s^2(X'X)^{-1}$，我们可用下式估计 $\hat{q}$ 的方差。

$$Est.Var[\hat{q}]=r'[s^2(X'X)^{-1}]r$$

（7）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。

2、随机 X 及正态 ε 下的检验统计量

现在，我们考虑当 X 是随机的，样本检验统计量和推断方法考虑（7）中检验 $H_0:\beta_k=\beta_k^0$ 的 t 统计量：

$$t\mid X=\frac{(b_k-\beta_k^0)}{[s^2(X'X)_{kk}^{-1}]^{1/2}} \qquad (8)$$

以 X 为条件，t|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，（7a）仅仅在以 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 X 是随机的情况下，在给定 X 的条件下，我们得到了（8）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 X 的分布是什么，甚至不论 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|X)不是 X 的函数这一事实，同样的原因可以用来推演不论 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。

3、**拟合优度和方差分析**

由方差分解公式，我们有：$Var(Y)=Var(E(Y\mid X))+E_x(Var(Y\mid X))$。我们用幂等矩阵 M^0 来表示：

$$Y'M^0Y=E'(Y\mid X)M_0E(Y\mid X)+e'e$$

$$Y'M^0Y=b'X'M_0Xb+Y'MY$$

$$SST=SSR+SSE$$

所以，$SSE=Y'MY$ 和 $SSR=SST-SSE=Y'(M^0-M)Y$

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β =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 $x^2(K-1)$；

b）残差平方和 $\frac{SSE}{\sigma^2}$ 服从自由度为 n－K 的卡方分布 $x^2(n-K)$；

c）**在β =0 的假设条件下，** $\frac{SSR/(K-1)}{SSE/(n-K)}$ 服从 F（k-1,n－k）分布。

证明：a）M^0－M 是幂等矩阵。先证明 $M^0M+MM^0=2M$。

M^0M+MM^0

$$=(I-\frac{1}{n}ii')M+M(I-\frac{1}{n}ii')$$

$$=M-\frac{1}{n}ii'M+M-\frac{1}{n}Mii'$$

$=2M$

从而 $(M^0-M)(M^0-M)=M^0-(M^0M+MM^0)+M$

$$=M^0-2M+M=M^0-M$$

所以，$r(M^0-M)=tr(M^0-M)=tr(M^0)-tr(M)=n-1-(n-K)=K-1$。

在β =0 的假设条件下， $\frac{SSR}{\sigma^2}$ 才服从自由度为 K－1 的卡方分布 $x^2(K-1)$（**为什么？**）

b）因为 M 是幂等矩阵而且 $r(M)=tr(M)=n-K$

c）只要验证 $M(M^0-M)=0$ 即可。

事实上，$M(M^0-M)=M(I-\frac{1}{n}ii'-M)$

$$=M-\frac{1}{n}Mii'-M^2=-\frac{1}{n}Mii'=0$$。

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。

决定系数 $R^2=\frac{SSR}{SST}=\frac{b'X'M^0Xb}{Y'M^0Y}=1-\frac{e'e}{Y'M^0Y}$

进一步推导，我们可以得到 R^2 另一个公式。

$b'X'M^0Xb=\hat{Y}'M^0\hat{Y}, Y=\hat{Y}+e$，以及 $M^0e=e$（表示残差已经具有零均值）和 X′ e=0。

$\bar{Y}=\bar{\hat{Y}}$

所以，$\hat{Y}'M^0\hat{Y}=\hat{Y}'M^0Y-\hat{Y}'M^0e=\hat{Y}'M^0Y-b'X'e=\hat{Y}'M^0Y$

$$R^2=\frac{\hat{Y}'M^0\hat{Y}}{Y'M^0Y}=\frac{(\hat{Y}'M_0\hat{Y})^2}{Y'M^0Y\cdot\hat{Y}'M_0\hat{Y}}$$

$$=\frac{(\hat{Y}'M_0M_0Y)^2}{Y'M^0Y\cdot\hat{Y}M^0\hat{Y}}$$

$$= \frac{\left[\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2}{\left[\sum_i (y_i - \bar{y})^2\right]\left[\sum_i (\hat{y}_i - \bar{y})^2\right]} = r_{Y\hat{Y}}{}^2$$

第一个方法度量了 y 的总变差中由回归变差所解释的部分，第二个是 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 $\overline{R}^2$，它的表达式为

$$\overline{R}^2 = 1 - \frac{SSE/(n-K)}{SST/(n-1)} = 1 - \frac{e'e/(n-K)}{SST/(n-1)}$$

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

这里 $\frac{e'e}{n-K}$ 是 σ^2 的无偏估计量，（**思考：**当 y 服从正态分布时，$\frac{y'y - n\bar{y}^2}{n-1}$ 也是 σ^2 的一个无偏估计量）。$\overline{R}^2$ 与 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 $M^0 Y = Y - \overline{Y}$

所以，$SSR = b'X'M^0 Xb = b'X'M^0 Y$

$$= b'X'Y - b'X'\overline{Y}$$

$$= b'X'Y - \hat{Y}'\overline{Y} = b'X'Y - n\overline{Y}^2$$

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 1　　多元线性回归模型方差分析

	来源	自由度	均方
回归	$b'X'y - n\bar{y}^2$	$K-1$	
残差	$e'e$	$n-K$	s^2
总	$y'y - n\bar{y}^2$	$n-1$	$S_{yy(n-1)}$
$F[K-1, n-K] = \frac{SSR/(K-1)}{SSE/(n-K)} = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$			

4、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。**

五、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

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

（$\varepsilon^0 \sim N(0,\sigma^2)$，且 $E\varepsilon^0\varepsilon_i=0$, i=1,…,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

$$e^0=y^0-\hat{y}^0=(\beta-b)'x^0+\varepsilon^0$$

（$\varepsilon^0 \sim N(0,\sigma^2)$，且 $E\varepsilon^0\varepsilon_i=0$, i=1,…,n）

这个估计的预测方差是

$$Var[e^0]=\sigma^2+Var[(\beta-b)'x^0]$$

$$=\sigma^2+x^{0'}[\sigma^2(X'X)^{-1}]x^0$$

若回归含有一个常数项，一个等价的表达式是

$$Var[e^0]=\sigma^2+\frac{\sigma^2}{n}+\sigma^2\left\{\sum_{j=2}^{K}\sum_{k=2}^{K}(x_j^0-\bar{x}_j)(x_k^0-\bar{x}_k)(\underline{X}'M^0\underline{X})_{jk}\right\}$$

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。

因此 $\frac{y_0-\hat{y}_0}{\sqrt{\sigma^2(1+x'^0(X'X)^{-1}x^0)}}\sim N(0,1)$

又因为 $\frac{(n-K)S^2}{\sigma^2}\sim x^2(n-K)$

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：

预测区间 $=\hat{y}^0\pm t_{\lambda/2}se(\hat{y}^0)$。

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

$$Var[e^0]=Var[(\beta-b)'x^0]$$

$$=x^{0'}[\sigma^2(X'X)^{-1}]x^0$$

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 $\frac{(n-K)S^2}{\sigma^2}\sim x^2(n-K)$

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：

预测区间 $=\hat{y}^0\pm t_{\lambda/2}se(\hat{y}^0)$。

六、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 X1,那么现在的回归方程包括两组变量 X_1 和 X_2，转换为:

$$y=X\beta+\varepsilon=X_1\beta_1+X_2\beta_2+\varepsilon$$

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

(1a)

(2a)
$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}\begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix}=\begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}.$$

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（2a）以求解 b_2^*。我们首先从（1a）求得解 b_1：

$$b_1^*=(X_1'X_1)^{-1}X_1'y-(X_1'X_1)^{-1}X_1'X_2b_2^*$$

$$=(X_1'X_1)^{-1}X_1'(y-X_2b_2^*). \qquad (9)$$

（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（2a）得到

$$X_2'X_1(X_1'X_1)^{-1}X_1'y-X_2'X_1(X_1'X_1)^{-1}X_1X_2b_2^*+X_2'X_2b_2^*=X_2'y.$$

整理各项后，

$$X_2'(I-X_1(X_1'X_1)^{-1}X_1')X_2b_2^*=X_2'(I-X_1(X_1'X_1)^{-1}X_1')y.$$

解是

$$b_2^*=[X_2'(I-X_1(X_1'X_1)^{-1}X_1')X_2]^{-1}[X_2'(I-X_1(X_1'X_1)^{-1}X_1')]y$$

$$=(X_2'M_1X_2)^{-1}(X_2'M_1y). \qquad (10)$$

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_1X_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（10）重写为

$$b_2^* = (X_2^{*/} X_2^*)^{-1} X_2^* y^*, \qquad (11)$$

其中

$$X_2^* = M_1 X_2 \qquad \text{和} \qquad y^* = M_1 y.$$

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。

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，（7）中用于检验系数等于 0 的t比率，可用于计算

$$r_{YE}^{*2}=\frac{t_k^2}{t_k^2+\text{自由度}} \qquad (12)$$

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。

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

$$X_1'X_1b_1+X'X_2b_2=X'y.$$

我们已经解出了b_2，所以，在求解b_1时可以使用它：

$$b_1=(X_1'X_1)^{-1}X_1'y-(X_1'X_2b_2=(X_1'X_1)^{-1}X_1'(y-X_2b_2). \qquad (13)$$

若X_1仅为一列，（13）中第一个将产生如下结果

$$b_1=y-x_2b_2-\cdots-x_kb_k. \qquad (14)$$

这我们以前已经见到过。

七、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。

$$\mu_r = E[\varepsilon^r]$$

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）：

$$S =: \frac{[E(X-u_x)^3]^2}{[E(X-u_x)^2]^3} \qquad (15)$$

$$= \frac{\text{三阶矩的平方}}{\text{两阶矩的立方}}$$

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）：

$$K =: \frac{E(X-u_x)^4}{[E(X-u_x)^2]^2} \qquad (16)$$

$$= \frac{\text{三阶矩的平方}}{\text{两阶矩的立方}}$$

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 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。

样本偏度与样本峰度

根据式（15）和式（16），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：

$$\frac{\sum(X-\bar{X})^3}{n-1} \qquad (17)$$

样本四阶矩为：

$$\frac{\sum(X-\bar{X})^4}{n-1} \qquad (18)$$

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 $E[\varepsilon^3]$ 为 0。分布对称性的标准量是偏态（Skewness）

$$\sqrt{\beta_1} = \frac{E[\varepsilon^3]}{(\sigma^2)^{3/2}}.$$

峰态（Kurtosis）是分布尾部厚度的度量。此度量是

$$\beta_2 = \frac{E[\varepsilon^4]}{(\sigma^2)^2}.$$

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

$$W = n\left[\frac{b_1}{6} + \frac{(b_2-3)^2}{24}\right] \sim x^2(2).$$

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 x^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 x^2 表。

由贝拉和哈尔克（1980a，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。

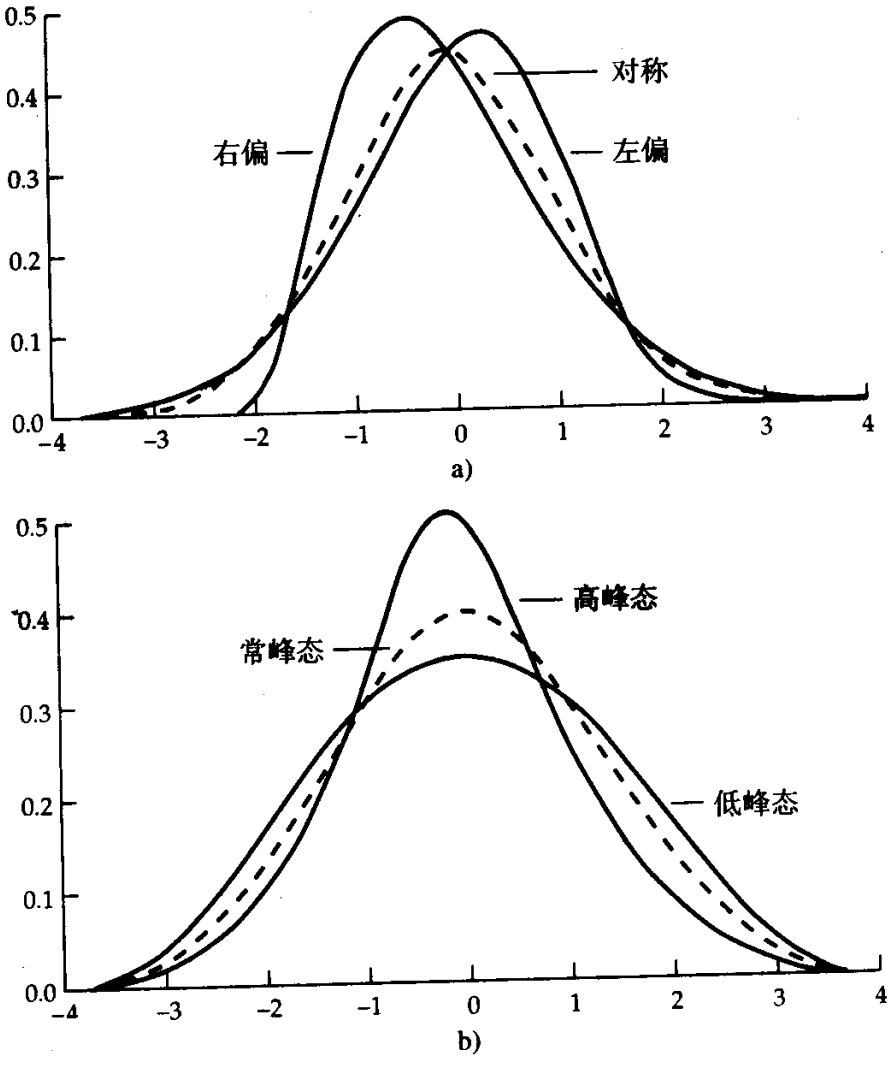


图 1

思考题

1、对于线性统计模型

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

假设$\varepsilon \sim N(0, \sigma^2 I_n), n = 13, K = 3$，最小化误差平方和$(y - X\beta)'(y - X\beta)$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$y'y = 53$,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 y 关于 X 的回归的最小二乘估计量，c 是另一 K×1 向量，证明两个残差平方和之差是

$$(y - Xc)'(y - Xc) - (y - Xb)'(y - Xb) = (c - b)'X'X(c - b)$$

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$\hat{\theta}_1$和$\hat{\theta}_2$，它们的方差分别为υ_1和υ_2,并且$\upsilon_1 \neq \upsilon_2$。那么什么样的线性组合$\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$是$\theta$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$\hat{\theta}_1 \cdots\cdots \hat{\theta}_n$，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$\theta$的最小方差无偏估计量？

第七章　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，Rβ =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，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。

第一节　线性约束的检验

从线性回归模型开始，

$$y = X\beta + \varepsilon \qquad (1)$$

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

$$R\beta = q \qquad (2)$$

作为我们的假设条件 H_0。

R 中每一行都是一个约束中的系数。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。因此，J 一定要小于或等于 K。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β ，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 b，我们的兴趣集中于“差异”向量 d=Rb－q。d 精确等于 0 是不可能的事件（因为其概率是 0），统计问题是 d 对 0 的离差是否可归因于抽样误差或它是否是显著的。

由于 b 是多元正态分布的，且 d 是 b 的一个线性函数，所以 d 也是多元正态分布的，若原假设为真，d 的均值为 0，方差为

$$Var[d]=Var[Rb-q]=R(Var[b])R'=\sigma^2R(X'X)^{-1}R' \qquad (3)$$

对 H_0 的检验我们可以将其基于沃尔德（Wald）准则：

$$W=\chi^2(J)=d'(Var[d])^{-1}d$$

$$=(Rb-q)'[\sigma^2R(X'X)^{-1}R']^{-1}(Rb-q) \qquad (4)$$

在假设正确时将服从自由度为 J 的 χ^2 分布(为什么?)。

直觉上，d 越大，即最小二乘满足约束的错误越大，则 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。

$$\frac{(n-K)s^2}{\sigma^2}=\frac{e'e}{\sigma^2}=\left(\frac{\varepsilon}{\sigma}\right)'M\left(\frac{\varepsilon}{\sigma}\right) \qquad (5)$$

由于σ 未知，（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

$$F=\frac{(Rb-q)'[\sigma^2R(X'X)^{-1}R']^{-1}(Rb-q)/J}{[(n-K)s^2/\sigma^2]/(n-K)} \qquad (6)$$

分子是（1/J）乘（4）中的 W，分母是 1/（n－K）乘（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 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（5）及 M 是幂等的这一事实，我们可以把 F 写为

$$F=\frac{\{R(b-\beta)/\sigma\}'[R(X'X)^{-1}R']^{-1}\{R(b-\beta)/\sigma\}/J}{[M(\varepsilon/\sigma)]'[M(\varepsilon/\sigma)]/(n-K)} \qquad (7)$$

由于

$$\frac{R(b-\beta)}{\sigma}=R(X'X)^{-1}X'\left(\frac{\varepsilon}{\sigma}\right)=T\left(\frac{\varepsilon}{\sigma}\right)$$

F 统计量是(ε/σ)的两个二次型的比率，由于 M(ε/σ)和 T(ε/σ)都服从正态分布且它们的协方差 TM 为 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（6）中的两个σ^2，剩下的是检验一个线性假设的 F 统计量，

$$F=\frac{(Rb-q)'[R(X'X)^{-1}R']^{-1}(Rb-q)/J}{e'e/(n-K)}$$

$$=\frac{(Rb-q)'[s^2R(X'X)^{-1}R']^{-1}(Rb-q)}{J} \qquad (8)$$

我们将检验统计量

$$F[J,n-K]=\frac{(Rb-q)'\{R[s^2(X'X)^{-1}]R'\}^{-1}(Rb-q)}{J}$$

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。

第二节　参数带有约束的最小二乘估计

一、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β 满足某特定的线性约束条件：Rβ =q，这里 R 是 J×K 矩阵（J＜K），并假定它的秩为 J 维向量，常常希望求β 的估计$\hat{\beta}$，使得

$$\left\|Y-X\hat{\beta}\right\|^2=\min_{\{\beta:R\beta=q\}}\left\|Y-X\beta\right\|^2 \qquad (9)$$

满足条件（9）的称为β 的具有线性约束 Rβ =q 的最小二乘估计。

解$\hat{\beta}$的问题实际上是在约束条件

Rβ =q

下求　　$f=\left\|Y-X\beta\right\|^2=\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{m}x_{ij}\beta_j\right)^2$

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

$$\frac{\partial S^*}{\partial\beta}=-2X'(y-Xb_*)+2R'\lambda=0$$

$$\frac{\partial S^*}{\partial \lambda} = 2(Rb_* - q) = 0$$

展开可以得到分块矩阵方程

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} b_* \\ \lambda \end{bmatrix} = \begin{bmatrix} X'y \\ q \end{bmatrix}$$

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

$$d_*=W^{-1}v$$

$$= \begin{bmatrix} b^* \\ \lambda \end{bmatrix}$$

where

$$W^{-1} = \begin{pmatrix} (X'X)^{-1} - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}R(X'X)^{-1} & (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1} \\ (R(X'X)^{-1}R')^{-1}R(X'X)^{-1} & -(R(X'X)^{-1}R')^{-1} \end{pmatrix}$$

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 b_*的协方差矩阵简单地就是σ^2 乘以 W^{-1} 的左上块，在 X′ X 是非奇异的通常情况下，再一次可以得到一个显性公式

$$Var[b_*] = \sigma^2(X'X)^{-1} - \sigma^2(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1},$$

这样，

$$Var[b_*] = Var[b] - \text{（一个非负定矩阵）},$$

Var[b_*]的方差比 Var[b]小的一个解释是约束条件提供了更多的信息价值。

二、对约束的检验的另一个方法

令 $e_* = y - Xb_*$，我们来计算新的离差平方和 $e_*'e_*$。

$$e_* = y - Xb - X(b_* - b) = e - X(b_* - b)$$

则新的离差平方和是

$$e_*'e_* = e'e + (b_* - b)'X'X(b_* - b) \geq e'e$$

$$\frac{e'e}{\sigma^2} \sim \chi^2_{n-k} \qquad \frac{e_*'e_*}{\sigma^2} \sim \chi^2_{n-(k-J)}$$

因为新的模型中参数的个数为 k-J 个，J 个约束条件是原模型中的 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 $X'e$，它是 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，

$$e_*'e_* - e'e = (Rb - q)'[R(X'X)^{-1}R']^{-1}(Rb - q)$$

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

$$F[J, n-K] = \frac{(e_*'e_* - e'e)/J}{e'e/(n-K)}$$

最后，以 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。

[实例]对数变换生产函数

所有科布—道格拉斯模型的一般化是如下的对数变换模型，

$$\ln Y = \beta_1 + \beta_2 \ln L + \beta_3 \ln K + \beta_4 \frac{\ln^2 L}{2} + \beta_5 \frac{\ln^2 K}{2} + \beta_6 \frac{\ln L \ln K}{2} + \varepsilon \qquad (10)$$

无约束回归的结果在表 1 中给出。

表 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	
变量	系数	标准误差	t 值
常数项	0.944216	2.911	0.324
LnL	3.61363	1.548	2.334
LnK	－1.89311	1.016	－1.863
$\frac{1}{2}\ln^2 L$	－0.96406	0.7074	－1.363
$\frac{1}{2}\ln^2 K$	0.08529	0.2926	0.291
lnL×lnK	0.31239	0.4389	0.71

系数估计量的估计协方差矩阵

	常数项	lnL	lnK	Ln2L/2	Ln2K/2	lnL×lnK
常数项	8.472					
LnL	－2.388	2.397				
LnK	－0.3313	－1.231	1.033			
$\frac{1}{2}\ln^2 L$	－0.08760	－0.6658	0.5231	0.5004		
$\frac{1}{2}\ln^2 K$	0.2332	0.03477	0.02637	0.1467	0.08562	
lnL×lnK	0.3635	0.1831	－0.2255	－0.2880	－0.1160	0.1927

考虑了约束条件 $\beta_4=\beta_5=\beta_6=0$ 的模型就可以得到科布一道格拉斯模型：

$$\ln Y=\beta_1+\beta_2\ln L+\beta_3\ln K+\varepsilon \qquad (11)$$

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：

$e_*'e_*=0.85163$，而且 n－K=21，则科布—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，所以我们不能拒绝科布—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科

布—道格拉斯生产函数。这个模型可以推导如下：

$$\begin{aligned}\ln Y &= \beta_1 + \beta_2 \ln L + \beta_3 \ln K + \varepsilon \\ &= \beta_1 + \beta_2 \ln L + (1-\beta_2)\ln K + \varepsilon\end{aligned} \qquad (12)$$

$$\therefore \ln Y - \ln L = \beta_1 + \beta_2(\ln L - \ln K) + \varepsilon$$

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，则科布—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，所以我们不能拒绝科布—道格拉斯模型是规模效应的生产函数的这一假设。

第三节　结构变化与邹至庄检验

(Structure Change and Chou-Test)

一、　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）。

二、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个回归方程

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \cdots\cdots (1)$$

即 $Y = Z\beta + \varepsilon$ 模型，其中 Y=$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$，Z=$\begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix}$，β =$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$，ε =$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$。

上述问题就转换成检验 $\begin{matrix} H_0 : \beta_1 = \beta_2 \\ H_1 : \beta_1 \neq \beta_2 \end{matrix}$ 的问题。

我们可以用两种方式来处理问题

一）用约束条件 $\beta_1 = \beta_2$，来检验。$\beta_1 = \beta_2$ 是更一般约束条件 Rβ =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）。

例题：用约束条件下，F 检验推导出邹至庄检验的表达式：

解：在约束条件 Rβ =q 下，F 检验

$$F(J,n-k)=\frac{(Rb-q)'[S^2R(Z'Z)^{-1}R']^{-1}(Rb-q)}{J}$$。

而邹至庄检验时约束条件 Rβ =q 的一种特殊形式，即 R=(I,-I)，而 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

$$F(k,n_1+n_2-2k)=\frac{(Rb-q)'[S^2R(Z'Z)^{-1}R']^{-1}(Rb-q)}{k}$$

$$=\frac{(b_1-b_2)'[S^2(I,-I)\begin{pmatrix}(Z_1'Z_1)^{-1} & 0\\ 0 & (Z_2'Z_2)^{-1}\end{pmatrix}\begin{pmatrix}I\\ -I\end{pmatrix}]^{-1}(b_1-b_2)}{k}$$

$$=\frac{(b_1-b_2)'[S^2\left((Z_1'Z_1)^{-1}+(Z_2'Z_2)^{-1}\right)]^{-1}(b_1-b_2)}{k}$$

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（1）改写成一个无约束的新的回归方程

$$\begin{bmatrix}Y_1\\ Y_2\end{bmatrix}=\begin{bmatrix}Z_1 & 0\\ 0 & Z_2\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_1\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix},\quad \text{即}$$

$$\begin{bmatrix}Y_1\\ Y_2\end{bmatrix}=\begin{bmatrix}Z_1\\ Z_2\end{bmatrix}\beta_1+\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix}\qquad (2)$$

即无约束的线性模型 $Y=Z\beta+\varepsilon$ 模型，其中 Y=$\begin{pmatrix}Y_1\\ Y_2\end{pmatrix}$，Z=$\begin{pmatrix}Z_1\\ Z_2\end{pmatrix}$，$\beta$ β =$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$，ε =$\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}$。

假如模型（2）的残差平方和是 $e_*'e_*$，在假设条件 $\beta_1=\beta_2$ 下， 我们可以得到 F 统计量可更简单地表示为：

$$F[k,n_1+n_2-2k]=\frac{(e_*'e_*-e'e)/k}{e'e/(n_1+n_2-2k)}$$。

二） 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（1）就转换为：

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \cdots\cdots（2）$$

由此我们推导出可以检验的邹至庄统计量 Chou-Test。

从模型（1）中，我们可以得到无约束最小二乘估计量是

$$b = (Z'Z)^{-1}Z'Y = \begin{pmatrix} Z_1'Z_1 & 0 \\ 0 & Z_2'Z_2 \end{pmatrix}^{-1} \begin{pmatrix} Z_1'Y_1 \\ Z_2'Y_2 \end{pmatrix} = \begin{pmatrix} (Z_1'Z_1)^{-1} & 0 \\ 0 & (Z_2'Z_2)^{-1} \end{pmatrix} \begin{pmatrix} Z_1'Y_1 \\ Z_2'Y_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

故

$$\begin{aligned} e = Y - Zb &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} - \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix} \begin{pmatrix} (Z_1'Z_1)^{-1}Z_1'Y_1 \\ (Z_2'Z_2)^{-1}Z_2'Y_2 \end{pmatrix} \\ &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} - \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix} \begin{pmatrix} (Z_1'Z_1)^{-1}Z_1' & 0 \\ 0 & (Z_2'Z_2)^{-1}Z_2' \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &= \left(I - \begin{pmatrix} Z_1(Z_1'Z_1)^{-1}Z_1' & 0 \\ 0 & Z_2(Z_2'Z_2)^{-1}Z_2' \end{pmatrix} \right) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \\ &\because M_1 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \end{aligned}$$

$$e'e = (Y_1'Y_2')M_1'M_1 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = (Y_1'Y_2')M_1 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

则 $\frac{e'e}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2k) \cdots\cdots$（3）

对于有约束条件 $\beta_1 = \beta_2$ 限制的模型（2）

$$\begin{aligned} e^* &= \left(I - \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \left((Z_1'Z_2') \cdot \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \right)^{-1} (Z_1'Z_2') \right) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &= \left(I - \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \left(Z_1'Z_1 + Z_2'Z_2 \right)^{-1} (Z_1'Z_2') \right) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &\because M_2 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \end{aligned}$$

$$e^{*'}e^* = (Y_1'Y_2')M_2'M_2 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = (Y_1'Y_2')M_2 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

则 $\frac{e^{*'}e^*}{\sigma^2} \sim \chi^2(n_1 + n_2 - k) \cdots\cdots$（4）

$$e^{*'}e^* - e'e = (Y_1'Y_2')(M_2 - M_1) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \because (Y_1'Y_2')M_3 \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

问 $\frac{e^{*'}e^* - e'e}{\sigma^2}$ 服从何分布？

首先证明： $M_3M_1 = 0$

$$
\begin{aligned}
(M_2 - M_1)M_1 &= M_2M_1 - M_1^2 = M_2M_1 - M_1 \\
&= \left(I - \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \left(Z_1^{'}Z_1 + Z_2^{'}Z_2 \right)^{-1} (Z_1^{'}Z_2^{'}) \right) \cdot M_1 - M_1 \\
&= -\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \left(Z_1^{'}Z_1 + Z_2^{'}Z_2 \right)^{-1} (Z_1^{'}Z_2^{'}) \left(I - \begin{pmatrix} Z_1(Z_1^{'}Z_1)^{-1}Z_1^{'} & 0 \\ 0 & Z_2(Z_2^{'}Z_2)^{-1}Z_2^{'} \end{pmatrix} \right) \\
&= -\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \left(Z_1^{'}Z_1 + Z_2^{'}Z_2 \right)^{-1} (Z_1^{'}Z_2^{'}) + -\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \left(Z_1^{'}Z_1 + Z_2^{'}Z_2 \right)^{-1} (Z_1^{'}Z_2^{'}) \\
&= 0
\end{aligned}
$$

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 $\frac{e^{*'}e^{*} - e'e}{\sigma^2} \sim \chi^2(k)$ 且与 $\frac{e'e}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2k)$

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）。

第八章　正态线性统计模型的最大似然估计

我们假定第五章的 6 个假设条件全部满足，我们就知道了 Y 的分布函数，我们也就可以用其他方法如最大似纵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纵然估计求出参数 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 $\varepsilon \sim N(0,\sigma^2 I_n)$，因此有 $y \sim N(X\beta,\sigma^2 I_n)$ 或者边际分布 $y_t \sim N(x_t'\beta,\sigma^2)$，这里 $x_t' = (x_{t1}, x_{t2}, \cdots, x_{tK})$，$t$=1,2,⋯,$n$, **{$y_t$}是相互独立的（Why?）**。$y_t$ 的密度函数为

$$f(y_t \mid \beta,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{(y_t - x_t'\beta)^2}{2\sigma^2}\right\}$$

所以 $y_1,y_2,\cdots,y_n$ 的联合概率密度函数为

$$f(y_t \mid \beta,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\sum_{t=1}^{n}\frac{(y_t - x_t'\beta)^2}{2\sigma^2}\right\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}\right\}$$

如果把 $y_1,y_2,\cdots,y_n$ 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

$$l(\beta,\sigma^2 \mid y) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}\right\}$$

两边取自然对数得到**对数似然函数**为

$$L(\beta,\sigma^2 \mid y) = \ln l(\beta,\sigma^2 \mid y)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}$$

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：

$$\max_{\beta,\sigma^2} L(\beta,\sigma^2 \mid y)$$

这个问题的一阶条件为

$$\frac{\partial L}{\partial \beta}=-\frac{1}{2\sigma^2}\left[\frac{\partial (y-X\beta)'(y-X\beta)}{\partial \beta}\right]=-\frac{1}{2\sigma^2}(-2X'y+2X'X\beta)=0$$

$$\frac{\partial L}{\partial \sigma^2}=-\frac{n}{2}\frac{1}{\sigma^2}+\frac{(y-X\beta)'(y-X\beta)}{2\sigma^4}=0$$

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

$$X'X\hat{\beta}=X'y$$

$$\tilde{\sigma}^2=\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n}$$

可以解得

$$\hat{\beta}=(X'X)^{-1}X'y=b\text{（与最小乘估计量一样）}$$

$$\hat{\sigma}^2=\frac{y'My}{n}$$

这里$M=I_n-X(X'X)^{-1}X'$。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$(y-X\beta)'(y-X\beta)$是等价的。如果记

$$\hat{e}=y-X\hat{\beta}=My$$

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$\hat{e}=e=y-Xb$。这样σ^2可表示为

$$\hat{\sigma}^2=\frac{y'My}{n}=\frac{\hat{e}'\hat{e}}{n}=\frac{e'e}{n}$$

对于$\tilde{\beta}$，我们已知知道$E[\hat{\beta}]=\beta,\text{cov}(\hat{\beta})=\sigma^2(X'X)^{-1}$。由于$\hat{\beta}=(X'X)^{-1}X'y=b$是 y 的线性函数，而 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$\sigma^2(X'X)^{-1}$，也即$\hat{\beta}\sim N(\beta,\sigma^2(X'X)^{-1})$。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$E[\hat{e}'\hat{e}]=E[(y-X\hat{\beta})'(y-X\hat{\beta})]=E[\varepsilon'M\varepsilon]=\sigma^2(n-K)$，所以其期望值为

$$E[\hat{\sigma}^2]=\frac{n-K}{n}\sigma^2$$

它是σ^2的一个有偏估计量，**其偏度为**$-\frac{K}{n}\sigma^2$。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。

第九章　非线性回归模型

回归模型的一般形式是

$$y_i = h(x_i, \beta) + \varepsilon_i \qquad (1)$$

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（1）。

例如，

$$y = \beta_1 + \beta_2 e^{\beta_3 x} + \varepsilon \qquad (2)$$

不能变换到线性形式。

1　线性化回归

非线性回归模型是

$$y = h(x, \beta) + \varepsilon$$

（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 $h(x, \beta)$ 的一个线性泰勒级数来近似：

$$h(x, \beta) \cong h(x, \beta^0) + \sum_k \left.\frac{\partial h(x, \beta)}{\partial \beta_k}\right|_{\beta=\beta^0} (\beta_k - \beta_k^0) \qquad (3)$$

这被称为**线性化回归模型**。整理各项可得

$$h(x, \beta) \cong h(x, \beta^0) - \sum_k \beta_k^0 \left.\frac{\partial h(x, \beta)}{\partial \beta_k}\right|_{\beta=\beta^0} + \sum_k \beta_k \left.\frac{\partial h(x, \beta)}{\partial \beta_k}\right|_{\beta=\beta^0}$$

令 $\tilde{x}_k^0$ 等于第 k 个偏微分 $\partial h(x, \beta^0) / \partial \beta_k^0$。**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

$$h(x, \beta) \cong [h^0 - \sum_k \tilde{x}_k^0 \beta_k^0] + \sum \tilde{x}_k^0 \beta_k$$

$$= h^0 - \tilde{x}^{0\prime} \beta^0 + \tilde{x}^{0\prime} \beta$$

或

$$y \cong h^0 - \tilde{x}^{0\prime} \beta^0 + \tilde{x}^{0\prime} \beta + \varepsilon$$

把已知项移到方程左边，可得回归模型：

$$\tilde{y} = y - h^0 + \tilde{x}^{0'}\beta^0 = \tilde{x}^{0'}\beta + \varepsilon \qquad (4)$$

有了 β^0 值，我们就可以计算 $\tilde{y}^0$ 和 $\tilde{x}^0$ 并通过线性最小二乘法估计（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。

［例］ 对于（2）所给的非线性回归模型，线性化方程中的回归量是

$$\tilde{x}_1^0 = \frac{\partial h(.)}{\partial \beta_1} = 1, \qquad \tilde{x}_2^0 = \frac{\partial h(.)}{\partial \beta_2} = e^{\beta_3^0 x},$$

$$\tilde{x}_3^0 = \frac{\partial h(.)}{\partial \beta_3} = \beta_2^0 x e^{\beta_3^0 x}$$

有了一组参数 β^0

$$\tilde{y}^0 = y - h(x, \beta_1^0, \beta_2^0, \beta_3^0) + \beta_1^0 \tilde{x}_1^0 + \beta_2^0 \tilde{x}_2^0 + \beta_3^0 \tilde{x}_3^0$$

可以对前面为估计 β_1, β_2 和 β_3 而定义的三个变量进行回归。

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）X′X 收敛于一个正定矩阵 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

$$p\lim\left(\frac{1}{n}\right)\tilde{X}\tilde{X} = p\lim\left(\frac{1}{n}\right)\sum_i \left[\frac{\partial h(x_i, \beta^0)}{\partial \beta^0}\right]\left[\frac{\partial h(x_i, \beta^0)}{\partial \beta^{0'}}\right] = \tilde{Q} \qquad (5)$$

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。

（5）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。

非线性最小二乘准则函数是

$$S(b)=\sum_i [y_i - h(x_i,b)]^2 = \sum_i e_i^2$$

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，计算可以继续进行。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：

$$\hat{\sigma}^2=\left(\frac{1}{n}\right)\sum_i [y_i - h(x_i,b)]^2 \qquad (6)$$

（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 $\frac{n-K}{n}\to 1$ 的事实。

$$b \xrightarrow{a} N[\beta, \frac{\sigma^2}{n}Q^{-1}]$$

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

$$Est.Asy.Var[b]=\hat{\sigma}^2(\tilde{X}'\tilde{X})^{-1}$$

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。

第十章　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。

第一节　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.$y=X\beta+\varepsilon$ 。

Ⅱ.X 是秩为 K 的 n×K 非随机矩阵。

Ⅲ.$E[\varepsilon]=0$。

Ⅳ.$E[\varepsilon\varepsilon']=\sigma^2 I$。

未知参数β 和σ^2的最小二乘估计量是

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

和

$$s^2=\frac{e'e}{(n-K)}$$

通过分析

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

并且

$$s^2=\frac{\varepsilon'M\varepsilon}{n-K}$$

我们可得下列精确的有限样本结果：

1. $E[b]=\beta$ （最小二乘估计是无偏的）

2. $Var[b]=\sigma^2(X'X)^{-1}$

3. 任意函数 $r'\beta$ 的最小方差线性无偏估计量是 $r'b$。（这就是高斯—马尔科夫定理）

4. $E[s^2]=\sigma^2$

5. $Cov[b,e]=0$

为了构造置信区间和检验假设，我们根据正态分布的假设

$V.\varepsilon \sim N[0,\sigma^2 I]$

推导额外了的结果，即

6. b 和 e 在统计上是相互独立的。相应的，b 和 s^2 无关并在统计上相互独立。

7. b 的精确分布依赖于 X，是 $N[\beta,\sigma^2(X'X)^{-1}]$。

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 2σ ^{4}/（n－K）。

9. 根据 6 至 8 结果，统计量 $t[n-K]=\dfrac{b_k-\beta_k}{s^2(X'X)^{-1}_{kk})}$ 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 Rβ =q 的检验统计量

$$\frac{(Rb-q)'[R(X'X)^{-1}R']^{-1}(Rb-q)/J}{e'e/(n-K)}=\frac{(Rb-q)'[Rs^2(X'X)^{-1}R']^{-1}(Rb-q)}{J}$$

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 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）**

第二节　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：

$W=n\left[\dfrac{b_1}{6}+\dfrac{(b_2-3)^2}{24}\right]\sim x^2(2).$ 通不过检验，即假设 $V.\varepsilon \sim N[0,\sigma^2 I]$ 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，Y 和 ε 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捉相应的调整，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，而不再只是强调无偏性。

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。

最小二乘系数向量的一致性

复习：依概率分布

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。

证明：$E[\bar{x}]=\mu$及$Var[\bar{x}]=\sigma^2/n$，所以，$\bar{x}$ 依均方收敛于μ ，或 $p\lim \bar{x}=\mu$。

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

$$\lim_{n\to\infty}\frac{1}{n}X'X=Q \qquad \text{是正定矩阵，} \qquad (1)$$

这个假设在大多数时候是不过份的，考虑一元的情况：

$$\mathrm{X}=\begin{pmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix}$$

$$\frac{1}{n}X'X=\frac{1}{n}\begin{pmatrix}1 & 1 & \cdots & 1\\ x_1 & x_2 & \cdots & x_n\end{pmatrix}\begin{pmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix}=\frac{1}{n}\begin{pmatrix}n & \sum x_i\\ \sum x_i & \sum x_i^2\end{pmatrix}=\begin{pmatrix}1 & \bar{x}\\ \bar{x} & \frac{n-1}{n}s^2+\bar{x}^2\end{pmatrix}$$

(我们知道，$\mathrm{p}\lim \bar{x}=\mu$　　$\mathrm{p}\lim\frac{\sum x_i^2}{n}=\lim\left[\frac{n-1}{n}s^2+\bar{x}^2\right]=\sigma^2+\mu^2$).

$$\therefore \lim\frac{1}{n}X'X=\begin{pmatrix}1 & \mu\\ \mu & \sigma^2+\mu^2\end{pmatrix}$$

which is positive definite as its principal submatrices all have positive determinants.

最小二乘估计量可以写成

$$b=\beta+\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\varepsilon\right) \qquad (2)$$

假设 Q^{-1} 存在，因为逆矩阵是原矩阵的连续函数，我们得到

$$p\lim b=\beta+Q^{-1}p\lim\left(\frac{1}{n}X'\varepsilon\right)$$

现在我们需要最后一项的概率极限。令

$$\overline{w}=\frac{1}{n}X'\varepsilon=\frac{1}{n}\sum_i x_i\varepsilon_i=\frac{1}{n}\sum_i w_i$$，其中 $x_i=\begin{pmatrix}x_{i1}\\x_{i2}\\\vdots\\x_{ik}\end{pmatrix}$,为 X' 的列向量

那么

$$b=\beta+\left(\frac{1}{n}X'X\right)^{-1}\overline{w}$$

且

$$p\lim b=\beta+Q^{-1}p\lim\overline{w}$$

因为，X 是非随机矩阵，所以

$$E[\overline{w}]=\frac{1}{n}X'E[\varepsilon]=0$$

且

$$Var[\overline{w}]=E[\overline{w}\,\overline{w}']=\frac{1}{n}X'E[\varepsilon\varepsilon']X\frac{1}{n}=\frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)$$

于是可得

$$\lim_{n\to\infty}Var[\overline{w}]=0\cdot Q=0$$

由于 $\overline{w}$ 的均值是 0，并且它的方差收敛于 0，所以 $\overline{w}$ 按均方收敛于 0，且 $p\lim\overline{w}=0$。

（下面定理揭示了 r-阶收敛与依概率收敛的关系

定理 8 $\xi_n\xrightarrow{r}\xi\Rightarrow\xi_n\xrightarrow{P}\xi$。）

因此

$$p\lim\left(\frac{1}{n}X'\varepsilon=0\right)\qquad(4)$$

所以

$$p\lim b=\beta+Q^{-1}\cdot 0=\beta\qquad(5)$$

这表明了在古典回归模型中，在假设（1）条件下 b 是β 的一致估计量。

二、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

$$\sqrt{n}(b-\beta)=\left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon$$

由于逆矩阵是原矩阵的连续函数，$\lim_{n\to\infty}(X'X/n)^{-1}=Q^{-1}$。因此，如果极限分布存在，则统计量的极限分布与下式相同：

$$\left[\lim_{n\to\infty}\left(\frac{X'X}{n}\right)^{-1}\right]\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon=Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \qquad (6a)$$

因此，我们必须建立下式的极限分布，

$$\frac{1}{\sqrt{n}}X'\varepsilon=\sqrt{n}(\overline{w}-E[\overline{w}])$$

其中 $E[\overline{w}]=0$。我们可以利用林德伯格-费勒形式的中心极限定理得到 $\sqrt{n}\overline{w}$ 的极限分布。利用定理中的表达式，

$$\overline{w}=\frac{1}{n}\sum_i x_i\varepsilon_i$$

是 n 个互不相关的随机向量 $x_i\varepsilon_i$ 的平均值，其中 $x_i=\begin{pmatrix}x_{i1}\\x_{i2}\\\vdots\\x_{ik}\end{pmatrix}$，

ε_i 的均值为 0，方差为

$$Var[x_i\varepsilon_i]=\sigma^2x_ix_i'=\sigma^2Q_i$$

$\sqrt{n}\overline{w}$ 的方差

$$\sigma^2\overline{Q}_n=\sigma^2\left(\frac{1}{n}\right)[Q_1+Q_2+\cdots+Q_n]$$

$$=\sigma^2\left(\frac{1}{n}\right)\sum_i x_ix_i'=\sigma^2\left(\frac{X'X}{n}\right)$$

只要总和不被任一特定项占据主导地位并且回归量表现良好，在这种情况中，这意味着（1）成立，则

$$\lim_{n\to\infty}\mathrm{Var}\left(\frac{1}{\sqrt{n}}X'\varepsilon\right)=\lim_{n\to\infty}\mathrm{Var}(\sqrt{n}\overline{w})=\lim_{n\to\infty}\sigma^2\overline{Q}_n=\sigma^2Q$$

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. X 的元素受到限制使得 $|x_{tk}|$ 有限并且 $\lim(X'X/n)=Q$ 是一个有限正定矩阵。则

$$\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \xrightarrow{d} N[0,\sigma^2 Q] \qquad (6)$$

(这也是为什么我们要假设 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：

$$Q^{-1}\left(\frac{1}{\sqrt{n}}\right)X'\varepsilon \xrightarrow{d} N[Q^{-1}0, Q^{-1}(\sigma^2 Q)Q^{-1}]$$

根据(6a):

$$\sqrt{n}(b-\beta) \xrightarrow{d} N[0,\sigma^2 Q^{-1}]$$

我们可以得到 b 的渐近分布(不加证明)：

$$b \xrightarrow{a} N\left[\beta, \frac{\sigma^2}{n}Q^{-1}\right]$$

三、标准检验统计量的渐近行为

如果没有 ε 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

$$b \xrightarrow{a} N\left[\beta, \frac{\sigma^2}{n}Q^{-1}\right]$$

由此得出

$$\theta_k = \frac{b_k-\beta_k}{[(\sigma^2/n)Q_{kk}^{-1}]^{1/2}}$$

的渐近分布是标准正态分布。

由于 $p\lim s^2(X'X/n)^{-1}=\sigma^2 Q^{-1}$（在下一节中将证明 $p\lim s^2=\sigma^2$ 这个结果）

$$t_k = \frac{b_k-\beta_k}{[s^2(X'X)_{kk}^{-1}]^{1/2}}$$

将与 θ_k 有同样的渐近分布。因此，我们可以认为，关于β 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，

$$F=\frac{(e_*'e_*-e'e)/J}{e'e/(n-K)}=\frac{(Rb-q)'[R(s^2(X'X)^{-1}R']^{-1}(Rb-q)}{J}$$

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。

定理　沃尔德统计量的极限分布定理

如果 $\sqrt{n}(b-\beta)\xrightarrow{d}N[0,\sigma^2Q^{-1}]$

以及 $H_0:R\beta-q=0$ 是正确的，那么

$$W=(Rb-q)'[R(s^2(X'X)^{-1}R']^{-1}(Rb-q)=JF$$

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 **R=I**，而 **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。

定理的证明：由于 R 是常数矩阵，

$$\sqrt{n}R(b-\beta)\xrightarrow{d}N[0,R(\sigma^2Q^{-1})R'] \qquad (1)$$

又 Rβ =q，因此

$$\sqrt{n}(Rb-q)\xrightarrow{d}N[0,R(\sigma^2Q^{-1})R'] \qquad (2)$$

为方便起见，将此写成

$$z\xrightarrow{d}N[0,P] \qquad (3)$$

令 T 满足 $T^2=P^{-1}$，并把 T 记为 $P^{-\frac{1}{2}}$，即 T 是 P 的逆平方根。

如果 $z\xrightarrow{d}N[0,P]$，那么 $P^{-1/2}z\xrightarrow{d}N[0,P^{-1/2}PP^{-1/2}]=N[0,I]$ （4）

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

$$(P^{-1/2}z)'(P^{-1/2}z) = z'P^{-1}z \xrightarrow{d} \chi^2(J) \qquad (5)$$

再结合前面的各部分， 不难证明：

$$z'P^{-1}z = n(Rb-q)'[R(\sigma^2 Q^{-1})R']^{-1}(Rb-q) \xrightarrow{d} \chi^2(J) \qquad (6)$$

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 $P\lim s^2(X'X/n)^{-1} = \sigma^2 Q^{-1}$（在下一节中将证明这个结果），这样：

$n(Rb-q)'[R(s^2(X'X/n)^{-1})R']^{-1}(Rb-q)$ 的极限分布式与（6）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。

三、s^2 的一致性和 Var[b]的估计量

本节证明上节用到的结果 $\text{plim}\, s^2(X'X/n)^{-1} = \sigma^2 Q^{-1}$ 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

$$s^2 = \frac{1}{n-K}\varepsilon' M\varepsilon$$

可得

$$s^2 = \frac{1}{n-K}[\varepsilon'\varepsilon - \varepsilon'X(X'X)^{-1}X'\varepsilon]$$

$$= \frac{n}{n-K}\left[\frac{\varepsilon'\varepsilon}{n} - \left(\frac{\varepsilon'X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right)\right]$$

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：$\frac{\varepsilon'\varepsilon}{n} = \frac{1}{n}\sum \varepsilon_i^2$

而且： $E(\varepsilon_i^2) = Var(\varepsilon_i) = \sigma^2$ $\qquad$ $Var(\varepsilon_i^2) = E(\varepsilon_i^4) - [E(\varepsilon_i^2)]^2 = E(\varepsilon_i^4) - \sigma^4$

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 $E(\varepsilon_i^4)$ 为有限的情况下， $\frac{\varepsilon'\varepsilon}{n}$ 是 σ^2 的一致估计量。

所以我们要假设 $E(\varepsilon_i^4)$ 是有限的。

这意味着

$$p\lim s^2 = \sigma^2 - p\lim\left(\frac{\varepsilon'X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right)$$

单独看 plims^2 的第二项，略微整理之后，我们有

$$\left(\frac{\varepsilon'X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right)=\left(\frac{1}{n}\right)\left(\frac{\varepsilon'X}{\sqrt{n}}\right)\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{\sqrt{n}}\right)$$

这个统计量的大样本特性与

$$q=\left(\frac{1}{n}\right)\left(\frac{\varepsilon'X}{\sqrt{n}}\right)Q^{-1}\left(\frac{X'\varepsilon}{\sqrt{n}}\right)$$

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$\left(\frac{X'\varepsilon}{\sqrt{n}}\right)$渐近方差矩阵是$\sigma^2$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

$$\frac{q}{\sigma^2}=\left(\frac{1}{n}\right)z'z,\qquad \text{其中,}\mathrm{z}=Q^{-1/2}\left(\frac{X'\varepsilon}{\sigma\sqrt{n}}\right),\text{所以,}z\xrightarrow{d}N[0,I]$$

这样

$$nq/\sigma^2\xrightarrow{d}\chi^2[K]$$

$E(nq/\sigma^2)\to$常数, 即$E(q)\to\sigma^2\times$常数$/n\to 0$

而且$Var(nq/\sigma^2)\to$常数,即$Var(q)\to\frac{\sigma^4\times\text{常数}}{n^2}\to 0$，q 是二阶收敛的，所以保证了概率收敛，即 $p\lim_{n\to\infty} q\to Eq\to 0$ 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：**

$$Est.Asy.Var[b]=\left(\frac{1}{n}\right)s^2\left(\frac{X'X}{n}\right)^{-1}=s^2(X'X)^{-1}$$

B 的函数的渐近分布——得尔塔方法

利用泰勒展开，把 f(x)线性化。

令 f(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

$$G=\frac{\partial f(b)}{\partial b'}$$

G 是 J×K 矩阵，其中第 j 行是第 j 个函数关于 b 的导数。利用斯拉茨基（Slutsky）定理，

$$p\lim f(b) = f(\beta)$$

并且

$$p\lim G = \frac{\partial f(\beta)}{\partial \beta'} = \Gamma ,$$

于是

$$f(b) \xrightarrow{a} N\left[f(\beta), \Gamma\left(\frac{\sigma^2}{n} Q^{-1}\right)\Gamma' \right] \qquad (2)$$

实际上，渐近协方差矩阵的估计量是

$$Est.Asy.Var[f(b)] = G[s^2 (X'X)^{-1}]G'$$

如果某个函数是非线性的，则 b 的无偏的性质不会传给 f(b)。不过从（2）中可得 f(b)是 f(β) 的一致估计量，而且渐近协方差矩阵很容易获得。

例 P324（小 Green）

小 结

有限样本和大样本的结果比较

有限样本	大样本
在条件$V.\varepsilon \sim N[0,\sigma^2 I]$下的结果	**在不满足条件$V.\varepsilon \sim N[0,\sigma^2 I]$下的结果**
1. E[b]=β	1. $p\lim b = \beta + Q^{-1}\cdot 0 = \beta$
最小二乘估计是无偏的	b 是β 的一致估计量
2. E[s^2]=σ ^{2}	2. s^2 是方差σ^2 的一致估计量
σ ^{2} 估计是无偏的	s^2 是σ^2 的一致估计量
3. $Est.$Var[b]=s^2(X′ X)$^{-1}$	3. $Est.Var[b] = \left(\frac{1}{n}\right)s^2\left(\frac{X'X}{n}\right)^{-1} = s^2(X'X)^{-1}$
4．b 的精确分布是	4. b 的渐近分布是正态分布

$N[\beta, \sigma^2 (X'X)^{-1}]$

$$b \xrightarrow{a} N\left[\beta, \frac{\sigma^2}{n} Q^{-1}\right]$$

5．统计量 $t[n-K] = \dfrac{b_k - \beta_k}{s^2 (X'X)^{-1}_{kk})}$

服从自由度为 n—K 的 t 分布

5. 统计量 $t_k = \dfrac{b_k - \beta_k}{[s^2 (X'X)^{-1}_{kk}]^{1/2}}$

服从标准正态分布，而不是 t 分布

6．用于检验一组 J 个线性约束 Rβ =q 的检验统计量

$$\frac{(Rb-q)'[R(X'X)^{-1}R']^{-1}(Rb-q)/J}{e'e/(n-K)} = \frac{(Rb-q)'[Rs^2(X'X)^{-1}R']^{-1}(Rb-q)}{J}$$

服从自由度为 J 和 n—K 的 F 分布

6. $W = (Rb-q)'[R(s^2(X'X)^{-1}R']^{-1}(Rb-q) = JF$

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**）

第十一章　非球形扰动项与广义最小二乘（GLS）

一） 问题的提出

In general, the ε_i's within a set of samples may have different variance and some degree of relationships, so it is not accurate for us to assume Cov($\varepsilon_i, \varepsilon_j$)=0.

We set Cov($\varepsilon_i, \varepsilon_j$)= ρ_{ij} , so the covariance matrix is:

$$\sum = \begin{pmatrix} \sigma_1^2 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \rho_{n-1,n} \\ \rho_{n1} & \cdots & \rho_{n,n-1} & \sigma_n^2 \end{pmatrix}$$

However, under this model, we are estimating 1+2+…+n = $\frac{1}{2}n(n+1)$ parameters for ε (or $\frac{1}{2}n(n+1)$+K parameters of total for ε and β) from only n observations, which is difficult to work out and won't be able to spot the most important properties. (because there are too many properties.)

So, in real life, we normally use the following models instead:

Model 1. $\sum = E(\varepsilon\varepsilon') = \sigma^2 I$ only need to estimate one parameter for ε.

Model 2. Heteroscedasticity:

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

But we still need to estimate n parameters for ε, which is still too many.

Therefore we develop the model further, where ε is divided into several groups according to some properties of group observations, the ones in each group have common variances. So now we only need to estimate g parameters for ε where g is the number of groups.

$$\sum = \begin{pmatrix} \sigma_1^2 & & & & & & & \\ & \ddots & & & & & & \\ & & \sigma_1^2 & & & & & \\ & & & \sigma_2^2 & & & & \\ & & & & \ddots & & & \\ & & & & & \sigma_2^2 & & \\ & & & & & & \ddots & \\ & & & & & & & \sigma_g^2 \end{pmatrix}$$

Model 3. Autocorrelation

The variance of ε_i in each observations are equal, but they are no longer uncorrelated, and in general, the size of the Cov($\varepsilon_i, \varepsilon_j$) depends on the distance between i & j. (e.g. the distance between the time of observations made in time series data) and decreases as the distance increases.(the longer between two observations, the less of their correlation).

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

We need to estimate 2 parameters (σ^2, ρ).

Model 4. ARCH (条件异方差) or GARCH（广义条件异方差）

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

All σ_i^2's are different from observations to observations, but there exist some relationships between them:

ARCH: (e.g. $\sigma_k^2 = a + b\sigma_{k-1}^2$)

GARCH: (e.g. σ_k^2=a+bσ_{k-1}^2+cσ_{k-2}^2+…)

Once we have set up such a model, we only need to estimate two parameters in ARCH.(a, b).

* In general, we check the suitability of models in the order of 4 to 1. (which has higher accuracy or less parameters).

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

$$y = X\beta + \varepsilon, \qquad E[\varepsilon] = 0, \qquad E[\varepsilon\varepsilon'] = \sigma^2\Omega$$

（1）

其中Ω 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Ω 两种特殊的情况是**异方差性**和**自相关**。

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$\sigma^2\Omega$ 是

$$\sigma^2\Omega=\begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ & & \vdots & \\ 0 & 0 & \cdots & \sigma_n^2\end{bmatrix}$$

自相关

自相关经常出现在时间序列数据中，经济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

$$\sigma^2\Omega=\sigma^2\begin{bmatrix}1 & \rho_1 & \cdots & \rho_{n-1}\\ \rho_1 & 1 & \cdots & \rho_{n-2}\\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & & 1\end{bmatrix}$$

非对角线上的值依赖于扰动项的模式。

普通最小二乘法（OLS）的结果

具有球形干扰项

$$E[\varepsilon]=0$$

和

$$E[\varepsilon\varepsilon']=\sigma^2 I$$

（2）

重申前面的内容，普通最小二乘估计量，

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

（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（1）模型中仍然成立。

有限样本特性

对（3）两边取期望，如果$E[\varepsilon|X]=0$，则

$$E[b]=E_X[E[b|X]]=\beta$$

（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（2）假设变化的影响。

最小二乘法估计量的样本方差是

$$Var[b-\beta]=E[(b-\beta)(b-\beta)']$$

$$=E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}]$$

$$=(X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1}$$

$$=\frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\Omega X}{n}\left(\frac{X'X}{n}\right)^{-1} \quad (5)$$

在（3）中，b 是ε的线性函数,因此，如果ε服从正态分布，则

$$b\sim N[\beta,\sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}]$$

由于最小二乘估计量的方差不再是$\sigma^2(X'X)^{-1}$，任何基于$s^2(X'X)^{-1}$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$\sigma^2(X'X)^{-1}$是比 b 的真正方差大还是小,因此即使有σ^2的一个好的估计，Var[b]的传统估计量也不会有用。

最小二乘法的渐近特性

如果 Var[b]收敛于 0，则 b 是一致的。使用表现良好的回归量，$(X'X/n)^{-1}$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$X'\Omega X/n$不一定收敛，如果它收敛，则从（5）式可推断普通最小二乘是一致的和无偏的。因此

如果$p\lim(X'X/n)$和$p\lim(X'\Omega X/n)$都是有限正定矩阵，则 b 是β 的一致估计量。

上述结论成立的条件依赖于 X 和Ω 。

另一种分离这两个组成部分的处理办法是：

如果

1、X′ X 最小的特征根当$n \to \infty$时无限制地增加，这意味着 $p\lim(X'X)^{-1} = 0$；

2、Ω 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Ω 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 $y = \mu + \varepsilon$，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

$$\Omega = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ & & & \vdots & \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

矩阵 X 是一列 1。μ 的普通最小二乘估计量是 $b = (X'X)^{-1}X'y = \bar{y}$。把Ω 代入（5），得

$$\begin{aligned} Var[\bar{y}] &= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1} \\ &= \frac{\sigma^2}{n}(1 - \rho + n\rho) \end{aligned} \qquad (5a)$$

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，$X'\Omega X / n = 1 + \rho(n-1)$ 不收敛。由于 X 是一列 1，因此 $XX' = n$ 是一个标

量，满足条件 1；但是，Ω 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

$$\sqrt{n}(b-\beta)=\left(\frac{X'X}{n}\right)^{-1}\frac{1}{\sqrt{n}}X'\varepsilon \qquad (5b)$$

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 $p\lim(X'X/n)=Q$，那么右边项的极限分布与

$$v=Q^{-1}\frac{1}{\sqrt{n}}X'\varepsilon=Q^{-1}\frac{1}{\sqrt{n}}\sum_i x_i\varepsilon_i \qquad (5c)$$

的分布相同，其中 x_i' 是 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 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（5）给出。对于大多数一般的情况，答案是否定的，因为（5c）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 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，

b 渐近地服从均值为β，方差矩阵由（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 b 的有效估计。

二） 广义最小二乘（GLS）

在广义回归模型中，β 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Ω是正定对称矩阵，它可以分解为

$$\Omega=C\wedge C' \qquad (6)$$

其中 C 的各列是Ω的特征向量经过正交化而得到，即 CC’=I，而且Ω的特征根被放在对角矩阵∧中。令 $\wedge^{1/2}$ 是对角元素为 $\sqrt{\lambda_i}$ 的对角矩阵。

如果令$P=C\wedge^{-1/2}$，则

$$\Omega^{-1}=PP'$$

用P’前乘（1）中的模型可得

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

或

$$y_*=X_*\beta+\varepsilon_* \qquad (7)$$

ε_*的方差是

$$E[\varepsilon_*\varepsilon_*']=P'\sigma^2\Omega P=\sigma^2 I$$

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_*

是可观测数据。在古典回归模型中，OLS 是有效的。

因此$\hat{\beta}=(X_*'X_*)^{-1}X_*'y_*=(X'PP'X)^{-1}X'PP'y=(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$

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$E[\varepsilon_*\mid X_*]=0$，GLS 估计量$\beta$是无偏的。这等价于$E[P'\varepsilon\mid P'X]=0$，但由于 P 是已知常数的矩阵，即要求$E[\varepsilon\mid X]=0$，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。

如果

$$p\lim\left(\frac{X_*'X_*}{n}\right)=Q_* \qquad (8)$$

GLS 估计量是一致的，其中Q_*是有限正定矩阵。进行替换可得

$$p\lim\left(\frac{X'\Omega^{-1}X}{n}\right)^{-1}=Q_*^{-1} \qquad (9)$$

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（9）的假设，GLS 估计量是渐近正态分布的，均值为β，样本方差为

$$Var[\hat{\beta}]=\sigma^2(X_*'X_*)^{-1}=\sigma^2(X'\Omega^{-1}X)^{-1} \qquad (10)$$

通过对（7）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。

$\hat{\beta}$ 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（7）中。为了检验 J 个线性约束 Rβ =q，相应的统计量是

$$F[J,n-K]=\frac{(R\hat{\beta}-q)'[R(\hat{\sigma}^2(X_*'X_*)^{-1}R']^{-1}(R\hat{\beta}-q)}{J}$$

$$=\frac{(\hat{\varepsilon}_c'\hat{\varepsilon}_c-\hat{\varepsilon}'\hat{\varepsilon})/J}{\hat{\sigma}^2},$$

其中残差向量是 $\varepsilon=y_*-X_*\hat{\beta},$

而

$$\hat{\sigma}^2=\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-K}=\frac{(y_*-X_*\hat{\beta})'(y_*-X_*\hat{\beta})}{n-K}=\frac{(y-X\hat{\beta})'PP'(y-X\hat{\beta})}{n-K}=\frac{(y-X\hat{\beta})'\Omega^{-1}(y-X\hat{\beta})}{n-K}$$

有约束的 GLS 残差 $\hat{\varepsilon}_c=y_*-X_*\beta_c$，基于

$$\hat{\beta}_c=\hat{\beta}\ -(X_*'X_*)^{-1}R'[R(X_*'X_*)^{-1}R']^{-1}(R\hat{\beta}\ -q)$$

$$=\hat{\beta}-[X'\Omega^{-1}X]^{-1}R'[R(X'\Omega^{-1}X)^{-1}R']^{-1}(R\hat{\beta}-q) \qquad (11)$$

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（7）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。

三） 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。

可行的最小二乘估计（FGLS）

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，其常见的表达形式是

$$\Omega=\begin{bmatrix}1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1}\\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2}\\ & & & & \vdots & \\ \rho^{n-1} & & \rho^{n-2} & & \cdots & 1\end{bmatrix},$$

一个也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

$$\hat{\Omega}=\Omega(\hat{\theta})$$

替代真正的Ω。我们所考虑的问题是利用 $\Omega(\hat{\theta})$ 是否要求我们改变上节的某些结果。

如果 $p\lim\hat{\theta}=\theta$，利用 $\hat{\Omega}$ 似乎渐近等价于利用真正的Ω（根据 slutsky 定理）。当然我们还需要满足一些其他的相应的条件。令可行广义最小二乘（或 FGLS）估计量记为

$$\hat{\hat{\beta}}=(X\hat{\Omega}^{-1}X)^{-1}X\hat{\Omega}^{-1}y$$

那么，$\hat{\hat{\beta}}$ 渐近等价于 $\hat{\beta}$ 的条件是

$$p\lim\left(\frac{X_*'X_*}{n}\right)=p\lim\frac{X\hat{\Omega}^{-1}X}{n}=p\lim\frac{X\Omega^{-1}X}{n} \qquad (18)$$

和

$$p\lim \frac{1}{\sqrt{n}} X'\hat{\Omega}^{-1}\varepsilon = p\lim \frac{1}{\sqrt{n}} X'\Omega^{-1}\varepsilon \qquad (19)$$

如果（7）中变换后的回归量表现良好，则（19）右边服务从极限正态分布。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（19）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。但在大多数情况中，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 FLGS 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。**

四）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。

一、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

$H_0 : \sigma_i^2 = \sigma^2$ 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：

$Var[b]=\sigma^2(X'X)^{-1}[X'\Omega X][X'X]^{-1}$

我们可用如下式对它加以估计

Est. $Var[b]=(X'X)^{-1}\left[\sum_{i=1}^{n}e_i^2(X_iX_i')\right](X'X)^{-1}$

如果不存在异方差性，最小二乘估计量（OLS）的协方差矩阵是：

Var[b]=$\sigma^2(X'X)^{-1}$

可得到 Var[b]的一个估计量，

$$Est.Var[b]=\left(\frac{1}{n}\right)s^2\left(\frac{X'X}{n}\right)^{-1}=s^2(X'X)^{-1}$$

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，即 $nR^2\sim x^2$(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。

二、戈德菲尔德一匡特检验(The Goldfeld-Quandt Test)

另外两个相对一般性的检验是戈尔德一匡特检验（1965）和布罗施一帕甘（1979）拉格朗日乘数检验。

对于戈德菲尔德一匡特检验，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 $\sigma_i^2=\sigma^2x_i^2$ 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：

$$F[n_1 - K, n_2 - K] = \frac{e_1' e_1 / n_1 - K}{e_2' e_2 / n_2 - K},$$

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，戈德菲尔德一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。