第六章

1. 记 $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2, S_1 = \sum_{i=1}^{n}(X_i - \bar{X})Y_i$

$$\sum_{i=1}^{n}(Y_i - \alpha_0 - \alpha_1(X_i - \bar{X}))^2$$

$$= S^2\alpha_1^2 - 2S^1\alpha_1 + n\alpha_0^2 - 2n\bar{Y}\alpha_0 + \sum_{i=1}^{n}Y_i^2$$

$$= (S\alpha_1 - S_1/S)^2 + n(\alpha_0 - \bar{Y})^2$$

$$+\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 - S_1^2/S^2$$

由此立即看出此平方和之最小值在 $\alpha_0 = \bar{Y} = \hat{\beta}_0$ 和 $\alpha_1 = S_1/S^2 = \hat{\beta}^1$ 处达到,且最小值为

$$\sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 - S_1^2/S^2$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - S_1^2/S^2$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \hat{\beta}_1 S_1$$

即(2.23)式,这个证明不仅简单,还有一个好处,即它确实肯定了达到最小值.用偏导数方法,理论上还有一个验证方程组(2.10),(2.11)的解确实是最小值点的问题.

2.(a)利用$\hat{\beta}_0,\hat{\beta}_1$ 的无偏性,因为

$$\delta_i = Y_i - \hat{Y}_i = \beta_0 + \beta_1(X_i - \bar{X}) + e_i - (\hat{\beta}_0 + \hat{\beta}_1(X_i - \bar{X}))$$

$$= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)(X_i - \bar{X}) + e_i$$

且 $E(e_i)=0$,即得 $E(\delta_i)=0$.

(b)这是因为,在(2.10)式中把 a_0, a_1 分别换成其解$\hat{\beta}_0,\hat{\beta}_1$,得到 $\delta_1+\cdots+\delta_n=0$,$\delta_1+\cdots+\delta_n$ 之间既然有这样一个函数关系,它不可能是相互独立的.

(c)在证明(2.21)式的过程中已得出

$$\delta_i = e_i - \bar{e} - t_i\sum_{j=1}^{n} t_j e_j / S^2 \qquad \left(t_j = X_j - \bar{X},\ \bar{e} = \sum_{j=1}^{n} e_j / n\right) \tag{1}$$

又由(a)有 $E(\delta_i)=0$($E(\delta_i)=0$ 也直接由上式得出),故

$$\mathrm{Var}(\delta_i) = E(\delta_i^2) = E(e_i - \bar{e})^2 + t_i^2 S^{-4} E\left(\sum_{j=1}^{n} t_j e_j\right)^2 \cdot$$

$$-2tS^{-2}E\left(\sum_{j=1}^{n}t_je_ie_j\right)-2t_iS^{-2}E\left(\sum_{j=1}^{n}t_j\bar{e}e_j\right) \tag{2}$$

注意到 $E(e_i-\bar{e})^2=E(e_i^2)+E(\bar{e}^2)-2E(e_i\bar{e})=\sigma^2+\sigma^2/n-2\sigma^2/n=\sigma^2-\sigma^2/n=(1-1/n)\sigma^2$,以及

$$E\left(\sum_{j=1}^{n}t_je_j\right)^2=\sigma^2\sum_{j=1}^{n}t_j^2=\sigma^2S^2,E\left(\sum_{j=1}^{n}t_je_ie_j\right)=t_i\sigma^2,$$

$$E\left(\sum_{j=1}^{n}t_j\bar{e}e_j\right)=\sum_{j=1}^{n}t_j\sigma^2/n=0$$

代入(2)式即得所要证的结果.

注:由这个结果,得

$$E\left(\sum_{i=1}^{n}\delta_i^2\right)=\sum_{i=1}^{n}E(\delta_i^2)=\sum_{i=1}^{n}\left[1-\frac{1}{n}-(X_i-\bar{X})^2/S^2\right]\sigma^2$$
$$=(n-2)\sigma^2$$

因而得到 $\sum_{i=1}^{n}\delta_i^2/(n-2)$ 为 σ^2 的无偏估计的另一证明.

(d)与(1)式类似写出 δ_j 的表达式,注意 $\mathrm{Cov}(\delta_i,\delta_j)=E(\delta_i\delta_j)$,把两式相乘逐项求均值,与($c$)完全类似地得到所要的结果.

3. 考虑线性回归模型

$$Y=\alpha_0+\alpha_1x+e,\alpha_0=a,\alpha_1=b-a \tag{3}$$

其中 $e\sim N(0,\sigma^2)$.在 $X=0$ 点重复观察 n 次,其 Y 值记为 $X_1,\cdots,X_n$;在 $X=1$ 点重复观察 m 次,其 Y 值记为 $Y_1,\cdots,Y_m$.这样按模型(3),$X_1,\cdots,X_n\sim N(a,\sigma^2)$,$Y_1,\cdots,Y_m\sim N(b,\sigma^2)$,如题中所设者.然自模型(3)观之,估计 $b-a$ 相当于估计回归系数a_1,检验亦然.而此处的平方和(2.9)为 $\sum_{i=1}^{n}(X_i-a_0)^2+\sum_{j=1}^{m}(Y_j-a_0-a_1)^2$,直接得出 a_0,a_1 的最小二乘估计为 $\bar{X}$ 和 $\bar{Y}-\bar{X}$.后者即 $b-a$ 的估计.残差平方和为 $\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{j=1}^{m}(Y_j-\bar{Y})$.自由度 $m+n-2$.又此处之 S^2(S^2 即(2.16)式中的 S_x^2)为(注意自变

量值中 n 个 0 和 m 个 1，其平均为 $m/(n+m)$)

$$S^2 = n(0-m/(n+m))^2 + m(1-m/(n+m))^2$$
$$= nm/(n+m)$$

由此，按(2.26)式求 $\alpha_1 = b-a$ 的区间估计，所得结果与两样本 t 区间估计一致.

4. 将平方和 $\sum_{i=1}^{n}(Y_i - bX_i)^2$ 按第 1 题的方式处理：$\sum_{i=1}^{n}(Y_i - bX_i)^2 = S_0^2 b^2 - 2S_{01}b + \sum_{i=1}^{n} Y_i^2 = (S_0 b - S_{01}/S_0)^2 + \sum_{i=1}^{n} Y_i^2 - S_{01}^2/S_0^2$，此处 $S_0^2 = \sum_{i=1}^{n} X_i^2$，$S_{01} = \sum_{i=1}^{n} X_i Y_i$. 由此式立即得出 b 的最小二乘估计 为

$$\hat{b} = S_{01}/S_0^2 = \sum_{i=1}^{n} X_i Y_i \Big/ \sum_{i=1}^{n} X_i^2$$

而残差平方和为 $\sum_{i=1}^{n} Y_i^2 - S_{01}^2/S_0^2$，暂记为 R. 由于

$$E(Y_i^2) = (EY_i)^2 + \mathrm{Var}(Y_i) = b^2 X_i^2 + \sigma^2$$
$$E(S_{01}^2) = (ES_{01})^2 + \mathrm{Var}(S_{01})$$
$$= \left(\sum_{i=1}^{n} bX_i^2\right)^2 + \sum_{i=1}^{n} X_i^2 \sigma^2 = b^2 S_0^4 + \sigma^2 S_0^2$$

得到

$$E(R) = \sum_{i=1}^{n}(b^2 X_i^2 + \sigma^2) - (b^2 S_0^4 + \sigma^2 S_0^2)/S_0^2$$
$$= n\sigma^2 + b^2 S_0^2 - b^2 S_0^2 - \sigma^2 = (n-1)\sigma^2$$

因而证明了 $R/(n-1)$是 σ^2 的无偏估计.

(c)只须作一个正交变换

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = A \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

其中 A 为正交方阵，第一行是$(X_1/S_0, \cdots, X_n/S_0)$. 则 $R = Z_2^2 +$

$\cdots + Z_n^2$,其中 $Z_2,\cdots,Z_n$ 独立同分布且有公共分布 $N(0,\sigma^2)$.

5. 若 $c_1=0$,则因 b 的区间估计问题已解决了,c_2b 当然直接由之得出.若 $c_1\neq 0$,把 c_1a+c_2b 表为 $c_1(a+xb)(x=c_2/c_1)$,即 $c_1m(x)$.因 $m(x)$的区间估计已在(2.27)式的基础上求得,故问题得到解决.

6.(a)取 $i=1$ 来讨论,因为把 $X_1,\cdots,X_n$ 作线性变换 $X'_i=aX_i+b(a\neq 0)$ 不影响$(X_1-\bar{X})^2/\sum_{j=1}^{n}(X_j-\bar{X})^2$ 之值,不妨设 $\bar{X}=0,X_1=1$.这时,为使上述比值最大,应在 $X_2+\cdots+X_n=-1$ 的约束下,使 $X_2^2+\cdots+X_n^2$ 达到最小.但易知后者的最小值在 $X_2=\cdots=X_n=-\dfrac{1}{n-1}$时达到,最小值为$\dfrac{1}{n-1}$.故所述比值不能大于$1\Big/\left(1+\dfrac{1}{n-1}\right)=1-\dfrac{1}{n}$,等号当且仅当对某个 i,有 $X_1=\cdots=X_{i-1}=X_{i+1}=\cdots=X_n\neq X_i$.

(b)分两种情况:若 $\bar{X}_n$ 保持有界,则因 $S_n^2\to\infty$,就有$(a-\bar{X}_n)^2/S_n^2\to 0$.若$|\bar{X}_n|\to\infty$,则注意到

$$(a-\bar{X}_n)^2/S_n^2\leqslant(a-\bar{X}_n)^2\Big/\sum_{i=1}^{m}(X_i-\bar{X}_n)^2,m\leqslant n$$

固定 m,令 $n\to\infty$.因为$|\bar{X}_n|\to\infty$,上式右端有极限 $1/m$.因 m 可取得任意大,知$(a-\bar{X}_n)^2/S_n^2$ 的极限可任意小,故只能为 0(若$|\bar{X}_n|$既不有界也不随 $n\to\infty$而趋于无穷,则通过抽取子序列的方法去讨论).

(c)先给出一个预备事实:在$[0,1]$上给出三个数 $x,c-x(0\leqslant c\leqslant 1)$及 a,记 $I=(x-a)^2+(c-x-a)^2$,则总可以改变 x 之值以增大I,使 $x,c-x$ 都仍在$[0,1]$上,且 x 及$c-x$ 中至少有一个为 0 或 1.如 x 和$c-x$ 分处a 的两边,这一点很清楚.若同在一边,例如 $0<x\leqslant c-x\leqslant a$,则 $\mathrm{d}I/\mathrm{d}x=4x-2c\leqslant 0$(因 $x\leqslant c-x$, $2x\leqslant c$).故让 x 下降能增大I.让 x 降为 0(这时 $c-x$ 升为c,仍在$[0,1]$上)即可.

现证明本题,不失普遍性可设区间$[A,B]$为$[0,1]$.证明分三段:1°为使 $S^2=\sum_{i=1}^{n}(X_i-\bar{X})^2$ 最大,诸 X_i 中至多只能有一个非0非1.,因为,若有两个,例如 X_1,X_2,非0非1.则据上述预备知识,可以在不改变 $X_3,\cdots,X_n$ 和 $\bar{X}$ 的条件下,使 X_1,X_2 中至少有一个为0或1,而$(X_1-\bar{X})^2+(X_2-\bar{X})^2$ 增大,即 S^2 增大.2°现设 $X_1,\cdots,X_n$ 中,有 n_0 个为0,n_1 个为1,还有一个为 a,$0\leqslant a\leqslant 1$.证明:总可以把 a 改为0或1以增大 S^2.此时

$$S^2=n_0\left(0-\frac{n_1+a}{n}\right)^2+n_1\left(1-\frac{n_1+a}{n}\right)^2+\left(a-\frac{n_1+a}{n}\right)^2$$

注意到 $n_0+n_1=n-1$,易算出

$$\mathrm{d}(S^2)/\mathrm{d}a=2(n-1)n^{-1}a+D$$

D 与 a 无关,若上式大于0,则把 a 增至1可增大 S^2;若上式不大于0,则把 a 减至0可以增大 S^2.总之,a 可改为0或1以增大 S^2.3°以上两步证明了:为使 S^2 最大,全部 X_i 必须只取0,1这两个值,设有 n_0 个0,n_1 个1.则

$$S^2=n_0\left(0-\frac{n_1}{n}\right)^2+n_1\left(1-\frac{n_1}{n}\right)=n_0n_1/n$$

在 $n_0+n_1=n$ 的约束下,要使 S^2 最大,n_0 和 n_1 之差距应尽量小,如 $n=2m$,应取 $n_0=n_1=m$:若 $n=2m+1$,则 n_0,n_1 中应有一个为 m,另一个为 $m+1$.

7.(a)由$\hat{\beta}_0=\bar{Y}$ 易得出$E(\hat{\beta}_0)=\beta_0$,为证 $E(\hat{\beta})=\beta$,暂把 p 行 n 列方阵$L^{-1}X$ 记为

$$L^{-1}X=\begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ l_{p1} & l_{p2} & \cdots & l_{pn} \end{pmatrix}$$

从 X 的每行元素之和为0,可推出此矩阵每行元素之和为0:$l_{i1}+\cdots+l_{in}=0,i=1,\cdots,p$.现有

$$E(\hat{\beta}_j)=E\left(\sum_{i=1}^{n}l_{ji}Y_i\right)=\sum_{i=1}^{n}l_{ji}\beta_0+\sum_{k=1}^{p}\beta_k\sum_{i=1}^{k}l_{ji}X_{ki}$$

据上述，β_0 之系数为 0，而 β_k 之系数，正是两矩阵 $L^{-1}X$ 和 X' 之积的 (j,k) 元，因 $L^{-1}XX' = L^{-1}L =$ 单位阵 I，知只有当 $k=j$ 时此系数为 1，k 为其他值时为 0. 故证明了 $E(\hat{\beta}_j)=\beta_j, j=1,\cdots,p$.

(b)因为 $\hat{\beta}_0 = \dfrac{1}{n}(Y_1+\cdots+Y_n)$，$\hat{\beta}_j = l_{j1}Y_1+\cdots+l_{jn}Y_n$. 由 $(Y_1,\cdots,Y_n)$ 独立且由等方差，易知

$$\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_j) = \frac{1}{n}\sigma^2(l_{j1}+\cdots+l_{jn}) = 0, j = 1,\cdots,p.$$

(c)与(b)相似，由 $\hat{\beta}_j = l_{i1}Y_1+\cdots+l_{in}Y_n$ 及 $\hat{\beta}_j = l_{j1}Y_1+\cdots+l_{jn}Y_n$. 得知

$$\mathrm{Cov}(\hat{\beta}_i,\hat{\beta}_j) = \sigma^2(l_{i1}l_{j1}+\cdots+l_{in}l_{jn})$$

右边括号内之量是矩阵 $L^{-1}X$ 及其转置之积的 (i,j) 元. 因 L 为对称方阵，故 L^{-1} 也是对称方阵，即 $(L^{-1})' = L^{-1}$. 故 $(L^{-1}X)\cdot(L^{-1}X)' = L^{-1}XX'L^{-1} = L^{-1}LL^{-1} = L^{-1}$. 因此，$\mathrm{Cov}(\hat{\beta}_i,\hat{\beta}_j) = \sigma^2\cdot L^{-1}$ 的 (i,j) 元，当 $i=j$ 时，得到 $\hat{\beta}_i$ 的方差.

8. 有关的理论考虑在题中已说了. 现在只须计算一下 $r\sqrt{n-2}/\sqrt{1-r^2}$：记 $S_x^2 = \sum\limits_{i=1}^{n}(X_i-\bar{X})^2$ 并注意 $\sum\limits_{i=1}^{n}(X_i-\bar{X})\cdot(Y_i-\bar{Y}) = \sum\limits_{i=1}^{n}(X_i-\bar{X})Y_i$，有

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\sqrt{n-2}\sum\limits_{i=1}^{n}(X_i-\bar{X})Y_i\Big/\left[S_x\sqrt{\sum\limits_{i=1}^{n}(Y_i-\bar{Y})^2}\right]}{\left(1-\sum\limits_{i=1}^{n}(X_i-\bar{X})Y_i\right)^2\Big/\left(S_x^2\sum\limits_{i=1}^{n}(Y_i-\bar{Y})^2\right)^{1/2}}$$

$$= \frac{\sqrt{n-2}\sum\limits_{i=1}^{n}(X_i-\bar{X})Y_i/S_x}{\left(\sum\limits_{i=1}^{n}(Y_i-\bar{Y})^2-\left(\sum\limits_{i=1}^{n}(X_i-\bar{X})Y_i\right)^2/S_x^2\right)^{1/2}}$$

因为 $\hat{\sigma}=\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\left(\sum_{i=1}^{n}(X_i-\bar{X})Y_i\right)^2\Big/S_x^2\right)^{1/2}\Big/\sqrt{n-2}$,而

$\hat{\beta}_1=\sum_{i=1}^{n}(X_i-\bar{X})Y_i/S_x^2$,又 $\beta_1=0$. 故即有

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=(\hat{\beta}_1-\beta_1)/(\hat{\sigma}S_X^{-1})$$

再用(2.26),即证得所要的结果.

9.(a)令 $Z_i=Y_i-X_i, i=1,\cdots,n, Z_1,\cdots Z_n$ 独立同分布,公共分布为 $N(b-a,2\sigma^2)$. 而 $H_0: b=a$ 成为一个检验正态分布 $N(\theta,\sigma_1^2)(\sigma_1^2=2\sigma^2,\theta=b-a)$ 中均值 $\theta=0$ 的问题,可用一样本 t 检验:当

$$\sqrt{n}\mid\bar{Z}\mid\Big/\left(\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2\right)^{1/2}>t_{n-1}(\alpha/2) \qquad (4)$$

时否定 H_0.

注:这个模型叫做“成对比较模型”,意即 X_i, Y_i 这一对可以比较,但当 $i\neq j$ 时,X_i, Y_j 无法比较,因为 $Y_j-X_i\sim N(b-a+d_j-d_i,2\sigma^2)$,不只与 $b-a$ 有关而 d_j-d_i 又不知道. 这与所谓“成组比较”不同:在成组比较模型中是 $d_1=\cdots=d_n=0$. 这时任意的 X_i, Y_j 都可比较,而我们可使用两样本 t 检验去检验 H_0,它有 $2n-2$个自由度. 而检验(4)只有 $n-1$ 个自由度,所损失的自由度,就是因为有了赘余参数 $d_1,\cdots,d_n$.

(b)可以把 $X_1,\cdots,X_n$ 和 $Y_1,\cdots,Y_n$ 分别视为一个两水平因素在其水平1的 n 个观察值和水平2的 n 个观察值. d_j 为区组效应,$j=1,\cdots,n$,而 a, b 则分别是这两个水平的效应. 为把模型写成(5.13)的形式,可令 $Y_{1j}=X_j, Y_{2j}=Y_j, j=1,\cdots,n$;而

$$\mu=\bar{d}+(a+b)/2 \qquad (\bar{d}=(d_1+\cdots d_n)/n)$$

$$a_1=a-(a+b)/2, a_2=b-(a+b)/2$$

$$b_j=d_j-\bar{d}, j=1,\cdots,n$$

则有

$$Y_{ij} = \mu + a_i + b_j + e_{ij}, i = 1,2; j = 1,\cdots,n \tag{5}$$

这里 $e_{ij}, i=1,2; j=1,\cdots,n$ 全体独立同分布并有公共分布 $N(0,\sigma^2)$. 模型(5)符合所要求的约束条件:

$$a_1 + a_2 = a - (a+b)/2 + b - (a+b)/2 = 0$$

$$\sum_{j=1}^{n} b_j = \sum_{j=1}^{n} (d_j - \bar{d}) = 0$$

原假设: $H_0: a = b$ 相应于检验(5)中的因子效应为 0, 即 $a_1 = a_2 = 0$.

(c)就模型(5)按(5.23)的分解式来计算 SS_A 和 SS_e:

$$\begin{aligned}
SS_e &= \sum_{i=1}^{2}\sum_{j=1}^{n}(Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})^2 \\
&= \sum_{j=1}^{n}(X_j - \bar{X} - (X_j + Y_j)/2 + (\bar{X} + \bar{Y})/2)^2 \\
&\quad + \sum_{j=1}^{n}(Y_j - \bar{Y} - (X_j + Y_j)/2 + (\bar{X} + \bar{Y})/2)^2 \\
&= \sum_{j=1}^{n}[(X_j - Y_j)/2 - (\bar{X} - \bar{Y})/2]^2 \\
&\quad + \sum_{j=1}^{n}[(Y_j - X_j)/2 - (\bar{Y} - \bar{X})/2]^2 \\
&= \frac{1}{2}\sum_{j=1}^{n}(Z_j - \bar{Z})^2 \quad (Z_j = Y_j - X_j)
\end{aligned}$$

自由度为$(2-1)(n-1) = n-1$. 而

$$\begin{aligned}
SS_A &= n\sum_{i=1}^{2}(Y_{i.} - Y_{..})^2 = n[(\bar{X} - (\bar{X} + \bar{Y})/2)^2 \\
&\quad + (\bar{Y} - (\bar{X} + \bar{Y})/2)^2] \\
&= \frac{n}{2}(\bar{X} - \bar{Y})^2 = \frac{n}{2}\bar{Z}^2
\end{aligned}$$

自由度为 $2-1=1$. 故 $H_0: a_1 = a_2 = 0$, 即 $a = b$ 的 F 检验为: 当

$$\frac{n}{2}\bar{Z}^2 \Big/ \left[\left(\frac{1}{2}\sum_{j=1}^{n}(Z_i - \bar{Z})^2/(n-1)\right] > F_{1,n-1}(\alpha)$$

时否定 H_0，即当

$$\sqrt{n}\,|\bar{Z}|\bigg/\left(\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2\right)^{1/2}>\sqrt{F_{1,n-1}(\alpha)}$$

时否定 H_0. 由于 $t_{n-1}^2(\alpha/2)=F_{1,n-1}(\alpha)$（这是因为，按定义，若 $X\sim t_{n-1}$，则 $X^2\sim F_{1,n-1}$），这个检验与(a)中得到的一致.

10. 这张正交表叫 $L_8(2^7)$正交表. 它只能排2水平因子，至多7个，试验一定做8次，不能多也不能少.

把因子 A,B,C 分别排在第1,2,4列头上，区组也视为一个因子 D，排在第5列头上，则得到如下的设计：

区组1：$A_1B_1C_1$，$A_1B_2C_1$，$A_2B_1C_2$，$A_2B_2C_2$，

区组2：$A_1B_1C_2$，$A_1B_2C_2$，$A_2B_1C_1$，$A_2B_2C_1$，

其中例如，$A_1B_2C_1$ 表示因子 A 取水平1，B 取水平2，C 取水平1，余类推.

其所以舍掉第3列不用，是为了避免某些组合做两次（如 $A_1B_1C_1$ 等），而某些组合（$A_1B_2C_1$ 等）则不出现，按上述设计，则8种可能的组合各出现了一次.

此设计 A,B,C 及区组各占一自由度，共4个自由度. 全部自由度为 $8-1=7$，故误差平方和 SS_e 尚有三个自由度.