6.3 多元线性回归

本节我们考虑有 p 个自变量 $X_1,\cdots,X_p$ 的情形,因变量仍记为 Y.模型为

$$Y = b_0 + b_1X_1 + \cdots + b_pX_p + e \tag{3.1}$$

其解释与(2.1)相同.这里也有自变量为随机或非随机的区别,今后我们一律把自变量视为非随机的.在(3.1)式中,b_0 为常数项或

截距，b_k 称为 Y 对 X_k 的回归系数，或称偏回归系数*. e 仍为随机误差.

现设对 $X_1,\cdots,X_p$ 和 Y 进行观察，第 i 次观察时它们的取值分别记为 $X_{1i},\cdots,X_{pi}$ 和 Y_i，随机误差为 e_i（注意 e_i 不可观察），则得到方程

$$Y_i = b_0 + b_1X_{1i} + \cdots + b_pX_{pi} + e_i, i = 1,\cdots,n \quad (3.2)$$

这里假定

$$e_1,\cdots,e_n \text{ 独立同分布}, E(e_i) = 0, 0 < \mathrm{Var}(e_i) = \sigma^2 < \infty \quad (3.3)$$

误差方差 σ^2 未知.

统计问题仍和一元回归时一样：要根据所得数据

$$(X_{1i},\cdots,X_{pi},Y_i), i = 1,\cdots,n \quad (3.4)$$

对 $b_0,\cdots,b_p$ 和误差方差 σ^2 进行估计，对回归函数 $b_0 + b_1x_1 + \cdots b_px_p$ 进行估计，在自变量的给定之值 $(x_1^0,\cdots,x_p^0)$ 处对因变量 Y 的取值进行预测，及有关的假设检验问题等. 在上节中对一元情况引进的不少方法和概念仍适用于此处多元的情况，但在计算和理论方面，都较一元的情况复杂. 就本课程而言，我们不能对这些进行仔细的论述，只能把一些重要的结果和公式不加证明地写出来.

在讨论一元的情况时我们曾实行“中心化”，即用(2.6)代替(2.4). 这一变换对多元的情况很有用，方法也一样：算出每个自变量 X_k 在 n 次观察中取值的算术平均 $\overline{X}_k = (X_{k1} + \cdots + X_{kn})/n$，而后令

$$X_{ki}^* = X_{ki} - \overline{X}_k, i = 1,\cdots,n; k = 1,\cdots,p \quad (3.5)$$

即可将(3.2)写为

$$Y_i = \beta_0 + \beta_1X_{1i}^* + \cdots + \beta_pX_{pi}^* + e_i, i = 1,\cdots,n \quad (3.6)$$

---

* 这“偏”字的意思，约略与微积分中偏导数中的“偏”字相当，其真实含义是：若只取一个自变量 X_k 而考虑 Y 与 X_k 之间的一元回归，则回归系数 b_k^* 将与(3.1)中的 b_k 不同.

β_k 等与 b_k 等的关系是：

$$\beta_k = b_k, k = 1, \cdots, p; \beta_0 = b_0 + b_1\overline{X}_1 + \cdots + b_p\overline{X}_p \quad (3.7)$$

如在模型(3.6)之下对 β_k 等作了估计，则可用(3.7)将其转化为对 b_k 等的估计. 在(3.6)中有

$$X_{k1}^{*} + \cdots + X_{kn}^{*} = 0, k = 1, \cdots, p$$

以后我们只讨论(3.6)，且为书写方便计，略去 X_{ki}^{*} 中的“*”号，即仍记为 X_{ki}：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + e_i, i = 1, \cdots, n \quad (3.8)$$

记住(3.8)中的 X_{ki} 已是经过中心化的，与(3.2)中的 X_{ki} 不同.

在讨论多元线性回归时，采用矩阵和向量的记号很方便. m 行 n 列的矩阵常用一个大写字母(如 X, A 等)去记，有时也记为 (a_{ij})，a_{ij} 为该矩阵的 (i, j) 元，即第 i 行第 j 列之元. 当 $m = n$ 时称为 n 阶方阵. n 阶方阵 $A = (a_{ij})$，若 $a_{ij} = 1$，当 $i = j$，$a_{ij} = 0$ 当 $i \neq j$，则称为 n 阶单位阵并记为 I 或 I_n. 方阵 A 的逆方阵(如存在)记为 A^{-1}. 矩阵 A 的转置矩阵将记为 A'.

向量 a 一般理解为列向量，如

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix}$$

为 k 维列向量，a_i 为其第 i 个分量. a' 则是行向量 $(a_1, \cdots, a_k)$. 在矩阵或向量运算中，0 表示各元皆为零的矩阵或向量，有相应的维数.

若 A 为 $m \times n$ 方阵，a 为 n 维向量，则按矩阵乘法定义，Aa 为 m 维向量，当 A 为 n 阶方阵，而 a 为 n 维向量时，$a'Aa$ 是一个数，这形式称为二次型. 一般在讨论二次型时总假定 A 为对称方阵，即其 (i, j) 元等于其 (j, i) 元，或 $A = A'$.

6.3.1 最小二乘估计

与一元的情形一样，令

$$Q(\alpha_0,\alpha_1,\cdots,\alpha_p)=\sum_{i=1}^{n}(Y_i-\alpha_0-X_{1i}\alpha_1-\cdots-X_{pi}\alpha_p)^2$$

然后找 $\alpha_0,\cdots,\alpha_p$ 之值，记为 $\hat{\beta}_0,\cdots,\hat{\beta}_p$，使上式达到最小. $\hat{\beta}_i$ 等就是 β_i 等的最小二乘估计. 作方程

$$\partial Q/\partial\alpha_0=0,\partial Q/\partial\alpha_1=0,\cdots,\partial Q/\partial\alpha_p=0$$

并加以简单的整理，即得

$$n\alpha_0=\sum_{i=1}^{n}Y_i,\text{解为}\hat{\beta}_0=\bar{Y} \tag{3.9}$$

$$\begin{cases} l_{11}\alpha_1+l_{12}\alpha_2+\cdots+l_{1p}\alpha_p=\sum_{i=1}^{n}X_{1i}Y_i \\ \cdots\cdots\cdots\cdots \\ l_{p1}\alpha_1+l_{p2}\alpha_2+\cdots+l_{pp}\alpha_p=\sum_{i=1}^{n}X_{pi}Y_i \end{cases} \tag{3.10}$$

此处 $l_{uv}=\sum_{i=1}^{n}X_{ui}X_{vi}$. 若引进以下的矩阵和向量*

$$X=\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{pmatrix},L=\begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ l_{p1} & l_{p2} & \cdots & l_{pp} \end{pmatrix} \tag{3.11}$$

$$Y_{(n)}=\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},\beta=\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\hat{\beta}=\begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix},\alpha=\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix}$$

则 $L=XX'$，方程组(3.10)右边各元分别是向量 $XY_{(n)}$ 的相应元. 于是方程组(3.10)可简写为

$$L\alpha=XY_{(n)} \tag{3.12}$$

方程组(3.10)，即(3.12)，称为正则方程. 其解，即 β 的最小二乘

---

* 矩阵 X 称为设计矩阵，但一般设计矩阵是指未经过中心化的，由原来的 X_{ij} 所构成的矩阵.

估计,可表为

$$\hat{\beta} = L^{-1}XY_{(n)} \tag{3.13}$$

一元情况中最小二乘估计的性质,在此也对*:

1. $\hat{\beta}_0, \hat{\beta}$ 分别是 β_0 和 β 的无偏估计.

2. $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_j) = 0, j = 1, \cdots, p$,即 $\hat{\beta}_0$ 与每个 $\hat{\beta}_j$ 都不相关.

3. $\mathrm{Var}(\hat{\beta}_0) = \sigma^2/n$;若记

$$C = (c_{ij}) = L^{-1} \tag{3.14}$$

则 $\mathrm{Var}(\hat{\beta}_j) = c_{jj}\sigma^2, \mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_k) = c_{jk}\sigma^2$. 由于这个性质,方阵 L^{-1} 在回归分析中有很大的重要性,一般都需要算出来. 不然的话,解方程(3.10)可用通常的消元法更简便,而无须用(3.13).

6.3.2 误差方差 σ^2 的估计

仍如一元回归一样,定义残差

$$\delta_i = Y_i - (\hat{\beta}_0 + X_{1i}\hat{\beta}_1 + \cdots + X_{pi}\hat{\beta}_p), i = 1, \cdots, n \tag{3.15}$$

及残差平方和 $\delta_1^2 + \cdots + \delta_n^2$. 可证明

$$\hat{\sigma}^2 = (\delta_1^2 + \cdots + \delta_n^2)/(n - p - 1) \tag{3.16}$$

是 σ^2 的一个无偏估计.

当随机误差服从正态分布时,可证明 $\sum_{i=1}^{n} \delta_i^2/\sigma^2$ 服从自由度 $n-p-1$ 的 χ^2 分布. 这里有 $p+1$ 个参数 $\beta_0, \beta_1, \cdots, \beta_p$ 要估计,故自由度减少了 $p+1$.

对此处多元的情况,类似于(2.23)式的结果也成立:

$$\sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \left(\hat{\beta}_1 \sum_{i=1}^{n} X_{1i}Y_i + \cdots + \hat{\beta}_p \sum_{i=1}^{n} X_{pi}Y_i\right) \tag{3.17}$$

---

* 证明见习题 7.

此式的方便之处在于:(3.17)右边括号内的各项,在列出正则方程组(3.10)时已算出了,而在估计 σ^2 时,一般先估计 β_j,故 $\hat{\beta}_1,\cdots,\hat{\beta}_p$ 等也已算出了.

6.3.3 区间估计与预测

在作区间估计和预测时,要假定随机误差服从正态分布,即要把(3.3)加强为

$$e_1,\cdots,e_n \text{ 独立同分布},e_i \sim N(0,\sigma^2),i=1,\cdots,n \tag{3.18}$$

这时,因 $\hat{\beta}_0,\cdots,\hat{\beta}_p$ 都是 $Y_1,\cdots,Y_n$ 的线性函数,它们都服从正态分布.

1. 回归系数 β_j 的区间估计

已知 $E(\hat{\beta}_j)=\beta_j$,$\mathrm{Var}(\hat{\beta}_j)=c_{jj}\sigma^2$,故有 $(\hat{\beta}_j-\beta_j)/(\sqrt{c_{jj}}\sigma)\sim N(0,1)$.以 σ 的估计 $\hat{\sigma}$ 代替上式中的 σ,则可以证明

$$(\hat{\beta}_j-\beta_j)/(\hat{\sigma}\sqrt{c_{jj}})\sim t_{n-p-1} \tag{3.19}$$

与一元情况相似,由此就可以作出 β_j 的区间估计

$$\hat{\beta}_j-\hat{\sigma}\sqrt{c_{jj}}t_{n-p-1}(\alpha/2)\leqslant\beta_j\leqslant\hat{\beta}_j+\hat{\sigma}\sqrt{c_{jj}}t_{n-p-1}(\alpha/2) \tag{3.20}$$

置信系数为 $1-\alpha$.类似地作出 β_j 的置信上、下界.

2. 回归函数的区间估计

仍记回归函数为

$$m(x)=\beta_0+\beta_1(x_1-\overline{X}_1)+\cdots+\beta_p(x_p-\overline{X}_p)$$

$\overline{X}_j$ 的意义前已指出,为 $\overline{X}_j=(X_{j1}+\cdots+X_{jn})/n$,$x=(x_1,\cdots,x_p)'$.

$m(x)$的点估计为

$$\hat{m}(x)=\hat{\beta}_0+\hat{\beta}_1(x_1-\overline{X}_1)+\cdots+\hat{\beta}_p(x_p-\overline{X}_p)$$

其期望值为 $m(x)$.其方差可根据 $\hat{\beta}_0,\cdots,\hat{\beta}_p$ 的方差与协方差算

出,结果为

$$\lambda^2(x)\sigma^2 = \left(\frac{1}{n} + \sum_{j,k=1}^{p}(x_j - \overline{X}_j)(x_k - \overline{X}_k)c_{jk}\right)\sigma^2$$

于是得到$(\hat{m}(x) - m(x))/(\lambda(x)\sigma) \sim N(0,1)$. 以 $\hat{\sigma}$ 代替 σ,得到

$$(\hat{m}(x) - m(x))/(\lambda(x)\hat{\sigma}) \sim t_{n-p-1} \tag{3.21}$$

由此就可作出 $m(x)$的区间估计为

$$\hat{m}(x) - \hat{\sigma}\lambda(x)t_{n-p-1}(\alpha/2)$$
$$\leqslant m(x) \leqslant \hat{m}(x) + \hat{\sigma}\lambda(x)t_{n-p-1}(\alpha/2) \tag{3.22}$$

置信系数为 $1-\alpha$.

在(3.22)式中令 $x_1 = \cdots = x_p = 0$,得到原模型(3.2)中的常数项 b_0 的区间估计.

3. 在自变量的值 $x_0 = (x_{10}, \cdots, x_{p0})$处预测因变量 Y 之取值 y_0

作为点预测,就用 $\hat{m}(x_0)$. 其区间预测与回归函数区间估计的差别,就在于方差多了一个 σ^2,故只须把(3.22)式中的 $\lambda(x)$改为$\sqrt{1+\lambda^2(x_0)}$即可:

$$\hat{m}(x_0) - \hat{\sigma}\sqrt{1+\lambda^2(x_0)}t_{n-p-1}(\alpha/2)$$
$$\leqslant y_0 \leqslant \hat{m}(x_0) + \hat{\sigma}\sqrt{1+\lambda^2(x_0)}t_{n-p-1}(\alpha/2) \tag{3.23}$$

其置信系数为 $1-\alpha$.

6.3.4 假设检验问题

在多元回归中,因包含了多个回归系数,可以考虑的假设检验问题,比一元情况要多些. 本段仍要假设随机误差服从正态分布.

1. 单个回归系数 β_j 的检验

考虑原假设 $H_0: \beta_j = c$, c 为给定常数,利用(3.19),仿照一元情况的处理方式,得 t 检验:

当$|\hat{\beta}_j - c| \leqslant \hat{\sigma}\sqrt{c_{jj}}t_{n-p-1}(\alpha/2)$ 时接受 H_0,不然就否定 H_0 (3.24)

类似地可考虑单边假设 $\beta_j \leqslant c$ 或 $\beta_j \geqslant c$ 的检验问题.

在应用上,主要考虑的一种情况是 $c=0$. 如果假设 $\beta_j=0$ 被接受,则可能解释为:自变量 X_j 对 Y 无影响,因而可以从回归函数中删去. 但这种解释要慎重. 一则是样本可能太少,二则还有其他原因,见 6.3.5.

2. 全体回归系数皆为 0 的检验

即原假设为

$$H_0:\beta_1=\beta_2=\cdots=\beta_p=0 \tag{3.25}$$

这个假设的检验常称为"回归显著性检验",其意思如下:若(3.25)通过了,则有可能,所选的自变量 $X_1,\cdots,X_p$ 其实对因变量 Y 无影响或影响很小. 这样,配出的经验回归方程也就没有多大意义. 在实用上,这有两种情况:一是确实 $\beta_1,\cdots,\beta_p$ 都为 0 或很小,这时我们选错了自变量;一是样本太少,随机误差的干扰太大,以致各自变量的作用显示不出来. 到底是哪种情况,当然须得对具体问题作具体分析. 但无论如何,如果假设 H_0 被接受,则总是显示,由数据配出的经验回归方程不理想,不宜迳直用于实际.

反之,若 H_0 被否定,则这说明了:所选定的自变量 $X_1,\cdots,X_p$,对因变量 Y 确有一定的影响,并非无的放矢. 通常把这说成回归达到了显著性,并进而引伸解释为:所配的回归方程成立,可以有效地使用了. 这样的解释还需慎重,因为检验的结果只是告诉我们:所选自变量中,至少有一部分是重要的,但也可能尚留有并非重要的;尤其是,并不能排斥遗漏了其他重要因素的可能性. 这一切要看前期工作做得如何,不能都委之于这个检验. 我们认为,这个检验的基本意义是事后验证性的:研究者在事前根据专业知识及经验,认为已把较重要的自变量选入了,且在一定的误差限度内,认为回归函数可取为线性的. 经过试验得出数据后,他可以通过这个检验验证一下,原来的考虑是否有毛病. 这时,若 H_0 被否定,他可以合理地解释为:数据与他事前(试验前)的设想并不矛盾. 反之,若 H_0 被接受,则提醒他,也许他事前的考虑有欠周到之处,值得再研究一下.

这里所谈的实质上涉及一个选择回归自变量的问题.在一项大型的研究中,看来与因变量 Y 有关的因素往往很多,而在回归方程中却只宜选进一部分关系最密切的,选多了反而不好.前面我们强调专业知识和经验在处理这个问题中的作用,但这并不排斥统计分析的作用.实际上,回归自变量的选择问题是回归分析中很受重视的一个课题,近 30 年来出现了大量的工作.这些在本书中无法细述了,有兴趣的读者,可参看陈希孺和王松桂所著《近代回归分析》的第三章.

现在我们回到假设(3.25)的检验问题.我们只能解释一下导出检验的思想,而不能仔细证明其中所涉及的分布问题.

前面我们在原模型(3.8)之下算出了残差平方和(3.17),其值暂记为 R_1.现如假设(3.25)成立,则无异乎说我们采纳新模型

$$Y_i = \beta_0 + e_i, i = 1, \cdots, n \tag{3.26}$$

在此模型下也计算其残差平方和 R_2,结果为

$$R_2 = \min_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{3.27}$$

对任一模型,残差平方和愈小,则说明数据对它的拟合愈好.容易看出:数据对模型(3.26)的拟合程度,决不能优于其对(3.8)的拟合程度,因为(3.8)中可供选择的余地比(3.26)大.但拟合程度相差多少,则取决于模型(3.26)是否正确,即假设(3.25)是否成立.若(3.26)正确,则差距要小些,否则就大些*.这样,R_2 和 R_1 之差 $R_2 - R_1$ 可作为假设 H_0 正确性的一种度量:$R_2 - R_1$ 愈小,H_0 愈像是成立.理论上可以证明:当 H_0 成立时有

$$\frac{1}{\sigma^2}(R_2 - R_1) \sim \chi_p^2, R_2 - R_1 \text{ 与 } \hat{\sigma}^2 \text{ 独立}$$

这样,再注意当随机误差服从正态分布时有

---

* 这是一种直观的想法,其根据在于:与数据拟合最好的模型,是在真模型附近而不是远离它——如果远离它(这并非不可能),则表示经验回归方程与理论回归方程差距很大,整个分析就没有多大意义了.

$$(n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p-1}$$

于是,由 F 分布的定义,知当原假设 H_0 成立时有

$$\frac{1}{p}(R_2-R_1)/\hat{\sigma}^2 \sim F_{p,n-p-1} \tag{3.28}$$

按(3.27)和(3.17),得

$$R_2-R_1=\hat{\beta}_1\sum_{i=1}^{n}X_{1i}Y_i+\cdots+\hat{\beta}_p\sum_{i=1}^{n}X_{pi}Y_i \tag{3.29}$$

于是得到(3.25)的 H_0 的下述检验法:

$$\text{当}\frac{1}{p}\sum_{j=1}^{p}\hat{\beta}_j\sum_{i=1}^{n}X_{ji}Y_i/\hat{\sigma}^2 \leqslant F_{p,n-p-1}(\alpha)\text{ 时接受 }H_0,$$
$$\text{不然就否定 }H_0 \tag{3.30}$$

检验水平为 α.这个检验称为 H_0 的 F 检验.

3. 一部分回归系数为 0,即

$$H_0:\beta_1=\cdots=\beta_r=0 \quad (1\leqslant r\leqslant p) \tag{3.31}$$

检验的背景是:全体自变量按其性质分成一些组,而 $X_1,\cdots,X_r$ 是反映某方面性质的因子.(3.31)的意义是:这方面的因子其实不影响因变量 Y 之值.

检验方法与(3.25)同:以 R_3 记当(3.31)成立时的残差平方和,即

$$R_3=\min_{\alpha_0,\alpha_{r+1},\cdots,\alpha_p}\sum_{i=1}^{n}(Y_i-\alpha_0-X_{r+1,i}\alpha_{r+1}-\cdots-X_{pi}\alpha_p)^2$$

然后,可以证明:当随机误差服从正态分布而 H_0 成立时,有

$$\frac{1}{r}(R_3-R_1)/\hat{\sigma}^2 \sim F_{r,n-p-1}$$

于是得到(3.31)的下述检验法:

$$\text{当}\frac{1}{r}(R_3-R_1)/\hat{\sigma}^2 \leqslant F_{r,n-p-1}(\alpha)\text{ 时接受 }H_0,\text{不然就否定 }H_0 \tag{3.32}$$

检验水平为 α.这个检验通称为假设(3.31)的 F 检验.称呼的来由显然是,所用的检验统计量有 F 分布.

直接计算 R_3 需要在新模型

$$Y_i = \beta_0 + \beta_{r+1}X_{r+1,i} + \cdots + \beta_pX_{pi} + e_i, i = 1, \cdots, n \tag{3.33}$$

之下算出 $\beta_0, \beta_{r+1}, \cdots, \beta_p$ 的最小二乘估计 $\beta_0^*, \beta_{r+1}^*, \cdots, \beta_p^*$. β_0^* 仍为 $\overline{Y}$，但 $\beta_{r+1}^*, \cdots, \beta_p^*$ 已与在原模型(3.8)之下求出的 $\beta_{r+1}, \cdots, \beta_p$ 的最小二乘估计 $\hat{\beta}_{r+1}, \cdots, \hat{\beta}_p$ 不同，因此涉及较多计算. 下面的公式则只须用到原模型(3.8)下有关的量，不须涉及新模型(3.33)，因此较为简单. 为引进这公式，把(3.11)式定义的方阵 L 分块为

$$L = \begin{pmatrix} L_{11} \vdots L_{12} \\ \cdots\cdots\cdots \\ L_{21} \vdots L_{22} \end{pmatrix}$$

其中 L_{11} 为 r 阶方阵，记方阵

$$D = (d_{ij}) = L_{11} - L_{12}L_{22}^{-1}L_{21}$$

则

$$R_3 - R_1 = \sum_{i,j=1}^{r} d_{ij}\hat{\beta}_i\hat{\beta}_j \tag{3.34}$$

(3.34)中，$\hat{\beta}_i$ 等是在原模型(3.8)之下已求得的.

线性回归是统计学应用中碰得最多的. 本节方法中涉及的运算，早已编入各种统计软件包，如有这种设备，则只须输入数据即可. 这类简化公式也就没有多大实际意义了.

例 3.1 本例引述自张启锐著《实用回归分析》p.60. 其目的纯粹是为了显示，本节提出的那些抽象公式是怎样使用的.

本例共有三个自变量 X_1, X_2, X_3，因变量 Y. 对这些变量进行了 $n = 48$ 次观测，原始数据 $(X_{1i}, X_{2i}, X_{3i}, Y_i)$, $i = 1, \cdots, 48$，没有写出，但与本节公式的应用有关的量的计算结果为

$$\overline{X}_1 = 18.98, \overline{X}_2 = 2.55, \overline{X}_3 = 3.125, \overline{Y} = 3.843$$

$$L = \begin{pmatrix} 2052.98 & 49.15 & 782.12 \\ 49.15 & 12.46 & 13.50 \\ 782.12 & 13.50 & 577.25 \end{pmatrix}, \sum_{i=1}^{48}(Y_i - \overline{Y})^2 = 74.15$$

$$\sum_{i=1}^{48}(X_{1i}-\overline{X}_1)Y_i=-257.59,\ \sum_{i=1}^{48}(X_{2i}-\overline{X}_2)Y_i=-11.72,$$

$$\sum_{i=1}^{48}(X_{3i}-\overline{X}_3)Y_i=-141.37$$

1. 常数项 β_0 的最小二乘估计为 $\overline{Y}=3.843$，而回归系数 β_1，β_2，β_3 的最小二乘估计则是下述方程组的解：

$$2052.98\alpha_1+49.15\alpha_2+782.12\alpha_3=-257.59$$

$$49.15\alpha_1+12.46\alpha_2+13.50\alpha_3=-11.72$$

$$782.12\alpha_1+13.50\alpha_2+577.25\alpha_3=-141.37$$

解 $\alpha_1,\alpha_2,\alpha_3$，即 β_1,β_2,β_3 的最小二乘估计 $\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$，结果为 $\hat{\beta}_1=-0.0488$，$\hat{\beta}_2=-0.5688$，$\hat{\beta}_3=-0.1655$. 而经验回归方程为

$$\begin{aligned}y&=3.843-0.0488(x_1-18.98)-0.5688(x_2-2.55)\\&\quad-0.1655(x_3-3.125)\\&=6.737-0.0488x_1-0.5688x_2-0.1655x_3\end{aligned}\tag{3.35}$$

为计算 $\hat{\beta}_0,\cdots,\hat{\beta}_3$ 的方差协方差，要算出 L 的逆方阵 $C=L^{-1}$，结果为

$$C=L^{-1}=10^{-3}\begin{pmatrix}1.0931&-2.7775&-1.4160\\-2.7775&89.4009&1.6725\\-1.4160&1.6725&3.6119\end{pmatrix}\tag{3.36}$$

于是得到

$$\mathrm{Var}(\hat{\beta}_0)=\sigma^2/48=0.0208\sigma^2,\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_j)=0,j=1,2,3$$

$$\mathrm{Var}(\hat{\beta}_1)=10^{-3}\times1.0931\sigma^2,\mathrm{Var}(\hat{\beta}_2)=10^{-3}\times89.4009\sigma^2$$

$$\mathrm{Var}(\hat{\beta}_3)=10^{-3}\times3.6119\sigma^2$$

$$\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)=10^{-3}\times(-2.7775)\sigma^2$$

$$\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_3)=10^{-3}\times(-1.4160)\sigma^2$$

$\mathrm{Cov}(\hat{\beta}_2,\hat{\beta}_3)=10^{-3}\times 3.6119\sigma^2$

2. 残差平方和按公式(3.17)计算,结果为

$$\begin{aligned}\sum_{i=1}^{48}\delta_i^2 &= 74.15-(-0.0488)(-257.59)\\ &\quad -(-0.5688)(-11.72)\\ &\quad -(-0.1655)(-141.37)\\ &= 31.5165\end{aligned}$$

自由度为 $n-p-1=48-3-1=44$,而得到误差方差 σ^2 的无偏估计 $\hat{\sigma}^2$ 为:$\hat{\sigma}^2=31.5216/44=0.7163$.

3. 各回归系数的区间估计,取置信系数 $1-\alpha=0.95$,查 t 分布表,$t_{44}(0.025)=2.02108$. 于是按(3.20),β_j 的区间估计有 $\hat{\beta}_j\pm\sqrt{0.7163}\times\sqrt{c_{jj}}\times 2.02108$ 的形式. 以(3.36)中的 c_{jj} 具体值代入,算出结果为:

$$\begin{aligned}&\beta_1:\ -0.0488\pm 0.0564;\beta_2:\ -0.5688\pm 0.5100;\\ &\beta_3:\ -0.1655\pm 0.1026\end{aligned}\qquad(3.37)$$

4. 回归函数

$m(x)=\beta_0+\beta_1(x_1-18.98)+\beta_2(x_2-2.55)+\beta_3(x_3-3.125)$的区间估计,按公式(3.22),应为 $\hat{m}(x)\pm\sqrt{0.7164}\times\lambda(x)\times 2.02108$. 其中 $\hat{m}(x)$即为方程(3.35)的右边的表达式,而

$$\begin{aligned}\lambda^2(x) &= 1/48+\{1.0931(x_1-18.98)^2+89.4009(x_2-2.55)^2\\ &\quad +3.6119(x_3-3.125)^2\\ &\quad -2\times 2.7775(x_1-18.98)(x_2-2.55)\\ &\quad -2\times 1.416(x_1-18.98)(x_3-3.125)\\ &\quad +2\times 1.6725(x_2-2.55)(x_3-3.125)\}\times 10^{-3}\end{aligned}$$

例如,对点 $x=(18,2.7,3)'$,上式计算结果为

$$\lambda^2(x)=0.02443,\hat{m}(x)=3.8263$$

而得到其置信系数 0.95 的区间估计为

$$3.8263 \pm \sqrt{0.7164} \times \sqrt{0.02443} \times 2.02108$$
$$= 3.8163 \pm 0.2674$$

在 x 点处 Y 的预测值 y_0 的 0.95 置信区间为 $\hat{m}(x) \pm \sqrt{0.7164}(1+\lambda^2(x))^{1/2} \times 2.02108$. 在点 $x=(18,2.7,3)'$ 处，结果为

$$3.8263 \pm \sqrt{0.7164} \times \sqrt{1.02443} \times 2.02108$$
$$= 3.8263 \pm 1.7314$$

看出预测的精度比回归函数估计的精度差得多.

5. 假设检验

一个回归系数为 0 的检验结果(取水平 $\alpha=0.05$)，从各回归系数的区间估计即得出：凡是 β_j 的置信区间包含 0 者，原假设 $\beta_j=0$ 就被接受，不然就被否定. 因此，从(3.37)看出，$\beta_1=0$ 被接受，而 $\beta_2=0$ 及 $\beta_3=0$ 都被否定.

$\beta_1=0$ 虽然被接受，但这并不等于说一定可以把自变量 X_1 去掉. 这个问题还要根据具体情况全面地去考虑，不能单凭这个检验就作出决定.

其次看原假设 $H_0:\beta_1=0,\beta_2=0$. 用检验(3.32)，要按(3.34)式算出 R_3-R_1. 有

$$L_{11}=\begin{pmatrix}2052.98 & 49.15\\ 49.15 & 12.46\end{pmatrix}, L_{12}=\begin{pmatrix}782.12\\ 13.50\end{pmatrix}$$
$$L_{21}=(782.12,13.50), L_{22}=(577.25)$$

于是

$$D=L_{11}-L_{12}L_{22}^{-1}L_{21}=\begin{pmatrix}2052.98 & 49.15\\ 49.15 & 12.46\end{pmatrix}$$
$$-\begin{pmatrix}782.12\\ 13.50\end{pmatrix}\left(\frac{1}{577.25}\right)(782.12,13.50)$$
$$=\begin{pmatrix}2052.98 & 49.15\\ 49.15 & 12.46\end{pmatrix}-\begin{pmatrix}1059.70 & 18.29\\ 18.29 & 0.32\end{pmatrix}$$
$$=\begin{pmatrix}993.18 & 30.86\\ 30.86 & 12.14\end{pmatrix}$$

于是，据 $\hat{\beta}_1=-0.0488,\hat{\beta}_2=-0.5688$，用(3.34)，得

$$R_3-R_1=993.18(0.0488)^2+12.14(0.5688)^2+2(30.86)(0.0488)(0.5688)=8.006$$

$r=2,\hat{\sigma}^2=0.7164$. 故

$$\frac{1}{r}(R_3-R_1)/\hat{\sigma}^2=\frac{1}{2}\times 8.006/0.7164=5.588$$

查 F 分布表，知 $F_{r,n-p-1}(\alpha)=F_{2,44}(0.05)\approx 3.21$. 故 H_0 被否定.

最后考虑检验问题 $H_0:\beta_1=\beta_2=\beta_3=0$. 用检验(3.30)，其检验统计量之分子为

$$\frac{1}{3}((-0.0488)(-257.59)+(-0.5688)(-11.72)+(-0.1655)(-141.347))=14.211$$

故(3.30)中的检验统计量之值为 $14.211/0.7164=19.837$.

因为 $F_{p,n-p-1}(\alpha)=F_{3,44}(0.05)\approx 2.82$，故 H_0 被否定.

6.3.5 应用上值得注意的几个问题

在一元回归应用上所曾提出过的那些值得注意之点，在此仍然有效. 多元回归情况更加复杂，在其结果的解释上更应慎重.

1. 设 Y 对自变量 X_j 的回归系数估计值为 $\hat{\beta}_j$，通常把它解释为：当 X_j 增减 1 单位时，平均说来因变量 Y 增减 $\hat{\beta}_j$ 单位. 如果 X_j 的取值能由人控制，其范围在建立经验回归方程时所用数据的范围内，且在尔后的使用时，其条件与建立回归方程时的条件相当，则这个解释可以认为是合理的.

如果 X_j 本身也是随机的，则情况复杂，不仅在一元情况下所讲的那些问题此处都存在，而且还有一个各自变量之间的相关问题. 如果自变量为随机的，它们一般不见得独立，即一个变量，例如 X_j，其值的变动往往会带动其他变量的值作变动. 这时，各回归系数的值，都是在全体自变量值的联合变动的格局内起作用，孤立地

抽一个去考察就不一定很现实了.在这种情况下,尤其不能人为地去设法变动其中一个(例如 X_j)之值而强行压住其他自变量值保持不变.在这样人为干预下所作的预测往往与实际相去甚远.

在使用线性回归时我们必须牢记一个基本点:真实的回归函数,特别在较大的范围内,很少是线性的.线性是一种近似.它包含了一种从实际角度看往往不一定合理的假定:它认为各变量的作用与其他变量取什么值无关,且各变量的作用可以叠加.因为,若 $y=b_0+b_1x_1+\cdots+b_px_p$,则不论你把 $x_2,\cdots,x_p$ 之值固定在何处,当 x_1 增减 1 单位时:y 总是增减 b_1 单位.事实常不如此.例如,以 Y 记某种农作物的亩产量,X_1,X_2,X_3 记每亩播种量,施肥量与耕作深度,则 X_1 起的作用如何,与 X_2,X_3 之值有关,其他亦然.这种现象称为各因素之间的"交互作用".如果专业知识或经验告诉我们,至少有一部分自变量之间有显著的交互作用存在,则在自变量值较大的范围内采用线性回归就不会有很好的效果.且在这种情况下,单个回归系数意义的解释,也应是基于其他变量的平均而言.

2.在实际应用中,一个回归模型内可包含为数甚多的自变量,其中难免有些是密切相关的.例如,若 X_1 和 X_2 高度线性相关,则 X_1 起的作用,基本上可由 X_2 挑起来.反之亦然.这样,如果你从方程中删除自变量 X_1,X_2 中的一个,而对剩下的 $p-1$ 个自变量再配出方程,实际效果与原来的相当.这就造成下述在假设检验上看来矛盾的现象:"$\beta_1=0$"或"$\beta_2=0$"都可以被接受,而"$\beta_1=\beta_2=0$"则被否定.

所以,如果自变量是随机的,则对它们之间的相关性的了解很重要.这有助于删去那些不需要的自变量,使配出的回归方程有更好的稳定性,并简化对回归方程的解释.

3.为得出回归系数的估计值,要解线性方程组(3.10),如果系数方阵 L 的行列式 $|L|=0$,则方程组(3.10)无解.在应用上可能碰到这样的情况:$|L|$ 不为 0 但很接近于 0.这时,诸系数 l_{uv} 在计算上一点点误差也可能导致方程组(3.10)的解的重大改变,因

而回归系数的估计值就失掉了其稳定性和可信性.

这种情况在统计上称为"复共线性",意指若干个自变量之间存在着高度的线性关系.在作多元线性回归分析时,复共线性是一个很有破坏性的东西.凡是可能,应极力予以避免.如果各自变量取值可人为控制,自可通过适当的设计达到这一点.如果自变量是随机的.通过分析其相关性并删去若干不必要的(可由其他自变量代替的)自变量,可能达到这一点.如这些都不成,则不宜强行使用最小二乘法,可考虑用其他更富稳定性的方法取代之.这个问题涉及太宽,不能在此细述.关于复共线性,张启锐的《实用回归分析》第六章可以参考.关于回归系数的种种估计方法(最小二乘法以外的方法),可参看陈希孺及王松桂的《近代回归分析》第四章,及上引张启锐的书第九章.

6.3.6 可转化为线性回归的模型

有时,回归函数并非自变量的线性函数,但通过取用新自变量,可以转化为线性回归去处理.举几个例子说明这一点.

例 3.2 设有一个自变量 X 和因变量 Y.如从某种理论考虑或数据的启示,认为回归模型有指数形式

$$Y = b_0 + b_1 e^{cX} + e$$

其中常数 c 已知,b_0, b_1 未知,e 为随机误差.则通过取新自变量 $Z = e^{cX}$ 将其转化为一元线性回归:

$$Y = b_0 + b_1 Z + e \tag{3.38}$$

若在原模型下对(X, Y)有了观测数据$(X_1, Y_1), \cdots, (X_n, Y_n)$,则等于在新模型下有了观测数据$(Z_1, Y_1), \cdots, (Z_n, Y_n)$,其中 $Z_i = e^{cX_i}, i = 1, \cdots, n$.若 c 也未知,则这一做法失效.

例 3.3 仍设有一个自变量 X 和因变量 Y,并认为回归函数为 X 的多项式:

$$Y = b_0 + b_1 X + b_2 X^2 + \cdots + b_p X^p + e \tag{3.39}$$

引进 p 个新自变量 $X_1, \cdots, X_p$,其中 $X_j = X^j, j = 1, \cdots, p$,则模型

(3.39)转化为有 p 个自变量 $X_1,\cdots,X_p$ 的多元线性回归

$$Y = b_0 + b_1X_1 + \cdots + b_pX_p + e \tag{3.40}$$

若在原模型下对 (X,Y) 有了观测数据 $(X_1,Y_1),\cdots,(X_n,Y_n)$,则等于在新模型(3.40)下有了观测数据

$$(X_{1i},\cdots,X_{pi},Y_i),i = 1,\cdots,n$$

其中 $X_{ji} = X_i^j, j = 1,\cdots,p, i = 1,\cdots,n$.

(3.39)称为"多项式回归",是一个应用较多的回归模型.经过转化后的回归模型(3.40)成为多元的.变换以后的自变量 $X_1,\cdots,X_p$ 之间有严格的函数关系,这没有关系.因为在前面讨论线性回归时,并没有对自变量之间可能有的关系作过任何限制.

在模型(3.39)之下,假设"$b_p = 0$"有特殊的意义,比方说,一开始我们较有把握认为取 2 阶多项式已够了,但还不太放心,希望检验一下.于是我们取模型(3.39)而令 $p = 3$.若假设"$b_3 = 0$"通过了,则数据不与我们原先的想法(回归取为 2 阶多项式已足)矛盾.否则就须调整原来的想法.

多个变元的多项式回归也一样变换.例如,包含两个自变量 X_1,X_2 的二次多项式回归模型

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_2^2 + b_5X_1X_2 + e$$

可通过采用新自变量

$$Z_1 = X_1, Z_2 = X_2, Z_3 = X_1^2, Z_4 = X_2^2, Z_5 = X_1X_2$$

化为多元线性模型

$$Y = b_0 + b_1Z_1 + \cdots + b_5Z_5 + e$$

在有些情况下,不仅自变量可施行变换,对因变量也这样做.例如 X,Y 有回归方程 $y = b_0e^{b_1x}$,b_0、b_1 未知,这不是线性的,也不能通过自变量的变换化为线性的.但若令 $Z = \log Y$,则 $Z = \log b_0 + b_1X = \beta_0 + \beta_1X(\beta_0 = \log b_0, \beta_1 = b_1)$,而化为线性的.

不过对因变量所作的变换,较之对自变量所作的变换,存在一个理论上的问题.即自变量的变换不改变模型中的随机误差 e 这一项.因此,有关 e 的假设(如均值为 0,方差非 0 有限,或 e 服从

正态分布之类)全都保持有效,对因变量之变换则不然.拿本例来说,原模型为

$$Y = b_0 e^{b_1 X} + e \tag{3.41}$$

把 Y 换成 $Z = \log Y$,得 $Z = \log(b_0 e^{b_1 X} + e)$.形式上可写为

$$Z = b_0 + b_1 X + \varepsilon, \varepsilon = \log(1 + e b_0^{-1} e^{-b_1 X}) \tag{3.42}$$

ε 已不能满足 e 原有的条件,甚至还和 X 有关.

因此,在对因变量作变换时,我们不是拘泥于从(3.41)到(3.42)这种形式运算.而是从头开始:我们觉得并认定,若取 $Z = \log Y$ 为因变量,则 X, Z 的回归很近似线性,不妨就认为它有(3.42)的形式而 ε 满足以往对 e 施加的条件.这有其道理可讲:因为反正原模型(3.41)中 e 的性质,也无非是一种假定而已,并非先天绝对无误.转化成(3.42)后,我们未尝不可对 ε 作出类似的假定,并无先天的理由认为:对 ε 的假定一定不如对 e 的假定那样符合事实.

更进一步,为达到线性回归,有时对自变量和因变量都要施加变换,其方法和道理与上同.例如,若回归方程为 $y = b_0 e^{b_1/x}$,则通过变换 $u = 1/x$, $v = \log y$,转化为线性型 $v = \log b_0 + b_1 u$.