

第四章 模型的推断与预测

上一章讨论了回归系数的几种点估计方法, 从而建立了回归方程. 但是, 所建立的回归方程是否刻画了因变量和自变量(整体)之间真实的相依关系呢? 从统计理论的角度, 可以通过假设检验进行分析, 称这部分内容为回归方程的显著性检验. 另外, 我们还希望研究因变量是否真正依赖于一个或几个特定的自变量, 称这部分内容为回归系数的显著性检验. 本章前四节将讨论上述假设检验问题, 第五节讨论异常点的假设检验, 第六节讨论模型误差的不相关性检验, 第七节讨论回归系数的区间估计, 最后一节讨论在给定自变量取值的情况下如何预测因变量的大小和范围.

4.1 一般线性假设

考虑正态线性回归模型

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (4.1.1)$$

其中 \mathbf{X} 为 $n \times (p+1)$ 列满秩设计矩阵. 下面关注比较一般的线性假设

$$H: \mathbf{A}\boldsymbol{\beta} = \mathbf{b} \quad (4.1.2)$$

的假设检验问题. 这里 \mathbf{A} 为 $m \times (p+1)$ 矩阵, 秩为 m , \mathbf{b} 为 $m \times 1$ 的已知向量. 实际应用中许多感兴趣的问题都可以归结为(4.1.2)的假设检验问题(例如, 后面介绍的回归方程的显著性检验和回归系数的显著性检验).

先了解检验方法的基本思想. 应用最小二乘法, 得到 $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ 和残差平方和

$$\text{RSS} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}. \quad (4.1.3)$$

RSS反映了实际数据与模型(4.1.1)的拟合程度. 现在在模型(4.1.1)上附加线性假设(4.1.2), 再应用最小二乘法得到约束最小二乘估计(见第三章)

$$\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b}) \quad (4.1.4)$$

和相应的残差平方和

$$\text{RSS}_H = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H)'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H). \quad (4.1.5)$$

加了约束条件(4.1.2)后, 回归系数 $\boldsymbol{\beta}$ 的参数空间变小了, 因而残差平方和 RSS_H 自然就变大了. 于是总有 $\text{RSS}_H \geq \text{RSS}$. 若回归系数确实满足约束条件(4.1.2), 那么加上约束条件和不加约束条件本质上是一样的, 此时 $\text{RSS}_H - \text{RSS}$ 应偏小. 若约束条件(4.1.2)不成立, 则 $\text{RSS}_H - \text{RSS}$ 应偏大. 所以, 当 $\text{RSS}_H - \text{RSS}$ 偏大到一定程度时, 我们就有充分的理由拒绝假设(4.1.2), 否则只能接受它.

当谈到一个统计量的大小的时候, 往往需要一个比较的标准. 在这里, 可以把RSS取为标准. 于是用

$$(\text{RSS}_H - \text{RSS})/\text{RSS}$$

的大小来决定接受还是拒绝(4.1.2). 值得一提的是, 也可根据似然比检验的方法推导出此统计量.

定理4.1.1(最小二乘法基本定理) 对于正态线性回归模型(4.1.1),

- (a) $RSS/\sigma^2 \sim \chi^2(n - \text{rk}(\mathbf{X}))$;
- (b) 若假设(4.1.2)成立, 则 $(RSS_H - RSS)/\sigma^2 \sim \chi^2(m)$;
- (c) RSS 与 $RSS_H - RSS$ 相互独立;
- (d) 若假设(4.1.2)成立, 则

$$F_H = \frac{(RSS_H - RSS)/m}{RSS/(n - \text{rk}(\mathbf{X}))} \sim F(m, n - \text{rk}(\mathbf{X})), \quad (4.1.6)$$

$F(m, n - \text{rk}(\mathbf{X}))$ 表示自由度为 $(m, n - \text{rk}(\mathbf{X}))$ 的 F 分布.

注 (1) 统计量 F_H 的分子 $(RSS_H - RSS)/m$ 表示每增加一个约束, 残差平方和的平均增量. 分母 $RSS/(n - \text{rk}(\mathbf{X}))$ 起着正则化的作用, 用来消除 F_H 的分子 $(RSS_H - RSS)/m$ 的量纲. (2) 定理中的两个残差平方和的几何意义见图4.1.1.

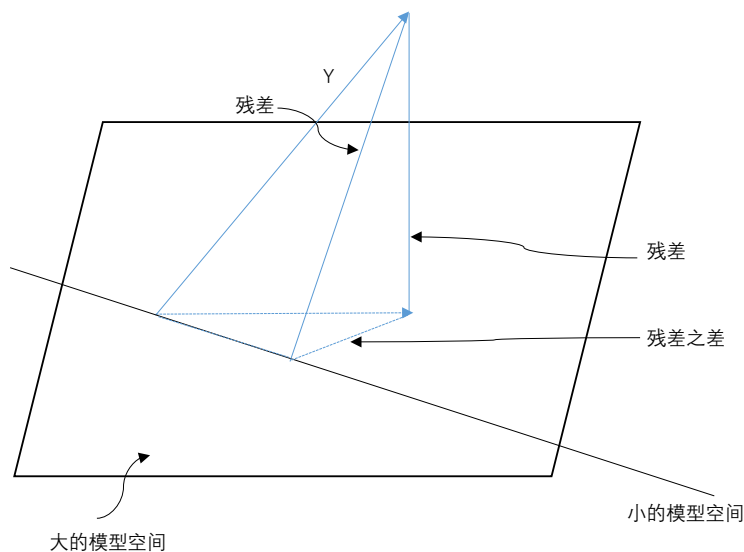


图4.1.1 最小二乘法基本定理的几何解释

证明 (a) 已在定理3.2.4中给出.

(b) 由第三章, 可知

$$\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H\|^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H)\|^2,$$

即

$$RSS_H = RSS + (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H)' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H). \quad (4.1.7)$$

利用 $\hat{\beta}_H$ 的表达式(4.1.4)可得

$$\text{RSS}_H - \text{RSS} = (\mathbf{A}\hat{\beta} - \mathbf{b})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta} - \mathbf{b}). \quad (4.1.8)$$

因为 $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, 所以由定理2.3.2知

$$\mathbf{A}\hat{\beta} - \mathbf{b} \sim N(\mathbf{A}\beta - \mathbf{b}, \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}').$$

若假设(4.1.2)成立, 则

$$\mathbf{A}\hat{\beta} - \mathbf{b} \sim N(\mathbf{0}, \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}').$$

再应用推论2.4.1可知 $(\text{RSS}_H - \text{RSS})/\sigma^2 \sim \chi^2(m)$.

(c) 因为

$$\begin{aligned} \mathbf{A}\hat{\beta} - \mathbf{b} &= \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e}) - \mathbf{b} \\ &= \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} + (\mathbf{A}\beta - \mathbf{b}), \end{aligned}$$

代入(4.1.8)得

$$\begin{aligned} \text{RSS}_H - \text{RSS} &= \mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \\ &\quad + 2(\mathbf{A}\beta - \mathbf{b})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \\ &\quad + (\mathbf{A}\beta - \mathbf{b})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\beta - \mathbf{b}) \\ &=: \mathbf{e}'\mathbf{M}\mathbf{e} + 2\mathbf{c}'\mathbf{e} + \Theta, \end{aligned}$$

其中

$$\begin{aligned} \mathbf{M} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \\ \mathbf{c}' &= (\mathbf{A}\beta - \mathbf{b})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \\ \Theta &= (\mathbf{A}\beta - \mathbf{b})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\beta - \mathbf{b}). \end{aligned}$$

注意 Θ 为非随机项. 记 $\mathbf{N} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (注意 $\mathbf{X}'\mathbf{N} = \mathbf{0}$), 于是

$$\text{RSS} = \mathbf{e}'\mathbf{N}\mathbf{e}.$$

因此, 为证 $\text{RSS}_H - \text{RSS}$ 与 RSS 独立, 只需证 $\mathbf{e}'\mathbf{M}\mathbf{e}$ 和 $\mathbf{c}'\mathbf{e}$ 都与 $\mathbf{e}'\mathbf{N}\mathbf{e}$ 独立. 由于

$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n), \quad \mathbf{N} \cdot \sigma^2\mathbf{I}_n \cdot \mathbf{M} = \mathbf{0}, \quad \mathbf{c}' \cdot \sigma^2\mathbf{I}_n \cdot \mathbf{N} = \mathbf{0},$$

根据推论2.4.10和推论2.4.11, 结论成立.

(d) 由(a)、(b)、(c)可直接推知(d)成立.

定理4.1.1(d)给出了线性假设 $\mathbf{A}\beta = \mathbf{b}$ 的检验统计量. 对于给定的显著性水平 α , 假设检验的拒绝域为

$$\{\text{样本}: F_H > F_\alpha(m, n - \text{rk}(\mathbf{X}))\}.$$

若 $[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}$ 的计算比较容易, 那么可通过(4.1.8)直接计算 $\text{RSS}_H - \text{RSS}$ 的大小. 但大部分情况下, 计算 $[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}$ 是不容易的, 此时我们可以分别计算 RSS 与 RSS_H 的大小, 然后再计算 $\text{RSS} - \text{RSS}_H$.

在计算RSS与RSS_H的时候, RSS可通过下列公式计算

$$\text{RSS} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}, \quad (4.1.9)$$

而计算RSS_H时可通过把约束条件 $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ “融入”到原来的模型(从而转化为一个无约束的模型, 称之为约简模型)来计算. 例如, 把模型(4.1.1)写成

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i, \quad i = 1, \cdots, n. \quad (4.1.10)$$

若要检验 $\beta_1 = \beta_2 = \beta_3$, 把线性假设写成线性约束 $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, 其中

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & -1 & \cdots & 0 \end{pmatrix},$$

其秩为2. 将 $\beta_1 = \beta_2 = \beta_3$ 融入到原模型得约简模型

$$y_i = \beta_0 + \beta_1(x_{i1} + x_{i2} + x_{i3}) + \beta_4 x_{i4} + \cdots + \beta_p x_{ip} + e_i, \quad i = 1, \cdots, n.$$

约简模型中的未知参数向量为 $\boldsymbol{\alpha} = (\beta_0, \beta_1, \beta_4, \cdots, \beta_p)'$, 设计矩阵是将原先的设计矩阵的第2-4列求和而得到的 $n \times (p-1)$ 矩阵 $\widetilde{\mathbf{X}}$. 对约简模型应用最小二乘法, 得 $\boldsymbol{\alpha}$ 的最小二乘估计

$$\hat{\boldsymbol{\alpha}} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Y},$$

相应的残差平方和为

$$\text{RSS}_H = (\mathbf{Y} - \widetilde{\mathbf{X}}\hat{\boldsymbol{\alpha}})'(\mathbf{Y} - \widetilde{\mathbf{X}}\hat{\boldsymbol{\alpha}}) = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\alpha}}'\widetilde{\mathbf{X}}'\mathbf{Y}. \quad (4.1.11)$$

例4.1.1 假设

$$\begin{cases} y_1 = \beta_1 + e_1, \\ y_2 = 2\beta_1 - \beta_2 + e_2, \\ y_3 = \beta_1 + 2\beta_2 + e_3, \end{cases}$$

其中 $\mathbf{e} = (e_1, e_2, e_3)' \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_3)$. 检验 $H: \beta_1 = \beta_2$.

将观测数据写成线性回归模型

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}.$$

把线性假设 H 写成线性约束 $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$:

$$(1, -1) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = 0.$$

因此, $\mathbf{A} = (1, -1)$, $m = \text{rk}(\mathbf{A}) = 1$, $n = 3$, $\text{rk}(\mathbf{X}) = 2$. 易求 $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ 的LSE为

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{6}(y_1 + 2y_2 + y_3) \\ \frac{1}{5}(-y_2 + 2y_3) \end{pmatrix},$$

相应的残差平方和为

$$\text{RSS} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \sum_{i=1}^3 y_i^2 - 6\hat{\beta}_1^2 - 5\hat{\beta}_2^2.$$

将 $\beta_1 = \beta_2 =: \alpha$ 融入模型, 得约简模型

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix} \alpha + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} =: \widetilde{\mathbf{X}}\alpha + \mathbf{e}.$$

从而得到 α 的LSE

$$\hat{\alpha} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Y} = \frac{1}{11}(y_1 + y_2 + 3y_3),$$

相应的残差平方和为

$$\text{RSS}_H = \mathbf{Y}'\mathbf{Y} - \hat{\alpha}'\widetilde{\mathbf{X}}'\mathbf{Y} = \sum_{i=1}^3 y_i^2 - 11\hat{\alpha}^2.$$

利用RSS和 RSS_H 的表达式, 通过一些简单的计算, 得到检验统计量

$$F_H = \frac{6\hat{\beta}_1^2 + 5\hat{\beta}_2^2 - 11\hat{\alpha}^2}{\sum_{i=1}^3 y_i^2 - 6\hat{\beta}_1^2 - 5\hat{\beta}_2^2}.$$

若 H 成立, 则

$$F_H \sim F(m, n - \text{rk}(\mathbf{X})) = F(1, 1).$$

因此, 给定显著性水平 α , 假设检验的拒绝域为

$$\{\text{样本}: F_H > F_\alpha(1, 1)\}.$$

例4.1.2(同一模型检验) 假设因变量 y 和自变量 x_1, \dots, x_p 有两批观测数据, 第一批数据来自线性回归模型

$$y_i = \beta_0^{(1)} + \beta_1^{(1)}x_{i1} + \dots + \beta_p^{(1)}x_{ip} + e_i, \quad i = 1, \dots, n_1;$$

第二批数据来自另一线性回归模型

$$y_i = \beta_0^{(2)} + \beta_1^{(2)}x_{i1} + \dots + \beta_p^{(2)}x_{ip} + e_i, \quad i = n_1 + 1, \dots, n_1 + n_2,$$

其中 $\{e_1, \dots, e_{n_1+n_2}\}$ 独立同分布, 服从 $N(0, \sigma^2)$. 问: 这两批数据是否来自同一个模型? 即检验

$$\beta_i^{(1)} = \beta_i^{(2)}, \quad i = 0, 1, \dots, p.$$

这个问题具有广泛的背景. 例如, 这两批数据可以是同一公司在两个不同时间段上的数据, y 表示公司经济效益的某项指标. 那么上述检验就是考察公司的效益指标对诸因素的相依关系在两个不同的时间段上是否有了变化. 再如, 在生

物理学研究中, 很多实验花费时间较长, 因此需要把实验分配在几个实验室同时进行. 这时前面讨论的两批数据可看成是来自两个不同实验室的数据, 而上述检验的目的就是考察两个实验室所得结论有没有差异.

下面来推导检验统计量. 把两个模型写成矩阵形式

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{e}_1, \quad \mathbf{e}_1 \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_1}), \\ \mathbf{Y}_2 &= \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{e}_2, \quad \mathbf{e}_2 \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_2}). \end{aligned}$$

将它们合并, 得原模型

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_1+n_2}). \quad (4.1.12)$$

注意这里的设计矩阵

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix},$$

$\text{rk}(\mathbf{X}) = 2(p+1)$, 回归系数向量 $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$, 样本容量 $n = n_1 + n_2$. 把要检验的假设写成

$$H: (\mathbf{I}_{p+1} \quad -\mathbf{I}_{p+1}) \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \mathbf{0}. \quad (4.1.13)$$

注意到 $m = \text{rk}(\mathbf{I}_{p+1} \quad -\mathbf{I}_{p+1}) = p+1$. 从(4.1.12)得到的 $\boldsymbol{\beta}_1$ 和 $\boldsymbol{\beta}_2$ 的LSE

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{X}_1' \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2' \mathbf{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2' \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{X}_1' \mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_2' \mathbf{X}_2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1' \mathbf{Y}_1 \\ \mathbf{X}_2' \mathbf{Y}_2 \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{Y}_1 \\ (\mathbf{X}_2' \mathbf{X}_2)^{-1} \mathbf{X}_2' \mathbf{Y}_2 \end{pmatrix}, \end{aligned}$$

相应的残差平方和为

$$\text{RSS} = \mathbf{Y}_1' \mathbf{Y}_1 + \mathbf{Y}_2' \mathbf{Y}_2 - \hat{\boldsymbol{\beta}}_1' \mathbf{X}_1' \mathbf{Y}_1 - \hat{\boldsymbol{\beta}}_2' \mathbf{X}_2' \mathbf{Y}_2. \quad (4.1.14)$$

将约束条件 $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 =: \boldsymbol{\alpha}$ 融入原模型, 得约简模型

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \boldsymbol{\alpha} + \mathbf{e}.$$

$\boldsymbol{\alpha}$ 的最小二乘估计为

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}_1' \mathbf{X}_1 + \mathbf{X}_2' \mathbf{X}_2)^{-1} (\mathbf{X}_1' \mathbf{Y}_1 + \mathbf{X}_2' \mathbf{Y}_2),$$

相应的残差平方和为

$$\text{RSS}_H = \mathbf{Y}_1' \mathbf{Y}_1 + \mathbf{Y}_2' \mathbf{Y}_2 - \hat{\boldsymbol{\alpha}}' (\mathbf{X}_1' \mathbf{Y}_1 + \mathbf{X}_2' \mathbf{Y}_2). \quad (4.1.15)$$

结合(4.1.14)和(4.1.15)得到

$$\begin{aligned} \text{RSS}_H - \text{RSS} &= \hat{\beta}'_1 \mathbf{X}'_1 \mathbf{Y}_1 + \hat{\beta}'_2 \mathbf{X}'_2 \mathbf{Y}_2 - \alpha'(\mathbf{X}'_1 \mathbf{Y}_1 + \mathbf{X}'_2 \mathbf{Y}_2) \\ &= (\hat{\beta}_1 - \alpha)' \mathbf{X}'_1 \mathbf{Y}_1 + (\hat{\beta}_2 - \alpha)' \mathbf{X}'_2 \mathbf{Y}_2. \end{aligned}$$

因此检验统计量为

$$F_H = \frac{[(\hat{\beta}_1 - \alpha)' \mathbf{X}'_1 \mathbf{Y}_1 + (\hat{\beta}_2 - \alpha)' \mathbf{X}'_2 \mathbf{Y}_2] / (p+1)}{[\mathbf{Y}'_1 \mathbf{Y}_1 + \mathbf{Y}'_2 \mathbf{Y}_2 - \hat{\beta}'_1 \mathbf{X}'_1 \mathbf{Y}_1 - \hat{\beta}'_2 \mathbf{X}'_2 \mathbf{Y}_2] / (n_1 + n_2 - 2p - 2)}.$$

若 H 成立, 则 $F_H \sim F(p+1, n_1 + n_2 - 2p - 2)$. 对给定的显著性水平 α , 若 $f_H > F_\alpha(p+1, n_1 + n_2 - 2p - 2)$ (f_H 为 F_H 的样本观测值), 则拒绝原假设, 即认为两批数据不是来自同一线性回归模型. 否则, 无充分的理由拒绝原假设, 即认为它们来自同一线性回归模型.

4.2 回归方程的显著性检验

将正态线性回归模型(4.1.1)写成

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i, \\ e_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \cdots, n. \end{cases} \quad (4.2.1)$$

所谓回归方程的显著性检验, 就是检验自变量这个整体 $\{x_1, \cdots, x_p\}$ 是否对因变量有显著的线性相依关系. 即检验

$$H: \beta_1 = \cdots = \beta_p = 0. \quad (4.2.2)$$

若拒绝原假设, 则认为至少有某一个 x_i 对因变量 y 有显著的线性相依关系. 若接受原假设, 则认为相对于模型误差而言, 自变量整体对因变量 y 的线性影响是可以忽略不计的. 把假设(4.2.2)写成 $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ 的形式, 其中

$$\mathbf{A} = (\mathbf{0} \quad \mathbf{I}_p), \quad \mathbf{b} = \mathbf{0},$$

这里的 $\mathbf{0}$ 是 p 维的零向量, $\text{rk}(\mathbf{A}) = p$. 定理4.1.1(d)给出的检验统计量可直接应用在回归方程的显著性检验. 下面针对这种特殊情形导出检验统计量的简单形式并解释其统计意义.

将假设(4.2.2)融入模型(4.2.1), 得约简模型

$$\begin{cases} y_i = \beta_0 + e_i, \\ e_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \cdots, n. \end{cases} \quad (4.2.3)$$

易求 β_0 的LSE为 $\beta_0^* = \bar{y}$, 相应的残差平方和为

$$\text{RSS}_H = \mathbf{Y}'\mathbf{Y} - \beta_0^* \mathbf{1}'_n \mathbf{Y} = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \sum_{i=1}^n (y_i - \bar{y})^2. \quad (4.2.4)$$

这就是回归分析中的总平方和TSS. 约简模型中不包含任何回归自变量, 残差平方和 RSS_H 完全是由 y_1, \dots, y_n 的变动引起的. 对于原模型(4.2.1), 残差平方和

$$RSS = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}.$$

于是

$$ESS = TSS - RSS = RSS_H - RSS = \hat{\beta}'\mathbf{X}'\mathbf{Y} - \beta_0^*\mathbf{1}'_n\mathbf{Y}, \quad (4.2.5)$$

它是由于在模型(4.2.3)中引入回归自变量 $\{x_1, \dots, x_p\}$ 后所引起的残差平方和的减少量. 所以根据定理4.1.1, 检验统计量为

$$F_H = \frac{ESS/p}{RSS/(n-p-1)}. \quad (4.2.6)$$

当原假设(4.2.2)成立时, $F_H \sim F(p, n-p-1)$. 记 f_H 为 F_H 的样本观测值. 给定显著性水平 α , 若 $f_H > F_\alpha(p, n-p-1)$, 则拒绝原假设, 否则接受原假设.

统计量 F_H 有清晰的统计解释. 由(4.2.4)知 RSS_H 就是通常的总平方和TSS, (4.2.5)可写为

$$TSS = RSS + ESS.$$

其中, 回归平方和ESS反映了回归自变量对因变量变动平方和的贡献, RSS反映了模型误差对因变量变动平方和的贡献. 因此, 检验统计量(4.2.6)是把自变量的平均贡献和模型误差的平均贡献进行比较. 当自变量的平均贡献显著大于模型误差的平均贡献时, 就有充分的理由相信回归自变量对因变量有显著的线性相依作用, 从而拒绝原假设. 当自变量的平均贡献没有显著大于模型误差的平均贡献时, 就没有充分的理由认为回归自变量比模型误差对因变量有更显著的线性相依作用, 因此接受原假设.

把上述假设检验的过程制成一张表, 称为方差分析表, 见下面的表4.2.1.

方差来源	平方和	自由度	均方	F值	$P(F > f_H)$
回归	ESS	p	ESS/p	F_H	p
误差	RSS	$n-p-1$	$RSS/(n-p-1)$		
总计	TSS	$n-1$			

表4.2.1 方差分析表

该表中的 $P(F > f_H)$ 表示 $P(F(p, n-p-1) > f_H)$. 当

$$p = P(F(p, n-p-1) > f_H) < \alpha \quad (4.2.7)$$

时拒绝原假设, 其中 α 为事先给定的显著性水平. 注意(4.2.7)等价于 $f_H > F_\alpha(p, n-p-1)$, 即样本落入拒绝域.

如果经过检验, 接受原假设 $H: \beta_1 = \dots = \beta_p = 0$, 这意味着和模型误差比起来, 诸自变量对因变量 y 的线性影响是不显著的. 这里可能有两种情况:

(1) 模型的误差太大, 虽然回归自变量对 y 有一定的影响, 但相对于较大的模型误差, 也不算大. 对这种情况, 要想办法缩小模型误差的影响(检查是否漏掉了重要的自变量, y 对模型中的自变量是否有非线性的相依关系, 等等);

(2) 回归自变量对 y 的影响确实很小, 这时应放弃 y 对诸回归自变量作线性回归.

例4.2.1 煤净化问题(来自Myers(2000)). 表4.2.2给出了煤净化的一组数据. y 表示净化后煤溶液中所含杂质的重量, x_1 表示输入净化过程的溶液所含的煤与杂质的比值, x_2 是溶液的pH值, x_3 表示溶液流量. 实验目的是通过这组实验数据, 建立净化效率 y 与三个因素 x_1, x_2, x_3 的经验关系.

编号	x_1	x_2	x_3	y
1	1.5	6	1315	243
2	1.5	6	1315	261
3	1.5	9	1890	244
4	1.5	9	1890	285
5	2	7.5	1575	202
6	2	7.5	1575	180
7	2	7.5	1575	183
8	2	7.5	1575	207
9	2.5	9	1315	216
10	2.5	9	1315	160
11	2.5	6	1890	104
12	2.5	6	1890	110

表4.2.2

利用R软件进行分析, 代码及分析结果如下:

```

1 > yx=read.table("***.txt")
2 > x1=yx[,1]
3 > x2=yx[,2]
4 > x3=yx[,3]
5 > y=yx[,4]
6 > coal=data.frame(x1,x2,x3,y)
7 > lm.sol=lm(y~x1+x2+x3,data=coal)
8 > summary(lm.sol)
9
10 Call:
11 lm(formula = y ~ x1 + x2 + x3, data = coal)
12
13 Residuals:
14     Min       1Q   Median       3Q      Max
15 -23.808 -17.193   0.904   8.143  32.192
16
17 Coefficients:
18             Estimate Std. Error t value Pr(>|t|)
19 (Intercept)  397.08738    62.75676   6.327 0.000226 ***
20 x1          -110.75000    14.76248  -7.502 6.91e-05 ***
21 x2           15.58333     4.92083   3.167 0.013258 *
22 x3           -0.05829     0.02563  -2.274 0.052565 .

```

```

23 ---
24 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
25
26 Residual standard error: 20.88 on 8 degrees of freedom
27 Multiple R-squared: 0.8993, Adjusted R-squared: 0.8616
28 F-statistic: 23.83 on 3 and 8 DF, p-value: 0.0002422

```

取显著性水平 $\alpha = 0.05$. 因为回归方程的显著性检验的 p 值为 $0.0002422 < \alpha$, 所以有充分的理由拒绝原假设, 认为回归自变量整体 $\{x_1, x_2, x_3\}$ 对因变量有显著的线性相依关系. 回归方程为

$$\hat{y} = 397.087 - 110.750x_1 + 15.583x_2 - 0.058x_3.$$

回归方程的 F 检验也可用`anova`函数来实现, 它同时给出了(约简模型与原模型比较的)方差分析表, R代码及分析结果如下:

```

1 > lm.reduction=lm(y~1,data=coal)
2 > anova(lm.reduction,lm.sol)
3 Analysis of Variance Table
4 Model 1: y ~ 1
5 Model 2: y ~ x1 + x2 + x3
6   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
7   1      11 34643
8   2       8 3487  3     31156 23.827 0.0002422 ***

```

注 Model 1 是约简模型(作为原假设), Model 2 是原模型(作为备择假设). 所以这里的 F 检验等价于回归方程的显著性检验.

以上的假设检验过程依赖于模型的正态性假设. 若无正态性假设, 则需要在大样本理论的框架下进行假设检验, 即要求出 F 统计量的极限分布, 然后用极限分布来构造拒绝域. 大样本理论在实际应用时需要样本量足够大, 但有时这个要求无法被满足. 置换检验提供了另一种解决方案.

若因变量与自变量整体无显著的相依关系, 那么可认为因变量的观测值是随机散布的. 我们知道(回归方程的显著性检验的) F 统计量可用来度量因变量与自变量整体的线性相依关系, F 值越大, 线性相依关系越显著. 那么有以下的问题: 比目前观测到的 F 统计量的样本值还要大的可能性有多大? 若这个可能性很小, 那么就有充分的理由拒绝“因变量与自变量整体无显著的线性相依关系”这一假设.

如何实现这一过程呢? 对于原样本我们计算出一个 F 值, 然后对因变量的 n 个观测值的 $n!$ 种全排列分别计算 $n!$ 个 F 值, 看一下这 $n!$ 个 F 值中有多少比例是大于由原样本计算出来的 F 值的. 最后基于这个比例大小进行统计决策(例如, 若这个比例小于5%, 则拒绝原假设: 因变量与自变量整体无显著的线性相依关系). 这就是置换检验.

接下来对煤净化例子做置换检验. 因为 $n! = 12! = 479001600$, 实在太大了! 所以只对其中的4000个全排列计算 F 值. R代码及分析结果如下:

```

1 > lms=summary(lm.sol)
2 > lms$fstatistic

```

```

3      value      numdf      dendf
4 23.82716  3.00000  8.00000
5 > nreps=4000
6 > set.seed(123)
7 > fstats=numeric(nreps)
8 > for (i in 1:nreps){
9     lm.sol=lm(sample(y)~x1+x2+x3,data=coal)
10    fstats[i]=summary(lm.sol)$fstatistic[1]
11  }
12 > mean(fstats>lms$fstatistic[1])
13 [1] 5e-04

```

因为这个比例值 5×10^{-4} 非常小, 所以有充分的理由拒绝原假设: $\beta_1 = \beta_2 = \beta_3 = 0$.

4.3 回归系数的显著性检验

回归方程的显著性检验是对回归自变量的一个整体性检验. 如果检验结果是拒绝原假设, 这意味着接受“因变量 y 与 $\{x_1, \dots, x_p\}$ 这个整体有线性相依关系”这个假设. 但是, 这不排除 y 与某些自变量无线性相依关系, 即某些 $\beta_i = 0$. 于是在回归方程的显著性检验被拒绝后, 还需对每个自变量逐一做显著性检验, 即对固定的 i , $1 \leq i \leq p$, 做如下检验

$$H_i: \beta_i = 0. \quad (4.3.1)$$

(4.3.1) 可等价地写成关于 β 的线性假设 $H: \mathbf{A}\beta = \mathbf{b}$, 其中

$$\begin{aligned} \beta &= (\beta_0, \beta_1, \dots, \beta_{i-1}, \beta_i, \beta_{i+1}, \dots, \beta_p)', \\ \mathbf{A} &= (0, 0, \dots, 0, 1, 0, \dots, 0), \quad \mathbf{b} = 0, \end{aligned}$$

\mathbf{A} 中第 $i+1$ 个元素为 1, 其它都为零. 注意 $m = \text{rk}(\mathbf{A}) = 1$.

考虑用最小二乘法基本定理推导(4.3.1)的检验统计量. 由公式(4.1.8)得

$$\text{RSS}_H - \text{RSS} = (\mathbf{A}\hat{\beta} - \mathbf{b})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta} - \mathbf{b}) = \frac{\hat{\beta}_i^2}{c_{i+1,i+1}},$$

其中 $c_{i+1,i+1}$ 表示 $(\mathbf{X}'\mathbf{X})^{-1}$ 的第 $i+1$ 个对角线元素. 记

$$\hat{\sigma}^2 = \text{RSS}/(n-p-1),$$

则检验统计量为

$$F_H = \frac{(\text{RSS}_H - \text{RSS})/m}{\text{RSS}/(n-p-1)} = \frac{\hat{\beta}_i^2}{\hat{\sigma}^2 c_{i+1,i+1}} \stackrel{H_i}{\sim} F(1, n-p-1).$$

给定显著性水平 α , 当 $f_H > F_\alpha(1, n-p-1)$ 时拒绝原假设 H_i , 否则接受 H_i .

对于假设检验问题(4.3.1), 除了 F 检验方法, 还有 t 检验方法. 根据 F 分布与 t 分布的关系, 可得 t 检验统计量为

$$T_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{c_{i+1,i+1}}}.$$

这个检验统计量也可通过定理3.2.4推得. 根据此定理, $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, 所以 $\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{i+1,i+1})$. 若(4.3.1)成立, 则

$$\frac{\hat{\beta}_i}{\sigma\sqrt{c_{i+1,i+1}}} \sim N(0, 1).$$

又因为

$$\text{RSS}/\sigma^2 \sim \chi^2(n-p-1) \text{ 且与 } \hat{\beta}_i \text{ 独立,}$$

所以

$$T_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{c_{i+1,i+1}}} \stackrel{H_i}{\sim} t(n-p-1). \quad (4.3.2)$$

给定显著性水平 α , 当 $|t_i| > t_{\alpha/2}(n-p-1)$ 时拒绝原假设 H_i , 否则接受 H_i . 因为 $\hat{\sigma}\sqrt{c_{i+1,i+1}}$ 为 $\hat{\beta}_i$ 的标准差 $\sigma\sqrt{c_{i+1,i+1}}$ 的一个估计, 所以称 $\hat{\sigma}\sqrt{c_{i+1,i+1}}$ 为 $\hat{\beta}_i$ 的标准误(standard error), 记为 $\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma}\sqrt{c_{i+1,i+1}}$.

实际中, 往往通过 p 值进行统计决策. 当

$$p_i = 2P(t(n-p-1) > |t_i|) < \alpha$$

时, 拒绝原假设 $H_i: \beta_i = 0$, 否则接受原假设 $H_i: \beta_i = 0$.

对于例4.2.1, 根据回归分析结果, $p_1 = 6.91 \times 10^{-5} < 0.05$, $p_2 = 0.013258 < 0.05$, $p_3 = 0.052565 \geq 0.05$, 所以拒绝 $H_1: \beta_1 = 0$ 和 $H_2: \beta_2 = 0$, 接受 $H_3: \beta_3 = 0$. 然后修改回归模型(剔除 x_3), 重新进行回归分析.

若想得到回归系数的显著性检验的方差分析表, 可通过anova函数实现. 以 β_1 的检验为例, R代码及分析结果如下:

```
1 > lm.sol=lm(y~x1+x2+x3,data=coal)
2 > lm.reduction=lm(y~x2+x3,data=coal)
3 > anova(lm.reduction,lm.sol)
4 Analysis of Variance Table
5 Model 1: y ~ x2 + x3
6 Model 2: y ~ x1 + x2 + x3
7   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
8 1       9 28018.0
9 2       8  3486.9  1    24531 56.282 6.914e-05 ***
```

在没有模型的正态性假设的情形下, 类似于回归方程的置换检验, 可以做回归系数的置换检验. 例如, 想检验 x_2 是否对因变量 y 有显著的线性相依关系, 其R代码及分析结果如下:

```

1 > lms$coef[3,]      #lms=summary(lm.sol)
2   Estimate Std. Error    t value    Pr(>|t|)
3 15.58333333  4.92082621  3.16681238  0.01325816
4 > nreps=4000
5 > tstats=numeric(nreps)
6 > set.seed(123)
7 > for (i in 1:nreps){
8   lm.sol=lm(y~x1+sample(x2)+x3,data=coal)
9   tstats[i]=summary(lm.sol)$coef[3,3]
10 }
11 > mean(abs(tstats)>abs(lms$coef[3,3]))
12 [1] 0.00825

```

因为这个比例值0.00825非常小, 所以有充分的理由拒绝原假设: $\beta_2 = 0$.

4.4 其它线性假设的检验

最小二乘法基本定理的应用非常广泛. 前面介绍了它在回归方程的显著性检验以及回归系数的显著性检验中的应用, 接下来介绍它在其它假设检验问题中的应用.

考虑检验成对自变量. 例如, 对于煤净化数据(见例4.2.1), 考虑检验 x_2 和 x_3 是否至少有一个对因变量 y 有显著的线性相依关系, 这等价于检验

$$H: \beta_2 = \beta_3 = 0.$$

记原模型的残差平方和为RSS, 约简模型

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i, \quad i = 1, \dots, n$$

的残差平方和为 RSS_H . 给定显著性水平 α , 由最小二乘法基本定理, 假设检验的拒绝域为

$$\left\{ \text{样本} : F = \frac{(RSS_H - RSS)/2}{RSS/(n-4)} > F_\alpha(2, n-4) \right\}.$$

下面用anova函数来实现这一分析过程, R代码及分析结果如下:

```

1 > lm.sol=lm(y~x1+x2+x3,data=coal)
2 > lm.reduction=lm(y~x1,data=coal)
3 > anova(lm.reduction,lm.sol)
4 Analysis of Variance Table
5 Model 1: y ~ x1
6 Model 2: y ~ x1 + x2 + x3
7   Res.Df    RSS Df Sum of Sq    F Pr(>F)
8 1      10 10111.8
9 2       8  3486.9  2    6624.9 7.5998 0.01414 *

```

由于 p 值= 0.01414 < 0.05, 所以有充分的理由拒绝原假设 $\beta_2 = \beta_3 = 0$, 即认为 x_2 和 x_3 至少有一个对因变量 y 有显著的线性相依关系.

接下来考虑检验回归参数的子空间. 例如, 检验 x_2 和 x_3 是否对因变量 y 具有相同程度的线性相依关系, 这等价于检验

$$H: \beta_2 = \beta_3.$$

记原模型的残差平方和为RSS, 约简模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2(x_{i2} + x_{i3}) + e_i, \quad i = 1, \dots, n$$

的残差平方和为 RSS_H . 给定显著性水平 α , 由最小二乘法基本定理, 假设检验的拒绝域为

$$\left\{ \text{样本} : F = \frac{RSS_H - RSS}{RSS/(n-4)} > F_\alpha(1, n-4) \right\}.$$

用anova函数来实现这一分析过程, R代码和分析结果如下:

```
1 > lm.sol=lm(y~x1+x2+x3,data=coal)
2 > lm.reduction=lm(y~x1+I(x2+x3),data=coal)
3 > anova(lm.reduction,lm.sol)
4 Analysis of Variance Table
5 Model 1: y ~ x1 + I(x2 + x3)
6 Model 2: y ~ x1 + x2 + x3
7   Res.Df    RSS Df Sum of Sq    F Pr(>F)
8 1       9 7890.7
9 2       8 3486.9  1   4403.8 10.104 0.01302 *
```

由于 p 值= 0.01302 < 0.05, 所以有充分的理由拒绝原假设 $\beta_2 = \beta_3$, 即认为 x_2 和 x_3 对因变量 y 具有不同程度的线性相依关系.

最后, 考虑检验某一参数值是否为一特殊值. 例如, 进行以下检验

$$H: \beta_3 = 1.$$

此时的约简模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + 1 \times x_{i3} + e_i, \quad i = 1, \dots, n,$$

残差平方和记为 RSS_H . 给定显著性水平 α , 由最小二乘法基本定理, 检验的拒绝域为

$$\left\{ \text{样本} : F = \frac{RSS_H - RSS}{RSS/(n-4)} > F_\alpha(1, n-4) \right\}.$$

用anova函数来实现这一分析过程, R代码和分析结果如下:

```
1 > lm.sol=lm(y~x1+x2+x3,data=coal)
2 > lm.reduction=lm(y~x1+x2+offset(1*x3),data=coal)
3 > anova(lm.reduction,lm.sol)
4 Analysis of Variance Table
5 Model 1: y ~ x1 + x2 + offset(1 * x3)
```

```

6 | Model 2: y ~ x1 + x2 + x3
7 | Res.Df    RSS Df Sum of Sq      F    Pr(>F)
8 | 1         9 746334
9 | 2         8 3487  1   742847 1704.3 1.305e-10 ***

```

由于 p 值 $= 1.305 \times 10^{-10} < 0.05$, 所以有充分的理由拒绝原假设 $\beta_3 = 1$.

4.5 异常点检验

在统计学中, 异常点(outlier)是泛指在一组数据中, 与它们的主体不是来自同一分布的那些少数点. 几何直观上, 异常点的“异常”之处就是它们远离数据的主体. 在上一章已用学生化残差来判断异常点, 现在通过假设检验的方法来作异常点的统计推断.

把正态线性回归模型(4.2.1)写成

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, \quad e_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n, \quad (4.5.1)$$

这里 \mathbf{x}_i' 表示设计矩阵 \mathbf{X} 的第 i 行. 如果第 j 组数据 (\mathbf{x}_j', y_j) 是一个异常点, 那么可假设 $\mathbf{E}(y_j)$ 发生了漂移 η , 因此有了一个新的模型

$$\begin{cases} y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, & i \neq j, \\ y_j = \mathbf{x}_j' \boldsymbol{\beta} + \eta + e_j, \\ e_1, \dots, e_n \stackrel{i.i.d.}{\sim} N(0, \sigma^2). \end{cases} \quad (4.5.2)$$

记 $\mathbf{d}_j = (0, \dots, 0, 1, 0, \dots, 0)'$ 是一个 n 维的列向量, 它的第 j 个元素为1, 其余为0. 将模型(4.5.2)写成矩阵形式

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{d}_j\eta + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n). \quad (4.5.3)$$

称模型(4.5.2)和(4.5.3)为均值漂移线性回归模型. 现在的目的是要判别 (\mathbf{x}_j', y_j) 是不是异常点, 这等价于检验假设 $H: \eta = 0$.

记 $\boldsymbol{\beta}^*$ 和 η^* 为模型(4.5.3)中 $\boldsymbol{\beta}$ 和 η 的最小二乘估计, 为了推导出具体表达式, 先介绍一个分块矩阵求逆公式.

引理4.5.1(分块矩阵求逆公式) 设 \mathbf{A} 为非奇异的对称矩阵, 将其分块为

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}' & \mathbf{D} \end{pmatrix},$$

则当 $\mathbf{B}^{-1}, \mathbf{D}^{-1}$ 都存在时有

$$\begin{aligned} \mathbf{A}^{-1} &= \begin{pmatrix} \mathbf{B}_1 & \mathbf{C}_1 \\ \mathbf{C}'_1 & \mathbf{D}_1 \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{B} - \mathbf{C}\mathbf{D}^{-1}\mathbf{C}')^{-1} & -\mathbf{B}_1\mathbf{C}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}'\mathbf{B}_1 & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}'\mathbf{B}_1\mathbf{C}\mathbf{D}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{C}\mathbf{D}_1\mathbf{C}'\mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{C}\mathbf{D}_1 \\ -\mathbf{D}_1\mathbf{C}'\mathbf{B}^{-1} & (\mathbf{D} - \mathbf{C}'\mathbf{B}^{-1}\mathbf{C})^{-1} \end{pmatrix}. \end{aligned}$$

定理4.5.1 对均值漂移线性回归模型(4.5.3), β 和 η 的最小二乘估计为

$$\beta^* = \hat{\beta}_{(j)}, \quad \eta^* = \frac{\hat{e}_j}{1 - h_{jj}},$$

其中 $\hat{\beta}_{(j)}$ 为从非均值漂移线性回归模型(4.5.1)中剔除第 j 组数据后得到的 β 的最小二乘估计, h_{jj} 为帽子矩阵 $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 的第 j 个对角线元素, \hat{e}_j 为从模型(4.5.1)导出的第 j 个残差.

注 这个定理给出了一个重要的事实: 如果因变量的第 j 个观测值发生了均值漂移, 那么在相应的均值漂移线性回归模型中, 回归系数 β 的最小二乘估计恰好等于在模型(4.5.1)中剔除第 j 组数据后得到的最小二乘估计.

证明 注意到 $\mathbf{d}_j'\mathbf{Y} = y_j$, $\mathbf{d}_j'\mathbf{d}_j = 1$. 记 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$, 则 $\mathbf{X}'\mathbf{d}_j = \mathbf{x}_j$. 首先, 易知

$$\begin{pmatrix} \beta^* \\ \eta^* \end{pmatrix} = \begin{pmatrix} (\mathbf{X}') \\ \mathbf{d}_j' \end{pmatrix} (\mathbf{X} \quad \mathbf{d}_j)^{-1} \begin{pmatrix} \mathbf{X}' \\ \mathbf{d}_j' \end{pmatrix} \mathbf{Y} = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{x}_j \\ \mathbf{x}_j' & 1 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{d}_j'\mathbf{Y} \end{pmatrix}.$$

应用分块矩阵求逆公式并注意 $h_{jj} = \mathbf{x}_j'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j$, 有

$$\begin{aligned} \begin{pmatrix} \beta^* \\ \eta^* \end{pmatrix} &= \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} + \frac{1}{1-h_{jj}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j'(\mathbf{X}'\mathbf{X})^{-1} & -\frac{1}{1-h_{jj}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j \\ -\frac{1}{1-h_{jj}}\mathbf{x}_j'(\mathbf{X}'\mathbf{X})^{-1} & \frac{1}{1-h_{jj}} \end{pmatrix} \begin{pmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{d}_j'\mathbf{Y} \end{pmatrix} \\ &= \begin{pmatrix} \hat{\beta} + \frac{1}{1-h_{jj}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j'\hat{\beta} - \frac{1}{1-h_{jj}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j y_j \\ -\frac{1}{1-h_{jj}}\mathbf{x}_j'\hat{\beta} + \frac{1}{1-h_{jj}}y_j \end{pmatrix} \\ &= \begin{pmatrix} \hat{\beta} - \frac{1}{1-h_{jj}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j\hat{e}_j \\ \frac{\hat{e}_j}{1-h_{jj}} \end{pmatrix}. \end{aligned}$$

再回忆上一章的公式(3.4.9), 可知定理成立.

接下来应用最小二乘法基本定理来推导出 $H: \eta = 0$ 的检验统计量. 首先把 $H: \eta = 0$ 写成线性约束

$$\mathbf{A} \begin{pmatrix} \beta \\ \eta \end{pmatrix} = \mathbf{b},$$

其中

$$\mathbf{b} = 0, \quad \mathbf{A} = (0, \dots, 0, 1), \quad m = \text{rk}(\mathbf{A}) = 1.$$

另外, 注意到在此约束条件下, 模型(4.5.3)退化到约简模型(4.5.1), 所以

$$\text{RSS}_H = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}.$$

而模型(4.5.3)的(无约束)残差平方和为

$$\text{RSS} = \mathbf{Y}'\mathbf{Y} - \beta^{*'}\mathbf{X}'\mathbf{Y} - \eta^*\mathbf{d}_j'\mathbf{Y}. \quad (4.5.4)$$

利用定理4.5.1得

$$\text{RSS}_H - \text{RSS} = (\beta^* - \hat{\beta})'\mathbf{X}'\mathbf{Y} + \eta^*\mathbf{d}_j'\mathbf{Y}$$

$$\begin{aligned}
&= -\frac{\hat{e}_j \mathbf{x}'_j}{1-h_{jj}} \hat{\boldsymbol{\beta}} + \frac{\hat{e}_j y_j}{1-h_{jj}} \\
&= \frac{\hat{e}_j^2}{1-h_{jj}}.
\end{aligned}$$

因此, RSS可进一步写成

$$\text{RSS} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \frac{\hat{e}_j^2}{1-h_{jj}} = (n-p-1)\hat{\sigma}^2 - \frac{\hat{e}_j^2}{1-h_{jj}},$$

其中 $\hat{\sigma}^2 = \text{RSS}/(n-p-1)$. 由最小二乘法基本定理, 检验统计量为

$$\begin{aligned}
F_H &= \frac{(\text{RSS}_H - \text{RSS})/1}{\text{RSS}/(n-p-2)} \\
&= \frac{\frac{\hat{e}_j^2}{1-h_{jj}}}{\frac{(n-p-1)\hat{\sigma}^2}{n-p-2} - \frac{\hat{e}_j^2}{(n-p-2)(1-h_{jj})}} \\
&= \frac{(n-p-2)r_j^2}{n-p-1-r_j^2} \stackrel{H}{\sim} F(1, n-p-2),
\end{aligned}$$

这里 $r_j = \frac{\hat{e}_j}{\hat{\sigma}\sqrt{1-h_{jj}}}$ 为学生化残差.

定理4.5.2 对于均值漂移线性回归模型(4.5.3), 若假设 $H: \eta = 0$ 成立, 则

$$F_j = \frac{(n-p-2)r_j^2}{n-p-1-r_j^2} \sim F(1, n-p-2).$$

根据此定理, 给定显著性水平 α , 若

$$f_j > F_\alpha(1, n-p-2), \quad (4.5.5)$$

则判定第 j 组数据 (\mathbf{x}'_j, y_j) 为异常点, 否则认为是正常数据点.

根据 F 分布与 t 分布的关系, 也可以用 t 检验法完成刚才的假设检验. 定义

$$T_j = \sqrt{\frac{n-p-2}{n-p-1-r_j^2}} r_j,$$

则对给定的显著性水平 α , 当

$$|t_j| > t_{\alpha/2}(n-p-2)$$

时拒绝原假设 $H: \eta = 0$, 即认为第 j 组数据 (\mathbf{x}'_j, y_j) 为异常点, 否则接受原假设.

例4.5.1(续例4.2.1) 检验12组数据中是否有异常点. R代码及分析结果如下:

```

1 > lm.sol=lm(y~x1+x2+x3,data=coal)
2 > library(car)
3 > outlierTest(lm.sol)

```

```

4 No Studentized residuals with Bonferroni p < 0.05
5 Largest |rstudent|:
6 rstudent unadjusted p-value Bonferroni p
7 9 2.869512          0.024009          0.28811

```

因为第9组数据的 p 值= 0.024009 < 0.05, 所有有充分的理由认为第9组数据为异常点.

如果样本量较大, 譬如 $n = 100$, 那么即使数据都是正常的, 也将会有5个左右的样本被判为异常点(假设每个检验的显著性水平都取为 $\alpha = 0.05$). 如果样本中含有异常点, 则被判为异常点的样本个数将会更多, 所以需要调整单个检验的显著性水平. 若每一个 $P(A_i) = 1 - \alpha$, 则

$$P\left(\bigcap_{i=1}^n A_i\right) = 1 - P\left(\bigcup_{i=1}^n \bar{A}_i\right) \geq 1 - \sum_{i=1}^n P(\bar{A}_i) = 1 - n\alpha.$$

这个公式就是著名的Bonferroni不等式. 所以, 只需把单个检验的显著性水平取为 α/n , 这样就可保证整体意义下的检验显著性水平为 α . 该方法被称为Bonferroni修正.

4.6 Durbin-Watson检验

Durbin-Watson检验是用来诊断线性模型的随机误差序列的不相关性假设的. 考虑相邻观测间存在的一种最简单的相关情况: 一阶自相关. 设 e_{i+1} 与 e_i 间有如下的关系,

$$e_{i+1} = \rho e_i + u_{i+1}, \quad i = 1, \dots, n-1,$$

且假设 $\{u_i, i \geq 1\}$ 为独立同分布随机变量序列, 服从 $N(0, \sigma^2)$. 这时检验 $\{e_i, i = 1, \dots, n\}$ 的不相关性问题就变成了检验

$$H_0: \rho = 0.$$

Durbin与Watson提出了一种D-W检验, 检验统计量为

$$DW = \frac{\sum_{i=2}^n (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^n \hat{e}_i^2}.$$

该统计量的渐近分布是 χ^2 分布的线性组合, 比较复杂. 为了给出假设检验的拒绝域, 先来了解一下DW统计量的意义.

由于 $\{e_i, i = 1, \dots, n\}$ 是不可观测的, 因此要考查 $\{e_i, i = 1, \dots, n\}$ 的相关性常用残差序列 $\{\hat{e}_i, i = 1, \dots, n\}$ 来考察. 将 $\{\hat{e}_1, \dots, \hat{e}_{n-1}\}$ 与 $\{\hat{e}_2, \dots, \hat{e}_n\}$ 看成两个序列, 称其样本相关系数 r 为一阶自相关系数,

$$r = \frac{\sum_{i=1}^{n-1} (\hat{e}_i - \bar{\hat{e}}_{1,n-1})(\hat{e}_{i+1} - \bar{\hat{e}}_{2,n})}{\sqrt{\sum_{i=1}^{n-1} (\hat{e}_i - \bar{\hat{e}}_{1,n-1})^2 \sum_{i=1}^{n-1} (\hat{e}_{i+1} - \bar{\hat{e}}_{2,n})^2}},$$

其中

$$\bar{\hat{e}}_{1,n-1} = \frac{1}{n-1} \sum_{i=1}^{n-1} \hat{e}_i, \quad \bar{\hat{e}}_{2,n} = \frac{1}{n-1} \sum_{i=2}^n \hat{e}_i.$$

由于 $|\hat{e}_i|$ 一般较小, 故可认为

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^{n-1} \hat{e}_i &\approx \frac{1}{n-1} \sum_{i=2}^n \hat{e}_i \approx \frac{1}{n} \sum_{i=1}^n \hat{e}_i = 0, \\ \sum_{i=1}^{n-1} \hat{e}_i^2 &\approx \sum_{i=2}^n \hat{e}_i^2 \approx \sum_{i=1}^n \hat{e}_i^2. \end{aligned}$$

将它们代入 r 的表达式, 得

$$r \approx \frac{\sum_{i=1}^{n-1} \hat{e}_i \hat{e}_{i+1}}{\sqrt{\sum_{i=1}^{n-1} \hat{e}_i^2 \sum_{i=1}^{n-1} \hat{e}_{i+1}^2}} \approx \frac{\sum_{i=1}^{n-1} \hat{e}_i \hat{e}_{i+1}}{\sum_{i=1}^n \hat{e}_i^2}.$$

若 $\{e_1, \dots, e_n\}$ 可观测, 则 ρ 的最小二乘估计为

$$\hat{\rho} = \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{\sum_{i=1}^{n-1} e_i^2} \approx \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{\sum_{i=1}^n e_i^2}.$$

因此, r 可看作 ρ 的一个点估计. 此外, 容易看出, DW统计量与 r 之间有如下的近似关系:

$$\begin{aligned} DW &= \frac{\sum_{i=2}^n \hat{e}_i^2 + \sum_{i=1}^{n-1} \hat{e}_i^2 - 2 \sum_{i=2}^n \hat{e}_{i-1} \hat{e}_i}{\sum_{i=1}^n \hat{e}_i^2} \\ &\approx \frac{2 \sum_{i=1}^n \hat{e}_i^2 - 2 \sum_{i=1}^{n-1} \hat{e}_i \hat{e}_{i+1}}{\sum_{i=1}^n \hat{e}_i^2} \\ &\approx 2 - 2r. \end{aligned} \tag{4.6.1}$$

由(4.6.1)知: 当 $r = -1$ 时, $DW \approx 4$; 当 $r = 1$ 时, $DW \approx 0$; 当 $r = 0$ 时, $DW \approx 2$. 因此当 $|DW - 2|$ 过大时拒绝原假设 $H_0: \rho = 0$.

根据DW的值, 可按下面的规则做统计决策($0 < d_L < d_U < 2$): $DW < d_L$, 认为 $\{e_i\}$ 存在正相关; $d_U < DW < 4 - d_U$, 认为 $\{e_i\}$ 不相关; $DW > 4 - d_L$, 认为 $\{e_i\}$ 存在负相关; $d_L < DW < d_U$ 或者 $4 - d_U < DW < 4 - d_L$ 时, 对 $\{e_i\}$ 是否相关不下结论. 以上的 d_L 和 d_U 与假设检验的显著性水平有关, 其值可查阅《线性统计模型》(王松桂等编著)的附录3.

例4.6.1 为研究某地居民对农产品的消费量 y 与居民收入 x 之间的关系, 现收集了16组数据, 见表4.6.1.

首先可求得一元线性回归方程:

$$\hat{y} = 27.912 + 0.3524x.$$

x_i	255.7	263.3	275.4	278.3	296.7	309.3	315.8	318.8
y_i	116.5	120.8	124.4	125.5	131.7	136.2	138.7	140.2
x_i	330.0	340.2	350.7	367.3	381.3	406.5	430.8	451.5
y_i	146.8	149.6	153.0	158.2	163.2	170.5	178.2	185.9

表4.6.1

计算残差序列并得到

$$DW = 0.6800.$$

取显著性水平 $\alpha = 0.05$, 查表得 $d_L = 1.10$, $d_U = 1.37$. 现 $DW < d_L$, 因此判断模型的误差序列存在正相关. R代码及分析结果如下:

```

1 > yx=read.table("***.txt")
2 > x=yx[,1]
3 > y=yx[,2]
4 > consumption=data.frame(x,y)
5 > lm.sol=lm(y~x,data=consumption)
6 > summary(lm.sol)
7
8 Call:
9 lm(formula = y ~ x, data = consumption)
10 Coefficients:
11             Estimate Std. Error t value Pr(>|t|)
12 (Intercept) 27.912296   1.869006   14.93 5.39e-10 ***
13 x           0.352372   0.005488   64.21 < 2e-16 ***
14 ---
15 Residual standard error: 1.255 on 14 degrees of freedom
16 Multiple R-squared:  0.9966, Adjusted R-squared:  0.9964
17 F-statistic: 4123 on 1 and 14 DF, p-value: < 2.2e-16
18 > library(car)
19 > durbinWatsonTest(lm.sol)
20 lag Autocorrelation D-W Statistic p-value
21 1      0.5801892      0.6799675      0
22 Alternative hypothesis: rho != 0

```

因为残差序列的自相关系数为0.5801892, 相应的 p 值= 0 < 0.05, 所以拒绝误差序列的不相关性假设, 且认为模型的误差序列是正相关的.

4.7 回归系数的区间估计

对于给定的显著性水平 α , 如果线性假设 $H: \mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ 的 F 检验是显著的, 这说明现有的数据不支持接受假设 $H: \mathbf{A}\boldsymbol{\beta} = \mathbf{b}$. 此时自然希望构造 $\mathbf{a}_i'\boldsymbol{\beta}, i = 1, \dots, m$ 的置信区间或者置信椭圆, 这里记 $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)'$, $\text{rk}(\mathbf{A}) = m$.

考虑构造置信椭圆. 对正态线性回归模型

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i, \quad e_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n, \quad (4.7.1)$$

记 $\Phi = A\beta = (\mathbf{a}'_1\beta, \dots, \mathbf{a}'_m\beta)'$, 则

$$\hat{\Phi} = A\hat{\beta} \sim N(\Phi, \sigma^2 V),$$

这里 $V = A(X'X)^{-1}A'$, $\text{rk}(V) = m$. 根据推论2.4.1,

$$\frac{(\hat{\Phi} - \Phi)'V^{-1}(\hat{\Phi} - \Phi)}{\sigma^2} \sim \chi^2(m).$$

另一方面, 由于

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

且与 $\hat{\Phi}$ 相互独立, 这里的 $\hat{\sigma}^2 = \text{RSS}/(n-p-1)$. 于是

$$\frac{(\hat{\Phi} - \Phi)'V^{-1}(\hat{\Phi} - \Phi)}{m\hat{\sigma}^2} \sim F(m, n-p-1). \quad (4.7.2)$$

因此对给定的置信水平 $1 - \alpha$, 有

$$P\left(\frac{(\hat{\Phi} - \Phi)'V^{-1}(\hat{\Phi} - \Phi)}{m\hat{\sigma}^2} \leq F_\alpha(m, n-p-1)\right) = 1 - \alpha. \quad (4.7.3)$$

定义

$$D = \left\{ \Phi : (\Phi - \hat{\Phi})'V^{-1}(\Phi - \hat{\Phi}) \leq m\hat{\sigma}^2 F_\alpha(m, n-p-1) \right\}, \quad (4.7.4)$$

这是一个中心在 $\hat{\Phi}$ 的椭球. 由(4.7.3)知 D 包含 $\Phi = A\beta$ 的概率为 $1 - \alpha$, 所以称 D 为 Φ 的置信水平为 $1 - \alpha$ 的置信椭球. 由(4.7.4)可知 β 的置信椭球为

$$\left\{ \beta : (\beta - \hat{\beta})'A'[A(X'X)^{-1}A']^{-1}A(\beta - \hat{\beta}) \leq m\hat{\sigma}^2 F_\alpha(m, n-p-1) \right\}.$$

特别地, 当 $m = 1$ 时, 改记 $A = \mathbf{a}'$, 上式变为

$$\left\{ \beta : (\mathbf{a}'\beta - \mathbf{a}'\hat{\beta})^2 \leq \mathbf{a}'(X'X)^{-1}\mathbf{a}\hat{\sigma}^2 F_\alpha(1, n-p-1) \right\}.$$

注意到 F 分布与 t 分布的关系, 可知 $\mathbf{a}'\beta$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\mathbf{a}'\hat{\beta} \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}} \right). \quad (4.7.5)$$

记 $\hat{\sigma}_{\mathbf{a}'\beta} = \hat{\sigma}\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}}$ 为 $\mathbf{a}'\hat{\beta}$ 的标准误, 因此(4.7.5)可简记为

$$\left(\mathbf{a}'\hat{\beta} \pm t_{\alpha/2}(n-p-1)\hat{\sigma}_{\mathbf{a}'\beta} \right). \quad (4.7.6)$$

若在(4.7.6)中取 $\mathbf{a} = (0, 0, \dots, 0, 1, 0, \dots, 0)'$, 其中第 $i+1$ 个元素为1, 其它均为0, 则 $\mathbf{a}'\beta = \beta_i$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\hat{\beta}_i \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}} \right) = \left(\hat{\beta}_i \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{c_{i+1, i+1}} \right),$$

这是已知的结果(可从回归系数的显著性检验的分析过程中推得). 若对回归函数 $f(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ 感兴趣, 则根据(4.7.5)可知 $\mathbf{x}'\boldsymbol{\beta}$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}} \right). \quad (4.7.7)$$

需指出的是, 对一个固定的 \mathbf{x} , (4.7.7)给出了 $\mathbf{x}'\boldsymbol{\beta}$ 的置信水平为 $1 - \alpha$ 的置信区间. 对多个 \mathbf{x} , $\mathbf{x}'\boldsymbol{\beta}$ 同时落在各自对应区间(4.7.7)的概率将不再是 $1 - \alpha$, 而是会低于 $1 - \alpha$. 理由如下: 对每个 $\mathbf{x}'_i\boldsymbol{\beta}$ 求置信水平为 $1 - \alpha$ 的置信区间

$$I_i = \left(\mathbf{x}'_i\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n-p-1)\hat{\sigma}_{\mathbf{x}'_i\hat{\boldsymbol{\beta}}} \right), \quad i = 1, \dots, m. \quad (4.7.8)$$

虽然每个区间 I_i 包含 $\mathbf{x}'_i\boldsymbol{\beta}$ 的概率为 $1 - \alpha$, 但是 $\{\mathbf{x}'_i\boldsymbol{\beta} \in I_i, i = 1, \dots, m\}$ 同时成立的概率却不再是 $1 - \alpha$. 回忆前面曾提到的Bonferroni不等式: 设 $\{E_i, i = 1, \dots, m\}$ 为 m 个随机事件, $P(E_i) = 1 - \alpha_i, i = 1, \dots, m$. 则

$$P\left(\bigcap_{i=1}^m E_i\right) = 1 - P\left(\bigcup_{i=1}^m \bar{E}_i\right) \geq 1 - \sum_{i=1}^m P(\bar{E}_i) = 1 - \sum_{i=1}^m \alpha_i.$$

若取 $E_i = \{\mathbf{x}'_i\hat{\boldsymbol{\beta}} \in I_i\}$, 则 $\alpha_i = \alpha$, 于是

$$P\left(\mathbf{x}'_i\hat{\boldsymbol{\beta}} \in I_i, \quad i = 1, \dots, m\right) \geq 1 - m\alpha.$$

当 m 较大时, $1 - m\alpha$ 将变得很小. 在(4.7.8)中把 α 换成 α/m , 即取

$$I_i = \left(\mathbf{x}'_i\hat{\boldsymbol{\beta}} \pm t_{\alpha/(2m)}(n-p-1)\hat{\sigma}_{\mathbf{x}'_i\hat{\boldsymbol{\beta}}} \right), \quad i = 1, \dots, m, \quad (4.7.9)$$

则依据Bonferroni不等式知

$$P\left(\mathbf{x}'_i\hat{\boldsymbol{\beta}} \in I_i, \quad i = 1, \dots, m\right) \geq 1 - \alpha.$$

通常称(4.7.9)为Bonferroni区间.

例4.7.1(续例4.2.1) 对于煤净化问题, 我们得到了回归系数的点估计并已建立了回归方程

$$y = 397.087 - 110.750x_1 + 15.583x_2 - 0.058x_3.$$

现在考虑: (1) $(\beta_1, \beta_2)'$ 的置信水平为0.95和0.99的置信椭球(为了可视化, 只考虑两个参数的置信椭球); (2) 各回归系数的置信水平为0.95和0.99的置信区间.

由(4.7.4)可知, 取

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$(\beta_1, \beta_2)'$ 的置信椭球为

$$\left\{ \boldsymbol{\beta} : (\mathbf{A}\boldsymbol{\beta} - \mathbf{A}\hat{\boldsymbol{\beta}})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\boldsymbol{\beta} - \mathbf{A}\hat{\boldsymbol{\beta}}) \leq m\hat{\sigma}^2 F_{\alpha}(m, n-p-1) \right\},$$

其中, $m = 2, p = 3, n = 12, \alpha = 0.05$ 或 0.01 . 利用R画出置信椭球, 代码如下:

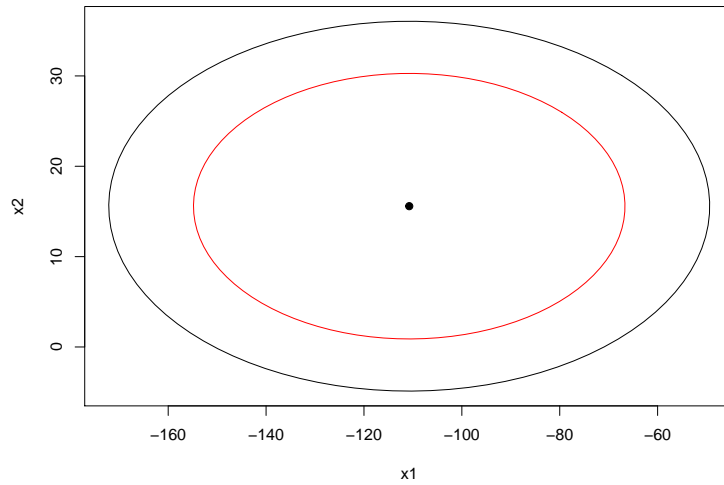


图4.7.1 置信椭球

```

1 > lm.sol=lm(y~x1+x2+x3,data=coal)
2 > lms=summary(lm.sol)
3 > library(ellipse)
4 > plot(ellipse(lm.sol,c(2,3),level=0.99),type="l")
5 > lines(ellipse(lm.sol,c(2,3)),type="l",col="red")
6 > points(coef(lm.sol)[2],coef(lm.sol)[3],pch=19)

```

置信椭球的图像见图4.7.1, 其中黑色和红色的椭球分别是置信水平为0.99和0.95的置信椭球.

若要求单个回归系数的置信区间, 则可使用confint命令. 例如, 求 β_0, \dots, β_3 的置信水平为0.95和0.99的置信区间, 代码和结果如下:

```

1 > confint(lm.sol)
2           2.5 %      97.5 %
3 (Intercept) 252.3700447 5.418047e+02
4 x1         -144.7923367 -7.670766e+01
5 x2           4.2358878 2.693078e+01
6 x3          -0.1174063 8.216995e-04
7 > confint(lm.sol,level=0.99)
8           0.5 %      99.5 %
9 (Intercept) 186.5141597 607.66060660
10 x1         -160.2838337 -61.21616625
11 x2          -0.9279446 32.09461125
12 x3          -0.1443070 0.02772244

```

在构造置信区间的过程中需要模型误差的正态性假设. Bootstrap方法可绕开正态性假设构造置信区间. 构造方法如下: 假设对于线性回归模型 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ 已通过最小二乘法得到 $\hat{\boldsymbol{\beta}}$ 和残差向量 $\hat{\mathbf{e}} = (\hat{e}_1, \dots, \hat{e}_n)'$, 然后采取下列的操作:

步骤1: 从 $\{\hat{e}_1, \dots, \hat{e}_n\}$ 中通过有放回抽样随机取出一个残差样本

$$\mathbf{e}^* = (\hat{e}_1^*, \dots, \hat{e}_n^*)';$$

步骤2: 构造 $\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}^*$;

步骤3: 从 $(\mathbf{X}, \mathbf{Y}^*)$ 计算出 $\hat{\boldsymbol{\beta}}^*$.

重复以上步骤 B 次(称 B 为Bootstrap次数), 则可得到 B 个 $\hat{\boldsymbol{\beta}}^*$. 对于每个 i , 根据 $\hat{\beta}_i^*$ 的 B 个样本值画出密度图: (1) 若图像看起来是与正态密度图像相似, 则用

$$\left(\hat{\beta}_i \pm 2\text{sd}(\hat{\beta}_i^*)\right) \quad (4.7.10)$$

作为 β_i 的置信水平为0.95的置信区间, 其中 $\text{sd}(\hat{\beta}_i^*)$ 表示根据 $\hat{\beta}_i^*$ 的 B 个样本值计算得到的样本标准差; (2) 若图像看起来与正态密度图像相去甚远, 则根据 $\hat{\beta}_i^*$ 的 B 个样本值计算其样本分位数来构造置信区间. 由此可见, 用Bootstrap方法构造置信区间除了无需考虑正态性假设以外, 也无需从理论角度计算置信区间的数学表达式.

仍以煤净化问题为例, 取 $B = 4000$, 来构造置信水平为0.95和0.99的Bootstrap置信区间, R代码如下:

```
1 > set.seed(123)
2 > nb=4000
3 > coefmat=matrix(NA,nb,4)
4 > resids=residuals(lm.sol)
5 > preds=fitted(lm.sol)
6 > for (i in 1:nb){
7 >   booty=preds+sample(resids,rep=T)
8 >   bmod=update(lm.sol,booty~.)
9 >   coefmat[i,]=coef(bmod)
10 > }
11 > colnames(coefmat)=c("Intercept",colnames(coal[,1:3]))
12 > coefmat=data.frame(coefmat)
13 > par(mfrow=c(2,2))
14 > plot(density(coefmat[,1]),xlab="intercpt",main="")
15 > plot(density(coefmat[,2]),xlab="beta1",main="")
16 > plot(density(coefmat[,3]),xlab="beta2",main="")
17 > plot(density(coefmat[,4]),xlab="beta3",main="")
```

所得密度图像见图4.7.2, 可以看出, 它们都与正态密度图像很相似.

接下来用(4.7.10)来构造Bootstrap置信区间. 例如, 构造置信水平为0.95的Bootstrap置信区间, R代码及结果如下:

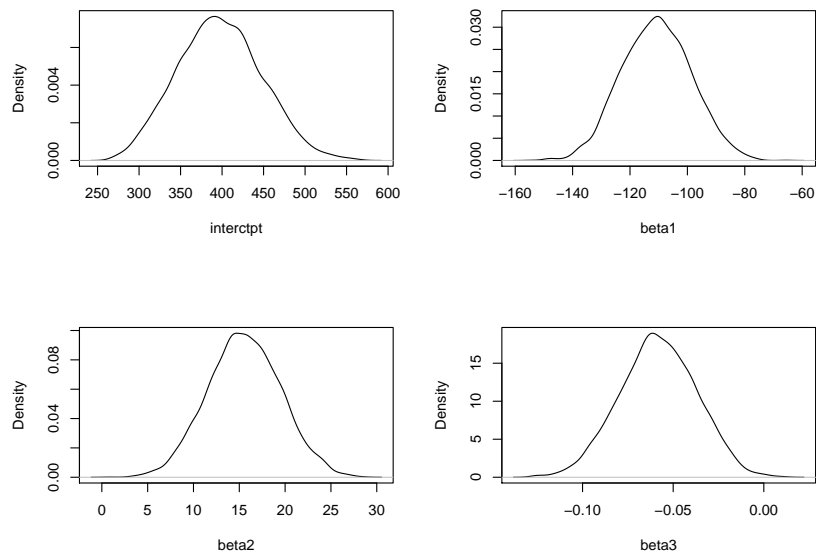


图4.7.2

```

1 > sd_est=apply(coefmat,2,function(x) sd(x))
2 > coef(lm.sol)[1]+c(-2,2)*sd_est[1]
3 [1] 295.8047 498.3701
4 > coef(lm.sol)[2]+c(-2,2)*sd_est[2]
5 [1] -134.87344 -86.62656
6 > coef(lm.sol)[3]+c(-2,2)*sd_est[3]
7 [1] 7.670079 23.496587
8 > coef(lm.sol)[4]+c(-2,2)*sd_est[4]
9 [1] -0.10053013 -0.01605446

```

若采用样本分位数来构造置信水平为0.95的Bootstrap置信区间, R代码及结果如下:

```

1 > apply(coefmat,2,function(x) quantile(x,c(0.025,0.975)))
2 Intercept x1 x2 x3
3 2.5% 301.8739 -133.96170 7.865427 -0.09945016
4 97.5% 498.9897 -86.84628 23.454765 -0.01823595
5 > apply(coefmat,2,function(x) quantile(x,c(0.005,0.995)))
6 Intercept x1 x2 x3
7 0.5% 283.6925 -140.42083 5.582808 -0.1127245
8 99.5% 532.2880 -80.92097 25.691882 -0.0054039

```

4.8 因变量的预测

所谓预测,就是对给定的回归自变量的值,预测对应的回归因变量的可能取值(点预测)或范围(区间预测),这是回归分析最重要的应用之一.因为在线性回归模型中,回归自变量往往代表一组实验条件、生产条件或社会经济条件,由于实验或者生产等方面的费用或花费时间长等原因,在有了回归方程后,我们希望对一些感兴趣的实验、生产条件不真正做实验,就可以对相应的因变量的取值能够作出一些统计推断.

考虑线性回归模型

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, \quad i = 1, \dots, n, \quad (4.8.1)$$

假设模型误差 $\{e_1, \dots, e_n\}$ 为i.i.d.序列且满足Gauss-Markov假设,这里

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)', \quad \mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})', \quad i = 1, \dots, n.$$

给定

$$\mathbf{x}_0 = (1, x_{01}, \dots, x_{0p})',$$

对应的因变量为 y_0 ,它可表示为

$$y_0 = \mathbf{x}_0' \boldsymbol{\beta} + e_0, \quad (4.8.2)$$

这里 e_0 与 e_1, \dots, e_n 不相关.我们对 y_0 的点预测和区间预测感兴趣.

先考虑点预测.注意到 y_0 由两部分组成: $\mathbf{x}_0' \boldsymbol{\beta}$ 和 e_0 .自然地,可以用 $\mathbf{x}_0' \hat{\boldsymbol{\beta}}$ 去估计 $\mathbf{x}_0' \boldsymbol{\beta}$.因为 e_0 是均值为零的随机变量,因此可以用0去估计它.由此, y_0 的一个点预测为

$$\hat{y}_0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}} + 0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}}. \quad (4.8.3)$$

点预测具有如下的两个性质:(1) \hat{y}_0 是 y_0 的无偏估计.这里“无偏”的含义是指预测量与被预测量具有相同的均值,这个概念不同于统计学中的参数估计的无偏性.这个性质是容易证明的,因为 $E(\hat{y}_0) = E(\mathbf{x}_0' \hat{\boldsymbol{\beta}}) = \mathbf{x}_0' \boldsymbol{\beta} = E(y_0)$.(2) 在 y_0 的一切线性无偏预测中, \hat{y}_0 具有最小方差.事实上,假设 $\mathbf{a}' \mathbf{Y}$ 是 y_0 的任一线性无偏预测,则 $E(\mathbf{a}' \mathbf{Y}) = E(y_0) = \mathbf{x}_0' \boldsymbol{\beta}$.因此 $\mathbf{a}' \mathbf{Y}$ 可看作 $\mathbf{x}_0' \boldsymbol{\beta}$ 的一个线性无偏估计.而点预测 $\hat{y}_0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}}$ 也可以看作 $\mathbf{x}_0' \boldsymbol{\beta}$ 的一个线性无偏估计.由Gauss-Markov定理知 $\text{Var}(\mathbf{a}' \mathbf{Y}) \geq \text{Var}(\mathbf{x}_0' \hat{\boldsymbol{\beta}}) = \text{Var}(\hat{y}_0)$.

值得注意的是,虽然从形式上讲, y_0 的点预测 $\hat{y}_0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}}$ 与参数函数 $\mu_0 = \mathbf{x}_0' \boldsymbol{\beta}$ 的最小二乘估计 $\hat{\mu}_0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}}$ 完全相同,但它们之间还是有本质区别的: $\hat{\mu}_0$ 是未知参数的点估计,而 \hat{y}_0 是随机变量的“点估计”.这也导致它们的估计/预测精度有所不同.引进预测偏差 $d_1 = y_0 - \hat{y}_0$ 和估计偏差 $d_2 = \mu_0 - \hat{\mu}_0$,然后计算它们的方差.由于 e_0 与 e_1, \dots, e_n 不相关,所以 y_0 与 $\hat{\boldsymbol{\beta}}$ 也不相关,因此

$$\text{Var}(d_1) = \text{Var}(y_0) + \text{Var}(\hat{y}_0) = \sigma^2[1 + \mathbf{x}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0].$$

而

$$\text{Var}(d_2) = \text{Var}(\mathbf{x}_0' \hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{x}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0.$$

所以总有 $\text{Var}(d_1) > \text{Var}(d_2)$.

再考虑区间预测. 所谓区间预测就是寻找一个区间, 使得被预测量落在这个区间内的概率达到预先给定的值. 假设模型误差还服从正态分布, 即

$$e_1, \dots, e_n, e_0 \stackrel{i.i.d.}{\sim} N(0, \sigma^2).$$

这时可知

$$y_0 - \hat{y}_0 \sim N(0, \sigma^2[1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]), \quad (4.8.4)$$

且

$$y_0 - \hat{y}_0 \text{ 与 } \hat{\sigma}^2 \text{ 相互独立.} \quad (4.8.5)$$

所以根据

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

可推得

$$\frac{y_0 - \hat{y}_0}{\hat{\sigma} \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \sim t(n-p-1).$$

因此对于给定的 α , 有

$$P\left(\frac{|y_0 - \hat{y}_0|}{\hat{\sigma} \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \leq t_{\alpha/2}(n-p-1)\right) = 1 - \alpha.$$

由此可得到 y_0 的概率为 $1 - \alpha$ 的预测区间为

$$\left(\hat{y}_0 \pm t_{\alpha/2}(n-p-1)\hat{\sigma} \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\right).$$

例4.8.1 考虑一元线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n.$$

自变量为 x_0 时对应的因变量 y_0 的点预测为

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

给定 $\mathbf{x}_0 = (1, x_0)'$, y_0 的概率为 $1 - \alpha$ 的预测区间为

$$\left(\hat{y}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right).$$

可以看出: 预测区间的长度在 $x_0 = \bar{x}$ 时达到最小, x_0 离 \bar{x} 越远, 预测区间就越长.

例4.8.2(续例4.2.1) 对于煤净化问题, 已建立了回归方程

$$\hat{y} = 397.087 - 110.750x_1 + 15.583x_2 - 0.058x_3.$$

假设现在要考察在 $\mathbf{x}_0 = (1, x_{01}, x_{02}, x_{03})' = (1, 1.5, 7.5, 1315)'$ 这一实验条件下的净化效率, 取 $\alpha = 0.05$, 计算得到

$$\hat{y}_0 = 397.087 - 110.750 \times 1.5 + 15.583 \times 7.5 - 0.058 \times 1315 = 271.565,$$

$$\hat{\sigma} = 20.88, n = 12, p = 3, t_{0.025}(12 - 3 - 1) = t_{0.025}(8) = 2.306.$$

于是, 经过简单的计算, 得到 y_0 的概率为 0.95 的预测区间为

$$(215.756, 326.609).$$

这一过程的 R 代码及分析结果如下 (同时给出了 y_0 的概率为 0.99 的预测区间):

```
1 > new=data.frame(x1=1.5,x2=7.5,x3=1315)
2 > predict(lm.sol,new,interval="prediction")
3      fit      lwr      upr
4 1 271.183 215.7633 326.6027
5 > predict(lm.sol,new,interval="prediction",level=0.99)
6      fit      lwr      upr
7 1 271.183 190.5437 351.8223
```

注 若把 interval="prediction" 改成 interval="confidence", 则可得到 $E(y_0)$ 的区间估计.

在作因变量的预测的时候, 还需关注下列几种情形, 他们可能会使预测变得糟糕或不可靠:

- (1) 不正确的模型. 这种情况发生在在数据建模方面做得很差的时候;
- (2) 定量外推. 当自变量的取值与我们在数据中看到的自变量的取值有较大差异的时候 (即此时我们在做样本外预测), 外推结果通常是不够理想的;
- (3) 定性外推. 当自变量的取值与统计建模的自变量数据来自不同的总体时, 外推结果是值得怀疑的. 例如, 用男性的身体数据进行体脂含量的统计建模, 却用这个模型去预测女性的体脂含量, 这显然是不合理的;
- (4) 过度拟合导致的过度自信. 数据分析师四处寻找适合于他们拥有的数据的好模型, 有时在寻找合适的模型方面做得太好了, 这可能会导致不切实际的过小的 $\hat{\sigma}$;
- (5) 黑天鹅事件. 有时, 模型误差可能乍看起来是来自正态分布的, 这其实是因为我们没有看到足够多的数据来了解极端情况. 在金融应用领域, 股票价格在大部分时间都在做小幅变化 (正态分布), 但可能也会出现不常见的大幅度波动 (通常是因为黑天鹅事件导致的下跌).

作业

1. 假设有一个大的模型 Ω , 它有 p 个自变量; 以及一个小的模型 ω , 它有 q 个自变量. 两个模型的残差平方和分别记为 RSS_{Ω} 和 RSS_{ω} . 我们的目的是根据样本选择其中的一个模型. 试从似然比检验的方法出发, 推导出检验统计量

$$F_H = \frac{(RSS_{\omega} - RSS_{\Omega}) / (p - q)}{RSS_{\Omega} / (n - p - 1)}.$$

2. 设

$$\begin{cases} y_1 = \beta_1 + e_1, \\ y_2 = 2\beta_1 - \beta_2 + e_2, \\ y_3 = \beta_1 + 2\beta_2 + e_3, \end{cases}$$

这里, $\mathbf{e} = (e_1, e_2, e_3)' \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_3)$. 试推导出检验 $H_0: \beta_1 = 2\beta_2$ 的统计量.

3. 设

$$\begin{aligned} y_{1i} &\sim N(\beta_{10} + \beta_{11}x_{1i}, \sigma^2), \quad i = 1, \dots, n, \\ y_{2i} &\sim N(\beta_{20} + \beta_{21}x_{2i}, \sigma^2), \quad i = 1, \dots, m \end{aligned}$$

且相互独立, 试推导出检验 $H_0: \beta_{11} = \beta_{21}$ 的统计量.

4. 设 $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, \dots, n$. $\{e_i, i = 1, \dots, n\}$ 相互独立且服从 $N(0, \sigma^2)$. 试推导出 $H_0: \beta_0 = 0$ (即回归直线经过原点) 的检验统计量.

5. 在一项关于收入、流动性和消费之间关系的研究中, 研究者将消费(记为 y)与收入(记为 x_1)和流动性(记为 x_2)做回归, 得到以下的回归结果:

$$\hat{y}_i = 100 + 0.85x_{i1} + 6.02x_{i2}, \quad i = 1, \dots, n.$$

此外, 判定系数 $R^2 = 0.65$, 样本容量 $n = 25$. 考虑收入和流动性的联合显著性, 构造检验统计量, 在显著性水平 0.05 下你的检验结论是什么?

6. 基于某一行业 $n = 22$ 家公司的数据, 计算得到

$$\begin{aligned} \bar{y} = 20, \quad \bar{x} = 10, \quad \sum_{i=1}^n (y_i - \bar{y})^2 &= 100, \\ \sum_{i=1}^n (x_i - \bar{x})^2 &= 60, \quad \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 30, \end{aligned}$$

其中 $y = \ln(\text{产出})$, $x = \ln(\text{劳动小时数})$, 两者之间的关系用 $y = \beta_0 + \beta_1 x + e$ 进行建模, 其中 $e \sim N(0, \sigma^2)$.

(1) 在显著性水平 0.05 下检验假设 $H_0: \beta_1 = 1$;

(2) 构造 σ^2 的置信水平为 0.95 的置信区间.

7. 对于一元线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \dots, n,$$

$E(e_i) = 0$, $\text{Var}(e_i) = \sigma^2$, 各 e_i 互不相关. 记 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别为 β_0 和 β_1 的最小二乘估计, 记 $\mu_0 = \beta_0 + \beta_1 x_0$ 为因变量 y 在 x_0 处的取值 y_0 的均值, 即 $E(y_0) = \mu_0$. 用 $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 估计 μ_0 , 证明

$$\text{Var}(\hat{\mu}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

8. 假设回归直线通过原点的一元线性回归模型为

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

$E(e_i) = 0$, $\text{Var}(e_i) = \sigma^2$, 各 e_i 互不相关.

- (1) 写出 β 和 σ^2 的最小二乘估计;
- (2) 记因变量 y 在 x_0 处的值 y_0 的预测值为 $\hat{y}_0 = \hat{\beta}x_0$, 求 $\text{Var}(\hat{y}_0 - y_0)$;
- (3) 记 $\mu_0 = \beta x_0$, $\hat{\mu}_0 = \hat{\beta}x_0$, 求 $\text{Var}(\hat{\mu}_0)$.

思考题: 考虑线性回归模型

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

其中 \mathbf{e} 满足Gauss-Markov假设, \mathbf{X} 为 $n \times (p+1)$ 列满秩设计矩阵. 关注线性假设

$$H: \mathbf{A}\boldsymbol{\beta} = \mathbf{b},$$

这里 \mathbf{A} 为 $m \times (p+1)$ 矩阵, 秩为 m ; \mathbf{b} 为 $m \times 1$ 的已知向量. 最小二乘法基本定理是在模型误差的正态性假设下推导出来的. 若没有正态性假设, 如何构造上述线性假设的渐近检验统计量? 试至少给出两种方法. (可作为大作业)

参考文献

Myers, R. H. (2000), Classical and Modern Regression with Applications (2nd Edition), Boston: Duxbury Press.