

第四章 假设检验与预测

Tianxiao Pang

Zhejiang University

November 17, 2016

内容

1 一般线性假设

内容

- 1 一般线性假设
- 2 回归方程的显著性检验

内容

- 1 一般线性假设
- 2 回归方程的显著性检验
- 3 回归系数的显著性检验

内容

- 1 一般线性假设
- 2 回归方程的显著性检验
- 3 回归系数的显著性检验
- 4 异常点检验和D-W检验

内容

- 1 一般线性假设
- 2 回归方程的显著性检验
- 3 回归系数的显著性检验
- 4 异常点检验和D-W检验
- 5 置信椭球和同时置信域

内容

- 1 一般线性假设
- 2 回归方程的显著性检验
- 3 回归系数的显著性检验
- 4 异常点检验和D-W检验
- 5 置信椭球和同时置信域
- 6 因变量的预测

在上一章, 我们讨论了回归系数的几种点估计方法, 从而建立了(经验)回归方程. 但是, 所建立的回归方程是否真正刻画了因变量和自变量之间的依赖关系呢? 从统计理论角度上, 我们可以做统计假设检验进行分析, 这叫做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, 这就导致了相应的回归系数的显著性检验. 本章前三节将讨论这些假设检验问题, 第四节讨论异常值的假设检验以及模型误差的不相关性检验, 第五节讨论回归系数的(同时)置信区间问题, 最后一节讨论在给定自变量的情况下如何预测因变量的值.

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. 我们将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 $m \times 1$ 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.

检验方法的基本思想:

应用最小二乘法, 得 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ 和残差平方和

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

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)

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

和相应的残差平方和

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

加了约束条件(4.1.2)后, 回归系数 β 的搜索范围变小了, 因而残差平方和 RSS_H 就自然变大了. 于是总有 $RSS_H \geq RSS$. 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, 因此 $RSS_H - RSS$ 应偏小. 若约束条件(4.1.2)不成立, 则 $RSS_H - RSS$ 应偏大. 所以, 当 $RSS_H - RSS$ 偏大到一定程度时, 我们有充分的理由拒绝假设(4.1.2), 否则只能接受它.

在统计学上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, 我们可以把 RSS 取为标准. 于是用

$$(RSS_H - RSS)/RSS$$

的大小来决定接受还是拒绝(4.1.2).

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) $RSS/\sigma^2 \sim \chi^2(n-p-1)$;

(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

$$F_H = \frac{(RSS_H - RSS)/m}{RSS/(n-p-1)} \sim F(m, n-p-1), \quad (4.1.6)$$

$F(m, n-p-1)$ 表示自由度为 $(m, n-p-1)$ 的 F 分布.

证明: (a). 已在定理3.2.4中给出.

(b). 由第三章, 我们已知道

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

即

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

利用 $\hat{\beta}_H$ 的表达式(4.1.4)可得

$$RSS_H - RSS = (A\hat{\beta} - b)' (A(X'X)^{-1}A')^{-1} (A\hat{\beta} - b). \quad (4.1.8)$$

因为 $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$, 所以根据定理2.3.2知

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

若假设(4.1.2)成立, 则

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

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

(c). 因为

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

代入(4.1.8)得

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

其中

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

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

$$RSS = e'Ne.$$

因此, 为证 $RSS_H - RSS$ 与 RSS 独立, 只需证 $e'Me$ 和 $c'e$ 都与 $e'Ne$ 独立. 由于

$$e \sim N(0, \sigma^2 I), \quad NM = 0, \quad c'N = 0,$$

根据定理2.4.5和2.4.6可推得结论成立.

(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 $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ 的检验统计量. 对于给定的显著水平 α , 假设检验的拒绝域为

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

RSS与RSS_H的计算: 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 $\beta_1 = \beta_2 = \beta_3$, 此时线性假设的形式为 $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, 其中

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

\mathbf{A} 的秩为2. 将 $\beta_1 = \beta_2 = \beta_3$ 代入原模型得

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

这个约简模型等价于原来的带约束条件的模型. 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2,3,4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

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

其残差平方和为

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

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

将观测数据写成线性回归模型

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

线性假设 H 等价于

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

已知 $\mathbf{A} = \begin{pmatrix} 1 & -1 \end{pmatrix}$, $m = \text{rk}(\mathbf{A}) = 1$, $n = 3$, $p + 1 = 2$. 易求 $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ 的LSE为

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

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 $\beta_1 = \beta_2 \triangleq \alpha$ 代入模型, 得约简模型

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

从而得 α 的LSE为

$$\hat{\alpha} = \hat{\beta}_H = \frac{1}{11}(y_1 + y_2 + 3y_3).$$

于是

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

然后可通过简单计算得到检验统计量 F_H . 若 H 成立, 则

$$F_H \sim F(m, n - p - 1) = F(1, 1).$$

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). 假设我们对因变量 y 和自变量 x_1, \dots, x_p 有两批观测数据. 对第一批数据, 有线性回归模型

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

对第二批数据, 有线性回归模型

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

其中 $e_1, \dots, e_{n_1+n_2}$ 独立同分布服从 $N(0, \sigma^2)$. 问题: 这两批数据所反映的因变量 y 与自变量 x_1, \dots, x_p 之间的依赖关系是否一样? 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. 那么我们所要做的检验就是考察公司的效益指标对诸因素的依赖关系在两个不同的时间段上是否有了变化.

再如, 在生物学研究中, 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观测数据, 而我们检验的目的就是考察两个实验室所得结论有没有差异.

推导检验统计量: 把两个模型写成矩阵形式

$$\begin{aligned} Y_1 &= X_1\beta_1 + e_1, \quad e_1 \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_1}), \\ Y_2 &= X_2\beta_2 + e_2, \quad e_2 \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_2}). \end{aligned}$$

将它们合并得

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 & \mathbf{0} \\ \mathbf{0} & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}, \quad \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_1+n_2}). \quad (4.1.12)$$

我们要检验的假设为

$$H: (\mathbf{I}_{p+1} \quad \vdots \quad -\mathbf{I}_{p+1}) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \mathbf{0}. \quad (4.1.13)$$

从(4.1.12)得到的 β_1, β_2 的LSE为

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

相应的残差平方和为

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

将约束条件 $\beta_1 = \beta_2 \triangleq \beta$ 代入原模型得约简模型

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \beta + \mathbf{e},$$

利用最小二乘法得约束最小二乘估计为

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

相应的残差平方和为

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

结合(4.1.14)和(4.1.15)得

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

因此检验统计量为

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

若 H 成立, 则 $F_H \sim F(p+1, n_1 + n_2 - 2p - 2)$. 对给定的显著水平 α , 若 $F_H > F_\alpha(p+1, n_1 + n_2 - 2p - 2)$, 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, 认为它们来自同一线性回归模型.

回归方程的显著性检验

将正态线性回归模型(4.1.1)写成

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

所谓回归方程的显著性检验, 就是检验自变量这个整体是否对因变量有显著的线性作用. 即检验假设:

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

若拒绝原假设, 则认为至少有某一个 x_i 对因变量 y 有显著的线性作用. 若接受原假设, 则认为相对于误差而言, 所有自变量对因变量 y 的影响是可以忽略不计的.

假设(4.2.2)是线性假设 $A\beta = b$ 的特例, 取

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

这里的 $\mathbf{0}$ 是 p 维的零列向量. 因此定理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. 我们将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)代入模型(4.2.1), 得约简模型

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

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

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

这就是回归分析中的总平方和TSS. 这是因为约简模型中不包含任何回归自变量, 残差平方和 RSS_H 完全是样本 y_1, \dots, y_n 的变动平方和.

对于原模型(4.2.1), 残差平方和

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

于是

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

它是由于在模型(4.2.3)中引入回归自变量后所引起的残差平方和的减少量, 所以称为回归平方和(或解释平方和, ESS). 注意到 $\text{rk}(A) = p$, 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, $F_H \sim F(p, n-p-1)$. 给定显著水平 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 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)是把自变量的贡献和模型误差的贡献进行比较.

当回归平方和相对于误差平方和比较大时, 那么我们就有充分的理由相信回归自变量对因变量有显著性的作用,从而拒绝原假设.

当回归平方和相对于误差平方和比较小时, 我们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作用, 因此接受原假设.

表4.2.1: 方差分析表

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	$n - 1$			

$P(F > F_H)$ 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, 一般取0.05. 注意(4.2.7)等价于 $F_H > F_\alpha(p, n - p - 1)$.

如果经过检验, 我们接受原假设 $H: \beta_1 = \cdots = \beta_p = 0$, 这意味着, 和模型误差比起来, 诸自变量对 y 的影响是不重要的. 这里可能有两种情况:

(1)模型的各种误差太大, 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, 我们要想办法缩小模型误差(检查是否漏掉了重要自变量, y 对模型自变量是否有非线性相依关系);

(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(取自Myers(1986)). 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, x_2 是溶液的pH 值, x_3 表示溶液流量. 实验目的是通过一组实验数据, 建立净化效率 y 与三个因素 x_1, x_2, x_3 的经验关系.

表4.2.2: 煤净化数据

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

```
yx=read.table("ex_p89_data.txt")
x1=yx[,1]
x2=yx[,2]
x3=yx[,3]
y=yx[,4]
myers=data.frame(x1,x2,x3,y)
myers
lm.sol=lm(y~x1+x2+x3)
summary(lm.sol)
```

```

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min       1Q   Median       3Q      Max
-23.808 -17.193   0.904   8.143  32.192

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  397.08738    62.75676     6.327 0.000226 ***
x1          -110.75000    14.76248    -7.502 6.91e-05 ***
x2             15.58333     4.92083     3.167 0.013258 *
x3             -0.05829     0.02563    -2.274 0.052565 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.88 on 8 degrees of freedom
Multiple R-squared:  0.8993,    Adjusted R-squared:  0.8616
F-statistic: 23.83 on 3 and 8 DF,  p-value: 0.0002422

```

$F_H = 23.83$, p -value为0.0002422 < 0.05 , 所以认为回归自变量对因变量有显著的线性影响. 回归方程为

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

回归系数的显著性检验

回归方程的显著性检验是对线性回归自变量的一个整体性检验. 如果我们检验的结果是拒绝原假设, 这意味着因变量 y 线性地依赖于 x_1, \dots, x_p 这个回归自变量的整体. 但是, 这不排除 y 不依赖于其中某些自变量, 即某些 $\beta_i = 0$. 于是在回归方程的显著性检验被拒绝后, 我们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 $\mathbf{b} = 0$,

$$\beta = (\beta_0, \beta_1, \dots, \beta_{i-1}, \beta_i, \beta_{i+1}, \dots, \beta_p)',$$

$$\mathbf{A} = (0, 0, \dots, 0, 1, 0, \dots, 0),$$

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

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

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

其中 $c_{i+1,i+1}$ 为 $(\mathbf{X}'\mathbf{X})^{-1}$ 的第 $i+1$ 个对角元. 记

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

所以检验统计量为

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

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 .

除了 F 检验, 我们还有 t 检验. 根据 F 分布于 t 分布的关系, 可立刻得到检验统计量为 $t_i = \hat{\beta}_i / \sqrt{\hat{\sigma}^2 c_{i+1, i+1}}$. 我们也可以通过如下途径得到这个检验统计量: 根据定理3.2.4知 $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, 所以

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{i+1, i+1}). \quad (4.3.2)$$

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

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

所以

$$t_i = \frac{\hat{\beta}_i}{\hat{\sigma} \sqrt{c_{i+1, i+1}}} \stackrel{H_i}{\sim} t(n-p-1). \quad (4.3.3)$$

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, 通常称 $\hat{\sigma}\sqrt{c_{i+1,i+1}}$ 为 $\hat{\beta}_i$ 的标准误(Std. Error), 记为 $\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma}\sqrt{c_{i+1,i+1}}$.

实际中, 我们往往通过 p -value进行假设检验. 当

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

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

```

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min       1Q   Median       3Q      Max
-23.808 -17.193   0.904   8.143  32.192

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  397.08738    62.75676   6.327 0.000226 ***
x1          -110.75000    14.76248  -7.502 6.91e-05 ***
x2           15.58333     4.92083   3.167 0.013258 *
x3           -0.05829     0.02563  -2.274 0.052565 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.88 on 8 degrees of freedom
Multiple R-squared:  0.8993,    Adjusted R-squared:  0.8616
F-statistic: 23.83 on 3 and 8 DF,  p-value: 0.0002422

```

因为 $p_1 < 0.05$, $p_2 < 0.05$, $p_3 \geq 0.05$, 所以我们拒绝 $H_1 : \beta_1 = 0$ 和 $H_2 : \beta_2 = 0$, 接受 $H_3 : \beta_3 = 0$. 然后修改回归模型, 重新进行回归分析.

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组的主体. 我们在上一章已用学生化残差来判断异常点, 现在我们将通过假设检验的方法来检验异常点.

把正态线性回归模型(4.2.1)写成

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, \quad e_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n. \quad (4.4.1)$$

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 $E(y_j)$ 发生了漂移 η , 我们因此有了一个新的模型

$$\begin{cases} y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, & i \neq j, \\ y_j = \mathbf{x}_j' \boldsymbol{\beta} + \eta + e_j, & e_1, \dots, e_n \stackrel{i.i.d.}{\sim} N(0, \sigma^2). \end{cases} \quad (4.4.2)$$

记 $\mathbf{d}_j = (0, \dots, 0, 1, 0, \dots, 0)'$ 是一个 n 维列向量, 它的第 j 个元素为 1, 其余为 0. 将模型(4.4.2)写成矩阵形式

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{d}_j\eta + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n). \quad (4.4.3)$$

模型(4.4.2)和(4.4.3)称为均值漂移线性回归模型. 我们的目的是要判别 (\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(4.4.3)中 $\boldsymbol{\beta}$ 和 η 的最小二乘估计. 我们来推导检验统计量.

引理 (分块矩阵求逆公式)

设 A 为非奇异的对称矩阵, 将其分块为

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

则当 B^{-1}, D^{-1} 都存在时有

$$\begin{aligned} A^{-1} &= \begin{pmatrix} B_1 & C_1 \\ C_1' & D_1 \end{pmatrix} \\ &= \begin{pmatrix} (B - CD^{-1}C')^{-1} & -B_1CD^{-1} \\ -D^{-1}C'B_1 & D^{-1} + D^{-1}C'B_1CD^{-1} \end{pmatrix} \\ &= \begin{pmatrix} B^{-1} + B^{-1}CD_1C'B^{-1} & -B^{-1}CD_1 \\ -D_1C'B^{-1} & (D - C'B^{-1}C)^{-1} \end{pmatrix}. \end{aligned}$$

定理 (4.4.1)

对均值漂移线性回归模型(4.4.3), 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(4.4.1)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 个对角元, \hat{e}_j 为从模型(4.4.1)导出的第 j 个残差.

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回归模型中, 回归系数 β 的最小二乘估计恰好等于在原来模型中剔除第 j 组数据后所获得的最小二乘估计.

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

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

应用分块矩阵求逆公式以及注意到 $h_{jj} = \mathbf{x}_j' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_j$, 我们有

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

再回忆公式(3.4.9), 可知定理成立.

我们应用最小二乘法基本定理来推导出检验 $H: \eta = 0$ 的检验统计量. 首先把 $H: \eta = 0$ 写成 $\mathbf{A}(\boldsymbol{\beta}) = \mathbf{b}$, 其中

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

另外, 注意到在约束条件 $\eta = 0$ 下, 模型(4.4.3)退化到约简模型(4.4.1), 所以

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

而模型(4.4.3)的无约束残差平方和为

$$\text{RSS} = \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}^*\mathbf{X}'\mathbf{Y} - \eta^* \mathbf{d}_j'\mathbf{Y}. \quad (4.4.4)$$

利用定理4.4.1得

$$\begin{aligned} \text{RSS}_H - \text{RSS} &= (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{Y} + \eta^* \mathbf{d}_j'\mathbf{Y} \\ &= -\frac{\hat{e}_j \mathbf{x}_j'}{1 - h_{jj}} \hat{\boldsymbol{\beta}} + \frac{\hat{e}_j y_j}{1 - h_{jj}} \\ &= \frac{\hat{e}_j^2}{1 - h_{jj}}. \end{aligned}$$

RSS可进一步写成

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

其中 $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p-1)$. 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 (4.4.2)

对于均值漂移线性回归模型(4.4.3), 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

$$F_j = \frac{(n-p-2)r_j^2}{n-p-1-r_j^2} > F_\alpha(1, n-p-2), \quad (4.4.5)$$

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, 我们也可以用 t 检验法完成上面的检验. 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

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

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

定理4.4.2提供了单个异常点的检验方法, 多个异常点的同时检验方法见陈希孺等(1987).

例4.4.1 (续例4.2.1)我们来检验12组数据中是否有异常点.

```
yx=read.table("ex_p89_data.txt")
x1=yx[,1];x2=yx[,2];x3=yx[,3];y=yx[,4]
myers=data.frame(x1,x2,x3,y)
myers
lm.sol=lm(y~x1+x2+x3)
summary(lm.sol)
```

```
> y.fit=predict(lm.sol)
> e=residuals(lm.sol)
> r.student=rstandard(lm.sol)
> h.x=matrix(c(rep(1,12),x1,x2,x3),12,4,byrow=F)
> h.value=hat(h.x)
> n=12;p=3
> t=sqrt((n-p-2)/(n-p-1-r.student^2))*r.student
> result=data.frame(y,y.fit,e,h.value,r.student,t)
> result
```

	y	y.fit	e	h.value	r.student	t
1	243	247.808	-4.8080209	0.45013318	-0.31057272	-0.29228153
2	261	247.808	13.1919791	0.45013318	0.85213208	0.83593622
3	244	261.040	-17.0399538	0.46602674	-1.11695166	-1.13724413
4	285	261.040	23.9600462	0.46602674	1.57055669	1.76647908
5	202	200.652	1.3479747	0.08384008	0.06745619	0.06311744
6	180	200.652	-20.6520253	0.08384008	-1.03348155	-1.03854449
7	183	200.652	-17.6520253	0.08384008	-0.88335367	-0.86981016
8	207	200.652	6.3479747	0.08384008	0.31766931	0.29904453
9	216	183.808	32.1919791	0.45013318	2.07943160	2.86951185
10	160	183.808	-23.8080209	0.45013318	-1.53787224	-1.71405315
11	104	103.540	0.4600462	0.46602674	0.03015556	0.02820955
12	110	103.540	6.4600462	0.46602674	0.42344947	0.40061579

因为 $t_{0.025}(7) = 2.3646$, 而 $|t_9| = 2.8695 > 2.3646$, 所以我们认为第9号数据为异常点.

outlierTest()也可以用来检测异常点.

```
> library(car)
> outlierTest(lm.sol)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferonni p
9 2.869512      0.024009      0.28811
> |
```

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.

我们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,

$$e_{i+1} = \rho e_i + u_{i+1}, \quad i = 1, 2, \dots, n-1,$$

且假设 $\{u_i\}$ 为独立同分布随机变量序列, 服从 $N(0, \sigma^2)$.

这时检验 $\{e_i\}$ 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

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

为了给出拒绝域, 我们先来看一下DW统计量的意义.

由于 $\{e_i\}$ 是不可观测的, 因此要考查 $\{e_i\}$ 间的相关性常用残差 $\{\hat{e}_i\}$ 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, 它是 ρ 的一个估计,

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

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

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \hat{e}_i \approx \frac{1}{n-1} \sum_{i=2}^n \hat{e}_i \approx \frac{1}{n} \sum_{i=1}^n \hat{e}_i = 0,$$

$$\sum_{i=1}^{n-1} \hat{e}_i^2 \approx \sum_{i=2}^n \hat{e}_i^2 \approx \sum_{i=1}^n \hat{e}_i^2.$$

将它们代入 r 的表达式, 得

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

容易看出, DW统计量与 r 之间有如下的近似关系:

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

由上式知: 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的值我们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.

d_L 和 d_U 的值查阅《线性统计模型》(王松桂等编著)的附录3.

例4.4.2 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, 见下表.

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

首先可求得一元线性回归方程:

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

由此可计算残差并得到

$$DW = 0.6800.$$

取 $\alpha = 0.05$, 查表得 $d_L = 1.10$, $d_U = 1.37$, 现 $DW < d_L$, 这表明 $\{e_i\}$ 存在正相关.

R程序:

```
yx=read.table("ex_442_data.txt")
```

```
x=yx[,1]
```

```
y=yx[,2]
```

```
mydata=data.frame(x,y)
```

```
mydata
```

```
lm.sol=lm(y~x,data=mydata)
```

```
summary(lm.sol)
```

```
library(car)
```

```
durbin.watson(lm.sol)
```

```
> durbin.watson(lm.sol)
lag Autocorrelation D-W Statistic p-value
  1          0.5801892      0.6799675    0.002
Alternative hypothesis: rho != 0
```

置信椭球和同时置信域

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, 这说明从现有数据看我们不能接受假设 $H: \mathbf{A}\boldsymbol{\beta} = \mathbf{b}$. 此时自然希望构造 $\mathbf{a}_i'\boldsymbol{\beta}, i = 1, \dots, m$ 的同时置信域, 这里 $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)'$, $\text{rk}(\mathbf{A}) = m$.

置信椭球:

考虑正态线性回归模型

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

记 $\boldsymbol{\Phi} = \mathbf{A}\boldsymbol{\beta} = (\mathbf{a}_1'\boldsymbol{\beta}, \dots, \mathbf{a}_m'\boldsymbol{\beta})'$, 则 $\hat{\boldsymbol{\Phi}} = \mathbf{A}\hat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\Phi}$ 的最小方差线性无偏估计且

$$\hat{\boldsymbol{\Phi}} \sim N(\boldsymbol{\Phi}, \sigma^2 \mathbf{V}),$$

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

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

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 $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2/(n-p-1)$. 于是

$$\frac{(\hat{\Phi} - \Phi)' \mathbf{V}^{-1}(\hat{\Phi} - \Phi)}{m\hat{\sigma}^2} \sim F(m, n-p-1). \quad (4.5.2)$$

因此对给定的置信水平 $1 - \alpha$, 有

$$\mathbf{P}\left(\frac{(\hat{\Phi} - \Phi)' \mathbf{V}^{-1}(\hat{\Phi} - \Phi)}{m\hat{\sigma}^2} \leq F_{\alpha}(m, n-p-1)\right) = 1 - \alpha \quad (4.5.3)$$

定义

$$D = \{\Phi : (\Phi - \hat{\Phi})' V^{-1} (\Phi - \hat{\Phi}) \leq m \hat{\sigma}^2 F_{\alpha}(m, n - p - 1)\}, \quad (4.5.4)$$

这是一个中心在 $\hat{\Phi}$ 的椭球. 由(4.5.3)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(4.5.4)可知 β 的置信椭球为

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

特别地, 当 $m = 1$ 时, 改记 $A = a'$, 上式变为

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

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

$$(\mathbf{a}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n - p - 1)\hat{\sigma}\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}). \quad (4.5.5)$$

记 $\hat{\sigma}_{\mathbf{a}'\hat{\boldsymbol{\beta}}} = \hat{\sigma}\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$ 为 $\mathbf{a}'\hat{\boldsymbol{\beta}}$ 的标准误, 因此(4.5.5)可简记为

$$(\mathbf{a}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n - p - 1)\hat{\sigma}_{\mathbf{a}'\hat{\boldsymbol{\beta}}}). \quad (4.5.6)$$

例4.5.1 若在(4.5.6)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 $\mathbf{a}'\boldsymbol{\beta} = \beta_i$ 的置信水平为 $1 - \alpha$ 的置信区间为

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

这是我们已知的结果(可从回归系数的显著性检验的分析中推得). 若我们对回归函数 $f(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ 感兴趣, 则依据(4.5.5)可知 $\mathbf{x}'\boldsymbol{\beta}$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$(\mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}). \quad (4.5.7)$$

需指出的是, 对一个固定的 \mathbf{x} , (4.5.7)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(4.5.7)的概率将不再是 $1 - \alpha$, 而是会低于 $1 - \alpha$.

同时置信域:

同时置信域也叫联立区间估计, 下面介绍两种求同时置信域的方法.

1. Scheffé区间

引理

设 \mathbf{a} 和 \mathbf{b} 均为 $n \times 1$ 向量, \mathbf{A} 为 $n \times n$ 正定矩阵. 则

$$\sup_{\mathbf{b} \neq 0} \frac{(\mathbf{a}'\mathbf{b})^2}{\mathbf{b}'\mathbf{A}\mathbf{b}} = \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}.$$

证明: 此结论可由以下不等式推得

$$|\mathbf{a}'\mathbf{b}| = |\mathbf{a}'\mathbf{A}^{-\frac{1}{2}} \cdot \mathbf{A}^{\frac{1}{2}}\mathbf{b}| \leq \|\mathbf{A}^{-\frac{1}{2}}\mathbf{a}\| \cdot \|\mathbf{A}^{\frac{1}{2}}\mathbf{b}\| = \sqrt{\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} \cdot \mathbf{b}'\mathbf{A}\mathbf{b}},$$

并注意 $\mathbf{A}^{-\frac{1}{2}}\mathbf{a} = \mathbf{A}^{\frac{1}{2}}\mathbf{b}$ 时可取到等号.

在(4.5.3)中视 $\Phi - \hat{\Phi}$ 和 V 分别为引理中的 α 和 A , 有

$$\begin{aligned}
 & 1 - \alpha \\
 = & P\left((\Phi - \hat{\Phi})' V^{-1} (\Phi - \hat{\Phi}) \leq m \hat{\sigma}^2 F_{\alpha}(m, n - p - 1)\right) \\
 = & P\left(\sup_{b \neq 0} \frac{(b' (\Phi - \hat{\Phi}))^2}{b' V b} \leq m \hat{\sigma}^2 F_{\alpha}(m, n - p - 1)\right) \\
 = & P\left(\frac{|b' (\Phi - \hat{\Phi})|}{\sqrt{b' V b}} \leq \sqrt{m \hat{\sigma}^2 F_{\alpha}(m, n - p - 1)}, \text{ 对一切 } b \neq 0\right) \\
 = & P\left(|b' (\Phi - \hat{\Phi})| \leq \sqrt{m F_{\alpha}(m, n - p - 1)} \hat{\sigma} \sqrt{b' A (X' X)^{-1} A' b}, \text{ 对一切 } b \neq 0\right).
 \end{aligned}$$

记 $l = A' b$, 则 $b' \Phi = l' \beta$, $b' \hat{\Phi} = l' \hat{\beta}$, 上式变形为

$$\begin{aligned}
 & 1 - \alpha \\
 = & P\left(|l' \beta - l' \hat{\beta}| \leq \sqrt{m F_{\alpha}(m, n - p - 1)} \hat{\sigma} \sqrt{l' (X' X)^{-1} l}, \text{ 对一切 } l \in \mathcal{M}(A')\right) \\
 = & P\left(l' \beta \in l' \hat{\beta} \pm \sqrt{m F_{\alpha}(m, n - p - 1)} \hat{\sigma} \sqrt{l' (X' X)^{-1} l}, \text{ 对一切 } l \in \mathcal{M}(A')\right),
 \end{aligned}$$

这里 $\mathcal{M}(A')$ 表示由 A' 的列向量张成的子空间.

定理

对于正态线性模型(4.5.1), 对一切 $l'\beta$, $l \in \mathcal{M}(A')$, 其置信水平为 $1 - \alpha$ 的同时置信域为

$$(l'\hat{\beta} \pm \sqrt{mF_{\alpha}(m, n-p-1)}\hat{\sigma}\sqrt{l'(X'X)^{-1}l}). \quad (4.5.8)$$

导出(4.5.8)的方法是由Scheffé于1953年提出的, 所以(4.5.8)通常被称为Scheffé区间. 需要强调的是, Scheffé区间并不是一个或若干个未知参数的联立区间估计, 而是无穷多个未知参数 $l'\beta$, $l \in \mathcal{M}(A')$ 的联立区间估计. 若只对有限个 $l'\beta$ 感兴趣, 这时若采用Scheffé 区间, 常常会嫌其偏长.

2. Bonferroni 区间

先用公式(4.5.6)对每个 $\mathbf{a}'_i \boldsymbol{\beta}$ 求置信水平为 $1 - \alpha$ 的置信区间

$$I_i = (\mathbf{a}'_i \hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n - p - 1) \hat{\sigma}_{\mathbf{a}'_i \hat{\boldsymbol{\beta}}}), \quad i = 1, \dots, m. \quad (4.5.9)$$

虽然每个区间 I_i 包含 $\mathbf{a}'_i \boldsymbol{\beta}$ 的概率为 $1 - \alpha$, 但是 $\mathbf{a}'_i \boldsymbol{\beta} \in I_i, i = 1, \dots, m$ 同时成立的概率却不再是 $1 - \alpha$.

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

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

这个公式就是著名的 Bonferroni 不等式.

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

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

当 m 较大时, $1 - m\alpha$ 将变得很小. 在(4.5.9)中把 α 换成 α/m , 即取

$$I_i = (\mathbf{a}_i' \hat{\boldsymbol{\beta}} \pm t_{\alpha/(2m)}(n - p - 1) \hat{\sigma}_{\mathbf{a}_i' \hat{\boldsymbol{\beta}}}), \quad i = 1, \cdots, m, \quad (4.5.10)$$

则依据Bonferroni不等式知

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

通常称(4.5.10)为Bonferroni区间.

例4.5.1 (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

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

现求回归系数的置信水平为0.95和0.99的区间估计.

```
> confint(lm.sol)
                2.5 %      97.5 %
(Intercept)  252.3700447  5.418047e+02
x1          -144.7923367 -7.670766e+01
x2           4.2358878  2.693078e+01
x3          -0.1174063  8.216995e-04
> confint(lm.sol,level=0.99)
                0.5 %      99.5 %
(Intercept)  186.5141597  607.66060660
x1          -160.2838337 -61.21616625
x2           -0.9279446  32.09461125
x3          -0.1443070  0.02772244
> |
```

Bootstrp方法可以构造更精确的区间估计. 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,

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

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

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

重复以上步骤 N (称为bootstrap次数)次, 则可得到 N 个 $\hat{\boldsymbol{\beta}}^*$, 由此可计算样本分位数作为置信区间.

仍以煤净化问题为例, 我们来构造0.95和0.99的bootstrap置信区间.

```
set.seed(123)
nb=4000
coefmat=matrix(NA,nb,4)
resids=residuals(lm.sol)
preds=fitted(lm.sol)
for (i in 1:nb){
  booty=preds+sample(resids,rep=T)
  bmod=update(lm.sol,booty~.)
  coefmat[i,]=coef(bmod)
}
colnames(coefmat)=c("Intercept",colnames(myers[,1:3]))
coefmat=data.frame(coefmat)
apply(coefmat,2,function(x) quantile(x,c(0.025,0.975)))
apply(coefmat,2,function(x) quantile(x,c(0.005,0.995)))
```

```
> apply(coefmat,2,function(x) quantile(x,c(0.025,0.975)))  
      Intercept      x1      x2      x3  
2.5%    300.9620 -134.79895  7.906286 -0.09850227  
97.5%    497.7901  -87.65044 23.631579 -0.01734166  
> apply(coefmat,2,function(x) quantile(x,c(0.005,0.995)))  
      Intercept      x1      x2      x3  
0.5%    277.6515 -142.12716  5.788559 -0.111365217  
99.5%    523.4043  -80.98948 25.926616 -0.001223387
```

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的原因,我们在有了经验回归方程后,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

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, \quad i = 1, \dots, n, \quad (4.6.1)$$

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

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

给定

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

对应的因变量值为 y_0 , 且 y_0 可表示为

$$y_0 = \mathbf{x}_0' \boldsymbol{\beta} + e_0, \quad (4.6.2)$$

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.

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, 我们可以用 $\mathbf{x}'_0\hat{\boldsymbol{\beta}}$ 去估计 $\mathbf{x}'_0\boldsymbol{\beta}$, 因为 e_0 是均值为零的随机变量, 因此我们可以用0去估计它. 由此, y_0 的点预测为

$$\hat{y}_0 = \mathbf{x}'_0\hat{\boldsymbol{\beta}} + 0 = \mathbf{x}'_0\hat{\boldsymbol{\beta}}. \quad (4.6.3)$$

点预测性质:

(1) \hat{y}_0 是 y_0 的无偏估计. 这里“无偏”的含义是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. 这个性质时容易证明的, 因为 $E(\hat{y}_0) = E(\mathbf{x}'_0 \hat{\beta}) = \mathbf{x}'_0 \beta = E(y_0)$.

(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 $E(\mathbf{a}'\mathbf{Y}) = E(y_0) = \mathbf{x}'_0 \beta$. 因此 $\mathbf{a}'\mathbf{Y}$ 可看作是 $\mathbf{x}'_0 \beta$ 的一个线性无偏估计. 而预测 $\hat{y}_0 = \mathbf{x}'_0 \hat{\beta}$ 也可以看作是 $\mathbf{x}'_0 \beta$ 的一个线性无偏估计. 根据 Gauss-Markov 定理知 $\text{Var}(\mathbf{a}'\mathbf{Y}) \geq \text{Var}(\mathbf{x}'_0 \hat{\beta})$.

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 $d_2 = \mathbf{x}_0' \boldsymbol{\beta} - \hat{\mu}_0$, 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 $\text{Var}(d_1) > \text{Var}(d_2)$.

区间预测:

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

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

这时可知

$$y_0 - \hat{y}_0 \sim N(0, \sigma^2[1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]) \quad (4.6.4)$$

以及(因为 $\hat{\beta}$ 与残差向量 \hat{e} 相互独立)

$$y_0 - \hat{y}_0 \text{ 与 } \hat{\sigma}^2 \text{ 相互独立.} \quad (4.6.5)$$

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4.6.1 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 时对应因变量 y_0 的点预测为

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

而概率为 $1 - \alpha$ 的区间预测为 $(\mathbf{x}_0 = (1, x_0)')$

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

因此, 预测区间的长度在 $x_0 = \bar{x}$ 时达到最小, 而当 x_0 离 \bar{x} 愈远, 预测区间就愈长.

例4.6.2 (续例4.2.1)对于煤净化问题, 我们已建立了回归方程

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

假设我们要考察 $\mathbf{x}_0 = (1, x_{01}, x_{02}, x_{03})' = (1, 1.5, 7.5, 1315)'$ 这一实验条件下的净化效率, 取 $\alpha = 0.05$,

$$\hat{y}_0 = 397.087 - 110.750 \times 1.5 + 15.583 \times 7.5 - 0.058 \times 1315 = 271.564.$$

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

于是通过计算可得 y_0 的概率为0.95的预测区间为

$$(215.756, 326.609).$$

```
new=data.frame(x1=1.5,x2=7.5,x3=1315)
lm.pred=predict(lm.sol,new,interval="prediction",level=0.95)
lm.pred
```

```
      fit      lwr      upr
1 271.183 215.7633 326.6027
```

注: 把interval="prediction" 改成interval="confidence" 可得到 $E(y_0)$ 的区间估计.