第四章　模型的推断与预测

Tianxiao Pang

Zhejiang University

November 15, 2023

内容

内容

内容

内容

内容

内容

内容

内容

上一章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.

上一章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.

上一章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.

本章前四节将讨论这些假设检验问题, 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置信区间, 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.

一般线性假设

考虑正态线性回归模型

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

其中$\boldsymbol{X}$为$n \times (p+1)$列满秩设计矩阵. 考虑比较一般的线性假设

$$H: \ \ \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b} \tag{4.1.2}$$

的检验问题. 这里$\boldsymbol{A}$为$m \times (p+1)$矩阵, 秩为m; $\boldsymbol{b}$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(如回归方程的显著性检验、回归系数的显著性检验).

检验方法的基本思想:

检验方法的基本思想:

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

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

RSS反映了实际数据与模型(4.1.1)的拟合程度.

检验方法的基本思想:

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

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

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)

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

和相应的残差平方和

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

加了约束条件(4.1.2)后, 回归系数$\boldsymbol{\beta}$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$\mathrm{RSS}_H \geq \mathrm{RSS}$.

加了约束条件(4.1.2)后, 回归系数$\boldsymbol{\beta}$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$\mathrm{RSS}_H \geq \mathrm{RSS}$. 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$\mathrm{RSS}_H - \mathrm{RSS}$应偏小. 若约束条件(4.1.2)不成立, 此时$\mathrm{RSS}_H - \mathrm{RSS}$应偏大.

加了约束条件(4.1.2)后, 回归系数$\boldsymbol{\beta}$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$\mathrm{RSS}_H \geq \mathrm{RSS}$. 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$\mathrm{RSS}_H - \mathrm{RSS}$应偏小. 若约束条件(4.1.2)不成立, 此时$\mathrm{RSS}_H - \mathrm{RSS}$应偏大. 所以, 当$\mathrm{RSS}_H - \mathrm{RSS}$偏大到一定程度时, 就有充分的理由拒绝假设(4.1.2), 否则只能接受它.

加了约束条件(4.1.2)后, 回归系数$\boldsymbol{\beta}$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$\mathrm{RSS}_H \geq \mathrm{RSS}$. 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$\mathrm{RSS}_H - \mathrm{RSS}$应偏小. 若约束条件(4.1.2)不成立, 此时$\mathrm{RSS}_H - \mathrm{RSS}$应偏大. 所以, 当$\mathrm{RSS}_H - \mathrm{RSS}$偏大到一定程度时, 就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, 可以把RSS取为标准. 于是用

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

的大小来决定接受还是拒绝(4.1.2). 也可从似然比检验的方法推导出此统计量.

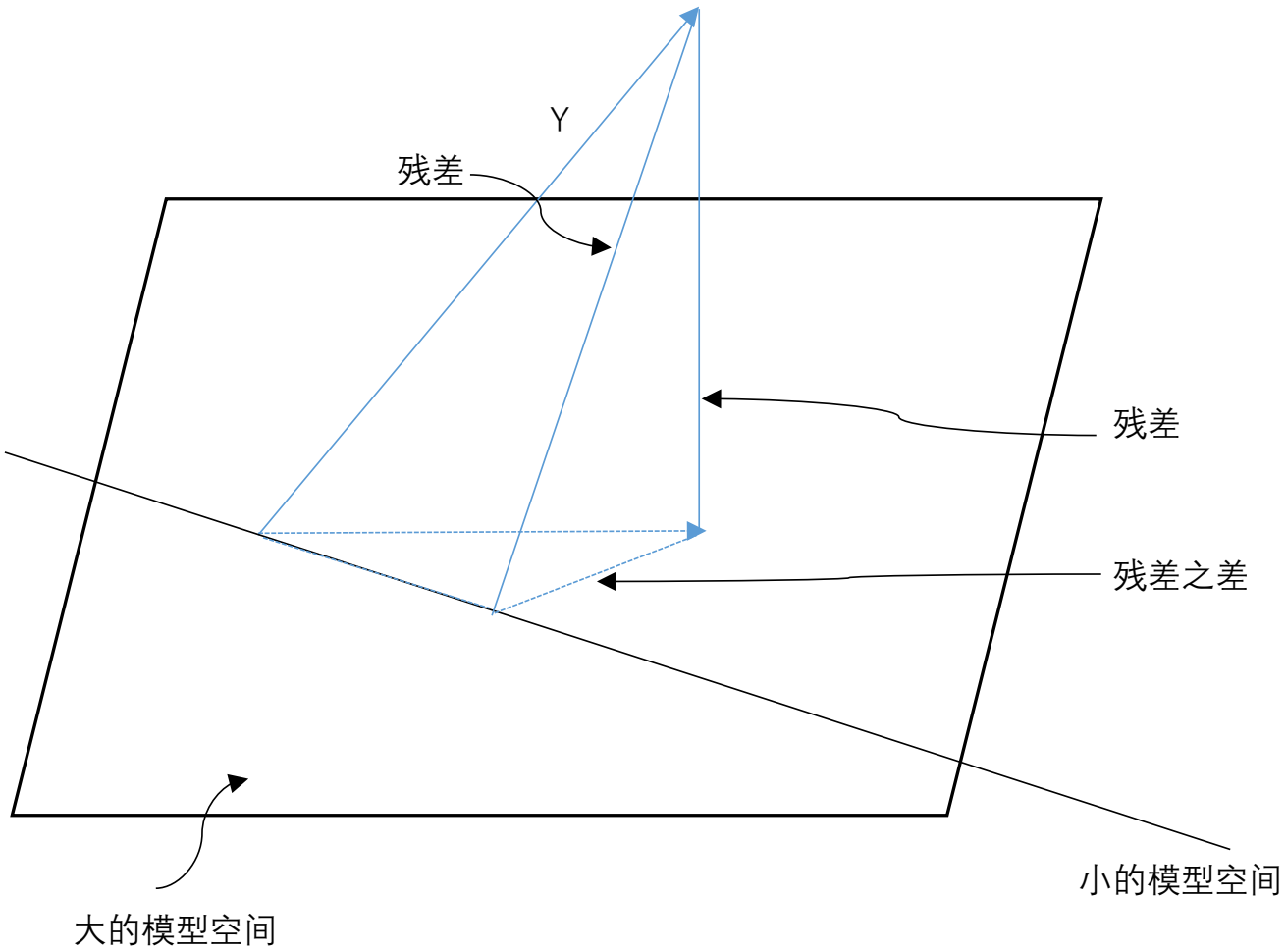
定理 (4.1.1, 最小二乘法基本定理)

对于正态线性回归模型(*4.1.1*),
(a) $RSS/\sigma^2 \sim \chi^2(n - rk(\boldsymbol{X}))$;
(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

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

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

注 (1) 统计量F_H的分子$(\mathrm{RSS}_H - \mathrm{RSS})/m$表示每增加一个约束, 残差平方和的平均变化量. 分母$\mathrm{RSS}/(n - \mathrm{rk}(\boldsymbol{X}))$起着正则化的作用, 用来消除$F_H$的分子$(\mathrm{RSS}_H - \mathrm{RSS})/m$的量纲. (3) 几何解释见下图.



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.

证明 (a). 已在定理3.2.4中给出.

(b). 由第三章, 我们已知道

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

即

$$\mathsf{RSS}_H=\mathsf{RSS}+(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H)'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H). \tag{4.1.7}$$

证明 (a). 已在定理3.2.4中给出.

(b). 由第三章, 我们已知道

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

即

$$\mathrm{RSS}_H=\mathrm{RSS}+(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H)'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H). \tag{4.1.7}$$

利用$\hat{\boldsymbol{\beta}}_H$的表达式(4.1.4)可得

$$\mathrm{RSS}_H-\mathrm{RSS}=(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{b})'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{b}). \tag{4.1.8}$$

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

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

若假设(4.1.2)成立, 则

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

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

若假设(4.1.2)成立, 则

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

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

(c). 因为

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

代入(4.1.8)得

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

其中

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

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

$$\mathsf{RSS}=\boldsymbol{e}'\boldsymbol{N}\boldsymbol{e}.$$

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

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

根据推论2.4.10和推论2.4.11, 结论成立.

其中

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

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

$$\mathrm{RSS}=\boldsymbol{e}'\boldsymbol{N}\boldsymbol{e}.$$

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

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

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

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

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

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

若$[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}$的计算比较容易, 那么可通过(4.1.8)直接计算$\mathrm{RSS}_H - \mathrm{RSS}$的大小. 但大部分情况下并非如此, 此时需分别计算RSS与RSS_H的大小.

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

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

若$[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}$的计算比较容易, 那么可通过(4.1.8)直接计算$\mathrm{RSS}_H - \mathrm{RSS}$的大小. 但大部分情况下并非如此, 此时需分别计算RSS与RSS_H的大小.

RSS与RSS_H的计算: RSS可通过下列公式计算

$$\mathrm{RSS} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}. \tag{4.1.9}$$

而计算RSS_H时可通过把约束条件$\boldsymbol{A\beta} = \boldsymbol{b}$“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

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

若要检验$\beta_1 = \beta_2 = \beta_3$, 此时线性假设的形式为$\boldsymbol{A}\boldsymbol{\beta} = \mathbf{0}$, 其中

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

$\boldsymbol{A}$的秩为2.

例如, 把模型(4.1.1)写成

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

若要检验$\beta_1 = \beta_2 = \beta_3$, 此时线性假设的形式为$\boldsymbol{A}\boldsymbol{\beta} = \mathbf{0}$, 其中

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

$\boldsymbol{A}$的秩为2. 将$\beta_1 = \beta_2 = \beta_3$融入原模型得约简模型

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

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$\widetilde{\boldsymbol{X}}$.

对约简模型应用最小二乘法得$\boldsymbol{\alpha}$的最小二乘估计

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

相应的残差平方和为

$$\mathsf{RSS}_H = (\boldsymbol{Y} - \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\alpha}})'(\boldsymbol{Y} - \widetilde{\boldsymbol{X}}\hat{\boldsymbol{\alpha}}) = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\alpha}}'\widetilde{\boldsymbol{X}}'\boldsymbol{Y}. \qquad (4.1.11)$$

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

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$\boldsymbol{e} = (e_1, e_2, e_3)' \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_3)$. 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

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

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

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

残差平方和为

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

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

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

残差平方和为

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

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

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

从而得α的LSE为

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

于是

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

然后可通过简单计算得到检验统计量

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

若H成立, 则

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

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$x_1, \cdots, x_p$有两批观测数据. 对第一批数据, 有线性回归模型

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

对第二批数据, 有线性回归模型

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

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

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

这个问题具有广泛的背景.

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相依关系在两个不同的时间段上是否有了变化.

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

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

将它们合并, 得原模型

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

注意到$\mathsf{rk}(\boldsymbol{X}) = 2(p+1)$, $n = n_1 + n_2$. 要检验的假设为

推导检验统计量: 把两个模型写成矩阵形式

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

将它们合并, 得原模型

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

注意到$\mathsf{rk}(\boldsymbol{X}) = 2(p+1)$, $n = n_1 + n_2$. 要检验的假设为

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

注意到$m = \mathsf{rk}\begin{pmatrix}\boldsymbol{I}_{p+1} & -\boldsymbol{I}_{p+1}\end{pmatrix} = p+1$.

从模型(4.1.12)得到的$\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$的LSE

从模型(4.1.12)得到的$\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$的LSE

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

相应的残差平方和为

从模型(4.1.12)得到的$\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$的LSE

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

相应的残差平方和为

$$\mathrm{RSS} = \boldsymbol{Y}_1{}'\boldsymbol{Y}_1 + \boldsymbol{Y}_2{}'\boldsymbol{Y}_2 - \hat{\boldsymbol{\beta}}_1'\boldsymbol{X}_1'\boldsymbol{Y}_1 - \hat{\boldsymbol{\beta}}_2'\boldsymbol{X}_2'\boldsymbol{Y}_2. \tag{4.1.14}$$

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

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

$\boldsymbol{\alpha}$的LSE为

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

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

$\boldsymbol{\alpha}$的LSE为

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

相应的残差平方和为

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

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

$\boldsymbol{\alpha}$的LSE为

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

相应的残差平方和为

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

结合(4.1.14)和(4.1.15)得

$$\mathrm{RSS}_H - \mathrm{RSS} \quad =$$

结合(4.1.14)和(4.1.15)得

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

因此检验统计量为

$$F_H =$$

结合(4.1.14)和(4.1.15)得

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

因此检验统计量为

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

若H成立, 则$F_H \sim F(p+1, n_1 + n_2 - 2p - 2)$.

结合(4.1.14)和(4.1.15)得

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

因此检验统计量为

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

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.

回归方程的显著性检验

将正态线性回归模型(4.1.1)写成

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

回归方程的显著性检验

将正态线性回归模型(4.1.1)写成

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

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相依关系. 即检验

$$H: \quad \beta_1 = \cdots = \beta_p = 0. \tag{4.2.2}$$

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, 所有自变量对因变量y的线性影响是可以忽略不计的.

假设(4.2.2)是线性假设$\boldsymbol{A\beta} = \boldsymbol{b}$的特例, 取

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{0} & \boldsymbol{I}_p \end{pmatrix}, \ \boldsymbol{b} = \mathbf{0},$$

这里的$\mathbf{0}$是p维的零列向量, $\mathrm{rk}(\boldsymbol{A}) = p$. 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. 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.

假设(4.2.2)是线性假设$\boldsymbol{A\beta}=\boldsymbol{b}$的特例, 取

$$\boldsymbol{A}=\begin{pmatrix}\mathbf{0} & \boldsymbol{I}_p\end{pmatrix},\ \boldsymbol{b}=\mathbf{0},$$

这里的$\mathbf{0}$是p维的零列向量, $\mathsf{rk}(\boldsymbol{A})=p$. 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. 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

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

易求β_0的LSE为$\beta_0^* =$

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

$$\mathrm{RSS}_H =$$

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

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

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$y_1, \cdots, y_n$的变动引起的.

对于原模型(4.2.1), 残差平方和

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

于是

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

它是由于在约简模型(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.

对于原模型(4.2.1), 残差平方和

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

于是

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

它是由于在约简模型(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. 所以根据定理4.1.1检验统计量为

$$F_H =$$

对于原模型(4.2.1), 残差平方和

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

于是

$$\mathsf{ESS} = \mathsf{TSS} - \mathsf{RSS} = \mathsf{RSS}_H - \mathsf{RSS} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - \beta_0^*\mathbf{1}_n'\boldsymbol{Y}. \tag{4.2.5}$$

它是由于在约简模型(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. 所以根据定理4.1.1检验统计量为

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

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的统计解释:

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

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS}.$$

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.

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

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS}.$$

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, 我们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.

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

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS}.$$

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, 我们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, 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线性相依作用, 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	$\mathsf{P}(F > f_H)$
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

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

$\mathsf{P}(F > f_H)$表示$\mathsf{P}(F(p, n-p-1) > f_H)$. 当

$$p = \mathsf{P}(F(p, n-p-1) > f_H) < \alpha \tag{4.2.7}$$

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, 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$y$的线性影响是不显著的. 这里可能有两种情况:

如果经过检验, 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);

如果经过检验, 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("***.txt")
x1=yx[,1]
x2=yx[,2]
x3=yx[,3]
y=yx[,4]
coal=data.frame(x1,x2,x3,y)
lm.sol=lm(y~x1+x2+x3, data=coal)
summary(lm.sol)
```

注 回归方程的F检验结果可在summary(lm.sol)中找到.

```
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
```

```
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值为$0.0002422 < 0.05$, 所以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(约简模型和原模型比较的)方差分析表:

```
lm.sol=lm(y~x1+x2+x3,data=coal)
lm.reduction=lm(y~1,data=coal)
anova(lm.reduction,lm.sol)
```

```
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2 + x3
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     11 34643
2      8  3487  3     31156 23.827 0.0002422 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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, 我们需要在大样本理论框架下完成假设检验, 即需要求出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.

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, 我们需要在大样本理论框架下完成假设检验, 即需要求出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相依关系, F 值越大, 相依关系越显著.

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, 我们需要在大样本理论框架下完成假设检验, 即需要求出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相依关系, F 值越大, 相依关系越显著.

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, 那么就有理由拒绝 “因变量与自变量整体无显著的相依关系” 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($n! = 12! = 479001600$, 太大了!):

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

因为这个比例值0.0005非常小, 所以有充分的理由拒绝原假设: $\beta_1 = \beta_2 = \beta_3 = 0$. (sample函数默认为不放回抽样)

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. 如果检验的结果是拒绝原假设, 这意味着接受“因变量y与$\{x_1, \cdots, x_p\}$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

$$H_i: \quad \beta_i = 0. \tag{4.3.1}$$

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. 如果检验的结果是拒绝原假设, 这意味着接受“因变量y与$\{x_1, \cdots, x_p\}$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

$$H_i: \quad \beta_i = 0. \tag{4.3.1}$$

(4.3.1)可等价地写成关于$\boldsymbol{\beta}$的线性假设$H: \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$, 其中$\boldsymbol{b} = 0$,

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

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

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

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

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

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

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

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

所以检验统计量为

$$F_H = \frac{(\mathrm{RSS}_H - \mathrm{RSS})/m}{\mathrm{RSS}/(n-p-1)} =$$

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

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

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

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

所以检验统计量为

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

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, 还有t检验.

除了F检验, 还有t检验. 根据F分布与t分布的关系, 可立刻得到检验统计量为

$$T_i = \hat{\beta}_i / \sqrt{\hat{\sigma}^2 c_{i+1,i+1}}.$$

除了F检验, 还有t检验. 根据F分布与t分布的关系, 可立刻得到检验统计量为

$$T_i = \hat{\beta}_i / \sqrt{\hat{\sigma}^2 c_{i+1,i+1}}.$$

或者通过如下途径得到这个检验统计量: 根据定理3.2.4知$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, 所以

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

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

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

所以

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

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

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

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-value进行统计决策. 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$H_i : \beta_i = 0$.

对于煤净化的例子:

```
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.

anova(lm.sol)可得到所有回归系数的F检验结果:

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 24531.1 24531.1 56.2819 6.914e-05 ***
x2         1  4371.1  4371.1 10.0287   0.01326 *
x3         1  2253.8  2253.8  5.1708   0.05257 .
Residuals  8  3486.9   435.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

若想得到回归系数的显著性检验的方差分析表, 仍可通过anova函数实现. 以β_1的检验为例:

```
lm.sol=lm(y~x1+x2+x3,data=coal)
lm.reduction=lm(y~x2+x3,data=coal)
anova(lm.reduction,lm.sol)
```

```
Analysis of Variance Table

Model 1: y ~ x2 + x3
Model 2: y ~ x1 + x2 + x3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1      9 28018.0
2      8  3486.9  1     24531 56.282 6.914e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

在没有模型的正态性假设情形下, 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是否对因变量有显著的相依关系:

```
> lm.sol=lm(y~x1+x2+x3,data=coal)
> lms=summary(lm.sol)
> lms$coef[3,]
   Estimate  Std. Error      t value     Pr(>|t|)
15.58333333  4.92082621  3.16681238  0.01325816
> nreps=4000
> tstats=numeric(nreps)
> set.seed(123)
> for (i in 1:nreps){
+   lm.sol=lm(y~x1+sample(x2)+x3,data=coal)
+   tstats[i]=summary(lm.sol)$coef[3,3]
+   }
> mean(abs(tstats)>abs(lms$coef[3,3]))
[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: $\beta_2 = 0$.

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. 接下来介绍在其它假设检验中的应用.

检验成对自变量. 例如, 对于煤净化数据, 我们想检验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

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

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

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

用anova函数来实现这一分析过程:

```
lm.sol=lm(y~x1+x2+x3,data=coal)
lm.reduction=lm(y~x1,data=coal)
anova(lm.reduction,lm.sol)
```

```
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2 + x3
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     10 10111.8
2      8  3486.9  2    6624.9 7.5998 0.01414 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由于p值< 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.

检验回归参数的子空间. 例如, 想检验x_2和x_3是否对因变量y具有相同程度的相依关系, 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

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2(x_{i2} + x_{i3}) + e_i, \ \ i = 1, \cdots, n$$

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

$$\left\{ F = \frac{\mathsf{RSS}_H - \mathsf{RSS}}{\mathsf{RSS}/(n-4)} > F_\alpha(1, n-4). \right\}$$

用anova函数来实现这一分析过程:

lm.sol=lm(y~x1+x2+x3,data=coal)
lm.reduction=lm(y~x1+I(x2+x3),data=coal) (I=Integrate)
anova(lm.reduction,lm.sol)

```
Analysis of Variance Table

Model 1: y ~ x1 + I(x2 + x3)
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      9 7890.7
2      8 3486.9  1    4403.8 10.104 0.01302 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由于p值< 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.

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

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i2} + 1 \times x_{i3} + e_i, \ \ i = 1, \cdots, n,$$

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

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

用anova函数来实现这一分析过程:

```
lm.sol=lm(y~x1+x2+x3,data=coal)
lm.reduction=lm(y~x1+x2+offset(1*x3),data=coal)
anova(lm.reduction,lm.sol)
```

```
Analysis of Variance Table

Model 1: y ~ x1 + x2 + offset(1 * x3)
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1      9 746334
2      8   3487  1    742847 1704.3 1.305e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由于p值< 0.05, 所以有充分的理由拒绝原假设$\beta_3 = 1$.

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, 现在将通过假设检验的方法来检验异常点.

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

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

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

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

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

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

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

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

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

称模型(4.5.2)和(4.5.3)为均值漂移线性回归模型. 我们的目的是要判别$(\boldsymbol{x}_j', y_j)$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. 来推导检验统计量.

引理（分块矩阵求逆公式）

设$\boldsymbol{A}$为非奇异的对称矩阵, 将其分块为

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

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

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

定理 (4.5.1)

对均值漂移线性回归模型(*4.5.3*), $\boldsymbol{\beta}$和η的最小二乘估计为

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

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

定理 (4.5.1)

对均值漂移线性回归模型(*4.5.3*), $\boldsymbol{\beta}$和η的最小二乘估计为

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

其中$\hat{\boldsymbol{\beta}}_{(j)}$为从非均值漂移线性回归模型(*4.5.1*)剔除第j组数据后得到的$\boldsymbol{\beta}$的最小二乘估计, h_{jj}为帽子矩阵$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$的第$j$个对角元, $\hat{e}_j$为从模型(*4.5.1*)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线性回归模型中, 回归系数$\boldsymbol{\beta}$的最小二乘估计恰好等于在模型(4.5.1)中剔除第j组数据后所获得的最小二乘估计.

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

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

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

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

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

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

再回忆公式(3.4.9), 可知定理成立.

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.

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$\boldsymbol{A}\binom{\boldsymbol{\beta}}{\eta}=\boldsymbol{b}$, 其中

$$\boldsymbol{b}=0, \boldsymbol{A}=(0,\cdots,0,1), m=\mathsf{rk}(\boldsymbol{A})=1.$$

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$\boldsymbol{A}\binom{\boldsymbol{\beta}}{\eta}=\boldsymbol{b}$, 其中

$$\boldsymbol{b}=0, \boldsymbol{A}=(0,\cdots,0,1), m=\mathsf{rk}(\boldsymbol{A})=1.$$

另外, 注意到在约束条件$\eta=0$下, 模型(4.5.3)退化到约简模型(4.5.1), 所以

$$\mathsf{RSS}_H=$$

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$\boldsymbol{A}\binom{\boldsymbol{\beta}}{\eta}=\boldsymbol{b}$, 其中

$$\boldsymbol{b}=0, \boldsymbol{A}=(0,\cdots,0,1), m=\mathsf{rk}(\boldsymbol{A})=1.$$

另外, 注意到在约束条件$\eta=0$下, 模型(4.5.3)退化到约简模型(4.5.1), 所以

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

而模型(4.5.3)的无约束残差平方和为

$$\mathsf{RSS}=$$

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

$$\boldsymbol{b}=0, \boldsymbol{A}=(0,\cdots,0,1), m=\mathsf{rk}(\boldsymbol{A})=1.$$

另外, 注意到在约束条件$\eta=0$下, 模型(4.5.3)退化到约简模型(4.5.1), 所以

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

而模型(4.5.3)的无约束残差平方和为

$$\mathsf{RSS}=\boldsymbol{Y}'\boldsymbol{Y}-\boldsymbol{\beta}^{*\prime}\boldsymbol{X}'\boldsymbol{Y}-\eta^{*}\boldsymbol{d}_j'\boldsymbol{Y}. \tag{4.5.4}$$

利用定理4.5.1得

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

RSS可进一步写成

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

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

由最小二乘法基本定理, 检验统计量为

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

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

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

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

$$f_j>F_\alpha(1,n-p-2), \tag{4.5.5}$$

则判定第j组数据$(\boldsymbol{x}_j',y_j)$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

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

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

例4.5.1(续例4.2.1) 来检验12组数据中是否有异常点.

```
yx=read.table("***.txt")
x1=yx[,1];x2=yx[,2];x3=yx[,3];y=yx[,4]
coal=data.frame(x1,x2,x3,y)
lm.sol=lm(y~x1+x2+x3,data=coal)
library(car)
outlierTest(lm.sol)
```

```
> lm.sol=lm(y~x1+x2+x3,data=coal)
> library(car)
> outlierTest(lm.sol)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferroni p
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.

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$\mathsf{P}(A_i) = 1 - \alpha$, 则

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

这个公式就是著名的Bonferroni不等式.

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$\mathsf{P}(A_i) = 1 - \alpha$, 则

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

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.

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.

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,

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

且假设$\{u_i, i \geq 1\}$为独立同分布随机变量序列, 服从$N(0, \sigma^2)$.

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,

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

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$\{e_i, i = 1, \cdots, n\}$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

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

该统计量的渐近分布是χ^2分布的线性组合, 比较复杂. 为了给出拒绝域, 先来看一下DW统计量的意义.

Durbin与Watson提出了一种D-W检验, 检验统计量为

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

该统计量的渐近分布是χ^2分布的线性组合, 比较复杂. 为了给出拒绝域, 先来看一下DW统计量的意义.

由于$\{e_i, i \geq 1\}$是不可观测的, 因此要考查$\{e_i, i \geq 1\}$间的相关性常用残差$\{\hat{e}_i, i \geq 1\}$来考察. 将$\{\hat{e}_1, \cdots, \hat{e}_{n-1}\}$与$\{\hat{e}_2, \cdots, \hat{e}_n\}$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

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

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

将它们代入r的表达式, 得

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

若$\{e_1, \cdots, e_n\}$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可看作是ρ的一个点估计. (把判断ρ是否等于零转化为判断r是否等于零)

容易看出, DW统计量与r之间有如下的近似关系:

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

容易看出, DW统计量与r之间有如下的近似关系:

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

由上式知: 当$r=-1$时, DW≈ 4; 当$r=1$时, DW≈ 0; 当$r=0$时, DW≈ 2. 因此当$|\text{DW}-2|$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$\{e_i, i = 1, \cdots, n\}$存在正相关;
- $d_U <$ DW $< 4 - d_U$, 认为$\{e_i, i = 1, \cdots, n\}$不相关;
- DW $> 4 - d_L$, 认为$\{e_i, i = 1, \cdots, n\}$存在负相关;
- $d_L <$ DW $< d_U$或者$4 - d_U <$ DW $< 4 - d_L$时, 对$\{e_i, i = 1, \cdots, n\}$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4.6.1 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, 见下表.

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

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

$$\mathsf{DW} = 0.6800.$$

取$\alpha = 0.05$, 查表得$d_L = 1.10$, $d_U = 1.37$, 现$\mathsf{DW} < d_L$, 这表明$\{e_i, i = 1, \cdots, n\}$存在正相关.

R程序:

```
yx=read.table("***.txt")
x=yx[,1]
y=yx[,2]
consumption=data.frame(x,y)
lm.sol=lm(y~x,data=consumption)
summary(lm.sol)
library(car)
durbinWatsonTest(lm.sol)
```

R程序:

```
yx=read.table("***.txt")
x=yx[,1]
y=yx[,2]
consumption=data.frame(x,y)
lm.sol=lm(y~x,data=consumption)
summary(lm.sol)
library(car)
durbinWatsonTest(lm.sol)
```

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

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$H:\ \boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$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$H:\ \boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$. 此时自然希望构造$\boldsymbol{a}_i'\boldsymbol{\beta}, i=1,\cdots,m$的置信区间或者置信椭球, 这里$\boldsymbol{A}=(\boldsymbol{a}_1,\cdots,\boldsymbol{a}_m)'$, $\mathsf{rk}(\boldsymbol{A})=m$.

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$H:\ \boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$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$H:\ \boldsymbol{A}\boldsymbol{\beta}=\boldsymbol{b}$. 此时自然希望构造$\boldsymbol{a}_i'\boldsymbol{\beta}, i=1,\cdots,m$的置信区间或者置信椭球, 这里$\boldsymbol{A}=(\boldsymbol{a}_1,\cdots,\boldsymbol{a}_m)'$, $\mathsf{rk}(\boldsymbol{A})=m$.

置信椭球:

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$H:\ \boldsymbol{A\beta}=\boldsymbol{b}$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$H:\ \boldsymbol{A\beta}=\boldsymbol{b}$. 此时自然希望构造$\boldsymbol{a}_i'\boldsymbol{\beta}, i=1,\cdots,m$的置信区间或者置信椭球, 这里$\boldsymbol{A}=(\boldsymbol{a}_1,\cdots,\boldsymbol{a}_m)'$, $\mathsf{rk}(\boldsymbol{A})=m$.

置信椭球:

考虑正态线性回归模型

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

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

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

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

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

另一方面, 由于

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

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

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

另一方面, 由于

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

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

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

因此对给定的置信水平$1-\alpha$, 有

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

定义

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

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

定义

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

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

$\boldsymbol{\beta}$的置信水平为$1-\alpha$的置信椭球为

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

定义

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

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

$\boldsymbol{\beta}$的置信水平为$1-\alpha$的置信椭球为

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

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

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

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

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

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

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

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

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

这是已知的结果(可从回归系数的显著性检验的分析中推得).

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

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

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

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

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

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

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

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

需指出的是, 对一个固定的$\boldsymbol{x}$, (4.7.7)给出了$\boldsymbol{x}'\boldsymbol{\beta}$的置信水平为$1-\alpha$的置信区间. 对多个$\boldsymbol{x}$, $\boldsymbol{x}'\boldsymbol{\beta}$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$\boldsymbol{x}_i'\boldsymbol{\beta}$求置信水平为$1-\alpha$的置信区间

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

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

对每个$\boldsymbol{x}_i'\boldsymbol{\beta}$求置信水平为$1-\alpha$的置信区间

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

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

设$E_i, i = 1, \cdots, m$为m个随机事件, $\mathsf{P}(E_i) = 1-\alpha_i, i = 1, \cdots, m$. 则

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

这就是前面出现过的Bonferroni不等式.

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

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

当m较大时, $1 - m\alpha$将变得很小.

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

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

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

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

则依据Bonferroni不等式知

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

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, 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: (1) (β_1, β_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的区间估计.

例4.7.1(续例4.2.1) 对于煤净化问题, 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: (1) (β_1, β_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的区间估计.

由(4.7.4)可知: 取

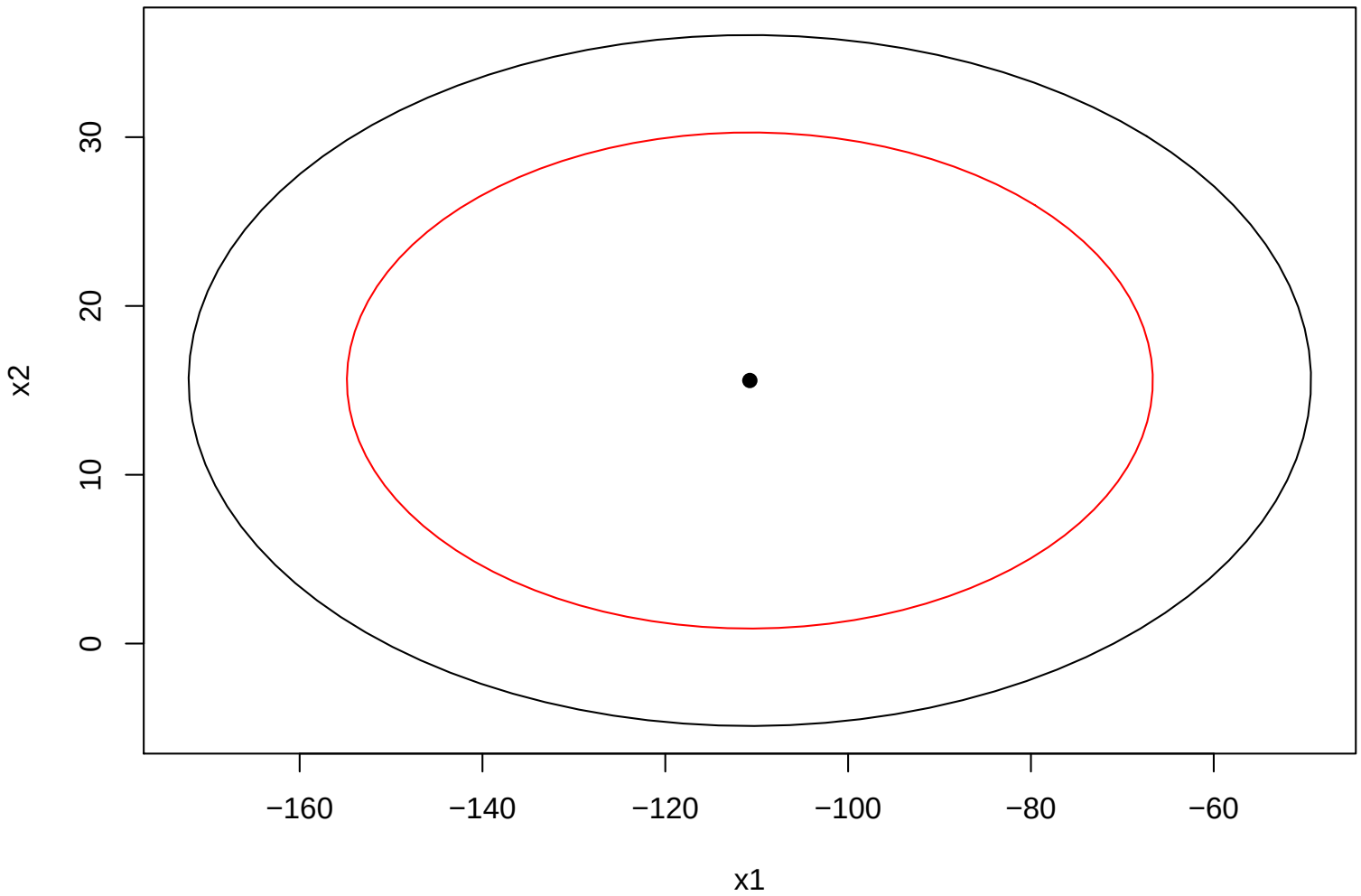
$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

(β_1, β_2)的置信椭球为

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

其中, $m = 2, p = 3, n = 12$, $\alpha = 0.05$或0.01.

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


x2
30
20
10
0
−160
−140
−120
−100
−80
−60
x1

```
> 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
>
```

在构造置信区间的过程中我们需要模型误差的正态性假设. Bootstrp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$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}$已通过最小二乘法得到$\hat{\boldsymbol{\beta}}$和残差向量$\hat{\boldsymbol{e}}=(\hat{e}_1,\cdots,\hat{e}_n)'$, 然后采取下列的操作:

步骤1: 从$\hat{e}_1,\cdots,\hat{e}_n$中通过有放回抽样随机取出一个残差样本$\boldsymbol{e}^*=(\hat{e}_1^*,\cdots,\hat{e}_n^*)'$;
步骤2: 构造$\boldsymbol{Y}^*=\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{e}^*=\hat{\boldsymbol{Y}}+\boldsymbol{e}^*$;
步骤3: 从$(\boldsymbol{X},\boldsymbol{Y}^*)$计算出$\hat{\boldsymbol{\beta}}^*=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^*$.

重复以上步骤B次(称B为Bootstrap次数), 则可得到B个$\hat{\boldsymbol{\beta}}^*$, 画出其密度图.

(1) 若图形看起来是与正态密度图像相似, 则用$(\hat{\beta}_i \pm 2\mathsf{sd}(\hat{\beta}_i^*))$作为$\beta_i$的95%置信区间, 其中$\mathsf{sd}(\hat{\beta}_i^*)$ 表示B个$\hat{\beta}_i^*$的样本标准差;

(1) 若图形看起来是与正态密度图像相似, 则用$(\hat{\beta}_i \pm 2\mathsf{sd}(\hat{\beta}_i^*))$作为$\beta_i$的95%置信区间, 其中$\mathsf{sd}(\hat{\beta}_i^*)$ 表示B个$\hat{\beta}_i^*$的样本标准差;

(2) 若图形看起来是与正态密度图像相去甚远, 则计算B个$\hat{\beta}_i^*$的样本分位数作为置信区间.

(1) 若图形看起来是与正态密度图像相似, 则用$(\hat{\beta}_i \pm 2\mathsf{sd}(\hat{\beta}_i^*))$作为$\beta_i$的95%置信区间, 其中$\mathsf{sd}(\hat{\beta}_i^*)$ 表示B个$\hat{\beta}_i^*$的样本标准差;

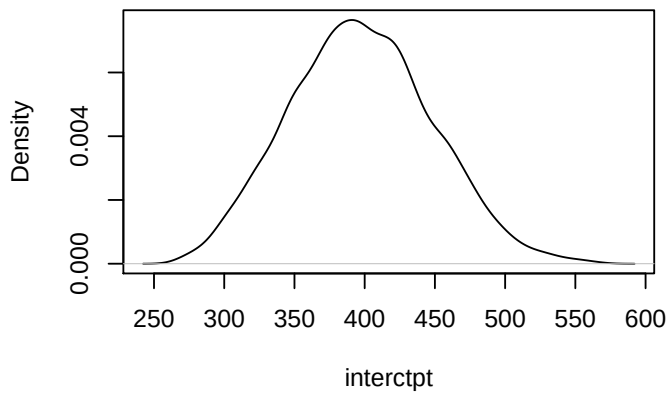
(2) 若图形看起来是与正态密度图像相去甚远, 则计算B个$\hat{\beta}_i^*$的样本分位数作为置信区间.

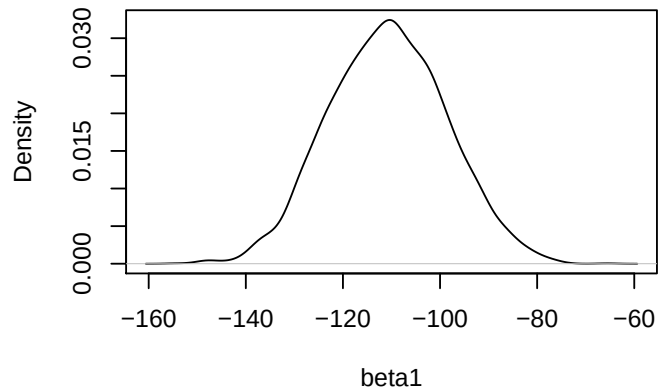
由此可见, 用Bootstrp方法构造置信区间除了无需考虑正态性假设以外, 也无需从理论角度计算置信区间的数学表达式.

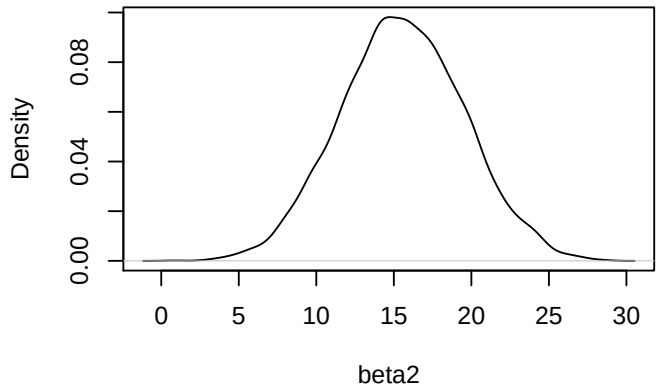
仍以煤净化问题为例, 来构造置信水平为0.95和0.99的Bootstrap置信区间.

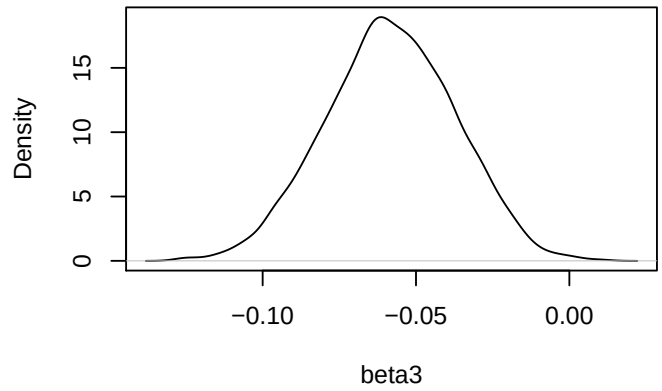
```
set.seed(123)
nb=4000
coefmat=matrix(NA,nb,4)
resids=residuals(lm.sol)
preds=fitted(lm.sol)
for (i in 1:nb){
booty=preds+sample(resids,rep=T)
bmod=update(lm.sol,booty~.)
coefmat[i,]=coef(bmod)
}
colnames(coefmat)=c("Intercept",colnames(coal[,1:3]))
coefmat=data.frame(coefmat)
```

```
plot(density(coefmat[,1]),xlab="Interctpt_boot",main="")
plot(density(coefmat[,2]),xlab="beta1_boot",main="")
plot(density(coefmat[,3]),xlab="beta2_boot",main="")
plot(density(coefmat[,4]),xlab="beta3_boot",main="")
```


Density
0.000
0.004
250
300
350
400
450
500
550
600
interctpt


Density
0.000
0.015
0.030
−160
−140
−120
−100
−80
−60
beta1


Density
0.00
0.04
0.08
0
5
10
15
20
25
30
beta2


Density
0
5
10
15
−0.10
−0.05
0.00
beta3

都与正态密度图像相似

构造置信水平为0.95的Bootstrap置信区间:

sd_est=apply(coefmat,2,function(x) sd(x))
coef(lm.sol)[1]+c(-2,2)*sd_est[1]
coef(lm.sol)[2]+c(-2,2)*sd_est[2]
coef(lm.sol)[3]+c(-2,2)*sd_est[3]
coef(lm.sol)[4]+c(-2,2)*sd_est[4]

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

若与正态密度图像不相似, 则采用样本分位数构造Bootstrap置信区间:

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%   301.8739 -133.96170  7.865427 -0.09945016
97.5%  498.9897  -86.84628 23.454765 -0.01823595
> apply(coefmat,2,function(x) quantile(x,c(0.005,0.995)))
      Intercept          x1        x2         x3
0.5%   283.6925 -140.42083  5.582808 -0.1127245
99.5%  532.2880  -80.92097 25.691882 -0.0054039
```

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, 我们在有了回归方程后, 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

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i, \quad i = 1, \cdots, n, \tag{4.8.1}$$

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

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

给定

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

对应的因变量为y_0, 它可表示为

$$y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + e_0, \tag{4.8.2}$$

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

点预测:

点预测:

注意到y_0由两部分组成: $\boldsymbol{x}_0'\boldsymbol{\beta}$和$e_0$. 自然地, 可以用$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$去估计$\boldsymbol{x}_0'\boldsymbol{\beta}$, 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

$$\hat{y}_0 =$$

点预测:

注意到y_0由两部分组成: $\boldsymbol{x}_0'\boldsymbol{\beta}$和$e_0$. 自然地, 可以用$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$去估计$\boldsymbol{x}_0'\boldsymbol{\beta}$, 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

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + 0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}. \tag{4.8.3}$$

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.

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$\mathsf{E}(\hat{y}_0) = \mathsf{E}(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0'\boldsymbol{\beta} = \mathsf{E}(y_0)$.

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$\mathsf{E}(\hat{y}_0) = \mathsf{E}(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0'\boldsymbol{\beta} = \mathsf{E}(y_0)$.

(2) 在y_0的一切线性无偏预测中, $\hat{y}_0$具有最小方差.

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$\mathsf{E}(\hat{y}_0) = \mathsf{E}(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0'\boldsymbol{\beta} = \mathsf{E}(y_0)$.

(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$\boldsymbol{a}'\boldsymbol{Y}$是$y_0$的某一线性无偏预测, 则$\mathsf{E}(\boldsymbol{a}'\boldsymbol{Y}) = \mathsf{E}(y_0) = \boldsymbol{x}_0'\boldsymbol{\beta}$. 因此$\boldsymbol{a}'\boldsymbol{Y}$可看作是$\boldsymbol{x}_0'\boldsymbol{\beta}$的一个线性无偏估计. 而预测$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$也可以看作是$\boldsymbol{x}_0'\boldsymbol{\beta}$的一个线性无偏估计. 由Gauss-Markov定理知$\mathsf{Var}(\boldsymbol{a}'\boldsymbol{Y}) \geq \mathsf{Var}(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) = \mathsf{Var}(\hat{y}_0)$.

值得注意的是, 虽然从形式上讲, y_0的点预测$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$与参数函数$\mu_0 = \boldsymbol{x}_0'\boldsymbol{\beta}$的最小二乘估计$\hat{\mu}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$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.

值得注意的是, 虽然从形式上讲, y_0的点预测$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$与参数函数$\mu_0 = \boldsymbol{x}_0'\boldsymbol{\beta}$的最小二乘估计$\hat{\mu}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$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$e_1, \cdots, e_n$不相关, 所以$y_0$与$\hat{\boldsymbol{\beta}}$也不相关, 因此

$$\mathsf{Var}(d_1) =$$

值得注意的是, 虽然从形式上讲, y_0的点预测$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$与参数函数$\mu_0 = \boldsymbol{x}_0'\boldsymbol{\beta}$的最小二乘估计$\hat{\mu}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$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$e_1, \cdots, e_n$不相关, 所以$y_0$与$\hat{\boldsymbol{\beta}}$也不相关, 因此

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

而

$$\mathsf{Var}(d_2) =$$

值得注意的是, 虽然从形式上讲, y_0的点预测$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$与参数函数$\mu_0 = \boldsymbol{x}_0'\boldsymbol{\beta}$的最小二乘估计$\hat{\mu}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$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$e_1, \cdots, e_n$不相关, 所以$y_0$与$\hat{\boldsymbol{\beta}}$也不相关, 因此

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

而

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

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

区间预测:

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.

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. 假设模型误差还满足服从正态分布, 即

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

这时可知

$$y_0 - \hat{y}_0 \sim$$

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. 假设模型误差还满足服从正态分布, 即

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

这时可知

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

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

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

所以根据

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

可推得

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

因此对于给定的α, 有

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

由此可得到y_0的概率为$1-\alpha$的预测区间为

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

例4.8.1 考虑一元线性回归模型

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

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$(\boldsymbol{x}_0 = (1, x_0)')$

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

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

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

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

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

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

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)
predict(lm.sol,new,interval="prediction",level=0.95)
```

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

	fit	lwr	upr
1	271.183	215.7633	326.6027

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

	fit	lwr	upr
1	271.183	215.7633	326.6027

注 "level=0.95"是默认参数, 可不写; 把interval="prediction"改成interval="confidence"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:

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;

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;

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$\hat{\sigma}$;

(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).