第6章　线性回归模型(二)
回归诊断

吴密霞

北京工业大学统计与数据科学系

E-mail:　*wumixia@bjut.edu.cn*

教材

- 吴密霞, 王松桂. 2024. 线性模型引论(第2版), 科学出版社.

回归诊断的内容目录

- 残差分析
  - 正态性诊断及处理
  - 方差等方差性诊断及处理
  - 序列相关性诊断及处理
- 影响分析
  - 高杠杆点诊断及处理
  - 异常值诊断及处理
  - 强影响点诊断及处理
- 复共线性
  - 复共线性的诊断
  - 复共线性的处理：岭估计，主成分估计

回归诊断

回归诊断：考察模型事先假定条件是否成立, 并给出处理办法

- 模型是线性的

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1} + e_i, \quad i = 1, 2, \cdots, n.$$

- 误差项满足不相关、等方差性、正态性
  - Gauss-Markov假设: $E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathbf{I}_n$
  - 参数的区间估计和假设检验$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$.
- 自变量是非随机、无测量误差、线性无关的
- 所有观测$(y_i, x_{i1}, \cdots, x_{ip-1})$是来自同一总体、且同等重要

回归诊断

- 残差分析
  - 与自变量的线性关系：
  - 误差正态性: $\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$
  - 误差方差齐性: $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathbf{I}_n$
  - 误差序列相关性$\mathrm{Cov}(e_i, e_{i+k}) \neq 0$
- 影响分析
  - 高杠杆点诊断
  - 异常值诊断
  - 强影响点诊断
- 自变量的复共线性

残差分析

残差分析：借助于残差来考察误差假设条件的合理性.

线性回归模型

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_n.$$

三种常用的残差：

(1) 普通残差: 即普通最小二乘残差

$$\widehat{\boldsymbol{e}} = (\hat{e}_1, \cdots, \hat{e}_n)' = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{y} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{y},$$

其中$\hat{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, 这里$\boldsymbol{x}_i$为$\mathbf{X}$的$X$的第$i$个行向量.

因$\mathbf{H}$将$\boldsymbol{y}$变成其 “戴帽子"的拟合值$\widehat{\boldsymbol{y}} = \mathbf{H}\boldsymbol{y} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ 故称$\mathbf{H}$ 为帽子矩阵

定理6.4.1

在Gauss-Markov假设下, 残差有如下结论.

(1) 若$E(\widehat{\boldsymbol{e}}) = \mathbf{0}$, 则$\mathrm{Cov}(\widehat{\boldsymbol{e}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$;

(2) $\mathrm{Cov}(\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{e}}) = \mathbf{0}$, 故残差$\widehat{\boldsymbol{e}}$ 与拟合值$\widehat{\boldsymbol{y}}$ 不相关;

(3) $\mathbf{1}_n'\widehat{\boldsymbol{e}} = \mathbf{0}$;

(4) 若$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, 则$\widehat{\boldsymbol{e}} \sim N(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$ 且与$\widehat{\boldsymbol{y}}$独立.

- 即使误差等方差不相关$\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathbf{I}_n$, 残差也相关且不等方差

$$\mathrm{Var}(\widehat{e}_i) = \sigma^2(1 - h_{ii}), \quad \mathrm{Cov}(\widehat{e}_i, \widehat{e}_j) = \sigma^2 h_{ij} \ \ (i \neq j)$$

其中h_{ii}为帽子矩阵$\mathbf{H}$ 的第i个对角元素.

标准化残差

- 由于普通残差$\widehat{e}_i$的方差$\sigma^2(1-h_{ii})$与h_{ii}有关, 彼此不相等, 因此, 不能直接通过比较残差$\widehat{e}_i$来诊断模型误差是否等方差.
- 对其进行标准化, 并用$\widehat{\sigma}$替换σ, 得

(2) 标准化残差

$$r_i = \frac{\widehat{e}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}}, \quad i = 1, \cdots, n,$$

其中

$$\widehat{\sigma} = (\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2/(n-p))^{1/2}.$$

也称r_i为内学生化残差(internally standardized residual).

定理6.4.2

若$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, 则

(1) $r_i^2/(n-p)$ 服从参数为$1/2$和$(n-p-1)/2$的Beta分布，记作

$$\frac{r_i^2}{n-p} \sim \text{Beta}\left(\frac{1}{2}, \frac{n-p-1}{2}\right);$$

(2) $E(r_i) = 0$, $\text{Var}(r_i) = 1$, 且

$$\text{Cov}(r_i, r_j) = -\frac{h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}, \quad i \neq j.$$

- 因为$\text{Cov}(r_i, r_j)$ 一般都很小，所以可近似地认为r_i 和r_j不相关；当n较大时, 可以近似地认为r_i相互独立且服从$N(0,1)$.

标准化预测残差

(3) 标准化预测残差

$$r_i^* = \frac{\sqrt{1-h_{ii}}e_i^*}{\widehat{\sigma}_{(i)}} = \frac{\widehat{e}_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \tag{6.17}$$

也称外学生化残差(externally studentized residual).

其中, $e_i^* = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(i)}$ 为$\boldsymbol{x}_i$处的预测残差,

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}'\boldsymbol{y}_{(i)}, \quad \widehat{\sigma}_{(i)}^2 = \frac{\|\boldsymbol{y}_{(i)} - \mathbf{X}_{(i)}\widehat{\boldsymbol{\beta}}_{(i)}\|^2}{(n-p-1)},$$

这里, $\boldsymbol{y}_{(i)}$和$\mathbf{X}_{(i)}$ 分别为剔除第i个观测后相应的$(n-1)\times 1$因变量观测向量和$(n-1)\times p$设计阵.

- $\widehat{\boldsymbol{\beta}}_{(i)}$ 和$\widehat{\sigma}_{(i)}^2$ 为模型$\boldsymbol{y}_{(i)} = \mathbf{X}_{(i)}\boldsymbol{\beta}_{(i)} + \boldsymbol{e}_{(i)}$下$\boldsymbol{\beta}$和方差$\sigma^2$的LS估计.

定理6.4.3

标准化预测残差r_i^*和标准化残差r_i有如下关系：

$$r_i^* = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}};$$

(2) 如果$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, 则

$$r_i^* \sim t_{n-p-1}.$$

可以分别采用R语言中的函数计算回归模型的残差：

- 普通残差：函数residuals()
- 标准化学生化(内学生化)：函数rstandard()
- 外学生化残差：函数rstudent()

残差图

所谓残差图就是以某种残差为纵坐标, 以任何其他的量为横坐标的散点图.

横坐标有多种选择，其中最常见的三种选择:

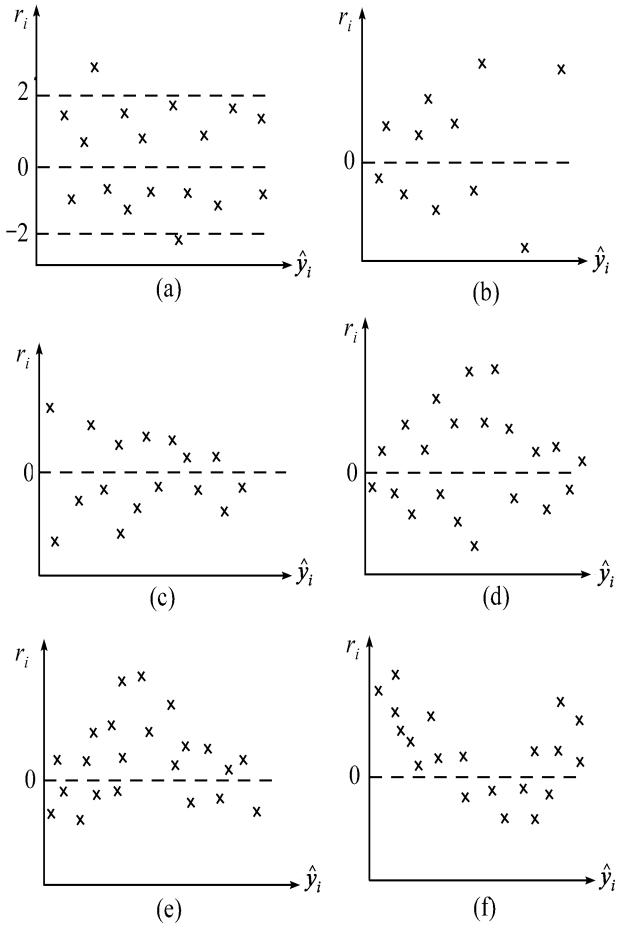
- 因变量Y的拟合值(fitted value)，也称预测值(predicted value)；
- 自变量x_j，$j = 1, \cdots, p-1$；
- 因变量的观测时间或观测序号（时间序列情形）.

这三类残差图分别侧重于考察随机误差的正态性和等方差性；单变量的线性性；误差不相关性等假定.

残差图

- 标准化残差与因变量拟合值的散点图:

- 正态性　若$\boldsymbol{e} \sim N_n(\mathbf{0},\ \sigma^2\mathbf{I}_n)$, 则当$n$较大时, 可近似地认为$r_i$为总体$N(0,1)$的一组简单随机样本. 故大约应有95%的点$(\hat{y}_i, r_i)$, $i = 1, \cdots, n$, 落在水平带$|r_i| \leq 2$区域内, 且不呈任何的趋势, 如图(a)

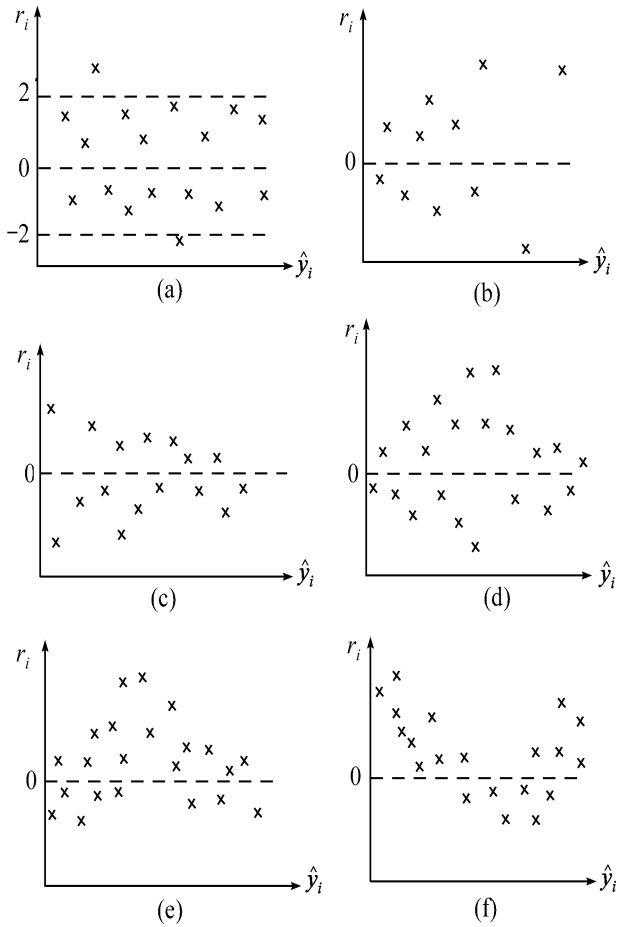


残差图

- 标准化残差与因变量拟合值的散点图:

- 等方差性　当误差e_i的方差不全相等时，残差图将大致关于直线$r = 0$呈现上下对称且有一定的趋势.

  图(b):误差随$\hat{y}_i$的增大有增加的趋势；图(c)：误差方差随$\hat{y}_i$的增大而减小；图(d):对较大或较小的$\hat{y}_i$，误差方差偏小，而对中等大小的$\hat{y}_i$，误差方差偏大.

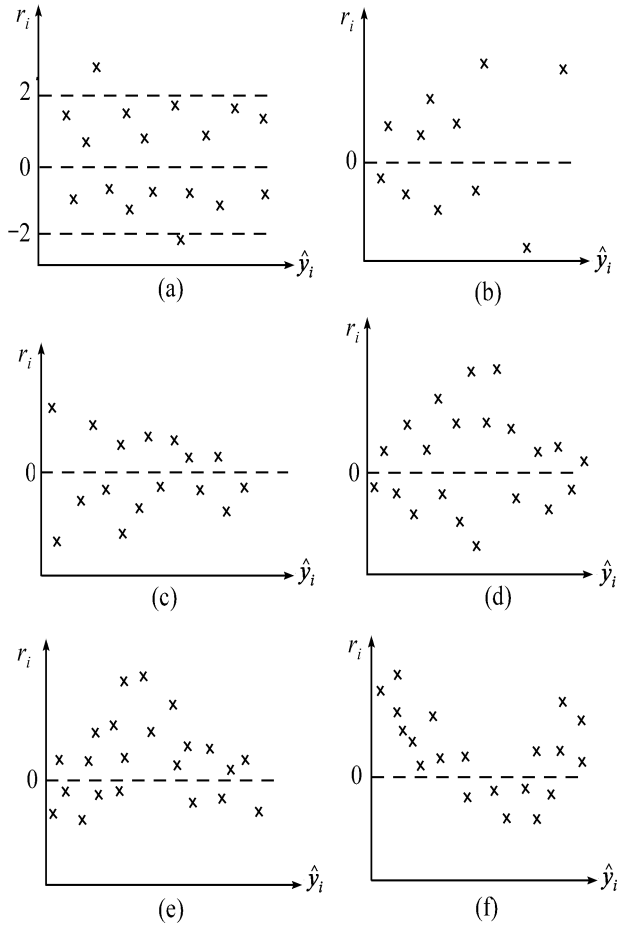


残差图

- 标准化残差与因变量拟合值的散点图:即$(r_i, \hat{y}_i)$散点图

- 线性性　当残差图呈现一定趋势时，图(e) 和(f)表明回归函数可能是非线性的, 或漏掉了一个或多个重要的回归自变量的二次项或交叉项, 或误差e_i之间有一定相关性.

  究竟属于何种情况, 还需作进一步的诊断.



残差图

- 以自变量为横坐标的残差图： 即$(x_{ij}, r_i), (1 \leq i \leq n)$ 散点图

在正态假设下，标准化残差与每个自变量都是不相关的. 若该假设成立，图中的点应随机散布, 若散点图有任何可辨别的模式都表明这些假设可能不成立.

- 如果呈现如图(a)的水平带状, 则说明关于自变量X_j与Y的线性假设合理;
- 如果呈现图(b)或(c)或(d)的形状，则说明误差方差不等方差，且受协变量X_j的影响;
- 如果呈现图(e)或(f)的形状，则需要在模型中增加自变量X_j的高次项，或需要对因变量Y作变换以得到线性关系.

残差图

- 以时间为横坐标的残差图：

假设$y_1, \cdots, y_n$是因变量Y分别是在$t_1, \cdots, t_n$时刻的观测值, 则可取时间t或观察序号为横坐标, 构造(t_i, r_i)或(i, r_i)的残差图, 用来检查误差的独立性假设.

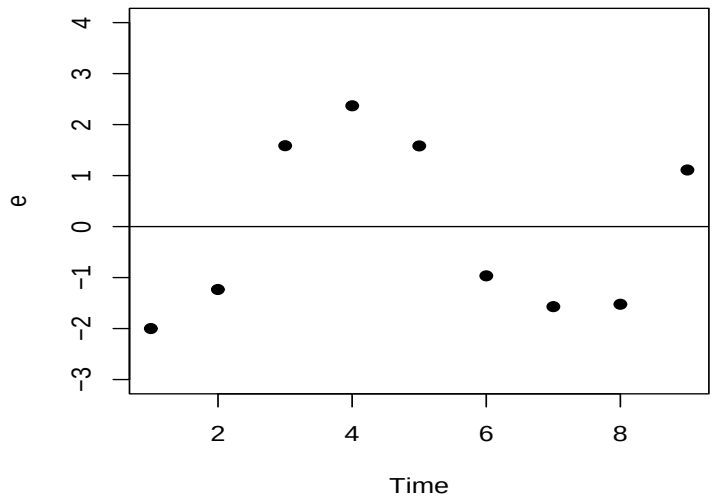
- 当误差是独立的, 这些点应该随机分布在一条过零的水平带状区域内, 反之，则会呈现一定趋势.
- 在经济、商业问题中, 误差e_i往往自相关, 如一阶自相关：

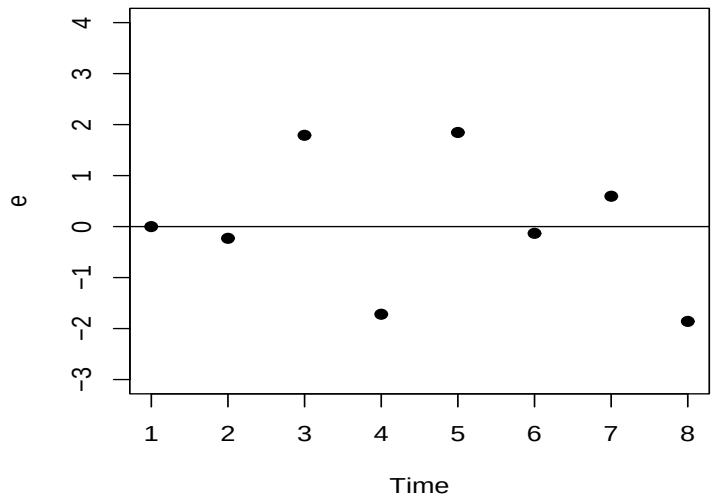
$$e_i = \rho e_i + u_i,$$

这里$u_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $\rho(|\rho| < 0)$ 是自相关参数.

残差图

- 当$\rho > 0$, 即正相关时, 残差符号具有“集团”性，即有一段全是正号，另一段全是负号，残差符号的改变不频繁，如左图.
- 当$\rho < 0$, 即负相关时, 残差符号大致有正负交错的趋势，符号改变非常频繁, 如图右.





残差图的R程序

```
plot(x, which = 1:6, caption = c("Residuals vs Fitted",
"Normal Q-Q plot", "Scale-Location plot", "Cook's distance plot")
panel = points, sub.caption = deparse(x$call), main = "",
ask = prod(par("mfcol"))<length(which)&&dev.interactive(),
id.n = 3, label.id = names(residuals(x)), cex.id = 0.75)
```

其中x是由lm生成的对象，which是1-6的全部或某个子集，1表示绘制普通残差与拟合值的残差图；2表示绘制残差的QQ图；3表示绘制标准化残差绝对值的开方与拟合值的残差图；4表示绘制Cook 距离图；5表示标准化残差对杠杆图；6表示Cook距离对杠杆图

正态性诊断

- 残差的Q-Q图

即$\{(q_{(i)}, \widehat{e}_{(i)})$散点图, 其中$\widehat{e}_{(i)}$表示残差$\widehat{e}_i$的次序统计量,

$$q_{(i)} = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right), \qquad i = 1, \cdots, n,$$

这里$\Phi(x)$为标准正态$N(0,1)$ 的分布函数, $\Phi^{-1}(x)$为其反函数.

若$e_i (i = 1, \cdots, n)$服从正态分布$N(0, \sigma^2)$，则点$\{(q_{(i)}, \widehat{e}_{(i)}), i = 1, \cdots, n\}$ 应在一条直线附近. 因此，若残差Q-Q图中点的大致趋势明显不在一条直线上, 则有理由怀疑对误差的正态性假设的合理性.

R语言中，可用函数plot(model, 2)绘制残差的Q-Q图

正态性诊断

- Shapiro-Wilk 检验

  又称W检验, 是由Shapiro和Wilk(1965)基于回归和相关提出的.

  记$\boldsymbol{u} = (u_1, \cdots, u_n)'$ 是随机变量U的顺序样本的观测向量.

  如果$U \sim N(\mu, \sigma^2)$, 则

$$u_i = \mu + \sigma m_i + \varepsilon_i, \quad i = 1, \cdots, n, \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 v_{ij},$$

  这里$\boldsymbol{m} = (m_1, \cdots, m_n)'$和$\mathbf{V} = (v_{ij})$分别为n个标准正态随机变量的顺序统计量的期望向量和协方差阵.

  - 对称分布下, $\boldsymbol{m}'\mathbf{V}^{-1}\mathbf{1}_n = 0$, 该模型下$\sigma$的BLU估计为

$$\widehat{\sigma} = \boldsymbol{m}'\mathbf{V}^{-1}\boldsymbol{u}/(\boldsymbol{m}'\mathbf{V}^{-1}\boldsymbol{m}).$$

正态性诊断

Shapiro-Wilk 检验

就是检验回归方程$u_i = \mu + \sigma m_i + \varepsilon_i$的显著性, 检验统计量

$$W = \frac{(\boldsymbol{m}'\mathbf{V}^{-1}\boldsymbol{m})^2\hat{\sigma}^2}{(\boldsymbol{m}'\mathbf{V}^{-2}\boldsymbol{m})S^2} = \frac{(\sum_i a_i u_i)^2}{\sum_i (u_i - \bar{u})^2}$$

这里$\bar{u} = \sum u_i/n$, $\boldsymbol{a} = (a_1, \cdots, a_n)' = \mathbf{V}^{-1}\boldsymbol{m}/\sqrt{(\boldsymbol{m}'\mathbf{V}^{-2}\boldsymbol{m})}$, $S^2 = \sum_{i=1}^2 (u_i - \bar{u})^2/(n-1)$. 可证$\boldsymbol{a}'\boldsymbol{a} = 1$, $W \leq 1$. 计算出的W值越接近1，样本服从正态分布的可能性越大.

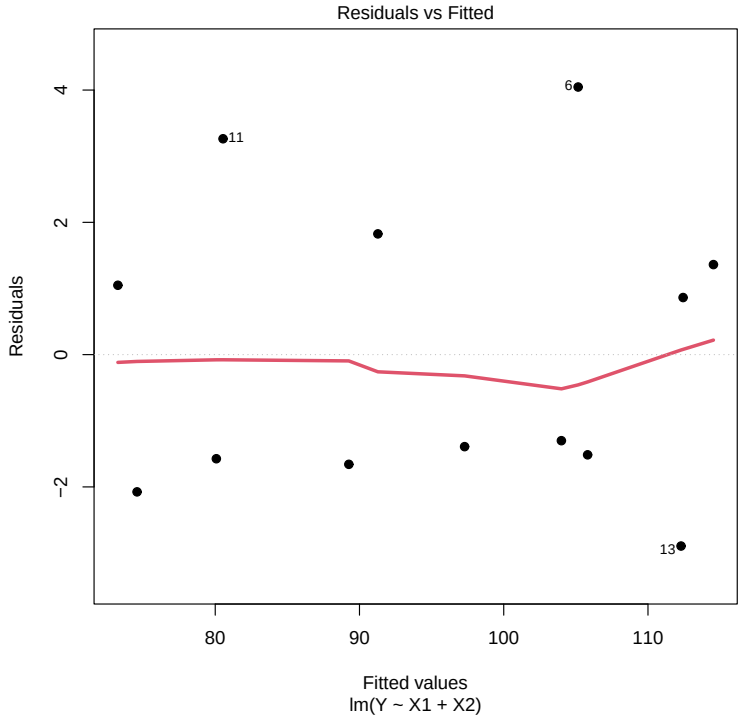
- R软件中函数shapiro.test() 提供W统计量和相应的p值.
- 误差向量$\boldsymbol{e}$不可被直接观测到, 借助于LS残差向量$\hat{\boldsymbol{e}}$作检验
- Shapiro-Wilk 检验适用于小样本$n \geq 7$

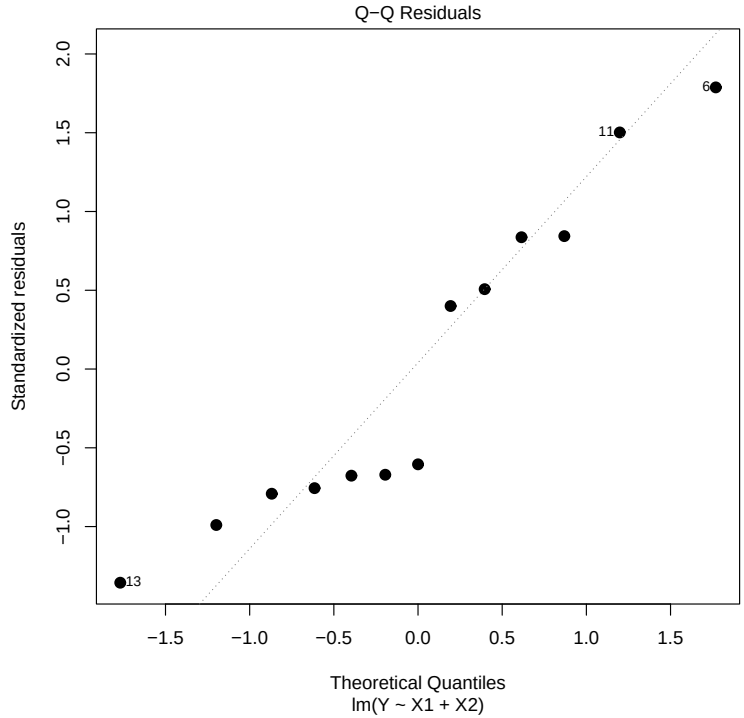
Hald水泥数据例子

例Hald水泥数据的BIC变量选择的结果：采用模型$Y = \beta_0 + X_1\beta_1 + + X_2\beta_2 + e$. 本例对模型的随机误差进行正态性检验.

采用R语言中的函数plot()绘制残差图和残差的Q-Q图.

```
lm.reg <- lm(Y ~ X1+X2, data = data1)  # 拟合BIC选模型
plot(lm.reg,1)   # 绘制残差图
plot(lm.reg, 2)  # 绘制Q-Q图
```





- 残差图: 模型的残差大都在[-2,2]之间，且与Y的拟合值没有明显的相依趋势;
- Q-Q图: 也可以看出点$\{(q_{(i)}, \widehat{e}_{(i)}), i = 1, \cdots, 13\}$ 大致在一条直线上附近, 因此，可以初步判定模型误差服从正态分布.

采用R语言中函数shapiro.test()对残差进行Shapiro-Wilk正态检验.

```
shapiro.test(y.res) ## 正态检验
Shapiro-Wilk normality test
data: y.res
W = 0.90527, p-value = 0.158
```

由于Shapiro-Wilk统计量W 较接近于1, 其p-值=$0.158 > 0.05$, 因此，接受模型误差的正态假设.

异方差性诊断

对于线性回归模型

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma},$$

其误差向量$\boldsymbol{e}$的协方差阵具有如下形式：

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix},$$

- 若$\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$不全相等, 则称模型为异方差线性回归模型.

异方差的影响

- 误差异方差时, LS估计仍是无偏的, 但不再有效.
- 基于LS估计的回归系数的区间估计、假设检验以及预测精度等都会受到影响.
- 一个有效的方法：加权最小二乘(weighted LS, WLS)估计

$$\beta^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{y},$$

其中$\mathbf{W} = \mathrm{Diag}(w_1, \cdots, w_n)$为权矩阵. 当

$$w_i = 1/\sigma_i^2, \quad i = 1, 2, \cdots, n$$

且已知时, 可证明$\boldsymbol{c}'\beta^*$为$\boldsymbol{c}'\beta$的BLUE.

异方差性产生的原因

(1) 模型中遗漏了某些重要的自变量.

假设正确的模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i,$$

如果遗漏了自变量x_{i2}, 则指定的模型为

$$y_i \beta_0 + \beta_1 x_{i1} + e_i,$$

此时y_i的方差自然与x_{i2}有关.

记$\boldsymbol{x}_j = (x_{1j}, \cdots, x_{nj})', j = 1, 2.$ 指定模型下的残差为

$$\widehat{\boldsymbol{e}} = (\widehat{e}_1, \cdots, \widehat{e}_n)' = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1_n}:\boldsymbol{x}_1})\boldsymbol{y} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{1_n}:\boldsymbol{x}_1})(\boldsymbol{x}_2\beta_2 + \boldsymbol{e}),$$

故$\widehat{e}_i$会呈现出随着x_{i2}的变化而变化的趋势

异方差性产生的原因

(2) 数据的测量误差.

样本数据的观测误差有可能随研究范围的扩大而增加，或随时间的推移逐步积累，也可能随着观测技术的提高而逐步减小.

如道格拉斯生产函数的对数模型：

$$y_t = \log A + \alpha_t \log K_t + \log L_t + \varepsilon_t,$$

其中y_t: 第t时刻某企业生产能力的对数，K_t: 第t时刻该企业的资本，L_t: 第t时刻该企业的劳动力，ε_t: 第t时刻除资本和劳动力的其他因素.

由于不同时期的观测技术、评价标准不同致使企业的投资环境、管理水平和生产规模（如L_t和K_t增大）的观测误差降低引起ε_t偏离均值的程度不同，从而产生异方差.

异方差性产生的原因

(3) 分组数据.

一般用分组数据估计线性回归模型往往会产生异方差性. 这是因为不同组数据往往受到不同程度随机因素的影响，使得模型误差的方差往往不等.

(4) 模型的函数形式存在设定误差.

如道格拉斯生产函数的对数模型中如果将y_t直接用第t时刻某企业生产能力取代,而不是其对数，则模型就会产生异方差.

(5) 异常点的存在也会产生异方差性.

(6) 一个或多个回归解释变量的分布是偏态（skewness).

正态线性回归模型起源于自变量和因变量的联合分布为正态分布. 假设$(Y, \boldsymbol{X}')' \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$, 其中

$$\boldsymbol{\mu} = \begin{pmatrix} E(Y) \\ E(\boldsymbol{X}) \end{pmatrix} = \begin{pmatrix} \mu_Y \\ \boldsymbol{\mu}_X \end{pmatrix}, \quad \boldsymbol{\Sigma}_0 = \mathrm{Cov} \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} = \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\sigma}_{Y,X} \\ \boldsymbol{\sigma}'_{Y,X} & \boldsymbol{\Sigma}_X \end{pmatrix}.$$

由多元正态向量的条件分布性质得

$$Y|\boldsymbol{X} = \boldsymbol{x} \sim N_n(\beta_0 + \boldsymbol{x}'\boldsymbol{\beta}, \sigma^2),$$

其中, $\beta_0 = \mu_Y - \boldsymbol{\sigma}'_{Y,X}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X$, $\boldsymbol{\beta} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\sigma}_{Y,X}$,

$$\sigma^2 = \mathrm{Cov}(Y|\boldsymbol{X} = \boldsymbol{x}) = \sigma_Y^2 - \boldsymbol{\sigma}'_{Y,X}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\sigma}_{Y,X}.$$

若模型中自变量$\boldsymbol{X}$的分布换成偏态分布，则其条件协方差$\mathrm{Cov}(Y|\boldsymbol{X} = \boldsymbol{x})$就可能为自变量的函数$\sigma^2 = g(\boldsymbol{x})$, 从而导致异方差.

异方差性检验

异方差性检验问题也称方差的齐性检验问题，检验的原假设为

$$H_0: \sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2.$$

常见的方差的齐性检验方法:

(1) 图示检验法

- 用前面介绍的残差图
- 采用X与Y的散点图:
  看是否存在明显的散点扩大、缩小或复杂性趋势
- 对自变量X与LS残差的平方$\hat{e}_i^2$ 的散点图:
  看是否形成一斜率为零的直线
  当模型中包含多个自变量时，可对自变量逐一进行以上判断.

(2) Spearman秩相关检验

Spearman秩相关检验也称等级相关检验. 此处我们通过Spearman秩相关系数(rank correlation coefficient)探究误差方差是否依赖于某个自变量, 从而体现异方差性是否存在.

计算残差的绝对值$|\widehat{e}_i|$与第j个自变量X_j的Spearman秩相关系数:

$$r_s(j) = 1 - \frac{6}{n(n^2)} \sum_{i=1}^{n} (R(|\widehat{e}_i|) - R(x_{ij}))^2,$$

$R(|\widehat{e}_i|)$和$R(x_{ij})$分别为$|\widehat{e}_i|$ 和x_{ij}的秩, 即$|\widehat{e}_i|$ 和x_{ij}分别在$\{|\widehat{e}_1, \cdots, |\widehat{e}_n|\}$和$\{x_{1j}, \cdots, x_{nj}\}$ 按升序（或降序）排列后所得序列中的位置.

异方差性检验

选择与绝对残差值Spearman秩相关系数最大的自变量所对应的Spearman秩相关系数$r_s = \max r_s(j)$ 构造t 检验统计量：

$$t = \frac{\sqrt{n-2}r_s}{1-r_s^2}.$$

对于给定的显著性水平α, 若$|t| > t_{n-2}(\alpha/2)$, 则认为该数据模型存在异方差性, 否则认为误差等方差. 应用中要求$n > 8$

- R语言提供了Spearman 检验程序，使用格式为

lm1<-lm(y x1+x2, data=data1); e<-resid(lm1); abse<-abs(e);

cor.test(data$x1, abse, method="spearman")

cor.test(data$x2, abse, method="spearman")

异方差性检验

(3) Goldfeld-Quanadt 检验

该检验的基本思想为：将样本分为两部分，然后分别对两个样本进行回归, 并计算两个子样的残差平方和所构成的比值:

$$F = \frac{\mathrm{SSE}_2/(n_2 - p)}{\mathrm{SSE}_1/(n_1 - p)} = \frac{\mathrm{SSE}_2}{\mathrm{SSE}_1},$$

若$F > F_{n_2-p,n_1-p}(\alpha)$, 则认为存在异方差.

这一检验需要满足两个前提条件：要求变量的观测值为大样本；除了同方差假定不成立外，其它假定均满足.

异方差性检验

(4) 参数检验方法

假定误差方差σ_i^2与自变量$\boldsymbol{x}_i$满足函数关系：

$$g(\sigma_i^2) = \delta_0 + \sum_l^h f_l(\boldsymbol{x}_i)\delta_l,$$

其中$g(\cdot), f_1(\cdot), \cdots, f_h(\cdot)$ 为已知的连续函数.

检验$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$ 等价于检验

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_h = 0.$$

用借助于辅助回归模型: $g(\widehat{e}_i^2) = \delta_0 + \sum_j^h f_j(\boldsymbol{x}_i)\delta_j + \varepsilon_i, i = 1, \cdots, n$

作该回归方程的显著性检验.

异方差性检验

常见的2种参数检验方法：

- Breusch-Pagan (B-P) 检验　Breusch和Pagan(1979)提出辅助回归模型

$$\widehat{e}_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_p x_{i(p-1)} + \varepsilon_i, \quad i = 1, \cdots, n.$$

- White 检验　B-P 检验的一种拓展, 辅助回归:

$$\widehat{e}_i^2 = \delta_0 + \sum_{j=1}^{p-1} \delta_j x_{ij} + \sum_{j=1}^{p-1} \sum_{l=j}^{p-1} \delta_{j,l} x_{ij} x_{il} + \varepsilon_i, \quad i = 1, \cdots, n,$$

此时，自由度为$h = 2p + (p-1)p/2$.

例6.4.2

该数据来自1974年美国汽车趋势杂志, 包括32辆汽车（1973-74款）的油耗和10个方面的汽车设计和性能, 我们选因变量Y (每加仑油能跑多少英里, miles/gallon, mpg), 自变量为X_1 (车的排量, displacement, disp)和X_2(总马力, gross horsepower, hp).

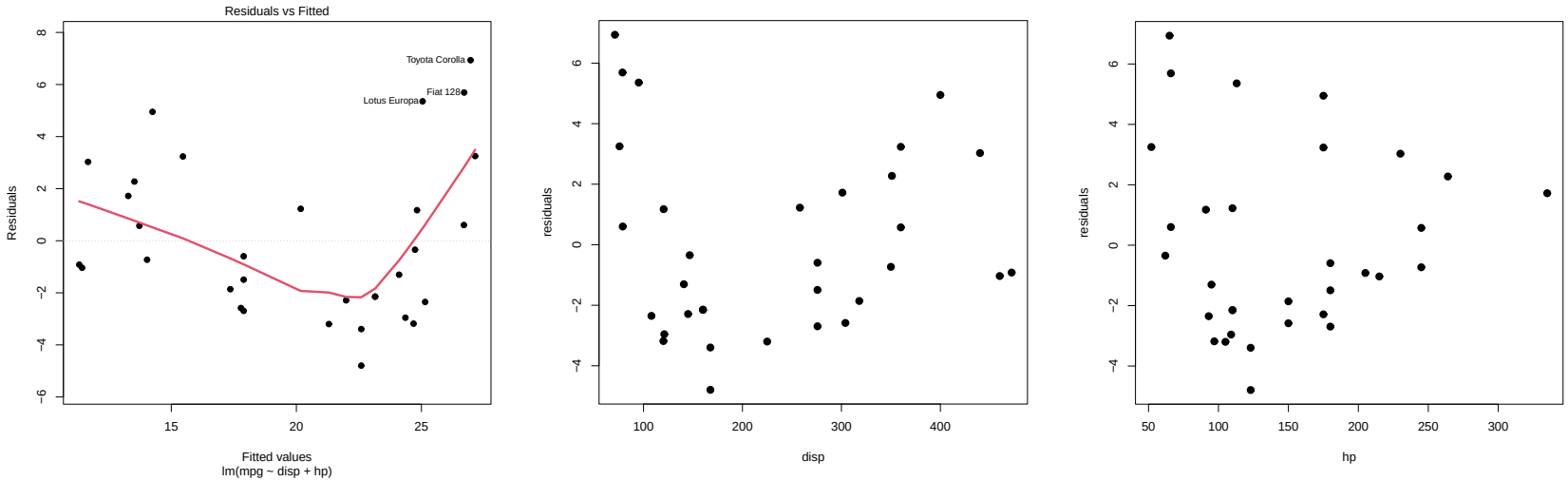
- 采用回归模型拟合数据，得经验回归方程为

$$Y = 30.7359 - 0.0303X_1 - 0.0248X_2.$$

经检验回归方程显著性, 调整的复相关系数为0.7309. X_2的回归系数的显著性检验的P值为0.074. 结合实际总马力对每加仑油能跑多少英里有影响的, 故将保留该变量.

例6.4.2

- 拟合值$\hat{y}_i$、自变量X_1 (disp)和X_2 (hp)为横轴的残差图



残差图直观上看，随拟合值和自变量的增加, 残差都有轻微的变化趋势.

例6.4.2

- 使用R语言的cor.test()函数执行残差与X_1 和X_2 的Spearman秩相关检验

  e<-abse(resid(model))

  cor.test(mtcars$disp, abse, method="spearman")

  cor.test(mtcars$hp, abse, method="spearman")

  - $|\hat{e}_i|$与X_1的Spearman秩相关系数为-0.3109
  - $|\hat{e}_i|$与X_2的Spearman秩相关系数为0.2572

在显著性水平$\alpha = 0.05$下, 两Spearman秩检验的p值分别为0.0833, 0.1554 认为残差与X_1和X_2的相关性都不显著, 不能排除模型误差同方差假设.

例6.4.2

- 使用Goldfeld-Quandt 异方差检验.
  - 删除总观测值的大约20%. mtcars总共有32个观测值, 选择删除中心7个观测值.

使用R语言的lmtest包中的gqtest()函数：

```
library(lmtest)
gqtest(model, order.by = disp+hp, data = mtcars, fraction = 7)
```

检验统计量的P值为0.486, 故该检验也不能排除模型误差同方差假设.

例6.4.2

- 应用Breusch-Pagan检验和White 检验了解异方差结构

  可使用lmtest包中的gqtest()函数, 相应程序如下：

  bptest(model)　♯ Breusch-Pagan test

  bptest(model, disp * hp + I(disp^2) + I(hp^2), $data = mtcars$)　♯ White 检验

  这两个检验统计量的的P值分别为0.1296和0.215, 故也不能排除模型误差同方差假设.

在样本量较少时, 常规统计检验方法通常倾向于作出保守结论的问题, 即在常用的显著性水平$\alpha = 0.05$ 和0.01下, 倾向于不拒绝原设. 因此, 关于少样本下的检验问题是一个值得关注的有待深入研究.

例6.4.2

- 例6.4.2的程序

```
####Ex 642
#load the dataset
 data(mtcars)
 #fit a regression model
 model<-lm(mpg~disp+hp, data=mtcars)
 summary(model)
plot(model, 1, pch=19, lwd=3)    残差图

plot(predict(model), model$residual, xlab="fitted value",ylab="residuals", pch=19, lwd=3)
plot(mtcars$disp, model$residual), xlab="disp",ylab="residuals", pch=19, lwd=3)
plot(mtcars$hp, model$residual, xlab="hp",ylab="residuals",  pch=19, lwd=3)

e<-resid(model) #获取残差
 abse<-abs(e)
 cor.test(mtcars$disp, abse, method="spearman" )
 cor.test(mtcars$hp, abse, method="spearman" )

#load lmtest library
library(lmtest)

#perform B-P test
bptest(model)

#perform White's test
bptest(model, ~ disp*hp + I(disp^2) + I(hp^2), data = mtcars) #采用二次项回归

#perform the Goldfeld Quandt test
gqtest(model, order.by = ~disp+hp, data = mtcars, fraction = 7)
```

异方差处理办法

(1) 加权LS估计方法

$$\widehat{\boldsymbol{\beta}}_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{y},$$

这里，$\mathbf{W} = \text{diag}(w_1, \cdots, w_n)$ 为权矩阵.

- 若σ_i^2已知, 则令

$$w_i = 1/\sigma_i^2, i = 1, \cdots, n,$$

此时$\widehat{\boldsymbol{\beta}}_w$为BLU估计

- σ_i^2往往未知, 可借助于辅助函数模型得到σ_i^2的估计, $\widehat{\sigma}_i^2$，令$w_i = 1/\widehat{\sigma}_i^2$, 得到可行的加权最小二乘估计.
- 特别地, 可用LS估计残差的平方$\widehat{e}_i$来估计σ_i^2.

异方差处理办法

(2) 方差稳定换变换

设$E(Y)=\mu$, $\mathrm{Var}(Y)=\sigma^2$, $\sigma=g(\mu)$, 这里μ未知, 函数$g(\cdot)$已知. 预寻找变换$U=f(Y)$, 使得U的方差等于或近似等于事先给定的常数σ_u^2.

假设f 函数可导，在$Y=\mu$ 附近作Taylor 展开, 得

$$U=f(\mu)+f'(\mu)(Y-\mu). \tag{6.18}$$

求方差得$\sigma_u^2=\mathrm{Var}(U)=(f'(\mu))^2\sigma^2$. 从而

$$f'(\mu)=\frac{\sigma_u}{\sigma}.$$

异方差处理办法

积分得

$$f(\mu)=\sigma_u\int\frac{d\mu}{\sigma}=\sigma_u\int\frac{d\mu}{g(\mu)},$$

于是所求的变换为

$$U=f(Y)=\sigma_u\int\frac{dY}{g(Y)}. \tag{6.19}$$

由(6.19), 容易推得下列方差稳定化变换：

- 若$\sigma^2\propto\mu(1-\mu)$, 则作变换$U=\sin^{-1}\sqrt{Y}$;
- 若$\sigma^2\propto\mu^2$, 则作变换$U=\ln(Y)$;
- 若$\sigma^2\propto\mu^3$, 则作变换$U=Y^{-1/2}$;
- 若$\sigma^2\propto\mu^4$, 则作变换$U=Y^{-1}$.

案例分析

在一项针对不同规模的27家工业企业的研究中, 统计了工人数X和主管人数Y. 我们希望研究两个变量之间的关系.

企业	X	Y	企业	X	Y	企业	X	Y
1	294	30	10	697	78	19	700	106
2	247	32	11	688	80	20	850	128
3	267	37	12	630	84	21	980	130
4	358	44	13	709	88	22	1025	160
5	423	47	14	627	97	23	1021	97
6	311	49	15	615	100	24	1200	180
7	450	56	16	999	109	25	1250	112
8	534	62	17	1434	114	26	1500	210
9	438	68	18	1015	117	27	1650	135

案例分析

- 用简单的线性模型拟合数据. 假设模型为

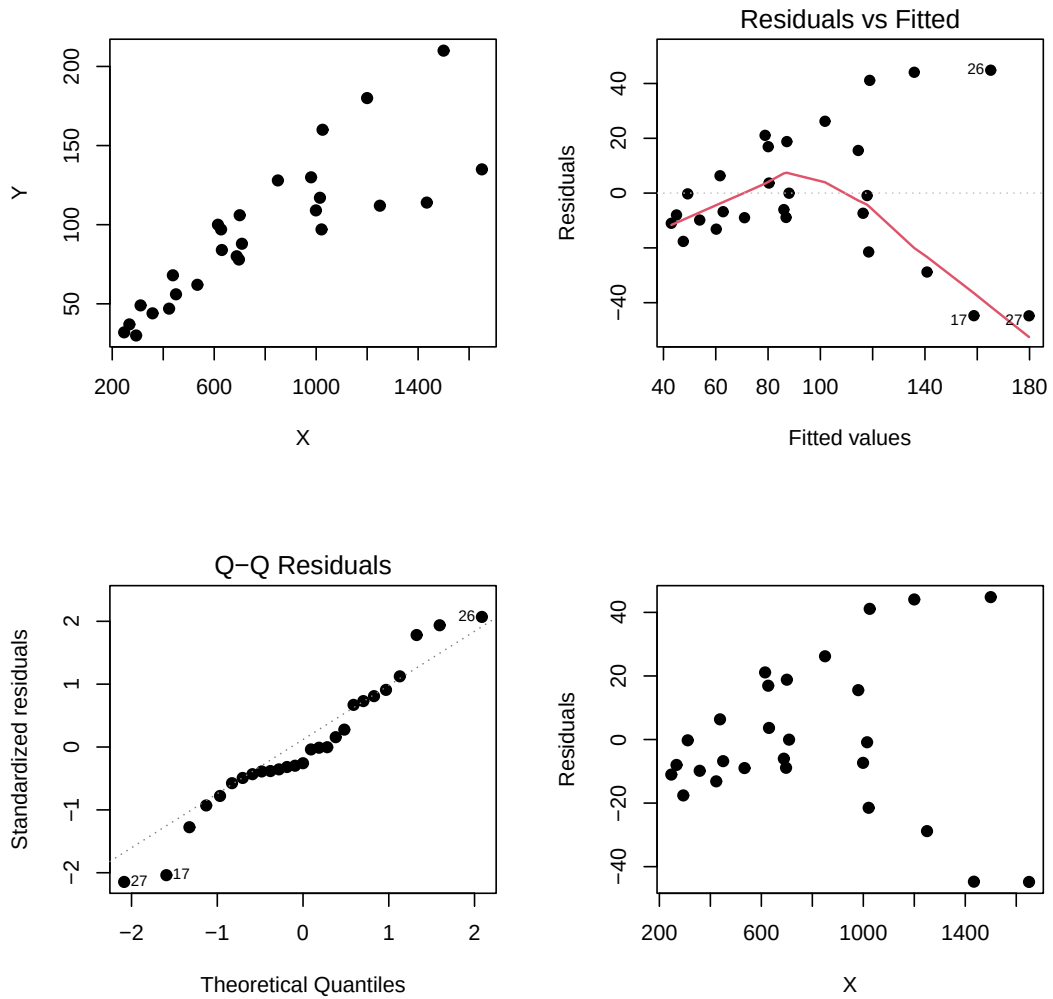
$$y_i = \beta_0 + x_i\beta_1 + e_i$$

应用R语言中函数lm(), 回归结果如下：

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.93883   10.19795   1.857   0.0751 .
X            0.09749    0.01177   8.280 1.25e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.73 on 25 degrees of freedom
Multiple R-squared:  0.7328,    Adjusted R-squared:  0.7221
F-statistic: 68.56 on 1 and 25 DF,  p-value: 1.247e-08
```

- (x_i, y_i) 散点图和$(\hat{y}_i, r_i)$散点图



案例分析

从X-Y散点图和拟合值-标准化残差的散点图可发现：

散点图皆呈现漏斗状, 表明模型误差具有异方差性.

- 对因变量作对数变换, 变换后模型为

$$\ln y_i = \beta_0 + x_i\beta_1 + e_i.$$

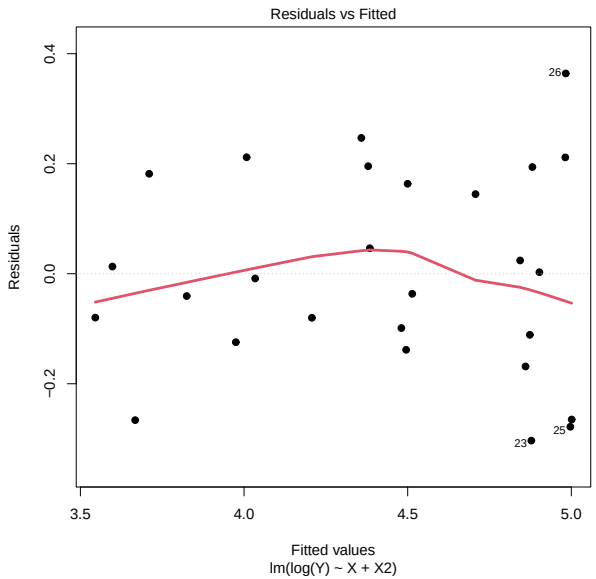
拟合结果表明：变换后模型的拟合效果更好, 复相关系数的平方由$R^2 = 0.7328$提升至$R^2 = 0.8767$.

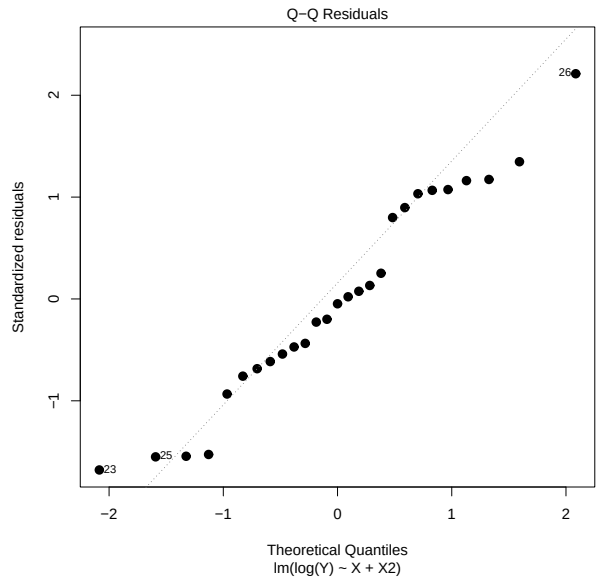
对数变换是回归分析中使用最为广泛的变换之一，由于对数变换常常起到降低数据的波动性额减少不对称性的作用，也能有效消除异方差性. 特别是当所分析变量的标准差相对于均值而言比较大时，这种变换特别有用.

- $(x_i, \ln y_i)$ 散点图和对数模型的残差图

拟合数据，回归结果和残差分析如下：

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.818e+00  1.644e-01  17.138 5.77e-15 ***
X            3.242e-03  4.192e-04   7.735 5.70e-08 ***
X2          -1.199e-06  2.312e-07  -5.188 2.58e-05 ***
---
```





案例分析

- 残差图显示该模型下的残差大都落在$[-0.2, 0.2]$内, 且随拟合值的变化没有明显的趋势, 其Q-Q图显示残差具有较好的正态性.
- Shapiro-Wilk正态检验也证实了这点, 检验统计量

$$W = 0.96058, \qquad P \text{ 值} = 0.381.$$

故工人数X和主管人数Y的关系可表示为如下经验回归方程:

$$\ln(Y) = 2.818 + 3.242X^* - 1.199(X^*)^2,$$

其中$X^* = (X/10^3)$ 表示以千为单位的工人人数.

综合处理：Box-Cox变换

试验数据集$(x_i', y_i), i = 1, \cdots, n$, 若经过回归诊断后得知, 它们不满足Gauss-Markov条件, 我们就要对数据采取“治疗"措施.

- 数据变换是处理有问题数据的一种好方法. 数据变换方法有多种, 其中最著名的Box-Cox变换.

Box-Cox变换可综合改善数据

正态性、对称性、方差相等性.

Box-Cox变换

Box-Cox变换

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \ln Y, & \lambda = 0, \end{cases} \tag{6.20}$$

这里λ是一个待定变换参数.

Box-Cox变换是一族变换, 它包括了许多常见的变换, 诸如

- 对数变换($\lambda = 0$)
- 倒数变换($\lambda = -1$)
- 和平方根变换($\lambda = 1/2$)

关键问题：如何确定λ?

Box-Cox变换

对因变量的n个观测值$y_1, \cdots, y_n$, 应用Box-Cox变换, 记变换后的向量为

$$\boldsymbol{y}^{(\lambda)} = (y_1^{(\lambda)}, \cdots, y_n^{(\lambda)})'.$$

我们的目的是确定变换参数λ, 使得$\boldsymbol{y}^{(\lambda)}$满足

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

Box-Cox变换是通过参数λ的选择, 达到对原来数据的“综合治理”, 使其满足一个正态线性回归模型的所有假设条件, 即

$$\boldsymbol{y}^{(\lambda)} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \ \sigma^2\mathbf{I}).$$

采用极大似然方法来确定变换参数λ.

对固定的λ, $\boldsymbol{\beta}$ 和σ^2 的似然函数为

$$L(\boldsymbol{\beta},\ \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})\right\} J,$$

这里J为Jacobi行列式的绝对值

$$J = \prod_{i=1}^{n}\left|\frac{dy_i^{(\lambda)}}{dy_i}\right| = \prod_{i=1}^{n} y_i^{\lambda-1}.$$

对给定的λ, 易证$\boldsymbol{\beta}$和σ^2的极大似然估计为

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}^{(\lambda)}, \quad \widehat{\sigma}^2(\lambda) = \frac{1}{n}\mathrm{SSE}(\lambda,\ \boldsymbol{y}^{(\lambda)}),$$

这里$\mathrm{SSE}(\lambda,\ \boldsymbol{y}^{(\lambda)}) = \boldsymbol{y}^{(\lambda)'}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{y}^{(\lambda)}$.

对应的似然函数最大值为

$$L(\widehat{\boldsymbol{\beta}}(\lambda),\ \widehat{\sigma}^2(\lambda)) = (2\pi e)^{-n/2} \cdot \boldsymbol{J} \cdot \left(\frac{\mathrm{SSE}(\lambda,\ \boldsymbol{y}^{(\lambda)})}{n} \right)^{-n/2}.$$

按照似然原理，我们选择

$$\lambda = \arg\max_{\lambda} L(\widehat{\boldsymbol{\beta}}(\lambda),\ \widehat{\sigma}^2(\lambda)) = \arg\min_{\lambda} \mathrm{SSE}(\lambda,\ \boldsymbol{z}^{(\lambda)})$$

其中

$$\boldsymbol{z}^{(\lambda)} = (z_1^{(\lambda)}, \cdots, z_n^{(\lambda)})' = \frac{\boldsymbol{y}^{(\lambda)}}{\boldsymbol{J}^{1/n}},$$

$$z_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda}}{(\prod_{i=1}^{n} y_i)^{(\lambda-1)/n}}, & \lambda \neq 0, \\ (\ln y_i)(\prod_{i=1}^{n} y_i)^{\frac{1}{n}}, & \lambda = 0. \end{cases} \tag{6.21}$$

Box-Cox变换的具体步骤

1. 对给定的λ值, 利用(6.21)计算$z_i^{(\lambda)}$;
2. 计算残差平方和$\mathrm{SSE}(\lambda,\ \boldsymbol{z}^{(\lambda)}) = \boldsymbol{z}^{(\lambda)'}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{z}^{(\lambda)}$;
3. 对一系列的λ值, 重复上述步骤, 得残差平方和$\mathrm{SSE}(\lambda,\ \boldsymbol{z}^{(\lambda)})$ 的一系列值, 以λ 为横轴, 作出相应的曲线. 用直观的方法, 找出使$\mathrm{SSE}(\lambda,\ \boldsymbol{z}^{(\lambda)})$ 达到最小值的点$\widehat{\lambda}$.
4. 求出$\widehat{\boldsymbol{\beta}}(\widehat{\lambda})$.

注 步骤3也可以换成对数似然函数$\ln L_{\max}(\lambda)$ 与λ 相应的曲线, 找使得$\ln L_{\max}(\lambda)$ 达到最大的$\widehat{\lambda}$. 这条曲线可以由R语言中函数boxcox()给出.

Box-Cox变换的案例

例6.4.4

一公司为了研究产品的营销策略, 对产品的销售情况进行了调查. 设Y表示某地区该产品的家庭人均购买量(单位:元), X表示家庭人均收入(单位:元). 下表记录了53个家庭的数据. 试对Y和X建模.

i	X	Y	i	X	Y	i	X	Y
1	679	0.790	2	292	0.440	3	1012	0.560
4	493	0.790	5	582	2.700	6	1156	3.640
7	997	4.730	8	2189	9.500	9	1097	5.340
10	2078	6.850	11	1818	5.840	12	1700	5.210
13	747	3.250	14	2030	4.430	15	1643	3.160
16	414	0.550	17	354	0.170	18	1276	1.880
19	745	0.770	20	435	1.390	21	540	0.560
22	874	1.560	23	1543	5.280	24	1029	0.640

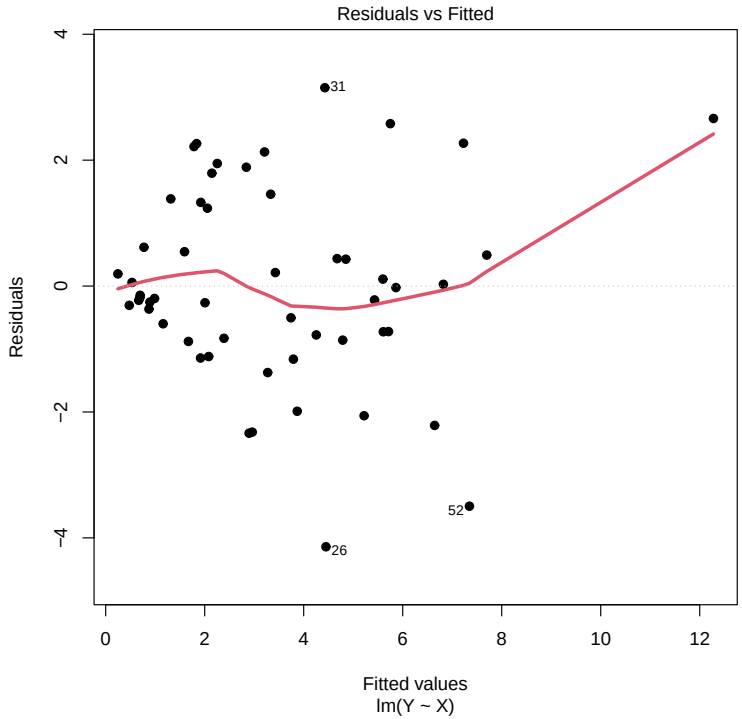
i	X	Y	i	X	Y	i	X	Y
25	710	4.000	26	1434	0.310	27	837	4.200
28	1255	2.630	29	1748	4.880	30	1381	3.480
31	1428	7.580	32	1777	4.990	33	370	0.590
34	2316	8.190	35	1130	4.790	36	463	0.510
37	770	1.740	38	724	4.100	39	808	3.940
40	790	0.960	41	783	3.290	42	406	0.440
43	1242	3.240	44	658	2.140	45	1746	5.710
46	468	0.640	47	1114	1.900	48	413	0.510
49	1787	8.330	50	3560	14.940	51	1495	5.110
52	2221	3.850	53	1526	3.930			

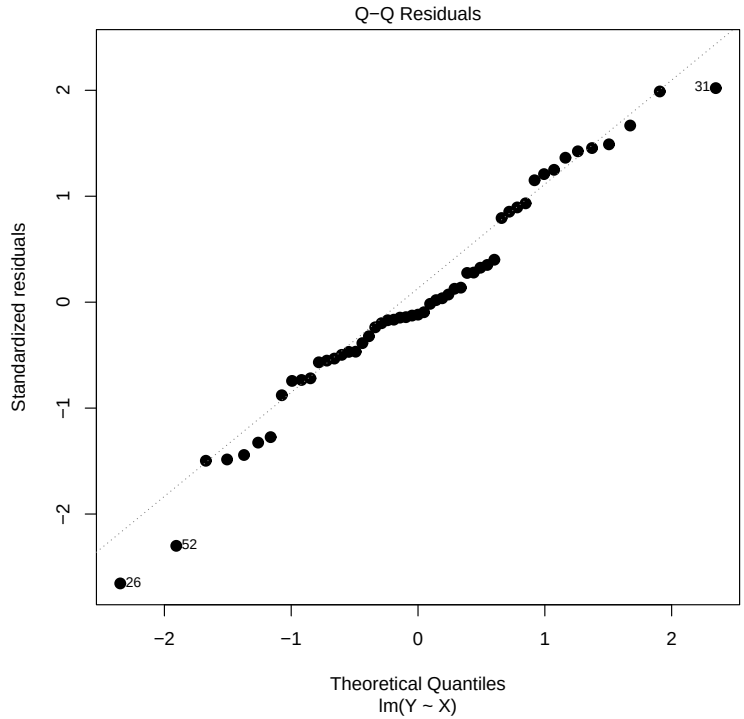
首先采用一元线性回归模型拟合数据, 得经验回归方程

$$\widehat{Y} = -0.828 + 0.004X.$$

回归方程显著性检验的P值为4.164e-15 , Y和X相关系数为0.839.

例6.4.4





残差图: 从左向右逐渐散开呈漏斗状

正态Q-Q图: 点$\{(q_{(i)}, r_{(i)})\}$大都在一条直线上附近

例6.4.4

- 采用Box-Cox变换对因变量Y的异方差性进行“治理".

  计算给出的12个不同λ值所对应的残差平方和$\mathrm{SSE}(\lambda,\ z^{(\lambda)})$：

Table: 残差平方和随着变换参数的变化趋势

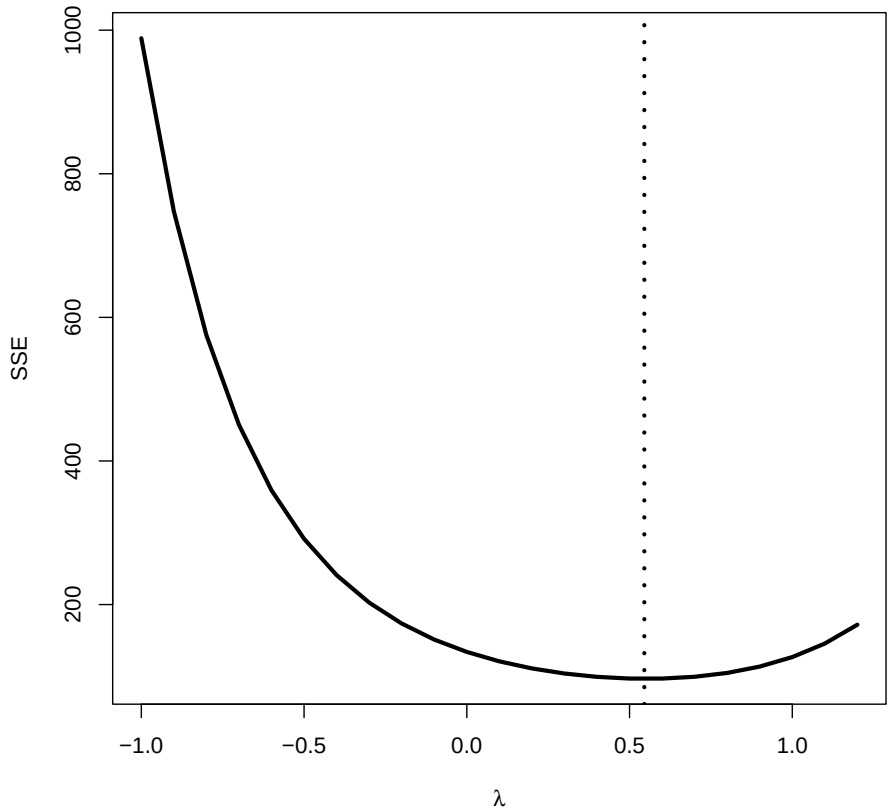
λ	-2	-1	-0.5	0	0.125	0.25
SSE	34460.43	988.71	291.56	134.04	118.170	107.20
λ	0.375	0.5	0.625	0.75	1	2
SSE	100.260	96.97	97.310	101.71	126.85	1271.04

- 当$\lambda = 0.5$时, 残差平方和$\mathrm{SSE}(\lambda,\ z^{(\lambda)})$ 达到最小, 因此我们可以初步认定认为最优λ 应在0.5附近.

例6.4.4

- 利用R语言中boxcox()函数来寻找最优λ (极大似然函数)

b<-boxcox(lm(Y $\sim$ X)); b$x[which.max(b$y)] ($\sharp\ \lambda = 0.5454545$)



例6.4.4

令$Z = Y^{1/2}$作为因变量, 对变换后所得新的因变量作回归，得到如下经验回归方程

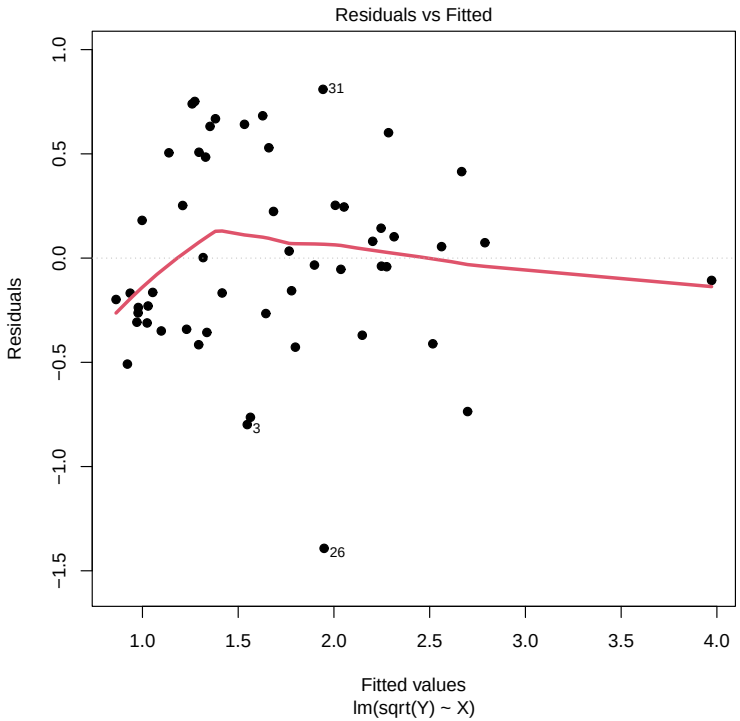
$$\sqrt{Y} = 0.5842 + 0.0009517X.$$

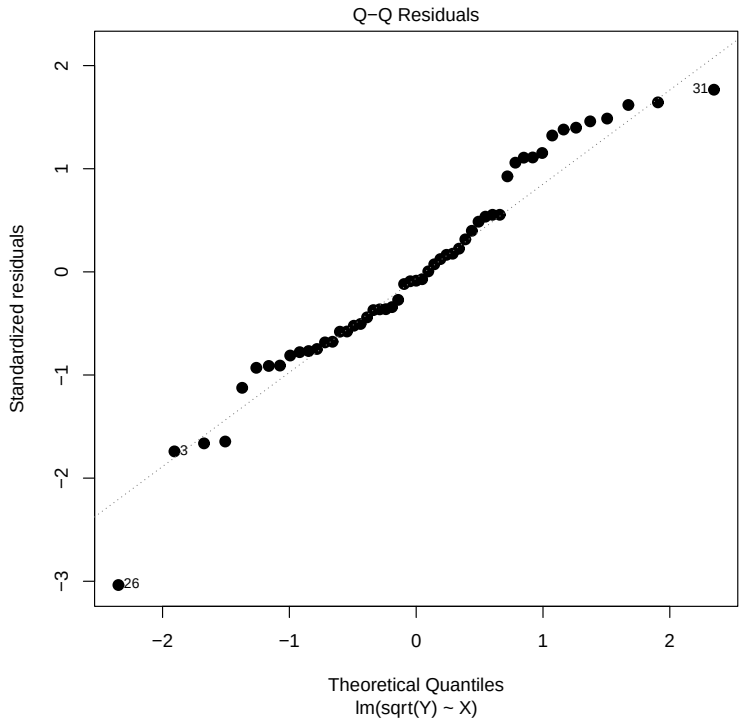
```
Coefficients:
             Estimate    Std.Error  t value  Pr(>|t|)
(Intercept) 5.842e-01  1.298e-01   4.500  3.96e-05 ***
X           9.517e-04  9.816e-05   9.695  3.66e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4637 on 51 degrees of freedom
Multiple R-squared:  0.6483,    Adjusted R-squared:  0.6414
F-statistic:    94 on 1 and 51 DF,  p-value: 3.663e-13
```

例6.4.4

- 根式变换后的残差图和残差Q-Q图





比原数据下的残差图有较大改善, 随着$\widehat{Z}$增大，残差基本已无明显变化趋势

例6.4.4

根式变换后的回归方程显著, 不过$\sqrt{Y}$与X的相关系数有所下降, 变为0.6483.

从Q-Q图看残差的正态性似乎有所减弱, 不过新残差的Shapiro-Wilk统计量W的P值为$0.138 > 0.05$, 在显著性水平$\alpha = 0.05$下, 作根式变换后的模型误差仍可以被认为服从正态分布.

经验回归方程:

$$\hat{Y} = \hat{Z}^2 = (0.5842 + 0.0009517X)^2$$

自相关性的诊断

考虑线性回归模型:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma} = (\sigma_{\mathbf{ij}}),$$

若$\sigma_{i,i+1} \neq 0$时，则误差项一定是自相关的.

本节介绍

- 产生自相关的原因
- 对数据分析的影响
- 常见的自相关的诊断方法以及自相关性数据分析法.

产生自相关的原因

主要有三个方面：

(1) 在时间上或空间上，相邻数据趋向于相似.

由于经济系统的经济行为都具有时间上的惯性、经济变量间影响的滞后性、微观经济学中的蛛网现象等，又如空间数据中，由于受共同的外部环境的影响，相邻地块的数据往往具有空间自相关性.

(2) 当回归方程设定不正确时, 也会出现自相关现象.

模型忽略某重要变量会产生系统误差, 模型误差之间就会出现相关性

(3) 因对数据进行加工整理而导致误差项之间产生自相关性.

如对缺失值采用特定的统计方法进行插值, 也可能是的插值后的数据自相关.

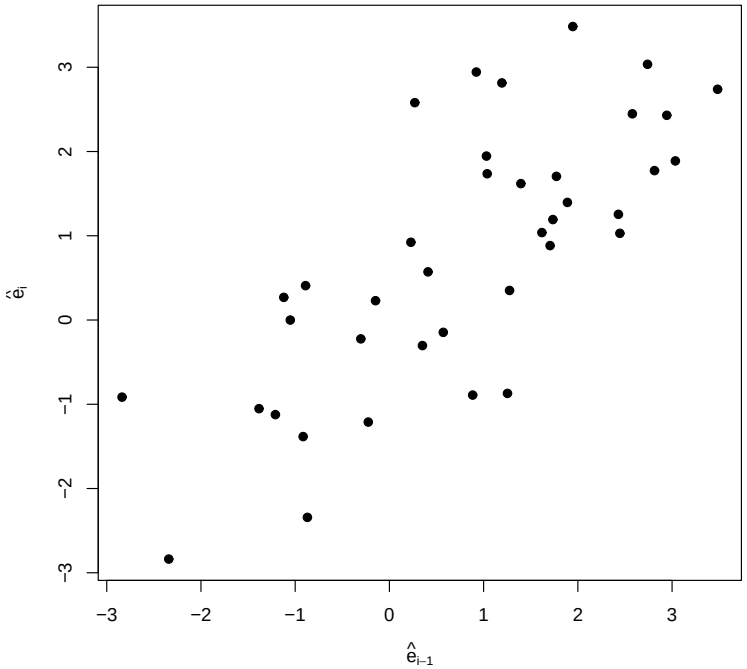
自相关现象对数据分析的影响

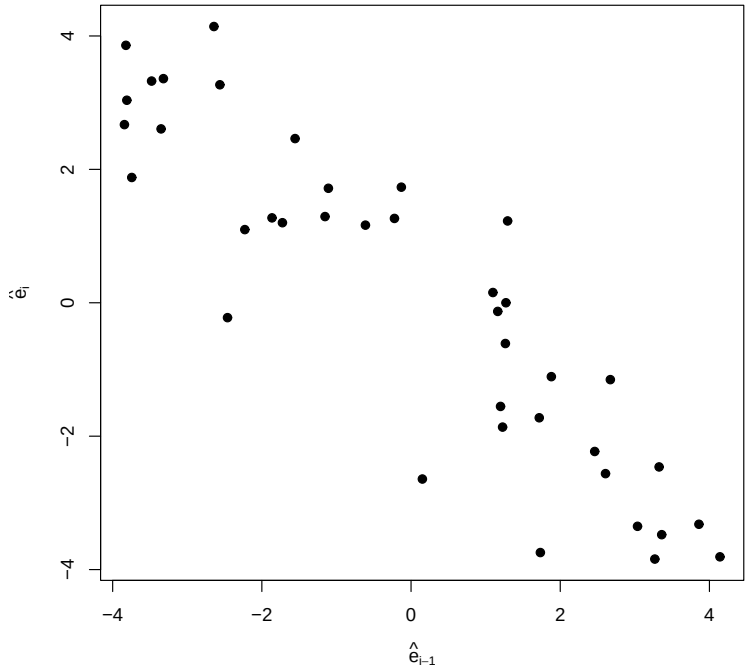
- 回归系数的最小二乘估计是无偏的，但是不再具有最小方差.
- 回归系数的标准差会被严重地低估；也就是说，由数据估得的标准差会比它的实际值大大地缩小，从而给出一个假想的精确估计.
- 置信区间和通常采用的各种显著性检验的结论，严格地说来不再是可信的.

自相关性的诊断3个基本方法: 图示法；游程检验法；Durbin-Watson (DW) 检验

自相关性的诊断方法

- 图示法 绘制$(\widehat{e}_{t-1}, \widehat{e}_t)$的散点图





序列正/负相关：散点大都集中在一条斜率为正/负的直线附近

自相关性的诊断方法: 游程检验法

- 游程检验法.

  从残差时序图判断模型的随机误差项间是否存在正或负的序列相关性, 主要依据是看残差序列的正负号出现是否有规律.

残差符号的序列图

将残差的正负号按时间顺序排列起来，形成一个符号的序列，称其为残差符号的序列图.

游程

按连续的符号可以将残差的符号序列图分解成若干子序列，称各子序列为一个游程. 记游程的个数为R.

设一个序列图共有n_1个正号，n_2个负号.

自相关性的诊断方法: 游程检验法

假设模型的随机误差项是独立同分布, 记$n = n_1 + n_2$. 则

$$P(R=r) = \begin{cases} \dfrac{2C_{n_1-1}^{k-1}C_{n_2-1}^{k-1}}{C_n^{n_1}}, & r = 2k, \\ \dfrac{C_{n_1-1}^{k-1}C_{n_2-1}^{k} + C_{n_1-1}^{k}C_{n_2-1}^{k-1}}{C_n^{n_1}}, & r = 2k+1, \end{cases}$$

$k = 1, 2, \cdots, [n/2]$. R的期望μ和方差σ^2:

$$E(R) = \mu = \frac{2n_1n_2}{n_1+n_2} + 1,$$

$$\mathrm{Var}(R) = \sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)} + 1,$$

自相关性的诊断方法: 游程检验法

当样本量小时，可以根据精确分布计算拒绝域；

当样本量小时很大时，我们可以依据R的渐近正态性检测误差序列的独立性，即若

$$\frac{|R-\mu|}{\sigma} \leq z_{\alpha/2},$$

则认为误差序列的独立性, 否则认为误差序列相关.

- 关于这个近似的游程检验，可用R语言数据中的函数runs.test().
- 近似的游程检验一般不可用于小样本的情况(n_1和n_2小于10).

游程检验法例子

例如残差的符号序列图:

$$- \ - \ - \ - \ - \ - \ - \ + \ + \ + \ + \ + \ + \ + \ +$$

共2个游程，$n_1 = 8, n_2 = 7,$

$$P(R \leq 2) = \frac{C_7^0 C_6^1 + C_7^1 C_6^0}{C_{15}^7} = \frac{6+7}{6435} \approx 0.002,$$

因此拒绝原假设, 认为误差序列是相关序列的.

若随即独立, $E(R) = 8.467$, R的标准差为1.857.

$$\frac{|R-\mu|}{\sigma} = \frac{|2-8.47|}{1.857} \approx 3.484 > z_{0.025} = 1.96,$$

近似的游程检验与精确游程检验结果相同.

游程检验法例子

但由

$$P\left(\frac{|R-\mu|}{\sigma} > \frac{|2-8.47|}{1.857}\right) \approx 1-\Phi(3.484) = 0.00025$$

可以反映出：小样本下, 近似的游程检验往往会更倾向于拒绝原假设, 使得犯第一类错误的概率大于名义显著性水平.

关于游程检验, 读者可参考有关非参数统计的著作：

Lehmann, E. L. Nonparametric Statistical Methods Based on Ranks, New York: McGraw-Hill, 1975.

Hollander, M. and Wollfe, D. A. Nonparametric Statistical Methods, New York: John Wiley & Sons, 1999.

自相关性的诊断方法

- Durbin-Watson (DW) 检验

  一阶自回归结构：

$$e_i = \rho e_{i-1} + u_i, \quad |\rho| < 1,$$

其中, ρ 是相邻误差e_{i-1} 与e_i的相关系数, $u_i \sim N(0, \sigma^2)$相互独立,

$$\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma} = \frac{\sigma^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}.$$

Durbin-Watson (DW) 检验

在经济学数据中, 倾向于正序列相关.

关于随机误差项的自相关性检验的原假设和备择假设：

$$H_0 : \rho = 0 \longleftrightarrow H_1 : \rho > 0.$$

DW统计量

$$D = \frac{\sum_{i=2}^{n} (\widehat{e}_i - \widehat{e}_{i-1})^2}{\sum_{i=1}^{n} \widehat{e}_i^2},$$

其中, $\widehat{e}_i$ 是第i 个普通最小二乘残差.

DW 统计量是著名的一阶自相关检验的统计量.

Durbin-Watson (DW) 检验

当n充分大时, 有

$$D \approx \frac{2\sum_{i=2}^{n}\widehat{e}_i^2 - 2\sum_{i=2}^{n}\widehat{e}_{i-1}\widehat{e}_i}{\sum_{i=2}^{n}\widehat{e}_i^2} = 2(1-\widehat{\rho}),$$

其中$\widehat{\rho} = \sum_{i=2}^{n}\widehat{e}_{i-1}\widehat{e}_i / \sum_{i=2}^{n}\widehat{e}_i^2$.

Table: DW值与$\widehat{\rho}$的对应关系

$\widehat{\rho}$	D	误差项的自相关性
-1	4	完全负自相关
(-1, 0)	(2,4)	负自相关
0	2	无自相关
(0, 1)	(0,2)	正自相关
1	0	完全正自相关

Durbin-Watson (DW) 检验

由表可知, 当$\rho = 0$ 时, D 接近于2.

- 对于假设

$$H_0 : \rho = 0 \longleftrightarrow H_1 : \rho > 0.$$

自相关的判别准则:

(I) 若$D < d_L$, 则拒绝原假设H_0, 认为误差序列为正自相关;

(II) 若$D > d_U$, 则接受原假设, 认为判断误差序列是无自相关;

(III) 若$d_L \leq D \leq d_U$, 则不能判断误差序列是否自相关,

- 对于假设

$$H_0 : \rho = 0 \longleftrightarrow H_1 : \rho < 0,$$

采用统计量$4 - \mathrm{D}$, 检验过程同上.

Durbin-Watson (DW) 检验

- 对于双边检验

$$H_0 : \rho = 0 \longleftrightarrow H_1 : \rho \neq 0.$$

自相关的判别准则如下:

(I) 若$D < d_L$或$D > 4 - d_U$ 则认为误差序列自相关;

(II) 若$d_U \leq D \leq 4 - d_U$, 则认为判断误差序列是无自相关;

(III) 若$d_L \leq D \leq d_U$或$4 - d_U \leq D \leq 4 - d_L$, 则待判;

关于临界值计算, 参见Durbin 和Watson (1971).

Durbin-Watson (DW) 检验

R语言提供了DW检验dwtest()函数

DW检验的以下局限性

- 存在待判区域, 需要增大样本量或其他方法进一步检验;
- 只给了$n > 15$时的临界值，不能用于$n \leq 15$的情形；
- 只针对一阶自相关问题构造的检验, 不能用于高阶序列相关的检验.
- 高阶序列自相关性检验, 需采用Breusch-Godfrey（BG）检验, 也称拉格朗日乘数检验
- 只适合随机误差方差相等，不适合异方差, 尤其因变量有滞后效应, 如自回归的情形：$y_i = \rho y_{i-1} + \boldsymbol{x}_i'\beta + e_i$.

消除自相关性的方法：广义差分变换

考虑模型:

$$y_i = \beta_0 + x_i\beta_1 + e_i, \quad e_i = \rho e_{i-1} + u_i, \quad i = 1, 2, \cdots, n,$$

这里, $u_1, \cdots, u_n$相互独立, 且$u_i \sim N(0, \sigma^2)$.

广义差分变换

$$y_i^* = y_i - \rho y_{i-1},$$

$$x_i^* = x_i - \rho x_{i-1},$$

$i = 2, \cdots, n$. 记

$$\beta_0^* = \beta_0(1-\rho), \quad \beta_1^* = \beta_1.$$

广义差分变换

一阶差分模型

$$y_i^* = \beta_0^* + \beta_1^* x_i^* + u_i, \quad i = 2, \cdots, n,$$

- 若ρ已知, 一阶差分模型下参数(β_0^*, β_1^*)的LS估计$(\widehat{\beta}_0^*(\rho), \widehat{\beta}_1^*(\rho))$就是BLU估计. 反解得

$$\widehat{\beta}_0 = \widehat{\beta}_0^*(\rho)/(1-\rho), \quad \widehat{\beta}_1 = \widehat{\beta}_1^*(\rho).$$

- ρ未知, 是一个待估参数. 两种常用的迭代估计
  - Cochrane-Orcutt迭代估计
  - Hildreth-Lu 估计

Cochrane-Orcutt迭代估计

- Cochrane-Orcutt迭代估计

(1) 估计ρ. 计算原模型的最小二乘残差$\widehat{e}_i$, 然后取ρ的初始估计为

$$\rho = \widehat{\rho} = \frac{\sum_{i=2}^{n} \widehat{e}_{i-1}\widehat{e}_i}{\sum_{i=2}^{n} \widehat{e}_i^2};$$

(2) 基于一阶差分模型求β_0和β_1的估计.

(3) 迭代停时检验. 采用DW检验对变换后的模型的误差项进行检验，若检验显示不相关，则停止迭代，输出原模型参数(β_0, β_1)的估计；否则在$\widehat{\beta}_0$ 和$\widehat{\beta}_1$的基础上, 计算原模型新的残差, 再更新ρ的估计；

(4) 重复步骤(2) 和(3) 直到DW检验显示模型的误差项不相关为止.

Hildreth-Lu 估计

- Hildreth-Lu 估计

直接估计参数β_0, β_1, ρ 方法, 即极小化模型的误差平方和

$$\text{SSE}(\beta_0, \beta_1, \rho) = \sum_{i=2}^{n} \left((y_i - \rho y_{i-1}) - \beta_0(1-\rho) - \beta_1(x_i - \rho x_{i-1})\right)^2 .$$

- 先将ρ当成已知, 可求得β_0 和β_1 估计为

$$\widehat{\beta}_0(\rho) = \widehat{\beta}_0^*(\rho)/(1-\rho), \quad \widehat{\beta}_1(\rho) = \widehat{\beta}_1^*(\rho);$$

- 求ρ的估计: $\widetilde{\rho} = \text{argmin}_\rho \text{SSE}(\widehat{\beta}_0(\rho), \widehat{\beta}_1(\rho), \rho)$;
- 回代得$\widehat{\beta}_0(\widetilde{\rho})$和$\widehat{\beta}_1(\widetilde{\rho})$.

R语言中的cochrane.orcutt()函数和hildreth.lu()函数实现

自相关性诊断案例

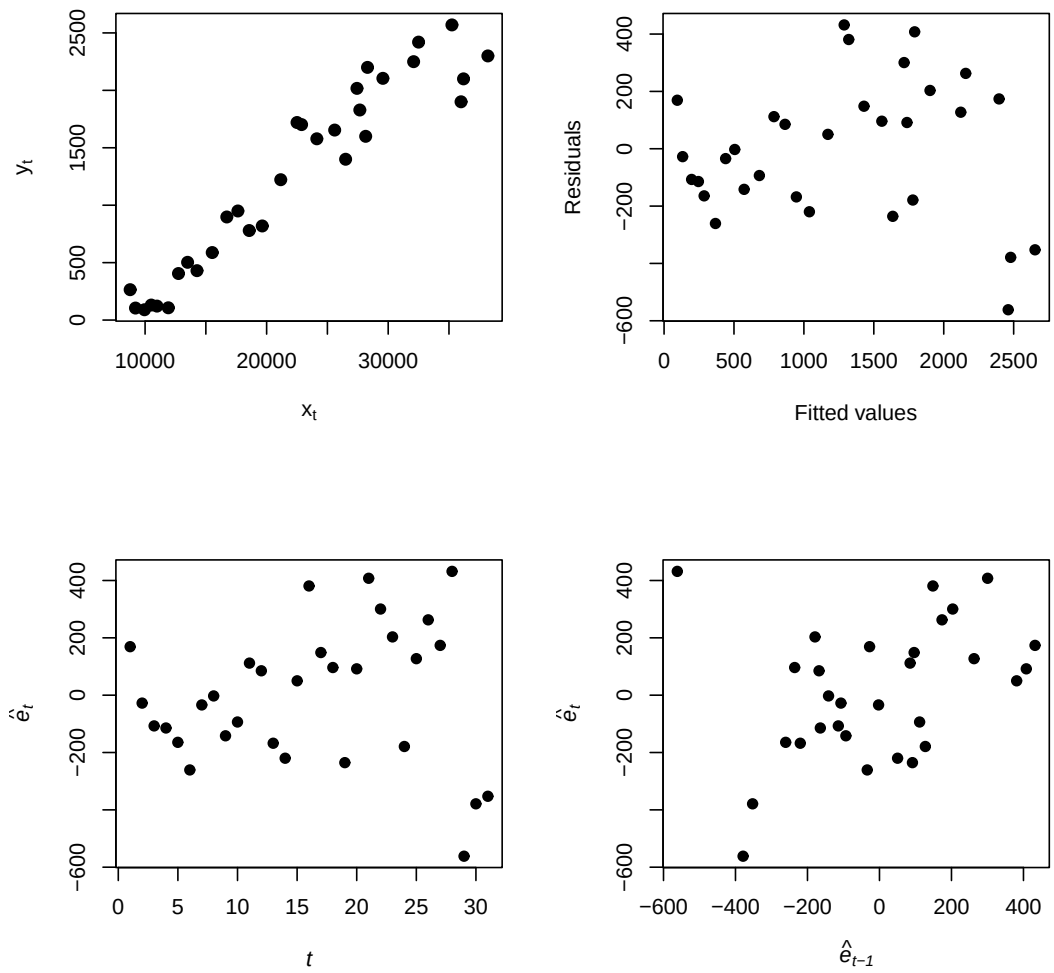
例6.4.5 (唐年胜,李会琼, 2014, 例5.8)

设某地区居民收入X（元）与储蓄额Y(元)的历史统计数据, 采用一元线性模型进行拟合数据并分析模型的随机误差是否存在自相关性.

t	y_t	x_t	t	y_t	x_t	t	y_t	x_t
1	264	8777	12	950	17633	23	2105	29560
2	105	9210	13	779	18575	24	1600	28150
3	90	9954	14	819	19635	25	2250	32100
4	131	10508	15	1222	21163	26	2420	32500
5	122	10979	16	1702	22880	27	2570	35250
6	107	11912	17	1578	24127	28	1720	22500
7	406	12747	18	1654	25604	29	1900	36000
8	503	13499	19	1400	26500	30	2100	36200
9	431	14269	20	1829	27670	31	2300	38200
10	588	15522	21	2200	28300			
11	898	16730	22	2017	27430			

例6.4.5

- 散点图、残差图、残差时序图、相邻时刻残差散点图



例6.4.5

- 图显示结果：
  - x_t和y_t具有强的线性关系；（计算得样本相关系数为0.956）
  - 随时间t或$\hat{y}_t$增大, 残差$\hat{e}_t$的波动范围有增大趋势;
  - $(\hat{e}_t, \hat{e}_{t-1})$大都落在一条斜率为正的直线附近(一阶正的自相关性).
- DW检验统计量$D = 1.2529$, P值= $0.008674 < 0.05$, 可认为误差序列一阶正自相关:$e_t = \rho e_{t-1} + u_i \ \ (\rho > 0)$.
- 一阶差分的经验回归方程（Cochrane-Orcutt方法）

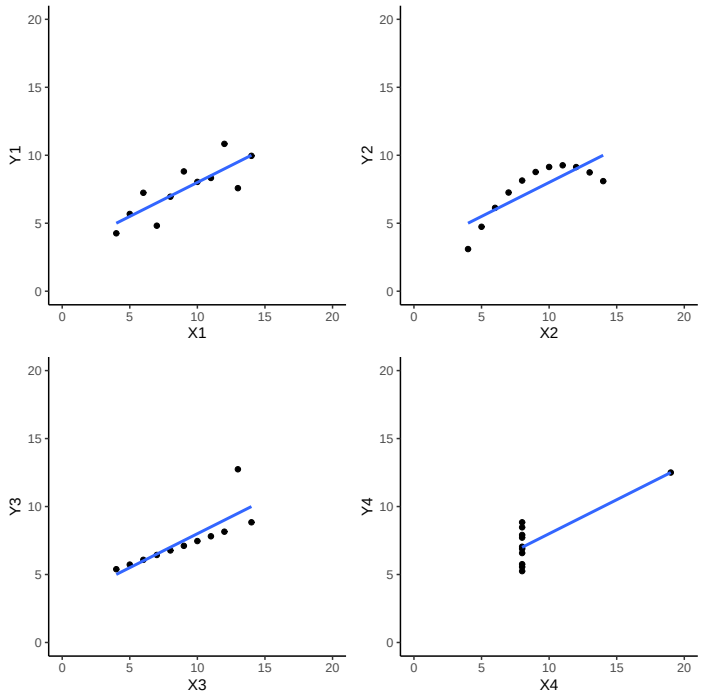
$$\hat{y}_t^* = -450.805032 + 0.076681 x_t^*,$$

其中$\hat{y}_t^* = \hat{y}_t - \hat{\rho}\hat{y}_{t-1}$, $x_t^* = x_t - \hat{\rho} x_{t-1}$, $\hat{\rho} = 0.518639$.

影响分析

样本$(y_i, \tilde{\boldsymbol{x}}_i',), i = 1, \cdots, n$只是$(Y, X_1, \cdots, X_{p-1})$的许多可能取值中取到的$n$组. 希望每组数据$(\tilde{\boldsymbol{x}}_i', y_i)$对未知参数的估计有一定的影响, 但这种影响不能过大, 否则得到的经验回归方程不具有稳定性.

- 如Anscombe四组数据：右下角的直线完全由一个点决定，如果去掉这个极端点，会得到完全不同的直线. 此时考察残差没有用，因为此点的残差是零.



影响分析

强影响点（Influential Points）

是指那些对统计模型的参数估计或预测结果有显著影响的观测值.

- 这些点通常具有极端的值或者与数据集中的其他点存在显著的差异, 可能对模型的准确性、稳定性和解释性产生重要影响.
- 可能是自变量的离群观测点, 也可能是因变量的离群观测点,
  - 高杠杆点(high leverage): 自变量的离群观测点
  - 异常点: 因变量的离群观测点
  - 高杠杆点+异常点

高杠杆点的诊断

- 考虑线性回归模型

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Var}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_n. \tag{6.22}$$

记$\mathbf{H} = (h_{ij}) = \mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

  - 第i个拟合值$\hat{y}_i$是Y的所有观测值的加权和:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n, \quad i = 1, \cdots, n,$$

    其中h_{ii} 是y_i对$\hat{y}_i$的权重, 被称为第i个观测的杠杆值.
  - 当h_{ii}越接近于1, $\hat{y}_i$越接近于y_i, 即残差$\hat{e}_i$ 越接近于零.
  - 残差$\hat{e}_i$的方差: $\mathrm{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$ 与σ^2和h_{ii}都有关.

高杠杆点的诊断

对于一元线性回归

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

记

$$\mathbf{X} = (\mathbf{1}_n, \boldsymbol{x}), \quad \boldsymbol{x} = (x_1, \cdots, x_n)'.$$

易证

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}.$$

- 当$x_i - \bar{x} = 0$ 时，h_{ii}达到最小值$1/n$;
- 随着x_i 远离中心点$\bar{x}$，h_{ii}增大. 当x_i离中心点$\bar{x}$足够远时，h_{ii}能都充分接近于1, 将回归直线拉向点(y_i, x_i).

h_{ii} 的性质与几何意义

定理6.5.1

(1) $0 \leq h_{ii} \leq 1$, 且当$h_{ii} = 0$时, $h_{ij} = 0$, $i \neq j$;

(2) $\sum_{i=1}^{n} h_{ii} = p$;

(3) $\sum_{j=1}^{n} h_{ij} = 1$;

(4) $h_{ii} = \frac{1}{n} + (\tilde{\boldsymbol{x}}_i - \bar{\tilde{\boldsymbol{x}}})'(\tilde{\mathbf{X}}_c'\tilde{\mathbf{X}}_c)^{-1}(\tilde{\boldsymbol{x}}_i - \bar{\tilde{\boldsymbol{x}}})$, 这里$\bar{\tilde{\boldsymbol{x}}} = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i/n$, $\tilde{\mathbf{X}}_c = \tilde{\mathbf{X}} - \mathbf{1}_n\bar{\tilde{\boldsymbol{x}}}'$为自变量观测矩阵$\tilde{\mathbf{X}}$的中心化.

由定理的(4)可知：

- h_{ii}的几何意义. $(n-1)(h_{ii} - \frac{1}{n})$表示在自变量空间中, 第$i$个试验点$\tilde{\boldsymbol{x}}_i$ 到试验中心$\bar{\tilde{\boldsymbol{x}}}$的Mahalanobis距离， 简称马氏距离, 刻画了第i个试验点到试验中心$\bar{\tilde{\boldsymbol{x}}}$的远近.

- 第i个观测的杠杆值h_{ii}与普通残差$\widehat{e}_i$都有满足关系:

$$h_{ii} + \frac{\widehat{e}_i^2}{\text{SSE}} \leq 1,$$

其中$\text{SSE} = \sum_i^n \widehat{e}_i^2$. 表明:

高杠杆点(h_{ii}值较大的点) 往往有较小的残差，因此，高杠杆点不能通过残差检测出来, 需要依据杠杆值h_{ii}来判断.

- 平均杠杆值为

$$\bar{h} = \frac{1}{n}\sum_{j=1}^{n} h_{jj} = \frac{p}{n}.$$

- 若$\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_n$是该正态总体的一个简单随机样本, 则

$$F = \frac{(n-p)(h_{ii} - 1/n)}{(p-1)(1-h_{ii})} \sim F_{p-1,n-p}.$$

高杠杆点方法

1. 杠杆值的2倍平均值法

Hoaglin 和Welsch(1978)提出了杠杆值的2倍平均值法，即将

$$h_{ii} > \frac{2p}{n} \tag{6.23}$$

的点视为高杠杆点.

- 当p和$n-p$较小时，杠杆值的2倍平均值法常常会失效.

2. F 检验法

统计量F为h_{ii}单调增函数，所以h_{ii}很大等价于F很大. 当$F > F_{p-1,n-p}(\alpha)$时，就认为h_{ii}很大，对应的点$\boldsymbol{x}_i$就是高杠杆点, 通常取$\alpha = 0.05$.

- 当$p > 10$, $n - p > 50$, 两方法等价. 事实上, 从F分布表知

$$P(F_{p-1,n-p} \leq 2) \geq 0.95,$$

而$F_{p-1,n-p} \leq 2$等价于$h_{ii} > 2p/n$.

- 或更直观的方法借助于画杠杆值图, 如顺序图、点图或箱线图.

异常点的诊断

异常点

一组数据$(\boldsymbol{x}_i', y_i)$如果它的残差($\widehat{e}_i$或r_i)较其它组数据的残差大得多, 则称此数据为异常点

- 正态线性回归模型

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

这里$e_i, i = 1, \cdots, n$ 相互独立.

- 均值漂移 假设异常点处$E(y_j) = \boldsymbol{x}_j'\boldsymbol{\beta} + \eta.$

如果第j组数据$(\boldsymbol{x}_j', y_j)$是一个异常点, 那么它的残差之所以很大是因为它的均值$E(y_j)$发生了非随机漂移η

异常点的诊断

均值漂移线性回归模型

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

矩阵形式

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{d}_j\eta + \boldsymbol{e}, \quad \boldsymbol{e} \sim N(\mathbf{0},\ \sigma^2\mathbf{I_n}), \tag{6.24}$$

其中$\boldsymbol{d}_j = (0, \cdots, 0, 1, 0, \cdots, 0)'$, 这是一个$n$维向量, 它的第$j$个元素为1, 其余元素为零.

- 要判定$(\boldsymbol{x}_j', y_j)$不是异常点, 等价于检验假设

$$H_0 : \eta = 0.$$

定理6.5.2

对均值漂移线性回归模型(6.24), β 和η 的LS估计分别为

$$\beta^* = \widehat{\beta}_{(j)} = \widehat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_j\hat{e}_j}{1-h_{jj}},$$

$$\eta^* = \frac{1}{1-h_{jj}}\widehat{e}_j,$$

这里, $\widehat{\beta}_{(j)}$ 为剔除第j 组数据模型

$$\boldsymbol{y}_{(j)} = \mathbf{X}_{(j)}\beta + \boldsymbol{e}_{(j)}$$

的β 的LS 估计, h_{jj} 为$\mathbf{H} = \mathbf{P}_{\mathbf{X}}$的第$j$个主对角元, $\widehat{e}_j$为从零假设模型导出的第j个残差.

定理6.5.2的证明提要

- 由$\mathbf{X}'_{(j)}\mathbf{X}_{(j)} = \mathbf{X}'\mathbf{X} - \boldsymbol{x}_j\boldsymbol{x}'_j$ 可得

$$(\mathbf{X}'_{(j)}\mathbf{X}_{(j)})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_j\boldsymbol{x}'_j(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{jj}}$$

故$\widehat{\boldsymbol{\beta}}_{(j)} = (\mathbf{X}'_{(j)}\mathbf{X}_{(j)})^{-1}\mathbf{X}_{(j)}\boldsymbol{y}_{(j)} = \widehat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_j\hat{e}_j}{1-h_{jj}}$.

- 由$\boldsymbol{d}'_j\boldsymbol{y} = y_j$, $\boldsymbol{d}'_j\boldsymbol{d}_j = 1$, 均值漂移线性回归模型的LS估计

$$\begin{pmatrix} \boldsymbol{\beta}^* \\ \eta^* \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \boldsymbol{x}_j \\ \boldsymbol{x}'_j & 1 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}'\boldsymbol{y} \\ y_j \end{pmatrix}.$$

由分块矩阵的逆可证$\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}}_{(j)}$.

异常点的诊断方法

判断y_j 是否是异常点等价于等价于检验

$$H_0 : \eta = 0.$$

- H_0成立下，约简模型的残差平方和：$\mathrm{SSE}_H = \boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\boldsymbol{y}$.
- 均值漂移模型的残差平方和：

$$\mathrm{SSE} = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}^{*\prime}\mathbf{X}'\boldsymbol{y} - \eta^*\boldsymbol{d}_j'\boldsymbol{y} = (n-p)\widehat{\sigma}^2 - \frac{\widehat{e}_j^2}{1-h_{jj}}$$

- 检验统计量

$$F = \frac{\mathrm{SSE}_H - \mathrm{SSE}}{\mathrm{SSE}/(n-p-1)} = \frac{(n-p-1)r_j^2}{n-p-r_j^2}, \quad r_j\text{标准化学生残差}$$

异常点的诊断方法

定理6.5.3

对于均值漂移线性回归模型(6.24), 如果假设$H:\eta=0$成立, 则

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

- 对给定的$\alpha(0<\alpha<1)$, 若

$$F_j=\frac{(n-p-1)r_j^2}{n-p-r_j^2}>F_{1,\,n-p-1}(\alpha), \tag{6.25}$$

则判定第j组数据$(\boldsymbol{x}_j',y_j)$为异常点.

- 以上方法仅适用于因变量的观测值中存在一个异常点的情形.

异常点的诊断方法

- 多个异常点的检验是一个复杂问题.

  因为异常点往往把回归方程拉向自身，从而使得其他点远离拟合方程. 这就会导致检测出的某些异常点并不是真正的异常点，而真正的异常点可能被淹没未被检测出来.

- 对异常点的识别中的伪装和掩没问题，参阅
  - Barnett, V. and Lewis, T. Outlier in Statistical Data, New York: John Wiley, 1978.
  - Hadi, A. S., Simonoff, J. S. Procedures for the identification of multiple outliers in linear models. Journal of the American Statistical Association, 1993, 88: 1264-1272.

强影响点的诊断

强影响点的诊断问题, 即探查对估计或预测有较大影响的数据.

记剔除第i组数据后, 剩余的$n-1$组数据的线性回归模型

$$\boldsymbol{y}_{(i)} = \mathbf{X}_{(i)}\boldsymbol{\beta} + \boldsymbol{e}_{(i)}, \quad E(\boldsymbol{e}_{(i)}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}_{(i)}) = \sigma^2\mathbf{I}_{n-1},$$

相应的LS估计

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\boldsymbol{y}_{(i)}.$$

- 向量$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}$反映了第$i$组数据对回归系数估计的影响大小.
- 用向量$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}$的某种数量化函数来定量比较影响的大小.

强影响点的诊断

1. Cook距离统计量

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{p\widehat{\sigma}^2}, \quad i = 1, \cdots, n.$$

- 若$D_i > F_{p,\, n-p}(\alpha)$, 则认为第$i$组数据的为强影响点.

  $\boldsymbol{\beta}$的置信系数为$1-\alpha$置信椭球

  $$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p\widehat{\sigma}^2} \leq F_{p,n-p}(\alpha).$$

  将$\widehat{\boldsymbol{\beta}}_{(i)}$ 代替$\boldsymbol{\beta}$, 就得到了Cook统计量.

在实际操作时，往往把$D_i > 1$的点视为潜在的强影响点，再结合D_i的点图或顺序图来判断.

Cook统计量

- Cook统计量D_i又刻画了第i组数据对模型拟合的影响，因为

$$D_i = \frac{(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)})'(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)})}{p\hat{\sigma}^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_{i(i)})^2}{p\hat{\sigma}^2}.$$

- D_i的简便计算公式

$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right)r_i^2, \quad i = 1, \cdots, n.$$

该公式将Cook统计量D_i分解成两部分:

- 势位函数: $P_i = \frac{h_{ii}}{1-h_{ii}}$, ($h_{ii}$的单调增函数);
- 学生化残差的平方 r_i^2.

高杠杆点和异常点都可能是强影响点，但又不一定都是强影响点

2. Welsch-Kuh 统计量(DFFITS)

$$W_i = \frac{\hat{y}_i - \hat{y}_i(i)}{(\widehat{\mathrm{Var}}(\hat{y}_i))^{1/2}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} r_i^*,$$

这里$\hat{y}_i(i) = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$, $r_i^* = \hat{e}_i/(\hat{\sigma}_{(i)}\sqrt{1-h_{ii}})$是的标准化预测残差.

- Welsch-Kuh(1977) 从预测(或单值拟合)的角度提供了另一种准则: DFFITS（difference in fits）准则. W_i^2 度量了第i组数据对$\boldsymbol{x}_i$处的预测影响大小.
- 问题: W_i^2与第i 组数据对其他点$\boldsymbol{x}$ 处的预测影响关系如何？

强影响点的诊断

定理6.5.5

(1)

$$W_i^2 = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{\widehat{\sigma}^2_{(i)}}; \quad (\text{与} D_i \text{仅是分母不同})$$

(2) W_i^2为第i组数据对任意点$\boldsymbol{x}$ 处预测影响的上界: 对于任意的$\boldsymbol{x}$, 有

$$\frac{(\boldsymbol{x}'\widehat{\boldsymbol{\beta}} - \boldsymbol{x}'\widehat{\boldsymbol{\beta}}_{(i)})^2}{\widehat{\sigma}^2_{(i)}\boldsymbol{x}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}} \leq W_i^2.$$

- 在应用中, 需根据样本量设定判别门限：

  判别门限为$|W_i| > 1$ (小、中等样本)；判别门限为$|W_i| > 2\sqrt{\frac{p}{n-p}}$ (大样本) .

- 可用W_i^2的顺序图、点图或箱线图等图工具识别强影响点.

3. 协方差比(covariance ratio, COVRATIO)

$$C_i = \frac{\det\left(\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{(-i)})\right)}{\det\left(\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\right)} = \frac{(\hat{\sigma}^2_{(-i)})^p}{(\hat{\sigma}^2)^p}\frac{1}{1-h_{ii}}.$$

- 如果第i点观测值所对应的C_i值离1越远, 则认为该点影响越大.

在R语言中函数influence.measures()可以直接计算

杠杆值h_{ii}, Cook距离D_i,

Welsch-Kuh 统计量W_i

COVRATIO 统计量C_i.

影响分析案例

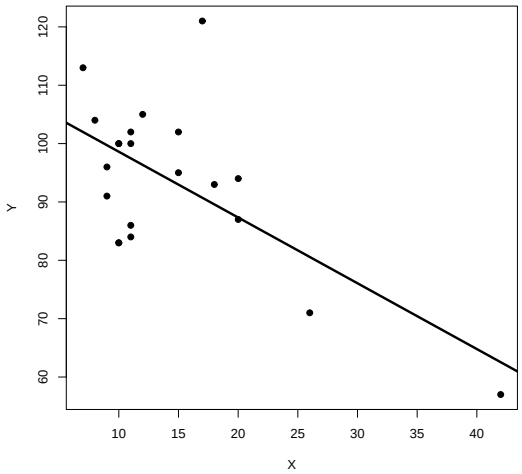
智力测试数据

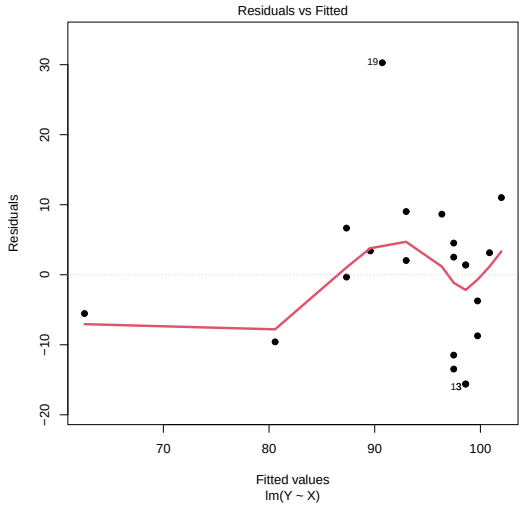
下表是教育学家测试的21个儿童的记录, 其中X为儿童的年龄(以月为单位), y表示某种智力指标, 通过这些数据, 我们要建立智力随年龄变化的关系.

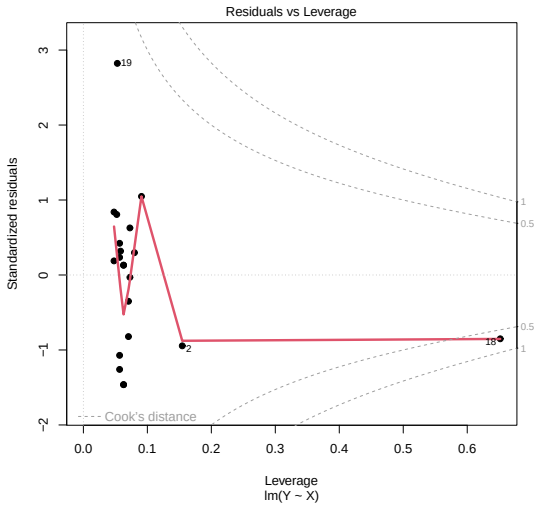
i	x	y	i	x	y	i	x	y
1	15	95	8	11	100	15	11	102
2	26	71	9	8	104	16	10	100
3	10	83	10	20	94	17	12	105
4	9	91	11	7	113	18	42	57
5	15	102	12	9	96	19	17	121
6	20	87	13	10	83	20	11	86
7	18	93	14	11	84	21	10	100

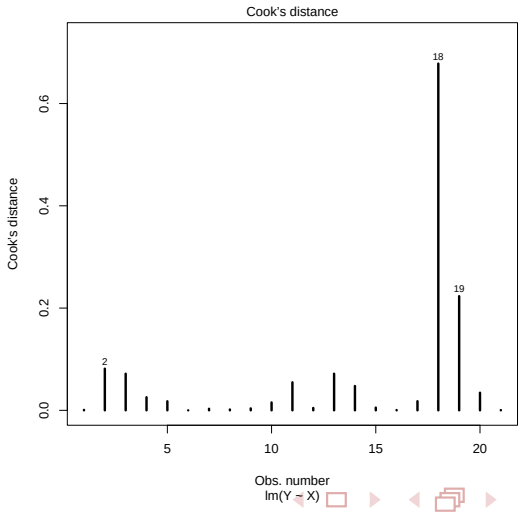
影响分析案例

经验回归直线为 $\widehat{Y} = 109.87 - 1.13X$









影响分析案例

- 计算$D_{18} = 0.6781$ 是所有D_i 中最大的, 而其它D_i值与D_{18} 相比也十分小. 因此, 第18号数据是一个对回归估计影响很大的数据.

  对给定的水平$\alpha = 0.05$, 只有$t_{19} = 3.6071 > t_{18}(0.025) = 2.101$.

  于是, 可认为第19号数据为异常点.

  计算W_i. W_{18} 和W_{19} 超过

$$2\sqrt{p/(n-p)} = 2\sqrt{2/(21-2)} = 0.6489.$$

  这两点的Cook距离C_{18}和C_{19}比其他点的大很多, 且远离1.

  综上, 可判断18号和19号组数据也是强影响点, 应被认真考察.

影响分析的注意事项

注意:

- 影响分析只是研究探查强影响数据的统计方法, 至于对已经确认的强影响数据如何处理, 这需要具体问题具体分析.
- 往往先要仔细核查数据获得全过程, 如果强影响数据是由于试验条件失控或纪录失误或其他一些过失所致, 那么这些数据应该剔除. 不然的话, 应该考虑收集更多的数据或采用一些稳健估计方法以缩小强影响数据对估计的影响, 从而获得较稳定的经验回归方程.

强影响的修正方法–稳健回归

LS估计对误差的正态假设、异常值和强影响点是敏感的. 处理异常点和强影响观测的常见方法：

- LS的方法就是删除异常值和强影响点, 用剩下的数据作回归.
- 稳健回归, 它是对高杠杆点赋予较小的权重的一种拟合方法.

  常见的稳健回归方法有
  - 最小绝对离差(least absolute deviation, LAD)估计
  - M-估计
  - 加权最小二乘（WLS）估计

三种稳健回归方法：

- LAD 估计

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i\boldsymbol{\beta}|.$$

- M-估计（当前最为流行的）

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i\boldsymbol{\beta}),$$

其中$\rho(\cdot)$为一选定的非负函数.

- WLS估计(第j次迭代)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}^{(j)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(j)}\boldsymbol{y},$$

其中$\mathbf{W}^{(j)} = \text{diag}(w_1^{(j)}, \cdots, w_n^{(j)})$, $w_i^{(j)}$ 为第j次迭代中第i观测的权.

稳健回归

文献中, 关于权函数w_i的选择有

- Chatterjee和Mächler (1997) 提出

$$w_i^{(j)} = \frac{(1-h_{ii})^2}{\max(|\widehat{e}_i^{(j-1)}|, m_e^{(j-1)})},$$

其中$m_e^{(j-1)}$ 是$|\widehat{e}_1^{(j-1)}|, \ldots, |\widehat{e}_n^{(j-1)}|$的中位数, 初始权重设为$w_i^{(0)} = \max(h_{ii}, (p-1)/n)$.

- DPS软件提供了Cauthy方法、Andrew方法等10 种不同权.

注1　稳健性与参数估计的最优性不同，不能说愈稳健就愈好.

注2　正态分布下, 过于强调稳健性会导致效率损失, 故应用中需平衡两者关系.

复共线性

Gauss-Markov定理保证了LS估计在线性无偏估计类中的方差最小性, 但大型线性回归问题中, LS估计有时表现不理想. 例如,

- 有时某些回归系数的估计的绝对值异常大,
- 有时回归系数的估计值的符号与问题的实际意义相违背等.

大量研究发现: 产生这些问题的原因之一是回归自变量之间存在着近似线性关系, 称为复共线性(multicolinearity).

本节研究

- 复共线性对LS估计的影响
- 复共线性的诊断
- 几类修正估计方法

均方误差准则

评价一个估计优劣的标准：

均方误差(Mean squares error, MSE)

$\widehat{\theta}$的均方误差为

$$\mathrm{MSE}(\widehat{\theta}) = E\|\widehat{\theta} - \theta\|^2 = E(\widehat{\theta} - \theta)'(\widehat{\theta} - \theta).$$

- 它度量了估计$\widehat{\theta}$与未知参数向量θ的平均偏离的大小, 一个好的估计应该有较小的均方误差.
- 这里参数向量θ的估计$\widehat{\theta}$不限于无偏估计.

复共线性对LS估计的影响

定理6.6.1

$$\mathrm{MSE}(\widehat{\boldsymbol{\theta}}) = \mathrm{trCov}(\widehat{\boldsymbol{\theta}}) + \|E(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\|^2,$$

这里$\mathrm{tr}(\mathbf{A})$表示$\mathbf{A}$的迹.

- $\widehat{\boldsymbol{\theta}}$的均方误差可以分解为两项之和, 其中一项为$\widehat{\boldsymbol{\theta}}$的各分量的方差之和, 另一项为$\widehat{\boldsymbol{\theta}}$的各分量的偏差的平方和.
- 一个估计的均方误差就是由它的各分量的方差和偏差所决定的. 一个好的估计应该有较小的方差和偏差.

复共线性对LS估计的影响

用均方误差这个标准来评价LS估计. 考虑线性回归模型

$$\boldsymbol{y} = \beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_n, \tag{6.26}$$

这里, $\mathbf{X} = (x_{ij})$ 已被标准化, 且$\mathrm{rk}(\mathbf{X}) = p - 1$. 由于$\mathbf{X}$是中心化的, 于是常数项$\beta_0$ 和回归系数$\boldsymbol{\beta}$的LS估计:

$$\widehat{\beta}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = \mathbf{R}_{\mathbf{X}}^{-1}\mathbf{X}'\boldsymbol{y},$$

其中$\mathbf{R}_{\mathbf{X}} = \mathbf{X}'\mathbf{X}$为变量$(X_1, \cdots, X_{p-1})$的样本相关系数矩阵.

复共线性对LS估计的影响

因为$\widehat{\beta}$是β的无偏估计, 所以

$$\mathrm{MSE}(\widehat{\beta}) = \mathrm{tr}(\mathrm{Cov}(\widehat{\beta})) = \sigma^2 \mathrm{tr}(\mathbf{R}_{\mathbf{X}}^{-1}) = \sigma^2 \sum_{i=1}^{p-1} \frac{1}{\lambda_i}, \tag{6.27}$$

其中$\lambda_1 \geq \cdots \geq \lambda_{p-1} > 0$为$\mathbf{R}_X$的特征值.

- 如果$\mathbf{R}_{\mathbf{X}}$至少有一个特征值非常接近于零, 则$\mathrm{MSE}(\widehat{\beta})$就会很大. 此时从MSE准则来看, LS估计$\widehat{\beta}$就不是一个好的估计.

这一点和Gauss-Markov定理并无抵触, 因为Gauss-Markov定理仅仅保证了LS估计在线性无偏估计类中的方差最小性, 但在$\mathbf{R}_{\mathbf{X}}$至少有一个特征值很小时, 这个最小的方差值本身却很大, 因而导致了很大的均方误差.

复共线性对LS估计的影响

- 只要$\mathbf{R_X}$有一个特征值很小, LS估计$\widehat{\beta}$的模长平均说来要比真正的β的模长大得多. 这就导致了 $\widehat{\beta}$的某些分量的绝对值太大.

  事实上, 由于

$$\mathrm{MSE}(\widehat{\beta}) = E((\widehat{\beta}-\beta)'(\widehat{\beta}-\beta)) = E\|\widehat{\beta}\|^2 - \|\beta\|^2,$$

  于是

$$E\|\widehat{\beta}\|^2 = \|\beta\|^2 + \mathrm{MSE}(\widehat{\beta}) = \|\beta\|^2 + \sigma^2 \sum_{i=1}^{p-1} \frac{1}{\lambda_i}.$$

当$\mathbf{R_X}$至少有一个特征值很小时, LS估计$\widehat{\beta}$就不再是一个好的估计.

复共线性的诊断

$\mathbf{R_X}$至少有一个特征值很小对设计阵$\mathbf{X}$意味着什么?

记$\mathbf{X} = (\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(p-1)})$, $\boldsymbol{x}_{(i)}$ 为设计阵$\mathbf{X}$的第i列. 设λ为$\mathbf{R_X}$的一个特征值, $\boldsymbol{\varphi} = (c_1, \cdots, c_{p-1})'$为其对应的标准化的特征向量. 若$\lambda \approx 0$, 则

$$\mathbf{R_X}\boldsymbol{\varphi} = \lambda\boldsymbol{\varphi} \approx \mathbf{0}.$$

用$\boldsymbol{\varphi}'$左乘上式, 得$\boldsymbol{\varphi}'\mathbf{R_X}\boldsymbol{\varphi} = \lambda\boldsymbol{\varphi}'\boldsymbol{\varphi} = \lambda \approx 0$. 于是, 有$\mathbf{X}\boldsymbol{\varphi} \approx \mathbf{0}$, 即

$$c_1\boldsymbol{x}_{(1)} + \cdots + c_{p-1}\boldsymbol{x}_{(p-1)} \approx 0.$$

这表明设计阵$\mathbf{X}$的列向量$\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(p-1)}$之间有近似的线性关系.

复共线性的诊断

复共线关系

称回归设计阵的列向量之间的关系

$$c_1\boldsymbol{x}_{(1)} + \cdots + c_{p-1}\boldsymbol{x}_{(p-1)} \approx 0 \tag{6.28}$$

为复共线关系. 相应地, 称设计阵$\mathbf{X}$或线性回归模型(6.26)存在复共线性, 有时也称设计阵X是病态的(ill-conditioned).

- $\mathbf{R_X}$的最大特征值与最小特征值之比就是度量多重共线性严重程度的一个重要指标.

条件数(conditional numbers)

矩阵$\mathbf{R_X}$的条件数(conditional numbers):

$$\kappa = \frac{\lambda_1}{\lambda_{p-1}},$$

其中λ_1和λ_2分别是矩阵$\mathbf{R_X}$的最大和最小特征值.

- 条件数刻画了$\mathbf{R_X}$的特征值差异性的大小, 可用来判断复共线性是否存在以及复共线性严重程度.
  - 一般若$k < 100$, 则认为复共线性的程度很小;
  - 若$100 \leq k \leq 1000$, 则认为存在中等程度或较强的复共线性;
  - 若$k > 1000$, 则认为存在严重的复共线性.

R语言提供了计算条件数的函数kappa().

方差扩大(膨胀)因子

故把矩阵$\mathbf{R}_{\mathbf{X}}^{-1}$中第$k$个对角线元素称为方差扩大(膨胀)因子(variance inflation factor, VIF), 记为$\mathrm{VIF}(\widehat{\alpha}_k)$, 其中$k = 1, \cdots, p-1$.

方差扩大(膨胀)因子的名称来源于

$$\mathrm{Var}(\widehat{\beta}_k) = \sigma^2 (\mathbf{R}_{\mathbf{X}}^{-1})_{kk},$$

其中$(\mathbf{R}_{\mathbf{X}}^{-1})_{kk}$为相关系数矩阵$(\mathbf{R}_{\mathbf{X}})^{-1}$ 中第k个对角线元素的乘积.

- 经验法则：当VIF的值超过5 或10, 则认为线性回归模型存在多重共线性.

可用R语言中程序包car中的函数vif()计算.

- 可以证明：

$$\mathrm{VIF}(\widehat{\beta}_k) = \frac{1}{1 - R^2_{X_k|\boldsymbol{X}_{(-k)}}}, \qquad k = 1, \cdots, p,$$

其中$R_{Z_k|\boldsymbol{Z}_{(-k)}} = R_{X_k|\boldsymbol{X}_{(-k)}}$是第$k$个协变量$X_k$与其余的$p-2$个协变量$\boldsymbol{X}_{(-k)}$之间的复相关系数.

  - 当第k个协变量与其余的协变量之间相关程度越高, 即$R^2_{X_k|\boldsymbol{X}_{(-k)}}$越接近于1时，$\mathrm{VIF}(\widehat{\beta}_k)$越大;
  - 第k个协变量与其余的协变量之间相关程度越小, $\mathrm{VIF}(\widehat{\beta}_k)$越小, 当表示完全不存在多重共线性时, VIF取得最小值1.

因此, 方差扩大因子VIF成为是衡量线性回归模型中自变量间的多重共线性严重程度的又一种度量.

例6.6.1

表中原始数据共有12 组数据, 除第一组外, 协变量$X_1, X_2, \cdots, X_6$ 的其余11 组数据满足线性关系: $X_1 + X_2 + X_3 + X_4 = 10$, 试用求矩阵条件数和方差膨胀因子判断自变量间是否存在多重共线性.

序号	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	10.006	8	1	1	1	0.541	−0.099
2	9.737	8	1	1	0	0.130	0.070
3	15.087	8	1	1	0	2.116	0.115
4	8.422	0	0	9	1	−2.397	0.252
5	8.625	0	0	9	1	−0.046	0.017
6	16.289	0	0	9	1	0.365	1.504
7	5.958	2	7	0	1	1.996	−0.865
8	9.313	2	7	0	1	0.228	−0.055
9	12.960	2	7	0	1	1.380	0.502
10	5.541	0	0	0	10	−0.798	−0.399
11	8.756	0	0	0	10	0.257	0.101
12	10.937	0	0	0	10	0.440	0.432

例6.6.1

- 计算$(\bar{x}_{.1}, \bar{x}_{.2}, \bar{x}_{.3}, \bar{x}_{.4}\bar{x}_{.5}, \bar{x}_{.6}) = (2.5, 2.0, 2.5, 3.0833, 0.351, 0.1313)$,

$$(s_1, s_2, s_3, s_4, s_5, s_6) = (3.4245, 3.0451, 3.9428, 41878, 1.2070, 0.5646),$$

$$\mathbf{R}_X = \begin{bmatrix} 1.000 & 0.052 & -0.343 & -0.498 & 0.417 & -0.192 \\ & 1.000 & -0.432 & -0.371 & 0.485 & -0.317 \\ & & 1.000 & -0.355 & -0.505 & 0.494 \\ & & & 1.000 & -0.215 & -0.087 \\ & & & & 1.000 & -0.123 \\ & & & & & 1.0000 \end{bmatrix}.$$

- kappa(RX, exact=TRUE)
  得到的条件数是$\kappa = 2195.908 > 1000$，认为有严重的多重共线性.
- 用函数vif()计算方差膨胀因子
  lm.fit = lm(Y ., data=collinear); round(vif(lm.fit), 3)

X1	X2	X3	X4	X5	X6
182.052	161.362	266.264	297.715	1.920	1.455

例6.6.1

- 计算矩阵$\mathbf{R_X}$的最小特征值和相应的特征向量, 得

$$\lambda_{\min} = 0.001106,$$

$$\phi = (0.4477, 0.4211, 0.5417, 0.5734, 0.0061, 0.0022)'.$$

于是$0.4477Z_1 + 0.4211Z_2 + 0.5417Z_3 + 0.5734Z_4 + 0.0061Z_5 + 0.0022Z_6 \approx 0$,
这里, $Z_j = (X_j - \bar{x}_{.j})/s_j$. 由于$Z_5$和$Z_6$的系数近似为0，因此有

$$0.4477Z_1 + 0.4211Z_2 + 0.5417Z_3 + 0.5734Z_4 \approx 0. \tag{6.29}$$

还原变量得

$$0.1308X_1 + 0.1383X_2 + 0.1374X_3 + 0.1369X_4 \approx 1.3691$$

与题目中给的变量关系$X_1 + X_2 + X_3 + X_4 = 10$ 大致相同.

复共线性产生的原因

- 由于数据“收集"的局限性所致.

  虽然这样产生的复共线性是非本质的, 原则上可以通过“收集"更多的数据来解决, 但具体实现起来会遇到许多困难. 例如,

  - 由于试验或生产过程已经完结或经费限制,不可能再产生新的数据.
  - 对于多于三个自变量的情况, 往往难于确定“收集"怎样的数据, 才能“打破" 复共线性.
  - 即便收集了一些新的数据, 但为了打破复共线性, 这些数据势必要远离原来的数据, 可能产生强影响点, 从而产生新问题.

- 由于自变量之间客观上就有近似的线性关系.

  由于人们往往对自变量之间的关系缺乏认识, 很可能把一些有复共线关系的自变量引入回归方程.

针对复共线的处理

当设计阵存在复共线关系时, LS估计不够理想, 甚至很坏.

Stein现象

Stein于1955年证明了, 当维数大于2时, 正态均值向量的LS估计的不可容许性, 即能够找到另外一个估计在某种意义下一致优于LS估计.

针对复共线的处理办法：

- 从模型或数据角度去考虑, 如变量选择和回归诊断.
- 寻求一些新的（有偏）估计方法.
  - 岭估计
  - 主成分估计

针对复共线的处理

岭估计

对于线性回归模型(6.26), 回归系数β的岭估计定义为

$$\widehat{\beta}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_{p-1})^{-1}\mathbf{X}'\boldsymbol{y} \tag{6.30}$$

这里$k > 0$是可选择参数, 称为岭参数或偏参数.

- 如果k取与试验数据y无关的常数, 则$\widehat{\beta}(k)$为线性估计, 不然的话, $\widehat{\beta}(k)$就是非线性估计.
- 岭估计$\widehat{\beta}(k)$是一个估计类, 不同的k得到不同的估计.
  - 取$k = 0$, $\widehat{\beta}(0) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$是通常的LS估计.
    但一般约定：当提到岭估计时, 不包括LS估计.

岭估计的性质

- 岭估计是β的有偏估计: $E(\widehat{\beta}(k)) = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_{p-1})^{-1}\mathbf{X}'\mathbf{X}\beta \neq \beta$.
- 岭估计$\widehat{\beta}(k)$ 的模长总比LS估计$\widehat{\beta}$ 的模长小.

对一切$k > 0$和$\widehat{\beta} \neq 0$, 有

$$\|\widehat{\beta}(k)\| = \|(\mathbf{X}'\mathbf{X} + k\mathbf{I}_{p-1})^{-1}\mathbf{X}'\mathbf{X}\widehat{\beta}\| = \| \left(\mathbf{I}_{p-1} - k(\mathbf{X}'\mathbf{X} + k\mathbf{I}_{p-1})^{-1}\right) \widehat{\beta}\| < \|\widehat{\beta}\|.$$

岭估计是一种压缩估计(shinked estimator), 将LS估计向原点压缩.

- 岭估计是惩罚最小二乘目标函数

$$\|\boldsymbol{y} - \mathbf{X}\beta\|^2 + k\|\beta\|^2 \tag{6.31}$$

的极小值点.

岭估计的性质

注1 第二条性质从一个侧面说明了当设计阵$\mathbf{X}$ 呈病态时岭估计的合理性. 因为此时LS估计的分量有偏大的趋势, 对它作适当的压缩是很有必要的.

注2 第三条性质表明岭估计的本质就是约束模长的最小二乘估计. 因为极小化惩罚最小二乘目标函数(6.31)，等价于约束的最小二乘问题：

$$\begin{cases} \min\limits_{\boldsymbol{\beta}} \dfrac{1}{n}\|\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2, \\ \text{s.t.} \quad \|\boldsymbol{\beta}\|^2 \le c, \end{cases}$$

其中c是非负常数, 作用与岭参数(调节参数) k 相同.

岭估计的性质

- 与LS估计$\widehat{\beta}$ 相比, 岭估计是把矩阵$\mathbf{X}'\mathbf{X}$ 换成矩阵$\mathbf{X}'\mathbf{X}+k\mathbf{I}_{p-1}$.
  - 直观上看这样作的理由也是明显的.
    因为当$\mathbf{X}$呈病态时, $\mathbf{X}'\mathbf{X}$的特征值至少有一个非常接近于零, 而$\mathbf{X}'\mathbf{X}+k\mathbf{I}_{p-1}$的特征值

$$\lambda_1+k,\cdots,\lambda_{p-1}+k$$

    接近于零的程度就会得到改善, 从而“打破"原来设计阵的复共线性, 使岭估计比LS估计有较小的均方误差.

定理6.7.1

存在$k>0$, 使得在均方误差意义下, 岭估计优于LS估计, 即

$$\mathrm{MSE}(\widehat{\beta}(k))<\mathrm{MSE}(\widehat{\beta}).$$

定理6.7.1的证明

证明 对$\mathbf{X}'\mathbf{X}$进行谱分解, 得

$$\mathbf{X}'\mathbf{X} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}',$$

其中$\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_{p-1})$, $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_{p-1})$, $\boldsymbol{\phi}_i$为$\mathbf{X}'\mathbf{X}$的特征值λ_i对应的标准化正交化的特征向量. 于是,

$$E(\widehat{\boldsymbol{\beta}}(k)) = \mathbf{\Phi}(\mathbf{\Lambda} + k\mathbf{I}_{p-1})^{-1}\mathbf{\Lambda}\mathbf{\Phi}'\boldsymbol{\beta},$$

$$\text{Cov}(\widehat{\boldsymbol{\beta}}(k)) = \sigma^2\mathbf{\Phi}(\mathbf{\Lambda} + k\mathbf{I}_{p-1})^{-1}\mathbf{\Lambda}(\mathbf{\Lambda} + k\mathbf{I}_{p-1})^{-1}\mathbf{\Phi}'.$$

再依定理6.6.1和矩阵迹运算的性质：$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, 得

$$\begin{aligned}\text{MSE}(\widehat{\boldsymbol{\beta}}(k)) &= \text{tr}\left(\text{Cov}(\widehat{\boldsymbol{\beta}}(k))\right) + \|E(\widehat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta}\|^2 \\ &= \sigma^2\sum_{i=1}^{p-1}\frac{\lambda_i}{(\lambda_i + k)^2} + k^2\sum_{i=1}^{p-1}\frac{\alpha_i^2}{(\lambda_i + k)^2} = f_1(k) + f_2(k) = f(k)\end{aligned}$$

定理6.7.1的证明

这里$\alpha_i = \phi_i'\beta$, $i = 1, \ldots, p-1$. 对$k > 0$, $f_1(k)$和$f_2(k)$存在连续的一阶导数, 且

$$f_1'(k) = -2\sigma^2 \sum_{i=1}^{p-1} \frac{\lambda_i}{(\lambda_i + k)^3}, \quad f_2'(k) = 2k \sum_{i=1}^{p-1} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}.$$

故$f'(k) = f_1'(k) + f_2'(k)$在$k \geq 0$时也连续. 注意到

$$f'(0) = f_1'(0) + f_2'(0) < 0,$$

故当$k > 0$且k充分小时, $f'(k) < 0$, 即$f(k)$ 严格单调函数. 因而存在$k_0 > 0$, 当$k \in (0,\ k_0)$时, 有

$$\mathrm{MSE}(\widehat{\beta}(k)) = f(k) < f(0) = \mathrm{MSE}(\widehat{\beta}).$$

定理证毕.

岭迹参数的选择方法

定理6.7.1只是表明：在均方误差意义下，优于LS估计的岭估计存在, 但不能求出最优解, 因为方程

$$f'(k) = 0$$

的最优值k^*依赖于未知参数β和σ^2.

对于给定的k, 记

$$\mathrm{Cov}(\widehat{\beta}(k)) = \sigma^2 \mathbf{D}(k) = \sigma^2 (d^{ij}(k)).$$

岭迹参数的选择方法

常用的4种方法:

1. Hoerl-Kennard公式: $\widehat{k} = \frac{\widehat{\sigma}^2}{\max_i \widehat{\alpha}_i^2}$.
2. 方差扩大因子: 使所有方差扩大因子$d^{jj}(k) < 10$.
3. 岭迹法: 选择k值, 使各回归系数的岭估计$\widehat{\beta}_j(k)$ 的岭迹大体稳定, 符号合理.
4. 广义交叉验证法: 极小化下面的广义交叉验证目标函数

$$\widehat{k}_{\text{gcv}} = \arg\min_k \text{GCV}(k) = \arg\min_k \frac{\|(\mathbf{I}_n - \mathbf{H}(k))\mathbf{y}\|_2^2/n}{[\text{tr}(\mathbf{I}_n - \mathbf{H}(k))/n]^2},$$

其中$\mathbf{H}(\lambda) = \mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'$.

岭估计

- R语言中，
  - 程序包MASS中的函数lm.ridge()、
  - 程序包ridge中的函数linearRidge()、
  - 程序包glmnet中的函数glmnet()

  都可以实现岭回归，其中在函数glmnet() 中，alpha=0，拟合岭回归模型, 函数cv.glmnet()是cv方法选择最优岭参数.

注 岭回归的优势是平衡了偏差和方差，随着λ的增加，岭回归拟合的光滑度降低，尽管方差变小，但是偏差变大.

注 如果$p > n$，则最小二乘没有唯一解，此时岭回归仍然能通过偏差小幅度的增加来换取方差大幅度的下降，通过这种权衡获得比较好的模型效果.

例6.6.2：外贸数据分析

因变量Y为进口总额, 自变量X_1为国内总产值, X_2为存储量, X_3为总消费量. 收集了11组数据, 列在下表.

序号	国内总产值(X_1)	存储量(X_2)	总消费量(X_3)	进口总额(y)
1	149.3	4.2	108.1	15.9
2	161.2	4.1	114.8	16.4
3	171.5	3.1	123.2	19.0
4	175.5	3.1	126.9	19.1
5	180.8	1.1	132.1	18.8
6	190.7	2.2	137.7	20.4
7	202.1	2.1	146.0	22.7
8	212.4	5.6	154.1	26.5
9	226.1	5.0	162.3	28.1
10	231.9	5.1	164.3	27.6
11	239.0	0.7	167.6	26.3

例6.6.2

解 将原始数据标准化, 计算得相关系数矩阵为

$$\mathbf{R_X} = \begin{bmatrix} 1 & 0.026 & 0.997 \\ 0.026 & 1 & 0.036 \\ 0.997 & 0.036 & 1 \end{bmatrix},$$

它的三个特征值为$\lambda_1 = 1.999, \lambda_2 = 0.998, \lambda_3 = 0.003.$

$\mathbf{R_X}$的条件数

$$\frac{\lambda_1}{\lambda_3} = 666.333 < 1000,$$

可知设计阵存在中等程度的复共线性.

λ_3对应的特征向量为 $\phi_3 = (-0.7070, \ -0.0070, \ 0.7072)$

例6.6.2

- 三个自变量之间存在复共线关系

$$-0.7070X_1 - 0.0070Z_2 + 0.7072Z_3 \approx 0.$$

注意到, Z_2的系数绝对值相对非常小, 可视为零, 而Z_1和Z_3的系数又近似相等, 故自变量之间的复共线关系可近似为

$$Z_1 = Z_3, \quad \text{即} \quad \frac{X_1 - \bar{x}_1}{s_1} = \frac{X_3 - \bar{x}_3}{s_3}.$$

可算出

$$\bar{x}_1 = 194.59, \quad s_1 = \left(\frac{1}{10} \sum_{i=1}^{11} (x_{i1} - \bar{x}_1)^2 \right)^{1/2} = 30.00,$$

$$\bar{x}_3 = 139.74, \quad s_1 = \left(\frac{1}{10} \sum_{i=1}^{11} (x_{i3} - \bar{x}_3)^2 \right)^{1/2} = 20.63.$$

例6.6.2

代入上式得

$$X_3 = 5.905 + 0.688X_1.$$

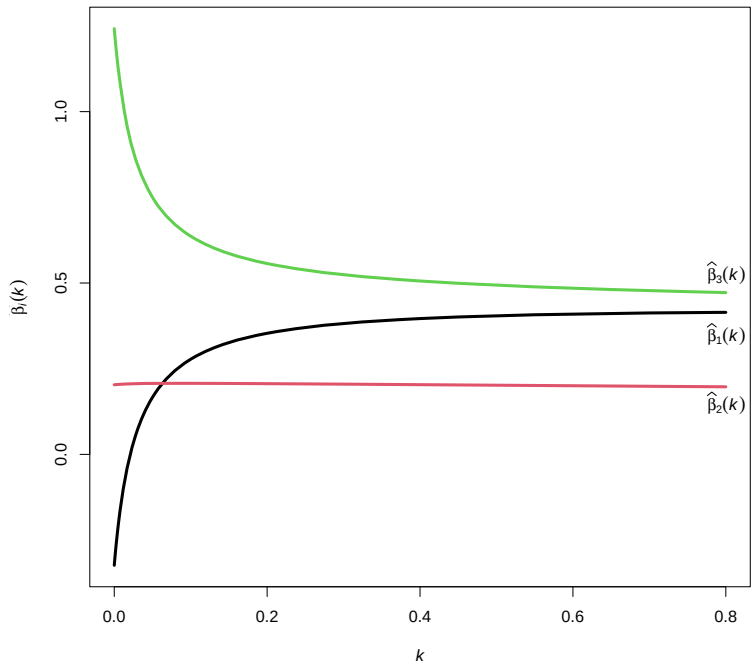
这就是总消费量和国内总产值之间的一个线性依赖关系.

由相关系数矩阵$\mathbf{R_X}$知: X_1的X_3相关系数为0.997.

与它们之间的复共线关系吻合, 都表明X_1与X_3有强的相关关系.

- 采用岭估计来估计回归系数. 对于标准化的变量, 计算出的岭迹, 并画对应的岭迹图

例6.6.2



从岭迹图上可以看出, 岭迹$\widehat{\beta}_1$随着k的增加, 很快增加, 大约在$k = 0.03$处从负值变为正值. 而$\widehat{\beta}_2$ 相对比较稳定, 但$\widehat{\beta}_3$随着k的增加, 骤然减少, 大约在$k = 0.4$ 以后就稳定下来.

例6.6.2

取$k = 0.4$, 对应的岭估计为

$$\widehat{\beta}_1(0.4) = 0.416, \quad \widehat{\beta}_2(0.4) = 0.213, \quad \widehat{\beta}_3(0.4) = 0.531.$$

平均值: $\bar{x}_1 = 194.59, \bar{x}_2 = 3.30, \bar{x}_3 = 139.74, \bar{y} = 21.89.$

标准差: $s_1 = 30.00, s_2 = 1.65, s_3 = 20.63, s_y = 4.54.$

代入经验回归方程化简后得到原变量的经验回归方程:

$$\widehat{Y} = -8.647 + 0.0630X_1 + 0.587X_2 + 0.117X_3.$$

例6.6.2

- 对标准化后的数据模型，采用GCV选择岭参数

```
data1=scale(dataF)
sol.ridge<-lm.ridge(y~0+x1+x2+x3, data=data.frame(data1),
lambda=c(seq(0,0.5,0.001)))
sol.ridge$lambda[which.min(sol.ridge$GCV)]
```

求得岭参数$k = 0.012$.

相应回归系数的岭估计为

$$\widehat{\boldsymbol{\beta}}(0.012) = (-0.103, 0.215, 1.066)'$$

注 此例中，由GCV选择的岭参数不合适，因为

- 第一个系数符号为负, 与实际意义不符
- 方差的扩大因子偏大, 最大的VIF为$6.848 > 5$

例6.6.2

- $\widehat{\beta}(k)$的方差扩大因子为$(\mathbf{R_X}+k\mathbf{I})^{-1}\mathbf{R_X}(\mathbf{R_X}+k\mathbf{I})^{-1}$的对角元素.

  当$k=\lambda=0.03$时, 采用如下命令

  solve(RX+0.03*diag(3))%*% RX%*% solve(RX+0.03*diag(3))

$$\begin{matrix} 1.601 & -0.005 & -1.115 \\ -0.005 & 0.943 & -0.023 \\ -1.115 & -0.023 & 1.602 \end{matrix}$$

  最大的VIF 为1.602<5

实际数据中, 岭参数需要结合系数正负号的实际意义、方差的扩大因子、GCV来选择和岭迹等来决定.

主成分估计

考虑已中心化的多元线性回归模型为

$$\boldsymbol{y} = \alpha_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_n.$$

令$\mathbf{Z} = \mathbf{X}\boldsymbol{\Phi}$, $\alpha_0 = \beta_0$, $\boldsymbol{\alpha} = \boldsymbol{\Phi}'\boldsymbol{\beta}$, 其中, $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_{p-1})$, $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_{p-1}$为$\mathbf{X}'\mathbf{X}$的特征值$\lambda_1, \cdots, \lambda_{p-1}$ 对应的标准正交化特征向量, 则线性回归模型可改写为

线性回归模型的典则形式

$$\boldsymbol{y} = \alpha_0 \mathbf{1}_n + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathbf{I_n}. \tag{6.32}$$

称$\boldsymbol{\alpha}$为典则回归系数.

主成分

- 因为$\mathbf{X}$是中心化的, 即$\mathbf{1}'\mathbf{X} = \mathbf{0}$, 所以$\mathbf{1}'\mathbf{Z} = \mathbf{1}'\mathbf{X}\Phi = \mathbf{0}$. 所以$\mathbf{Z}$也是中心化的.
- 不失一般性, 假设$\lambda_1 \geq \cdots \geq \lambda_{p-1}$.

$$\mathbf{Z} = (z_{(1)}, \cdots, z_{(p-1)}) = (\mathbf{X}\phi_1, \cdots, \mathbf{X}\phi_{p-1},)$$

其中$z_{(i)} = \mathbf{X}\phi_i$ 为第i 主成分, $i = 1, \cdots, p-1$.

注 $\mathbf{Z}$的第i 列$z_{(i)}$ 是原来$p-1$ 个回归自变量的线性组合, 其组合系数为$\mathbf{X}'\mathbf{X}$ 的第i 个特征值对应的特征向量ϕ_i.

主成分性质

- 矩阵$\mathbf{Z}$的各列元满足

$$\bar{z}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij} = 0, \quad j = 1, \cdots, p-1,$$

$$\boldsymbol{z}'_{(i)}\boldsymbol{z}_{(i)} = \boldsymbol{\phi}'_i \mathbf{X}'\mathbf{X}\boldsymbol{\phi}_i = \lambda_i,$$

$$\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2 = \boldsymbol{z}'_{(i)}\boldsymbol{z}_{(i)} = \lambda_i, \quad j = 1, \cdots, p-1.$$

$\mathbf{X}'\mathbf{X}$的第i 个特征值λ_i就度量了第i个主成分取值变动大小. 当设计阵$\mathbf{X}$存在复共线关系时, 有一些$\mathbf{X}'\mathbf{X}$的特征值很小, 对应的主成分取值变动就很小, 它们的均值都为零. 因而这些主成分取值近似为零.

主成分回归

当设计阵$\mathbf{X}$存在复共线关系时, 用主成分作为新的回归自变量, 并将对应变量的影响很小的主成分从回归模型中剔除, 用最小二乘法做剩下主成分的回归, 然后再变回到原来的自变量.

- 记$\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_{p-1})$, 对$\mathbf{\Lambda}, \boldsymbol{\alpha}, \mathbf{Z}$和$\mathbf{\Phi}$作相应分块:

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_2 \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}, \quad \mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2), \quad \mathbf{\Phi} = (\mathbf{\Phi}_1, \mathbf{\Phi}_2).$$

- 剔除影响很小的$\mathbf{Z}_2\boldsymbol{\alpha}_2$项, 得到回归模型

$$\boldsymbol{y} = \alpha_0 \mathbf{1}_n + \mathbf{Z}_1\boldsymbol{\alpha}_1 + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \mathbf{0}, \quad \text{Cov}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_n, \tag{6.33}$$

主成分回归

- 应用最小二乘法, 得到α_0和$\boldsymbol{\alpha}_1$的LS估计:

$$\widehat{\alpha}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\widehat{\boldsymbol{\alpha}}_1 = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\boldsymbol{y} = \boldsymbol{\Lambda}_1^{-1}\mathbf{Z}_1'\boldsymbol{y}.$$

- 令前面剔除了后面$p-r-1$个主成分对的系数$\boldsymbol{\alpha}_2$的估计为$\tilde{\boldsymbol{\alpha}}_2 = \mathbf{0}$.

  利用关系$\boldsymbol{\beta} = \boldsymbol{\Phi}\boldsymbol{\alpha}$, 可以获得原来参数$\boldsymbol{\beta}$的估计

$$\widetilde{\boldsymbol{\beta}} = \Phi \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_1 \\ \widehat{\boldsymbol{\alpha}}_2 \end{pmatrix} = (\boldsymbol{\Phi}_1,\ \boldsymbol{\Phi}_2) \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} = \boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Phi}_1'\mathbf{X}'\boldsymbol{y}, \qquad (6.34)$$

  这就是$\boldsymbol{\beta}$ 的主成分估计.

主成分估计的性质

- 主成分估计是有偏估计, 因为根据(6.34)式有

$$E(\tilde{\boldsymbol{\beta}}) = (\boldsymbol{\Phi}_1,\ \boldsymbol{\Phi}_2)\begin{pmatrix} \boldsymbol{\alpha}_1 \\ \mathbf{0} \end{pmatrix} = \boldsymbol{\Phi}_1\boldsymbol{\alpha}_1 \neq \boldsymbol{\Phi}_1\boldsymbol{\alpha}_1 + \boldsymbol{\Phi}_2\boldsymbol{\alpha}_2 = \boldsymbol{\Phi}\boldsymbol{\alpha} = \boldsymbol{\beta}.$$

- 优于LS估计的主成分估计的存在性.

定理6.7.2

当设计阵存在复共线关系时, 适当选择保留的主成分个数可致主成分估计比LS估计有较小的均方误差, 即

$$\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}) < \mathrm{MSE}(\widehat{\boldsymbol{\beta}}).$$

证明 因为$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \mathrm{tr}(\boldsymbol{\Lambda}^{-1})$,

$$\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}) = \mathrm{MSE}\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_1 \\ \mathbf{0} \end{pmatrix} = \sigma^2 \mathrm{tr}(\boldsymbol{\Lambda}_1^{-1}) + \|\boldsymbol{\alpha}_2\|^2,$$

所以

$$\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}) < \mathrm{MSE}(\widehat{\boldsymbol{\beta}}),$$

当且仅当

$$\|\alpha_2\|^2 < \sigma^2 \mathrm{tr}(\boldsymbol{\Lambda}_2^{-1}) = \sigma^2 \sum_{i=r+1}^{p-1} \frac{1}{\lambda_i}. \tag{6.35}$$

由于假定$\mathbf{X}'\mathbf{X}$的后面$p-r-1$个特征值接近于零, 于是上式右端很大, 故不等式(6.35)成立. 定理得证.

主成分估计

在主成分估计应用中, 一个重要的问题是如何选择保留主成分个数.

- 通常有两种方法:
  - 保留对应的特征值相对比较大的那些主成分.
  - 选择r, 使得$\sum_{i=1}^{r}\lambda_i$与全部$p-1$个特征值之和$\sum_{i=1}^{p-1}\lambda_i$的比值(称这个比值为前$r$个主成分的贡献率)达到预先给定的值, 譬如75%或80%等.
- 主成分作为原来变量的线性组合, 是一种“人造变量", 一般不易解释它的实际含义, 特别是当回归自变量具有不同度量单位时.

续例6.7.1

对外贸数据分析问题, 来求它的主成分估计.

解　$\mathbf{X}'\mathbf{X}$的3个特征值分别为$\lambda_1 = 1.999,\ \lambda_2 = 0.998,\ \lambda_3 = 0.003,$

对应的3个标准正交化特征向量分别为

$$\phi_1 = (0.7063, 0.043, 0.7065)',$$

$$\phi_2 = (-0.0357, 0.9990, -0.0258)',$$

$$\phi_3 = (-0.7070, -0.0070, 0.7072)',$$

3个主成分分别为

$$z_1 = 0.7063X_1 + 0.0435X_2 + 0.7065X_3,$$

$$z_2 = -0.0357X_1 + 0.9990X_2 - 0.0258X_3,$$

$$z_3 = -0.7070X_1 - 0.0070X_2 + 0.7072X_3$$

因为$\lambda_3 \approx 0$，且前两个主成分的贡献率

$$\sum_{i=1}^{2} \lambda_i / \sum_{i=1}^{3} = 0.999 = 99.9\%.$$

因此，我们剔除第3个主成分，只保留前两个主成分，它们的回归系数的LS估计分别为

$$\widehat{\alpha}_1 = 0.690, \widehat{\alpha}_2 = 0.1913.$$

还原到原来变量，得到经验回归方程

$$\widehat{Y} = -9.1057 + 0.0727X_1 + 0.6091X_2 + 0.1062X_3.$$

Table: 外贸数据分析问题的三种估计

变量	常数项	x_1	x_2	x_3
主成分估计(r=2)	-9.1057	0.0727	0.6091	0.1062
LS估计	-10.1300	-0.0514	0.5869	0.2868
岭估计(k=0.04)	-8.5537	0.0635	0.5859	0.1156

总的来讲,

- 主成分估计和岭估计比较相近.
- 而与LS估计相比, 复共线关系所包含的X_1和X_3的回归系数变化较大, 并且X_1的回归系数的符号也发生了变化.

第6章 第二部分总结

线性回归诊断：

- 残差分析：对误差假设的诊断及处理
- 影响分析：高杠杆点、异常值和强影响点的诊断及处理
- 协变量复共线性：复共线性对回归的影响、诊断、处理方法