

第3章 线性回归模型

李高荣

北京师范大学统计学院

E-mail: ligaorong@bnu.edu.cn



本章纲要

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

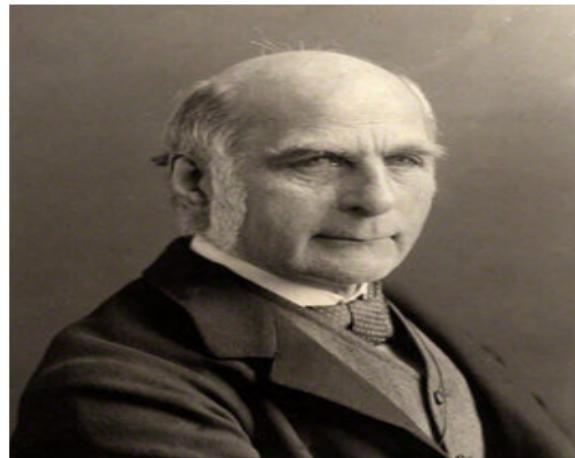
4 Box-Cox变换

5 定性协变量建模



- 扫二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

- “回归”的概念是1886年由英国统计学家Galton在研究父代身高与子代身高之间的关系时提出的.
- 回归分析已经成为现代统计学中应用最为广泛的方法之一, 主要用于探索和检验协变量 X 与响应变量 Y 之间的相关关系, 也可以通过协变量 X 的取值变化来预测响应变量 Y 的取值, 进一步可以描述协变量 X 和响应变量 Y 之间的相关关系.



Francis Galton (1822–1911)

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断?
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

■ 客观世界中变量之间的关系包括：

✚ 确定性关系：变量之间的关系能用函数来表达

✚ 非确定性关系：相关关系

■ 回归分析：研究相关关系的数学工具，可帮助人们从一个变量的取值去估计另一个变量的值。

■ 假设 Y 为响应变量, X_1, \dots, X_p 为 p 个协变量(或预测变量), 这时响应变量与协变量之间有如下的关系

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

其中 β_0 为截距项, β_1, \dots, β_p 为回归系数, ε 为随机模型误差.

■ 对给定 $\mathbf{X} = \mathbf{x}$ 时, 则回归函数为

$$g(\mathbf{x}) = \text{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

其中 $\mathbf{X} = (X_1, \dots, X_p)^T$ 和 $\mathbf{x} = (x_1, \dots, x_p)^T$.

■ 假设对 Y, X_1, \dots, X_p 进行了 n 次独立的试验，得到 n 组观测值，即

$$y_i, x_{i1}, \dots, x_{ip}, \quad i = 1, \dots, n$$

多元线性回归中的数据结构

响应变量		协变量			
序号	Y	X_1	X_2	\dots	X_p
1	y_1	x_{11}	x_{12}	\dots	x_{1p}
2	y_2	x_{21}	x_{22}	\dots	x_{2p}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	y_n	x_{n1}	x_{n2}	\dots	x_{np}

■ 观测样本 $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$ 满足

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n.$$

■ 模型误差 ε_i 满足如下的Gauss–Markov假设:

- ① $E(\varepsilon_i) = 0, \quad i = 1, \dots, n;$
- ② $\text{Var}(\varepsilon_i) = \sigma^2, \quad i = 1, \dots, n;$
- ③ $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j, \quad i, j = 1, \dots, n.$

■ 引进矩阵记号

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

■ 多元线性模型写成如下矩阵形式

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\varepsilon}_{n \times 1}.$$

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断?
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

■ 最小二乘目标函数为

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}^T \mathbf{Y} - 2\mathbf{Y}^T \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

■ 对 $\boldsymbol{\beta}$ 求偏导数，并令其为零，则可以得到关于 $\boldsymbol{\beta}$ 的 **正规方程**:

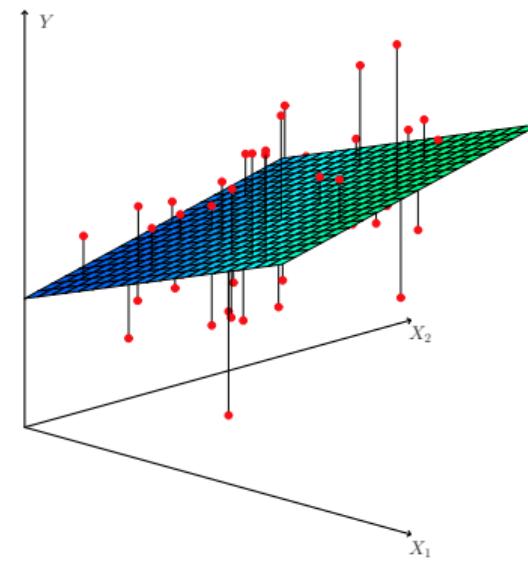
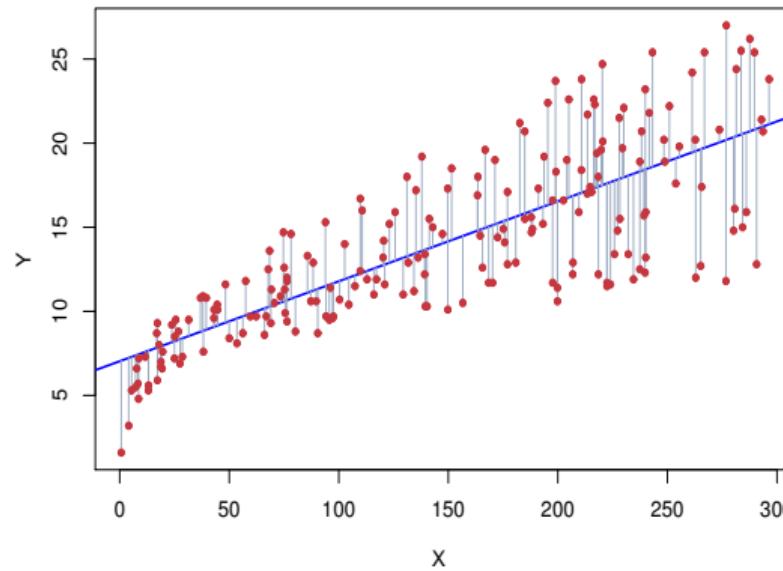
$$\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}.$$

■ 正规方程有唯一解的充要条件是 $\mathbf{X}^T \mathbf{X}$ 的秩为 $p + 1$ ，或者矩阵 $\mathbf{X}^T \mathbf{X}$ 的逆存在。

最小二乘估计

■ 解正规方程, 可得 β 的最小二乘估计为

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$



最小二乘估计

♠ 问题: $\hat{\beta}$ 是最小二乘目标函数 $Q(\beta)$ 的最小值吗?

■ 事实上, 对任意一个 β , 有

$$\begin{aligned}\|Y - X\beta\|_2^2 &= \|Y - X\hat{\beta} + X(\hat{\beta} - \beta)\|_2^2 \\ &= \|Y - X\hat{\beta}\|_2^2 + (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \\ &\quad + 2(\hat{\beta} - \beta)^T X^T (Y - X\hat{\beta}).\end{aligned}$$

■ 因为 $\hat{\beta}$ 是正规方程的解, 则有 $X^T(Y - X\hat{\beta}) = 0$.

■ 因此, 对于任意的 β , 有

$$\|Y - X\beta\|_2^2 = \|Y - X\hat{\beta}\|_2^2 + (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta).$$

■ 因为 $X^T X$ 是一个正定矩阵, 故上式第二项总是非负的, 则有

$$Q(\beta) = \|Y - X\beta\|_2^2 \geq \|Y - X\hat{\beta}\|_2^2 = Q(\hat{\beta}).$$

■ 等号成立当且仅当

$$(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) = 0.$$

最小二乘估计

定理3.1.1

对于多元线性回归模型, 则最小二乘估计 $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ 具有下列性质:

- ① $E(\hat{\beta}) = \beta;$
- ② $\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.$

■ 有了LS估计 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^T$, 可得下面的经验线性回归方程

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p.$$

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断?
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

■ 由LS估计 $\hat{\beta}$, 可得Y的拟合值为

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} =: \mathbf{H}\mathbf{Y}.$$

■ $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ 称为**投影矩阵或帽子矩阵**.

■ 最小二乘估计的**有效自由度**可定义为

$$\begin{aligned} df(ols) &= \text{tr}(\mathbf{H}) = \text{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \text{tr}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) \\ &= \text{tr}(\mathbf{I}_{p+1}) = p + 1. \end{aligned}$$

■ 模型误差向量 $\varepsilon = \mathbf{Y} - \mathbf{X}\beta$ 是不可观测的随机向量, 可考虑下面的残差向量:

$$\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

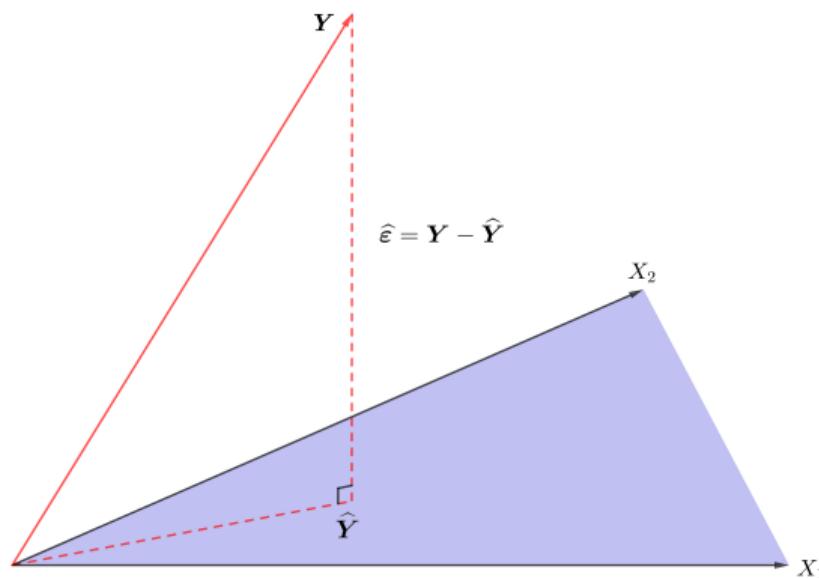
■ 对于第 i 个残差, 可以定义为: $\hat{\varepsilon}_i = y_i - \mathbf{x}_i^T \hat{\beta}$, 其中 $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$.

■ 可以证明:

$$\mathbf{H}^2 = \mathbf{H}, \quad \mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \mathbf{0},$$

$$\mathbf{X}^T \hat{\varepsilon} = \mathbf{0}, \quad \hat{\mathbf{Y}}^T \hat{\varepsilon} = \mathbf{0}.$$

σ^2 的估计



最小二乘方法的几何解释，拟合值 \hat{Y} 可以看成是投影到由 X_1 和 X_2 张成的空间上

■ 残差平方和(residual sum of squares, RSS), 定义为

$$\text{RSS} = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}.$$

■ 残差平方和(RSS)的大小反映了实际数据与理论模型的偏离程度或者拟合程度. RSS越小, 说明模型对数据的拟合变得越好.

定理3.1.2

由残差平方和RSS的定义, 可以得到 σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1}.$$

定理3.1.3

对于多元线性回归模型，进一步假设随机模型误差向量 $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ ，则

- ① $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$;
- ② $\frac{\text{RSS}}{\sigma^2} = \frac{(n - p - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$;
- ③ $\hat{\beta}$ 和 RSS 相互独立.

推论3.1.1

对于多元线性回归模型, 若 $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, 则

- ① $\hat{\beta}_k \sim N(\beta_k, \sigma^2 c_{k+1,k+1})$, 其中 $c_{k+1,k+1}$ 表示矩阵 $(\mathbf{X}^T \mathbf{X})^{-1}$ 的第 $(k+1, k+1)$ 个元素;
- ② 在 β_k 的一切线性无偏估计中, $\hat{\beta}_k$ 是唯一方差最小者, 其中 $k = 0, 1, \dots, p$.

■ 由推论3.1.1, 可得 β_k 的 $100(1 - \alpha)\%$ 的置信区间如下

$$\hat{\beta}_k \pm t_{n-p-1} \left(\frac{\alpha}{2} \right) \sqrt{\widehat{\text{Var}}(\hat{\beta}_k)}, \quad k = 0, 1, \dots, p,$$

其中

- ▶ $\widehat{\text{Var}}(\hat{\beta}_k) = \hat{\sigma}^2 c_{k+1,k+1}$;
- ▶ $c_{k+1,k+1}$ 是矩阵 $(\mathbf{X}^T \mathbf{X})^{-1}$ 的第 $(k + 1, k + 1)$ 个元素;
- ▶ $t_{n-p-1}(\alpha/2)$ 为自由度为 $n - p - 1$ 的 t 分布的上侧 $\alpha/2$ 分位数.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断?
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

回归系数的显著性检验

■ 在实际问题中，需要检验第 k 个协变量 X_k 对响应变量 Y 的影响是否显著，即考虑下面的假设检验

$$H_{k0} : \beta_k = 0 \longleftrightarrow H_{k1} : \beta_k \neq 0, \quad k = 1, \dots, p.$$

■ 由推论3.1.1, 定理3.1.2和定理3.1.3, 可以证明, 当原假设 H_{k0} 成立时, 统计量

$$T_k = \frac{\widehat{\beta}_k}{\widehat{\sigma} \sqrt{c_{k+1,k+1}}} \sim t_{n-p-1}, \quad k = 1, \dots, p.$$

■ 如果 $|T_k| \geq t_{n-p-1}(\alpha/2)$ 或者 p 值 $p_k = \mathbb{P}(t_{n-p-1} \geq |T_k|) < \alpha/2$ 时, 则拒绝原假设 H_{k0} , 认为 $\beta_k \neq 0$.

♠ **问题:** 如何鉴别用回归方程对观测值 $\{(x_i, y_i), i = 1, \dots, n\}$ 的拟合程度呢?

■ 为了解决这个问题, 需要考虑回归方程或回归模型的**拟合优**

度(goodness of fit) 方法.

■ 考虑下面的假设检验问题:

$$H_0 : \beta_1 = \dots = \beta_p = 0$$

$$H_1 : \beta_1, \dots, \beta_p \text{ 不全为 } 0.$$

回归方程的拟合优度

■ 考虑总平方和(SST)的分解：

$$\begin{aligned} \text{SST} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2 \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i), \end{aligned}$$

其中 $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ 和 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$

■ 由 $\mathbf{X}^T \hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ 和 $\hat{\mathbf{Y}}^T \hat{\boldsymbol{\varepsilon}} = 0$, 可证得上式中交叉项等于0.

回归方程的拟合优度

- 总平方和SST可以分解为

$$SST = SSReg + RSS,$$

其中

$$SSReg = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- SSReg称为回归平方和, 表示总平方和中被回归方程解释的那部分变异或离差, 反映了协变量X对响应变量Y变动平方和的贡献.
- RSS反映的是随机误差的变动对总平方和的贡献.

■ 当原假设 H_0 成立时, 可证得统计量

$$F = \frac{\text{SSReg}/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}.$$

- 当 $F > F_{p,n-p-1}(\alpha)$ 时, 则拒绝原假设 H_0 , 否则就接受原假设 H_0 .
- 检验统计量是把SSReg和RSS进行比较, 当SSReg相对RSS比较大时, 就拒绝原假设, 认为回归直线与样本观测值的拟合效果是显著的.

回归方程的拟合优度

- 对线性回归的拟合优度检验可使用下面的方差分析表进行解释.
- 对给定的显著性水平 α , 当 $F \geq F_{p,n-p-1}(\alpha)$ 或 $p_v = \mathbb{P}(F_{p,n-p-1} \geq F) < \alpha$ 时, 拒绝原假设 H_0 , 否则就接受原假设 H_0 .

方差来源	平方和	自由度	均方	F 比
回归	SSReg	p	$\overline{\text{SSReg}} = \text{SSReg}/p$	$F = \overline{\text{SSReg}}/\overline{\text{RSS}}$
误差	RSS	$n - p - 1$	$\overline{\text{RSS}} = \text{RSS}/(n - p - 1)$	
总和	SST	$n - 1$		

- 在实际应用中, 更关心的问题是部分协变量对响应变量是否显著, 即需要检验部分回归系数是否为0, 或者检验某个子模型对数据拟合是否和全模型一样显著.
- 检验包含 k 个协变量的**简约模型**(reduced model, RM)对数据的拟合是否显著($0 \leq k < p$), 即考虑如下的假设检验

H'_0 : 简约模型对数据拟合显著 $\longleftrightarrow H'_1$: 全模型对数据拟合显著.

- **注意:** 简约模型是全模型的一个子模型, 具有嵌套关系.

回归方程的拟合优度

- 假设 \hat{y}_i 是基于 p 个协变量全模型对 y_i 的拟合值, \hat{y}_i^* 是基于 k 个协变量简约模型对 y_i 的拟合值.
- 全模型和简约模型的残差平方和分别定义为

$$\text{RSS}_{\text{FM}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \text{RSS}_{\text{RM}} = \sum_{i=1}^n (y_i - \hat{y}_i^*)^2.$$

- 全模型需要估计 $p + 1$ 个未知参数, 而简约模型仅需要估计 $k + 1$ 个未知参数, 且有 $\text{RSS}_{\text{RM}} \geq \text{RSS}_{\text{FM}}$.

回归方程的拟合优度

■ 构造下面的检验统计量

$$F' = \frac{(\text{RSS}_{\text{RM}} - \text{RSS}_{\text{FM}})/(p - k)}{\text{RSS}_{\text{FM}}/(n - p - 1)}.$$

■ $\text{RSS}_{\text{RM}} - \text{RSS}_{\text{FM}}$ 表示用简约模型拟合数据时残差平方和的增量，自由度为 $(n - k - 1) - (n - p - 1) = p - k$.

■ 当原假设 H'_0 成立时，可以证明

$$F' = \frac{(\text{RSS}_{\text{RM}} - \text{RSS}_{\text{FM}})/(p - k)}{\text{RSS}_{\text{FM}}/(n - p - 1)} \sim F_{p-k, n-p-1}.$$

回归方程的拟合优度

■ 对给定的显著性水平 α , 当 $F' > F_{p-k,n-p-1}(\alpha)$ 或 $p_v = \mathbb{P}(F_{p-k,n-p-1} \geq F') < \alpha$ 时, 拒绝原假设 H'_0 , 否则就接受原假设 H'_0 .

■ 当 $k = 0$ 时, 假设检验退化为

$$H'_0 : Y = \beta_0 + \varepsilon \leftrightarrow H'_1 : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

■ 这时, 检验统计量 F' 退化为检验统计量 F .

判定系数 R^2

■ 判定系数 R^2 定义为SSReg 占SST 的比例, 即

$$R^2 = \frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

■ 判定系数 R^2 的取值在[0,1]之间:

■ 如果在SST 中SSReg所占的比重越大, 这时判定系数 R^2 越接近于1, 则线性回归效果就越好, 说明回归方程与样本观测值的拟合效果就越好.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断?
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

Y 的点预测和预测区间

■ 给定 $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0p})^T$, 回归方程的真实值为

$$y_0 = \beta_0 + \beta_1 x_{01} + \dots + \beta_p x_{0p} + \varepsilon_0.$$

■ 忽略未知的 ε_0 , 可得响应变量的点估计为

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_p x_{0p} = \tilde{\mathbf{x}}_0^T \hat{\boldsymbol{\beta}},$$

其中 $\tilde{\mathbf{x}}_0 = (1, x_{01}, x_{02}, \dots, x_{0p})^T$.

■ 可以证明: $\hat{y}_0 \sim N(g(\mathbf{x}_0), \sigma^2 \tilde{\mathbf{x}}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \tilde{\mathbf{x}}_0)$, 其中 $g(\mathbf{x}_0) = \beta_0 + \beta_1 x_{01} + \dots + \beta_p x_{0p}$.

■ 进一步, 可得

$$\hat{y}_0 - y_0 \sim N \left(0, \sigma^2 \left[1 + \tilde{\mathbf{x}}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \tilde{\mathbf{x}}_0 \right] \right).$$

■ 由定理3.1.3(2), 以及 y_0, \hat{y}_0 和RSS 的相互独立性, 则有

$$\begin{aligned} \frac{\hat{y}_0 - y_0}{\hat{\sigma} \sqrt{1 + \tilde{\mathbf{x}}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \tilde{\mathbf{x}}_0}} &= \frac{\hat{y}_0 - y_0}{\sigma \sqrt{1 + \tilde{\mathbf{x}}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \tilde{\mathbf{x}}_0}} \sqrt{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}} / (n-p-1) \\ &\sim t_{n-p-1}. \end{aligned}$$

■ 给定置信水平为 $1 - \alpha$, 则可得 y_0 的预测区间为

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

■ 当 $p = 1$ 时, 退化为一元线性回归模型: $Y = \beta_0 + \beta_1 x + \varepsilon$.

■ 给定 $x = x_0$ 时, 则 y_0 的置信水平为 $1 - \alpha$ 预测区间为

$$\left[\hat{y}_0 \pm t_{n-2}(\alpha/2) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right].$$

■ 容易看到, 该预测区间的长度是 x_0 的函数, 它随着 $|x_0 - \bar{x}|$ 的增加而增加.

回归函数 $g(\mathbf{x})$ 的点估计和置信区间

■ 回归函数 $g(\mathbf{x})$ 在 \mathbf{x}_0 的点估计为经验回归函数

$$\hat{y}_0 = \hat{g}(\mathbf{x}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}.$$

■ 由 $\hat{y}_0 \sim N\left(g(\mathbf{x}_0), \sigma^2 \tilde{\mathbf{x}}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \tilde{\mathbf{x}}_0\right)$ 和定理3.1.3(2), 可得

$$\frac{\hat{y}_0 - g(\mathbf{x}_0)}{\hat{\sigma} \sqrt{\tilde{\mathbf{x}}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \tilde{\mathbf{x}}_0}} \sim t_{n-p-1}.$$

■ 可得回归函数 $g(\mathbf{x}_0)$ 置信水平为 $1 - \alpha$ 的置信区间为

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

回归函数 $g(\mathbf{x})$ 的点估计和置信区间

■ 当 $p = 1$ 时, 可得到 $g(x_0) = \beta_0 + \beta_1 x_0$ 置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\hat{y}_0 \pm t_{n-2}(\alpha/2) \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right].$$

■ 注意: 在相同的置信水平 $1 - \alpha$ 下, y_0 的预测区间要比回归函数 $g(x_0)$ 的置信区间要长.

■ 理由: $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \varepsilon_0$ 比回归函数 $g(x_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$ 多了误差项 ε_0 .

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

■ 对于多元线性回归模型的应用，在R语言中，可用**函数lm()**进行计算，其调用格式为：

```
lm(formula, data, subset, weights, na.action,  
method = "qr", model = TRUE, x = FALSE,  
y = FALSE, qr = TRUE, singular.ok = TRUE,  
contrasts = NULL, offset, ...)
```

其中formula为模型公式；data为数据框数据；subset为可选择向量，表示观测值的子集；weights为可选择向量，表示用于数据拟合的权重；其余参数见在线帮助。

■ 下面再介绍几个常用的函数：

`anova(object, ...)`

其中`object`为函数`lm()`和`glm()`得到的对象，其返回值为模型的方差分析表。

`predict(object, newdata, se.fit=F, scale=NULL, df=Inf,`
`interval=c("none", "confidence", "prediction"),`
`level = 0.95, type = c("response", "terms"),`
`terms = NULL, na.action = na.pass,`
`pred.var = res.var/weights, weights = 1, ...)`

其中`object`是由函数`lm()`得到的对象；`newdata`是预测点的数据框数据；选`interval`为`"confidence"`，返回值为回归函数的置信区间；选`interval`为`"prediction"`，返回值为Y的预测区间；其余参数见在线帮助。

`plot(object, ...)`

其中`object`是由函数`lm()`得到的对象，绘制模型诊断的几种图形，显示残差、拟合值和一些诊断情况。

`confint(object, ...)`

其中`object`是由函数`lm()`得到的对象，返回值为截距和回归系数的置信区间。

`residuals(object, type = c("working", "response", "deviance", "pearson", "partial"))`

其中`object`是由`lm`或`aov`构成的对象，`type`是返回值的类型，返回值为模型的残差。

`summary(object, ...)`

其中`object`是由`lm`构成的对象，返回值是显示较为详细的模型拟合结果。

R语言函数及应用—前列腺癌症数据集prostate

例：前列腺癌症数据集prostate

考虑程序包faraway中的前列腺癌症数据集prostate, 该数据集包含97个样本和9个变量: lpsa (PSA的对数)、lcavol (癌体积的对数)、lweight (前列腺重量的对数)、age (患者年龄)、lbph (良性前列腺增生量的对数)、svi (精囊浸润)、lcp (包膜穿透的对数)、gleason (格里森分数)和pgg45 (格里森分数为4或5的比例). 考虑下面的多元线性回归模型

$$\begin{aligned} \text{lpsa} = & \beta_0 + \beta_1 \text{lcavol} + \beta_2 \text{lweight} + \beta_3 \text{age} + \beta_4 \text{lbph} + \beta_5 \text{svi} \\ & + \beta_6 \text{lcp} + \beta_7 \text{gleason} + \beta_8 \text{pgg45} + \varepsilon. \end{aligned}$$

```

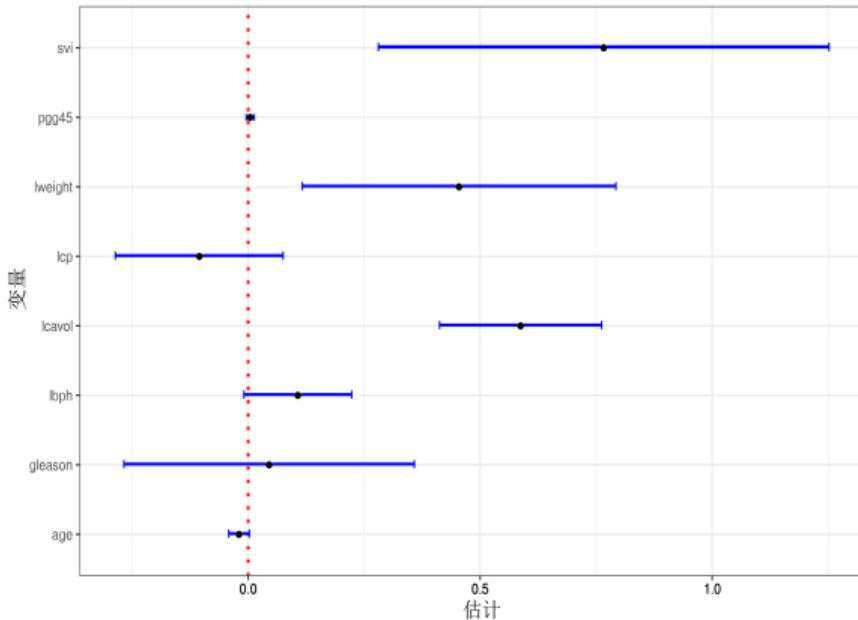
data(prostate, package = "faraway")
> summary(lm.reg = lm(lpsa ~ ., data = prostate))    ## 输出结果

```

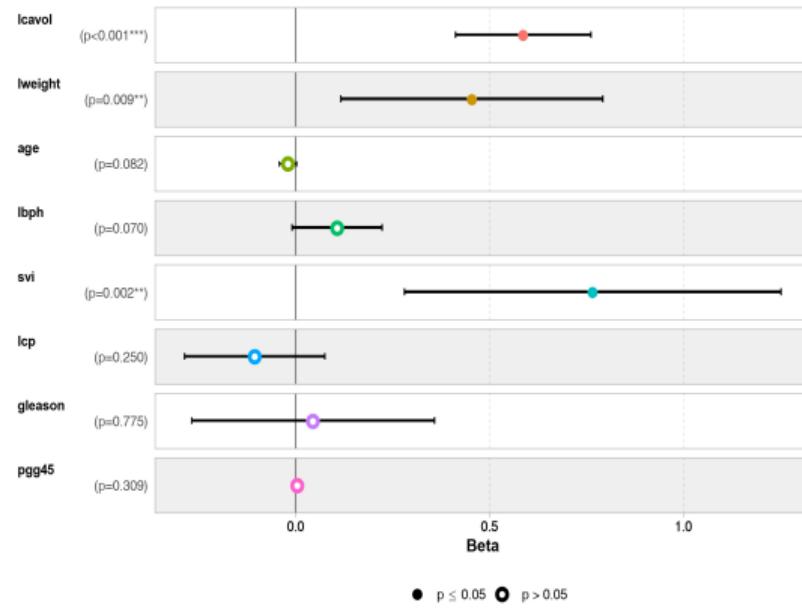
变量	系数估计	标准误差	t统计量	p值
β_0	0.669337	1.296387	0.516	0.60693
lcavol	0.587022	0.087920	6.677	2.11e-09
lweight	0.454467	0.170012	2.673	0.00896
age	-0.019637	0.011173	-1.758	0.08229
lbph	0.107054	0.058449	1.832	0.07040
svi	0.766157	0.244309	3.136	0.00233
lcp	-0.105474	0.091013	-1.159	0.24964
gleason	0.045142	0.157465	0.287	0.77503
pgg45	0.004525	0.004421	1.024	0.30886
$n = 97$	$R^2 = 0.6548$	$R_{adj}^2 = 0.6234$	残差的SE=0.7084	d.f. =88
F统计量	20.86	d.f.=(8, 88)	p-value < 2.2e-16	

```
> confint(lm.reg)          ## 输出结果
                               2.5 %           97.5 %
(Intercept)      -1.906960983   3.245634379
lcavol          0.412298699   0.761744954
lweight          0.116603435   0.792331414
age              -0.041840618   0.002566267
lbph             -0.009101499   0.223209561
svi              0.280644232   1.251670420
lcp              -0.286344443   0.075395916
gleason          -0.267786053   0.358069248
pgg45            -0.004260932   0.013311395
## 利用程序包GGally和ggstats对回归系数置信区间进行可视化
library(GGally); library(ggstats)
ggcoef(lm.reg, exclude_intercept = T, vline_color = "red",
       errorbar_color="blue", errorbar_height=0.1)+theme_bw()
ggcoef_model(model = lm.reg)
```

R语言函数及应用—前列腺癌数据集prostate



(a)



(b)

(a) 程序包GGally对回归系数置信区间的可视化; (b) 程序包ggstats对回归系数置信区间的可视化

R语言函数及应用—前列腺癌症数据集prostate

- ① 经验回归方程为: $\text{lpsa} \approx 0.6693 + 0.5870 \times \text{lcavol} + 0.4545 \times \text{lweight} - 0.0196 \times \text{age} + 0.1071 \times \text{lbph} + 0.7662 \times \text{svi} - 0.1055 \times \text{lcp} + 0.0451 \times \text{gleason} + 0.0045 \times \text{pgg45}$;
- ② 从 p 值可知: 变量lcavol, lweight和svi 是线性回归显著的, 而变量age, lbph, lcp, gleason和pgg45是不显著的变量;
- ③ $R^2 = 0.6548$ 说明: 经验回归方程对数据的拟合效果较为显著;
- ④ 回归方程的检验, F 统计量的 p 值为 2.2×10^{-16} , 远远小于显著性水平 $\alpha = 0.05$, 说明经验回归方程是显著的.

■ 考虑下面两个简约模型对数据拟合的假设检验问题.

$$\text{RM1 : } \text{lpsa} = \beta_0 + \beta_2 \text{lweight} + \beta_5 \text{svi} + \varepsilon,$$

$$\text{RM2 : } \text{lpsa} = \beta_0 + \beta_1 \text{lcavol} + \beta_2 \text{lweight} + \beta_5 \text{svi} + \varepsilon.$$

- 取2个协变量lweight和svi, $p_v = 5.389 \times 10^{-11} \ll 0.05 = \alpha$, 则拒绝原假设 H'_0 , 认为仅仅考虑2个协变量lweight 和svi 时, 拟合效果不好.
- 取3个协变量lcavol, lweight和svi, $p_v = 0.2167 > 0.05 = \alpha$, 则接受原假设 H'_0 , 认为用3个协变量lcavol, lweight和svi的模型和全模型的拟合效果是没有显著差异的.

R语言函数及应用—前列腺癌数据集prostate

```
lm.reg2 = lm(lpsa ~ lweight + svi, data = prostate)
lm.reg3 = lm(lpsa ~ lcavol + lweight + svi, data = prostate)
> anova(lm.reg2, lm.reg3, lm.reg)      ## 输出结果
Analysis of Variance Table
Model 1: lpsa ~ lweight + svi
Model 2: lpsa ~ lcavol + lweight + svi
Model 3: lpsa ~ lcavol + lweight + age + lbph + svi + lcp
          + gleason + pgg45
Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     94 75.829
2     93 47.785  1   28.0445 55.8820 5.389e-11 ***
3     88 44.163  5    3.6218  1.4434  0.2167
---
Signif. codes: 0  '***' 0.001  '**' 0.01  '*' 0.05 '.' 0.1
```

■ 给定 $x_0 = (0.88, 3.50, 45, -1.25, 1, -1.35, 6, 10)^T$, 使用函数predict(), 求 y_0 的估计值、 y_0 的置信水平为95% 的预测区间和回归函数 $g(x_0)$ 的置信水平为95% 的置信区间.

```
x.0 = data.frame(lcavol=0.88, lweight=3.50, age=45, lbph=-1.25,
                  svi=1, lcp=-1.35, gleason=6, pgg45=10)
Y.pred=predict(lm.reg, x.0, interval="prediction", level=0.95)
> Y.pred
            fit        lwr        upr
1 2.983711 1.435591 4.531831
g.conf = predict(lm.reg, x.0, interval="confidence", level=0.95)
> g.conf
            fit        lwr        upr
1 2.983711 2.339739 3.627683
```

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

什么是回归诊断？

■ 回归诊断是对回归分析中的假设以及数据的检验与分析。

- ① 误差项是否满足独立性、等方差性、正态性；
- ② 选择多元线性模型是否合适；
- ③ 样本数据中是否存在异常值；
- ④ 回归分析的结果是否对某些样本的依赖性过重，即回归模型是否具备稳健性；
- ⑤ 自变量之间是否存在高度相关，即是否存在多重共线性问题。

什么是回归诊断？

■ 在R语言中，下面函数与回归诊断有关。

influence.measures

rstandard

rstudent

dfits

cooks.distance

dfbeta

dfbetas

covratio

hatvalues

hat

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

普通残差

- 针对多元线性回归模型, 残差为:

$$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

- 帽子矩阵: $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.
- 在R语言中, 函数`residuals()`提供了模型残差的计算.
- 因此, 得到残差后, 可对残差进行检验, 如正态性检验等.

```
y.res = residuals(lm.reg)
> shapiro.test(y.res)          ## 输出结果
      Shapiro-Wilk normality test
data: y.res
W = 0.99113, p-value = 0.7721
```

标准化(内学生化)残差

■ 由模型误差 ε 的性质, 可知

$$\mathbf{E}(\widehat{\boldsymbol{\varepsilon}}) = \mathbf{0}, \quad \text{Cov}(\widehat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

■ 对每个残差 $\widehat{\varepsilon}_i$, 有

$$\frac{\widehat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0, 1), \quad i = 1, \dots, n.$$

■ 杠杆统计量: 帽子矩阵 \mathbf{H} 对角线上的元素 h_{ii} , 称为杠杆统计量.

■ 作业: 证明: 杠杆值 h_{ii} 满足: $0 < h_{ii} < 1$, $i = 1, \dots, n$, 且

$$\sum_{i=1}^n h_{ii} = p + 1, \quad 1/n \leq h_{ii} < 1.$$

标准化(内学生化)残差

■ 用 $\hat{\sigma}^2$ 作为 σ^2 的估计值, 称

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

为**标准化残差**(standardized residual), 或称为**内学生化残差**(internally studentized residual).

```
rstandard(model, infl = lm.influence(model, do.coef = FALSE),  
          sd = sqrt(deviance(model)/df.residual(model)), ...)
```

其中model是由lm或glm生成的对象, infl是由lm.influence返回值得到的影响结构, sd是模型的标准差.

外学生化残差

- 首先定义 σ^2 的估计为

$$\hat{\sigma}_{(-i)}^2 = \frac{1}{n-p-2} \sum_{j \neq i} (y_j - \mathbf{x}_j \hat{\boldsymbol{\beta}}_{(-i)})^2,$$

其中 $\hat{\boldsymbol{\beta}}_{(-i)}$ 为删除第*i*个样本后的最小二乘估计.

- 对 $i = 1, \dots, n$, 称

$$\frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_{ii}}}$$

为**学生化残差**, 或者称为**外学生化残差**.

- 函数`rstudent()`用来计算回归模型的(外)学生化残差.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

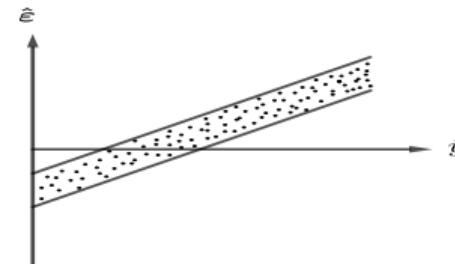
5 定性协变量建模

6 参考文献

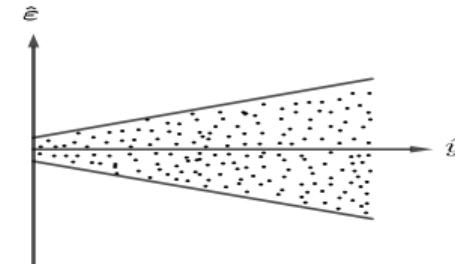
7 作业

残差图

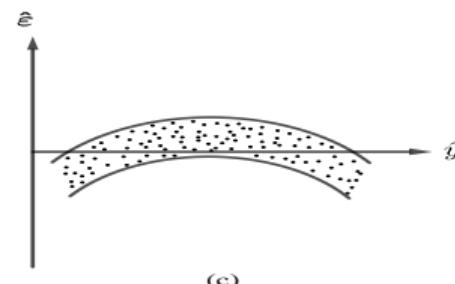
■ 以残差 $\hat{\varepsilon}_i$ 为纵坐标, 以拟合值 \hat{y}_i 或对应的数据观测序号 i , 或数据观测时间为横坐标的散点图统称为 **残差图**.



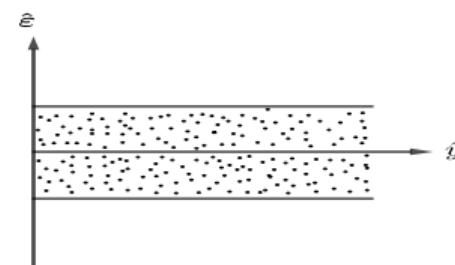
(a)



(b)



(c)



(d)

回归值 \hat{y} 与残差的散点图. (a) 模型错误的情形; (b) 异方差情形; (c) 非线性情形; (d) 正常情形

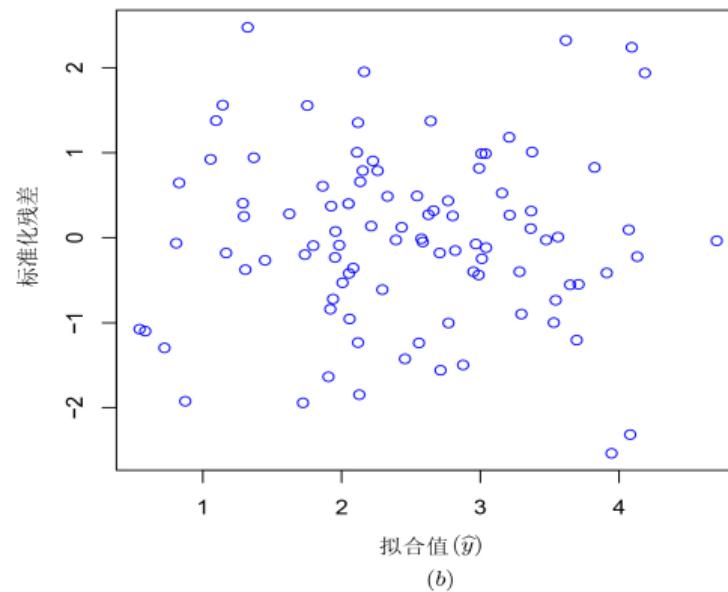
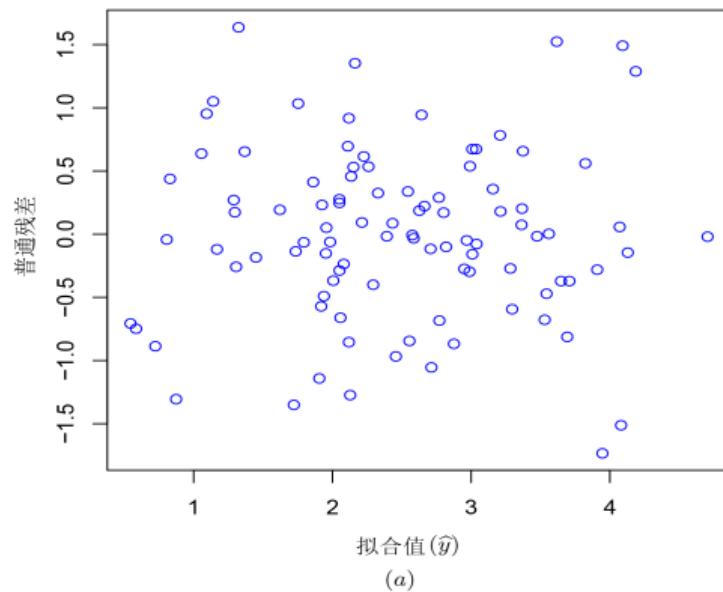
残差图

- 当残差服从正态分布的假设时，标准化残差应该近似服从标准化正态分布。
- 根据正态分布的性质，若随机变量 $X \sim N(\mu, \sigma^2)$ ，则有

$$\mathbb{P}(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954.$$

- 对于标准化残差，应该有95.4%的样本点落在区间 $[-2, 2]$ 内。
- 所以，如果以拟合值 \hat{y}_i 为横坐标，标准化残差 r_i 为纵坐标，那么平面上的点 $\{(y_i, r_i), i = 1, \dots, n\}$ 大致应落在宽度为4的水平带 $|r_i| \leq 2$ 的区域内，且不呈现任何趋势。

残差图



前列腺癌数据集多元线性回归分析的残差图. (a) 普通残差的残差图; (b) 标准化残差的残差图

残差的QQ图

- 可用残差的QQ图检验残差的正态性, 设 $\hat{\varepsilon}_{(i)}$ 表示残差 $\hat{\varepsilon}_i$ 的次序统计量, 其中 $i = 1, \dots, n$.
- 令

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

其中

- ▶ $\Phi(x)$ 为标准正态分布 $N(0, 1)$ 的分布函数, $\Phi^{-1}(x)$ 为其反函数;
- ▶ $q_{(i)}$ 为 $\hat{\varepsilon}_{(i)}$ 的期望值.

- R 语言中, 可用函数`plot(model, 2)`绘制残差的QQ图, 其中`model`是由`lm`生成的对象.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

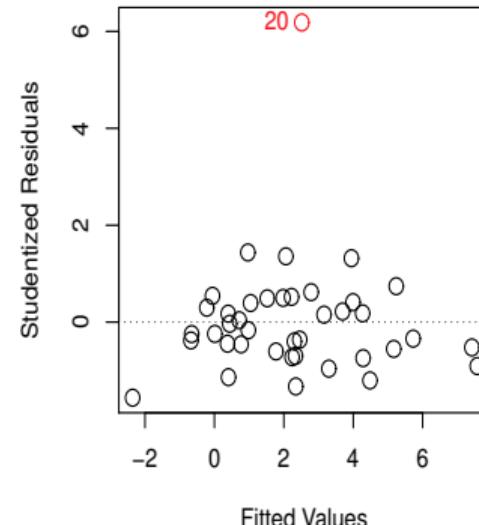
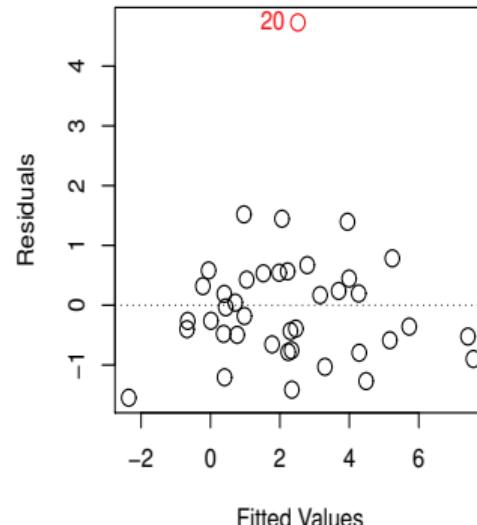
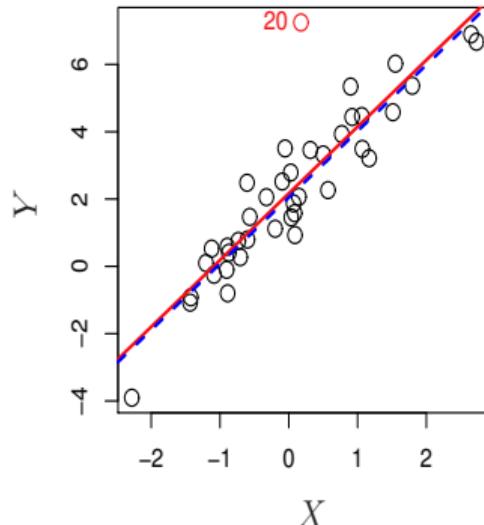
5 定性协变量建模

6 参考文献

7 作业

强影响点

■ 强影响点: 如果某个样本不遵从回归模型, 但是其余数据都遵从这个回归模型, 则称该样本点为强影响点, 也称为异常点.



红线: 最小二乘回归直线; 蓝线: 删除异常点后的最小二乘回归直线

- 残差图可以用来识别异常点或离群点. 但在实践中, 确定残差多大的点可以被认为是一个异常点或离群点会十分困难.
- 响应变量的拟合值为: $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{H}\mathbf{Y}$.
- 从几何上看, $\hat{\mathbf{Y}}$ 是 \mathbf{Y} 在 \mathbf{X} 的列向量张成子空间内的投影, 且满足

$$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}.$$

- h_{ii} 称为**杠杆统计量**, h_{ii} 的大小可表示第*i*个样本值对 \hat{y}_i 影响的大小.
- 由 $\text{Var}(\hat{y}_i) = \sigma^2 h_{ii}$ 可知, h_{ii} 反映了拟合值 \hat{y}_i 的波动情况.

杠杆统计量

■ 进一步, 对观测点 \mathbf{x}_i , 杠杆统计量定义为

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i,$$

其中 $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ 为矩阵 \mathbf{X} 的第*i*行的*p+1*维列向量.

■ 计算所有观测值的平均杠杆统计量为

$$\begin{aligned}\bar{h} &= \frac{1}{n} \sum_{i=1}^n h_{ii} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = \frac{1}{n} \text{tr} \left(\sum_{i=1}^n \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \right) \\ &= \frac{1}{n} \text{tr} \left((\mathbf{X}^T \mathbf{X})^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \right) = \frac{1}{n} \text{tr}(\mathbf{I}_{p+1}) = \frac{p+1}{n}.\end{aligned}$$

杠杆统计量

■ 因此, 可知

$$0 \leq h_{ii} \leq 1, \quad i = 1, \dots, n, \quad \sum_{i=1}^n h_{ii} = p + 1.$$

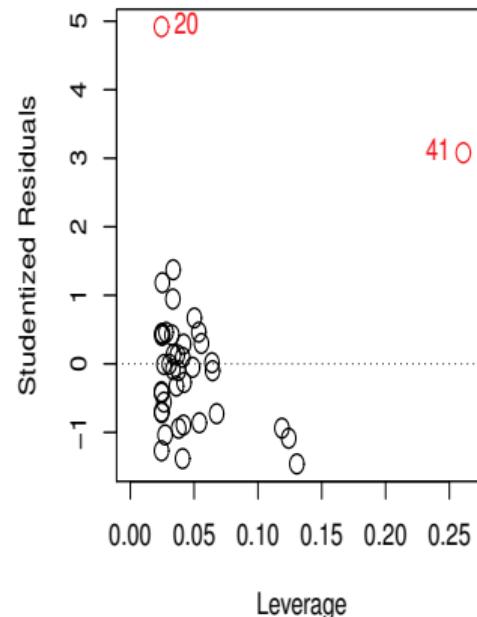
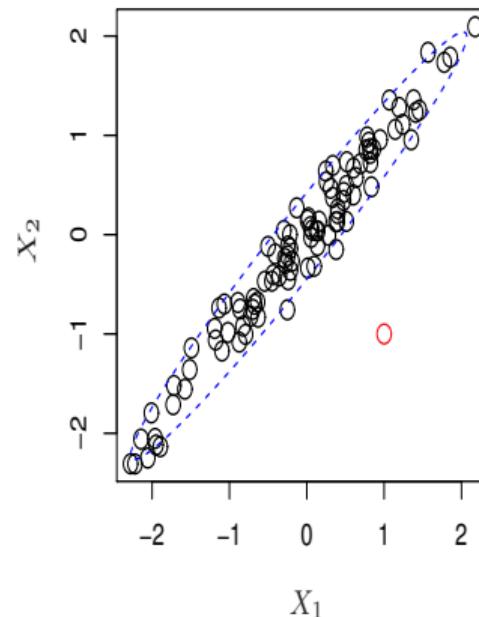
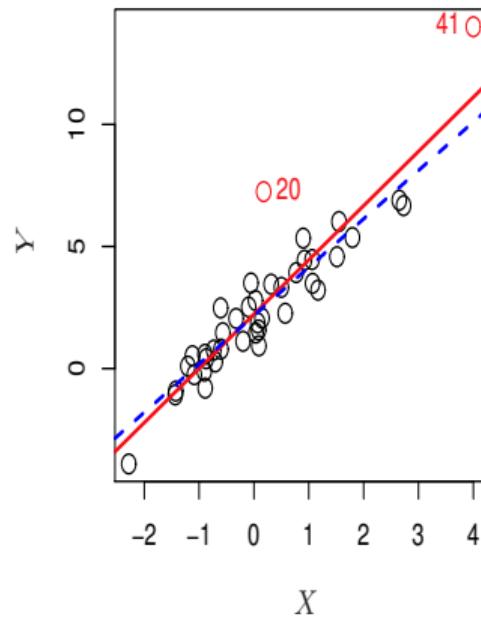
■ 对一元线性模型, 即 $p = 1$ 时, 杠杆统计量为

$$h_{ii} = \begin{pmatrix} 1 & x_i \end{pmatrix} (\mathbf{X}^T \mathbf{X})^{-1} \begin{pmatrix} 1 \\ x_i \end{pmatrix} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2},$$

其中 $\mathbf{X}^T = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}$.

杠杆统计量

可见, h_{ii} 随着 $x_i - \bar{x}$ 的增加而增加, h_{ii} 的取值总是在 $1/n$ 和1之间.



红线: 所有数据的拟合; 蓝线: 去掉观测点41后的拟合

■ Hoaglin和Welsch (1978)给出了一种判断异常点的方法，当

$$h_{i_0 i_0} \geq \frac{2(p+1)}{n},$$

则可认为第 i_0 个样本点影响较大，可结合其他准则，考虑是否将其剔除。

```
hatvalues(model, infl=lm.influence(model, do.coef=FALSE), ...)  
hat(x, intercept = TRUE)
```

其中model是由lm或glm生成的对象，x是设计矩阵X。

DFFITS准则

- Belsley等(1980)提供了另一种准则, 考虑统计量

$$D_i(\hat{\sigma}_{(-i)}) = \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_{ii}}}.$$

- 对于第*i*个样本, 如果有

$$|D_i(\hat{\sigma}_{(-i)})| > 2\sqrt{\frac{p + 1}{n}},$$

则认为第*i*个样本的影响比较大.

- 计算DFFITS准则的函数`dffits()`.

Cook距离统计量

- Cook (1977)提出了Cook距离统计量, 定义为

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \hat{\beta}_{(-i)})}{(p+1)\hat{\sigma}^2}, \quad i = 1, \dots, n.$$

- Cook统计量可以改写为

$$C_i = \frac{1}{p+1} \left(\frac{h_{ii}}{1-h_{ii}} \right) r_i^2, \quad i = 1, \dots, n,$$

其中 r_i 是标准化残差.

- Cook距离统计量 C_i 越大的样本点, 越可能是强影响点或异常点.
- 计算Cook距离统计量的函数`cooks.distance()`.

COVRATIO准则

- 分别计算 $\hat{\beta}$ 和 $\hat{\beta}_{(-i)}$ 的协方差矩阵

$$\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}, \quad \text{Cov}(\hat{\beta}_{(-i)}) = \sigma^2(\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1}.$$

- 对 $i = 1, \dots, n$, 计算

$$C_i = \frac{\det\left(\hat{\sigma}_{(-i)}^2 (\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1}\right)}{\det\left(\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}\right)} = \frac{(\hat{\sigma}_{(-i)}^2)^{p+1}}{(\hat{\sigma}^2)^{p+1}} \frac{1}{1 - h_{ii}}.$$

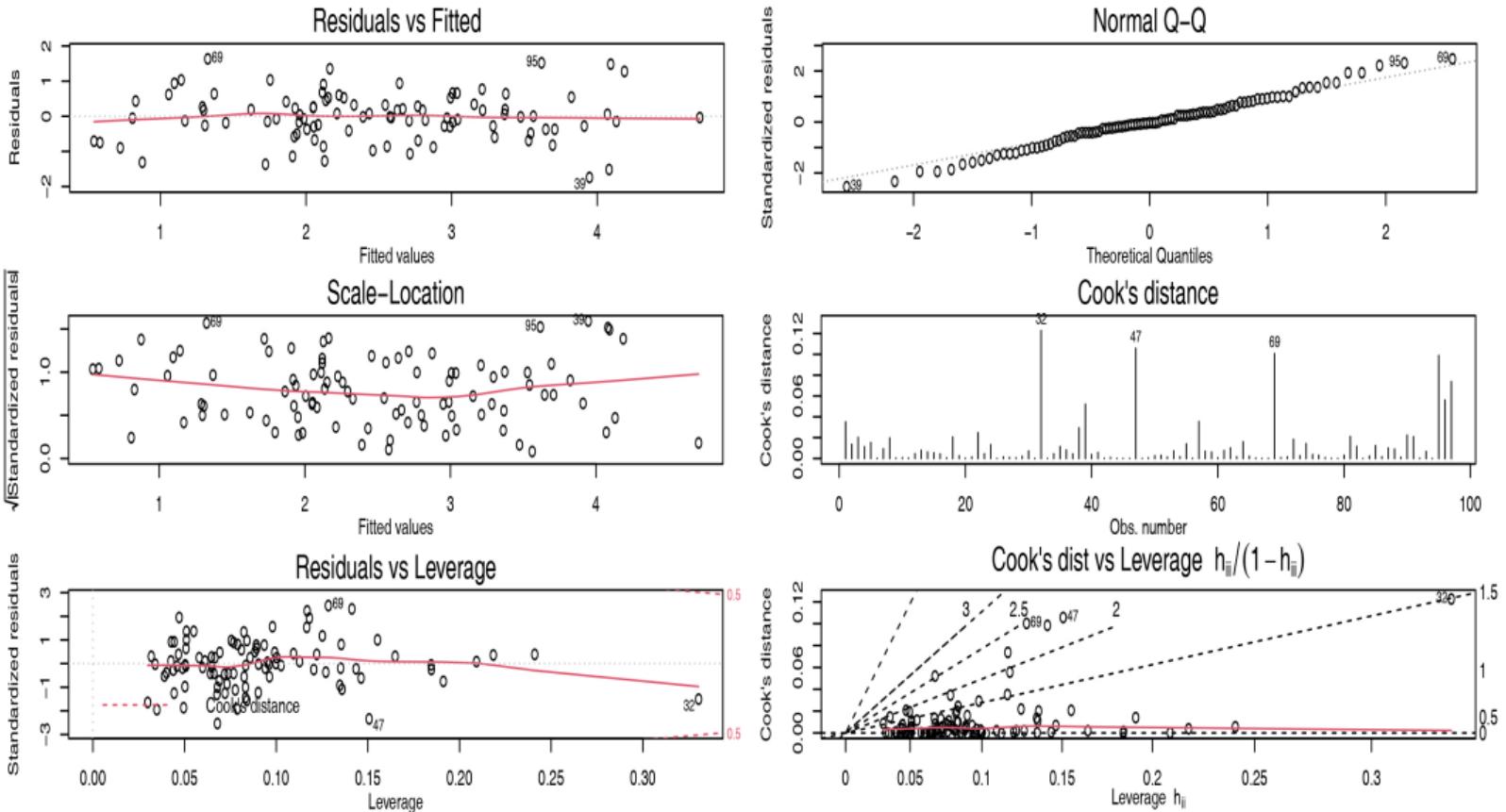
- 如果第 i 个样本所对应的 C_i 值离 1 越远, 则认为该样本影响越大.
- 计算 COVRATIO 统计量的函数 covratio().

前列腺癌数据的回归诊断和影响分析

■ 用函数`plot()`绘制前列腺癌数据的回归诊断图，共展示6个图：

- ① 普通残差与拟合值的残差图；
- ② 残差的QQ图；
- ③ 标准化残差绝对值的开方与拟合值的残差图；
- ④ Cook距离图；
- ⑤ 标准化残差对杠杆图；
- ⑥ Cook距离对杠杆图.

■ 利用函数`hatvalues()`、`dffits()`、`cooks.distance()` 和`covratio()`进行影响分析，确定强影响点或异常点.



前列腺癌数据的模型诊断图

前列腺癌症数据的回归诊断和影响分析

```
n = nrow(prostate); p = ncol(prostate)-1  
> round(prostate[hatvalues(lm.reg)>2*(p+1)/n, ], 4)  
    lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa  
32  0.1823   6.1076  65   1.7047   0 -1.3863       6      0 2.0082  
37  1.4231   3.6571  73  -0.5798   0  1.6582       8     15 2.1576  
41  0.6206   3.1420  60  -1.3863   0 -1.3863       9     80 2.2976  
74  1.8390   3.2367  60   0.4383   1  1.1786       9     90 3.0750  
92  2.5329   3.6776  61   1.3481   1 -1.3863       7     15 4.1296  
> round(prostate[dffits(lm.reg)>2*sqrt((p+1)/n), ], 4)  
    lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa  
69  -0.4463   4.4085  69  -1.3863   0 -1.3863       6      0 2.9627  
95   2.9074   3.3962  52  -1.3863   1  2.4638       7     10 5.1431  
96   2.8826   3.7739  68   1.5581   1  1.5581       7     80 5.4775  
97   3.4720   3.9750  68   0.4383   1  2.9042       7     20 5.5829
```

前列腺癌症数据的回归诊断和影响分析

```
> round(prostate[cooks.distance(lm.reg)>0.1, ], 4)
    lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
32   0.1823   6.1076  65   1.7047   0 -1.3863       6      0 2.0082
47   2.7279   3.9954  79   1.8795   1  2.6568       9    100 2.5688
69  -0.4463   4.4085  69  -1.3863   0 -1.3863       6      0 2.9627
s = rep("", n); co = covratio(lm.reg)
abs.co = abs(co-1); s[abs.co==max(abs.co)] = "*"
> data.frame(COVRATIO = co, s)

COVRATIO s
41 1.4363072 *
```

■ 对于这些结果是否合理, 还需要作进一步的研究.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

- **多重共线性**是指线性回归模型中的协变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确.
- 如果存在不全为0的常数 $a_0, a_1, a_2, \dots, a_p$, 使得

$$a_1X_1 + a_2X_2 + \cdots + a_pX_p = a_0.$$

- 如果数据中所有样本都满足上式, 则称协变量 X_1, X_2, \dots, X_p 存在**精确共线性**.
- 如果上式对观测数据近似成立, 则有**近似共线性**, 也就表示这 p 个协变量存在**多重共线性**.

- 假设 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ 是协变量 X_1, X_2, \dots, X_p 经过中心化或标准化得到的观测向量, 记 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$ 为 $n \times p$ 的设计矩阵, 其中 $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{ni})^T$.
- 考虑多元线性回归模型: $\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$, 其中 $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.
- 设 λ 为 $\mathbf{X}^T \mathbf{X}$ 的一个特征值, $\boldsymbol{\phi}$ 为其对应的单位特征向量, 即 $\boldsymbol{\phi}^T \boldsymbol{\phi} = 1$.
- 若 $\lambda \approx 0$, 则

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\phi} = \lambda \boldsymbol{\phi} \approx \mathbf{0}.$$

■ 用 ϕ^T 左乘上式, 得到

$$\phi^T \mathbf{X}^T \mathbf{X} \phi = \lambda \phi^T \phi = \lambda \approx 0.$$

■ 则有, $\mathbf{X}\phi \approx \mathbf{0}$, 其中 $\phi = (\phi_1, \phi_2, \dots, \phi_p)^T$, 即

$$\phi_1 \mathbf{x}_1 + \phi_2 \mathbf{x}_2 + \dots + \phi_p \mathbf{x}_p \approx \mathbf{0},$$

■ 若矩阵的某个特征值接近零, 就意味着矩阵 \mathbf{X} 的列向量 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ 之间存在近似线性关系.

- 下面讨论多重共线性问题对回归系数最小二乘估计的影响.
- 计算 $\hat{\beta}$ 的均方误差为

$$\text{MSE}(\hat{\beta}) = \text{E}[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)] = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i},$$

其中假设 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ 是矩阵 $\mathbf{X}^T \mathbf{X}$ 的特征值.

- 显然, 如果矩阵 $\mathbf{X}^T \mathbf{X}$ 至少有一个特征值非常接近于零, 则 $\text{MSE}(\hat{\beta})$ 就会很大, 则最小二乘估计 $\hat{\beta}$ 也就不再是回归系数 β 的一个好的估计.

■ 判断多重共线性及其严重程度的方法有：

- ① 条件数方法
- ② 方差膨胀因子(VIF)方法

■ **条件数方法：**度量多重共线性严重程度的一个重要指标是矩阵 $\mathbf{X}^T\mathbf{X}$ 的条件数，定义为

$$\kappa(\mathbf{X}^T\mathbf{X}) = \|\mathbf{X}^T\mathbf{X}\| \cdot \|(\mathbf{X}^T\mathbf{X})^{-1}\| = \frac{\lambda_{\max}(\mathbf{X}^T\mathbf{X})}{\lambda_{\min}(\mathbf{X}^T\mathbf{X})}.$$

■ $\lambda_{\max}(\mathbf{X}^T\mathbf{X})$ 和 $\lambda_{\min}(\mathbf{X}^T\mathbf{X})$ 分别是矩阵 $\mathbf{X}^T\mathbf{X}$ 的最大和最小特征值。

■ 条件数刻画了矩阵 $\mathbf{X}^T\mathbf{X}$ 特征值差异性的大小：

- 若 $\kappa < 100$, 则认为多重共线性的程度很小;
- 若 $100 \leq \kappa \leq 1000$, 则认为存在中等程度或较强的多重共线性;
- 若 $\kappa > 1000$, 则认为存在严重的多重共线性.

■ R 语言提供了计算条件数的函数kappa().

- 方差膨胀因子(VIF)是衡量多元线性回归模型中多重共线性严重程度的又一种度量.
- 假设模型中数据已进行标准化，则回归系数最小二乘估计的协差矩阵为 $\sigma^2 \mathbf{R}_X^{-1}$ ，其中 \mathbf{R}_X 是协变量 X 的样本相关系数矩阵.
- 方差膨胀因子定义为

$$\text{VIF}(\hat{\beta}_k) = \frac{1}{1 - R_{X_k|X_{(-k)}}^2}, \quad k = 1, \dots, p,$$

其中 $R_{X_k|X_{(-k)}}^2$ 是第 k 个协变量 X_k 与其余的 $p - 1$ 个协变量 $X_{(-k)}$ 之间的判定系数.

■ VIF的最小可能的值是1, 表示完全不存在多重共线性.

■ 当方差膨胀因子VIF($\hat{\beta}_k$) 越大, 则表示越存在多重共线性.

原因: 当第 k 个协变量与其余的协变量之间相关程度越高, 即判定系数 $R_{X_k|X_{(-k)}}^2$ 越接近于1, 相应的方差膨胀因子VIF($\hat{\beta}_k$)也就越大.

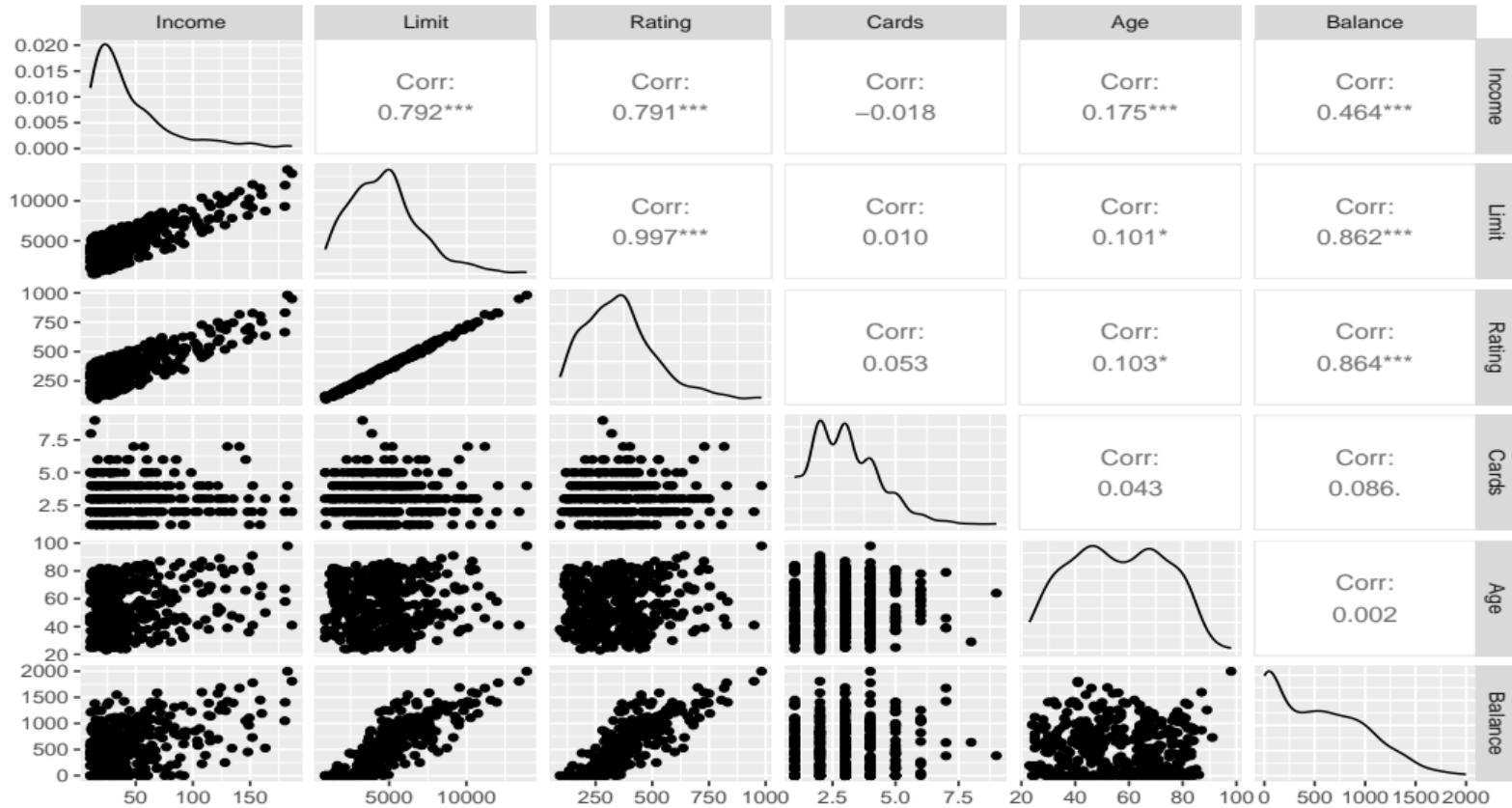
■ 实际应用中, 一个经验法则是当方差膨胀因子VIF 的值超过5或10, 就表示存在多重共线性问题.

■ 在R语言中, 可用程序包car中的函数vif()计算方差膨胀因子.

例: Credit数据

考虑来自James等(2021)的Credit数据, 可从程序包ISLR中获取该数据集, 该数据集包含400个样本和11个变量. 本例仅考虑5个定量协变量Income, Limit, Rating, Cards 和Age对响应变量Balance的多元线性回归问题.

试用求矩阵条件数和方差膨胀因子的方法, 分析协变量间是否存在多重共线性问题.



多重共线性— Credit数据

```
library(ISLR); library(car); attach(Credit)
XX = cor(Credit[2:6])
> kappa(XX, exact = TRUE)
[1] 1256.042
lm.C=lm(Balance~Income + Limit + Rating + Cards + Age, data=Credit)
> round(vif(lm.C), 3)
      Income       Limit       Rating       Cards       Age
    2.773     227.843     229.588      1.434     1.038
```

- 条件数: $\kappa = 1256.042 > 1000$, 认为有严重的多重共线性.
- 变量Limit和Rating的方差膨胀因子表明它们之间有共线性.

- 协变量Income, Limit, Rating, Cards 和Age的相关系数矩阵的最小特征值和相应的特征向量为

$$\lambda_{\min} = 0.00219, \quad \phi = (0.0022, 0.7054, -0.7081, 0.0306, 0.0002)^T.$$

- 对每个样本点, 近似都有

$$0.0022\text{Income} + 0.7054\text{Limit} - 0.7081\text{Rating} + 0.0306\text{Cards} + 0.0002\text{Age} \approx 0.$$

- 由于Income, Cards和Age的系数近似为0, 因此有

$$0.7054\text{Limit} - 0.7081\text{Rating} \approx 0.$$

■ 所以, 存在不全为0的常数 a_0, a_1, a_2 , 使得

$$a_1 \times \text{Limit} + a_2 \times \text{Rating} \approx a_0.$$

■ 说明变量Limit和Rating存在多重共线性.

■ 从VIF得到的结果可以看出, 变量Limit和Rating 的VIF都远远大于10, 而变量Income, Cards和Age的VIF小于5, 则表明变量Limit 和Rating存在多重共线性.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

♠ **问题:** 如何处理模型误差 $\varepsilon_1, \dots, \varepsilon_n$ 是相关且异方差的情形?

♠ **问题:** 如何处理模型误差 $\varepsilon_1, \dots, \varepsilon_n$ 是相关且异方差的情形?

■ 针对多元线性回归模型, 但是假设模型误差满足

$$E(\varepsilon|\mathbf{X}) = \mathbf{0}, \quad \text{Cov}(\varepsilon|\mathbf{X}) = \sigma^2 \mathbf{W},$$

其中 \mathbf{W} 是一个已知的 $n \times n$ 正定矩阵.

■ 如果 $\mathbf{W} = \mathbf{I}_n$ 时, 上式假设退化为 Gauss-Markov 条件.

■ 如果 $\mathbf{W} = \text{diag}\{w_{11}, \dots, w_{nn}\}$ 时, 模型变为异方差的多元线性回归模型.

加权最小二乘方法

■ 令 $\mathbf{W}^{-1/2}$ 是 \mathbf{W}^{-1} 的平方根, 即有 $(\mathbf{W}^{-1/2})^T \mathbf{W}^{-1/2} = \mathbf{W}^{-1}$, 则

$$\text{Cov}(\mathbf{W}^{-1/2} \boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n.$$

■ 可见: $\mathbf{W}^{-1/2} \boldsymbol{\varepsilon}$ 是不相关且同方差的.

■ 令 $\mathbf{Y}^* = \mathbf{W}^{-1/2} \mathbf{Y}$, $\mathbf{X}^* = \mathbf{W}^{-1/2} \mathbf{X}$, $\boldsymbol{\varepsilon}^* = \mathbf{W}^{-1/2} \boldsymbol{\varepsilon}$.

■ 这时, 可得如下的多元线性回归模型

$$\mathbf{Y}^* = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}^*,$$

其中 $\boldsymbol{\varepsilon}^*$ 满足 Gauss-Markov 条件.

加权最小二乘方法

■ 定义如下的**加权最小二乘目标函数**

$$Q(\boldsymbol{\beta}) = \|Y^* - \mathbf{X}^* \boldsymbol{\beta}\|_2^2 = (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta})^T \mathbf{W}^{-1} (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta}).$$

■ 极小化加权最小二乘目标函数 $Q(\boldsymbol{\beta})$, 可得 $\boldsymbol{\beta}$ 的**加权最小二乘估计**为

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^{*T} \mathbf{X}^*)^{-1} \mathbf{X}^{*T} Y^* = (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{Y}.$$

■ 可证明, 即使 \mathbf{W} 被错误指定, $\hat{\boldsymbol{\beta}}_{WLS}$ 仍然是**无偏的**, 即

$$E(\hat{\boldsymbol{\beta}}_{WLS} | \mathbf{X}) = (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}.$$

■ 当 \mathbf{W} 被正确指定时, 即 $\mathbf{W} = \mathbf{W}_0$ 时, $Cov(\hat{\boldsymbol{\beta}}_{WLS}) = \sigma^2 (\mathbf{X}^T \mathbf{W}_0^{-1} \mathbf{X})^{-1}$,
则 $\hat{\boldsymbol{\beta}}_{WLS}$ 是 $\boldsymbol{\beta}$ 的**最有效估计**.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

♠ **问题:** 当协变量和响应变量之间是非线性的, 如何用线性回归模型对这些问题进行分析?

Box-Cox变换

♠ **问题:** 当协变量和响应变量之间是非线性的, 如何用线性回归模型对这些问题进行分析?

■ Box和Cox(1964)提出了**Box-Cox 变换方法**, 即考虑下面的变换函数

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \text{如果 } \lambda \neq 0, \\ \log(Y), & \text{如果 } \lambda = 0, \end{cases}$$

其中 λ 是一个未知参数.

■ 考虑如下的Box-Cox模型

$$Y^{(\lambda)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

Box-Cox变换

■ 给定i.i.d.的样本 $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$, Box-Cox模型的似然函数为

$$L(\lambda, \boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \|\mathbf{Y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\} J(\lambda, \mathbf{Y}),$$

其中

- ▶ $\mathbf{Y}^{(\lambda)} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})^T$ 是 $n \times 1$ 的向量
- ▶ \mathbf{X} 是 $n \times (p + 1)$ 的设计矩阵
- ▶ $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ 是 $p + 1$ 维的未知参数向量
- ▶ $J(\lambda, \mathbf{Y}) = \prod_{i=1}^n \left| \frac{\partial y_i^{(\lambda)}}{\partial y_i} \right| = \left(\prod_{i=1}^n |y_i| \right)^{\lambda-1}$

■ 对给定的 λ , 可得 β 和 σ^2 的极大似然估计分别为

$$\hat{\beta}(\lambda) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}^{(\lambda)}, \quad \hat{\sigma}^2(\lambda) = \frac{1}{n} \|\mathbf{Y}^{(\lambda)} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}^{(\lambda)}\|_2^2.$$

■ 把估计量 $\hat{\beta}(\lambda)$ 和 $\hat{\sigma}^2(\lambda)$ 代入到似然函数 $L(\lambda, \beta, \sigma^2)$, 可得 λ 的对数似然函数为

$$\log(L(\lambda)) = (\lambda - 1) \sum_{i=1}^n \log(|y_i|) - \frac{n}{2} \log(\hat{\sigma}^2(\lambda)) - \frac{n}{2}.$$

■ 这时, 可得 λ 的极大似然估计为: $\hat{\lambda} = \arg \max_{\lambda} \log(L(\lambda)).$

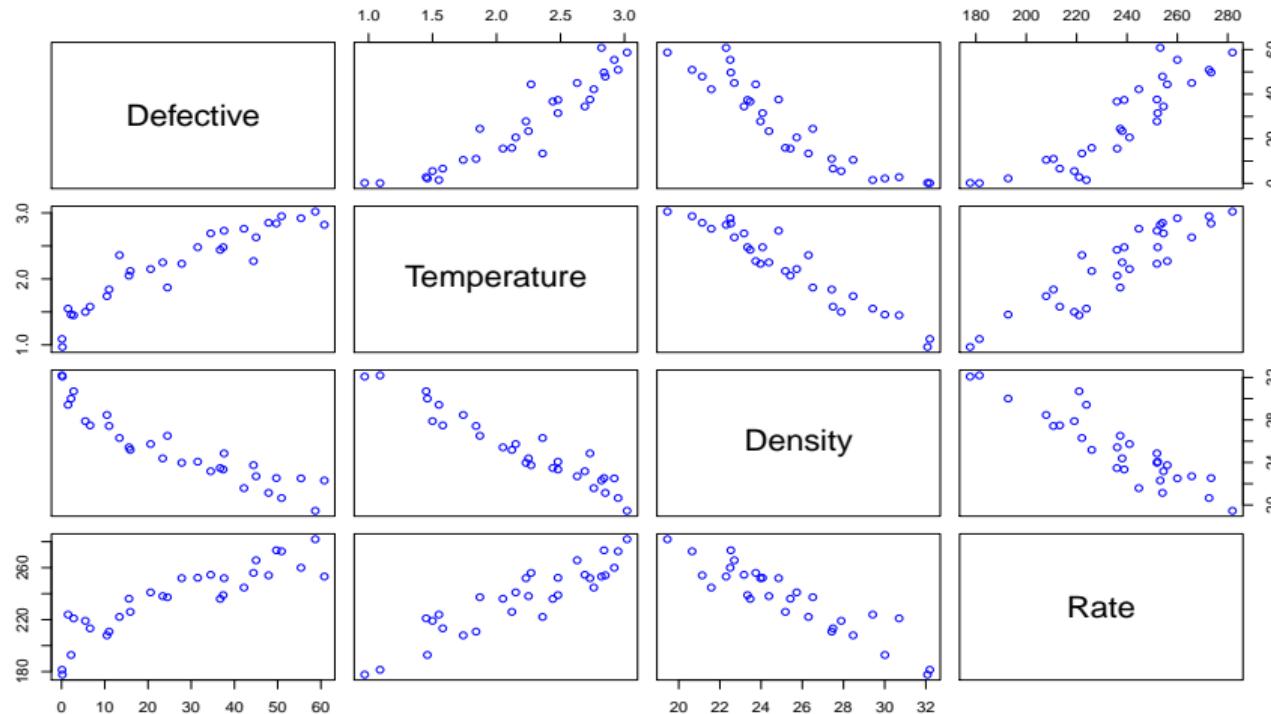
■ 最后, 可得 β 和 σ^2 的极大似然估计分别为 $\hat{\beta}(\hat{\lambda})$ 和 $\hat{\sigma}^2(\hat{\lambda})$.

例: 产品零件缺陷率数据分析

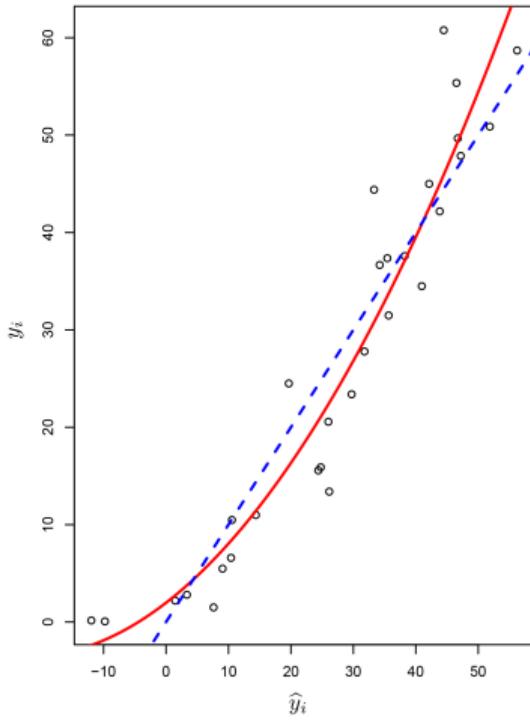
对Siegel(1997)提供的某产品零件缺陷率(defective rates)数据进行举例说明, 该数据也被Sheather(2009)进行了详细的分析. Y 表示每生产1000个零件的平均缺陷率(Defective), X_1 表示温度(Temperature), X_2 表示密度(Density), X_3 表示生产速度(Rate).

编号	X_1	X_3	X_5	Y	编号	X_1	X_3	X_5	Y
1	0.97	32.08	177.7	0.2	16	2.76	21.58	244.7	42.2
2	2.85	21.14	254.1	47.9	17	2.36	26.30	222.1	13.4
3	2.95	20.65	272.6	50.9	18	1.09	32.19	181.4	0.1
4	2.84	22.53	273.4	49.7	19	2.15	25.73	241.0	20.6
5	1.84	27.43	210.8	11.0	20	2.12	25.18	226.0	15.9
6	2.05	25.42	236.1	15.6	21	2.27	23.74	256.0	44.4
7	1.50	27.89	219.1	5.5	22	2.73	24.85	251.9	37.6
8	2.48	23.34	238.9	37.4	23	1.46	30.01	192.8	2.2
9	2.23	23.97	251.9	27.8	24	1.55	29.42	223.9	1.5
10	3.02	19.45	281.9	58.7	25	2.92	22.50	260.0	55.4
11	2.69	23.17	254.5	34.5	26	2.44	23.47	236.0	36.7
12	2.63	22.70	265.7	45.0	27	1.87	26.51	237.3	24.5
13	1.58	27.49	213.3	6.6	28	1.45	30.70	221.0	2.8
14	2.48	24.07	252.2	31.5	29	2.82	22.30	253.2	60.8
15	2.25	24.38	238.1	23.4	30	1.74	28.47	207.9	10.5

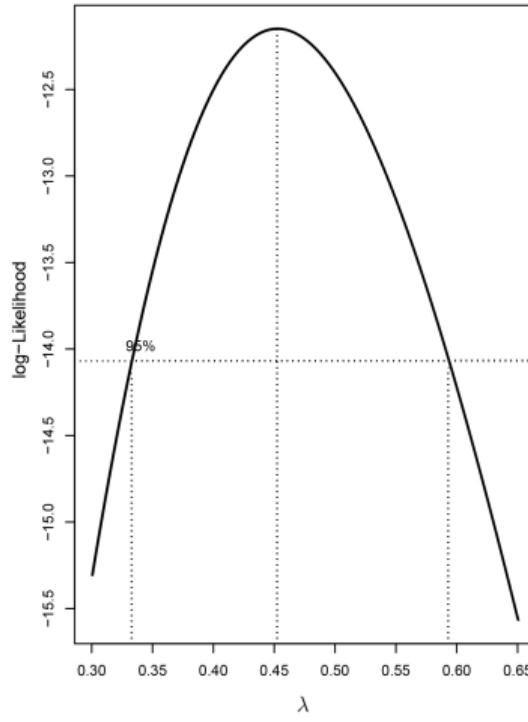
Box-Cox变换—产品零件缺陷率数据分析



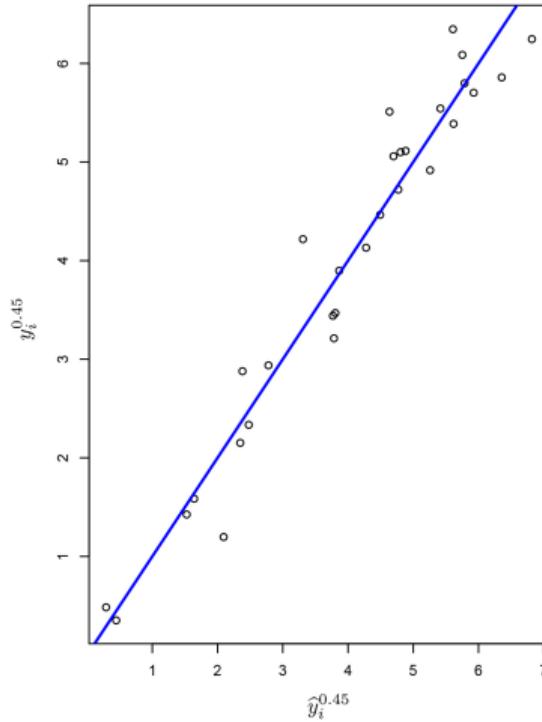
Box-Cox变换—产品零件缺陷率数据分析



(a)



(b)



(c)

- 从散点图可看出, 三个协变量之间具有一定的线性关系, 但是每个协变量和响应变量 Y 之间具有非线性关系.
- 考虑下面的多元线性回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad i = 1, \dots, 30.$$

- 图(a)显示, 拟合值 \hat{y}_i 和 y_i 之间用线性函数拟合, 效果较差, 而用二次函数拟合效果更好.
- 用Box-Cox变换对响应变量进行变换, 为了选取最优的 λ , 采用程序包MASS中的函数`boxcox()`选取最优的 λ .

- 图(b)提供了Box-Cox变换方法的对数似然, λ 的最优取值约为0.45, 且 λ 的95%置信区间不包含0.
- 考虑如下的多元线性回归模型

$$y_i^{0.45} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad i = 1, \dots, 30.$$

- 图(c)提供了所得拟合值 $\hat{y}_i^{0.45}$ 和 $y_i^{0.45}$ 的散点图, 以及拟合直线.
- 经过Box-Cox变换后, 模型对数据的拟合效果更好, 并有好的预测效果.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

♠ **问题:** 协变量中存在定性变量时, 如何进行建模和分析?

♠ **问题:** 协变量中存在定性变量时, 如何进行建模和分析?

例: 工资调查数据

考虑Chatterjee和Hadi (2006) 中计算机专业人员的工资调查数据, 该数据集包含46个样本和4 个变量, 其中响应变量 S 表示雇员工资(单位: 美元), X 表示工作经历(单位: 年), E 表示教育水平(1为高中学历, 2为学士学位, 3为高级学位), M 表示是否为管理人员(1为管理人员, 0为其他). 在这个数据中, X 是定量协变量, E 和 M 是定性协变量, 工资调查的目的是确定和量化决定工资差异的变量.

编号	S	X	E	M	编号	S	X	E	M
1	13876	1	1	1	24	22884	6	2	1
2	11608	1	3	0	25	16978	7	1	1
3	18701	1	3	1	26	14803	8	2	0
4	11283	1	2	0	27	17404	8	1	1
5	11767	1	3	0	28	22184	8	3	1
6	20872	2	2	1	29	13548	8	1	0
7	11772	2	2	0	30	14467	10	1	0
8	10535	2	1	0	31	15942	10	2	0
9	12195	2	3	0	32	23174	10	3	1
10	12313	3	2	0	33	23780	10	2	1
11	14975	3	1	1	34	25410	11	2	1
12	21371	3	2	1	35	14861	11	1	0
13	19800	3	3	1	36	16882	12	2	0
14	11417	4	1	0	37	24170	12	3	1
15	20263	4	3	1	38	15990	13	1	0
16	13231	4	3	0	39	26330	13	2	1
17	12884	4	2	0	40	17949	14	2	0
18	13245	5	2	0	41	25685	15	3	1
19	13677	5	3	0	42	27837	16	2	1
20	15965	5	1	1	43	18838	16	2	0
21	12336	6	1	0	44	17483	16	1	0
22	21352	6	3	1	45	19207	17	2	0
23	13839	6	2	0	46	19346	20	1	0



- 假设忽略其他变量, 只考虑协变量 M 对工资 S 的影响, 即调查管理人员和非管理人员的工资差异.
- 对于定性变量, 也被称为**因子变量**(factor variable).
- 定性变量 M 是一个**二值变量**, 表示**两个水平**(levels).
- 在建模时, 只需给二值变量 M 创建一个指标, 称为**哑变量** (dummy variable), 即

$$M_i = \begin{cases} 1, & \text{第 } i \text{ 个雇员为管理人员,} \\ 0, & \text{其他.} \end{cases}$$

- 在回归模型中仅使用定性变量 M , 并考虑下面的模型

$$S_i = \beta_0 + \beta_1 M_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{第 } i \text{ 个雇员为管理人员,} \\ \beta_0 + \varepsilon_i, & \text{其他.} \end{cases}$$

- β_0 可解释为非管理人员的平均工资;
- $\beta_0 + \beta_1$ 为管理人员的平均工资;
- 因此, β_1 是管理人员和其他人员之间工资的平均差异.

定性协变量建模

仅考虑定性变量 M 的回归分析结果

变量	系数估计	标准误差	t统计量	p值
β_0	14285.3	639.8	22.327	< 0.0001
M	6865.2	970.3	7.075	< 0.0001
$n = 46$	$R^2 = 0.5322$	$R^2_{\text{adj}} = 0.5216$	3262.41	d.f. = 44

- ① 非管理人员的平均工资为14285.3美元；
- ② 管理人员的平均工资比非管理人员的平均工资多6865.2美元，共为14285.3美元+6865.2美元=21150.5美元.
- ③ 定性变量 M 的p值远远小于显著性水平0.05, 故认为是否为管理人员的平均工资具有显著性的差异.

- 当一个定性协变量有两个以上水平时，单个哑变量不能代表所有可能的取值。这时，需要考虑更多的哑变量。
- 例如，对定性协变量 E ，取值为三种情况：1为高中学历，2为学士学位，3为高级学位，可以考虑两个哑变量。
- 第一个哑变量和第二个哑变量分别为

$$E_{i1} = \begin{cases} 1, & \text{第 } i \text{ 个雇员为学士学位,} \\ 0, & \text{其他.} \end{cases}, \quad E_{i2} = \begin{cases} 1, & \text{第 } i \text{ 个雇员为高级学位,} \\ 0, & \text{其他.} \end{cases}$$

■ 仅考虑定性变量 E , 并考虑上面两个哑变量建模, 则回归模型为

$$S_i = \beta_0 + \beta_2 E_{i1} + \beta_3 E_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_2 + \varepsilon_i, & \text{第 } i \text{ 个雇员为学士学位,} \\ \beta_0 + \beta_3 + \varepsilon_i, & \text{第 } i \text{ 个雇员为高级学位,} \\ \beta_0 + \varepsilon_i, & \text{其他.} \end{cases}$$

- β_0 可解释为高中学历雇员的平均工资;
- β_2 是学士学位雇员和高中学历雇员之间工资的平均差异;
- β_3 是高级学位雇员和高中学历雇员之间工资的平均差异;
- $\beta_0 + \beta_2$ 为具有学士学位雇员的平均工资; $\beta_0 + \beta_3$ 为具有高级学位雇员的平均工资.

仅考虑定性变量 E 的回归分析结果

变量	系数估计	标准误差	t统计量	p值
β_0	14942	1217	12.275	< 0.0001
E_1	3345	1604	2.085	0.0430
E_2	3351	1754	1.910	0.0628
$n = 46$	$R^2 = 0.109$	$R^2_{\text{adj}} = 0.06757$	4554.48	d.f. = 43

- ① 高中学历雇员的平均工资为14942美元;
- ② 学士学位雇员的平均工资比高中学历雇员的平均工资多3345美元;
- ③ 高级学位雇员的平均工资比高中学历雇员的平均工资多3351美元;
- ④ 仅考虑定性协变量 E 时的判定系数为 $R^2 = 0.109$, 表示仅考虑定性协变量 E 时的回归方程不是很显著.

定性协变量建模

当定性协变量和定量协变量同时存在时, 可以考虑如下的多元线性回归模型

$$S_i = \beta_0 + \beta_1 M_i + \beta_2 E_{i1} + \beta_3 E_{i2} + \beta_4 X_i + \varepsilon_i, \quad i = 1, \dots, 46.$$

变量	系数估计	标准误差	t统计量	p值
β_0	8035.60	386.69	20.781	< 0.0001
M	6883.53	313.92	21.928	< 0.0001
E_1	3144.04	361.97	8.686	< 0.0001
E_2	2996.21	411.75	7.277	< 0.0001
X	546.18	30.52	17.896	< 0.0001
$n = 46$	$R^2 = 0.9568$	$R^2_{\text{adj}} = 0.9525$	1027.44	d.f. = 41

- ① 判定系数为 $R^2 = 0.9568$, 说明回归方程是非常显著的;
- ② 变量 X 回归系数的估计为 $\hat{\beta}_4 = 546.18$, 说明工作经历增加1年, 雇员的工资将平均增加**546.18**美元;
- ③ $\hat{\beta}_1 = 6883.53$, 说明管理人员认工的平均增长值;
- ④ $\hat{\beta}_2 = 3144.04$ 度量了具有学士学位雇员相对高中学历雇员平均工资的差异;
- ⑤ $\hat{\beta}_3 = 2996.21$ 度量了具有高级学位雇员相对高中学历雇员平均工资的差异;
- ⑥ $\hat{\beta}_2 - \hat{\beta}_3 = 147.83$ 表示学士学位雇员比高级学位雇员的平均工资多**147.83**美元.

■ 进一步, 可以在模型中考虑变量之间的**交互效应**, 即

$$S_i = \beta_0 + \beta_1 M_i + \beta_2 E_{i1} + \beta_3 E_{i2} + \beta_4 X_i + \beta_5 (E_{i1} \times M_i) + \beta_6 (E_{i2} \times M_i) + \varepsilon_i.$$

变量	系数估计	标准误差	t统计量	p值
β_0	9472.685	80.344	117.90	< 0.0001
M	3981.377	101.175	39.35	< 0.0001
E_1	1381.671	77.319	17.87	< 0.0001
E_2	1730.748	105.334	16.43	< 0.0001
X	496.987	5.566	89.28	< 0.0001
$E_1 \times M$	4902.523	131.359	37.32	< 0.0001
$E_2 \times M$	3066.035	149.330	20.53	< 0.0001
$n = 46$	$R^2 = 0.9988$	$R^2_{\text{adj}} = 0.9986$	173.8	d.f. = 39

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

- Belsley, D., Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons.
- Box, G. and Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B, 26: 211–252.
- Chatterjee, S. and Hadi, A. S. (2006). Regression Analysis by Example (4th Edition). New Jersey: John Wiley & Sons.
- Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics, 19: 15–18.

- Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32: 17–22.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*. 2nd Ed. New York: Springer-Verlag.
- Seber, G. A. F. and Lee, A. J. (2003) *Linear Regression Analysis* (2nd Edition). New York: John Wiley & Sons.
- Sheather, S. J. (2009). *A Modern Approach to Regression with R*. New York: Springer.
- Siegel, A. (1997). *Practical Business Statistics* (3rd Edition). Boston: Irwin McGraw-Hill.

1 线性回归模型

- 模型介绍
- 最小二乘估计
- σ^2 的估计
- 假设检验
- 预测区间与置信区间
- R语言函数及应用

2 回归诊断

- 什么是回归诊断？
- 残差
- 残差图
- 影响分析
- 多重共线性

3 加权最小二乘方法

4 Box-Cox变换

5 定性协变量建模

6 参考文献

7 作业

作业

[习题见教材: 统计学习(R语言版) — 习题3]

- **课后思考题:** 第4题、第5题、第6题、第9题、第15题
- **需要完成的课后作业:** 第2题、第3题、第7题、第10题
- **应用:** 第14题、第17题、第18题. 具体要求:
 - ① 能使用R语言把数据读入, 并对数据中的每个变量进行了解;
 - ② 能用学过的一些统计方法, 按照题目要求, 利用R语言对数据进行一些简单的分析, 并思考数据分析的结果.



谢谢，请多提宝贵意见！