



北京航空航天大学

— 经济管理学院 —

BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第 3 讲：多元线性回归模型

康雁飞

数量经济与商务统计系

Outline

- 1 多元线性回归模型
- 2 模型估计及估计的性质
- 3 模型的统计推断
- 4 模型诊断
- 5 变量选择
- 6 扩展
- 7 Key points & 上机实验

Outline

- 1 多元线性回归模型
- 2 模型估计及估计的性质
- 3 模型的统计推断
- 4 模型诊断
- 5 变量选择
- 6 扩展
- 7 Key points & 上机实验

多元线性回归模型

- 因变量只受单一自变量影响的情况非常少见。
- 通常影响一个变量的变量有多个。
 - ▶ 房价 (y) 预测: $x_1 = \text{size}(\text{ft}^2)$, $x_2 = \text{age}(\text{years})$, $x_3 = \text{numbers of rooms}$,
 $x_4 = \text{number of bathrooms}$
- 一元线性回归 \Rightarrow 多元线性回归
- 注: 后续对于多元回归模型和分析, 使用矩阵和向量形式更为简洁
(请自行回顾矩阵和向量相关内容)

多元线性回归模型

模型

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \dots, n$$

其中 $\epsilon_i \sim N(0, \sigma^2)$, 且独立.

- p 个自变量: x_1, \cdots, x_p
- $p + 2$ 个待估参数: $\beta_0, \beta_1, \cdots, \beta_p, \sigma^2$
- 回归方程 $E(Y_i | x_1, \cdots, x_p) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$
- β_j 的含义: 在控制 $x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p$ 下, x_j 变动一个单位对 Y 的平均影响

同一元线性回归模型假设 (LINE):

- 线性关系 (linear function)
- 独立性 (independence)
- 正态性 (normal distributed)
- 同方差性 (equal variance)

Outline

- 1 多元线性回归模型
- 2 模型估计及估计的性质
- 3 模型的统计推断
- 4 模型诊断
- 5 变量选择
- 6 扩展
- 7 Key points & 上机实验

最小二乘估计

- 最小二乘法：最小化残差平方和
- 极小化得到最小二乘估计

$$f(\beta_0, \beta_1, \beta_2, \dots, \beta_p) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}))^2,$$

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \arg \min f(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$$

- 对 $\beta_0, \beta_1, \dots, \beta_p$ 求偏导, 得到

$$n\beta_0 + \left(\sum_{i=1}^n x_{i1}\right)\beta_1 + \dots + \left(\sum_{i=1}^n x_{ip}\right)\beta_p = \sum_{i=1}^n Y_i$$

$$\left(\sum_{i=1}^n x_{i1}\right)\beta_0 + \left(\sum_{i=1}^n x_{i1}^2\right)\beta_1 + \dots + \left(\sum_{i=1}^n x_{i1}x_{ip}\right)\beta_p = \sum_{i=1}^n x_{i1}Y_i$$

⋮

$$\left(\sum_{i=1}^n x_{ip}\right)\beta_0 + \left(\sum_{i=1}^n x_{i1}x_{ip}\right)\beta_1 + \dots + \left(\sum_{i=1}^n x_{ip}^2\right)\beta_p = \sum_{i=1}^n x_{ip}Y_i.$$

- 回归方程

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \dots + \hat{\beta}_px_p$$

矩阵形式

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$Y = X\beta + \epsilon$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

估计方程

$$\hat{\beta} = \arg \min_{\beta} \|Y - X\beta\|_2^2$$

然后求导数得到方程：

$$X^T X \beta = X^T Y.$$

然后我们可以得到参数估计：

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

几点注意

$\hat{\beta}$ 的计算中需要有 $(X^T X)^{-1}$

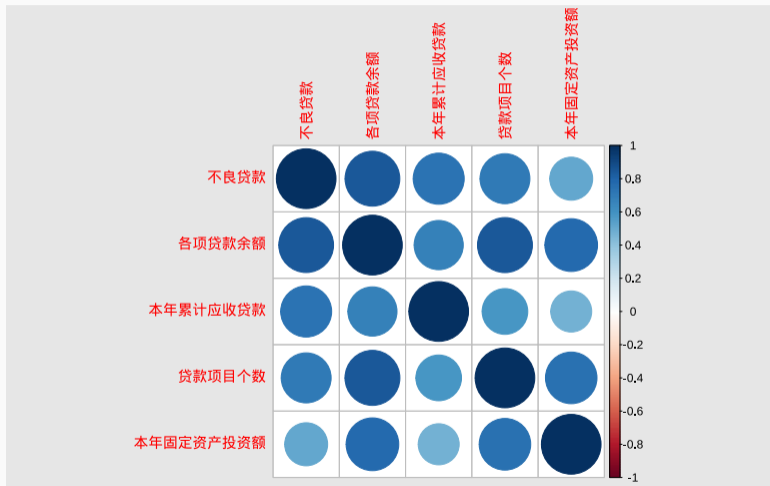
- 1 要求 X 矩阵必须为列满秩，即 $\text{rank}(X) = p + 1$
- 2 理论上： X 的列变量线性独立，不能存在线性相关的列变量
- 3 事实上：实际数据中无法保证 x_1, x_2, \dots, x_p 线性无关，可容许有低相关的协变量。在进行多元线性回归之前，需要进行变量之间的多重共线性检验（或者通过相关系数矩阵初步观察是否存在多重共线性的变量）

R 中的多元线性回归

例：某商业银行 25 家分行主要业务数据

```
## tibble [25 x 5] (S3: tbl_df/tbl/data.frame)
## $ 不良贷款          : num [1:25] 0.9 1.1 4.8 3.2 7.8 2.7 1.6 12.5 1 2.6 ...
## $ 各项贷款余额      : num [1:25] 67.3 111.3 173 80.8 199.7 ...
## $ 本年累计应收贷款  : num [1:25] 6.8 19.8 7.7 7.2 16.5 2.2 10.7 27.1 1.7 9.1 ...
## $ 贷款项目个数      : num [1:25] 5 16 17 10 19 1 17 18 10 14 ...
## $ 本年固定资产投资额: num [1:25] 51.9 90.9 73.7 14.5 63.2 2.2 20.2 43.8 55.9 64.3 ...
```

相关性



相关系数矩阵

##	不良贷款	各项贷款余额	本年累计应收贷款	贷款项目个数
## 不良贷款	1.0000000	0.8435714	0.7315050	0.7002815
## 各项贷款余额	0.8435714	1.0000000	0.6787718	0.8484164
## 本年累计应收贷款	0.7315050	0.6787718	1.0000000	0.5858315
## 贷款项目个数	0.7002815	0.8484164	0.5858315	1.0000000
##	本年固定资产投资额			
## 不良贷款	0.5185181			
## 各项贷款余额	0.7797022			
## 本年累计应收贷款	0.4724310			
## 贷款项目个数	0.7466458			
## 本年固定资产投资额	1.0000000			

回归模型

```
loan.model <- lm(不良贷款~各项贷款余额+  
                本年累计应收贷款+  
                贷款项目个数+  
                本年固定资产投资额,  
                data = loan)  
loan.model.summary <- summary(loan.model)
```

回归模型

```
##
## Call:
## lm(formula = 不良贷款 ~ 各项贷款余额 + 本年累计应收贷款 +
##      贷款项目个数 + 本年固定资产投资额, data = loan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.9198 -0.9507 -0.2880  1.0334  3.1037
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.02164    0.78237  -1.306  0.20643
## 各项贷款余额    0.04004    0.01043   3.837  0.00103 **
## 本年累计应收贷款  0.14803    0.07879   1.879  0.07494 .
## 贷款项目个数    0.01453    0.08303   0.175  0.86285
## 本年固定资产投资额 -0.02919    0.01507  -1.937  0.06703 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.779 on 20 degrees of freedom
## Multiple R-squared:  0.7976, Adjusted R-squared:  0.7571
## F-statistic: 19.7 on 4 and 20 DF, p-value: 1.035e-06
```

估计误差项的方差

■ 残差平方和 $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

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

2 误差项方差 σ^2 可以通过 s_e^2 来估计:

$$s_e^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1} = \frac{\mathbf{e}^\top \mathbf{e}}{n - p - 1}$$

■ 注意: $E(s_e^2) = \sigma^2$.

估计误差项的标准差

```
loan.model.summary$sigma
```

```
## [1] 1.778752
```

$\hat{\beta}$ 的分布

我们可以得到：

$$\hat{\beta} \sim N\left(\beta, \sigma^2 (X^T X)^{-1}\right), \text{Cov}(\hat{\beta}, e) = 0$$

因此：

$$E[\hat{\beta}_j] = \beta_j, \text{Var}[\hat{\beta}_j] = \sigma^2 C_{jj},$$

其中 $C = (X^T X)^{-1}$ 。

Gauss-Markov 定理

如果基本假设成立，最小二乘估计量 $\hat{\beta}$ 是总体参数 β 的线性最小方差无偏估计量。

$\hat{\beta}$ 的标准误 (Standard Error)

- $\hat{\beta}$ 的标准误为:

$$\text{SE}[\hat{\beta}] = s_e \sqrt{(X^T X)^{-1}}$$

- 对每一个 $\hat{\beta}_j$,

$$\text{SE}[\hat{\beta}_j] = s_e \sqrt{C_{jj}}.$$

- β_j 与 s_e 独立

拟合优度

通过方差分解我们有： $SST = SSE + SSR$.

拟合优度 (Goodness of Fit), 又叫测定系数 (Coefficient of Determination) :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

拟合优度越大模型就越好吗? 拟合优度小模型就一定差吗?

R^2 是 p 的单调递增函数。

调整的拟合优度

$$R_{\text{adj}}^2 = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-p-1} \frac{\text{SSE}}{\text{SST}}$$

- R_{adj}^2 一般小于 R^2
- R_{adj}^2 的值远远小于 R^2 时，这表明所选择的模型中预测因子的数量相对于数据量过多，这就是一个警示信号。

调整的拟合优度

```
loan.model.summary$r.squared
```

```
## [1] 0.797604
```

```
loan.model.summary$adj.r.squared
```

```
## [1] 0.7571248
```

Outline

- 1 多元线性回归模型
- 2 模型估计及估计的性质
- 3 模型的统计推断**
- 4 模型诊断
- 5 变量选择
- 6 扩展
- 7 Key points & 上机实验

两部分的内容：

1 变量的显著性检验： x_j 是否对模型有效果？

$$H_0 : \beta_j = 0, j = 1, 2, \dots, p$$

2 模型的显著性检验： x_1, \dots, x_p 是否对模型预测有效果？

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0.$$

对 $\hat{\beta}$ 的推断

因为:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{jj}).$$

我们得到:

$$\frac{\hat{\beta}_j - \beta_j}{s_e \sqrt{C_{jj}}} \sim t(n - p - 1).$$

$\hat{\beta}$ 的区间估计

对每一个 $\hat{\beta}_j$, 其区间估计为:

$$\hat{\beta}_j \pm t_{1-\alpha/2}(n-p-1) \cdot s_e \sqrt{C_{jj}}$$

$\hat{\beta}$ 的区间估计: R 实现

```
confint(loan.model, level = 0.99)
```

```
##              0.5 %      99.5 %  
## (Intercept) -3.24775491 1.20447538  
## 各项贷款余额 0.01035187 0.06972683  
## 本年累计应收贷款 -0.07616275 0.37223054  
## 贷款项目个数 -0.22172819 0.25078690  
## 本年固定资产投资额 -0.07208059 0.01369486
```

- 给定未知的 \mathbf{x}_0 , 点预测为:

$$\begin{aligned}\hat{y}(\mathbf{x}_0) &= \mathbf{x}_0^\top \hat{\beta} \\ &= \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_{01} + \hat{\beta}_2 \mathbf{x}_{02} + \cdots + \hat{\beta}_p \mathbf{x}_{0p}.\end{aligned}$$

- 这是一个无偏估计:

$$\begin{aligned}E[\hat{y}(\mathbf{x}_0)] &= \mathbf{x}_0^\top \beta \\ &= \beta_0 + \beta_1 \mathbf{x}_{01} + \beta_2 \mathbf{x}_{02} + \cdots + \beta_p \mathbf{x}_{0p}\end{aligned}$$

■ 标准误为: $SE[\hat{y}(x_0)] = s_e \sqrt{x_0^\top (X^\top X)^{-1} x_0}$.

■ 置信区间:

$$\hat{y}(x_0) \pm t_{1-\alpha/2}(n-p-1) \cdot s_e \sqrt{x_0^\top (X^\top X)^{-1} x_0}.$$

置信区间

```
new_loan = data.frame(各项贷款余额 = c(100, 150, 200), 本年累计应收贷款 = c(7, 20, 13),  
                      贷款项目个数 = c(10, 6, 5), 本年固定资产投资额 = c(40, 76, 5))  
predict(loan.model, newdata = new_loan, interval = "confidence", level = 0.99)
```

```
##           fit           lwr           upr  
## 1 2.996112 1.631683 4.360541  
## 2 5.813459 2.089203 9.537715  
## 3 8.837354 3.896133 13.778574
```

- 计算 y 的预测区间，标准误为：

$$\sqrt{\text{SE}^2[\hat{y}(x_0)] + \text{SE}^2[\epsilon]} = s_e \sqrt{\mathbf{1} + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}.$$

- 预测区间为：

$$\hat{y}(x_0) \pm t_{1-\alpha/2}(n - p - 1) \cdot s_e \sqrt{\mathbf{1} + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}.$$

预测区间

```
predict(loan.model, newdata = new.loan, interval = "prediction", level = 0.99)
```

```
##           fit           lwr           upr  
## 1 2.996112 -2.2457345  8.237958  
## 2 5.813459 -0.4702792 12.097198  
## 3 8.837354  1.7640983 15.910609
```

β_j 的显著性检验

我们要检验：

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

统计量

$$t = \frac{\hat{\beta}_j - \beta_j}{\text{SE}[\hat{\beta}_j]} = \frac{\hat{\beta}_j - 0}{s_e \sqrt{C_{jj}}}$$

在零假设成立的情况下，服从 $t(n - p - 1)$ 分布。

β_j 的显著性检验

```
loan.model.summary$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-1.02163976	0.78237236	-1.3058229	0.206433969
## 各项贷款余额	0.04003935	0.01043372	3.8374953	0.001028464
## 本年累计应收贷款	0.14803389	0.07879433	1.8787378	0.074935421
## 贷款项目个数	0.01452935	0.08303316	0.1749825	0.862852686
## 本年固定资产投资额	-0.02919287	0.01507297	-1.9367689	0.067030078

回归模型的显著性检验

回顾：通过方差分解我们有： $SST = SSE + SSR$.

多元回归中，回归模型的显著性检验的零假设为：

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

也就是说零假设为：

$$Y_i = \beta_0 + \epsilon_i.$$

备择假设为：

$$H_1 : \text{至少存在一个 } \beta_j \neq 0, j = 1, 2, \cdots, p$$

回归模型的 F 检验

我们有方差分析表：

来源	平方和	自由度	均方	F
回归	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	p	SSR/p	MSR/MSE
误差	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - (p + 1)$	$SSE/(n - (p + 1))$	
总	$\sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$		

回归模型的 F 检验

F 统计量为:

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p - 1)} \sim F(p, n - p - 1), \text{ under } H_0$$

拒绝域为:

$$W = \{F \geq F_{1-\alpha}(p, n - p - 1)\}$$

p -值为:

$$p = P(F(p, n - p - 1) > F_0)$$

F 检验的另一个表达

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - (p + 1)}{p}$$

- F 正比于 $\frac{R^2}{1 - R^2}$, 即可解释部分与未解释部分的占比
- $\frac{n - (p + 1)}{p}$ 随着 p 的变大而变小, 当 $p > n - 1$ 时, F 检验不再适用, 需要发展新的模型显著性检验。

回归模型的 F 检验

```
loan.model.summary$fstatistic
```

```
##      value      numdf      dendf  
## 19.70404    4.00000   20.00000
```

Outline

- 1 多元线性回归模型
- 2 模型估计及估计的性质
- 3 模型的统计推断
- 4 模型诊断
- 5 变量选择
- 6 扩展
- 7 Key points & 上机实验

是否符合模型假设? LINE

- 线性关系 (linear function)
- 独立性 (independence)
- 正态性 (normal distributed)
- 同方差性 (equal variance)

标准化残差

- $e = y - \hat{y} = (I - X(X^T X)^{-1} X^T) y = (I - H) y$
- 帽子矩阵 (hat matrix) $H = X(X^T X)^{-1} X^T$
- $\text{Cov}(e) = \sigma^2 (I - H)$
- 标准化残差

$$e_i^* = \frac{e_i}{s_e \sqrt{1 - h_{ii}}}$$

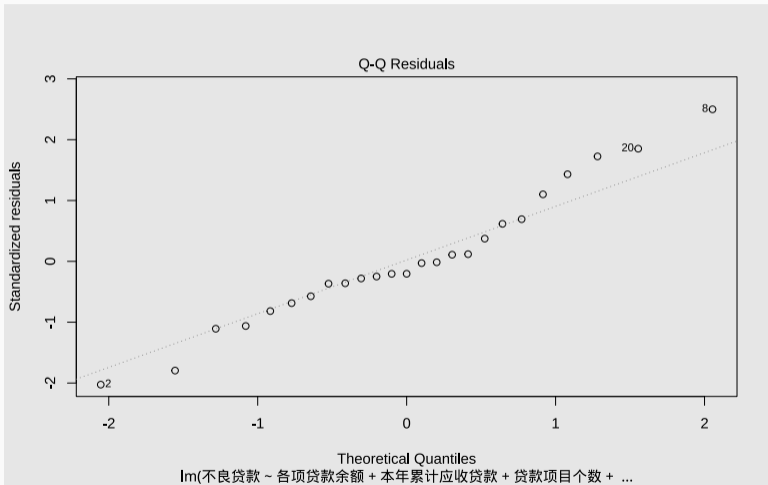
类似于简单一元线性回归的模型诊断：残差图、QQ图等

1 $(\hat{y}_i, e_i)_{i=1}^n$ 的散点图

- e_i 在 0 附近波动 (linear)
- 波动稳定, 且不随着 \hat{y}_i 的变化而变化 (equal variance)
- 没有大的残差点 ($|e_i| \geq 3$) (outlier)

2 $(x_{ij}, e_i)_{i=1}^n$ 的散点图

- e_i 在 0 附近波动 (linear)
- 波动稳定, 且不随着 x_{ij} 的变化而变化 (equal variance)



正态性检验（扩展）

- Anderson-Darling Test
- Shapiro-Wilk Test
- Kolmogorov-Smirnov Test

R 实现

- 1 `ad.test()` in 'nortest' package
- 2 `shapiro.test()`
- 3 `ks.test()`

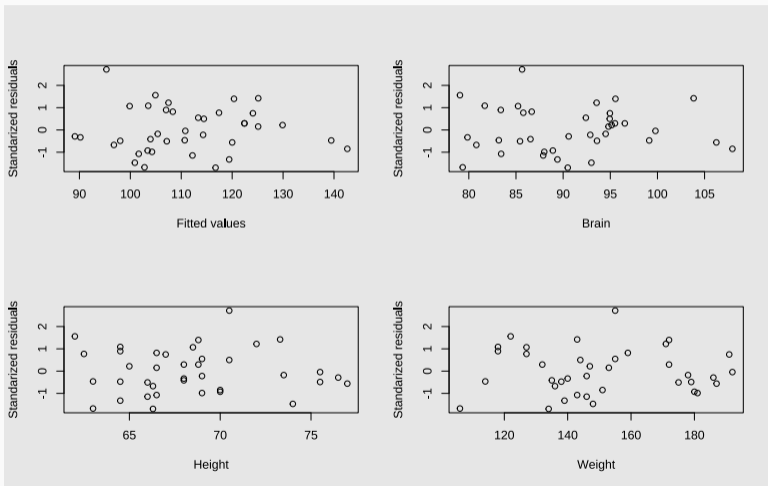
另外一个例子: PIQ 与身体指标

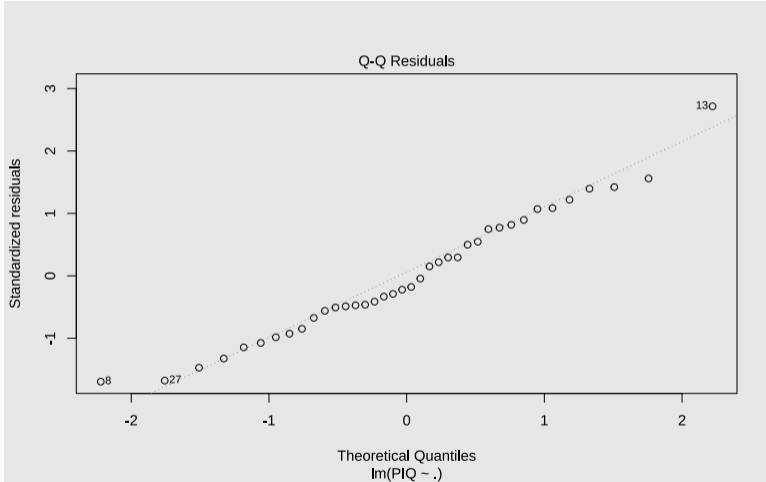
- iqsize 数据集: 分析 PIQ 与身体指标 (Brain、Weight、Height) 之间的关系

标准化残差图

```
# 标准化残差图
par(mfrow = c(2, 2))
plot(iq.lm$fitted, iq.res, xlab = 'Fitted values', ylab = 'Standarized residuals')
plot(iq$Brain, iq.res, xlab = 'Brain', ylab = 'Standarized residuals')
plot(iq$Height, iq.res, xlab = 'Height', ylab = 'Standarized residuals')
plot(iq$Weight, iq.res, xlab = 'Weight', ylab = 'Standarized residuals')
```

标准化残差图





正态性检验

```
##  
## Exact one-sample Kolmogorov-Smirnov test  
##  
## data: iq.res  
## D = 0.096875, p-value = 0.8344  
## alternative hypothesis: two-sided
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: iq.res  
## W = 0.97492, p-value = 0.5402
```

```
##  
## Anderson-Darling normality test  
##  
## data: iq.res  
## A = 0.26917, p-value = 0.6614
```

不正常点

- 除了检查模型假设之外，还应该注意“不正常点” (unusual observations)。
- 因为有时少量的不正常点对回归的影响是非常大的。

常见的不正常点：

- 1 异常点 (Outliers)
- 2 高杠杆点 (Points with high leverage)
- 3 强影响点 (Influential points)

异常点是没有被模型很好的拟合的点，通常是有很大的标准化残差 (standardized residual) 的观测值。

判断标准如：标准化残差的绝对值大于 3 或者大于 2。

R: 异常值检验

```
outlierTest(loan.model)
```

```
## No Studentized residuals with Bonferroni p < 0.05
```

```
## Largest |rstudent|:
```

```
##   rstudent unadjusted p-value Bonferroni p
```

```
## 8 2.937128          0.0084586          0.21146
```

高杠杆点，对预测点影响大的点。

杠杆值

h_{ii} 为样本 i 对预测值 \hat{y}_i 的影响 $h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$

- 特殊地，在一元线性回归中

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

- 杠杆值为帽子矩阵 H 的对角线元素，由自变量取值的来决定，也就是说杠杆点只度量了自变量对回归的影响
- $0 \leq h_{ii} \leq 1, \sum_{i=1}^n h_{ii} = p + 1$

多高的杠杆值应该引起注意呢？

高杠杆点

通常如果某个观测值的杠杆值大于二倍的平均杠杆值 $((p + 1)/n)$ ，被认为是高杠杆值。即

$$h_{ii} > \frac{2(p + 1)}{n}.$$

杠杆值

```
hatvalues(loan.model)
```

```
##           1           2           3           4           5           6           7
## 0.14699817 0.37700617 0.11985647 0.10804892 0.14506018 0.16892163 0.16380175
##           8           9          10          11          12          13          14
## 0.51247764 0.14807691 0.11234041 0.10369810 0.13773405 0.13179714 0.18982442
##          15          16          17          18          19          20          21
## 0.28661792 0.10235522 0.21113050 0.07722630 0.11290567 0.35554736 0.57824701
##          22          23          24          25
## 0.10095553 0.05756815 0.26663244 0.28517194
```

高杠杆点

```
hatvalues(loan.model) > 2 * mean(hatvalues(loan.model))
```

```
##      1      2      3      4      5      6      7      8      9     10     11     12     13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE FALSE
##     14     15     16     17     18     19     20     21     22     23     24     25
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  FALSE FALSE FALSE FALSE
```

有影响点

有影响点是对回归模型有显著影响的观测值，一般同时具有大残差和高杠杆值的特点。可以通过库克距离（Cook's Distance） D_i 来计算。一般当 $D_i > \frac{4}{n}$ 时，判断为有影响点。

库克距离（Cook's Distance）

$$D_i = \frac{e_i^{*2}}{p + 1} \cdot \frac{h_{ii}}{1 - h_{ii}},$$

其中 e_i^* 表示标准化残差。

事实上， D_i 是通过度量 $\hat{\beta}$ 与 $\hat{\beta}_{-(i)}$ （去掉第 i 个样本点得到的估计）之间的差距得到，即

$$D_i = \frac{\|\hat{y}_{-(i)} - \hat{y}\|_2^2}{(p+1)s_e^2} = \frac{(\hat{\beta}_{-(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{-(i)} - \hat{\beta})}{(p+1)s_e^2}$$

强影响点

```
cooks.distance(loan.model)[8] > 4/nrow(loan)
```

```
##      8
```

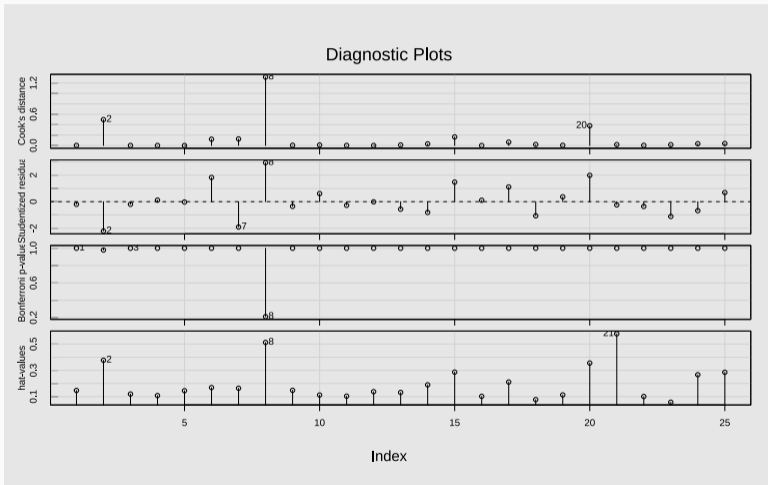
```
## TRUE
```

```
cooks.distance(loan.model)[21] > 4/nrow(loan)
```

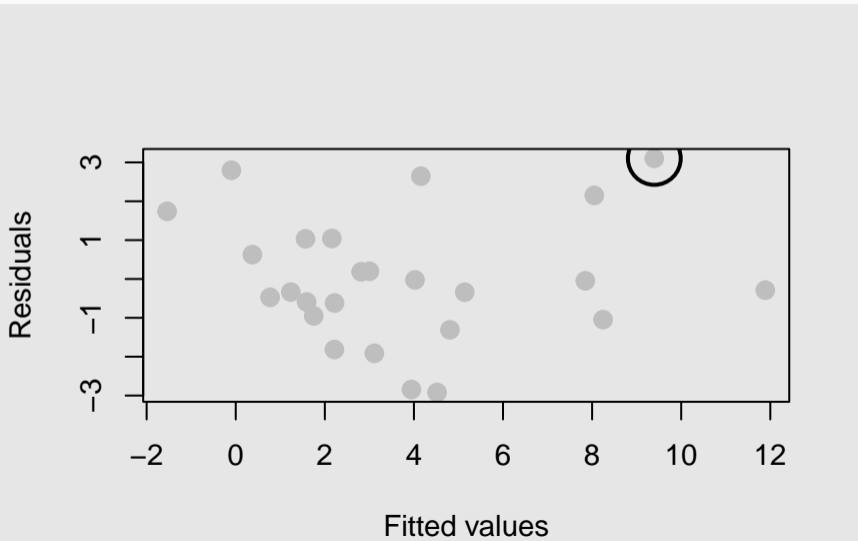
```
##     21
```

```
## FALSE
```

可视化：强影响点



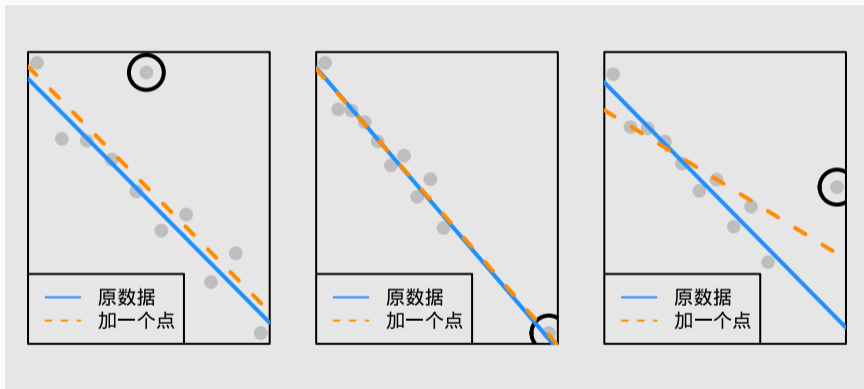
强影响点



去除有影响点?

```
##  
## Call:  
## lm(formula = 不良贷款 ~ 各项贷款余额 + 本年累计应收贷款 +  
##      贷款项目个数 + 本年固定资产投资额, data = loan.new)  
##  
## Residuals:  
##      Min      1Q   Median      3Q      Max  
## -2.09473 -1.24993 -0.09849  0.98472  2.77202  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)    -0.674348   0.676097  -0.997  0.331098  
## 各项贷款余额     0.039071   0.008884   4.398  0.000309 ***  
## 本年累计应收贷款 -0.014968   0.087032  -0.172  0.865271  
## 贷款项目个数     0.039881   0.071174   0.560  0.581803  
## 本年固定资产投资额 -0.017078   0.013472  -1.268  0.220223  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##
```

例子



更多内容，请参考

Chapter 8 of An R Companion to Applied Regression.

Outline

- 1 多元线性回归模型
- 2 模型估计及估计的性质
- 3 模型的统计推断
- 4 模型诊断
- 5 变量选择**
- 6 扩展
- 7 Key points & 上机实验

- 回归模型中重要的预测变量应该在模型中，实际情况往往不是如此
- 有大量候选预测变量，应该选择哪些预测变量在模型中？
- 变量选择：从较大的候选预测变量集中选择一个小的子集，以便得到的回归模型既简单又有用

预测变量越多越好吗？

- 随着预测变量个数的增多，尽管残差平方和 SSE 随之变小，但并非多多益善
- 导致过拟合 (overfitting):
 - 1 导致模型样本外的预测能力差
 - 2 模型复杂，可解释性差
 - 3 MSE 的自由度下降，导致置信区间变宽、假设检验的功效降低

模型中缺少某重要预测变量

- 模型估计有偏
- MSE 高估 (overestimate) σ^2 , 导致置信区间和假设检验失效
- 例子: $y=\text{weight}$, $x_1=\text{height}$, $x_2=\text{water}$ (0,10,20 cups per day)
 - 1 模型 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
 - 2 模型 2: $y = \beta_0 + \beta_1 x_1 + \varepsilon$

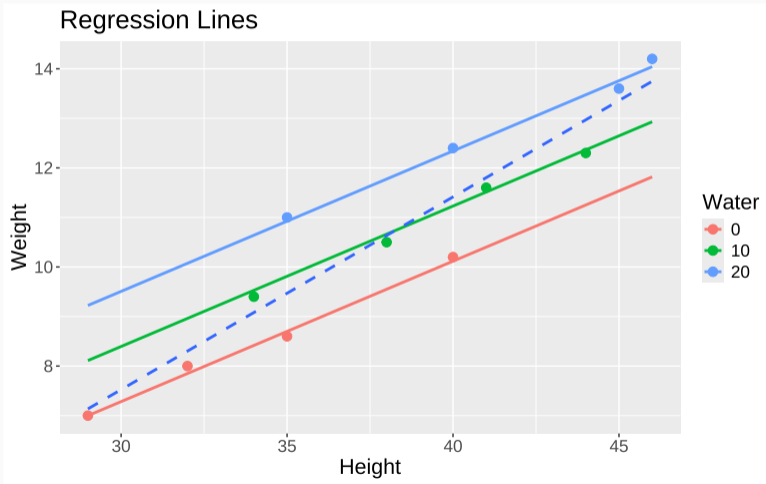
weight~height+water

```
##  
## Call:  
## lm(formula = weight ~ height + water, data = mdata)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -0.16247 -0.10722  0.02955  0.08388  0.15792   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept) -1.220194   0.320978  -3.801  0.00421 **     
## height      0.283436   0.009142  31.003 1.85e-10 ***   
## water       0.111212   0.005748  19.348 1.22e-08 ***   
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.1305 on 9 degrees of freedom  
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9966  
## F-statistic: 1592 on 2 and 9 DF.  p-value: 3.353e-12
```

weight~height

```
##  
## Call:  
## lm(formula = weight ~ height, data = mdata)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1.2140 -0.3943 -0.1359  0.3528  1.5307   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept) -4.14335    1.75340  -2.363  0.0397 *      
## height       0.38893    0.04543   8.561 6.48e-06 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.808 on 10 degrees of freedom  
## Multiple R-squared:  0.8799, Adjusted R-squared:  0.8679   
## F-statistic: 73.28 on 1 and 10 DF,  p-value: 6.475e-06
```

Visualization



变量选择

变量选择通常是一个迭代过程，在每一步中，以某种预先定好的标准来决定是否加入或剔除某个自变量。这个标准可以是：

- 假设检验，如 F 检验或 t 检验
- R_{adj}^2
- Akaike information criterion (AIC) 或 Bayesian information criterion (BIC)
- Mallows's C_p
- 其他标准

向后筛选法 (Backward Elimination)

在一开始假设模型中包含所有自变量，然后依据某种标准逐渐剔除不显著的变量，重复直到现存变量均不符合剔除条件。以 t 检验为例，向后筛选法为：

- 1 所有自变量 X_1, X_2, \dots, X_{p-1} 均包含在模型中；
 - ▶ 如果 t 检验都显著，则 X_1, X_2, \dots, X_{p-1} 均包含在模型中；
 - ▶ 若存在若干 t 检验不通过的参数，则先把 p -值最大的变量删除；
- 2 对剩余的 $p - 2$ 个变量做回归方程，删除 t 检验不通过中 p -值最大的变量；
- 3 重复以上步骤。直到模型中所有变量均通过 t 检验。

向后筛选法

```
step(loan.model, direction = "backward")
```

```
## Start: AIC=33.22
```

```
## 不良贷款 ~ 各项贷款余额 + 本年累计应收贷款 +  
## 贷款项目个数 + 本年固定资产投资额
```

```
##
```

##		Df	Sum of Sq	RSS	AIC
##	- 贷款项目个数	1	0.097	63.376	31.255
##	<none>			63.279	33.217
##	- 本年累计应收贷款	1	11.168	74.447	35.280
##	- 本年固定资产投资额	1	11.868	75.147	35.514
##	- 各项贷款余额	1	46.594	109.873	45.011

```
##
```

```
## Step: AIC=31.26
```

```
## 不良贷款 ~ 各项贷款余额 + 本年累计应收贷款 +  
## 本年固定资产投资额
```

```
##
```

##		Df	Sum of Sq	RSS	AIC
##	<none>			63.376	31.255

向前选择法 (Forward Selection)

在一开始假设模型中没有变量，计算 Y 和每一个 X_i 的一元线性回归模型，选择 p -值最小的变量，然后依据某种标准逐渐加入一个变量，重复直到剩下的变量均不符合加入条件。

向前选择法

```
step(lm(不良贷款 ~ 1, data = loan),  
      scope = 不良贷款 ~ 各项贷款余额 + 本年累计应收贷款 +  
              贷款项目个数 + 本年固定资产投资额,  
      direction = "forward")
```

```
## Start:  AIC=65.16  
## 不良贷款 ~ 1  
##  
##  
##           Df Sum of Sq    RSS    AIC  
## + 各项贷款余额      1    222.49  90.164 36.069  
## + 本年累计应收贷款  1    167.30 145.351 48.007  
## + 贷款项目个数      1    153.32 159.328 50.302  
## + 本年固定资产投资额 1     84.06 228.591 59.326  
## <none>                    312.650 65.155  
##  
## Step:  AIC=36.07  
## 不良贷款 ~ 各项贷款余额  
##  
##           Df Sum of Sq    BSS    AIC
```

逐步回归法 (Stepwise Regression)

- 前进法的问题：一旦某自变量进入模型后，它就永远留在模型中。然而，随着其他自变量的引入，一些先进入模型的变量的作用会变得不再显著。
- 向后法的问题：一旦某自变量被删除后，就永远不再进入模型。然而，随着其他自变量被删除，它的作用有可能会显著起来。

逐步回归法 (Stepwise Regression)

- 对于模型外部的变量，只要还能提供显著的解释作用，则可以再次进入模型。而在模型内部的变量，只要它的 t 检验不再显著，则可以从模型中删除。
- 方法：边进边退
- 起始：同前进法
- 结束：模型外所有变量均不能通过 t 检验

逐步回归法

```
step(lm(不良贷款 ~ 1, data = loan),  
      scope = 不良贷款 ~ 各项贷款余额 + 本年累计应收贷款 +  
              贷款项目个数 + 本年固定资产投资额,  
      direction = "both")
```

```
## Start:  AIC=65.16  
## 不良贷款 ~ 1  
##  
##  
##           Df Sum of Sq    RSS    AIC  
## + 各项贷款余额      1    222.49  90.164 36.069  
## + 本年累计应收贷款  1    167.30 145.351 48.007  
## + 贷款项目个数      1    153.32 159.328 50.302  
## + 本年固定资产投资额 1     84.06 228.591 59.326  
## <none>                                312.650 65.155  
##  
## Step:  AIC=36.07  
## 不良贷款 ~ 各项贷款余额  
##  
##           Df Sum of Sq    RSS    AIC
```

Outline

- 1 多元线性回归模型
- 2 模型估计及估计的性质
- 3 模型的统计推断
- 4 模型诊断
- 5 变量选择
- 6 扩展
- 7 Key points & 上机实验

1 非线性回归

- 有些时候自变量和因变量之间的关系是非线性的，可以通过从原始解释变量中创建新的预测变量，比如二次项或者交互作用项，改善多元回归模型的拟合效果
- 多项式回归模型、Box-Cox 变换后再进行线性回归等
- 若无法确定非线性回归关系，还可以采用半参数或者非参数回归方法（后续课程）

2 带类别变量的回归

- 引入哑变量编码将分类预测变量纳入到多元回归模型中

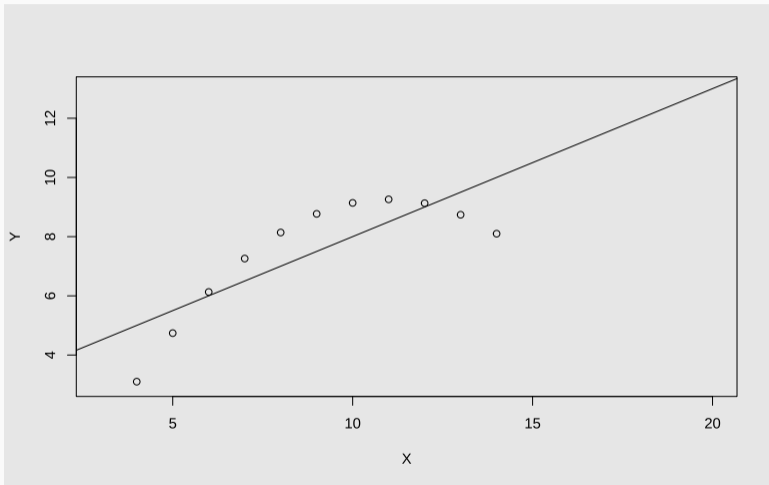
多项式回归

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_h X^h + \epsilon$$

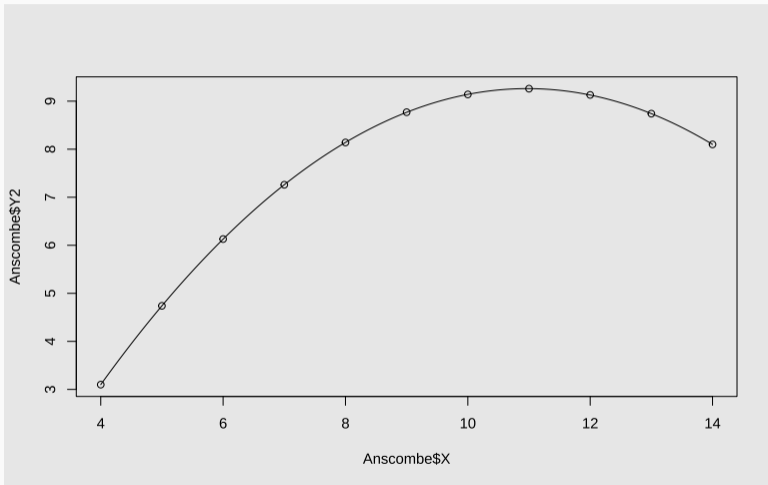
h 为多项式的阶数.

- $h = 2$: 二次多项式
- $h = 3$: 三次多项式
- $h = 4$: 四次多项式
- 不推荐使用高阶多项式回归

一个例子：线性回归模型拟合不好



模型更新：添加 x^2

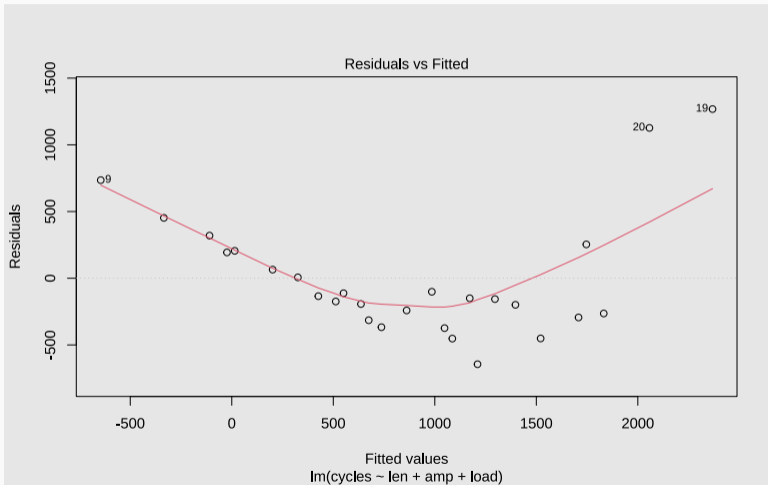


- 选用公开数据集 Wool (需要加载 car 程序包): 对含有不同水平的三个因素的精梳纱样品施加循环载荷, 直到样品失效。研究目标是了解失效循环数如何取决于这些因素。

线性回归模型

```
##           (Intercept)  len  amp  load
## Estimate             4521 13.2 -536 -62.2
## Std. Error           1622  2.3  115  23.0
##
## Residual SD = 488 on 23 df, R-squared = 0.729
```

残差图

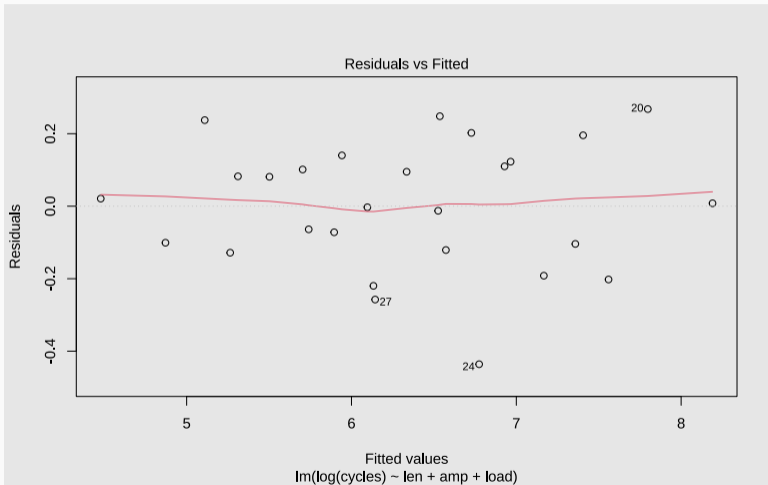


对因变量做 log 变换

```
wool.mod.log <- lm(log(cycles) ~ len + amp + load, data=Wool)
```

```
##           (Intercept)      len      amp      load
## Estimate           10.552 0.016648 -0.6309 -0.07852
## Std. Error           0.617 0.000875  0.0438  0.00875
##
## Residual SD = 0.186 on 23 df, R-squared = 0.966
```

变换后的残差图



带类别变量的回归

- birthsmokers 数据集：研究新生儿出生体重 (Weight) 和怀孕期间是否吸烟 (Smoking)、妊娠持续时间 (Gest) 的周数之间的关系

$$Weight = \beta_0 + \beta_1 Smooking + \beta_2 Gest + \varepsilon$$

- 二类别可以用一个哑变量进行编码：Smooking=1, 若怀孕期间吸烟；否则 Smooking=0

- 若 Smooking=1, 回归方程为

$$\widehat{Weight} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 Gest$$

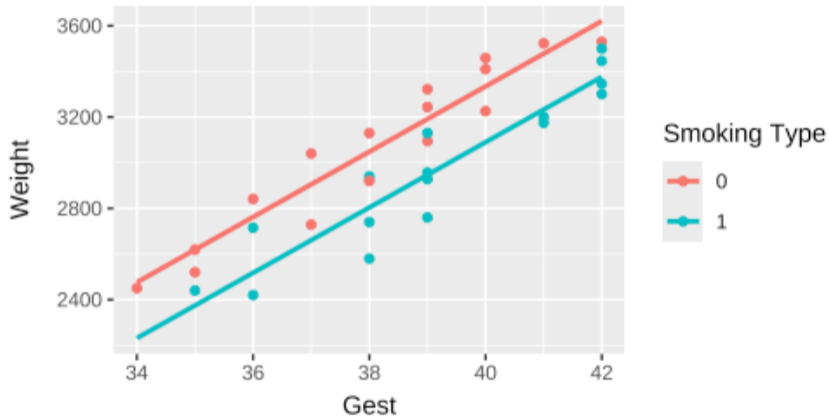
- 若 Smooking=0, 回归方程为

$$\widehat{Weight} = \hat{\beta}_0 + \hat{\beta}_2 Gest$$

回归方程

```
##  
## Call:  
## lm(formula = Wgt ~ ., data = birthsmokers)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -223.693  -92.063   -9.365   79.663  197.507  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2389.573    349.206  -6.843 1.63e-07 ***  
## Gest         143.100      9.128  15.677 1.07e-15 ***  
## Smoke       -244.544     41.982  -5.825 2.58e-06 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 115.5 on 29 degrees of freedom  
## Multiple R-squared:  0.8964, Adjusted R-squared:  0.8892  
## F-statistic: 125.4 on 2 and 29 DF.  p-value: 5.289e-15
```

Regression Lines



更多内容，请参考

- 1 Modern Mathematical Statistics with Applications: 12.8
- 2 Lesson 8: Categorical Predictors
- 3 Lesson 9: Data Transformations

Outline

- 1 多元线性回归模型
- 2 模型估计及估计的性质
- 3 模型的统计推断
- 4 模型诊断
- 5 变量选择
- 6 扩展
- 7 Key points & 上机实验

Key points

- 熟练掌握多元线性回归模型的建立（四个假设）、估计（最小二乘估计）、估计性质、统计推断（模型的显著性检验和变量的显著性检验）、模型诊断、变量选择（向前、向后、逐步回归法）等
- 熟练掌握利用 R 软件进行多元线性回归分析的全过程
- 实际数据分析中，线性回归模型的一般步骤：模型建议、模型估计、模型推断、模型诊断、模型更新（可选）、模型预测与应用

上机实验：hospital_infct 数据集分析

美国 113 家医院的数据，评估与住院患者在医院内感染风险相关的因素

- y : 感染风险
- x_1 : 患者平均住院天数
- x_2 : 患者平均年龄
- x_4 : 医院进行 X-光检查的数量

要求完成：

- 1 计算和分析各变量之间的相关关系，并可视化展示
- 2 建立回归模型，并完成多元线性回归的建模、推断、诊断及预测过程。