多元线性回归分析

原俊青，浙江工业大学

http://homepage.zjut.edu.cn/yjq/

I. 多元线性回归模型

设变量Y与X_1，X_2， ……，X_p之间有线性关系

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

其中ε均值为0，方差为σ^2 ，$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$和$\sigma^2$ 是未知参数。

当$p \geq 2$，称以上公式为 **多元线性回归模型。**

- n次独立观测值记为 $(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i), \quad i = 1, 2 \cdots, n.$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- 矩阵形式

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon}\text{均值为0,等方差,不相关(协方差矩阵为}\boldsymbol{\sigma}^2\boldsymbol{I_n}\text{)}$$

$$\boldsymbol{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \quad \boldsymbol{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

误差的平方和

$$\boldsymbol{Q}(\boldsymbol{\beta})=\sum_{i=1}^{n}\varepsilon_i^2=\boldsymbol{\varepsilon}^{\boldsymbol{T}}\boldsymbol{\varepsilon}=(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$$

$$=\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y}-\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta}$$

- 矩阵的求导公式

列向量$\alpha_{n\times1}, x_{n\times1}$ 和矩阵 $A_{n\times n}$， 标量对向量的求导:

$$\frac{\partial\alpha^T x}{\partial x}=\alpha,\qquad \frac{\partial x^T\alpha}{\partial x}=\alpha,\qquad \frac{\partial x^T A x}{\partial x}=(A^T+A)x.$$

2. 最小二乘估计

$$\begin{aligned}\boldsymbol{Q}(\boldsymbol{\beta})&=\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y}-\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta}\\ \frac{\partial\boldsymbol{Q}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}&=-(\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{X})^{\boldsymbol{T}}-\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}+\left((\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{\boldsymbol{T}}+\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)\boldsymbol{\beta}\\ &=-\boldsymbol{2}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}+\boldsymbol{2}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta}=0\end{aligned}$$

如果$(\boldsymbol{X}^T\boldsymbol{X})_{(p+1)\times(p+1)}$是列满秩的，那么

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

- 估计量的性质

$$\begin{aligned}
\boldsymbol{E}(\hat{\boldsymbol{\beta}}) &= \left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\boldsymbol{E}(\boldsymbol{Y}) = \left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\beta} \\
\boldsymbol{Cov}(\hat{\boldsymbol{\beta}}) &= \left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\boldsymbol{Cov}(\boldsymbol{Y})\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)^{\boldsymbol{T}} \\
&= \left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)(\sigma^2\boldsymbol{I}_n)\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)^{\boldsymbol{T}} \\
&= \sigma^2\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\left(\boldsymbol{X}^{T^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1^T}\right) \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}$$

- 经验回归方程 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$
- 回代分析 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y}) = \boldsymbol{H}\boldsymbol{Y}.$

  - $\boldsymbol{H}$是**对称幂等矩阵**，$\boldsymbol{H}^T = \boldsymbol{X}^{T^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1^T}\boldsymbol{X}^T = \boldsymbol{H}$。

$$\boldsymbol{H}^2 = \left(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\left(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right) = \left(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right) = \boldsymbol{H}.$$

  - $(\boldsymbol{I}_n - \boldsymbol{H})$也是**对称幂等矩阵**

$$(\boldsymbol{I}_n - \boldsymbol{H})^T = \boldsymbol{I}_n^T - \boldsymbol{H}^T = \boldsymbol{I}_n - \boldsymbol{H}.$$
$$(\boldsymbol{I}_n - \boldsymbol{H})^2 = \boldsymbol{I}_n - \boldsymbol{I}_n\boldsymbol{H} - \boldsymbol{H}\boldsymbol{I}_n + \boldsymbol{H}^2 = \boldsymbol{I}_n - \boldsymbol{H}.$$
$$(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{I}_n\boldsymbol{X} - \left(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}.$$

3. 残差向量

$$
\begin{aligned}
\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} &= \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{Y} \\
&= (\boldsymbol{I_n} - \boldsymbol{H})(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = (\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{X\beta} + (\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{\varepsilon} \\
&= (\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{\varepsilon}.
\end{aligned}
$$

- 残差向量的性质 $E(\hat{\boldsymbol{\varepsilon}}) = (\boldsymbol{I_n} - \boldsymbol{H})E(\boldsymbol{\varepsilon}) = \mathbf{0}$.

$$
\begin{aligned}
Cov(\hat{\boldsymbol{\varepsilon}}) &= E(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^{\boldsymbol{T}}) = E((\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{T}}(\boldsymbol{I_n} - \boldsymbol{H})^{\boldsymbol{T}}) \\
&= (\boldsymbol{I_n} - \boldsymbol{H})E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{T}})(\boldsymbol{I_n} - \boldsymbol{H})^{\boldsymbol{T}} \\
&= (\boldsymbol{I_n} - \boldsymbol{H})(\sigma^2\boldsymbol{I_n})(\boldsymbol{I_n} - \boldsymbol{H}) \\
&= \sigma^2(\boldsymbol{I_n} - \boldsymbol{H})^{\mathbf{2}} = \sigma^2(\boldsymbol{I_n} - \boldsymbol{H}).
\end{aligned}
$$

- 矩阵的**迹**的性质

-对于标量b, 和方阵$A_{n\times n}$及其特征值$\lambda_1, \cdots, \lambda_n$,

$$\text{迹} \quad tr(A) = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i$$

$$tr(A) = tr(A^T), \qquad tr(b) = b.$$

-对于矩阵$B_{n\times m}$ 和 $C_{m\times n}$,

$$tr(BC) = tr(CB).$$

$$tr(ABC) = tr(CAB) = tr(BCA).$$

-对于标量w_1, w_2 和n阶方阵A_1, A_2,

$$tr(w_1 A_1 + w_2 A_2) = w_1 tr(A_1) + w_2 tr(A_2).$$

-对于列满秩矩阵$B_{m\times q}$, $(B^T B)$ 是可逆的，且

$$tr\left((B^T B)^{-1}(B^T B)\right) = q = tr\left(B(B^T B)^{-1} B^T\right).$$

-对于行满秩矩阵$C_{k\times n}$, (CC^T) 是可逆的，且

$$tr\left((CC^T)^{-1}(CC^T)\right) = k = tr\left(C^T (CC^T)^{-1} C\right).$$

- 残差平方和(SSE)

$$
\begin{aligned}
\boldsymbol{Q}(\hat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}^{\boldsymbol{T}} \hat{\boldsymbol{\varepsilon}} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\boldsymbol{T}} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\
&= ((\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{\varepsilon})^{\boldsymbol{T}} ((\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^{\boldsymbol{T}}(\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{\varepsilon} \\
&= tr\left(\boldsymbol{\varepsilon}^{\boldsymbol{T}}(\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{\varepsilon}\right) = tr\left((\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{T}}\right) \\
E\left(\boldsymbol{Q}(\hat{\boldsymbol{\beta}})\right) &= tr\left((\boldsymbol{I_n} - \boldsymbol{H})E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{T}})\right) = tr\left((\boldsymbol{I_n} - \boldsymbol{H})Cov(\boldsymbol{\varepsilon})\right) \\
&= tr\left((\boldsymbol{I_n} - \boldsymbol{H})\sigma^2 \boldsymbol{I_n}\right) = \sigma^2 tr(\boldsymbol{I_n} - \boldsymbol{H}) \\
&= \sigma^2 \left(tr(\boldsymbol{I_n}) - tr(\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}})\right) \\
&= \sigma^2 \left(n - tr(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1})\right) = \sigma^2 \left(n - (p+1)\right)
\end{aligned}
$$

- 误差项的方差估计

$$\hat{\sigma}^2 = \frac{SSE}{n-(p+1)} = \frac{\hat{\boldsymbol{\varepsilon}}^{\boldsymbol{T}}\hat{\boldsymbol{\varepsilon}}}{n-(p+1)} = \frac{Y^T(I_n - H)Y}{n-(p+1)}.$$

- 估计量的性质

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}, \quad E(\hat{\sigma}^2) = \sigma^2, \qquad \text{无偏估计}$$

$$Cov(\hat{\boldsymbol{\beta}}) = E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\boldsymbol{T}}) = \sigma^2(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}$$

$$Cov(\hat{\boldsymbol{\varepsilon}}) = E(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^{\boldsymbol{T}}) = \sigma^2(\boldsymbol{I_n} - \boldsymbol{H}).$$

4. 多元正态线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I_n})$$

- 估计量的性质

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}(\boldsymbol{X^T Y}) \sim N(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X^T X})^{-1});$$

$$\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{Y} \sim N(0, \sigma^2 (\boldsymbol{I_n} - \boldsymbol{H}))$$

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

$\hat{\boldsymbol{\beta}}$ 和 $\hat{\sigma}^2$ 相互独立;

5. 单个回归系数的显著性检验（*t*检验）

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

- 检验统计量 $T = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{C_{jj}}} \overset{H_0\text{真}}{\sim} t(n-p-1)$,

  其中 $\hat{\sigma} = \sqrt{\frac{SSE}{n-p-1}}, \quad C = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

- 拒绝域 $|T| > t_{\alpha/2}(n-p-1)$

6. 回归方程的显著性检验（F检验）

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0 \text{ vs. } H_1 : \text{至少有一个系数不为0.}$$

- 检验统计量 $F = \frac{SSR/p}{SSE/(n-p-1)} \overset{H_0\text{真}}{\sim} F(p, n-p-1)$
- 拒绝域 $F > F_\alpha(p, (n-p-1))$
- 方差分析表 $SST = SSR + SSE$

$$SST = \sum(y_i - \bar{y})^2, \text{ 反映了数据 } y_i's \text{ 波动性的大小；}$$
$$SSR = \sum(\hat{y}_i - \bar{y})^2, \text{ 反映了由于 } X_1, \cdots, X_p \text{ 的变化引起的数据 } y_i's \text{ 的波动；}$$
$$SSE = \sum(y_i - \hat{y}_i)^2, \text{ 反映了除去 } X_i's \text{ 的影响外，其他因素引起的 } y_i's \text{ 的波动}$$

7. 拟合程度（R^2）

- 决定系数 $R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$

- 修正的决定系数 $R^{2*}=\frac{SSR/p}{SST/(n-1)}$

8. 预测及统计推断

- 点估计，新的观测值$\mathbf{x}=(1,x_1,x_2,\cdots,x_p)$,

$$\hat{y}=\mathbf{x}\hat{\boldsymbol{\beta}}\sim N(\mathbf{x}\boldsymbol{\beta},\mathbf{x}Cov(\hat{\boldsymbol{\beta}})\mathbf{x}^T)=N(\mathbf{x}\boldsymbol{\beta},\sigma^2\mathbf{x}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\mathbf{x}^T).$$

- 区间估计，$\hat{y}\pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{\mathbf{x}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\mathbf{x}^T}$

例：使用datarium包中的marketing数据集，我们将建立一个多元回归模型。根据在三种广告媒体（youtube，facebook和newspaper）上投入的预算来预测sales。计算公式如下：

$$sales = b_0 + b_1 * youtube + b_2 * facebook + b_3 * newspaper.$$

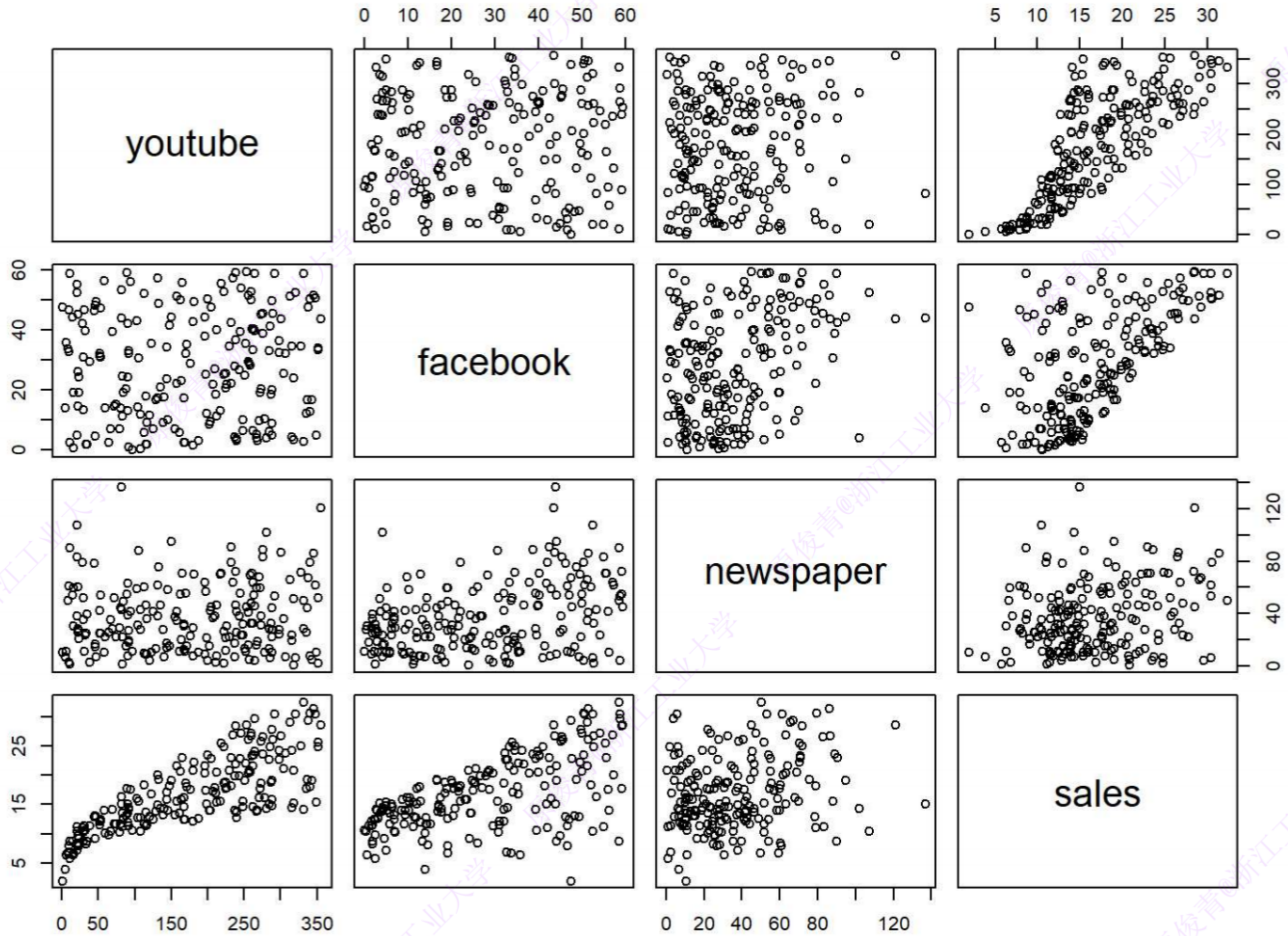
```r
data("marketing", package = "datarium")
dim(marketing)
```

```
## [1] 200   4
```

```r
head(marketing)
```

```
##   youtube facebook newspaper sales
## 1  276.12    45.36     83.04 26.52
## 2   53.40    47.16     54.12 12.48
## 3   20.64    55.08     83.16 11.16
## 4  181.80    49.56     70.20 22.20
## 5  216.96    12.96     70.08 15.48
## 6   10.44    58.68     90.00  8.64
```

```
plot(marketing)
```



- 检查二变量相关关系

```
cor(marketing)
```

```
##              youtube   facebook  newspaper     sales
## youtube   1.00000000 0.05480866 0.05664787 0.7822244
## facebook  0.05480866 1.00000000 0.35410375 0.5762226
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## sales     0.78222442 0.57622257 0.22829903 1.0000000
```

- 将数据分为训练组和测试组，按 8:2 的比例

```
set.seed(2022) # 设置随机种子使结果可重复
idx = sample(nrow(marketing), 0.8 * nrow(marketing))
trainData  <- marketing[idx, ]
testData   <- marketing[-idx, ]
```

```r
model_1 <- lm(sales ~ youtube + facebook + newspaper, data = trainData)
summary(model_1) #summary(model_1)$coef
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = trainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1666  -1.1738   0.3923   1.5835   3.4380
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.369249   0.437066   7.709 1.39e-12 ***
## youtube     0.046567   0.001595  29.203  < 2e-16 ***
## facebook    0.182270   0.010363  17.588  < 2e-16 ***
## newspaper   0.001606   0.007048   0.228     0.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.137 on 156 degrees of freedom
## Multiple R-squared:  0.8902, Adjusted R-squared:  0.8881
## F-statistic: 421.7 on 3 and 156 DF,  p-value: < 2.2e-16
```

```r
model_2 <- lm(sales ~ youtube + facebook, data = trainData)
summary(model_2) #summary(model_1)$coef
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook, data = trainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2223  -1.2030   0.3522   1.5695   3.4318
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.400910   0.413150   8.232 6.69e-14 ***
## youtube     0.046591   0.001586  29.370  < 2e-16 ***
## facebook    0.183128   0.009627  19.022  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.13 on 157 degrees of freedom
## Multiple R-squared:  0.8902, Adjusted R-squared:  0.8888
## F-statistic: 636.3 on 2 and 157 DF,  p-value: < 2.2e-16
```

- 回归模型 $sales = 3.4 + 0.047 * youtube + 0.183 * facebook$

- 预测, 点估计和区间估计

```
newdata <- data.frame(youtube = 2000, facebook = 1000, newspaper = 1000) # New advertising budgets
predict(model_2, newdata, interval="prediction", level=0.95) # Predict sales values
```

```
##        fit      lwr      upr
## 1 279.7107 260.1261 299.2953
```

```
test<-data.frame(trainData, predict(model_2, trainData))
y_true<-as.vector(test[,4])
SST <- crossprod(y_true-mean(y_true)); SST
```

```
##          [,1]
## [1,] 6488.245
```

```r
y_hat<-as.vector(test[,5]); SSR <- crossprod(y_hat - mean(y_true)); SSR
```

```
##          [,1]
## [1,] 5775.737
```

```r
err<- y_true - y_hat; SSE <- t(err) %*% err; SSE
```

```
##          [,1]
## [1,] 712.5088
```

```r
sig_hat <- sqrt(SSE/(nrow(test)-length(model_2$coef)-1)); sig_hat
```

```
##          [,1]
## [1,] 2.137139
```

```r
R_sq <- SSR/SST; R_sq
```

```
##           [,1]
## [1,] 0.8901847
```

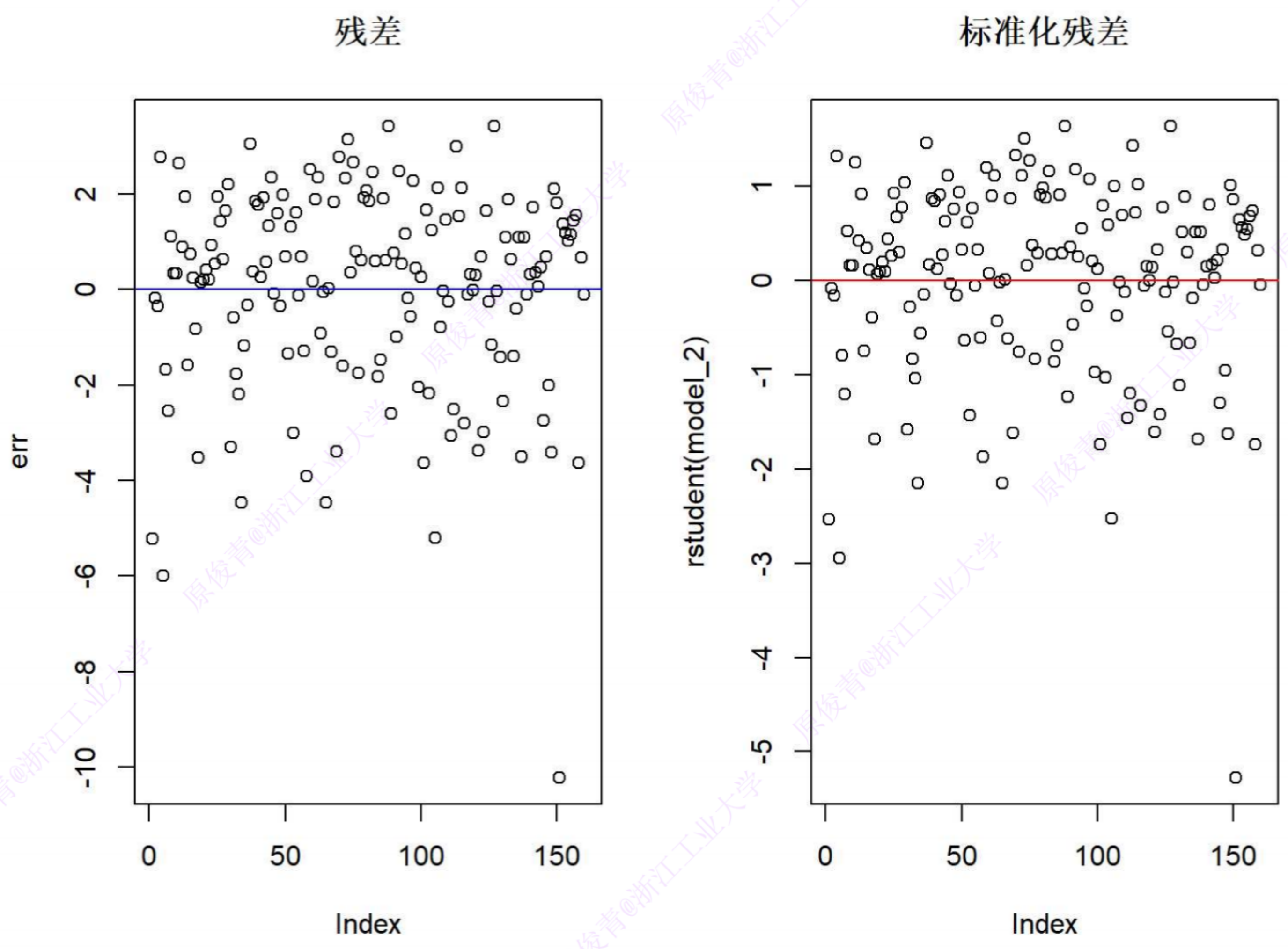
9. 回归模型诊断

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I_n})$$

- 借助残差探测样本中的异常值, 等方差

-①. 绘制残差的**散点图**, 检验等方差, 异常值;

```
opar=par(mfrow=c(1,2))
plot(err, main='残差'); abline(h=mean(err), col='blue')
plot(rstudent(model_2), main='标准化残差'); abline(h=0, col='red')
```

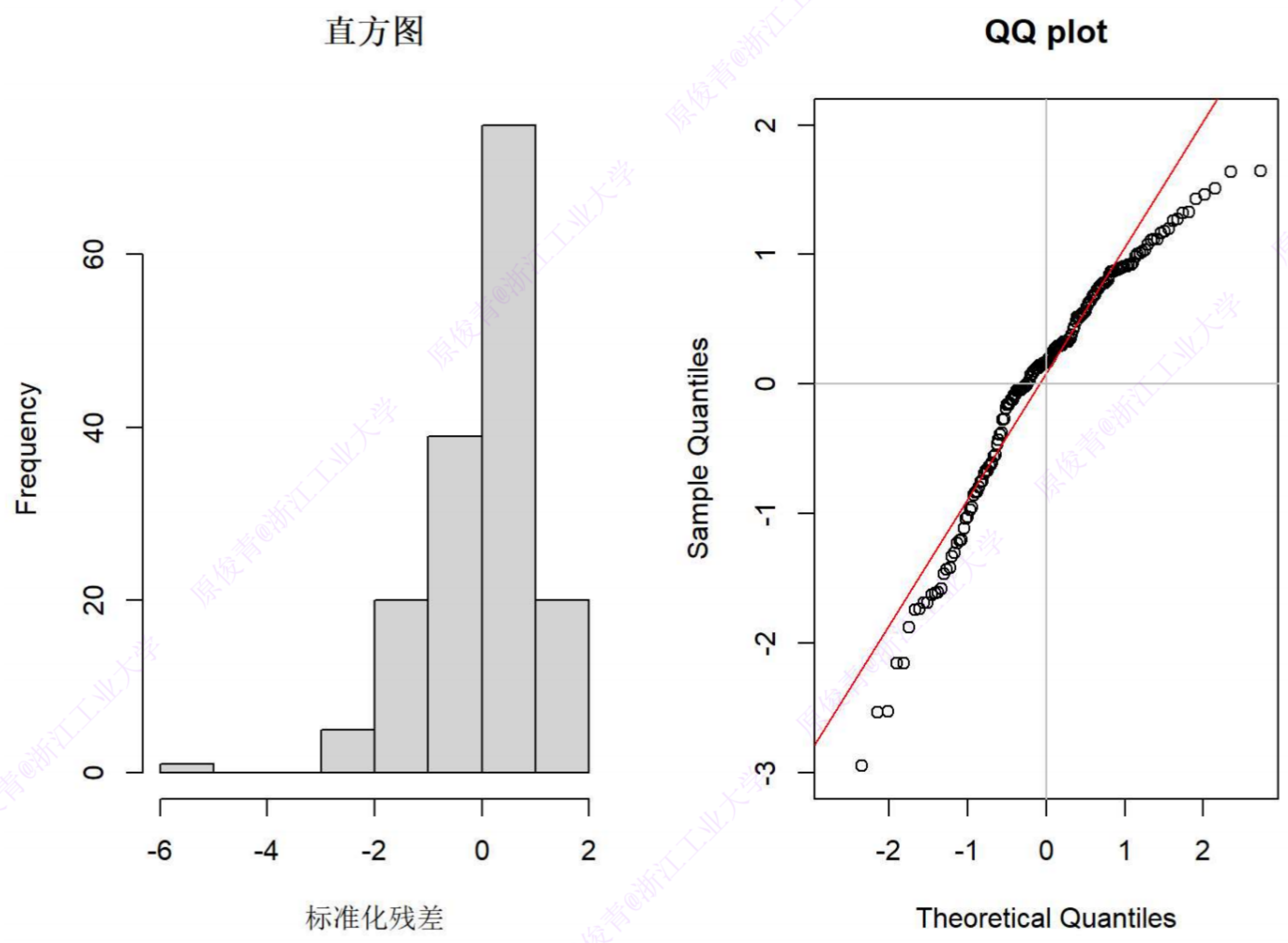



```
par(opar)
```

- 残差是否服从均值为零,等方差的正态分布

-①. 绘制残差的**直方图, QQ图**, 看分布, 检验正态性;

```
opar=par(mfrow=c(1,2))
hist(rstudent(model_2), xlab='标准化残差',main='直方图')
qqnorm(rstudent(model_2),ylim=c(-3,2),main='QQ plot')
qqline(rstudent(model_2),col='red')
abline(h=0,pch=3,col='gray'); abline(v=0,pch=3,col='gray')
```


直方图
Frequency
0
20
40
60
-6
-4
-2
0
2
标准化残差
QQ plot
Sample Quantiles
-3
-2
-1
0
1
2
Theoretical Quantiles

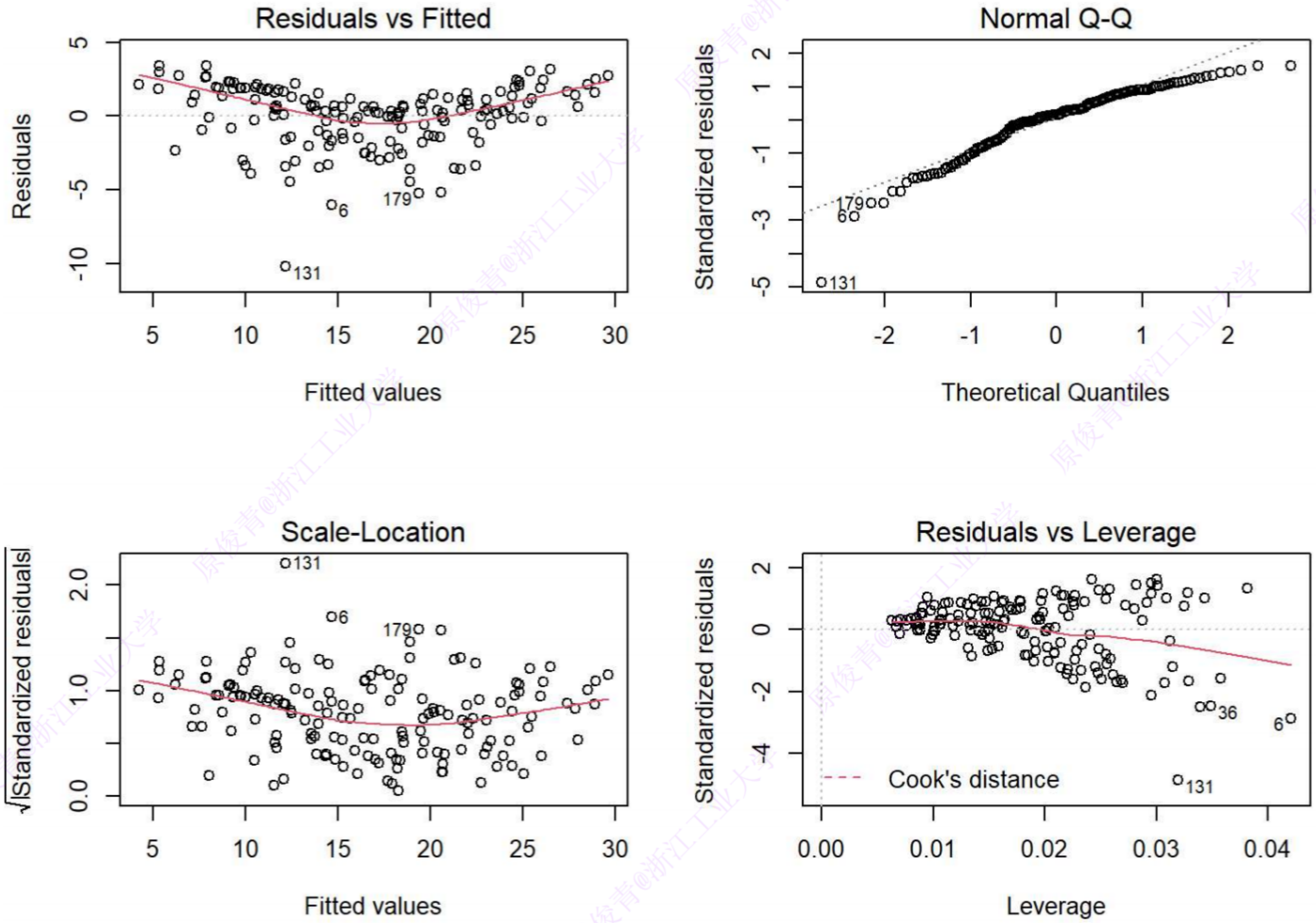
```
par(opar)
```

-基础绘图系统的plot函数输入对象为回归模型时，返回对象是四幅模型诊断图。

-四幅图依次为残差-拟合图、正态Q-Q图、尺度-位置图、残差-杠杆图；

-除残差-杠杆图用于检验异常点外，其余三幅图均用于检验模型结果是否符合假设。

```
opar=par(mfrow=c(2,2))
plot(model_2)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

```
par(opar)
```

$$\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{Y} \sim N(0, \sigma^2(\boldsymbol{I_n} - \boldsymbol{H}))$$

-②. 零均值t检验，$H_0 : E(\hat{\boldsymbol{\varepsilon}}) = 0$;

```
t.test(err)        ##H_0: E(err)=0
```

```
##
##  One Sample t-test
##
## data:  err
## t = -5.4399e-14, df = 159, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.3305239  0.3305239
## sample estimates:
##     mean of x
## -9.103829e-15
```

-③. Shapiro-Wilk正态性检验， H_0：正态分布；

检验统计量：$W=\frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$

其中，$(a_1,a_2,\cdots,a_n)=\frac{M^TV^{-1}}{C}$，　$M=(m_1,m_2,\cdots,m_n)^T$，　$m_i=E(Y_{(i)})$，

$$V=Cov(\mathbf{Y},\mathbf{Y}),\quad \mathbf{Y}=(Y_1,Y_2,\cdots,Y_n)^T,\quad Y_i\sim N(0,1).$$

$$C=||V^{-1}M||=(M^TV^{-1}V^{-1}M)^{1/2}.$$

H_0 为真时， W接近1。

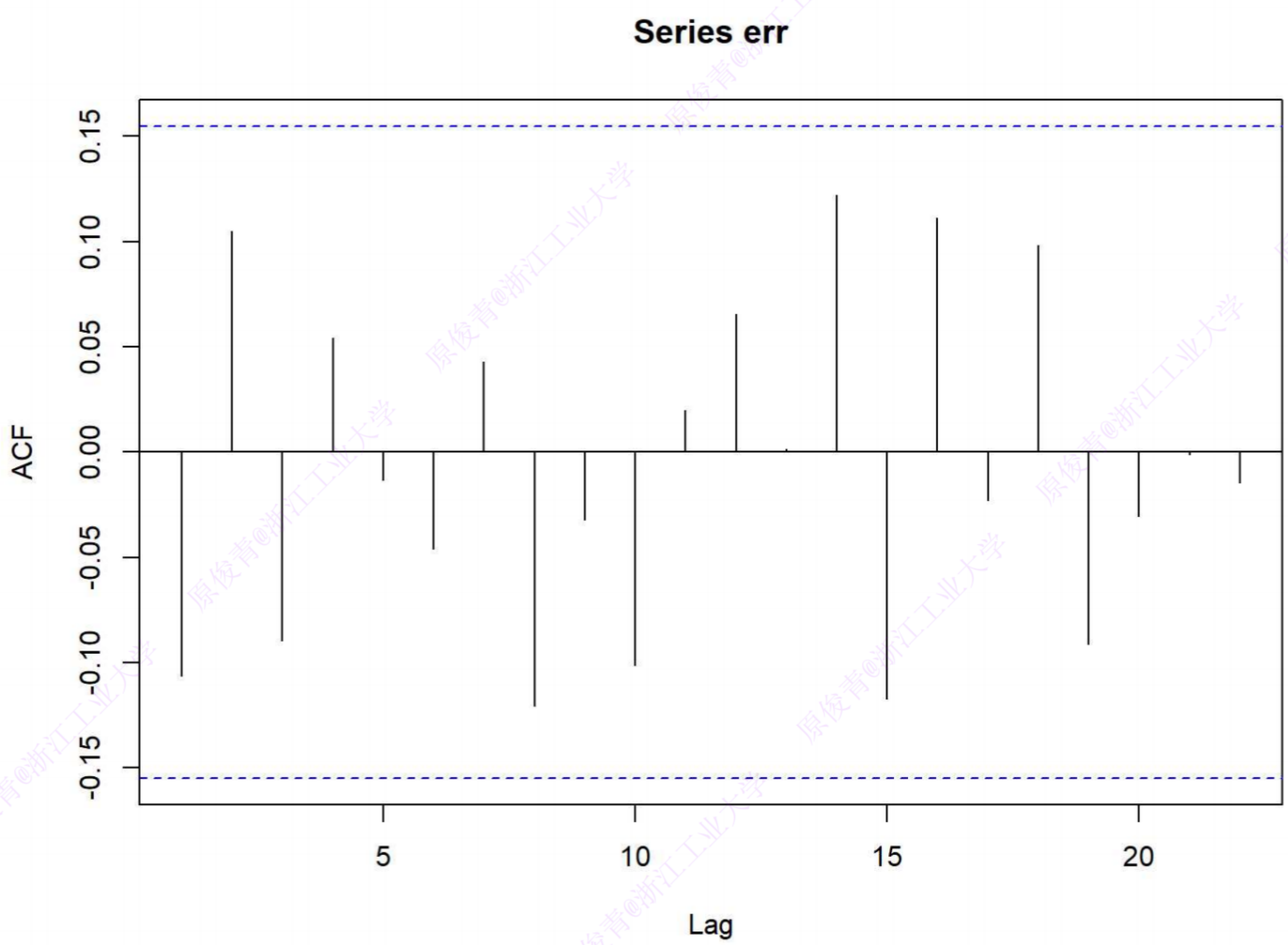
```
shapiro.test(err) ##H_0: 正态分布
```

```
##
##  Shapiro-Wilk normality test
##
## data:  err
## W = 0.92185, p-value = 1.321e-07
```

- 残差序列是否独立

-④. 绘制时间序列的自相关函数图ACF， 检验相关性；

```
acf(err)  ##自相关函数图
```


Series err
ACF
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
5
10
15
20
Lag

-⑤. Durbin-Watson独立性检验， H_0 : 不相关 $\rho = 0$;

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n}(e_i)^2}$$

若残差项间是正相关，则其差异必小；若残差项间是负相关，则其差异必大。

- 当DW值愈接近2时，残差项间愈无相关
- 当DW值愈接近0时，残差项间正相关愈强
- 当DW值愈接近4时，残差项间负相关愈强

```
##DW检验，H_0: 不相关
DW.test<-function(tserr){
Z_err <- zlag(tserr)
D_err <- tserr[-1]-Z_err[-1]
crossprod(D_err,D_err)/crossprod(tserr[-1],tserr[-1])
}

tserr <- residuals(model_2);
DW.test(tserr)
```

```
##              [,1]
## [1,] 2.261129
```

-⑥. 游程检验， H_0 : 独立性;

https://baike.baidu.com/item/%E6%B8%B8%E7%A8%8B%E6%A3%80%E9%AA%8C/9163984?fr=aladdin

```r
runs(err) ##游程检验, H_0: 独立
```

```
## $pvalue
## [1] 0.962
##
## $observed.runs
## [1] 79
##
## $expected.runs
## [1] 78.1875
##
## $n1
## [1] 65
##
## $n2
## [1] 95
##
## $k
## [1] 0
```