

Multivariate Statistical Analysis

多元统计分析

2026年3月19日

第 3 章 多元中的一些概念与模型

3.1 协方差

- ▶ 协方差是相关性的一种度量.
- ▶ 协方差仅度量线性相关程度.
- ▶ 协方差与度量单位有关.
- ▶ 协方差为零也可能存在非线性相关.
- ▶ 协方差为零不一定独立.
- ▶ 独立协方差一定为零.
- ▶ 协方差为负对应着向下倾斜的散点图.
- ▶ 协方差为正对应着向上倾斜的散点图.
- ▶ 变量与自身的协方差就是方差 $\text{Cov}(X, X) = \sigma_{XX} = \sigma_X^2$.
- ▶ 对于较小的 n , 在计算协方差时, 我们用 $\frac{1}{n-1}$ 代替 $\frac{1}{n}$.

第 3 章 多元中的一些概念与模型

3.2 相关系数

- ▶ 相关系数是标准化之后对相关性的—种度量.
- ▶ 相关系数的绝对值不超过 1.
- ▶ 相关系数仅度量线性相关程度.
- ▶ 不相关亦可能存在非线性的相关.
- ▶ 不相关不一定独立. 但对两个正态分布的变量而言, 不相关则意味着独立.
- ▶ 独立—定不相关.
- ▶ 负相关对应着向下倾斜的散点图.
- ▶ 正相关对应着向上倾斜的散点图.
- ▶ Fisher 的 Z 变换 $W = \frac{1}{2} \log \left(\frac{1 + r_{XY}}{1 - r_{XY}} \right)$ 可用于相关系数的假设检验.
- ▶ 对小样本, 可用修正后的 Fisher 的 Z 变换 $W^* = W - \frac{3W + \tanh(W)}{4(n - 1)}$.

描述性统计量 (Summary Statistics)

- 讨论基本的描述性统计量 (均值向量、协方差矩阵和相关矩阵) 的矩阵表示.

▶ 随机向量 $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$. 数据矩阵 $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$.

▶ 重心: $\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n} \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}_{p \times n} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times 1} = \frac{1}{n} \mathcal{X}^T \mathbf{1}_n$ (均值向量)

描述性统计量 (Summary Statistics)

▶ 样本协方差矩阵:

$$S = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

$$= \frac{1}{n} \begin{pmatrix} \sum_{k=1}^n (x_{k1} - \bar{x}_1)^2 & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{kp} - \bar{x}_p) \\ \sum_{k=1}^n (x_{k2} - \bar{x}_2)(x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{k2} - \bar{x}_2)^2 & \cdots & \sum_{k=1}^n (x_{k2} - \bar{x}_2)(x_{kp} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n (x_{kp} - \bar{x}_p)(x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{kp} - \bar{x}_p)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^n (x_{kp} - \bar{x}_p)^2 \end{pmatrix}$$

$$= \frac{1}{n} \begin{pmatrix} \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{kp} - \bar{x}_p) \\ \sum_{k=1}^n (x_{k2} - \bar{x}_2)(x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{k2} - \bar{x}_2)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^n (x_{k2} - \bar{x}_2)(x_{kp} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n (x_{kp} - \bar{x}_p)(x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{kp} - \bar{x}_p)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^n (x_{kp} - \bar{x}_p)(x_{kp} - \bar{x}_p) \end{pmatrix}$$

描述性统计量 (Summary Statistics)

▶ 样本协方差矩阵:

$$S = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

$$= \frac{1}{n} \sum_{k=1}^n \begin{pmatrix} (x_{k1} - \bar{x}_1)(x_{k1} - \bar{x}_1) & (x_{k1} - \bar{x}_1)(x_{k2} - \bar{x}_2) & \cdots & (x_{k1} - \bar{x}_1)(x_{kp} - \bar{x}_p) \\ (x_{k2} - \bar{x}_2)(x_{k1} - \bar{x}_1) & (x_{k2} - \bar{x}_2)(x_{k2} - \bar{x}_2) & \cdots & (x_{k2} - \bar{x}_2)(x_{kp} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{kp} - \bar{x}_p)(x_{k1} - \bar{x}_1) & (x_{kp} - \bar{x}_p)(x_{k2} - \bar{x}_2) & \cdots & (x_{kp} - \bar{x}_p)(x_{kp} - \bar{x}_p) \end{pmatrix}$$

$$= \frac{1}{n} \begin{pmatrix} \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{kp} - \bar{x}_p) \\ \sum_{k=1}^n (x_{k2} - \bar{x}_2)(x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{k2} - \bar{x}_2)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^n (x_{k2} - \bar{x}_2)(x_{kp} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n (x_{kp} - \bar{x}_p)(x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{kp} - \bar{x}_p)(x_{k2} - \bar{x}_2) & \cdots & \sum_{k=1}^n (x_{kp} - \bar{x}_p)(x_{kp} - \bar{x}_p) \end{pmatrix}$$

描述性统计量 (Summary Statistics)

▶ 样本协方差矩阵:

$$S = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

$$= \frac{1}{n} \sum_{k=1}^n \begin{pmatrix} (x_{k1} - \bar{x}_1)(x_{k1} - \bar{x}_1) & (x_{k1} - \bar{x}_1)(x_{k2} - \bar{x}_2) & \cdots & (x_{k1} - \bar{x}_1)(x_{kp} - \bar{x}_p) \\ (x_{k2} - \bar{x}_2)(x_{k1} - \bar{x}_1) & (x_{k2} - \bar{x}_2)(x_{k2} - \bar{x}_2) & \cdots & (x_{k2} - \bar{x}_2)(x_{kp} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{kp} - \bar{x}_p)(x_{k1} - \bar{x}_1) & (x_{kp} - \bar{x}_p)(x_{k2} - \bar{x}_2) & \cdots & (x_{kp} - \bar{x}_p)(x_{kp} - \bar{x}_p) \end{pmatrix}$$

$$= \frac{1}{n} \sum_{k=1}^n \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix} (x_{k1} - \bar{x}_1, x_{k2} - \bar{x}_2, \cdots, x_{kp} - \bar{x}_p)$$

描述性统计量 (Summary Statistics)

▶ 样本协方差矩阵:

$$S = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

$$= \frac{1}{n} \sum_{k=1}^n \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix} \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix}^T$$

$$= \frac{1}{n} \sum_{k=1}^n \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix} \left(x_{k1} - \bar{x}_1, x_{k2} - \bar{x}_2, \cdots, x_{kp} - \bar{x}_p \right)$$

$$\text{Data matrix: } \mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

描述性统计量 (Summary Statistics)

▶ 样本协方差矩阵:

$$S = \begin{pmatrix} s_{X_1X_1} & s_{X_1X_2} & \cdots & s_{X_1X_p} \\ s_{X_2X_1} & s_{X_2X_2} & \cdots & s_{X_2X_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{X_pX_1} & s_{X_pX_2} & \cdots & s_{X_pX_p} \end{pmatrix}$$

$$= \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k - \bar{\mathbf{x}})^T$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

$$= \frac{1}{n} \sum_{k=1}^n \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix} \begin{pmatrix} x_{k1} - \bar{x}_1 \\ x_{k2} - \bar{x}_2 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix}^T = \frac{1}{n} \sum_{k=1}^n \left(\begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} \right) \cdot \left(\begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} \right)^T$$

描述性统计量 (Summary Statistics)

▶ 样本协方差矩阵:

$$\mathcal{S} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k - \bar{\mathbf{x}})^T$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

$$= \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k^T - \bar{\mathbf{x}}^T) = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k \mathbf{x}_k^T - \mathbf{x}_k \bar{\mathbf{x}}^T - \bar{\mathbf{x}} \mathbf{x}_k^T + \bar{\mathbf{x}} \bar{\mathbf{x}}^T)$$

$$= \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k \mathbf{x}_k^T) - \left(\frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \right) \bar{\mathbf{x}}^T - \bar{\mathbf{x}} \left(\frac{1}{n} \sum_{k=1}^n \mathbf{x}_k^T \right) + \frac{1}{n} \sum_{k=1}^n (\bar{\mathbf{x}} \bar{\mathbf{x}}^T)$$

$$= \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k \mathbf{x}_k^T) - \bar{\mathbf{x}} \bar{\mathbf{x}}^T - \bar{\mathbf{x}} \bar{\mathbf{x}}^T + \bar{\mathbf{x}} \bar{\mathbf{x}}^T$$

$$= \frac{1}{n} \left[\sum_{k=1}^n (\mathbf{x}_k \mathbf{x}_k^T) \right] - (\bar{\mathbf{x}} \bar{\mathbf{x}}^T) = \frac{1}{n} (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} - \bar{\mathbf{x}} \bar{\mathbf{x}}^T = \frac{1}{n} \mathcal{X}^T \mathcal{X} - \bar{\mathbf{x}} \bar{\mathbf{x}}^T$$

描述性统计量 (Summary Statistics)

- ▶ 样本协方差矩阵:

$$\mathcal{S} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k - \bar{\mathbf{x}})^T$$

$$= \frac{1}{n} \mathcal{X}^T \mathcal{X} - \bar{\mathbf{x}} \bar{\mathbf{x}}^T \quad \leftarrow \quad \bar{\mathbf{x}} = \frac{1}{n} \mathcal{X}^T \mathbf{1}_n$$

$$= \frac{1}{n} \mathcal{X}^T \mathcal{X} - \frac{1}{n} \mathcal{X}^T \mathbf{1}_n \left(\frac{1}{n} \mathcal{X}^T \mathbf{1}_n \right)^T = \frac{1}{n} \mathcal{X}^T \mathcal{X} - \frac{1}{n^2} \mathcal{X}^T \mathbf{1}_n \mathbf{1}_n^T \mathcal{X}$$

$$= \frac{1}{n} \left(\mathcal{X}^T \mathcal{X} - \frac{1}{n} \mathcal{X}^T \mathbf{1}_n \mathbf{1}_n^T \mathcal{X} \right)$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

描述性统计量 (Summary Statistics)

- ▶ 样本协方差矩阵:

$$\mathcal{S} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k - \bar{\mathbf{x}})^T$$

$$= \frac{1}{n} \mathcal{X}^T \mathcal{X} - \bar{\mathbf{x}} \bar{\mathbf{x}}^T$$

$$= \frac{1}{n} \left(\mathcal{X}^T \mathcal{X} - \frac{1}{n} \mathcal{X}^T \mathbf{1}_n \mathbf{1}_n^T \mathcal{X} \right) = \frac{1}{n} \mathcal{X}^T \left(\mathcal{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \mathcal{X}$$

$$= \frac{1}{n} \mathcal{X}^T \mathcal{H} \mathcal{X}$$

Data matrix: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

中心化矩阵 $\mathcal{H} = \mathcal{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$

描述性统计量 (Summary Statistics)

- ▶ 中心化矩阵 $\mathcal{H} = \mathcal{J}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ 是对称幂等阵.

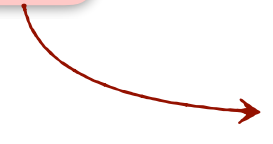
$$\mathcal{H}^T = \left(\mathcal{J}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \right)^T = \mathcal{J}_n^T - \frac{1}{n} (\mathbf{1}_n^T)^T \mathbf{1}_n^T = \mathcal{J}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T = \mathcal{H}$$

$$\mathcal{H}^2 = \left(\mathcal{J}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \right)^2 = \left(\mathcal{J}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \right) \left(\mathcal{J}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \right)$$

$$= \mathcal{J}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T + \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \right) \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \right)$$

$$= \mathcal{J}_n - \frac{2}{n}\mathbf{1}_n\mathbf{1}_n^T + \frac{1}{n^2}\mathbf{1}_n \mathbf{1}_n^T \mathbf{1}_n \mathbf{1}_n^T$$

$$= \mathcal{J}_n - \frac{2}{n}\mathbf{1}_n\mathbf{1}_n^T + \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$$


 $\mathbf{1}_n^T \mathbf{1}_n = n$

$$= \mathcal{J}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T = \mathcal{H}$$

描述性统计量 (Summary Statistics)

- ▶ 样本协方差矩阵 \mathcal{S} 是非负定矩阵, 即 $\mathcal{S} \geq 0$.

$$\forall \mathbf{a} \in \mathbb{R}^p,$$

$$\begin{aligned} \mathbf{a}^T \mathcal{S} \mathbf{a} &= \frac{1}{n} \mathbf{a}^T \mathcal{X}^T \mathcal{H} \mathcal{X} \mathbf{a} && \longleftarrow \mathcal{S} = \frac{1}{n} \mathcal{X}^T \mathcal{H} \mathcal{X} \\ &= \frac{1}{n} \mathbf{a}^T \mathcal{X}^T (\mathcal{H}^T \mathcal{H}) \mathcal{X} \mathbf{a} && \longleftarrow \mathcal{H} \mathcal{H} = \mathcal{H}^T \mathcal{H} = \mathcal{H} \\ &= \frac{1}{n} (\mathbf{a}^T \mathcal{X}^T \mathcal{H}^T) (\mathcal{H} \mathcal{X} \mathbf{a}) \\ &= \frac{1}{n} \mathbf{y}^T \mathbf{y} && \longleftarrow \mathbf{y} = \mathcal{H} \mathcal{X} \mathbf{a} \\ &= \frac{1}{n} \sum_{k=1}^n y_k^2 \geq 0 \end{aligned}$$

描述性统计量 (Summary Statistics)

- ▶ 由数理统计中的知识我们知道，一元情形中， $S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ 是

$\sigma^2 = \text{Var}(X)$ 的有偏估计量，而 $\frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ 是 $\sigma^2 = \text{Var}(X)$ 的

无偏估计量.

- ▶ 在多元情形中，可以证明 $\mathcal{S}_u = \frac{n}{n-1} \mathcal{S}$ 是总体协方差矩阵的一个无偏估计量.

描述性统计量 (Summary Statistics)

- ▶ 变量 i 与变量 j 之间的**样本相关系数**为

$$r_{X_i X_j} = \frac{S_{X_i, X_j}}{\sqrt{S_{X_i, X_i} \cdot S_{X_j, X_j}}}$$

- ▶ 相关矩阵 \mathcal{R} 则为

$$\mathcal{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} = \mathcal{D}^{-1/2} \mathcal{S} \mathcal{D}^{-1/2}$$

$$\mathcal{D}^{-1/2} = \text{diag} \left(\sqrt{S_{X_1 X_1}}, \sqrt{S_{X_2 X_2}}, \dots, \sqrt{S_{X_p X_p}} \right) = \begin{pmatrix} \sqrt{S_{X_1 X_1}} & & & \\ & \sqrt{S_{X_2 X_2}} & & \\ & & \ddots & \\ & & & \sqrt{S_{X_p X_p}} \end{pmatrix}$$

描述性统计量 (Summary Statistics)

- 例: 套头衫数据集的样本协方差矩阵的计算.

```

pullover
pullover.mean <- sapply(pullover, mean)
pullover.mean
  
```

```

> pullover.mean
  X1    X2    X3    X4
172.7 104.6 104.0  93.8
  
```

```

> pullover
  X1  X2  X3  X4
1 230 125 200 109
2 181  99  55 107
3 165  97 105  98
4 150 115  85  71
5  97 120  0  82
6 192 100 150 103
7 181  80  85 111
8 189  90 120  93
9 172  95 110  86
10 170 125 130  78
  
```

$$\Rightarrow \bar{\mathbf{x}} = \begin{pmatrix} 172.7 \\ 104.6 \\ 104.0 \\ 93.8 \end{pmatrix}$$

$$\mathcal{X} =$$

```

pullover <- as.matrix(pullover)
pullover.mean <- as.matrix(pullover.mean)
pullover.cov <- (1/10) * t(pullover) %*% pullover - pullover.mean %*% t(pullover.mean)
pullover.cov
  
```

$$\mathcal{S} = \begin{pmatrix} 1037.21 & -80.02 & 1430.7 & 271.44 \\ -80.02 & 219.84 & 92.1 & -91.58 \\ 1430.70 & 92.10 & 2624.0 & 210.30 \\ 271.44 & -91.58 & 210.3 & 177.36 \end{pmatrix}$$

$$\mathcal{S} = \frac{1}{n} \mathcal{X}^T \mathcal{X} - \bar{\mathbf{x}} \bar{\mathbf{x}}^T$$

描述性统计量 (Summary Statistics)

- 例: 套头衫数据集的样本协方差矩阵的计算.
 - ▶ 协方差矩阵的无偏估计为

```
pullover.cov.u <- (10/9) * pullover.cov
pullover.cov.u
```

```
> pullover.cov.u
      X1      X2      X3      X4
X1 1152.45556 -88.91111 1589.6667 301.6000
X2 -88.91111 244.26667 102.33333 -101.7556
X3 1589.66667 102.33333 2915.5556 233.6667
X4 301.60000 -101.75556 233.6667 197.0667
```

$$\Rightarrow \mathcal{S}_u = \frac{n}{n-1} \mathcal{S} = \begin{pmatrix} 1152.45556 & -88.91111 & 1589.6667 & 301.6000 \\ -88.91111 & 244.26667 & 102.3333 & -101.7556 \\ 1589.66667 & 102.33333 & 2915.5556 & 233.6667 \\ 301.60000 & -101.75556 & 233.6667 & 197.0667 \end{pmatrix}$$

```
pullover.cov.u2 <- cov(pullover)
pullover.cov.u2
```

```
> pullover.cov.u2
      X1      X2      X3      X4
X1 1152.45556 -88.91111 1589.6667 301.6000
X2 -88.91111 244.26667 102.33333 -101.7556
X3 1589.66667 102.33333 2915.5556 233.6667
X4 301.60000 -101.75556 233.6667 197.0667
```

描述性统计量 (Summary Statistics)

- 例: 套头衫数据集的样本协方差矩阵的计算.
 - ▶ 样本相关矩阵为

```
pullover.cor <- cor(pullover)  
round(pullover.cor, digits = 2)
```

```
> round(pullover.cor, digits = 2)  
      X1    X2    X3    X4  
X1  1.00 -0.17  0.87  0.63  
X2 -0.17  1.00  0.12 -0.46  
X3  0.87  0.12  1.00  0.31  
X4  0.63 -0.46  0.31  1.00
```

$$\Rightarrow \mathcal{R} = \begin{pmatrix} 1.00 & -0.17 & 0.87 & 0.63 \\ -0.17 & 1.00 & 0.12 & -0.46 \\ 0.87 & 0.12 & 1.00 & 0.31 \\ 0.63 & -0.46 & 0.31 & 1.00 \end{pmatrix}$$

描述性统计量 (Summary Statistics)

- 线性变换: $Y = \mathcal{A}_{q \times p} X$

随机向量: $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$

数据矩阵: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \triangleq \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

变换矩阵: $\mathcal{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \cdots & a_{qp} \end{pmatrix}$

变换后得随机向量: $Y = \mathcal{A}X = \begin{pmatrix} a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + a_{q2}X_2 + \cdots + a_{qp}X_p \end{pmatrix} \triangleq \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$

变换后的数据矩阵: $\mathcal{Y} = \mathcal{X}\mathcal{A}^T = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix} \triangleq \begin{pmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix}$

描述性统计量 (Summary Statistics)

- 线性变换: $Y = \mathcal{A}_{q \times p} X$

变换后的数据矩阵: $y = x \mathcal{A}^T$

$$\bar{y} = \frac{1}{n} y^T \mathbf{1}_n = \frac{1}{n} \left(x \mathcal{A}^T \right)^T \mathbf{1}_n = \frac{1}{n} \mathcal{A} x^T \mathbf{1}_n$$

$$= \mathcal{A} \left(\frac{1}{n} x^T \mathbf{1}_n \right) = \mathcal{A} \bar{x}$$

$$S_y = \frac{1}{n} y^T \mathcal{H} y = \frac{1}{n} \left(x \mathcal{A}^T \right)^T \mathcal{H} \left(x \mathcal{A}^T \right)$$

中心化矩阵 $\mathcal{H} = \mathcal{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$

$$= \frac{1}{n} \mathcal{A} x^T \mathcal{H} x \mathcal{A}^T$$

$$= \mathcal{A} \left(\frac{1}{n} x^T \mathcal{H} x \right) \mathcal{A}^T$$

$$= \mathcal{A} S_x \mathcal{A}^T$$

描述性统计量 (Summary Statistics)

- 线性变换: $Y = \mathcal{A}_{q \times p} X \implies \begin{cases} \bar{y} = \mathcal{A} \bar{x} \\ \mathcal{S}_y = \mathcal{A} \mathcal{S}_x \mathcal{A}^T \end{cases}$

- ▶ 如果是非齐次的线性变换 $Y = \mathcal{A}X + b$, $b \in \mathbb{R}^q$

$$\implies \begin{cases} \bar{y} = \mathcal{A} \bar{x} + b \\ \mathcal{S}_y = \mathcal{A} \mathcal{S}_x \mathcal{A}^T \end{cases}$$

- ▶ 特别: $q = 1$ 时, $Y = a^T X$, 其中 $a \in \mathbb{R}^p$, 变换后的数据 $y_{n \times 1} = \mathcal{X} a$,

即 $y_i = a^T x_i, i = 1, 2, \dots, n$

$$\implies \begin{cases} \bar{y} = a^T \bar{x} \\ \mathcal{S}_y = a^T \mathcal{S}_x a \end{cases}$$

描述性统计量 (Summary Statistics)

- 例:** 设 \mathcal{X} 是套头衫的数据集. 经理欲计算广告费用 (X_3) 加促销员费用 (X_4) 的平均支出. 假设促销员的时薪为 10 欧元.

pullover

```

> pullover
      X1  X2  X3  X4
[1,] 230 125 200 109
[2,] 181  99  55 107
[3,] 165  97 105  98
[4,] 150 115  85  71
[5,]  97 120   0  82
[6,] 192 100 150 103
[7,] 181  80  85 111
[8,] 189  90 120  93
[9,] 172  95 110  86
[10,] 170 125 130  78
    
```

$$Y = X_3 + 10X_4 = (0 \ 0 \ 1 \ 10) \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \mathcal{A}X$$

```
A <- matrix(c(0, 0, 1, 10), nrow=1)
```

```
A %*% pullover.mean
```

```

> A %*% pullover.mean
      [,1]
[1,] 1042
    
```

$$\Rightarrow \bar{y} = \mathcal{A}\bar{x} = (0 \ 0 \ 1 \ 10) \begin{pmatrix} 172.7 \\ 104.6 \\ 104.0 \\ 93.8 \end{pmatrix} = 1042.0$$

描述性统计量 (Summary Statistics)

- 例:** 设 \mathcal{X} 是套头衫的数据集. 经理欲计算广告费用 (X_3) 加促销员费用 (X_4) 的平均支出. 假设促销员的时薪为 10 欧元.

```
pullover
```

$$Y = X_3 + 10X_4 = (0 \quad 0 \quad 1 \quad 10) \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \mathcal{A}X$$

```
> pullover
```

	X1	X2	X3	X4
[1,]	230	125	200	109
[2,]	181	99	55	107
[3,]	165	97	105	98
[4,]	150	115	85	71
[5,]	97	120	0	82
[6,]	192	100	150	103
[7,]	181	80	85	111
[8,]	189	90	120	93
[9,]	172	95	110	86
[10,]	170	125	130	78

```
S_Y <- A %*% pullover.cov.u2 %*% t(A)
```

```
S_Y
```

```
> S_Y
```

	[,1]
[1,]	27295.56

$$\Rightarrow S_y = \mathcal{A} S_x \mathcal{A}^T$$

$$= (0 \quad 0 \quad 1 \quad 10) \begin{pmatrix} 1152.45556 & -88.91111 & 1589.66667 & 301.60000 \\ -88.91111 & 244.26667 & 102.33333 & -101.75556 \\ 1589.66667 & 102.33333 & 2915.55556 & 233.66667 \\ 301.60000 & -101.75556 & 233.66667 & 197.06667 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 10 \end{pmatrix}$$

$$= 27295.56$$

描述性统计量 (Summary Statistics)

- Mahalanobis 变换

- ▶ 线性变换的一种特殊情形

$$\mathbf{z}_i = \mathcal{S}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, 2, \dots, n$$

$$\mathcal{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)^T = \begin{pmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{pmatrix} = \begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \mathcal{S}^{-1/2} \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^T \mathcal{S}^{-1/2} \\ \vdots \\ (\mathbf{x}_n - \bar{\mathbf{x}})^T \mathcal{S}^{-1/2} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{x}_1^T - \bar{\mathbf{x}}^T \\ \mathbf{x}_2^T - \bar{\mathbf{x}}^T \\ \vdots \\ \mathbf{x}_n^T - \bar{\mathbf{x}}^T \end{pmatrix} \mathcal{S}^{-1/2} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \mathcal{S}^{-1/2} - \begin{pmatrix} \bar{\mathbf{x}}^T \\ \bar{\mathbf{x}}^T \\ \vdots \\ \bar{\mathbf{x}}^T \end{pmatrix} \mathcal{S}^{-1/2}$$

$$= \mathcal{X} \mathcal{S}^{-1/2} - \mathbf{1}_n \bar{\mathbf{x}}^T \mathcal{S}^{-1/2} = \mathcal{X} \mathcal{S}^{-1/2} - \mathbf{1}_n \left(\frac{1}{n} \mathcal{X}^T \mathbf{1}_n \right)^T \mathcal{S}^{-1/2}$$

$$= \mathcal{X} \mathcal{S}^{-1/2} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \mathcal{X} \mathcal{S}^{-1/2} = \mathcal{H} \mathcal{X} \mathcal{S}^{-1/2}$$

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \triangleq \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n} \mathcal{X}^T \mathbf{1}_n$$

$$\mathcal{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$= \frac{1}{n} \mathcal{X}^T \mathcal{X} - \bar{\mathbf{x}} \bar{\mathbf{x}}^T$$

$$= \frac{1}{n} \mathcal{X}^T \mathcal{H} \mathcal{X}$$

$$\mathcal{S} \geq 0$$

$$\mathcal{H} = \mathcal{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$$

描述性统计量 (Summary Statistics)

- Mahalanobis 变换

- ▶ 线性变换的一种特殊情形

$$\mathbf{z}_i = \mathcal{S}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, 2, \dots, n$$

$$\mathcal{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)^T = \mathcal{H} \mathcal{X} \mathcal{S}^{-1/2}$$

$$\mathcal{S}_{\mathcal{Z}} = \frac{1}{n} \mathcal{Z}^T \mathcal{H} \mathcal{Z} = \frac{1}{n} (\mathcal{H} \mathcal{X} \mathcal{S}^{-1/2})^T \mathcal{H} (\mathcal{H} \mathcal{X} \mathcal{S}^{-1/2})$$

$$= \frac{1}{n} \mathcal{S}^{-1/2} \mathcal{X}^T \mathcal{H}^T \mathcal{H} \mathcal{H} \mathcal{X} \mathcal{S}^{-1/2}$$

$$= \frac{1}{n} \mathcal{S}^{-1/2} \mathcal{X}^T \mathcal{H} \mathcal{X} \mathcal{S}^{-1/2}$$

$$= \mathcal{S}^{-1/2} \mathcal{S} \mathcal{S}^{-1/2} = \mathcal{I}_p$$

对称、幂等

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \triangleq \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n} \mathcal{X}^T \mathbf{1}_n$$

$$\mathcal{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$= \frac{1}{n} \mathcal{X}^T \mathcal{X} - \bar{\mathbf{x}} \bar{\mathbf{x}}^T$$

$$= \frac{1}{n} \mathcal{X}^T \mathcal{H} \mathcal{X}$$

$$\mathcal{H} = \mathcal{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$$

描述性统计量 (Summary Statistics)

- Mahalanobis 变换

- ▶ 线性变换的一种特殊情形

$$\mathbf{z}_i = \mathcal{S}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}) , \quad i = 1, 2, \dots, n$$

$$\mathcal{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)^T = \mathcal{H} \mathcal{X} \mathcal{S}^{-1/2}$$

$$\mathcal{S}_{\mathcal{Z}} = \mathcal{I}_p$$

- ▶ 因此, Mahalanobis 变换消除了变量之间的相关性, 并对每一个变量实施了标准化变换.

两个变量的线性模型

- 简单线性回归模型

- ▶ 两个变量 X 与 Y 满足

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$\varepsilon_i \rightarrow E(\varepsilon_i) = 0, \text{Var}(\varepsilon_i) = \sigma^2, \text{ or } \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

- ▶ 估计

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \implies \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\text{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{n \cdot s_{XX}} \implies \text{SE}(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{n \cdot s_{XX}}}$$

两个变量的线性模型

- 简单线性回归模型

- 假设检验 $H_0: \beta_1 = 0 \leftrightarrow H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} \stackrel{H_0}{\sim} t_{n-2} \implies \text{当 } |t| \geq t_{n-2} \left(\frac{\alpha}{2} \right) \text{ 时拒绝 } H_0$$

\implies 当 $p\text{值} = 2 \times P(t_{n-2} > |t|) < \alpha$ 时拒绝 H_0

- 可以证明

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

总平方和 = 回归平方和 + 残差平方和

- 决定系数

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\text{回归平方和}}{\text{总平方和}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

两个变量的线性模型

- 例: 我们对套头衫数据集应用线性回归模型.

```

pullover
plot(X1 ~ X2, data = pullover, xlab = "Price (X2)", ylab = "Sales (X1)", pch = 16, cex = 1.5)
pullover.lm <- lm(X1 ~ X2, data = pullover)
abline(pullover.lm, col = "red", lwd = 2)
summary(pullover.lm)
  
```

```

> pullover
      X1  X2  X3  X4
1  230 125 200 109
2  181  99  55 107
3  165  97 105  98
4  150 115  85  71
5   97 120   0  82
6  192 100 150 103
7  181  80  85 111
8  189  90 120  93
9  172  95 110  86
10 170 125 130  78
  
```

```

> summary(pullover.lm)

Call:
lm(formula = X1 ~ X2, data = pullover)

Residuals:
    Min       1Q   Median       3Q      Max
-70.095  -8.898   2.036   9.805  64.725

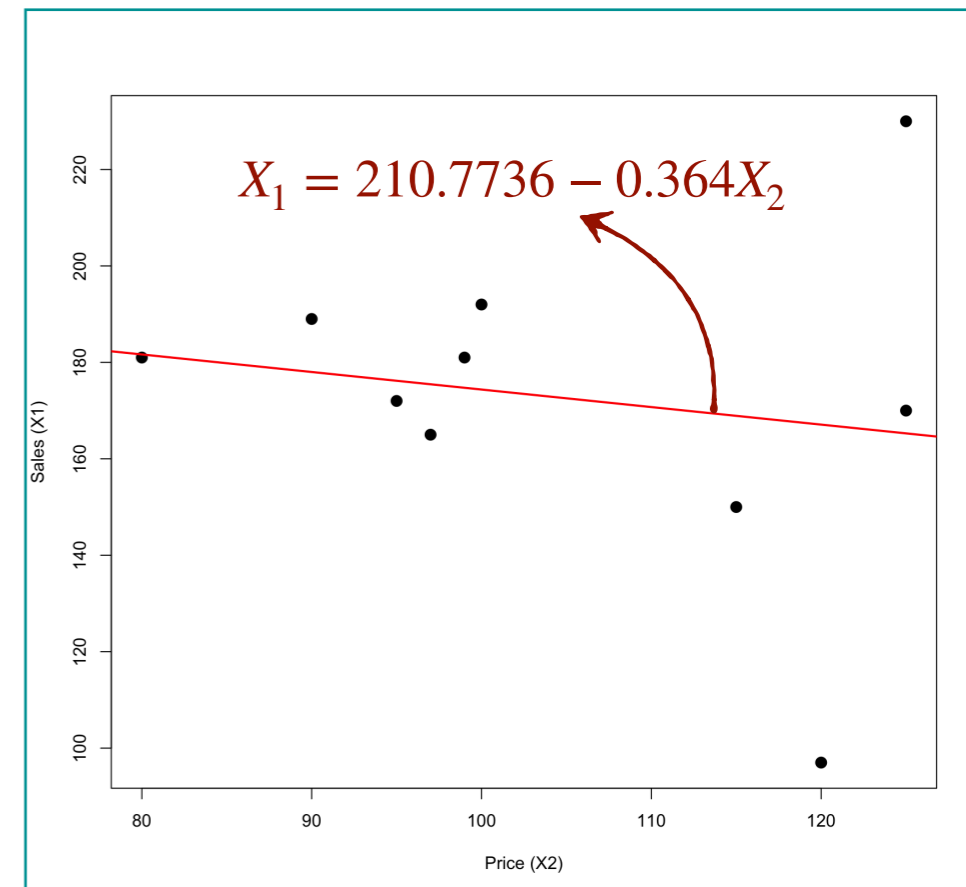
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.7736   79.9837   2.635  0.0299 *
X2          -0.3640    0.7571  -0.481  0.6435
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.5 on 8 degrees of freedom
Multiple R-squared:  0.02808,    Adjusted R-squared:  -0.09341
F-statistic: 0.2311 on 1 and 8 DF,  p-value: 0.6435
  
```

$H_0: \beta_1 = 0 \leftrightarrow H_1: \beta_1 \neq 0$

接受 H_0

$r^2 = 0.02808$



单因子方差分析 (Simple Analysis of Variance)

- 单因子方差分析 (ANOVA)

- ▶ 假设响应变量 y 的平均值仅受单一因子影响.

- ▶ 该因子有 p 个不同值 (水平).

- ▶ 对该因子的每一个水平观测 m 次.

- ▶ 假设所有的观测结果相互独立.

- ▶ 目的: 分析不同水平的均值是否有显著差异

$$\Leftrightarrow \begin{cases} H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0 \\ H_1 : \text{at least one of } \alpha_j \neq 0 \end{cases}$$

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_p = \mu \\ H_1 : \mu_k \neq \mu_l \text{ for some } k \text{ and } l \end{cases}$$

$$y_{ij} = \mu_j + \varepsilon_{ij} = (\mu + \alpha_j) + \varepsilon_{ij}, \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, p$$

Sample element i	Factor levels j				
	1	...	j	...	p
1	y_{11}	...	y_{1j}	...	y_{1p}
2	y_{21}	...	y_{2j}	...	y_{2p}
\vdots	\vdots		\vdots		\vdots
i	y_{i1}	...	y_{ij}	...	y_{ip}
\vdots	\vdots		\vdots		\vdots
m	y_{m1}	...	y_{mj}	...	y_{mp}

单因子方差分析 (Simple Analysis of Variance)

- 单因子方差分析 (ANOVA)

$$y_{ij} = \mu_j + \varepsilon_{ij} = (\mu + \alpha_j) + \varepsilon_{ij}, \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, p$$

第 j 个水平的均值 \leftarrow μ_j \leftarrow $(\mu + \alpha_j)$ \leftarrow $\varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

- 当 $m = 1, p = n, \mu_j = \beta_0 + \beta_1 x_j$ 时, 上述模型就是线性回归模型, 其中 x_j 是因子的第 j 个水平的值.

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

Sample element i	Factor levels j				
	x_1	...	x_j	...	x_n
1	y_1	...	y_j	...	y_n

单因子方差分析 (Simple Analysis of Variance)

● **例:** 销售套头衫的公司欲分析三种营销策略的效果.

1. 本地报纸作广告

2. 有促销员

将这些策略在 10 家不同的商店进行了试验.

3. 橱窗里的奢华展示

```

Sales <- c(9, 11, 10, 12, 7, 11, 12, 10, 11, 13, 10, 15, 11, 15, 15, 13, 7, 15, 11, 13, 10, 18,
          14, 17, 9, 14, 17, 16, 14, 17, 15)
Strategy <- as.factor(c(rep(1, 10), rep(2, 10), rep(3, 10)))
Pullover <- data.frame(Sales = Sales, Strategy = Strategy)
Pullover
  
```

> Pullover		
	Sales	Strategy
1	9	1
2	11	1
3	10	1
4	12	1
5	7	1
6	11	1
7	12	1
8	10	1
9	11	1
10	13	1
11	10	2
12	15	2
13	11	2
14	15	2
15	15	2
16	13	2
17	7	2
18	15	2
19	13	2
20	10	2
21	18	3
22	14	3
23	17	3
24	9	3
25	14	3
26	17	3
27	16	3
28	14	3
29	17	3
30	15	3

▶ 公司欲分析这三种营销策略的平均效果是否相同, 或是否存在差异.

$$\begin{cases} H_0 & : \mu_1 = \mu_2 = \mu_3 = \mu \\ H_1 & : \mu_k \neq \mu_l \text{ for some } k \text{ and } l \end{cases}$$

$$\Leftrightarrow \begin{cases} H_0 & : \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ H_1 & : \text{at least one of } \alpha_j \neq 0 \end{cases}$$

单因子方差分析 (Simple Analysis of Variance)

● 例: 销售套头衫的公司欲分析三种营销策略的效果.

1. 本地报纸作广告

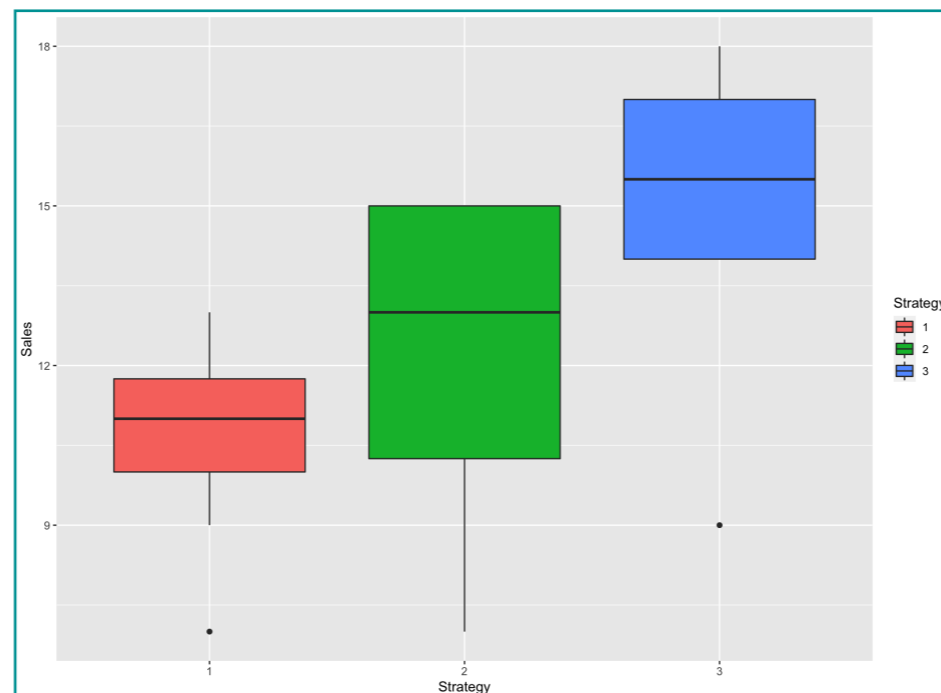
2. 有促销员

3. 橱窗里的奢华展示

将这些策略在 10 家不同的商店进行了试验.

可视化

```
library(ggplot2)
p1 <- ggplot(data = Pullover) +
  geom_boxplot(aes(x = Strategy, y = Sales, fill = Strategy))
p1
```



> Pullover	Sales	Strategy
1	9	1
2	11	1
3	10	1
4	12	1
5	7	1
6	11	1
7	12	1
8	10	1
9	11	1
10	13	1
11	10	2
12	15	2
13	11	2
14	15	2
15	15	2
16	13	2
17	7	2
18	15	2
19	13	2
20	10	2
21	18	3
22	14	3
23	17	3
24	9	3
25	14	3
26	17	3
27	16	3
28	14	3
29	17	3
30	15	3

单因子方差分析 (Simple Analysis of Variance)

- **例:** 销售套头衫的公司欲分析三种营销策略的效果.

1. 本地报纸作广告

2. 有促销员

3. 橱窗里的奢华展示

将这些策略在 10 家不同的商店进行了试验.

因子 (营销策略) 各水平的均值、标准差

```

library(dplyr)
group_by(Pullover, Strategy) %>%
  summarise(
    n_i = n(),
    hat_mean_i = mean(Sales),
    hat_sd_i = sd(Sales)
  )

```

```

# A tibble: 3 × 4
  Strategy  n_i hat_mean_i hat_sd_i
  <fct>    <int>    <dbl>    <dbl>
1 1         10     10.6     1.71
2 2         10     12.4     2.80
3 3         10     15.1     2.60

```

> Pullover		
	Sales	Strategy
1	9	1
2	11	1
3	10	1
4	12	1
5	7	1
6	11	1
7	12	1
8	10	1
9	11	1
10	13	1
11	10	2
12	15	2
13	11	2
14	15	2
15	15	2
16	13	2
17	7	2
18	15	2
19	13	2
20	10	2
21	18	3
22	14	3
23	17	3
24	9	3
25	14	3
26	17	3
27	16	3
28	14	3
29	17	3
30	15	3

单因子方差分析 (Simple Analysis of Variance)

- 例: 销售套头衫的公司欲分析三种营销策略的效果.

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu \\ H_1 : \mu_k \neq \mu_l \text{ for some } k \text{ and } l \end{cases} \iff \begin{cases} H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ H_1 : \text{at least one of } \alpha_j \neq 0 \end{cases}$$

► 模型拟合

```

sales.lm <- lm(Sales ~ Strategy, data = Pullover)
summary(sales.lm)
  
```

```

> summary(sales.lm)

Call:
lm(formula = Sales ~ Strategy, data = Pullover)

Residuals:
    Min       1Q   Median       3Q      Max
 -6.1    -1.1     0.4     1.9     2.9

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6000    0.7642  13.870 8.44e-14 ***
Strategy2     1.8000    1.0808   1.665 0.107392
Strategy3     4.5000    1.0808   4.164 0.000287 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.417 on 27 degrees of freedom
Multiple R-squared:  0.3942,    Adjusted R-squared:  0.3493
F-statistic: 8.783 on 2 and 27 DF,  p-value: 0.001153
  
```

单因子方差分析 (Simple Analysis of Variance)

Sample element i	Factor levels j				
	1	...	j	...	p
1	y_{11}	...	y_{1j}	...	y_{1p}
2	y_{21}	...	y_{2j}	...	y_{2p}
...
i	y_{i1}	...	y_{ij}	...	y_{ip}
...
m	y_{m1}	...	y_{mj}	...	y_{mp}

- 单因子方差分析 (ANOVA)

- 总偏差平方和 (H_0 为真时)

$$SS_{\text{reduced}} = \sum_{i=1}^m \sum_{j=1}^p (y_{ij} - \bar{y})^2$$

$\bar{y} = \frac{1}{mp} \sum_{i=1}^m \sum_{j=1}^p y_{ij}$ 总均值

- H_1 为真时的偏差平方和

$$SS_{\text{full}} = \sum_{i=1}^m \sum_{j=1}^p (y_{ij} - \bar{y}_j)^2$$

$\bar{y}_j = \frac{1}{m} \sum_{i=1}^m y_{ij}$ 第 j 个水平的均值

- 可以证明

$$\sum_{i=1}^m \sum_{j=1}^p (y_{ij} - \bar{y})^2 = m \sum_{j=1}^p (\bar{y}_j - \bar{y})^2 + \sum_{i=1}^m \sum_{j=1}^p (y_{ij} - \bar{y}_j)^2$$

$\xrightarrow{SS_{\text{treatment}}}$

总偏差平方和 = 组间偏差平方和 + 组内偏差平方和

$$SS_{\text{reduced}} = SS_{\text{treatment}} + SS_{\text{full}}$$

单因子方差分析 (Simple Analysis of Variance)

- 单因子方差分析 (ANOVA)

- 单因子方差分析模型中的大部分信息都包含在方差分析表中

Source	SS	df	MS	E(MS)
Treatment	$SS_{\text{treatment}} = m \sum_{j=1}^p (\bar{y}_j - \bar{y})^2$	$p - 1$	$\frac{SS_{\text{treatment}}}{p - 1}$	$\sigma^2 + \frac{m}{p - 1} \sum_{j=1}^p \alpha_j^2$
Error	$SS_{\text{full}} = \sum_{i=1}^m \sum_{j=1}^p (y_{ij} - \bar{y}_j)^2$	$p(m - 1)$	$\frac{SS_{\text{full}}}{p(m - 1)}$	σ^2

- 方差分析表中的行通常是独立的。

- 当 $H_0 : \mu_1 = \mu_2 = \dots = \mu_p = \mu$ 或 $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ 为真时

$$F = \frac{MS_{\text{treatment}}}{MS_{\text{full}}} = \frac{\frac{SS_{\text{treatment}}}{df_{\text{treatment}}}}{\frac{SS_{\text{full}}}{df_{\text{full}}}} \sim F_{df_{\text{treatment}}, df_{\text{full}}}$$

- 显著水平取 α , 则当 $F \geq F_{df_{\text{treatment}}, df_{\text{full}}}(\alpha)$ 时拒绝 H_0 .

单因子方差分析 (Simple Analysis of Variance)

- 例: 销售套头衫的公司欲分析三种营销策略的效果.

$$\begin{cases} H_0 & : \mu_1 = \mu_2 = \mu_3 = \mu \\ H_1 & : \mu_k \neq \mu_l \text{ for some } k \text{ and } l \end{cases} \iff \begin{cases} H_0 & : \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ H_1 & : \text{at least one of } \alpha_j \neq 0 \end{cases}$$

```
anova(sales.lm)
```

```

> anova(sales.lm)
Analysis of Variance Table

Response: Sales
      Df Sum Sq Mean Sq F value    Pr(>F)
Strategy  2  102.6   51.300   8.7831 0.001153 **
Residuals 27  157.7    5.841
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  
```

- 我们可以用 \mathcal{R} 中的 `aov()` 函数来作方差分析

```

sales.aov <- aov(Sales ~ Strategy, data = Pullover)
summary(sales.aov)
  
```

```

> summary(sales.aov)
      Df Sum Sq Mean Sq F value    Pr(>F)
Strategy  2  102.6   51.30    8.783 0.00115 **
Residuals 27  157.7    5.84
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  
```

多重线性模型 (Multiple Linear Model)

- 响应变量 Y , 数据向量 $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ 是响应变量 Y 的一组观测值.
- 解释变量 $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$, 数据矩阵 $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$ 是解释变量 \mathbf{X} 的一组观测值.
 $\begin{matrix} \uparrow & \uparrow & \cdots & \uparrow \\ X_1 & X_2 & \cdots & X_p \end{matrix}$

多重线性模型 (Multiple Linear Model)

- 线性模型 (无截距项) 可以表示为

$$y = X\beta + \varepsilon$$

回归系数: $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$

误差项: $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

- 已有理论的重要应用: 最小二乘拟合.

$$\hat{\beta} = \arg \min_{\beta} (y - X\beta)^T (y - X\beta)$$

- ▶ 如果矩阵 $(X^T X)$ 满秩, 因此可逆.

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

- ▶ 拟合值

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = \mathcal{P} y$$

向量 y 在列空间 $C(X)$ 上的投影

多重线性模型 (Multiple Linear Model)

- 线性模型 (无截距项) 可以表示为

$$y = X\beta + \varepsilon$$

误差项: $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

回归系数: $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$

- 已有理论的重要应用: 最小二乘拟合.

$$\hat{\beta} = \arg \min_{\beta} (y - X\beta)^T (y - X\beta)$$

- ▶ 如果矩阵 $(X^T X)$ 满秩, 因此可逆.

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

- ▶ 最小二乘残差为

$$e = y - \hat{y} = y - X\hat{\beta} = y - \mathcal{P}y = (\mathcal{I}_n - \mathcal{P})y = Qy$$

向量 y 在 $C(X)$ 的正交补上的投影

多重线性模型 (Multiple Linear Model)

注 3.5 具有截距项 β_0 的线性模型也有类似的形式. 模型的表达式为

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

亦可表示为矩阵形式

$$\mathbf{y} = \mathcal{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$$

其中 $\mathcal{X}^* = (\mathbf{1}_n \quad \mathcal{X})$ (给数据矩阵添加了一列1向量). 从而我们有

$$\widehat{\boldsymbol{\beta}}^* = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}} \end{pmatrix} = \left(\mathcal{X}^{*\top} \mathcal{X}^* \right)^{-1} \mathcal{X}^{*\top} \mathbf{y}$$

多重线性模型 (Multiple Linear Model)

- 例：我们依然讨论套头衫数据集的例子。

pullover

```
> pullover
```

	X1	X2	X3	X4
1	230	125	200	109
2	181	99	55	107
3	165	97	105	98
4	150	115	85	71
5	97	120	0	82
6	192	100	150	103
7	181	80	85	111
8	189	90	120	93
9	172	95	110	86
10	170	125	130	78

- 我们来作销售量 X_1 关于价格 X_2 ，广告费 X_3 ，以及有促销员 X_4 的一个线性回归拟合。

```
pullover.lm2 <- lm(X1 ~ ., data = pullover)
summary(pullover.lm2)
```

```
> summary(pullover.lm2)
```

```
Call:
lm(formula = X1 ~ ., data = pullover)

Residuals:
    Min       1Q   Median       3Q      Max
-13.369  -9.406   1.599   5.151  19.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.66956   57.12507    1.150  0.29407
X2          -0.21578    0.32194   -0.670  0.52764
X3           0.48519    0.08678    5.591  0.00139 **
X4           0.84373    0.37400    2.256  0.06491 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.7 on 6 degrees of freedom
Multiple R-squared: 0.9067    Adjusted R-squared: 0.8601
F-statistic: 19.44 on 3 and 6 DF, p-value: 0.001713
```

$$X_1 = 65.66956 - 0.21578X_2 + 0.48519X_3 + 0.84373X_4$$

$$r^2 = 0.9067$$

X_1 的变化可由一个线性关系很好地近似。

多重线性模型 (Multiple Linear Model)

注 3.6 决定系数受解释变量个数的影响. 给定样本容量 n , r^2 会随着在线性模型中引入更多解释变量而增加. 因此, 即使引入可能不相关的解释变量, r^2 的值也可能很高. 有 p 个解释变量以及含截距项 (共有 $p + 1$ 个参数) 的调整后的决定系数为

$$r_{\text{adj}}^2 = r^2 - \frac{p(1 - r^2)}{n - p - 1}$$

多重线性模型 (Multiple Linear Model)

- 例：我们依然讨论套头衫数据集的例子。

```
> pullover
```

	X1	X2	X3	X4
1	230	125	200	109
2	181	99	55	107
3	165	97	105	98
4	150	115	85	71
5	97	120	0	82
6	192	100	150	103
7	181	80	85	111
8	189	90	120	93
9	172	95	110	86
10	170	125	130	78

pullover

- ▶ 我们来作销售量 X_1 关于价格 X_2 ，广告费 X_3 ，以及有促销员 X_4 的一个线性回归拟合。

```
pullover.lm2 <- lm(X1 ~ ., data = pullover)
summary(pullover.lm2)
```

```
> summary(pullover.lm2)

Call:
lm(formula = X1 ~ ., data = pullover)

Residuals:
    Min       1Q   Median       3Q      Max
-13.369  -9.406   1.599   5.151  19.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.66956   57.12507   1.150  0.29407
X2          -0.21578    0.32194  -0.670  0.52764
X3           0.48519    0.08678   5.591  0.00139 **
X4           0.84373    0.37400   2.256  0.06491 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.7 on 6 degrees of freedom
Multiple R-squared: 0.9067    Adjusted R-squared: 0.8601
F-statistic: 19.44 on 3 and 6 DF, p-value: 0.001713
```

$$X_1 = 65.66956 - 0.21578X_2 + 0.48519X_3 + 0.84373X_4$$

$$r^2 = 0.9067$$

X_1 的变化可由一个线性关系很好地近似。

$$r_{adj}^2 = 0.8601$$

多重线性模型 (Multiple Linear Model)

- $\hat{\beta}$ 的性质

$$E(\hat{\beta}) = \beta$$

$$\text{Var}(\hat{\beta}) = \sigma^2 (\mathcal{X}^T \mathcal{X})^{-1}$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1} (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}})$$

- 假设检验问题

- ▶ 对假设 $H_0: \beta_j = 0$, 检验统计量: $t = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$, 当 $|t| > t_{n-p-1} \left(\frac{\alpha}{2} \right)$ 时拒绝 H_0

检验的显著水平

- ▶ 对假设 $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$, 我们用 F 统计量进行检验.

多重线性模型 (Multiple Linear Model)

- 例：我们依然讨论套头衫数据集的例子。

pullover

- 我们来作销售量 X_1 关于价格 X_2 ，广告费 X_3 ，以及有促销员 X_4 的一个线性回归拟合。

```
pullover.lm2 <- lm(X1 ~ ., data = pullover)
summary(pullover.lm2)
```

```
> pullover
```

	X1	X2	X3	X4
1	230	125	200	109
2	181	99	55	107
3	165	97	105	98
4	150	115	85	71
5	97	120	0	82
6	192	100	150	103
7	181	80	85	111
8	189	90	120	93
9	172	95	110	86
10	170	125	130	78

```
> summary(pullover.lm2)
```

```
Call:
lm(formula = X1 ~ ., data = pullover)

Residuals:
    Min       1Q   Median       3Q      Max
-13.369  -9.406   1.599   5.151  19.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.66956   57.12507   1.150  0.29407
X2          -0.21578    0.32194  -0.670  0.52764
X3           0.48519    0.08678   5.591  0.00139 **
X4           0.84373    0.37400   2.256  0.06491 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.7 on 6 degrees of freedom
Multiple R-squared:  0.9067,    Adjusted R-squared:  0.8601
F-statistic: 19.44 on 3 and 6 DF,  p-value: 0.001713
```

$H_0 : \beta_1 = 0 \leftrightarrow H_1 : \beta_1 \neq 0 \implies$ 接受 H_0

$H_0 : \beta_2 = 0 \leftrightarrow H_1 : \beta_2 \neq 0 \implies$ 拒绝 H_0

$H_0 : \beta_3 = 0 \leftrightarrow H_1 : \beta_3 \neq 0 \implies$ 拒绝 H_0

$\left\{ \begin{array}{l} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 : \text{at least one } \beta_j \neq 0 \end{array} \right. \implies$ 拒绝 H_0

多重线性模型 (Multiple Linear Model)

- 例：我们依然讨论套头衫数据集的例子。

pullover

- 我们来作销售量 X_1 关于价格 X_2 ，广告费 X_3 ，以及有促销员 X_4 的一个线性回归拟合。

```
pullover.lm2 <- lm(X1 ~ ., data = pullover)
summary(pullover.lm2)
```

```
> pullover
```

	X1	X2	X3	X4
1	230	125	200	109
2	181	99	55	107
3	165	97	105	98
4	150	115	85	71
5	97	120	0	82
6	192	100	150	103
7	181	80	85	111
8	189	90	120	93
9	172	95	110	86
10	170	125	130	78

```
> summary(pullover.lm2)
```

Call:
lm(formula = X1 ~ ., data = pullover)

Residuals:
Min 1Q Median 3Q Max
-13.369 -9.406 1.599 5.151 19.729

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.66956 57.12507 1.150 0.29407
X2 -0.21578 0.32194 -0.670 0.52764
X3 0.48519 0.08678 5.591 0.00139 **
X4 0.84373 0.37400 2.256 0.06491 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.7 on 6 degrees of freedom
Multiple R-squared: 0.9067, Adjusted R-squared: 0.8601
F-statistic: 19.44 on 3 and 6 DF, p-value: 0.001713

$H_0 : \beta_1 = 0 \leftrightarrow H_1 : \beta_1 \neq 0 \implies$ 接受 H_0

$H_0 : \beta_2 = 0 \leftrightarrow H_1 : \beta_2 \neq 0 \implies$ 拒绝 H_0

$H_0 : \beta_3 = 0 \leftrightarrow H_1 : \beta_3 \neq 0 \implies$ 拒绝 H_0

$\left\{ \begin{array}{l} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 : \text{at least one } \beta_j \neq 0 \end{array} \right. \implies$ 拒绝 H_0

Boston 住房数据集

- Boston 住房数据集由 Boston 大区的每个人口普查地区的共 506 个观测值构成。

```
library(MASS)  
str(Boston)
```

```
> str(Boston) 506 个观测值 14 个变量  
'data.frame': 506 obs. of 14 variables:  
 $ crim : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...  
 $ zn : num 18 0 0 0 0 12.5 12.5 12.5 12.5 ...  
 $ indus : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...  
 $ chas : int 0 0 0 0 0 0 0 0 0 ...  
 $ nox : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  
 $ rm : num 6.58 6.42 7.18 7 7.15 ...  
 $ age : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...  
 $ dis : num 4.09 4.97 4.97 6.06 6.06 ...  
 $ rad : int 1 2 2 3 3 3 5 5 5 5 ...  
 $ tax : num 296 242 242 222 222 222 311 311 311 311 ...  
 $ ptratio: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...  
 $ black : num 397 397 393 395 397 ...  
 $ lstat : num 4.98 9.14 4.03 2.94 5.33 ...  
 $ medv : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Boston 住房数据集

- Boston 住房数据集由 Boston 大区的每个人口普查地区的共 506 个观测值构成。

head(Boston)

X_{14} : 自住房屋价格的中位数(单位: 1000美元)

X_6 : 每个住宅的平均房间数

X_4 : Charles 河 (1 指河边, 0 为其它)

X_8 : 到波士顿五个就业中心的加权距离

X_{10} : 每一万美元的全额财产税税率

> head(Boston)

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

X_1 : 人均犯罪率

X_2 : 划作大型住宅用地的比例

X_3 : 非零售业务占地的比例

X_5 : 一氧化氮浓度

X_7 : 1940年之前建造的自住房的比例

X_9 : 辐射状高速公路的可达性指数

X_{11} : 生师比

X_{12} : $1000 (B - 0.63)^2 I(B < 0.63)$

其中 B 是非裔美国人的比例

X_{13} : 社会底层人口的百分比

Boston 住房数据集

- 描述性统计

```
Boston_mean <- sapply(Boston, mean)
Boston_median <- sapply(Boston, median)
Boston_var <- sapply(Boston, var)
Boston_sd <- sapply(Boston, sd)
Boston_descriptive <- tibble(Mean = Boston_mean, Median = Boston_median, Variance = Boston_var,
                             Std = Boston_sd)
Boston_descriptive
```

```
> Boston_descriptive
# A tibble: 14 × 4
  Mean Median Variance Std
  <dbl> <dbl> <dbl> <dbl>
1 3.61 0.257 74.0 8.60
2 11.4 0 544. 23.3
3 11.1 9.69 47.1 6.86
4 0.0692 0 0.0645 0.254
5 0.555 0.538 0.0134 0.116
6 6.28 6.21 0.494 0.703
7 68.6 77.5 792. 28.1
8 3.80 3.21 4.43 2.11
9 9.55 5 75.8 8.71
10 408. 330 28405. 169.
11 18.5 19.0 4.69 2.16
12 357. 391. 8335. 91.3
13 12.7 11.4 51.0 7.14
14 22.5 21.2 84.6 9.20
```

Boston 住房数据集

- 样本协方差矩阵的无偏估计

```
S <- cov(Boston)  
round(S, digits = 2)
```

```
> round(S, digits = 2)
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
crim	73.99	-40.22	23.99	-0.12	0.42	-1.33	85.41	-6.88	46.85	844.82	5.40	-302.38	27.99	-30.72
zn	-40.22	543.94	-85.41	-0.25	-1.40	5.11	-373.90	32.63	-63.35	-1236.45	-19.78	373.72	-68.78	77.32
indus	23.99	-85.41	47.06	0.11	0.61	-1.89	124.51	-10.23	35.55	833.36	5.69	-223.58	29.58	-30.52
chas	-0.12	-0.25	0.11	0.06	0.00	0.02	0.62	-0.05	-0.02	-1.52	-0.07	1.13	-0.10	0.41
nox	0.42	-1.40	0.61	0.00	0.01	-0.02	2.39	-0.19	0.62	13.05	0.05	-4.02	0.49	-0.46
rm	-1.33	5.11	-1.89	0.02	-0.02	0.49	-4.75	0.30	-1.28	-34.58	-0.54	8.22	-3.08	4.49
age	85.41	-373.90	124.51	0.62	2.39	-4.75	792.36	-44.33	111.77	2402.69	15.94	-702.94	121.08	-97.59
dis	-6.88	32.63	-10.23	-0.05	-0.19	0.30	-44.33	4.43	-9.07	-189.66	-1.06	56.04	-7.47	4.84
rad	46.85	-63.35	35.55	-0.02	0.62	-1.28	111.77	-9.07	75.82	1335.76	8.76	-353.28	30.39	-30.56
tax	844.82	-1236.45	833.36	-1.52	13.05	-34.58	2402.69	-189.66	1335.76	28404.76	168.15	-6797.91	654.71	-726.26
ptratio	5.40	-19.78	5.69	-0.07	0.05	-0.54	15.94	-1.06	8.76	168.15	4.69	-35.06	5.78	-10.11
black	-302.38	373.72	-223.58	1.13	-4.02	8.22	-702.94	56.04	-353.28	-6797.91	-35.06	8334.75	-238.67	279.99
lstat	27.99	-68.78	29.58	-0.10	0.49	-3.08	121.08	-7.47	30.39	654.71	5.78	-238.67	50.99	-48.45
medv	-30.72	77.32	-30.52	0.41	-0.46	4.49	-97.59	4.84	-30.56	-726.26	-10.11	279.99	-48.45	84.59

Boston 住房数据集

- 样本相关矩阵

```
R <- cor(Boston)
round(R, digits = 2)
```

```
> R <- cor(Boston)
> round(R, digits = 2)
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
$X_1 \rightarrow$ crim	1.00	-0.20	0.41	-0.06	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	-0.39	0.46	-0.39
$X_2 \rightarrow$ zn	-0.20	1.00	-0.53	-0.04	-0.52	0.31	-0.57	0.66	-0.31	-0.31	-0.39	0.18	-0.41	0.36
$X_3 \rightarrow$ indus	0.41	-0.53	1.00	0.06	0.76	-0.39	0.64	-0.71	0.60	0.72	0.38	-0.36	0.60	-0.48
$X_4 \rightarrow$ chas	-0.06	-0.04	0.06	1.00	0.09	0.09	0.09	-0.10	-0.01	-0.04	-0.12	0.05	-0.05	0.18
$X_5 \rightarrow$ nox	0.42	-0.52	0.76	0.09	1.00	-0.30	0.73	-0.77	0.61	0.67	0.19	-0.38	0.59	-0.43
$X_6 \rightarrow$ rm	-0.22	0.31	-0.39	0.09	-0.30	1.00	-0.24	0.21	-0.21	-0.29	-0.36	0.13	-0.61	0.70
$X_7 \rightarrow$ age	0.35	-0.57	0.64	0.09	0.73	-0.24	1.00	-0.75	0.46	0.51	0.26	-0.27	0.60	-0.38
$X_8 \rightarrow$ dis	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75	1.00	-0.49	-0.53	-0.23	0.29	-0.50	0.25
$X_9 \rightarrow$ rad	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.49	1.00	0.91	0.46	-0.44	0.49	-0.38
$X_{10} \rightarrow$ tax	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1.00	0.46	-0.44	0.54	-0.47
$X_{11} \rightarrow$ ptratio	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1.00	-0.18	0.37	-0.51
$X_{12} \rightarrow$ black	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18	1.00	-0.37	0.33
$X_{13} \rightarrow$ lstat	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	-0.37	1.00	-0.74
$X_{14} \rightarrow$ medv	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	1.00

\uparrow
 X_6
 \uparrow
 X_{10}
 \uparrow
 X_{11}
 \uparrow
 X_{13}

变量 X_{14} 与其它变量的相关系数

Boston 住房数据集

- 对假设 $H_0: \rho_{X_{14}X_i} = 0, i = 1, 2, \dots, 13$, 应用 Fisher 的 Z 变换

$$W = \frac{1}{2} \log \left(\frac{1 + r_{XY}}{1 - r_{XY}} \right)$$

$$E(W) \approx \frac{1}{2} \log \left(\frac{1 + \rho_{XY}}{1 - \rho_{XY}} \right), \quad \text{Var}(W) \approx \frac{1}{n - 3}$$

```

R_14 <- R[14, 1:13]
W_14 <- 0.5 * log((1 + R_14) / (1 - R_14))
Z_14 <- W_14 / sqrt(1/(dim(Boston)[1]-3))
p_value <- 2 * pnorm(abs(Z_14), lower.tail = FALSE)
p_value
    
```

拒绝 $H_0: \rho_{X_{14}X_i} = 0$

```

> p_value
      crim          zn          indus          chas          nox          rm          age
3.895604e-20  2.580349e-17  2.479294e-32  7.136774e-05  1.301673e-24  1.445868e-82  5.965212e-19
      dis          rad          tax          ptratio          black          lstat
1.024590e-08  1.967604e-19  4.299209e-30  3.792391e-36  7.482300e-15  9.230431e-100
    
```

- ▶ 相关系数与 Fisher 的 Z 变换不适用于二值变量，所以，上述检验结果不适用于变量 X_{14} 与 X_4 .

Boston 住房数据集

- 如果我们希望用变量 X_1, X_2, \dots, X_{13} 来解释价格 X_{14} 的变化，我们可以拟合下述线性模型

$$X_{14} = \beta_0 + \sum_{j=1}^{13} \beta_j X_j + \varepsilon$$

```
Boston.lm <- lm(medv ~ ., data = Boston)
summary(Boston.lm)
```

```
> summary(Boston.lm)

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
 crim       -1.080e-01  3.286e-02  -3.287 0.001087 **
  zn         4.642e-02  1.373e-02   3.382 0.000778 ***
  indus      2.056e-02  6.150e-02   0.334 0.738288
  chas       2.687e+00  8.616e-01   3.118 0.001925 **
  nox       -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
   rm       3.810e+00  4.179e-01   9.116 < 2e-16 ***
   age      6.922e-04  1.321e-02   0.052 0.958229
   dis     -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
   rad      3.060e-01  6.635e-02   4.613 5.07e-06 ***
   tax     -1.233e-02  3.760e-03  -3.280 0.001112 **
 ptratio   -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
  black     9.312e-03  2.686e-03   3.467 0.000573 ***
  lstat    -5.248e-01  5.072e-02  -10.347 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

对 X_{14} 的影响不显著

$r^2 = 0.7406$

$r^2_{adj} = 0.7338$