

Multivariate Statistical Analysis

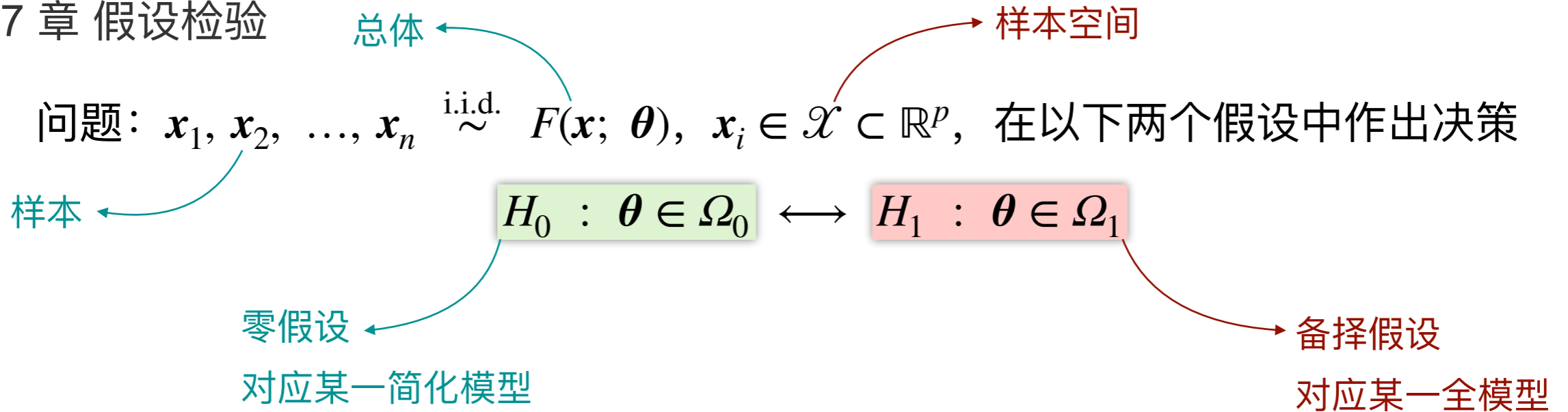
多元统计分析

2026年4月21日

已学知识点 (Recap)

第7章 假设检验

- ▶ 问题: $x_1, x_2, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} F(x; \theta), x_i \in \mathcal{X} \subset \mathbb{R}^p$, 在以下两个假设中作出决策



- ▶ **拒绝域** (rejection region) R : 由样本空间 $\mathcal{X} \subset \mathbb{R}^p$ 中的某些值构成的集合,

犯第 I 类错误的概率

$$\sup_{\theta \in \Omega_0} \mathbb{P}(\mathcal{X} \in R \mid \theta) \leq \alpha$$

检验的显著水平 (事先指定)

- ▶ 似然比:

$$\lambda(\mathcal{X}) \triangleq \frac{L_0^*}{L_1^*} = \frac{\max_{\theta \in \Omega_0} L(\mathcal{X}; \theta)}{\max_{\theta \in \Omega_1} L(\mathcal{X}; \theta)}$$

似然函数

确定临界值 c

- ▶ **似然比检验**: 似然比的值很小时我们拒绝 H_0 , 拒绝域为 $R = \{ \mathcal{X} : \lambda(\mathcal{X}) < c \}$.

因为 $\lambda(\mathcal{X})$ 的抽样分布复杂, 如何将 c 表示为 α 的函数是难点!

已学知识点 (Recap)

第 7 章 假设检验

- 问题: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} F(\mathbf{x}; \boldsymbol{\theta}), \mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^p$, 在以下两个假设中作出决策
- $H_0 : \boldsymbol{\theta} \in \Omega_0 \iff H_1 : \boldsymbol{\theta} \in \Omega_1$
- 总体 \leftarrow 样本空间
 样本 \leftarrow 零假设 \leftarrow 对应某一简化模型
 备择假设 \leftarrow 对应某一全模型

- 对数似然比: $-2 \ln \lambda = 2 \left(\ell_1^* - \ell_0^* \right)$.
- $$\begin{cases} \ell_0^* = \ln L_0^* = \max_{\boldsymbol{\theta} \in \Omega_0} \ln L(\mathcal{X}; \boldsymbol{\theta}) \\ \ell_1^* = \ln L_1^* = \max_{\boldsymbol{\theta} \in \Omega_1} \ln L(\mathcal{X}; \boldsymbol{\theta}) \end{cases}$$
- 拒绝域为 $R = \left\{ \mathcal{X} : -2 \ln \lambda(\mathcal{X}) > k \right\}$.

- Wilks 定理:** 如果 $\Omega_1 \subset \mathbb{R}^q$ 是一个 q 维空间, 而 $\Omega_0 \subset \Omega_1$ 是 r 维的子空间, 则在正则条件下

$$\forall \boldsymbol{\theta} \in \Omega_0 : -2 \ln \lambda \xrightarrow{\mathcal{L}} \chi_{q-r}^2 \quad \text{as } n \rightarrow \infty$$

\leftarrow 上侧 α 分位数

- 渐进地, 似然比检验的拒绝域为 $R = \left\{ \mathcal{X} : -2 \ln \lambda(\mathcal{X}) > \chi_{q-r}^2(\alpha) \right\}$.

已学知识点 (Recap)

第 7 章 假设检验

- ▶ **检验问题 1:** 总体 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 其中 $\boldsymbol{\Sigma}$ 已知, 对假设 $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, 当

$$-2 \ln \lambda = n (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \chi_p^2(\alpha) \text{ 时拒绝 } H_0.$$

已学知识点 (Recap)

第 7 章 假设检验

► **检验问题 2:** 总体 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 其中 $\boldsymbol{\Sigma}$ 未知, 对假设 $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$,

① 精确检验: 当 $\frac{n-p}{p} \cdot (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) > F_{p, n-p}(\alpha)$ 时拒绝 H_0 .

② 渐进检验: 当 $-2 \ln \lambda = n \ln \left[1 + (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \right] > \chi_p^2(\alpha)$ 时拒绝 H_0 .

③ $\boldsymbol{\mu}$ 的置信度为 $1 - \alpha$ 的置信域:

$$\left\{ \boldsymbol{\mu} \in \mathbb{R}^p \mid (\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \leq \frac{p}{n-p} \cdot F_{p, n-p}(\alpha) \right\}$$

④ $\boldsymbol{a}^T \boldsymbol{\mu}$ 的联合置信区间: $\forall \boldsymbol{a} \in \mathbb{R}^p$, 下述区间同时以概率 $1 - \alpha$ 包含 $\boldsymbol{a}^T \boldsymbol{\mu}$ 在内

$$K_\alpha = \frac{p}{n-p} \cdot F_{p, n-p}(\alpha) \quad \left(\boldsymbol{a}^T \bar{\boldsymbol{x}} - \sqrt{K_\alpha \cdot \boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a}}, \quad \boldsymbol{a}^T \bar{\boldsymbol{x}} + \sqrt{K_\alpha \cdot \boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a}} \right)$$

⑤ $\mu_j, j = 1, 2, \dots, p$ 的联合置信区间


$$\left(\bar{x}_j - \sqrt{\frac{p}{n-p} \cdot F_{p, n-p}(\alpha) \cdot s_{jj}}, \quad \bar{x}_j + \sqrt{\frac{p}{n-p} \cdot F_{p, n-p}(\alpha) \cdot s_{jj}} \right)$$

已学知识点 (Recap)

第 7 章 假设检验

- ▶ **检验问题 3:** 总体 $X \sim N_p(\mu, \Sigma)$, 其中 μ 未知, 对假设 $H_0: \Sigma = \Sigma_0$, 当

$$-2 \ln \lambda = n \operatorname{tr}(\Sigma_0^{-1} \mathcal{S}) - n \ln |\Sigma_0^{-1} \mathcal{S}| - np > \chi_m^2(\alpha) \text{ 时拒绝 } H_0.$$


$$m = \frac{1}{2}p(p+1)$$

似然比检验 (Likelihood Ratio Test)

检验问题 4: 假设 Y_1, Y_2, \dots, Y_n 相互独立的随机变量, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$$H_0: \beta = \beta_0 \longleftrightarrow H_1: \text{无约束}$$

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

$$f(y, \mathcal{X}; \beta^T, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left\{ -\frac{1}{2\sigma^2} (y - \beta^T x_i)^2 \right\}$$

$$L(y, \mathcal{X}; \beta^T, \sigma) = \prod_{i=1}^n f(y_i, \mathcal{X}; \beta^T, \sigma) = \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi} \sigma} \exp \left[-\frac{1}{2\sigma^2} (y_i - \beta^T x_i)^2 \right] \right\}$$

$$\begin{aligned}
 \ell(y, \mathcal{X}; \beta^T, \sigma) &= \log L(y, \mathcal{X}; \beta^T, \sigma) = -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^T x_i)^2 \\
 &= -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} (y - \mathcal{X}\beta)^T (y - \mathcal{X}\beta)
 \end{aligned}$$

$$\Rightarrow \hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T y, \quad \hat{\sigma}^2 = \frac{1}{n} \left\| y - \mathcal{X} \hat{\beta} \right\|^2$$

似然比检验 (Likelihood Ratio Test)

检验问题 4: 假设 Y_1, Y_2, \dots, Y_n 相互独立的随机变量, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$$H_0: \beta = \beta_0 \longleftrightarrow H_1: \text{无约束}$$

▶ 当 H_0 为真时, $\beta = \beta_0$, $\hat{\sigma}_0^2 = \frac{1}{n} \|y - \mathcal{X}\beta_0\|^2$

▶ 当 H_1 为真时, $\hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T y$, $\hat{\sigma}^2 = \frac{1}{n} \|y - \mathcal{X}\hat{\beta}\|^2$

$$-2 \log \lambda = 2 \left(\ell_1^* - \ell_0^* \right) = n \log \left(\frac{\|y - \mathcal{X}\beta_0\|^2}{\|y - \mathcal{X}\hat{\beta}\|^2} \right) \xrightarrow{\mathcal{L}} \chi_p^2, \quad n \rightarrow \infty$$

定理 7.1 (Wilks 定理) 如果 $\Omega_1 \subset \mathbb{R}^q$ 是一个 q 维空间, 而 $\Omega_0 \subset \Omega_1$ 是 r 的子空间, 则在正则条件下

$$\forall \theta \in \Omega_0: -2 \log \lambda \xrightarrow{\mathcal{L}} \chi_{q-r}^2 \quad \text{as } n \rightarrow \infty$$

$$= n \log(2\pi) + n \log \left(\frac{1}{n} \|y - \mathcal{X}\beta_0\|^2 \right) + \frac{n}{\|y - \mathcal{X}\beta_0\|^2} (y - \mathcal{X}\beta_0)^T (y - \mathcal{X}\beta_0)$$

$$- n \log(2\pi) - n \log \left(\frac{1}{n} \|y - \mathcal{X}\hat{\beta}\|^2 \right) - \frac{n}{\|y - \mathcal{X}\hat{\beta}\|^2} (y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})$$

似然比检验 (Likelihood Ratio Test)

检验问题 4: 假设 Y_1, Y_2, \dots, Y_n 相互独立的随机变量, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$$H_0: \beta = \beta_0 \longleftrightarrow H_1: \text{无约束}$$

- ▶ 当 H_0 为真时, $\beta = \beta_0$, $\hat{\sigma}_0^2 = \frac{1}{n} \|\mathbf{y} - \mathcal{X}\beta_0\|^2$
- ▶ 当 H_1 为真时, $\hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathbf{y}$, $\hat{\sigma}^2 = \frac{1}{n} \|\mathbf{y} - \mathcal{X}\hat{\beta}\|^2$

$$-2 \log \lambda = 2 \left(\ell_1^* - \ell_0^* \right) = n \log \left(\frac{\|\mathbf{y} - \mathcal{X}\beta_0\|^2}{\|\mathbf{y} - \mathcal{X}\hat{\beta}\|^2} \right) \xrightarrow{\mathcal{L}} \chi_p^2, \quad n \rightarrow \infty$$

- ▶ 还可以证明

$$F = \frac{n-p}{p} \cdot \left(\frac{\|\mathbf{y} - \mathcal{X}\beta_0\|^2}{\|\mathbf{y} - \mathcal{X}\hat{\beta}\|^2} - 1 \right) \sim F_{p, n-p}$$

似然比检验 (Likelihood Ratio Test)

- 例: 我们再来讨论“经典蓝色”帽衫数据.

```
x <- c(230, 125, 200, 109, 181, 99, 55, 107, 165, 97, 105, 98, 150, 115, 85, 71, 97, 120, 0, 82,
      192, 100, 150, 103, 181, 80, 85, 111, 189, 90, 120, 93, 172, 95, 110, 86, 170, 125, 130, 78)
pullover <- matrix(x, ncol = 4, byrow = TRUE)
pullover <- as.data.frame(pullover)
colnames(pullover) <- c("X1", "X2", "X3", "X4")
pullover
X1_2.lm <- lm(X1 ~ X2, data = pullover)
summary(X1_2.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(X1_2.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

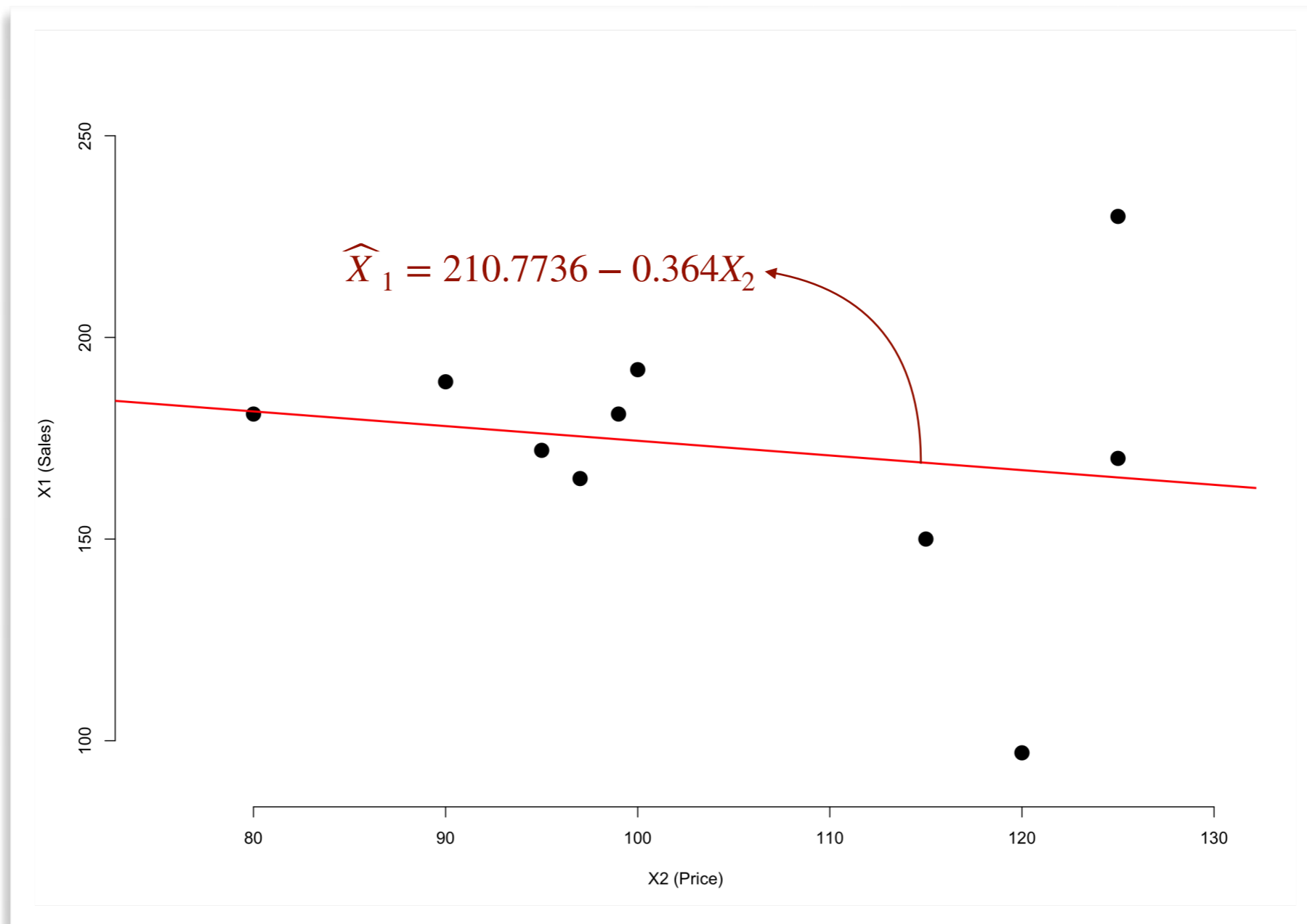
Multiple R-squared: 0.02808, Adjusted R-squared: -0.09341

F-statistic: 0.2311 on 1 and 8 DF, p-value: 0.6435

似然比检验 (Likelihood Ratio Test)

- 例: 我们再来讨论“经典蓝色”帽衫数据.

```
plot(pullover[,2:1], xlab='X2 (Price)', ylab='X1 (Sales)', axes=FALSE, pch=16, ylim=c(90, 250), xlim=c(75, 130))  
axis(1)  
axis(2)  
abline(X1_2.lm, col='red', lwd=2)
```



似然比检验 (Likelihood Ratio Test)

- 例: 我们再来讨论“经典蓝色”帽衫数据.

▶ 也许我们会问 $H_0: \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 211 \\ 0 \end{pmatrix}$.

▶ 似然比检验:

$$-2 \log \lambda = n \log \left(\frac{\|y - X\beta_0\|^2}{\|y - X\hat{\beta}\|^2} \right) = 9.098785$$

```
# LR test statistic
```

```
y <- as.matrix(pullover$X1)
```

```
X <- matrix(c(rep(1, 10), pullover$X2), ncol=2, byrow=FALSE)
```

```
beta_0 <- matrix(c(211, 0), nrow=2)
```

```
beta_hat <- as.matrix(X1_2.lm$coefficients)
```

```
n <- 10
```

```
(test.stat <- n * log(t(y - X %*% beta_0) %*% (y - X %*% beta_0) / t(y - X %*% beta_hat) %*% (y - X %*% beta_hat)))
```

```
> (test.stat <- n * log(t(y - X %*% beta_0) %*% (y - X %*% beta_0) / t(y - X %*% beta_hat) %*% (y - X %*% beta_hat)))
```

```
 [,1]
```

```
[1,] 9.098785
```

似然比检验 (Likelihood Ratio Test)

- 例: 我们再来讨论“经典蓝色”帽衫数据.

▶ 也许我们会问 $H_0: \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 211 \\ 0 \end{pmatrix}$.

▶ 似然比检验:

$$-2 \log \lambda = n \log \left(\frac{\|y - X\beta_0\|^2}{\|y - X\hat{\beta}\|^2} \right) = 9.098785$$

```
# p-value
```

```
pchisq(test.stat, 2, lower.tail = FALSE)
```

```
> pchisq(test.stat, 2, lower.tail = FALSE)
      [,1]
[1,] 0.01057362
```

- ▶ **结论:** 因为检验的 p 值 = 0.01057362 < 0.05, 我们拒绝 H_0 .

似然比检验 (Likelihood Ratio Test)

- 例: 我们再来讨论“经典蓝色”帽衫数据.

▶ 也许我们会问 $H_0: \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 211 \\ 0 \end{pmatrix}$.

▶ F 检验:

$$F = \frac{n-p}{p} \cdot \left(\frac{\|y - X\beta_0\|^2}{\|y - X\hat{\beta}\|^2} - 1 \right) = 5.936083$$

F test statistic

```
(ftest.stat <- ((n - 2)/2) * (t(y - X %*% beta_0) %*% (y - X %*% beta_0) / t(y - X %*% beta_hat) %*% (y - X %*% beta_hat) - 1))
```

```
> (fctest.stat <- ((n - 2)/2) * (t(y - X %*% beta_0) %*% (y - X %*% beta_0) / t(y - X %*% beta_hat) %*% (y - X %*% beta_hat) - 1))
      [,1]
[1,] 5.936083
```

p-value

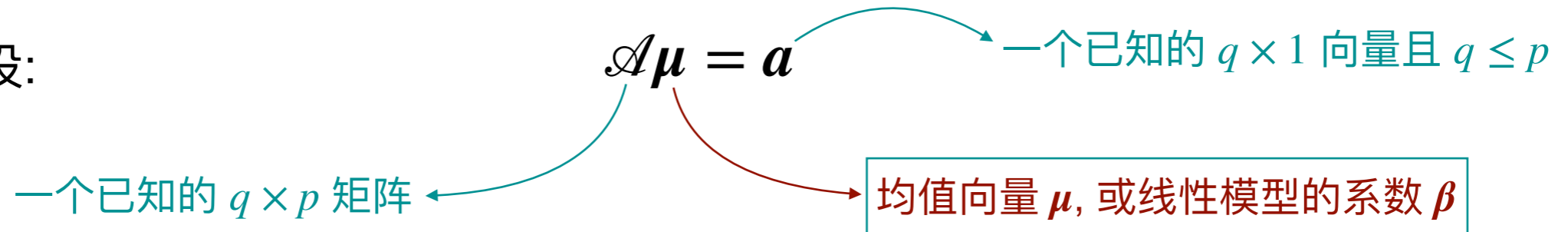
```
pf(ftest.stat, 2, n-2, lower.tail = FALSE)
```

```
> pf(ftest.stat, 2, n-2, lower.tail = FALSE)
      [,1]
[1,] 0.0262651
```

- ▶ 结论: 因为检验的 p 值 = 0.0262651 < 0.05, 我们拒绝 H_0 .

线性假设

- 线性假设:



▶ 设 $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$,

$$H_0: \mu_1 = \mu_2 \iff H_0: A\mu = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = 0 = a$$

线性假设

检验问题 5: 假设 X_1, X_2, \dots, X_n 是取自总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一个独立同分布的简单随机样本, 其中 $\boldsymbol{\Sigma}$ 已知.

$$H_0: \mathcal{A}\boldsymbol{\mu} = \boldsymbol{a} \iff H_1: \text{无约束}$$

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$$

$$\mathcal{A}\bar{\boldsymbol{x}} \sim N_q\left(\mathcal{A}\boldsymbol{\mu}, \frac{1}{n}\mathcal{A}\boldsymbol{\Sigma}\mathcal{A}^T\right)$$

定理 4.7 设 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则二次型变量 $U = (X - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (X - \boldsymbol{\mu})$ 服从 χ_p^2 分布.

$$(\mathcal{A}\bar{\boldsymbol{x}} - \mathcal{A}\boldsymbol{\mu})^T \left(\frac{1}{n}\mathcal{A}\boldsymbol{\Sigma}\mathcal{A}^T\right)^{-1} (\mathcal{A}\bar{\boldsymbol{x}} - \mathcal{A}\boldsymbol{\mu}) \sim \chi_q^2$$

$$n(\mathcal{A}\bar{\boldsymbol{x}} - \mathcal{A}\boldsymbol{\mu})^T (\mathcal{A}\boldsymbol{\Sigma}\mathcal{A}^T)^{-1} (\mathcal{A}\bar{\boldsymbol{x}} - \mathcal{A}\boldsymbol{\mu}) \sim \chi_q^2$$

- ▶ 当 H_0 为真时:

$$n(\mathcal{A}\bar{\boldsymbol{x}} - \boldsymbol{a})^T (\mathcal{A}\boldsymbol{\Sigma}\mathcal{A}^T)^{-1} (\mathcal{A}\bar{\boldsymbol{x}} - \boldsymbol{a}) \sim \chi_q^2$$

- ▶ 在给定的显著水平下, 如果该统计量的值很大则我们拒绝 H_0 .

线性假设

- 例: 考虑均值向量 $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ 分块的假设检验.

- ▶ 总体 $N_{2p} \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{pmatrix} \right)$, 其中 $\boldsymbol{\Sigma}$ 已知.

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \iff H_1: \text{无约束}$$

- ▶ 设

$$\mathcal{A} = \begin{pmatrix} \mathcal{I}_p & -\mathcal{I}_p \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{p \times 1}$$

$$\mathcal{A}\boldsymbol{\mu} = \mathbf{a} \iff \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

$$\implies n (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T (2\boldsymbol{\Sigma})^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \sim \chi_p^2$$

线性假设

- 例: 考虑均值向量 $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ 分块的假设检验.

- ▶ 总体 $N_{2p} \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Sigma} \end{pmatrix} \right)$, 其中 $\boldsymbol{\Sigma}$ 已知.

$$H_0: \boldsymbol{\mu}_1 = \mathbf{0} \iff H_1: \text{无约束}$$

- ▶ 设

$$\mathcal{A} = \begin{pmatrix} \mathcal{I}_p & \mathbf{O} \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{p \times 1}$$

$$\mathcal{A}\boldsymbol{\mu} = \mathbf{a} \iff \boldsymbol{\mu}_1 = \mathbf{0}$$

$$\implies n\bar{\mathbf{x}}_1^T \boldsymbol{\Sigma}^{-1} \bar{\mathbf{x}}_1 \sim \chi_p^2$$

线性假设

检验问题 6: 假设 X_1, X_2, \dots, X_n 是取自总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一个独立同分布的简单随机样本, 其中 $\boldsymbol{\Sigma}$ 未知.

$$H_0: \mathcal{A}\boldsymbol{\mu} = \boldsymbol{a} \longleftrightarrow H_1: \text{无约束}$$

推论 5.4 考虑 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一个线性变换 $Y = \mathcal{C}X$, 其中 $\mathcal{C}_{q \times p}$ 满足 $q \leq p$. 如果 $\bar{\boldsymbol{x}}$ 与 \mathcal{S}_X 分别为样本均值向量与样本协方差矩阵, 则有

$$\bar{\boldsymbol{y}} = \mathcal{C}\bar{\boldsymbol{x}} \sim N_q\left(\mathcal{C}\boldsymbol{\mu}, \frac{1}{n}\mathcal{C}\boldsymbol{\Sigma}\mathcal{C}^T\right)$$

$$n\mathcal{S}_Y = n\mathcal{C}\mathcal{S}_X\mathcal{C}^T \sim W_q(\mathcal{C}\boldsymbol{\Sigma}\mathcal{C}^T, n-1)$$

$$(n-1)(\mathcal{C}\bar{\boldsymbol{x}} - \mathcal{C}\boldsymbol{\mu})^T (\mathcal{C}\mathcal{S}_X\mathcal{C}^T)^{-1} (\mathcal{C}\bar{\boldsymbol{x}} - \mathcal{C}\boldsymbol{\mu}) \sim T_{q, n-1}^2$$

线性假设

检验问题 6: 假设 X_1, X_2, \dots, X_n 是取自总体 $N_p(\boldsymbol{\mu}, \Sigma)$ 的一个独立同分布的简单随机样本, 其中 Σ 未知.

$$H_0: \mathcal{A}\boldsymbol{\mu} = \mathbf{a} \iff H_1: \text{无约束}$$

▶ 当 H_0 为真时:

$$\mathcal{A}\bar{\mathbf{x}} \sim N_q\left(\mathbf{a}, \frac{1}{n}\mathcal{A}\Sigma\mathcal{A}^T\right)$$

$$n\mathcal{A}\mathcal{S}\mathcal{A}^T \sim W_q(\mathcal{A}\Sigma\mathcal{A}^T, n-1)$$

$$(n-1)(\mathcal{A}\bar{\mathbf{x}} - \mathbf{a})^T \left(\mathcal{A}\mathcal{S}\mathcal{A}^T\right)^{-1} (\mathcal{A}\bar{\mathbf{x}} - \mathbf{a}) \sim T_{q, n-1}^2$$

推论 5.4 考虑 $X \sim N_p(\boldsymbol{\mu}, \Sigma)$ 的一个线性变换 $Y = \mathcal{C}X$, 其中 $\mathcal{C}_{q \times p}$ 满足 $q \leq p$. 如果 $\bar{\mathbf{x}}$ 与 \mathcal{S}_X 分别为样本均值向量与样本协方差矩阵, 则有

$$\bar{\mathbf{y}} = \mathcal{C}\bar{\mathbf{x}} \sim N_q\left(\mathcal{C}\boldsymbol{\mu}, \frac{1}{n}\mathcal{C}\Sigma\mathcal{C}^T\right)$$

$$n\mathcal{S}_Y = n\mathcal{C}\mathcal{S}_X\mathcal{C}^T \sim W_q(\mathcal{C}\Sigma\mathcal{C}^T, n-1)$$

$$(n-1)(\bar{\mathbf{y}} - \mathcal{C}\boldsymbol{\mu})^T (\mathcal{C}\mathcal{S}_X\mathcal{C}^T)^{-1} (\bar{\mathbf{y}} - \mathcal{C}\boldsymbol{\mu}) \sim T_{q, n-1}^2$$

线性假设

- **例:** 我们再来讨论钞票数据集 (伪钞). 假设我们希望检验 $\mu_4 = \mu_5$.

▶ 此时有
$$\mathcal{A} = (0 \ 0 \ 0 \ 1 \ -1 \ 0), \quad \mathbf{a} = \mathbf{0}$$

▶ 检验统计量为
$$(n - 1)(\mathcal{A}\bar{\mathbf{x}})^T (\mathcal{A} \mathcal{S}_f \mathcal{A}^T)^{-1} (\mathcal{A}\bar{\mathbf{x}}) = 13.63797 \sim T_{1, 99}^2 = F_{1, 99}$$

```
# clear variables and close windows
```

```
rm(list = ls(all = TRUE))
```

```
library(mclust)
```

```
data(banknote)
```

```
banknote_counterfeit = subset(banknote, Status == 'counterfeit')[, 2:7]
```

```
x_bar <- as.matrix(sapply(banknote_counterfeit, mean)); x_bar
```

```
n <- dim(banknote_counterfeit)[1]
```

```
S_f <- (n-1) * cov(banknote_counterfeit) / n; S_f
```

```
A <- matrix(c(0, 0, 0, 1, -1, 0), nrow = 1, byrow = TRUE); A
```

```
# Test statistic
```

```
test.stat <- (n-1) * t(A %*% x_bar) %*% solve(A %*% S_f %*% t(A)) %*% (A %*% x_bar); test.stat
```

```
# p-value
```

```
pf(test.stat, 1, n-1, lower.tail = FALSE)
```

```

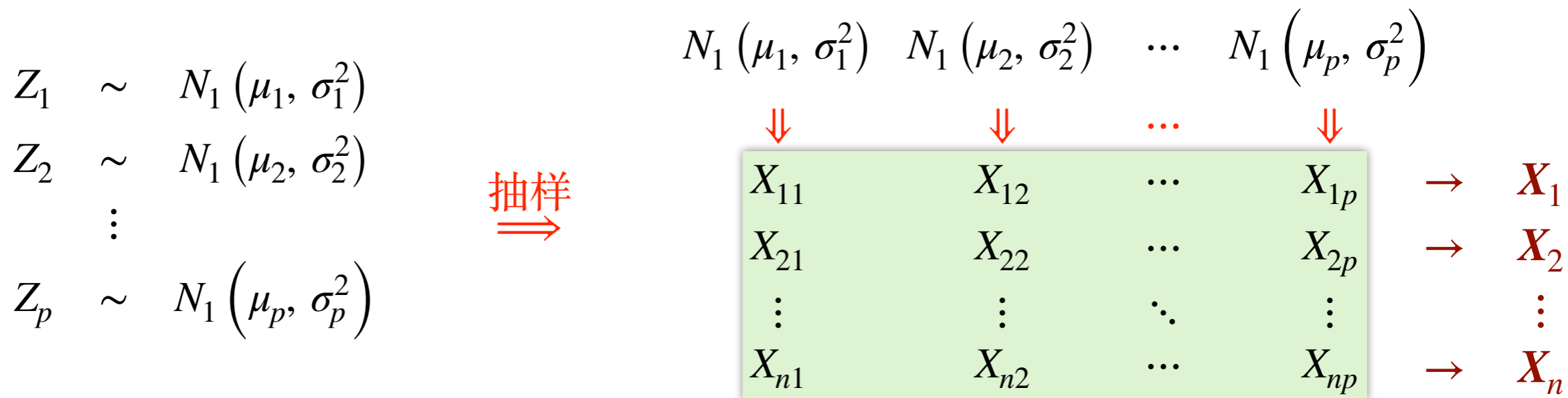
> # Test statistic
> test.stat <- (n-1) * t(A %*% x_bar) %*% solve(A %*% S_f %*% t(A)) %*% (A %*% x_bar); test.stat
      [,1]
[1,] 13.63797
> # p-value
> pf(test.stat, 1, n-1, lower.tail = FALSE)
      [,1]
[1,] 0.0003630897
  
```

- ▶ **结论:** 因为检验的 p 值 = 0.0003630897 < 0.05, 我们拒绝 $H_0 : \mu_4 = \mu_5$.

线性假设

- 重复观测

- ▶ 在很多情形中，我们会在 p 个不同时刻、或者 p 种不同试验条件 (不同处理方式等) 下进行抽样，分别得到容量为 n 的独立样本.



- ▶ 上述问题可看作如下情形： $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- ▶ 欲检验的假设为 $H_0: \mu_1 = \mu_2 = \dots = \mu_p$.

线性假设

- 重复观测

- ▶ 注意到在很多情形中，试验条件之一会受“控制” (安慰剂、标准药物或参考条件).

$$\mathcal{C} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}_{(p-1) \times p}$$

控制 (μ_1)

- ▶ 记 $\mathcal{C} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}_{(p-1) \times p}$

- ▶ 欲检验的假设为 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p. \iff H_0 : \mathcal{C}\boldsymbol{\mu} = \mathbf{0}$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

线性假设

- 重复观测

- ▶ 当 $H_0: \mathcal{C}\mu = \mathbf{0}$ 真时

$$(n-1) \bar{\mathbf{x}}^T \mathcal{C}^T \left(\mathcal{C} \mathcal{S} \mathcal{C}^T \right)^{-1} \mathcal{C} \bar{\mathbf{x}} \sim T_{p-1, n-1}^2$$

推论 5.4 考虑 $X \sim N_p(\mu, \Sigma)$ 的一个线性变换 $Y = \mathcal{C}X$, 其中 $\mathcal{C}_{q \times p}$ 满足 $q \leq p$. 如果

$\bar{\mathbf{x}}$ 与 \mathcal{S}_X 分别为样本均值向量与样本协方差矩阵, 则有

$$\bar{\mathbf{y}} = \mathcal{C} \bar{\mathbf{x}} \sim N_q \left(\mathcal{C} \mu, \frac{1}{n} \mathcal{C} \Sigma \mathcal{C}^T \right)$$

$$n \mathcal{S}_Y = n \mathcal{C} \mathcal{S}_X \mathcal{C}^T \sim W_q(\mathcal{C} \Sigma \mathcal{C}^T, n-1)$$

$$(n-1)(\mathcal{C} \bar{\mathbf{x}} - \mathcal{C} \mu)^T (\mathcal{C} \mathcal{S}_X \mathcal{C}^T)^{-1} (\mathcal{C} \bar{\mathbf{x}} - \mathcal{C} \mu) \sim T_{q, n-1}^2$$

定理 5.9 Hotelling T^2 与 F 分布的关系为 $T_{p, n}^2 = \frac{np}{n-p+1} F_{p, n-p+1}$.

$$\frac{(n-1) - (p-1) + 1}{(n-1)(p-1)} \cdot (n-1) \bar{\mathbf{x}}^T \mathcal{C}^T \left(\mathcal{C} \mathcal{S} \mathcal{C}^T \right)^{-1} \mathcal{C} \bar{\mathbf{x}} \sim F_{p-1, n-p+1}$$

$$\Rightarrow \frac{n-p+1}{p-1} \bar{\mathbf{x}}^T \mathcal{C}^T \left(\mathcal{C} \mathcal{S} \mathcal{C}^T \right)^{-1} \mathcal{C} \bar{\mathbf{x}} \sim F_{p-1, n-p+1}$$

- ▶ 零假设 $H_0: \mathcal{C}\mu = \mathbf{0}$ 的拒绝域为

$$\frac{n-p+1}{p-1} \bar{\mathbf{x}}^T \mathcal{C}^T \left(\mathcal{C} \mathcal{S} \mathcal{C}^T \right)^{-1} \mathcal{C} \bar{\mathbf{x}} > F_{p-1, n-p+1}(\alpha)$$

上 α 分位数

线性假设

- 重复观测

- ▶ 记 $Y_i = \mathcal{C}_{(p-1) \times p} X_i$, 则 $Y_i \sim N_{p-1}(\mathcal{C}\mu, \mathcal{C}\Sigma\mathcal{C}^T)$. $X_i \sim N_p(\mu, \Sigma)$

- μ 的置信域 (confidence region)

- ▶ 于是, 对所有 $a \in \mathbb{R}^p$, 下述区间同时以概率 $1 - \alpha$ 包含 $a^T \mu$ 在内

$$\left(a^T \bar{x} - \sqrt{K_\alpha \cdot a^T S a}, \quad a^T \bar{x} + \sqrt{K_\alpha \cdot a^T S a} \right)$$

$K_\alpha = \frac{p}{n-p} \cdot F_{p, n-p}(\alpha)$

- ▶ 对所有 $a \in \mathbb{R}^{p-1}$, 下述联合置信区间将以概率 $1 - \alpha$ 包含 $a^T \mathcal{C}\mu$

$$\left(a^T \mathcal{C}\bar{x} - \sqrt{\frac{p-1}{n-p+1} F_{p-1, n-p+1}(\alpha) \cdot a^T \mathcal{C} S \mathcal{C}^T a}, \quad a^T \mathcal{C}\bar{x} + \sqrt{\frac{p-1}{n-p+1} F_{p-1, n-p+1}(\alpha) \cdot a^T \mathcal{C} S \mathcal{C}^T a} \right)$$

线性假设

- 重复观测

$$\Rightarrow \mathcal{C}\mathbf{1}_p = \mathbf{0}$$

- ▶ 看看该问题的性质

行元素之和为零

$$\mathcal{C} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}_{(p-1) \times p}, \mathcal{C} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}_{(p-1) \times p}$$

- ▶ 若 $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{p-1} \end{pmatrix}$, 则 $\mathbf{a}^T \mathcal{C} = (a_1, a_2, \dots, a_{p-1}) \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$

元素之和为零

$$= (a_1, a_2 - a_1, a_3 - a_2, \dots, a_{p-1} - a_{p-2}, -a_{p-1})$$

- ▶ 这表明如果我们记 $\mathbf{b} = \mathcal{C}^T \mathbf{a} = (\mathbf{a}^T \mathcal{C})^T$, 则有 $\mathbf{b}^T \mathbf{1}_p = \sum_{j=1}^p b_j = 0$.

线性假设

- 重复观测

▶ 于是，在置信度为 $1 - \alpha$ 时，对 $\boldsymbol{\mu}$ 的所有对比， $\boldsymbol{b}^T \boldsymbol{\mu}$ 的联合置信区间为

$$\left(\boldsymbol{b}^T \bar{\boldsymbol{x}} - \sqrt{\frac{p-1}{n-p+1} F_{p-1, n-p+1}(\alpha) \cdot \boldsymbol{b}^T \boldsymbol{S} \boldsymbol{b}}, \boldsymbol{b}^T \bar{\boldsymbol{x}} + \sqrt{\frac{p-1}{n-p+1} F_{p-1, n-p+1}(\alpha) \cdot \boldsymbol{b}^T \boldsymbol{S} \boldsymbol{b}} \right)$$

线性假设

- 重复观测
- **例:** Bock (1975) 研究了从八年级到十一年级儿童词汇的演变. 该数据集包含随机选择的 40 名儿童的词汇测试成绩.
 - ▶ 由于对同一个孩子的八年级至十一年级进行了观测, 所以这是重复观测的情形, $n = 40, p = 4$.
 - ▶ 检验统计量为

$$\bar{\mathbf{x}} = \begin{pmatrix} 1.086 \\ 2.544 \\ 2.851 \\ 3.420 \end{pmatrix}, \quad \mathcal{S} = \begin{pmatrix} 2.902 & 2.438 & 2.963 & 2.183 \\ 2.438 & 3.049 & 2.775 & 2.319 \\ 2.963 & 2.775 & 4.281 & 2.939 \\ 2.183 & 2.319 & 2.939 & 3.162 \end{pmatrix}$$

线性假设

- 重复观测
- **例:** Bock (1975) 研究了从八年级到十一年级儿童词汇的演变. 该数据集包含随机选择的 40 名儿童的词汇测试成绩.

▶ 假设我们对孩子们每年的进步情况感兴趣, 那么给出 μ_j 的相邻差的矩阵 \mathcal{C} 为

$$\begin{pmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \\ \mu_3 - \mu_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad \mathcal{C} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

▶ 对 $H_0: \mathcal{C}\boldsymbol{\mu} = \mathbf{0}$, 我们有 $\frac{n-p+1}{p-1} \bar{\mathbf{x}}^T \mathcal{C}^T (\mathcal{C} \mathcal{S} \mathcal{C}^T)^{-1} \mathcal{C} \bar{\mathbf{x}} \sim F_{p-1, n-p+1}$

▶ 零假设 $H_0: \mathcal{C}\boldsymbol{\mu} = \mathbf{0}$ 的拒绝域为

$$\frac{n-p+1}{p-1} \bar{\mathbf{x}}^T \mathcal{C}^T (\mathcal{C} \mathcal{S} \mathcal{C}^T)^{-1} \mathcal{C} \bar{\mathbf{x}} > F_{p-1, n-p+1}(\alpha)$$

线性假设

- 重复观测
- **例:** Bock (1975) 研究了从八年级到十一年级儿童词汇的演变. 该数据集包含随机选择的 40 名儿童的词汇测试成绩.

Example: Bock

```
rm(list = ls(all = TRUE))
n <- 40
p <- 4
x_bar <- matrix(c(1.086, 2.544, 2.851, 3.420), nrow = 4)
x <- c(2.902, 2.438, 2.963, 2.183, 2.438, 3.049, 2.775, 2.319, 2.963, 2.775, 4.281, 2.939, 2.183, 2.319, 2.939, 3.162)
S <- matrix(x, nrow = 4, byrow = TRUE)
C <- matrix(c(1, -1, 0, 0, 0, 1, -1, 0, 0, 0, 1, -1), nrow = 3, byrow = TRUE)
(F_obs <- ((n-p-1) / (p-1)) * t(x_bar) %*% t(C) %*% solve(C %*% S %*% t(C)) %*% C %*% x_bar)
(p_value <- pf(F_obs, p-1, n-p+1, lower.tail = FALSE))
```

```
> (F_obs <- ((n-p-1) / (p-1)) * t(x_bar) %*% t(C) %*% solve(C %*% S %*% t(C)) %*% C %*% x_bar)
      [,1]
[1,] 50.29575
> (p_value <- pf(F_obs, p-1, n-p+1, lower.tail = FALSE))
      [,1]
[1,] 3.930106e-13
```

- ▶ **结论:** 相邻的均值之间存在显著差异.

线性假设

- 重复观测
- **例:** Bock (1975) 研究了从八年级到十一年级儿童词汇的演变. 该数据集包含随机选择的 40 名儿童的词汇测试成绩.

▶ 对 μ 的所有对比, $\mathbf{b}^T \mu$ 的联合置信区间为

$$\left(\mathbf{b}^T \bar{\mathbf{x}} - \sqrt{\frac{p-1}{n-p+1} F_{p-1, n-p+1}(\alpha) \cdot \mathbf{b}^T \mathcal{S} \mathbf{b}}, \mathbf{b}^T \bar{\mathbf{x}} + \sqrt{\frac{p-1}{n-p+1} F_{p-1, n-p+1}(\alpha) \cdot \mathbf{b}^T \mathcal{S} \mathbf{b}} \right)$$

▶ 取 $\mathbf{b}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}$, $\mathbf{b}_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \end{pmatrix}$, $\mathbf{b}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}$

```

> b1 <- matrix(c(1, -1, 0, 0), nrow=4)
> (t(b1) %*% x_bar - sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b1) %*% S %*% b1))
[1,] -1.957178
> (t(b1) %*% x_bar + sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b1) %*% S %*% b1))
[1,] -0.9588218
>
> b2 <- matrix(c(0, 1, -1, 0), nrow=4)
> (t(b2) %*% x_bar - sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b2) %*% S %*% b2))
[1,] -0.9493347
> (t(b2) %*% x_bar + sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b2) %*% S %*% b2))
[1,] 0.3353347
>
> b3 <- matrix(c(0, 0, 1, -1), nrow=4)
> (t(b3) %*% x_bar - sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b3) %*% S %*% b3))
[1,] -1.171294
> (t(b3) %*% x_bar + sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b3) %*% S %*% b3))
[1,] 0.03329398
    
```

```

b1 <- matrix(c(1, -1, 0, 0), nrow=4)
(t(b1) %*% x_bar - sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b1) %*% S %*% b1))
(t(b1) %*% x_bar + sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b1) %*% S %*% b1))
b2 <- matrix(c(0, 1, -1, 0), nrow=4)
(t(b2) %*% x_bar - sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b2) %*% S %*% b2))
(t(b2) %*% x_bar + sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b2) %*% S %*% b2))
b3 <- matrix(c(0, 0, 1, -1), nrow=4)
(t(b3) %*% x_bar - sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b3) %*% S %*% b3))
(t(b3) %*% x_bar + sqrt((p-1)/(n-p+1) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b3) %*% S %*% b3))
    
```

线性假设

- 重复观测
- **例:** Bock (1975) 研究了从八年级到十一年级儿童词汇的演变. 该数据集包含随机选择的 40 名儿童的词汇测试成绩.

▶ 对 μ 的所有对比, $\mathbf{b}^T \mu$ 的联合置信区间为

$$\left(\mathbf{b}^T \bar{\mathbf{x}} - \sqrt{\frac{p-1}{n-p+1} F_{p-1, n-p+1}(\alpha) \cdot \mathbf{b}^T \mathcal{S} \mathbf{b}}, \mathbf{b}^T \bar{\mathbf{x}} + \sqrt{\frac{p-1}{n-p+1} F_{p-1, n-p+1}(\alpha) \cdot \mathbf{b}^T \mathcal{S} \mathbf{b}} \right)$$

▶ 取 $\mathbf{b}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}$, $\mathbf{b}_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \end{pmatrix}$, $\mathbf{b}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix} \Rightarrow \begin{cases} -1.957 \leq \mu_1 - \mu_2 \leq -0.959 \\ -0.949 \leq \mu_2 - \mu_3 \leq 0.335 \\ -1.171 \leq \mu_3 - \mu_4 \leq 0.033 \end{cases}$

▶ 因此, 拒绝 $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ 的主要原因在于儿童们在八年级和九年级的表现有差异.

线性假设

- 重复观测
- **例:** Bock (1975) 研究了从八年级到十一年级儿童词汇的演变. 该数据集包含随机选择的 40 名儿童的词汇测试成绩.

▶ 若取 $b_1 = \begin{pmatrix} 1 \\ -\frac{1}{3} \\ -\frac{1}{3} \\ -\frac{1}{3} \end{pmatrix}$, $b_2 = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ -1 \end{pmatrix}$, $b_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix}$

```

> b4 <- matrix(c(1, -1/3, -1/3, -1/3), nrow=4)
> (t(b4) %*% x_bar - sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b4) %*% S %*% b4)
[1,] -2.282117
> (t(b4) %*% x_bar + sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b4) %*% S %*% b4)
[1,] -1.422549
>
> b5 <- matrix(c(1/3, 1/3, 1/3, -1), nrow=4)
> (t(b5) %*% x_bar - sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b5) %*% S %*% b5)
[1,] -1.777111
> (t(b5) %*% x_bar + sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b5) %*% S %*% b5)
[1,] -0.742228
>
> b6 <- matrix(c(0, 1, 0, -1), nrow=4)
> (t(b6) %*% x_bar - sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b6) %*% S %*% b6)
[1,] -1.479831
> (t(b6) %*% x_bar + sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b6) %*% S %*% b6)
[1,] -0.2721686
  
```

```

b4 <- matrix(c(1, -1/3, -1/3, -1/3), nrow=4)
(t(b4) %*% x_bar - sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b4) %*% S %*% b4)
(t(b4) %*% x_bar + sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b4) %*% S %*% b4)
b5 <- matrix(c(1/3, 1/3, 1/3, -1), nrow=4)
(t(b5) %*% x_bar - sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b5) %*% S %*% b5)
(t(b5) %*% x_bar + sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b5) %*% S %*% b5)
b6 <- matrix(c(0, 1, 0, -1), nrow=4)
(t(b6) %*% x_bar - sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b6) %*% S %*% b6)
(t(b6) %*% x_bar + sqrt((p-1)/(n-p+1)) * qf(0.05, p-1, n-p+1, lower.tail = FALSE) * t(b6) %*% S %*% b6)
  
```

线性假设

- 重复观测
- **例:** Bock (1975) 研究了从八年级到十一年级儿童词汇的演变. 该数据集包含随机选择的 40 名儿童的词汇测试成绩.

$$\text{▶ 若取 } \mathbf{b}_1 = \begin{pmatrix} 1 \\ -\frac{1}{3} \\ -\frac{1}{3} \\ -\frac{1}{3} \end{pmatrix}, \quad \mathbf{b}_2 = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ -1 \end{pmatrix}, \quad \mathbf{b}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix} \Rightarrow \begin{cases} -2.282 \leq \mu_1 - \frac{1}{3}(\mu_2 + \mu_3 + \mu_4) \leq -1.423 \\ -1.777 \leq \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \mu_4 \leq -0.742 \\ -1.480 \leq \mu_2 - \mu_4 \leq -0.272 \end{cases}$$

- ▶ 因此, μ_1 与其它三个年级的平均值不等.
- ▶ μ_4 也与其它三个年级的平均值不等.
- ▶ μ_4 的值要大于 μ_2 的值.

线性假设

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$$H_0: \mathcal{A}\beta = a \iff H_1: \text{无约束}$$

- ▶ 当 H_0 真时, 带有约束条件的极大似然函数为

$$f(\beta, \lambda) = (y - x\beta)^T (y - x\beta) - \lambda^T (\mathcal{A}\beta - a), \quad \lambda \in \mathbb{R}^q$$

- ▶ 求解

$$\begin{cases} \frac{\partial f(\beta, \lambda)}{\partial \beta} = 0 \\ \frac{\partial f(\beta, \lambda)}{\partial \lambda} = 0 \end{cases} \implies \tilde{\beta} = \hat{\beta} - (x^T x)^{-1} \mathcal{A}^T \left[\mathcal{A} (x^T x)^{-1} \mathcal{A}^T \right]^{-1} (\mathcal{A} \hat{\beta} - a)$$

无约束的极大似然估计量 $\hat{\beta} = (x^T x)^{-1} x^T y$

$$\tilde{\sigma}^2 = \frac{1}{n} (y - x\tilde{\beta})^T (y - x\tilde{\beta})$$

线性假设

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$$H_0: \mathcal{A}\beta = a \iff H_1: \text{无约束}$$

▶ 于是, 似然比统计量为

$$\begin{aligned}
 -2 \log \lambda &= 2 \left(\ell_1^* - \ell_0^* \right) \\
 &= n \log \left(\frac{\|y - \mathcal{X}\tilde{\beta}\|^2}{\|y - \mathcal{X}\hat{\beta}\|^2} \right) \\
 &\xrightarrow{\mathcal{L}} \chi_q^2
 \end{aligned}$$

q 是 a 中元素的个数

▶ 则 H_0 的似然比检验的拒绝域为

$$R = \left\{ \mathcal{X} : -2 \log \lambda > \chi_q^2(\alpha) \right\}$$

$\chi_q^2(\alpha)$ 为上 α 分位数

线性假设

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$$H_0: \mathcal{A}\beta = a \iff H_1: \text{无约束}$$

- ▶ 利用检验问题 4 中我们曾用到的结果, 该问题也有精确的 F 检验

$$F = \frac{n-p}{q} \left(\frac{\|y - \mathcal{X}\tilde{\beta}\|^2}{\|y - \mathcal{X}\hat{\beta}\|^2} - 1 \right) = \frac{n-p}{q} \frac{(\mathcal{A}\hat{\beta} - a)^T [\mathcal{A}(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{A}^T]^{-1} (\mathcal{A}\hat{\beta} - a)}{(y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})} \sim F_{q, n-p}$$

- ▶ 此时 H_0 的拒绝域为

$$R = \left\{ \mathcal{X} : F > F_{q, n-p}(\alpha) \right\}$$

$F_{q, n-p}(\alpha)$ 是上 α 分位数

线性假设

- **例:** 我们再来讨论“经典蓝”帽衫数据.

```
# regression of sales X1 on price X2
summary(X1_2.lm)
```

$$X_1 = \alpha + \beta X_2 = 210.7736 - 0.3640X_2$$

- ▶ 对 $H_0: \beta = 0$, F 统计量的值为

$$F = 0.2311$$

- ▶ **结论:** 我们可以认为销量与价格 (一个变量) 独立.

- ▶ 但应当意识到, 如果我们将价格与广告成本 (X3) 和促销员工时 (X4) 一起考虑, 则必须修正这一结论.

```
> summary(X1_2.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
```

线性假设

- **例:** 我们再来讨论“经典蓝”帽衫数据.
- ▶ 在**检验问题 4** 中, 我们对 $H_0: \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 211 \\ 0 \end{pmatrix}$ 进行了检验并拒绝了 H_0 .
- ▶ 之所以拒绝是因为这一对值均被拒绝. 事实上, 如果 $\beta = 0$, 则 α 的估计量为 $\bar{y} = 172.70$, 这与 211 相去甚远.
- ▶ 现在我们来考虑帽衫数据的多元回归.

```
#### Classic pullover data, regression of sales X1 on X2, X3, and X4
```


```
X1_234 <- lm(X1 ~ X2 + X3 + X4, data = pullover)
```

```
summary(X1_234)
```

线性假设

- 例: 我们再来讨论“经典蓝”帽衫数据.

$$\begin{aligned}
 X_1 &= \alpha + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_4 + \varepsilon \\
 &= 65.670 - 0.216X_2 + 0.485X_3 + 0.844X_4 + \varepsilon
 \end{aligned}$$



$$\beta_1 \approx -\frac{1}{2}\beta_2$$

- 在实践中, 这意味着若将价格提高 20 欧元, 为保持套头衫的销售数量不变, 则需要广告成本增加 10 欧元.

```
> summary(X1_234)
```

```
Call:
lm(formula = X1 ~ X2 + X3 + X4, 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
```

- 现在我们来检验 $H_0: \beta_1 = -\frac{1}{2}\beta_2$. $\iff \begin{pmatrix} 0 & 1 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = 0$

线性假设

- **例:** 我们再来讨论“经典蓝”帽衫数据.
- ▶ 此时, 似然比统计量的值为

$$-2 \log \lambda = n \log \left(\frac{\| \mathbf{y} - \mathcal{X} \tilde{\boldsymbol{\beta}} \|^2}{\| \mathbf{y} - \mathcal{X} \hat{\boldsymbol{\beta}} \|^2} \right) = 0.01237415$$

Likelihood Ratio statistic

```

n <- 10
A <- matrix(c(0, 1, 1/2, 0), nrow = 1)
X <- as.matrix(cbind(X_1 = rep(1, 10), pullover[, 2:4]))
beta_hat <- as.matrix(coefficients(X1_234))
beta_tilde <- beta_hat - solve(t(X) %*% X) %*% t(A) %*% solve(A %*% solve(t(X) %*% X) %*% t(A)) %*% A %*% beta_hat
LR <- n * log(sum((pullover$X1 - X %*% beta_tilde)^2) / sum((pullover$X1 - X %*% beta_hat)^2))
LR
  
```

```

> LR
[1] 0.01237415
  
```

- ▶ 现在我们来检验 $H_0 : \beta_1 = -\frac{1}{2}\beta_2$. $\iff \begin{pmatrix} 0, 1, \frac{1}{2}, 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = 0$

线性假设

- **例:** 我们再来讨论“经典蓝”帽衫数据.
- ▶ 此时, 似然比统计量的值为

$$-2 \log \lambda = n \log \left(\frac{\| \mathbf{y} - \mathcal{X} \tilde{\boldsymbol{\beta}} \|^2}{\| \mathbf{y} - \mathcal{X} \hat{\boldsymbol{\beta}} \|^2} \right) = 0.01237415$$

```
# p_value
```

```
q <- 4
```

```
p_value <- pchisq(LR, q, lower.tail = FALSE)
```

```
p_value
```

```
> p_value  
[1] 0.9999809
```

- ▶ **结论:** 接受 H_0 .

- ▶ 现在我们来检验 $H_0: \beta_1 = -\frac{1}{2}\beta_2$. $\iff \begin{pmatrix} 0 & 1 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = 0$

线性假设

- **例:** 我们再来讨论“经典蓝”帽衫数据.

- ▶ F 统计量的值为

$$F = \frac{n-p}{q} \frac{(\mathcal{A}\hat{\beta} - a)^T [\mathcal{A}(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{A}^T]^{-1} (\mathcal{A}\hat{\beta} - a)}{(y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})} = 0.002166816$$

F statistic

p <- 3

```
F <- ((n-p)/q) * (t(A %*% beta_hat) %*% solve(A %*% solve(t(X) %*% X) %*% t(A)) %*% A %*% beta_hat) / (sum((pullover$X1 - X %*% beta_hat)^2))
```

F

```
> F
      [,1]
[1,] 0.002166816
```

- ▶ 现在我们来检验 $H_0: \beta_1 = -\frac{1}{2}\beta_2$. $\iff \begin{pmatrix} 0 & 1 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = 0$

线性假设

- **例:** 我们再来讨论“经典蓝”帽衫数据.

- ▶ F 统计量的值为

$$F = \frac{n-p}{q} \frac{\left(\mathcal{A}\hat{\beta} - a\right)^T \left[\mathcal{A} \left(\mathcal{X}^T \mathcal{X}\right)^{-1} \mathcal{A}^T\right]^{-1} \left(\mathcal{A}\hat{\beta} - a\right)}{\left(y - \mathcal{X}\hat{\beta}\right)^T \left(y - \mathcal{X}\hat{\beta}\right)} = 0.002166816$$

```
# p_value
```

```
p_value <- pf(F, q, n-p, lower.tail = FALSE)
```

```
p_value
```

```
> p_value
```

```
      [,1]
```

```
[1,] 0.999988
```

- ▶ **结论:** 接受 H_0 .

- ▶ 现在我们来检验 $H_0 : \beta_1 = -\frac{1}{2}\beta_2$. $\iff \begin{pmatrix} 0 & 1 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = 0$

线性假设

- 两个均值向量的比较

检验问题 8: 假设 $X_{i1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, $j = 1, 2, \dots, n_2$, 其中所有的变量相互独立.

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \iff H_1: \text{无约束}$$

- 记 $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, 则有

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \sim N_p\left(\boldsymbol{\delta}, \frac{n_1 + n_2}{n_1 n_2} \boldsymbol{\Sigma}\right) \implies \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \boldsymbol{\delta} \right] \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$$

$$n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2 \sim W_p(\boldsymbol{\Sigma}, n_1 + n_2 - 2) \implies (n_1 + n_2) \mathcal{S} \sim W_p(\boldsymbol{\Sigma}, n_1 + n_2 - 2)$$

- 用 $\mathcal{S} = \frac{1}{n_1 + n_2} (n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2)$ 表示 \mathcal{S}_1 与 \mathcal{S}_2 的加权平均.

定理 5.8 如果 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 与 $\mathcal{A} \sim W_p(\boldsymbol{\Sigma}, n)$ 相互独立, 则

$$n(X - \boldsymbol{\mu})^T \mathcal{A}^{-1} (X - \boldsymbol{\mu}) \sim T_{p, n}^2$$

$$(n_1 + n_2 - 2) \left\{ \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \boldsymbol{\delta} \right] \right\}^T \left[(n_1 + n_2) \mathcal{S} \right]^{-1} \left\{ \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \boldsymbol{\delta} \right] \right\} \sim T_{p, n_1 + n_2 - 2}^2$$

$$\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)^2} \left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \boldsymbol{\delta} \right]^T \mathcal{S}^{-1} \left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \boldsymbol{\delta} \right] \sim T_{p, n_1 + n_2 - 2}^2$$

线性假设

- 两个均值向量的比较

检验问题 8: 假设 $X_{i1} \sim N_p(\mu_1, \Sigma)$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\mu_2, \Sigma)$, $j = 1, 2, \dots, n_2$, 其中所有的变量相互独立.

$$H_0 : \mu_1 = \mu_2 \longleftrightarrow H_1 : \text{无约束}$$

定理 5.9 Hotelling T^2 与 F 分布的关系为 $T_{p,n}^2 = \frac{np}{n-p+1} F_{p, n-p+1}$.

$$\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)^2} \left[(\bar{x}_1 - \bar{x}_2) - \delta \right]^T \mathcal{S}^{-1} \left[(\bar{x}_1 - \bar{x}_2) - \delta \right] \sim \frac{(n_1 + n_2 - 2) p}{n_1 + n_2 - 2 - p + 1} F_{p, n_1 + n_2 - 2 - p + 1}$$

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} \left[(\bar{x}_1 - \bar{x}_2) - \delta \right]^T \mathcal{S}^{-1} \left[(\bar{x}_1 - \bar{x}_2) - \delta \right] \sim F_{p, n_1 + n_2 - p - 1}$$

- ▶ 当 $H_0 : \delta = \mathbf{0}$ 真时, 则有

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^T \mathcal{S}^{-1} (\bar{x}_1 - \bar{x}_2) \sim F_{p, n_1 + n_2 - p - 1}$$

$$\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)^2} \left[(\bar{x}_1 - \bar{x}_2) - \delta \right]^T \mathcal{S}^{-1} \left[(\bar{x}_1 - \bar{x}_2) - \delta \right] \sim T_{p, n_1 + n_2 - 2}^2$$

线性假设

- 两个均值向量的比较

检验问题 8: 假设 $X_{i1} \sim N_p(\mu_1, \Sigma)$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\mu_2, \Sigma)$, $j = 1, 2, \dots, n_2$, 其中所有的变量相互独立.

$$H_0: \mu_1 = \mu_2 \longleftrightarrow H_1: \text{无约束}$$

- ▶ $H_0: \mu_1 = \mu_2$ 的拒绝域为

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^T \mathcal{S}^{-1} (\bar{x}_1 - \bar{x}_2) \geq F_{p, n_1 + n_2 - p - 1}(\alpha)$$

上 α 分位数

- ▶ 置信度为 $1 - \alpha$ 时, $\delta = \mu_1 - \mu_2$ 的置信域是以 $(\bar{x}_1 - \bar{x}_2)$ 为中心的椭球

$$\left[\delta - (\bar{x}_1 - \bar{x}_2) \right]^T \mathcal{S}^{-1} \left[\delta - (\bar{x}_1 - \bar{x}_2) \right] \leq \frac{p (n_1 + n_2)^2}{n_1 n_2 (n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha)$$

线性假设

- 两个均值向量的比较

检验问题 8: 假设 $X_{i1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, $j = 1, 2, \dots, n_2$, 其中所有的变量相互独立.

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \iff H_1: \text{无约束}$$

- ▶ $\boldsymbol{\delta}$ 的元素的所有线性组合 $\mathbf{a}^T \boldsymbol{\delta}$ 的联合置信区间为

$$\mathbf{a}^T \boldsymbol{\delta} \in \mathbf{a}^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \pm \sqrt{\frac{p(n_1 + n_2)^2}{n_1 n_2 (n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha)} \mathbf{a}^T \mathcal{S} \mathbf{a}$$

- ▶ 特别, 对 $j = 1, 2, \dots, p$, 置信度为 $1 - \alpha$ 时, 我们有

$$\delta_j = \mu_{1j} - \mu_{2j} \in (\bar{x}_{1j} - \bar{x}_{2j}) \pm \sqrt{\frac{p(n_1 + n_2)^2}{n_1 n_2 (n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha)} \cdot s_{jj}$$

线性假设

- 例:** 再来考虑美国公司数据集. 对于能源行业 (group 1) 和制造业 (group 2), 我们来比较其资产 (X_1) 和销量 (X_2) 的均值.

```
# energy sector data
```

```
x = rbind(c(13621, 4848, 4572, 485, 898.9, 23.4), c(1117, 1038, 478, 59.7, 91.7, 3.8),
          c(1633, 701, 679, 74.3, 135.9, 2.8), c(5651, 1254, 2002, 310.7, 407.9, 6.2),
          c(5835, 4053, 1601, -93.8, 173.8, 10.8), c(3494, 1653, 1442, 160.9, 320.3, 6.4),
          c(1654, 451, 779, 84.8, 130.4, 1.6), c(1679, 1354, 687, 93.8, 154.6, 4.6),
          c(1257, 355, 181, 167.5, 304, 0.6), c(1743, 597, 717, 121.6, 172.4, 3.5),
          c(1440, 1617, 639, 81.7, 126.4, 3.5), c(14045, 15636, 2754, 418, 1462, 27.3),
          c(3010, 749, 1120, 146.3, 209.2, 3.4), c(3086, 1739, 1507, 202.7, 335.2, 4.9),
          c(1995, 2662, 341, 34.7, 100.7, 2.3))
```

```
n1 <- dim(x)[1]; p <- 2
```

```
# mean vector for energy sector
```

```
(x_mean <- as.matrix(apply(x[, 1:2], 2, mean)))
```

```
# MLE/empirical variance matrix for energy sector
```

```
(S_1 = cov(x[, 1:2]) * (n1-1)/(n1))
```

```
> (x_mean <- as.matrix(apply(x[, 1:2], 2, mean)))
```

		[,1]	
[1,]	4084.000		
[2,]	2580.467		

 $\Rightarrow \bar{x}_1 = \begin{pmatrix} 4084.000 \\ 2580.467 \end{pmatrix}$

```
> (S_1 = cov(x[, 1:2]) * (n1-1)/(n1))
```

		[,1]	[,2]
[1,]	16634749	12409637	
[2,]	12409637	13747417	



$$S_1 = \begin{pmatrix} 16634749 & 12409637 \\ 12409637 & 13747417 \end{pmatrix}$$

线性假设

- 例:** 再来考虑美国公司数据集. 对于能源行业 (group 1) 和制造业 (group 2), 我们来比较其资产 (X_1) 和销量 (X_2) 的均值.

```
# manufacturing sector data
```

```
y = rbind(c(1093, 1679, 1070, 100.9, 164.5, 20.8), c(1128, 1516, 430, -47, 26.7, 13.2),
          c(1804, 2564, 483, 70.5, 164.9, 26.6), c(4662, 4781, 2988, 28.7, 371.5, 66.2),
          c(6307, 8199, 598, -771.5, -524.3, 57.5), c(2366, 3305, 1117, 131.2, 256.5, 25.2),
          c(4084, 4346, 3023, 302.7, 521.7, 37.5), c(10348, 5721, 1915, 223.6, 322.5, 49.5),
          c(752, 2149, 101, 11.1, 15.2, 2.6), c(10528, 14992, 5377, 312.7, 710.7, 184.8))
```

```
n2 <- dim(y)[1]
```

```
# mean vector for manufacturing sector
```

```
(y_mean <- as.matrix(apply(y[, 1:2], 2, mean)))
```

```
# MLE/empirical variance matrix for manufacturing sector
```

```
(S_2 = cov(y[, 1:2]) * (n2-1)/(n2))
```

```
> (y_mean <- as.matrix(apply(y[, 1:2], 2, mean)))
```

	[,1]
[1,]	4307.2
[2,]	4925.2

 $\implies \bar{\mathbf{x}}_2 = \begin{pmatrix} 4307.2 \\ 4925.2 \end{pmatrix}$

```
> (S_2 = cov(y[, 1:2]) * (n2-1)/(n2))
```

	[,1]	[,2]
[1,]	12247663	11425398
[2,]	11425398	15111585



$$\mathcal{S}_2 = \begin{pmatrix} 12247663 & 11425398 \\ 11425398 & 15111585 \end{pmatrix}$$

线性假设

- **例:** 再来考虑美国公司数据集. 对于能源行业 (group 1) 和制造业 (group 2), 我们来比较其资产 (X_1) 和销量 (X_2) 的均值.

Weighted S

```
(S <- (n1 * S_1 + n2 * S_2) / (n1 + n2))
```

```
> (S <- (n1 * S_1 + n2 * S_2) / (n1 + n2))  
      [,1] [,2]  
[1,] 14879915 12015941  
[2,] 12015941 14293085
```

$$\Rightarrow \mathcal{S} = \frac{1}{n_1 + n_2} (n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2) = \begin{pmatrix} 14879915 & 12015941 \\ 12015941 & 14293085 \end{pmatrix}$$

线性假设

- **例:** 再来考虑美国公司数据集. 对于能源行业 (group 1) 和制造业 (group 2), 我们来比较其资产 (X_1) 和销量 (X_2) 的均值.

- ▶ 检验 $H_0: \mu_1 = \mu_2$ 的 F 统计量为

$$F = \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathcal{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = 2.703567$$

F statistic

```
(F <- n1 * n2 * (n1 + n2 - p - 1) / (p * (n1 + n2)^2) * t(x_mean - y_mean) %% solve(S) %% (x_mean - y_mean))
```

```
> (F <- n1 * n2 * (n1 + n2 - p - 1) / (p * (n1 + n2)^2) * t(x_mean - y_mean) %% solve(S) %% (x_mean - y_mean))  
      [,1]  
[1,] 2.703567
```

- ▶ 检验的 p 值为 0.08915582.

p-value

```
(pf(F, p, n1+n2-p-1, lower.tail = FALSE))
```

```
> (pf(F, p, n1+n2-p-1, lower.tail = FALSE))  
      [,1]  
[1,] 0.08915582
```

线性假设

- **例:** 再来考虑美国公司数据集. 对于能源行业 (group 1) 和制造业 (group 2), 我们来比较其资产 (X_1) 和销量 (X_2) 的均值.

- ▶ **结论:** 显著水平 $\alpha = 0.05$ 时, 因为检验的 p 值 $0.08915582 > 0.05$, 所以我们接受

$$H_0 : \mu_1 = \mu_2.$$

- ▶ **结论:** 显著水平 $\alpha = 0.10$ 时, 因为检验的 p 值 $0.08915582 < 0.10$, 所以我们拒绝

$$H_0 : \mu_1 = \mu_2.$$

- ▶ 检验的 p 值为

p-value

(pf(F, p, n1+n2-p-1, lower.tail = FALSE))

```
> (pf(F, p, n1+n2-p-1, lower.tail = FALSE))  
      [,1]  
[1,] 0.08915582
```

线性假设

- **例:** 再来考虑美国公司数据集. 对于能源行业 (group 1) 和制造业 (group 2), 我们来比较其资产 (X_1) 和销量 (X_2) 的均值.

- ▶ 均值差的置信度为 95% 的联合置信区间为

$$\delta_j = \mu_{1j} - \mu_{2j} \in \left(\bar{x}_{1j} - \bar{x}_{2j} \right) \pm \sqrt{\frac{p(n_1 + n_2)^2}{n_1 n_2 (n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha) \cdot s_{jj}}, \quad j = 1, 2$$

$$-761.5132 \leq \mu_{1a} - \mu_{2a} \leq 315.1132$$

$$-2872.3248 \leq \mu_{1s} - \mu_{2s} \leq -1817.1418$$

The 95% simultaneous confidence intervals for the differences

```
a <- sqrt(((p * (n1 + n2)^2) / (n1 * n2 * (n1 + n2 - p - 1))) * qf(0.05, p, n1+n2-p-1) * S[1, 1])
```

```
b <- sqrt(((p * (n1 + n2)^2) / (n1 * n2 * (n1 + n2 - p - 1))) * qf(0.05, p, n1+n2-p-1) * S[2, 2])
```

```
x_mean - y_mean - matrix(c(a, b), nrow=2)
```

```
x_mean - y_mean + matrix(c(a, b), nrow=2)
```

```
> x_mean - y_mean - matrix(c(a, b), nrow=2)
      [,1]
[1,] -761.5132
[2,] -2872.3248
```

```
> x_mean - y_mean + matrix(c(a, b), nrow=2)
      [,1]
[1,]  315.1132
[2,] -1817.1418
```

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.

▶ 我们从 $N_4(\mu_i, \Sigma_1)$ 拟合两个样本, 其中 $\Sigma_1 = \mathcal{I}_4$, $\mu_1 = \begin{pmatrix} 8 \\ 6 \\ 10 \\ 10 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 6 \\ 6 \\ 10 \\ 13 \end{pmatrix}$, $n_1 = 30$, $n_2 = 20$.

```

library(MASS)
(mu_1 = c(8, 6, 10, 10))
(mu_2 = c(6, 6, 10, 13))
(Sigma_1 = matrix(c(1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1), nrow = 4, byrow = TRUE))
(n1 = 30)
(n2 = 20)
(p = 4)
(X_1 <- mvrnorm(n1, mu_1, Sigma_1))
(X_2 <- mvrnorm(n2, mu_2, Sigma_1))
  
```

```

> (X_1 <- mvrnorm(n1, mu_1, Sigma_1))
      [,1] [,2] [,3] [,4]
[1,] 8.260837 6.671906 11.117292 10.256293
[2,] 8.082771 5.759805 11.125805 10.589469
[3,] 8.903574 7.753947 9.950393 9.518630
[4,] 6.290549 6.790796 10.642292 8.479732
[5,] 6.676542 6.368791 9.411594 10.158326
[6,] 6.745329 6.808953 9.964471 10.586075
[7,] 6.885432 4.937738 11.135152 11.199041
[8,] 9.526924 4.332007 9.001943 12.631406
[9,] 9.166748 6.184875 10.331208 10.359070
[10,] 7.320884 4.913665 10.022148 10.532201
[11,] 8.114750 6.377355 9.318121 9.491029
[12,] 10.430525 5.422526 10.408491 10.025123
[13,] 6.176544 7.244428 11.335207 10.800919
[14,] 6.138845 8.108788 10.056912 9.857842
[15,] 6.779966 4.125711 10.160490 9.419439
[16,] 8.310867 5.059859 10.118258 8.426746
[17,] 8.538591 7.096629 9.581178 9.530868
[18,] 8.185146 4.730059 9.449271 9.302461
[19,] 8.359620 6.255100 8.901019 9.877490
[20,] 9.065371 6.931014 8.605386 9.706008
[21,] 9.141920 5.609571 11.689098 8.870538
[22,] 8.830803 6.282785 9.994154 11.882147
[23,] 7.424494 4.800010 8.841714 10.437428
[24,] 10.089527 4.871879 9.720482 10.015031
[25,] 7.179824 5.592972 9.533401 11.009305
[26,] 7.818662 6.225895 8.008966 8.784138
[27,] 7.401145 5.662974 8.052528 9.755740
[28,] 8.734859 6.179385 11.563029 9.541995
[29,] 8.356537 5.209258 10.678182 10.047856
[30,] 6.417151 4.095609 10.192849 10.624464
  
```

```

> (X_2 <- mvrnorm(n2, mu_2, Sigma_1))
      [,1] [,2] [,3] [,4]
[1,] 5.125250 6.226469 11.850497 14.13727
[2,] 6.738880 9.083297 11.280650 13.39498
[3,] 5.815085 6.991109 9.844850 14.56964
[4,] 5.612698 6.975234 9.827844 12.85726
[5,] 6.337115 6.875820 8.979938 13.71038
[6,] 6.452551 4.538974 10.113856 13.20398
[7,] 7.419900 6.110562 10.117898 12.78523
[8,] 5.932880 5.363696 7.989238 11.25792
[9,] 6.464152 7.013614 9.372163 13.53824
[10,] 5.519181 4.376781 10.078456 12.39057
[11,] 5.552706 8.868371 9.064187 11.12131
[12,] 5.393396 4.167856 7.862729 15.25098
[13,] 7.797070 6.994050 9.124206 12.92307
[14,] 6.758487 5.706366 10.174298 11.41367
[15,] 5.920911 6.661731 9.385993 14.06337
[16,] 5.114027 6.988115 11.886501 14.75661
[17,] 5.888629 5.518362 8.954602 14.00841
[18,] 8.222050 4.354827 10.535476 13.58807
[19,] 5.732962 6.147791 9.681520 13.00775
[20,] 6.940529 4.915214 10.067072 13.22897
  
```

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.

- ▶ 我们从 $N_4(\mu_i, \Sigma_1)$ 拟合两个样本, 其中 $\Sigma_1 = \mathcal{I}_4$, $\mu_1 = \begin{pmatrix} 8 \\ 6 \\ 10 \\ 10 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 6 \\ 6 \\ 10 \\ 13 \end{pmatrix}$, $n_1 = 30$, $n_2 = 20$.

- ▶ 计算得样本统计量如下:

```
# sample mean vectors
```

```
mean_X1 <- apply(X_1, 2, mean)
round(mean_X1, digits = 3)
mean_X2 <- apply(X_2, 2, mean)
round(mean_X2, digits = 3)
```

```
> mean_X1 <- apply(X_1, 2, mean)
> round(mean_X1, digits = 3)
[1] 7.978 5.910 9.964 10.057
> mean_X2 <- apply(X_2, 2, mean)
> round(mean_X2, digits = 3)
[1] 6.237 6.194 9.810 13.260
```

$$\Rightarrow \bar{\mathbf{x}}_1 = \begin{pmatrix} 7.978 \\ 5.910 \\ 9.963 \\ 10.057 \end{pmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} 6.237 \\ 6.194 \\ 9.810 \\ 13.260 \end{pmatrix}$$

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.

- ▶ 我们从 $N_4(\mu_i, \Sigma_1)$ 拟合两个样本, 其中 $\Sigma_1 = \mathcal{I}_4$, $\mu_1 = \begin{pmatrix} 8 \\ 6 \\ 10 \\ 10 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 6 \\ 6 \\ 10 \\ 13 \end{pmatrix}$, $n_1 = 30$, $n_2 = 20$.

- ▶ 计算得样本统计量如下:

sample covariance matrix

```

S_1 = cov(X_1) * (n1-1)/(n1)
round(S_1, digits = 3)
S_2 = cov(X_2) * (n2-1)/(n2)
round(S_2, digits = 3)
    
```

```

> S_1 = cov(X_1) * (n1-1)/(n1)
> round(S_1, digits = 3)
      [,1] [,2] [,3] [,4]
[1,] 1.311 -0.188 -0.046 0.023
[2,] -0.188 0.981 0.053 -0.177
[3,] -0.046 0.053 0.897 0.017
[4,] 0.023 -0.177 0.017 0.824
> S_2 = cov(X_2) * (n2-1)/(n2)
> round(S_2, digits = 3)
      [,1] [,2] [,3] [,4]
[1,] 0.711 -0.126 0.006 -0.179
[2,] -0.126 1.797 0.276 -0.178
[3,] 0.006 0.276 1.078 0.222
[4,] -0.179 -0.178 0.222 1.181
    
```



$$\mathcal{S}_1 = \begin{pmatrix} 1.311 & -0.188 & -0.046 & 0.023 \\ -0.188 & 0.981 & 0.053 & -0.177 \\ -0.046 & 0.053 & 0.897 & 0.017 \\ 0.023 & -0.177 & 0.017 & 0.824 \end{pmatrix}$$



$$\mathcal{S}_2 = \begin{pmatrix} 0.711 & -0.126 & 0.006 & -0.179 \\ -0.126 & 1.797 & 0.276 & -0.178 \\ 0.006 & 0.276 & 1.078 & 0.222 \\ -0.179 & -0.178 & 0.222 & 1.181 \end{pmatrix}$$

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.

- ▶ 我们从 $N_4(\mu_i, \Sigma_1)$ 拟合两个样本, 其中 $\Sigma_1 = \mathcal{I}_4$, $\mu_1 = \begin{pmatrix} 8 \\ 6 \\ 10 \\ 10 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 6 \\ 6 \\ 10 \\ 13 \end{pmatrix}$, $n_1 = 30$, $n_2 = 20$.

- ▶ 检验 $H_0: \mu_1 = \mu_2$ 的 F 统计量为

$$F = \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathcal{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = 36.55469$$

```
# F statistic for testing H_0: mu_1 = mu_2
```

```
S <- (n1 * S_1 + n2 * S_2) / (n1 + n2)
```

```
F <- n1 * n2 * (n1 + n2 - p - 1) / (p * (n1 + n2)^2) * t(mean_X1 - mean_X2) %*% solve(S) %*% (mean_X1 - mean_X2)
```

```
F
```

```
# p-value
```

```
pf(F, p, n1+n2-p-1, lower.tail = FALSE)
```

```
> F
      [,1]
[1,] 36.55469
```

```
> pf(F, p, n1+n2-p-1, lower.tail = FALSE)
      [,1]
[1,] 1.327509e-13
```

- ▶ **结论:** 拒绝 $H_0: \mu_1 = \mu_2$.
- ▶ 即使在中等大小的样本容量下, 小的方差也可以检测到差异的存在.

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.

- ▶ 我们从 $N_4(\mu_i, \Sigma_1)$ 拟合两个样本, 其中 $\Sigma_1 = \mathcal{I}_4$, $\mu_1 = \begin{pmatrix} 8 \\ 6 \\ 10 \\ 10 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 6 \\ 6 \\ 10 \\ 13 \end{pmatrix}$, $n_1 = 30$, $n_2 = 20$.

- ▶ 均值差的置信度为 95% 的联合置信区间为

$$\delta_j = \mu_{1j} - \mu_{2j} \in \left(\bar{x}_{1j} - \bar{x}_{2j} \right) \pm \sqrt{\frac{p(n_1 + n_2)^2}{n_1 n_2 (n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha) \cdot s_{jj}}, \quad j = 1, 2, 3, 4$$

The 95% simultaneous confidence intervals for the differences

```
a1 <- sqrt(((p * (n1 + n2)^2) / (n1 * n2 * (n1 + n2 - p - 1))) * qf(0.05, p, n1+n2-p-1) * S[1, 1])
```

```
a2 <- sqrt(((p * (n1 + n2)^2) / (n1 * n2 * (n1 + n2 - p - 1))) * qf(0.05, p, n1+n2-p-1) * S[2, 2])
```

```
a3 <- sqrt(((p * (n1 + n2)^2) / (n1 * n2 * (n1 + n2 - p - 1))) * qf(0.05, p, n1+n2-p-1) * S[3, 3])
```

```
a4 <- sqrt(((p * (n1 + n2)^2) / (n1 * n2 * (n1 + n2 - p - 1))) * qf(0.05, p, n1+n2-p-1) * S[4, 4])
```

```
mean_X1 - mean_X2 - matrix(c(a1, a2, a3, a4), nrow=4)
```

```
[1,] 1.47790823
```

```
[2,] -0.57504658
```

```
[3,] -0.09673054
```

```
[4,] -3.45360215
```

```
mean_X1 - mean_X2 + matrix(c(a1, a2, a3, a4), nrow=4)
```

```
[1,] 2.005228521
```

```
[2,] 0.007514841
```

```
[3,] 0.404935224
```

```
[4,] -2.952711955
```

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.

- ▶ 我们从 $N_4(\mu_i, \Sigma_1)$ 拟合两个样本, 其中 $\Sigma_1 = \mathcal{I}_4$, $\mu_1 = \begin{pmatrix} 8 \\ 6 \\ 10 \\ 10 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 6 \\ 6 \\ 10 \\ 13 \end{pmatrix}$, $n_1 = 30$,
 $n_2 = 20$.

- ▶ 均值差的置信度为 95% 的联合置信区间为

$$1.47790823 \leq \delta_1 = \mu_{11} - \mu_{21} \leq 2.005228521$$

$$-0.57504658 \leq \delta_2 = \mu_{12} - \mu_{22} \leq 0.007514841$$

$$-0.09673054 \leq \delta_2 = \mu_{12} - \mu_{22} \leq 0.404935224$$

$$-3.45360215 \leq \delta_1 = \mu_{11} - \mu_{21} \leq -2.952711955$$

确认了 X_1 与 X_4 的均值不同.

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.
- ▶ 我们再来模拟来自 $N_4(\mu_i, \Sigma_2)$ 的两个样本, 其中 $\Sigma_2 = 16 \times \mathcal{I}_4$, 均值向量保持不变.

Simulate with different Sigma

```
Sigma_2 = matrix(c(16, 0, 0, 0, 0, 16, 0, 0, 0, 0, 16, 0, 0, 0, 0, 16), nrow = 4, byrow = TRUE)
```

```
Y_1 <- mvrnorm(n1, mu_1, Sigma_2); Y_1
```

```
Y_2 <- mvrnorm(n2, mu_2, Sigma_2); Y_2
```

```

> Y_1 <- mvrnorm(n1, mu_1, Sigma_2); Y_1
      [,1]      [,2]      [,3]      [,4]
[1,]  8.3974618  5.6239717 11.0430480  9.1577758
[2,]  2.5655640  6.4732461  6.8699196  6.4663611
[3,] 11.6269392  4.7602155 15.9520012  9.3999412
[4,]  2.9622076  7.9008808 14.0142753 11.1761885
[5,]  4.4609419  3.7831869 11.6210735 19.1144205
[6,]  9.0532925 13.3975639  9.9757708  0.5685187
[7,] 13.8054527  0.6812262 18.1700756  6.6383017
[8,] 18.6778096  3.7646728 -0.4937562  5.8048799
[9,] 10.9239779  9.1130861 13.3113333  8.6132512
[10,]  0.9766640  5.8239364  4.6726798  9.4461028
[11,]  0.8677006  8.3602153  4.3238356  7.3307191
[12,]  4.9780011  5.4006436  7.4792162  8.8289769
[13,]  4.8823786  6.1921179  8.2168797  8.7604994
[14,]  7.4326080 11.4293103 15.8503835  4.6419381
[15,]  5.6320427  9.1945914 11.5040166  8.9562280
[16,] 13.1343026  6.7090997  8.6112216 10.5773574
[17,] -1.7694523  4.0295208  8.7826784  7.8686858
[18,] 10.7978931  8.7803181 10.4981390  6.7867055
[19,] 15.1951108  7.2516345 11.0498941  3.9106981
[20,]  1.4278909  8.2678676 14.7626042 14.8588964
[21,]  7.7636468  3.4093564 15.2820005 12.8585235
[22,] 12.0895838 13.6414836  5.2765707  8.5743802
[23,]  5.0028268 10.0349296 12.7990420  8.5908834
[24,]  4.6203294  4.3800721  6.7605634  6.5961305
[25,]  4.6732495  6.5886095  9.2502411 19.5157791
[26,]  3.1829271  6.8609971 14.0103484 11.4241146
[27,]  8.7853168  6.3196780 11.8560324  8.5628778
[28,]  4.7488134  0.6448099  7.2487858 12.1769575
[29,]  8.7762873  3.3993835  4.6385889 15.3355914
[30,]  3.7925491  9.3927687 10.1319640  9.7514829
  
```

```

> Y_2 <- mvrnorm(n2, mu_2, Sigma_2); Y_2
      [,1]      [,2]      [,3]      [,4]
[1,]  9.6880895  0.6107765 10.576171 13.972106
[2,]  0.4286959 10.3560186  6.843536 13.946251
[3,]  6.5455939 10.7287525 12.856401 13.888709
[4,]  6.2194654 13.1424701  9.719947 14.989641
[5,] -2.1446008  7.4583453 15.369491 12.645088
[6,]  3.8734797  7.9093282 12.906593  6.741463
[7,]  9.2947745 11.6054377  7.945354 11.337787
[8,] 14.0846093  8.8132397  9.873295 14.333612
[9,] 12.5657412  6.4085165 12.465166 16.791857
[10,]  1.1174743  3.2295778 13.628226 16.551058
[11,]  9.9642556  6.1565453  4.083274 12.810890
[12,]  8.9792592 11.0616746  8.514730  7.796829
[13,]  8.6532576 -1.5300757 10.217014  8.143847
[14,]  3.9539232 11.3752375 13.097941  2.357502
[15,]  5.6849686  6.7640401 15.251841 16.642179
[16,]  4.3162601  2.8386859 14.505072  2.730159
[17,]  6.1136579  4.7240511 13.412611  7.398367
[18,]  8.7485605 14.5282507  5.798741  8.797023
[19,] 10.8835587 10.6198283  7.148565 14.691502
[20,]  1.4001456  8.5506892 16.137652 11.388991
  
```

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.
- ▶ 我们再来模拟来自 $N_4(\mu_i, \Sigma_2)$ 的两个样本, 其中 $\Sigma_2 = 16 \times \mathcal{I}_4$, 均值向量保持不变.
- ▶ 计算得样本统计量如下:

```
# sample mean vectors
```

```
mean_Y1 <- apply(Y_1, 2, mean)
```

```
round(mean_Y1, digits = 3)
```

```
mean_Y2 <- apply(Y_2, 2, mean)
```

```
round(mean_Y2, digits = 3)
```

```
> mean_Y1 <- apply(Y_1, 2, mean)
> round(mean_Y1, digits = 3)
[1] 6.982 6.720 10.116 9.410
> mean_Y2 <- apply(Y_2, 2, mean)
> round(mean_Y2, digits = 3)
[1] 6.519 7.768 11.018 11.398
```



$$\bar{y}_1 = \begin{pmatrix} 6.982 \\ 6.720 \\ 10.116 \\ 9.410 \end{pmatrix}, \quad \bar{y}_2 = \begin{pmatrix} 6.519 \\ 7.768 \\ 11.018 \\ 11.398 \end{pmatrix}$$

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.
- ▶ 我们再来模拟来自 $N_4(\mu_i, \Sigma_2)$ 的两个样本, 其中 $\Sigma_2 = 16 \times \mathcal{F}_4$, 均值向量保持不变.
- ▶ 计算得样本统计量如下:

empirical covariance matrices

S_1 = cov(Y_1) * (n1-1)/(n1)

round(S_1, digits = 3)

S_2 = cov(Y_2) * (n2-1)/(n2)

round(S_2, digits = 3)

```

> S_1 = cov(Y_1) * (n1-1)/(n1)
> round(S_1, digits = 3)
      [,1] [,2] [,3] [,4]
[1,] 21.919 -0.152 -0.031 -5.841
[2,] -0.152  9.641  0.737 -4.341
[3,] -0.031  0.737 17.050  1.261
[4,] -5.841 -4.341  1.261 15.835
> S_2 = cov(Y_2) * (n2-1)/(n2)
> round(S_2, digits = 3)
      [,1] [,2] [,3] [,4]
[1,] 17.248  0.354 -7.239  2.859
[2,]  0.354 16.794 -4.511  0.543
[3,] -7.239 -4.511 11.421 -1.547
[4,]  2.859  0.543 -1.547 18.163
    
```

$$\mathcal{S}_1 = \begin{pmatrix} 21.919 & -0.152 & -0.031 & -5.841 \\ -0.152 & 9.641 & 0.737 & -4.341 \\ -0.031 & 0.737 & 17.050 & 1.261 \\ -5.841 & -4.341 & 1.261 & 15.835 \end{pmatrix}$$

$$\mathcal{S}_2 = \begin{pmatrix} 17.248 & 0.354 & -7.239 & 2.859 \\ 0.354 & 16.794 & -4.511 & 0.543 \\ -7.239 & -4.511 & 11.421 & -1.547 \\ 2.859 & 0.543 & -1.547 & 18.163 \end{pmatrix}$$

线性假设

- **例:** 以下我们利用模拟数据来指明, 在检验均值时协方差矩阵的重要性.
- ▶ 我们再来模拟来自 $N_4(\mu_i, \Sigma_2)$ 的两个样本, 其中 $\Sigma_2 = 16 \times \mathcal{I}_4$, 均值向量保持不变.
- ▶ 检验 $H_0: \mu_1 = \mu_2$ 的 F 统计量为

$$F = \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{y}_1 - \bar{y}_2)^T \mathcal{S}^{-1} (\bar{y}_1 - \bar{y}_2) = 1.226174$$

```
# F statistic for testing H_0: mu_1 = mu_2
```

```
S <- (n1 * S_1 + n2 * S_2) / (n1 + n2)
```

```
F <- n1 * n2 * (n1 + n2 - p - 1) / (p * (n1 + n2)^2) * t(mean_Y1 - mean_Y2) %*% solve(S) %*% (mean_Y1 - mean_Y2)
```

```
F
```

```
# p-value
```

```
pf(F, p, n1+n2-p-1, lower.tail = FALSE)
```

```
> F
      [,1]
[1,] 1.226174
```

```
> pf(F, p, n1+n2-p-1, lower.tail = FALSE)
      [,1]
[1,] 0.3131732
```

- ▶ **结论:** 接受 $H_0: \mu_1 = \mu_2$.
- ▶ 方差的增大使我们无法检测到这种幅度的差异.

线性假设

- **例:** 我们来比较伪钞和真钞的均值向量.

```

rm(list = ls(all = TRUE))
library(mclust)
data(banknote)
banknote_genuine = subset(banknote, Status == 'genuine')[, 2:7]
banknote_counterfeit = subset(banknote, Status == 'counterfeit')[, 2:7]
  
```

- ▶ 计算样本统计量如下:

```

# sample statistics
mean_g <- apply(banknote_genuine, 2, mean)
round(mean_g, digits = 2)
mean_f <- apply(banknote_counterfeit, 2, mean)
round(mean_f, digits = 2)
  
```

```

> round(mean_g, digits = 2)
  Length      Left      Right      Bottom      Top Diagonal
  214.97  129.94  129.72      8.30    10.17    141.52
> round(mean_f, digits = 2)
  Length      Left      Right      Bottom      Top Diagonal
  214.82  130.30  130.19     10.53    11.13    139.45
  
```



$$\bar{\mathbf{x}}_g = \begin{pmatrix} 214.97 \\ 129.94 \\ 129.72 \\ 8.30 \\ 10.17 \\ 141.52 \end{pmatrix}, \quad \bar{\mathbf{x}}_f = \begin{pmatrix} 214.82 \\ 130.30 \\ 130.19 \\ 10.53 \\ 11.13 \\ 139.45 \end{pmatrix}$$

线性假设

- **例:** 我们来比较伪钞和真钞的均值向量.
- ▶ 计算样本统计量如下:

sample statistics

```

Sigma_g <- (n_g - 1) * cov(banknote_genuine) / n_g
round(Sigma_g, digits = 4)

Sigma_f <- (n_f - 1) * cov(banknote_counterfeit) / n_f
round(Sigma_f, digits = 4)
    
```

```

> round(Sigma_g, digits = 4)
      Length  Left  Right  Bottom  Top  Diagonal
Length  0.1487  0.0574  0.0567  0.0566  0.0143  0.0054
Left    0.0574  0.1313  0.0850  0.0561  0.0486 -0.0426
Right   0.0567  0.0850  0.1250  0.0576  0.0303 -0.0235
Bottom  0.0566  0.0561  0.0576  0.4091 -0.2608 -0.0002
Top     0.0143  0.0486  0.0303 -0.2608  0.4170 -0.0746
Diagonal 0.0054 -0.0426 -0.0235 -0.0002 -0.0746  0.1978

> round(Sigma_f, digits = 4)
      Length  Left  Right  Bottom  Top  Diagonal
Length  0.1228  0.0312  0.0238 -0.0996  0.0192  0.0114
Left    0.0312  0.0644  0.0463 -0.0238 -0.0118 -0.0050
Right   0.0238  0.0463  0.0881 -0.0184  0.0001  0.0339
Bottom -0.0996 -0.0238 -0.0184  1.2685 -0.4853  0.2361
Top     0.0192 -0.0118  0.0001 -0.4853  0.4004 -0.0219
Diagonal 0.0114 -0.0050  0.0339  0.2361 -0.0219  0.3081
    
```

$$\Rightarrow \mathcal{S}_g = \begin{pmatrix} 0.1487 & 0.0574 & 0.0567 & 0.0566 & 0.0143 & 0.0054 \\ 0.0574 & 0.1313 & 0.0850 & 0.0561 & 0.0486 & -0.0426 \\ 0.0567 & 0.0850 & 0.1250 & 0.0576 & 0.0303 & -0.0235 \\ 0.0566 & 0.0561 & 0.0576 & 0.4091 & -0.2608 & -0.0002 \\ 0.0143 & 0.0486 & 0.0303 & -0.2608 & 0.4170 & -0.0746 \\ 0.0054 & -0.0426 & -0.0235 & -0.0002 & -0.0746 & 0.1978 \end{pmatrix}$$

$$\Rightarrow \mathcal{S}_f = \begin{pmatrix} 0.1228 & 0.0312 & 0.0238 & -0.0996 & 0.0192 & 0.0114 \\ 0.0312 & 0.0644 & 0.0463 & -0.0238 & -0.0118 & -0.0050 \\ 0.0238 & 0.0463 & 0.0881 & -0.0184 & 0.0001 & 0.0339 \\ -0.0996 & -0.0238 & -0.0184 & 1.2685 & -0.4853 & 0.2361 \\ 0.0192 & -0.0118 & 0.0001 & -0.4853 & 0.4004 & -0.0219 \\ 0.0114 & -0.0050 & 0.0339 & 0.2361 & -0.0219 & 0.3081 \end{pmatrix}$$

线性假设

- **例:** 我们来比较伪钞和真钞的均值向量.

- ▶ 检验 $H_0 : \mu_1 = \mu_2$ 的 F 统计量为 $F = 391.9217$

```
# F statistic for testing H_0: mu_1 = mu_2
```

```
S <- (n_g * Sigma_g + n_f * Sigma_f) / (n_g + n_f)
```

```
F <- n_g * n_f * (n_g + n_f - p - 1) / (p * (n_g + n_f)^2) * t(mean_g - mean_f) %*% solve(S) %*% (mean_g - mean_f)
```

```
F
```

```
# p-value
```

```
pf(F, p, n_g + n_f - p - 1, lower.tail = FALSE)
```

```
> F  
[1,] 391.9217
```

```
> # p-value  
> pf(F, p, n_g + n_f - p - 1, lower.tail = FALSE)  
[1,] 3.378887e-105
```

- ▶ **结论:** 拒绝 $H_0 : \mu_1 = \mu_2$.

线性假设

- **例:** 我们来比较伪钞和真钞的均值向量.

- ▶ 均值差 $\delta_j = \mu_{gj} - \mu_{fj}$ ($j = 1, 2, \dots, 6$) 的置信度为 95% 的联合置信区间为

$$\delta_j = \mu_{1j} - \mu_{2j} \in \left(\bar{x}_{1j} - \bar{x}_{2j} \right) \pm \sqrt{\frac{p \left(n_g + n_f \right)^2}{n_g n_f \left(n_g + n_f - p - 1 \right)} F_{p, n_g + n_f - p - 1}(\alpha) \cdot s_{jj}}, \quad j = 1, 2, 3, 4, 5, 6$$

```

> mean_g - mean_f - matrix(c(a1, a2, a3, a4, a5, a6), nrow=6)
      [,1]
[1,] 0.0783729
[2,] -0.4144075
[3,] -0.5329059
[4,] -2.3931004
[5,] -1.0823388
[6,] 1.9746866
  
```

```

> mean_g - mean_f + matrix(c(a1, a2, a3, a4, a5, a6), nrow=6)
      [,1]
[1,] 0.2136271
[2,] -0.2995925
[3,] -0.4130941
[4,] -2.0568996
[5,] -0.8476612
[6,] 2.1593134
  
```

The 95% simultaneous confidence intervals for the differences

```

a1 <- sqrt(((p * (n_g + n_f)^2) / (n_g * n_f * (n_g + n_f - p - 1))) * qf(0.05, p, n_g + n_f - p - 1) * S[1, 1])
a2 <- sqrt(((p * (n_g + n_f)^2) / (n_g * n_f * (n_g + n_f - p - 1))) * qf(0.05, p, n_g + n_f - p - 1) * S[2, 2])
a3 <- sqrt(((p * (n_g + n_f)^2) / (n_g * n_f * (n_g + n_f - p - 1))) * qf(0.05, p, n_g + n_f - p - 1) * S[3, 3])
a4 <- sqrt(((p * (n_g + n_f)^2) / (n_g * n_f * (n_g + n_f - p - 1))) * qf(0.05, p, n_g + n_f - p - 1) * S[4, 4])
a5 <- sqrt(((p * (n_g + n_f)^2) / (n_g * n_f * (n_g + n_f - p - 1))) * qf(0.05, p, n_g + n_f - p - 1) * S[5, 5])
a6 <- sqrt(((p * (n_g + n_f)^2) / (n_g * n_f * (n_g + n_f - p - 1))) * qf(0.05, p, n_g + n_f - p - 1) * S[6, 6])
mean_g - mean_f - matrix(c(a1, a2, a3, a4, a5, a6), nrow=6)
mean_g - mean_f + matrix(c(a1, a2, a3, a4, a5, a6), nrow=6)
  
```

线性假设

- **例:** 我们来比较伪钞和真钞的均值向量.

- ▶ 均值差 $\delta_j = \mu_{gj} - \mu_{fj}$ ($j = 1, 2, \dots, 6$) 的置信度为 95% 的联合置信区间为

$$\delta_j = \mu_{1j} - \mu_{2j} \in \left(\bar{x}_{1j} - \bar{x}_{2j} \right) \pm \sqrt{\frac{p \left(n_g + n_f \right)^2}{n_g n_f \left(n_g + n_f - p - 1 \right)} F_{p, n_g + n_f - p - 1}(\alpha) \cdot s_{jj}}, \quad j = 1, 2, 3, 4, 5, 6$$

$$0.0783729 \leq \delta_1 \leq 0.2136271$$

$$-0.4144075 \leq \delta_2 \leq -0.2995925$$

$$-0.5329059 \leq \delta_3 \leq -0.4130941$$

$$-2.3931004 \leq \delta_4 \leq -2.0568996$$

$$-1.0823388 \leq \delta_5 \leq -0.8476612$$

$$1.9746866 \leq \delta_6 \leq 2.1593134$$

- ▶ 所有分量的均值都表现出了显著的差异. 最为显著的是 X_4 (lower border) 与 X_6 (diagonal).

线性假设

检验问题 9: (协方差矩阵的比较) 设 $X_{ih} \sim N_p(\mu_h, \Sigma_h)$, $i = 1, 2, \dots, n_h$, $h = 1, 2, \dots, k$ 是相互独立的随机向量.

$$H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_k \iff H_1 : \text{无约束}$$

$$\begin{array}{l}
 N_p(\mu_1, \Sigma_1) \xrightarrow{\text{sampling}} X_{1\ 1}, X_{2\ 1}, \dots, X_{n_1\ 1} \mapsto n_1 \mathcal{S}_1 = \sum_{i=1}^{n_1} (x_{i\ 1} - \bar{x}_1)(x_{i\ 1} - \bar{x}_1)^T \sim W_p(\Sigma_1, n_1 - 1) \\
 N_p(\mu_2, \Sigma_2) \xrightarrow{\text{sampling}} X_{1\ 2}, X_{2\ 2}, \dots, X_{n_2\ 2} \mapsto n_2 \mathcal{S}_2 = \sum_{i=1}^{n_2} (x_{i\ 2} - \bar{x}_2)(x_{i\ 2} - \bar{x}_2)^T \sim W_p(\Sigma_2, n_2 - 1) \\
 \vdots \\
 N_p(\mu_k, \Sigma_k) \xrightarrow{\text{sampling}} X_{1\ k}, X_{2\ k}, \dots, X_{n_k\ k} \mapsto n_k \mathcal{S}_k = \sum_{i=1}^{n_k} (x_{i\ k} - \bar{x}_k)(x_{i\ k} - \bar{x}_k)^T \sim W_p(\Sigma_k, n_k - 1)
 \end{array}$$

相互独立

▶ 当 $H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_k \triangleq \Sigma$ 真时, 我们有

$$\sum_{h=1}^k n_h \mathcal{S}_h \sim W_p(\Sigma, n - k)$$

$n = n_1 + n_2 + \dots + n_k$

线性假设

检验问题 9: (协方差矩阵的比较) 设 $X_{ih} \sim N_p(\mu_h, \Sigma_h)$, $i = 1, 2, \dots, n_h$, $h = 1, 2, \dots, k$ 是相互独立的随机向量.

$$H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_k \longleftrightarrow H_1: \text{无约束}$$

- ▶ 记 \mathcal{S}_h 的加权平均为 \mathcal{S} :

$$\mathcal{S} = \frac{1}{n} \sum_{h=1}^k n_h \mathcal{S}_h = \frac{n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2 + \dots + n_k \mathcal{S}_k}{n_1 + n_2 + \dots + n_k}$$

- ▶ 似然比检验得到以下统计量

$$-2 \log \lambda = n \log |\mathcal{S}| - \sum_{h=1}^k n_h \log |\mathcal{S}_h|$$

- ▶ 当 $H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$ 真时, 我们有

$$-2 \log \lambda = n \log |\mathcal{S}| - \sum_{h=1}^k n_h \log |\mathcal{S}_h| \stackrel{\text{approximately}}{\sim} \chi_m^2$$

$$m = \frac{1}{2} (k-1) p (p+1)$$

线性假设

- **例:** 我们来比较伪钞和真钞数据的协方差矩阵.
- ▶ 检验 $H_0: \Sigma_g = \Sigma_f$, 我们有

```
# likelihood ratio statistic
```

```
LR_statistic <- log(det(S)) * (n_g + n_f) - (log(det(Sigma_g)) * n_g + log(det(Sigma_f)) * n_f)  
LR_statistic
```

```
> LR_statistic  
[1] 127.2149
```



$-2 \log \lambda = 127.2149$

```
# p-value
```

```
k <- 2  
m <- 0.5 * (k-1) * p * (p+1)  
pchisq(LR_statistic, m, lower.tail = FALSE)
```

```
> pchisq(LR_statistic, m, lower.tail = FALSE)  
[1] 3.338319e-17
```



p 值 = 3.338319×10^{-17}

- ▶ **结论:** 拒绝 $H_0: \Sigma_g = \Sigma_f$.

线性假设

检验问题 10: (协方差矩阵不等时, 大样本情形下, 两个均值向量的比较)

假设 $X_{i1} \sim N_p(\mu_1, \Sigma_1)$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\mu_2, \Sigma_2)$, $j = 1, 2, \dots, n_2$,
是相互独立的随机向量.

$$H_0: \mu_1 = \mu_2 \iff H_1: \text{无约束}$$

▶ 记 $\delta = \mu_1 - \mu_2$, 我们有

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \sim N_p\left(\delta, \frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right)$$

$$\implies (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \left(\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \sim \chi_p^2$$

▶ 因为 \mathcal{S}_i 是 Σ_i 的一致估计量, $i = 1, 2$, 我们有

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \left(\frac{\mathcal{S}_1}{n_1} + \frac{\mathcal{S}_2}{n_2}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \xrightarrow{\mathcal{L}} \chi_p^2$$

线性假设

检验问题 10: (协方差矩阵不等时, 大样本情形下, 两个均值向量的比较)

假设 $X_{i1} \sim N_p(\mu_1, \Sigma_1)$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\mu_2, \Sigma_2)$, $j = 1, 2, \dots, n_2$, 是相互独立的随机向量.

$$H_0: \mu_1 = \mu_2 \iff H_1: \text{无约束}$$

- ▶ 显著水平为 α 时, $H_0: \mu_1 = \mu_2$ 的拒绝域为

$$(\bar{x}_1 - \bar{x}_2)^T \left(\frac{\mathcal{S}_1}{n_1} + \frac{\mathcal{S}_2}{n_2} \right)^{-1} (\bar{x}_1 - \bar{x}_2) > \chi_p^2(\alpha)$$

上 α 分位数

- ▶ $\delta_j = \mu_{1j} - \mu_{2j}$, $j = 1, 2, \dots, p$ 的置信度为 $1 - \alpha$ 的联合置信区间为

$$\delta_j \in (\bar{x}_1 - \bar{x}_2) \pm \sqrt{\chi_p^2(\alpha) \left(\frac{s_{jj}^{(1)}}{n_1} + \frac{s_{jj}^{(2)}}{n_2} \right)}$$

\mathcal{S}_1 的 (j, j) 元素

\mathcal{S}_2 的 (j, j) 元素

线性假设

- **例:** 我们用上述检验来比较伪钞和真钞的协方差矩阵 (n_1 和 n_2 都足够大).
- ▶ 对于 $H_0: \Sigma_g = \Sigma_f$, 我们有

```
# test statistic
```

```
mean_g <- as.matrix(mean_g)
```

```
mean_f <- as.matrix(mean_f)
```

```
TS <- t(mean_g - mean_f) %*% solve((1/n_g) * Sigma_g + (1/n_f) * Sigma_f) %*% (mean_g - mean_f)
```

```
TS
```

```
> TS  
      [,1]  
[1,] 2436.819
```

```
# p-value
```

```
pchisq(TS, p, lower.tail = FALSE)
```

```
> pchisq(TS, p, lower.tail = FALSE)  
      [,1]  
[1,] 0
```

- ▶ **结论:** 拒绝 $H_0: \Sigma_g = \Sigma_f$.

线性假设

- **例:** 我们用上述检验来比较伪钞和真钞的协方差矩阵 (n_1 和 n_2 都足够大).
- ▶ 置信度为 95% 的联合置信区间为

The 95% simultaneous confidence intervals

```

a <- array(0, dim = p)
for (j in 1:p) {
  a[j] <- sqrt(qchisq(0.05, p, lower.tail = FALSE) * (Sigma_g[j, j] / n_g + Sigma_f[j, j] / n_f))
}
a <- as.matrix(a)
Lower <- mean_g - mean_f - a
Upper <- mean_g - mean_f + a
SCI <- data.frame(Lower = Lower, Upper = Upper)
SCI
  
```

```

> round(SCI, digits = 4)
      Lower  Upper
Length -0.0389  0.3309
Left   -0.5140 -0.2000
Right  -0.6368 -0.3092
Bottom -2.6846 -1.7654
Top    -1.2858 -0.6442
Diagonal 1.8146  2.3194
  
```

$$-0.0389 \leq \delta_1 = \mu_{g1} - \mu_{f1} \leq 0.3309$$

$$-0.5140 \leq \delta_2 = \mu_{g2} - \mu_{f2} \leq -0.2000$$

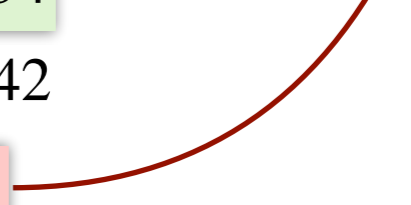
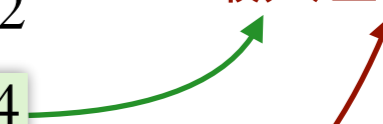
$$-0.6368 \leq \delta_3 = \mu_{g3} - \mu_{f3} \leq -0.3092$$

$$-2.6846 \leq \delta_4 = \mu_{g4} - \mu_{f4} \leq -1.7654$$

$$-1.2858 \leq \delta_5 = \mu_{g5} - \mu_{f5} \leq -0.6442$$

$$1.8146 \leq \delta_6 = \mu_{g6} - \mu_{f6} \leq 2.3194$$

最大差异

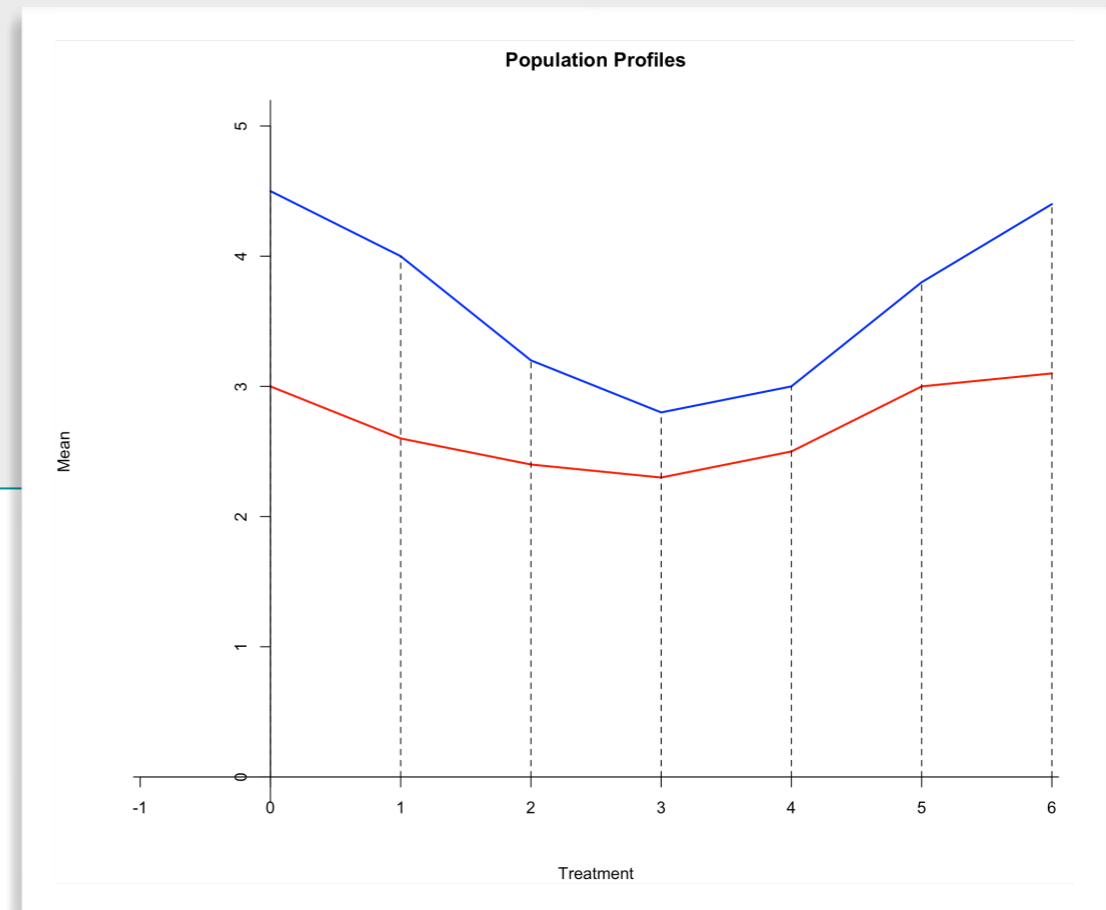


线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 问题源自实践中, 当我们对某一特征进行重复观测 (或在不同实验条件下进行相同类型的测量), 且需要对不同群体进行比较时.

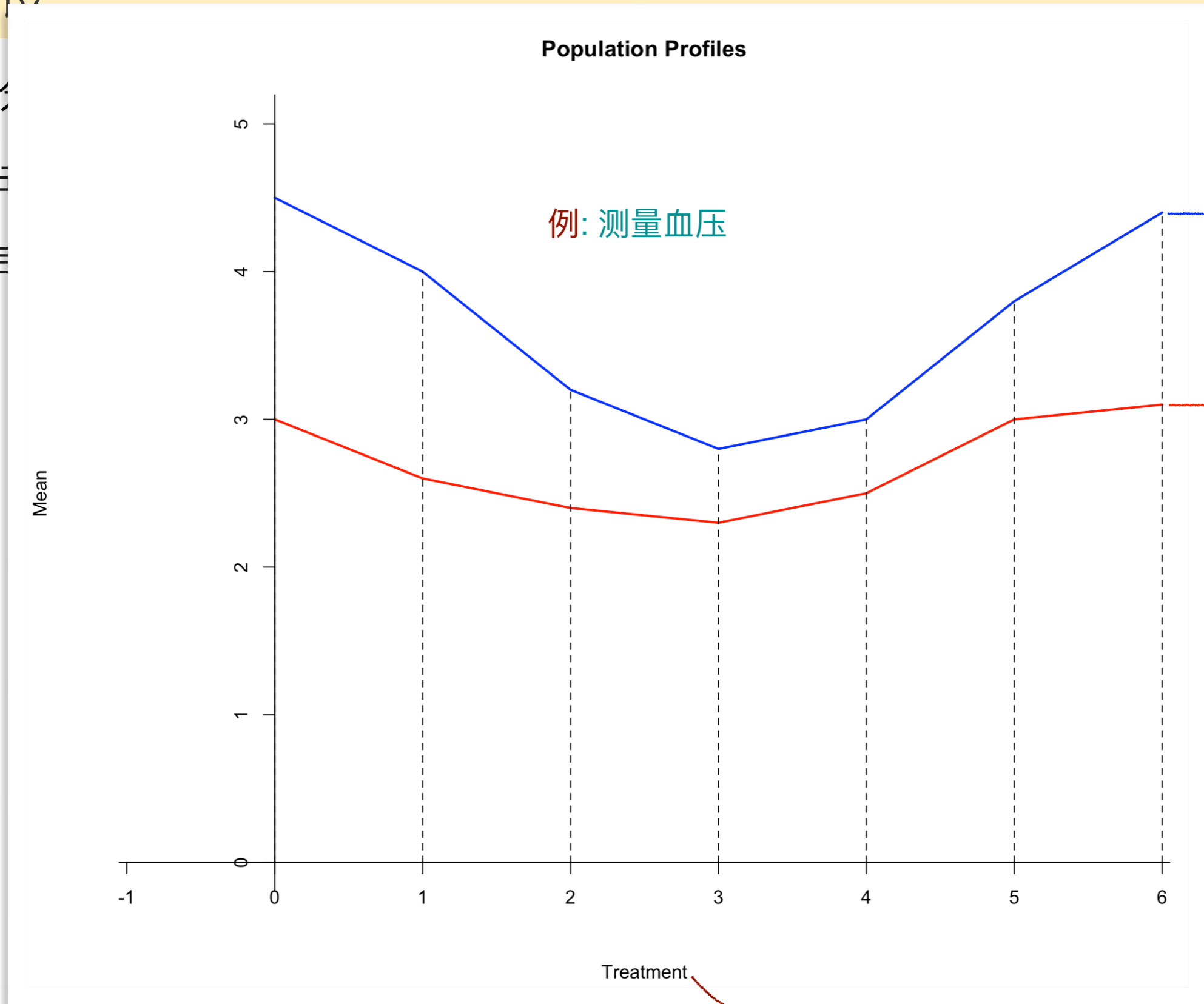
```
x <- c(0, 1, 2, 3, 4, 5, 6)
y1 <- c(4.5, 4.0, 3.2, 2.8, 3.0, 3.8, 4.4)
y2 <- c(3.0, 2.6, 2.4, 2.3, 2.5, 3.0, 3.1)
plot(c(0, 5), c(0, 5), type = 'n', axes = FALSE, xlab = 'Treatment', ylab = 'Mean',
     main = 'Population Profiles', asp = 1)
abline(h = 0, v = 0)
lines(x, y1, lty = 1, lwd = 2, col = 'blue')
lines(x, y2, lty = 1, lwd = 2, col = 'red')
axis(1, pos = 0)
axis(2, pos = 0)
for (i in 1:7) lines(c(x[i], x[i]), c(0, y1[i]), lty = 2)
```



线性假设

- 轮廓分

- ▶ 问题同



例: 测量血压

进行相

新药

安慰剂

Treatment

不同时间点

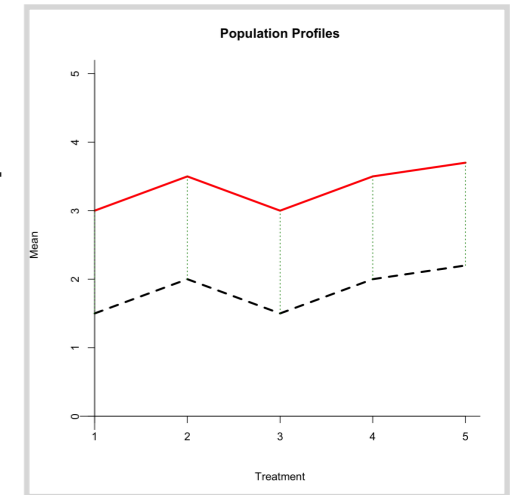
线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 这与两个均值向量的比较属于相同的统计问题:

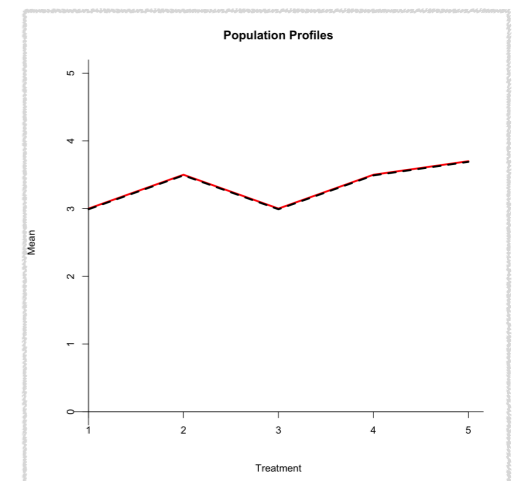
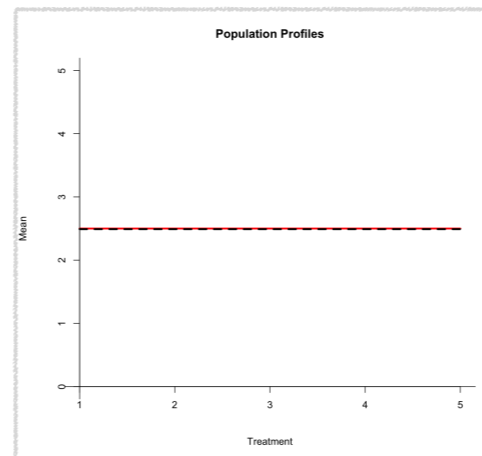
$$\begin{aligned}
 X_{i1} &\sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad i = 1, 2, \dots, n_1 \\
 X_{i2} &\sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \quad i = 1, 2, \dots, n_2
 \end{aligned}$$

所有变量相互独立



- ▶ 我们关心的问题有:

- ✓ 从平行的角度看, 这些轮廓是否相似? (这说明治疗 and 分组之间无交互作用)
- ✓ 如果轮廓是平行的, 它们是否处于相同的水平?
- ✓ 如果轮廓是平行的, 是否有分组效应, 即轮廓是否为水平的? (无论接受哪种治疗方法, 效果都相同)



线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 平行

$$\mathcal{C} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}_{(p-1) \times p} \begin{pmatrix} (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) \\ (\mu_{21} - \mu_{22}) - (\mu_{31} - \mu_{32}) \\ \vdots \\ (\mu_{p-11} - \mu_{p-12}) - (\mu_{p1} - \mu_{p2}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

✓ 欲检验的假设为 $H_0^{(1)} : \mathcal{C} (\mu_1 - \mu_2) = \mathbf{0}$

```
y3 <- y1 + 1.5
```

```
plot(c(1, 5), c(0, 5), type = 'n', axes = FALSE, xlab = 'Treatment', ylab = 'Mean', main = 'Population Profiles')
```

```
abline(h=0, v=1)
```

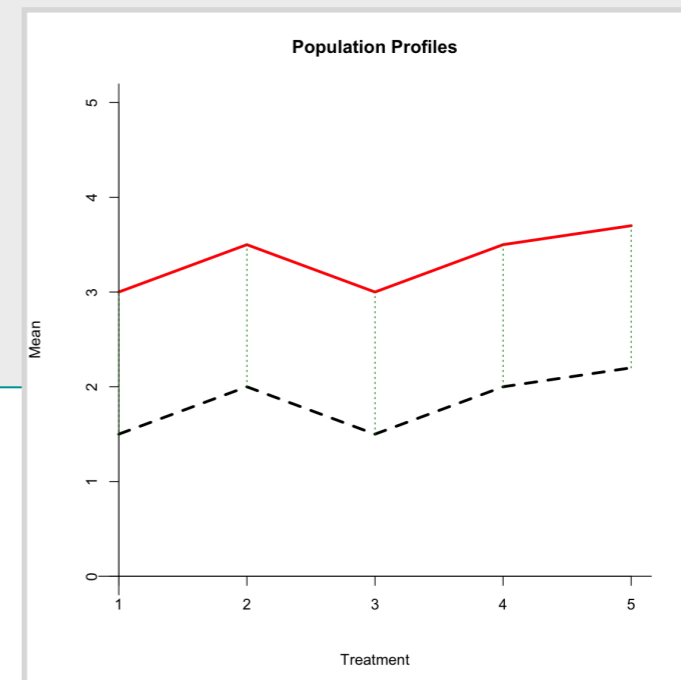
```
lines(x, y1, lty=2, lwd=3)
```

```
lines(x, y3, lty=1, lwd=3, col = 'red')
```

```
for (i in 1:5) lines(c(x[i], x[i]), c(y1[i], y3[i]), col = 'green4', lty=3)
```

```
axis(1, pos = 0)
```

```
axis(2, pos = 1)
```



线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 平行

$$\mathcal{C} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}_{(p-1) \times p} \begin{pmatrix} (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) \\ (\mu_{21} - \mu_{22}) - (\mu_{31} - \mu_{32}) \\ \vdots \\ (\mu_{p-11} - \mu_{p-12}) - (\mu_{p1} - \mu_{p2}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

✓ 欲检验的假设为 $H_0^{(1)} : \mathcal{C}(\mu_1 - \mu_2) = \mathbf{0}$

✓ 应用检验问题 8 的结果, 我们有

$$\mathcal{S} = \frac{1}{n_1 + n_2} (n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2)$$

$$\frac{n_1 n_2}{(n_1 + n_2)^2} (n_1 + n_2 - 2) \left[\mathcal{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \right]^T \left(\mathcal{C} \mathcal{S} \mathcal{C}^T \right)^{-1} \mathcal{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \stackrel{H_0 \text{ True}}{\sim} T_{p-1, n_1+n_2-2}^2$$

✓ 拒绝域:

$$\frac{n_1 n_2 (n_1 + n_2 - p)}{(n_1 + n_2)^2 (p - 1)} (\mathcal{C} \bar{\mathbf{x}})^T \left(\mathcal{C} \mathcal{S} \mathcal{C}^T \right)^{-1} \mathcal{C} \bar{\mathbf{x}} > F_{p-1, n_1+n_2-p}(\alpha)$$

线性假设

- 轮廓分析 (Profile Analysis)

$$\implies \mu_{1,1} = \mu_{1,2}, \mu_{2,1} = \mu_{2,2}, \dots, \mu_{p,1} = \mu_{p,2}$$

- ▶ 两个相同的水平

$$(\mu_{1,1} - \mu_{1,2}) + (\mu_{2,1} - \mu_{2,2}) + \dots + (\mu_{p,1} - \mu_{p,2}) = 0$$

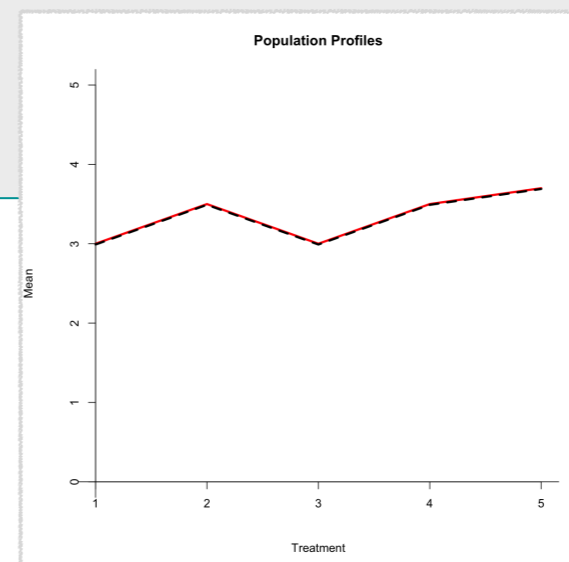
- ✓ 该问题只有当两个轮廓平行 (接受 $H_0^{(1)}$) 时才有意义

- ✓ 两个水平相同可表述为

$$H_0^{(2)} : \mathbf{1}_p^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0 \quad (1 \quad 1 \quad \dots \quad 1) \begin{pmatrix} \mu_{1,1} - \mu_{1,2} \\ \mu_{2,1} - \mu_{2,2} \\ \vdots \\ \mu_{p,1} - \mu_{p,2} \end{pmatrix} = 0$$

```

plot(c(1, 5), c(0, 5), type = 'n', axes = FALSE, xlab = 'Treatment', ylab = 'Mean', main = 'Population Profiles')
abline(h=0, v=1)
lines(x, y3, lty=1, lwd=3, col = 'red')
lines(x, y3-0.01, lty=2, lwd=3, col = 'black')
axis(1, pos = 0)
axis(2, pos = 1)
    
```



线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 两个相同的水平

✓ 该问题只有当两个轮廓平行 (接受 $H_0^{(1)}$) 时才有意义

✓ 两个水平相同可表述为

$$H_0^{(2)} : \mathbf{1}_p^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$$

✓ 因为

$$\mathbf{1}_p^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \sim N_1 \left(\mathbf{1}_p^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \frac{n_1 + n_2}{n_1 n_2} \mathbf{1}_p^T \boldsymbol{\Sigma} \mathbf{1}_p \right)$$

$$(n_1 + n_2) \mathbf{1}_p^T \boldsymbol{\Sigma} \mathbf{1}_p \sim W_1 \left(\mathbf{1}_p^T \boldsymbol{\Sigma} \mathbf{1}_p, n_1 + n_2 - 2 \right)$$

$$\frac{n_1 n_2}{(n_1 + n_2)^2} (n_1 + n_2 - 2) \frac{\left[\mathbf{1}_p^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \right]^2}{\mathbf{1}_p^T \boldsymbol{\Sigma} \mathbf{1}_p} \sim T_{1, n_1 + n_2 - 2}^2 = F_{1, n_1 + n_2 - 2}$$

推论 5.4 考虑 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一个线性变换 $Y = \mathcal{C}X$, 其中 $\mathcal{C}_{q \times p}$ 满足 $q \leq p$. 如果

$\bar{\mathbf{x}}$ 与 \mathcal{S}_X 分别为样本均值向量与样本协方差矩阵, 则有

$$\bar{\mathbf{y}} = \mathcal{C}\bar{\mathbf{x}} \sim N_q \left(\mathcal{C}\boldsymbol{\mu}, \frac{1}{n} \mathcal{C}\boldsymbol{\Sigma}\mathcal{C}^T \right)$$

$$n\mathcal{S}_Y = n\mathcal{C}\mathcal{S}_X\mathcal{C}^T \sim W_q(\mathcal{C}\boldsymbol{\Sigma}\mathcal{C}^T, n-1)$$

$$(n-1)(\mathcal{C}\bar{\mathbf{x}} - \mathcal{C}\boldsymbol{\mu})^T (\mathcal{C}\mathcal{S}_X\mathcal{C}^T)^{-1} (\mathcal{C}\bar{\mathbf{x}} - \mathcal{C}\boldsymbol{\mu}) \sim T_{q, n-1}^2$$

线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 两个相同的水平

- ✓ 该问题只有当两个轮廓平行 (接受 $H_0^{(1)}$) 时才有意义

- ✓ 两个水平相同可表述为

$$H_0^{(2)} : \mathbf{1}_p^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$$

- ✓ 拒绝域为

$$\frac{n_1 n_2 (n_1 + n_2 - 2) \left[\mathbf{1}_p^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \right]^2}{(n_1 + n_2)^2 \mathbf{1}_p^T \mathcal{S} \mathbf{1}_p} > F_{1, n_1 + n_2 - 2}(\alpha)$$

上 α 分位数

线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 组间效应

✓ 若 $H_0^{(1)}$ 与 $H_0^{(2)}$ 均已接受, 则, 我们可以利用两组中包含的信息来检验组间效应, 即, 两个轮廓是否均为水平线. 这可以表示为

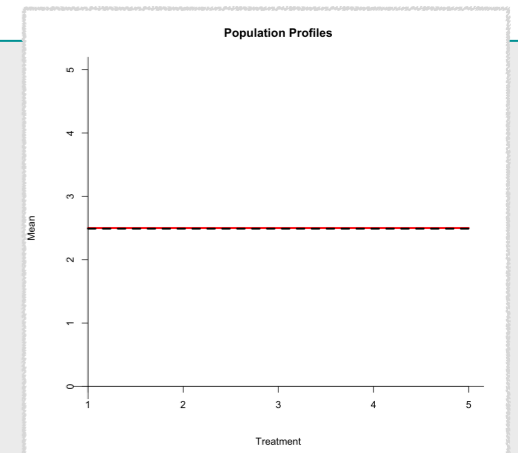
$$H_0^{(3)} : \mathcal{C}(\mu_1 + \mu_2) = 0$$

$$\begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}_{(p-1) \times p} \begin{pmatrix} 2\mu_{11} \\ 2\mu_{21} \\ \vdots \\ 2\mu_{p1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \implies \begin{pmatrix} \mu_{11} - \mu_{21} \\ \mu_{21} - \mu_{31} \\ \vdots \\ \mu_{p-11} - \mu_{p1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\implies \mu_{11} = \mu_{21} = \dots = \mu_{p-11} = \mu_{p1}$$

```

y4 <- rep(2.5, 5)
plot(c(1, 5), c(0, 5), type = 'n', axes = FALSE, xlab = 'Treatment', ylab = 'Mean', main = 'Population Profiles')
abline(h=0, v=1)
lines(x, y4, lty=1, lwd=3, col = 'red')
lines(x, y4-0.01, lty=2, lwd=3, col = 'black')
axis(1, pos = 0)
axis(2, pos = 1)
    
```



线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 组间效应

- ✓ 若 $H_0^{(1)}$ 与 $H_0^{(2)}$ 均已接受, 则, 我们可以利用两组中包含的信息来检验组间效应, 即, 两个轮廓是否均为水平线. 这可以表示为

$$H_0^{(3)} : \mathcal{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \mathbf{0}$$

- ✓ 考虑平均轮廓 $\bar{\mathbf{x}}$

$$\bar{\mathbf{x}} = \frac{n_1 \bar{\mathbf{x}}_1 + n_2 \bar{\mathbf{x}}_2}{n_1 + n_2} \implies \bar{\mathbf{x}} \sim N_p \left(\frac{n_1 \boldsymbol{\mu}_1 + n_2 \boldsymbol{\mu}_2}{n_1 + n_2}, \frac{1}{n_1 + n_2} \boldsymbol{\Sigma} \right)$$

在 $H_0^{(3)}$ 与 $H_0^{(1)}$ 为真时 $\implies \mathcal{C} \left(\frac{n_1 \boldsymbol{\mu}_1 + n_2 \boldsymbol{\mu}_2}{n_1 + n_2} \right) = \mathbf{0}$

在 $H_0^{(3)}$ 与 $H_0^{(1)}$ 为真时 $\implies \sqrt{n_1 + n_2} \mathcal{C} \bar{\mathbf{x}} \sim N_p(\mathbf{0}, \mathcal{C} \boldsymbol{\Sigma} \mathcal{C}^T)$

线性假设

- 轮廓分析 (Profile Analysis)

- ▶ 组间效应

✓ 若 $H_0^{(1)}$ 与 $H_0^{(2)}$ 均已接受, 则, 我们可以利用两组中包含的信息来检验组间效应, 即, 两个轮廓是否均为水平线. 这可以表示为

$$H_0^{(3)} : \mathcal{C} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = 0$$

$$(n_1 + n_2 - 2) (\mathcal{C}\bar{\mathbf{x}})^T (\mathcal{C}\mathcal{S}\mathcal{C}^T)^{-1} \mathcal{C}\bar{\mathbf{x}} \sim T_{p-1, n_1+n_2-p}^2$$

✓ $H_0^{(3)}$ 的拒绝域为

$$\frac{n_1 + n_2 - p}{p - 1} (\mathcal{C}\bar{\mathbf{x}})^T (\mathcal{C}\mathcal{S}\mathcal{C}^T)^{-1} \mathcal{C}\bar{\mathbf{x}} > F_{p-1, n_1+n_2-p}(\alpha)$$

上 α 分位数

在 $H_0^{(3)}$ 与 $H_0^{(1)}$ 为真时



$$\sqrt{n_1 + n_2} \mathcal{C}\bar{\mathbf{x}} \sim N_p(\mathbf{0}, \mathcal{C}\Sigma\mathcal{C}^T)$$

推论 5.4 考虑 $X \sim N_p(\boldsymbol{\mu}, \Sigma)$ 的一个线性变换 $Y = \mathcal{C}X$, 其中 $\mathcal{C}_{q \times p}$ 满足 $q \leq p$. 如果

$\bar{\mathbf{x}}$ 与 \mathcal{S}_X 分别为样本均值向量与样本协方差矩阵, 则有

$$\bar{\mathbf{y}} = \mathcal{C}\bar{\mathbf{x}} \sim N_q\left(\mathcal{C}\boldsymbol{\mu}, \frac{1}{n}\mathcal{C}\Sigma\mathcal{C}^T\right)$$

$$n\mathcal{S}_Y = n\mathcal{C}\mathcal{S}_X\mathcal{C}^T \sim W_q(\mathcal{C}\Sigma\mathcal{C}^T, n-1)$$

$$(n-1)(\mathcal{C}\bar{\mathbf{x}} - \mathcal{C}\boldsymbol{\mu})^T (\mathcal{C}\mathcal{S}_X\mathcal{C}^T)^{-1} (\mathcal{C}\bar{\mathbf{x}} - \mathcal{C}\boldsymbol{\mu}) \sim T_{q, n-1}^2$$

Boston 房屋数据

```
library(MASS)
```

```
str(Boston)
```

506 个观测值 ←

→ 14 个变量

```
> str(Boston)
'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 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 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 房屋数据

X_{14} : 自住房屋价格的中位数(单位: 1000美元)

head(Boston)

```
> 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_4 : Charles 河 (1 指河边, 0 为其它)

X_6 : 每个住宅的平均房间数

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

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 房屋数据

- 由于大多数变量都呈现出不对称，且左侧的密度更高，所以进行了以下变换：

$$\widetilde{X}_1 = \log(X_1)$$

$$\widetilde{X}_2 = \frac{X_2}{10}$$

$$\widetilde{X}_3 = \log(X_3)$$

$$\widetilde{X}_4 = X_4, \text{ 不作变换, 因其为二值变量}$$

$$\widetilde{X}_5 = \log(X_5)$$

$$\widetilde{X}_6 = \log(X_6)$$

$$\widetilde{X}_7 = \frac{X_7^{2.5}}{10000}$$

$$\widetilde{X}_8 = \log(X_8)$$

$$\widetilde{X}_9 = \log(X_9)$$

$$\widetilde{X}_{10} = \log(X_{10})$$

$$\widetilde{X}_{11} = \frac{\exp(0.4 \times X_{11})}{1000}$$

$$\widetilde{X}_{12} = \frac{X_{12}}{100}$$

$$\widetilde{X}_{13} = \sqrt{X_{13}}$$

$$\widetilde{X}_{14} = \log(X_{14})$$

```

X_1 = log(Boston$crim)
X_2 = Boston$zn / 10
X_3 = log(Boston$indus)
X_4 = Boston$chas
X_5 = log(Boston$nox)
X_6 = log(Boston$rm)
X_7 = Boston$age^(2.5) / 10000
X_8 = log(Boston$dis)
X_9 = log(Boston$rad)
X_10 = log(Boston$tax)
X_11 = (exp(0.4 * Boston$ptratio)) / 1000
X_12 = Boston$black / 100
X_13 = sqrt(Boston$lstat)
X_14 = log(Boston$medv)
  
```

以下使用变换后的数据

Boston 房屋数据

- 分组：我们曾利用 $X_{14} \leq \text{median}(X_{14})$ 的中位数或 $X_{14} > \text{median}(X_{14})$ 的中位数将数据集分为两组。
- 检验两组均值是否相等：仅考虑变量 $X_1, X_5, X_8, X_{11}, X_{13}$
 - ▶ 欲检验 $H_0: \mu_1 = \mu_2$, 其中 $\mu_1, \mu_2 \in \mathbb{R}^5$

检验问题 8: 假设 $X_{i1} \sim N_p(\mu_1, \Sigma)$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\mu_2, \Sigma)$, $j = 1, 2, \dots, n_2$, 其中所有的变量相互独立.

$$H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \text{无约束}$$

▶ $H_0: \mu_1 = \mu_2$ 的拒绝域为

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^T \mathcal{S}^{-1} (\bar{x}_1 - \bar{x}_2) \geq F_{p, n_1 + n_2 - p - 1}(\alpha)$$

上 α 分位数

$$\mathcal{S} = \frac{1}{n_1 + n_2} (n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2)$$

```

boston = data.frame(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_10, X_11, X_12, X_13, X_14)
names(boston) = names(Boston)
boston_sub_group1 = subset(boston, boston$medv <= median(boston$medv), select = c(1, 5, 8, 11, 13))
boston_sub_group2 = subset(boston, boston$medv > median(boston$medv), select = c(1, 5, 8, 11, 13))
n1 = dim(boston_sub_group1)[1]
p = dim(boston_sub_group1)[2]
n2 = dim(boston_sub_group2)[1]
mu_1 = sapply(boston_sub_group1, mean)
mu_2 = sapply(boston_sub_group2, mean)
S_1 = cov(boston_sub_group1) * (n1 - 1) / n1
S_2 = cov(boston_sub_group2) * (n2 - 1) / n2
S = (n1 * S_1 + n2 * S_2) / (n1 + n2)
S = as.matrix(S)
S_inv = solve(S)
F_test = t(as.matrix(xbar_1 - xbar_2)) %*% S_inv %*% as.matrix(xbar_1 - xbar_2) * n1 * n2 * (n1 + n2 - p - 1) / (p * (n1 + n2)^2)
F_test
  
```

```

> F_test
      [,1]
[1,] 126.3005
  
```

Boston 房屋数据

- 分组：我们曾利用 $X_{14} \leq \text{中位数}$ 或 $X_{14} > \text{中位数}$ 将数据集分为两组。
- 检验两组均值是否相等：仅考虑变量 $X_1, X_5, X_8, X_{11}, X_{13}$

- ▶ 欲检验 $H_0: \mu_1 = \mu_2$, 其中 $\mu_1, \mu_2 \in \mathbb{R}^5$
- ▶ 计算得检验统计量的值为
- ▶ 查得检验的临界值为
- ▶ **结论**：拒绝 $H_0: \mu_1 = \mu_2$

检验问题 8: 假设 $X_{i1} \sim N_p(\mu_1, \Sigma)$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\mu_2, \Sigma)$, $j = 1, 2, \dots, n_2$, 其中所有的变量相互独立。

$$H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \text{无约束}$$

- ▶ $H_0: \mu_1 = \mu_2$ 的拒绝域为

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^T \mathcal{S}^{-1} (\bar{x}_1 - \bar{x}_2) \geq F_{p, n_1 + n_2 - p - 1}(\alpha)$$

上 α 分位数

$$\mathcal{S} = \frac{1}{n_1 + n_2} (n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2)$$

$$F = 126.3005 > 2.232042 = F_{p, n_1 + n_2 - p - 1}(0.05)$$

```
> F_test
      [,1]
[1,] 126.3005
```

```
qf(0.05, p, n1 + n2 - p - 1, lower.tail = FALSE)
```

```
> qf(0.05, p, n1 + n2 - p - 1, lower.tail = FALSE)
[1] 2.232042
```

Boston 房屋数据

- 分组：我们曾利用 $X_{14} \leq X_{14}$ 的中位数或 $X_{14} > X_{14}$ 的中位数将数据集分为两组。
- 检验两组均值是否相等：仅考虑变量 $X_1, X_5, X_8, X_{11}, X_{13}$

▶ $\delta_j = \mu_{1j} - \mu_{2j}, j = 1, 5, 8, 11, 13$ 的置信度为 95% 的联合置信区间为

$$\delta_1 \in (1.4019611, 2.5498855)$$

$$\delta_5 \in (0.1314898, 0.2383455)$$

$$\delta_8 \in (-0.5243907, -0.2221791)$$

$$\delta_{11} \in (1.0374981, 1.7383789)$$

$$\delta_{13} \in (1.1576946, 1.5818046)$$

- ▶ 联合置信区间确认了所有的 δ_j 均显著不为零。
- ▶ 注意： X_8 (到波士顿五个就业中心的加权距离) 具有负效应。

检验问题 8: 假设 $X_{i1} \sim N_p(\mu_1, \Sigma), i = 1, 2, \dots, n_1, X_{j2} \sim N_p(\mu_2, \Sigma), j = 1, 2, \dots, n_2$, 其中所有的变量相互独立。

$$H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \text{无约束}$$

▶ 特别, 对 $j = 1, 2, \dots, p$, 置信度为 $1 - \alpha$ 时, 我们有

$$\delta_j = \mu_{1j} - \mu_{2j} \in (\bar{x}_{1j} - \bar{x}_{2j}) \pm \sqrt{\frac{p(n_1 + n_2)^2}{n_1 n_2 (n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha) \cdot s_{jj}}$$

```

> data.frame(mu_L = mu_L, mu_R = mu_R)
      mu_L      mu_R
crim  1.4019611  2.5498855
nox    0.1314898  0.2383455
dis   -0.5243907 -0.2221791
ptratio 1.0374981  1.7383789
lstat  1.1576946  1.5818046
    
```

```

mu_L = as.matrix(xbar_1 - xbar_2) -
  sqrt(qf(0.05, p, n1 + n2 - p - 1, lower.tail = FALSE) * (p * (n1 + n2)^2) / (n1 * n2 * (n1 + n2 - p - 1))) *
  as.matrix(c(S[1, 1], S[2, 2], S[3, 3], S[4, 4], S[5, 5]))
mu_R = as.matrix(xbar_1 - xbar_2) +
  sqrt(qf(0.05, p, n1 + n2 - p - 1, lower.tail = FALSE) * (p * (n1 + n2)^2) / (n1 * n2 * (n1 + n2 - p - 1))) *
  as.matrix(c(S[1, 1], S[2, 2], S[3, 3], S[4, 4], S[5, 5]))
data.frame(mu_L = mu_L, mu_R = mu_R)
    
```

Boston 房屋数据

- 分组：我们还可以检验是否“河景房” X_4 对其它变量是否有影响。
- 检验两组均值是否相等：考虑变量 $X_5, X_8, X_9, X_{12}, X_{13}, X_{14}$

▶ 欲检验 $H_0: \mu_1 = \mu_2$, 其中 $\mu_1, \mu_2 \in \mathbb{R}^6$

检验问题 8: 假设 $X_{i1} \sim N_p(\mu_1, \Sigma)$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\mu_2, \Sigma)$, $j = 1, 2, \dots, n_2$, 其中所有的变量相互独立。

$$H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \text{无约束}$$

▶ $H_0: \mu_1 = \mu_2$ 的拒绝域为

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^T \mathcal{S}^{-1} (\bar{x}_1 - \bar{x}_2) \geq F_{p, n_1 + n_2 - p - 1}(\alpha)$$

上 α 分位数

$$\mathcal{S} = \frac{1}{n_1 + n_2} (n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2)$$

```

n3 = dim(boston_sub_group3)[1]
p = dim(boston_sub_group3)[2]
n4 = dim(boston_sub_group4)[1]
xbar_3 = sapply(boston_sub_group3, mean)
xbar_4 = sapply(boston_sub_group4, mean)
S_3 = cov(boston_sub_group3) * (n3 - 1) / n3
S_4 = cov(boston_sub_group4) * (n4 - 1) / n4
S = (n3 * S_3 + n4 * S_4) / (n3 + n4)
S = as.matrix(S)
S_inv = solve(S)
F_test = t(as.matrix(xbar_3 - xbar_4)) %*% S_inv %*% as.matrix(xbar_3 - xbar_4) * n3 * n4 * (n3 + n4 - p - 1) / (p * (n3 + n4)^2)
F_test
    
```

```

> F_test
      [,1]
[1,] 5.814844
    
```

Boston 房屋数据

- 分组：我们还可以检验是否“河景房” X_4 对其它变量是否有影响。
- 检验两组均值是否相等：考虑变量 $X_5, X_8, X_9, X_{12}, X_{13}, X_{14}$

- ▶ 欲检验 $H_0 : \mu_1 = \mu_2$ ，其中 $\mu_1, \mu_2 \in \mathbb{R}^6$
- ▶ 计算得检验统计量的值为
- ▶ 查得检验的临界值为
- ▶ **结论**：拒绝 $H_0 : \mu_1 = \mu_2$

检验问题 8: 假设 $X_{i1} \sim N_p(\mu_1, \Sigma)$, $i = 1, 2, \dots, n_1$, $X_{j2} \sim N_p(\mu_2, \Sigma)$, $j = 1, 2, \dots, n_2$, 其中所有的变量相互独立。

$$H_0 : \mu_1 = \mu_2 \leftrightarrow H_1 : \text{无约束}$$

- ▶ $H_0 : \mu_1 = \mu_2$ 的拒绝域为

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p (n_1 + n_2)^2} (\bar{x}_1 - \bar{x}_2)^T \mathcal{S}^{-1} (\bar{x}_1 - \bar{x}_2) \geq F_{p, n_1 + n_2 - p - 1}(\alpha)$$

上 α 分位数

$$\mathcal{S} = \frac{1}{n_1 + n_2} (n_1 \mathcal{S}_1 + n_2 \mathcal{S}_2)$$

$$F = 5.814844 > 2.116738 = F_{p, n_1 + n_2 - p - 1}(0.05)$$

```
> F_test
      [,1]
[1,] 5.814844
```

```
qf(0.05, p, n3 + n4 - p - 1, lower.tail = FALSE)
```

```
> qf(0.05, p, n3 + n4 - p - 1, lower.tail = FALSE)
[1] 2.116738
```

Boston 房屋数据

- 分组：我们还可以检验是否“河景房” X_4 对其它变量是否有影响。
- 检验两组均值是否相等：考虑变量 $X_5, X_8, X_9, X_{12}, X_{13}, X_{14}$

▶ $\delta_j = \mu_{1j} - \mu_{2j}, j = 5, 8, 9, 12, 13, 14$ 的置信度为 95% 的联合置信区间为

$$\delta_5 \in (-0.060322746, 0.1919118)$$

$$\delta_8 \in (-0.522497204, 0.1527192)$$

$$\delta_9 \in (-0.505094923, 0.5937961)$$

$$\delta_{12} \in (-0.397384925, 0.7481127)$$

$$\delta_{13} \in (-0.859549918, 0.3781680)$$

$$\delta_{14} \in (0.001432809, 0.5084371)$$

▶ 联合置信区间表明，只有 X_{14} 导致了拒绝 H_0 。

检验问题 8: 假设 $X_{i1} \sim N_p(\mu_1, \Sigma), i = 1, 2, \dots, n_1, X_{j2} \sim N_p(\mu_2, \Sigma), j = 1, 2, \dots, n_2$, 其中所有的变量相互独立。

$$H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \text{无约束}$$

▶ 特别，对 $j = 1, 2, \dots, p$, 置信度为 $1 - \alpha$ 时，我们有

$$\delta_j = \mu_{1j} - \mu_{2j} \in (\bar{x}_{1j} - \bar{x}_{2j}) \pm \sqrt{\frac{p(n_1 + n_2)^2}{n_1 n_2 (n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha) \cdot s_{jj}}$$

```

> data.frame(mu_L = mu_L, mu_R = mu_R)
      mu_L      mu_R
nox -0.060322746 0.1919118
dis -0.522497204 0.1527192
rad -0.505094923 0.5937961
black -0.397384925 0.7481127
lstat -0.859549918 0.3781680
medv 0.001432809 0.5084371
    
```

```

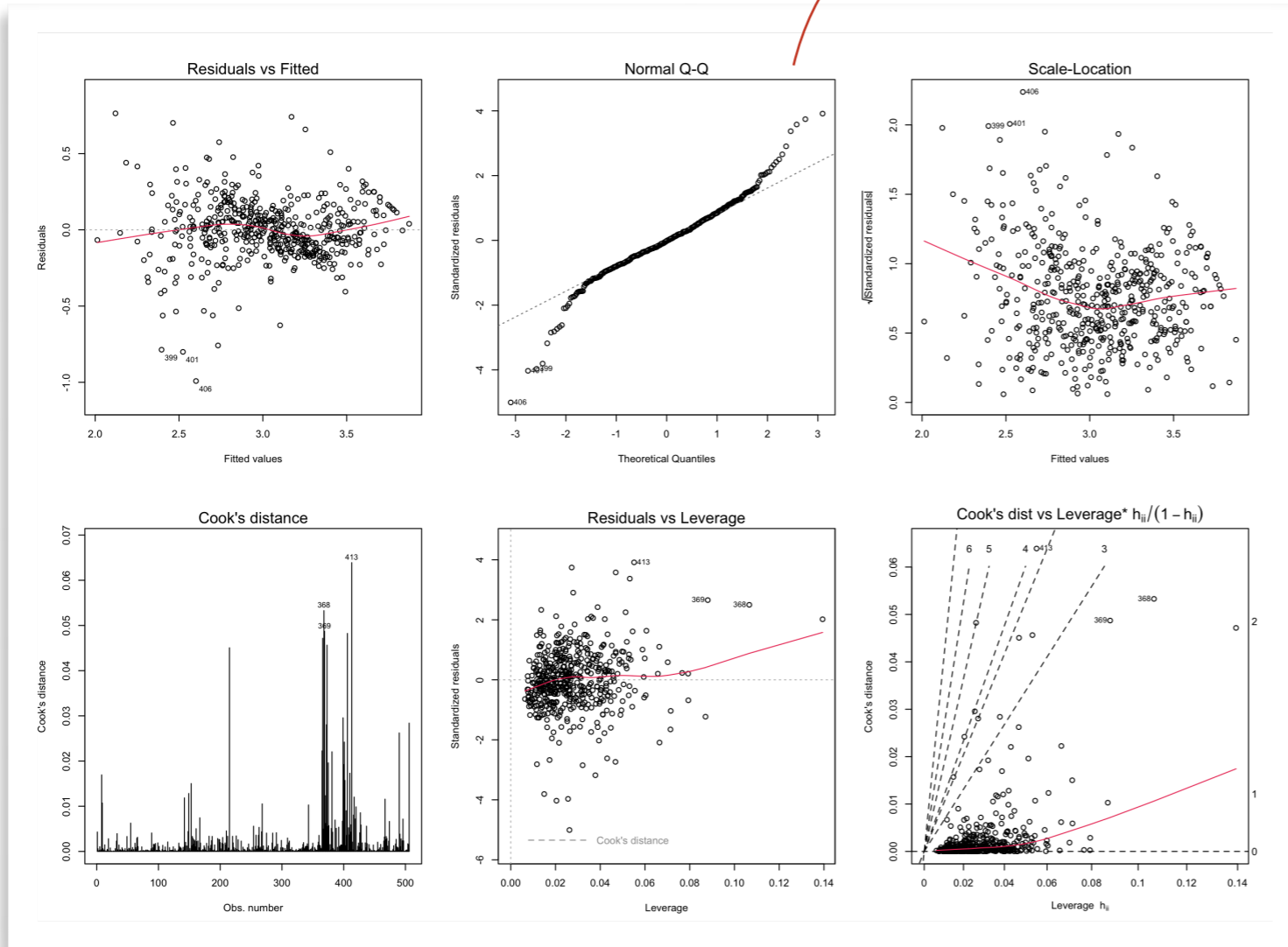
mu_L = as.matrix(xbar_3 - xbar_4) -
  sqrt(qf(0.05, p, n3 + n4 - p - 1, lower.tail = FALSE) * (p * (n3 + n4)^2) / (n3 * n4 * (n3 + n4 - p - 1))) *
  as.matrix(c(S[1, 1], S[2, 2], S[3, 3], S[4, 4], S[5, 5], S[6, 6]))
mu_R = as.matrix(xbar_3 - xbar_4) +
  sqrt(qf(0.05, p, n3 + n4 - p - 1, lower.tail = FALSE) * (p * (n3 + n4)^2) / (n3 * n4 * (n3 + n4 - p - 1))) *
  as.matrix(c(S[1, 1], S[2, 2], S[3, 3], S[4, 4], S[5, 5], S[6, 6]))
data.frame(mu_L = mu_L, mu_R = mu_R)
    
```

Boston 房屋数据

● 检验线性约束

▶ 我们在第 3 章建立了房价 X_{14} 利用其它变量解释的一个线性模型。

独立、正态假设基本成立



```
> summary(X_14_lm)
```

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

```
Residuals:
    Min       1Q   Median       3Q      Max
-0.9918 -0.1002 -0.0034  0.1117  0.7640
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.176874   0.379017  11.020 < 2e-16 ***
    crim      -0.014606   0.011650  -1.254 0.210527
    zn         0.001392   0.005639   0.247 0.805121
    indus     -0.012709   0.022312  -0.570 0.569195
    chas       0.109980   0.036634   3.002 0.002817 **
    nox       -0.283112   0.105340  -2.688 0.007441 **
    rm         0.421108   0.110175   3.822 0.000149 ***
    age        0.006403   0.004863   1.317 0.188536
    dis       -0.183154   0.036804  -4.977 8.97e-07 ***
    rad        0.068362   0.022473   3.042 0.002476 **
    tax       -0.201832   0.048432  -4.167 3.64e-05 ***
    ptratio   -0.040017   0.008091  -4.946 1.04e-06 ***
    black     0.044472   0.011456   3.882 0.000118 ***
    lstat     -0.262615   0.016091 -16.320 < 2e-16 ***
```

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

```
Residual standard error: 0.2008 on 492 degrees of freedom
Multiple R-squared:  0.765,    Adjusted R-squared:  0.7588
F-statistic: 123.2 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
X_14_lm = lm(medv ~ ., data = boston)
```

```
summary(X_14_lm)
```

```
par(mfrow = c(2, 3))
```

```
plot(X_14_lm, which = 1:6)
```

Boston 房屋数据

- 检验线性约束

▶ 全部回归系数的显著性检验 $H_0: (\beta_1, \beta_2, \dots, \beta_{13})^T = \mathbf{0} \iff \mathcal{A}\beta = \mathbf{0}$

$$\beta \triangleq \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{13} \end{pmatrix}, \quad \mathcal{A} \triangleq (\mathbf{0}_{13}, \mathcal{F}_{13}) = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}_{13 \times 14}$$

```
X = as.matrix(model.matrix(X_14_lm))
```

```
n = dim(boston)[1]
```

```
p = dim(boston)[2] - 1
```

```
q = p
```

```
beta_hat = as.matrix(X_14_lm$coefficients)
```

```
y = as.matrix(boston$medv)
```

```
A = array(0, dim = c(13, 14))
```

```
for (i in 1:13) A[i, i+1] = 1
```

```
F_test = (t(A %*% beta_hat) %*% solve(A %*% solve(t(X) %*% X) %*% t(A)) %*% (A %*% beta_hat) /
```

```
t(y - X %*% beta_hat) %*% (y - X %*% beta_hat)) * (n - p) / q
```

```
F_test
```

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$H_0: \mathcal{A}\beta = a \iff H_1: \text{无约束}$

▶ 利用检验问题 4 中我们曾用到的结果, 该问题也有精确的 F 检验

$$F = \frac{n-p}{q} \left(\frac{\|y - \mathcal{X}\hat{\beta}\|^2}{\|y - \hat{\beta}\|^2} - 1 \right) = \frac{n-p}{q} \frac{(\mathcal{A}\hat{\beta} - a)^T [\mathcal{A}(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{A}^T]^{-1} (\mathcal{A}\hat{\beta} - a)}{(y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})} \sim F_{q, n-p}$$

▶ 此时 H_0 的拒绝域为

$$R = \{ \mathcal{X} : F > F_{q, n-p}(\alpha) \}$$

→ $F_{q, n-p}(\alpha)$ 是上 α 分位数

```
> F_test
      [,1]
[1,] 123.452
```

Boston 房屋数据

- 检验线性约束

▶ 全部回归系数的显著性检验 $H_0 : (\beta_1, \beta_2, \dots, \beta_{13})^T = \mathbf{0} \iff \mathcal{A}\beta = \mathbf{0}$

$$\beta \triangleq \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{13} \end{pmatrix}, \quad \mathcal{A} \triangleq (\mathbf{0}_{13}, \mathcal{F}_{13}) = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}_{13 \times 14}$$

▶ 计算得检验统计量的值为

▶ 查得检验的临界值为

▶ **结论:** 拒绝 $H_0 : (\beta_1, \beta_2, \dots, \beta_{13})^T = \mathbf{0}$

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$H_0 : \mathcal{A}\beta = a \leftrightarrow H_1 : \text{无约束}$

▶ 利用检验问题 4 中我们曾用到的结果, 该问题也有精确的 F 检验

$$F = \frac{n-p}{q} \left(\frac{\|y - \mathcal{X}\hat{\beta}\|^2}{\|y - \hat{\beta}\|^2} - 1 \right) = \frac{n-p}{q} \frac{(\mathcal{A}\hat{\beta} - a)^T [\mathcal{A}(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{A}^T]^{-1} (\mathcal{A}\hat{\beta} - a)}{(y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})} \sim F_{q, n-p}$$

▶ 此时 H_0 的拒绝域为

$$R = \{ \mathcal{X} : F > F_{q, n-p}(\alpha) \}$$

$F_{q, n-p}(\alpha)$ 是上 α 分位数

$$F = 123.452 > 1.740034 = F_{q, n-p}(0.05)$$

qf(0.05, q, n - p, lower.tail = FALSE)

```
> qf(0.05, q, n - p, lower.tail = FALSE)
[1] 1.740034
```

```
> F_test
      [,1]
[1,] 123.452
```

Boston 房屋数据

- 检验线性约束

是否位于湖边 (X_4) 对房价的影响, 我们来检验 $H_0: \beta_4 = 0 \iff \mathcal{A}\beta = 0$

$$\beta \triangleq \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{13} \end{pmatrix}, \quad \mathcal{A} \triangleq (0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0)$$

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$H_0: \mathcal{A}\beta = a \iff H_1: \text{无约束}$

利用检验问题 4 中我们曾用到的结果, 该问题也有精确的 F 检验

$$F = \frac{n-p}{q} \left(\frac{\|y - \mathcal{X}\hat{\beta}\|^2}{\|y - \mathcal{X}\hat{\beta}\|^2} - 1 \right) = \frac{n-p}{q} \frac{(\mathcal{A}\hat{\beta} - a)^T [\mathcal{A}(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{A}^T]^{-1} (\mathcal{A}\hat{\beta} - a)}{(y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})} \sim F_{q, n-p}$$

此时 H_0 的拒绝域为

$$R = \{ \mathcal{X} : F > F_{q, n-p}(\alpha) \}$$

$\rightarrow F_{q, n-p}(\alpha)$ 是上 α 分位数

```

q = 1
A = array(0, dim = c(1, 14))
A[5] = 1
A = as.matrix(A)
F_test = (t(A %*% beta_hat) %*% solve(A %*% solve(t(X) %*% X) %*% t(A)) %*% (A %*% beta_hat) /
          t(y - X %*% beta_hat) %*% (y - X %*% beta_hat)) * (n - p) / q
F_test
    
```

```

> F_test
           [,1]
[1,] 9.031265
    
```

Boston 房屋数据

- 检验线性约束

是否位于湖边 (X_4) 对房价的影响, 我们来检验 $H_0: \beta_4 = 0 \iff \mathcal{A}\beta = 0$

$$\beta \triangleq \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{13} \end{pmatrix}, \quad \mathcal{A} \triangleq (0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0)$$

计算得检验统计量的值为

查得检验的临界值为

结论: 拒绝 $H_0: \beta_4 = 0$

$$F = 9.031265 > 3.86039 = F_{q, n-p}(0.05)$$

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$$H_0: \mathcal{A}\beta = a \iff H_1: \text{无约束}$$

利用检验问题 4 中我们曾用到的结果, 该问题也有精确的 F 检验

$$F = \frac{n-p}{q} \left(\frac{\|y - \mathcal{X}\hat{\beta}\|^2}{\|y - \hat{\beta}\|^2} - 1 \right) = \frac{n-p}{q} \frac{(\mathcal{A}\hat{\beta} - a)^T [\mathcal{A}(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{A}^T]^{-1} (\mathcal{A}\hat{\beta} - a)}{(y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})} \sim F_{q, n-p}$$

此时 H_0 的拒绝域为

$$R = \{ \mathcal{X} : F > F_{q, n-p}(\alpha) \}$$

$\rightarrow F_{q, n-p}(\alpha)$ 是上 α 分位数

```
> F_test
      [,1]
[1,] 9.031265
```

```
qf(0.05, q, n - p, lower.tail = FALSE)
```

```
> qf(0.05, q, n - p, lower.tail = FALSE)
[1] 3.86039
```

Boston 房屋数据

- 检验线性约束

- ▶ 我们看到全模型当中有部分变量不显著 (p 值较高).
- ▶ 我们来检验 $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_7 = 0$

```
> summary(X_14_lm)

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

Residuals:
    Min       1Q   Median       3Q      Max
-0.9918 -0.1002 -0.0034  0.1117  0.7640

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.176874   0.379017  11.020 < 2e-16 ***
crim         -0.014606   0.011650  -1.254 0.210527
zn           0.001392   0.005639   0.247 0.805121
indus       -0.012709   0.022312  -0.570 0.569195
chas         0.109980   0.036634   3.002 0.002817 **
nox         -0.283112   0.105340  -2.688 0.007441 **
rm           0.421108   0.110175   3.822 0.000149 ***
age          0.006403   0.004863   1.317 0.188536
dis         -0.183154   0.036804  -4.977 8.97e-07 ***
rad          0.068362   0.022473   3.042 0.002476 **
tax         -0.201832   0.048432  -4.167 3.64e-05 ***
ptratio     -0.040017   0.008091  -4.946 1.04e-06 ***
black        0.044472   0.011456   3.882 0.000118 ***
lstat       -0.262615   0.016091 -16.320 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2008 on 492 degrees of freedom
Multiple R-squared:  0.765,    Adjusted R-squared:  0.7588
F-statistic: 123.2 on 13 and 492 DF,  p-value: < 2.2e-16
```

Boston 房屋数据

- 检验线性约束

- ▶ 我们看到全模型当中有部分变量不显著 (p 值较高).

- ▶ 我们来检验 $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_7 = 0 \iff \mathcal{A}\beta = \mathbf{0}_{4 \times 1}$

$$\beta \triangleq \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{13} \end{pmatrix}, \quad \mathcal{A} \triangleq \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

q = 4

```
A = array(0, dim = c(4, 14))
```

```
for (i in 1:3) A[i, i+1] = 1
```

```
A[4, 8] = 1
```

```
A = as.matrix(A)
```

```
F_test = (t(A %*% beta_hat) %*% solve(A %*% solve(t(X) %*% X) %*% t(A)) %*% (A %*% beta_hat) /
```

```
t(y - X %*% beta_hat) %*% (y - X %*% beta_hat)) * (n - p) / q
```

```
F_test
```

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$H_0 : \mathcal{A}\beta = a \iff H_1 : \text{无约束}$

- ▶ 利用检验问题 4 中我们曾用到的结果, 该问题也有精确的 F 检验

$$F = \frac{n-p}{q} \left(\frac{\|y - \mathcal{X}\tilde{\beta}\|^2}{\|y - \mathcal{X}\hat{\beta}\|^2} - 1 \right) = \frac{n-p}{q} \frac{(\mathcal{A}\hat{\beta} - a)^T [\mathcal{A}(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{A}^T]^{-1} (\mathcal{A}\hat{\beta} - a)}{(y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})} \sim F_{q, n-p}$$

- ▶ 此时 H_0 的拒绝域为

$$R = \{ \mathcal{X} : F > F_{q, n-p}(\alpha) \}$$

$\rightarrow F_{q, n-p}(\alpha)$ 是上 α 分位数

```
> F_test
```

```

          [,1]
[1,] 0.9363377
    
```

Boston 房屋数据

● 检验线性约束

▶ 我们看到全模型当中有部分变量不显著 (p 值较高).

▶ 我们来检验 $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_7 = 0 \iff \mathcal{A}\beta = \mathbf{0}_{4 \times 1}$

$$\beta \triangleq \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{13} \end{pmatrix}, \quad \mathcal{A} \triangleq \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

▶ 计算得检验统计量的值为

▶ 查得检验的临界值为

▶ **结论:** 接受 $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_7 = 0$

检验问题 7: 假设 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N_1(\beta^T x_i, \sigma^2)$, $x_i \in \mathbb{R}^p$, 其中 σ^2 未知.

$$H_0 : \mathcal{A}\beta = a \iff H_1 : \text{无约束}$$

▶ 利用检验问题 4 中我们曾用到的结果, 该问题也有精确的 F 检验

$$F = \frac{n-p}{q} \left(\frac{\|y - \mathcal{X}\tilde{\beta}\|^2}{\|y - \mathcal{X}\hat{\beta}\|^2} - 1 \right) = \frac{n-p}{q} \frac{(\mathcal{A}\hat{\beta} - a)^T [\mathcal{A}(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{A}^T]^{-1} (\mathcal{A}\hat{\beta} - a)}{(y - \mathcal{X}\hat{\beta})^T (y - \mathcal{X}\hat{\beta})} \sim F_{q, n-p}$$

▶ 此时 H_0 的拒绝域为

$$R = \{ \mathcal{X} : F > F_{q, n-p}(\alpha) \}$$

$\rightarrow F_{q, n-p}(\alpha)$ 是上 α 分位数

$$F = 0.9363377 < 2.390021 = F_{q, n-p}(0.05)$$

```
qf(0.05, q, n - p, lower.tail = FALSE)
```

```
> qf(0.05, q, n - p, lower.tail = FALSE)
[1] 2.390021
```

```
> F_test
           [,1]
[1,] 0.9363377
```

Boston 房屋数据

- 检验线性约束

- ▶ 从而得到一个可能的简化模型如下

$$\begin{aligned}
 X_{14} = & 4.158186 + 0.108655 X_4 \\
 & -0.305541 X_5 + 0.466807 X_6 \\
 & -0.185537 X_8 + 0.049184 X_9 \\
 & -0.209594 X_{10} - 0.041047 X_{11} \\
 & +0.048139 X_{12} - 0.258773 X_{13}
 \end{aligned}$$

- ▶ **结论:** 接受 $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_7 = 0$

```
> summary(X_14_lm_reduced)
```

Call:

```
lm(formula = medv ~ chas + nox + rm + dis + rad + tax + ptratio +
    black + lstat, data = boston)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-1.01206 -0.10847 -0.00662  0.11665  0.77767
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.158186	0.362788	11.462	< 2e-16	***
chas	0.108655	0.036227	2.999	0.00284	**
nox	-0.305541	0.097307	-3.140	0.00179	**
rm	0.466807	0.105917	4.407	1.28e-05	***
dis	-0.185537	0.032671	-5.679	2.31e-08	***
rad	0.049184	0.018286	2.690	0.00739	**
tax	-0.209594	0.044550	-4.705	3.30e-06	***
ptratio	-0.041047	0.007774	-5.280	1.94e-07	***
black	0.048139	0.011180	4.306	2.01e-05	***
lstat	-0.258773	0.014875	-17.396	< 2e-16	***

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

Residual standard error: 0.2007 on 496 degrees of freedom
 Multiple R-squared: 0.7632, Adjusted R-squared: 0.7589
 F-statistic: 177.6 on 9 and 496 DF, p-value: < 2.2e-16

```
X_14_lm_reduced = lm(medv ~ chas + nox + rm + dis + rad + tax + ptratio + black + lstat, data = boston)
summary(X_14_lm_reduced)
```