中国石油大学(北京)克拉玛依校区
CHINA UNIVERSITY OF PETROLEUM - BEIJING AT KARAMAY

Multivariate Statistical Analysis

多元统计分析

2026 年 4 月 23 日

已学知识点 (Recap)

第 7 章 假设检验

- 问题：$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} F(\boldsymbol{x};\ \boldsymbol{\theta}),\ \boldsymbol{x}_i \in \mathscr{X} \subset \mathbb{R}^p$，在以下两个假设中作出决策

总体　样本空间　样本

$$H_0 : \boldsymbol{\theta} \in \Omega_0 \longleftrightarrow H_1 : \boldsymbol{\theta} \in \Omega_1$$

零假设　对应某一简化模型　　备择假设　对应某一全模型

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

犯第 I 类错误的概率　检验的显著水平 (事先指定)

$$\sup_{\boldsymbol{\theta} \in \Omega_0} \mathrm{P}\left(\mathscr{X} \in R \middle| \boldsymbol{\theta}\right) \leq \alpha$$

确定临界值 c

- 似然比：

$$\lambda(\mathscr{X}) \triangleq \frac{L_0^*}{L_1^*} = \frac{\max\limits_{\boldsymbol{\theta} \in \Omega_0} L(\mathscr{X};\ \boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta} \in \Omega_1} L(\mathscr{X};\ \boldsymbol{\theta})}$$

似然函数

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

因为 $\lambda(\mathscr{X})$ 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：$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} F(\boldsymbol{x};\ \boldsymbol{\theta}),\ \boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^p$，在以下两个假设中作出决策

总体 ← $F(\boldsymbol{x};\ \boldsymbol{\theta})$；样本空间 ← $\mathcal{X}$；样本 ← $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$

$$H_0 : \boldsymbol{\theta} \in \Omega_0 \longleftrightarrow H_1 : \boldsymbol{\theta} \in \Omega_1$$

零假设：对应某一简化模型 ← H_0

备择假设：对应某一全模型 ← H_1

- 对数似然比：$-2\ln\lambda = 2\left(\ell_1^* - \ell_0^*\right)$.

$$\begin{cases} \ell_0^* = \ln L_0^* = \max\limits_{\boldsymbol{\theta}\in\Omega_0} \ln L\left(\mathcal{X};\ \boldsymbol{\theta}\right) \\ \ell_1^* = \ln L_1^* = \max\limits_{\boldsymbol{\theta}\in\Omega_1} \ln L\left(\mathcal{X};\ \boldsymbol{\theta}\right) \end{cases}$$

- 拒绝域为 $R = \left\{\mathcal{X} : -2\ln\lambda\left(\mathcal{X}\right) > k\right\}$.

- 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

$$\forall \boldsymbol{\theta} \in \Omega_0 : -2\ln\lambda \xrightarrow{\mathscr{L}} \chi^2_{q-r} \qquad \text{as} \quad n \to \infty$$

上侧 α 分位数

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

已学知识点 (Recap)

第 7 章 假设检验

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

$$-2\ln\lambda = n\left(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0\right)^{\mathrm{T}} \Sigma^{-1}\left(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0\right) > \chi_p^2(\alpha)$$ 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 $\boldsymbol{X} \sim N_p\left(\boldsymbol{\mu}, \Sigma\right)$，其中 Σ 未知，对假设 $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$，

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

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

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

$$\left\{ \boldsymbol{\mu} \in \mathbb{R}^p \,\middle|\, \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)^{\mathrm{T}} \mathcal{S}^{-1} \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \le \frac{p}{n-p} \cdot F_{p,\, n-p}(\alpha) \right\}$$

④ $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\mu}$ 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 $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\mu}$ 在内

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

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

⑤ $\mu_j, j = 1, 2, \ldots, p$ 的联合置信区间

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

已学知识点 (Recap)

第 7 章 假设检验

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

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

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

已学知识点 (Recap)

第 7 章 假设检验

- 检验问题 4：总体 $Y \sim N_1\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}, \sigma^2\right)$，其中 σ^2 未知，$\boldsymbol{x} \in \mathbb{R}^p$，对假设 $H_0: \boldsymbol{\beta}=\boldsymbol{\beta}_0$

① 精确检验：当

$$F=\frac{n-p}{p} \cdot\left(\frac{\left\|\boldsymbol{y}-\mathscr{X} \boldsymbol{\beta}_0\right\|^2}{\left\|\boldsymbol{y}-\mathscr{X} \hat{\boldsymbol{\beta}}\right\|^2}-1\right)>F_{p,\, n-p}(\alpha)$$

$$\hat{\boldsymbol{\beta}}=\left(\mathscr{X}^{\mathrm{T}} \mathscr{X}\right)^{-1} \mathscr{X}^{\mathrm{T}} \boldsymbol{y}$$

时拒绝 H_0.

② 渐进检验：当

$$-2 \ln \lambda=n \ln \left(\frac{\left\|\boldsymbol{y}-\mathscr{X} \boldsymbol{\beta}_0\right\|^2}{\left\|\boldsymbol{y}-\mathscr{X} \hat{\boldsymbol{\beta}}\right\|^2}\right)>\chi_p^2(\alpha) \qquad (n \rightarrow \infty)$$

时拒绝 H_0.

已学知识点 (Recap)

第 7 章 假设检验

$\mathscr{A}_{q\times p},\ q \le p$

- 检验问题 5：(线性假设) 总体 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$，其中 Σ 已知，对假设 H_0：$\mathscr{A}\boldsymbol{\mu} = \boldsymbol{a}$，

$\boldsymbol{a}_{q\times 1}$

当

$$n(\mathscr{A}\bar{\boldsymbol{x}} - \boldsymbol{a})^{\mathrm{T}} \left(\mathscr{A}\Sigma\mathscr{A}^{\mathrm{T}}\right)^{-1} (\mathscr{A}\bar{\boldsymbol{x}} - \boldsymbol{a}) > \chi_q^2(\alpha)$$

时拒绝 H_0 .

已学知识点 (Recap)

第 7 章 假设检验

$\mathscr{A}_{q\times p},\ q \le p$

- 检验问题 6：(线性假设) 总体 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$，其中 Σ 未知，对假设 $H_0: \mathscr{A}\boldsymbol{\mu} = \boldsymbol{a}$，

$\boldsymbol{a}_{q\times 1}$

① 渐进检验：当

$$-2\ln\lambda = n\ln\left\{1 + (\mathscr{A}\overline{\boldsymbol{x}} - \boldsymbol{a})^{\mathrm{T}}\left(\mathscr{A}\mathcal{S}\mathscr{A}^{\mathrm{T}}\right)^{-1}(\mathscr{A}\overline{\boldsymbol{x}} - \boldsymbol{a})\right\} > \chi_q^2(\alpha)$$

时拒绝 H_0.

② 精确检验：当

$$\frac{n-q}{q}(\mathscr{A}\overline{\boldsymbol{x}} - \boldsymbol{a})^{\mathrm{T}}\left(\mathscr{A}\mathcal{S}\mathscr{A}^{\mathrm{T}}\right)^{-1}(\mathscr{A}\overline{\boldsymbol{x}} - \boldsymbol{a}) > F_{q,n-q}(\alpha)$$

时拒绝 H_0.

③ 检验 $H_0: \mathscr{C}\boldsymbol{\mu} = \boldsymbol{0} \iff H_0: \mu_1 = \mu_2 = \cdots = \mu_p$：当

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

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

时拒绝 H_0.

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

已学知识点 (Recap)

第 7 章 假设检验

$$\boldsymbol{b}^{\mathrm{T}}\mathbf{1}_p = \sum_{j=1}^{p} b_j = 0$$

- 检验问题 6：(线性假设) 总体 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$，其中 Σ 未知，对假设 $H_0: \mathscr{A}\boldsymbol{\mu} = \boldsymbol{a}$，

  ④ 置信度为 $1-\alpha$ 时，$\boldsymbol{\mu}$ 的所有对比 $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\mu}$ 的联合置信区间为

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

已学知识点 (Recap)

第 7 章 假设检验

- 检验问题 7：总体 $Y \sim N_1\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}, \sigma^2\right)$，其中 σ^2 未知，$\boldsymbol{x} \in \mathbb{R}^p$，对假设 H_0：$\mathscr{A}\boldsymbol{\beta} = \boldsymbol{a}$

$\mathscr{A}_{q\times p},\ q \le p$　　$\boldsymbol{a}_{q\times 1}$

① 渐进检验：当

$$-2\ln\lambda = n\ln\left(\frac{\left\|\boldsymbol{y}-\mathscr{X}\tilde{\boldsymbol{\beta}}\right\|^2}{\left\|\boldsymbol{y}-\mathscr{X}\hat{\boldsymbol{\beta}}\right\|^2}\right) > \chi_q^2(\alpha) \qquad (n\to\infty)$$

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{A}^{\mathrm{T}}\left[\mathscr{A}\left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{A}^{\mathrm{T}}\right]^{-1}\left(\mathscr{A}\hat{\boldsymbol{\beta}}-\boldsymbol{a}\right)$$

$$\hat{\boldsymbol{\beta}} = \left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{X}^{\mathrm{T}}\boldsymbol{y}$$

时拒绝 H_0.

② 精确检验：当

$$\frac{n-p}{q}\cdot\frac{\left(\mathscr{A}\hat{\boldsymbol{\beta}}-\boldsymbol{a}\right)^{\mathrm{T}}\left[\mathscr{A}\left(\mathscr{X}^{\mathrm{T}}\mathscr{X}\right)^{-1}\mathscr{A}^{\mathrm{T}}\right]^{-1}\left(\mathscr{A}\hat{\boldsymbol{\beta}}-\boldsymbol{a}\right)}{\left(\boldsymbol{y}-\mathscr{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{y}-\mathscr{X}\hat{\boldsymbol{\beta}}\right)} > F_{q,\,n-p}(\alpha)$$

时拒绝 H_0.

线性假设

- 两个均值向量的比较

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

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

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

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

$$n_1\mathcal{S}_1+n_2\mathcal{S}_2 \sim W_p\left(\Sigma\,,\,n_1+n_2-2\right) \Longrightarrow \left(n_1+n_2\right)\mathcal{S} \sim W_p\left(\Sigma\,,\,n_1+n_2-2\right)$$

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

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

$$n\left(\boldsymbol{X}-\boldsymbol{\mu}\right)^{\mathrm{T}}\mathscr{A}^{-1}\left(\boldsymbol{X}-\boldsymbol{\mu}\right) \sim T^2_{p,\,n}$$

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

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

线性假设

- 两个均值向量的比较

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

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

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

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

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

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

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

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

线性假设

- 两个均值向量的比较

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

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

- $H_0: \boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ 的拒绝域为

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

上 α 分位数

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

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

线性假设

- 两个均值向量的比较

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

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

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

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

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

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

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
```

$$\Longrightarrow \quad \bar{\boldsymbol{x}}_1 = \begin{pmatrix} 4084.000 \\ 2580.467 \end{pmatrix}$$

```
> (S_1 = cov(x[, 1:2]) * (n1-1)/(n1))
          [,1]     [,2]
[1,] 16634749 12409637
[2,] 12409637 13747417
```

$$\Longrightarrow \quad \mathcal{S}_1 = \begin{pmatrix} 16634749 & 12409637 \\ 12409637 & 13747417 \end{pmatrix}$$

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
```

$$\Longrightarrow \quad \bar{\boldsymbol{x}}_2 = \begin{pmatrix} 4307.2 \\ 4925.2 \end{pmatrix}$$

```
> (S_2 = cov(y[, 1:2]) * (n2-1)/(n2))
          [,1]     [,2]
[1,] 12247663 11425398
[2,] 11425398 15111585
```

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

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
```

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

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 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ 的 F 统计量为

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

```
# 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

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2.$$

▶ 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

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2.$$

▶ 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

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

$$-761.5132 \le \mu_{1a} - \mu_{2a} \le 315.1132$$
$$-2872.3248 \le \mu_{1s} - \mu_{2s} \le -1817.1418$$

```
# 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 $N_4\left(\boldsymbol{\mu}_i\ ,\ \Sigma_1\right)$ 拟合两个样本，其中 $\Sigma_1 = \mathcal{I}_4$，$\boldsymbol{\mu}_1 = \begin{pmatrix} 8 \\ 6 \\ 10 \\ 10 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 6 \\ 6 \\ 10 \\ 13 \end{pmatrix}$, $n_1 = 30$, $n_2 = 20.$

```
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 4.995699 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 $N_4\left(\boldsymbol{\mu}_i\,,\,\Sigma_1\right)$ 拟合两个样本，其中 $\Sigma_1=\mathcal{I}_4$，$\boldsymbol{\mu}_1=\begin{pmatrix}8\\6\\10\\10\end{pmatrix}$, $\boldsymbol{\mu}_2=\begin{pmatrix}6\\6\\10\\13\end{pmatrix}$，$n_1=30$，$n_2=20$.

  - 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
```

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

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 $N_4\left(\boldsymbol{\mu}_i\,,\,\Sigma_1\right)$ 拟合两个样本，其中 $\Sigma_1=\mathcal{I}_4$，$\boldsymbol{\mu}_1=\begin{pmatrix}8\\6\\10\\10\end{pmatrix}$, $\boldsymbol{\mu}_2=\begin{pmatrix}6\\6\\10\\13\end{pmatrix}$，$n_1=30$，$n_2=20.$

  - 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
```

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

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

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 $N_4\left(\boldsymbol{\mu}_i\ ,\ \Sigma_1\right)$ 拟合两个样本，其中 $\Sigma_1=\mathscr{I}_4$，$\boldsymbol{\mu}_1=\begin{pmatrix}8\\6\\10\\10\end{pmatrix}$, $\boldsymbol{\mu}_2=\begin{pmatrix}6\\6\\10\\13\end{pmatrix}$，$n_1=30$，$n_2=20$.

  - 检验 $H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ 的 F 统计量为

$$F=\frac{n_1n_2\left(n_1+n_2-p-1\right)}{p\left(n_1+n_2\right)^2}\left(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\right)^{\mathrm{T}}\mathcal{S}^{-1}\left(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\right)=36.55469$$

```
# 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 $H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2$.

  - 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 $N_4\left(\boldsymbol{\mu}_i\,,\,\Sigma_1\right)$ 拟合两个样本，其中 $\Sigma_1=\mathcal{I}_4$，$\boldsymbol{\mu}_1=\begin{pmatrix}8\\6\\10\\10\end{pmatrix}$, $\boldsymbol{\mu}_2=\begin{pmatrix}6\\6\\10\\13\end{pmatrix}$，$n_1=30$，$n_2=20.$

  - 均值差的置信度为 95% 的联合置信区间为

$$\delta_j=\mu_{1j}-\mu_{2j}\in\left(\bar{x}_{1j}-\bar{x}_{2j}\right)\pm\sqrt{\frac{p\left(n_1+n_2\right)^2}{n_1n_2\left(n_1+n_2-p-1\right)}F_{p,\,n_1+n_2-p-1}(\alpha)\cdot s_{jj}}\quad,\quad j=1,\,2,\,3,\,4$$

```
# 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 <-                                                                                    05, p
mean
mean
```

```
> mean_X1 - mean_X2 - matrix(c(a1, a2, a3, a4), nrow=4)
            [,1]
[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]
[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 $N_4\left(\boldsymbol{\mu}_i\,,\,\Sigma_1\right)$ 拟合两个样本，其中 $\Sigma_1=\mathcal{I}_4$，$\boldsymbol{\mu}_1=\begin{pmatrix}8\\6\\10\\10\end{pmatrix}$, $\boldsymbol{\mu}_2=\begin{pmatrix}6\\6\\10\\13\end{pmatrix}$，$n_1=30$，$n_2=20.$

  - 均值差的置信度为 95% 的联合置信区间为

$$1.47790823 \le \delta_1=\mu_{11}-\mu_{21} \le 2.005228521$$
$$-0.57504658 \le \delta_2=\mu_{12}-\mu_{22} \le 0.007514841$$
$$-0.09673054 \le \delta_2=\mu_{12}-\mu_{22} \le 0.404935224$$
$$-3.45360215 \le \delta_1=\mu_{11}-\mu_{21} \le -2.952711955$$

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 $N_4\left(\boldsymbol{\mu}_i\ ,\ \Sigma_2\right)$ 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 $N_4\left(\boldsymbol{\mu}_i\ ,\ \Sigma_2\right)$ 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
```

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

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 $N_4\left(\boldsymbol{\mu}_i\,,\,\Sigma_2\right)$ 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:

```
# 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
```

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

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

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 $N_4\left(\boldsymbol{\mu}_i\ ,\ \Sigma_2\right)$ 的两个样本，其中 $\Sigma_2 = 16 \times \mathcal{I}_4$，均值向量保持不变.

- 检验 $H_0:\ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ 的 F 统计量为

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

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

```r
# 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)
```

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

- 结论: 接受 $H_0:\ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

- 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
```

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

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
```

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

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

线性假设

- 例: 我们来比较伪钞和真钞的均值向量.

- 检验 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ 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]
[1,] 391.9217
```

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

- 结论: 拒绝 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

线性假设

- 例: 我们来比较伪钞和真钞的均值向量.

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

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

$$\begin{aligned}
0.0783729 &\le \delta_1 \le 0.2136271 \\
-0.4144075 &\le \delta_2 \le -0.2995925 \\
-0.5329059 &\le \delta_3 \le -0.4130941 \\
-2.3931004 &\le \delta_4 \le -2.0568996 \\
-1.0823388 &\le \delta_5 \le -0.8476612 \\
1.9746866 &\le \delta_6 \le 2.1593134
\end{aligned}$$

- 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 $\boldsymbol{X}_{ih} \sim N_p\left(\boldsymbol{\mu}_h\,,\,\Sigma_h\right)$，$i=1,2,\ldots,n_h$，$h=1,2,\ldots,k$ 是相互独立的随机向量.

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

$$N_p\left(\boldsymbol{\mu}_1\,,\,\Sigma_1\right) \xrightarrow{\text{sampling}} \boldsymbol{X}_{1\,1},\boldsymbol{X}_{2\,1},\ \ldots,\ \boldsymbol{X}_{n_1\,1} \longmapsto n_1\,\mathcal{S}_1=\sum_{i=1}^{n_1}\left(\boldsymbol{x}_{i\,1}-\bar{\boldsymbol{x}}_1\right)\left(\boldsymbol{x}_{i\,1}-\bar{\boldsymbol{x}}_1\right)^{\mathrm{T}} \sim W_p\left(\Sigma_1\,,\,n_1-1\right)$$

$$N_p\left(\boldsymbol{\mu}_2\,,\,\Sigma_2\right) \xrightarrow{\text{sampling}} \boldsymbol{X}_{1\,2},\boldsymbol{X}_{2\,2},\ \ldots,\ \boldsymbol{X}_{n_2\,2} \longmapsto n_2\,\mathcal{S}_2=\sum_{i=1}^{n_2}\left(\boldsymbol{x}_{i\,2}-\bar{\boldsymbol{x}}_2\right)\left(\boldsymbol{x}_{i\,2}-\bar{\boldsymbol{x}}_2\right)^{\mathrm{T}} \sim W_p\left(\Sigma_2\,,\,n_2-1\right)$$

$$\vdots \qquad\qquad \vdots \qquad \text{相互独立} \qquad\qquad \vdots$$

$$N_p\left(\boldsymbol{\mu}_k\,,\,\Sigma_k\right) \xrightarrow{\text{sampling}} \boldsymbol{X}_{1\,k},\boldsymbol{X}_{2\,k},\ \ldots,\ \boldsymbol{X}_{n_k\,k} \longmapsto n_k\,\mathcal{S}_k=\sum_{i=1}^{n_k}\left(\boldsymbol{x}_{i\,k}-\bar{\boldsymbol{x}}_k\right)\left(\boldsymbol{x}_{i\,k}-\bar{\boldsymbol{x}}_k\right)^{\mathrm{T}} \sim W_p\left(\Sigma_k\,,\,n_k-1\right)$$

- 当 $H_0:\ \Sigma_1=\Sigma_2=\cdots=\Sigma_k \triangleq \Sigma$ 真时，我们有

$$\sum_{h=1}^{k} n_h\,\mathcal{S}_h \sim W_p\left(\Sigma\,,\,n-k\right), \quad n=n_1+n_2+\cdots+n_k$$

线性假设

检验问题 9：(协方差矩阵的比较) 设 $\boldsymbol{X}_{ih} \sim N_p\left(\boldsymbol{\mu}_h, \Sigma_h\right)$, $i = 1, 2, \ldots, n_h$, $h = 1, 2, \ldots, k$ 是相互独立的随机向量.

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

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

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

- 似然比检验得到以下统计量

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

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

$$-2\log\lambda = n\log\left|\mathcal{S}\right| - \sum_{h=1}^{k} n_h \log\left|\mathcal{S}_h\right| \overset{\text{approximately}}{\sim} \chi^2_m$$

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

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
```

$\Longrightarrow \quad -2\log\lambda = 127.2149$

```
# 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
```

$\Longrightarrow \quad p\text{ 值} = 3.338319 \times 10^{-17}$

- 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 $\boldsymbol{X}_{i\,1} \sim N_p\left(\boldsymbol{\mu}_1\,,\,\Sigma_1\right),\ i=1,\,2,\,\ldots,\,n_1,\ \boldsymbol{X}_{j\,2} \sim N_p\left(\boldsymbol{\mu}_2\,,\,\Sigma_2\right),\ j=1,\,2,\,\ldots,\,n_2,$ 是相互独立的随机向量.

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

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

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

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

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

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

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 $\boldsymbol{X}_{i\,1} \sim N_p\left(\boldsymbol{\mu}_1\,,\,\Sigma_1\right),\ i=1,\,2,\,\ldots,\,n_1,\ \boldsymbol{X}_{j\,2} \sim N_p\left(\boldsymbol{\mu}_2\,,\,\Sigma_2\right),\ j=1,\,2,\,\ldots,\,n_2,$ 是相互独立的随机向量.

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

- 显著水平为 α 时，$H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ 的拒绝域为

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

上 α 分位数

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

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

$\mathcal{S}_1$ 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
```

$\Longrightarrow$

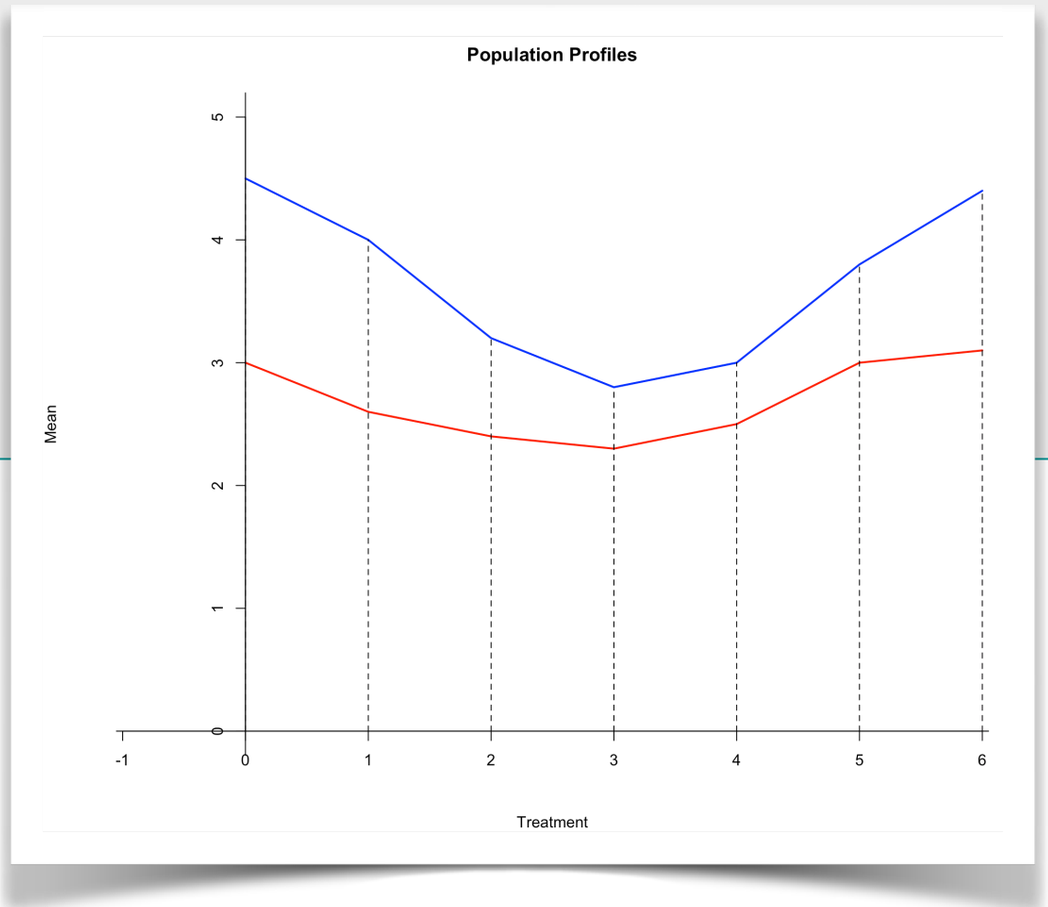
$$-0.0389 \le \delta_1 = \mu_{g\,1} - \mu_{f\,1} \le 0.3309$$
$$-0.5140 \le \delta_2 = \mu_{g\,2} - \mu_{f\,2} \le -0.2000$$
$$-0.6368 \le \delta_3 = \mu_{g\,3} - \mu_{f\,3} \le -0.3092$$
$$-2.6846 \le \delta_4 = \mu_{g\,4} - \mu_{f\,4} \le -1.7654$$
$$-1.2858 \le \delta_5 = \mu_{g\,5} - \mu_{f\,5} \le -0.6442$$
$$1.8146 \le \delta_6 = \mu_{g\,6} - \mu_{f\,6} \le 2.3194$$

最大差异

线性假设

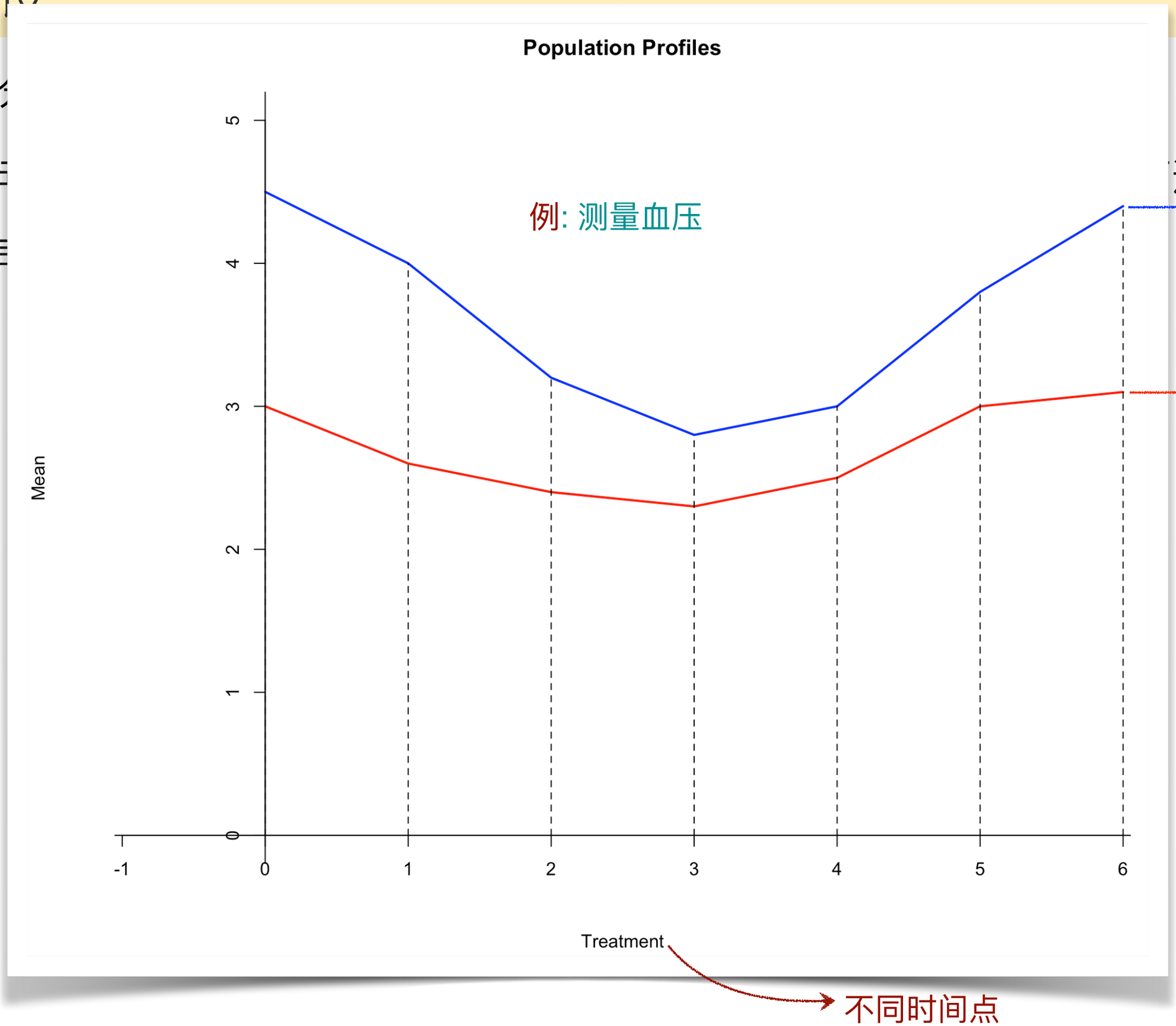
- 轮廓分析 (Profile Analysis)
  - 问题源自实践中，当我们对某一特征进行重复观测 (或在不同实验条件下进行相同类型的测量)，且需要对不同群体进行比较时.

```
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
    同


Population Profiles
例: 测量血压
进行相
新药
安慰剂
Mean
0
1
2
3
4
5
-1
0
1
2
3
4
5
6
Treatment
不同时间点

线性假设

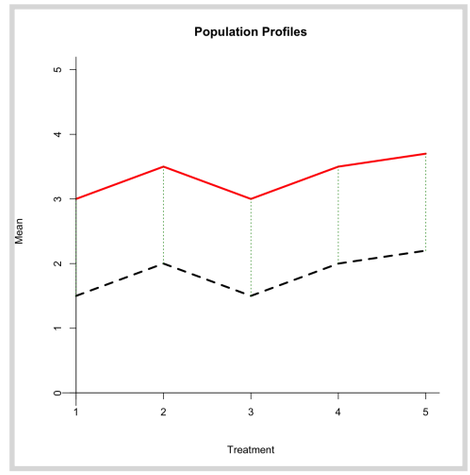
- 轮廓分析 (Profile Analysis)
  - 这与两个均值向量的比较属于相同的统计问题：

$$\boldsymbol{X}_{i1} \sim N_p(\boldsymbol{\mu}_1, \Sigma), \quad i = 1, 2, \ldots, n_1$$
$$\boldsymbol{X}_{i2} \sim N_p(\boldsymbol{\mu}_2, \Sigma), \quad i = 1, 2, \ldots, n_2$$

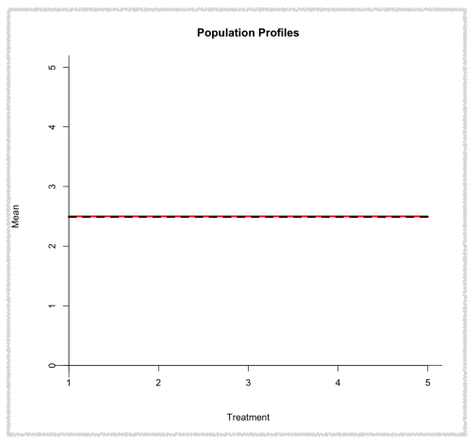
$\bar{x}_1, S_1$

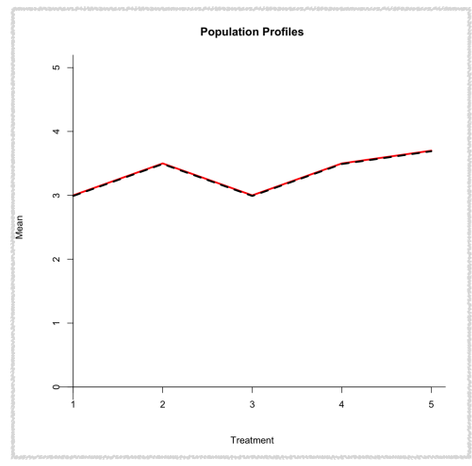
$\bar{x}_2, S_2$

，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？(这说明治疗和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

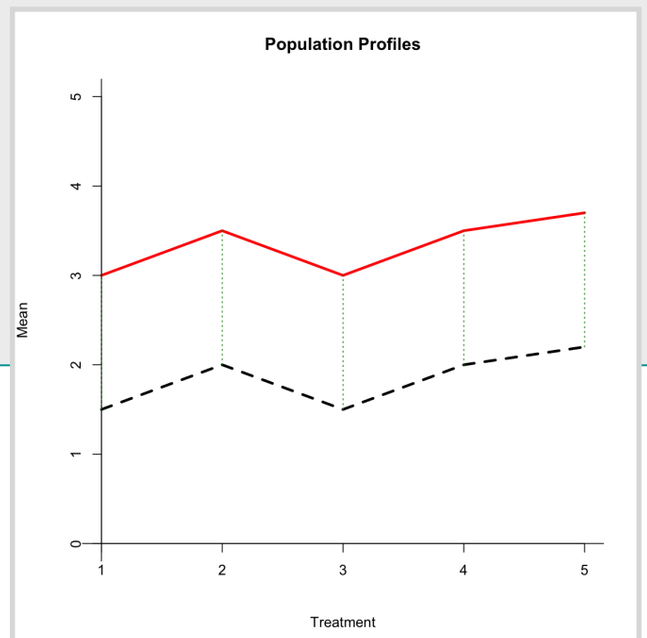
- 轮廓分析 (Profile Analysis)
  - 平行

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

$$\begin{pmatrix} (\mu_{1\,1} - \mu_{1\,2}) - (\mu_{2\,1} - \mu_{2\,2}) \\ (\mu_{2\,1} - \mu_{2\,2}) - (\mu_{3\,1} - \mu_{3\,2}) \\ \vdots \\ (\mu_{p-1\,1} - \mu_{p-1\,2}) - (\mu_{p\,1} - \mu_{p\,2}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

✓ 欲检验的假设为 $H_0^{(1)}: \mathscr{C}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$

```
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

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

$$\begin{pmatrix}(\mu_{1\,1}-\mu_{1\,2})-(\mu_{2\,1}-\mu_{2\,2})\\ (\mu_{2\,1}-\mu_{2\,2})-(\mu_{3\,1}-\mu_{3\,2})\\ \vdots\\ (\mu_{p-1\,1}-\mu_{p-1\,2})-(\mu_{p\,1}-\mu_{p\,2})\end{pmatrix}=\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}$$

✓ 欲检验的假设为 $H_0^{(1)}:\ \mathscr{C}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)=\mathbf{0}$

✓ 应用检验问题 8 的结果，我们有

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

$$\frac{n_1n_2}{\left(n_1+n_2\right)^2}\left(n_1+n_2-2\right)\left[\mathscr{C}\left(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\right)\right]^{\mathrm{T}}\left(\mathscr{C}\mathcal{S}\mathscr{C}^{\mathrm{T}}\right)^{-1}\mathscr{C}\left(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\right)\overset{H_0\ \text{True}}{\sim}T^2_{p-1,\,n_1+n_2-2}$$

✓ 拒绝域：

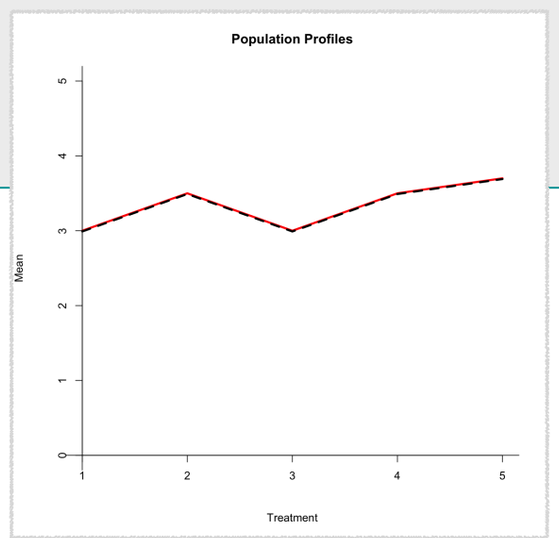
$$\frac{n_1n_2\left(n_1+n_2-p\right)}{\left(n_1+n_2\right)^2(p-1)}\left(\mathscr{C}\bar{\boldsymbol{x}}\right)^{\mathrm{T}}\left(\mathscr{C}\mathcal{S}\mathscr{C}^{\mathrm{T}}\right)^{-1}\mathscr{C}\bar{\boldsymbol{x}}>F_{p-1,\,n_1+n_2-p}(\alpha)$$

线性假设

- 轮廓分析 (Profile Analysis) $\Longrightarrow \quad \mu_{1\,1} = \mu_{1\,2},\ \mu_{2\,1} = \mu_{2\,2},\ \cdots,\ \mu_{p\,1} = \mu_{p\,2}$
  - 两个相同的水平 $\left(\mu_{1\,1} - \mu_{1\,2}\right) + \left(\mu_{2\,1} - \mu_{2\,2}\right) + \cdots + \left(\mu_{p\,1} - \mu_{p\,2}\right) = 0$
    - ✓ 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

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

```
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

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

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

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

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

$$(n-1)\left(\mathscr{C}\bar{\boldsymbol{x}}-\mathscr{C}\boldsymbol{\mu}\right)^{\mathrm{T}}\left(\mathscr{C}\mathcal{S}_X\mathscr{C}^{\mathrm{T}}\right)^{-1}\left(\mathscr{C}\bar{\boldsymbol{x}}-\mathscr{C}\boldsymbol{\mu}\right)\sim T^2_{q,\,n-1}$$

✓ 因为

$$\mathbf{1}_p^{\mathrm{T}}\left(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\right)\sim N_1\left(\mathbf{1}_p^{\mathrm{T}}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)\ ,\frac{n_1+n_2}{n_1n_2}\mathbf{1}_p^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{1}_p\right)$$

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

$$\frac{n_1n_2}{\left(n_1+n_2\right)^2}\left(n_1+n_2-2\right)\frac{\left[\mathbf{1}_p^{\mathrm{T}}\left(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2\right)\right]^2}{\mathbf{1}_p^{\mathrm{T}}\mathcal{S}\mathbf{1}_p^{\mathrm{T}}}\sim T^2_{1,\,n_1+n_2-2}=F_{1,\,n_1+n_2-2}$$

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

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

    - ✓ 拒绝域为

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

上 α 分位数

线性假设

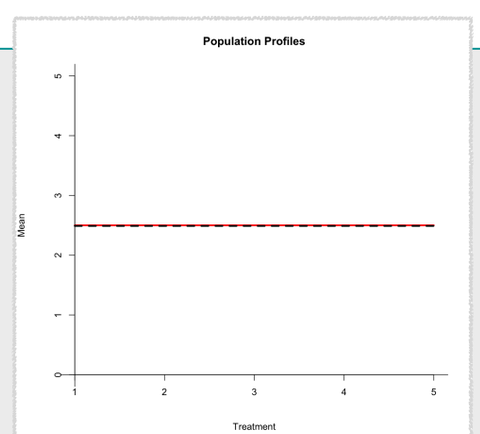
- 轮廓分析 (Profile Analysis)
  - 组间效应
    - ✓ 若 $H_0^{(1)}$ 与 $H_0^{(2)}$ 均已接受，则，我们可以利用两组中包含的信息来检验组间效应，即，两个轮廓是否均为水平线. 这可以表示为

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

$$\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}_{(p-1)\times p} \begin{pmatrix} 2\mu_{1\,1} \\ 2\mu_{2\,1} \\ \vdots \\ 2\mu_{p\,1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} \mu_{1\,1}-\mu_{2\,1} \\ \mu_{2\,1}-\mu_{3\,1} \\ \vdots \\ \mu_{p-1\,1}-\mu_{p\,1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\Longrightarrow \quad \mu_{1\,1}=\mu_{2\,1}=\cdots=\mu_{p-1\,1}=\mu_{p\,1}$$

```
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

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

    ✓ 考虑平均轮廓 $\bar{\boldsymbol{x}}$

$$\bar{\boldsymbol{x}}=\frac{n_1\bar{\boldsymbol{x}}_1+n_2\bar{\boldsymbol{x}}_2}{n_1+n_2} \quad\Longrightarrow\quad \bar{\boldsymbol{x}}\sim N_p\left(\frac{n_1\boldsymbol{\mu_1}+n_2\boldsymbol{\mu_2}}{n_1+n_2}\ ,\ \frac{1}{n_1+n_2}\Sigma\right)$$

在 $H_0^{(3)}$ 与 $H_0^{(1)}$ 为真时 $\Longrightarrow$ $\mathscr{C}\left(\dfrac{n_1\boldsymbol{\mu_1}+n_2\boldsymbol{\mu_2}}{n_1+n_2}\right)=\mathbf{0}$

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

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

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

$$\left(n_1+n_2-2\right)\left(\mathscr{C}\bar{\boldsymbol{x}}\right)^{\mathrm{T}}\left(\mathscr{C}\mathcal{S}\mathscr{C}^{\mathrm{T}}\right)^{-1}\mathscr{C}\bar{\boldsymbol{x}} \ \sim \ T^2_{p-1,\, n_1+n_2-p}$$

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

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

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

$$(n-1)\left(\mathscr{C}\bar{\boldsymbol{x}}-\mathscr{C}\boldsymbol{\mu}\right)^{\mathrm{T}}\left(\mathscr{C}\mathcal{S}_X\mathscr{C}^{\mathrm{T}}\right)^{-1}\left(\mathscr{C}\bar{\boldsymbol{x}}-\mathscr{C}\boldsymbol{\mu}\right)\sim T^2_{q,\,n-1}$$

✓ $H_0^{(3)}$ 的拒绝域为

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

上 α 分位数

在 $H_0^{(3)}$ 与 $H_0^{(1)}$ 为真时 $\Longrightarrow$

$$\sqrt{n_1+n_2}\ \mathscr{C}\bar{\boldsymbol{x}}\ \sim N_p\left(\mathbf{0}\ ,\ \mathscr{C}\Sigma\mathscr{C}^{\mathrm{T}}\right)$$

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)
```

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

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

X_1：人均犯罪率

X_2：划作大型住宅用地的比例

X_3：非零售业务占地的比例

X_5：一氧化氮浓度

X_7：1940年之前建造的自住房的比例

X_9：辐射状高速公路的可达性指数

X_{11}：生师比

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

其中 B是非裔美国人的比例

X_{13}：社会低层人口的百分比

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 $X_{14} \leq X_{14}$ 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 $X_1, X_5, X_8, X_{11}, X_{13}$
  - 欲检验 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$，其中 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^5$

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

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

- $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ 的拒绝域为

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

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 $X_{14} \le X_{14}$ 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 $X_1, X_5, X_8, X_{11}, X_{13}$
  - 欲检验 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$，其中 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^5$
  - 计算得检验统计量的值为
  - 查得检验的临界值为
  - 结论：拒绝 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$

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

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

- $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ 的拒绝域为

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

上 α 分位数

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

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

```
> 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 $X_{14} \le X_{14}$ 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

$$\delta_1 \in (1.4019611, 2.5498855)$$
$$\delta_5 \in (0.1314898, 0.2383455)$$
$$\delta_8 \in (-0.5243907, -0.2221791)$$
$$\delta_{11} \in (1.0374981, 1.7383789)$$
$$\delta_{13} \in (1.1576946, 1.5818046)$$

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

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

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

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

- 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.

```
> 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 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$，其中 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^6$

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

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

- $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ 的拒绝域为

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

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 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$，其中 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^6$
  - 计算得检验统计量的值为
  - 查得检验的临界值为
  - 结论：拒绝 $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$

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

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

- $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ 的拒绝域为

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

上 α 分位数

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

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

```
> 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 $\boldsymbol{X}_{i1} \sim N_p(\boldsymbol{\mu}_1, \Sigma)$，$i = 1, 2, \ldots, n_1$，$\boldsymbol{X}_{j2} \sim N_p(\boldsymbol{\mu}_2, \Sigma)$，$j = 1, 2, \ldots, n_2$，其中所有的变量相互独立.

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

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

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

```
> 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
```

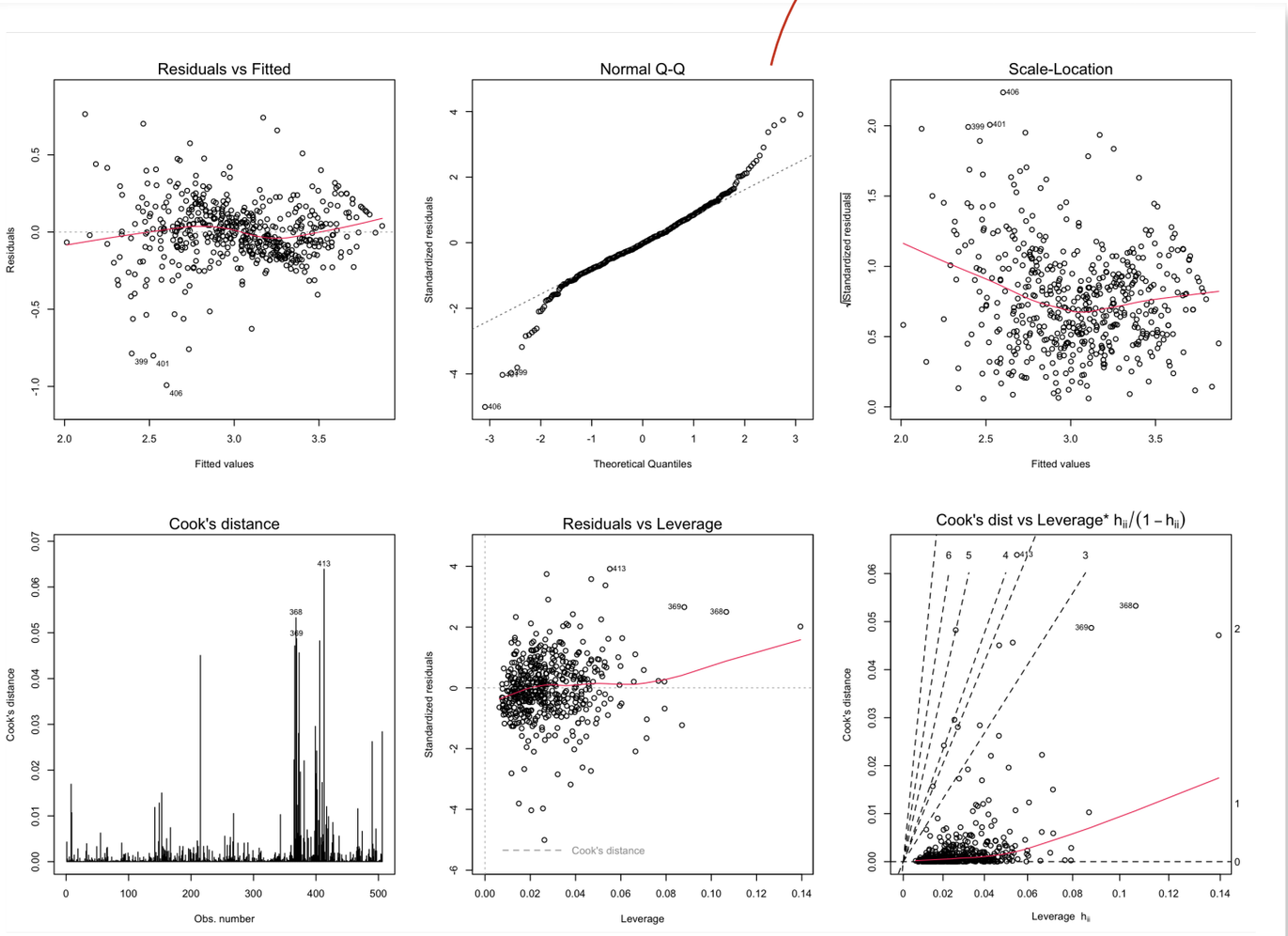
```
mu_L = as.matrix(xbar_3 - xbar_4) -
  sqrt(qf(0.05, p, n3 + n4 - p - 1, lower.tail = FALSE) * (p * (n3 + n4)^2) / (n3 * n4 * (n3 + n4 - p - 1)) *
         as.matrix(c(S[1, 1], S[2, 2], S[3, 3], S[4, 4], S[5, 5], S[6, 6])))
mu_R = as.matrix(xbar_3 - xbar_4) +
  sqrt(qf(0.05, p, n3 + n4 - p - 1, lower.tail = 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 $H_0: (\beta_1, \beta_2, \dots, \beta_{13})^{\mathrm{T}} = \mathbf{0} \iff \mathscr{A}\boldsymbol{\beta} = \mathbf{0}$

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

```
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 $Y_i \sim N_1(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2)$，$\boldsymbol{x}_i \in \mathbb{R}^p$，其中 σ^2 未知.

$H_0: \mathscr{A}\boldsymbol{\beta} = \boldsymbol{a} \longleftrightarrow H_1:$ 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

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

- 此时 H_0 的拒绝域为

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

$F_{q,\, n-p}(\alpha)$ 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 $H_0: (\beta_1, \beta_2, \cdots, \beta_{13})^{\mathrm{T}} = \mathbf{0} \iff \mathscr{A}\boldsymbol{\beta} = \mathbf{0}$

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

  - 计算得检验统计量的值为
  - 查得检验的临界值为
  - 结论：拒绝 $H_0: (\beta_1, \beta_2, \cdots, \beta_{13})^{\mathrm{T}} = \mathbf{0}$

检验问题 7：假设 $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ 相互独立，且 $\boldsymbol{Y}_i \sim N_1(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2)$，$\boldsymbol{x}_i \in \mathbb{R}^p$，其中 σ^2 未知.

$H_0: \mathscr{A}\boldsymbol{\beta} = \boldsymbol{a} \longleftrightarrow H_1:$ 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

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

- 此时 H_0 的拒绝域为

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

$F_{q,\, n-p}(\alpha)$ 是上 α 分位数

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

```
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 $H_0: \beta_4 = 0 \iff \mathscr{A}\boldsymbol{\beta} = 0$

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

检验问题 7：假设 $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ 相互独立，且 $\boldsymbol{Y}_i \sim N_1(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2)$，$\boldsymbol{x}_i \in \mathbb{R}^p$，其中 σ^2 未知.

H_0： $\mathscr{A}\boldsymbol{\beta} = \boldsymbol{a}$ $\longleftrightarrow$ H_1： 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

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

- 此时 H_0 的拒绝域为

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

$F_{q,\,n-p}(\alpha)$ 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 $H_0: \beta_4 = 0 \iff \mathscr{A}\boldsymbol{\beta} = 0$

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

  - 计算得检验统计量的值为
  - 查得检验的临界值为
  - 结论：拒绝 $H_0: \beta_4 = 0$

检验问题 7：假设 $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ 相互独立，且 $\boldsymbol{Y}_i \sim N_1(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2)$，$\boldsymbol{x}_i \in \mathbb{R}^p$，其中 σ^2 未知.

$H_0: \mathscr{A}\boldsymbol{\beta} = \boldsymbol{a} \longleftrightarrow H_1:$ 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

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

- 此时 H_0 的拒绝域为

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

$F_{q,\,n-p}(\alpha)$ 是上 α 分位数

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

```
> 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 $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_7 = 0 \iff \mathscr{A}\boldsymbol{\beta} = \mathbf{0}_{4\times 1}$

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

检验问题 7：假设 $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ 相互独立，且 $\boldsymbol{Y}_i \sim N_1\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2\right)$，$\boldsymbol{x}_i \in \mathbb{R}^p$，其中 σ^2 未知.

H_0： $\mathscr{A}\boldsymbol{\beta} = \boldsymbol{a}$ $\longleftrightarrow$ H_1： 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

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

- 此时 H_0 的拒绝域为

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

$F_{q,\, n-p}(\alpha)$ 是上 α 分位数

```
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
```

```
> 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 $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_7 = 0 \iff \mathscr{A}\boldsymbol{\beta} = \mathbf{0}_{4\times 1}$

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

  - 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 $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ 相互独立，且 $\boldsymbol{Y}_i \sim N_1\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2\right)$，$\boldsymbol{x}_i \in \mathbb{R}^p$，其中 σ^2 未知.

H_0：$\mathscr{A}\boldsymbol{\beta} = \boldsymbol{a}$ $\longleftrightarrow$ H_1：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

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

- 此时 H_0 的拒绝域为

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

$F_{q,\, n-p}(\alpha)$ 是上 α 分位数

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

```
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

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

```
> 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
```

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

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