4.11 HW 11

作业 11 链接

本次作业，很多同学用的方法相当复杂，没有选出简单的公式。涉及 F 检验的内容，对于具体的模型，核心是把全模型和零模型的简单形式写出来，计算 $\hat{y}$ 与 $\hat{y}_0$, 再考虑问题，具体做法参见解答。一味地套原始公式会面对复杂的矩阵，难以处理!

✍ **练习 4.1** 假设 n 个个体被随机分配服用某种药物, 按照剂量从小到达分成 K 个组, 假设服用第 k 种剂量的个体数为 n_k, 响应的均值为 μ_k, 具体如下:

$$y_1, \ldots, y_{n_1} \text{ iid } \sim N\left(\mu_1, \sigma^2\right); \quad y_{n_1+1}, \ldots, y_{n_1+n_2} \text{ iid } \sim N\left(\mu_2, \sigma^2\right) \text{ 等等}$$

以线性模型表示如下:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N_n\left(0, \sigma^2 I_n\right), \boldsymbol{\beta} = \left(\mu_1, \mu_2, \ldots, \mu_K\right)^\top,$$

设计阵为

$$X_{n\times K} = \begin{pmatrix} \mathbb{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbb{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \cdots & \mathbb{1}_{n_K} \end{pmatrix}$$

其中 $\mathbb{1}_{n_i}$ 代表长度为 n_i 的分量全是 1 的向量, $\mathbf{0}_{n_i}$ 代表长度为 n_i 的分量全是 0 的向量, $n = n_1 + \ldots + n_K, K \geq 2$. 求 $H_0 : \mu_k = k\mu_1 (k = 1, 2, \ldots, K)$ 的 F 检验统计量.

解 由题知, 线性模型表示如下:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \epsilon \sim N_n\left(0, \sigma^2 I_n\right), \beta = \left(\mu_1, \cdots, \mu_K\right)^T$$

设计阵记为 $X_{n\times K} = (g_1, \cdots, g_k)$, 其中 g_k 为列向量。

在 H_0 假设下，模型为 $y = (g_1 + 2g_2 + \cdots + kg_k)\,\mu_1 + \epsilon = Z\mu_1 + \epsilon$ 。因此

$$\hat{\mu_1} = \frac{\sum_{i=1}^m i n_i \overline{y_i}}{\sum_{i=1}^m i^2 n_i}$$

于是

$$\hat{y}_0 = \left(\hat{\mu_1}, \cdots, k\hat{\mu_1}\right)^T$$

$$\hat{y} = P_{g_1} y + \cdots + P_{g_K} y = \left(\overline{y_1}, \cdots, \overline{y_1}, \cdots, \overline{y_k}, \cdots, \overline{y_k}\right)^T$$

将上式代入 F 检验的一般形式: $F = \frac{n-p}{q} \times \frac{\|\hat{y}-\hat{y_0}\|^2}{\|y-\hat{y}\|^2}$ 。即得:

$$F_{K-1,n-K} = \frac{n-K}{K-1} \times \frac{\sum_{i=1}^K \left(\overline{y_i} - i\hat{\mu}_1\right)^2}{\sum_{k=1}^K \sum_{i=1}^{n_k} \left(y_{ki} - \bar{y}_k\right)^2}$$

✍ **练习 4.2** 为了估计两个物品的重量 α, β, 用天平称量三次, 三次测量分别测的是 $\alpha, \alpha - \beta$ （天平一边放一个物品), $\alpha + \beta$ (两个物品都放天平的同一边), 得到的测量值分别为 y_1, y_2, y_3 。假设天平的测量误差服从 $N\left(0, \sigma^2\right)$ (与被测物品的真实重量无关)。

(a) 写出回归模型, 并求出 α, β, σ^2 的估计。

(b) 求 $H_0 : \alpha = \beta$ 的 F 检验统计量。

解

1. 由题知, 线性模型如下:

$$\boldsymbol{y}=\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}=\begin{pmatrix} 1 & 0 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}+\boldsymbol{\epsilon}=A\boldsymbol{\eta}+\boldsymbol{\epsilon}$$

由前已知的最小二乘估计，可知:

$$\hat{\boldsymbol{\eta}}=\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}=\left(A^TA\right)^{-1}A^Ty=\begin{pmatrix} \frac{1}{3} & 0 \\ \frac{1}{2} & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}=\begin{pmatrix} \frac{1}{3}\left(y_1+y_2+y_3\right) \\ \frac{1}{2}\left(y_3-y_2\right) \end{pmatrix}$$

同理 $\hat{\sigma}^2=\frac{RSS}{n-p}=\frac{\sum_{i=1}^3\|y_i-\hat{y_i}\|^2}{3-2}$ 。其中:

$$y-\hat{y}=\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}-\begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{3}\left(y_1+y_2+y_3\right) \\ \frac{1}{2}\left(y_3-y_2\right) \end{pmatrix}=\begin{pmatrix} \frac{2}{3}y_1-\frac{1}{3}y_2-\frac{1}{3}y_3 \\ -\frac{1}{3}y_1+\frac{1}{6}y_2+\frac{1}{6}y_3 \\ -\frac{1}{3}y_1+\frac{1}{6}y_2+\frac{1}{6}y_3 \end{pmatrix}$$

化简即得: $\hat{\sigma}^2=\frac{1}{6}\left(2y_1-y_2-y_3\right)^2$ 。

2. 求 $H_0:\alpha=\beta$ 的 F 检验统计量: 由前已知, $\hat{y}=(\hat{\alpha},\hat{\alpha}-\hat{\beta},\hat{\alpha}+\hat{\beta})^T$, 即

$$\hat{y}=\begin{pmatrix} \frac{1}{3}\left(y_1+y_2+y_3\right) \\ \frac{1}{3}y_1+\frac{5}{6}y_2-\frac{1}{6}y_3 \\ \frac{1}{3}y_1-\frac{1}{6}y_2+\frac{5}{6}y_3 \end{pmatrix}$$

在零假设下，模型为:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_2 \end{pmatrix}=\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}\alpha+\epsilon$$

求得 α 在 V_0 下的 LS 估计:

$$\hat{\alpha_0}=\frac{y_1+2y_3}{5}$$

于是:

$$\hat{y}_0=\left(\frac{y_1+2y_3}{5},0,\frac{2y_1+4y_3}{5}\right)^T$$

代入 F 检验的一般形式:

$$F_{1,1}=\frac{n-p}{q}\times\frac{\|\hat{y}-\hat{y_0}\|^2}{\|y-\hat{y}\|^2}=\frac{\|\hat{y}-\hat{y_0}\|^2}{\hat{\sigma^2}}=\frac{\left(2y_1+5y_2-y_3\right)^2}{5\left(2y_1-y_2-y_3\right)^2}$$

其中 $\hat{y}-\hat{y}_0=\left(\frac{1}{15},\frac{1}{6},-\frac{1}{30}\right)^T\left\{2y_1+5y_2-y_3\right\}$

练习 4.3 为了比较处理 1 (对照、安慰剂) 和处理 2 (药物) 是否有差异, 我们将研究对象进行配对匹配使得他们尽量相似, 其中一个（随机决定）接受处理 1 , 另一个接受处理 2 , 他们组成一个区组（这里称为配对)。假设有两个配对, 第 j 对研究对象的响应为 $\left(y_{1j},y_{2j}\right),j=1,2$ 其中下标中的 1,2 分别代表处理 1 和 2 。数据列表如下:

假设正态模型: $y_{ij}\sim N\left(\mu_{ij},\sigma^2\right),i,j=1,2$, 其中 $y_{11},y_{21},y_{12},y_{22}$ 独立, 假设处理的效果在两个区组中相同: $\mu_{21}-\mu_{11}=\mu_{22}-\mu_{12}$, 记之为 δ 。

(a) 验证 $H_0:\delta=0$ 的 F 检验 $F=t_{\text{pair}}^2$, 其中 t_{pair} (成对 t 检验统计量) 如下

		区组/配对 1	区组/配对 2
处理	1	y_{11}	y_{12}
	2	y_{21}	y_{22}

$$t_{\text{pair}} = \frac{y_{21}+y_{22}-y_{11}-y_{12}}{|y_{11}+y_{22}-y_{12}-y_{21}|} \overset{H_0}{\sim} t_1$$

(评注: t_1 分布也称作 Cauchy 分布)

(b) 假设 r.v. x 关于 0 对称, $P(x=0)=0, P(\text{sgn}(x)=\pm 1)=1/2$, 其中 $\text{sgn}(x)$ 代表 x 的符号。证明 $x=\text{sgn}(x)|x|$ 中 $\text{sgn}(x)$ 与 $|x|$ 独立。假设 σ 是符号随机变量, $P(\sigma=\pm 1)=1/2)$, 且 σ 与 x 独立, 证明 $x \overset{d}{=} \sigma|x| \overset{d}{=} \sigma x$, 其中 $\overset{d}{=}$ 表示同分布。

(c) 假设 x_1, x_2 iid $\sim N(0,1)$, 证明 $\frac{x_1}{x_2}$ 与 $\frac{x_1}{|x_2|}$ 分布相同 (都服从 t_1 分布)。这说明 (1) 式中 t_{pair} 统计量也可取为

$$t_{\text{pair}} = \frac{y_{21}+y_{22}-y_{11}-y_{12}}{y_{11}+y_{22}-y_{12}-y_{21}}$$

(d) 假如列表中的数据不是配对数据而是随机样本, 即接受处理 1 的 y_{11}, y_{12} iid $\sim N\left(\mu_1, \sigma^2\right)$, 接受处理 2 的 y_{21}, y_{22} iid $\sim N\left(\mu_2, \sigma^2\right)$, 数据表示如下 (注意中间没有竖线, y_{11} 和 y_{12} 地位是对称的, 即每一行内的数据没有次序):

		id 1	id 2
处理	1	y_{11}	y_{12}
	2	y_{21}	y_{22}

此时 $H_0: \mu_1=\mu_2$ 的检验为两样本 t 检验, 验证

$$t_{two-sample} = \frac{y_{21}+y_{22}-y_{11}-y_{12}}{\sqrt{(y_{11}-y_{12})^2+(y_{21}-y_{22})^2}} \overset{H_0}{\sim} t_2$$

比较 t_{pair} 和 $t_{\text{two-sample}}$ 的差别, 并讨论为什么有这种差别。

证明

1. 成对比较的假设检验: 令 $z_1=y_{21}-y_{11}, z_2=y_{22}-y_{12}$. 因此:

$$t_{\text{pair}} = \frac{\bar{z}}{s_z/\sqrt{n}} = \frac{y_{22}+y_{21}-y_{11}-y_{12}/2}{\sqrt{\sum_{i=1}^{2}\left(y_{2i}-y_{1i}-(\overline{y_2}-\overline{y_1})\right)^2/2}} = \frac{y_{22}+y_{21}-y_{11}-y_{12}}{|y_{11}+y_{22}-y_{21}-y_{12}|} \sim t_1$$

2. 先证明 $\text{sgn}(x)$ 与 $|x|$ 独立, 已知 x 是对称的随机变量:

$$P(x=a \mid |x|=a) = P(x=-a \mid |x|=a) = \frac{1}{2}, \quad \forall a$$

因此:

$$P(\text{sgn}(x)=1 \mid |x|=t) = P(x=t \mid |x|=t) = \frac{1}{2} = P(\text{Sgn}(x)=1)$$

$\text{Sgn}(x)$ 与 $|x|$ 独立。

可知 $Sgn(x)$ 与 σ 独立同分布, 于是由 $x=Sgn(x)|x|$ 可知, $x \overset{d}{=} \sigma|x|$ 。同样的, $\text{Sgn}(x)\sigma \overset{d}{=} \sigma$, 相互独立, 均与 x 独立。所以 $x \overset{d}{=} \sigma|x| = \sigma\,\text{Sgn}(x)x \overset{d}{=} \sigma x$

3. 我们反复使用上问结论, 注意到 x_1, x_2 独立同分布且为对称随机变量:

$$\frac{x_1}{x_2} \overset{d}{=} \frac{\operatorname{Sgn}(x_1)|x_1|}{\operatorname{Sgn}(x_2)|x_2|} \overset{d}{=} \frac{\sigma|x_1|}{|x_2|} \overset{d}{=} \frac{x_1}{|x_2|}$$

4. 由于没有分组, 样本地位对称, 此为方差相同的两样本 t 检验:
检验统计量为:

$$T_w = \frac{\bar{Y} - \bar{X}}{S_w}\sqrt{\frac{mn}{m+n}} \sim t_{n+m-2}$$

即

$$t_{two-sample} = \frac{y_{21} + y_{22} - y_{11} - y_{12}}{\sqrt{2(y_{11} - \overline{y_1})^2 + 2(y_{12} - \overline{y_1})^2 + 2(y_{21} - \overline{y_2})^2 + 2(y_{22} - \overline{y_2})^2}}$$

化简即得:

$$t_{two-sample} = \frac{y_{21} + y_{22} - y_{11} - y_{12}}{\sqrt{(y_{11} - y_{12})^2 + (y_{21} - y_{22})^2}} \sim t_2$$

成对比较与两样本 t 检验的区别在于假设的样本情形不同: 在后者中, 我们假设来自两个正态总体的样本相互独立, 但现实情形中, 有可能两个正态的样本是来自于同一个总体上的重复观察, 于是我们引进了成对比较检验, 通过作差消除其他方面的影响, 转而研究引起变化的量, 譬如药效等。

✍ **练习 4.4** (参见第 11 讲例 3) 假设 5 个中心标准化的随机变量 Y, W, U, X, V 的协方差矩阵 (即相关系数矩阵) 如下

		Y Son's occ	W Son's 1^{st} job	U Son's ed	X Dad's occ	V Dad's ed
$\bar{Y}$	Son's occ	1.000	.541	.596	.405	.322
W	Son's 1^{st} job	.541	1.000	.538	.417	.332
U	Son's ed	.596	.538	1.000	.438	.453
X	Dad's occ	.405	.417	.438	1.000	.516
V	Dad's ed	.322	.332	.453	.516	1.000

假设线性模型（因为所有变量已经中心化，故无截距）

$$U = \beta_1 X + \beta_2 V + \epsilon, \epsilon \sim N(0, \sigma^2)$$

(a) 试求 β_1, β_2, σ 的 LS 估计。
(b) 求该模型的决定系数 R^2, 回归方程的显著性检验统计量 F 及其 p 值 ($n = 20000$).
(c) 求 $H_0 : \beta_1 = 0$ 的 t 检验统计量及其 p 值。

注 **注意中心化模型的自由度参数!**

解

1. 已知线性模型为: $U = \beta_1 X + \beta_2 V + \epsilon, \epsilon \sim N(0, \sigma^2)$.
对应的相关系数矩阵为:

$$S = \begin{pmatrix} 1 & 0.438 & 0.453 \\ 0.438 & 1 & 0.516 \\ 0.453 & 0.516 & 1 \end{pmatrix} = \begin{pmatrix} S_{UU} & S_{U\boldsymbol{Z}} \\ S_{\boldsymbol{Z}U} & S_{\boldsymbol{ZZ}} \end{pmatrix}$$

由前已知的 LS 估计:

$$\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = S_{ZZ}^{-1}S_{ZU} = \begin{pmatrix}1 & 0.516\\ 0.516 & 1\end{pmatrix}^{-1}\begin{pmatrix}0.438\\ 0.453\end{pmatrix} = \begin{pmatrix}0.279\\ 0.310\end{pmatrix}$$

$$\hat{\sigma^2} = S_{UU} - S_{UZ}S_{ZZ}^{-1}S_{ZU} = 0.737, \quad \hat{\sigma} = 0.858$$

2. 代入回归方程的绝对系数公式有:

$$R^2 = \frac{S_{UZ}S_{ZZ}^{-1}S_{ZU}}{S_{UU}} = 1 - \hat{\sigma^2} = 0.263$$

显著性检验统计量为:(注意这里一个自由度为约束个数 $q = 2$, 另一个为 $n - p = 20000 - 2 = 19998$, 中心化模型没有截距项，自变量个数和参数个数都是 $p = 2$).

$$F = \frac{n-p}{q} \times \frac{R^2}{1-R^2} \sim F_{q,n-p}$$

代入数值得

$$F = 3549.78$$

p 值计算为:

$$p = P\left(F_{2,19998} > 3549.78\right)$$

3. 在 $H_0 : \beta_1 = 0$ 下的 t 检验统计量为:

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\left\|X^{\perp}\right\|} = \sqrt{n-p}\frac{r_p}{\sqrt{1-r_p^2}} \sim t_{n-p}$$

其中 r_p 为 U 与 X 的偏相关系数, 计算为:

$$r_p = \frac{r_{UX} - r_{uv}r_{vx}}{\sqrt{1-r_{uv}^2}\sqrt{1-r_{vx}^2}} = \frac{0.438 - 0.453 \times 0.516}{\sqrt{1-0.453^2}\sqrt{1-0.516^2}} = 0.267$$

p 值为 $p = P\left(t_{19998} > 39.18\right)$

注 线性模型的 F 检验中，约束 $A_{q\times p}\beta_{p\times 1} = c_0$。$p$ 是模型参数的个数，在中心化模型里是自变量的个数，在一般模型里是自变量个数 +1 (多了一个截距参数)。这里 q 是约束的个数，是 A 的行数，A 的每一行给出一个约束方程。由于每一个约束相当于少了一个参数，所以 q 也可以表示零假设下的模型与原模型的参数个数之差。只有模型和课件上的完全相同才可以直接代公式。n 是样本量。

检验统计量

$$F = \frac{\left\|\hat{y} - \hat{y_0}\right\|^2/q}{\|y - \hat{y}\|^2/(n-p)}$$

可以理解为分子分母均除以各自的自由度：q 是 $\hat{y}$ 与 $\hat{y}_0$ 对应的模型参数个数之差，$n-p$ 是 y 与 $\hat{y}$ 对应的模型参数个数之差。