作业 **11**

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：

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

以线性模型表示如下：

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

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，$n = n_1 + ... + n_K, K \geq 2$. 求 $H_0 : \mu_k = k\mu_1 (k = 1,2,...,K)$ 的 F 检验统计量.

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 α, $\alpha-\beta$ （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 $N(0,\sigma^2)$ （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。

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 (y_{1j},y_{2j})，$j = 1,2$. 其中下标中的 1，2 分别代表处理 1 和 2。数据列表如下：

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

假设正态模型:$y_{ij} \sim N(\mu_{ij},\sigma^2), i,j = 1,2$, 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 $F = t_{pair}^2$，其中 t_{pair} (成对 t 检验统计量) 如下

$$t_{pair} = \frac{y_{21} + y_{22} - y_{11} - y_{12}}{|y_{11} + y_{22} - y_{12} - y_{21}|} \overset{H_0}{\sim} t_1 \tag{1}$$

   （评注：t_1 分布也称作 Cauchy 分布）

(b) 假设 r.v. x 关于 0 对称，$P(x=0)=0$, $P(sgn(x)=\pm 1)=1/2$, 其中 $sgn(x)$ 代表 x 的符号。证明 $x=sgn(x)|x|$ 中 $sgn(x)$ 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

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

(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 $\sim N(\mu_1, \sigma^2)$, 接受处理 2 的 y_{21}, y_{22} iid $\sim N(\mu_2, \sigma^2)$, 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	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 $t_{two-sample}$ 的差别，并讨论为什么有这种差别。

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
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。