作业 **7**

下列各题都基于简单线性回归模型：假设独立样本 $(x_i, y_i), i = 1, ..., n$ 满足下述模型

$$y_i = a + bx_i + \epsilon_i, \epsilon_i \sim (0, \sigma^2), \text{且 } x_i \text{ 与 } \epsilon_i \text{ 独立}, i = 1, ..., n.$$

未知参数 $a, b, \sigma^2$ 的 LS 估计分别记为 $\hat{a}, \hat{b}, \hat{\sigma}^2$。记 $\mathbf{x} = (x_1, ..., x_n)^\top$。

1. 对于简单线性模型，证明 $r_{\hat{y}y} = |r_{xy}|$ （前者为拟合值与响应变量的相关系数）。

2. 定义最小二乘得到的残差 $e_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$, 证明

$$E(e_i) = 0, \quad var(e_i|\mathbf{x}) = \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{s_{xx}}\right)\sigma^2,$$

   由此证明 $\hat{\sigma}^2$ 是 $\sigma^2$ 的无偏估计。

3. 证明第六讲命题 5。

4. 对任一给定的 $x_0 \in R$, 均值函数或回归函数 $m(x_0) = E(y|x = x_0) = a + bx_0$ 的 LS 估计为

$$\hat{m}(x_0) = \hat{a} + \hat{b}x_0,$$

   其中 $\hat{a}, \hat{b}$ 是 $a, b$ 的 LS 估计。

   (a) 证明 $E(\hat{m}(x_0)) = m(x_0)$，$\mathrm{var}(\hat{m}(x_0)|\mathbf{x}) = \frac{\sigma^2}{n} + \frac{(x_0-\bar{x})^2\sigma^2}{s_{xx}}$.

   (b) 假设误差 $\epsilon_1, ..., \epsilon_n$ iid $\sim N(0, \sigma^2)$, 证明 $\hat{m}(x_0)$ 与 $\hat{\sigma}^2$ 独立，且

$$\frac{\hat{m}(x_0) - m(x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \sim t_{n-2}.$$

5. 简单线性模型的斜率估计 $\hat{b} = s_{xy}/s_{xx} = \sum(x_i - \bar{x})y_i/s_{xx} = \sum c_{0i}y_i$ 是 $y_1, ..., y_n$ 的线性组合，其中 $c_{0i} = (x_i - \bar{x})/s_{xx}$。假设 $\widetilde{b} = \sum c_i y_i$ 是 $b$ 的任一线性无偏估计，其中 $c_1, ..., c_n$ 只与 $\mathbf{x} = (x_1, ..., x_n)^\top$ 有关，与 $y_i$'s 无关。

   (a) 由 $\widetilde{b}$ 的无偏性，证明 $c_1, ..., c_n$ 满足约束 $\sum c_i = 0, \sum c_i x_i = 1$。

   (b) (Gauss-Markov 定理的特殊情况) 证明 $\mathrm{var}(\widetilde{b}|\mathbf{x}) \geq \sigma^2/s_{xx} = \mathrm{var}(\hat{b}|\mathbf{x})$.

6. 2000 年联合国的关于 193 个国家或地区的人口统计数据, 包括每个国家 (或地区) 的女性人均生育数目 (Fertility) 和人均国民生产总值 (PPgdp, 单位：千美元)。部分数据如下。

|  | Fertility | PPgdp |
|---|---|---|
| Afghanistan | 6.80 | 98 |
| Albania | 2.28 | 1317 |
| Algeria | 2.80 | 1784 |
| Angola | 7.20 | 739 |
| Argentina | 2.44 | 7163 |
| Armenia | 1.15 | 687 |
| Australia | 1.70 | 18788 |
| ... | | |

(完整数据集参见 R package alr4 中的 UN1）

考虑线性模型

$$\text{Fertility} = a + b \times \text{PPgdp} + \epsilon, \quad \epsilon \sim (0, \sigma^2)$$

下面是 R 软件的部分输出结果:

```
Call: lm(formula = Fertility ~ PPgdp, data = UN1)

Coefficients:
              Estimate   Std. Error   t value    Pr(>|t|)
 (Intercept)    3.733       ①         28.040    < 2e-16 ***
   PPgdp       -0.085      0.012        ②       1.15e-11 ***

Residual standard error: 1.526 on  ③  degrees of freedom
Multiple R-squared:  ④
```

(a) 填写①-④处的数字。

(b) 计算 Fertility 和 PPgdp 的样本方差和样本相关系数。

(c) 已知所有 193 个国家或地区的 PPgdp 的平均值为 6408 美元，求全世界（即 193 个国家或地区）的 Fertility 的平均值。

(d) 试解释 PPgdp 的回归系数估计值 $-0.085$ 的含义。

7. 老忠实 (或老实) 喷泉 (Old faithful geyser) 是美国黄石公园的一个间歇式热喷泉。除了每天零点到清晨 6 点之间，1980 年 10 月份的所有喷水持续时间 (Duration, 单位：秒）以及到下一次喷发的间隔时间 (Interval, 单位：分钟) 被记录了下来，共有 270 条记录 (数据集 alr4: oldfaith), 例如前 5 条记录如下：

| Duration | Interval |
|---|---|
| 216 | 79 |
| 108 | 54 |
| 200 | 74 |
| 137 | 62 |
| 272 | 85 |
| ... | |

其中第一次喷水持续 216 秒，其后经过 79 分钟再次喷水并持续了 108 秒，等等。Duration ($y$) 和 Interval ($x$) 的平均值分别是 209.9 秒和 71.1 分钟, (Duration,Interval) 的样本协方差矩阵为

$$S = \begin{pmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{pmatrix} = \frac{1}{n-1} \begin{pmatrix} s_{yy} & s_{yx} \\ s_{xy} & s_{xx} \end{pmatrix} = \begin{pmatrix} 4677.5 & 827.3 \\ 827.3 & 182.2 \end{pmatrix}$$

注意区分其中的记号，其中小写 $s_{ab} = \sum(a_i - \bar{a})(b_i - \bar{b})$，大写 $S_{ab} = s_{ab}/(n-1)$ 为样本协方差或方差。假设如下线性模型

$$\text{Duration} = a + b \times \text{Interval} + \epsilon, \quad \epsilon \sim (0, \sigma^2),$$

(a) 试求 LS 估计 $\hat{a}, \hat{b}$。如果某次喷水时间很短，你预期等待下次喷水的时间较长还是较短？

(b) 求 LS 估计 $\hat{\sigma}^2$ 及其标准差，以及 $H_0 : b = 0$ 的 $t$ 检验统计量；

(c) 求回归方程的决定系数 $R^2$;

(d) 如果某次喷水时间为 200 秒，试预测为了观看下次喷水需要等待多长时间。