## 4.7 HW 7

作业 7 链接

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

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

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

✍ **练习 4.1** 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)。

**证明** 由

$$\hat{y}_i - \bar{\hat{y}} = \left(\hat{a} + \hat{b}x_i\right) - \left(\hat{a} + \hat{b}\bar{x}\right) = \hat{b}\left(x_i - \bar{x}\right),$$

得

$$\begin{aligned} r_{\hat{y}y} &= \frac{\sum_{i=1}^n \left(\hat{y}_i - \bar{\hat{y}}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^n \left(\hat{y}_i - \bar{\hat{y}}\right)^2}\sqrt{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2}} \\ &= \frac{\hat{b}\sum_{i=1}^n \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\hat{b}^2\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2}} \\ &= \operatorname{Sgn}(\hat{b}) \cdot \frac{\sum_{i=1}^n \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2}} \\ &= \operatorname{Sgn}\left(r_{xy}\right) \cdot r_{xy} = \left|r_{xy}\right|. \end{aligned}$$

✍ **练习 4.2** 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

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

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

**证明** 由 HW6 第三题：

$$\operatorname{var}(\hat{a}) = \sigma^2/n + \bar{x}^2\sigma^2/s_{xx}, \quad \operatorname{var}(\hat{b}) = \sigma^2/s_{xx}, \quad \operatorname{cov}(\hat{a}, \hat{b}) = -\bar{x}\sigma^2/s_{xx}.$$

$$\begin{aligned} cov(y_i, \hat{b}) &= cov\left(y_i, \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{s_{xx}}\right) \\ &= (x_i - \bar{x})/s_{xx} var(y_i) \\ &= (x_i - \bar{x})/s_{xx}\sigma^2. \\ cov(y_i, \hat{a}) &= cov(y_i, \bar{y} - \hat{b}\bar{x}) \\ &= cov(y_i, \bar{y}) - \bar{x}cov(y_i, \hat{b}) \\ &= \frac{1}{n}\sigma^2 - \bar{x}(x_i - \bar{x})/s_{xx}\sigma^2. \end{aligned}$$

于是

$$\begin{aligned} var(e_i) &= var(y_i - \hat{a} - \hat{b}x_i) \\ &= cov(y_i - \hat{a} - \hat{b}x_i, y_i - \hat{a} - \hat{b}x_i) \\ &= var(y_i) + var(\hat{a}) + x_i^2 var(\hat{b}) - 2cov(y_i, \hat{a}) - 2x_i cov(y_i, \hat{b}) + 2x_i cov(\hat{a}, \hat{b}) \\ &= \sigma^2 + \sigma^2/n + \bar{x}^2\sigma^2/s_{xx} + x_i^2/s_{xx}\sigma^2 - 2\frac{1}{n}\sigma^2 + 2\bar{x}(x_i - \bar{x})/s_{xx}\sigma^2 - 2x_i(x_i - \bar{x})/s_{xx}\sigma^2 - 2x_i\bar{x}\sigma^2/s_{xx} \\ &= \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{s_{xx}}\right)\sigma^2. \end{aligned}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2,$$

故

$$E\hat{\sigma}^2 = \frac{1}{n-2}E\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} var(e_i) = \sigma^2.$$

✍ **练习 4.3** 证明第六讲命题 5 。

**证明** 命题和证明参见 week 06 P20-21.

✍ **练习 4.4** 对任一给定的 $x_0 \in R$, 均值函数或回归函数 $m(x_0) = E(y \mid x = x_0) = a + bx_0$ 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 $E(\hat{m}(x_0)) = m(x_0)$, $\text{var}(\hat{m}(x_0) \mid \mathbf{x}) = \frac{\sigma^2}{n} + \frac{(x_0-\bar{x})^2\sigma^2}{s_{xx}}$.

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

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

**证明**

- $$\begin{aligned} E(\hat{m}(x_0)) &= E\hat{a} + E\hat{b}x_0 = a + bx_0 = m(x_0) \\ var(\hat{m}(x_0)) &= var(\hat{a} + \hat{b}x_0) = var(\hat{a}) + var(\hat{b})x_0^2 + 2x_0 cov(\hat{a}, \hat{b}) = \sigma^2(1/n + (x_0 - \bar{x})^2/s_{xx}). \end{aligned}$$

- 接下来的证明类似上题, 只是在记号上稍作改变。
  给定 $x_0$ 和 $\boldsymbol{x}$, 由于 $\hat{\sigma}^2$ 与 $\hat{a}$ 和 $\hat{b}$ 独立 (我们可以从上一题中学到这一点), 我们有

$$\hat{m}(x_0) = \hat{a} + \hat{b}x_0 = g(\hat{a}, \hat{b}) \tag{4.1}$$

  与 $\hat{\sigma^2}$ 无关。此外, 请注意

$$\hat{m}(x_0) - m(x_0) = \sum_{i=1}^{n}\left(\frac{(x_i - \bar{x})(x_0 - \bar{x})}{s_{xx}} + \frac{1}{n}\right)\varepsilon_i. \tag{4.2}$$

  这给出

$$\frac{\hat{m}(x_0) - m(x_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} = \frac{\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})(x_0-\bar{x})}{s_{xx}} + \frac{1}{n}\right)\varepsilon_i}{\sqrt{\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})(x_0-\bar{x})}{s_{xx}} + \frac{1}{n}\right)^2\sigma^2}} \sim \mathcal{N}(0,1). \tag{4.3}$$

  因此,

$$\frac{\hat{m}(x_0) - m(x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} = \frac{\hat{m}(x_0) - m(x_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}} \Big/ \sqrt{\frac{(n-2)\hat{\sigma}^2}{(n-2)\sigma^2}} \sim t_{n-2}. \tag{4.4}$$

✍ **练习 4.5**

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 $y_1, \ldots, y_n$ 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 $c_1, \ldots, c_n$ 只与 $\mathbf{x} = (x_1, \ldots, x_n)^\top$ 有关, 与 $y_i'$'s 无关。

(a) 由 $\widetilde{b}$ 的无偏性, 证明 $c_1, \ldots, c_n$ 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 $\operatorname{var}(\widetilde{b} \mid \mathbf{x}) \geq \sigma^2/s_{xx} = \operatorname{var}(\hat{b} \mid \mathbf{x})$.

**证明** (a) 由于对于任意 $b \in \mathbb{R}$ 都成立

$$b = \mathbb{E}\tilde{b} = \sum_{i=1}^{n} c_i \mathbb{E} y_i = \sum_{i=1}^{n} c_i \left(a + b x_i\right) = a + b \sum_{i=1}^{n} c_i x_i$$

我们有

$$\sum_{i=1}^{n} c_i = 0, \quad \sum_{i=1}^{n} c_i x_i = 1$$

(b) 注意到

$$\tilde{b} - b = \sum_{i=1}^{n} c_i \left(y_i - a - b x_i\right) = \sum_{i=1}^{n} c_i \varepsilon_i$$

由此得到

$$\operatorname{Var}(\tilde{b} \mid x) = \sigma^2 \sum_{i=1}^{n} c_i^2$$

应用 Cauchy-Schwarz 不等式，我们有

$$\left(\sum_{i=1}^{n} c_i^2\right)\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right) \geq \left(\sum_{i=1}^{n} c_i (x_i - \bar{x})\right)^2 = \left(\sum_{i=1}^{n} c_i x_i\right)^2 = 1$$

因此

$$\operatorname{Var}(\tilde{b} \mid x) \geq \frac{\sigma^2}{s_{xx}} = \operatorname{Var}(\hat{b} \mid x)$$

**以下两题不是编程题，要求不借助 R 语言，只利用题目中的数据来算出其他的量。**

**练习 4.6** 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

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

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

下面是 R 软件的部分输出结果:

(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。

**解**

- ①$std.error = \frac{Estimate}{t-value} = 0.133$,
  ②$t-value = \frac{Estimate}{std.error} = -7.083$,
  ③自由度 $= n - p = 191$,
  ④$t = \sqrt{n-2}\frac{r}{1-r^2} \Rightarrow R^2 = \frac{t^2/(n-2)}{1+t^2/(n-2)} = \frac{7.083^2/191}{1+7.083^2/191} = 0.208$.
- 由于

$$\mathrm{se}(\hat{b}) = \sqrt{\frac{\hat{\sigma^2}}{s_{xx}}}, \quad \hat{b} = \frac{s_{xy}}{s_{xx}}, \quad r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}},$$

我们有

$$s_{xx} = \frac{\hat{\sigma^2}}{\mathrm{se}(\hat{b})^2} = 16171.36, \quad s_{xy} = \hat{b}s_{xx} = -1374.566,$$
$$s_{yy} = \frac{s_{xy}^2}{s_{xx}\cdot r^2} = 561.6152, \quad r = -\sqrt{r^2} = -0.456$$

由此得到

$$\mathcal{S}_x = \frac{1}{n-1}s_{xx} = 84.226, \quad \mathcal{S}_y = \frac{1}{n-1}s_{yy} = 2.925, \quad r_{xy} = r = -0.456$$

- 由于 $\bar{y} = \hat{a} + \hat{b}\bar{x}$，我们有

$$\overline{\text{Fertility}} = \hat{a} + \hat{b}\,\overline{\text{Education}} = 3.188$$

- 若一个国家比另一个国家的 PPgdp 高 1000 美元，则前者比后者的女性人均剩余数目大约低 -0.085 个.

注 关联而非因果！这是观察数据的回归结果。

### 练习 4.7

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 $s_{ab}=\sum\left(a_i-\bar{a}\right)\left(b_i-\bar{b}\right)$, 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

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

(a) (原题有误) 试求 LS 估计 $\hat{a},\hat{b}$ 。如果等待了很久，你预期下次喷水时间较长还是较短?

(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) (原题有误) 如果某人已经等待了一小时，他预测下次喷水不少于多长时间?

**解**

(a) 由于 $\hat{b}=\frac{s_{xy}}{s_{xx}}$ 且 $\hat{a}=\bar{y}-\hat{b}\bar{x}$，我们有 $\hat{b}=4.541$ 和 $\hat{a}=-112.938$。如果这个模型工作良好，那么在等待了很久的情况下，下一次喷水时间较长。(因为 $\hat{b}>0$).

(b) 由于 $\hat{\sigma^2}=\frac{1}{n-2}RSS=\frac{1}{n-2}\left(s_{yy}-\frac{s_{xy}^2}{s_{xx}}\right)=924.90$。此外，我们知道 $H_0:b=0$ 的 $t$ 统计量是

$$t=\frac{\hat{b}}{se(\hat{b})}=\frac{\sqrt{s_{xx}}\hat{b}}{\hat{\sigma}}=33.061$$

(c) 有

$$R^2=\frac{t^2}{t^2+(n-2)}=0.803$$

(d) 从线性模型中我们可以得知

$$\text{Duration}=\hat{a}+\hat{b}\times\text{Interval}=159.522\ \text{（秒）}$$