Homework 14

Put your name and student ID here

2020-12-09

Q1: Consider the multiple linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$, $\boldsymbol{X}$ is the $n \times p$ design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Assume that $\text{rank}(X) = p < n$, $E[\boldsymbol{\epsilon}] = \boldsymbol{0}$, and $\text{Var}[\boldsymbol{\epsilon}] = \sigma^2 I_n$ with $\sigma > 0$.

(a). Show that the covariance matrix of the least squares estimates is diagonal if and only if the columns of $\boldsymbol{X}$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, are orthogonal, that is $\boldsymbol{X}_i^\top \boldsymbol{X}_j = 0$ for $i \neq j$.

(b). Let $\hat{y}_i$ and $\hat{\epsilon}_i$ be the fitted values and the residuals, respectively. Show that $n\sigma^2 = \sum_{i=1}^n \text{Var}[\hat{y}_i] + \sum_{i=1}^n \text{Var}[\hat{\epsilon}_i]$.

(c). Suppose further that $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n)$, and you use F test to handle the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \; vs. \; H_1 : \sum_{i=1}^{p-1} \beta_i^2 \neq 0.$$

If the coefficient of determination $R^2 = 0.58$, $p = 5$ and $n = 15$, is the null rejected at the significance level $\alpha = 0.05$? $(F_{0.95}(4, 10) = 3.48, F_{0.95}(5, 10) = 3.33, t_{0.95}(10) = 1.81)$

Q2: Consider the multiple linear model $Y = X\beta + \epsilon$, where X is the $n \times p$ design matrix, $\beta = (\beta_0, \ldots, \beta_{p-1})^\top$ is a vector of p parameters, and the error $\epsilon \sim N(0, \sigma^2 I_n)$. Now consider the problem of estimating $\theta = \beta_0 + \beta_1 + \cdots + \beta_{p-1}$. Assume that $\text{rank}(X) = p < n$. Let $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_{p-1})^\top$ be the least squares estimate of β. Let $\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 + \cdots + \hat{\beta}_{p-1}$.

(a) Show that $\hat{\theta}$ is an unbaised estimate of θ.

(b) Find the variance of the estimate $\hat{\theta}$.

(c) Let $\hat{\theta}_c = c^\top Y$ be an unbiased estimate of θ for any $\beta \in \mathbb{R}^{p \times 1}$, where $c \in \mathbb{R}^{n \times 1}$ is any fixed vector. Prove that $\text{Var}(\hat{\theta}_c) \geq \text{Var}(\hat{\theta})$. (Notice that $\hat{\theta}$ is also a linear combination of y_i. This result implies that $\hat{\theta}$ is the best linear unbiased estimator for θ.)

Q3: Consider the simple linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Use the F-test method derived in the multiple linear model to test the hypothesis $H_0 : \beta_1 = 0 \; vs. \; H_1 : \beta_1 \neq 0$, and see whether the F-test agrees with the earlier t-test derived in the simple linear models.

Q4: Consider the linear model in matrix formalism

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$, $\boldsymbol{X}$ is the $n \times p$ design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \sim N(\boldsymbol{0}, \sigma^2 I_n)$ with unknown $\sigma > 0$. Assume that $\text{rank}(\boldsymbol{X}) = r < p$.

(a) Show that the least squares estimator (LSE) for $\boldsymbol{\beta}$ is not unique.

(b) Show that there exists an $n \times r$ submatrix $\boldsymbol{X}^*$ of $\boldsymbol{X}$ with rank r such that $\boldsymbol{X} = \boldsymbol{X}^* \boldsymbol{Q}$, where $\boldsymbol{Q}$ is a $r \times p$ matrix.

(c) Let $\beta^* = Q\beta$. Then the linear model becomes $Y = X^*\beta^* + \epsilon$. Find an LSE for β^* and show that the LSE is unique. Find an unbiased estimate of σ^2 and show its variance.

Q5: 这是一份 1994 年收集 1379 个对象关于收入 (earn)、身高 (height，单位：英寸)、性别 (sex)、教育水平 (ed，单位：年)、年龄 (age) 等信息的数据集。下面展示前 **6** 组数据 **(**不是完整数据**)**。

earn	height	sex	race	ed	age
79571.30	73.89	male	white	16	49
96396.99	66.23	female	white	16	62
48710.67	63.77	female	white	16	33
80478.10	63.22	female	other	16	95
82089.35	63.08	female	white	17	43
15313.35	64.53	female	white	15	30

根据完整的数据集: http://www.hezhijian.com.cn/hw/wages.csv，建立回归模型探索收入与其他变量之间的关系. 根据回归的结果回答以下问题:

1. 长的越高的人挣钱越多？男性是否比女性挣得多？是否年龄越大挣得越多？高学历是不是挣得更多？
2. 请根据你自己的个人数据预测下你在未来不同年龄段收入的置信区间和预测区间。PS: 这里身高单位为英寸（1 厘米 =0.3937008 英寸），教育年长是指你一共读了多少年的书（不算幼儿园时间, 例如大部分大三学生读书时间 $=6+3+3+3=15$, 当然留级或者跳级除外.）。

在 R 中你可以这样导入数据:

```
wages <- read.csv(url('http://www.hezhijian.com.cn/hw/wages.csv'))
head(wages) # 展示前 6 行
```

```
##       earn height    sex ed age
## 1 79571.30  73.89   male 16  49
## 2 96396.99  66.23 female 16  62
## 3 48710.67  63.77 female 16  33
## 4 80478.10  63.22 female 16  95
## 5 82089.35  63.08 female 17  43
## 6 15313.35  64.53 female 15  30
```