Homework 15

Put your name and student ID here

2021-06-10

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 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.

Q3: 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**)**。

根据完整的数据集: http://www.hezhijian.com.cn/media/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.）。

如果使用 R，下面这种方式可以得到该数据集。

```
data=read.csv('http://www.hezhijian.com.cn//media//hw//wages.csv')
knitr::kable(head(data),caption = ' 收入数据集前六行')
```

Table 1: 收入数据集前六行

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

Q4: Pima 数据集来自美国国立糖尿病、消化与肾脏疾病研究所，用于研究比马 (Pima) 印第安妇女糖尿病情况。该数据集中的自变量为怀孕次数 (npreg)、血糖浓度 (glu)、血压 (bp)、皮肤厚度 (skin)、体重指数 (bmi)、糖尿病谱系功能 (ped)、年龄 (age)，因变量 (type) 为该妇女是否患有糖尿病。下面展示前 **6** 组数据 **(**不是完整数据**)**。根据完整的数据集: http://www.hezhijian.com.cn/media/hw/pima.csv，分别用 Logit 和 Probit 二值回归模型分析糖尿病患病率与其他变量之间的关系.

如果使用 R，下面这种方式可以得到 Pima 数据集。

```
library(MASS)
pima = rbind(Pima.tr,Pima.te) ## 完整数据集
knitr::kable(head(pima),caption = 'Pima 数据集前六行')
```

Table 2: Pima 数据集前六行

npreg	glu	bp	skin	bmi	ped	age	type
5	86	68	28	30.2	0.364	24	No
7	195	70	33	25.1	0.163	55	Yes
5	77	82	41	35.8	0.156	35	No
0	165	76	43	47.9	0.259	26	No
0	107	60	25	26.4	0.133	23	No
5	97	76	27	35.6	0.378	52	Yes