Statistical linear Model

Assignment 5.

牛圣杰 11910901

Problem 1:

(a) From the problem. We have that

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \underset{\sim}{y} = \begin{pmatrix} P_0(x_1) & \cdots & P_{p-1}(x_1) \\ \vdots & & \vdots \\ P_0(x_n) & \cdots & P_{p-1}(x_n) \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_{p-1} \end{pmatrix} + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{P} = \begin{pmatrix} P_0(x_1) & \cdots & P_{p-1}(x_1) \\ \vdots & & \vdots \\ P_0(x_n) & \cdots & P_{p-1}(x_n) \end{pmatrix}_{n*p}$ $\underset{\sim}{a} = \begin{pmatrix} a_0 \\ \vdots \\ a_{p-1} \end{pmatrix}_{p*1}$ $\underset{\sim}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$

Consider $\underset{\sim}{P_i} = \begin{pmatrix} P_i(x_1) \\ \vdots \\ P_i(x_n) \end{pmatrix}$ $\Rightarrow$, then we have $\underset{\sim}{P} = \left(\underset{\sim}{P_0}, \underset{\sim}{P_1}, \cdots, \underset{\sim}{P_{p-1}} \right)$

Then, we have

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (\underset{\sim}{y} - \underset{\sim}{P}\underset{\sim}{a})' (\underset{\sim}{y} - \underset{\sim}{P}\underset{\sim}{a}) = (\underset{\sim}{y}' - \underset{\sim}{a}'\underset{\sim}{P}')(\underset{\sim}{y} - \underset{\sim}{P}\underset{\sim}{a})$$

$$= \underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{a}'\underset{\sim}{P}'\underset{\sim}{y} - \underset{\sim}{y}'\underset{\sim}{P}\underset{\sim}{a} + \underset{\sim}{a}'\underset{\sim}{P}'\underset{\sim}{P}\underset{\sim}{a} = \underset{\sim}{y}'\underset{\sim}{y} - 2\underset{\sim}{a}'\underset{\sim}{P}'\underset{\sim}{y} + \underset{\sim}{a}'\underset{\sim}{P}'\underset{\sim}{P}\underset{\sim}{a}$$

$$\Rightarrow \frac{d\,SSE}{d\underset{\sim}{a}} = -2\underset{\sim}{P}'\underset{\sim}{y} + 2\underset{\sim}{P}'\underset{\sim}{P}\underset{\sim}{a} \quad \cdots (*)$$

Note that $\sum_{i=1}^{n} P_l(x_i) P_m(x_i) = 0$. $l \neq m$ for all l and m

$\Rightarrow$ then we have $\underset{\sim}{P_l}'\underset{\sim}{P_m} = \sum_{i=1}^{n} P_l(x_i) P_m(x_i) = 0$.

$\Rightarrow P_l \perp P_m$ for $l \neq m$

Then let $\frac{d\,SSE}{d\underset{\sim}{a}} = 0$. from $(*)$ we have $\underset{\sim}{P}'\underset{\sim}{P}\underset{\sim}{a} = \underset{\sim}{P}'\underset{\sim}{y} \Rightarrow \hat{\underset{\sim}{a}} = (\underset{\sim}{P}'\underset{\sim}{P})^{-1}\underset{\sim}{P}'\underset{\sim}{y}$

Then we have $\underset{\sim}{P}'\underset{\sim}{P} = \begin{pmatrix} P_0' \\ \vdots \\ P_{p-1}' \end{pmatrix} (P_0, \cdots, P_{p-1}) = \begin{pmatrix} P_0'P_0 & P_0'P_1 & \cdots & P_0'P_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ P_{p-1}'P_0 & P_{p-1}'P_1 & \cdots & P_{p-1}'P_{p-1} \end{pmatrix}$

$$= \begin{pmatrix} \sum_{i=1}^{n} P_0^2(x_i) & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^{n} P_1^2(x_i) & & \vdots \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sum_{i=1}^{n} P_{p-1}^2(x_i) \end{pmatrix}$$

Since $P_k^2(x) = (1 + x + \cdots + x^k)^2 \geq 0$

$\Rightarrow \underset{\sim}{P}'P$ is nonsingular and $(\underset{\sim}{P}'\underset{\sim}{P})^{-1} = \begin{pmatrix} (\sum_{i=1}^{n} P_0^2(x_i))^{-1} & 0 & \cdots & 0 \\ 0 & (\sum_{i=1}^{n} P_1^2(x_i))^{-1} & & \vdots \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & (\sum_{i=1}^{n} P_{p-1}^2(x_i))^{-1} \end{pmatrix}$

And we have $\underset{\sim}{P}'\underset{\sim}{y} = \begin{pmatrix} P_0(x_1) & P_0(x_2) & \cdots & P_0(x_n) \\ \vdots & \vdots & & \vdots \\ P_{p-1}(x_1) & P_{p-1}(x_2) & \cdots & P_{p-1}(x_n) \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n P_0(x_i)y_i \\ \vdots \\ \sum_{i=1}^n P_{p-1}(x_i)y_i \end{pmatrix}$

$$\hat{\underset{\sim}{a}} = (\underset{\sim}{P}'\underset{\sim}{P})^{-1}\underset{\sim}{P}'\underset{\sim}{y} = \begin{pmatrix} (\sum_{i=1}^n P_0^2(x_i))^{-1} & 0 & \cdots & 0 \\ 0 & (\sum_{i=1}^n P_1^2(x_i))^{-1} & & \vdots \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & (\sum_{i=1}^n P_{p-1}^2(x_i))^{-1} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n P_0(x_i)y_i \\ \vdots \\ \sum_{i=1}^n P_{p-1}(x_i)y_i \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^n P_0(x_i)y_i / \sum_{i=1}^n P_0^2(x_i) \\ \vdots \\ \sum_{i=1}^n P_{p-1}(x_i)y_i / \sum_{i=1}^n P_{p-1}^2(x_i) \end{pmatrix}$$

$\Rightarrow \hat{a}_j = \sum_{i=1}^n P_j(x_i)y_i / \sum_{i=1}^n P_j^2(x_i)$ for $j = 0, 1, \cdots, p-1$

Then, want to prove that $\hat{a}_j$'s are uncorrelated for $j=0, 1, \cdots, p-1$

Since $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, \cdots, n$.

then we have $\underset{\sim}{y} \sim N(\underset{\sim}{P}\underset{\sim}{a}, \sigma^2 \underset{\sim}{I}) \Rightarrow \hat{\underset{\sim}{a}} \sim N(\underset{\sim}{a}, (\underset{\sim}{P}'\underset{\sim}{P})^{-1}\sigma^2)$

Note that the $cov(\hat{\underset{\sim}{a}})$ is $(\underset{\sim}{P}'\underset{\sim}{P})^{-1}\sigma^2$, which is a diagonal matrix.

$\Rightarrow$ which means that $Cov(\hat{a}_i, \hat{a}_j) = 0$ for any $i \neq j$

$\Rightarrow$ $\hat{a}_j$'s are uncorrelated for $j = 0, 1, \cdots, p-1$

(b). By the problem, we have that $\hat{\underset{\sim}{a}} \sim N(\underset{\sim}{a}, (\underset{\sim}{P}'\underset{\sim}{P})^{-1}\sigma^2)$

$\hat{a}_j \sim N(a_j, \sigma^2 / \sum_{i=1}^n P_j^2(x_i))$

under the null hypothesis $H_0: a_j = 0$, $\Rightarrow \hat{a}_j \sim N(0, \sigma^2 / \sum_{i=1}^n P_j^2(x_i))$

At the same time, $\hat{\sigma}^2 = \frac{SSE}{n-p}$, where $SSE = (\underset{\sim}{y} - \underset{\sim}{P}\hat{\underset{\sim}{a}})'(\underset{\sim}{y} - \underset{\sim}{P}\hat{\underset{\sim}{a}})$

Note that $SSE = (\underset{\sim}{y} - \underset{\sim}{P}\hat{\underset{\sim}{a}})'(\underset{\sim}{y} - \underset{\sim}{P}\hat{\underset{\sim}{a}}) = (\underset{\sim}{y} - \underset{\sim}{P}(\underset{\sim}{P}'\underset{\sim}{P})^{-1}\underset{\sim}{P}'\underset{\sim}{y})'(\underset{\sim}{y} - \underset{\sim}{P}(\underset{\sim}{P}'\underset{\sim}{P})^{-1}\underset{\sim}{P}'\underset{\sim}{y})$

$= \underset{\sim}{y}'(\underset{\sim}{I} - \underset{\sim}{P}(\underset{\sim}{P}'\underset{\sim}{P})^{-1}\underset{\sim}{P}')(\underset{\sim}{I} - \underset{\sim}{P}(\underset{\sim}{P}'\underset{\sim}{P})^{-1}\underset{\sim}{P}')\underset{\sim}{y}$

$= \underset{\sim}{y}'(\underset{\sim}{I} - \underset{\sim}{P}(\underset{\sim}{P}'\underset{\sim}{P})^{-1}\underset{\sim}{P}')\underset{\sim}{y}$ $\cdots$ (**)

$\Rightarrow \dfrac{\sqrt{\sum_{i=1}^n P_j^2(x_i)} \cdot \hat{a}_j}{\sqrt{SSE/n-p}} \sim t_{n-p}$ where $SSE = \underset{\sim}{y}'(\underset{\sim}{I} - \underset{\sim}{P}(\underset{\sim}{P}'\underset{\sim}{P})^{-1}\underset{\sim}{P}')\underset{\sim}{y}$

Hence. we will reject H_0 if $\left| \dfrac{\sqrt{\sum_{i=1}^{n} P_j^2(x_i)} \cdot \hat{a}_j}{\sqrt{SSE/n-p}} \right| > t_{\frac{\alpha}{2}, n-p}$

or the p-value $= P_r \left(t_{n-p} > \left| \dfrac{\sqrt{\sum_{i=1}^{n} P_j^2(x_i)} \cdot \hat{a}_j}{\sqrt{SSE/n-p}} \right| \right) \leq \alpha$

(c). Given that $\underset{\sim}{x} = \underset{\sim}{x}^*$. Suppose $\underset{\sim}{P}^* = \underset{\sim}{P}\big|_{x=x^*} = \begin{pmatrix} P_0(x^*) \\ \vdots \\ P_{p-1}(x^*) \end{pmatrix}$

$\Rightarrow$ $E(y^*) = \underset{\sim}{P}^{*\prime} \underset{\sim}{a}$, $\widehat{E(y^*)} = \underset{\sim}{P}^{*\prime} \underset{\sim}{\hat{a}}$

$Var\left(E(\underset{\sim}{y}^*) - \widehat{E(\underset{\sim}{y}^*)} \right) = \left[\underset{\sim}{P}^{*\prime} (\underset{\sim}{P}' \underset{\sim}{P})^{-1} \underset{\sim}{P}^* \right] \sigma^2$

hence. a $100(1-\alpha)\%$ confidence interval for $E(y^*)$ is

$$\underset{\sim}{P}^{*\prime} \underset{\sim}{\hat{a}} \pm t_{\frac{\alpha}{2}, n-p} \cdot \hat{\sigma} \cdot \sqrt{\underset{\sim}{P}^{*\prime} (\underset{\sim}{P}' \underset{\sim}{P})^{-1} \underset{\sim}{P}^*}$$

Problem2

2(a)

```r
# import the data
address = getwd()
files = paste(address, "6data.csv", sep = "/")
data = read.csv(file = files)

# use the multiple linear regression model to fit

fit <- lm(Y~., data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17355 -0.55425 -0.00316  0.61569  2.02727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.22211    0.71119  17.185  < 2e-16 ***
## X1          -0.18698    0.02497  -7.489 9.04e-09 ***
## X2           0.29510    0.07349   4.016 0.000298 ***
## X3          -1.21196    1.40668  -0.862 0.394786
## X4           0.07479    0.01637   4.569 5.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9353 on 35 degrees of freedom
## Multiple R-squared:  0.7541, Adjusted R-squared:  0.726
## F-statistic: 26.84 on 4 and 35 DF,  p-value: 3.088e-10
```

Conduct a hypothesis test for the overall utility of the model

Hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ v.s. H_1 : at least one $\beta_j \neq 0$ for $j = 1, 2, 3, 4$

Then under H_0, we have

$$F(\boldsymbol{H}) \sim F_{\{q,N-k-1\}} = F_{\{4,35\}}$$

Then according to the ANOVA table, we have F statistic: 26.84 on 4 and 35 DF, p-value: 3.088e-10, which is very significant.

We can not reject the $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ And we can conclude that the model is useful for the prediction of rental rates.

2(b)

```r
# Studentized residuals
r1<-rstandard(fit)

# Studentized deleted residuals
r2<-rstudent(fit)
```

```r
# Leverage values
h<-hatvalues(fit)

# Dffits
dff<-dffits(fit)

# Cook's distance
co<-cooks.distance(fit)

# tabulate the results
table<-data.frame(Studentized=r1,Stud_deleted=r2,Leverage_val=h,Dffits=dff,Cooks_dist=co)
table
```

```
##     Studentized Stud_deleted Leverage_val        Dffits   Cooks_dist
## 1  -0.881313929 -0.878434210   0.13469748 -0.3465811939 2.418147e-02
## 2  -0.547027035 -0.541475420   0.14719530 -0.2249577359 1.032980e-02
## 3   0.114526120  0.112899333   0.36914451  0.0863623876 1.534990e-03
## 4   0.001844146  0.001817610   0.08380724  0.0005497282 6.221786e-08
## 5   0.908126276  0.905794106   0.05964149  0.2281166588 1.046110e-02
## 6  -2.520284183 -2.745620746   0.14979177 -1.1524494177 2.238163e-01
## 7  -0.239868320 -0.236611361   0.10028484 -0.0789952047 1.282644e-03
## 8   2.137663634  2.259566038   0.52194545  2.3610156478 9.978296e-01
## 9   2.429263793  2.625895934   0.20392176  1.3290197065 3.023341e-01
## 10  0.503208583  0.497771708   0.05951632  0.1252196844 3.204873e-03
## 11  0.730853920  0.725897877   0.16180864  0.3189369120 2.062290e-02
## 12 -0.303250984 -0.299280866   0.07711837 -0.0865136968 1.536902e-03
## 13  0.775534861  0.771029052   0.07742031  0.2233553000 1.009447e-02
## 14  0.271122006  0.267501818   0.15000359  0.1123748224 2.594443e-03
## 15 -0.131615597 -0.129753863   0.10232066 -0.0438067804 3.948997e-04
## 16 -0.941069237 -0.939490166   0.14157544 -0.3815356611 2.921184e-02
## 17  0.400112460  0.395260142   0.06830628  0.1070229134 2.347370e-03
## 18 -1.151851331 -1.157426559   0.14920288 -0.4846955654 4.653439e-02
## 19 -0.586051309 -0.580473602   0.18019839 -0.2721469992 1.509883e-02
## 20 -0.608227275 -0.602668830   0.05929007 -0.1513010173 4.663243e-03
## 21  0.067067672  0.066106867   0.09300209  0.0211684851 9.224501e-05
## 22  0.338608468  0.334284136   0.09243832  0.1066850849 2.335616e-03
## 23 -0.008970208 -0.008841144   0.09775913 -0.0029102204 1.743692e-06
## 24  0.270208563  0.266598685   0.08846972  0.0830557620 1.417267e-03
## 25 -0.431894646 -0.426818898   0.12784686 -0.1634151456 5.468686e-03
## 26 -1.776371827 -1.835507218   0.17601682 -0.8483479432 1.348136e-01
## 27  1.247561875  1.257897088   0.08369103  0.3801575295 2.843094e-02
## 28  0.174568468  0.172131513   0.08252251  0.0516236723 5.481995e-04
## 29 -0.733720368 -0.728789274   0.07953229 -0.2142246442 9.303065e-03
## 30  0.730635239  0.725677315   0.10296404  0.2458563427 1.225482e-02
## 31 -1.249953962 -1.260421576   0.08660188 -0.3881051666 2.962683e-02
## 32 -0.108416507 -0.106874423   0.07103079 -0.0295526507 1.797489e-04
## 33 -1.043192193 -1.044548682   0.07718264 -0.3020860161 1.820382e-02
## 34  0.688281672  0.683015946   0.09531188  0.2216944949 9.981838e-03
## 35  1.002134178  1.002197155   0.08919865  0.3136320696 1.967054e-02
## 36  1.600626622  1.638711484   0.05806995  0.4068823945 3.158951e-02
## 37  1.721521774  1.773496669   0.13203018  0.6916950670 9.016202e-02
## 38 -1.693003760 -1.741473039   0.17171092 -0.7929115605 1.188398e-01
## 39 -0.109289486 -0.107735278   0.07536270 -0.0307574597 1.947026e-04
```

```
## 40 -0.640939528 -0.635457160   0.12206680 -0.2369487002 1.142353e-02
```

2(c)

Identification criterion:

Using the Studentized residuals, considered a point to be an outlier if absolute value of studentized residual is greater than 2.

```r
abr1<-abs(r1)
sort(abr1, decreasing = TRUE)[1:5]
```

```
##        6        9        8       26       37
## 2.520284 2.429264 2.137664 1.776372 1.721522
```

By the table, we have that the 6th, 8th and 9th observations are high leverage points

2(d)

Identification criterion:

The ith is an outlier in X point if the leverage value satisfy$h_{ii} \geq 2\frac{k+1}{n}$

```r
# note that mean(h)=(k+1)/n
sort(h, decreasing=TRUE)[1:5]/mean(h)
```

```
##        8        3        9       19       26
## 4.175564 2.953156 1.631374 1.441587 1.408135
```

By the table, we have that the 3th and 8th observations are high leverage points

2(e)

Consider the Dffits:

The ith is a influential observation if $|diffits_i| \geq 2\sqrt{\frac{k+1}{n}}$

```r
# note that mean(h)=(k+1)/n
f<-sqrt(mean(h))
abdff<-abs(dff)
sort(abdff, decreasing = TRUE)[1:6]/f
```

```
##        8        9        6       26       38       37
## 6.677961 3.759035 3.259619 2.399490 2.242693 1.956409
```

By the table, we have that the 6th, 8th, 9th, 26th and 38th are observations are influential points

Consider the Cook's distance:

The ith is a influential if its Cook's distance is larger than $F_{\{0.05,k+1,n-(k+1)\}} = F_{\{0.05,5,35\}}$

```r
g<-qf(0.5,5,35)
g
```

```
## [1] 0.8873122
```

```r
sort(co, decreasing = TRUE)[1:5]-g
```

```
##          8          9          6         26         38
##  0.1105174 -0.5849781 -0.6634959 -0.7524986 -0.7684724
```

By the table, we have that the 8th observation are influential point