Statistical Linear Models

Assignment 5

Hanbin Liu 11912410

1

(a)

Let

$$\mathbf{P} = \begin{pmatrix} 1 & P_1(x_1) & \dots & P_{p-1}(x_1) \\ 1 & P_1(x_2) & \dots & P_{p-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & P_1(x_n) & \dots & P_{p-1}(x_n) \end{pmatrix}$$

Then the model can be rewritten as

$$\mathbf{y} = \mathbf{P'a} + \epsilon,$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)'$, $\mathbf{a} = (a_0, a_1, ..., a_{p-1})'$ and $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)'$.
Since

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

we have

$$\begin{aligned}
\mathbf{P'P} &= \begin{pmatrix} n & \sum_{i=1}^{n} P_1(x_i) & \dots & \sum_{i=1}^{n} P_{p-1}(x_i) \\ \sum_{i=1}^{n} P_1(x_i) & \sum_{i=1}^{n} P_1^2(x_i) & \dots & \sum_{i=1}^{n} P_1(x_i)P_{p-1}(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} P_{p-1}(x_i) & \sum_{i=1}^{n} P_{p-1}(x_i)P_1(x_i) & \dots & \sum_{i=1}^{n} P_{p-1}^2(x_i) \end{pmatrix} \\
&= \begin{pmatrix} n & 0 & \dots & 0 \\ 0 & \sum_{i=1}^{n} P_1^2(x_i) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sum_{i=1}^{n} P_{p-1}^2(x_i) \end{pmatrix}
\end{aligned}$$

Hence,

$$\begin{aligned}
\hat{\mathbf{a}} &= (\mathbf{P'P})^{-1}\mathbf{P'y} \\
&= \begin{pmatrix} 1/n & 0 & \dots & 0 \\ 0 & 1/\sum_{i=1}^{n} P_1^2(x_i) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sum_{i=1}^{n} P_{p-1}^2(x_i) \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \\ P_1(x_1) & P_1(x_2) & \dots & P_1(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ P_{p-1}(x_1) & P_{p-1}(x_2) & \dots & P_{p-1}(x_n) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \\
&= \begin{pmatrix} 1/n & 1/n & \dots & 1/n \\ P_1(x_1)/\sum_{i=1}^{n} P_1^2(x_i) & P_1(x_2)/\sum_{i=1}^{n} P_1^2(x_i) & \dots & P_1(x_n)/\sum_{i=1}^{n} P_1^2(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ P_{p-1}(x_1)/\sum_{i=1}^{n} P_{p-1}^2(x_i) & P_{p-1}(x_2)/\sum_{i=1}^{n} P_{p-1}^2(x_i) & \dots & P_{p-1}(x_n)/\sum_{i=1}^{n} P_{p-1}^2(x_i) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \\
&= \begin{pmatrix} \sum_{j=1}^{n} y_j/n \\ \sum_{j=1}^{n} y_j P_1(x_j)/\sum_{i=1}^{n} P_1^2(x_i) \\ \vdots \\ \sum_{j=1}^{n} y_j P_{p-1}(x_j)/\sum_{i=1}^{n} P_{p-1}^2(x_i) \end{pmatrix}
\end{aligned}$$

That is,

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

Let $\mathbf{A} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'$, then

$$Cov(\hat{\mathbf{a}}) = Cov(\mathbf{A}\mathbf{y}) = \mathbf{A}Cov(\mathbf{y})\mathbf{A}' = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\sigma^2\mathbf{I}\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1} = \sigma^2(\mathbf{P}'\mathbf{P})^{-1},$$

which implies that $a_j's$ are uncorrelated since $(\mathbf{P}'\mathbf{P})^{-1}_{(i,j)} = 0,\ i \neq j.$

(b)

Since $\hat{a}_j$ is the linear combination of $y_i's$ and each y_i follows a normal distribution, then $\hat{a}_j$ follows a normal distribution as well. Moreover,

$$E(\hat{a}_j) = \frac{1}{\sum_{i=1}^n P_j^2(x_i)} \sum_{i=1}^n P_j(x_i)E(y_i) = \frac{1}{\sum_{i=1}^n P_j^2(x_i)} \sum_{i=1}^n P_j(x_i) \sum_{l=0}^{p-1} a_l P_l(x_i)$$

Since

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

it follows that

$$E(\hat{a}_j) = \frac{1}{\sum_{i=1}^n P_j^2(x_i)} \sum_{i=1}^n a_j P_j(x_i)P_j(x_i) = \frac{1}{\sum_{i=1}^n P_j^2(x_i)} a_j \sum_{i=1}^n P_j^2(x_i) = a_j,\ j = 0, 1, ..., p-1$$

Therefore, $\hat{\mathbf{a}} \sim N(\mathbf{a}, \sigma^2(\mathbf{P}'\mathbf{P})^{-1})$. Particularly, $\hat{a}_j \sim N(a_j, \sigma^2/\sum_{i=1}^n P_j^2(x_i))$. Then

$$\frac{\hat{a}_j - a_j}{\sqrt{\sigma^2/\sum_{i=1}^n P_j^2(x_i)}} \sim N(0,1)$$

and thus

$$\frac{\hat{a}_j - a_j}{\sqrt{\hat{\sigma}^2/\sum_{i=1}^n P_j^2(x_i)}} \sim t_{n-p},$$

where $\hat{\sigma}^2 = \frac{\text{SSE}_\text{p}}{n-p}$ and $\text{SSE}_\text{p} = \mathbf{y}'(\mathbf{I} - \mathbf{P}(\mathbf{P}'\mathbf{P})^{-\mathbf{1}}\mathbf{P}')\mathbf{y}$. This is obvious if we view $\mathbf{P}$ as $\mathbf{X}$ and $\mathbf{a}$ as β. Under $H_0 : a_j = 0$, we have

$$\frac{\hat{a}_j}{\sqrt{\hat{\sigma}^2/\sum_{i=1}^n P_j^2(x_i)}} \sim t_{n-p}.$$

We reject H_0 if

$$\left|\frac{\hat{a}_j}{\sqrt{\hat{\sigma}^2/\sum_{i=1}^n P_j^2(x_i)}}\right| \geq t_{\frac{\alpha}{2},n-p}$$

The $(1-\alpha)\%$ confidence interval of $\hat{a}_j$ is

$$\left[-t_{\frac{\alpha}{2},n-p}\hat{\sigma}\Big/\sqrt{\sum_{i=1}^n P_j^2(x_i)}, t_{\frac{\alpha}{2},n-p}\hat{\sigma}\Big/\sqrt{\sum_{i=1}^n P_j^2(x_i)}\right]$$

and p-value is given by

$$2 \times \Pr\left(t_{n-p} \geq \left|\frac{\hat{a}_j}{\sqrt{\hat{\sigma}^2/\sum_{i=1}^n P_j^2(x_i)}}\right|\right)$$

(c)

Given that $x = x^*$, we have
$$y^* = {p^*}'\mathbf{a} + \epsilon^*,$$
where $p^* = (P_0(x^*), P_1(x^*), ..., P_{p-1}(x^*))'$. Then
$$E(y^*) = {p^*}'\mathbf{a}$$
$$\widehat{E(y^*)} = {p^*}'\hat{\mathbf{a}}$$
and
$$Var(E(y^*) - \widehat{E(y^*)}) = Var({p^*}'\hat{\mathbf{a}}) = {p^*}'Cov(\hat{\mathbf{a}})p^* = {p^*}'(\mathbf{P}'\mathbf{P})^{-1}p^*\sigma^2$$
Hence, a $100(1-\alpha)\%$ confidence interval for $E(y^*)$ is
$$\left[{p^*}'\hat{\mathbf{a}} - t_{\frac{\alpha}{2},n-p}\hat{\sigma}\sqrt{{p^*}'(\mathbf{P}'\mathbf{P})^{-1}p^*}, {p^*}'\hat{\mathbf{a}} + t_{\frac{\alpha}{2},n-p}\hat{\sigma}\sqrt{{p^*}'(\mathbf{P}'\mathbf{P})^{-1}p^*}\right],$$
where $t_{\frac{\alpha}{2},n-p}$ is the $\frac{\alpha}{2}$ upper quantile of t_{n-p} distribution.

2

```
mydata <- read.csv("6data.csv", header = T)
```

(a)

Hypothesis:
$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_1 : \beta_j \neq 0 \text{ for at least one value of } j,\ \ j = 1, 2, 3, 4.$$

```
model <- lm(Y ~ X1 + X2 + X3 + X4, data = mydata)

summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = mydata)
##
## 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
```

The F-statistics is 26.84 and the degree of freedom is 4 and 35. The p-value is 3.088×10^{-10}. We should reject H_0. i.e., the model has an overall utility.

(b)

```
## (i)
studentized.residuals <- rstandard(model)
studentized.residuals
```

```
##            1            2            3            4            5            6
## -0.881313929 -0.547027035  0.114526120  0.001844146  0.908126276 -2.520284183
##            7            8            9           10           11           12
## -0.239868320  2.137663634  2.429263793  0.503208583  0.730853920 -0.303250984
##           13           14           15           16           17           18
##  0.775534861  0.271122006 -0.131615597 -0.941069237  0.400112460 -1.151851331
##           19           20           21           22           23           24
## -0.586051309 -0.608227275  0.067067672  0.338608468 -0.008970208  0.270208563
##           25           26           27           28           29           30
## -0.431894646 -1.776371827  1.247561875  0.174568468 -0.733720368  0.730635239
##           31           32           33           34           35           36
## -1.249953962 -0.108416507 -1.043192193  0.688281672  1.002134178  1.600626622
##           37           38           39           40
##  1.721521774 -1.693003760 -0.109289486 -0.640939528
```

```
## (ii)
studentized.deleted.residuals <- rstudent(model)
studentized.deleted.residuals
```

```
##            1            2            3            4            5            6
## -0.878434210 -0.541475420  0.112899333  0.001817610  0.905794106 -2.745620746
##            7            8            9           10           11           12
## -0.236611361  2.259566038  2.625895934  0.497771708  0.725897877 -0.299280866
##           13           14           15           16           17           18
##  0.771029052  0.267501818 -0.129753863 -0.939490166  0.395260142 -1.157426559
##           19           20           21           22           23           24
## -0.580473602 -0.602668830  0.066106867  0.334284136 -0.008841144  0.266598685
##           25           26           27           28           29           30
## -0.426818898 -1.835507218  1.257897088  0.172131513 -0.728789274  0.725677315
##           31           32           33           34           35           36
## -1.260421576 -0.106874423 -1.044548682  0.683015946  1.002197155  1.638711484
##           37           38           39           40
##  1.773496669 -1.741473039 -0.107735278 -0.635457160
```

```r
## (iii)
h <- hatvalues(model)
h
```

```
##          1          2          3          4          5          6          7
## 0.13469748 0.14719530 0.36914451 0.08380724 0.05964149 0.14979177 0.10028484
##          8          9         10         11         12         13         14
## 0.52194545 0.20392176 0.05951632 0.16180864 0.07711837 0.07742031 0.15000359
##         15         16         17         18         19         20         21
## 0.10232066 0.14157544 0.06830628 0.14920288 0.18019839 0.05929007 0.09300209
##         22         23         24         25         26         27         28
## 0.09243832 0.09775913 0.08846972 0.12784686 0.17601682 0.08369103 0.08252251
##         29         30         31         32         33         34         35
## 0.07953229 0.10296404 0.08660188 0.07103079 0.07718264 0.09531188 0.08919865
##         36         37         38         39         40
## 0.05806995 0.13203018 0.17171092 0.07536270 0.12206680
```

```r
## (iv)
dffits(model)
```

```
##             1             2             3             4             5
## -0.3465811939 -0.2249577359  0.0863623876  0.0005497282  0.2281166588
##             6             7             8             9            10
## -1.1524494177 -0.0789952047  2.3610156478  1.3290197065  0.1252196844
##            11            12            13            14            15
##  0.3189369120 -0.0865136968  0.2233553000  0.1123748224 -0.0438067804
##            16            17            18            19            20
## -0.3815356611  0.1070229134 -0.4846955654 -0.2721469992 -0.1513010173
##            21            22            23            24            25
##  0.0211684851  0.1066850849 -0.0029102204  0.0830557620 -0.1634151456
##            26            27            28            29            30
## -0.8483479432  0.3801575295  0.0516236723 -0.2142246442  0.2458563427
##            31            32            33            34            35
## -0.3881051666 -0.0295526507 -0.3020860161  0.2216944949  0.3136320696
##            36            37            38            39            40
##  0.4068823945  0.6916950670 -0.7929115605 -0.0307574597 -0.2369487002
```

```r
## (v)
cooks.distance(model)
```

```
##            1            2            3            4            5            6
## 2.418147e-02 1.032980e-02 1.534990e-03 6.221786e-08 1.046110e-02 2.238163e-01
##            7            8            9           10           11           12
## 1.282644e-03 9.978296e-01 3.023341e-01 3.204873e-03 2.062290e-02 1.536902e-03
##           13           14           15           16           17           18
## 1.009447e-02 2.594443e-03 3.948997e-04 2.921184e-02 2.347370e-03 4.653439e-02
##           19           20           21           22           23           24
## 1.509883e-02 4.663243e-03 9.224501e-05 2.335616e-03 1.743692e-06 1.417267e-03
##           25           26           27           28           29           30
## 5.468686e-03 1.348136e-01 2.843094e-02 5.481995e-04 9.303065e-03 1.225482e-02
```

```
##            31           32           33           34           35           36
## 2.962683e-02 1.797489e-04 1.820382e-02 9.981838e-03 1.967054e-02 3.158951e-02
##            37           38           39           40
## 9.016202e-02 1.188398e-01 1.947026e-04 1.142353e-02
```

(c)

```
abs.studentized.residuals <- abs(studentized.residuals)

which(abs.studentized.residuals > 2)
```

```
## 6 8 9
## 6 8 9
```

y_6, y_8, y_9 are outlying Y observations.

Identification criterion: $h_{ii} \geq 2\frac{k+1}{n}$

(d)

```
k <- length(coef(model)) - 1
n <- length(mydata$Y)

which(h >= 2 * (k + 1) / n)
```

```
## 3 8
## 3 8
```

x_3 and x_8 are outlying observations.

(e)

Identification criterion:

Influential observations: $|\text{Dffits}| \geq 2\sqrt{\frac{k+1}{n}}$ or Cook's distance $\geq F_{0.5,k+1,n-k-1}$

```
diff <- which(abs(dffits(model)) >= 2 * sqrt((k + 1) / n))

cook <- which(cooks.distance(model) >= pf(0.5, k + 1, n - k - 1))

diff
```

```
##  6  8  9 26 38
##  6  8  9 26 38
```

```r
cook
```

```
## 8 9
## 8 9
```

x_6, x_8, x_9, x_{26} and x_{38} are the influential observations.