Homework Chapter 10

Jinhong Du 15338039

10.9 Refer to Brand preference Problem 6.5.

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .10$. State the decision rule and conclusion.

```
data1 <- read.table("CH06PR05.txt",head=FALSE,col.names = c('Y',
'X1','X2'))
Y <- data1$Y
X1 <- data1$X1
X2 <- data1$X2
n = length(Y)
X <- cbind(rep(1,n),X1,X2)
fit1 <- lm('Y~X1+X2',data=data1)
e <- fit1$residuals
df_ <- fit1$df.residual
SSE <- sum(e^2)
h <- diag(X%*%solve(crossprod(X))%*%t(X))
t <- e*sqrt(df_-1)/sqrt(SSE*(1-h)-e^2)
tab <- as.table(cbind(
  'e' = e,
  't' = t
))
round(t(tab), 4)
```

```
##         1       2       3       4       5       6       7       8       9
## e -0.1000  0.1500 -3.1000  3.1500 -0.9500 -1.7000 -1.9500  1.3000  1.2000
## t -0.0409  0.0613 -1.3606  1.3860 -0.3669 -0.6649 -0.7672  0.5046  0.4651
##        10      11      12      13      14      15      16
## e -1.5500  4.2000  2.4500 -2.6500 -4.4000  3.3500  0.6000
## t -0.6044  1.8230  0.9778 -1.1397 -2.1027  1.4897  0.2457
```

```
print(sprintf('t(%f;%d)=%f',1-0.1/2/n,df_,qt(1-0.1/2/n,df_)))
```

```
## [1] "t(0.996875;13)=3.256463"
```

```
print(sprintf('max{|t_i|}=%f',max(abs(t))))
```

```
## [1] "max{|t_i|}=2.102726"
```

$$
\begin{aligned}
d_i &= Y_i - \hat{Y}_{i(i)} \\
&= \frac{e_i}{1-h_{ii}} \\
s\{d_i\} &= MSE_{(i)}[1 + X_i^T(X_{(i)}^T X_{(i)})^{-1}X_i] \\
&= \frac{MSE_{(i)}}{1-h_{ii}} \\
t_i &= \frac{d_i}{s\{d_i\}} \\
&= \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \\
&= \frac{e_i\sqrt{n-p-1}}{\sqrt{SSE(1-h_{ii})-e_i^2}} \sim t(n-p-1)
\end{aligned}
$$

$$
H_0 : (X_i, Y_i) \text{ is an outlier} \qquad H_a : (X_i, Y_i) \text{ is not an outlier}
$$

the test statistic is t_i.

Given α, the decision rule is

If $|t_i| \geqslant t(1-\frac{\alpha}{2n}; n-p)$, then conclude H_0;

If $|t_i| < t(1-\frac{\alpha}{2n}; n-p)$, then conclude H_a;

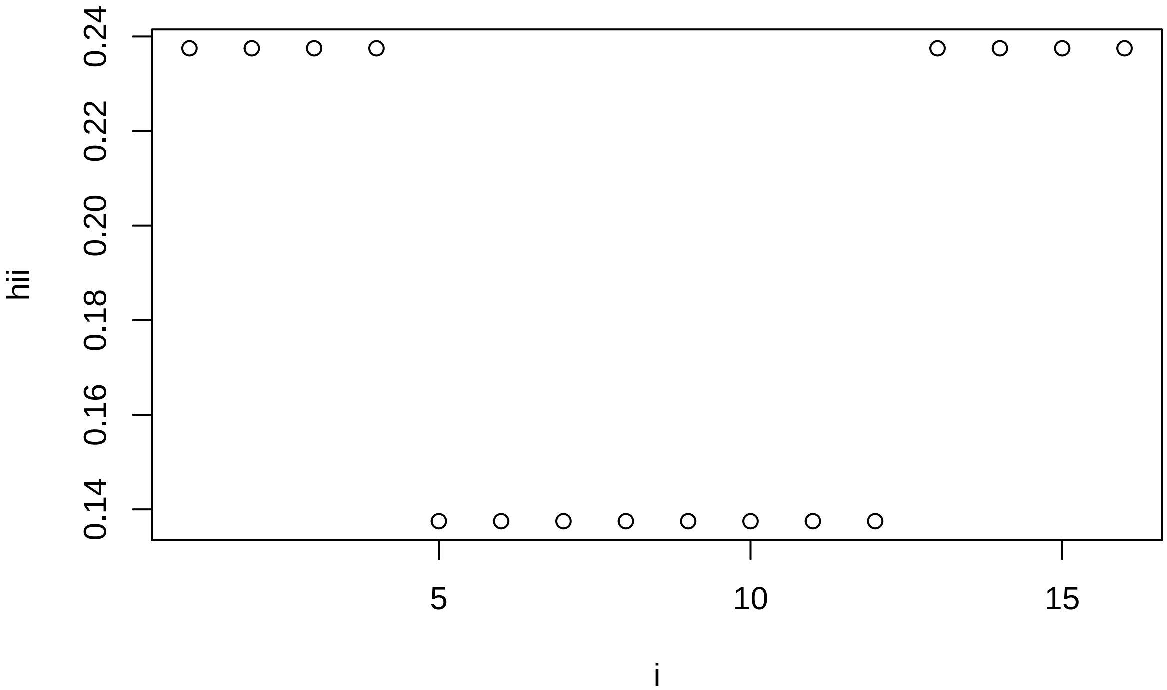
Here, $\max_i\{|t_i|\} = 2.102726 < t(0.996875; 13) = 3.256463$, therefore, conclude H_a, i.e. there is no outlier.

b. Obtain the diagonal elements of the hat matrix, and provide an explanation for the pattern in these elements.

```
print(h)
```

```
##  [1] 0.2375 0.2375 0.2375 0.2375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375
## [11] 0.1375 0.1375 0.2375 0.2375 0.2375 0.2375
```

```
plot(c(1:n),h,xlab = 'i',ylab='hii')
```



Half $h_{ii} = 0.2375$ and the others equal to 0.1375. It means that the data can be equally divided into 2 groups. In each of these group, the distances between the data and the regression surface are the same.

c. Are any of the observations outlying with regard to their X values according to the rule of thumb stated in the chapter?

```
p = n-df_
print(2*p/n)
```

```
## [1] 0.375
```

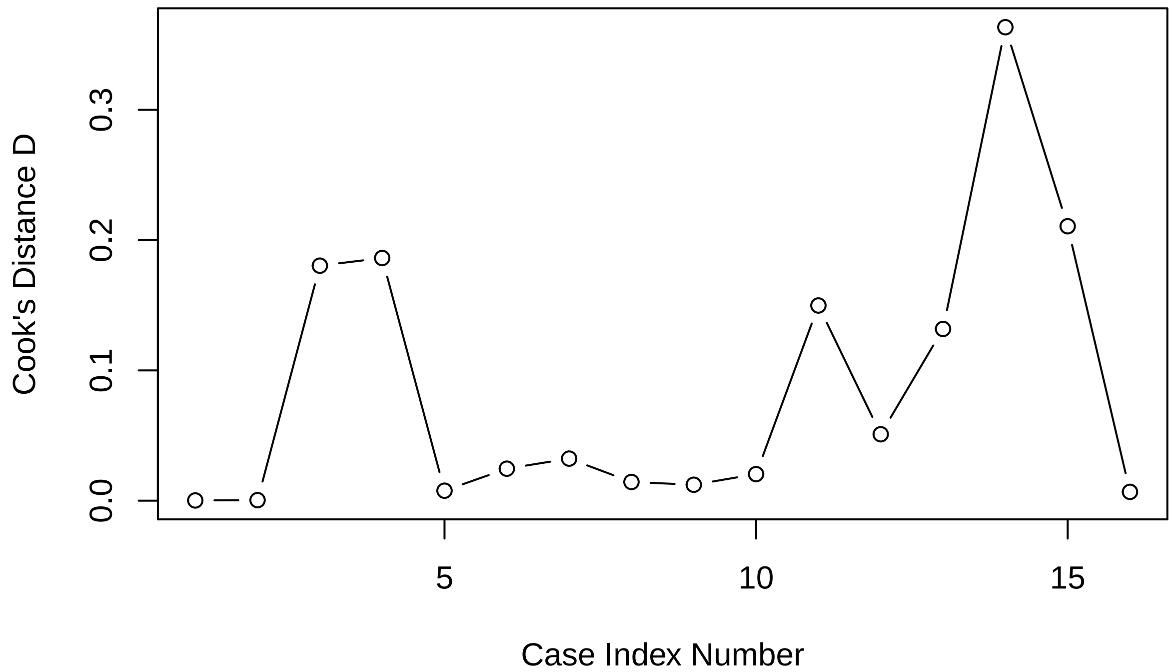
Since $\frac{2p}{n} = 0.375 > 0.2375 = \max_i\{h_{ii}\}$, there is no observation outlying with regard to their X values.

g. Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

```
MSE <- SSE/df_
D <- e^2 /p/MSE*h/((1-h)^2)
print(D)
```

```
##            1            2            3            4            5
## 0.0001877130 0.0004223542 0.1803921815 0.1862582123 0.0076655286
##            6            7            8            9           10
## 0.0245466787 0.0322971439 0.0143542862 0.0122308711 0.0204060192
##           11           12           13           14           15
## 0.1498281704 0.0509831969 0.1318214458 0.3634123447 0.2106609008
##           16
## 0.0067576676
```

```
plot(c(1:n),D,'b',xlab = 'Case Index Number',ylab='Cook\'s Distance D')
```



$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}$$
$$= \frac{e_i^2 h_{ii}}{pMSE(1-h_{ii})^2}$$

For a small data set, all D_i is less than 1. Therefore, there is no cases influential according to this measure.

10.13 Cosmetics sales. An assistant in the district sales office of a national cosmetics firm obtained data, shown below, on advertising expenditures and sales last year in the district's 44 territories. X_1 denotes expenditures for point-of-sale displays in beauty salons and department stores (in thousand dollars), and X_2 and X_3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Y denotes sales (in thousand cases). The assistant was instructed to estimate the increase in expected sales when X_1 is increased by 1 thousand dollars and X_2 and X_3 are held constant, and was told to use an ordinary multiple regression model with linear terms for the predictor variables and with independent normal error terms.

a. State the regression model to be employed, and fit it to the data.

```
data2 <- read.table("CH10PR13.txt",head=FALSE,col.names = c('Y',
'X1','X2','X3'))
Y <- data1$Y
X1 <- data2$X1
X2 <- data2$X2
X3 <- data2$X3
n = length(Y)
fit2 <- lm('Y~X1+X2+X3',data=data2)
summary(fit2)
```

```
##
## Call:
## lm(formula = "Y~X1+X2+X3", data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4217 -0.9115  0.0703  1.1420  3.5479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0233     1.2029   0.851   0.4000
## X1            0.9657     0.7092   1.362   0.1809
## X2            0.6292     0.7783   0.808   0.4237
## X3            0.6760     0.3557   1.900   0.0646 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.825 on 40 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7223
## F-statistic: 38.28 on 3 and 40 DF,  p-value: 7.821e-12
```

```r
print(fit2$fitted.values)
```

```
##         1         2         3         4         5         6         7
## 12.523331 11.247489  9.232069 12.005132  8.594978 12.861599 15.557943
##         8         9        10        11        12        13        14
##  9.971563  7.134733  7.893822 11.358116  7.858766  5.889261 11.179547
##        15        16        17        18        19        20        21
##  8.541178  2.974351  6.328273 10.691463  7.226361 10.817552  9.962447
##        22        23        24        25        26        27        28
##  8.217182  4.404481 12.971120  8.623391  9.705286 11.580732  9.295027
##        29        30        31        32        33        34        35
## 12.812187  8.381651  5.989922  5.777189 11.658598  3.133836  9.459454
##        36        37        38        39        40        41        42
## 10.213518 16.930693  7.134775  6.559452  6.221696  8.324271  9.802153
##        43        44
##  9.014906 13.198505
```

The regression model is

$$Y = 1.0233 + 0.9657X_1 + 0.6292X_2 + 0.6760X_3 + \epsilon$$

b. Test whether there is a regression relation between sales and the three predictor vruiables; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

```r
fit2.aov <- anova(fit2)
SSR <- sum(fit2.aov[1:2, 2])
SSE <- sum(fit2$residuals^2)
F <- (SSR)/(3-1)/(SSE/fit2$df.residual)
print(sprintf('F*=%f',F))
```

```
## [1] "F*=55.613439"
```

```r
print(sprintf('F(0.95;%d,%d)=%f',3-1,fit2$df.residual,df(0.95,3-1,fit2$df.residual)))
```

```
## [1] "F(0.95;2,40)=0.377368"
```

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \qquad H_a: \text{not all } \beta_k = 0 \ (k = 1, 2, 3)$$

The test statistic is

$$\begin{aligned} F^* &= \frac{\frac{SSR(X_1,X_2,X_3)}{p-1}}{\frac{SSE(X_1,X_2,X_3)}{n-p}} \\ &= \frac{MSR}{MSE} \end{aligned}$$

Given α, the decision rule is

If $F^* \leqslant F(1-\alpha; 2, 40)$, then conclude H_0;

If $F^* > F(1-\alpha; 2, 40)$, then conclude H_a;

Here, $F^* = 55.613439 > F(0.95; 2, 40) = 0.377368$, therefore, conclude H_a, i.e. not all $\beta_k = 0$ $(k = 1, 2, 3)$.

c. Test for each of the regression coefticients $\beta_k (k = 1, 2, 3)$ individually whether or not $\beta_k = 0$; use $\alpha = .05$ each time. Do the conclusions of these tests correspond to that obtained in part (b)?

```
SSE123 <- sum(fit2$residuals^2)
SSR1_23 <- sum(lm('Y~X2+X3',data=data2)$residuals^2)-SSE123
SSR2_13 <- sum(lm('Y~X1+X3',data=data2)$residuals^2)-SSE123
SSR3_12 <- sum(lm('Y~X1+X2',data=data2)$residuals^2)-SSE123
F1 <- SSR1_23/(SSE123/fit2$df.residual)
F2 <- SSR2_13/(SSE123/fit2$df.residual)
F3 <- SSR3_12/(SSE123/fit2$df.residual)
print(sprintf('When k=1, F*=%f',F1))
```

```
## [1] "When k=1, F*=1.854008"
```

```
print(sprintf('When k=2, F*=%f',F2))
```

```
## [1] "When k=2, F*=0.653481"
```

```
print(sprintf('When k=3, F*=%f',F3))
```

```
## [1] "When k=3, F*=3.611251"
```

```
print(sprintf('F(0.95;%d,%d)=%f',1,fit2$df.residual,df(0.95,1,fit2$df.residual)))
```

```
## [1] "F(0.95;1,40)=0.251396"
```

For $k = 1, 2, 3$,

$$H_0: \beta_k = 0 \qquad H_a: \beta_k \neq 0$$

The test statistic is

$$\begin{aligned} F^* &= \frac{\frac{SSR(X_k|X_j, \text{ for } j=1,2,3, \ j\neq k)}{1}}{\frac{SSE(X_1,X_2,X_3)}{n-p}} \\ &= \frac{MSR(X_k|X_j, \text{ for } j=1,2,3, \ j\neq k)}{MSE(X_1,X_2,X_3)} \end{aligned}$$

Given α, the decision rule is

If $F^* \leqslant F(1-\alpha;1,40)$, then conclude H_0;

If $F^* > F(1-\alpha;1,40)$, then conclude H_a;

Here, $F^* = \begin{cases} 1.854008 & ,k=1 \\ 0.653481 & ,k=2 \\ 3.611251 & ,k=3 \end{cases} > F(0.95;1,40) = 0.251396$, therefore, conclude H_a, i.e. $\beta_k \neq 0$ $(k=1,2,3)$.

d. Obtain the correlation matrix of the X variables.

```
cor(cbind(X1,X2,X3))
```

```
##           X1        X2        X3
## X1 1.0000000 0.9744313 0.3759509
## X2 0.9744313 1.0000000 0.4099208
## X3 0.3759509 0.4099208 1.0000000
```

e. What do the results in parts (b), (c), and (d) suggest about the suitability of the data for the research objective?

Under this model, the expect sales when X_1 is increased by 1 thousand dollars and X_2 and X_3 are held constant, is β_1. It can be estimated by b_1. Since X_1 and X_2 may be linear related from (d), when X_2 is fixed, X_1 is almost fixed, i.e. $b_1 \approx 0$. There is a contradiction and therefore, the data may not be suitable for the research objective.