Linear Regression Homework 7

Name: Eric Yuan

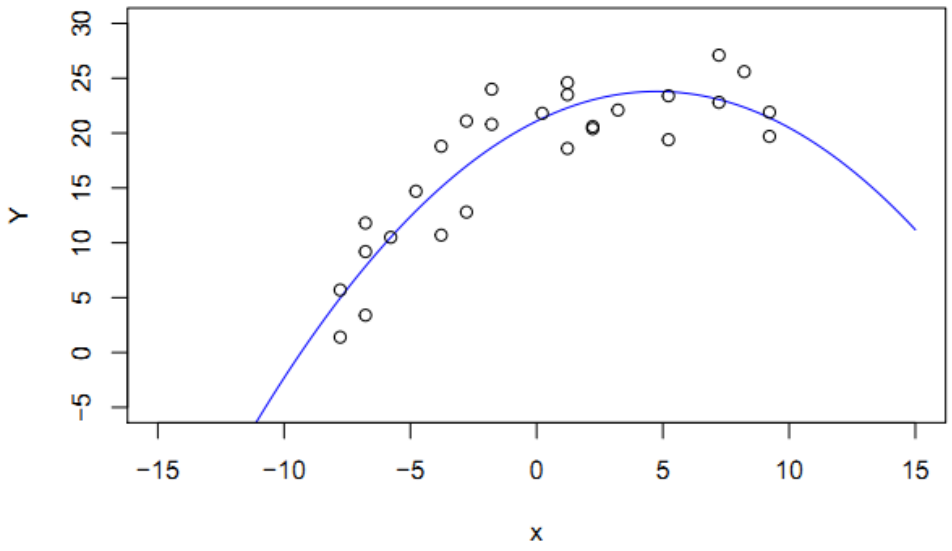
UNI: qy2205

1. Chapter 8 problem 6

(a)

```{r}
data = read.table('CH08PR06.txt')
colnames(data) <- c('y', 'x')
X <- data$x
Y <- data$y
x1 <- X - mean(X)
x11 <- x1**2
lm1 <- lm(Y~x1+x11)
lm1
```

```
Call:
lm(formula = Y ~ x1 + x11)

Coefficients:
(Intercept)           x1          x11
    21.0942       1.1374      -0.1184
```




```{r}
SSR <- sum((lm1$fitted.values-mean(Y))^2)
SST <- sum((Y-mean(Y))^2)
R.square <- SSR/SST
R.square
```

```
[1] 0.8143372
```

Based on the result, we could see quadratic function seems a good fit.

(b)

$Y = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2$

$H_0: \beta_1 = \beta_{11} = 0$

H_1: either of $\beta_1, \beta_{11} \neq 0$

$$F^* = \frac{SSR(x_1, {x_1}^2)}{p} \div \frac{SSE(x_1, {x_1}^2)}{n - p - 1}$$

$$= \frac{1046.266}{2} \div \frac{238.5408}{24}$$

$= 52.63$

Since $F^* = 52.633 > F(0.99; 2, 24) = 5.614$

So we reject H_0 which means there is a regression relation

```
p.value <- 1-pf(52.633,2,24)
p.value
```

```
[1] 1.677734e-09
```

(c)

```
```{r}
y<-data[,1]
x<-data[,2]-mean(data[,2])

xx<-x^2
g <- 3
B <- qt(1-0.01/(2*g),24)
x_mat <- t(rbind(1,x,xx))
mse=sum((lm1$residuals)^2)/24
m=mean(data[,2])

x10<-matrix(c(1,10-m,(10-m)^2),nrow=3)
y10 <- coef(lm1)%*%x10
sd10 <- sqrt(mse*t(x10)%*%solve(t(x_mat)%*%x_mat)%*%x10)
cat('CI for X=10: ', c(y10-B*sd10,y10+B*sd10))

x15<-matrix(c(1,15-m,(15-m)^2),nrow=3)
y15 <- coef(lm1)%*%x15
sd15 <- sqrt(mse*t(x15)%*%solve(t(x_mat)%*%x_mat)%*%x15)
cat('CI for X=10: ', c(y15-B*sd15,y15+B*sd15))

x20<-matrix(c(1,20-m,(20-m)^2),nrow=3)
y20 <- coef(lm1)%*%x20
sd20 <- sqrt(mse*t(x20)%*%solve(t(x_mat)%*%x_mat)%*%x20)
cat('CI for X=10: ', c(y20-B*sd20,y20+B*sd20))
```
```

99% confidence interval for x=10: [7.559977,13.580437]

99% confidence interval for x=15: [17.22897,23.04688]

99% confidence interval for x=20: [20.99140,26.57975]

(d)

```r
t <- qt(1-0.01/2,24)
sd_pred <- sqrt(mse*(1+t(x15)%*%solve(t(x_mat)%*%x_mat)%*%x15))
c(y15-t*sd_pred,y15+sd_pred)
```

```
[1] 10.97342 23.41454
```

99% prediction interval for x=15 is [10.97342,23.41454]

(e)

```r
anova(lm(y~x))
F=(491.53-238.54)/mse
F>qf(0.99,1,24)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 793.28  793.28  40.348 1.196e-06 ***
Residuals 25 491.53   19.66
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] TRUE
```

$Y = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2$

$H_0: \beta_{11} = 0$

$H_1: \beta_{11} \neq 0$

$$F^* = \frac{SSR(x_1{}^2|x_1)}{1} \div \frac{SSE(x_1, x_1{}^2)}{n-p-1}$$

$$= \frac{252.9853}{1} \div \frac{238.5408}{24}$$

$= 25.45329$

Since $F^* = 25.45 > F(0.99; 1, 24) = 7.82$

So we reject H_0

(f)

```r
b0 <- lm1$coefficients[1]
b1 <- lm1$coefficients[2]
b11 <- lm1$coefficients[3]
c(b0-b1*m+b11*m^2, b1-2*m*b11,b11)
```

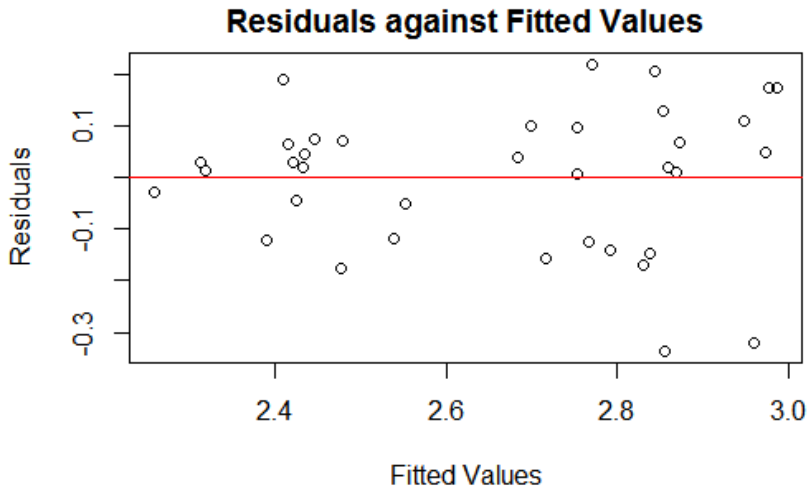
```
(Intercept)          x1         x11
-26.3254125   4.8735744  -0.1184012
```

2. Chapter 8 problem 42

(a)

```
data <- read.table("APPENC03.txt", stringsAsFactors=F)
names(data) <- c("id","y","f1","f2","f3","f4","Month","year")
data$i1999 <- 1*(data$year == 1999)
data$i2001 <- 1*(data$year == 2001)
data$i2002 <- 1*(data$year == 2002)

lm2 = lm(y~f1+f2+f3+f4+i1999+i2001+i2002,data=data)
lm2_sum<-summary(fit)
lm2_sum

plot(fitted(lm2), lm2$residuals, xlab="Fitted Values",
     ylab = "Residuals",
     main = "Residuals against Fitted Values")
abline(h=0,col="red")
```

```
Call:
lm(formula = y ~ f1 + f2 + f3 + f4 + i1999 + i2001 + i2002, data
= data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33558 -0.11872  0.02459  0.08020  0.21952

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.021e+00  4.705e-01   6.421 5.94e-07 ***
f1          -2.470e-01  1.982e-01  -1.246   0.2229
f2          -9.653e-05  1.914e-04  -0.504   0.6181
f3           4.093e-01  5.385e-02   7.601 2.80e-08 ***
f4           1.240e-01  5.484e-02   2.261   0.0317 *
i1999        1.324e-02  9.304e-02   0.142   0.8879
i2001       -1.088e-01  7.133e-02  -1.525   0.1385
i2002       -8.306e-02  8.657e-02  -0.959   0.3456
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1529 on 28 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6657
F-statistic: 10.96 on 7 and 28 DF,  p-value: 1.382e-06
```



The first-order model is good since the residual seems follow the normal distribution.

(b)

```
relm2 <- lm(y~f1+f2+I(f1^2)+I(f2^2)+f1*f2+f3+f4+i1999+i2001+i2002, data=data)
summary(relm2)
```

```
Call:
lm(formula = y ~ f1 + f2 + I(f1^2) + I(f2^2) + f1 * f2 + f3 +
    f4 + i1999 + i2001 + i2002, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33455 -0.08692  0.01892  0.07039  0.23931

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.698e+00  6.818e+00   1.276   0.2138
f1          -4.803e+00  5.380e+00  -0.893   0.3805
f2          -9.508e-04  3.492e-03  -0.272   0.7877
I(f1^2)      9.221e-01  1.069e+00   0.863   0.3965
I(f2^2)      5.518e-07  7.375e-07   0.748   0.4613
f3           3.941e-01  6.098e-02   6.463 9.09e-07 ***
f4           1.149e-01  5.772e-02   1.991   0.0575 .
i1999        1.236e-02  1.006e-01   0.123   0.9031
i2001       -1.006e-01  7.476e-02  -1.345   0.1906
i2002       -5.807e-02  9.541e-02  -0.609   0.5483
f1:f2        1.629e-04  1.393e-03   0.117   0.9078
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1583 on 25 degrees of freedom
Multiple R-squared:  0.744,     Adjusted R-squared:  0.6417
F-statistic: 7.267 on 10 and 25 DF,  p-value: 2.837e-05
```

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{11} {X_1}^2 + \beta_{22} {X_2}^2 + \beta_{12} X_1 X_2 + \beta_{5,99} X_{5,99} + \beta_{5,01} X_{5,01} + \beta_{5,02} X_{5,02} b + \varepsilon$$

$H_0: \beta_{11} = \beta_{22} = \beta_{12} = 0$

$H_1: at\ least\ one\ of\ them\ not\ equal\ to\ 0$

$$F^* = \frac{SSR(x_{11}, x_{22}, x_{12} | x_1, x_2, X_3, X_4, X_{5,99}, X_{5,01}, X_{5,02})}{1} \div \frac{SSE}{n - p - 1}$$

$$= \frac{0.0281}{3} \div \frac{0.6261}{36 - 11}$$

$$= 0.374$$

Since $F^* = 0.374 < F(0.95; 3, 25) = 2.991$

So we reject H_0 and conclude H_1

(c)

```
relm21 <- lm(y~f1+f3+f4, data=data)
summary(relm21)
```

```
Call:
lm(formula = y ~ f1 + f3 + f4, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.286376 -0.100465 -0.002259  0.104174  0.240020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.18527    0.36505   8.726  5.7e-10 ***
f1          -0.35269    0.15738  -2.241   0.0321 *
f3           0.39914    0.05125   7.787  7.0e-09 ***
f4           0.11803    0.05149   2.292   0.0286 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1498 on 32 degrees of freedom
Multiple R-squared:  0.7065,    Adjusted R-squared:  0.679
F-statistic: 25.68 on 3 and 32 DF,  p-value: 1.191e-08
```

$H_0: \beta_2 = \beta_{5,99} = \beta_{5,01} = \beta_{5,02} = 0$

$H_1: at\ least\ one\ of\ them\ not\ equal\ to\ 0$

$$F^* = \frac{SSR(x_2, X_{5,99}, X_{5,01}, X_{5,02} | X_1, X_3, X_4)}{4} \div \frac{SSE}{n-p-1}$$

$$= \frac{0.0637}{4} \div \frac{0.6542}{28}$$

$$= 0.6817$$

Since $F^* = 0.6817 < F(0.95; 4, 28) = 2.7141$

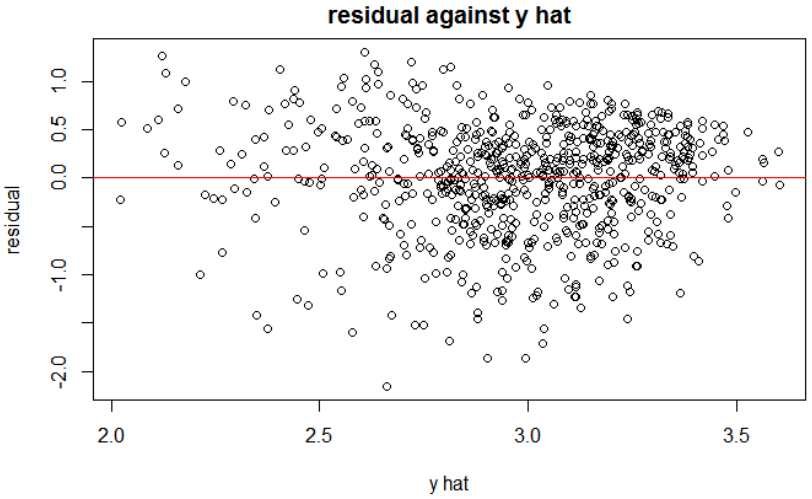
So we conclude H_0

3. Chapter 8 problem 43

Fit the first order regression model first.

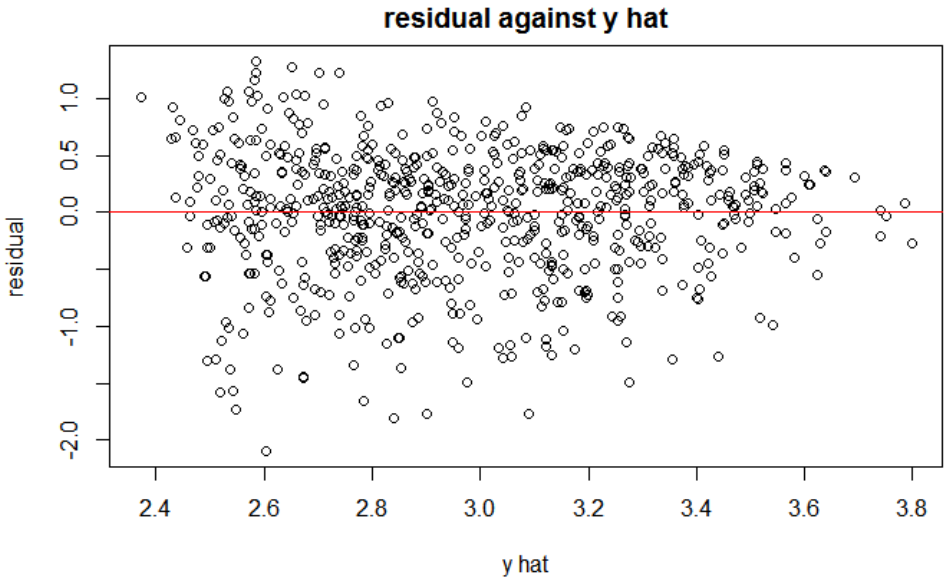
```
data<-read.table("APPENC04.txt", stringsAsFactors=F)
names(data) <- c("id","y","x1","x2","year")
data$i1996 <- 1*(data$year == 1996)
data$i1997 <- 1*(data$year == 1997)
data$i1998 <- 1*(data$year == 1998)
data$i1999 <- 1*(data$year == 1999)

fit=lm(y~x1+x2+i1996+i1997+i1998+i1999,data=data)
summary(fit)
```

```
Call:
lm(formula = y ~ x1 + x2 + i1996 + i1997 + i1998 + i1999, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15048 -0.28873  0.07655  0.39619  1.30415

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.273880   0.143571   8.873  < 2e-16 ***
x1           0.010124   0.001285   7.878 1.28e-14 ***
x2           0.037188   0.005951   6.248 7.21e-10 ***
i1996       -0.056007   0.068013  -0.823    0.411
i1997        0.027651   0.069417   0.398    0.691
i1998        0.059333   0.066657   0.890    0.374
i1999        0.024065   0.068197   0.353    0.724
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5674 on 698 degrees of freedom
Multiple R-squared:  0.2071,    Adjusted R-squared:  0.2003
F-statistic: 30.39 on 6 and 698 DF,  p-value: < 2.2e-16
```



It seems the first order model is not good. Let’s try second order model.

```
Call:
lm(formula = y ~ x1 + x2 + I(x1^2) + I(x2^2) + x1 * x2 + i1996 +
    i1997 + i1998 + i1999, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09218 -0.31352  0.07167  0.37913  1.32764

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.823e+00  6.396e-01   4.414 1.17e-05 ***
x1          -2.224e-02  7.482e-03  -2.973  0.00305 **
x2          -4.476e-03  5.100e-02  -0.088  0.93009
I(x1^2)      1.476e-04  5.625e-05   2.623  0.00890 **
I(x2^2)     -7.625e-05  1.148e-03  -0.066  0.94707
i1996       -4.295e-02  6.723e-02  -0.639  0.52309
i1997        1.698e-02  6.849e-02   0.248  0.80423
i1998        5.689e-02  6.595e-02   0.863  0.38863
i1999        2.238e-02  6.763e-02   0.331  0.74087
x1:x2        5.652e-04  3.561e-04   1.587  0.11288
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5595 on 695 degrees of freedom
Multiple R-squared:  0.2324,    Adjusted R-squared:  0.2224
F-statistic: 23.38 on 9 and 695 DF,  p-value: < 2.2e-16
```



Based on the summary result, we can see that X_2^2, X_1X_2 *and* X_3 could be dropped from the model.

H_0: $\beta_{22} = \beta_{12} = \beta_{3,97} = \beta_{3,98} = \beta_{3,99} = \beta_{3,00} = 0$

H_1: *at least one of them not equal to* 0

$$F^* = \frac{1.719789}{6} \div \frac{217.5797}{704 - 10}$$

$$= 0.91425$$

Since $F^* = 0.91425 < F(0.99; 6, 694) = 2.828$

So we conclude H_0

```
refit2=lm(y~x1+x2+I(x1^2), data = data)
summary(refit2)
```

```
Call:
lm(formula = y ~ x1 + x2 + I(x1^2), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04480 -0.29380  0.08396  0.38842  1.31625

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.145e+00  2.304e-01   9.311  < 2e-16 ***
x1          -1.735e-02  6.130e-03  -2.830  0.00478 **
x2           3.530e-02  5.872e-03   6.012 2.96e-09 ***
I(x1^2)      2.075e-04  4.547e-05   4.563 5.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5593 on 701 degrees of freedom
Multiple R-squared:  0.2263,    Adjusted R-squared:  0.223
F-statistic: 68.35 on 3 and 701 DF,  p-value: < 2.2e-16
```

So the final model will be:

$$Y = 2.15 - 0.0173X_1 + 0.0353X_2 + 0.0002X_1^2$$

4. Chapter 10, problem 9 abcdg

(a)

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

$$t_i = e_i\left[\frac{n-p-1}{SSE(1-h_{ii})-{e_i}^2}\right]^{1/2}$$

Decision rule:

If $|t_i| > t^*(0.9969; 12) = 3.307783$, the i^{th} data is an outlier

```
data <- read.table('CH06PR05.txt')
colnames(data) <- c('y', 'x1', 'x2')
Y = data$y
X1 = data$x1
X2 = data$x2
L <- lm(Y~X1+X2)
e <- L$residuals
SSE <- sum((e)^2)
n <- length(Y)
X.Mat <- t(rbind(1,X1,X2))
H <- X.Mat%*%solve(t(X.Mat)%*%X.Mat)%*%t(X.Mat)
H.diag <- diag(H)
t <- e*sqrt((n-3-1)/(SSE*(1-H.diag)-e^2))
df <- data.frame(i=1:n,e=e,t=t)
t.star <- qt(1-0.1/(2*n),n-3-1)
outlier.place <- c()
for(i in 1:n)
{
if(abs(t[i])>t.star)
{outlier.place <- c(outlier.place,i)}
}
df
```

	i <int>	e <dbl>	t <dbl>
1	1	-0.10	-0.04085498
2	2	0.15	0.06128781
3	3	-3.10	-1.36059879
4	4	3.15	1.38602483
5	5	-0.95	-0.36694571
6	6	-1.70	-0.66490618
7	7	-1.95	-0.76716157
8	8	1.30	0.50461264
9	9	1.20	0.46506694
10	10	-1.55	-0.60436295

No outlying y observations.

(b)

$$H = X(X^T X)^{-1} X^T$$

```
X.Mat <- t(rbind(1,X1,X2))
H <- X.Mat%*%solve(t(X.Mat)%*%X.Mat)%*%t(X.Mat)
H.diag <- diag(H)
H.diag
```

```
 [1] 0.2375 0.2375 0.2375 0.2375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375
[12] 0.1375 0.2375 0.2375 0.2375 0.2375
```

(c)

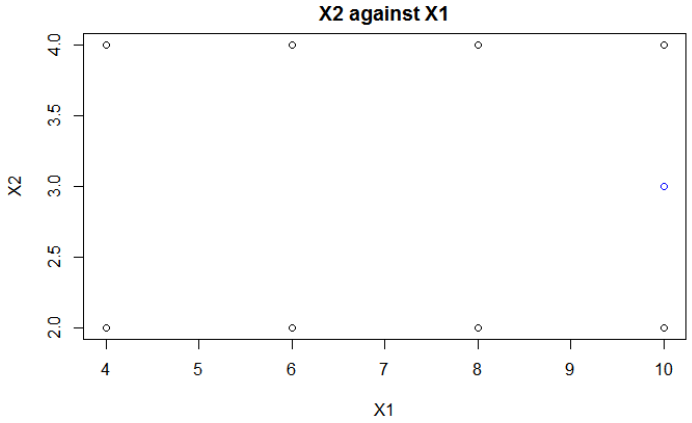
If $H_{ii} > \frac{2p}{n} = 0.375$ which means ith data is an outlier. Write R code to test that.

```
Critical.Point <- 2*3/n
Outlier.X <- c()
for(i in 1:n){ if(H.diag[i]>Critical.Point) {Outlier.X <- c(Outlier.X,i)} }
Outlier.X
```

```
NULL
```

No outliers

(d)



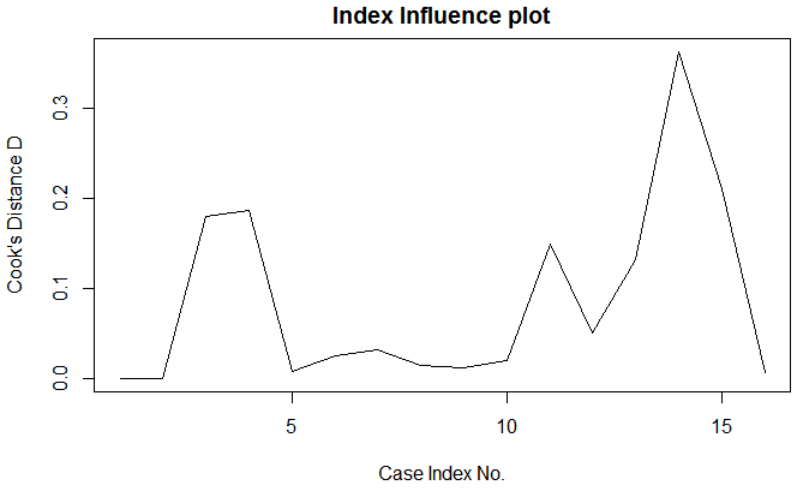
$H_{ii} = 0.175 < \frac{2p}{n} = 0.375, (X_1, X_2) = (10, 3)$ is not an outlier

(g)

$$D_i = \frac{e_i^2}{pMSE} \frac{h_{ii}}{(1 - h_{ii})^2}$$

```
MSE <- SSE/(n-3)
D <- e^2/(3*MSE)*H.diag/(1-H.diag)^2
df <- data.frame(i=1:n,D=D,Percentile=pf(D,3,n-3))
df
```

	i <int>	D <dbl>	Percentile <dbl>
1	1	0.0001877130	3.753988e-06
2	2	0.0004223542	1.266642e-05
3	3	0.1803921815	9.220276e-02
4	4	0.1862582123	9.615114e-02
5	5	0.0076655286	9.715666e-04
6	6	0.0245466787	5.464829e-03
7	7	0.0322971439	8.177954e-03
8	8	0.0143542862	2.471311e-03
9	9	0.0122308711	1.948307e-03
10	10	0.0204060192	4.161023e-03




```
df[df$Percentile>0.1,1]
```

```
[1] 14 15
```

Case 14 and 15 is influential.

5.

$\hat{\beta} = \widehat{\beta_{(i)}}$

$$D_i = \frac{\left(\hat{\beta} - \widehat{\beta_{(i)}}\right)^T X^T X \left(\hat{\beta} - \widehat{\beta_{(i)}}\right)}{pMSE} = 0$$

$$D_i = \frac{e_i^2}{pMSE} \frac{h_{ii}}{(1-h_{ii})^2} = 0$$

$e_i = 0$

This is because $h_{ii} \geq \frac{1}{n} > 0$

$\mathrm{Y_i} = \widehat{\mathrm{Y}}_\mathrm{i}$