Homework Seven

Yi Chen(yc3356)

November 19, 2017

Homework seven

8.6

problem:a

Model: $Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$

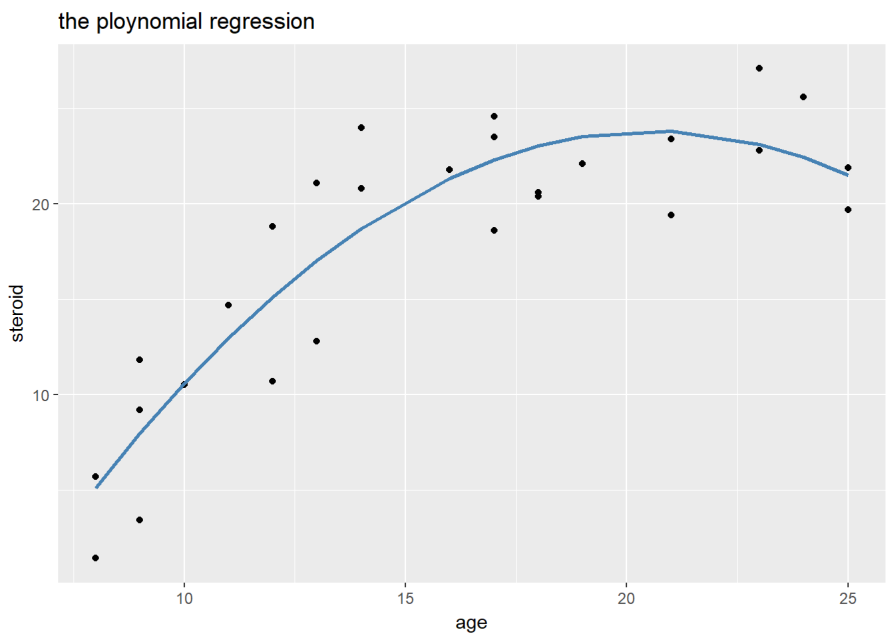
```
# solution one (use the R funtion)
setwd("C:/Users/cheny/Desktop/study/linear regression model/homework/homework record/Homework seven")
data_8.6 <- read.table('8.6.txt',header = FALSE,col.names = c('Y','X'))

data_8.6$X_centered <- scale(data_8.6$X,center = TRUE,scale = FALSE)
data_8.6$X_2 <- (data_8.6$X_centered)^2

reg_8.6 <- lm(data = data_8.6, Y~X_centered+X_2)
summary(reg_8.6)
```

```
##
## Call:
## lm(formula = Y ~ X_centered + X_2, data = data_8.6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5463 -2.5369  0.3868  2.1973  5.3020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.09416    0.91415  23.075  < 2e-16 ***
## X_centered   1.13736    0.11546   9.851 6.59e-10 ***
## X_2         -0.11840    0.02347  -5.045 3.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.153 on 24 degrees of freedom
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.7989
## F-statistic: 52.63 on 2 and 24 DF,  p-value: 1.678e-09
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
ggplot(data = data_8.6)+
        geom_point(mapping = aes(x=X,y=Y))+
        geom_line(mapping = aes(x=X,y=fitted(reg_8.6)),col='steelblue',lwd=1)+
        labs(title='the ploynomial regression',x='age',y='steroid')
```



```
# solution two (show every detail in the process)

## x matrix
x <- matrix(ncol = 3,nrow = nrow(data_8.6))
x[,1] <- 1
x[,2] <- data_8.6[,3]
x[,3] <- data_8.6[,4]
head(x)
```

```
##      [,1]      [,2]     [,3]
## [1,]    1  7.222222 52.16049
## [2,]    1  3.222222 10.38272
## [3,]    1  9.222222 85.04938
## [4,]    1 -3.777778 14.27160
## [5,]    1 -7.777778 60.49383
## [6,]    1 -3.777778 14.27160
```

```
## y matrix
y <- as.matrix(data_8.6$Y)
head(y)
```

```
##      [,1]
## [1,] 27.1
## [2,] 22.1
## [3,] 21.9
## [4,] 10.7
## [5,]  1.4
## [6,] 18.8
```

```
## estimate the parameters
b <- solve(t(x)%*%x) %*% t(x) %*% y
b
```

```
##            [,1]
## [1,] 21.0941598
## [2,]  1.1373573
## [3,] -0.1184012
```

analysis

The regression result:

$\widehat{Y} = 21.09416 + 1.13736x - 0.11840x^2, x = X - \bar{X}$

As we can see in the plot, the line fit the data very well. And according to $R^2 = 0.8143$, the regression seems to be a good fit of the data.

problem b

$H0: \beta_1 = \beta_{11} = 0$ and $Ha:*10 or* {11}0 $

```
anova(reg_8.6)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X_centered  1 793.28  793.28  79.813 4.236e-09 ***
## X_2         1 252.99  252.99  25.453 3.708e-05 ***
## Residuals  24 238.54    9.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

as we can see $MSR = \frac{SSR(x)+SSR(x^2)}{df} = \frac{793.28+252.99}{2} = 523.135$,while $MSE = 9.94$

Thus: $F^* = \frac{MSR}{MSE} = \frac{523.135}{9.94} = 52.63$

```
qf(0.99,2,24)
```

```
## [1] 5.613591
```

Clearly,$F(0.99, 2, 24) = 5.613591$, thus$F^* \leq F(0.99, 2, 24)$.Conclude Ha.

problem c

```
x_h <- matrix(c(1,10,10^2,1,15,15^2,1,20,20^2),ncol=3)
x_h <- t(x_h)
```

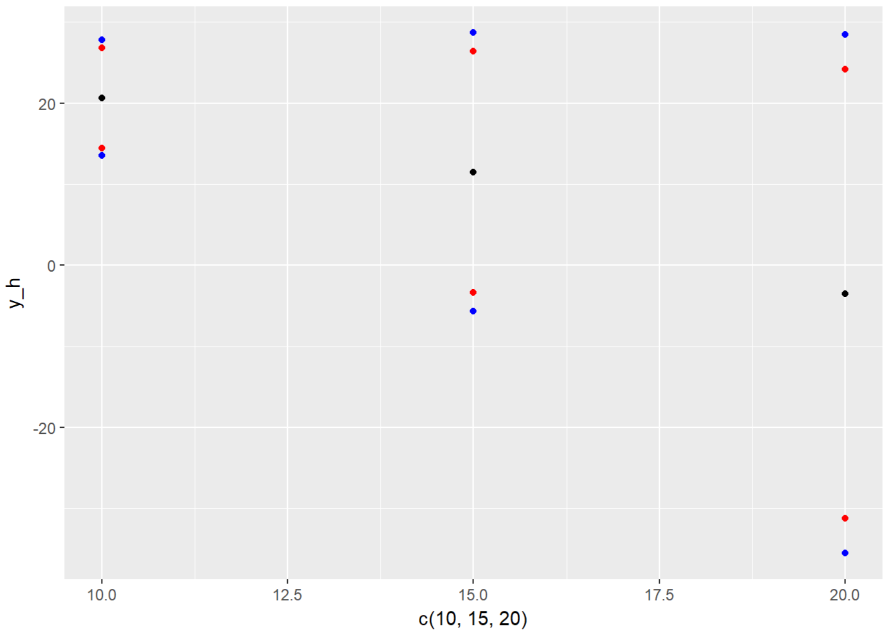
```
# here g=3,n=27,p=3
W <- sqrt(3*qf(0.99,3,24))

B <- qt(1-0.01/(2*3),24)


y_h <- x_h %*% b

MSE <- anova(reg_8.6)$`Mean Sq`[3]

s_2_b <- MSE * solve(t(x)%*%x)

s_2_y_h <- x_h %*% s_2_b %*% t(x_h)

s_2_y_h <- matrix(c(sqrt(s_2_y_h[1,1]),sqrt(s_2_y_h[2,2]),sqrt(s_2_y_h[3,3])),ncol = 1)

# solution one (Bnoferroni)
CI1 <- cbind(y_h-B*s_2_y_h,y_h+B*s_2_y_h)
CI1
```

```
##             [,1]     [,2]
## [1,]  14.454595 26.80062
## [2,]  -3.366579 26.39506
## [3,] -31.218189 24.17980
```

```
# solution two (Working-Hotelling)
CI2 <- cbind(y_h-W*s_2_y_h,y_h+W*s_2_y_h)
CI2
```

```
##             [,1]     [,2]
## [1,]  13.500109 27.75511
## [2,]  -5.667487 28.69596
## [3,] -35.501074 28.46269
```

```
library(ggplot2)
ggplot()+
        geom_point(aes(x=c(10,15,20),y=y_h))+
        geom_point(aes(x=c(10,15,20),y=CI1[,1]),col='red') +
        geom_point(aes(x=c(10,15,20),y=CI1[,2]),col='red') +
        geom_point(aes(x=c(10,15,20),y=CI2[,1]),col='blue') +
        geom_point(aes(x=c(10,15,20),y=CI2[,2]),col='blue')
```



problem d

```
# calculate the point estimate of y
Y_new <- reg_8.6$coefficients[1] + reg_8.6$coefficients[2]*15 + reg_8.6$coefficients[3]*15^2
Y_new
```

```
## (Intercept)
##    11.51424
```

```
t <- qt(0.995,24)
# MSE=9.94
#calculate the standard deviation of y_new
S_2_pred_15 <- sqrt(s_2_y_h[2,1]^2 + MSE)


CI3 <- c((Y_new + t*S_2_pred_15),(Y_new - t*S_2_pred_15))
CI3
```

```
## (Intercept) (Intercept)
##   27.035640   -4.007162
```

problem e:

T-TEST

```
summary(reg_8.6)
```

```
##
## Call:
## lm(formula = Y ~ X_centered + X_2, data = data_8.6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5463 -2.5369  0.3868  2.1973  5.3020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.09416    0.91415  23.075  < 2e-16 ***
## X_centered   1.13736    0.11546   9.851 6.59e-10 ***
## X_2         -0.11840    0.02347  -5.045 3.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.153 on 24 degrees of freedom
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.7989
## F-statistic: 52.63 on 2 and 24 DF,  p-value: 1.678e-09
```

As we can see, the t-test for the x_2.

$H0 = \beta_{11} = 0, Ha = \beta_{11} \neq 0$

$s(b_{11}) = 0.02347$ and $t^* = \frac{b_{11}}{s(b_{11})} = \frac{-0.11840}{0.02347} = -5.045$

while

```
qt(0.995,24)
```

```
## [1] 2.79694
```

Thus$|t^*| \geq t(0.995, 24)$. Conclude Ha. Or we can easily see the result in the summary that the p-value of the quadratic term is less that 0.01 which indicate that $\beta_{11} \neq 0$

partial F-TEST

```
anova(reg_8.6)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X_centered  1 793.28  793.28  79.813 4.236e-09 ***
## X_2         1 252.99  252.99  25.453 3.708e-05 ***
## Residuals  24 238.54    9.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see: $SSR(x^2|x) = 252.99$ and $SSE(x, x^2) = 238.54$, $F^* = \frac{\frac{252.99}{1}}{\frac{238.54}{24}} = 25.453$

```
qf(0.99,1,24)
```

```
## [1] 7.822871
```

Clearly$F^* \leq F(0.99, 1, 24)$, conclude H0

problem f:

the model we used for regression: $\widehat{Y} = b_0 + b_1 x_i + b_{11} x_i^2, x = X - \bar{X}$

the orginal model: $\widehat{Y} = b_0^{'} + b_1^{'} x_i + b_{11}^{'} x_i^2$

Thus: $b_0^{'} = b_0 - b_1 \bar{X} + b_{11} \bar{X}^2$ and $b_1^{'} = b_1 - 2b_{11} \bar{X}$ and $b_{11}^{'} = b_{11}$

```
x_bar <- mean(data_8.6$X)

orginal_bo <- reg_8.6$coefficients[1] -reg_8.6$coefficients[2]*x_bar+reg_8.6$coefficients[3]*x_bar^2

orginal_b1 <- reg_8.6$coefficients[2]-2*reg_8.6$coefficients[3]*x_bar

orginal_b11 <- reg_8.6$coefficients[3]

orginal_bo;orginal_b1;orginal_b11
```

```
## (Intercept)
##   -26.32541
```

```
## X_centered
##   4.873574
```

```
##        X_2
## -0.1184012
```

clearly, the original model is : $\widehat{Y} = -26.32541 + 4.873574X - 0.1184012X^2$

8.42

problem a.

```
data_8.42 <- read.table('8.42.txt',header = FALSE,col.names=c('index','Y','X1','X2','X3','X4','month','X5'))
data_8.42 <- data_8.42[,c(-1,-7)]
as.factor(data_8.42$X3)
```

```
##  [1] 1 0 1 0 1 0 1 1 0 0 1 0 1 0 0 1 0 1 1 1 0 0 1 1 1 0 0 1 1 1 1 0 1 1 0
## [36] 1
## Levels: 0 1
```

```
as.factor(data_8.42$X4)
```

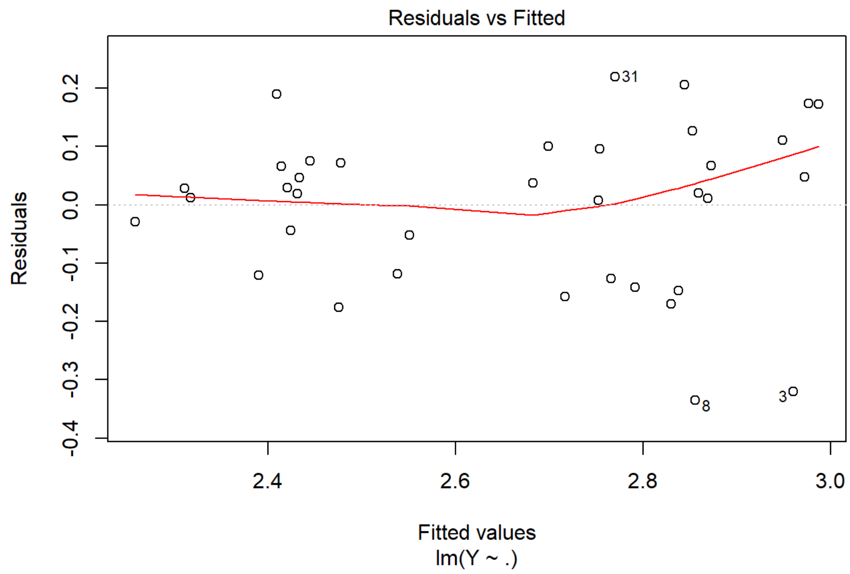
```
##  [1] 1 0 1 1 0 0 1 0 0 1 1 0 0 1 0 1 0 1 1 0 1 0 0 1 1 1 0 1 0 1 0 1 1 0 1
## [36] 1
## Levels: 0 1
```

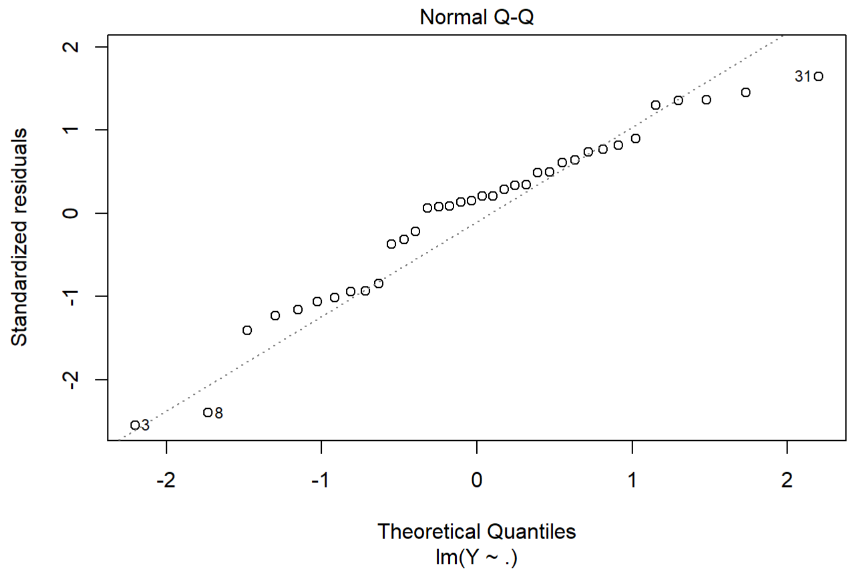
```
# use 2000 as reference year
data_8.42$x5_1 <- as.numeric(data_8.42$X5==1999)
data_8.42$x5_2 <- as.numeric(data_8.42$X5==2001)
data_8.42$x5_3 <- as.numeric(data_8.42$X5==2002)
data_8.42 <- data_8.42[,c(-6)]

reg_8.42 <- lm(data=data_8.42,Y~.)
summary(reg_8.42)
```

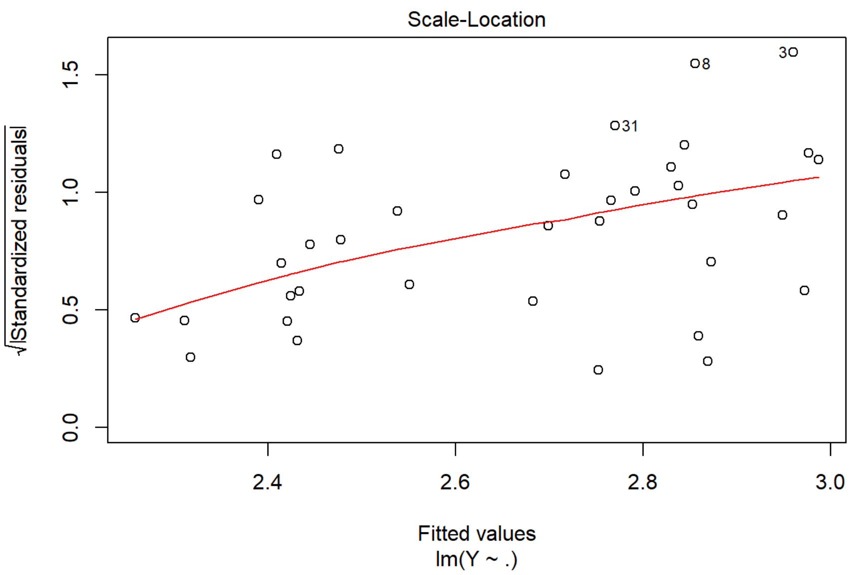
```
##
## Call:
## lm(formula = Y ~ ., data = data_8.42)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33558 -0.11872  0.02459  0.08020  0.21952
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.021e+00  4.705e-01   6.421 5.94e-07 ***
## X1          -2.470e-01  1.982e-01  -1.246   0.2229
## X2          -9.653e-05  1.914e-04  -0.504   0.6181
## X3           4.093e-01  5.385e-02   7.601 2.80e-08 ***
## X4           1.240e-01  5.484e-02   2.261   0.0317 *
## x5_1         1.324e-02  9.304e-02   0.142   0.8879
## x5_2        -1.088e-01  7.133e-02  -1.525   0.1385
## x5_3        -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
```

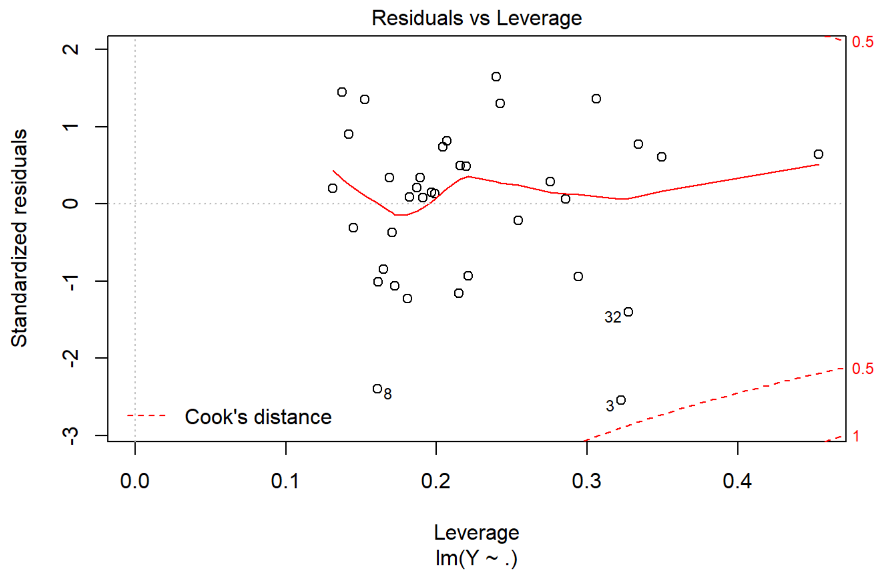
** analysis** the regression result:
$\widehat{Y} = 3.0211 - 0.247X_1 - 0.000097X_2 + 0.4093X_3 + 0.124X_4 - 0.1324X_{5(1)}(1999) - 0.1088X_{5(1)}(2001) - 0.8306X_{5(3)}(2002)$

```
plot(reg_8.42)
```


Residuals vs Fitted
Residuals
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
31
8
3
2.4
2.6
2.8
3.0
Fitted values
lm(Y ~ .)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
31
8
3
-2
-1
0
1
2
Theoretical Quantiles
lm(Y ~ .)


Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
8
3
31
2.4
2.6
2.8
3.0
Fitted values
lm(Y ~ .)



problem b.

```
data_8.42$x1_2 <- scale(data_8.42$X1^2,center = TRUE,scale = FALSE)
data_8.42$x2_2 <- scale(data_8.42$X2^2,center = TRUE,scale = FALSE)
data_8.42$X1 <- scale(data_8.42$X1,center = TRUE,scale = FALSE)
data_8.42$X2 <- scale(data_8.42$X2,center = TRUE,scale = FALSE)

reg_8.42_2 <- lm(data=data_8.42,Y~X1+X2+X3+X4+x5_1+x5_2+x5_3+x1_2+x2_2+X1:X2)
summary(reg_8.42_2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + x5_1 + x5_2 + x5_3 + x1_2 +
##     x2_2 + X1:X2, data = data_8.42)
##
## 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)  2.417e+00  6.368e-02  37.954  < 2e-16 ***
## X1          -4.739e+00  5.161e+00  -0.918   0.3672
## X2          -5.721e-04  6.494e-04  -0.881   0.3867
## X3           3.941e-01  6.098e-02   6.463 9.09e-07 ***
## X4           1.149e-01  5.772e-02   1.991   0.0575 .
## x5_1         1.236e-02  1.006e-01   0.123   0.9031
## x5_2        -1.006e-01  7.476e-02  -1.345   0.1906
## x5_3        -5.807e-02  9.541e-02  -0.609   0.5483
## x1_2         9.221e-01  1.069e+00   0.863   0.3965
## x2_2         5.518e-07  7.375e-07   0.748   0.4613
## X1:X2        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
```

the regression result:
$\widehat{Y} = 2.417 - 0.453x_1 - 0.000144x_2 + 0.394X_3 + 0.115X_4 - 0.012X_{5(1)}(1999) - 0.101X_{5(1)}(2001) - 0.0581X_{5(3)}(2002) + 0.00016x_1x_2 + 0.$
where$x_i = X_i - \bar{X}i, i = 1, 2$

** analysis **

in order to determine whether we need to keep the quadratic and interaction term we need to use the method of F test to find whether all the relative parameter equal to zero

$H0 : \beta_{11} = \beta_{22} = \beta_{12} = \beta_6 = 0$

and

Ha : not all β in $H0$ equal to 0

```
anova(reg_8.42)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 0.08693 0.08693  3.7203   0.06395 .
## X2         1 0.00000 0.00000  0.0000   0.99531
## X3         1 1.53369 1.53369 65.6386 8.044e-09 ***
## X4         1 0.10800 0.10800  4.6221   0.04035 *
## x5_1       1 0.00898 0.00898  0.3841   0.54041
## x5_2       1 0.03292 0.03292  1.4088   0.24522
## x5_3       1 0.02151 0.02151  0.9205   0.34555
## Residuals 28 0.65424 0.02337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(reg_8.42_2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 0.08693 0.08693  3.4708   0.07425 .
## X2         1 0.00000 0.00000  0.0000   0.99547
## X3         1 1.53369 1.53369 61.2361 3.506e-08 ***
## X4         1 0.10800 0.10800  4.3121   0.04827 *
## x5_1       1 0.00898 0.00898  0.3584   0.55480
## x5_2       1 0.03292 0.03292  1.3143   0.26247
## x5_3       1 0.02151 0.02151  0.8588   0.36294
## x1_2       1 0.01402 0.01402  0.5596   0.46139
## x2_2       1 0.01374 0.01374  0.5487   0.46575
## X1:X2      1 0.00034 0.00034  0.0137   0.90784
## Residuals 25 0.62614 0.02505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSE_F = 0.62614, SSE_R = 0.65424$ So $SSR_R = SSE_R - SSE_F = 0.65424 - 0.62614 = 0.0281$ and $df_R = 4, df_F = 25$

Thus: $F^* = \frac{SSR(x_1^2, x_2^2, x_1x_2, x_1:x_2 | x_1, x_2, x_3, x_4, x_{5(1)}, x_{5(2)}, x_{5(3)})}{df_R} \div \frac{SSE_F}{df_F} = \frac{0.0281}{4} \div \frac{0.62614}{25} = 0.2804884$

```
qf(0.95,4,25)
```

```
## [1] 2.75871
```

clearly, $F^* \leq F(0.95, 4, 25)$, thus we conclude H0, and we can say that it is not need for all the quadratic and interaction terms.

problem c.

```
reg_8.42_3 <- lm(data=data_8.42,Y~X1+X3+X4)
anova(reg_8.42_3)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 0.08693 0.08693  3.8745   0.05774 .
## X3         1 1.52347 1.52347 67.9027 2.057e-09 ***
## X4         1 0.11791 0.11791  5.2552   0.02860 *
## Residuals 32 0.71795 0.02244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(reg_8.42)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 0.08693 0.08693  3.7203   0.06395 .
## X2         1 0.00000 0.00000  0.0000   0.99531
## X3         1 1.53369 1.53369 65.6386 8.044e-09 ***
## X4         1 0.10800 0.10800  4.6221   0.04035 *
## x5_1       1 0.00898 0.00898  0.3841   0.54041
## x5_2       1 0.03292 0.03292  1.4088   0.24522
## x5_3       1 0.02151 0.02151  0.9205   0.34555
## Residuals 28 0.65424 0.02337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

** analysis **

$H0: \beta_2 = \beta_{5(1)} = \beta_{5(2)} = \beta_{5(3)} = 0$

and

Ha : not all β in $H0$ equal to 0

$SSE_F = 0.65424, SSE_R = 0.71795$ So $SSR_R = SSE_R - SSE_F = 0.71795 - 0.65424 = 0.06371$ and $df_R = 4, df_F = 28$

Thus: $F^* = \frac{SSR(x_2, x_{5(1)}, x_{5(2)}, x_{5(3)}|x_1, x_3, x_4)}{df_R} \div \frac{SSE_F}{df_F} = \frac{0.06371}{4} \div \frac{0.65424}{28} = 0.6816612$

```
qf(0.95,4,28)
```

```
## [1] 2.714076
```

clearly, $F^* \leq F(0.95, 4, 28)$, thus we conclude H0, and we can say that it is not need for x2 and x5 term.

8.43

```
data_8.43 <- read.table('8.43.txt',header = FALSE,col.names = c('index','y','x1','x2','x3'))

data_8.43 <- data_8.43[,-1]
#take 1996 as the reference year
data_8.43$x3_1 <- as.numeric(data_8.43$x3==1997)
data_8.43$x3_2 <- as.numeric(data_8.43$x3==1998)
data_8.43$x3_3 <- as.numeric(data_8.43$x3==1999)
data_8.43$x3_4 <- as.numeric(data_8.43$x3==2000)
data_8.43 <- data_8.43[,-4]

reg1 <- lm(data=data_8.43,y~x1+x2)
reg2 <- lm(data=data_8.43,y~.)
summary(reg1);summary(reg2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = data_8.43)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10265 -0.29862  0.07311  0.40355  1.31336
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.292793   0.136725   9.455  < 2e-16 ***
## x1          0.010022   0.001279   7.835 1.74e-14 ***
## x2          0.037210   0.005939   6.266 6.48e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5672 on 702 degrees of freedom
## Multiple R-squared:  0.2033, Adjusted R-squared:  0.2011
## F-statistic: 89.59 on 2 and 702 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = y ~ ., data = data_8.43)
##
## 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) 1.217874   0.144598   8.422  < 2e-16 ***
## x1          0.010124   0.001285   7.878 1.28e-14 ***
## x2          0.037188   0.005951   6.248 7.21e-10 ***
## x3_1        0.083657   0.068816   1.216   0.2245
## x3_2        0.115339   0.066158   1.743   0.0817 .
## x3_3        0.080071   0.067475   1.187   0.2358
## x3_4        0.056007   0.068013   0.823   0.4105
## ---
## 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
```

```
anova(reg1);anova(reg2)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## x1          1  45.007  45.007 139.918 < 2.2e-16 ***
## x2          1  12.628  12.628  39.257 6.479e-10 ***
## Residuals 702 225.813   0.322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## x1          1  45.007  45.007 139.7837 < 2.2e-16 ***
## x2          1  12.628  12.628  39.2195 6.618e-10 ***
## x3_1        1   0.041   0.041   0.1283    0.7203
## x3_2        1   0.553   0.553   1.7166    0.1906
## x3_3        1   0.259   0.259   0.8050    0.3699
## x3_4        1   0.218   0.218   0.6781    0.4105
## Residuals 698 224.742   0.322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

** analysis **

First, let's discuss the whether we need to keep the x3(year). As a brive analysis. We can see add through add x3, we create 4 new parameter, however only very little imporvement in adjusted R square.

Second, we can see the t test for x3 prove that these parameter are all highly possiblily equal to 0.

To determine whether we need to keep the x3, i need to do a f test.

$H0 : \beta_{3(4)} = \beta_{3(1)} = \beta_{3(2)} = \beta_{3(3)} = 0$

and

Ha : not all β in $H0$ equal to 0

$SSE_F = 224.742, SSE_R = 225.813$ So $SSR_R = SSE_R - SSE_F = 225.813 - 224.742 = 1.071$ and $df_R = 4, df_F = 698$

Thus: $F^* = \frac{SSR(x_2,x_{5(1)},x_{5(2)},x_{5(3)}|x_1,x_3,x_4)}{df_R} \div \frac{SSE_F}{df_F} = \frac{1.071}{4} \div \frac{224.742}{698} = 0.8315735$

```
qf(0.95,4,698)
```

```
## [1] 2.384693
```

clearly, $F^* \leq F(0.95, 4, 698)$, thus we conclude H0, and we can say that it is not need for x3 term.

```
data_8.43 <- data_8.43[,c(-4,-5,-6,-7)]
data_8.43$x1 <- scale(data_8.43$x1,scale = FALSE)
data_8.43$x2 <- scale(data_8.43$x2,scale = FALSE)

reg3 <- lm(data=data_8.43,y~.+ I(x1^2) + I(x2^2) + x1:x2)
summary(reg3);summary(reg1)
```

```
##
## Call:
## lm(formula = y ~ . + I(x1^2) + I(x2^2) + x1:x2, data = data_8.43)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05113 -0.30469  0.07794  0.38071  1.33711
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.906e+00  3.021e-02  96.207  < 2e-16 ***
## x1          1.432e-02  1.618e-03   8.849  < 2e-16 ***
## x2          3.539e-02  5.926e-03   5.972 3.73e-09 ***
## I(x1^2)     1.508e-04  5.604e-05   2.691  0.00729 **
## I(x2^2)     1.069e-05  1.139e-03   0.009  0.99252
## x1:x2       5.621e-04  3.556e-04   1.580  0.11446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5589 on 699 degrees of freedom
## Multiple R-squared:  0.2297, Adjusted R-squared:  0.2242
## F-statistic: 41.68 on 5 and 699 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = data_8.43)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10265 -0.29862  0.07311  0.40355  1.31336
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.292793   0.136725   9.455  < 2e-16 ***
## x1          0.010022   0.001279   7.835 1.74e-14 ***
## x2          0.037210   0.005939   6.266 6.48e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5672 on 702 degrees of freedom
## Multiple R-squared:  0.2033, Adjusted R-squared:  0.2011
## F-statistic: 89.59 on 2 and 702 DF,  p-value: < 2.2e-16
```

** analysis **

as a brive analysis, as we can see, x1^2 passed the t test, but other new added parameters don't. And ofter three new parameters have beed added, the adjusted R square changed a little.

To futhre determine whether we need to keep all these square term. Again we need to do a f test.

```
anova(reg3);anova(reg1)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## x1          1  45.007  45.007 144.0811 < 2.2e-16 ***
## x2          1  12.628  12.628  40.4253 3.688e-10 ***
## I(x1^2)     1   6.514   6.514  20.8518 5.864e-06 ***
## I(x2^2)     1   0.169   0.169   0.5395    0.4629
## x1:x2       1   0.780   0.780   2.4977    0.1145
## Residuals 699 218.351   0.312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## x1          1  45.007  45.007 139.918 < 2.2e-16 ***
## x2          1  12.628  12.628  39.257 6.479e-10 ***
## Residuals 702 225.813   0.322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H0 : \beta_{11} = \beta_{22} = \beta_{12} = 0$

and

Ha : not all β in $H0$ equal to 0

$SSE_F = 218.351, SSE_R = 225.813$ So $SSR_R = SSE_R - SSE_F = 225.813 - 218.351 = 7.462$ and $df_R = 4, df_F = 699$

Thus: $F^* = \frac{SSR(x_2, x_{5(1)}, x_{5(2)}, x_{5(3)} | x_1, x_3, x_4)}{df_R} \div \frac{SSE_F}{df_F} = \frac{7.462}{3} \div \frac{218.351}{699} = 7.96262$

```
qf(0.95,3,699)
```

```
## [1] 2.617645
```

clearly, $F^* \geq F(0.95, 4, 698)$, thus we conclude Ha, and we can say still need to keep the square terms.Based on the t test, we choose only keep the x1^2

To sum up,the model we get is:

$\widehat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 X_2 + \beta_{11} x_1$, where $x_1 = X_1 - \bar{X}_1$

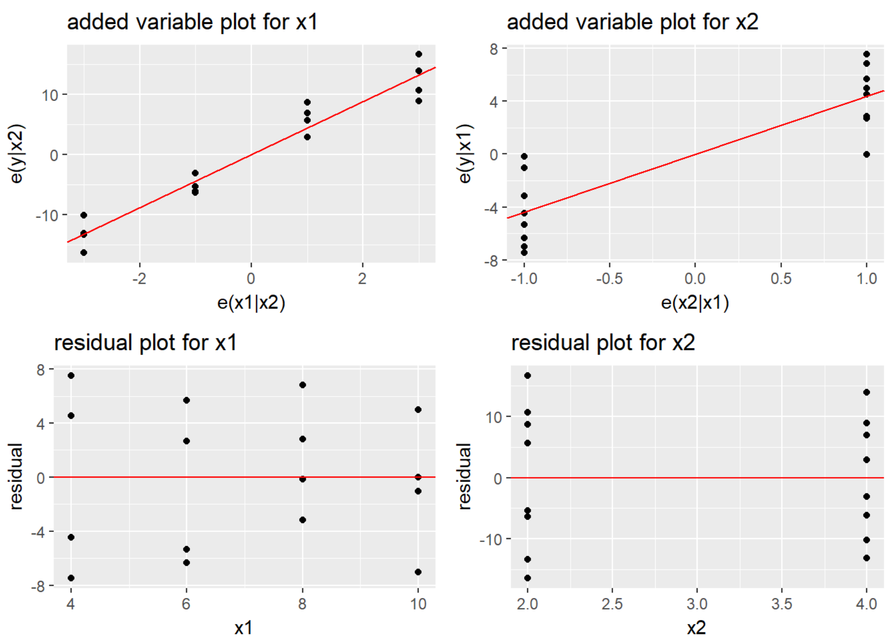
10.5

problem a

```
Brand <- read.table('6.5.txt',header = FALSE, col.names = c('y','x1','x2'))

#  added value plot for x1

reg10_1 <- lm(data = Brand,y~x2)
residual_Y_x2 <- reg10_1$residuals

reg10_2 <- lm(data = Brand,x1~x2)
residual_x1_x2 <- reg10_2$residuals

p1 <- ggplot()+
        geom_point(mapping = aes(x=residual_x1_x2,y=residual_Y_x2))+
        geom_abline(slope=4.425, intercept=0,col='red')+
        labs(title="added variable plot for x1",x="e(x1|x2)",y="e(y|x2)")

#  added value plot for x2
reg10_3 <- lm(data = Brand,y~x1)
residual_Y_x1 <- reg10_3$residuals

reg10_4 <- lm(data = Brand,x2~x1)
residual_x2_x1 <- reg10_4$residuals

p2 <- ggplot()+
        geom_point(mapping = aes(x=residual_x2_x1,y=residual_Y_x1))+
        geom_abline(slope=4.375, intercept=0,col='red')+
        labs(title="added variable plot for x2",x="e(x2|x1)",y="e(y|x1)")

p3 <- ggplot()+
        geom_point(mapping = aes(x=Brand$x1,y=reg10_3$residuals))+
        geom_abline(slope=0, intercept=0,col='red')+
        labs(title='residual plot for x1',x='x1',y='residual')

p4 <- ggplot()+
        geom_point(mapping = aes(x=Brand$x2,y=reg10_1$residuals))+
        geom_abline(slope=0, intercept=0,col='red')+
        labs(title='residual plot for x2',x='x2',y='residual')

library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.4.2
```

```
grid.arrange(p1,p2,p3,p4,ncol=2)
```



problem b

** analysis **

our model in 6.5(b) is $\widehat{Y} = 37.650 + 4.425X_1 + 4.375X_2$. In the problem a, I have indicated the solve of x1 and x2 in the plot.

And in the plot we draw in the problem a we can see that : from the added varibale plot for x1, we can see that :

relatively, when x2 has already in the model, the x1 provide a lot additional help in the regression model. While ,when x1 has already in the model, the x2 provide little additional help.

Besides, both added variables tends to be adequate because no curvilinear relation is suggested by the scatter of points.

problem c

According to the reg10_1 we have already regress the y on x2 and according to the reg10_3 we have already regress the y on x3.

```
reg10_1;reg10_3
```

```
##
## Call:
## lm(formula = y ~ x2, data = Brand)
##
## Coefficients:
## (Intercept)           x2
##      68.625        4.375
```

```
##
## Call:
## lm(formula = y ~ x1, data = Brand)
##
## Coefficients:
## (Intercept)           x1
##      50.775        4.425
```

Clearly, the result is : $\widehat{Y}(X_2) = 68.625 + 4.375X_2$ and $\widehat{Y}(X_1) = 50.775 + 4.425X_1$.

Then, according to the result in the problem (b) we know ,when x2 has already in the model, x1 provide a lot of additional help. Thus, it is appropriate to include x1 into the model with x2.

Firstly we need to calculate the $e(\widehat{Y|X_1}) = Y - \widehat{Y}(X_1)$ and $e(\widehat{X_2|X_1}) = X - \widehat{X_2}(X_1)$. And make a regression baed on them.

```
summary(reg10_4)
```

```
##
## Call:
## lm(formula = x2 ~ x1, data = Brand)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.000e+00  8.783e-01   3.416  0.00418 **
## x1          -2.483e-17  1.195e-01   0.000  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.069 on 14 degrees of freedom
## Multiple R-squared:  6.163e-32,  Adjusted R-squared:  -0.07143
## F-statistic: 8.628e-31 on 1 and 14 DF,  p-value: 1
```

As we can see $\widehat{X_2}(X_1) = 3 + -2.483e - 17X_1 \approx 3$

Thus: $e(\widehat{Y|X_1}) = Y - \widehat{Y}(X_1) = Y - (50.775 + 4.425X_1)$ and $e(\widehat{X_2|X_1}) = X_2 - \widehat{X_2}(X_1) = X_2 - 3$. And baed on this we can make a regression

```
e_x <- Brand$x2-3
e_y <- Brand$y-(50.775 + 4.425*Brand$x1)
reg10_5 <- lm(e_y~e_x-1)
reg10_5
```

```
##
## Call:
## lm(formula = e_y ~ e_x - 1)
##
## Coefficients:
##   e_x
## 4.375
```

Thus, we have: $[Y - \widehat{Y}(X_1)] = 4.375[X_2 - \widehat{X_2}(X_1)]$ After some basic regulation, we get: $\widehat{Y} = 37.650 + 4.425X_1 + 4.375X_2$

10.9

problem a

```
# read the data
Brand <- read.table('6.5.txt',header = FALSE, col.names = c('y','x1','x2'))

# construct x matrix
x <- matrix(nrow = 16,ncol = 3)
x <- Brand
x[,1] <- 1
x <- as.matrix(x)
colnames(x) <- NULL

# calculate the standard deleted residuals
reg_10_9 <- lm(data = Brand,y~.)
standard_delected_residual  <- rstandard(reg_10_9)
standard_delected_residual <- as.data.frame(standard_delected_residual)

# calculate the t critical value
t_critical <-  qt(1-0.1/(2*16),16-3-1)

# result
standard_delected_residual$test <- ifelse( abs(standard_delected_residual$standard_delected_residual)< t_critical,"no outlie
rs","outliers")
head(standard_delected_residual)
```

```
##   standard_delected_residual        test
## 1                -0.04252026 no outliers
## 2                 0.06378039 no outliers
## 3                -1.31812804 no outliers
## 4                 1.33938817 no outliers
## 5                -0.37980431 no outliers
## 6                -0.67964981 no outliers
```

```
# solution two:use the function
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.2
```

```
outlierTest(reg_10_9)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 14 -2.102726           0.057267      0.91627
```

** analysis **

$H0$: there is no outliers AND Ha : there is outliers

if $|t_i| < t(1 - \alpha/2n, n - p - 1)$ we conclude H0.

As we can see there is no outliers in the data set.

problem b

```
influence(reg_10_9)$hat
```

```
##      1      2      3      4      5      6      7      8      9     10
## 0.2375 0.2375 0.2375 0.2375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375
##     11     12     13     14     15     16
## 0.1375 0.1375 0.2375 0.2375 0.2375 0.2375
```

** analysis **

$h_{ii}$is a measure of the distance between the X values for the i-th case and the means of the X values for all n cases. Thus, a large value h_{ii} indicates that the ith case is distant from the center of all X observations.

problem c

```
# calculate the h_bar
h_bar <- sum(influence(reg_10_9)$hat) / 16
h_bar > (2*3)/16
```

```
## [1] FALSE
```

** analysis **

clearly, as we can see : there is no outliers to x value.

problem d

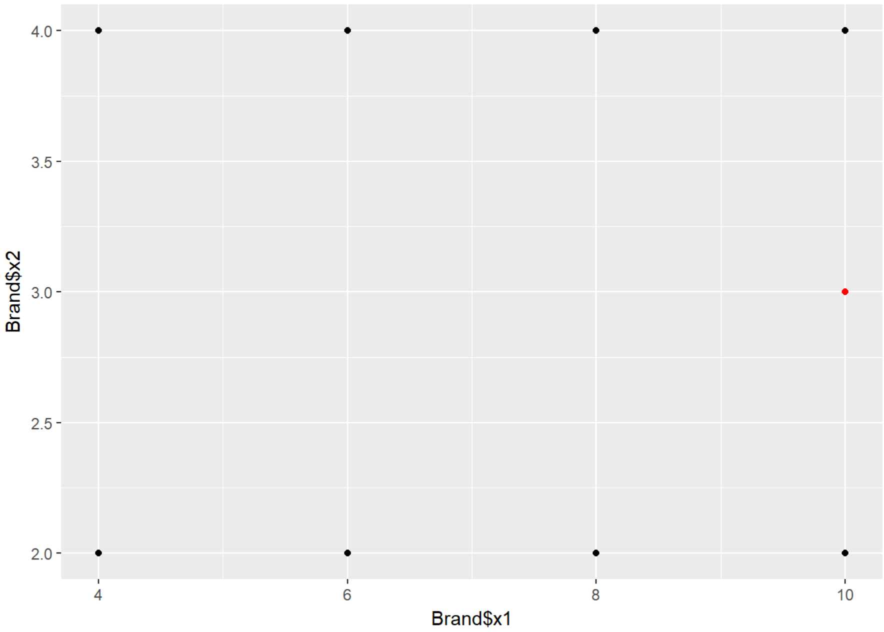
```
x_new <- matrix(c(1,10,3),ncol = 1)
h_new <- t(x_new)%*%solve(t(x)%*%x)%*%x_new
h_new < max(influence(reg_10_9)$hat) & h_new > min(influence(reg_10_9)$hat)
```

```
##      [,1]
## [1,] TRUE
```

** analysis **

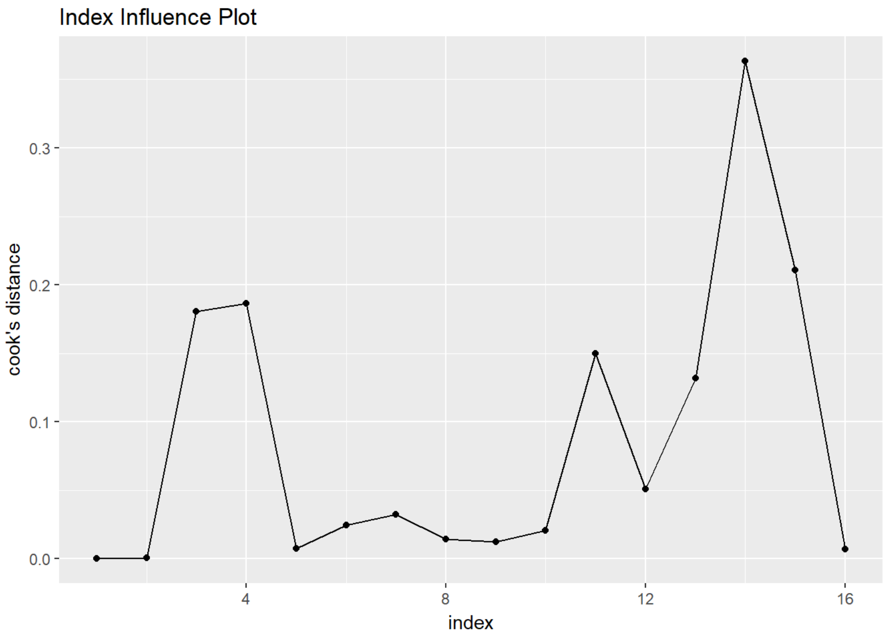
clearly the no extrapolation is involved.

```
ggplot()+
        geom_point(mapping = aes(x=Brand$x1,y=Brand$x2)) +
        geom_point(aes(x=10,y=3),col='red')
```



problem g

```
D <- cooks.distance(reg_10_9)
ggplot()+
        geom_line(aes(x=1:16,y=D))+
        geom_point(aes(x=1:16,y=D))+
        labs(x='index',y='cook\'s distance',title='Index Influence Plot')
```



** analysis **

as we can see the index 14 has the highest cook's distance with the value 0.3634.

```
pf(0.3634,3,13)
```

```
## [1] 0.2194597
```

As, we can see 0.3634 is the 21.9-th percentile of this distribution. Hence, it appears that case 14 does influence the regression fit,but the extent of the influence may not be large enough to call for consideration of remedial measures.

problem 6

① Basic Solution

i) if we delet i-th observation. and then $\hat{\vec{\beta}} = \hat{\vec{\beta}}_{(i)}$

we can conclude that:

$Q = ||\vec{y} - \vec{X}\vec{\beta}||$ do not change.

thus: $y_i - \vec{X_i}\hat{\vec{\beta}}$ keep same

thus $\vec{X_i}$ must on the fitted space.

ii) if $\vec{X_i}$ on the fitted space, then Q not change.

the we get the same estimate of parameters.

② Solution based on DFFITS.

For $D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$

if $\vec{X_i}$ on the fitted space. then $e_i^2 = 0$

the $D_i = 0$ which can not be any influential at all.

solution of problem 6