HW5 GR5205 - Multiple Linear Regression Model

Mathieu Sauterey - UNI: mjs2364

5 December 2017

Problem 1 (35p) (Problem 9.10 b,c, Problem 9.11 a & Problem 9.18 a,b)

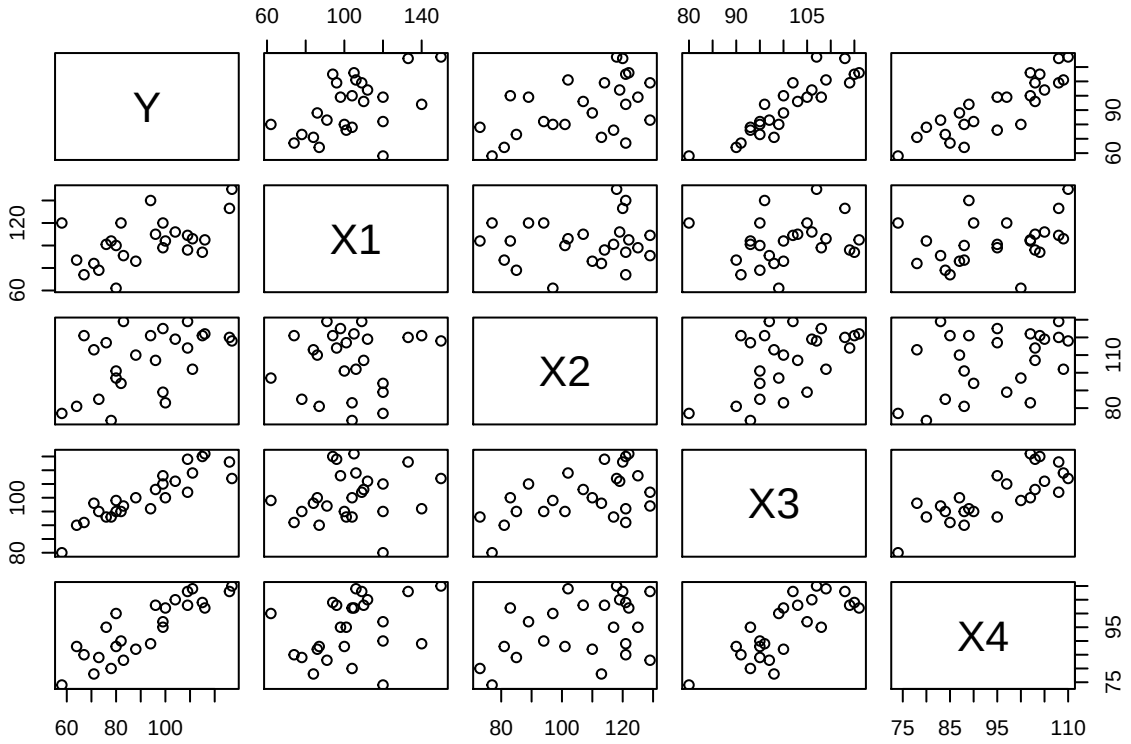
A personnel officer in a governmental agency administrated four newly developed aptitude test to each of 25 applicants for entry-level clerical positions in the agency. For purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job.

(a)(5p) Obtain the scatter plot matrix, and the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between the response variable Y and each of the predictor variables? Are any serious multi-collinearity problems evident? Explain.

```
#Read data and renames the column
data <- read.table("Homework_data_5_Problem1and2.txt", header = FALSE, as.is =TRUE)
names(data) <- c("Y","X1", "X2", "X3", "X4")

# Prints the scatter plot matrix and correlation matrix
library(graphics)
cor(data)
```

```
##            Y        X1        X2        X3        X4
## Y  1.0000000 0.5144107 0.4970057 0.8970645 0.8693865
## X1 0.5144107 1.0000000 0.1022689 0.1807692 0.3266632
## X2 0.4970057 0.1022689 1.0000000 0.5190448 0.3967101
## X3 0.8970645 0.1807692 0.5190448 1.0000000 0.7820385
## X4 0.8693865 0.3266632 0.3967101 0.7820385 1.0000000
```

```
pairs(data)
```



The scatterplot matrix suggests that all four explanatory variables are linearly associated with the response variable Y. Based on the correlation matrix we notice that the X2 and X3 are positively correlated with correlation equal to 0.51, and most importantly that X3 and X4 have a strong positive correlation of 0.78. Thus, multi-colinearity problems may arise between X3 and X4.

(b)(5p) Fit the multiple regression function containing all four predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

$$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \beta_3 * X_{i3} + \beta_4 * X_{i4} + \epsilon_i$$

Considering a multiple linear regression model of the form above, we compute the estimated regression coefficients using the function lm.

```
#Stores each variables in a new vector for more clarity
X1 <- data$X1
X2 <- data$X2
X3 <- data$X3
X4 <- data$X4
Y  <- data$Y

# Regresses Y on all four X variables
lm1 <- lm(Y ~ X1 + X2 + X3 + X4)
beta_0 <- lm1$coefficients[1]
beta_1 <- lm1$coefficients[2]
beta_2 <- lm1$coefficients[3]
```

```
beta_3 <- lm1$coefficients[4]
beta_4 <- lm1$coefficients[5]

# Renames the estimated regression coefficients and prints them
names(lm1$coefficients) <- c("Intercept","B1","B2","B3","B4")
lm1$coefficients[1]
```

```
## Intercept
## -124.3818
```

```
lm1$coefficients[2]
```

```
##        B1
## 0.2957254
```

```
lm1$coefficients[3]
```

```
##         B2
## 0.04828772
```

```
lm1$coefficients[4]
```

```
##       B3
## 1.306011
```

```
lm1$coefficients[5]
```

```
##        B4
## 0.5198191
```

We finally obtain the estimated regression function

$$\hat{Y_i} = -124.382 + 0.296 * X_{i1} + 0.0483 * X_{i2} + 1.306 * X_{i3} + 0.52 * X_{i4}$$

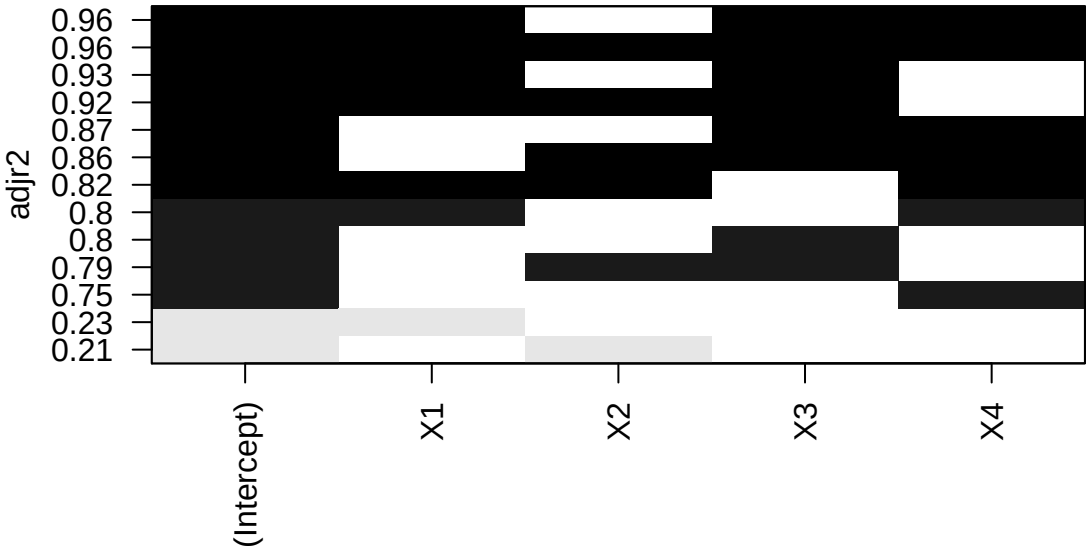
X2 is assigned a regression coefficient β_2=0.0483 which has a very low value, thus low predictive power and it probably shouldn't be retained.

(c)(5p) Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression models according to the adjusted R2 criterion.

```
# Plots a graph of X subsets vs adjusted R-squared
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.4.3
```

```
leaps=regsubsets(Y ~ X1 + X2 + X3 + X4, data=data, nbest=4)
plot(leaps, scale="adjr2")
```



According to the adjusted R2 criterion, the 4 best subset models are (best model listed first):

{X1, X3, X4}

{X1, X2, X3, X4}

{X1, X3}

{X1, X2, X3}

(d)(5p) Using forward stepwise regression find the best subset of predictor variables to predict

job proficiency. Use alpha limits of 0.05 and 0.10 for adding or deleting a variable, respectively.

The 'step' package can only do stepwise regression using AIC criteria, therefore we must add and drop variables manually to use the p-value as criteria.

```
library(MASS)

# Starts by doing regressions with all, and without any coeffcients
all <- lm(Y ~ X1 + X2 + X3 + X4)
none <- lm(Y ~ 1)


addterm(none, scope=all, test="F")
```

```
## Single term additions
```

```
##
## Model:
## Y ~ 1
##        Df Sum of Sq    RSS    AIC F Value     Pr(F)
## <none>              9054.0 149.30
## X1      1    2395.9 6658.1 143.62   8.276  0.008517 **
## X2      1    2236.5 6817.5 144.21   7.545  0.011487 *
## X3      1    7286.0 1768.0 110.47  94.782 1.264e-09 ***
## X4      1    6843.3 2210.7 116.06  71.198 1.699e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# adding X3 has the lowest p-value < 0.05 so we add it to the model
best_subset1 <- update(none, ~. +X3)

dropterm(none, test="F")
```

```
## Single term deletions
##
## Model:
## Y ~ 1
##        Df Sum of Sq  RSS   AIC F Value Pr(F)
## <none>              9054 149.3
```

```
# There are no variables to exclude, so we move on to add one more

addterm(best_subset1, scope=all, test="F")
```

```
## Single term additions
##
## Model:
## Y ~ X3
##        Df Sum of Sq     RSS     AIC F Value     Pr(F)
## <none>              1768.02 110.469
## X1      1   1161.37  606.66  85.727  42.116 1.578e-06 ***
## X2      1     12.21 1755.81 112.295   0.153   0.69946
## X4      1    656.71 1111.31 100.861  13.001   0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# adding X1 has the lowest p-value < 0.05 so we add it to the model
best_subset2 <- update(best_subset1, ~. +X1)

dropterm(best_subset1, test="F")
```

```
## Single term deletions
##
## Model:
## Y ~ X3
##        Df Sum of Sq  RSS    AIC F Value     Pr(F)
## <none>              1768 110.47
## X3      1      7286 9054 149.30  94.782 1.264e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# There are no p-values > 0.1 so we do not exclude any variable
```

```r
addterm(best_subset2, scope=all, test="F")
```

```
## Single term additions
##
## Model:
## Y ~ X3 + X1
##        Df Sum of Sq    RSS    AIC F Value     Pr(F)
## <none>              606.66 85.727
## X2      1     9.937 596.72 87.314  0.3497 0.5605965
## X4      1   258.460 348.20 73.847 15.5879 0.0007354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# adding X4 has the lowest p-value < 0.05 so we add it to the model
best_subset3 <- update(best_subset2, ~. +X4)

dropterm(best_subset2, test="F")
```

```
## Single term deletions
##
## Model:
## Y ~ X3 + X1
##        Df Sum of Sq    RSS     AIC F Value     Pr(F)
## <none>               606.7  85.727
## X3      1    6051.5 6658.1 143.618 219.453 6.313e-13 ***
## X1      1    1161.4 1768.0 110.469  42.116 1.578e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# There are no p-values > 0.1 so we do not exclude any variable

addterm(best_subset3, scope=all, test="F")
```

```
## Single term additions
##
## Model:
## Y ~ X3 + X1 + X4
##        Df Sum of Sq    RSS    AIC F Value  Pr(F)
## <none>              348.20 73.847
## X2      1     12.22 335.98 74.954  0.7274 0.4038
```

```r
# There are no p-values < 0.05 so we do not add any variable

# We cannot add or remove any variable anymore, so the algorithm ends

best_subset3
```

```
##
## Call:
## lm(formula = Y ~ X3 + X1 + X4)
##
## Coefficients:
## (Intercept)           X3           X1           X4
##   -124.2000       1.3570       0.2963       0.5174
```

Using forward stepwise regression with p-value criteria, the best subset of predictor variables found is {X1, X3, X4}.

(e)(5p) How does the best subset according to forward stepwise regression compare with the best subset according to the adjusted R2 criterion from (c) above?

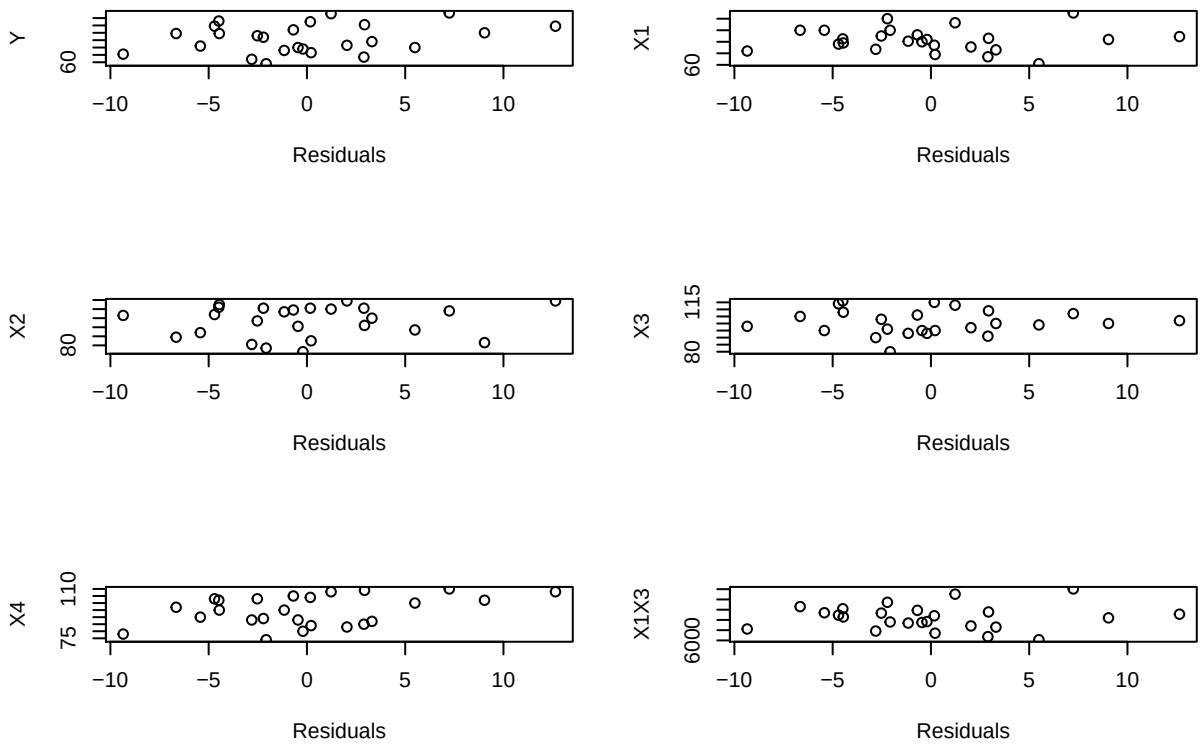
We obtain the same best subset {X1, X3, X4} with forward stepwise regression as with the adjusted R2 criterion from (c) above.

Problem 2 (65p) (Problem 10.19 a,b,c,d,e (first part),f,g)

The subset model from Problem 1 containing only first-order terms in X1 and X3 is to be evaluated in detail.

(a)(5p) Obtain the residuals and plot them separately against Y, each of the four predictor variables, and the cross-product term X1X3. On the basis of these plots, should any modifications in the regression model be investigated?

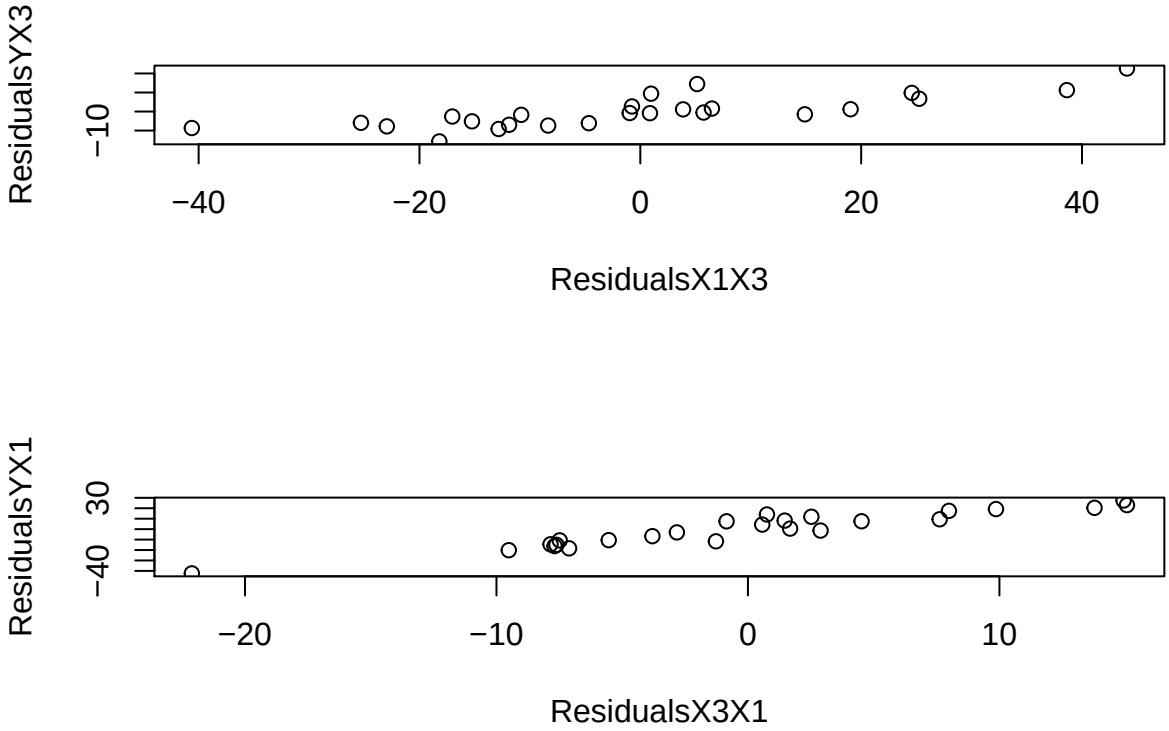
```
# Regresses Y on X1 and X3 variables
lm13 <- lm(Y ~ X1 + X3)

# Obtain the residuals from this regression
Residuals <- lm13$residuals

# Creates a new cross-product variable term
X1X3 <- X1*X3

# Plots the residuals against Y, each of the four X variables, and X1X3
par(mfrow=c(3,2))
plot(Residuals, Y)
plot(Residuals,X1)
plot(Residuals,X2)
plot(Residuals,X3)
plot(Residuals,X4)
plot(Residuals,X1X3)
```



We clearly see that the residuals have a linear association with X4 so we could add X4 to the model.

(b)(10p) Prepare separate added-variable plots against e(X1|X3) and e(X3|X1). Do these plots suggest that any modifications in the model form are warranted?

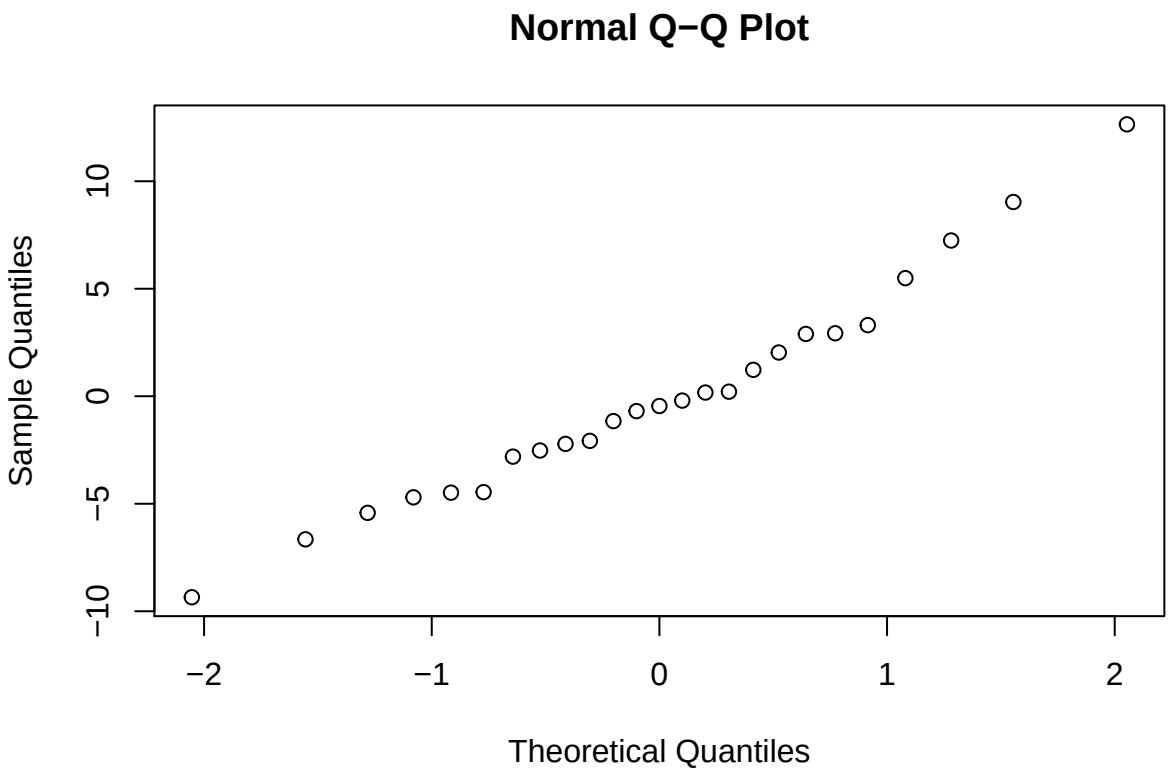
```
# Regresses Y on X3 only, and Y on X1 only
lm3 <- lm(Y ~ X3)
lm1 <- lm(Y ~ X1)

# Obtains the residuals for each regression
ResidualsYX1 <- lm1$residuals
ResidualsYX3 <- lm3$residuals

# Regresses X1 on X3, and X3 on X1
lmX1X3 <- lm(X1 ~ X3)
lmX3X1 <- lm(X3 ~ X1)

# Obtains the residuals for each regression
ResidualsX1X3 <- lmX1X3$residuals
ResidualsX3X1 <- lmX3X1$residuals

# Plots the residuals of Y|X3 against X1|X3, Y|X1 against X3|X1,
par(mfrow=c(2,1))
plot(ResidualsX1X3, ResidualsYX3)
plot(ResidualsX3X1, ResidualsYX1)
```



Both added-variable plots suggest that adding X1 to the model containing X3, and adding X3 to the model containing X1 may be helpful. As our model already contains X1 and X3, there is no need to modify it.

(c)(10p) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumptions, using alpha = 0.1. What do you conclude?

```
# QQplot of the residuals
sample.quants <- qqnorm(lm13$residuals)
```




```
# Correlation btween ordered residuals and their expected values under normality
cor(lm13$residuals, sample.quants$x)
```

```
## [1] 0.9840739
```

```
# Uses Shapiro-Wilk to test normality of residuals distribution
shapiro.test(lm13$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm13$residuals
## W = 0.97113, p-value = 0.6738
```

The Shapiro-Wilk's method assumes that the null hypothesis is H0: "Sample is normally distributed". Here, the p-value returned is 0.6738 which is larger than 0.1. Therefore we cannot reject the null hypothesis and normality can reasonably be assumed.

(d)(10p) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test (a t-test with t(1-alpha/(2N);N-P-1) critical value) with alpha = 0.05. State the decision rule and conclusion.

```
# Calculates the deleted studentized residuals
student.del.residuals <- rstudent(lm13)

N <- dim(data)[1]
```

```
P <- dim(data)[2]-1
qt(1-0.05/(2*N),N-P-1)
```

```
## [1] 3.551808
```

```
# Uses Bonferroni to test for any outlying Y observations
abs(student.del.residuals) <= qt(1-0.05/(2*N),N-P-1)
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##   16   17   18   19   20   21   22   23   24   25
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

We performed the Bonferroni test for outliers with critical value t* = 3.552 and decision rule as follows: If $|t_i|$ <= t* then conclude that t_i is not an outlier If $|t_i|$ > t* then conclude that t_i is an outlier.

All values $|t_i|$ <= t* therefore we conclude that there are no outliers in our Y observations.

(e)(10p) Obtain the diagonal elements of the hat matrix. Using the rule of thumb (Xi is outlying case with regard to the X values, if hii > 2P=N), identify any outlying X observations.

```
# Obtains the diagonal elements of the hat matrix
influence(lm13)$hat
```

```
##          1          2          3          4          5          6
## 0.07065696 0.21418067 0.04600259 0.07240679 0.05797031 0.11297480
##          7          8          9         10         11         12
## 0.33658569 0.16503154 0.05876518 0.07258372 0.11674502 0.18227927
##         13         14         15         16         17         18
## 0.20825309 0.07879682 0.07578796 0.04353035 0.04043874 0.26349128
##         19         20         21         22         23         24
## 0.07482306 0.09578960 0.15795012 0.14979081 0.07389471 0.17143610
##         25
## 0.05983483
```

```
# Uses rule of thumb (hii > 2P=N) to identify any outlying X observations
influence(lm13)$hat > 2*P/N
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##    13    14    15    16    17    18    19    20    21    22    23    24
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    25
## FALSE
```

Based on the rule of thumb, we identify the 7th X observation as a potential outlier.

(f)(10p) Case 7 and 18 appear to be moderately outlying with respect to their X values, and case 16 is reasonably far outlying with respect to its Y value. Obtain DFFITS, DFBETAS, and Cook's distance values for these cases to asses their influence. What do you conclude?

```
# Obtains DFFITS, DFBETAS, and Cook's distance values for cases 7,16,18

dffits(lm13)[c(7,16,18)]
```

```
##          7         16         18
## -0.3395422  0.6030725  0.9998970
```

```
abs(dffits(lm13)[c(7,16,18)]) > 1
```

```
##     7    16    18
## FALSE FALSE FALSE
```

```
dfbetas(lm13)[c(7,16,18),]
```

```
##    (Intercept)         X1         X3
## 7  -0.24018835 -0.1511532 0.30330857
## 16 -0.06867086  0.1519977 0.05116114
## 18 -0.46436969  0.8777907 0.11512640
```

```
abs(dfbetas(lm13)[c(7,16,18)]) > 1
```

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

```
cooks.distance(lm13)[c(7,16,18)]
```

```
##          7         16         18
## 0.03982863 0.09199677 0.30812948
```

```
abs(cooks.distance(lm13)[c(7,16,18)]) > 1
```

```
##     7    16    18
## FALSE FALSE FALSE
```

After passing DFFITS, DFBETAS, and Cook's distance for these 3 observations, we conclude that they are not influential.

(g)(10p) Obtain the variance inflation factors. What do they indicate?

```
# Obtains the variance inflation factors
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3
```

```
vif(lm13)
```

```
##       X1       X3
## 1.033781 1.033781
```

There are no VIF > 10 which indicates that multicollinearity is not influencing the estimates.