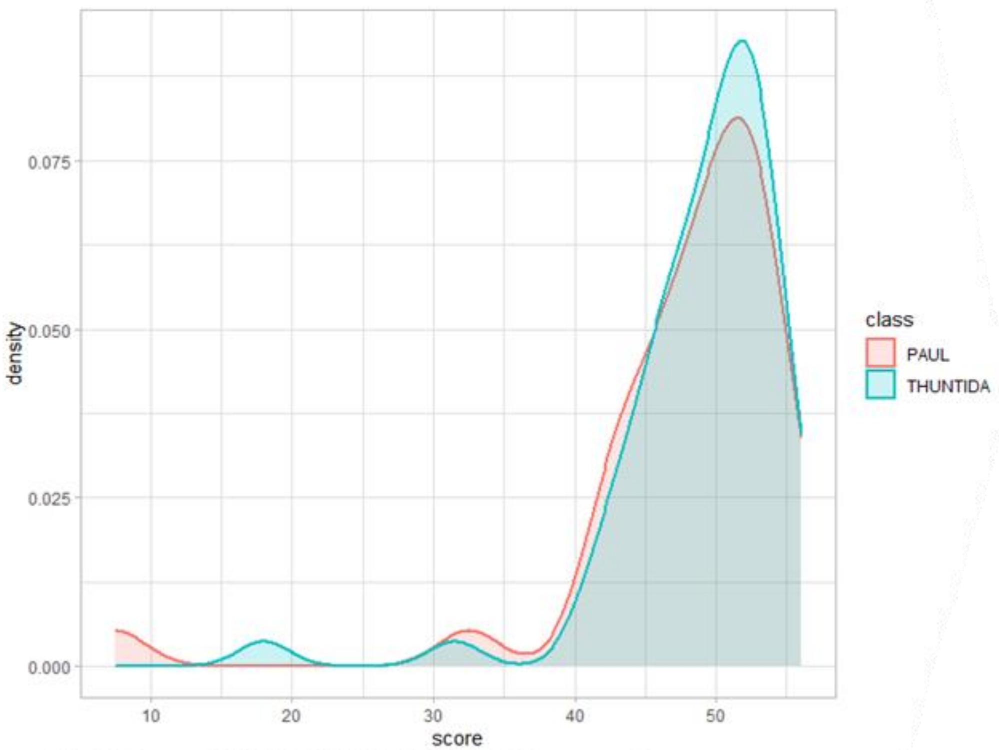
TOPIC4 In Class Problems



	class	meanScore	sdScore	n
	<fct>	<dbl>	<dbl>	<int>
1	PAUL	47.8	8.61	35
2	THUNTIDA	49.0	6.04	62

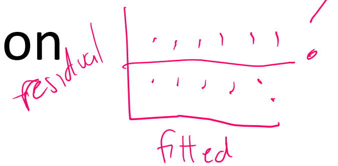
PROBLEM 16

Use the CLERICAL.CSV data.

BEGIN with the model `Y ~ X2 + X4 + X5`

$E(\epsilon_i) = 0$?

(a) Check whether this model meets the linearity assumption



(b) If it doesn't (it doesn't), use ggpairs() to identify potential terms that might be transformed in a higher-order model.

(c) Fit that higher-order model and evaluate.

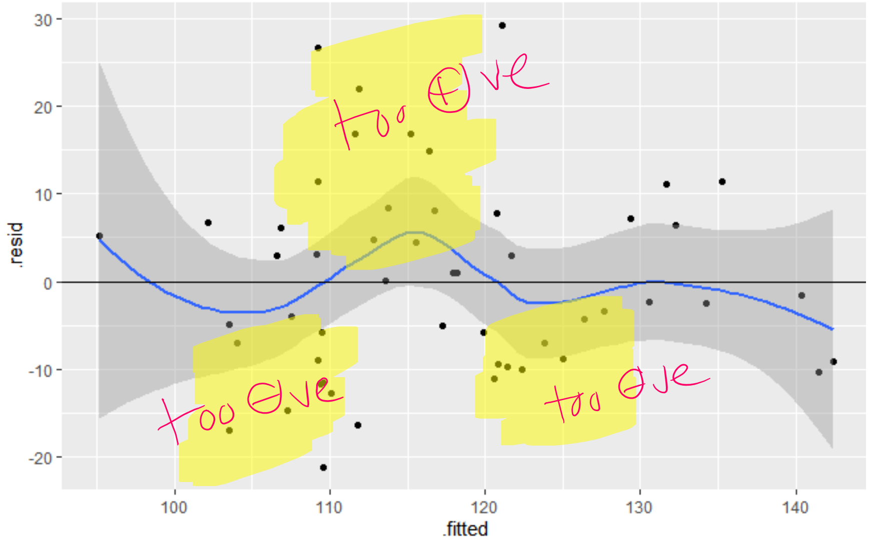
ANSWER TO PROBLEM 16

(a) Check whether this model meets the linearity assumption

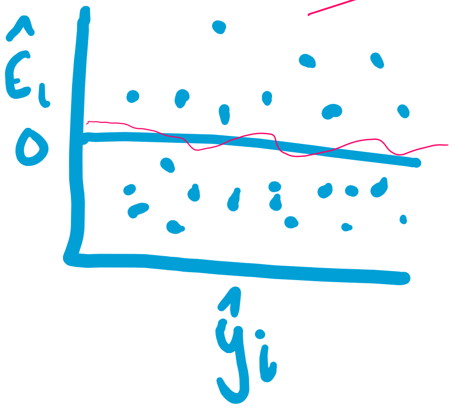
(b) If it doesn't (it doesn't), use ggpairs() to identify potential terms that might be transformed in a higher-order model.

(c) Fit that higher-order model and evaluate.

Doesn't meet linearity assumption



what we want

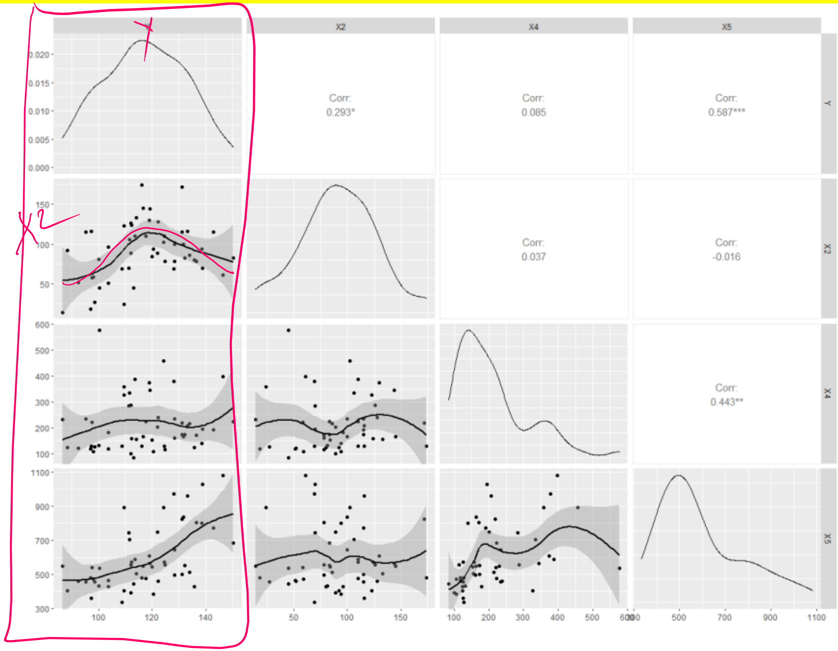


ANSWER TO PROBLEM 16

(a) Check whether this model meets the linearity assumption

(b) If it doesn't (it doesn't), use ggpairs() to identify potential terms that might be transformed in a higher-order model.

(c) Fit that higher-order model and evaluate.



ANSWER TO PROBLEM 16

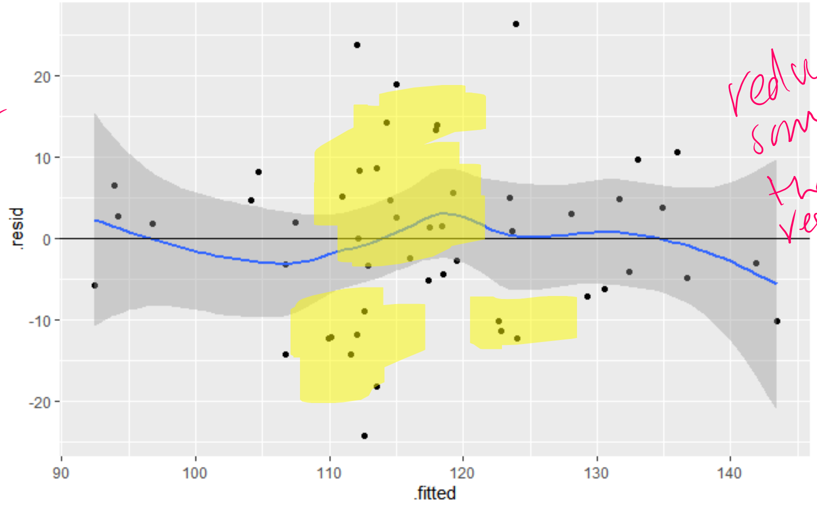
(a) Check whether this model meets the linearity assumption

(b) If it doesn't (it doesn't), use ggpairs() to identify potential terms that might be transformed in a higher-order model.

(c) Fit that higher-order model and evaluate.

```
Call:
lm(formula = Y ~ X2 + I(X2^2) + X4 + X5, data = workhours)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.315  -6.480   1.185   5.320  26.482 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 61.0933183  8.7297596   6.998 8.22e-09 ***
X2           0.5762076  0.1611431   3.576 0.000821 ***
I(X2^2)     -0.0024326  0.0008596  -2.830 0.006827 ** 
X4          -0.0326852  0.0160268  -2.039 0.047054 *  
X5           0.0571700  0.0090822   6.295 9.62e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.78 on 47 degrees of freedom
Multiple R-squared:  0.5562,	Adjusted R-squared:  0.5184 
F-statistic: 14.73 on 4 and 47 DF,  p-value: 7.196e-08
```

$X2^2$



reduced some of the residuals

A model that meets linearity assumption

PROBLEM 17

Use the CLERICAL.CSV data.

BEGIN with the best model from PROBLEM 16

```
Y ~ X2 + I(X2^2) + X4 + X5
```

Does this model meet the equal-variance assumption?

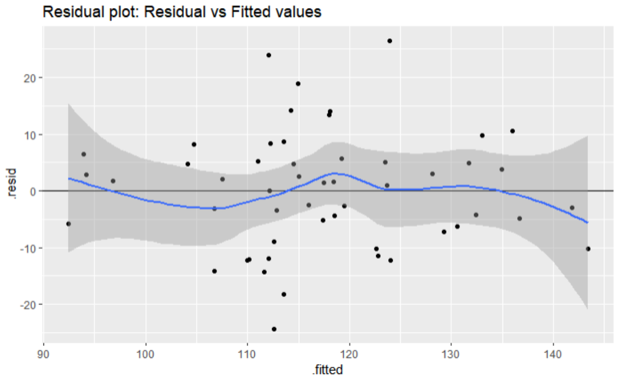
(a) Examine residual plot and scale-location plot

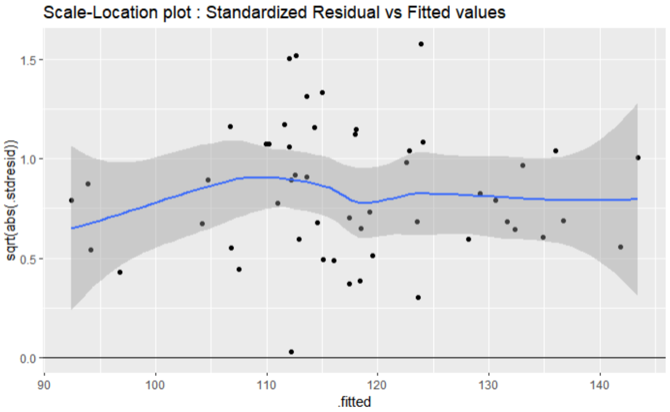
(b) Conduct the Breusch-Pagan test.

ANSWER TO PROBLEM 17

(a) Examine residual plot and scale-location plot

(b) Conduct the Breusch-Pagan test.






```
studentized Breusch-Pagan test

data:  improvemodel
BP = 6.7107, df = 4, p-value = 0.152
```

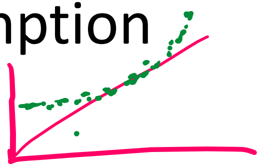
PROBLEM 18

Use the CLERICAL.CSV data.

BEGIN with the best model from PROBLEM 16

```
Y ~ X2 + I(X2^2) + X4 + X5
```

Does this model meet the normality assumption



QQ Plots

(a) Examine histograms and ~~qqplots~~ of the residuals

(b) Conduct the Shapiro-Wilk test. bptest(——)

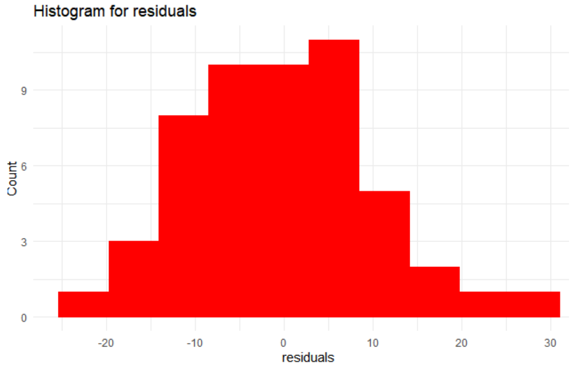
ggplot(——)+
stat_qq()+
stat_qq_line()

ANSWER TO PROBLEM 18

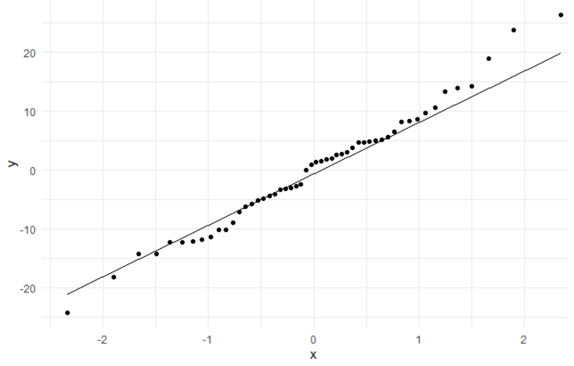
(a) Examine histograms and qqplots of the residuals

(b) Conduct the Shapiro-Wilk test.

```
ggplot(data=workhours, aes(residuals(improvemodel)))+
  geom_histogram(bins=10, col="red", fill="red") +
  labs(title="Histogram for residuals") +
  labs(x="residuals", y="Count") +
  theme_minimal()
```




```
	Shapiro-Wilk normality test

data:  residuals(improvemodel)
W = 0.9875, p-value = 0.8576
```



PROBLEM 19

Use the CREDIT.CSV data.

Consider the model:

```
BALANCE ~ Income + Rating + Age + Limit + Cards + factor(Student)
```

(a) Examine scatter plots among the variables

(b) Test for multicollinearity using VIF.

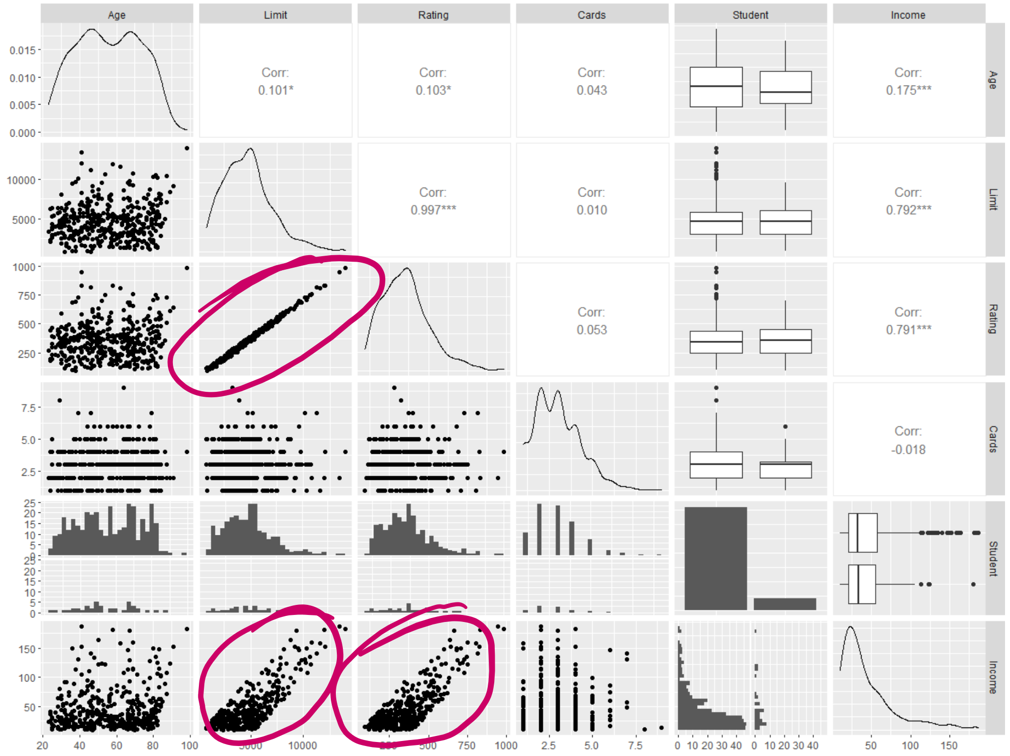
(c) Discuss what we should do

ANSWER TO PROBLEM 19

(a) Examine scatter plots among the variables

(b) Test for multicollinearity using VIF.

(c) Discuss what we should do



ANSWER TO PROBLEM 19

```
> multimodel<-lm(Balance~Income+Rating+Age+Limit+Cards+factor(Student),data=credit)
```

(a) Examine scatter plots among the variables

(b) Test for multicollinearity using VIF.

(c) Discuss what we should do

```
> imcdiag(multimodel, method="VIF")

Call:
imcdiag(mod = multimodel, method = "VIF")


 VIF Multicollinearity Diagnostics

                        VIF detection
Income               2.7769         0
Rating             230.8695         1
Age                  1.0397         0
Limit              229.2385         1
Cards                1.4390         0
factor(Student)Yes   1.0091         0

Multicollinearity may be due to Rating Limit regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

```
> library(car)
> vif(multimodel)
         Income          Rating             Age           Limit           Cards factor(Student)
       2.776906      230.869514        1.039696      229.238479        1.439007        1.009064
```

PROBLEM 20

Use the CREDIT.CSV data.

We found multicollinearity in this model

`BALANCE ~ Income + Rating + Age + Limit + Cards + factor(Student)`

Let’s remove `Limit` as it is clearly highly correlated with `Rating`

(a) Rerun the model and check the VIF

ANSWER TO PROBLEM 20

(a) Rerun the model and check the VIF

```
> nomultimodel<-lm(Balance~Income+Rating+Age+Cards+factor(Student),data=credit)
> summary(nomultimodel)

Call:
lm(formula = Balance ~ Income + Rating + Age + Cards + factor(Student),
    data = credit)

Residuals:
    Min      1Q  Median      3Q     Max
-214.37  -79.91  -12.38   66.19  295.23

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -557.62738   23.37703 -23.854   <2e-16 ***
Income               -7.76925    0.24346 -31.912   <2e-16 ***
Rating                3.97382    0.05492  72.354   <2e-16 ***
Age                  -0.64215    0.30441  -2.110   0.0355 *
Cards                 4.20917    3.78584   1.112   0.2669
factor(Student)Yes  417.90477   17.17026  24.339   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 102.9 on 394 degrees of freedom
Multiple R-squared:  0.9506,    Adjusted R-squared:  0.9499
F-statistic:  1515 on 5 and 394 DF,  p-value: < 2.2e-16
```

```
> imcdiag(nomultimodel, method="VIF")

Call:
imcdiag(mod = nomultimodel, method = "VIF")


 VIF Multicollinearity Diagnostics

                      VIF detection
Income             2.7760         0
Rating             2.7226         0
Age                1.0396         0
Cards              1.0161         0
factor(Student)Yes 1.0029         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test
```

PROBLEM 21

Use the CLERICAL.CSV data.

Use this model: `Y ~ X2 + I(X2^2) + X4 + X5`

(a) Plot the residuals versus leverage plot

(b) Explore the leverages by observation number

(c) Examine the Cook’s distances

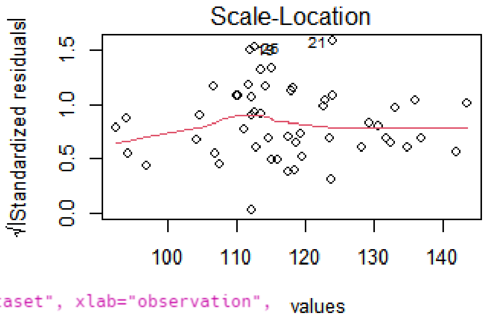
ANSWER TO PROBLEM 21

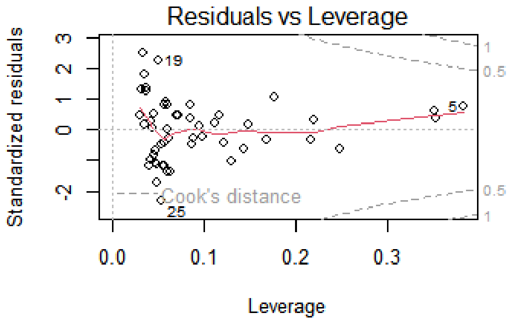
(a) Plot the residuals versus leverage plot

(b) Explore the leverages by observation number

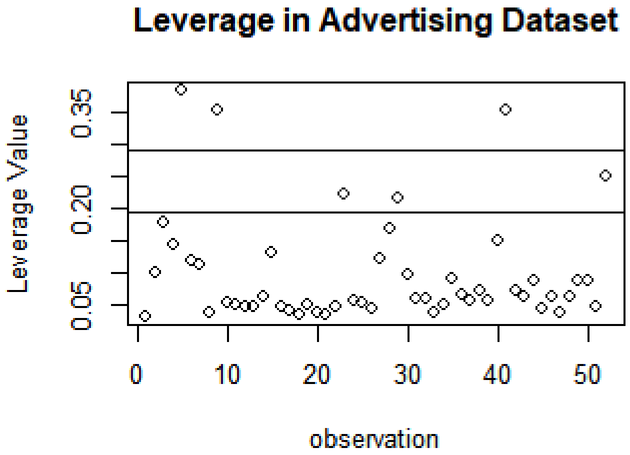
(c) Examine the Cook's distances

```
improvemodel<-lm(Y~X2+I(X2^2)+X4+X5,data=workhours)
plot(improvemodel)
```

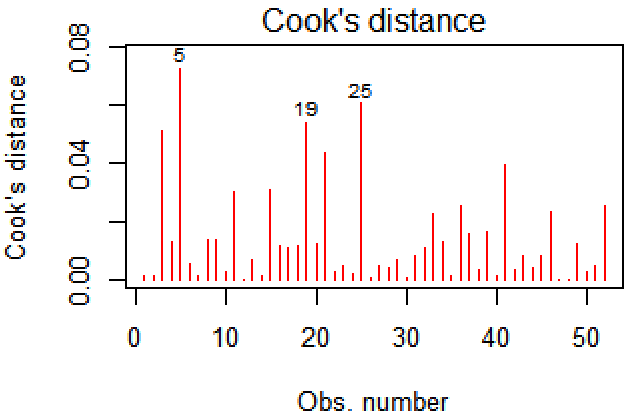





```
lev=hatvalues(improvemodel)
p = length(coef(improvemodel))
n = nrow(workhours)
plot(rownames(workhours),lev, main = "Leverage in Advertising Dataset", xlab="observation",
     ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
abline(h = 3 *p/n, lty = 1)
```




```
> plot(improvemodel,pch=18,col="red",which=c(4))
```



PROBLEM 22

Use the EXECSAL2.CSV data.

Use this model:

```
bestmodel <- lm(log(Y) ~ X1 + I(X1^2) + X2 + factor(X3) + X4 + X5 + factor(X3)*X4)
```

Check the following the assumptions:

(a) Linearity – use a plot

(b) Normality – use a plot and a test

(c) Heteroscedasticity – use a plot and a test

(d) Multicollinearity – use a test

(e) Outliers – use plots involving Cook’s Distance and Leverage

ANSWER TO PROBLEM 22

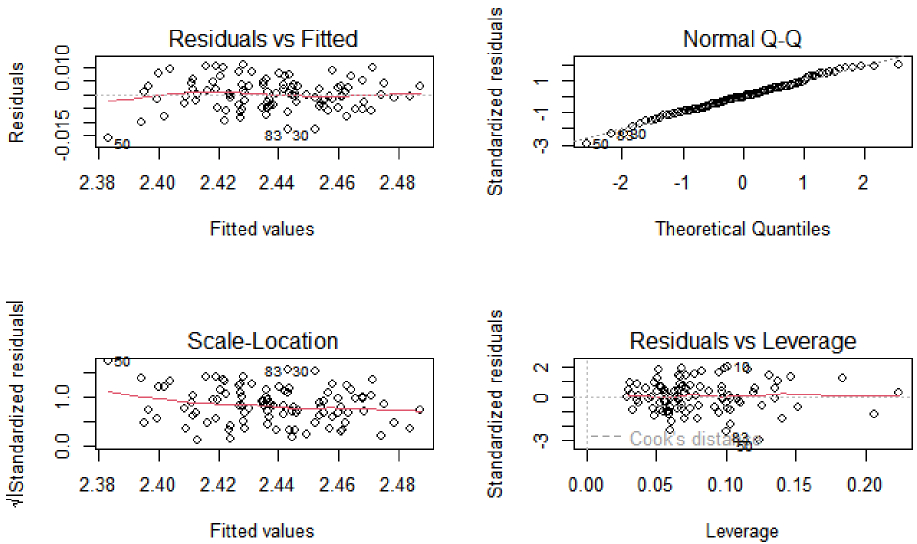
(a) Linearity – use a plot

(b) Normality – use a plot and a test

(c) Heteroscedasticity – use a plot and a test

(d) Multicollinearity – use a test

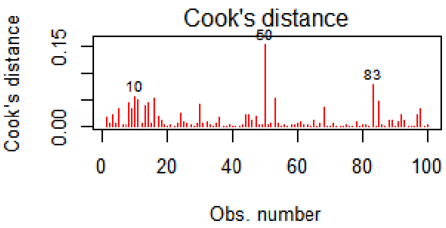
(e) Outliers – use plots involving Cook's Distance and Leverage

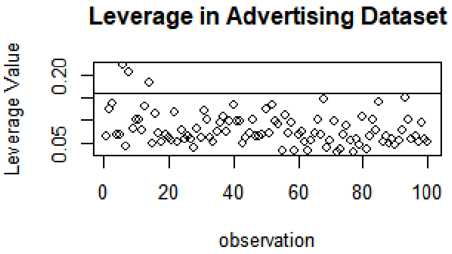



```
        studentized Breusch-Pagan test

data:  bestmodel
BP = 19.667, df = 7, p-value = 0.006336
```

```
        Shapiro-Wilk normality test

data:  residuals(bestmodel)
W = 0.98893, p-value = 0.579
```

```
> firstordermodel<-lm(Y~X2+X4+X5,data=workhours)
> vif(firstordermodel)
      X2       X4       X5 
1.002657 1.246889 1.245509 
```






```
# Leverage Point
lev=hatvalues(bestmodel)
p = length(coef(bestmodel))
n = nrow(salary)
outlier = lev[lev>(2*p/n)]
print(outlier)
        6         8        14 
0.2242153 0.2065964 0.1832883 
```