HW3 - 250620601

Ravin Lathigra

November 2, 2018

R Packages & Libraries

```
library(corrplot)    #Visualize Correlation between variables
library(kableExtra)  #Style tables
library(tidyverse)   #contains ggplot2,dplyr,tidyr, readr,purr,tibble,stringr,forcats
library(formatR)     #Improve readability of code
library(e1071)       #Functions for latent class analysis, Fourier transform ect.
library(VIM)         #Knn
library(ggfortify)   #Add on to ggplot2 to allow for more plot types
library(Rtsne)       #Dimension reduction classification
library(caret)       #streamlined model development
library(RColorBrewer)#Control colours of visualizations
library(GGally)      #Contains ggpairs plots
library(lmtest)
```

Question 1a

	Predicted Value
Question 1a	19.95

If `weight` = 3 and `cylinder` = 6, the fitted value for `mpg` is **19.95**.

Question 1b

```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8544 -1.7440 -0.4468  1.2646  6.6174
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.8988     2.3390  14.065  3.7e-12 ***
## wt           -3.0606     0.9136  -3.350  0.00303 **
## cyl6         -3.7623     1.7639  -2.133  0.04490 *
## cyl8         -5.4415     1.8085  -3.009  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.579 on 21 degrees of freedom
## Multiple R-squared:  0.8129, Adjusted R-squared:  0.7862
## F-statistic: 30.42 on 3 and 21 DF,  p-value: 7.818e-08
```

Considering a signficance level

$$\alpha = 0.05$$

it can be shown that if `weight` is considered, `cylinder` **is an important** predictor as the p-value is less that 0.05.

Question 1c

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + wt:cyl, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6507 -1.1242 -0.5088  1.4086  5.2918
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.6787     3.7624  10.280 3.37e-09 ***
## wt           -5.4880     1.5419  -3.559  0.00209 **
## cyl6         -4.3800    16.9168  -0.259  0.79849
## cyl8        -16.2269     5.7241  -2.835  0.01059 *
## wt:cyl6       0.8649     5.2116   0.166  0.86995
## wt:cyl8       3.7042     1.8856   1.964  0.06427 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.466 on 19 degrees of freedom
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8045
## F-statistic: 20.75 on 5 and 19 DF,  p-value: 4.241e-07
```

	Predicted Value
Question 1c	17.1

Considering `weight` and `cylinder` as predictors as well as their interaction, the fitted value assuming `weight` = 3 and `cylinder` = 8 is **17.1**

Question 1d

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + wt:cyl, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6507 -1.1242 -0.5088  1.4086  5.2918
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.6787     3.7624  10.280 3.37e-09 ***
## wt           -5.4880     1.5419  -3.559  0.00209 **
## cyl6         -4.3800    16.9168  -0.259  0.79849
## cyl8        -16.2269     5.7241  -2.835  0.01059 *
## wt:cyl6       0.8649     5.2116   0.166  0.86995
## wt:cyl8       3.7042     1.8856   1.964  0.06427 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.466 on 19 degrees of freedom
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8045
## F-statistic: 20.75 on 5 and 19 DF,  p-value: 4.241e-07
```

Considering the following Null and Alternative Hypothesis:

Null: There is no significant interaction effect between two predictors.

Alt : There is a significant interaction effect between two predictors.

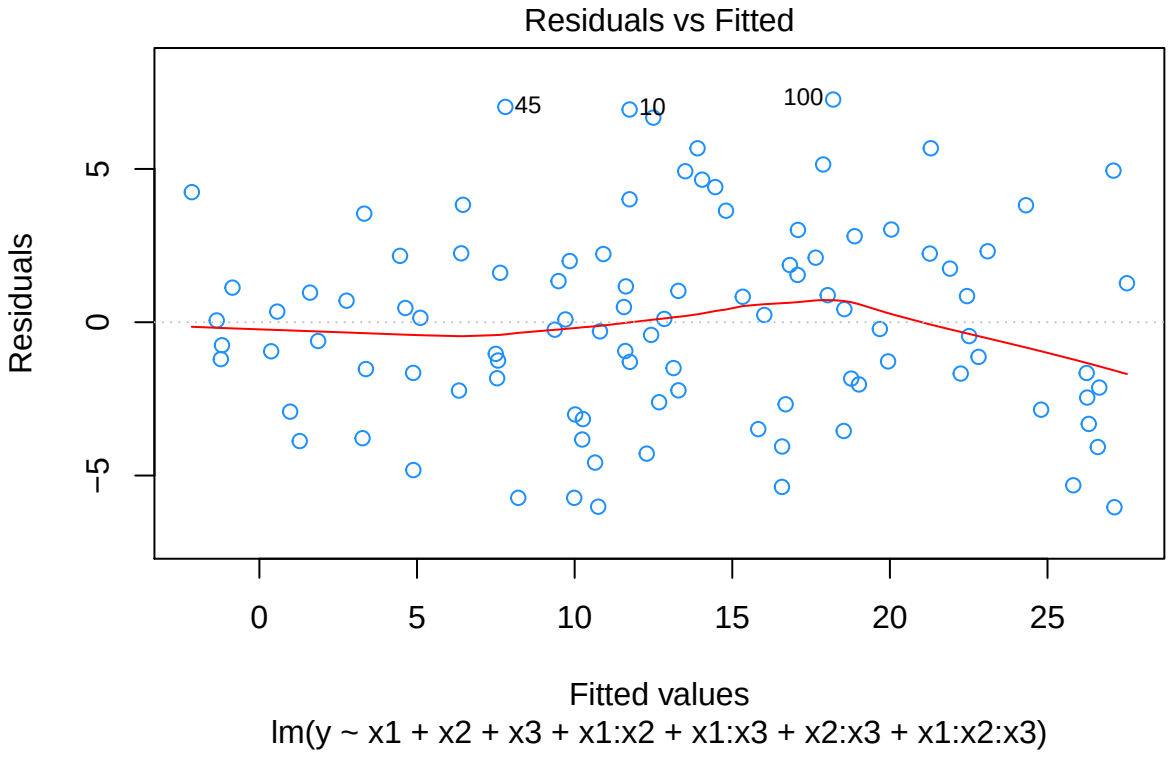
At a significance level of 0.05, there is no significant evidence that suggests the interaction between `weight` and `cylinder` aids in modelling shown by p-values greater than the significance level.

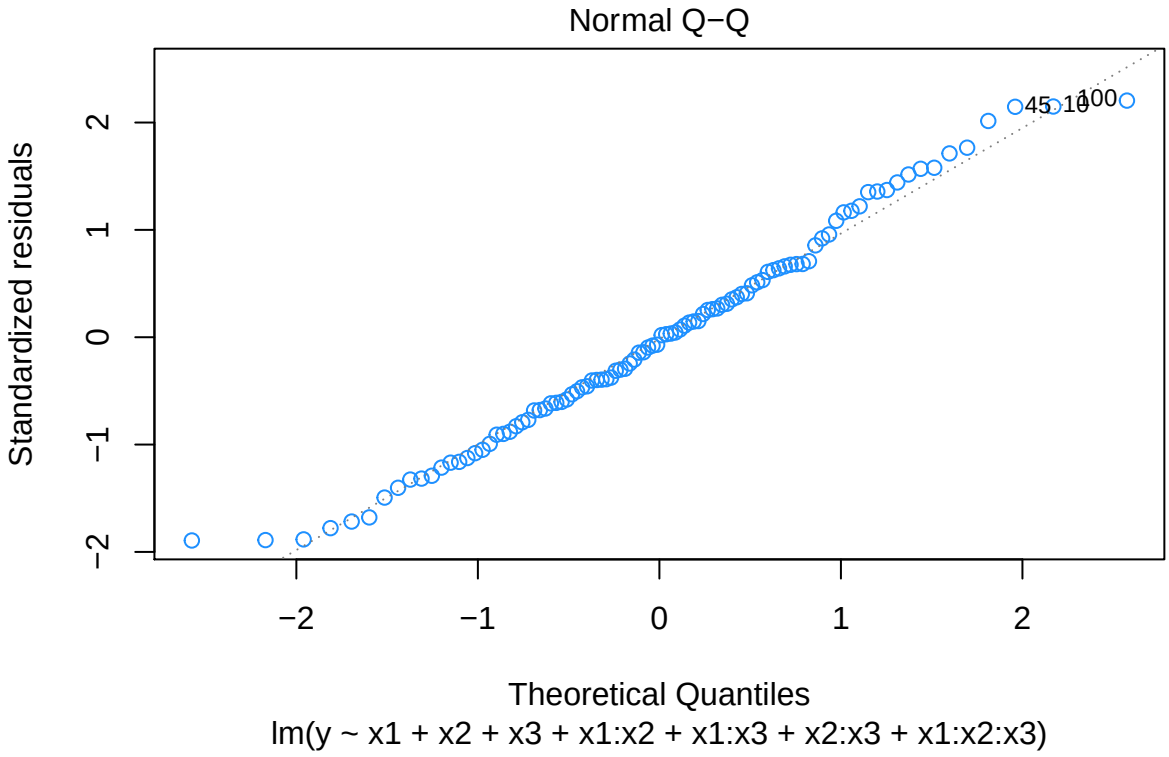
Question 2a

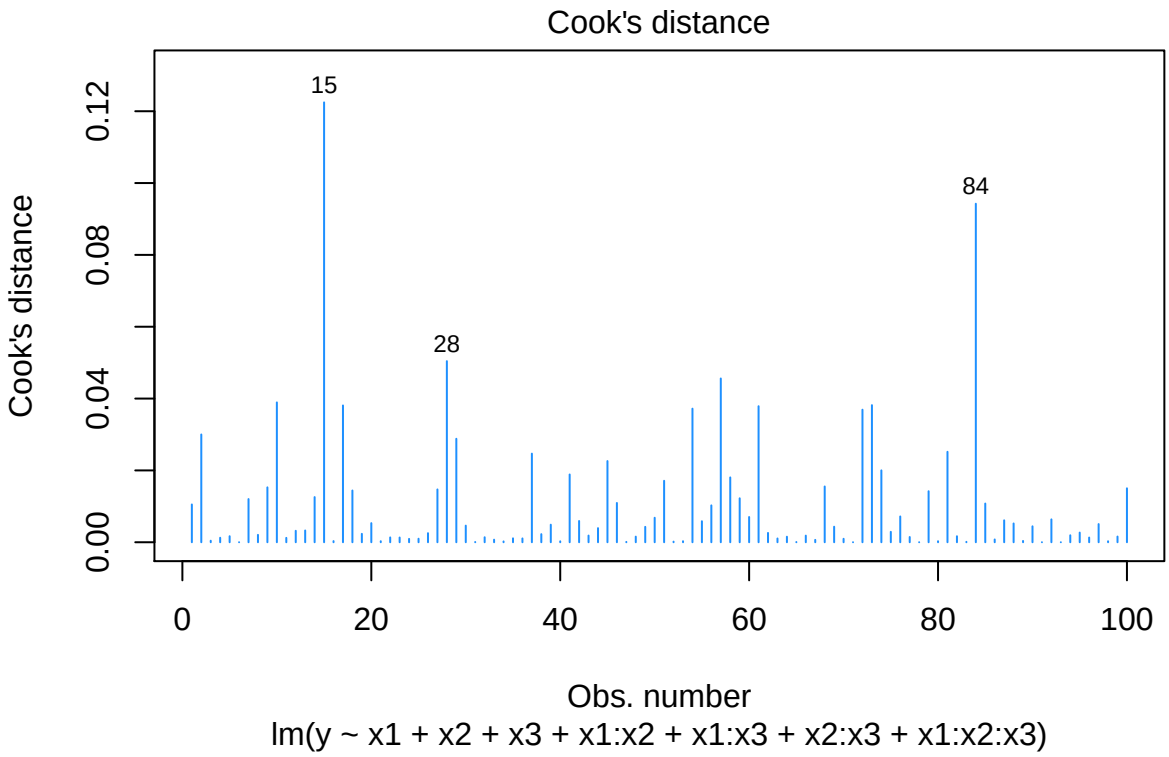
	Increase in mean
A	4

Table 1.0 shows that given `x2` = 50 and `x3` = 7, **one unit increase** in `x1` increases the estimated mean by 4 units i.e `A` = 4.

Question 2b


Residuals vs Fitted
45
10
100
Residuals
5
0
−5
0
5
10
15
20
25
Fitted values
lm(y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3)


Normal Q−Q
45
10
100
Standardized residuals
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3)



The above plots display the following:

- Resiudals vs Fitted Values
- Normal qq plot
- Cook's Distance

Linear Model Appropriateness:

Linearity - Inspecting the plot "Residuals Vs Fitted" we see that the residuals are randomly scattered around the zero line indicating that a linear model may be an appropriate model. Labeled on the plot are 3 observations with large observations that can be investigated if desired.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see that at any subset of the fitted values, there is a constant variance as there is no defining trends. There is nothing to suuggest a linear model is not appropriate considering the variance of the residuals alone.

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals closely correspond to the theoretical quantiles of a normal distribution suggesting that the residuals are approximately normally distributed. The plot identifies 3 points that have the largest residuals.

Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation.

Visualization suggests that a linear model is appropriate.

Question 2c

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  lm_2
## BP = 6.4252, df = 7, p-value = 0.4911
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_2)
## W = 0.98441, p-value = 0.2875
```

To assess if equal variance assumption holds, we use the **BP test** and to tests to see if the normal assumption holds we use the **Shapiro test**. A significance level of 5% was used.

BP Test: p-value $> 5\%$ **No evidence against equal variance**

Shapiro Test: p-value $> 5\%$ **No evidence against normality**

Question 2d

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3,
##     data = data_q2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.034 -2.224 -0.081  2.121  7.264
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.327393   3.559242   2.059   0.0424 *
## x1           1.709184   1.251519   1.366   0.1754
## x2          -0.166497   0.059186  -2.813   0.0060 **
## x3           0.561826   0.312254   1.799   0.0753 .
## x1:x2        0.038134   0.020579   1.853   0.0671 .
## x1:x3        0.121700   0.110824   1.098   0.2750
## x2:x3       -0.003239   0.005007  -0.647   0.5193
## x1:x2:x3    -0.001350   0.001735  -0.778   0.4385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.336 on 92 degrees of freedom
## Multiple R-squared:  0.8574, Adjusted R-squared:  0.8466
## F-statistic: 79.04 on 7 and 92 DF,  p-value: < 2.2e-16
```

Summary of the linear model with 3 way interaction shows that the 3 way interaction is insignificant.

Question 2e

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     96 1240.8
## 2     92 1023.6  4    217.16 4.8795 0.001297 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

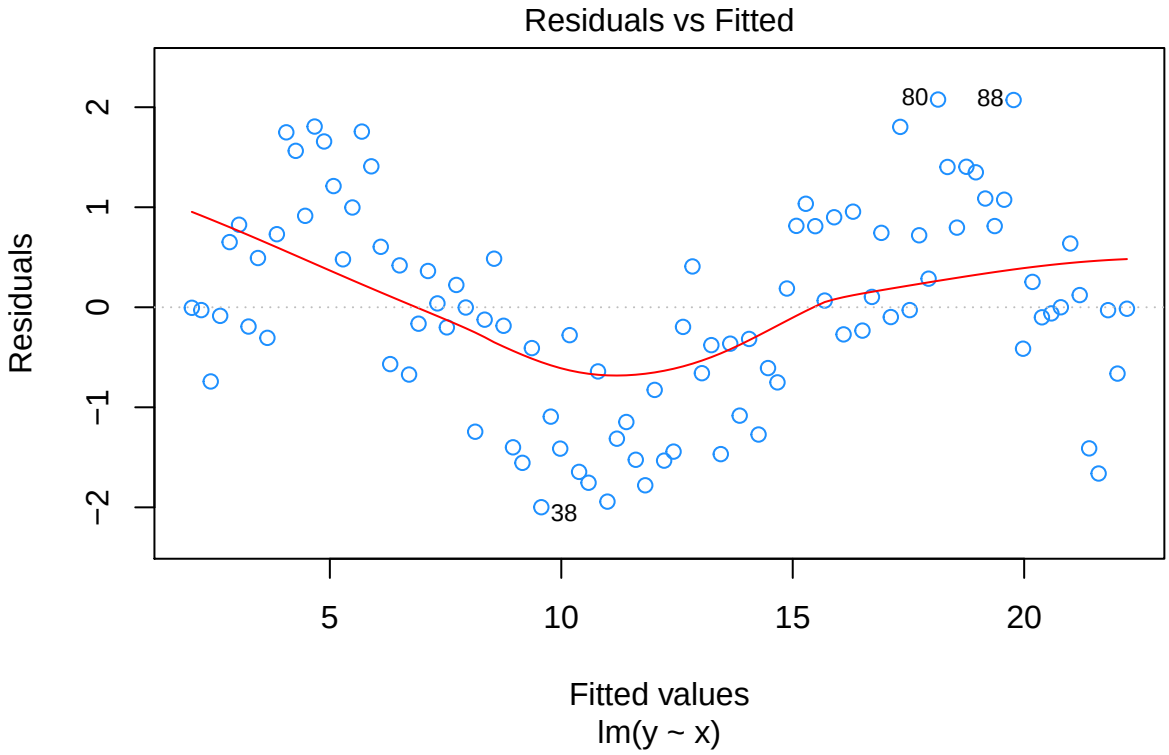
Considering the following null and alternative hypothesis and a 5% significance level

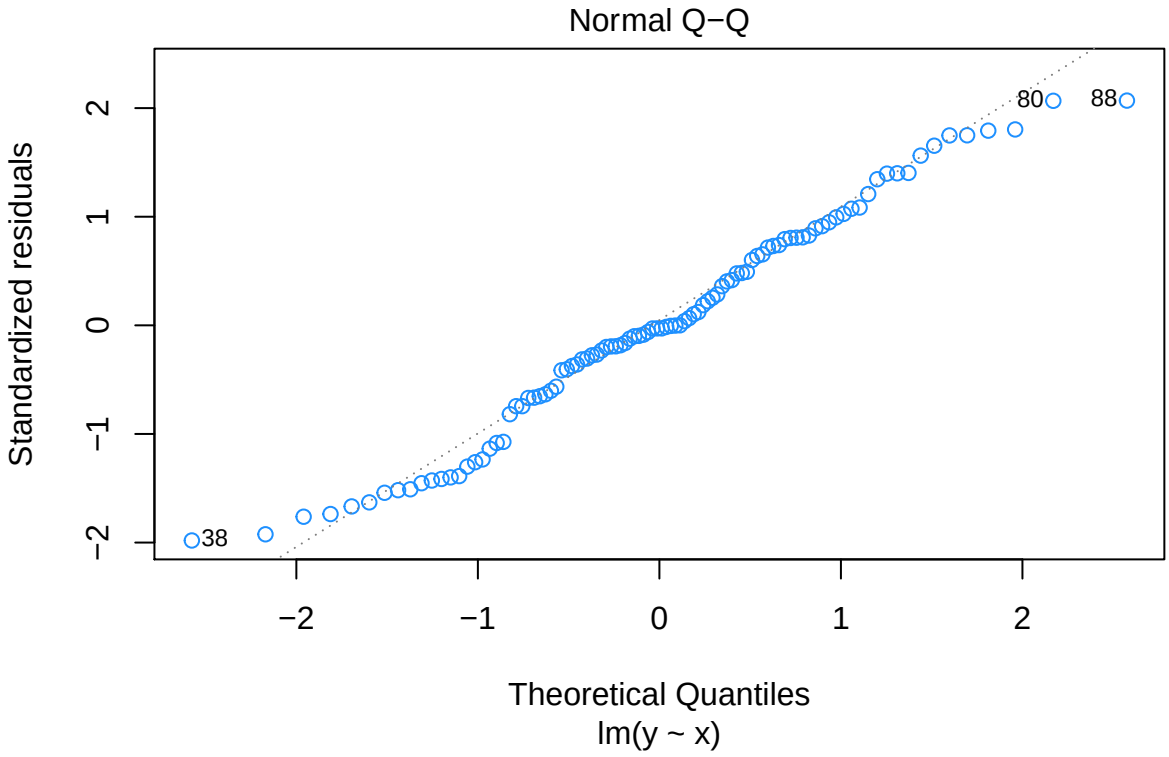
Null : B4 = B5 = B6 = B7 = 0

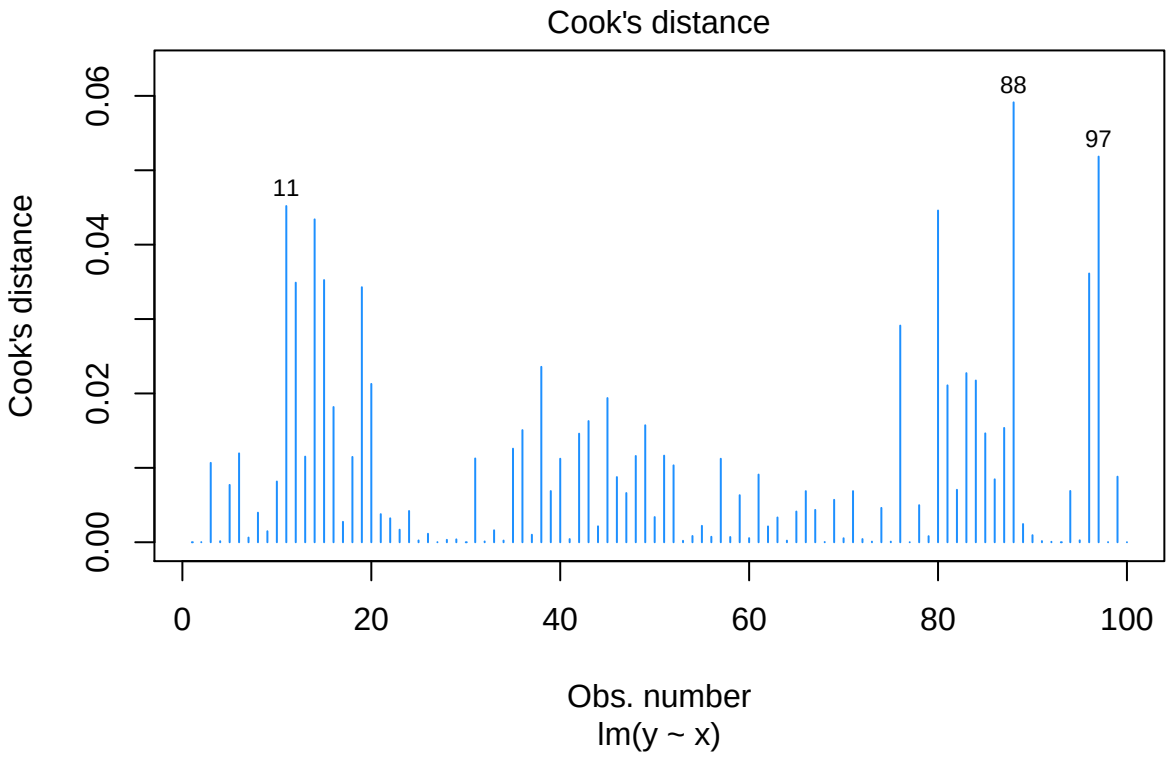
Alt : At least one of B4,B5,B6,B7 is non-zero.

Using anova to assess the importance of the interaction terms (B4:B7) shows support against the null hypothesis.

Question 3


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x)
80
88
38
5
10
15
20
2
1
0
−1
−2


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
80
88
38
−2
−1
0
1
2




```
##
##  studentized Breusch-Pagan test
##
## data:  lm3
## BP = 0.0090726, df = 1, p-value = 0.9241
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm3)
## W = 0.97905, p-value = 0.1121
```

The above plots display the following:

- Resiudals vs Fitted Values
- Normal qq plot
- Cook's Distance

Linear Model Appropriateness:

Linearity - Inspecting the plot "Residuals Vs Fitted" has a trend line that helps illustrate that there is a slight parabolic relationship between fitted values and residuals. Furthermore, the residuals do not exhibit zero mean suggesting that a linear model may not be the most appropriate model and perhaps transformations should be considered.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see that at any subset of the fitted values, there is a constant variance.

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals moderately correspond to the theoretical quantiles of a normal distribution. To properly assess if the normality assumption

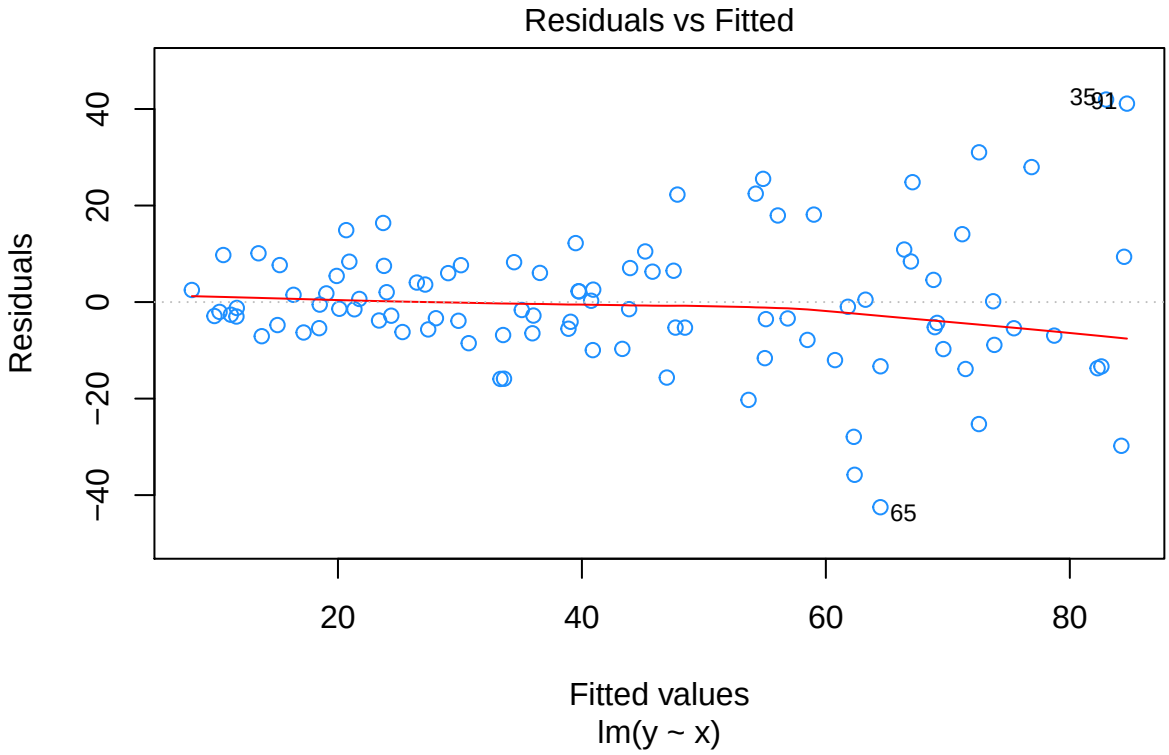
is violated the Shapiro test will be carried out. Additionally, the plot identifies 3 points that have the largest residuals.

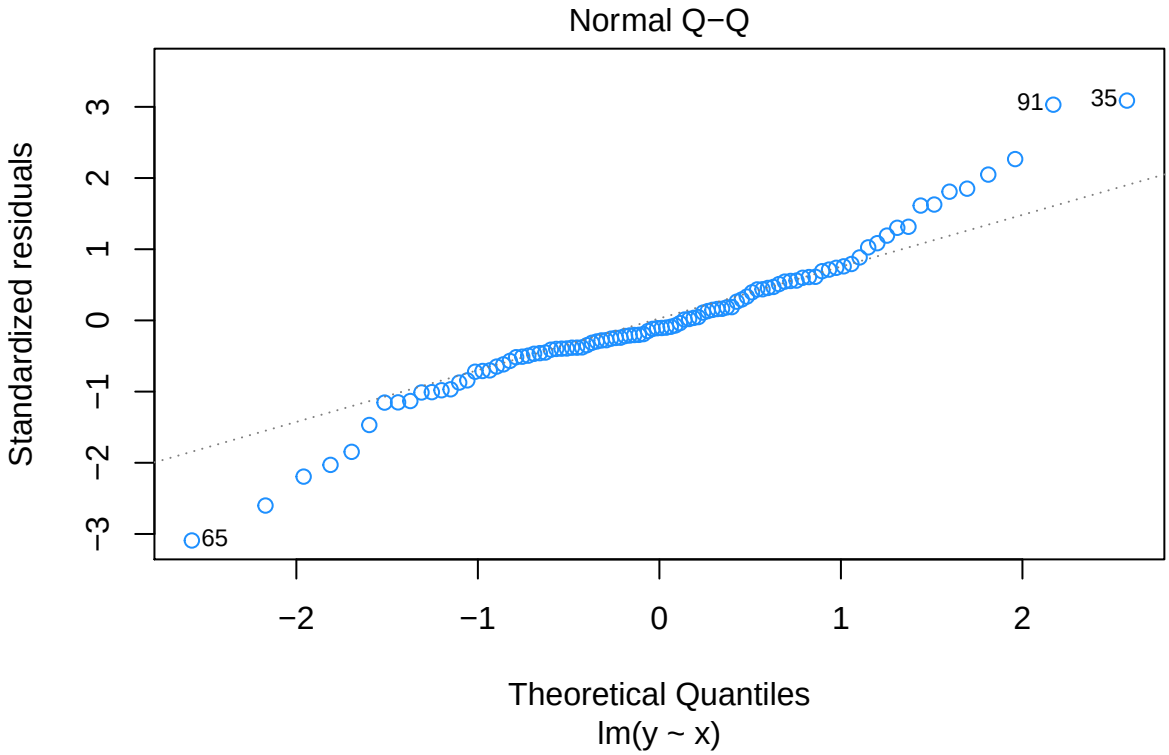
Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may high influence or require further investigation.

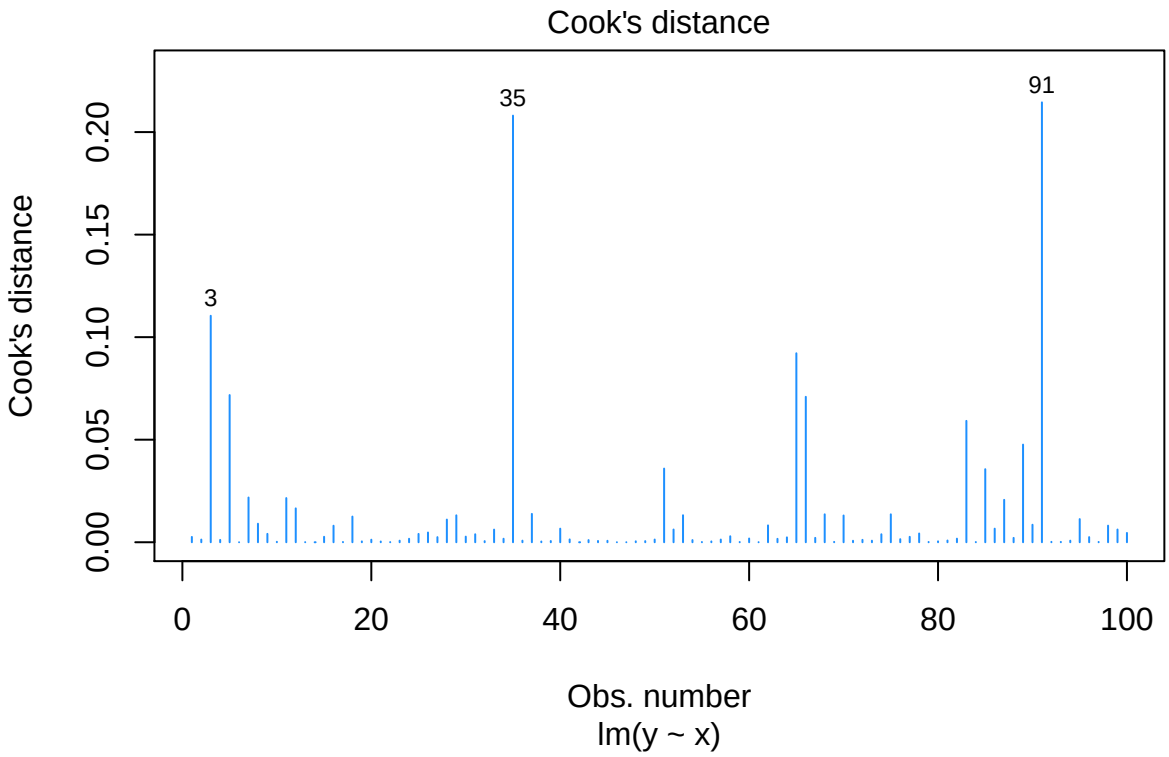
BP Test: p-value $>$ 5% significance level this suggests that the equal variance assumptions holds for this model.

Shapiro Test: p-value $>$ 5% significance level this suggests that the equal variance assumptions holds for this model, though it could be argued that there is margnial support that it is violated as the p-value is quite low.

Question 4


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x)
35
91
65
40
20
0
−20
−40
20
40
60
80


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
91
35
65
3
2
1
0
−1
−2
−3
−2
−1
0
1
2




```
##
##  studentized Breusch-Pagan test
##
## data:  lm4
## BP = 22.542, df = 1, p-value = 2.056e-06
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm4)
## W = 0.95913, p-value = 0.003487
```

The above plots display the following:

- Resiudals vs Fitted Values
- Normal qq plot
- Cook's Distance

Linear Model Appropriateness:

Linearity - Inspecting the plot "Residuals Vs Fitted" has residuals seemingly randomly distributed about the zero line suggesting that the linearity assumption holds.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see that at any subset of the fitted values, there is **not constant** variance. Residuals seem to diverge as the fitted values increase perhaps indicative of heteroskedasitiy

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals closely correspond to the theoretical quantiles at points particularly the middle however there is indication at end

points that show normality may be violated. To properly assess if the normality assumption is violated the Shapiro test will be carried out. Additionally, the plot identifies 3 points that have the largest residuals.

Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have influence or require further investigation.

BP Test: p-value $< 5\%$ significance level this suggests that the equal variance assumptions is violated.

Shapiro Test: p-value $< 5\%$ significance level this suggests that the normality assumption is violated.

Question 5a

Points with high leverage
C
E

The table above shows which observations have high leverage. Points **C** and **E** exceed 2*p/n (0.4) therefore are considered to have high leverage.

Question 5b

If Yb change to 50 the leverage of the observation would be unchanged as it is calculated independently of the response.

Question 5c

ID	x	y	diag_h	resid.lm	Standardized Residual
B	23	120	0.13	53.29	2.0263427
C	5	20	0.47	-12.55	-0.6114101
E	35	50	0.55	-39.49	-2.0878903

To calculate the standardize residuals we do the following:

$$r_i = \frac{e_i}{\sqrt{(1-h_{ii})\hat{\sigma}^2}}$$

. Note, we use sigma hat as the true variance is unknow so it needs to be estimated.

$$\hat{\sigma}^2 = \frac{e}{n-p}$$

. The table above shows the calculated Standardized Residuals for points **B**, **C**, and **E**.

Question 5d

ID	x	y	diag_h	resid.lm	Standardized Residual	Cooks Distance
B	23	120	0.13	53.29	2.0263427	0.3067750
C	5	20	0.47	-12.55	-0.6114101	0.1657514
E	35	50	0.55	-39.49	-2.0878903	2.6640082

To calculate the Cooks Distance we do the following:

$$CooksDistance = \frac{resid^2_{std}(h_{ii})}{p(1-h_{ii})}$$

. The table above shows the calculated Cooks Distances for points **B**, **C**, and **E**.

Point E exceeds the criteria of 4/n (.4) therefore it is a point of high influence. ## Question 6

Question A - Point B is the observation furthest from the regression line therefore it has the largest absolute residual.

Question B - Point E is the observation that has the largest leverage because it is the furthest difference from the mean of x.

Question C - Point D is the observation that has the potential to have the largest influence because it has high leverage and moderately large absolute residual.