STAT 3032 - Homework 3

Saksham Goel

March 1, 2018

Problem 1 - Airfare Dataset

Part A - Linear Model Fitting and Interpretting the Intercept

The following segment of R Code is used to load the data through the file - "airfare.txt" and then fit a linear model.

```r
file = read.table("airfares.txt", header=T)
xCol = file$Distance
yCol = file$Fare
mod = lm(yCol ~ xCol)
mod
```

```
##
## Call:
## lm(formula = yCol ~ xCol)
##
## Coefficients:
## (Intercept)         xCol
##     48.9718       0.2197
```

The fitted model looks like -

$$\hat{y} = 48.9718 + 0.2197 \cdot x_i$$

$$Fare = 48.9718 + 0.2197 \cdot Distance$$

This means that the values of β_0 and β_1 are given as follows:

$$\hat{\beta}_0 = 48.9718$$

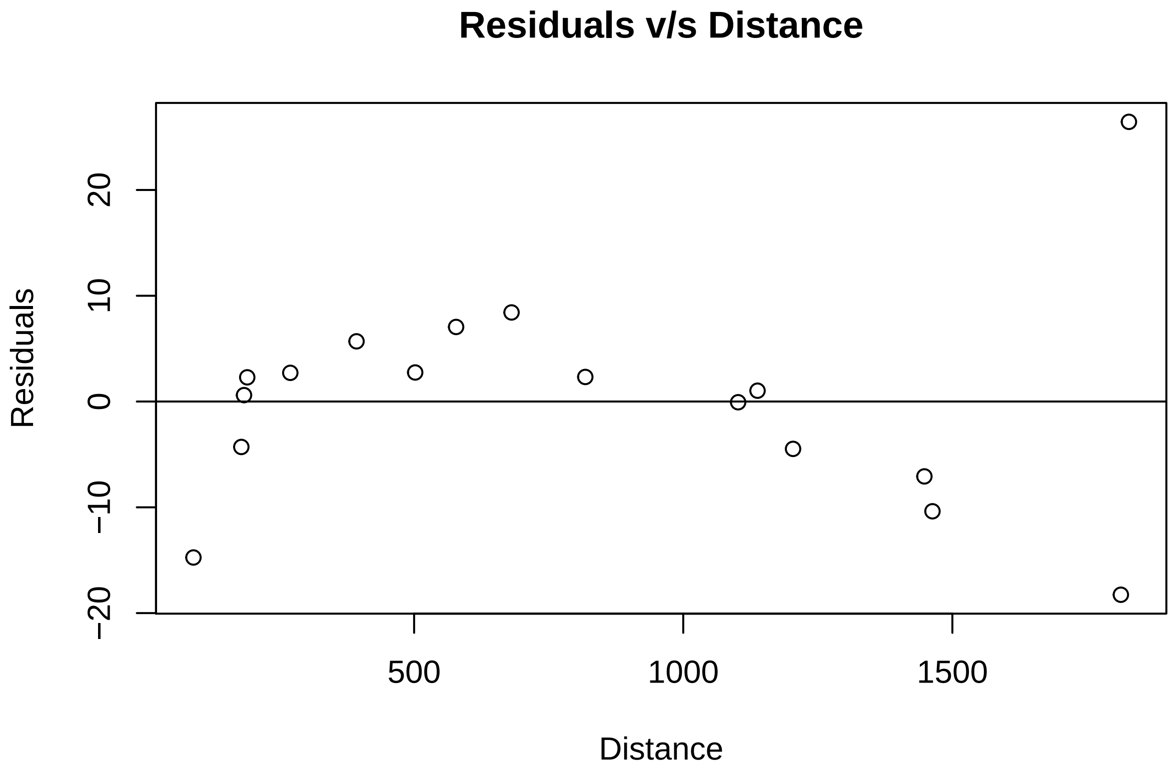
$$\hat{\beta}_1 = 0.2197$$

Using the model we can interpret the meaning of the intercept as the base price of a flight. By base price, I mean what a person has to pay just for being able to board a flight/for a flight that just travels 0 miles.

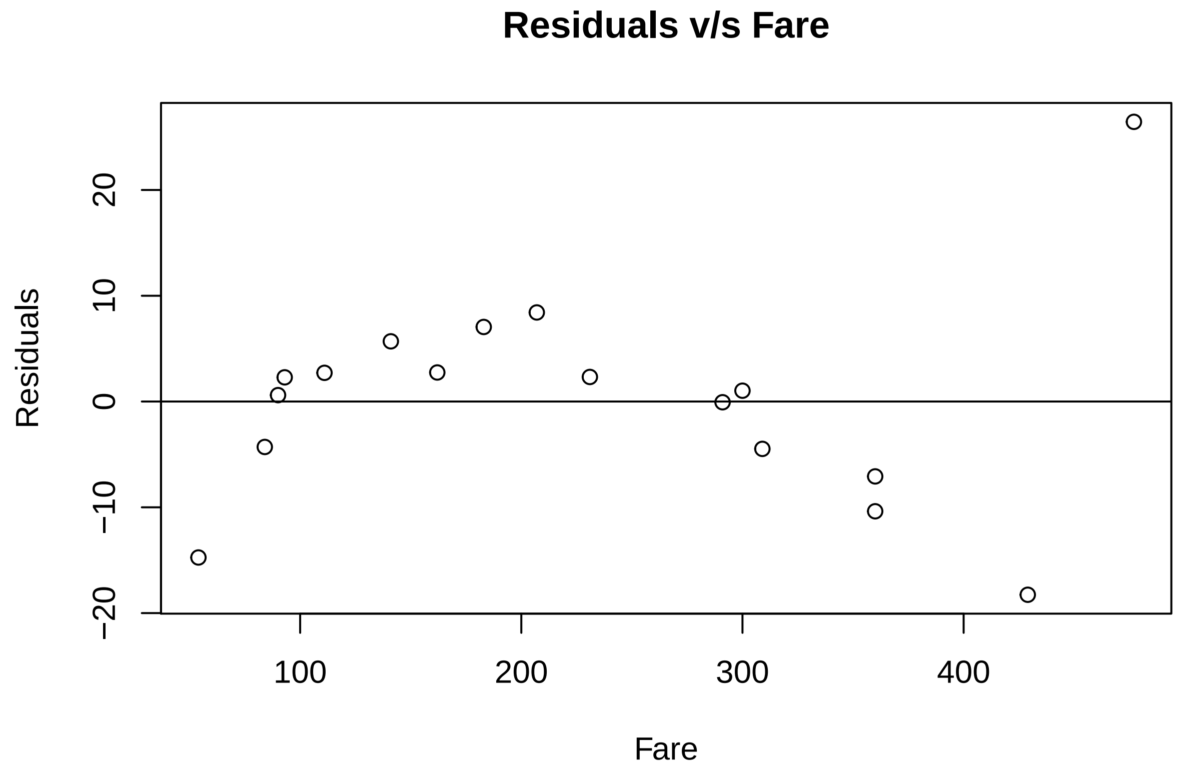
Part B - Residuals Plots

The following segment of R Code is used to create scatter plots for the residuals of the dataset fitted in the linear model.

```r
res = resid(mod)
plot(y = res, x = xCol, ylab = "Residuals", xlab = "Distance", main = "Residuals v/s Distance")
abline(0, 0)
```



```
plot(y = res, x = yCol, ylab = "Residuals", xlab = "Fare", main = "Residuals v/s Fare")
abline(0, 0)
```



The above two residual plots depict scatter plots of Residuals v/s Distance and Residuals v/s Fare.

Similarities:

1. Both of these plots have the same range of Y values (Which is expected because the residual value doesnt change)
2. Both of these plots have the same overall distribution of the Residuals. There is not much difference in the positions of the points.

Different:

1. The range of values on the X Axis is really different (which is expected because the values plotted in the x axis are different, i.e. range of Distance is different from range of Fare)

Part C - Calculate Leverages

The following segment of R Code is used to calculate leverage values of all the individual data points.

```
leverages = hatvalues(mod)
leverages
```

```
##          1          2          3          4          5          6
## 0.13416776 0.13071191 0.06213499 0.11267276 0.12959128 0.09116201
##          7          8          9         10         11         12
## 0.07351534 0.06908091 0.08588990 0.07745448 0.13095319 0.07665861
##         13         14         15         16         17
## 0.24326517 0.13209804 0.05882392 0.15398450 0.23783524
```

The largest leverage values is $h_{ii} = 0.24326517$. This leverage value corresponds to observation number 13.

According to the textbook the threshold of the leverage value is given by $h_{ii} < 4/n$. Here n = 17, hence $h_{ii} < 4/17 = 0.235294$. Now as the leverage of obsrvation number 13 (h_{ii} = 0.24326517) is greater then the threshold (0.235294) the observation is a leverage point.

Part D - Calculate Standardized Residual

The following segment of R Code is used to calculate standard residual values of all the individual data points.

```
sResiduals = rstandard(mod)
sResiduals
```

```
##            1            2            3            4            5 
## -1.070782935 -0.729205219  0.835139842  0.276572788  0.235495724 
##            6            7            8            9           10 
##  0.573338687 -0.006706541  0.701659153 -0.449554976  0.102395666 
##           11           12           13           14           15 
##  0.062407480  0.274377543  2.919064802 -0.442860375  0.230068107 
##           16           17 
## -1.539476397 -2.009328033
```

The largest standardized residual value is $\hat{r}_i = 2.919064802$. This standardized residual value corresponds to observation number 13.

According to the textbook the threshold of the standardized residual value is given by $-2 < \hat{r}_i < 2$ for small datasets, which is true for this case as n = 17. Now as the standardized residual of obsrvation number 13 ($\hat{r}_i$ = 2.919064802) is greater then 2 the observation is an outlier point.

Part E - Calculate Cooks Distance

The following segment of R Code is used to calculate cook's distance of all the individual data points.

```
cooksDistance = cooks.distance(mod)
cooksDistance
```

```
##            1            2            3            4            5 
## 8.883565e-02 3.997799e-02 2.310385e-02 4.856507e-03 4.128465e-03 
##            6            7            8            9           10 
## 1.648618e-02 1.784460e-06 1.826705e-02 9.494655e-03 4.401410e-04 
##           11           12           13           14           15 
## 2.934379e-04 3.125113e-03 1.369600e+00 1.492552e-02 1.654116e-03 
##           16           17 
## 2.156824e-01 6.299398e-01
```

The largest cook's distance is $D_i = 1.3696$. This cook's distance corresponds to observation number 13.

According to the textbook the threshold of the cook's distance is given by $D_i < 4/(n-2)$. Here n = 17, hence threshold is given by $D_i < 4/15 = 0.26666$. Now as the cook's distance of obsrvation number 13 ($D_i = 1.3696$) is greater then the threshold the observation is a bad point (outlier and leverage). This conclusion aligns with our earlier observations as well. Also the value for cook's distance is rather big, in magnitude of 1, which clearly states that the point is a bad point (has very much effect on the regression line).

Problem 2 - Calculating Leverages and Cooks Distance

Part A - Estimate Leverage

I think that Observation 11 has higher leverage because the value of x_i for observation 11 is much far from the mean value ($\bar{x}$) as compared to the value of x_i for observation 69.

Part B - Calculate Leverage

The formula for leverage is given by:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}$$

Here we have:

n = 100,

$\bar{x} = 20$,

$S_{XX} = 400$

so plugging the values for Observation 11, $x_i = 30.5$, we get:

```r
n = 100
xbar = 20
sxx = 400
xi = 30.5
hii = (1/n) + (((xi-xbar)^2)/sxx)
hii
```

```
## [1] 0.285625
```

$$h_{11} = 0.285625$$

similarly plugging the values for Observation 69, $x_i = 26.5$, we get:

```r
n = 100
xbar = 20
sxx = 400
xi = 26.5
hii = (1/n) + (((xi-xbar)^2)/sxx)
hii
```

```
## [1] 0.115625
```

$$h_{69} = 0.115625$$

The results found in part b) align with what was estimated in part a).

Part C - Calculate Residuals

The formula for Residuals is given by:

$$\hat{e}_i = y_i - \hat{y}_i$$

Finding the residuals for the two observations as follows:

```r
y11 = 5
yhat11 = 5.3
y69 = 4.5
```

```
yhat69 = 4.3
residual11 = y11-yhat11
residual69 = y69-yhat69
residual11
```

```
## [1] -0.3
```

```
residual69
```

```
## [1] 0.2
```

The two residuals are as follows

$$\hat{e_{11}} = -0.3$$

$$\hat{e_{69}} = 0.2$$

Part D - Calculate Standard Residuals

The formula for Standard Residuals is given by:

$$\hat{r}_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Finding the standard residuals for the two observations as follows:

```
e11 = -0.3
h11 = 0.285625
e69 = 0.2
h69 = 0.115625
RSS = 24.5
n = 100
sigma = sqrt(RSS/(n-2))
r11 = e11/(sigma*sqrt(1-h11))
r69 = e69/(sigma*sqrt(1-h69))
r11
```

```
## [1] -0.7098852
```

```
r69
```

```
## [1] 0.4253454
```

The two standard residuals are as follows:

$$\hat{r}_{11} = -0.7098852$$

$$\hat{r}_{69} = 0.4253454$$

Part E - Calculate Cook's Distance

The formula for Cook's Distance is given by:

$$D_i = \frac{r_i^2}{2}\frac{h_{ii}}{1-h_{ii}}$$

Finding the Cook's Distance for the two observations as follows:

```
r11 = -0.7098852
h11 = 0.285625
r69 = 0.4253454
h69 = 0.115625
D11 = (r11^2)*h11/(2*(1-h11))
```

```
D69 = (r69^2)*h69/(2*(1-h69))
D11
```

```
## [1] 0.1007433
```

```
D69
```

```
## [1] 0.01182684
```

The two Cook's Distance are as follows:

$$D_{11} = 0.1007433$$

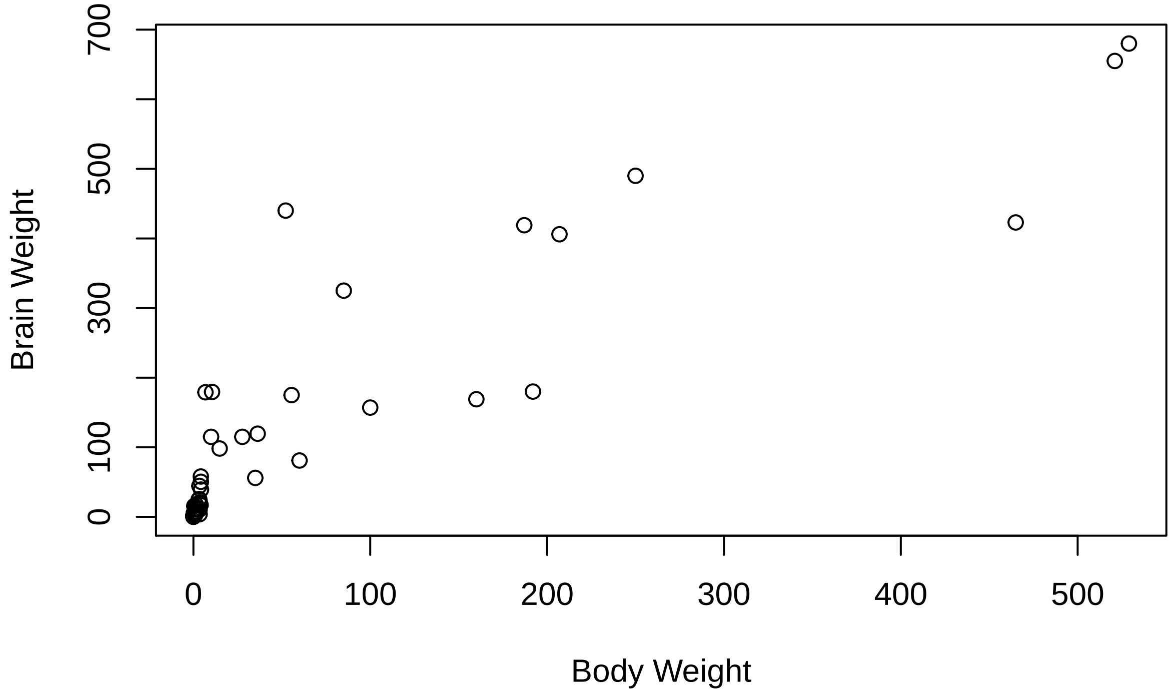
$$D_{69} = 0.01182684$$

Problem 3 - Body and Brain Weight Dataset

Part A - Scatter Plot

The following segment of R Code is used to draw a scatter plot for the dataset

```
file = read.csv("brainPartial.csv", header=T)
xCol = file$BodyWt
yCol = file$BrainWt
plot(x = xCol, y = yCol, ylab = "Brain Weight", xlab = "Body Weight")
```



After carefully observing the scatter plot it seems that the linearity assumption is violated because the points seems to follow a curvature of form squreroot or log scale. Also the linearity assumption doesnt seem to fit because there seems to multiple different slope lines that would fit different parts of the plot in a good way yet no single line that seems to fit the model in a good way.

Part B - Fitting a linear model

The following segment of R Code is used to fit a linear model

```r
mod = lm(yCol ~ xCol)
summary(mod)
```

```
##
## Call:
## lm(formula = yCol ~ xCol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -184.77  -34.52  -27.16    0.67  339.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.57276   10.95090    3.34  0.00148 **
## xCol         1.22846    0.08408   14.61  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.15 on 57 degrees of freedom
## Multiple R-squared:  0.7893, Adjusted R-squared:  0.7856
## F-statistic: 213.5 on 1 and 57 DF,  p-value: < 2.2e-16
```

If we carefully observe the summary output of the model, we will realise that the Hypothesis test for slope $= 0$ yields an extremely small p-value. Which directly shows that the slope ($\hat{\beta}_1$) cannot be removed from the model when fitting the linear model.

I am not sure whether the question about removing the slope required to actually fit a model like that, so I still did try my best to fit a model as follows:

However if I wish to remove the slope and just fit a model that has information about intercept, I would transform the x variable such that all the observations lie on the same point. To do that I would use a simple power transformation such that I raise every observation to a power of 0, which makes each observations' predictor value $= 1$. After doing this if I fit a model the expected value for the slope should be 0 and the expected value of the intercept should be the mean of all the y values. The following segment or R code will help us achive this as follows:

```r
slopezeromod = lm(yCol ~ I((xCol)^0))
summary(slopezeromod)
```

```
##
## Call:
## lm(formula = yCol ~ I((xCol)^0))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -100.19  -96.38  -84.83   16.92  579.68
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   100.33      21.69   4.625 2.14e-05 ***
## I((xCol)^0)       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 166.6 on 58 degrees of freedom
```

```
# for checking purposes finding mean of the y values
mean(yCol)
```

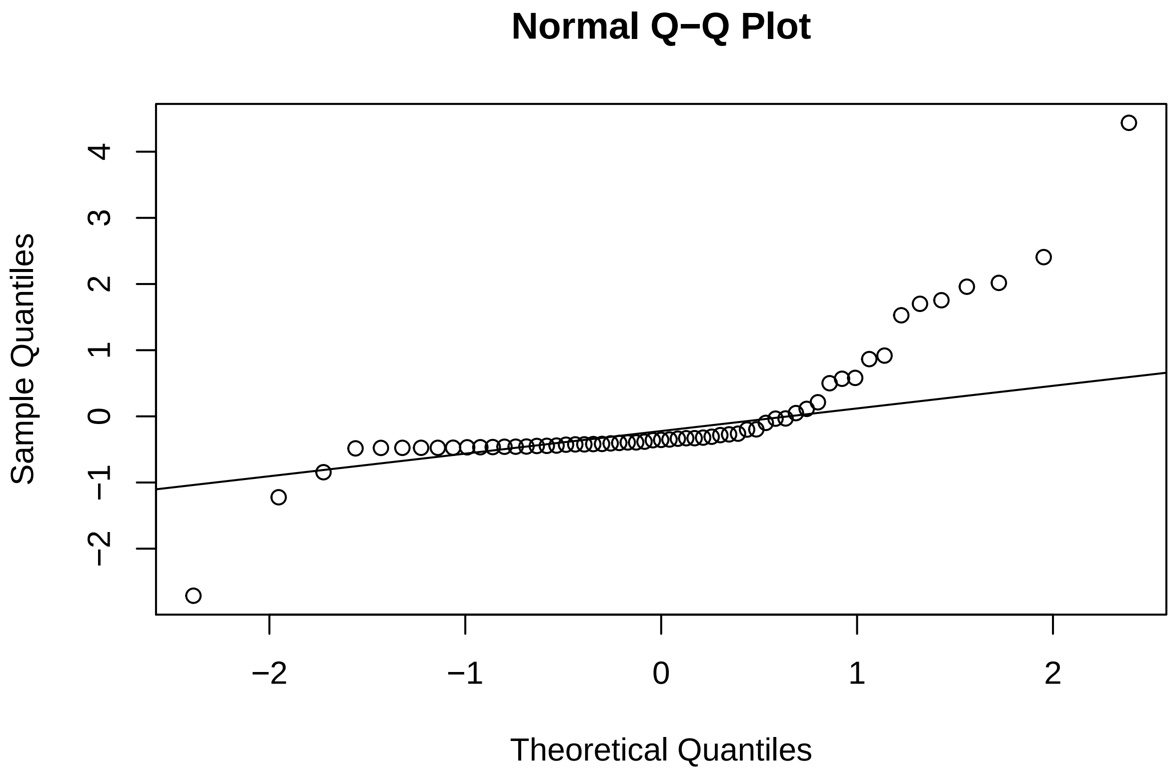
```
## [1] 100.3278
```

Observing the summary output we see that the Slope doesnt have any values, because slope doesnt matter in this linear model. Also we can observe that the value of the Intercept in the fitted model is approximately equal to the actual mean found which shows that our initial prediction was correct.

Part C - Normality Assumption

The following segment of R Code is used to generte the QQ Plot

```
qqnorm(rstandard(mod))
qqline(rstandard(mod))
```

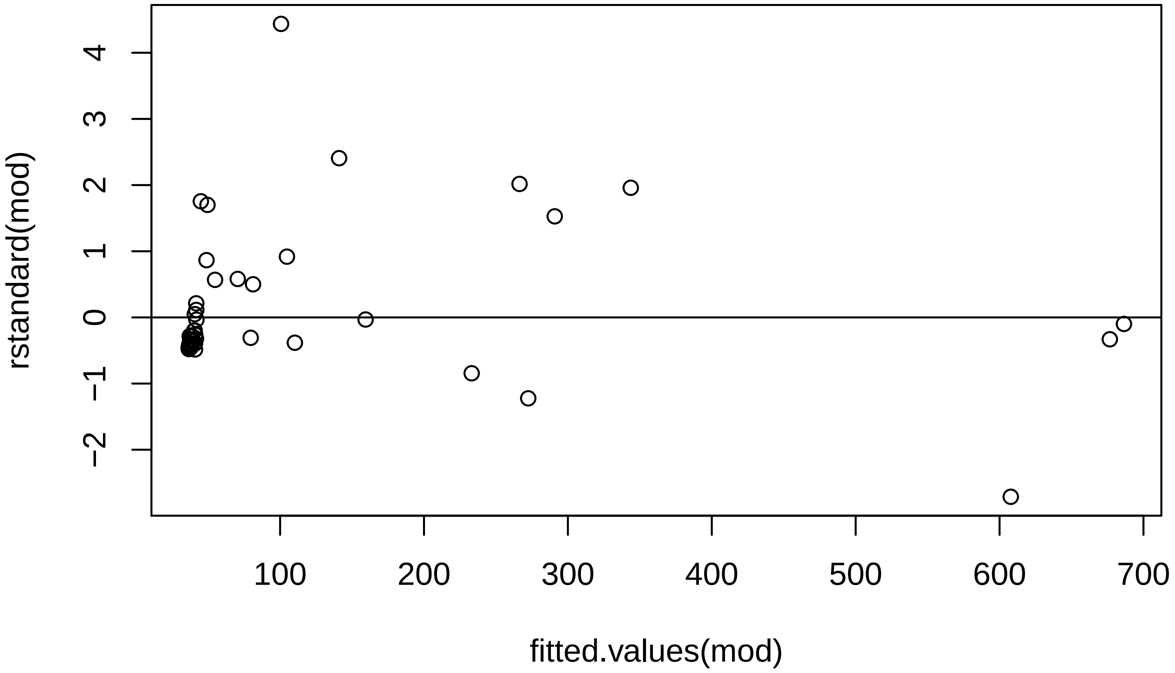


The QQ plot clearly indicates that the normality assumption has not been held. We can say that because a lot of points on the QQ plot do not lie on the normal line.

Part D - Equal Variance Assumption

The following segment of R Code is used to generate the Standard Residual Plot

```
plot(x = fitted.values(mod), y = rstandard(mod))
abline(0, 0)
```



The Standard Residual Plot clearly indicates that the equal variance assumption has not been held. We can say that because variance seems to decrease from lower values of fitted values to higher values of fitted values.

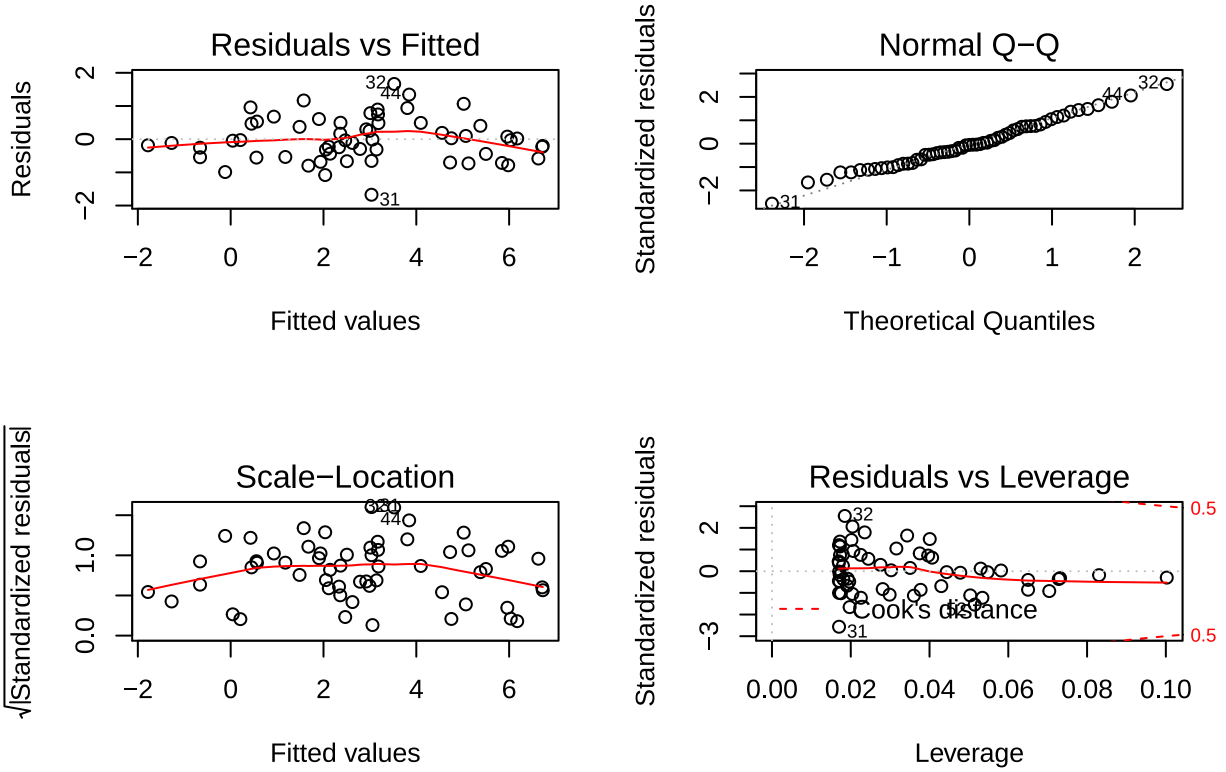
Part E - Transforming Variables

The following segment of R Code is used to fit a new linear model on transformed variables.

```
newmod = lm(log(yCol)~log(xCol))
summary(newmod)
```

```
##
## Call:
## lm(formula = log(yCol) ~ log(xCol))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67407 -0.48995 -0.03502  0.47547  1.66401
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11392    0.09138   23.13   <2e-16 ***
## log(xCol)    0.73528    0.02993   24.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6588 on 57 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9122
## F-statistic: 603.5 on 1 and 57 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(newmod)
```



After carefully observing all the Linear Model plots we see that there is no exact assumption violated. We see that in the QQ plot the values closely align to the normal line hence the normal assumption is not violated. Also in the residual plots we see that the points seems to follow no trend in their variance which depitcs that they follow Equal Variance Assumption. Also in the Resiudals vs Leverage plot we see that there is no point which is a bad point. Also if we compare the r^2 values of the older model (simple linear) and the new model (transformed model) we see that the value increased from 0.7893 to 0.9137, which is a direct evidence of much more strongly and better fitting model. Combining all the points stated we can say confidently that this linear model with transformed variables fits the dataset better.