Final Project

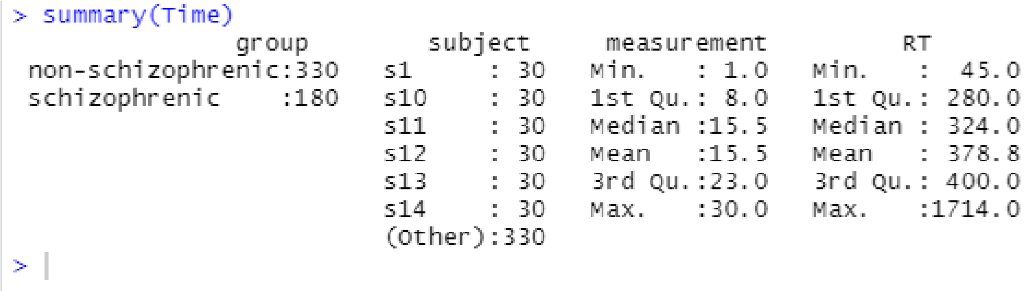
Question 1:

Dataset: reaction_times.csv

Description: The dataset contains data collected from a psychological experiment. Each of 17 people performed a cognitive task and had their reaction times measured 30 times. The experimental participants included 11 patients who were non-schizophrenics and 6 patients who were schizophrenics. Psychological theory suggests that schizophrenics suffer from an attentional deficit on some trials, as well as a general motor reflex retardation; both aspects lead to relatively slower responses for the schizophrenics.

Key question: Are there meaningful differences in reaction times between the schizophrenic and non-schizophrenic patients, adjusting for correlation between patients' repeated measurements?

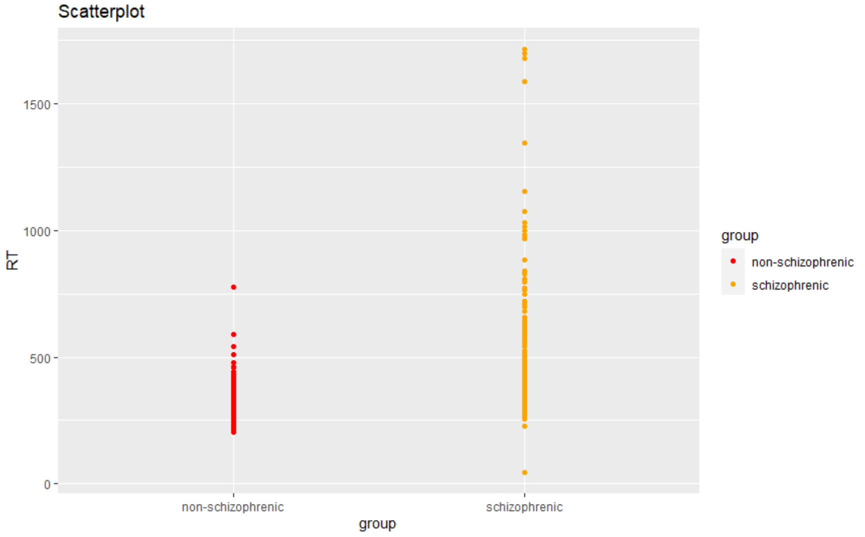
Analysis: To start off, the summary of the dataset we are using is shown below.



```
> summary(Time)
                group           subject      measurement           RT        
 non-schizophrenic:330   s1     : 30   Min.   : 1.0   Min.   :  45.0  
 schizophrenic    :180   s10    : 30   1st Qu.: 8.0   1st Qu.: 280.0  
                         s11    : 30   Median :15.5   Median : 324.0  
                         s12    : 30   Mean   :15.5   Mean   : 378.8  
                         s13    : 30   3rd Qu.:23.0   3rd Qu.: 400.0  
                         s14    : 30   Max.   :30.0   Max.   :1714.0  
                         (Other):330                                  
> 
```

From the summary above, we can infer that there are 330 observations for non-schizophrenic group and 180 observations for the schizophrenic group. We can also observe that for non-schizophrenic group, the minimum reaction time is 1 and the maximum reaction time is 30 which is much less than the schizophrenic group whose minimum reaction time is 45 and the maximum reaction time is 1714. This proves that the reaction time for schizophrenic group is much more spread out.

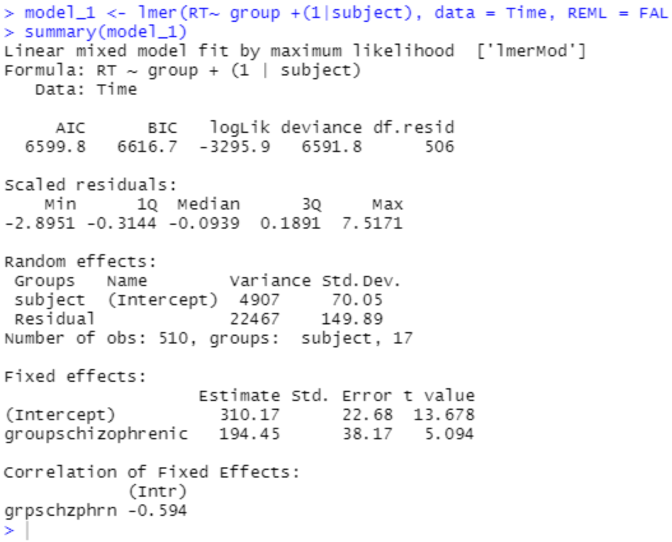
And the scatterplot for the variables RT vs group is shown below.



From this plot above, we can further confirm that the non-schizophrenic data points of RT are very clustered together whereas the schizophrenic data points of RT are comparatively spread out, which proves that the reaction time of the non-schizophrenic group is much less than that for schizophrenic group.

Now, lets move on to choosing a right random effects model for this. From all the factors in the data collected, group and subject seem to be two of the efficient factors out of everything and the most suitable and appropriate factor to take into consideration is subject as the random effect because the data collected says that reaction time observations are collected from 17 subjects randomly. For the same reasons we will be using one-factor random effect model.

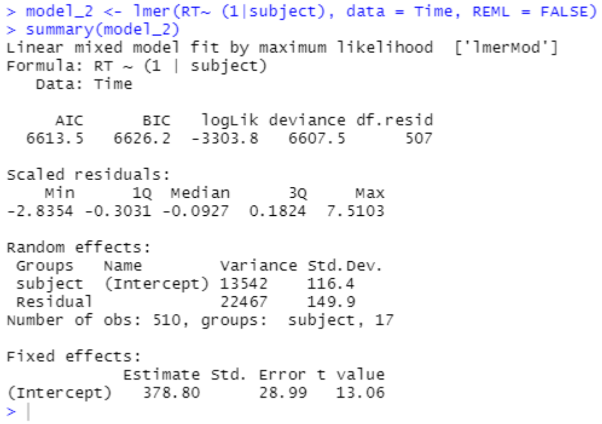
Summary for the random effect model is shown below.



```
> model_1 <- lmer(RT~ group +(1|subject), data = Time, REML = FAL
> summary(model_1)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: RT ~ group + (1 | subject)
   Data: Time

     AIC      BIC   logLik deviance df.resid 
  6599.8   6616.7  -3295.9   6591.8      506 

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-2.8951 -0.3144 -0.0939  0.1891  7.5171 

Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept)  4907     70.05  
 Residual             22467    149.89  
Number of obs: 510, groups:  subject, 17

Fixed effects:
                   Estimate Std. Error t value
(Intercept)          310.17      22.68  13.678
groupschizophrenic   194.45      38.17   5.094

Correlation of Fixed Effects:
            (Intr)
grpschzphrn -0.594
> 
```

From the above summary, we can infer that the estimated variance of subject which is our random effect is 4907 and thus the variance of Yij is 27374. We can also infer that around 17% of the total variance of reaction time is due to the random effect.

Now, let’s check if there is a significant effect of the factor group. We do that by performing the inferential analysis between this model and a model taken without the group as a factor as shown below. The summary for the reduced model is shown below.



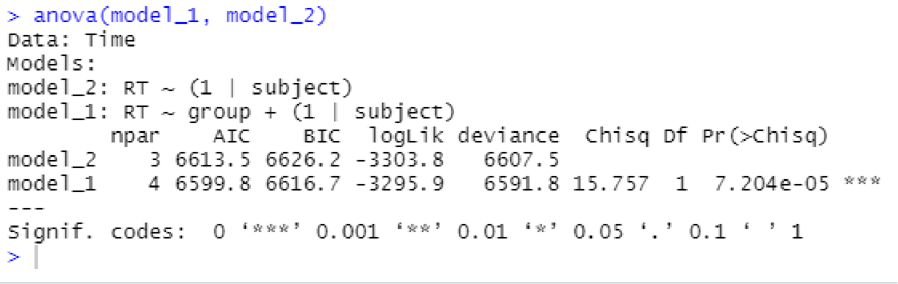
```
> model_2 <- lmer(RT~ (1|subject), data = Time, REML = FALSE)
> summary(model_2)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: RT ~ (1 | subject)
   Data: Time

     AIC      BIC   logLik deviance df.resid 
  6613.5   6626.2  -3303.8   6607.5      507 

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-2.8354 -0.3031 -0.0927  0.1824  7.5103 

Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 13542    116.4   
 Residual             22467    149.9   
Number of obs: 510, groups:  subject, 17

Fixed effects:
            Estimate Std. Error t value
(Intercept)   378.80      28.99   13.06
> 
```

Hypothesis testing:

Let’s say β1, β2 is the coefficients of both the groups.

Null hypothesis - H0: β1 = β2

Alternate hypothesis – Ha: Both the groups are different, i.e. β1 is not equal to β2

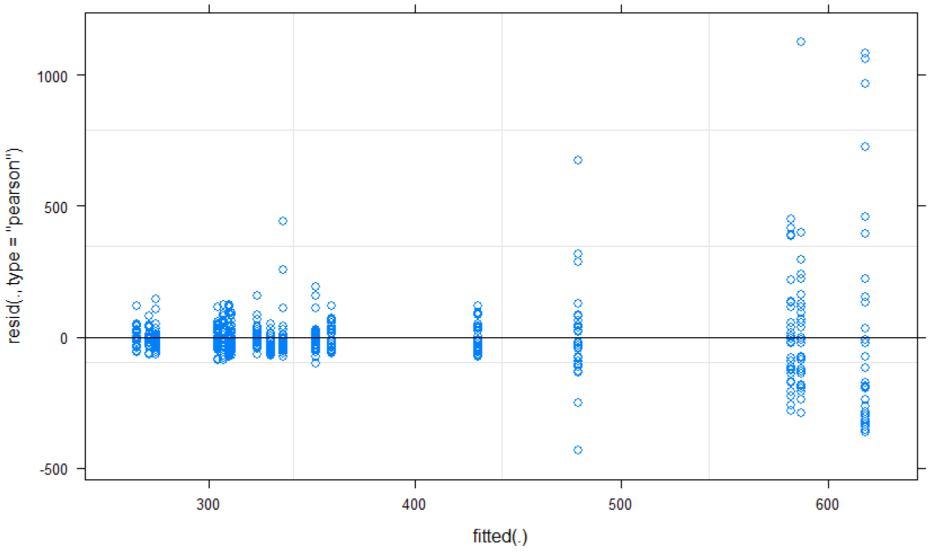


```
> anova(model_1, model_2)
Data: Time
Models:
model_2: RT ~ (1 | subject)
model_1: RT ~ group + (1 | subject)
        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)    
model_2    3 6613.5 6626.2 -3303.8   6607.5                         
model_1    4 6599.8 6616.7 -3295.9   6591.8 15.757  1  7.204e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

From the above screenshot we can see that the test statistic value is 15.757 and the p-value is 7.204e-05 which is less than the alpha value 0.05. Thus, we reject the null hypothesis. Hence, we can say that there is evidence to prove that there is a significant effect of the predictor group on the reaction time and that both the groups are different.

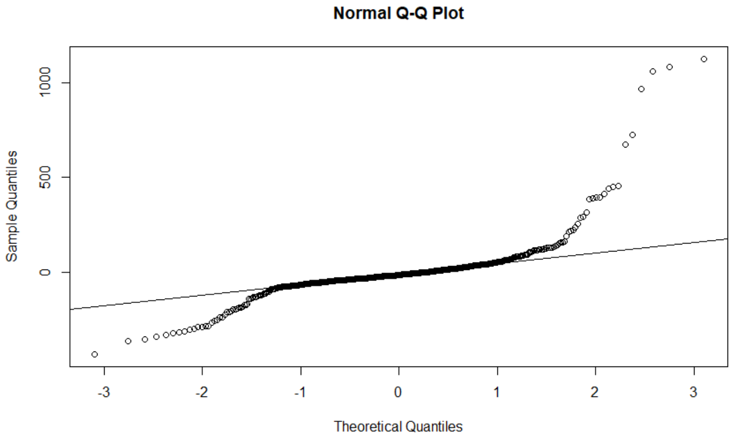
Model Assessment:

Let’s see the Residual vs fitted plot for our model with group and subject factors below.



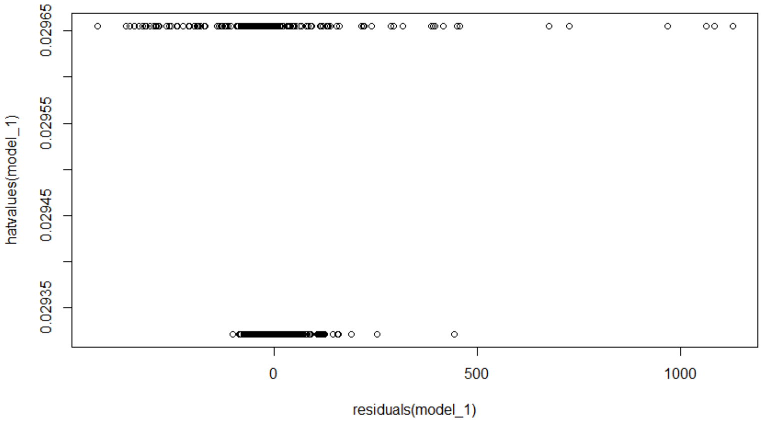
We can see from the plot above that most of the data points lie in at the origin and the rest of the data points lie away from the origin which are not many when compared. All the datapoints however show a pattern of being symmetric showing that it satisfies the assumption of normality of residuals.

Now, let’s see the normal Q-Q plot of our model below

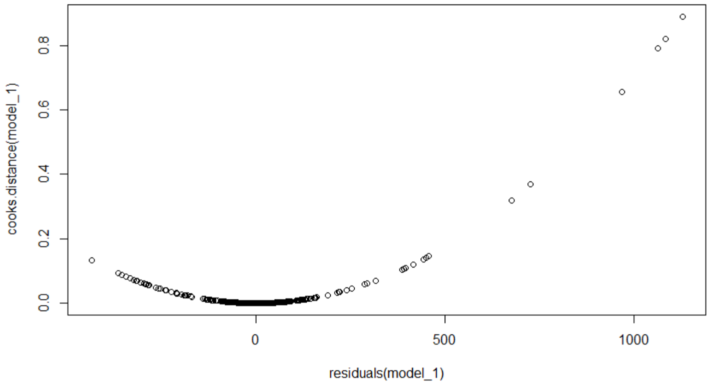


From the above plot we see that the ends of the plot on both the sides are deviated from the straight line while the rest of the part in the between lies on the straight line. The majority of the data points can be seen lying on the straight line. As a whole it looks like a fat tailed distribution and thus, we can say that there is no normal distribution between the data points.

Now, moving on let's see the leverage vs residual plot below.



Cooks distance vs residual plot for our model below.



We can see from the plots above that there are 33 points that have more than 3 times of the mean and that removing these points from the model could give us an improved model with better fit.

From the overall analysis done above, we can now conclude that there are meaningful differences in the reaction times between the schizophrenic and non-schizophrenic patients and that the reaction times of the schizophrenic group is much more than that of non-schizophrenic group and the inferential analysis further proves that the factor group has influential effect in the fitting of the model. One of the limitations we observed in the model is that the datapoints are not normally distributed and that removing these points from the model could give us an improved model with better fit.

Question 2:

Dataset: abalone.csv

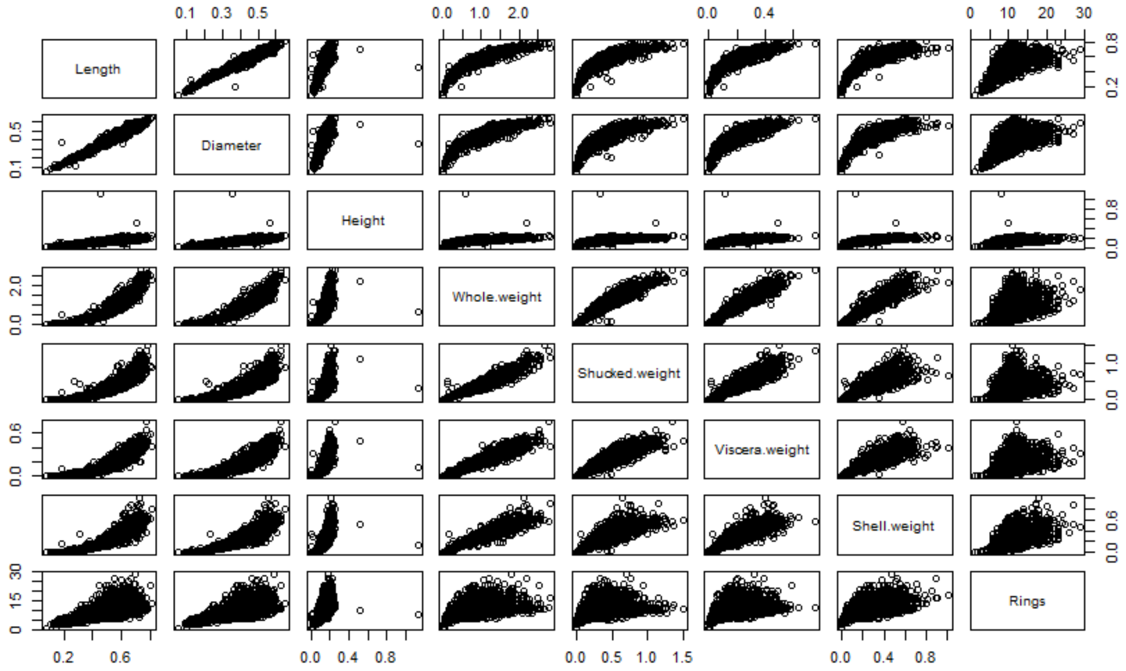
Dataset Source: https://www.kaggle.com/rodolfomendes/abalone-dataset

Description: The file *abalone.csv* contains data collected from the field of zoology. It is said that the age of Abalone shells (a type of mollusk) could be determined by cutting through their cone, staining them, and counting the number of rings inside the shell under a microscope. However, in a data scientist and a statistician point of view it is said that the number of rings that determines the age of Abalone could be predicted, depending on the length, diameter, height, whole weight, shucked weight, viscera weight, shell weight of the Abalone shells. For these reasons, the data collected by the zoologist regarding the Abalone shells is stored in the abalone.csv file.

Key question: Are there any meaningful predictors to predict the number of rings such that we could know the age of Abalone shells by keeping the response variable as Rings. Show which model is the strongest in predicting the number of rings.

Analysis: The original data contains the following variables: Sex, Length, Diameter, Height, whole weight, shucked weight, viscera weight, shell weight, Rings. Here, Rings is the response variable, and all the other variables are the predictor variables. However, after carefully inspecting of all the variables, I have decided to remove the Sex as a predictor variable since it is irrelevant for predicting the response variable.

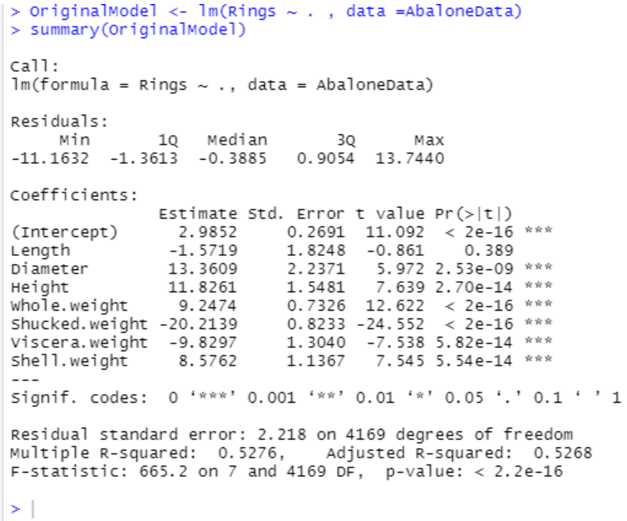
Now, lets first check the scatter plot for all the predictors to see if we can find the variables that have a good co-relationship with Rings. The scatterplot for them is shown below.



From the image above, we can infer the following:

1. Length and Diameter has the strongest positive correlation with Rings.
2. Shucked weight, Shell weight seems to be positively correlated with Rings.
3. Viscera weight, Whole weight are the other variable that seem to be the least positively correlated with Rings.

Now let us also check the standard error for the variables. The summary of the whole model is shown below that also includes the standard error:

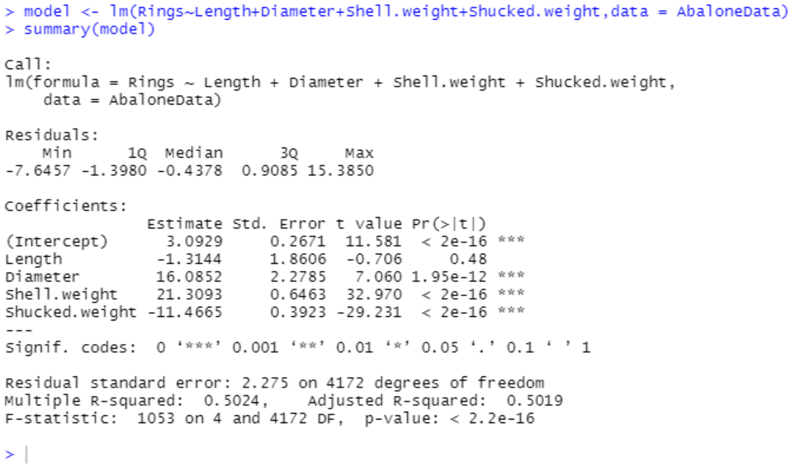


```
> OriginalModel <- lm(Rings ~ . , data =AbaloneData)
> summary(OriginalModel)

Call:
lm(formula = Rings ~ ., data = AbaloneData)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1632  -1.3613  -0.3885   0.9054  13.7440 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      2.9852     0.2691  11.092  < 2e-16 ***
Length          -1.5719     1.8248  -0.861    0.389    
Diameter        13.3609     2.2371   5.972 2.53e-09 ***
Height          11.8261     1.5481   7.639 2.70e-14 ***
Whole.weight     9.2474     0.7326  12.622  < 2e-16 ***
Shucked.weight -20.2139     0.8233 -24.552  < 2e-16 ***
Viscera.weight  -9.8297     1.3040  -7.538 5.82e-14 ***
Shell.weight     8.5762     1.1367   7.545 5.54e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.218 on 4169 degrees of freedom
Multiple R-squared:  0.5276,	Adjusted R-squared:  0.5268 
F-statistic: 665.2 on 7 and 4169 DF,  p-value: < 2.2e-16

> 
```

From the summary above we also see that the standard error is close to zero for all the variables and that all the predictor variables do not seem to have much standard error.

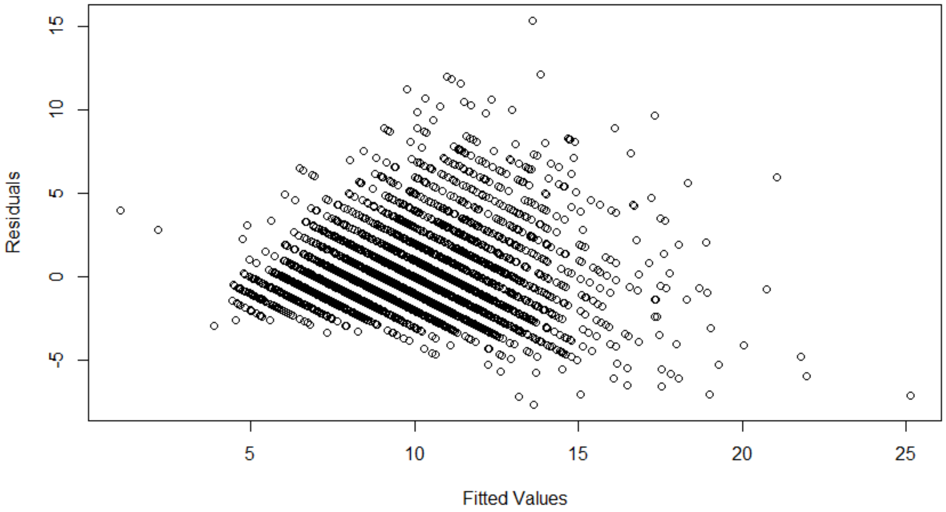
So, we move on and by inferring from the scatter plot and standard error. We see that the predictors Length, Diameter, Shucked weight, Shell weight are the most correlated to predict the Rings than the other 3 predictors. We use the combination of these predictors further to make our model.

Model: With the predictors Length, Diameter and Shucked weight, Shell weight that we found to be correlated strongly positive and moderately positive respectively. The summary for the model is shown below.



```
> model <- lm(Rings~Length+Diameter+Shell.weight+Shucked.weight,data = AbaloneData)
> summary(model)

Call:
lm(formula = Rings ~ Length + Diameter + Shell.weight + Shucked.weight,
    data = AbaloneData)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6457 -1.3980 -0.4378  0.9085 15.3850

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.0929     0.2671  11.581  < 2e-16 ***
Length          -1.3144     1.8606  -0.706     0.48
Diameter        16.0852     2.2785   7.060 1.95e-12 ***
Shell.weight    21.3093     0.6463  32.970  < 2e-16 ***
Shucked.weight -11.4665     0.3923 -29.231  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.275 on 4172 degrees of freedom
Multiple R-squared:  0.5024,    Adjusted R-squared:  0.5019
F-statistic:  1053 on 4 and 4172 DF,  p-value: < 2.2e-16

> |
```

From the summary we see that the models R-squared value is 0.50 which means that only 50% dataset fits the model perfectly. Now, lets see the residual plot for the model below. We can observe that there is no equal variance with which the points are distributed throughout the plot. In order to formally test the variance, we move on to perform the Breusch-pagan test.

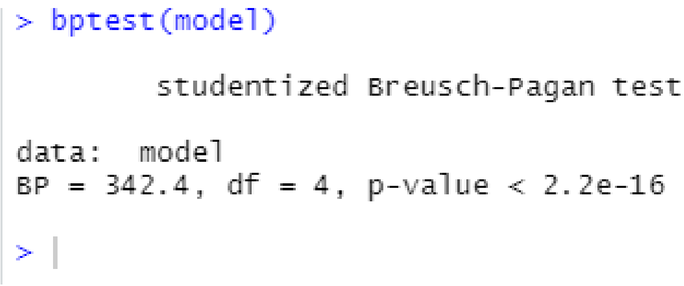


Hypothesis testing:

Null hypothesis – Residuals are distributed with equal variance

Alternate hypothesis – Residuals do not have an equal variance distribution

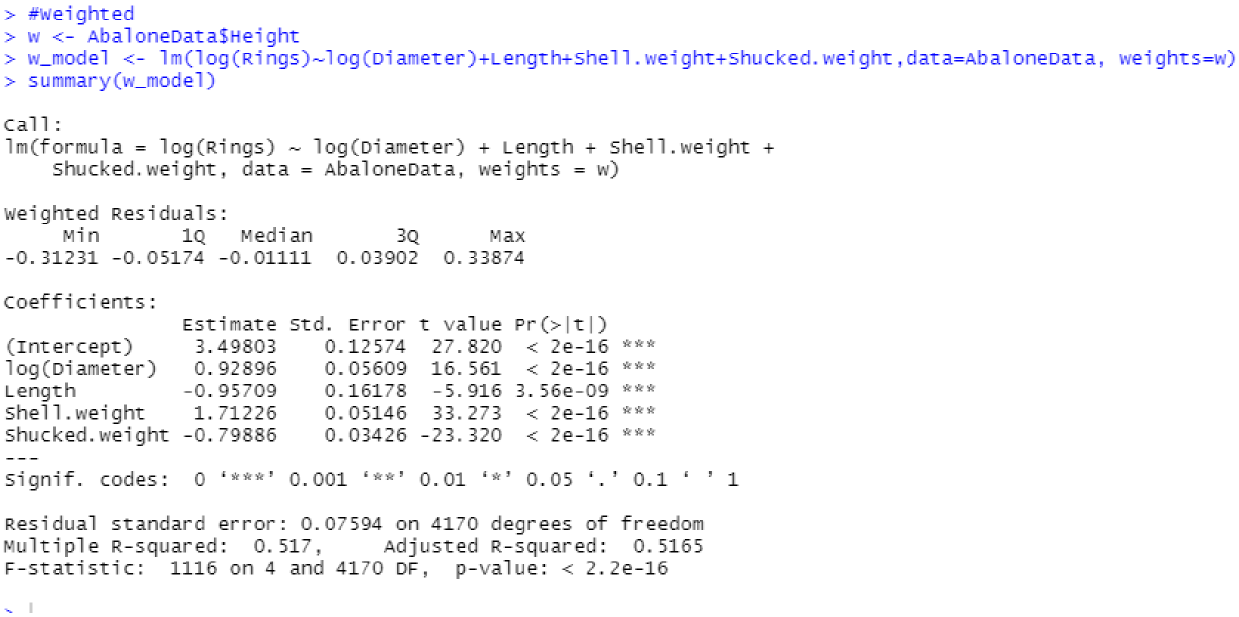
BP test is shown below.



```
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 342.4, df = 4, p-value < 2.2e-16

> |
```

From the screenshot above, we can see that in the test the p-value is < 2.2e-16 which is less than the alpha value 0.05. Hence, we reject the null hypothesis. Thus, we can conclude that there is a variance issue in the model and that the residuals do not have an equal variance distribution.

Now, lets move on to making a model with weighted least squares. Here, we use the Height variable as the weight since its correlation with Rings could not be inferred visually from the scatterplot. The summary for the same is shown below.

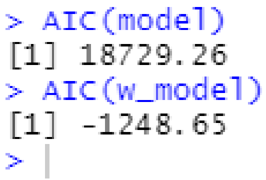


```
> #weighted
> w <- AbaloneData$Height
> w_model <- lm(log(Rings)~log(Diameter)+Length+Shell.weight+Shucked.weight,data=AbaloneData, weights=w)
> summary(w_model)

Call:
lm(formula = log(Rings) ~ log(Diameter) + Length + Shell.weight + 
    Shucked.weight, data = AbaloneData, weights = w)

Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-0.31231 -0.05174 -0.01111  0.03902  0.33874 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.49803    0.12574  27.820  < 2e-16 ***
log(Diameter)   0.92896    0.05609  16.561  < 2e-16 ***
Length         -0.95709    0.16178  -5.916 3.56e-09 ***
Shell.weight    1.71226    0.05146  33.273  < 2e-16 ***
Shucked.weight -0.79886    0.03426 -23.320  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07594 on 4170 degrees of freedom
Multiple R-squared:  0.517,	Adjusted R-squared:  0.5165 
F-statistic:  1116 on 4 and 4170 DF,  p-value: < 2.2e-16

> |
```

From the summary above, we can see that the overall fit of the model has improved. The residual standard error for the weighted least squares model is 0.517 which when compared to the original simple linear regression model, is more. This means that the predicted values produced by weighted least squares model are closer to the actual observations than the ones by the simple linear regression model.

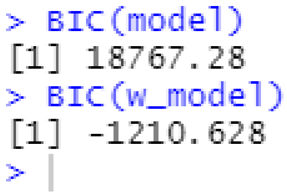
Now, based on the R-squared values for now we can say that the weighted least squares model is better since the R-squared value of that model is more than the R-squared value of the simple linear regression model. But to further confirm our claim, lets proceed with doing AIC and BIC for both the models.

The AIC values of both the models is shown below:



```
> AIC(model)
[1] 18729.26
> AIC(w_model)
[1] -1248.65
> 
```

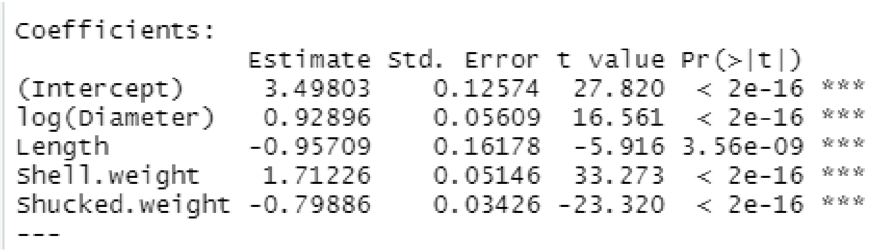
The BIC values of both the models is shown below:



```
> BIC(model)
[1] 18767.28
> BIC(w_model)
[1] -1210.628
> 
```

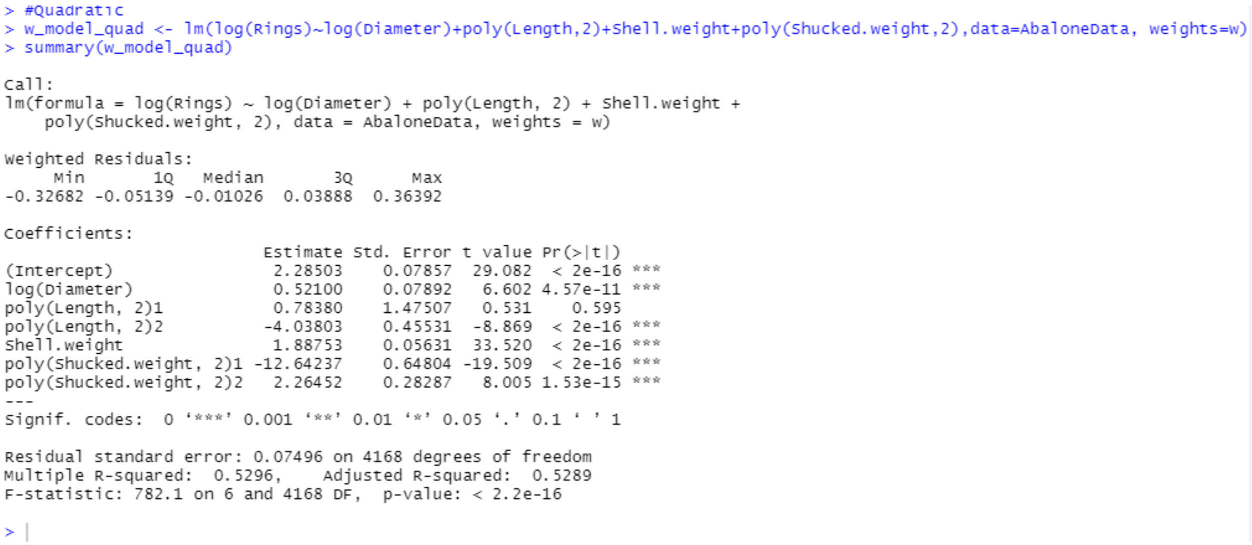
We can see from the screenshots above that both the AIC and BIC values of the weighted least squares model is lesser than the AIC and BIC values of the linear regression model. This further confirms our claim that the weighted least squares model is a better fit for predicting the Rings as of now.

The coefficients for the weighted least squares model is shown below:



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.49803    0.12574  27.820  < 2e-16 ***
log(Diameter)    0.92896    0.05609  16.561  < 2e-16 ***
Length          -0.95709    0.16178  -5.916 3.56e-09 ***
Shell.weight     1.71226    0.05146  33.273  < 2e-16 ***
Shucked.weight  -0.79886    0.03426 -23.320  < 2e-16 ***
---
```

Now, moving on, lets check if adding the quadratic terms to this selected model further improves the model. The summary for the same is shown below:



```
> #Quadratic
> w_model_quad <- lm(log(Rings)~log(Diameter)+poly(Length,2)+Shell.weight+poly(Shucked.weight,2),data=AbaloneData, weights=w)
> summary(w_model_quad)

Call:
lm(formula = log(Rings) ~ log(Diameter) + poly(Length, 2) + Shell.weight + 
    poly(Shucked.weight, 2), data = AbaloneData, weights = w)

Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-0.32682 -0.05139 -0.01026  0.03888  0.36392 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                2.28503    0.07857  29.082  < 2e-16 ***
log(Diameter)              0.52100    0.07892   6.602 4.57e-11 ***
poly(Length, 2)1           0.78380    1.47507   0.531    0.595    
poly(Length, 2)2          -4.03803    0.45531  -8.869  < 2e-16 ***
Shell.weight               1.88753    0.05631  33.520  < 2e-16 ***
poly(Shucked.weight, 2)1 -12.64237    0.64804 -19.509  < 2e-16 ***
poly(Shucked.weight, 2)2   2.26452    0.28287   8.005 1.53e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07496 on 4168 degrees of freedom
Multiple R-squared:  0.5296,	Adjusted R-squared:  0.5289 
F-statistic: 782.1 on 6 and 4168 DF,  p-value: < 2.2e-16

> 
```

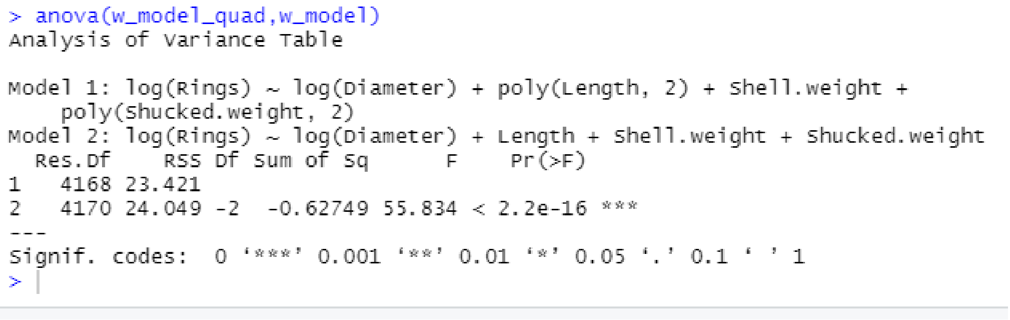
From the summary we can see that the R-squared value of this model has increased which tells us that this model might be a better fit. To further check if this is true, lets do the inferential analysis for it.

Inferential analysis:

Let’s say that β1, β2, β3 are the coefficients of Diameter, Length, Shucked weight respectively.

Null hypothesis: β1 =β2=β3=0

Alternate hypothesis: At least one of the above coefficients is non-zero

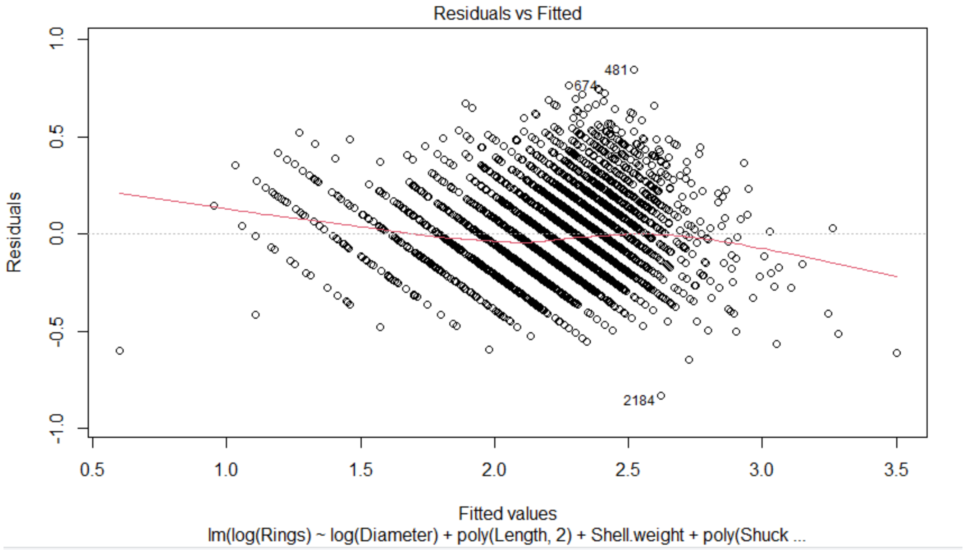


```
> anova(w_model_quad,w_model)
Analysis of Variance Table

Model 1: log(Rings) ~ log(Diameter) + poly(Length, 2) + Shell.weight +
    poly(Shucked.weight, 2)
Model 2: log(Rings) ~ log(Diameter) + Length + Shell.weight + Shucked.weight
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1   4168 23.421                                  
2   4170 24.049 -2  -0.62749 55.834 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

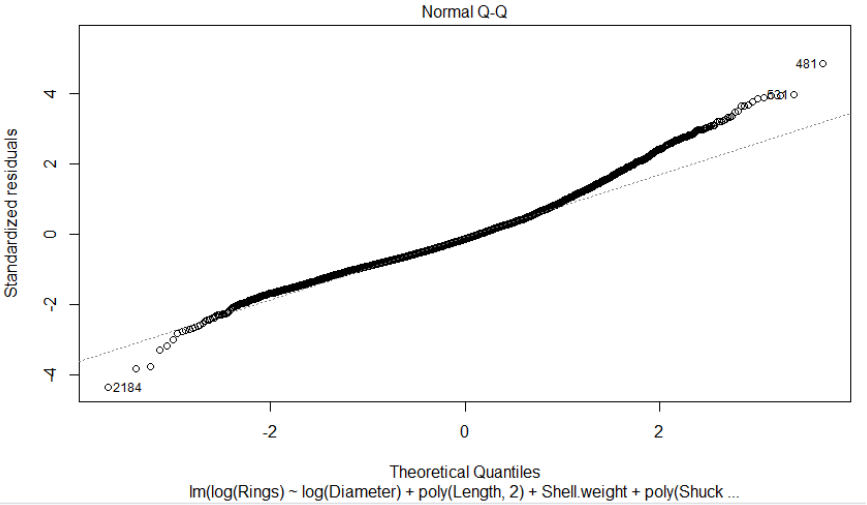
From the above table in the screenshot, we can see that the F-statistic is 55.834 and the p-value is <2.2e-16 which is less than the alpha value 0.05. Thus, we reject the null hypothesis. Hence, we can conclude that there is evidence to prove that the quadratic terms will improve the model fit.

Model assessment:

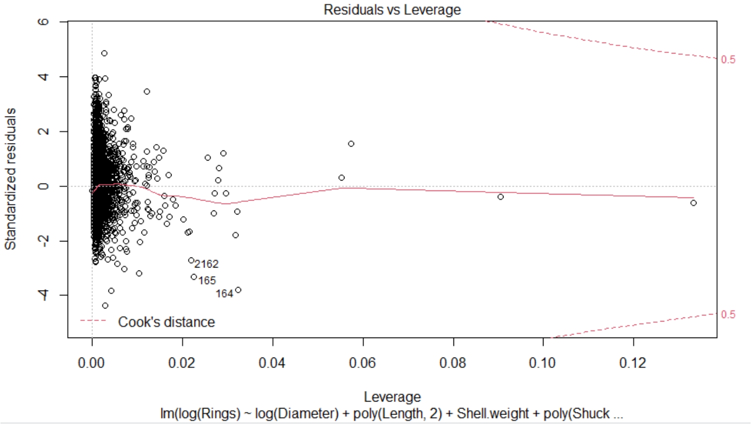
Residual vs fitted plot is shown below



Q-Q plot is shown below:



Leverage vs residual plot is shown below



Though the residual plot above looks similar to the original model plot, we need to take note that it is however different. In the residual plot above, the points that are closer to the reference line are much more than that of the original model. Though there are a few outliers in the plot, it is much lesser when compared to the original model.

From the Q-Q plot, we can see that the observation 531 and 481 are outside which means that these two observations could be influential. From the leverage values, we see that the highest value is 0.13

Taking all this analysis into consideration, we can now for sure say that Diameter, Length, Shucked weight effect the model fitting. We also observe that adding quadratic terms to the weighted least squares model further improves the model fit. By also taking all the analysis of the plot above we can conclude that the weighted least squares model with added quadratic terms is the best fit model for predicting the Rings. However, one limitation observed in the final quadratic model is that the datapoints are not normally distributed and there is a possibility that removing these points from the model could give us a further improved model with better fit.