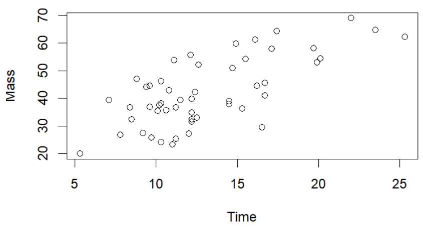
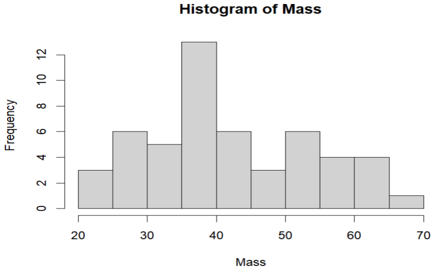
Statistics Assignment 2

Exploring the data:

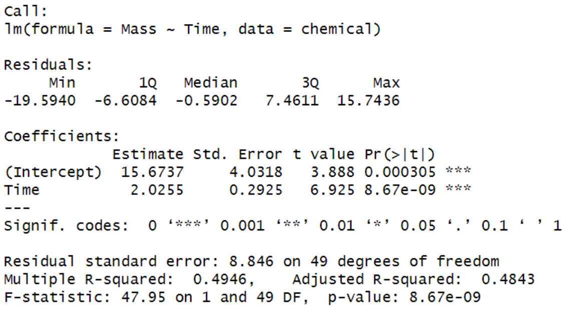




To start to understand the dataset I loaded the csv file into R plotting the Explanatory Variable (Time) against the response variable (Mass). For this graph I'm looking for any potential relationship and which model I could use as an approach. I'm also looking at possible residuals which could later alter my predictions. For this graph, I'm observing a possible positive linear relationship. This gives the prediction that as the Time increases for the chemical reaction, so does the mass of the chemical.

Reading the numbers:

After creating a linear regression model, summarising the data in R can give a more accurate relationship between the two given variables and the number output allows me to draw conclusions later on.



```
Call:
lm(formula = Mass ~ Time, data = chemical)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.5940  -6.6084  -0.5902   7.4611  15.7436 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  15.6737     4.0318   3.888 0.000305 ***
Time          2.0255     0.2925   6.925 8.67e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.846 on 49 degrees of freedom
Multiple R-squared:  0.4946,	Adjusted R-squared:  0.4843 
F-statistic: 47.95 on 1 and 49 DF,  p-value: 8.67e-09
```

Straight away from my coefficients I can see that there's a statistical significant relationship between my two variables. This is shown by the P-values given by hypothesis testing. The Estimates show me that there's a positive relationship in my dataset where the response variable is on average increased by 2 units whenever the explanatory variable is increased by 1 unit. Leading to my prediction being correct. For every second the chemical reaction takes place, on average the mass of the chemical compound is increased by 2 grams. As Time is increasing the Mass is also increasing.

Checking assumptions:

I’ve decided to use a Linear Regression Model for this dataset because of my observation of the shape from the Time against Mass plot therefore I have to check the assumptions…

- Data is independent
- Data is Identically distributed with mean 0 and constant variance
- Data follows a normal distribution

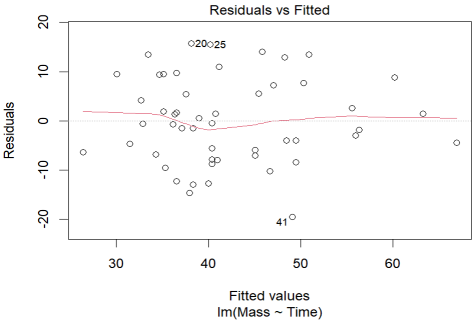
Is the data independent?

Yes, I will assume that the data is independent as the data gathered is from 51 individual experiments, of course this doesn’t mean that no error occurred in gathering the data but given the lack of information provided, I will assume it’s independent.

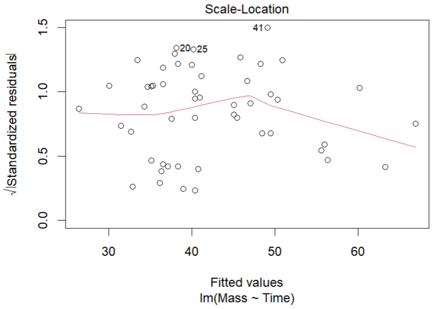
Is the data Identically distributed, constant variance?

The observed observations aren’t more than we would expect by chance.

Which = 1

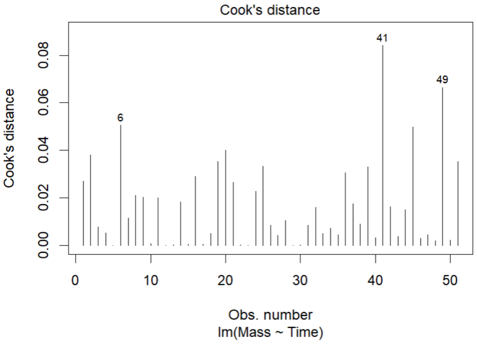


Which = 3



Creating a linear regression model and using the command plot() in R, outputs graphs which allow me to check my assumptions, the graph where I’ve plotted which=1 shows a Fitted values against residuals graph which I observe to be a loose flat line with no patterns or big curves. From this graph I therefore can say that a linear relationship is correct for my dataset and that the residual scatter is fairly consistent. In which=3 I have a Scale-Location plot which is used to check constant variance of residuals. Because the red-line in my Scale-Location graph is relatively flat, I can say that I have constant variance here. There’s a minor bend upwards of the line though which may indicate that there’s residuals which could be affecting the variance, R has highlighted a few at the (20, 25, 41) marks which I can check in the Cook’s Distance Plot.

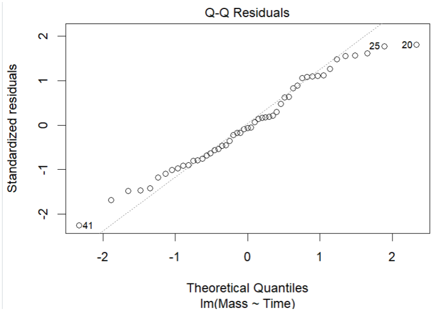
No point in the Cook’s Distance graph exceeds 0.4 therefore there’s no notable residuals that skew my data and my model is stable.



Does the data follow a normal distribution?

Yes, the data follows an approximate normal distribution. The Histogram plot of my Response Variable shows that although not perfect, the variable holds for an approximate Normal Distribution. Therefore it’s okay to use the Linear Regression Model. I can also tell it’s a normal distribution from the following graph if the Histogram wasn’t a clear indication.

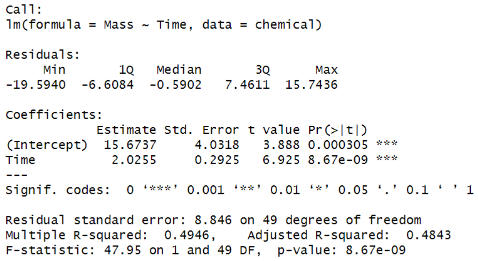
Which = 2



From the Which=2 plot (Q-Q Residuals plot) it shows me that my residuals follow a clear Normal distribution with the majority of the data being centered around the 0 point. There’s only a few outliers at the extreme left and right tails which is nothing to be concerned about.

Predictions:

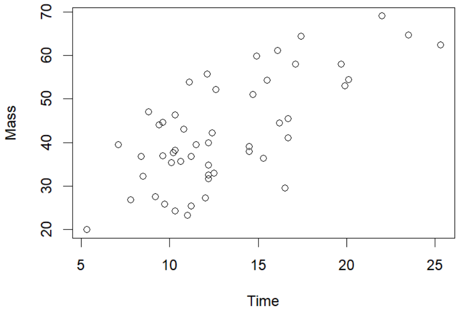
To predict by hand what the mass at 20 seconds is, I look at the summary data for my Linear Model and calculate what the Mass should be at the point. For functionality, I will round the estimates to 15.7 and 2



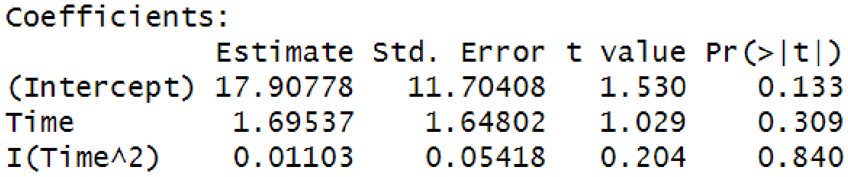
```
Call:
lm(formula = Mass ~ Time, data = chemical)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.5940  -6.6084  -0.5902   7.4611  15.7436 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  15.6737     4.0318   3.888 0.000305 ***
Time          2.0255     0.2925   6.925 8.67e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.846 on 49 degrees of freedom
Multiple R-squared:  0.4946,	Adjusted R-squared:  0.4843 
F-statistic: 47.95 on 1 and 49 DF,  p-value: 8.67e-09
```

Mass at 20 seconds = 15.7 + (2 * 20)
= 15.7 + (40)
= 55.7 grams

Looking at the Explanatory against response plot I can see that it’s roughly accurate. In conclusion when the time of the experiment is 20 seconds, the mass of the chemical will be 55.7 grams on average.

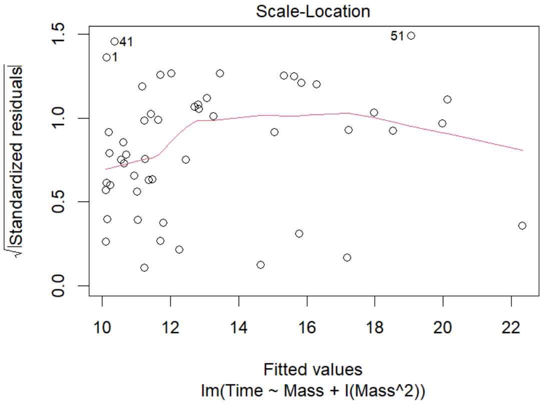


Quadratic Model:

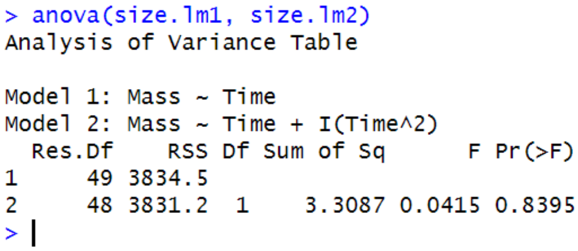


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.90778   11.70408   1.530    0.133
Time         1.69537    1.64802   1.029    0.309
I(Time^2)    0.01103    0.05418   0.204    0.840
```

After summarising the data of a quadratic model for the same dataset, the estimates outputted in R show the relationship between the Explanatory and Response variable and the Quadratic increase of 0.011. Looking at the P-values of the Quadratic Model, neither the Time or Time^2 are significant being at 0.31 and 0.84 meaning that introducing the quadratic model causes the linear model to be non significant.



The Scale-Location for this Quadratic Model indicates a clear curve which violates the assumption for constant variance. It also shows the scatter centered around the 10-12 region which further guarantees this.



```
> anova(size.lm1, size.lm2)
Analysis of Variance Table

Model 1: Mass ~ Time
Model 2: Mass ~ Time + I(Time^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     49 3834.5
2     48 3831.2  1    3.3087 0.0415 0.8395
> |
```

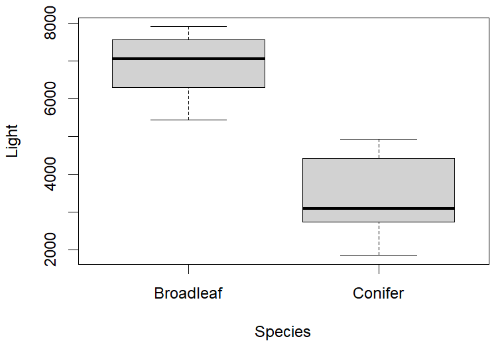
After an ANOVA test comparison between the Linear and Quadratic Models, it shows that there’s no significance of the Quadratic model as the P-value is at 0.84.

Because of the violations in constant variance and the insignificance of the Quadratic Model as well as the shape of the Explanatory against Response plot, I believe that the Linear Model is better for this dataset to find a relationship between the two variables and reject the statement that the Quadratic Model is a better fit.

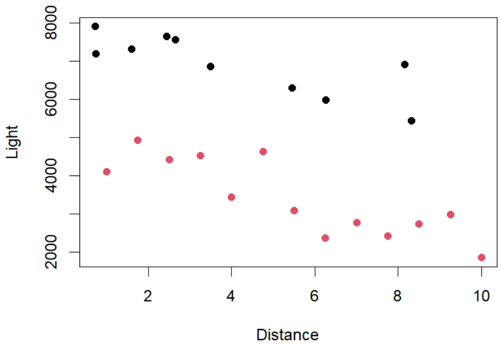
Forest Dataset:

Exploring the data:

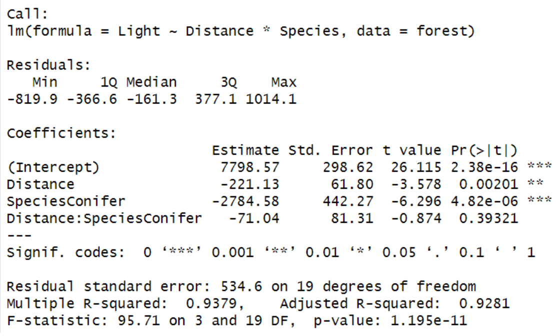
This boxplot graph shows that the median light intensity is higher for the Broadleaf species in comparison to the Conifer species. The upper quartile of the conifer doesn't overlap with the Broadleaf forest lower quartile range meaning the lowest light forests of the Broadleaf species still have more light than the highest of the conifer species.



For this graph Black = Broadleaf, Red = Conifer I see two negative slopes for both variables which look to be decreasing in a linear relationship. From just looking at the points, it seems that as the distance from the top of the canopy is increasing, the light level decreases.
As seen in the box plot comparison, the Broadleaf points in this graph have higher measured Lux than the Conifer. This may be because of the shape of the trees, which part of the forest the light is measured in, placement of the forests, and many other factors. The graph shows me all individual observations from this sample. Such a small sample size (13 Conifer and 10 Broadleaf) observations is noted when drawing conclusions about this data as it could cause inaccuracy in the concluded result.



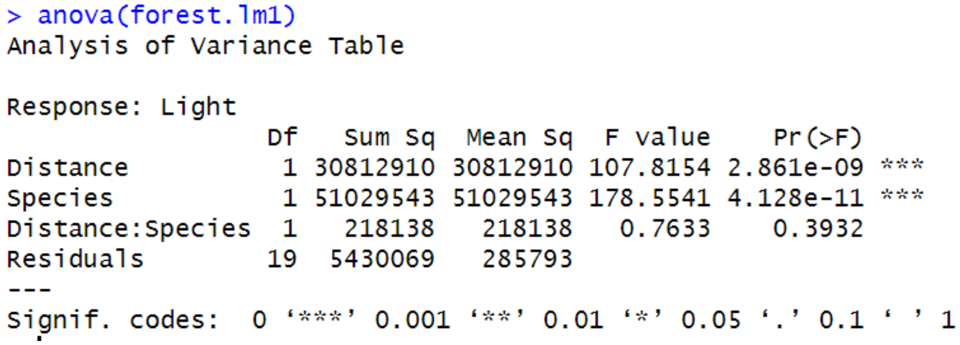
Reading the numbers:



```
Call:
lm(formula = Light ~ Distance * Species, data = forest)

Residuals:
   Min     1Q Median     3Q    Max 
-819.9 -366.6 -161.3  377.1 1014.1 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              7798.57     298.62  26.115 2.38e-16 ***
Distance                 -221.13      61.80  -3.578  0.00201 ** 
SpeciesConifer          -2784.58     442.27  -6.296 4.82e-06 ***
Distance:SpeciesConifer   -71.04      81.31  -0.874  0.39321    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 534.6 on 19 degrees of freedom
Multiple R-squared:  0.9379,	Adjusted R-squared:  0.9281 
F-statistic: 95.71 on 3 and 19 DF,  p-value: 1.195e-11
```

The summary of this linear regression model shows me that the intercept for the broadleaf species is 7798 and that the distance is -221.13 so for every unit increase of distance from the tree (I will assume meters), the measured Light amount will decrease by 221.13 Lux on average. This has a P-value of 0.00201 meaning it's highly significant. If instead the species is Conifer, the Estimate is -2784.58 meaning that the intercept is 7798.57 - 2784.58 = 5013.99 which means the Conifer species allows less light than the Broadleaf species. When looking at the Distance:SpeciesConifer result I see a -71.04 value which means that the rate at which the light decreases for the Conifer species may be different to that of the Broadleaf. However, the P-value for this is 0.39 meaning it isn’t significant. Because of this significance, there’s no clear evidence to determine that distance affects light differently through species.

Different Model?



```
> anova(forest.lm1)
Analysis of Variance Table

Response: Light
                 Df   Sum Sq  Mean Sq  F value    Pr(>F)    
Distance          1 30812910 30812910 107.8154 2.861e-09 ***
Species           1 51029543 51029543 178.5541 4.128e-11 ***
Distance:Species  1   218138   218138   0.7633    0.3932    
Residuals        19  5430069   285793                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Above is the ANOVA output for the linear regression model. Because of the significance levels, I can see that Distance has a significant relationship to the Light variable meaning as distance is changing the light is decreasing. The species variable is also highly significant meaning that species also has an impact on the amount of lighting. The Distance:Species is not significant meaning that the impact of distance on light does not depend on the Species.

Since Distance:Species isn’t significant (0.3932) and there’s no higher order terms, we can simply remove it to avoid unnecessary complexity. A better model for this would be Light ~ Distance + Species as this avoids the interaction.