STAT201 Assignment 3

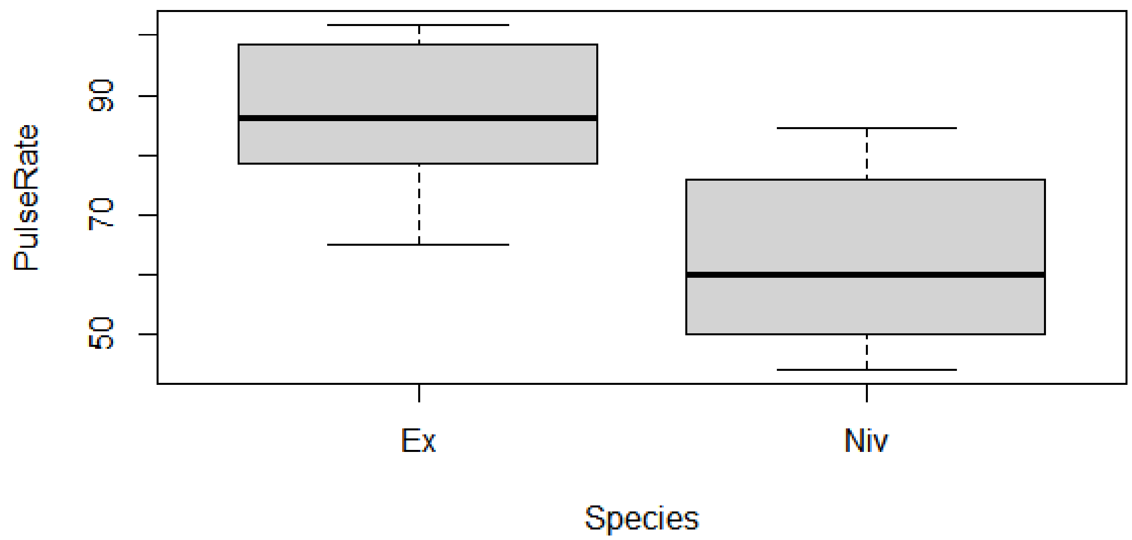
Robert Ivill 46012819

The dataset, crickets.csv, is a study about the pulse rates of the wings of tree crickets and how they differ between two species of tree cricket: Oecanthus exclamationis (Ex) and Oecanthus niveus (Niv). For this report, all the available code and plots are placed in the appendix in the order that they were performed and are mentioned. After importing the dataset into RStudio and checking the import has been read in correctly using head() and tail(), the dataset was explored by creating a boxplot of the pulse rate for the two species. This boxplot shows a clear difference between the two species, despite having similar looking boxplots. Ex has a median pulse rate of about 85, whereas Niv has a median of about 60, which is less than the lower bound of Ex's plot. Ex also has a shorter range and IQR than Niv, meaning it is has a much tighter distribution. The next step was to create a dot plot that shows how temperature and the crickets' pulse rates were related for both species (Black dots are Ex and red are Niv). We can see from the plot that both species are similarly affected by temperature, where they both have similar slopes and correlation, however Ex appears to have a higher intercept. It will be useful to fit two different lines, dependent on species, to solidify this. Ex also doesn't have any data points for temperatures below 20C, perhaps they don't live in climates as cold as Niv. A linear regression model for the plot was created using the lm() function, and the data for this regression model was shown from the summary() and anova() functions. The output shows two levels of the factor 'Species', where Ex is the baseline and Niv is compared to it. The regression model shows coefficients for the baseline (Ex) and then the coefficients for the difference between the baseline and the other factor (Niv). Therefore the Ex interpretation is -11.0408+3.7514 * Temperature and the Niv interpretation is (-11.0408-4.3484)+(3.7514-0.2340)*Temperature = -15.3892+3.5174*Temperature. This shows that they have similar slopes, however Niv has a lower intercept than Ex. From the anova() summary, we see that the interaction between temperature and species is not significant (p=0.2542), therefore we don't need the interaction term. We can remove the interaction term and have a simpler model with two parallel lines for the species, each with their own intercept. We create this model using the lm() function and again use summary and anova to print the model. The model now has two parallel lines, both with a slope of 3.60275. The interpretation for Ex is -7.21091 + 3.60275 * Temperature and -17.2762 +3.60275 * Temperature for Niv. We can see from anova that both temperature (p<2.2e-16) and species (p=6.272e-14) are significant to the model. The p-value for the Niv coefficient in the model is very small (p=6.27e-14), meaning that there is strong evidence that it doesn't differ from the baseline (Ex), thus this is the simplest model. As this is the simplest model, we can look at the residual plots for the model. Our residuals vs fitted model shows that the residuals have an even spread around zero. The Normal Q-Q model shows have little deviation from the normal distribution. However, the Cook's distance model shows that we have 4 or 5 high influence points in the model. From the final model and the plots that were made, we can confidently say that we can use the pulse rate of tree crickets' wings to distinguish between the two species Ex and Niv. This is because Ex always has a higher pulse rate at every temperature than Niv, as the intercept of their slope is higher than that of the Niv.

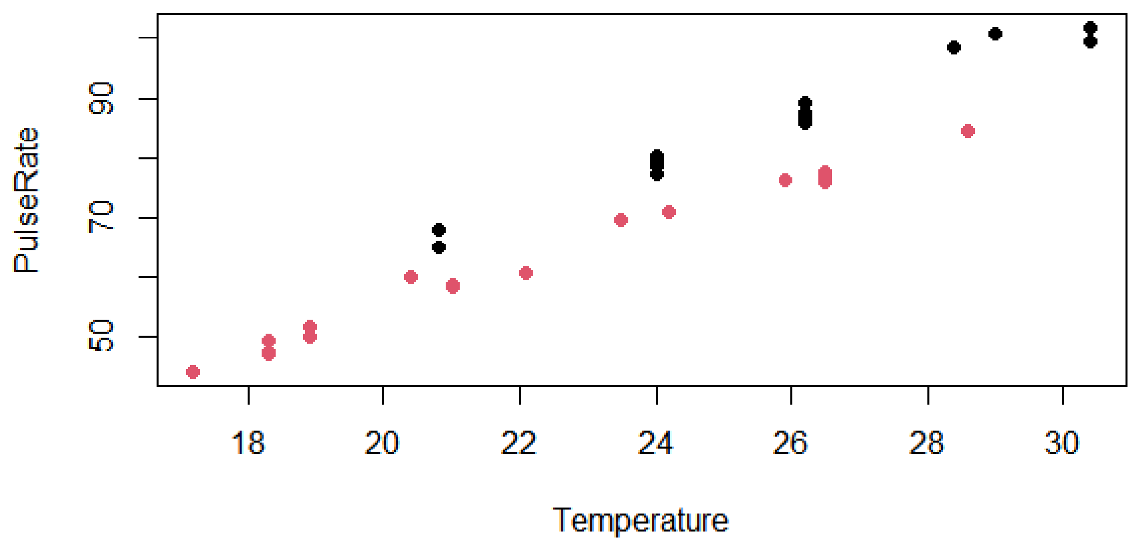
For this next part of the report, we are using the crimes.csv dataset which has data on crime rates for different counties in the USA. It also has data on average county income and whether the county is considered a northern or southern one. The csv file was imported into RStudio and confirmed to be correct with head() and tail(). Then, a dot plot was created to compare the income and crime rates of north counties (black dots) to south counties (red dots). We can see from this plot that the south have on average less income than the north. It appears that the relationship between crime and income is different for the north and south. The north has a more scattered correlation and appears to have a higher slope than the south. To understand more about these relationships, a linear regression model was created and the summary and anova was printed. This time, using a significance level of 0.1 instead of 0.05, we see that the model cannot be simplified. This is because the p-value for the interaction term Income:South is 0.086849, which is less than the significance level chosen. This means that this interaction term is significant for the model as the difference between some of the means is statistically significant, so we can reject the null hypothesis and conclude that not all of the population means are equal. As this model cannot be simplified more, we can look at the three residual plots for the model. We see in the residuals vs fitted model that the residuals are evenly spread around zero except for at the end of the fitted values where it deviates to have predictions that are too low. The Normal Q-Q model shows that the data, for the most part, does not deviate from the normal distribution too much, apart from at the ends of the model again. Cook’s distance shows again that there a few values that are anomalies to the model and perhaps removing them from it could yield better results for the model.

Appendix:

```
> head(crickets)
  Species Temperature PulseRate
1      Ex        20.8      67.9
2      Ex        20.8      65.1
3      Ex        24.0      77.3
4      Ex        24.0      78.7
5      Ex        24.0      79.4
6      Ex        24.0      80.4
> tail(crickets)
   Species Temperature PulseRate
26     Niv        24.2      70.9
27     Niv        25.9      76.2
28     Niv        26.5      76.1
29     Niv        26.5      77.0
30     Niv        26.5      77.7
31     Niv        28.6      84.7
> boxplot(PulseRate~Species, data=crickets)
```

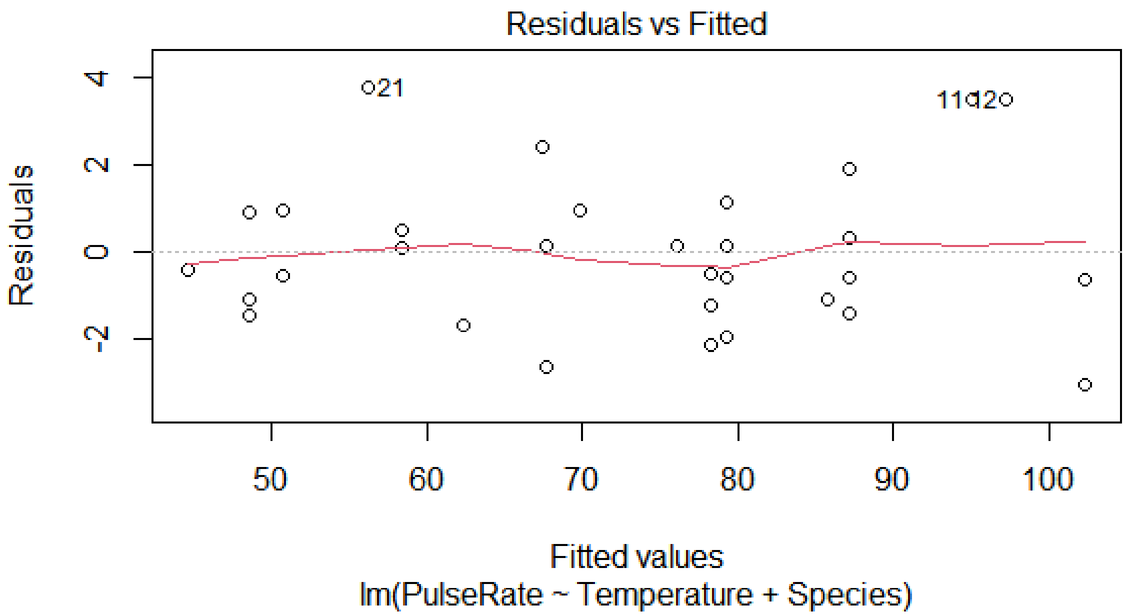



```
> plot(PulseRate~Temperature, data=crickets, pch = 19, col =
+              Species)
```

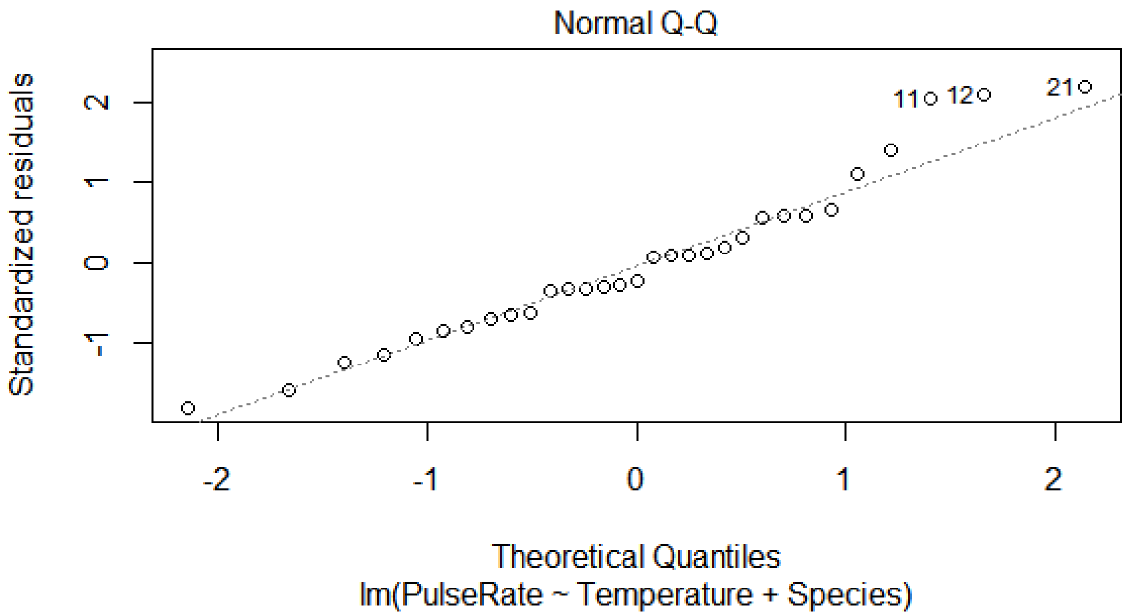



```
> crickets.lm1<- lm(PulseRate ~ Temperature * Species, data =
+                        crickets)
> summary(crickets.lm1)

Call:
lm(formula = PulseRate ~ Temperature * Species, data = crickets)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7031 -1.3417 -0.1235  0.8100  3.6330 

Coefficients:
```

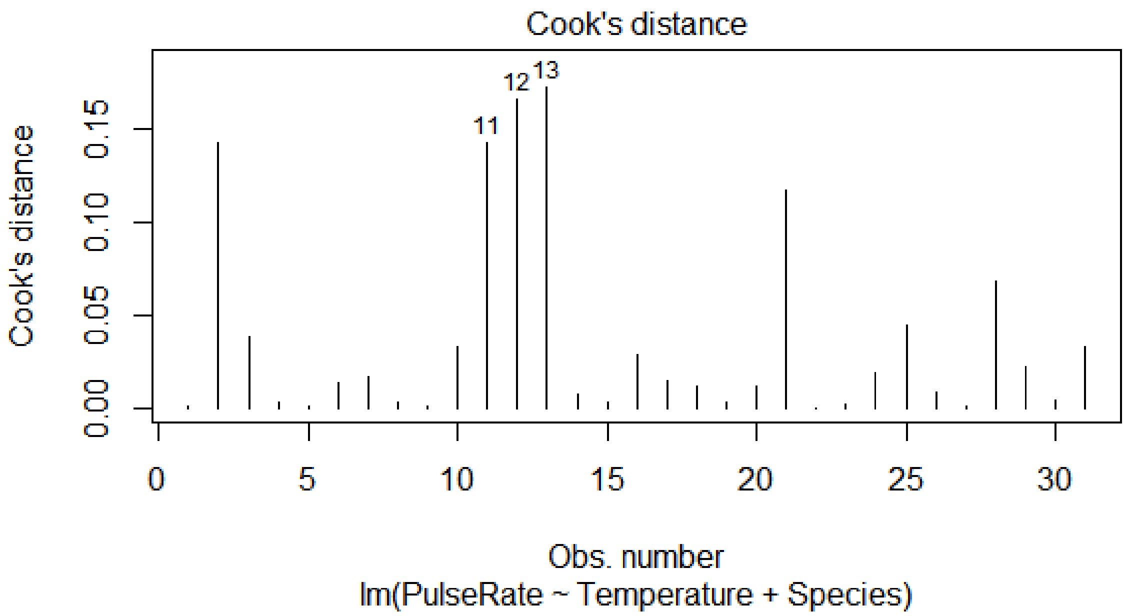
```
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -11.0408     4.1515  -2.659    0.013 *  
Temperature                3.7514     0.1601  23.429   <2e-16 ***
SpeciesNiv                -4.3484     4.9617  -0.876    0.389    
Temperature:SpeciesNiv    -0.2340     0.2009  -1.165    0.254    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.775 on 27 degrees of freedom
Multiple R-squared:  0.9901,	Adjusted R-squared:  0.989 
F-statistic: 898.9 on 3 and 27 DF,  p-value: < 2.2e-16

> anova(crickets.lm1)
Analysis of Variance Table

Response: PulseRate
                    Df Sum Sq Mean Sq  F value    Pr(>F)    
Temperature          1 7894.8  7894.8 2505.583 < 2.2e-16 ***
Species              1  598.0   598.0  189.789 9.907e-14 ***
Temperature:Species  1    4.3     4.3    1.357    0.2542    
Residuals           27   85.1     3.2                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> crickets.lm2<- lm(PulseRate ~ Temperature + Species, data =
+                        crickets)
> summary(crickets.lm2)

Call:
lm(formula = PulseRate ~ Temperature + Species, data = crickets)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0128 -1.1296 -0.3912  0.9650  3.7800 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -7.21091    2.55094  -2.827  0.00858 ** 
Temperature   3.60275    0.09729  37.032  < 2e-16 ***
SpeciesNiv  -10.06529    0.73526 -13.689 6.27e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.786 on 28 degrees of freedom
Multiple R-squared:  0.9896,	Adjusted R-squared:  0.9888 
F-statistic:  1331 on 2 and 28 DF,  p-value: < 2.2e-16

> anova(crickets.lm2)
Analysis of Variance Table

Response: PulseRate
            Df Sum Sq Mean Sq F value    Pr(>F)    
Temperature  1 7894.8  7894.8  2474.0 < 2.2e-16 ***
Species      1  598.0   598.0   187.4 6.272e-14 ***
Residuals   28   89.3     3.2                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(crickets.lm2, which = 1)
```




```
> plot(crickets.lm2, which = 2)
```

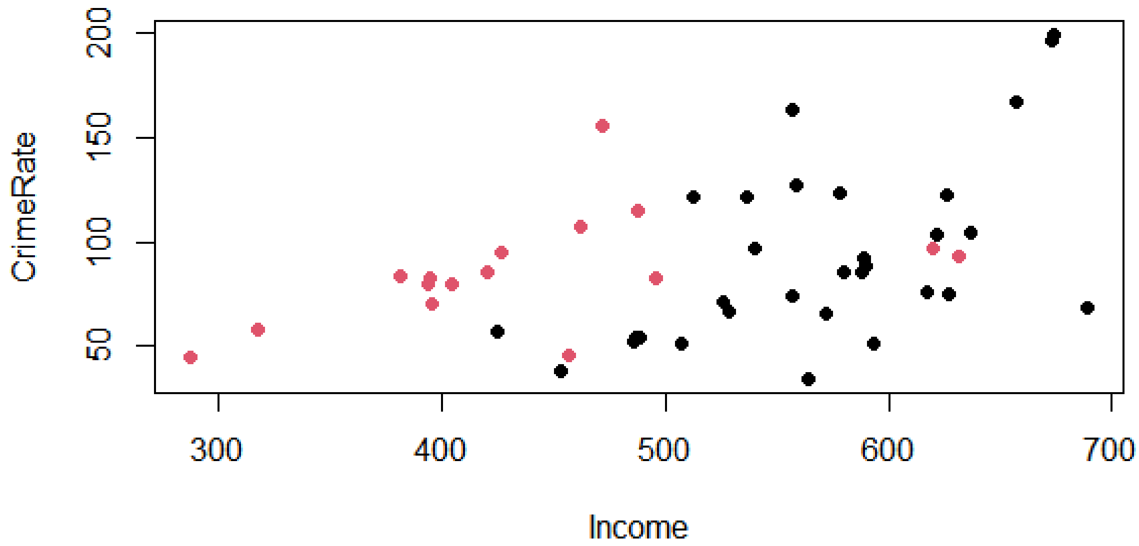



```
> plot(crickets.lm2, which = 4)
```

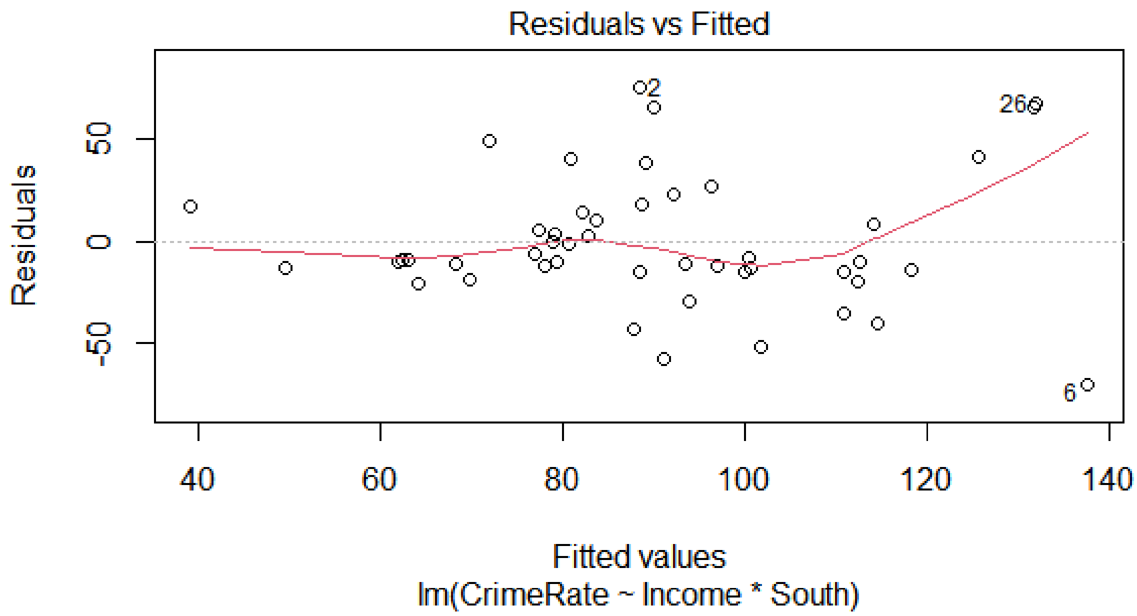



```
> head(crime)
  South Income CrimeRate
1   Yes    394      79.1
2    No    557     163.5
3   Yes    318      57.8
4    No    673     196.9
5    No    578     123.4
6    No    689      68.2
> tail(crime)
   South Income CrimeRate
42    No    489      54.2
43   Yes    496      82.3
44    No    622     103.0
45   Yes    457      45.5
46    No    593      50.8
47    No    588      84.9
```

```
> plot(CrimeRate~Income, data = crime, pch = 19, col = South)
```

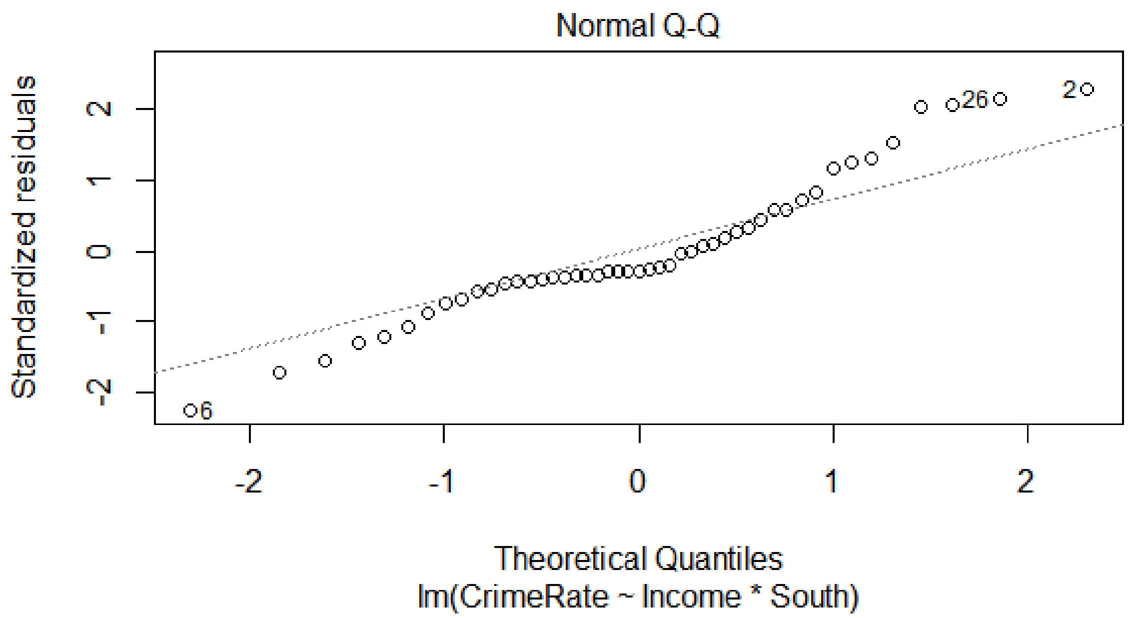



```
> crime.lm1
lm(CrimeRate~Income*South, data = crime)
> summary(crime.lm1)

Call:
lm(formula = CrimeRate ~ Income * South, data = crime)

Residuals:
    Min      1Q  Median      3Q     Max
-69.504 -14.489  -8.958  15.965  74.997

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -119.10748   53.11335  -2.243 0.030138 *
Income             0.37273    0.09274   4.019 0.000231 ***
SouthYes         142.59545   67.92520   2.099 0.041697 *
Income:SouthYes   -0.23162    0.13218  -1.752 0.086849 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.34 on 43 degrees of freedom
Multiple R-squared:  0.3054,   Adjusted R-squared:  0.2569
F-statistic: 6.302 on 3 and 43 DF,  p-value: 0.00122

> anova(crime.lm1)
Analysis of Variance Table

Response: CrimeRate
             Df Sum Sq Mean Sq F value    Pr(>F)
Income        1  13402 13401.5 12.0571 0.001189 **
South         1   4200  4200.0  3.7786 0.058473 .
Income:South  1   3413  3413.0  3.0706 0.086849 .
Residuals    43  47795  1111.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(crime.lm1, which = 1)
```



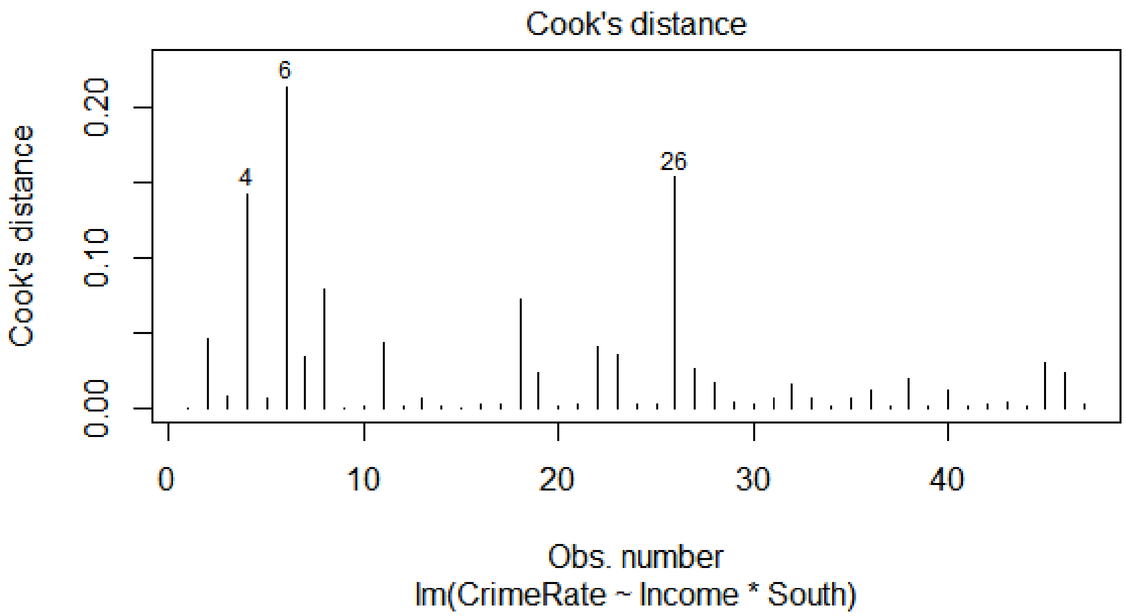
> plot(crim

lm1, which = 2)



> plot(crim

lm1, which = 4)


Cook's distance
Cook's distance
0.00
0.10
0.20
4
6
26
0
10
20
30
40
Obs. number
lm(CrimeRate ~ Income * South)