STAT201 Assignment 4

Robert Ivill 46012819

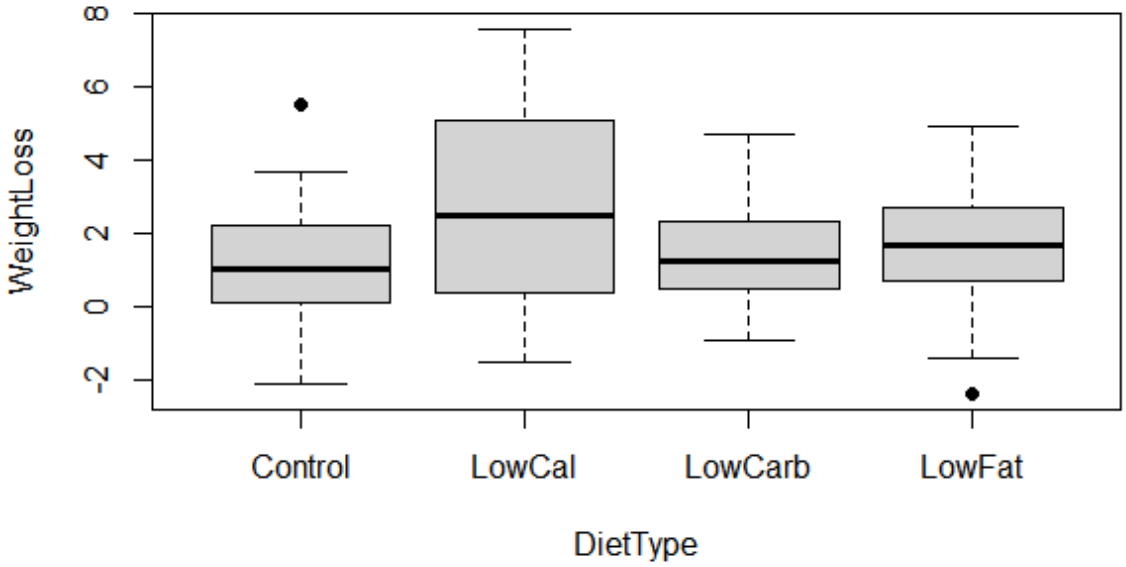
The dataset, diet.csv, has measurements of weight loss resulting from four different diets: low calorie, low carbohydrate, low fat and control. After importing the dataset into R and verifying the correct import with the head and tail functions to compare the first and last values, we could look at a plot to see the differences in the diets. Using the boxplot function, a plot was created (see Appendix). This boxplot clearly shows that the low calorie diet had the highest average diet, at just above 2 units, and the largest spread of result, whilst the other three diets all yielded similar looking results. To investigate if there is statistically significant differences in the diets, a linear model was created and printed. The analysis of variance shows that there is roughly a total of 445 sum of squares, 410 of which is explained by the regression model and 35 from the difference in diet types from their group means, meaning that our data does not have statistically significant differences across their groups. We can confirm this from looking at the residual plots. The residuals vs fitted plot shows that the groups have a constant scatter around zero with no patterns. The normal Q-Q plot shows that the model has a roughly normal distribution, and the cook's distance plot shows there are a few data points that are above 0.04, so these should be left out of the model. We then make a new linear model using the aov function and confirm that it gives the same ANOVA model as lm. We can then use Tukey's honest significance difference, we see that for low carb, low fat and control diets have little differences among them, as they have high adjusted p-values and the lower and upper bounds include 0 respectively. Low calorie diets have low p-values when compared with all other diets, meaning that the low calorie diet is likely to be the only diet that makes a difference in the weight loss of the participants in the study. We can thus propose that a low-calorie diet is the only diet in the study that differs from the control and affects a person's weight loss more, so it would be the one to recommend to someone trying to lose weight.

The dataset, cooking.csv, has information about different cooking methods for baking biscuits and records the consumer satisfaction score of each of the methods. It uses different mixing methods along with using either brown or white sugar during baking. The score was measured using several factors, such as taste, shelf-life, crumbliness, etc. After loading the dataset into R, we, again, confirmed that it was read in correctly with the head and tail functions. We then created two plots using the interactive.plot function to show the differences in scores for the different methods. We see in these plots that there was a high difference in the groups, with brown sugar being better than white, regardless of the mixing method. The mixing methods also had high differences in score, where A had the highest average score, then B had a slightly lower mean score, and C had a much lower score than both other methods. From these results, we can assume that the mixing and sugar variables don't affect each other a lot, and therefore have little to no relationship. To see the if there were statistically significant differences, we created a linear regression model. In the anova summary, we see that the relationship between mixing and sugar has a high p-value ($0.1792 > 0.05$), meaning that there is likely not a significant relationship between the two variables. As the interaction variables have high p-values and are not statistically significant, these can be removed from the model and we can simplify it. With the new model, we print out the summary and anova tables and can comment on them. In the summary table, we see that all

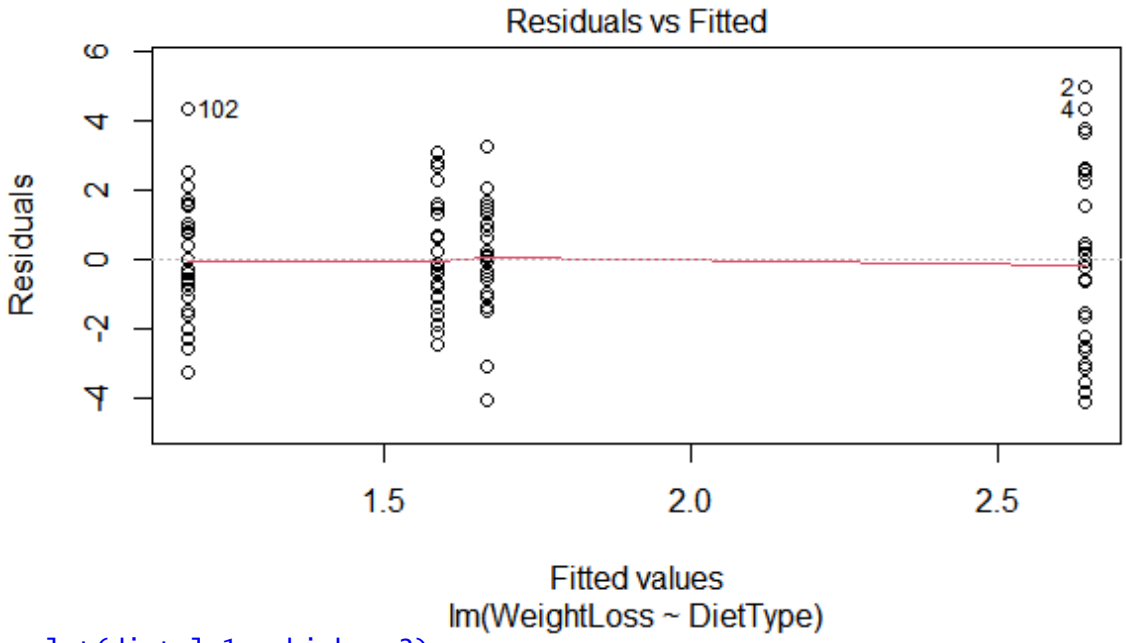
of the variables have low p-values, meaning that they are significant to the model. The p-values in the anova table are also low, meaning that it was a good decision to remove the interaction variable as now the model is simpler and only has statistically significant variables. Looking at the residual plots for the new model, we see that the residuals vs fitted is mostly scattered evenly around zero. The normal Q-Q plot shows that the points in our model follow a normal distribution and the cook's distance plot shows that there are several data points with a distance above 0.04 that have high influence on the model. We cannot reduce the model further by removing either of the effects left in the model as they both are statistically significant to it. We then create another linear model using aov and use it to produce Tukey's HSD test. This shows that every comparison has very high difference in our model, none of which have a range that includes zero and has very low adjusted p-values. From all the data and models that have been produced, it is safe to say that the recommended best way to bake biscuits should be to use brown sugar and mixing method A.

Appendix:

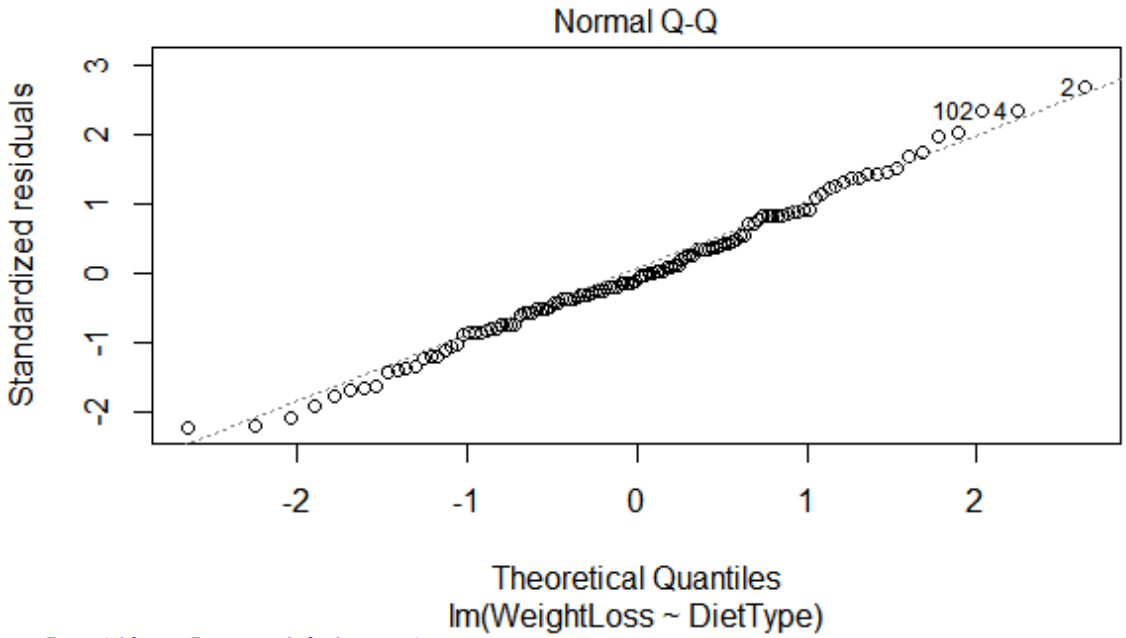
```
> head(diet)
  DietType WeightLoss
1   LowCal        6.3
2   LowCal        7.6
3   LowCal        5.2
4   LowCal        7.0
5   LowCal        3.1
6   LowCal        5.1
> tail(diet)
    DietType WeightLoss
115  Control       -0.4
116  Control       -0.8
117  Control        0.9
118  Control        0.6
119  Control       -0.4
120  Control       -2.1
> boxplot(WeightLoss~DietType, data = diet, pch = 19)
```



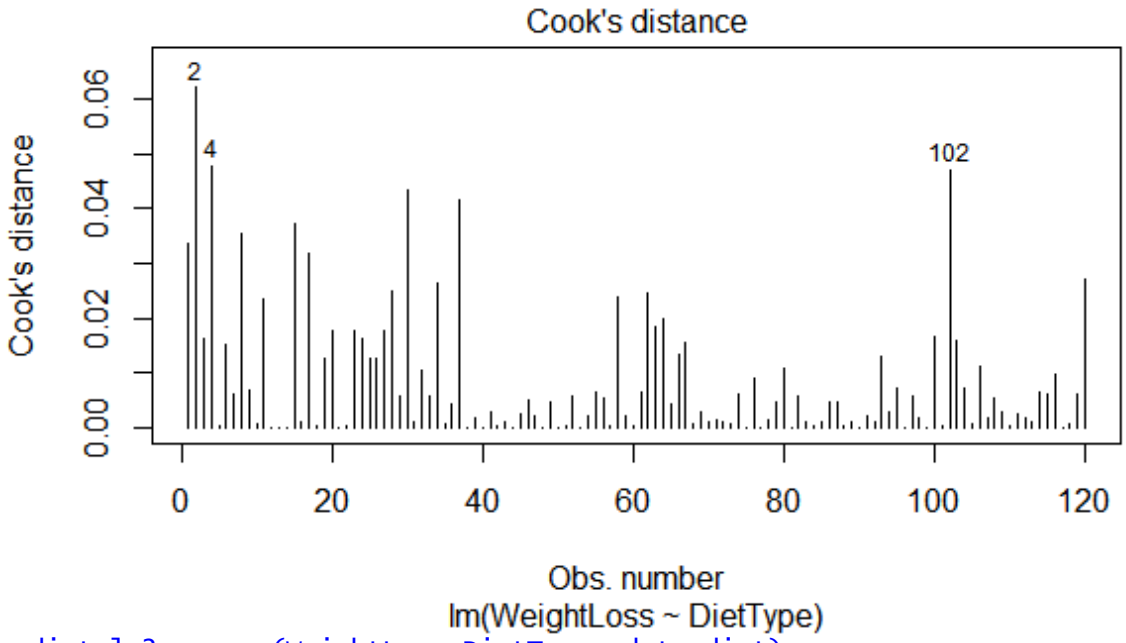

```
> diet.lm1 <- lm(WeightLoss~DietType,data=diet)
> summary(diet.lm1)

Call:
lm(formula = WeightLoss ~ DietType, data = diet)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1433 -1.0817 -0.1367  1.3183  4.9567

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.1800     0.3433   3.437 0.000817 ***
DietTypeLowCal    1.4633     0.4855   3.014 0.003168 **
DietTypeLowCarb   0.4067     0.4855   0.838 0.403991
DietTypeLowFat    0.4867     0.4855   1.002 0.318263
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.88 on 116 degrees of freedom
Multiple R-squared:  0.0779,   Adjusted R-squared:  0.05405
F-statistic: 3.267 on 3 and 116 DF,  p-value: 0.02389

> anova(diet.lm1)
Analysis of Variance Table

Response: WeightLoss
           Df Sum Sq Mean Sq F value  Pr(>F)
DietType    3  34.65 11.5510  3.2666 0.02389 *
Residuals 116 410.18  3.5361
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(diet.lm1, which = 1)
```




```
> plot(diet.lm1, which = 2)
```

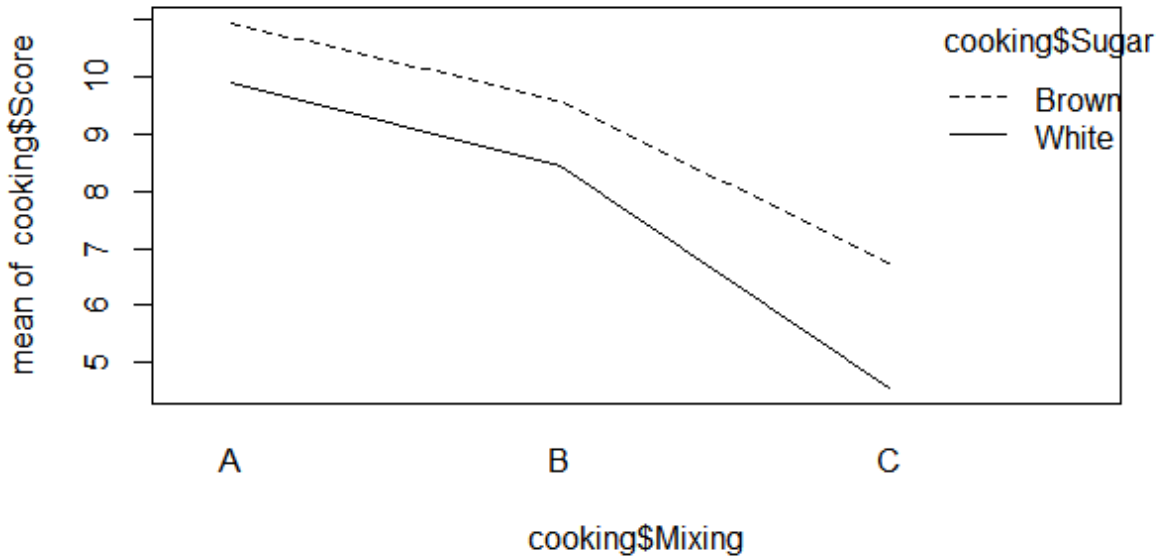


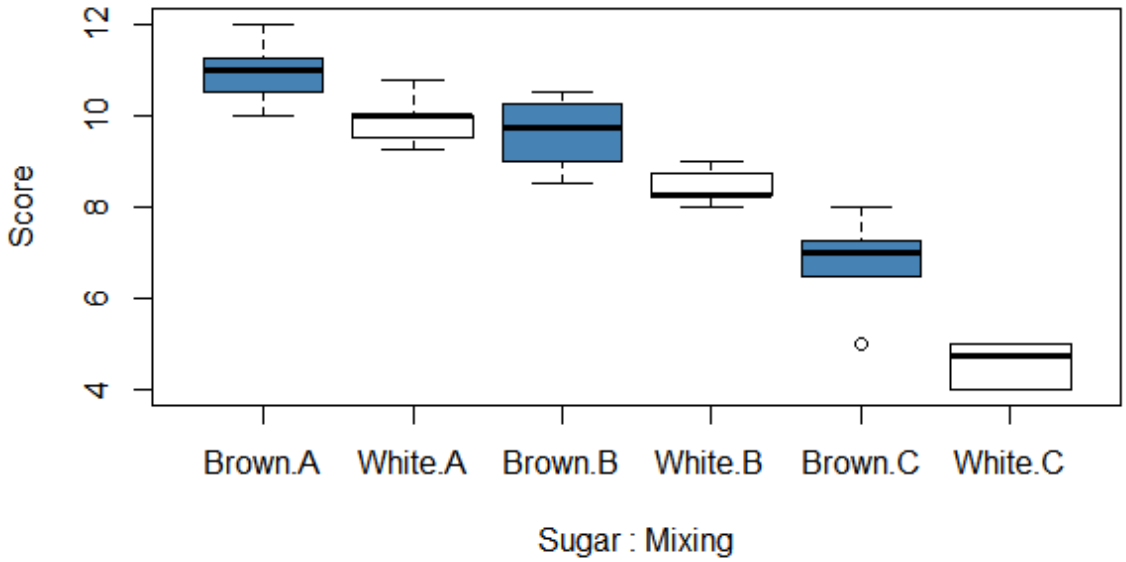

```
> plot(diet.lm1, which = 4)
```




```
> diet.lm2 <- aov(WeightLoss~DietType, data=diet)
> anova(diet.lm1)
Analysis of Variance Table

Response: WeightLoss
           Df Sum Sq Mean Sq F value  Pr(>F)
DietType    3  34.65 11.5510  3.2666 0.02389 *
Residuals 116 410.18  3.5361
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(diet.lm2)
Analysis of Variance Table

Response: WeightLoss
           Df Sum Sq Mean Sq F value  Pr(>F)
DietType    3  34.65 11.5510  3.2666 0.02389 *
Residuals 116 410.18  3.5361
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(diet.lm2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = WeightLoss ~ DietType, data = diet)

$DietType
                     diff        lwr       upr     p adj
LowCal-Control   1.4633333  0.1977247 2.7289419 0.0164995
LowCarb-Control  0.4066667 -0.8589419 1.6722753 0.8364493
LowFat-Control   0.4866667 -0.7789419 1.7522753 0.7483316
LowCarb-LowCal  -1.0566667 -2.3222753 0.2089419 0.1359736
LowFat-LowCal   -0.9766667 -2.2422753 0.2889419 0.1896841
LowFat-LowCarb   0.0800000 -1.1856086 1.3456086 0.9983997
> head(cooking)
  Mixing Sugar Score
1      A White 10.75
2      A White  9.50
3      A White 10.00
4      A White 10.00
5      A White  9.25
6      A Brown 12.00
```

```
> tail(cooking)
   Mixing Sugar Score
25      C White  5.00
26      C Brown  7.00
27      C Brown  7.25
28      C Brown  6.50
29      C Brown  5.00
30      C Brown  8.00
> interaction.plot(cooking$Mixing, cooking$Sugar, cooking$Score)
> boxplot(Score~Sugar*Mixing, col=c("steelblue", "white"), data
+         =cooking)
>
```

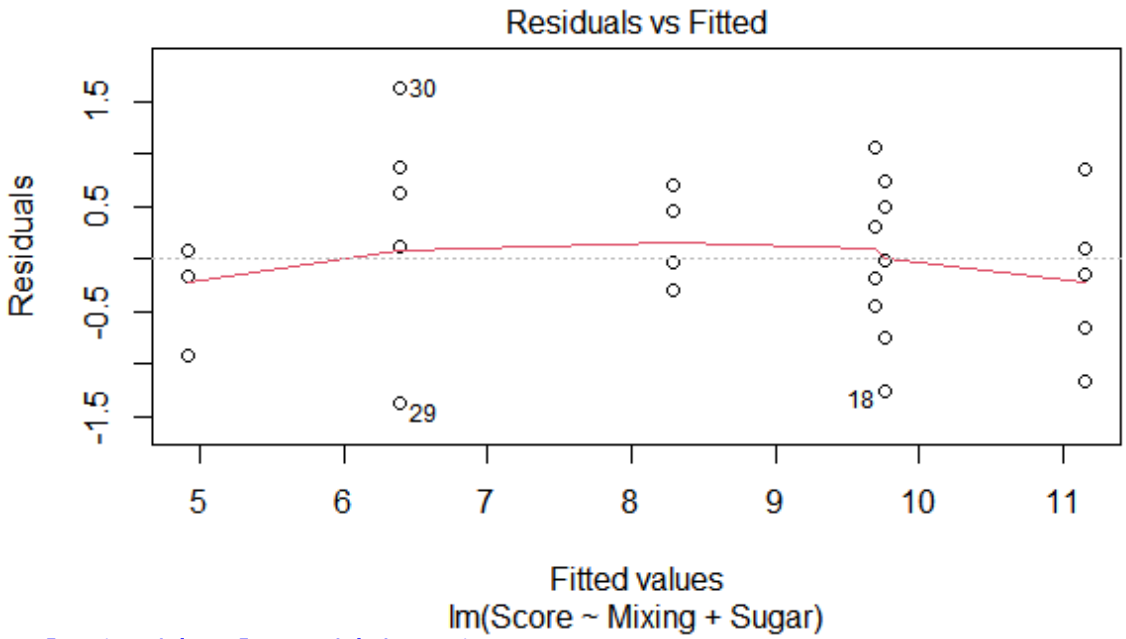




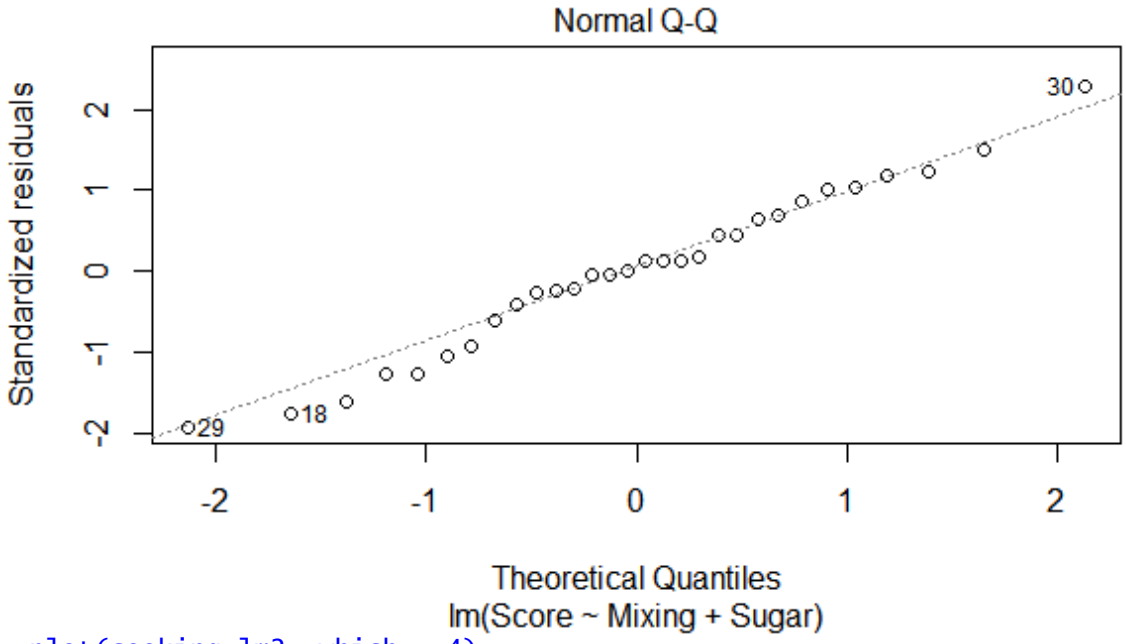

```
> cooking.lm1<-lm(Score~Mixing*Sugar, data = cooking)
```

```
>
> summary(cooking.lm1)

Call:
lm(formula = Score ~ Mixing * Sugar, data = cooking)

Residuals:
   Min     1Q Median     3Q    Max
 -1.75  -0.45   0.10   0.45   1.25

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          10.9500     0.3313  33.047  < 2e-16 ***
MixingB              -1.3500     0.4686  -2.881  0.00822 **
MixingC              -4.2000     0.4686  -8.963 3.99e-09 ***
SugarWhite           -1.0500     0.4686  -2.241  0.03456 *
MixingB:SugarWhite   -0.1000     0.6627  -0.151  0.88132
MixingC:SugarWhite   -1.1500     0.6627  -1.735  0.09551 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7409 on 24 degrees of freedom
Multiple R-squared:  0.9132,   Adjusted R-squared:  0.8952
F-statistic: 50.52 on 5 and 24 DF,  p-value: 5.827e-12

> anova(cooking.lm1)
Analysis of Variance Table

Response: Score
             Df  Sum Sq Mean Sq  F value    Pr(>F)
Mixing        2 120.504  60.252 109.7571 8.399e-13 ***
Sugar         1  16.133  16.133  29.3890 1.436e-05 ***
Mixing:Sugar  2   2.029   1.015   1.8482    0.1792
Residuals    24  13.175   0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> cooking.lm2<-lm(Score~Mixing+Sugar, data = cooking)
>
> summary(cooking.lm2)

Call:
lm(formula = Score ~ Mixing + Sugar, data = cooking)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3833 -0.4042  0.0375  0.4833  1.6167

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.1583     0.2792  39.961  < 2e-16 ***
MixingB      -1.4000     0.3420  -4.094 0.000366 ***
MixingC      -4.7750     0.3420 -13.963 1.36e-13 ***
SugarWhite   -1.4667     0.2792  -5.253 1.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7647 on 26 degrees of freedom
Multiple R-squared:  0.8999,   Adjusted R-squared:  0.8883
F-statistic: 77.89 on 3 and 26 DF,  p-value: 4.054e-13

> anova(cooking.lm2)
Analysis of Variance Table

Response: Score
          Df  Sum Sq Mean Sq F value    Pr(>F)
Mixing     2 120.504  60.252 103.034 4.382e-13 ***
Sugar      1  16.133  16.133  27.589 1.728e-05 ***
Residuals 26  15.204   0.585
---
```

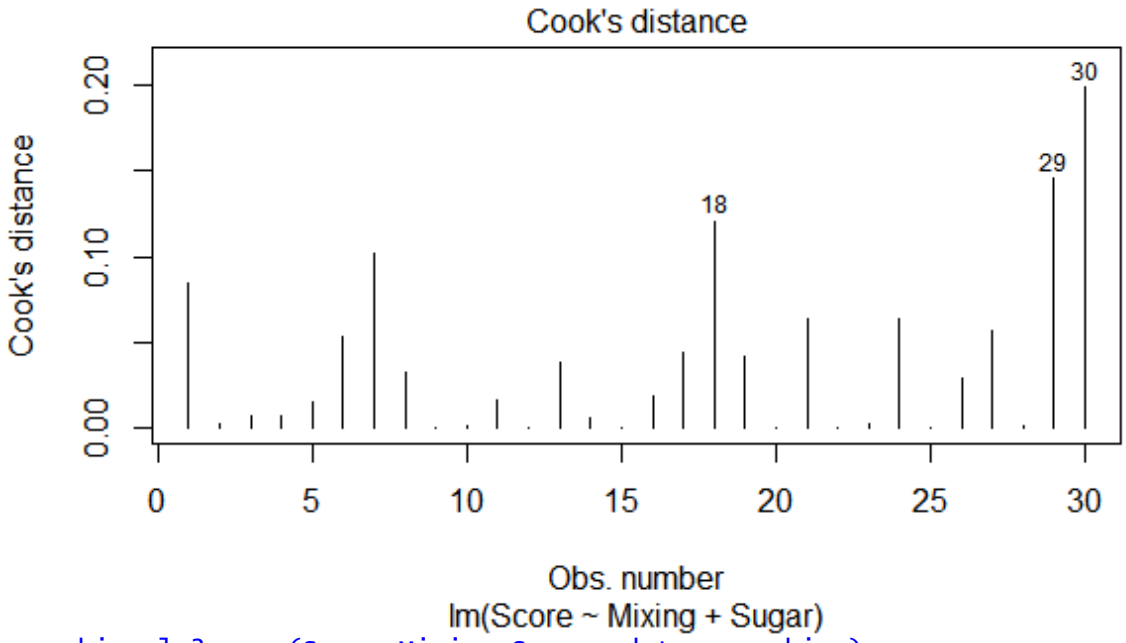
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(cooking.lm2, which =1)
```




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




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




```
> cooking.lm3<-aov(Score~Mixing+Sugar, data = cooking)
> anova(cooking.lm3)
Analysis of Variance Table

Response: Score
          Df  Sum Sq Mean Sq F value    Pr(>F)
Mixing     2 120.504  60.252 103.034 4.382e-13 ***
Sugar      1  16.133  16.133  27.589 1.728e-05 ***
Residuals 26  15.204   0.585
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(cooking.lm3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Score ~ Mixing + Sugar, data = cooking)

$Mixing
      diff       lwr        upr     p adj
B-A -1.400 -2.249801 -0.5501985 0.0010327
C-A -4.775 -5.624801 -3.9251985 0.0000000
C-B -3.375 -4.224801 -2.5251985 0.0000000

$Sugar
                 diff       lwr        upr    p adj
White-Brown -1.466667 -2.040635 -0.8926986 1.73e-05
```