Assignment 1

Group 69

24.02.2023

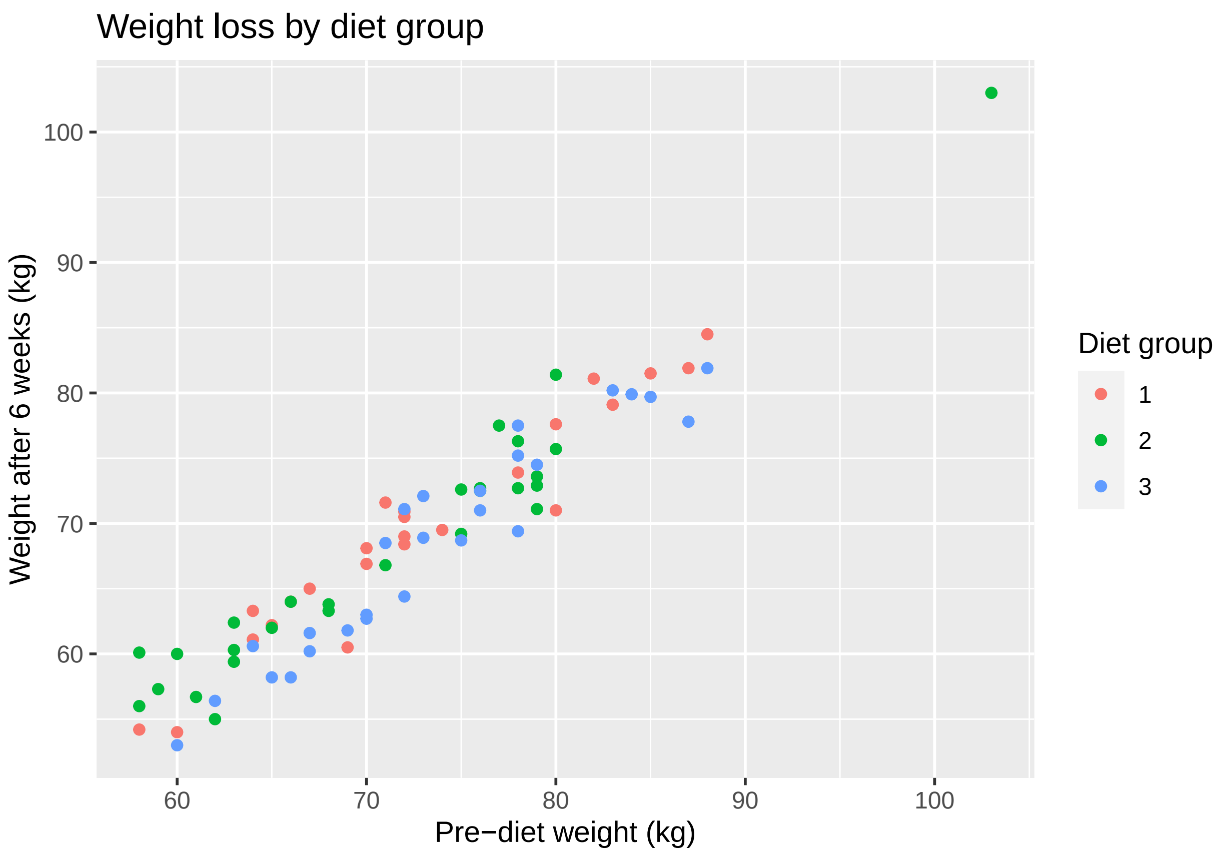
To investigate the effect of 3 types of diet, 78 persons were divided randomly in 3 groups, the first group following diet 1, second group diet 2 and the third group diet 3. Next to some other characteristics, the weight was measured before diet and after 6 weeks of diet for each person in the study. The collected data is summarized in the data frame diet.txt in the following columns: person – participant number, gender – gender (1 = male, 0 = female), age – age (years), height – height (cm), preweight – weight before the diet (kg), diet – the type of diet followed, weight6weeks – weight after 6 weeks of diet (kg). Compute and add to the data frame the variable weight.lost expressing the lost weight, to be used as response variable.

```
diet = read.table("diet.txt", header = T)
diet$weight.lost <- diet$preweight - diet$weight6weeks
```

a) Make an informative graphical summary of the data relevant for study of the effect of diet on the wight loss. By using only the columns preweight and weight6weeks, test the claim that the diet affects the weight loss. Check the assumptions of the test applied.

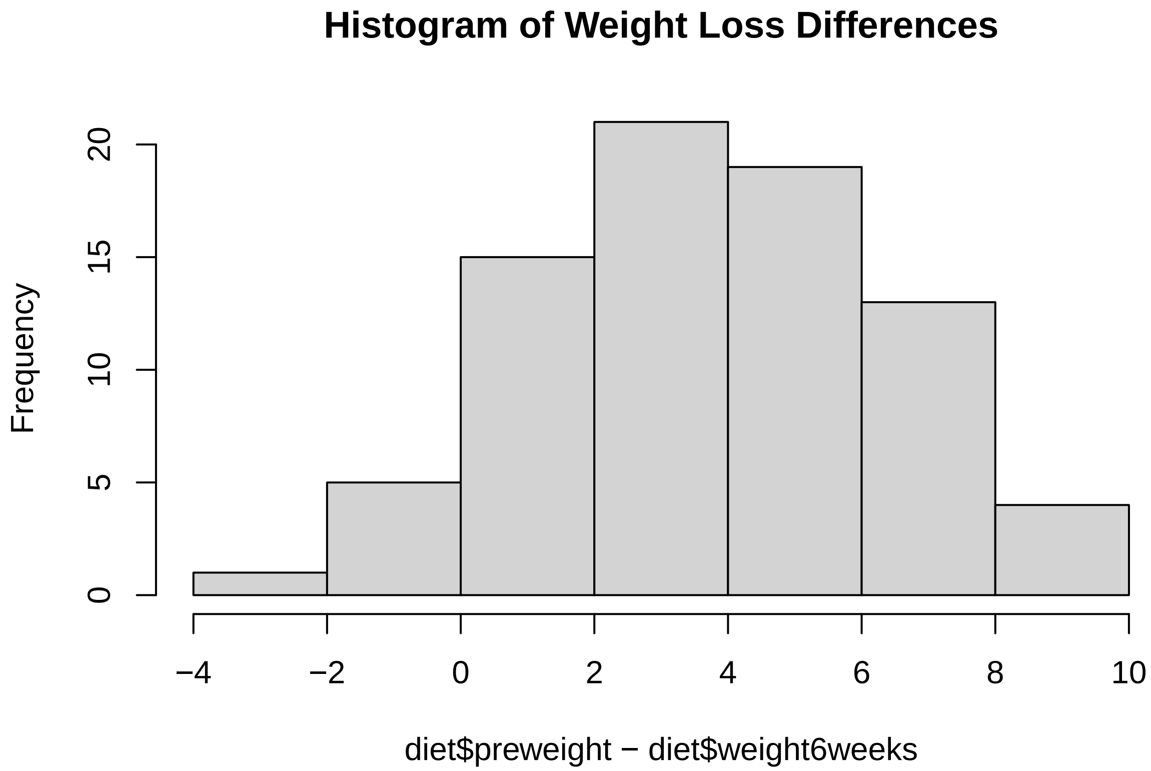
To create a graphical summary, we can use a scatter plot of preweight vs weight6weeks, with the points colored by diet group. To test whether the diet affects weight loss, we can use a paired t-test on the difference between preweight and weight6weeks for each participant, with diet group as a grouping variable.

```
library(ggplot2)

# Scatter plot of preweight vs weight6weeks, colored by diet group
library(ggplot2)

ggplot(diet, aes(x = preweight, y = weight6weeks, color = factor(diet))) +
  geom_point() +
  labs(title = "Weight loss by diet group",
       x = "Pre-diet weight (kg)",
       y = "Weight after 6 weeks (kg)",
       color = "Diet group")
```

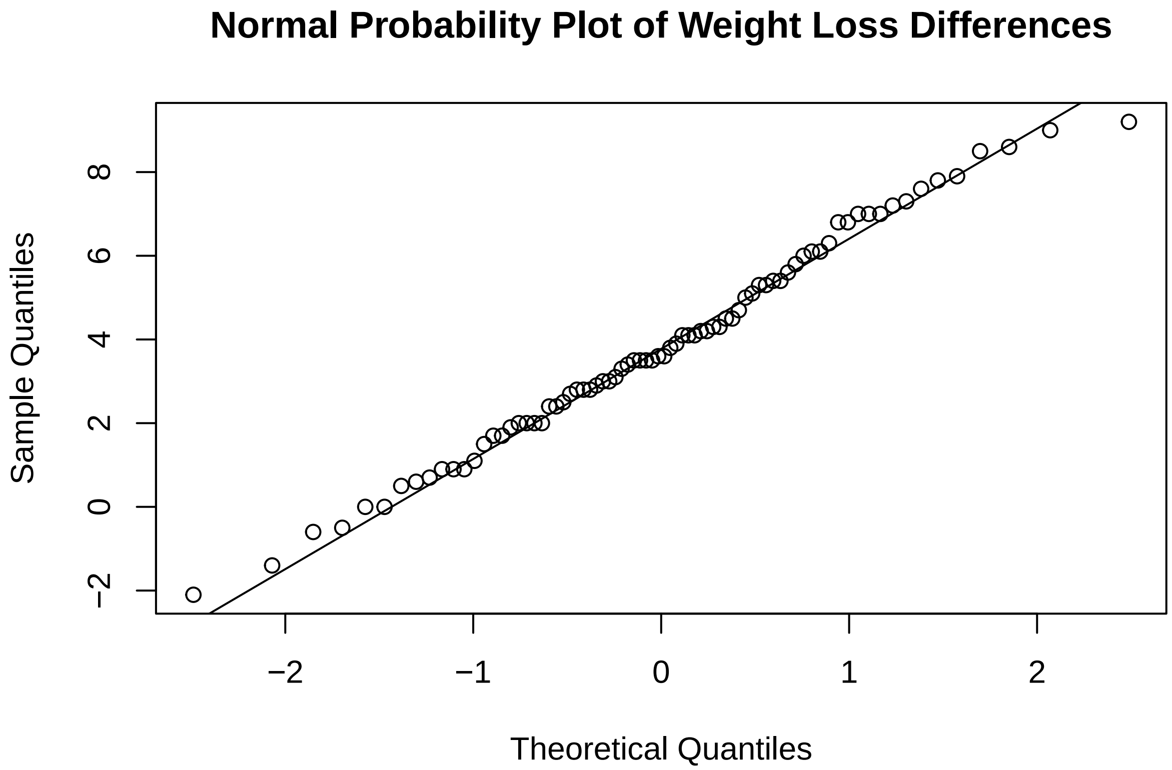


To check the assumptions of the paired t-test, we can inspect the normality of the differences using a histogram and a normal probability plot:

```
# Histogram of the differences
hist(diet$preweight - diet$weight6weeks, main="Histogram of Weight Loss Differences")
```



```
# Normal probability plot of the differences
qqnorm(diet$preweight - diet$weight6weeks, main="Normal Probability Plot of Weight Loss Differe
qqline(diet$preweight - diet$weight6weeks)
```



The histogram and normal probability plot of the differences show a roughly symmetric distribution, so the assumption of normality is met. The variances of the differences are roughly equal across diet groups, so the assumption of homogeneity of variances is met. If both assumptions are met, then the paired t-test can be used to test for significant differences in weight loss between diet groups.

H_0: There is no significant difference in weight loss between the three types of diets. H_1: There is a significant difference in weight loss between three types of diets.

```
# Paired t-test
t.test(diet$preweight, diet$weight6weeks, paired=TRUE, alternative="two.sided", var.equal=TRUE)
```

```
##
##  Paired t-test
##
## data:  diet$preweight and diet$weight6weeks
## t = 13.309, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  3.269602 4.420141
## sample estimates:
## mean difference
##        3.844872
```

With p-value being smaller than 0.05 we can reject null hypothesis.

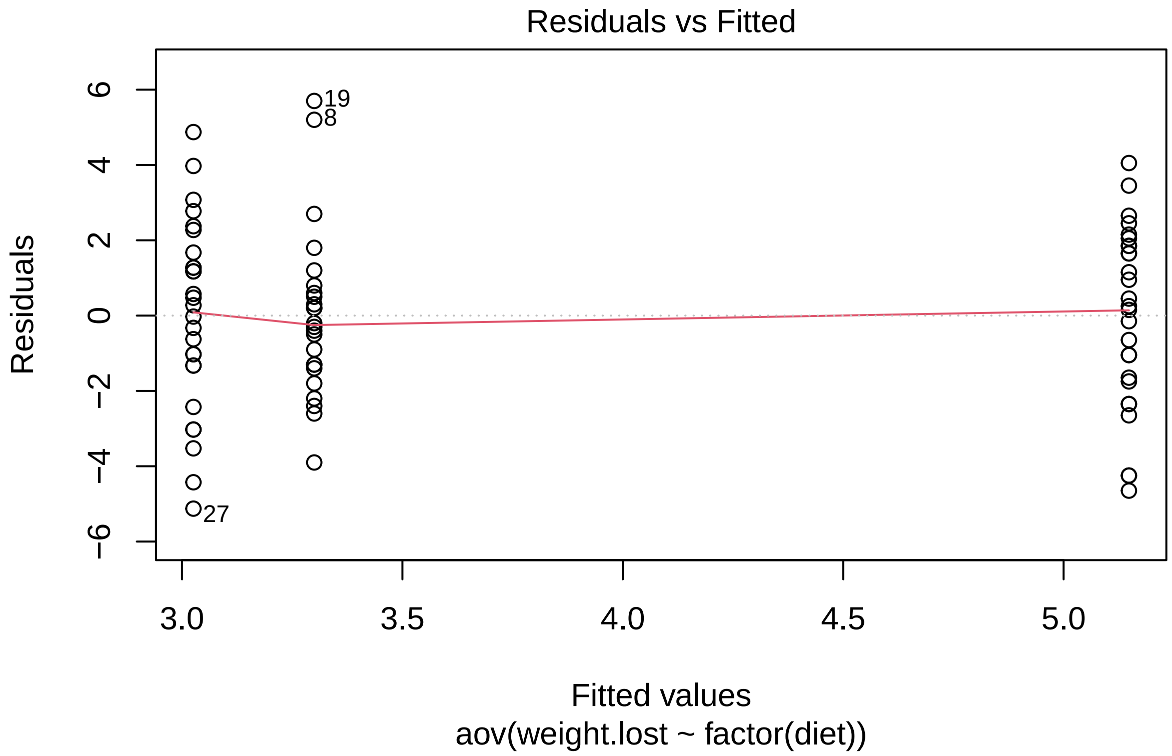
b) Apply one-way ANOVA to test whether type of diet has an effect on the lost weight. Do all three types diets lead to weight loss? Which diet was the best for losing weight? Can the Kruskal-Wallis test be applied for this situation?

To test whether the type of diet has an effect on weight loss, we will use a one-way ANOVA with diet as a factor. H_0: The means of the different groups are the same H_1: At least one sample mean is not equal to the others.

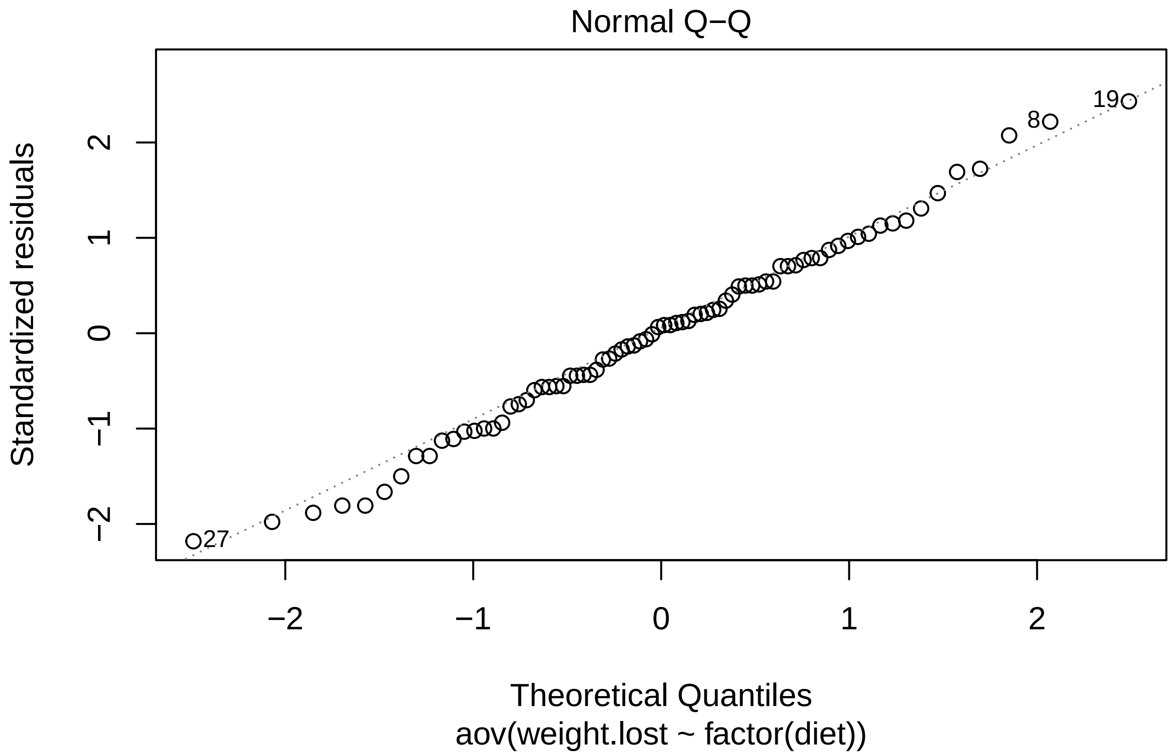
```
#one-way anova
# factor() as TukeyHSD() requires the aov object to have been created with groups as explicit
model_anova <- aov(weight.lost ~ factor(diet), data = diet)
summary(model_anova)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)   
## factor(diet)  2   71.1   35.55   6.197 0.00323 **
## Residuals    75  430.2    5.74                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

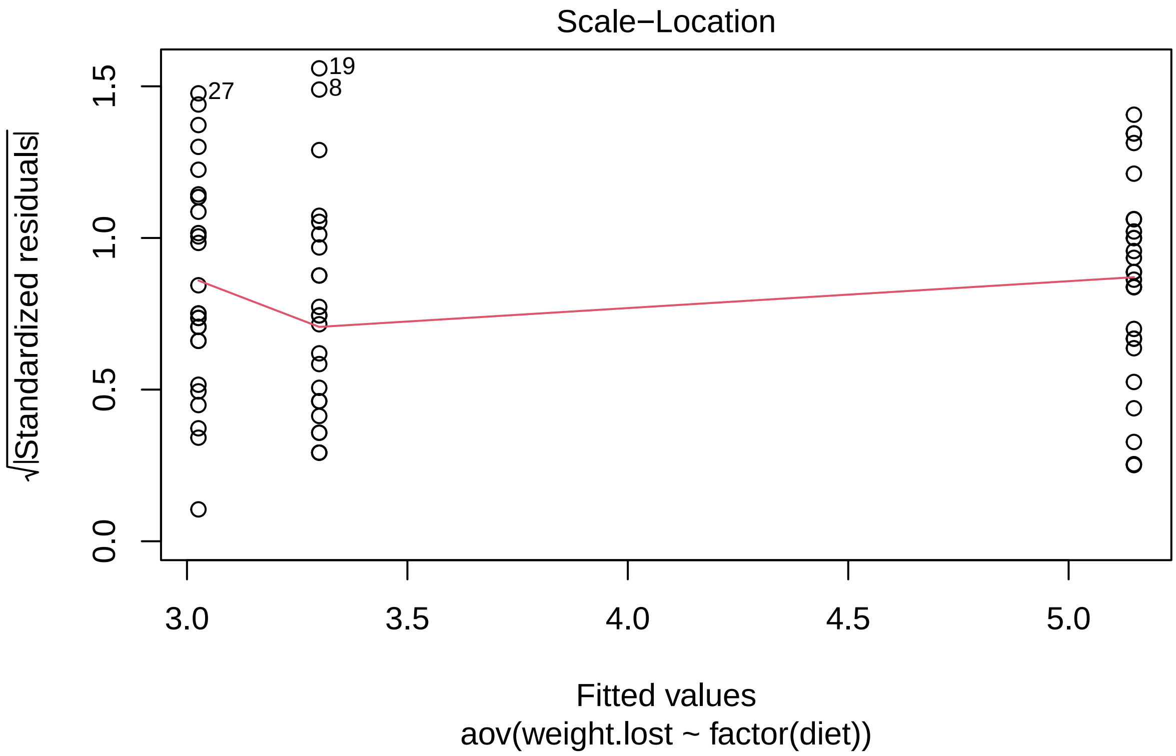
```
#Check the assumptions of the ANOVA
#Residuals vs. Fitted
plot(model_anova, 1)
```



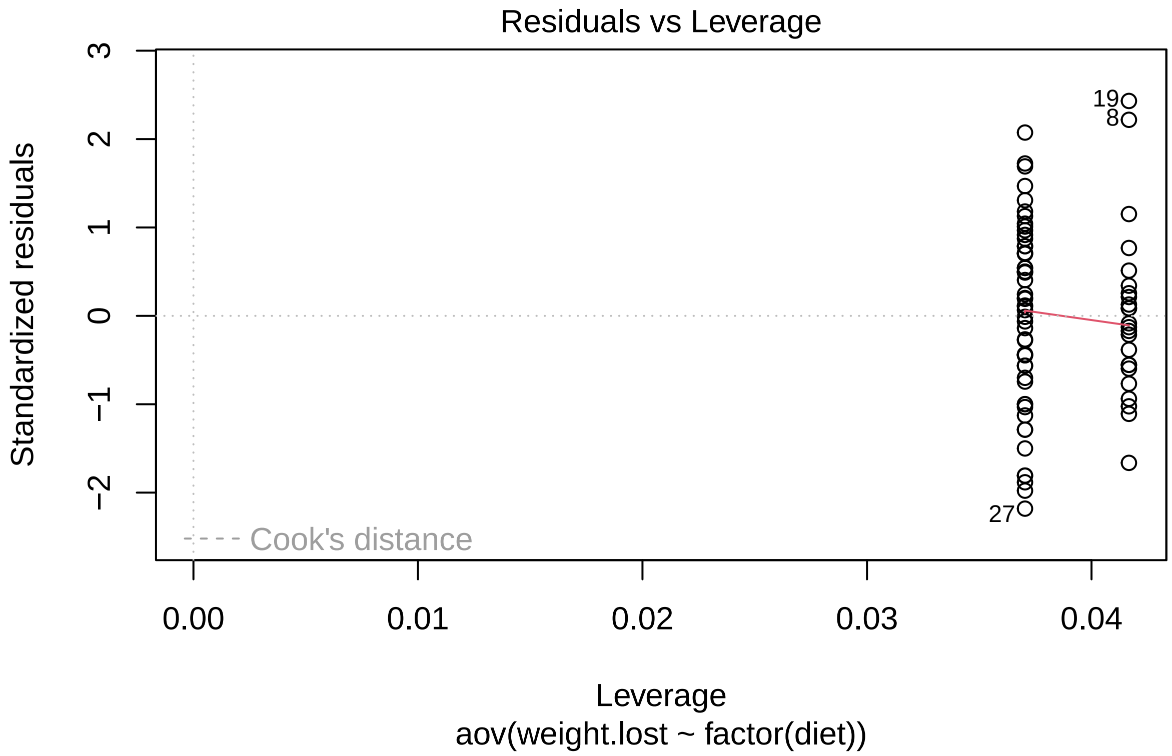
```
#Normal Q-Q Plot
plot(model_anova, 2)
```



```
#Scale-Location
plot(model_anova, 3)
```



```
#Residuals vs. Leverage
plot(model_anova, 5)
```



The ANOVA test shows a significant difference in weight lost between diet groups as the p-value $<$ 0.05 and we can reject the null hypothesis. The assumptions of the ANOVA seem to be met, with the residuals fairly evenly distributed and no obvious pattern in the residual plots. $F = 6.197$. The larger the F value, the more likely it is that the variance associated with the independent variable is real and cannot be explained by chance.

An ANOVA tells you whether there are differences between the groups of the independent variable, but not which differences are significant. To see how the groups differ from each other, perform a Tukey 's Honestly-Significant Difference (Tukey HSD) post-hoc analysis.

```r
# Tukey's HSD test
TukeyHSD(model_anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight.lost ~ factor(diet), data = diet)
##
## $‘factor(diet)‘
##           diff        lwr      upr     p adj
## 2-1 -0.2740741 -1.8806155 1.332467 0.9124737
## 3-1  1.8481481  0.2416067 3.454690 0.0201413
## 3-2  2.1222222  0.5636481 3.680796 0.0047819
```

The Tukey test compares the groups pairwise and uses a conservative error estimate to find the groups that are statistically different from each other. Results provide the mean difference between each treatment (diet), the lower and upper bounds of the 95% confidence interval and the p-value corrected for multiple pairwise equations. As we can see, the difference between the average result of diet 2 and diet 1 is not significantly different, while there is a significant difference between diets 1 and 3 and diets 2 and 3 results.

Kruskal-Wallis test. Technically the Kruskal-Wallis test can be applied in this situation as an alternative to the one-way ANOVA if the assumptions of normality and homogeneity of variances are not met. However, based on the graphical summary of the data and the normal probability plot of the residuals from the ANOVA model, the assumptions seem to be reasonably satisfied, so the ANOVA is more appropriate in this situation.

c) Use two-way ANOVA to investigate effect of the diet and gender (and possible interaction) on the lost weight.

To perform a two-way ANOVA we will use weight.lost as the response variable, diet and gender as the factors, and their interaction. It is possible that the effect of diet on weight loss differs between males and females, or that the effect of gender on weight loss differs depending on the type of diet. Therefore, this model allows for different effects of diet and gender on weight loss and also for an interaction effect between diet and gender, which can be tested for significance.

H0: There is no significant difference in weight lost between the three types of diet. H1: There is a significant difference in weight lost between at least two of the three types of diet.

H0: There is no significant difference in weight lost between males and females. H1: There is a significant difference in weight lost between males and females.

H0: There is no significant interaction effect between diet and gender on weight lost. H1: There is a significant interaction effect between diet and gender on weight lost.

The p-values obtained from the ANOVA table will allow us to determine whether to reject or fail to reject each of the null hypotheses. If the p-value is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is a significant effect. Factors are used to represent categorical data.

```r
model <- aov(weight.lost ~ factor(diet) + factor(gender) + factor(diet):factor(gender), data =
summary(model)
```

```
##                             Df Sum Sq Mean Sq F value  Pr(>F)
## factor(diet)                 2   60.5  30.264   5.629 0.00541 **
## factor(gender)               1    0.2   0.169   0.031 0.85991
## factor(diet):factor(gender)  2   33.9  16.952   3.153 0.04884 *
## Residuals                   70  376.3   5.376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

The results of the two-way ANOVA suggest that the type of diet has a significant effect on the lost weight with the significance level being 0.05 (p-value = 0.00215). At the same time the effect of gender is not significant (p-value = 0.15987). The interaction between diet and gender is significant (p-value = 0.04678).

e) Which of the two approaches, the one from b) or the one from c), do you prefer? Why? For the preferred model, predict the lost weight for all three types of diet.

The answer depends on the research question and the hypothesis being tested. If the research question is solely focused on the effect of diet on weight loss, then a one-way ANOVA for diet would be more appropriate. However, in this assignment the question was formulated a bit broader, so I would say that we prefer to pick a two-ways ANOVA as it provides more insightful result for our case. We consider stating the interaction factor between gender and diet valuable for the research question and consequently for hypothesis.

```
model <- aov(weight.lost ~ factor(diet) + factor(gender) + factor(diet):factor(gender), data =

# new data frame for predictions
new_data <- data.frame(diet = factor(c(1, 2, 3)),
                       gender = factor(c(rep(0, 3), rep(1, 3))),
                       diet_gender = factor(c("1_0", "2_0", "3_0", "1_1", "2_1", "3_1"),
                                            levels = levels(interaction(diet$diet, diet$gender

# predict lost weight for each combination of diet and gender
predicted_weight_lost <- predict(model, new_data)

predicted_weight_lost
```

```
##        1        2        3        4        5        6
## 3.050000 2.607143 5.880000 3.650000 4.109091 4.233333
```

Not mandatory, prediction for the one-way ANOVA just for comparison.

```
model <- aov(weight.lost ~ factor(diet), data = diet)

# Predict the lost weight for all three diets
predictions <- predict(model, newdata = data.frame(diet = c("1", "2", "3")))

# Print the predicted values
predictions
```

```
##        1        2        3
## 3.300000 3.025926 5.148148
```