Assignment 1

Group 69

24.02.2023

Assignment 1

Exercise 1

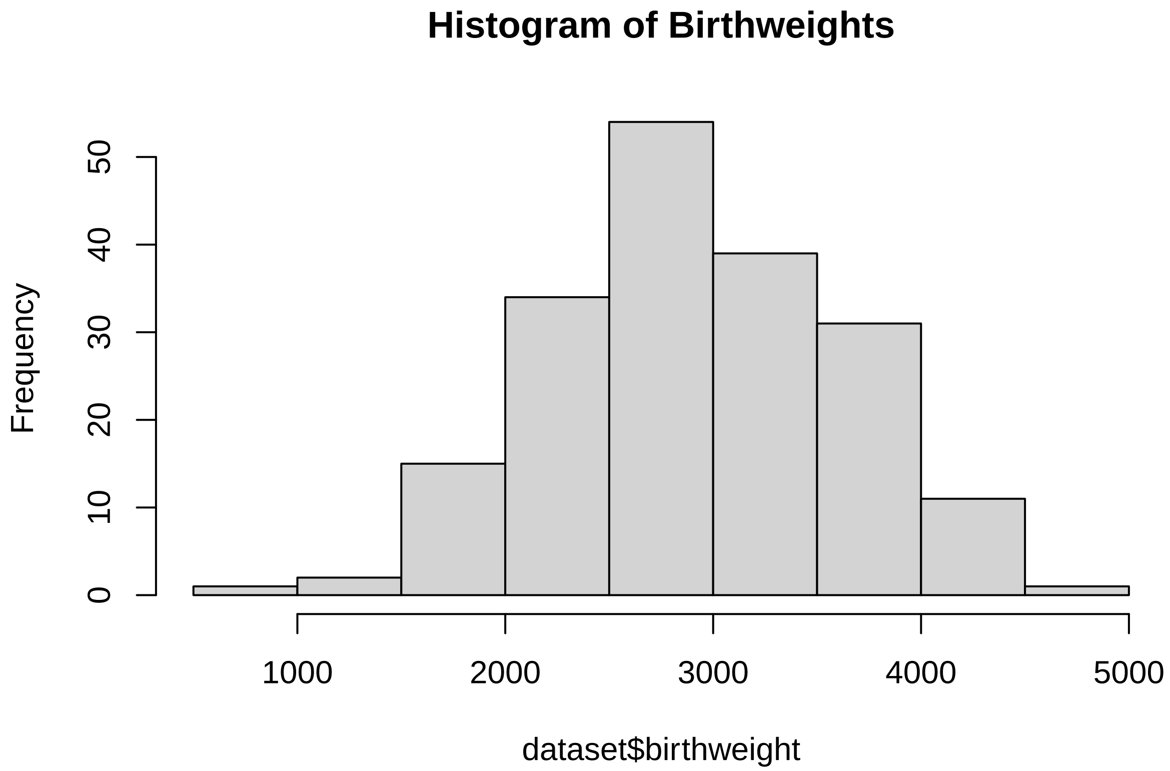
The data set birthweight.txt contains the birthweights (in grams) of 188 newborn babies. Denote the underlying mean birthweight by .

a) Check normality of the data. Assuming normality (irrespective of your conclusion about normality), construct a bounded 96%-CI for . Evaluate the sample size needed to provide that the length of the 96%-CI is at most 100. Compute a bootstrap 96%-CI for and compare it to the above CI.

To start with we import the birthweight.txt dataset and store it to the respective variable.

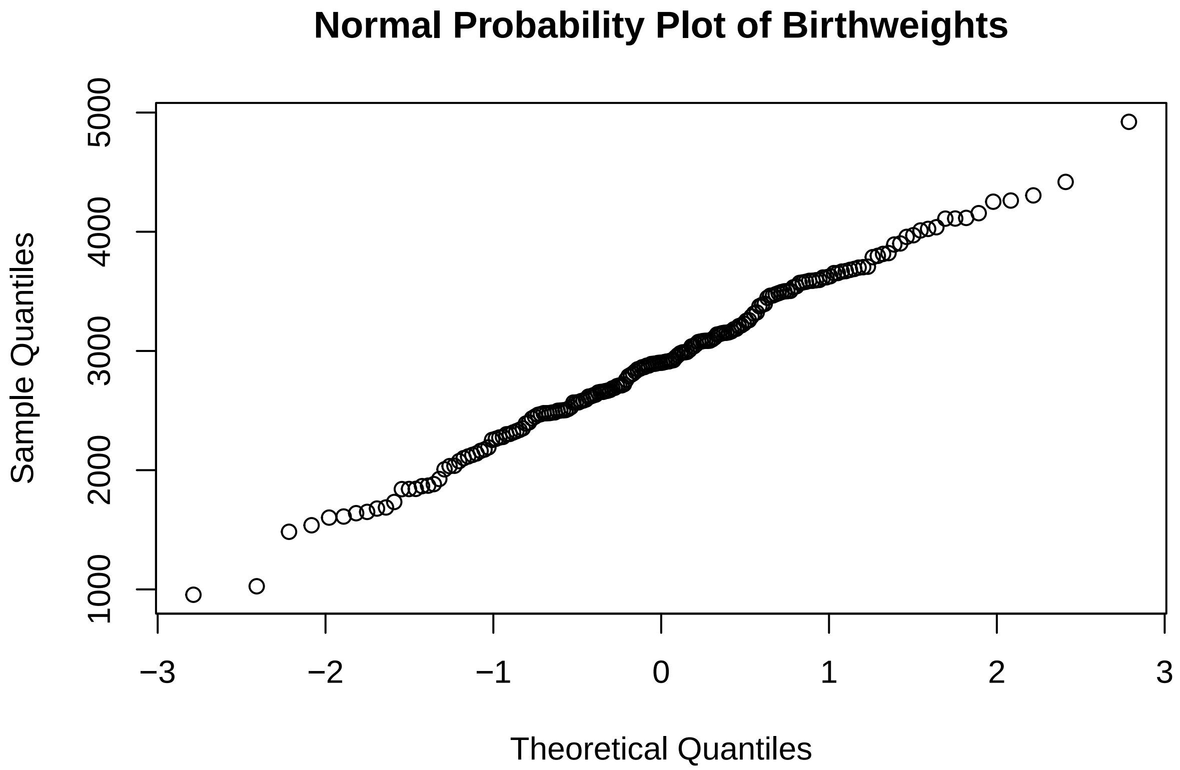
To make an assumption about the distribution we display the data using the following function.

```
dataset = read.table("birthweight.txt", header = T)
hist(dataset$birthweight, main="Histogram of Birthweights")
```



After running the histogram we observe a bell-shaped form, which indicates that the data is normally distributed. Although we can clearly see the shape, the histogram alone isn't a good indicator of normally distributed data.

```
qqnorm(dataset$birthweight, main="Normal Probability Plot of Birthweights")
```



We run the qqnorm function to confirm the distribution. A nearly straight line indicates that the distribution is indeed normal.

To construct the 96%-CI for the birthweight the following calculations are performed.

```
n <- length(dataset$birthweight)
se <- sd(dataset$birthweight) / sqrt(n)
t_star <- qt(0.98, df=n-1)
ci <- c(mean(dataset$birthweight) - t_star*se, mean(dataset$birthweight) + t_star*se)
```

To get the 96%-CI of the mean we find the sample size (n), standard error (se), and critical value of t-distribution (t_star). We apply the formula¯ ± t * (se/√n).

To evaluate the sample size needed to provide that the length of the 96%-CI is at most 100.

```
t_star <- qt(0.98, df=1e6)
n <- ceiling(((t_star * sd(dataset$birthweight) / 50) ^ 2))
```

Here the t_star uses other df value that denotes an arbitrary large sample size that we need for our calculation.

To compute a bootstrap 96%-CI for define a function to generate a bootstrap sample and compute the mean of the sample.

```
boot_mean <- function(data) {
  boot_sample <- sample(data, replace=TRUE, size=length(data))
  mean(boot_sample)
}
```

Generate 10000 bootstrap samples and store the means in a vector.

```
boot_means <- replicate(10000, boot_mean(dataset$birthweight))
```

Calculate the 2.5th and 97.5th percentiles of the bootstrapped means to obtain the 96% CI.

```
boot_ci <- quantile(boot_means, c(0.025, 0.975))
```

So the values of the bootstrap CI to the one obtained using the t-distribution could be compared.

```
ci
```

```
## [1] 2808.084 3018.501
```

```
boot_ci
```

```
##     2.5%    97.5%
## 2816.939 3013.859
```

b) An expert claims that the mean birthweight is bigger than 2800 gram. Verify this claim by using a relevant t-test, explain the meaning of the CI in the R-output for this test. Also propose and perform a suitable sign tests for this problem.

To verify the claim a one-sample t-test could be performed with null hypothesis H0: = 2800 and alternative hypothesis Ha: > 2800. Here, represents the true population mean birthweight.

```
t.test(dataset, mu=2800, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  dataset
## t = 2.2271, df = 187, p-value = 0.01357
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829.202      Inf
## sample estimates:
## mean of x
##  2913.293
```

The output of this test includes the test statistic (t-value), the p-value, and a confidence interval for the population mean. The p-value represents the probability of observing a sample mean at least as extreme as the one we obtained, assuming that the null hypothesis is true.

The p-value is 0.01357 which is less than the significance level, so we reject the null hypothesis and conclude that there is strong evidence to suggest that the mean birthweight is indeed greater than 2800 gram. The confidence interval also supports this conclusion, as it does not include 2800 gram.

Alternatively a sign test can be performed to confirm the hypothesis.

```
n <- length(dataset$birthweight)
sign_test <- binom.test(sum(dataset$birthweight > 2800), n, p=0.5, alternative="greater")
sign_test
```

```
##
##  Exact binomial test
##
## data:  sum(dataset$birthweight > 2800) and n
## number of successes = 107, number of trials = 188, p-value = 0.03399
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5065781 1.0000000
## sample estimates:
## probability of success
##              0.5691489
```

The output of this test will include the test statistic (the number of observations above 2800 gram), the p-value, and a confidence interval for the true proportion of observations that are above 2800 gram. Assuming the output p-value we reject the null hypothesis in favor of alternative hypothesis.

c) Propose a way to compute the powers of the t-test and sing test from

    b) at some > 2800, comment.

To compute the power of the t-test, we need to specify a hypothetical value of the mean birthweight, denoted by 1, which is greater than 2800g. Let's say we choose 1 = 2850g. Then we can use the following approach to compute the power.

For the t-test, we can compute the t-value and p-value for the null hypothesis H0: = 2800g versus the alternative hypothesis Ha: > 2800g, using the formula: t = (- 0)/(sd/sqrt(n)).

The p-value can then be obtained from the t-distribution with n - 1 degrees of freedom. Once we have the p-value, we can compute the power of the test for the specified value of 1.

For the sign test the p-value can be computed for the null hypothesis H0: P(X > 2800) = 0.5 against the alternative hypothesis Ha: P(X > 2800) > 0.5 using the binomial distribution: p = sum(dbinom(k:n, n, 0.5)),

k - number of observations greater than 2800g.

The p-value is then compared to the significance level to determine whether to reject or fail to reject the null hypothesis.

d) Let p be the probability that birthweight of a newborn baby is less than 2600 gram. Using asymptotic normality, the expert computed the left end p^l = 0.25 of the confidence interval [p^{l,p}r] for p. Recover the whole confidence interval and its confidence level.

If the probability that birthweight of a newborn baby is less than 2600 gram is p, then the number of newborn babies out of 188 with birthweights less than 2600 gram follows a binomial distribution with parameters n = 188 and p. Assuming that the sample size is large enough, we can approximate the distribution of the sample proportion by a normal distribution with mean p and variance p(1-p)/n.

The expert computed the left end p^l = 0.25 of the confidence interval [p^l, p^r] for p. Since the normal distribution is symmetric, we have p^r = 1 - p^l = 1 - 0.25 = 0.75.

To recover the whole confidence interval, we need to find the z-value that corresponds to the left tail probability of 0.25. We can use the qnorm() function in R to find this value.

```
z_star <- qnorm(0.25)
z_star
```

```
## [1] -0.6744898
```

Using the normal approximation to the binomial distribution, the confidence interval for p is.

```
p_hat <- sum(dataset$birthweight < 2600) / nrow(dataset)
se_p <- sqrt(p_hat * (1 - p_hat) / nrow(dataset))
ci_p <- p_hat + z_star * se_p + c(0, 1) * 1.96 * se_p
ci_p
```

```
## [1] 0.3066602 0.3738650
```

e) The expert also reports that there were 34 male and 28 female babies among 62 who weighted less than 2600 gram, and 61 male and 65 female babies among the remaining 126 babies. The expert claims that the mean weight is different for male and female babies. Verify this claim by an appropriate test.

To test if the mean birthweight is different for male and female babies, we can perform a two-sample t-test.

First, we can split the data into two groups based on the gender of the newborns.

Exercise 2.

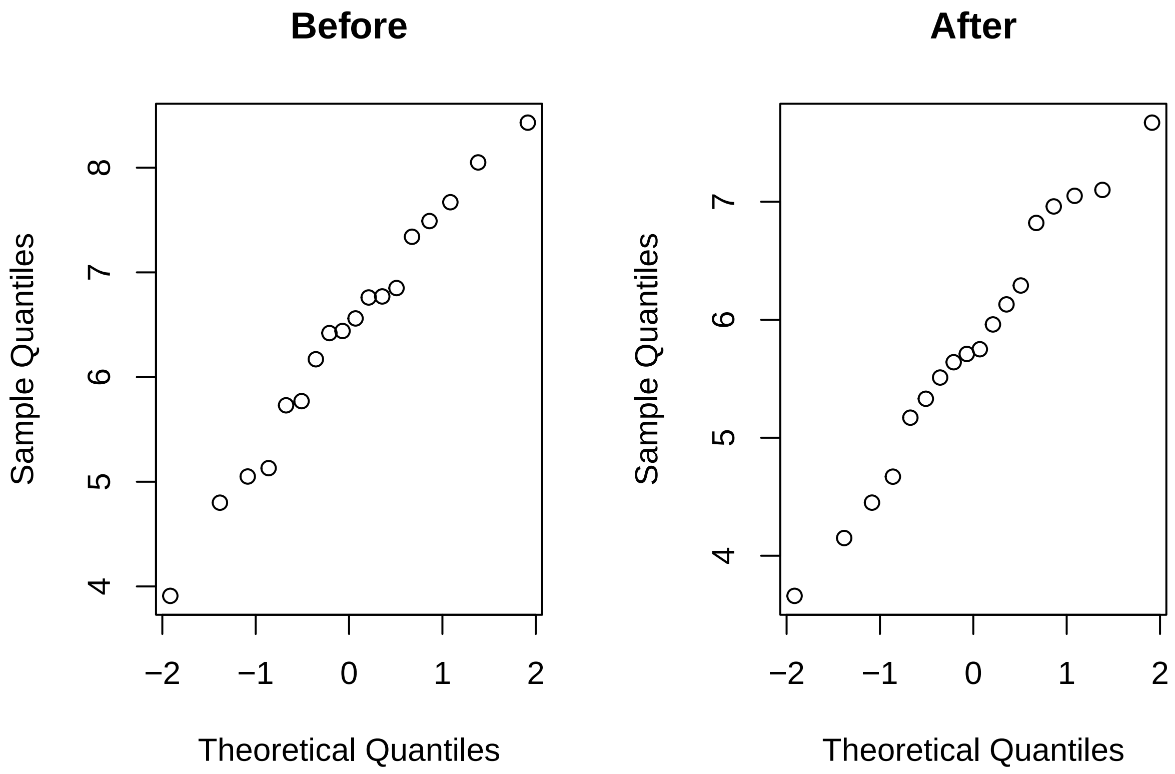
A study tested whether cholesterol was reduced after using a certain brand of margarine as part of a low fat low cholesterol diet. The data set cholesterol.txtcontains information on 18 people using margarine to reduce cholesterol: columns Before and After8weeks contain the cholesterol level (mmol/L) respectively before the diet and after 8 weeks on the diet.

a) Make some relevant plots of this data set, comment on normality. Are there any inconsistencies in the data? Investigate whether the columns Before and After8weeks are correlated.

```
data = read.table("cholesterol.txt", header = T)
attach(data)
head(data)
```

```
##   Before After8weeks
## 1   6.42        5.75
## 2   6.76        6.13
## 3   6.56        5.71
## 4   4.80        4.15
## 5   8.43        7.67
## 6   7.49        7.05
```

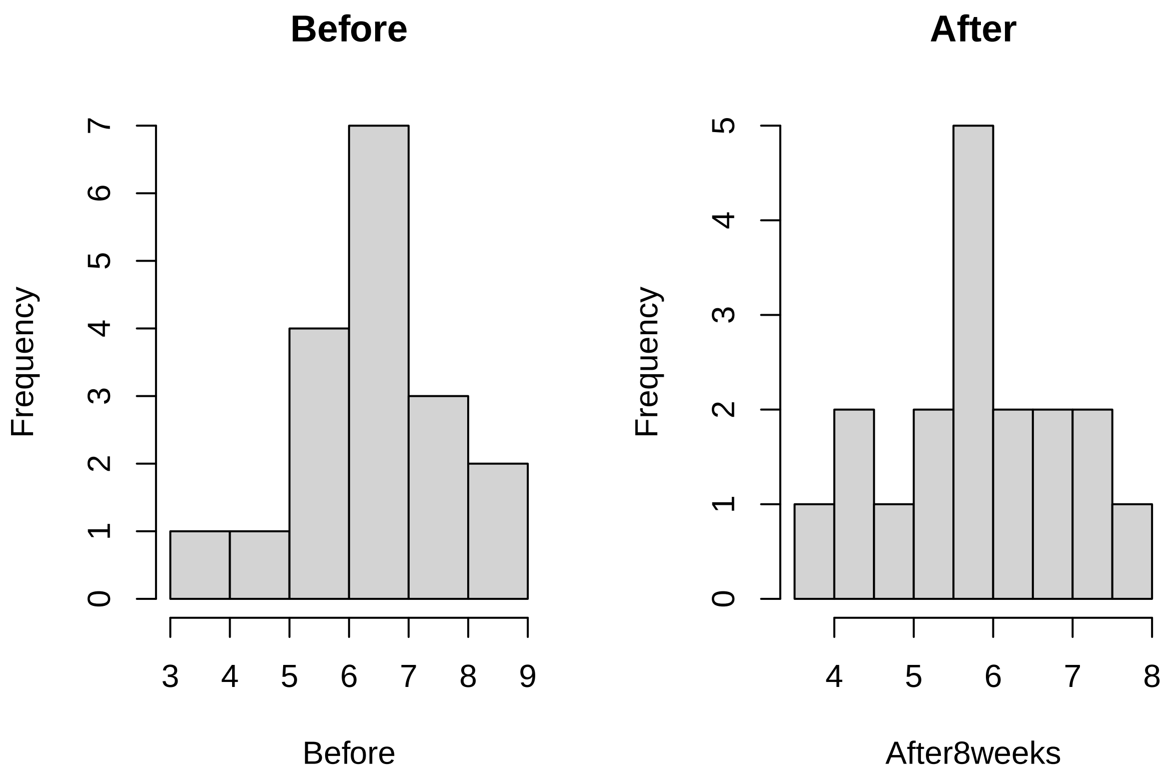
```
par(mfrow=c(1,2))
qqnorm(Before, main="Before")
qqnorm(After8weeks, main="After")
```



```
par(mfrow=c(1,1))
```

Both before and after follow the normal distribution.

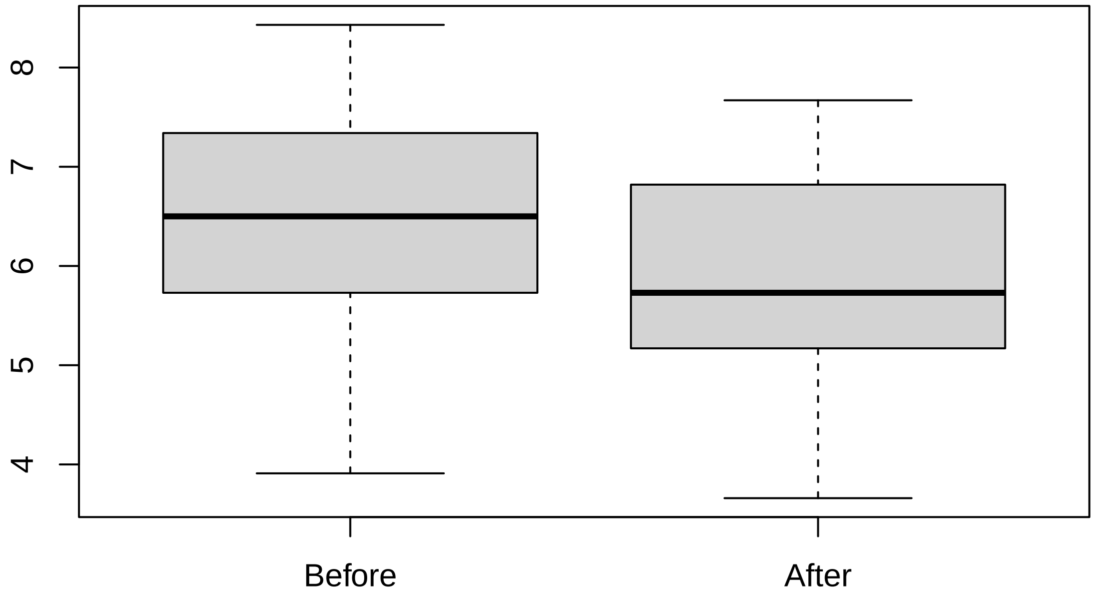
```
par(mfrow=c(1,2))
hist(Before, main="Before")
hist(After8weeks, main="After")
```



```
par(mfrow=c(1,1))
```

According to the histograms and boxplots, we can see that both Before and After columns are approximately normal. There are no major inconsistencies in the data.

```
boxplot(Before, After8weeks, names = c("Before", "After"))
```

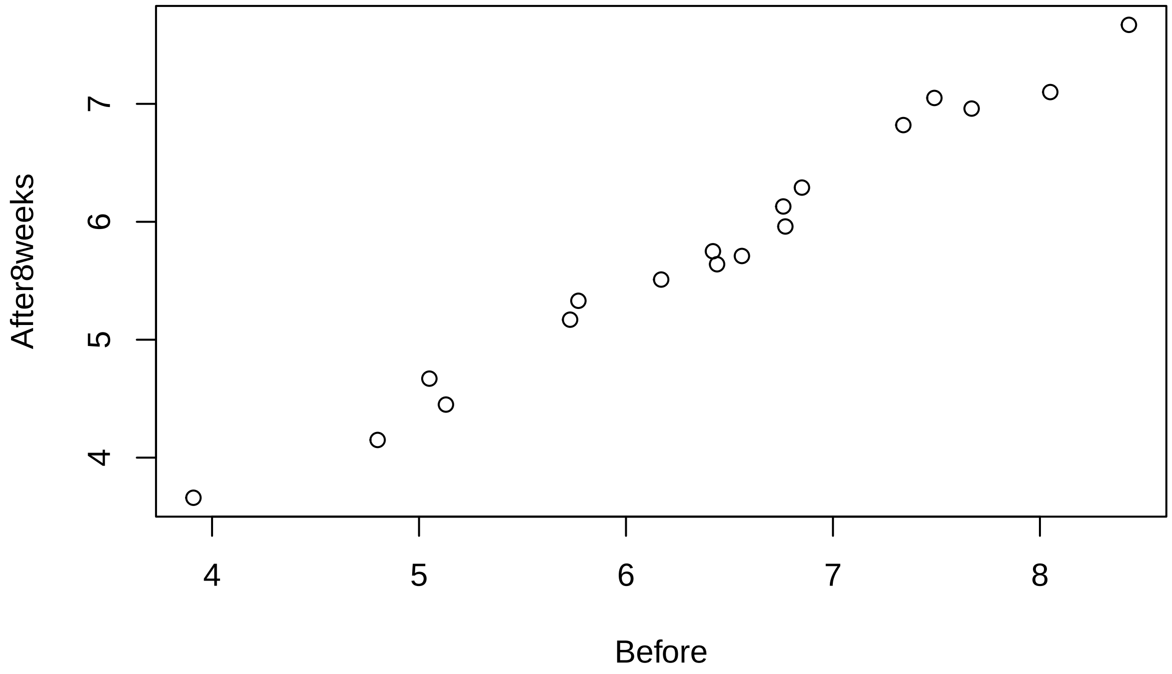


```
cor.test(Before, After8weeks)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Before and After8weeks
## t = 29.428, df = 16, p-value = 2.321e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9751289 0.9966788
## sample estimates:
##       cor
## 0.9908885
```

And have a high correlation

```
plot(Before, After8weeks)
```



b) Apply two relevant tests (cf. Lectures 2, 3) to verify whether the diet with low fat margarine has an effect (argue whether the data are paired or not). Is a permutation test applicable?

As each row represents the cholesterol level (mmol/L) respectively before the diet and after 8 weeks on the diet for every person, paired tests can be conducted.

```
t.test(Before, After8weeks, paired=T)
```

```
##
##  Paired t-test
##
## data:  Before and After8weeks
## t = 14.946, df = 17, p-value = 3.279e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.5401131 0.7176646
## sample estimates:
## mean difference
##       0.6288889
```

```
# symmetric data is required
wilcox.test(Before, After8weeks, paired=T)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  Before and After8weeks
## V = 171, p-value = 7.629e-06
## alternative hypothesis: true location shift is not equal to 0
```

Both tests showed p values $< .05$ which indicates that the we reject H_0 and accept H_a , so **the low fat margarine diet has an effect on reducing cholesterol levels**.

Additionally, permutation test can be conducted, however due to the small sample size of 18 entries, test may not be reliable and paired t-test and Wilcoxon signed-rank test may be more appropriate.

c) Let $X_1, \dots, X_{18}$ *be the column After8weeks. Assume* $X_1, \dots, X_{18} \sim \mathrm{Unif}[3, \theta]$, then use the central limit theorem to find an estimate $\hat{\theta}$ *for* θ *and construct a 95%-CI for* θ*. Can you improve this CI?*

```r
x.bar <- mean(After8weeks)
theta.hat <- x.bar * 2 - 3
z.val <- qnorm(.975) # z-value for a 95% CI
n <- length(After8weeks)
se <- sd(After8weeks) / sqrt(n)
CI <- theta.hat - c(1,-1) * z.val * se

theta.hat
```

```
## [1] 8.557778
```

```r
CI
```

```
## [1] 8.048730 9.066826
```

To improve the CI, we could try using a bootstrap approach to estimate the distribution of $\hat{\theta}$ and calculate a percentile CI.

Another way is to increase sample size. More realistic way is to change normal distribution to a t-distribution, as the sample has only 18 records.

d) By using a bootstrap test with test statistic $T = \max(X_1, \dots, X_{18})$*, determine those* $\theta \in [3, 12]$*, for which the hypothesis* $X_1, \dots, X_{18} \sim \mathrm{Unif}[3, \theta]$ *is not rejected. Can the Kolmogorov-Smirnov test be also applied for this situation?*

```r
n = 18
t = max(After8weeks)
nRep = 100

unifTest = function(thetaMax) {
  tstar = numeric(nRep)
  for (i in 1:nRep) {
    tstar[i] = max(runif(18, 3, thetaMax))
  }
  # manually calculate p-value
  pl=mean(tstar<t)
  pr=mean(tstar>t)
  p=2*min(pl,pr)
  p
}

# generate the sequence of possible thetas with the precision 0.001
# in a given range [3,12]
thetas = seq(3,12,.001)

pvalues = numeric(length(thetas))
for (i in 1:length(thetas)) {
  thetaMax = thetas[i]
  # get p-value of the bootstrap test
  pvalues[i] = unifTest(thetaMax)
}

# retrieve thetas that did not reject H0 of the bootstrap test
# and calculate the range
range(thetas[pvalues > .05])
```

```
## [1] 7.678 9.285
```

For these θ ranges the hypothesis $X_1, \dots, X_{18} \sim \mathrm{Unif}[3, \theta]$ is not rejected (p value $> .05$).

It is also possible to determine the range using Kolmogorov-Smirnov test.

```
testIfUnif = function(thetaMax) {
  ks.test(After8weeks, 'punif', 3, thetaMax)$p.value
}

thetas = seq(3,12,.001)
pvalues = numeric(length(thetas))
for (i in 1:length(thetas)) {
  thetaMax = thetas[i]
  pvalues[i] = testIfUnif(thetaMax)
}

range(thetas[pvalues > .05])
```

```
## [1] 7.083 9.455
```

With a similar test setup we end up getting more precise θ range.

e) Using an appropriate test, verify whether the median cholesterol level after 8 weeks of low fat diet is less than 6. Next, design and perform a test to check whether the fraction of the cholesterol levels after 8 weeks of low fat diet less than 4.5 is at most 25%.

In order to test median, one can use wilcoxon signed-rank test.

```
wilcox.test(After8weeks, mu = 6, alternative="l", exact = T)
```

```
## Warning in wilcox.test.default(After8weeks, mu = 6, alternative = "l", exact =
## T): cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  After8weeks
## V = 67.5, p-value = 0.223
## alternative hypothesis: true location is less than 6
```

Based on the output, we can see that the p-value is 0.223, which is greater than 0.05. Therefore, we fail to reject the null hypothesis that that the median cholesterol level is ≥ 6. We also do not have enough evidence to conclude that that the median cholesterol level after 8 weeks of low fat diet is less than 6.

Let us take another approach and test the hypothesis with the binomial test.

```
n = length(After8weeks)
xSuccess = sum(After8weeks > 6)
binom.test(xSuccess, n, .5, alternative = 'l')
```

```
##
##  Exact binomial test
##
## data:  xSuccess and n
## number of successes = 7, number of trials = 18, p-value = 0.2403
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.6078447
## sample estimates:
## probability of success
##              0.3888889
```

H_0 is not rejected. The confidence interval of the true proportion is [0.0000000 0.6078447] which does not help in this case, as it includes 0.5 value.

```
n = length(After8weeks)
xSuccess = sum(After8weeks < 4.5)
binom.test(xSuccess, n, .25, alternative = 'l')
```

```
##
##  Exact binomial test
##
## data:  xSuccess and n
## number of successes = 3, number of trials = 18, p-value = 0.3057
## alternative hypothesis: true probability of success is less than 0.25
## 95 percent confidence interval:
##  0.0000000 0.3766792
## sample estimates:
## probability of success
##              0.1666667
```

We fail to reject the null hypothesis (p>0.3057) that the true proportion of cholesterol levels less than 4.5 is 0.25, and do not have evidence to conclude that the true proportion is less than 0.25.

The 95% confidence interval for the true proportion is [0.0000000 0.3766792], which includes the value 0.25, providing further support for the null hypothesis. Therefore, we do not have strong evidence to conclude that the fraction of cholesterol levels after 8 weeks of low-fat diet less than 4.5 is at most 25%.

```
detach(data)
```

Exercise 3

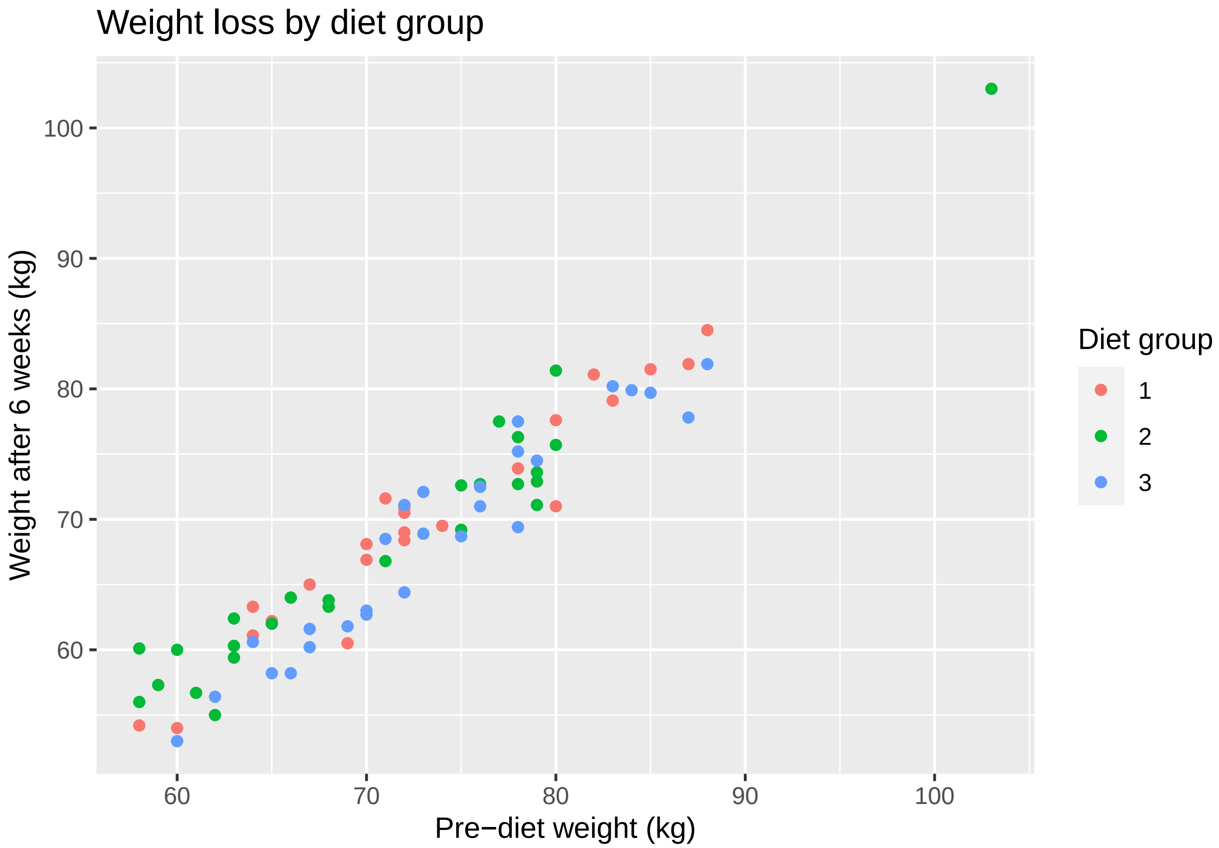
To investigate the effect of 3 types of diet, 78 persons were divided randomly in 3 groups, the first group following diet 1, second group diet 2 and the third group diet 3. Next to some other characteristics, the weight was measured before diet and after 6 weeks of diet for each person in the study. 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.

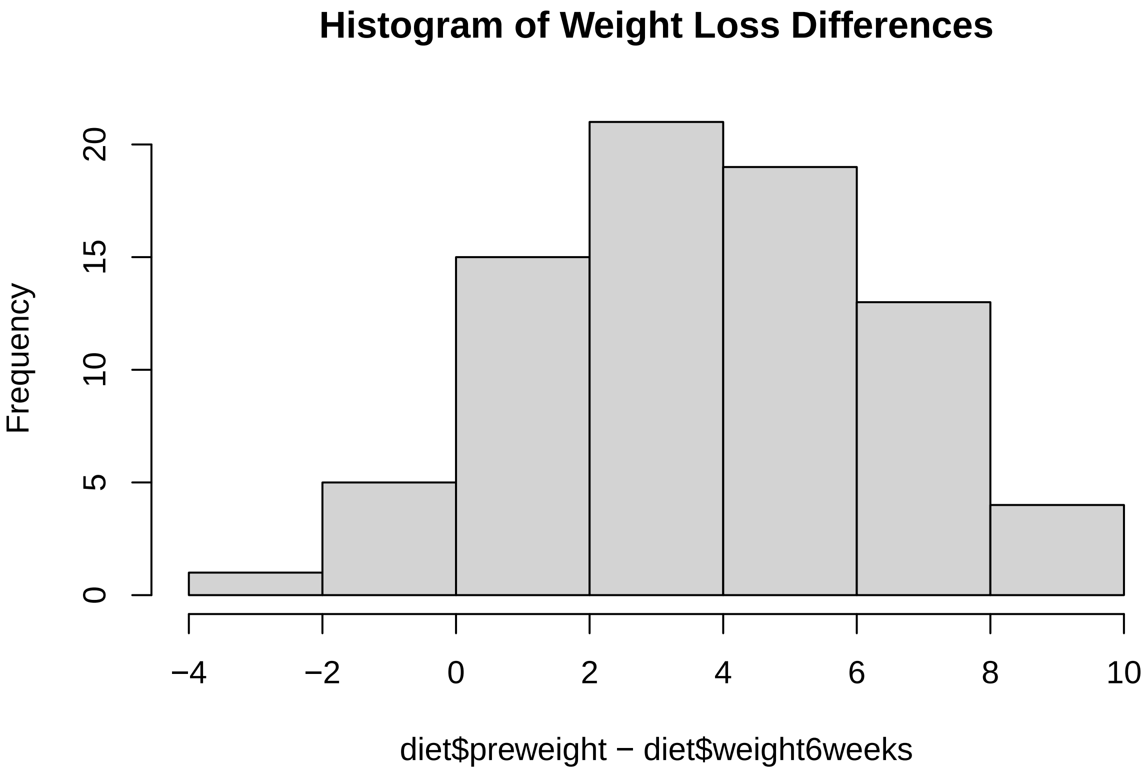
To create a graphical summary, we can use a scatter plot of preweight vs weight6weeks, with the points colored by diet group. To test whether the diet affects weight loss, we can use a paired t-test on the difference between preweight and weight6weeks for each participant, with diet group as a grouping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

```r
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")
```

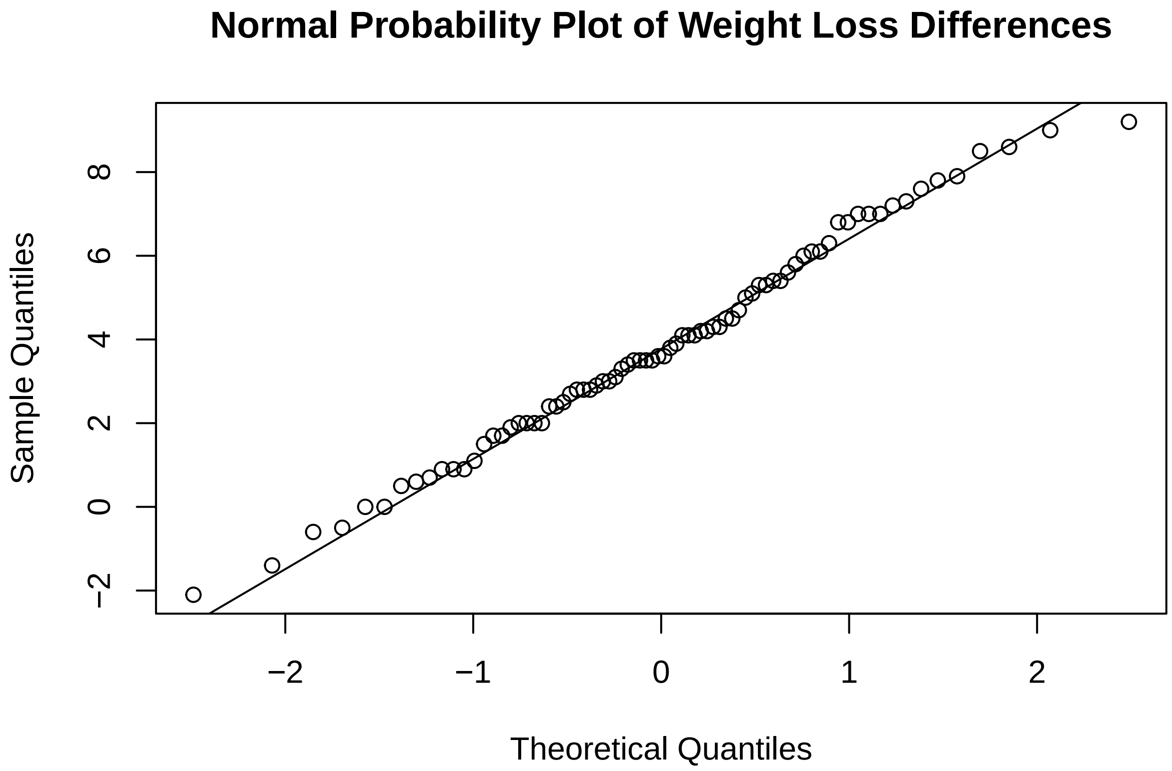


To check the assumptions of the paired t-test, we can inspect the normality of the differences using a histogram and a normal probability plot:

```r
# 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.

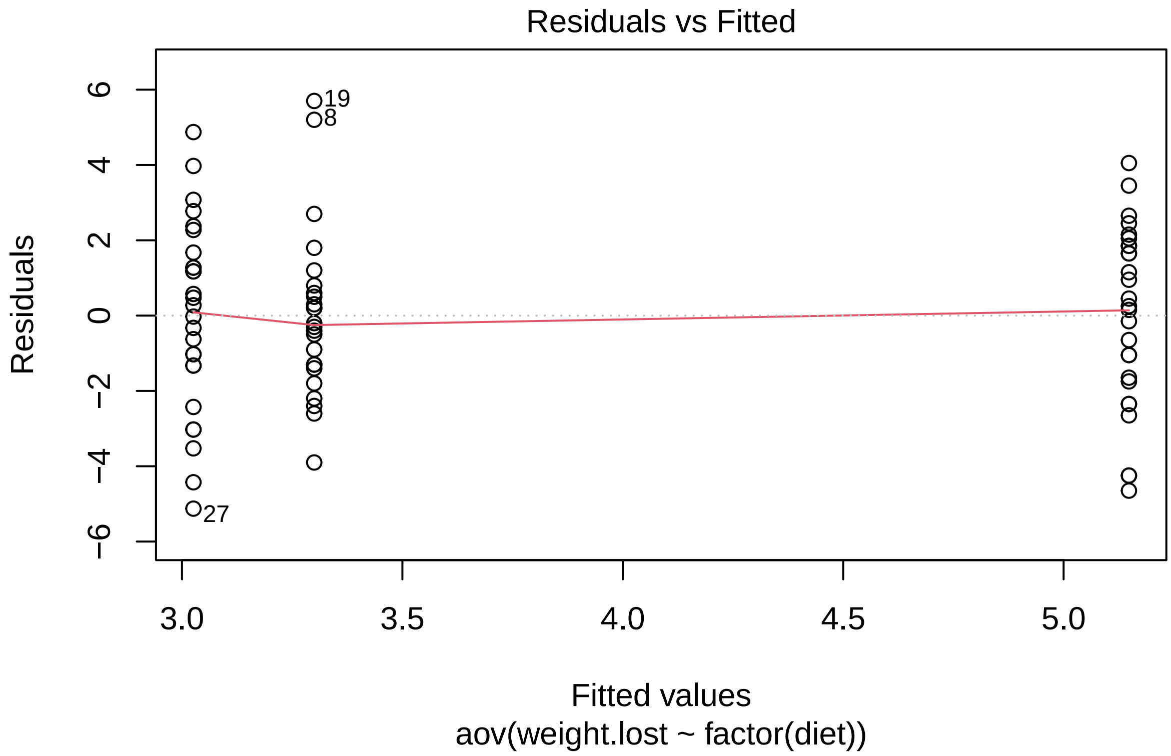
b) Apply one-way ANOVA to test whether type of diet has an effect on the lost weight. Do all three types diets lead to weight loss? Which diet was the best for losing weight? Can the Kruskal-Wallis test be applied for this situation?

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.

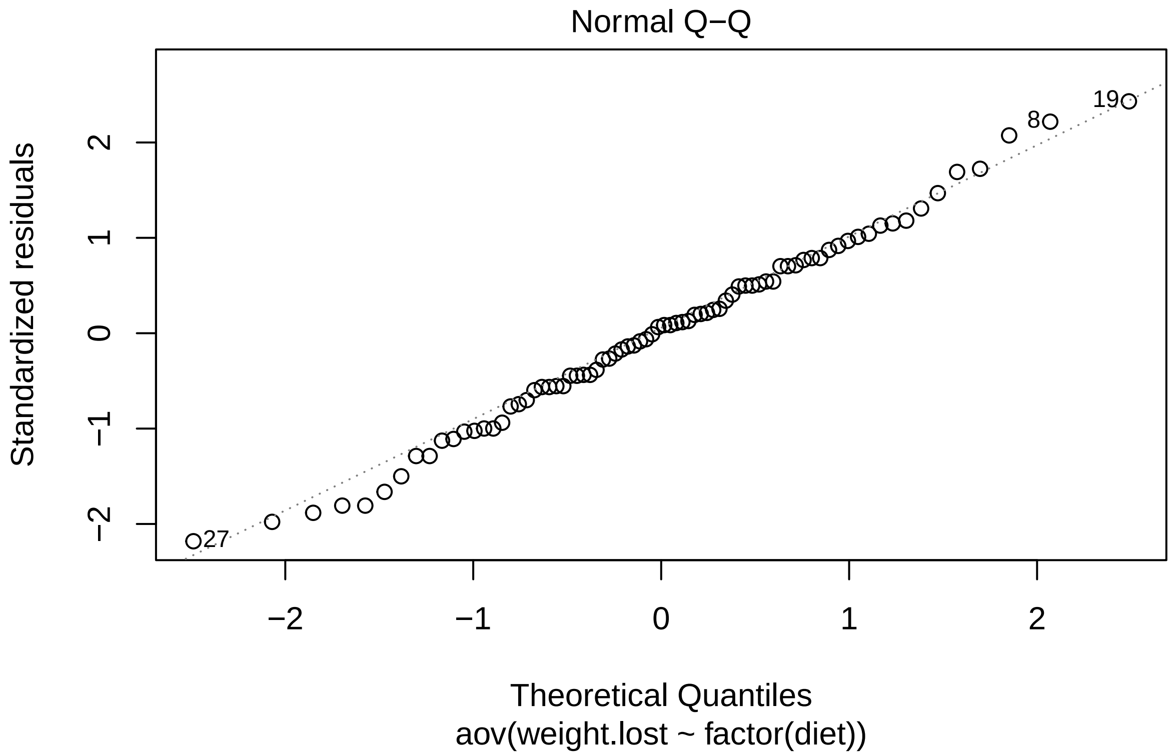
```
#one-way anova
# factor() as TukeyHSD() requires the aov object to have been created with groups as explicit
model_anova <- aov(weight.lost ~ factor(diet), data = diet)
summary(model_anova)
```

```
##              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
```

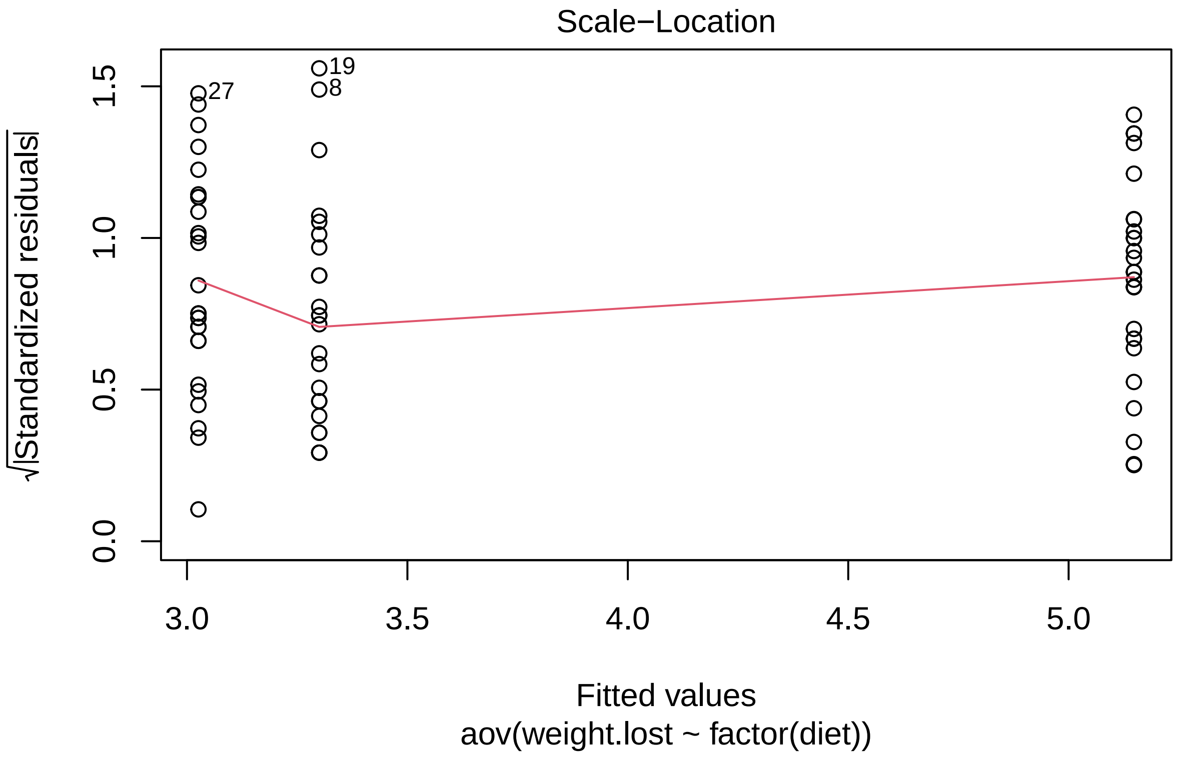
```
#Check the assumptions of the ANOVA
#Residuals vs. Fitted
plot(model_anova, 1)
```



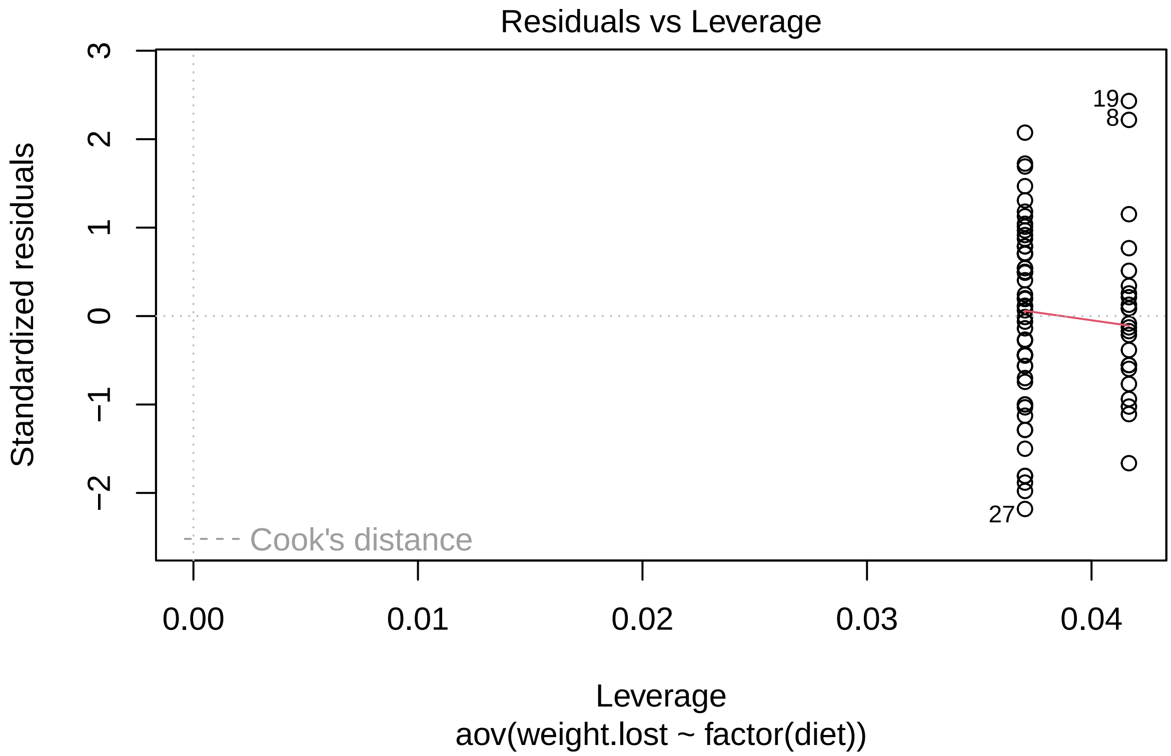
```
#Normal Q-Q Plot
plot(model_anova, 2)
```



```
#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.

```
# 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
```

Exercise 4

The dataset npk is available in the R package MASS. After loading the package MASS, type npk at the prompt to view this dataset. This dataset gives the yield of peas in pounds per plot, based on four factors: in which block the plot was located (labeled 1 through 6), and whether nitrogen (N), phosphate (P) or potassium (K) was applied to the soil (1 = applied, 0 = not applied). There are 24 plots, 4 per block. This is incomplete block design but balanced in the sense that within each block each soil additive is received by two plots. Our main question of interest is whether nitrogen N has an effect on yield.

a) Present an R-code for the randomization process to distribute soil additives over plots in such a way that each soil additive is received exactly by two plots within each block.

We create separate vectors for the block, N, P, and K variables, and then combine them into a data frame. The rep function is used to repeat each treatment vector twice. The resulting data frame has 24 rows (one for each plot) and four columns (one for each variable). In the loop we randomly permute the order of plots within each block. We identify the indices of the rows corresponding to each block, and randomly permute these rows.

```
library(MASS)
npk
```

```
##    block N P K yield
## 1      1 0 1 1  49.5
## 2      1 1 1 0  62.8
## 3      1 0 0 0  46.8
## 4      1 1 0 1  57.0
## 5      2 1 0 0  59.8
## 6      2 1 1 1  58.5
## 7      2 0 0 1  55.5
## 8      2 0 1 0  56.0
## 9      3 0 1 0  62.8
## 10     3 1 1 1  55.8
## 11     3 1 0 0  69.5
## 12     3 0 0 1  55.0
## 13     4 1 0 0  62.0
## 14     4 1 1 1  48.8
## 15     4 0 0 1  45.5
## 16     4 0 1 0  44.2
## 17     5 1 1 0  52.0
## 18     5 0 0 0  51.5
## 19     5 1 0 1  49.8
## 20     5 0 1 1  48.8
## 21     6 1 0 1  57.2
## 22     6 1 1 0  59.0
## 23     6 0 1 1  53.2
## 24     6 0 0 0  56.0
```

```
block <- rep(1:6, each = 4)
N <- rep(c(1, 1, 0, 0), 6)
```

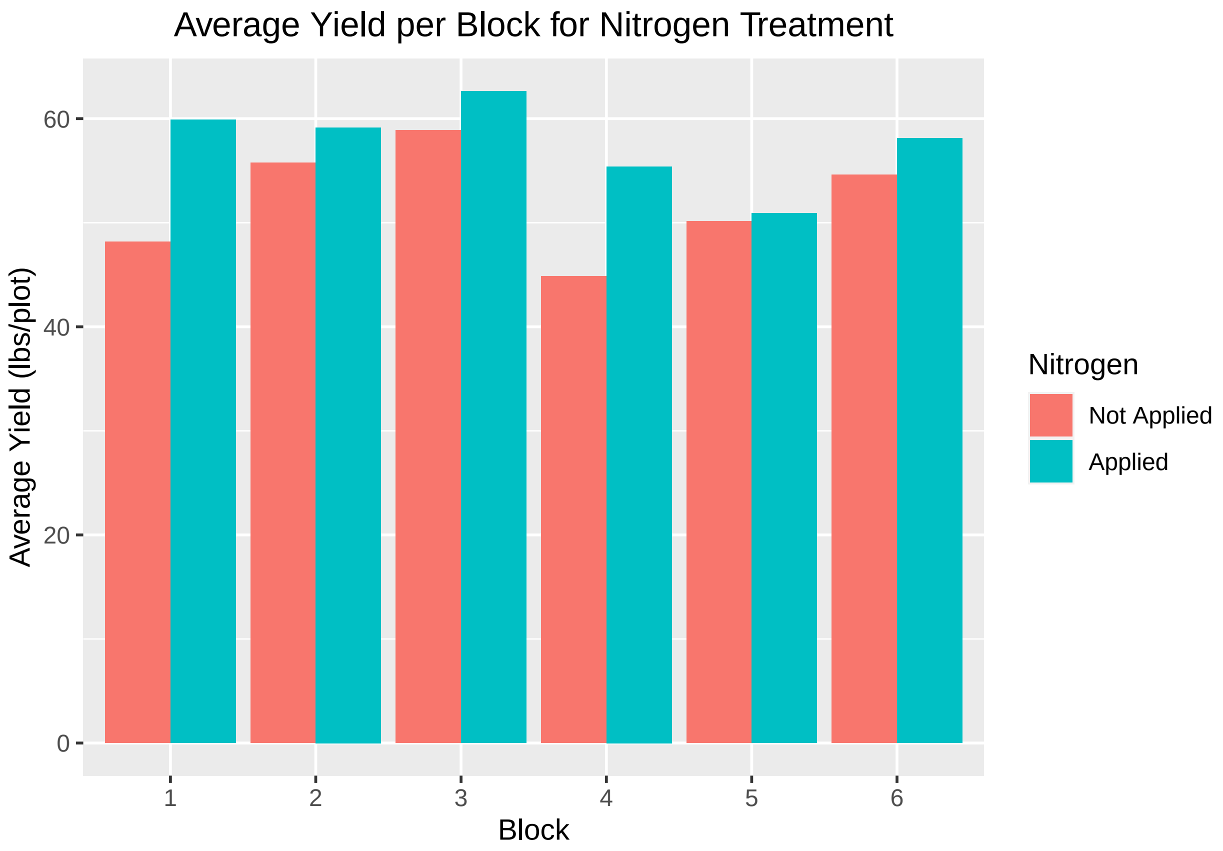
```r
P <- rep(c(1, 0), each = 12)
K <- rep(c(0, 1, 1, 0), 6)

npk_df <- data.frame(block = block, N = N, P = P, K = K)

# random permutations the order of plots within each block
for(i in unique(block)) {
  ind <- which(block == i)
  npk_df[ind,] <- npk_df[sample.int(length(ind)),][,]
}

npk_df
```

```
##    block N P K
## 1      1 0 1 1
## 2      1 1 1 0
## 3      1 0 1 0
## 4      1 1 1 1
## 5      1 1 1 1
## 6      1 0 1 1
## 7      1 0 1 0
## 8      1 1 1 0
## 9      1 0 1 0
## 10     1 1 1 1
## 11     1 1 1 0
## 12     1 0 1 1
## 13     1 0 1 1
## 14     1 1 1 1
## 15     1 1 1 0
## 16     1 0 1 0
## 17     1 1 1 1
## 18     1 1 1 0
## 19     1 0 1 1
## 20     1 0 1 0
## 21     1 0 1 0
## 22     1 1 1 1
## 23     1 1 1 0
## 24     1 0 1 1
```

b) Make a plot to show the average yield per block for the soil treated with nitrogen and for the soil that did not receive nitrogen, and comment. What is the purpose to take the factor block into account?

We first calculate the mean yield for each combination of block and nitrogen treatment using the aggregate() function. Then, we use ggplot() to create the plot.

```r
library(ggplot2)

npk_means <- aggregate(yield ~ block + N, data = npk, mean)

ggplot(npk_means, aes(x = block, y = yield, fill = factor(N))) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_discrete(name = "Nitrogen", labels = c("Not Applied", "Applied")) +
  labs(x = "Block", y = "Average Yield (lbs/plot)") +
  ggtitle("Average Yield per Block for Nitrogen Treatment") +
  theme(plot.title = element_text(hjust = 0.5))
```



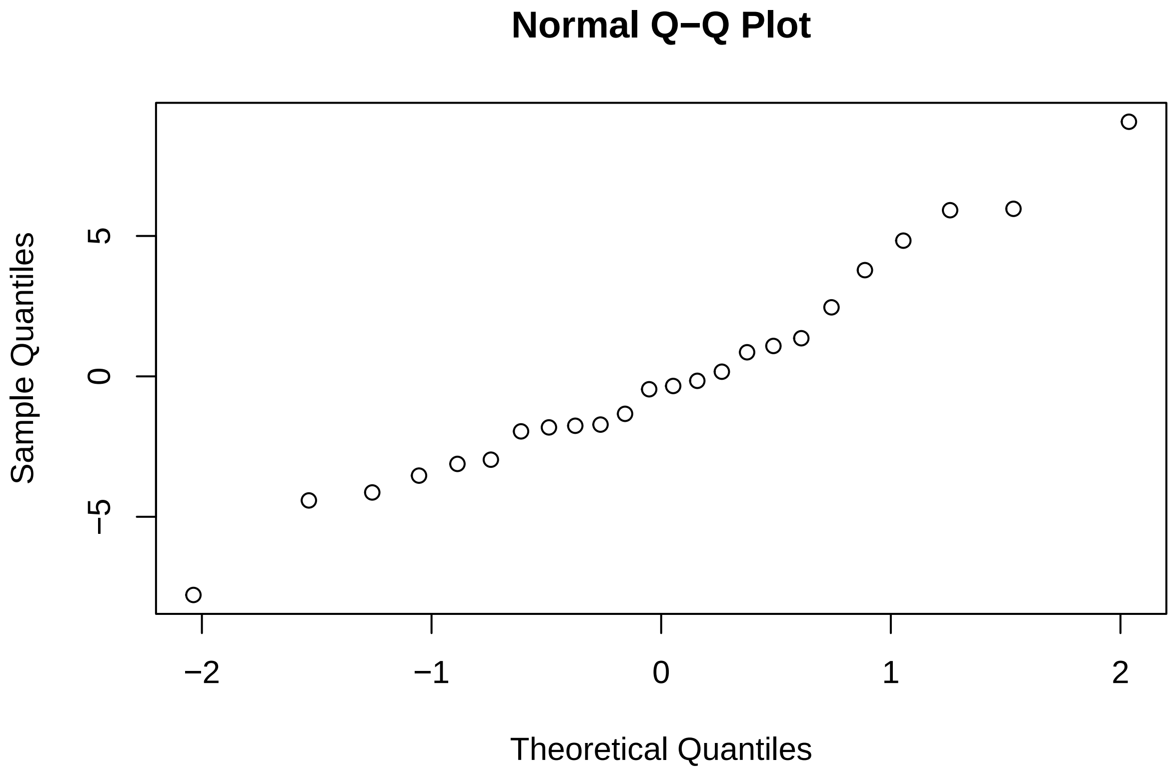
c) Conduct a full two-way ANOVA with the response variable yield and the two factors block and N. Was it sensible to include factor block into this model? Can we also apply the Friedman test for this situation? Comment.

```r
yieldlm = lm(yield ~ block + N,data = npk)
yieldav = anova(yieldlm)
yieldav
```

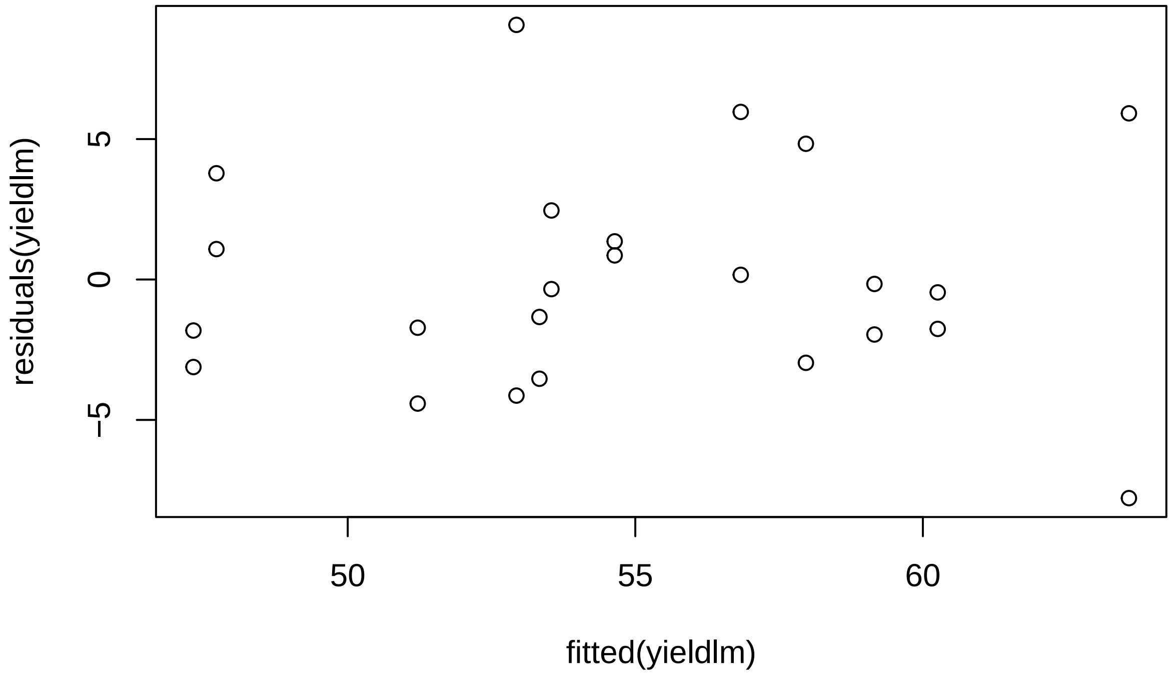
```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## block      5 343.29  68.659  3.3951 0.026173 *
## N          1 189.28 189.282  9.3598 0.007095 **
## Residuals 17 343.79  20.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen from the two-way anova test, factor block is statistically significant (p<.05), thus it was important to include it.

```
qqnorm(residuals(yieldlm));
```



```
plot(fitted(yieldlm),residuals(yieldlm))
```



Residuals are approximately normal, the spread in residuals seems to be consistent, however some points are not symmetric.

Friedman test is appropriate on non-normally distributed data with a repeated measure design, however, data is normally distributed in this case and does not require a rank-based testing.

d) Investigate other possible models with all the factors combined, restricting to only one (pairwise) interaction term of factors N, P and K with block in one model (no need to check the model assumptions for all the models). Test for the presence of main effects of N, P and K, possibly taking into account factor block. Give your favorite model and motivate your choice.

```
{# {r} # npk_df2 = cbind(npk_df, yield=npk$yield) # npklm = lm(yield ~ factor(N)+factor(P)+fac
data=npk_df2) # anova(npklm)
```

e) Recall the main question of interest. In this light, repeat c) by performing a mixed effects analysis, modeling the block variable as a random effect by using the function lmer. Compare your results to the results found by using the fixed effects model in c). (You will need to install the R-package lme4, which is not included in the standard distribution of R.)

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
yieldlmer = lmer(yield ~ N + (1|block),REML=F, data = npk)
anova(yieldlmer, yieldlm)
```

```
## Data: npk
## Models:
## yieldlmer: yield ~ N + (1 | block)
## yieldlm: yield ~ block + N
##           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## yieldlmer    4 153.48 158.20 -72.742   145.48
## yieldlm      8 148.00 157.42 -65.998   132.00 13.487  4   0.009124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen from the results, factor block has a significant effect (p>0.05).