Assignment 1 - Group 27

Joost Driessen, Emma van Lipzig, Rohan Zonneveld

2023-02-22

Excercise 1. Birthweight

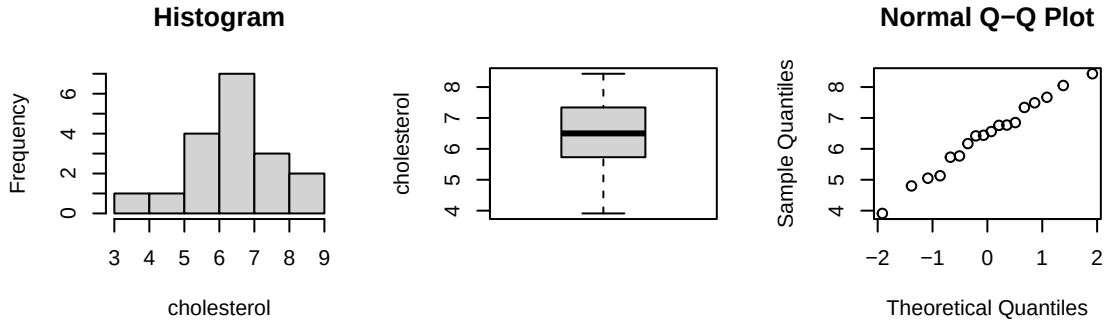
Excercise 2. Cholesterol

a)

We summarise the data by giving a summary and by making a histogram, a boxplot and a QQ-plot of both columns.

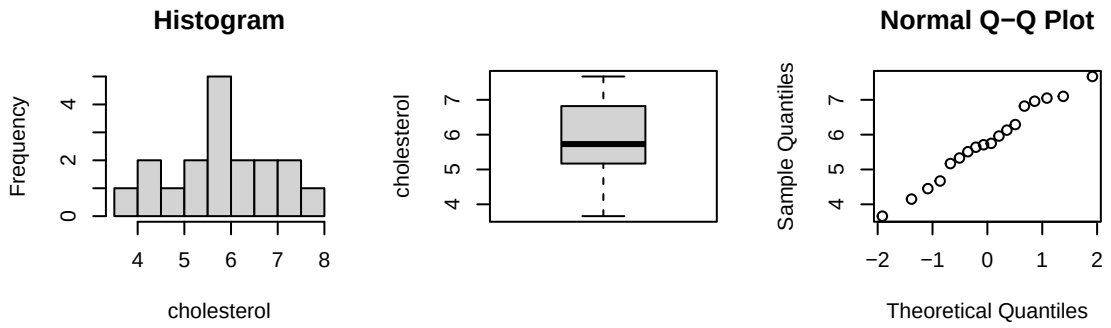
Before Diet

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.91    5.74    6.50    6.41    7.22    8.43
```



After 8 weeks of diet

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.66    5.21    5.73    5.78    6.69    7.67
```

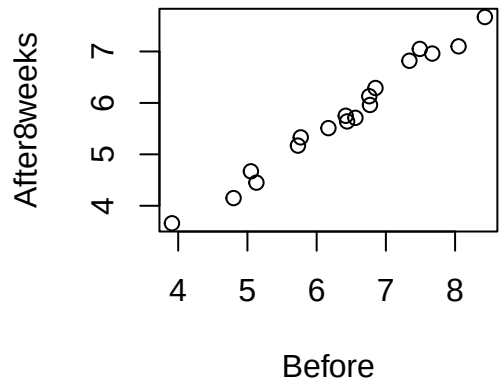


Both histograms look normally distributed, the boxplots are symmetrical with the mean in the center of the box and the QQ-plots are linear. In conclusion, the plots give no reason to suspect that the data is drawn from a non-normal distribution.

To investigate if the columns are correlated we calculate the correlation according to the literature. To visualize the correlation a plot is created with before on the x-axis and after8weeks on the y-axis.

```
cor(Before,After8weeks)
```

```
## [1] 0.991
```



The columns are strongly correlated.

b)

Since the study measured two sets of observations obtained from the same individuals the data is paired. In this case, each individual's before and after cholesterol levels are paired observations. Two relevant statistical tests for paired samples are: paired t-test and Wilcoxon signed-rank test.

A *paired t-test* is performed when the data is normally distributed. Looking back to *a)* we can see that the data is indeed normally distributed.

```
p=t.test(After8weeks, Before, paired = TRUE)[[3]]; p
```

```
## [1] 3.28e-11
```

The p-value is lower than 0.05, which means H0 can be rejected meaning that there is indeed a true difference in means between the two variables.

The *Wilcoxon signed-rank test* does not assume that the data is normally distributed and is therefore a useful alternative to the paired t-test when this assumption is violated.

```
p=wilcox.test(After8weeks, Before, paired = TRUE)[[3]]; p
```

```
## [1] 7.63e-06
```

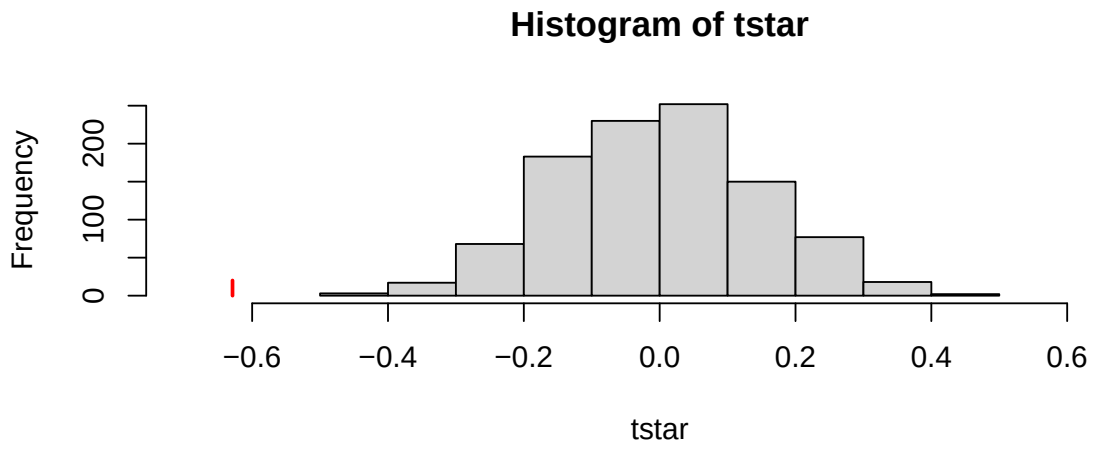
The p-value is lower than 0.05, so H0 is rejected meaning that the true location of the population is different between columns.

Since the data is paired, we can apply a permutation test to determine whether the diet has an effect. Also, the permutation test does not assume any distribution for the data. The R code to perform this test is as follows:

```
mystat=function(x,y) {mean(x-y)}
B=1000; tstar=numeric(B)
for (i in 1:B) {
  dietstar=t(apply(cbind(After8weeks,Before),1,sample))
  tstar[i]=mystat(dietstar[,1],dietstar[,2]) }
myt=mystat(After8weeks,Before)

myt
```

```
## [1] -0.629
```

In the histogram below the distribution of t* is depicted. The distribution is normal, due to CLT. The actual value of the T statistic is depicted in red.



```
pl=sum(tstar<myt)/B
pr=sum(tstar>myt)/B
p=2*min(pl,pr); p
```

```
## [1] 0
```

The value of p is equal to zero, which means H0 can be rejected. This means that there is a significant difference between before and after the diet.

c)

The mean of a uniform sample is given by $E[\bar{X}] = (1/n) * \sum Xi = 1/18 * \sum Xi$.The variation of a uniform sample is also given, $Var(Xi) = ((\theta - 3)^2)/12$. So the variance of the sample mean is given by $Var(\bar{X}) = Var((1/18) * \sum Xi) = (1/18^2) * \sum Var(Xi) = ((\theta - 3)^2)/(12 * 18 * 18)$. According to the central limit theorem, as n goes to infinity, the distribution of $\bar{X}$ approaches a normal distribution with mean $E[\bar{X}]$ and variance $Var(\bar{X})$. The mean of a uniform distribution is in the center, so to estimate the maximum of a uniform distribution ($\bar{\theta}$) one can simply multiply the mean by 2 and subtract the start of the interval.

```
x.bar = mean(After8weeks)
theta.hat = 2 * x.bar - 3
```

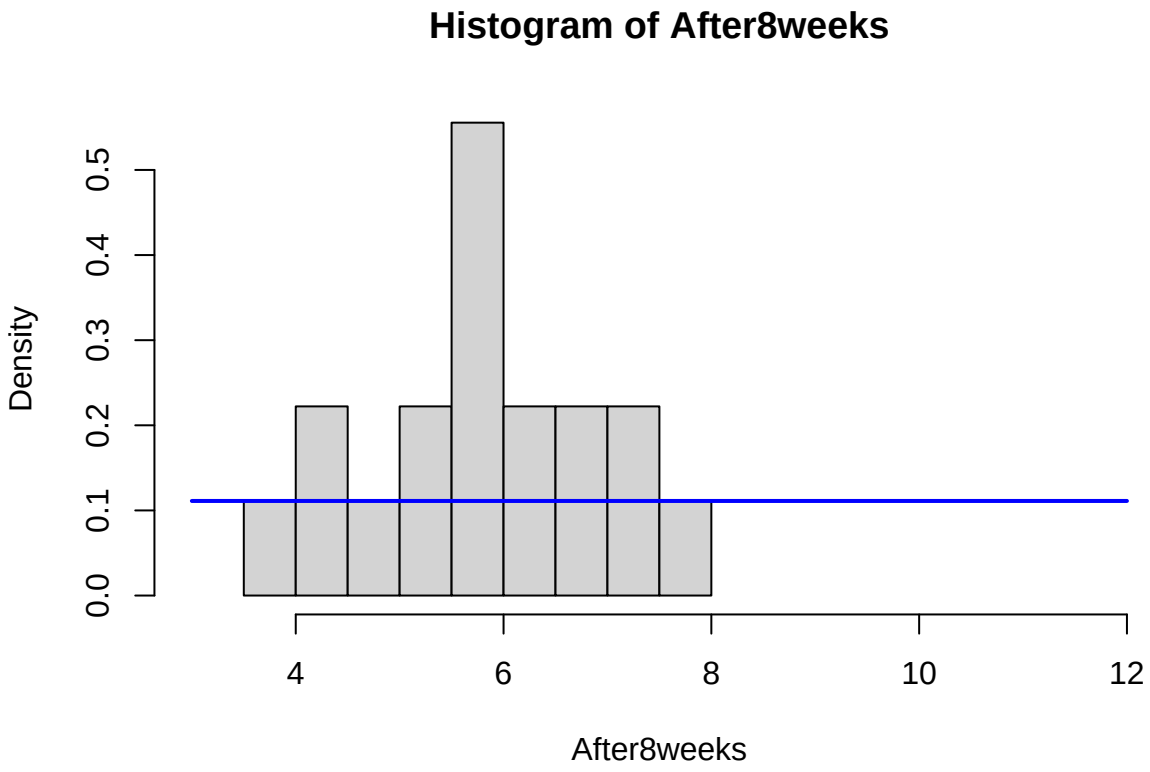
A 95% confidence interval for θ is found using the central limit theorem by calculating the margin of error and adding and subtracting it to $\bar{\theta}$.

```
z = qnorm(0.975)
sigma = sqrt(((theta.hat - 3)^2)/(12*18))
me = z * sigma

ci = c(theta.hat - me, theta.hat + me)
ci
```

```
## [1] 7.82 9.30
```

Can you improve this CI?

d)



The After8weeks column does not appear to be uniformly distributed. Since there are no values bigger than 8 in the data it is expected that values of $\theta > 8$ will reject the null hypothesis. To investigate if After8weeks is drawn from an uniform distribution in [3,θ] a bootstrap test is performed. Surrogate T-values are generated that are representative of values of T under H0. The test statistic (in this case the maximum) is computed. This process is repeated 1000 times. Finally, the T-value is compared to the surrogate T*-values to determine a p-value.

```
H0=rep(1, 9)
for(theta in 4:12){
  n=length(After8weeks); t=max(After8weeks)
  B=1000; Tstar=numeric(B)
  for(i in 1:B) {
   Xstar=runif(n,3,theta)
   Tstar[i]=max(Xstar)}
  pl=sum(Tstar<t)/B;pr=sum(Tstar>t)/B
  p=2*min(pl,pr)
  if(p<0.05){H0[theta-3]=0}}
H0
```

```
## [1] 0 0 0 0 1 0 0 0 0
```

The H0 is rejected for all values of θ except 8. This result was to be expected as the data ranges from 3 to 8 as can be seen in the histogram.

The *Kolmogorov-Smirnov* is used to test if two samples are from the same distribution. The actual data can be compared to data generated from a uniform distribution. So it is perfectly suited to use in this situation.

```
H0=rep(1,9)
for(theta in 4:12){
  n=length(After8weeks)
  sample=runif(n,3,theta)
  p=ks.test(After8weeks,sample)$p.value
  if(p<0.05){H0[theta-3]=0}
}
H0
```

```
## [1] 0 0 0 1 1 1 1 1 1
```

The null hypothesis is rejected in cases where θ<8.

e)

To test whether the median cholesterol level after 8 weeks of low fat diet is less than 6 a *Wilcoxon signed-rank test* is performed.

```
p=wilcox.test(After8weeks,mu=6,alt="l")[[3]]; p
```

```
## [1] 0.223
```

The p-value is bigger than 0.05 so the null hypothesis can not be rejected. This means that it is not statistically significant that the population median is lower than 6.

```
n=sum(After8weeks<4.5)
prop=n/length(After8weeks);prop
```

```
## [1] 0.167
```

The proportion of the values lower than 4.5 is less than 25%

Excercise 3. Diet

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 Download diet.txt with 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.

```
data <- read.table(file= 'diet.txt', header=TRUE)
#View(data)
```

```
data$weight.lost <- 0
for (i in 1:78) {
  data$weight.lost[i] <- data$preweight[i] - data$weight6weeks[i]
}
head(data) #shows that weight.lost column has been added
```

```
##   person gender age height preweight diet weight6weeks weight.lost
## 1      1      0  22    159        58    1         54.2         3.8
## 2      2      0  46    192        60    1         54.0         6.0
## 3      3      0  55    170        64    1         63.3         0.7
## 4      4      0  33    171        64    1         61.1         2.9
## 5      5      0  50    170        65    1         62.2         2.8
## 6      6      0  50    201        66    1         64.0         2.0
```

```
summary(data) #inspect data
```

```
##      person         gender           age           height      preweight
##  Min.   : 1.0   Min.   :0.000   Min.   :16.0   Min.   :141   Min.   : 58.0
##  1st Qu.:20.2   1st Qu.:0.000   1st Qu.:32.2   1st Qu.:164   1st Qu.: 66.0
##  Median :39.5   Median :0.000   Median :39.0   Median :170   Median : 72.0
##  Mean   :39.5   Mean   :0.434   Mean   :39.2   Mean   :171   Mean   : 72.5
##  3rd Qu.:58.8   3rd Qu.:1.000   3rd Qu.:46.8   3rd Qu.:175   3rd Qu.: 78.0
##  Max.   :78.0   Max.   :1.000   Max.   :60.0   Max.   :201   Max.   :103.0
##                 NA's   :2
##       diet       weight6weeks    weight.lost
##  Min.   :1.00   Min.   : 53.0   Min.   :-2.10
##  1st Qu.:1.00   1st Qu.: 61.9   1st Qu.: 2.00
##  Median :2.00   Median : 69.0   Median : 3.60
##  Mean   :2.04   Mean   : 68.7   Mean   : 3.84
##  3rd Qu.:3.00   3rd Qu.: 73.8   3rd Qu.: 5.55
##  Max.   :3.00   Max.   :103.0   Max.   : 9.20
##
```

a)

Make an informative graphical summary of the data relevant for study of the effect of diet on the weight 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.

The qqplot and histogram of the weigth before diet show that this data does not appear to be normally distributed, as there is one large outlier of 103. The same goes for the weight after six weeks, so after the diet. This again does not seem normally distributed because of the outlier 103. However, the differences between the two, represented as weight.lost, do appear to be distributed normally. This can be seen in both the QQ plot and histogram.

```
qq_pre <- ggplot(data) + geom_qq(aes(sample=preweight)) + theme_light() + labs(title="QQ plot Preweight
qq_post <- ggplot(data) + geom_qq(aes(sample=weight6weeks))  + theme_light()  + labs(title="QQ plot Wei

hist_pre <- ggplot(data, aes(x=preweight))+ geom_histogram(binwidth=1, color="seagreen", fill="lightgre
hist_post <- ggplot(data, aes(x=weight6weeks))+ geom_histogram(binwidth=1, color="seagreen", fill="ligh

grid.arrange(qq_pre, hist_pre, qq_post, hist_post)
```

```
max(data$preweight)
max(data$weight6weeks)

qq_diff <- ggplot(data) + geom_qq(aes(sample=weight.lost)) + theme_light() + labs(title="QQ plot Lost w
hist_diff <- ggplot(data, aes(x=weight.lost))+ geom_histogram(binwidth=1, color="seagreen", fill="light

grid.arrange(qq_diff, hist_diff)
```

The Shapiro-Wilk test does not reject normality (if we take alpha = 0.05), for the preweight data. As this test does not always reject normality for non-normal distributions, this distribution may still be non-normal (as indicated by the histogram and qqplot).

The Shapiro-Wilk test rejects normality (if we take alpha = 0.05), for the weight at six weeks data. When this test rejects normality, the distribution should indeed be non-normal, which corresponds with the qqplot and histogram of the data.

```
shapiro.test(data$preweight)
shapiro.test(data$weight6weeks)
```

The boxplots below shows that there may be a difference in the weight before and after dieting. The first boxplot shows the distribution of weight data before and after a diet, the second boxplot shows the distribution of the difference between these weights, which is the distribution of the weight loss.

```
data_subset <- select(data,preweight, weight6weeks)
data_long <- melt(data_subset)

weight_boxplt <- ggplot(data_long, aes(x = variable, y = value)) +
  geom_boxplot(color="seagreen", fill="lightgreen") + labs(title = 'Weights before and after diet', x =

diet_boxplt <- ggplot(data, aes( y=weight.lost)) + geom_boxplot(color="seagreen", fill="lightgreen") + l

grid.arrange(weight_boxplt, diet_boxplt,  ncol = 2)
```

Because the differences between the pre and post diet weight data is distributed normally, a t-test can be used to determine whether these two groups differ statistically significantly from each other. Because the data is the pre and post weight from the same subjects, in this situation a paired t-test is applicable. H0 is that the groups are not different.

```
t.test(data$preweight, data$weight6weeks, data = data, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data$preweight and data$weight6weeks
## t = 13, df = 77, p-value <2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  3.27 4.42
## sample estimates:
## mean difference
##            3.84
```

The p value of the t-test is < 0.05, therefore H0 is rejected. This means that the means of preweight and weight6weeks statistically differ. As the weightloss over the six weeks of dieting was significant, it could be that diet has an influence on weightloss. However, this does not take into account other factors included in the experiment like gender and age.

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?

The boxplots below show that that different diets may have different effects on weightloss.

```
diet_g_boxplt <- ggplot(data, aes( y=weight.lost, group = diet)) + geom_boxplot(color="seagreen", fill="
diet_g_boxplt
```

The summary below shows that the study is not a balanced design. Diet 1 has less participants than diet 2 and 3.

The ANOVA with treatment parametrization and diet 1 as base level shows that the other diets differ significantly from diet 1. This means that some diets are more effective for weightloss than others. If the study had a 'no diet' treatment, this would have been a good base level to use, however, this is not the case. As it is not clear which of the diets should be the base level, it is better to use sum parametrization. A summary of the treatment parametrization model shows that diet 3 is better for weightloss than diet 1. A summary of the model with sum parametrization shows that diet 2 differs significantly from the global diet mean. Although all diets can be used for weightloss, diet 2 and 3 are better than diet 1. A t-test shows that diet 2 and 3 are significantly different, which implies that diet 3 is best for losing weight.

```
data$diet <- as.factor(data$diet) # diet is numerical but needs to be factor for ANOVA
is.factor(data$diet) # True
```

```
## [1] TRUE
```

```
summary(data$diet) #design is not balanced
```

```
##  1  2  3
## 24 27 27
```

```
# treatment parametrization, diet1 is base level
dietaov <- lm(weight.lost ~ diet, data=data)
anova(dietaov)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## diet       2     71    35.5     6.2 0.0032 **
## Residuals 75    430     5.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(dietaov)
```

```
##
## Call:
## lm(formula = weight.lost ~ diet, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.126 -1.381  0.176  1.652  5.700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.300      0.489    6.75  2.7e-09 ***
## diet2         -0.274      0.672   -0.41   0.6845
## diet3          1.848      0.672    2.75   0.0075 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.39 on 75 degrees of freedom
## Multiple R-squared:  0.142,  Adjusted R-squared:  0.119
## F-statistic:  6.2 on 2 and 75 DF,  p-value: 0.00323
```

```
# sum parametrization
contrasts(data$diet)=contr.sum
dietaov2 <- lm(weight.lost ~ diet, data=data)
anova(dietaov2)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## diet       2     71    35.5     6.2 0.0032 **
## Residuals 75    430     5.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(dietaov2)
```

```
##
## Call:
## lm(formula = weight.lost ~ diet, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.126 -1.381  0.176  1.652  5.700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.825      0.272   14.08   <2e-16 ***
## diet1         -0.525      0.392   -1.34    0.184
## diet2         -0.799      0.380   -2.10    0.039 *
## ---
```

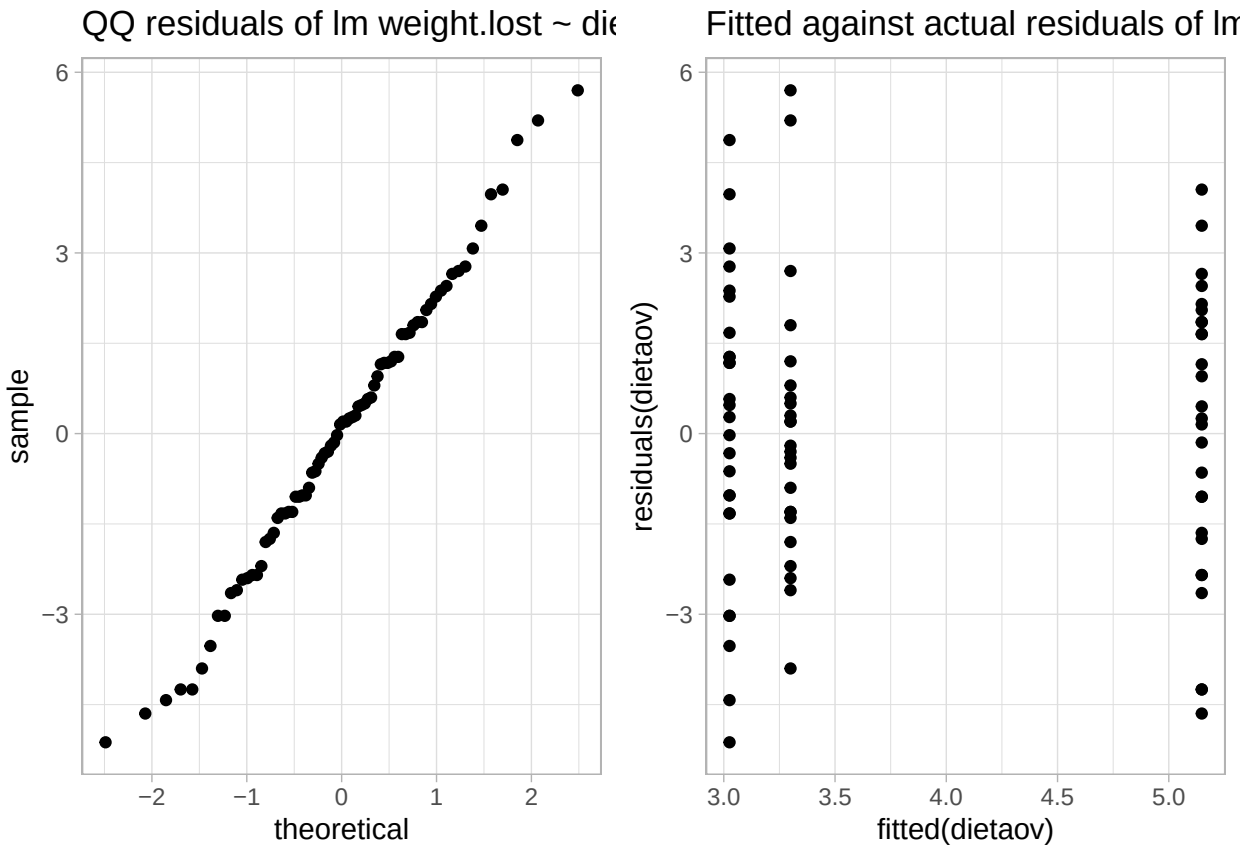
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.39 on 75 degrees of freedom
## Multiple R-squared:  0.142,  Adjusted R-squared:  0.119
## F-statistic:  6.2 on 2 and 75 DF,  p-value: 0.00323
```

```
# t-test
data_diet2 <- data[data$diet == 2,]
data_diet3 <- data[data$diet == 3,]
names(data_diet3)[8]<- paste("weight.lost2")
data_diet23 <- cbind(data_diet2,data_diet3)
t.test(data_diet23$weight.lost, data_diet23$weight.lost2, data = data_diet23, paired = FALSE)
```

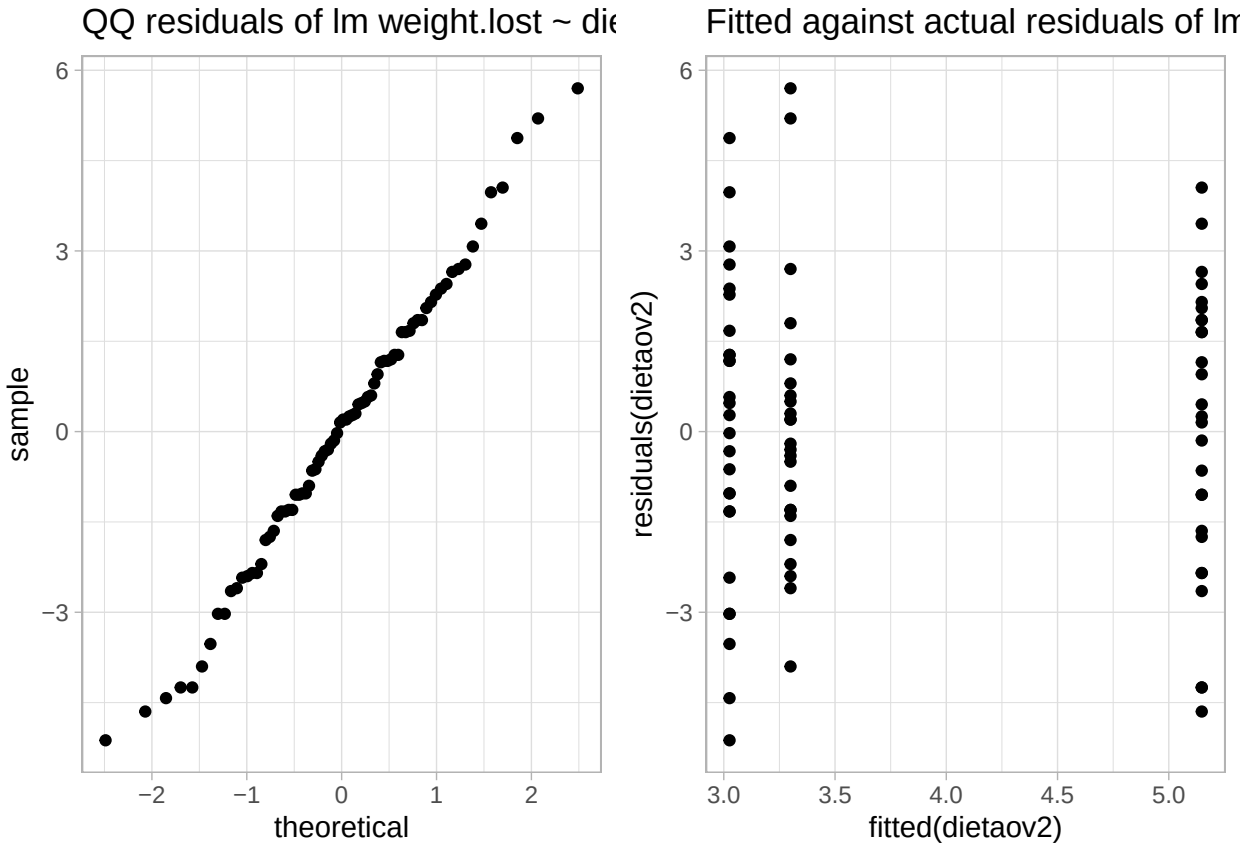
```
##
##  Welch Two Sample t-test
##
## data:  data_diet23$weight.lost and data_diet23$weight.lost2
## t = -3, df = 52, p-value = 0.003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.466 -0.778
## sample estimates:
## mean of x mean of y
##      3.03      5.15
```

Below the model assumptions for the two ANOVA's are tested. The qqplots show that the residuals for both models seem normally distributed, and neither plot of the fitted against actual residuals shows particular patterns. Therefore the model assumptions are met for both ANOVA's. Because these assumptions are met, a non-parametric test like Kruskall-Wallis would be weaker to use. This is based on ranks, which means a lot of information is thrown out when using Kruskal-Wallis. It is still possible to use this test, but in this case it would not provide better insights than ANOVA.

```
qq_aovres1 <- ggplot(data) + geom_qq(aes(sample=residuals(dietaov)))  + theme_light()  + labs(title="QQ
scatter_aovres1 <- ggplot(dietaov, aes(fitted(dietaov), residuals(dietaov))) + geom_point() + theme_ligh

grid.arrange(qq_aovres1, scatter_aovres1, ncol=2)
```




```
qq_aov2res2 <- ggplot(data) + geom_qq(aes(sample=residuals(dietaov2)))  + theme_light()  + labs(title="(
scatter_aov2res2 <- ggplot(dietaov2, aes(fitted(dietaov2), residuals(dietaov2))) + geom_point() + theme_

grid.arrange(qq_aov2res2, scatter_aov2res2, ncol=2)
```



c)

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

The two-way ANOVAs below show a main effect of diet on weight loss, no main effect of gender on weight loss and an interaction effect between gender and diet on weight loss. The factor diet is sum parametrized because there is no clear reason to pick one of the diets as a base level.

```
data$gender <- as.factor(data$gender)
is.factor(data$gender) # TRUE
```

```
## [1] TRUE
```

```
genderaovm <- lm(weight.lost ~ diet + gender, data = data) # main effects
anova(genderaovm)
```

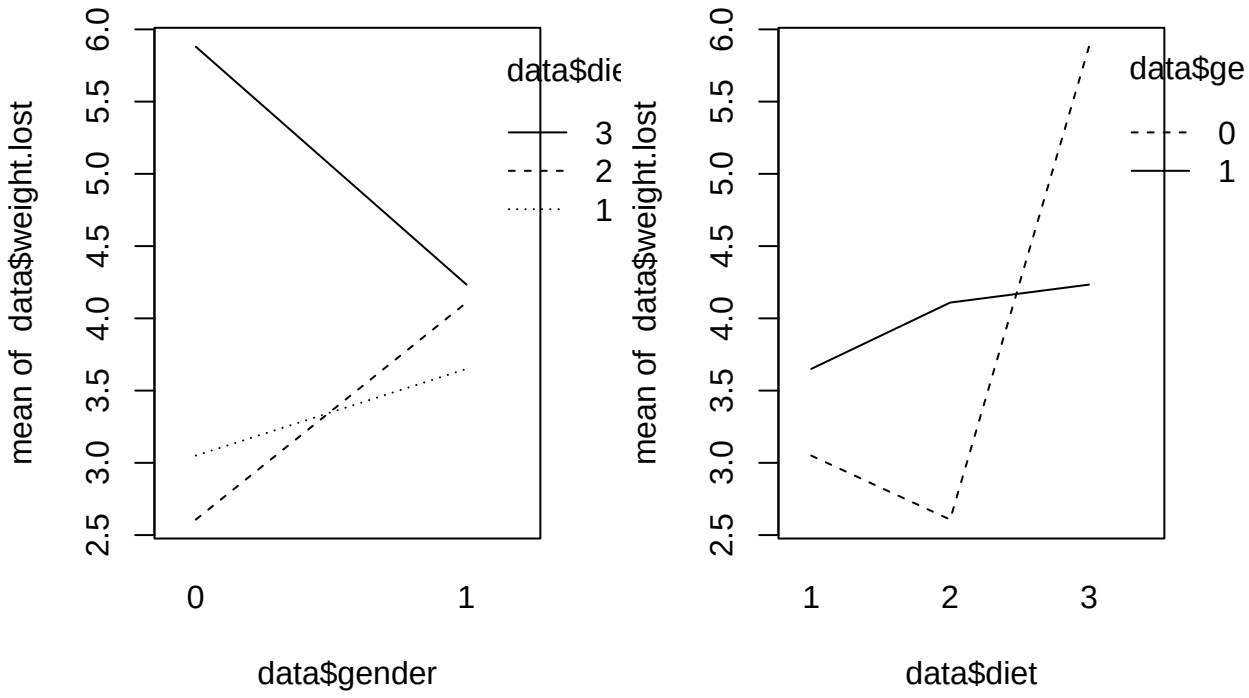
```
## Analysis of Variance Table
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## diet       2     61   30.26    5.31 0.0071 **
## gender     1      0    0.17    0.03 0.8639
## Residuals 72    410    5.70
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
genderaovi <- lm(weight.lost ~ diet * gender, data = data) # interaction effects
anova(genderaovi)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##             Df Sum Sq Mean Sq F value Pr(>F)
## diet         2     61   30.26    5.63 0.0054 **
## gender       1      0    0.17    0.03 0.8599
## diet:gender  2     34   16.95    3.15 0.0488 *
## Residuals   70    376    5.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction plots show that indeed diet matters a lot. The specifically for diet 2 and 3 the difference between the genders is very pronounced. It looks like gender matters less as the lines for diet 1 and 2 look somewhat similar, but the line for diet 3 is very different.

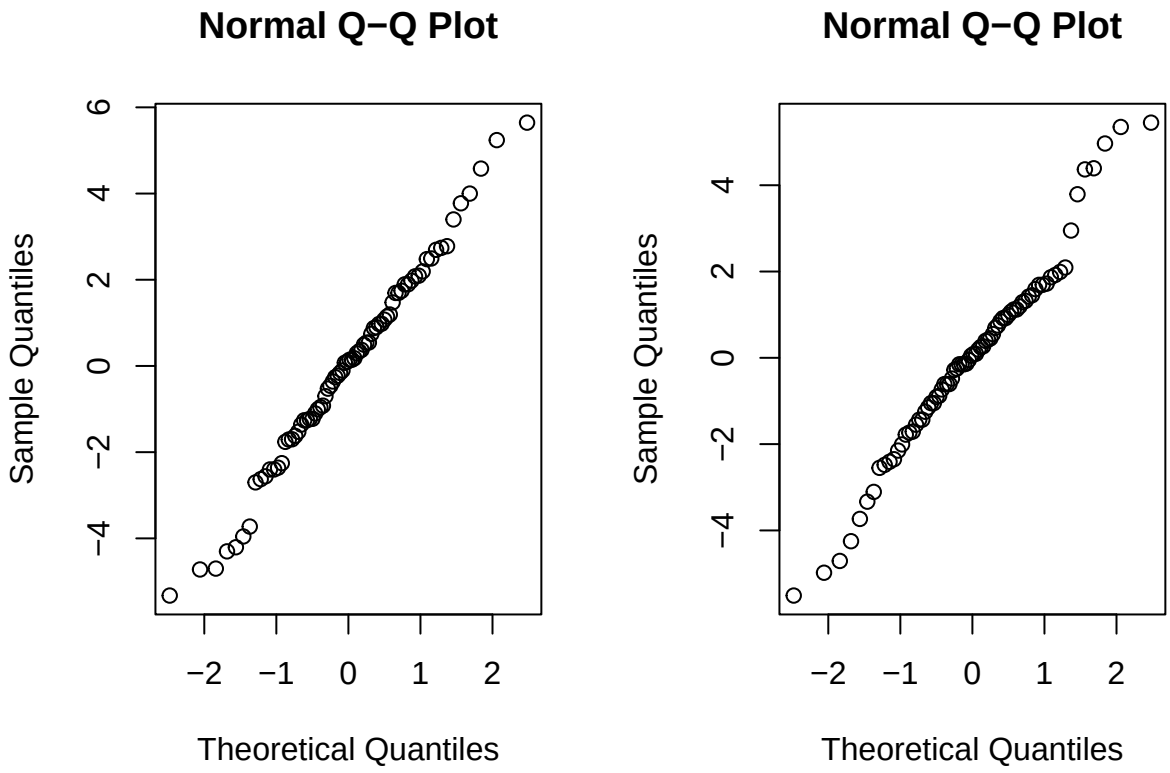
```
par(mfrow=c(1,2))
interaction.plot(data$gender, data$diet, data$weight.lost)
interaction.plot(data$diet, data$gender, data$weight.lost)
```



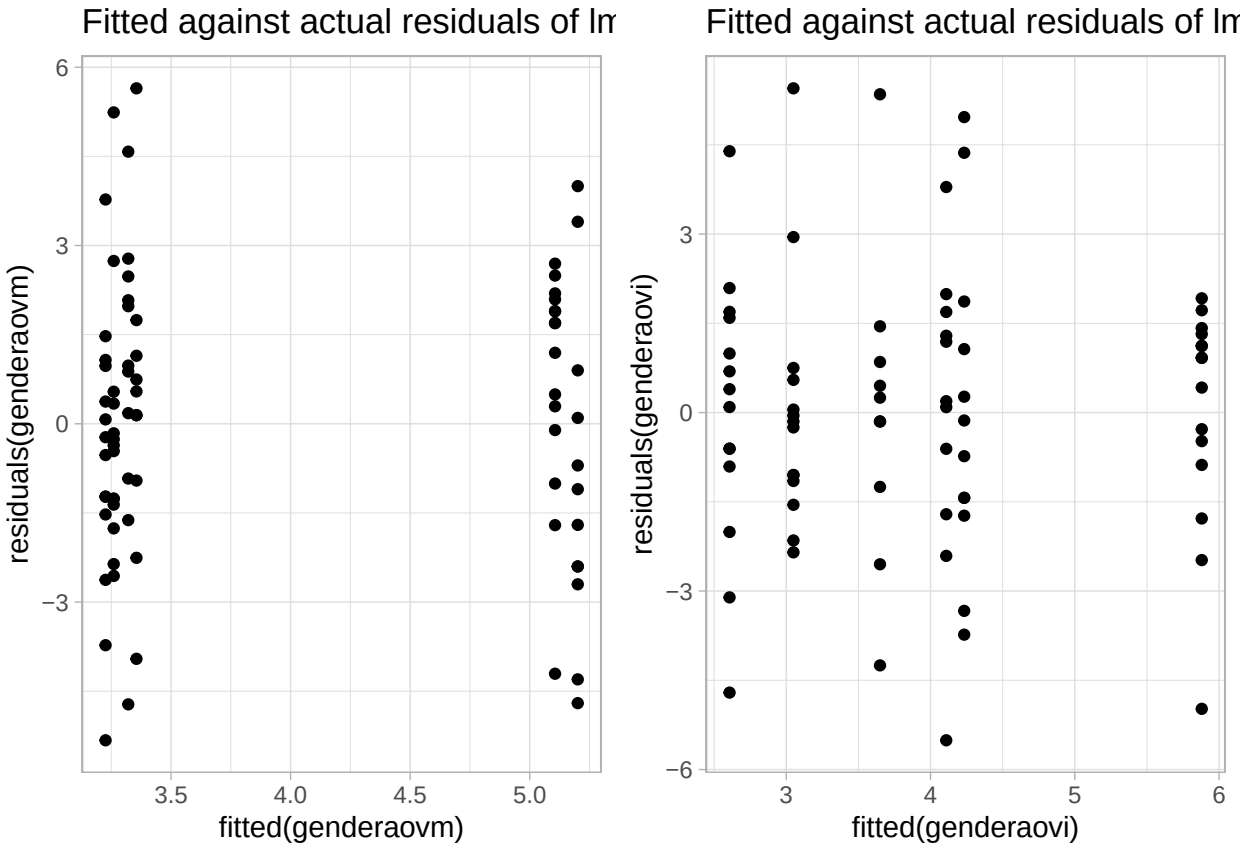
Both the fitted against actual residual plot and qq residual plot indicate that the additive model does not violate the assumption of normality. However, the spread of the residuals looks different in the last column

from the other columns for the interaction model. The QQ residual plot also does not look completely normal. Therefore it is doubtful that the assumption of normality is met for the interaction model.

```
par(mfrow=c(1,2))
qqnorm(residuals(genderaovm))
scatter_genderaovm <- ggplot(genderaovm, aes(fitted(genderaovm), residuals(genderaovm))) + geom_point()

qqnorm(residuals(genderaovi))
```




```
scatter_genderaovi <- ggplot(genderaovi, aes(fitted(genderaovi), residuals(genderaovi))) + geom_point()

grid.arrange(scatter_genderaovm, scatter_genderaovi, ncol=2)
```



d)

Apply an appropriate model to investigate effects of diet and height (and possibly their interaction) on the lost weight. Is the effect of height the same for all 3 types of diet?

The two-way ANOVAs below show a main effect of diet on weight loss, no main effect of height on weight loss and no interaction effect between height and diet on weight loss. The factor diet is sum parametrized because there is no clear reason to pick one of the diets as a base level.

```
data$height <- as.factor(data$height)
is.factor(data$height) # TRUE
```

```
## [1] TRUE
```

```
heightaovm <- lm(weight.lost ~ diet + height, data = data) # main effects
anova(heightaovm)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## diet       2   71.1    35.5    6.86 0.0028 **
## height    36  228.2     6.3    1.22 0.2680
## Residuals 39  202.0     5.2
```

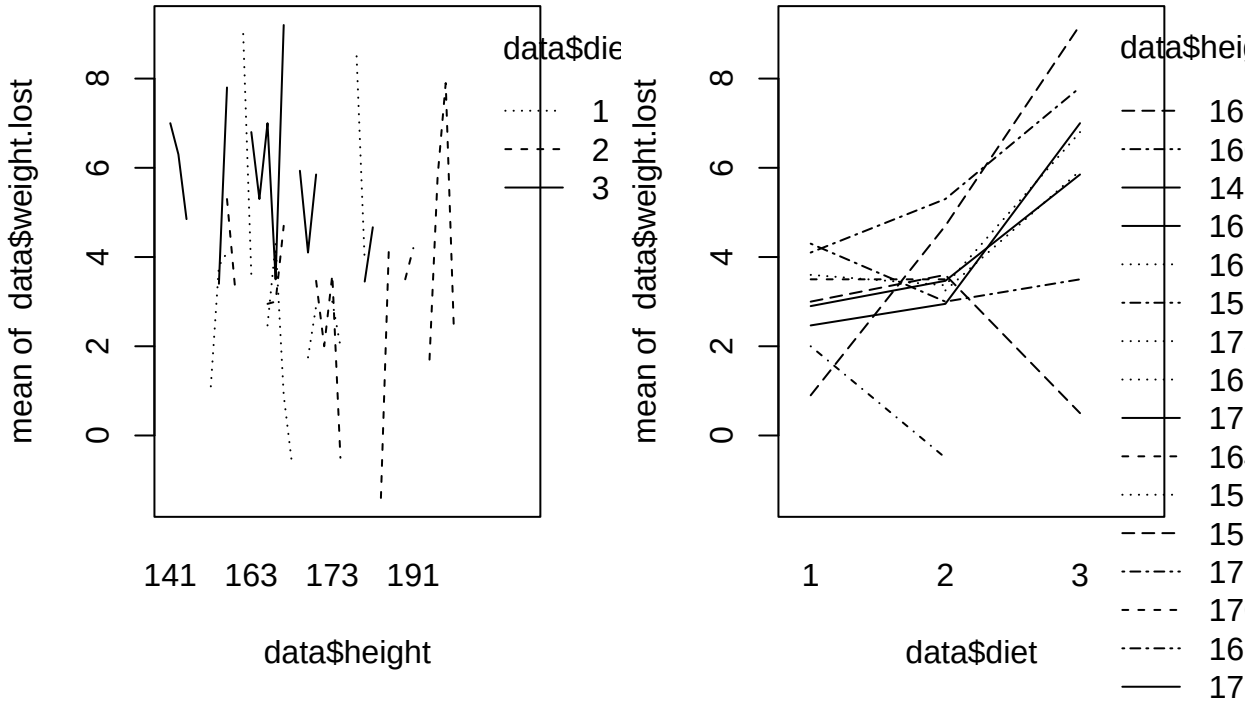
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
heightaovi <- lm(weight.lost ~ diet * height, data = data) # interaction effects
anova(heightaovi)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##             Df Sum Sq Mean Sq F value Pr(>F)
## diet         2   71.1    35.5    5.35  0.013 *
## height      36  228.2     6.3    0.95  0.560
## diet:height 18   62.6     3.5    0.52  0.915
## Residuals   21  139.4     6.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

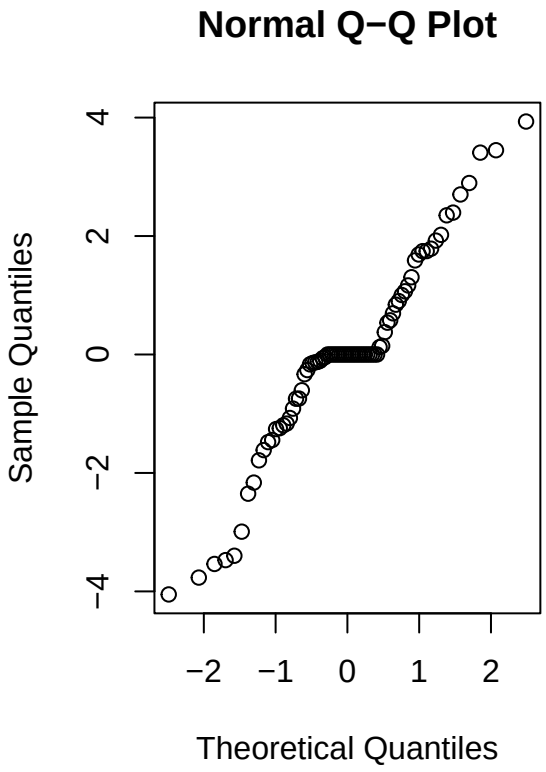
The interaction plots show that lines where diet is fixed vary a lot. No interaction effect and no main effect of height was found, so this variebility is probably because of noise. Some lines where height is fixed behave similarly, but there are also lines which look very different, mostly for diet 2 and 3 they are very spread.

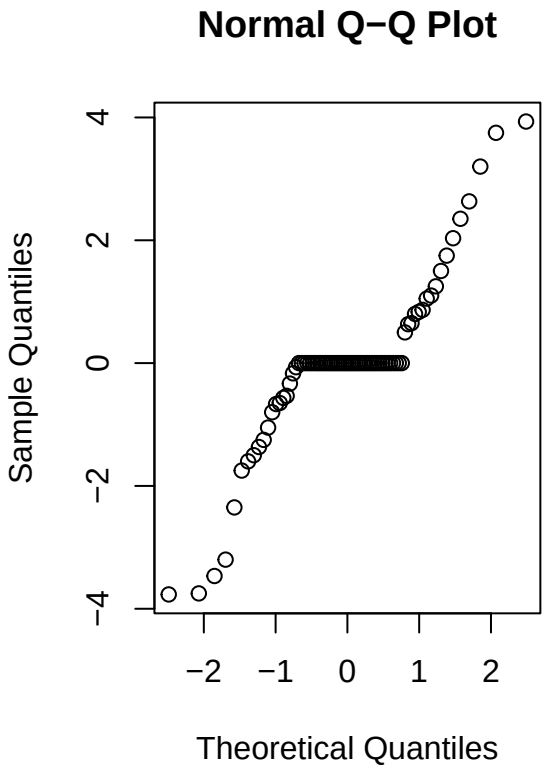
```
par(mfrow=c(1,2))
interaction.plot(data$height, data$diet, data$weight.lost)
interaction.plot(data$diet, data$height, data$weight.lost)
```



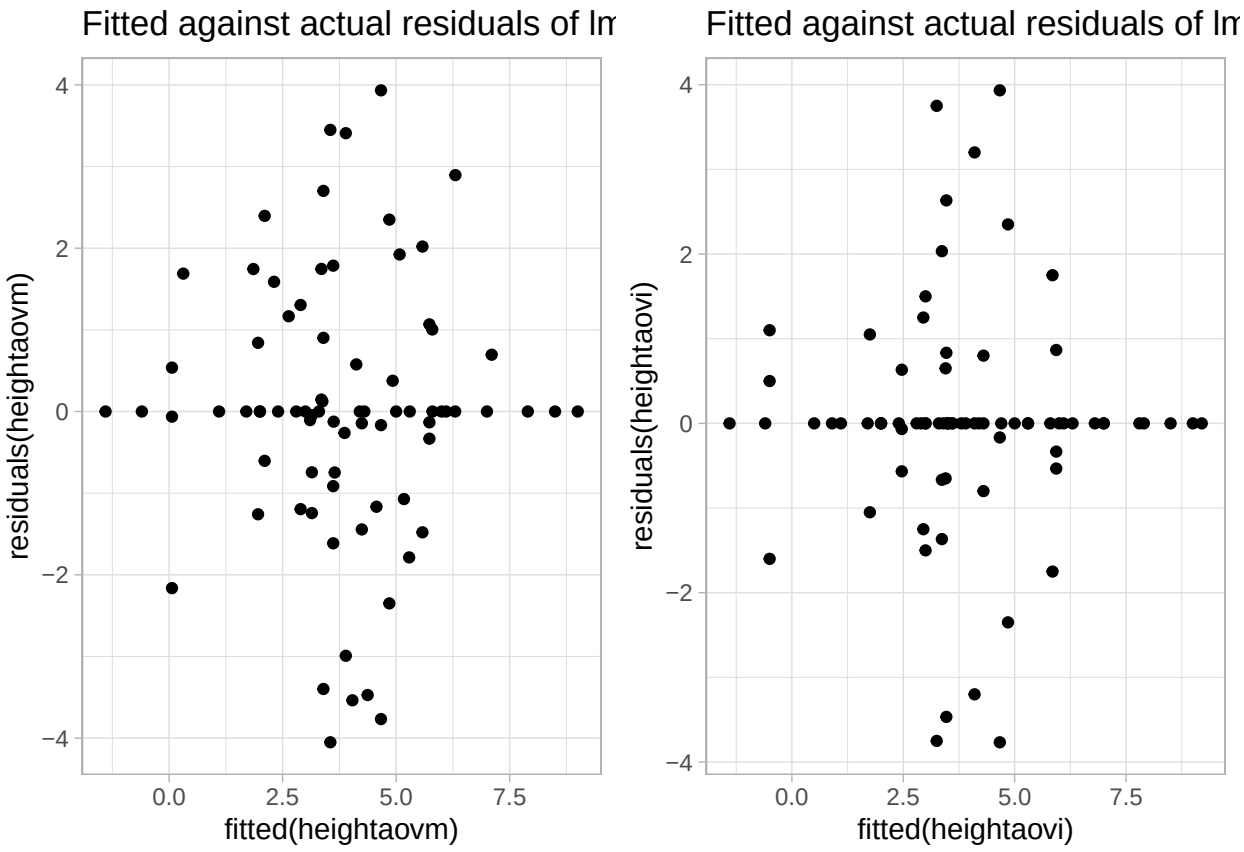
QQ residual plots of the main effect and the interaction models show that both models violate the assumption of normality, which can also be seen om tje fitted against actual residual plots, where there are systematic spread differences.

```
par(mfrow=c(1,2))
qqnorm(residuals(heightaovm))
scatter_heightaovm <- ggplot(heightaovm, aes(fitted(heightaovm), residuals(heightaovm))) + geom_point()

qqnorm(residuals(heightaovi))
```






```
scatter_heightaovi <- ggplot(heightaovi, aes(fitted(heightaovi), residuals(heightaovi))) + geom_point()

grid.arrange(scatter_heightaovm, scatter_heightaovi, ncol=2)
```



###Is the effect of height the same for all 3 types of diet?

e)

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

Excercise 4. Yield of peas