Assignment 4 - Final Assignment

Roel Rotteveel (271547)

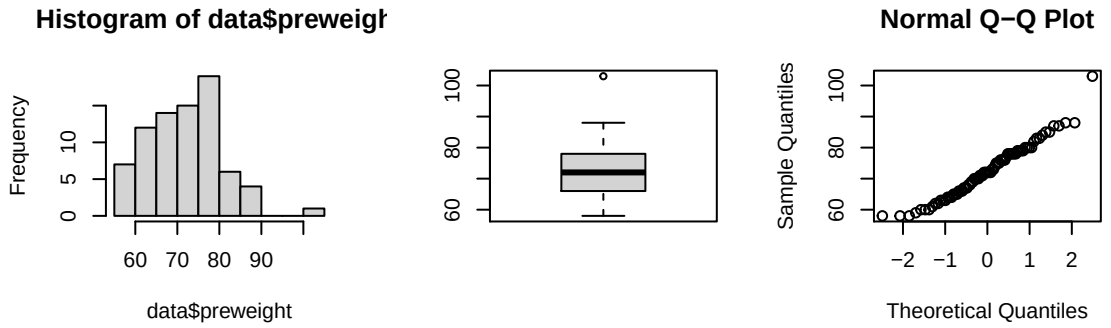
22-3-2021

EDDA Assignment 3

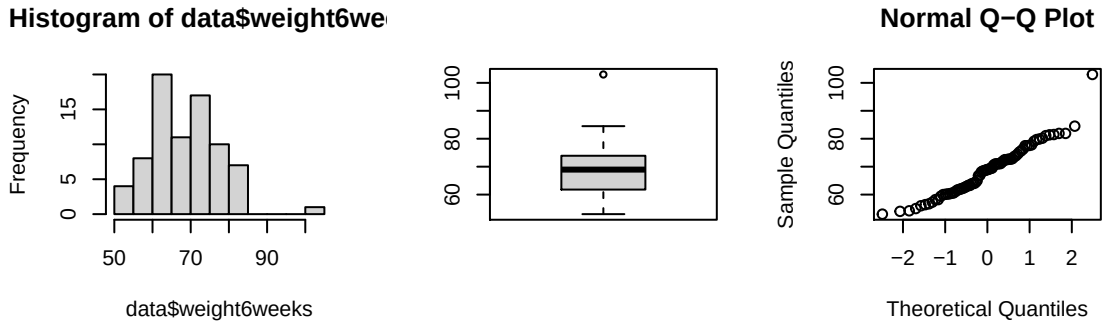
Exercise 1: DIET

A)

```
# delete suspicious rows
# save in different data frame because otherwise creating new variable in b will give issues
# Check normality of data:
par(mfrow=c(1,3))
hist(data$preweight); boxplot(data$preweight); qqnorm(data$preweight)
```




```
hist(data$weight6weeks); boxplot(data$weight6weeks); qqnorm(data$weight6weeks)
```



```r
# Doubtfully normal and paired so paired t test
t.test(data$preweight, data$weight6weeks, paired = TRUE)
```

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

```r
mean(data$preweight) - mean(data$weight6weeks)
```

```
## [1] 3.84
```

When looking at the data there are 2 rows (25 & 26) which contain NA values for gender and are the only rows out of the dataset where there is no difference in 'preweight' and 'weight6weeks'. It seems like something has gone wrong here, which is why the data is deleted.

When looking at the normality of the data we see that for preweight the histogram looks doubtful, (although not completely different than bell-shaped), the boxplot looks normal (with one outlier) and the qqplot looks quite straight but not really from corner to corner. Overall we can say that the data is doubtfully normal.

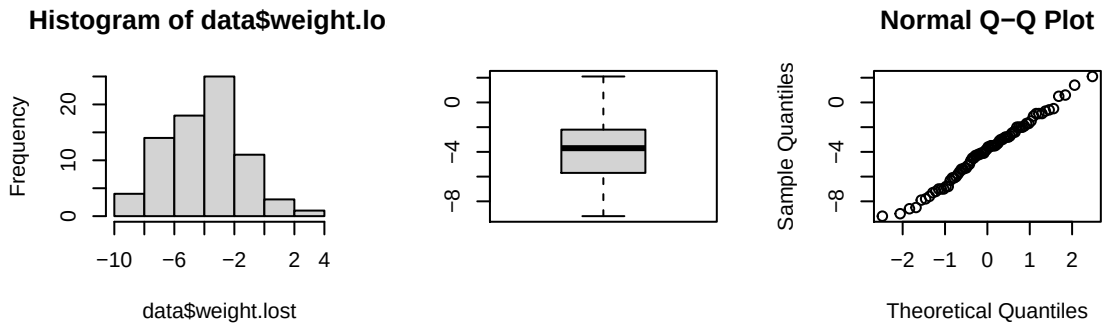
For weight6weeks the histograms looks more normal apart from one extra peak, the boxplot looks similar and the qqplot is more from corner to corner, but not as straight. We can again conclude that this data looks doubtfully normal.

For this last reason apply the (paired) two-sample t-test. The p-value for this test is below 0.05 (<2e-16) meaning it rejects the H0 that the mean of the differences = 0. There is thus a significant difference between both columns. The mean of the differences is estimated to be 3.84, meaning that people lost on average 3.95 kilos due to the diet.

B)

```r
# Add variable weight.lost
data$weight.lost <- data$weight6weeks - data$preweight
# Now apply na.omit to diet, because otherwise line above gives error
data = na.omit(data)
```

```r
# Test normality
par(mfrow=c(1,3))
hist(data$weight.lost); boxplot(data$weight.lost) ; qqnorm(data$weight.lost)
```




```
mean_weight_lost = mean(data$weight.lost)

B=1000 # Number of bootstrap approximations
t = median(data$weight.lost) # sample statistic
tstar=numeric(B)
n = length(data$weight.lost)
for (i in 1:B){
  xstar=rnorm(n,mean = mean_weight_lost) # Generate n pseudo observations according to H0, values that
  tstar[i]= median(xstar) # Compute relevant statistic
}
# hist(data,prob=T)
pl=sum(tstar<t)/B; # Probability(pseudo statistic below sample statistic)
pr=sum(tstar>t)/B; # Probability(pseudo statistic above sample statistic)
p=2*min(pl,pr) # Pick minimal probability*2 (two-tailed) and observe p-value. Lower than 0.05? -> rejec
```

To test the claim if the median is greater than 3 we need to bootstrap. The data looks fairly normal so we sample from normal data. The p-value is higher than 0.05 (0.328) so we do not reject the H0 that

C)

```
diet_aov = lm(data$weight.lost ~ data$diet)
anova(diet_aov)
```

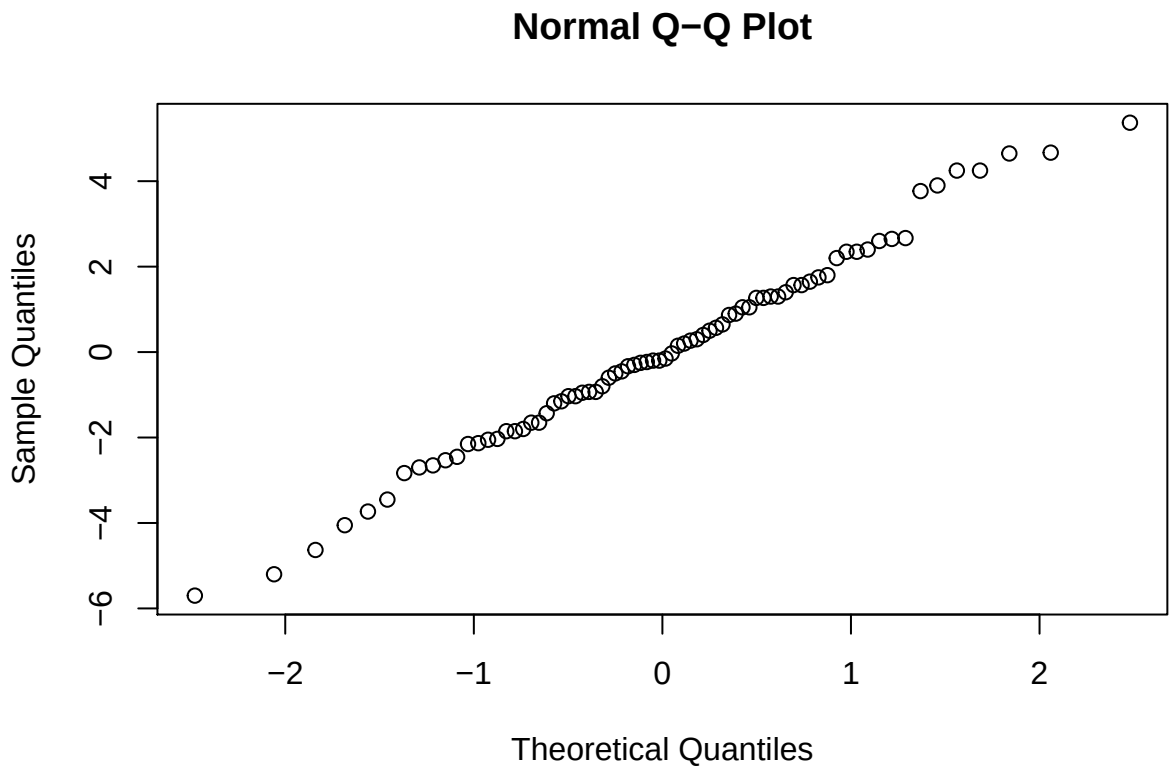
```
## Analysis of Variance Table
##
## Response: data$weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## data$diet  2     61   30.26    5.38 0.0066 **
## Residuals 73    410    5.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(diet_aov)
```

```
##
## Call:
## lm(formula = data$weight.lost ~ data$diet)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.700 -1.652 -0.176  1.442  5.368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.300      0.484   -6.82  2.3e-09 ***
## data$diet2     0.032      0.678    0.05   0.9625
## data$diet3    -1.848      0.665   -2.78   0.0069 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.37 on 73 degrees of freedom
## Multiple R-squared:  0.129,  Adjusted R-squared:  0.105
## F-statistic: 5.38 on 2 and 73 DF,  p-value: 0.0066
```

```
qqnorm(residuals(diet_aov)) # check normality of residuals
```




```
-3.3-1.848
```

```
## [1] -5.15
```

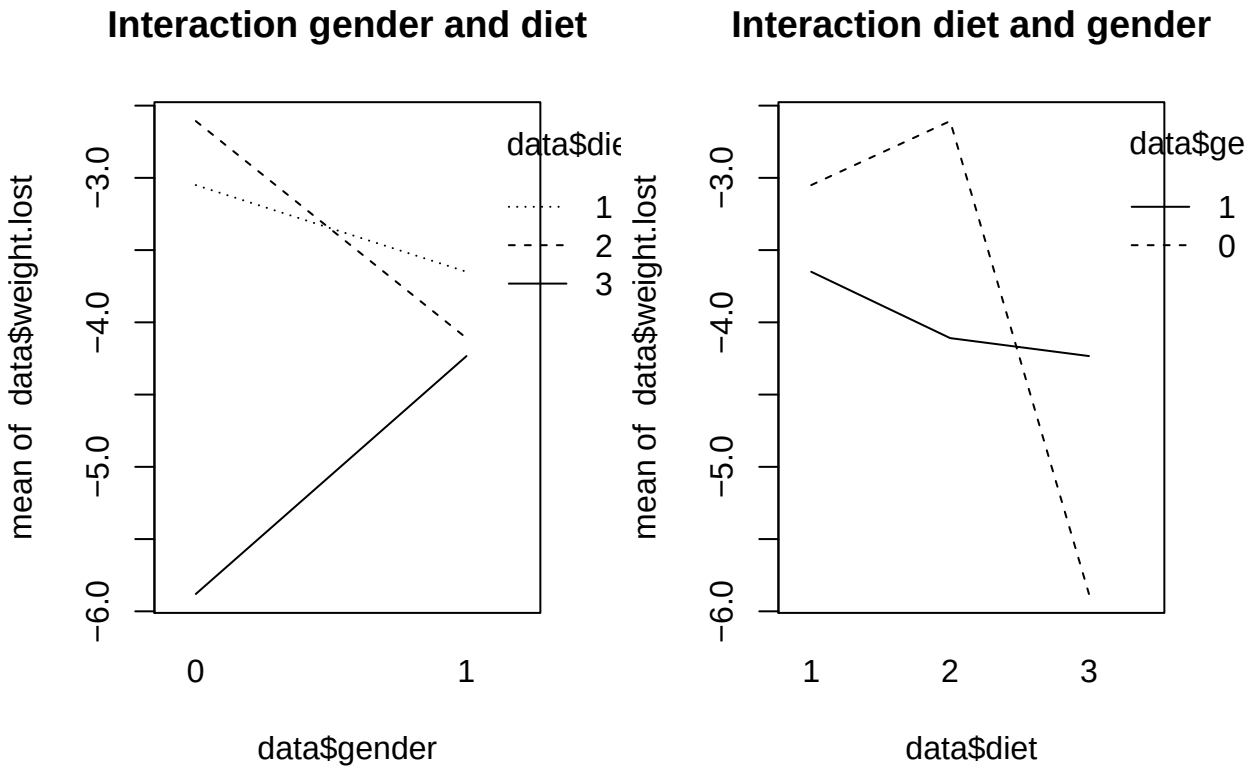
When performing the anova we see that that the p-value for diet is below 0.05 (0.0032) so significant, the H0 that the means for all factor levels are the same (no factor effect) is thus rejected. When looking at the

summary we see that the p-value for diet 2 is insignificant (0.6845), showing that not all levels of the factors are significant. Diet 3 is significant (p value < 0.05: 0.0069) thus rejecting the H0 that the level effect is zero. All three types of diets lead to weight loss (since we have negative weight loss), with diet 3 leading to the most weight loss of -3.3-1.848 = -5.15 kg of weight loss.

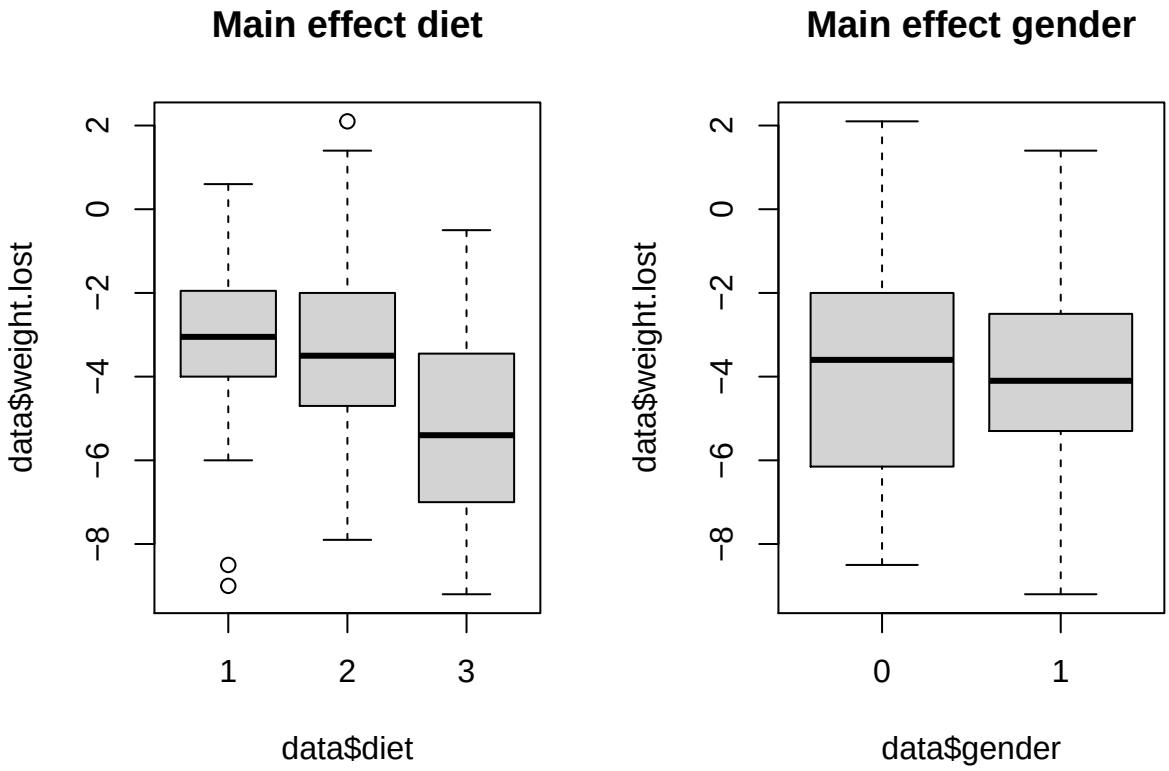
When looking at the qqplot of the residuals we see that the residuals look normal, with a straight line from corner to corner. This assumption of the model thus holds

D)

```
par(mfrow=c(1,2))
data = na.omit(data)
rownames(data) = NULL
# Interaction plots
interaction.plot(data$gender, data$diet, data$weight.lost, main = "Interaction gender and diet")
interaction.plot(data$diet,data$gender,data$weight.lost, main = "Interaction diet and gender")
```




```
# Check main effects
boxplot(data$weight.lost ~ data$diet, main = "Main effect diet")
boxplot(data$weight.lost ~ data$gender, main = "Main effect gender")
```




```
diet_2aov = lm (data$weight.lost ~ data$diet * data$gender)
anova(diet_2aov)
```

```
## Analysis of Variance Table
##
## Response: data$weight.lost
##                       Df Sum Sq Mean Sq F value Pr(>F)
## data$diet              2     61   30.26    5.63 0.0054 **
## data$gender            1      0    0.17    0.03 0.8599
## data$diet:data$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
```

```
summary(diet_2aov)
```

```
##
## Call:
## lm(formula = data$weight.lost ~ data$diet * data$gender)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.45  -1.22  -0.07   1.30   5.51
##
## Coefficients:
```

```
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -3.050      0.620   -4.92  5.5e-06 ***
## data$diet2                   0.443      0.876    0.51   0.6149
## data$diet3                  -2.830      0.862   -3.28   0.0016 **
## data$gender1                -0.600      0.960   -0.62   0.5340
## data$diet2:data$gender1     -0.902      1.340   -0.67   0.5030
## data$diet3:data$gender1      2.247      1.315    1.71   0.0919 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.32 on 70 degrees of freedom
## Multiple R-squared:  0.201,  Adjusted R-squared:  0.144
## F-statistic: 3.52 on 5 and 70 DF,  p-value: 0.00677
```

Looking at the interaction plots the lines are not parallel and intersecting, suspecting interaction between the two factors. We thus test the interactive model. Looking at the boxplots the levels seem pretty similar (for both boxplots) apart from diet 3, which is a bit lower than the other two diets.

The anova test indeed shows that there is interaction between diet and gender, with a p-value (just) below 0.05 (0.0488), thus rejecting the null hypothesis that there is no effect from the interaction. We thus keep this model.

When looking at the summary of this model we see that for for a woman using diet 1 the average weight loss is -3.050. Diet 2 is insignificant (p: value 0.6149 $>$ 0.05, failing to rejecting H0 that effect of this level is 0) but adds weight (thus reducing absolute weight loss: 0.443) while diet 3 is significant (p-value: smaller than 0.05 (0.0016) thus rejecting abovementioned H0) and is good for another reduction in weight of -2.83 kilos on average. The p-values of gender and the interactions are all above 0.05 and thus insignificant, but lead to a further weight loss for man, except when men use diet3, which leads to an increase of 2.247 kilos.

The friedman test is not relevant here as it assumes that the data is not from a normal population (which weight.lost is, as proven in b). If the data was non-parametric the only way we could use it was if we assumed Gender to be a block (friedman test assumes we have outcome, factor and block). This is of course a design choice but in this case it seems logical that gender is a factor and not a block.

E)

Since we now have a factor and an explanatory variable we use ANCOVA:

```r
# Test for interaction
diet_acov = lm (data$weight.lost~data$diet * data$height)
drop1(diet_acov, test = "F")
```

```
## Single term deletions
##
## Model:
## data$weight.lost ~ data$diet * data$height
##                       Df Sum of Sq RSS AIC F value Pr(>F)
## <none>                             396 138
## data$diet:data$height  2      13.8 410 136    1.22    0.3
```

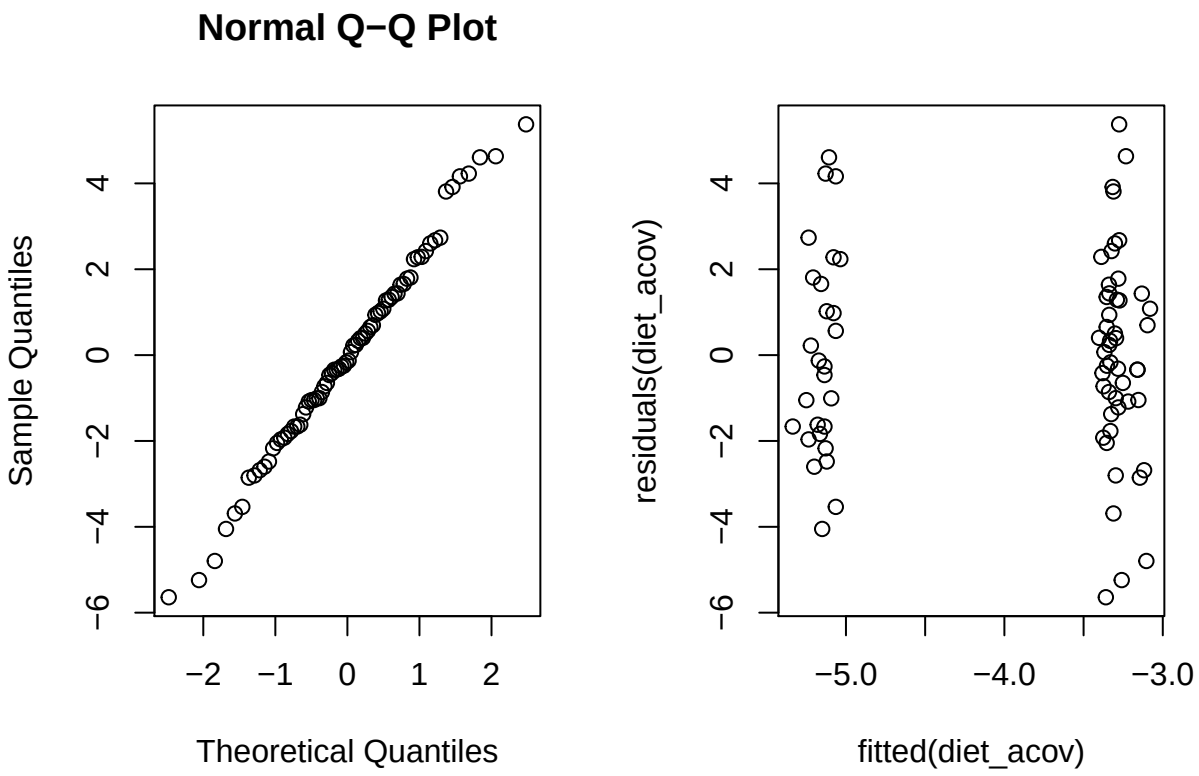
```r
# Additive model
diet_acov = lm (data$weight.lost~data$diet + data$height)
drop1(diet_acov, test = "F")
```

```
## Single term deletions
##
## Model:
## data$weight.lost ~ data$diet + data$height
##             Df Sum of Sq RSS AIC F value Pr(>F)
## <none>                   410 136
## data$diet    2      54.9 465 142    4.82  0.011 *
## data$height  1       0.5 410 134    0.08  0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

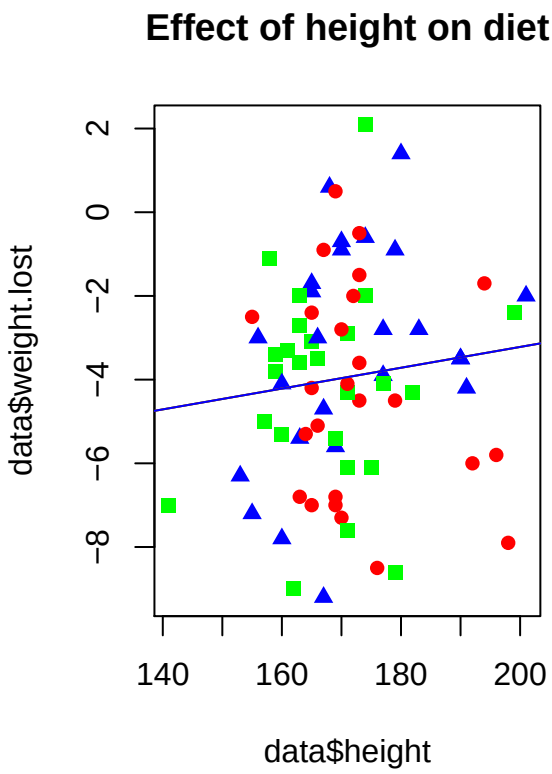
```
summary(diet_acov)
```

```
##
## Call:
## lm(formula = data$weight.lost ~ data$diet + data$height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.641 -1.632 -0.149  1.434  5.375
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.51848    4.31384   -1.05   0.2984
## data$diet2  -0.00198    0.69229    0.00   0.9977
## data$diet3  -1.82645    0.67375   -2.71   0.0084 **
## data$height  0.00716    0.02517    0.28   0.7770
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.39 on 72 degrees of freedom
## Multiple R-squared:  0.13,   Adjusted R-squared:  0.0932
## F-statistic: 3.57 on 3 and 72 DF,  p-value: 0.0181
```

```
par(mfrow=c(1,2))
qqnorm(residuals(diet_acov))
plot(fitted(diet_acov),residuals(diet_acov))
```




```
# Test if
plot(data$weight.lost~data$height, pch= c(15, 16, 17), col=c("green", "red", "blue"))
legend(0.64, 4.5, legend=c("diet1", "diet2", "diet3"), pch= c(15, 16, 17), col=c("green", "red", "blue"))
title("Effect of height on diet")
abline(lm(data$weight.lost~data$height,data=data[diet=="1",]), col = "green")
abline(lm(data$weight.lost~data$height,data=data[diet=='2',]), col = "red")
abline(lm(data$weight.lost~data$height,data=data[diet=="3",]), col = "blue")
```



When testing for interaction the p-value is above 0.05 (0.3) thus the H0 that there is no interaction effect is not rejected. We thus use the additive model. The qqplot of the residuals looks pretty normal and the fitted values vs the residuals looks fairly homogenous. When looking at the summary only the third level of diet is significant (p value of 0.0084, smaller than 0.05, thus rejecting H0 that the effect of this factor level is null) and height is insignificant (p value of 0.770, greater than 0.05, thus not rejecting H0 that the effect of height is zero) .

To test if the lost weight is the same for all 3 types of diet we plot the 3 different levels.

My computer unfortunately only plots the last line but I have plotted them one-for-one and they seem fairly parallel. I therefore conclude the effect of height is the same for all 3 types of diet (since we can also hand in R-code I hope this is good enough, as this worked earlier)

F)

I prefer the one from c, as this is a simpler model where the factor diet is significant, whereas the explanatory variable height is insignificant for e. I thus use this model to predict the lost weight:

```
diet_aov = lm(data$weight.lost ~ data$diet)
summary(diet_aov)
```

```
##
## Call:
## lm(formula = data$weight.lost ~ data$diet)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -5.700 -1.652 -0.176  1.442  5.368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.300      0.484   -6.82  2.3e-09 ***
## data$diet2     0.032      0.678    0.05   0.9625
## data$diet3    -1.848      0.665   -2.78   0.0069 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.37 on 73 degrees of freedom
## Multiple R-squared:  0.129,  Adjusted R-squared:  0.105
## F-statistic: 5.38 on 2 and 73 DF,  p-value: 0.0066
```

```
# Diet 1: -3.300
# Diet 2:
-3.300 + 0.032
```

```
## [1] -3.27
```

```
# Diet 3:
-3.300 - 1.848
```

```
## [1] -5.15
```

Table 1: Average weight loss per diet (for average person)

Diet 1	Diet 2	Diet 3
-3.3	-3.27	-5.15

G)

```
data$lost.4kg = abs(data$weight.lost)>4
```

b) no we cannot test this since we cannot test median in the same way as before, as we only have the binary information whether the weight loss is more or less than 4 kg

c) We can do a logistic regression instead of an anova since we have a binary response variable now (if we assume weight lost is replace with lost.4kg since that is the most logical way to subtract weight6weeks from pre-weight)

```
logisticmodel = glm(data$lost.4kg~data$diet,family=binomial)
drop1(logisticmodel, test="Chisq")
```

```
## Single term deletions
##
## Model:
```

```
## data$lost.4kg ~ data$diet
##           Df Deviance AIC LRT Pr(>Chi)
## <none>           94.1 100
## data$diet  2    105.1 107  11    0.004 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(logisticmodel)
```

```
##
## Call:
## glm(formula = data$lost.4kg ~ data$diet, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.560  -1.077  -0.758   0.838   1.665
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.099      0.471   -2.33   0.0198 *
## data$diet2     0.857      0.620    1.38   0.1668
## data$diet3     1.964      0.632    3.11   0.0019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.148  on 75  degrees of freedom
## Residual deviance:  94.104  on 73  degrees of freedom
## AIC: 100.1
##
## Number of Fisher Scoring iterations: 4
```

```
exp(-1.099) # diet 1
```

```
## [1] 0.333
```

```
exp(-1.099 + 0.857) # diet 2
```

```
## [1] 0.785
```

```
exp(-1.099 + 1.964) # diet 3
```

```
## [1] 2.38
```

When testing the anova with chisq-test we see diet is significant (p-value smaller than 0.05: 0.004, thus rejecting H0 that the factor effect is 0). When looking at the summary we see that now the odds for a successful outcome (losing more than 4 kg) : diet 1 (0.33), diet 2(0.785), diet 3(2.38). The 3rd diet is thus again the best diet for losing weight

d)

```
logisticmodel_gender = glm(data$lost.4kg~data$diet * data$gender,family=binomial)
drop1(logisticmodel_gender, test="Chisq")
```

```
## Single term deletions
##
## Model:
## data$lost.4kg ~ data$diet * data$gender
##                       Df Deviance   AIC LRT Pr(>Chi)
## <none>                      84.5  96.5
## data$diet:data$gender  2     93.7 101.7 9.2     0.01 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(logisticmodel_gender)
```

```
##
## Call:
## glm(formula = data$lost.4kg ~ data$diet * data$gender, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.007  -0.820  -0.555   0.951   1.973
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.7918     0.7638   -2.35  0.01898 *
## data$diet2                 0.8755     0.9661    0.91  0.36483
## data$diet3                 3.6636     1.0772    3.40  0.00067 ***
## data$gender1               1.3863     1.0000    1.39  0.16566
## data$diet2:data$gender1    0.0896     1.3202    0.07  0.94588
## data$diet3:data$gender1   -3.2581     1.3821   -2.36  0.01841 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.148  on 75  degrees of freedom
## Residual deviance:  84.531  on 70  degrees of freedom
## AIC: 96.53
##
## Number of Fisher Scoring iterations: 4
```

```
exp(-1.7918)
```

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

```
exp(-1.7918-3.2581)
```

```
## [1] 0.00641
```

```
exp(-1.7918+3.6636)
```

```
## [1] 6.5
```

When performing the anova for the interactive model with gender we now see that the interaction is significant (p-value below 0.05: 0.01, thus rejecting H0 of no interaction effect). We thus use the interactive model.

From the summary of the model we witness that the odds of a successful outcome are 0.167 for a woman using diet 3. The odds also decrease for a man using diet 3 (same finding as before) but for women using diet 3 only increases their odds. Being a woman indeed seems to be a beneficial combination as the odds for losing 4 kg are 6.5.

e) it is possible to test for height but is is impossible to plot the 3 different levels as before, as we would now plot against a binary variable. The first part can be examined however:

```
logisticmodel_height = glm(data$lost.4kg~data$diet * data$height,family=binomial)
drop1(logisticmodel_height, test="Chisq")
```

```
## Single term deletions
##
## Model:
## data$lost.4kg ~ data$diet * data$height
##                        Df Deviance AIC  LRT Pr(>Chi)
## <none>                       92.7 105
## data$diet:data$height  2     94.1 102 1.31     0.52
```

```
logisticmodel_height = glm(data$lost.4kg~data$diet + data$height,family=binomial)
drop1(logisticmodel_height, test="Chisq")
```

```
## Single term deletions
##
## Model:
## data$lost.4kg ~ data$diet + data$height
##             Df Deviance AIC   LRT Pr(>Chi)
## <none>             94.1 102
## data$diet    2    104.7 109 10.67   0.0048 **
## data$height  1     94.1 100  0.04   0.8364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(logisticmodel_height)
```

```
##
## Call:
## glm(formula = data$lost.4kg ~ data$diet + data$height, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.586  -1.053  -0.738   0.855   1.711
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.30078    3.90050   -0.08   0.9385
## data$diet2   0.88014    0.63038    1.40   0.1627
## data$diet3   1.95043    0.63521    3.07   0.0021 **
## data$height -0.00469    0.02277   -0.21   0.8369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105.148  on 75  degrees of freedom
## Residual deviance:  94.061  on 72  degrees of freedom
## AIC: 102.1
##
## Number of Fisher Scoring iterations: 4
```

When looking at the anova of the interactive model we see that the p value is insignificant for the interaction (p-value > 0.05: 0.52, thus not rejecting H0 of no interaction effect). We thus now use the additive model and find that height is not significant now (p-value > 0.05: 0.84, thus not rejecting H0 of no effect), so it is better to look at the model mentioned above this.