# EDDA - Final Assignment

Ignas Krikštaponis

## Diet

```
# read the data
data <- read.table(file="data/diet.txt", header=TRUE)
data$diet <- as.factor(data$diet)
data$gender <- as.factor(data$gender)
# create new variable
data <- data %>% mutate(weight.lost = preweight - weight6weeks)
```

There seems to be two rows with NA values for gender. Upon closer inspection both of these rows have 0 weight loss and one of the rows has substantially higher weight than the rest. Based on this and the fact that there is no possibility to gather information on what went wrong with data collection, these data points will be removed from the rest of the analysis.

```
# remove missing data points
data <- data %>% filter(!is.na(gender))
```

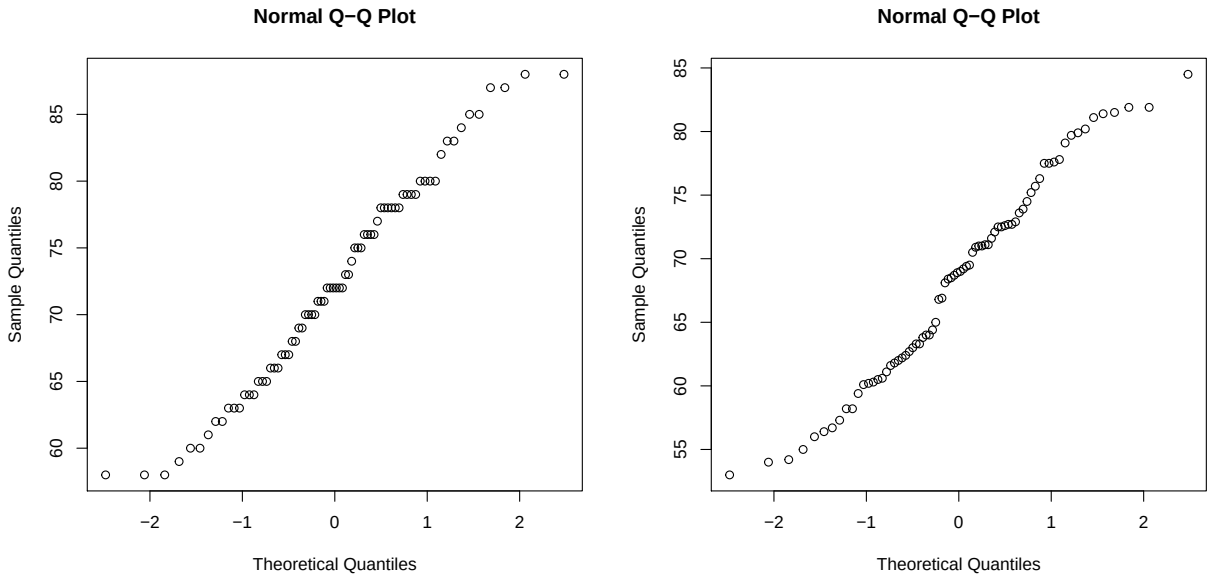
**a)**

```
# perform a paired t-test to see if the weight before and after are significantly different
t.test(data$preweight, data$weight6weeks, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data$preweight and data$weight6weeks
## t = 14, df = 75, p-value <2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.37 4.52
## sample estimates:
## mean of the differences
##                    3.95
```

We perform a paired t-test to check whether there is significant difference between the population means. The test is paired since the weight measures were carried out on the same person. From the p-value above $< 0.05$ we can conclude that there is significant affect of the diet.

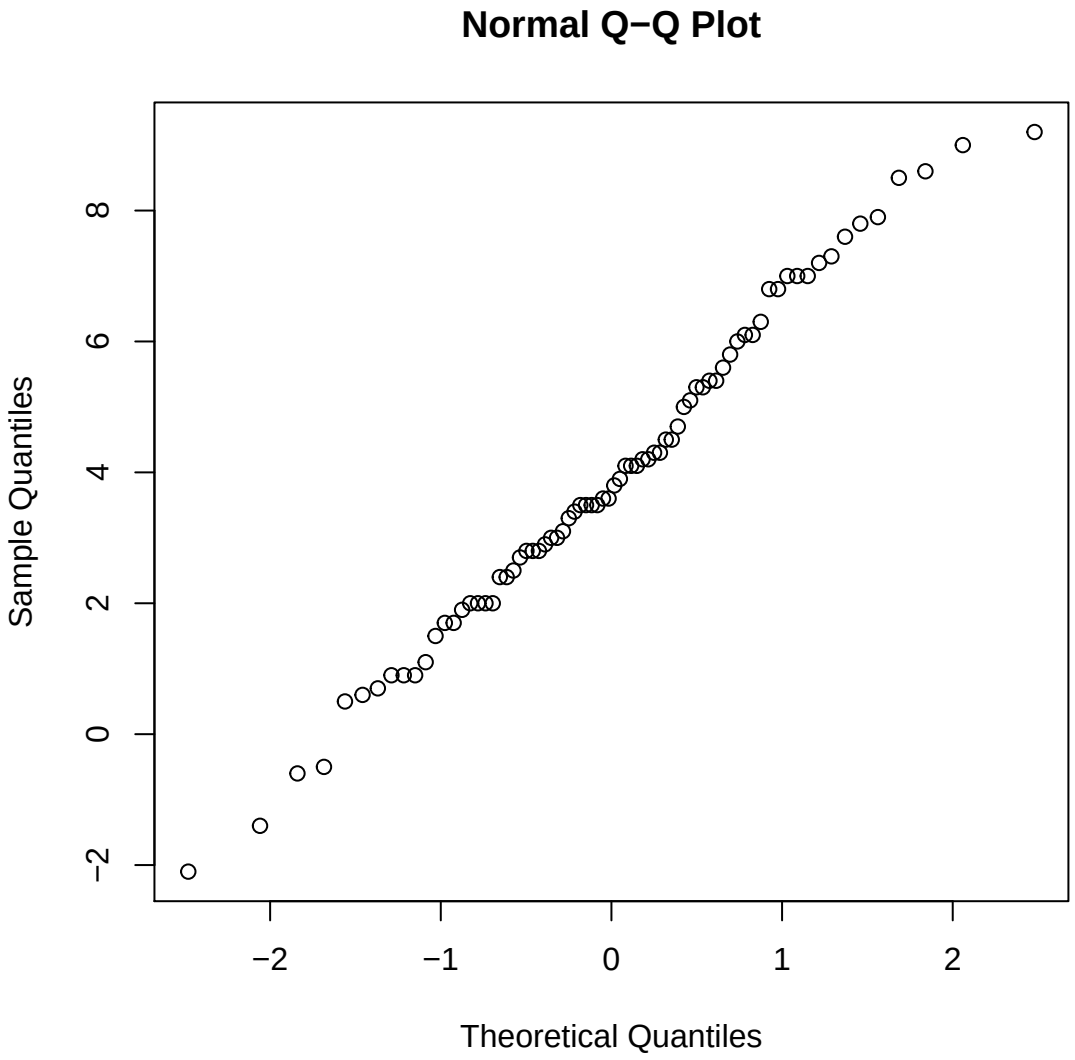
```
# diagnostics
par(mfrow=c(1,2)); qqnorm(data$preweight); qqnorm(data$weight6weeks)
```




QQ-plots above follow a straight line well, therefore confirming the normality assumption required in the t-test.

**b)**

```
# diagnostics
qqnorm(data$weight.lost)
```




```
shapiro.test(data$weight.lost)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$weight.lost
## W = 1, p-value = 0.8
```

```
# perform one-tailed t-test
round(t.test(data$weight.lost, mu=3, alternative = "greater", conf.level = 0.95)$p.value, 3)
```

```
## [1] 0.001
```

From the normality diagnostics we see that the data can be assumed to be normally distributed, therefore we can apply a one-tailed t-test on the sample mean. As the data is normally distributed the median is assumed

to be the same as the mean. From the p-value above we can conclude that the median is significantly higher than 3.

**c)**

```
# perform one-way ANOVA
model <- lm(weight.lost ~ diet, data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## 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(model)$coefficients
```

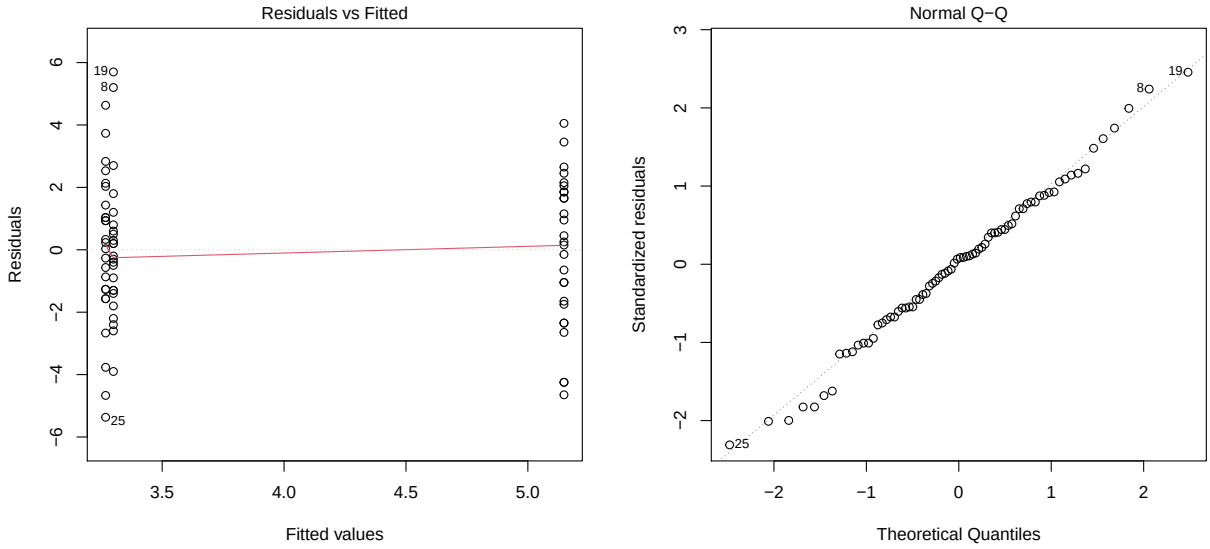
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.300      0.484  6.8183 2.26e-09
## diet2         -0.032      0.678 -0.0472 9.62e-01
## diet3          1.848      0.665  2.7784 6.94e-03
```

```
paste("r-squared:", round(summary(model)$r.squared, 4))
```

```
## [1] "r-squared: 0.1285"
```

From the one-way ANOVA analysis above we can conclude that diet has a significant effect on weight loss (p-value $< 0.05$). From the summary table we see that all diet types result in weight loss. Diet3 results in the best weight loss as its estimate is the highest $3.3 + 1.848 = 5.15$.

```
# diagnostics
par(mfrow=c(1,2));
plot(model, 1); plot(model, 2)
```




From the diagnostics above we do not observe any obvious relationship in the Fitted vs Residuals plot - which is the desired behaviour here. From the normal QQ-plot we see that the residuals follow a straight line very well. Therefore, we can conclude that the model assumptions have been met.

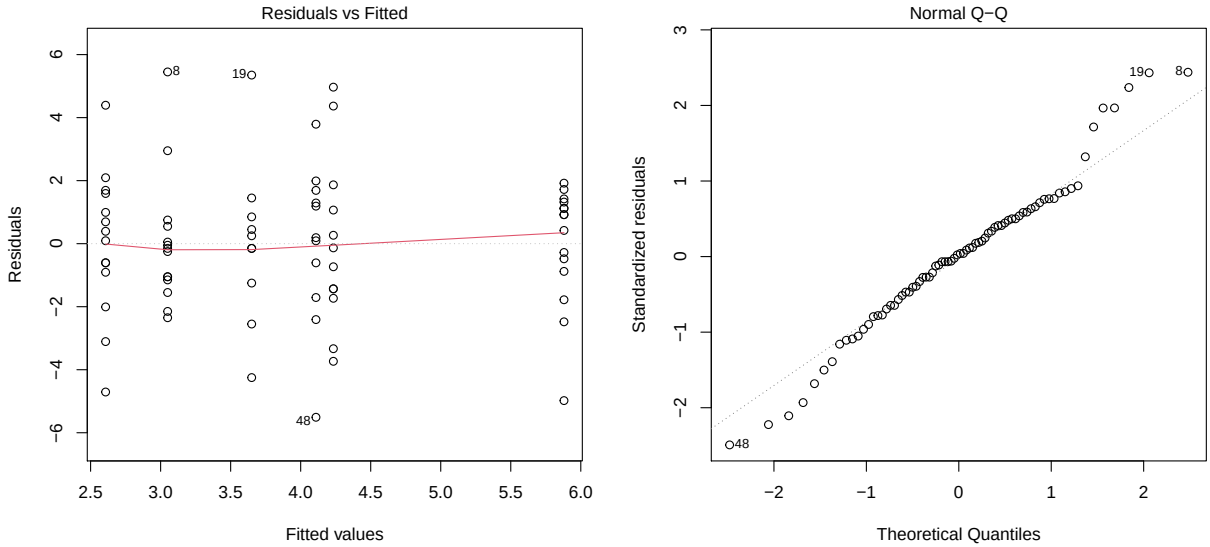
**d)**

```
# perform two-way ANOVA with interaction
model <- lm(weight.lost ~ diet * gender, data = data)
anova(model)
```

```
## 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
```

From the results above we can see that there is significant interaction between gender and diet. For this reason the additive ANOVA model will not be investigated as this would not be valid.

```
# diagnostics
par(mfrow=c(1,2));
plot(model, 1); plot(model, 2)
```

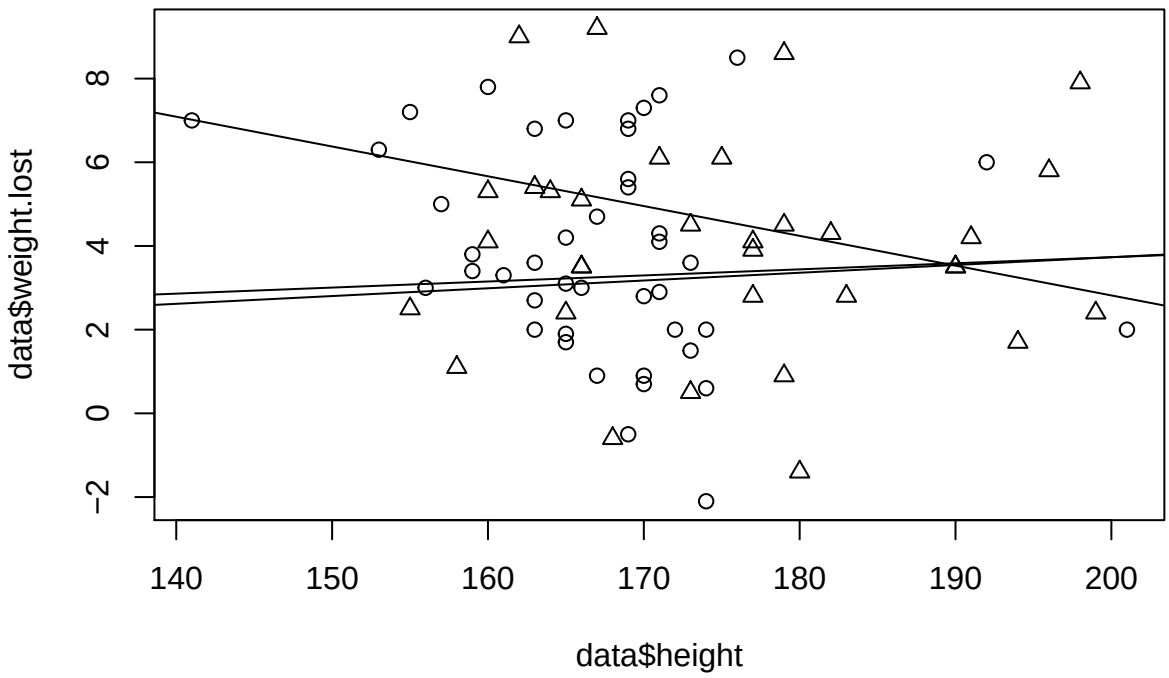



Diagnostics: From the diagnostics above we do not observe any obvious relationship in the Fitted vs Residuals plot - which is the desired behaviour here. From the normal QQ-plot we see that the residuals somewhat follows a straight line, but there are some outlier at the extremes that raise question about the normality assumption.

Friedman test is applicable to randomised block design and repeated measures. As the question here asked for investigation into the effects of both gender and diet (with interaction) the Friedman test would not be relevant as it would only be able to answer the question if diet has an effect (if we take gender as the blocking factor). However, the Friedman test could still be applied here.

**e)**

```
# investigate visualy
plot(data$weight.lost~data$height,pch=unclass(data$gender))
diets <- unique(as.character(data$diet))
for (i in diets) {
    abline(lm(weight.lost~height
    ,data=data[data$diet==i,]))}
```

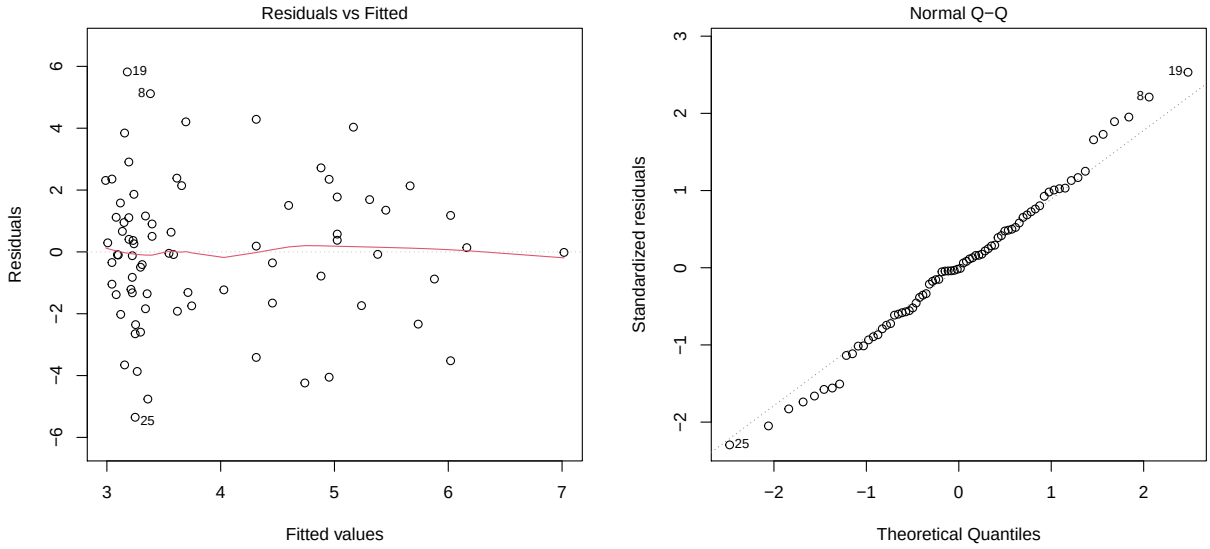



From the graph above we can not observe any obvious relationship between weight lost and height. There seems to be some difference in the slopes between the two gender, but it still need to be investigate via ANCOVA to determine whether it is significant.

```
# perform ANCOVA with interaction
model <- lm(weight.lost ~ diet*height, data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##             Df Sum Sq Mean Sq F value Pr(>F)
## diet         2     61   30.26    5.35 0.0069 **
## height       1      0    0.46    0.08 0.7764
## diet:height  2     14    6.92    1.22 0.3007
## Residuals   70    396    5.66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# diagnostics
par(mfrow=c(1,2));
plot(model, 1); plot(model, 2)
```




First we perform the ANCOVA analysis with interaction to investigate if the effect of height is the same under all diet types. From the analysis above we can see that p-value for interaction component is $> 0.05$, meaning it is insignificant. From this we can conclude that the effect of height is the same under all diet types.

From the diagnostics above we do not observe any obvious relationship in the Fitted vs Residuals plot - which is the desired behaviour here. From the normal QQ-plot we see that the residuals follow a straight line very well. Therefore, we can conclude that the model assumptions have been met.

Now let's move on to additive ANCOVA:

```
# perform additive ANCOVA
model <- lm(weight.lost ~ height + diet, data = data)
anova(model)
```

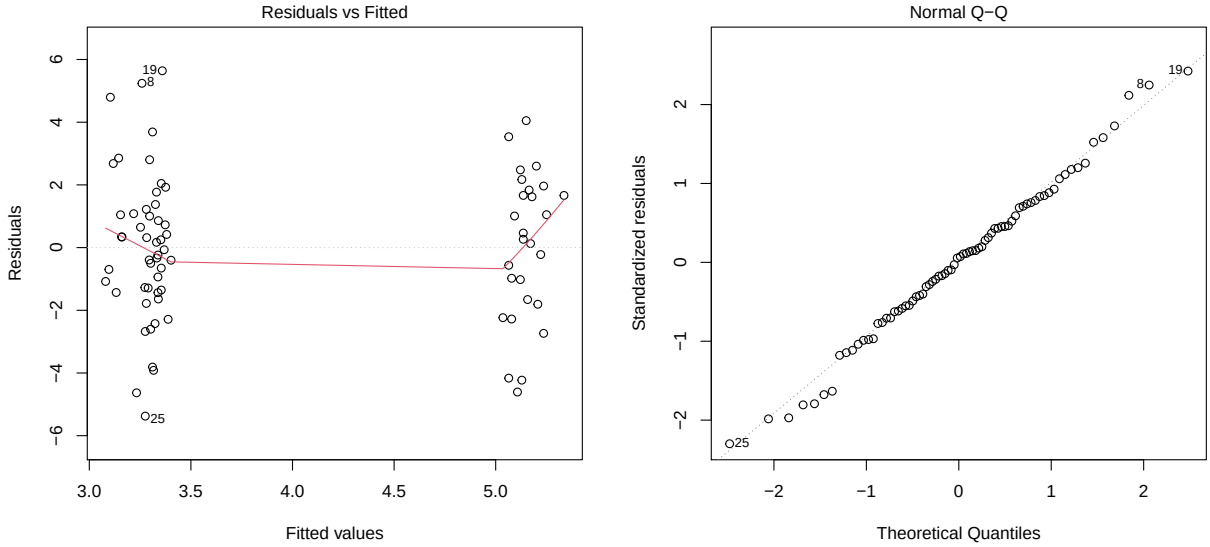
```
## Analysis of Variance Table
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value Pr(>F)
## height     1      6    6.05    1.06  0.306
## diet       2     55   27.47    4.82  0.011 *
## Residuals 72    410    5.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
paste("r-squared:", round(summary(model)$r.squared, 4))
```

```
## [1] "r-squared: 0.1295"
```

From the results above we can conclude that there is no significant effect of height on weight.lost as the p-value for height is $> 0.05$.

```
# diagnostics
par(mfrow=c(1,2));
plot(model, 1); plot(model, 2)
```




From the diagnostics above we do not observe any obvious relationship in the Fitted vs Residuals plot - which is the desired behaviour here. From the normal QQ-plot we see that the residuals follow a straight line very well. Therefore, we can conclude that the model assumptions have been met.

f)

From the analysis in *e*) we concluded that there is no significant effect of height on weight.lost. Furthermore looking at the r-squared values of both models in *c*) and in *e*) we do not observe a significant improvement in r-squared from *c*) to *e*). Based on this and also the Occom's razor principle the model from *c*) is the preferred one.

```
# perform predictions
model <- lm(weight.lost ~ diet, data = data)
diets <- data.frame(diet = unique(data$diet))
predict(model, diets, type = "response")
```

```
##    1    2    3
## 3.30 3.27 5.15
```

The chosen model was used to perform predictions. The predicted weight losses for corresponding diet numbers can be observed above.

g)

```
# create new variable
data <- data %>% mutate(lost.4kg = as.factor(ifelse(weight.lost > 4, 1, 0)))
```

As the new observed variable is now a binary lost.4kg, the most logical model to use here is the logistic regression model.

***Replicating c)***

```
logistic <- glm(lost.4kg~diet, data = data, family = binomial)
drop1(logistic, test="Chisq")
```

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

To replicate the analysis in $c$) we create a logistic model with diet as explanatory variable. From drop1 results we can conclude that diet does have a significant effect (p-value $< 0.05$). Same conclusion as in $c$).

***Replicating d)***

```
logistic <- glm(lost.4kg~diet*gender, data = data, family = binomial)
summary(logistic)$coefficients
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.7918      0.764 -2.3460 0.018978
## diet2           0.8755      0.966  0.9062 0.364832
## diet3           3.6636      1.077  3.4012 0.000671
## gender1         1.3863      1.000  1.3863 0.165657
## diet2:gender1   0.0896      1.320  0.0679 0.945882
## diet3:gender1  -3.2581      1.382 -2.3573 0.018407
```

To replicate the analysis in $d$) we create a logistic model with diet and gender (with interaction) as explanatory variable. From the summary table we can see that diet and diet/gender interaction has a significant effect. Gender does not seem to have a significant effect on the results. This is the same conclusion as in $d$).

***Replicating e)***

```
logistic <- glm(lost.4kg~height * diet, data = data, family = binomial)
summary(logistic)$coefficients
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.39605     7.2607  -0.468    0.640
## height        0.01346     0.0424   0.318    0.751
## diet2         1.73693     9.2584   0.188    0.851
## diet3        13.06237    11.2187   1.164    0.244
## height:diet2 -0.00536     0.0535  -0.100    0.920
## height:diet3 -0.06578     0.0659  -0.998    0.318
```

To replicate the analysis in $e$) we create a logistic model with diet and height (with interaction) as explanatory variable. From the summary table we can see that none of the variables (including interaction factors) are significant. This is different from $e$ where we saw that the estimate for diet was still significant. The reason for this could be that we do not take into consideration people that have lost less than 4 kg while creating the model - these people could be of high significance is the model from $e$). Let's move on to a model without interaction:

```r
logistic <- glm(lost.4kg~ height + diet, data = data, family = binomial)
summary(logistic)$coefficients
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.30078     3.9005 -0.0771  0.93853
## height      -0.00469     0.0228 -0.2059  0.83688
## diet2        0.88014     0.6304  1.3962  0.16266
## diet3        1.95043     0.6352  3.0705  0.00214
```

From the summary table results above we can conclude that diet has a significant influence and height does not. This is the same conclusion from *e*.

***Replicating b)*** We could use the column created here to run a binomial test. However, the test here would actually be to test the hypothesis that the median is larger than 4, not 3. This is different from *b*), however if the test concludes that it is bigger than 4 then automatically we can conclude that it is bigger than 3.

```r
success <- sum(data$weight.lost > 4)
trials <- nrow(data)
binom.test(success, trials, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  success and trials
## number of successes = 36, number of trials = 76, p-value = 0.7
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.375 1.000
## sample estimates:
## probability of success
##                  0.474
```

From the p-value above (p-value $> 0.05$) we can not confidently say that the median of weight.lost is higher than 4. Therefore, we can not replicate the results obtained in *b*) - based on the binomial.test results on lost.4kg we can neither reject nor accept the hypothesis that the median of the weight.lost is higher than 3.