QDA-Lab-5-Part-2

Isabel Sassoon

October 2020

Loading the libraries used in the notebook

```
library(ggplot2)
```

(1) Data

Loading the data into the notebook

```
diet.df<-read.csv("Diet_r.csv")
```

(2) Exploring the data

Numerical summaries of all the attributes in the data

```
summary(diet.df)
```

```
##      Person          gender            Age            Height
##  Min.   : 1.00   Min.   :0.0000   Min.   :16.00   Min.   :141.0
##  1st Qu.:20.25   1st Qu.:0.0000   1st Qu.:32.25   1st Qu.:164.2
##  Median :39.50   Median :0.0000   Median :39.00   Median :169.5
##  Mean   :39.50   Mean   :0.4342   Mean   :39.15   Mean   :170.8
##  3rd Qu.:58.75   3rd Qu.:1.0000   3rd Qu.:46.75   3rd Qu.:174.8
##  Max.   :78.00   Max.   :1.0000   Max.   :60.00   Max.   :201.0
##                  NA's   :2
##    pre.weight          Diet        weight6weeks
##  Min.   : 58.00   Min.   :1.000   Min.   : 53.00
##  1st Qu.: 66.00   1st Qu.:1.000   1st Qu.: 61.85
##  Median : 72.00   Median :2.000   Median : 68.95
##  Mean   : 72.53   Mean   :2.038   Mean   : 68.68
##  3rd Qu.: 78.00   3rd Qu.:3.000   3rd Qu.: 73.83
##  Max.   :103.00   Max.   :3.000   Max.   :103.00
##
```

Notice that the data has not been "interpreted" as intended by the import. For example gender has two values and is binary (in this dataset) yet here it was treated as a number.

Also note that there are 2 missing values for gender.

The as.factor R function can help fix this:

```
diet.df$gender<-as.factor(diet.df$gender)
```

Lets look at the Diet variable:

```
table(diet.df$Diet)
```

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

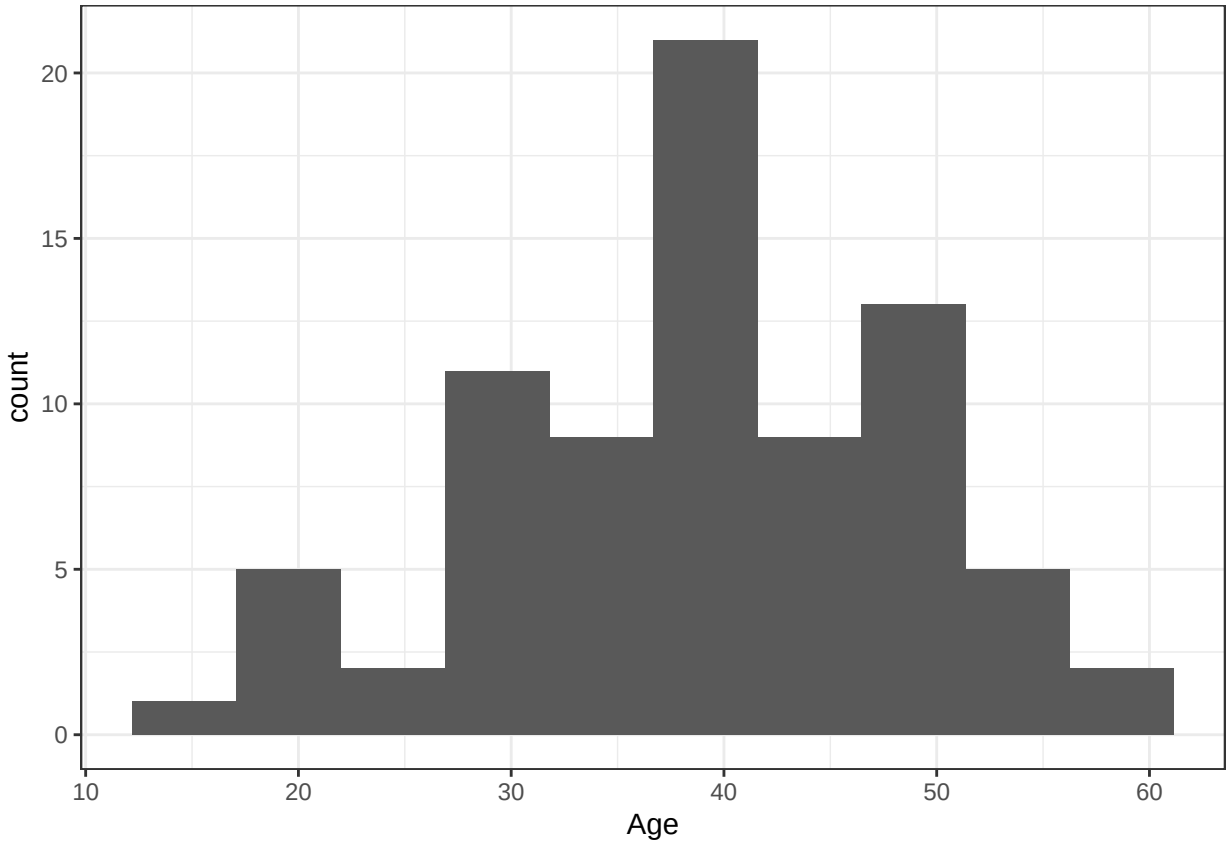
This also is a categorical variable that has been imported as a number so lets fix this too

```
diet.df$Diet<-as.factor(diet.df$Diet)
```

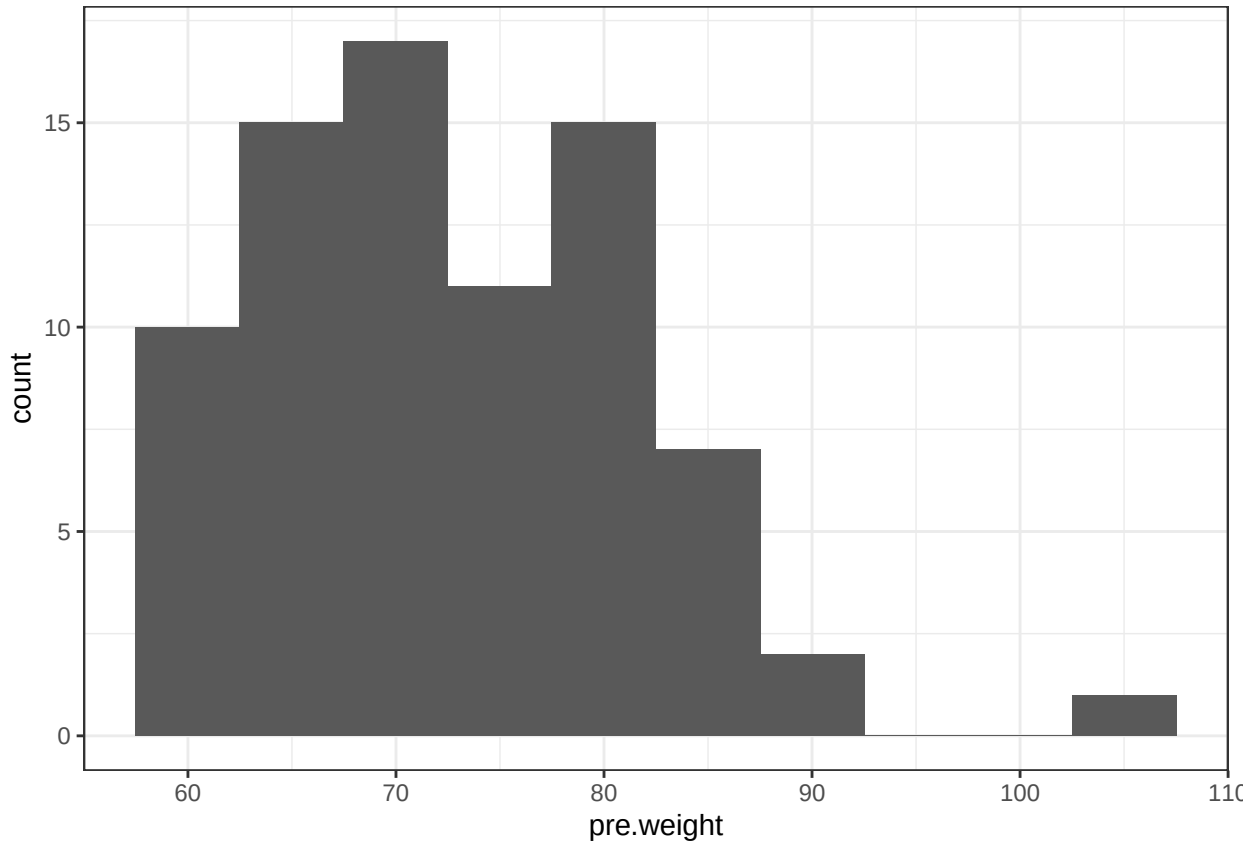
Graphical exploration of the data

Initially here are histograms for each variable in turn. This helps us see if there are any strage skews (lack of symmetry) in the data.

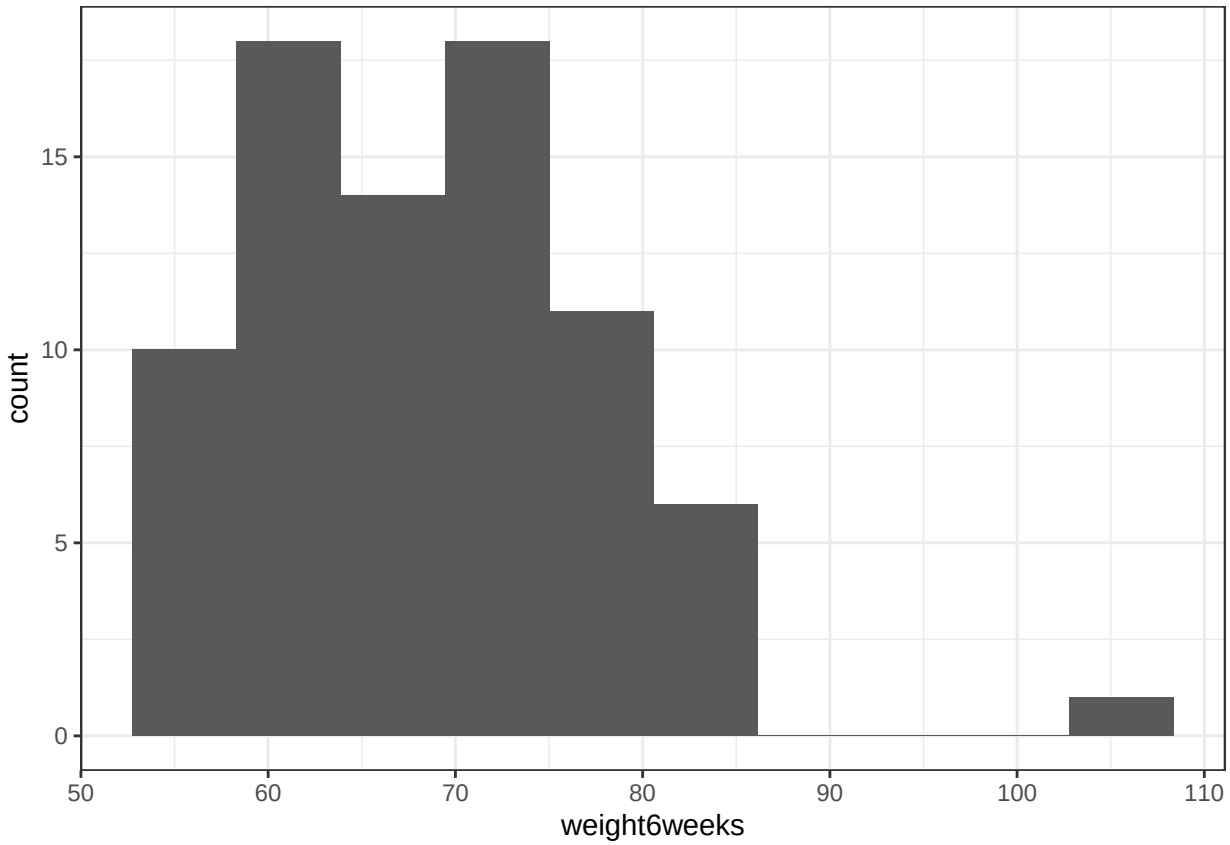
```
ggplot(diet.df, aes(x=Age)) + geom_histogram(bins=10) + theme_bw()
```




```
ggplot(diet.df, aes(x=pre.weight)) + geom_histogram(bins=10) + theme_bw()
```



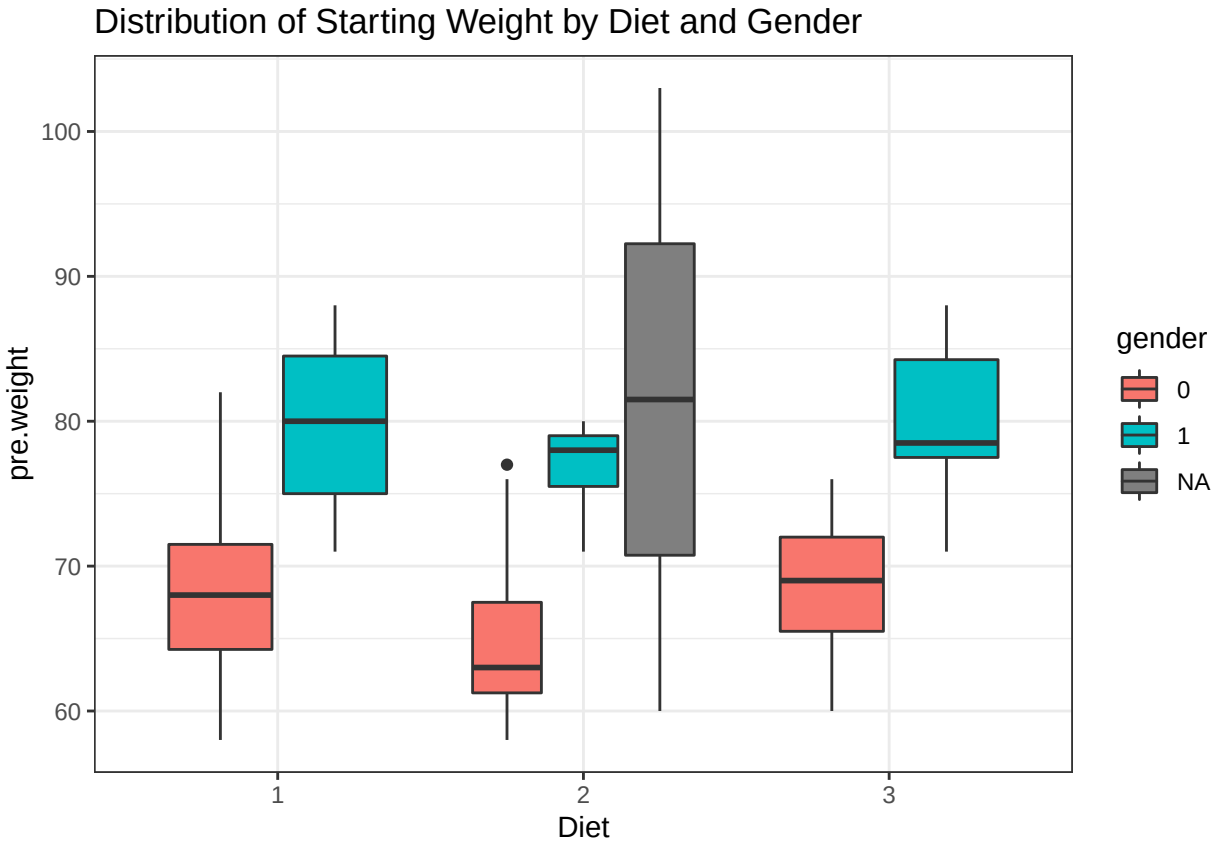

```
ggplot(diet.df, aes(x=weight6weeks)) + geom_histogram(bins=10) + theme_bw()
```



The histograms do not point to any problems in the data. The values for age and weight all seem plausible. There are no extreme values on either end, no negative values or any missing values.

Two way plots

```
ggplot(diet.df, aes(x=Diet, y=pre.weight, fill=gender)) +geom_boxplot() + theme_bw() +
  ggtitle("Distribution of Starting Weight by Diet and Gender")
```



Interestingly there is some missing data, but only for Gender. If we were planning to use Gender in our models we may want to remove the rows with these missing values. We may also scrutinize those rows of data, in case there are any anomalies in the whole row of data.

(3) Hypothsis testing that the mean weight before the diet is the same as after

For this we can use the t-test to test the hypothesis H_0 mean weight before the diet is the same as the mean weight after the diet vs the H_1 the means are different.

```
t.test(diet.df$pre.weight, diet.df$weight6weeks)
```

```
##
##  Welch Two Sample t-test
##
## data:  diet.df$pre.weight and diet.df$weight6weeks
## t = 2.721, df = 153.92, p-value = 0.007259
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.053396 6.636348
## sample estimates:
## mean of x mean of y
##  72.52564  68.68077
```

Looks like there is a significant difference before vs after.

(4) Compute the weight lost

In order to model the effect of diet type on weight lost we need to compute the latter.

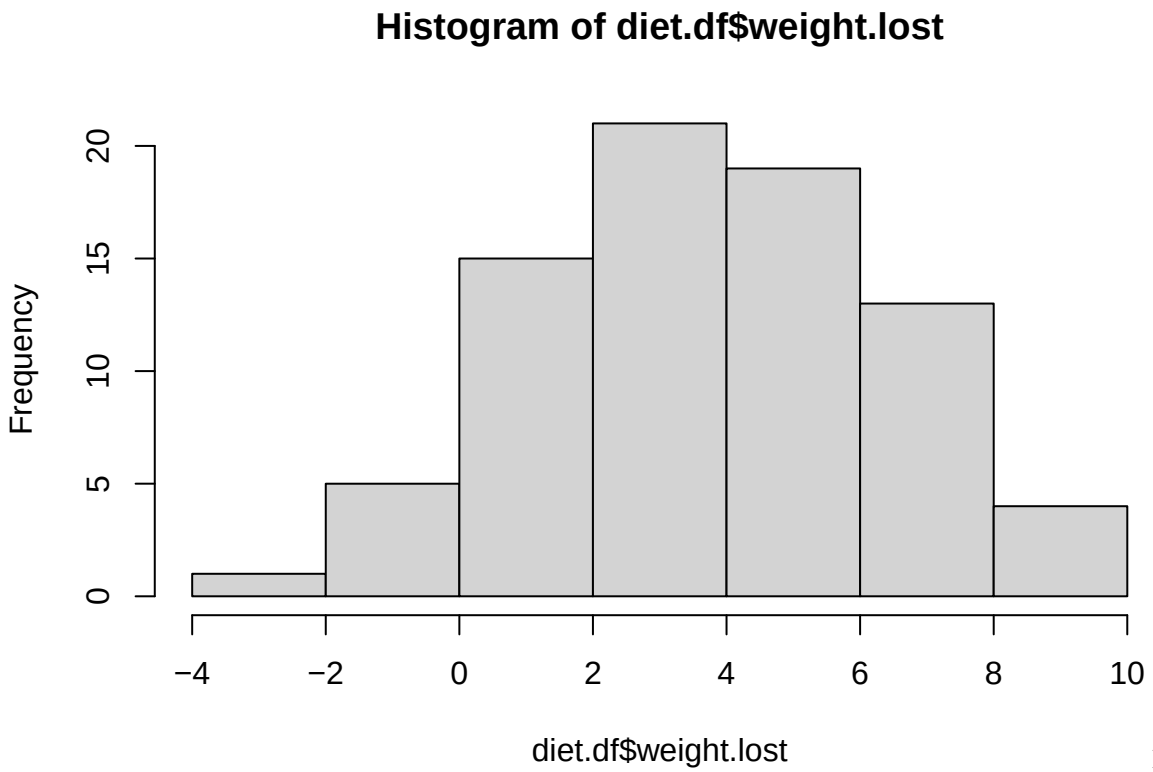
```
diet.df$weight.lost<-diet.df$pre.weight-diet.df$weight6weeks
```

Lets see what this looks like:

```
summary(diet.df$weight.lost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -2.100   2.000   3.600   3.845   5.550   9.200
```

```
hist(diet.df$weight.lost)
```



From the hostogram is looks like most of the people in the sample have lost weight, as the bulk of the data points are above 0.

But can this weight loss be explained by one specific diet?

(5) ANOVA

Lets model the weight lost vs the diet used.

```
summary(aov(diet.df$weight.lost~diet.df$Diet))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## diet.df$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
```

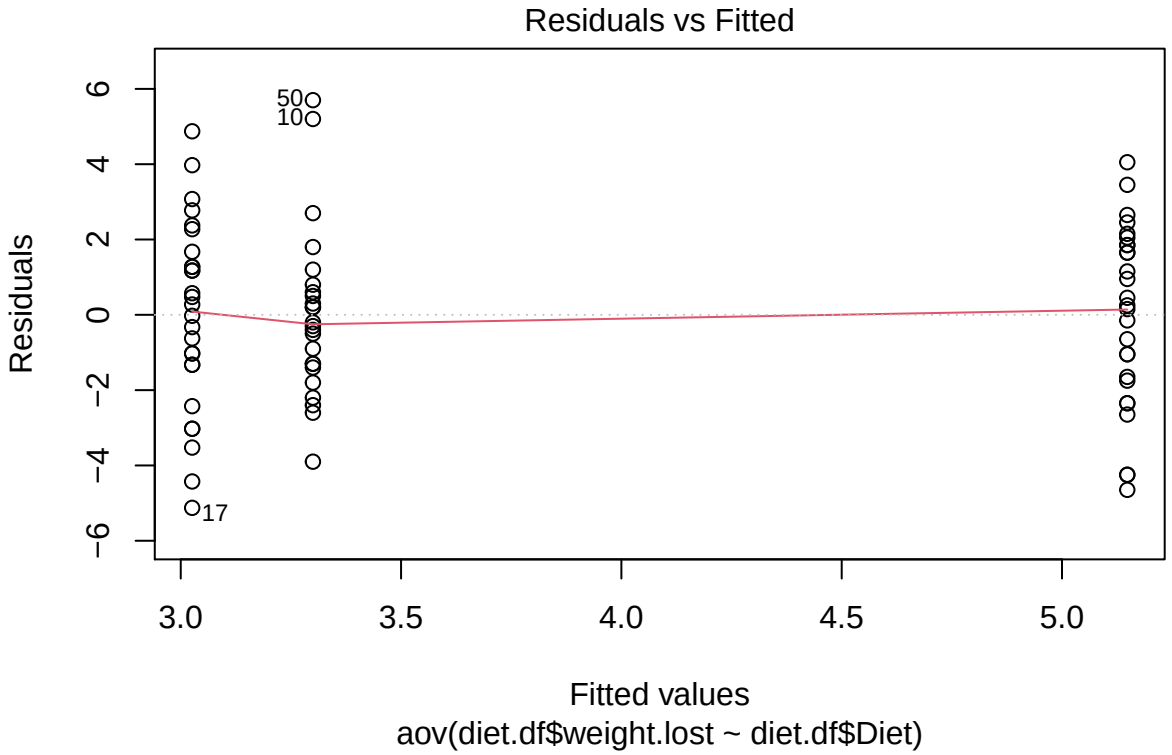
We can see that the F-value is significant, so there is a difference between the mean weight lost for each of the three diets.

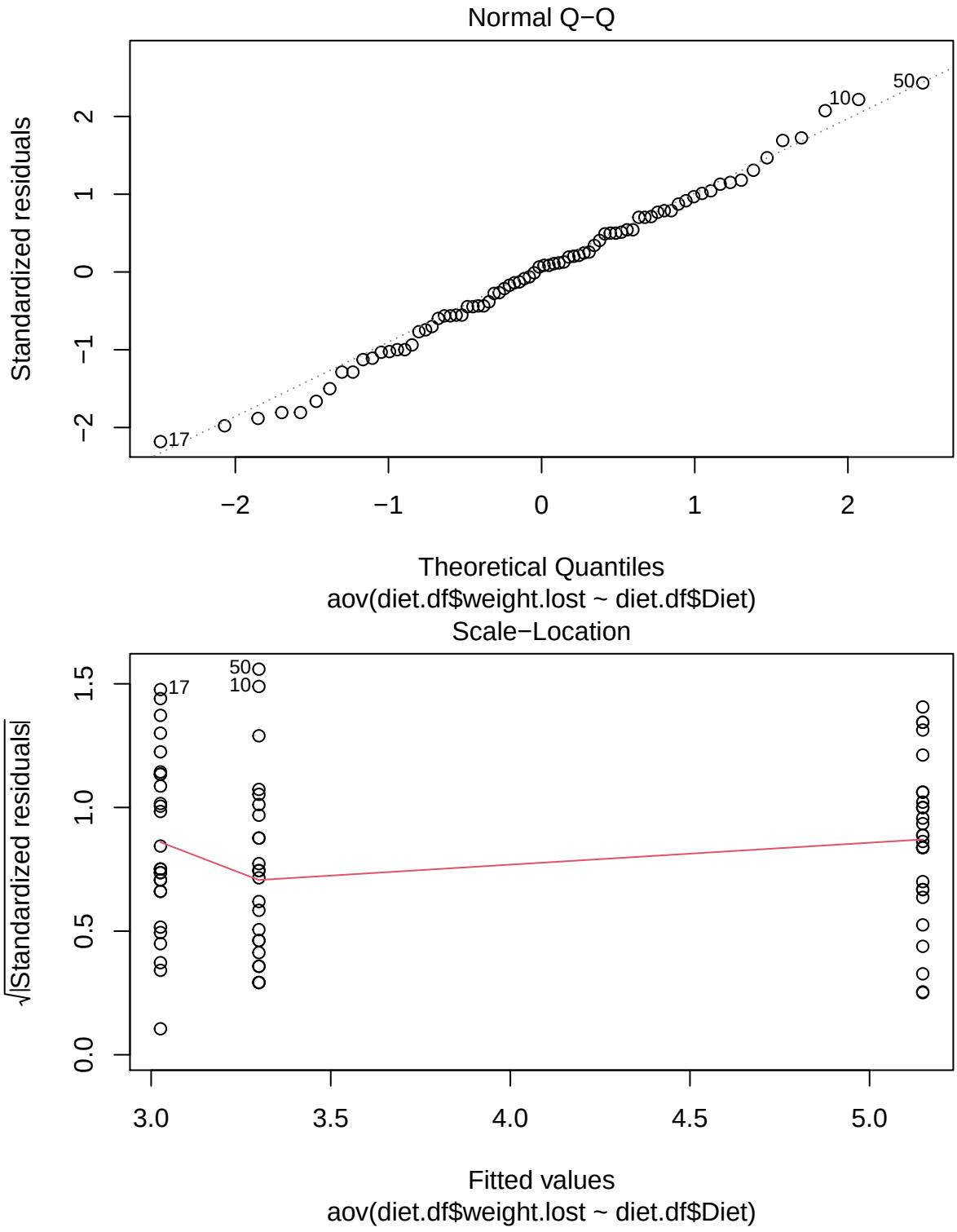
```
summary.lm(aov(diet.df$weight.lost~diet.df$Diet))
```

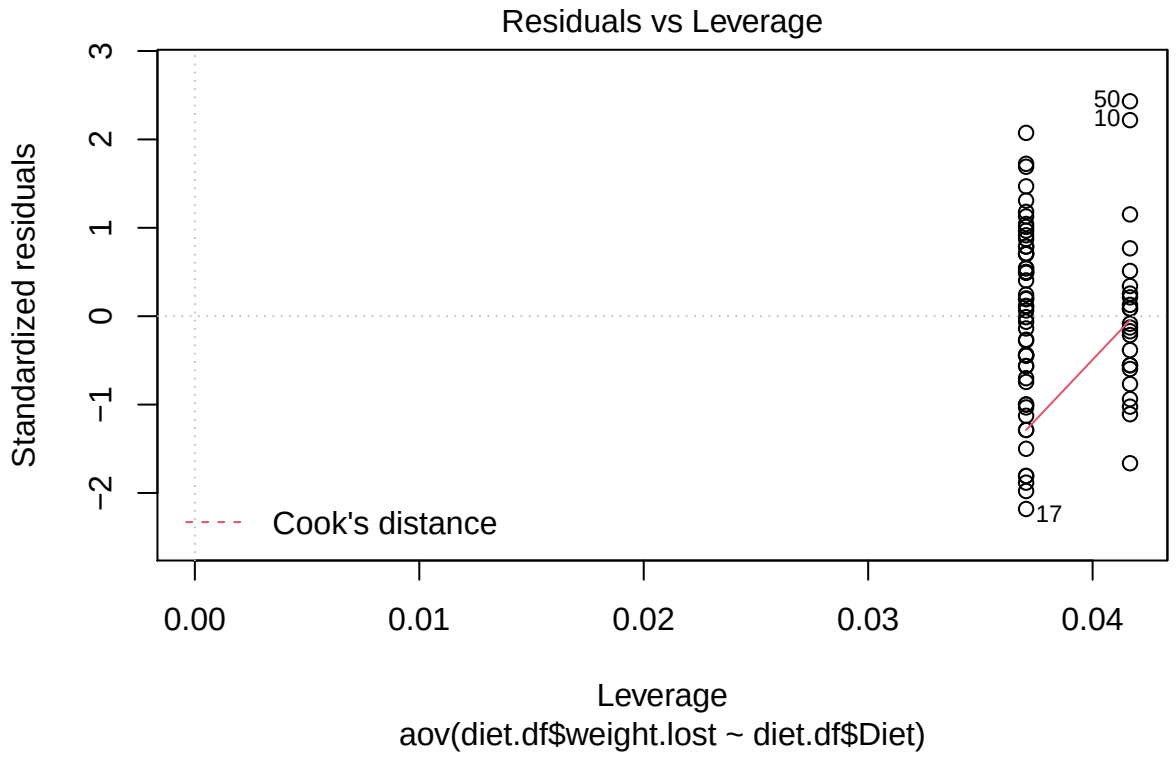
```
##
## Call:
## aov(formula = diet.df$weight.lost ~ diet.df$Diet)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1259 -1.3815  0.1759  1.6519  5.7000
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.3000     0.4889   6.750 2.72e-09 ***
## diet.df$Diet2  -0.2741     0.6719  -0.408  0.68449
## diet.df$Diet3   1.8481     0.6719   2.751  0.00745 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.395 on 75 degrees of freedom
## Multiple R-squared:  0.1418, Adjusted R-squared:  0.1189
## F-statistic: 6.197 on 2 and 75 DF,  p-value: 0.003229
```

From this output we can see that although the F was significant our r^2 is low. The difference in the diets seem to be between Diets 1,2 and diet 3.

```
plot(aov(diet.df$weight.lost~diet.df$Diet))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(diet.df$weight.lost ~ diet.df$Diet)
Scale−Location
√|Standardized residuals|
Fitted values
aov(diet.df$weight.lost ~ diet.df$Diet)



The plots for the residuals do not point to any issues. We see constant variance for the residuals plot (plot1) and normally distributed residuals in plot 2.

Going back to the means for the different diets, lets compute them:

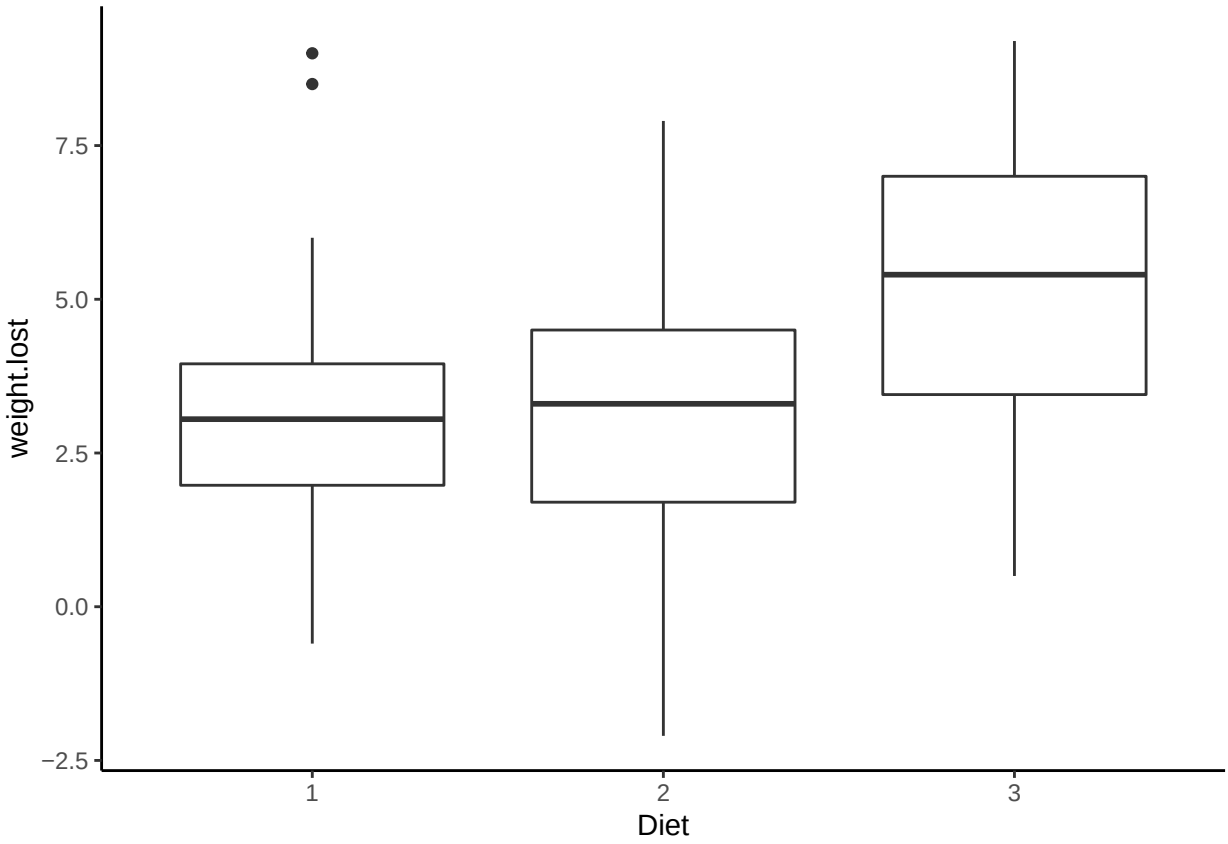
```
aggregate(diet.df$weight.lost~diet.df$Diet, FUN="mean")
```

```
##   diet.df$Diet diet.df$weight.lost
## 1            1            3.300000
## 2            2            3.025926
## 3            3            5.148148
```

Seems like diet 3 causes higher weight loss, but the difference between diet 1 and 2 is much smaller - as we saw in the anova model summary.

We can look at this also with a boxplot.

```
ggplot(diet.df, aes(x=Diet, y=weight.lost)) + geom_boxplot() + theme_classic()
```



What to look for in an ANOVA analysis output?

- In the summary(aov()) check if the F statistic is significant. If it is significant then the means are different for the categorical variable values.
- In the summary.lm(aov()) check the direction and the significance of the estimates for the different levels of the categorical variables. Are they significant? Are they positive or negative? Also pay attention to the default category - which is represented by the intercept.
- In plot(aov()) check the first two plots. In the first plot "residuals vs fitted" we want constant variance - this means a straightish red line and the spread of points that is similar for all the x values. In the second plot "QQplot" we want to see the points (represented by round circles) to be as close to the straignt line as possible. Check if there are some suspicious outliers.

---

other attributes may affect the weight loss

Lets introduce Age, and in this case we want to do this we are going to apply an ANCOVA model because our target variable is numeric, and we have both numeric and categorical explanatory variables. An ANCOVA model uses the same lm syntax as the linear regression.

```
model2<-lm(diet.df$weight.lost~diet.df$Diet+diet.df$Age)
summary(model2)
```

```
##
```

```
## Call:
## lm(formula = diet.df$weight.lost ~ diet.df$Diet + diet.df$Age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.096 -1.443  0.118  1.615  5.713
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.548832   1.254023   2.830  0.00599 **
## diet.df$Diet2 -0.285488   0.678259  -0.421  0.67504
## diet.df$Diet3  1.829293   0.681817   2.683  0.00900 **
## diet.df$Age   -0.006088   0.028220  -0.216  0.82980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 74 degrees of freedom
## Multiple R-squared:  0.1424, Adjusted R-squared:  0.1076
## F-statistic: 4.095 on 3 and 74 DF,  p-value: 0.009569
```

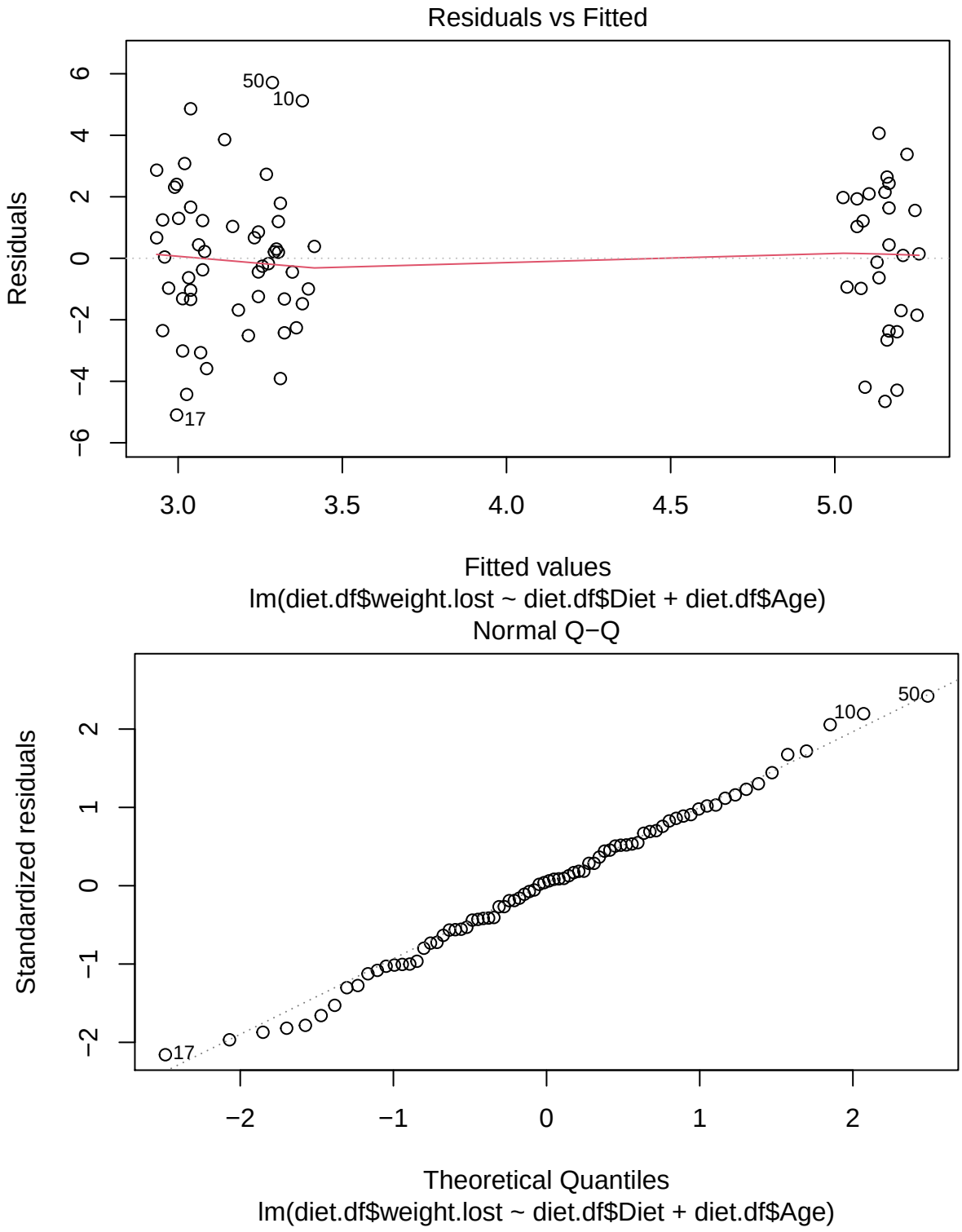
This model's performance is not good at all, the F statistic is significant but the r^2 is still very low. Adding Age does not help us better model weight loss.

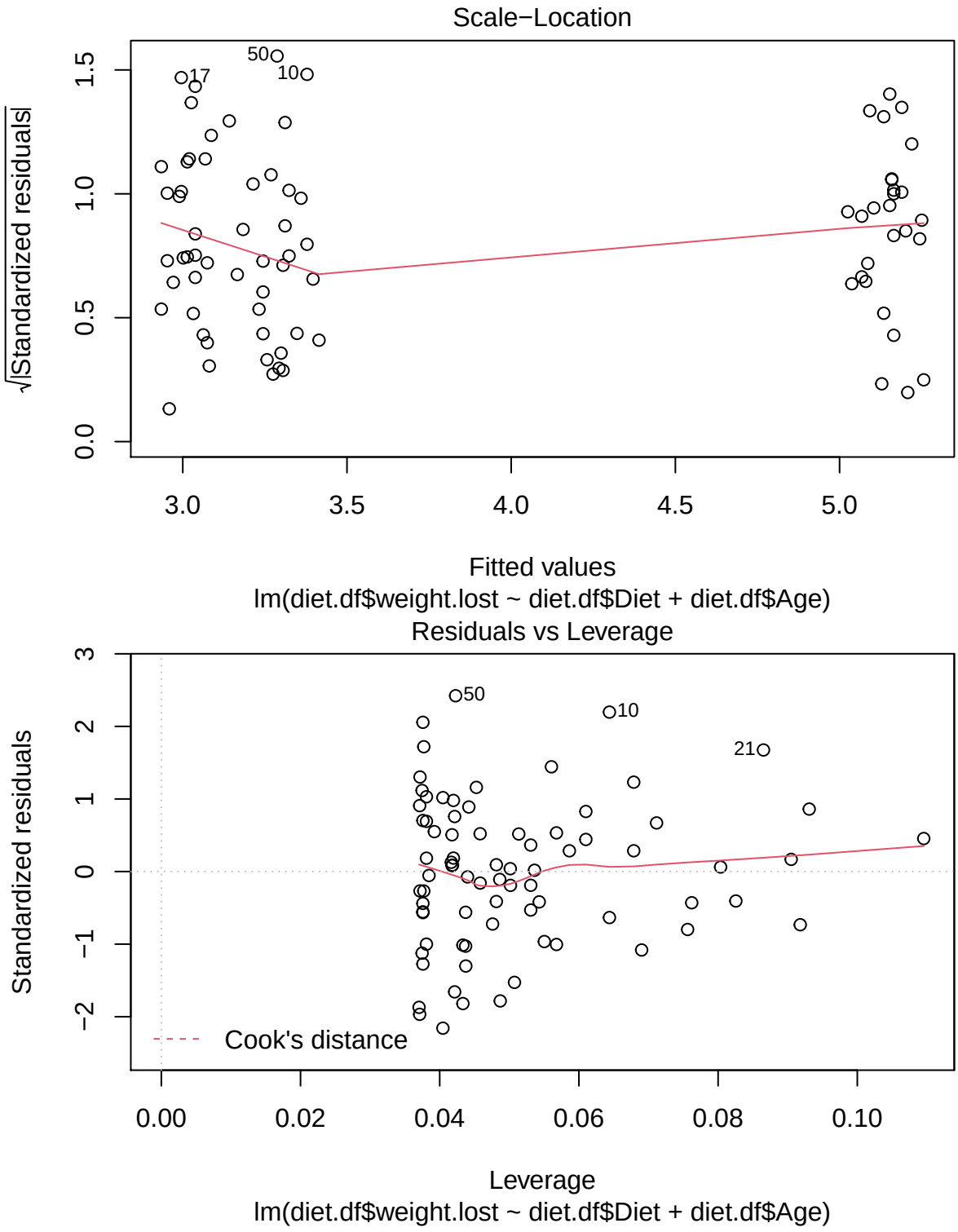
Explain the relation between the variables

This model does not have a good r^2 but *how would we explain the realtion ebtween the dependent and the independent variables?*

From the table summary(model2) we can learn that there is a weak negative relationship between age and weightloss. This is reflected in the value of the estimate for the effect of age which is -0.006. We can also see that diet 2 has a negative effect on weight loss (-0.285) and diet 3 has a more positive effect on weightloss.

```
plot(model2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(diet.df$weight.lost ~ diet.df$Diet + diet.df$Age)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(diet.df$weight.lost ~ diet.df$Diet + diet.df$Age)



The plots for the residuals vs fitted and the QQ plot do not raise any concerns.

Perhaps the weight before the diet is a better predictor?

Lets see, as the pre.weight is a numerical attribute. We are again using an ANCOVA model.

```
model3<-lm(diet.df$weight.lost~diet.df$Diet+ diet.df$gender + diet.df$pre.weight)
summary(model3)
```

```
##
## Call:
## lm(formula = diet.df$weight.lost ~ diet.df$Diet + diet.df$gender +
##     diet.df$pre.weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9842 -1.3278 -0.1511  1.5109  5.6365
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.04217    3.70905  -0.011  0.99096
## diet.df$Diet2       0.10553    0.70059   0.151  0.88069
## diet.df$Diet3       1.82492    0.67110   2.719  0.00822 **
## diet.df$gender1    -0.48514    0.84959  -0.571  0.56978
## diet.df$pre.weight  0.04864    0.05404   0.900  0.37117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.39 on 71 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.1387, Adjusted R-squared:  0.09019
## F-statistic: 2.859 on 4 and 71 DF,  p-value: 0.02954
```

This model is no better than our previous one. Low r^2 and not many estimates are significant. The only one seems to be the different between diet 3 and the rest.

What to look out for in an ANCOVA output?

An ANCOVA model is similar to a linear regression model

- In the summary(lm()) look at the r^2 which should be as close as 1 as possible, look at the F statistic which should be significant and finally look at the estimates for the coefficients and see which ones are significant.
- In the plot(lm()) look for equal variances in the residuals vs fitted plot (first plot) and for the residuals to alogn along the straight line in the qqplot (the second plot). ***

Joining categories

Looking at weight loss we saw that the difference seemed to be more marked between Diet 3 and the other two. So lets join diet 1 and 2 together then model this instead.

```
diet.df$diet.ind<-ifelse(diet.df$Diet=="3", "3", "1 or 2")
table(diet.df$diet.ind)
```

```
##
## 1 or 2      3
##     51     27
```

```
model3<-aov(diet.df$weight.lost~diet.df$diet.ind)
summary.lm(model3)
```

```
##
## Call:
## aov(formula = diet.df$weight.lost ~ diet.df$diet.ind)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2549 -1.4549  0.1485  1.6252  5.8451
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.1549     0.3335   9.460 1.77e-14 ***
## diet.df$diet.ind3   1.9932     0.5669   3.516 0.000742 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.382 on 76 degrees of freedom
## Multiple R-squared:  0.1399, Adjusted R-squared:  0.1286
## F-statistic: 12.36 on 1 and 76 DF,  p-value: 0.0007418
```

There are many other ways of proceeding with the analysis. One possible option is to check whether there are significant differences in the starting weights of the people allocated to the three diets, or their ages.

If you chose to use the Gender as an explanatory variable, then you may want to exclude the rows of data that have NA. It is unlikely that we will be able to "guess" or impute the missing values for gender in a meaningful way.

Removing Missing values

We can try to see if gender is a better explainer of weight loss, but before we do we should remove the missing values for gender, this means we will have a smaller data set to use.

```
diet2.df<-subset(diet.df, !is.na(diet.df$gender))
```

We now removed the two rows with missing values for gender.

```
model.gender.aov<-aov(diet2.df$weight.lost~diet2.df$gender)
summary.lm(model.gender.aov)
```

```
##
## Call:
## aov(formula = diet2.df$weight.lost ~ diet2.df$gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9930 -1.6846 -0.2041  1.7264  5.1848
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.8930     0.3846  10.123  1.3e-15 ***
## diet2.df$gender1    0.1221     0.5836   0.209    0.835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.522 on 74 degrees of freedom
## Multiple R-squared:  0.0005914,  Adjusted R-squared:  -0.01291
## F-statistic: 0.04379 on 1 and 74 DF,  p-value: 0.8348
```

Clearly this has not helped at all.

Overall the best model remains the one that uses either only the diet type only.