Assignment 1

2022-10-05

Question 1

a)

We have $\hat{\alpha}_\text{i} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$ so

$$\text{E}[\hat{\alpha}_\text{i}] = \frac{1}{n}\sum_{j=1}^{n}\text{E}[Y_{ij}] = \frac{1}{n}\sum_{j=1}^{n}\alpha_i = \alpha_i.$$

Thus $\hat{\alpha}_\text{i}$ is an ubiased estimator.

b)

We have

$$\begin{aligned}
\text{Var}(\hat{\alpha}_\text{i}) &= \text{Var}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{ij}\right) \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\text{Var}(Y_{ij}) && \text{since } Y_{ij} \text{ are i.i.d} \\
&= \frac{1}{n^2}\sum_{j=1}^{n}[\text{Var}(\alpha_i) + \sigma^2] \\
&= \frac{\text{Var}(\alpha_i) + \sigma^2}{n}.
\end{aligned}$$

If $i = j$, we have

$$\text{Var}(\hat{\alpha}_\text{i} - 2\,\hat{\alpha}_\text{j}) = \text{Var}(-\,\hat{\alpha}_\text{i}) = \text{Var}(\hat{\alpha}_\text{i}) = \frac{\text{Var}(\alpha_i + \sigma^2)}{n}.$$

Otherwise if $i \neq j$, then

$$\begin{aligned}
\text{Var}(\hat{\alpha}_\text{i} - 2\,\hat{\alpha}_\text{j}) &= \text{Var}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{ij} - \frac{2}{n}\sum_{i=1}^{n} Y_{ij}\right) \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\text{Var}(Y_{ij}) + \frac{4}{n^2}\sum_{j=1}^{n}\text{Var}(Y_{ij}) \\
&= \frac{5}{n^2}\sum_{j=1}^{n}\text{Var}(Y_{ij}) \\
&= \frac{5(\text{Var}(\hat{\alpha}_\text{i}) + \sigma^2)}{n}.
\end{aligned}$$

c)

We have

$$\begin{aligned}
\mathrm{E}[S_\Omega - 3S_\omega] &= \mathrm{E}[S_\Omega - S_\omega - 2S_\omega] \\
&= \mathrm{E}[S_\Omega - S_\omega] - 2\,\mathrm{E}[S_\omega] \\
&= I - 1 - 2(n-1).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}[3S_\Omega - 5S_\omega] &= \mathrm{E}[3(S_\Omega - S_\omega) - 2S_\omega] \\
&= 3\,\mathrm{E}[S_\Omega - S_\omega] - 2\,\mathrm{E}[S_\omega] \\
&= 3(I-1) - 2(n-1).
\end{aligned}$$

Question 2

a)

Using the $\mu^{st} = 0$ parametrization, we have

$$\hat{\alpha}_{\mathrm{i}}{}^{st} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{i\cdot} = \hat{\alpha}_{\mathrm{i}}{}^{tr} + \hat{\mu}^{tr}$$

and

$$\hat{\mu}^{st} = 0.$$

b)

We have

$$\hat{\mu}^{sum} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{I}\sum_{i=1}^{I}\hat{\alpha}_{\mathrm{i}}{}^{tr}$$

and

$$\hat{\alpha}_{\mathrm{i}}{}^{sum} = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} = \hat{\alpha}_{\mathrm{i}}{}^{tr} - \hat{\mu}^{sum}.$$

c)

We have

$$\hat{\mu}^{sum} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{I}\sum_{i=1}^{I}\hat{\alpha}_{\mathrm{i}}{}^{tr}$$

and

$$\hat{\alpha}_{\mathrm{i}}{}^{sum} = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} = \hat{\alpha}_{\mathrm{i}}{}^{tr} - \hat{\mu}^{sum}.$$

Question 3

a)

Using the formula $\hat{\gamma_{ij}} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$, we can estimate $\hat{\gamma}$. Using the properties of the expectation operator and the assumption that $\mathrm{E}[e_{ij}] = 0$ for any i and j we can obtain this expectation as, yet this gives us $\mathrm{E}[Y_{ijk} = \eta_{ij}]$:

$$
\begin{aligned}
\mathrm{E}[\hat{\gamma}_{ij}] &= \mathrm{E}[Y_{ij.} + Y_{i..} - Y_{.j.} + Y_{...}] \\
&= \mathrm{E}[Y_{ij.}] + \mathrm{E}[Y_{i..}] - \mathrm{E}[Y_{.j.}] + \mathrm{E}[Y_{...}] \\
&= \eta_{ij} - \eta_{i.} - \eta_{.j} + \eta_{..} \\
&= \gamma_{ij} \ .
\end{aligned}
$$

b)

The main reason why we would use a non-parametric test such as the F-test is that we do not know the parameter σ^2, yet in that test it cancels out in the derivation of the F-statistic.

A more suitable test, when we know the variance, would be the χ^2-test. The intuition behind this test achieving a better performance is that it incorporates more information, i.e. we know exactly the value of σ^2, so it will reduce the uncertainty.

Question 4

a) Using the parametrization $\mu = 0$:

```
data("iris");

Y <- iris[order(iris$Species), "Sepal.Width"];

X <- diag(3) %x% rep(1, 50);

n = 150;
I = 3;
```

Then we calculate the estimated $\hat{\beta} = (X^T X)^{-1} X^T Y$ as

```
beta = solve((t(X) %*% X)) %*% t(X) %*% Y;
```

The residual sum of squares S_Ω and S_ω of the full and reduced models respectively are

```
s1 = norm(Y - X %*% beta, type="2")^2;
s2 = norm(Y - matrix(rep(1, n),ncol=1) * mean(Y), type="2")^2;
```

The unbiased estimator of σ^2 are $\frac{S_\Omega}{n-1} = 16.962$ and $\frac{S_\omega}{n-1} = 28.3069333$.

```
unb_est = s1/(n - I);
bet_ss = s2 - s1;
bet_means = (s2 - s1)/(I);
f_val = ((s2 - s1)/(I - 1))/(s2/(n - I));

within_means = s1/(n - I);
```

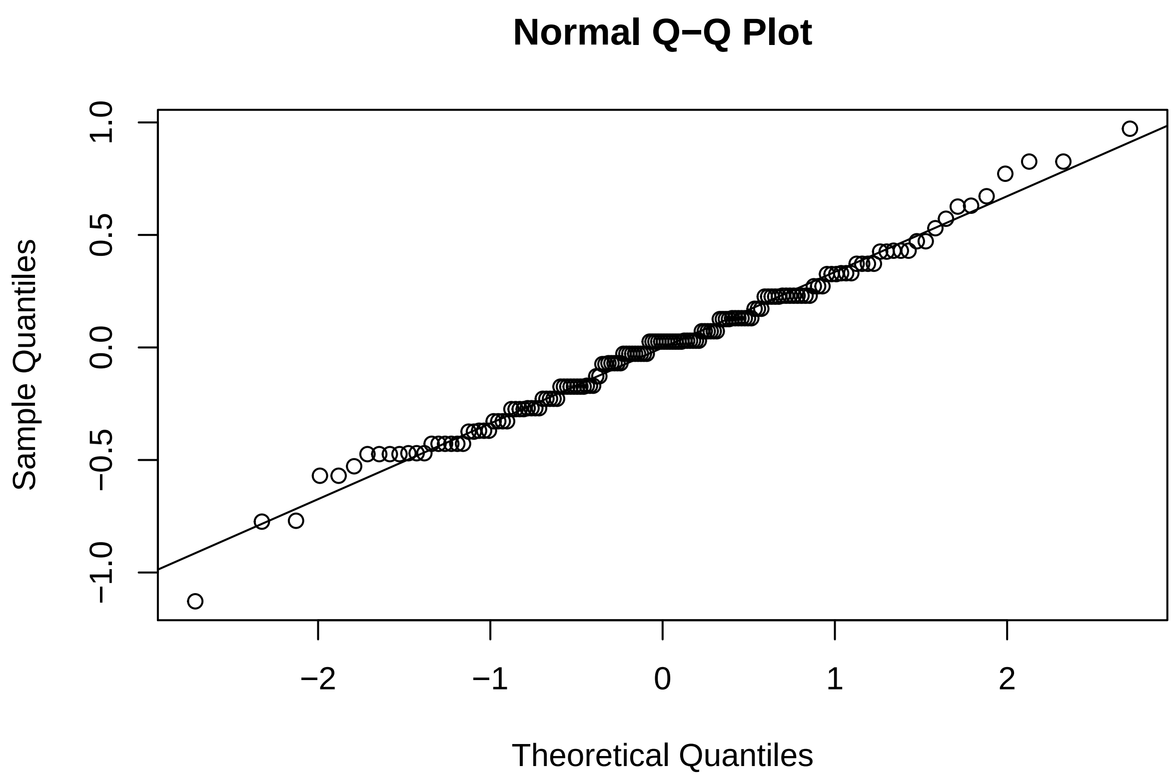
The quantities needed to complete an ANOVA table are :

- Between groups sum of square: $S_\omega - S_\Omega = 11.3449333$.
- Between groups mean square: $\frac{(S_\omega - S_\Omega)/(I-1)}{S_\Omega/(n-I)} = 3.7816444$.
- Within groups sum of square: $S_\Omega = 16.962$.
- F value $= 29.4575393$.

b) We first check for the model assumptions:

The normality of residuals with expectation zero are checked using QQ-plot, Shapiro-Wilk test, and one-sample t-test because the true standard deviation is not known.

```
plot.new()
res <- Y - c(rep(beta, 1, each=50));
qqnorm(res)
qqline(res)
```




```
shapiro.test(res)
```

```
## [...]
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.98948, p-value = 0.323
```

```
t.test(res, mu=0, alternative = "greater")
```

```
## [...]
##  One Sample t-test
##
## data:  res
## t = -1.2788e-14, df = 149, p-value = 0.5
## alternative hypothesis: true mean is greater than 0
## [...]
```

Since the p-values for both test are larger than 0.05, with the mean of the residuals being extremely close, we can say that the normality and zero mean assumptions hold. Next we check that $Var(e_{ij}) = \sigma^2$ using Bartlett test.

```
species <- factor(c(rep(1, 50), rep(2, 50), rep(3, 50)),
                  labels=c("setosa", "versicolor", "virginica"));
data_iris <- data.frame(Y, species);

bartlett.test(Y ~ species, data=data_iris)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Y by species
## Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

Thus, the model assumptions hold. Now we test for the mean of iris sepal width of the three species using the produced F-statistic above.

```
pv <- pf(f_val, I - 1, n - I, lower.tail = FALSE);
```

The p-value is $1.7447978 \times 10^{-11} < 0.05$ so we can reject the null hypothesis that the means are statistically the same.

c)

```
model <- aov(Y ~ species, data=data_iris);

summary(model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## species       2  11.35   5.672   49.16 <2e-16 ***
## Residuals   147  16.96   0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It can be seen that the results from ANOVA agree with the final conclusion although some quantities are a bit off.

d)

```
kruskal.test(Y ~ species, data=data_iris);
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Y by species
## Kruskal-Wallis chi-squared = 63.571, df = 2, p-value = 1.569e-14
```

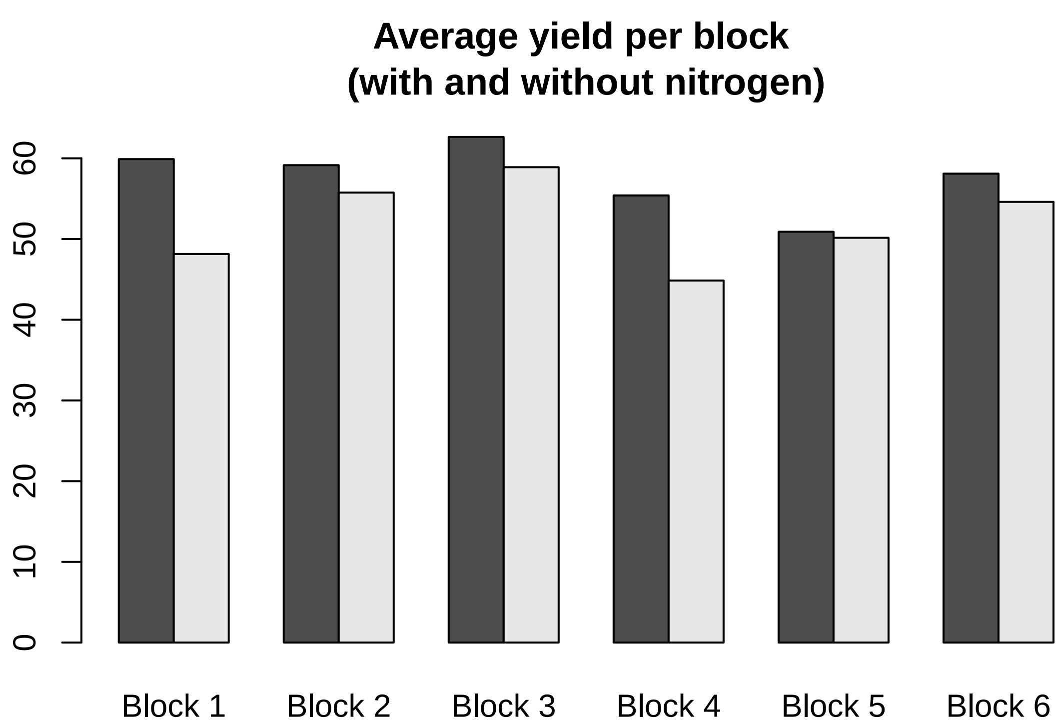
Thus, the Kruskal-Wallis test agrees with our findings since its p-value is smaller than 0.05, and because we the normal distribution assumption holds, the location parameters are the means.

Question 5

a)

Plot of average yield per block, distinguishing between using or not using nitrogen

```r
utils::data(npk, package="MASS")
utils::data(npk, package="MASS")
npk_nitro = npk[npk$N == 1,]
npk_nonnitro = npk[npk$N == 0,]
bck = matrix(0,6,2)

for (i in 1:6)
  {
  bck[i,1] = mean(npk_nitro[npk_nitro$block == i,]$yield)
  bck[i,2] = mean(npk_nonnitro[npk_nonnitro$block == i,]$yield)
}

df = data.frame(bck[1,], bck[2,], bck[3,], bck[4,], bck[5,], bck[6,])
colnames(df) = c("Block 1", "Block 2", "Block 3", "Block 4", "Block 5", "Block 6")

barplot((as.matrix(df)),
        beside=TRUE,
        main = "Average yield per block \n(with and without nitrogen)")
```



b) Two way ANOVA full test

```r
mod.full=lm(yield ~ block*N, data = npk)
anova(mod.full)
```

```
## Analysis of Variance Table
##
## Response: yield
```

```
##           Df Sum Sq Mean Sq F value  Pr(>F)
## block      5 343.29  68.659  3.3592 0.03967 *
## N          1 189.28 189.282  9.2607 0.01021 *
## block:N    5  98.52  19.704  0.9640 0.47690
## Residuals 12 245.27  20.439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## H0: no interaction versus H1: there is an interaction
## H0: smaller model holds versus H1: smaller model is not true
```

The p-values for both "block" and "N" are small enough for being significant for us. However, the value for the interaction is clearly above the significant level, so we cannot reject that the interaction between "block" and "N" does not exist (i.e. we do not have enough evidence of the existence of interaction).

```
mod.full=lm(yield ~ block, data = npk)   # full model
anova(mod.full)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value  Pr(>F)
## block      5 343.29  68.659  2.3184 0.08607 .
## Residuals 18 533.07  29.615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod.full=lm(yield ~ block, data = npk)   # full model
anova(mod.full)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value  Pr(>F)
## block      5 343.29  68.659  2.3184 0.08607 .
## Residuals 18 533.07  29.615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that if we conduct the one way ANOVA test with the two variables separately, the "block" variable does not seem to be significant enough for our analysis. Moreover, the analogous test for the "N" variable outputs a p-value that indicates that the use of nitrogen is really significant.

Question 6

```
diet <- read.table("diet.txt", header = TRUE);
diet["weight.loss"] <- diet$preweight - diet$weight6weeks;
```

a) A short summary of the data is given:

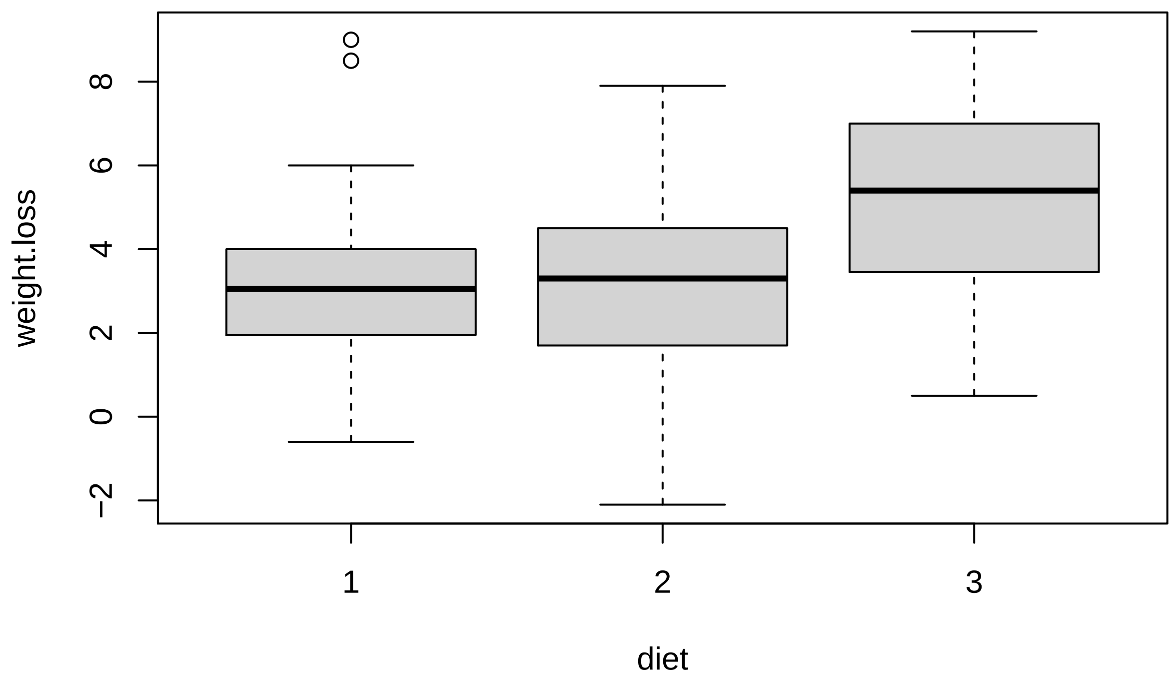
```
summary(diet);
```

```
##      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
##    preweight          diet         weight6weeks     weight.loss
##  Min.   : 58.00   Min.   :1.000   Min.   : 53.00   Min.   :-2.100
##  1st Qu.: 66.00   1st Qu.:1.000   1st Qu.: 61.85   1st Qu.: 2.000
##  Median : 72.00   Median :2.000   Median : 68.95   Median : 3.600
##  Mean   : 72.53   Mean   :2.038   Mean   : 68.68   Mean   : 3.845
##  3rd Qu.: 78.00   3rd Qu.:3.000   3rd Qu.: 73.83   3rd Qu.: 5.550
##  Max.   :103.00   Max.   :3.000   Max.   :103.00   Max.   : 9.200
##
```

To further see the effects of the diets on weight loss, we use boxplots.

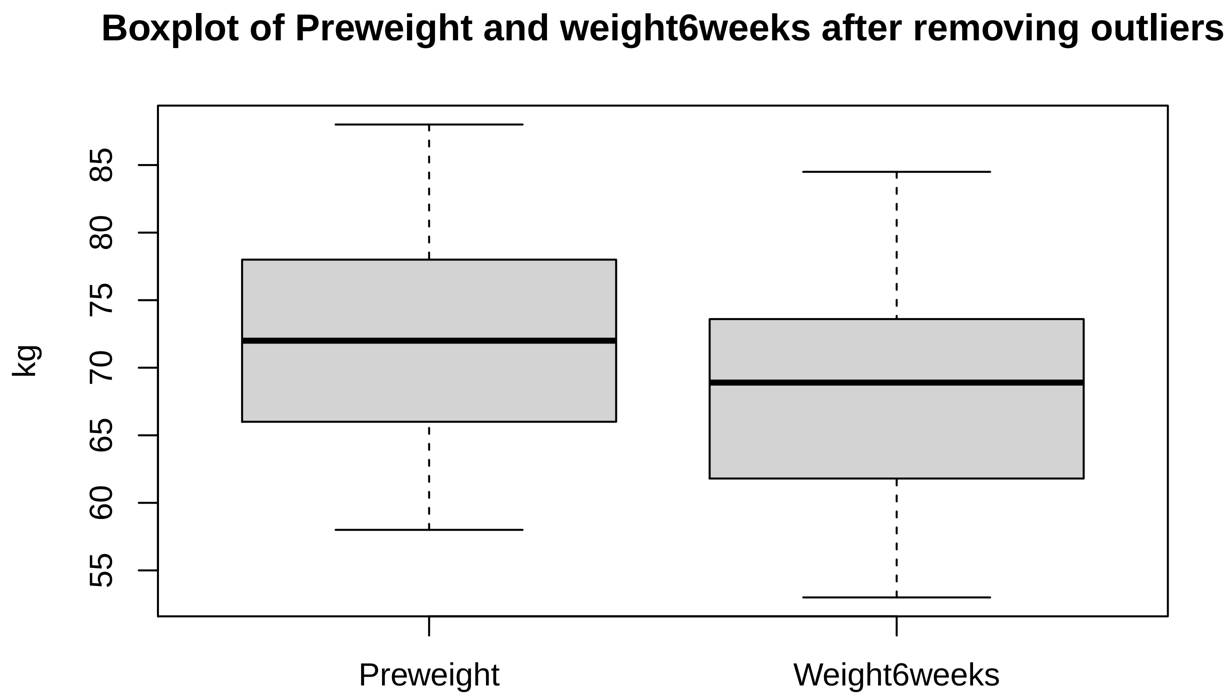
```
boxplot(weight.loss ~ diet, data=diet)
```



It can be seen that there are a few outliers within the samples and it may affect our tests later thus we shall remove them.

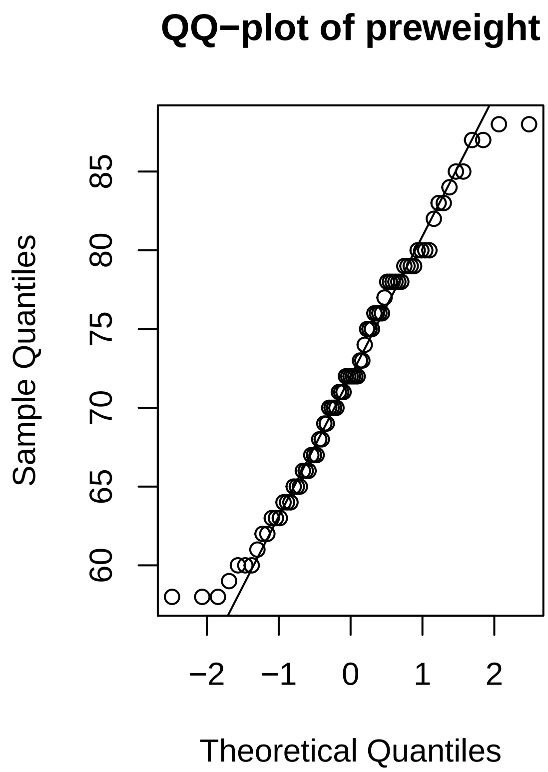
```
Q1 <- quantile(diet$preweight, probs=c(.25, .75), na.rm = FALSE)
Q2 <- quantile(diet$weight6weeks, probs=c(.25, .75), na.rm = FALSE)

iqr_pre <- IQR(diet$preweight);
iqr_aft <- IQR(diet$weight6weeks);

diet_elim <- subset(diet, (preweight > (Q1[1] - 1.5*iqr_pre) &
                          preweight < (Q1[2]+1.5*iqr_pre)) |
                          (weight6weeks < (Q2[1] - 1.5*iqr_aft) &
```

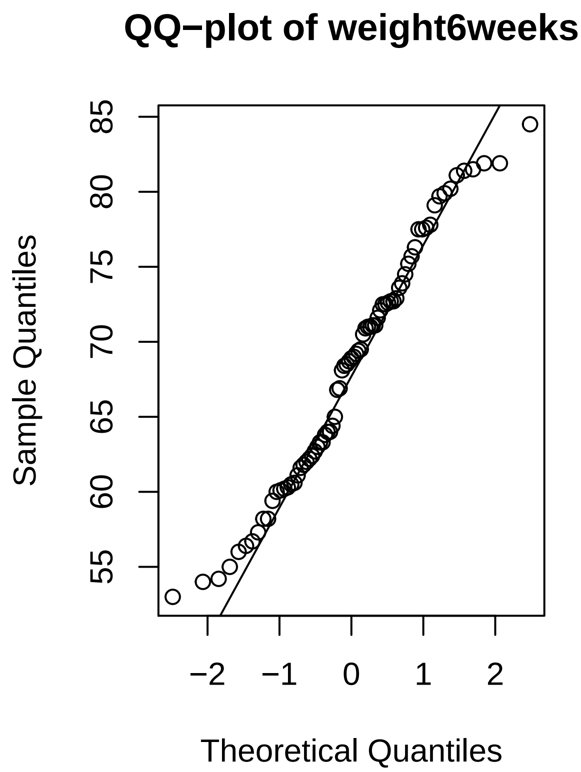
```
                                        weight6weeks > (Q2[1] + 1.5*iqr_aft)));



boxplot(diet_elim$preweight, diet_elim$weight6weeks,
        names=c("Preweight", "Weight6weeks"),
        ylab="kg",
        main="Boxplot of Preweight and weight6weeks after removing outliers")
```



To check whether the diets affect the weight loss, we can test for statistical difference between *preweight* and *weight6weeks*, if the diet does not affect then the mean is approximately the same and vice versa. We first use QQ-plot and Shapiro-Wilk test to check the normality of the samples.

```
par(mfrow=(c(1,2)))

qqnorm(diet_elim$preweight, main="QQ-plot of preweight")
qqline(diet_elim$preweight)

qqnorm(diet_elim$weight6weeks, main="QQ-plot of weight6weeks")
qqline(diet_elim$weight6weeks)
```






```
shapiro.test(diet_elim$preweight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet_elim$preweight
## W = 0.97376, p-value = 0.1117
```

```
shapiro.test(diet_elim$weight6weeks)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet_elim$weight6weeks
## W = 0.97222, p-value = 0.08964
```

The two p-values are higher than 0.05 thus we can safely assume that they do not significantly differ from normal distribution and the two samples t-test can be used.

```
t.test(diet_elim$preweight, diet_elim$weight6weeks)
```

```
## [...]
##  Welch Two Sample t-test
##
## data:  diet_elim$preweight and diet_elim$weight6weeks
## t = 3.0008, df = 152, p-value = 0.003148
## alternative hypothesis: true difference in means is not equal to 0
## [...]
```

Since the resulting p-value from the t-test is smaller than 0.05, we can reject the null hypothesis that their means are the same, i.e there is a statistical significant difference in the means and the diets do affect the weight loss.

b)

To check whether any type of diet has an effect on the lost weight, we use ANOVA to test the null hypothesis that across all three diets, the means of lost weights are the same.

```
an_mod <- aov(weight.loss ~ diet, data=diet_elim)

summary(an_mod)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## diet         1   45.5   45.50   7.741 0.00682 **
## Residuals   75  440.8    5.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is smaller than 0.05 so we can reject the null hypothesis and say that the diets have an effect on losing weight.

To check which diet is best for losing weight, we test for the their repspective means to see which has the highest means i.e expected lost weight. By definition,

$$weight.loss = preweight - weight6weeks,$$

, but $preweight$ and $weight6weeks$ are normally distributed so we can assume that $weight.loss$ is also normally distributed and the t-test can be used. We check if diet 3 is more effective than 1 and 2.

```
wl1 = subset(diet_elim, diet == "1");
wl2 = subset(diet_elim, diet == "2");
wl3 = subset(diet_elim, diet == "3");

t.test(wl3$weight.loss, wl1$weight.loss, alternative = "greater")
```

```
## [...]
##  Welch Two Sample t-test
##
## data:  wl3$weight.loss and wl1$weight.loss
## t = 2.8462, df = 48.862, p-value = 0.003225
## alternative hypothesis: true difference in means is greater than 0
```

```
t.test(wl3$weight.loss, wl2$weight.loss, alternative = "greater")
```

```
## [...]
##  Welch Two Sample t-test
##
## data:  wl3$weight.loss and wl2$weight.loss
## t = 2.9815, df = 50.672, p-value = 0.0022
## alternative hypothesis: true difference in means is greater than 0
```

Since the p-values are smaller than 0.05, we reject the null hypothesis that the means are the same so diet 3 is more effective than 1 and 2.

c)

We use two-way anova to investigate the effect of diet, gender, and their interaction on weight loss

```r
tw_aov1 <- aov(weight.loss ~ diet * gender, data=diet_elim);

summary(tw_aov1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## diet         1   45.2   45.21   7.957 0.00619 **
## gender       1    0.1    0.14   0.025 0.87521
## diet:gender  1   16.5   16.47   2.898 0.09300 .
## Residuals   72  409.1    5.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

The p-values for gender and interaction between diet and gender are larger than 0.05 so there are no statistical significance for their effects on weight loss as opposed to diet alone.

d)

We investigate the effect of diet and height using ANCOVA. We test the hypothesis $H_A : \alpha_i = \cdots = \alpha_I = 0$

```r
anc1 <- lm(weight.loss ~ height + diet, data=diet_elim);

anova(anc1)
```

```
## Analysis of Variance Table
##
## Response: weight.loss
##           Df Sum Sq Mean Sq F value   Pr(>F)
## height     1   6.09   6.091  1.0292 0.313658
## diet       1  42.24  42.240  7.1370 0.009281 **
## Residuals 74 437.97   5.918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is smaller than 0.05 so we reject the null hypothesis that the diet does not affect the weight loss. Similarly, we test for $H_\beta : \beta = 0$.

```r
anc2 <- lm(weight.loss ~ diet + height, data=diet_elim);

anova(anc2)
```

```
## Analysis of Variance Table
##
## Response: weight.loss
##           Df Sum Sq Mean Sq F value   Pr(>F)
## diet       1  45.50  45.499  7.6876 0.007031 **
## height     1   2.83   2.832  0.4786 0.491230
## Residuals 74 437.97   5.918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is larger than 0.49 so we can not reject the null hypothesis that the height does not have an effect. The interaction between diet and height us subsequently tested.

```r
anc3 <-  lm(weight.loss ~ height * diet, data=diet_elim);

anova(anc3)
```

```
## Analysis of Variance Table
##
## Response: weight.loss
##             Df Sum Sq Mean Sq F value   Pr(>F)
## height       1   6.09   6.091  1.0425 0.310602
## diet         1  42.24  42.240  7.2297 0.008879 **
## height:diet  1  11.46  11.458  1.9611 0.165629
## Residuals   73 426.51   5.843
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the interaction effect is larger than 0.05 so it does not bear any statistical significant effect towards weight loss. Furthermore, for the 3 types of diet, the effect of height is the same because of the hypothesis $H_{A\beta} : \beta_i = \cdots = \beta_I$ and we did not reject it.

e)

Out of two approaches, we prefer the d) one because in b), we did not test for the significance of height's effect on weight loss. Since diet is the only (tested) factor to have a significant effect on weight loss, we can do a simple linear regression model.

```r
lm.model <- lm(weight.loss ~ diet, data=diet_elim)

summary(lm.model)
```

```
## [...]
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9667     0.7460   2.636  0.01018 *
## diet          0.9456     0.3399   2.782  0.00682 **
## [...]
```

So, based on the model, the lost weight of an average person only depend on their chosen diet and can be given as

$$lost\_weight = 1.9 + 0.9 \times diet.$$