Exercise 2

Exercise 2: Hemoglobin in trout

In this exercise, the influence of varying amounts of the antibacterial sulfamerazine on the hemoglobin levels in the blood of brown trout is explored. The statistical analysis is performed on the data set `hemoglobin.txt`, which is examined below.

```
hg_data = read.table("data/hemoglobin.txt", header=TRUE)
summary(hg_data)
```

```
##    hemoglobin          rate         method
##  Min.   : 5.500   Min.   :1.00   Length:80
##  1st Qu.: 7.350   1st Qu.:1.75   Class :character
##  Median : 8.700   Median :2.50   Mode  :character
##  Mean   : 8.736   Mean   :2.50
##  3rd Qu.:10.200   3rd Qu.:3.25
##  Max.   :11.900   Max.   :4.00
```

```
mu <- mean(hg_data$hemoglobin)
```

The set consists of 80 entries. As can be read from the data, the administering of the treatment is conducted using two different methods, A and B, and in four types of doses, labeled 1, 2, 3 and 4 and corresponding to 0, 5, 10 and 15 g of the drug. The rate (dose) and method are the factors which will be analysed, and should therefore be transformed into factor datatypes.

```
hg_data$rate = as.factor(hg_data$rate); hg_data$method = as.factor(hg_data$method)
```

a)

Upon first inspection of the data, it could be identified that the sample in the data set exhibits *balanced design*, i.e. with an equal amount of observations per factor level combination. This can be verified by creating a contingency table of the counts at each factor level combination:

```
table(hg_data$rate, hg_data$method)
```

```
##
##      A  B
##   1 10 10
##   2 10 10
##   3 10 10
##   4 10 10
```

It is evident that each factor combination contains $N = 10$ experimental units. This makes the data suitable for a two-way ANOVA test on the factors `rate` and `method`. The code below simulates how such a balanced-design sample can be acquired from a large fish population. For this purpose, a population of $M = 1000$ fish is generated, each assigned a random `rate` and `method` factor, and a hemoglobin level sampled from a normal distribution with a mean and a standard deviation equivalent to those of the original data set. The resulting data set is then sampled to obtain a balanced design of 80 fish with $N = 10$ units per combination, and the contingency table is calculated to verify the balance of the sample.

```r
set.seed(123)
M=1000; I=4; J=2

# Generate population
rates <- sample(1:I, M, replace=TRUE); methods <- sample(c('A', 'B'), M, replace=TRUE)
hemoglobin <- rnorm(M, mean(hg_data$hemoglobin), sd(hg_data$hemoglobin))
dummy_pop <- data.frame(rate=rates, method=methods, hemoglobin=hemoglobin)

# Create factor combinations
N=10
dummy_pop$combos <- interaction(dummy_pop$rate, dummy_pop$method)
combo_groups <- split(dummy_pop, dummy_pop$combos)

# Sample balanced design
balanced_groups <- lapply(combo_groups,
                          function(group) group[sample(nrow(group),min(N, nrow(group))), ])
balanced_data <- do.call(rbind, balanced_groups)
balanced_data <- subset(balanced_data, select=-combos)

# Verify balance
table(balanced_data$rate, balanced_data$method)
```

```
##
##      A  B
##   1 10 10
##   2 10 10
##   3 10 10
##   4 10 10
```

b)

Using the provided data set, a two-way ANOVA test is performed to investigate the influence of the factors `rate` and `method` on the hemoglobin levels in the trout. The test is conducted by constructing a linear model from the two factors and using it as input for the `anova` function. By default, a treatment parameterization is used.

```r
model <- lm(hemoglobin~rate*method, data=hg_data); anova(model)
```

```
## Analysis of Variance Table
##
## Response: hemoglobin
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## rate         3  90.560 30.1868 19.4689 2.404e-09 ***
## method       1   2.415  2.4151  1.5576    0.2161
## rate:method  3   4.872  1.6241  1.0475    0.3769
## Residuals   72 111.637  1.5505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

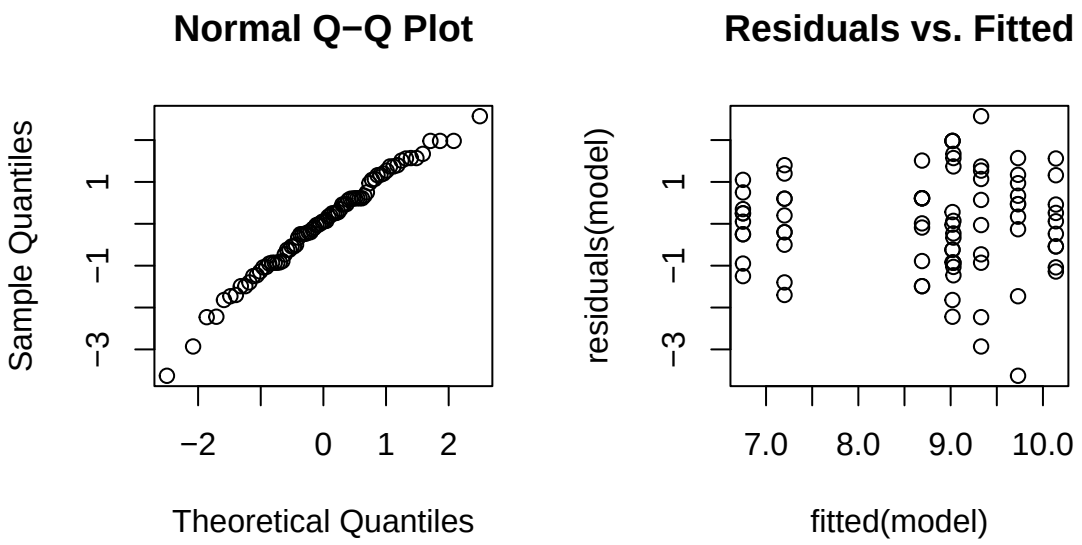
The first observation to be made is whether the H_{AB} hypothesis holds, i.e. whether an interaction between the two factors is absent. This is done by examining the p-value of the interaction term in the ANOVA table. In this case, $p = 0.3769$, which is not significant at the 5% level, and thus the null hypothesis is not rejected. This suggests that there is no evidence for interaction between the two factors. Considering the separate effect of the factors, the p-value for the `rate` factor is below the significance level, indicating a rejection of the null-hypothesis, while the p-value of the `method` factor is above the significance level, indicating a failure

to reject the null-hypothesis.

c)

To validate the test, the assumptions of the ANOVA test should be checked. The residuals of the model are examined for normality using a Q-Q-plot and are plotted against the fitted model values to ensure there is no pattern in the residuals.

```
par(mfrow=c(1,2))
qqnorm(residuals(model)); plot(fitted(model), residuals(model)); title("Residuals vs. Fitted")
```



A qualitative inspection of the plots confirms that these assumptions are met to a satisfactory degree. The validation of the model and the lack of significant interaction between the factors confirm that the indicated contribution of the `rate` factor is indeed representative for its greater effect on the response value `hemoglobin`.

Furthermore, the insignificance of the interaction allows the use of an additive model, which allows more specific comparisons between the factor levels. The `summary` function is used to extract the coefficients of the model.

```
model2 <- lm(hemoglobin~rate+method, data=hg_data); invisible(anova(model2))
summary(model2)
```

```
##
## Call:
## lm(formula = hemoglobin ~ rate + method, data = hg_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4538 -0.8881  0.0050  0.8406  2.3388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.8012     0.3116  21.827  < 2e-16 ***
## rate2         2.7600     0.3941   7.003 9.18e-10 ***
## rate3         2.4050     0.3941   6.102 4.24e-08 ***
## rate4         1.8800     0.3941   4.770 8.86e-06 ***
```

```
## methodB       0.3475     0.2787   1.247    0.216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.246 on 75 degrees of freedom
## Multiple R-squared:  0.4438, Adjusted R-squared:  0.4142
## F-statistic: 14.96 on 4 and 75 DF,  p-value: 4.919e-09
```

Due to the treatment parameterization, the first level of both factors is used as a reference level, so the contributions α_1 and β_1 contributions are equal to zero. As the hemoglobin levels can be seen as the sum $Y_{i,j} = \mu + \alpha_i + \beta_j + e_{j,k}$, the highest yield can be calculated by adding the highest α_i and β_j contributions to the mean μ. From the summary it is evident that these are rate 2 (5g dose) and method B. On the other hand, when extracting the `rate` and `method` factors for the highest hemoglobin measurement in the data set, the other method is present.: 11.9, 2, 1 However, this does not show contradiction, as the `method` factor is not significant in this case. Rate 2, on the other hand, is confirmed as the highest yielding rate, resulting in a mean hemoglobin level of 11.84375.

Another sample measurement is the mean hemoglobin yield under a combination of rate 3 and method A. This can be calculated by adding the contributions α_3 and β_1 (which is zero due to treatment parameterization) to the mean μ and obtaining 11.14125.

d)

Ignoring the factor `method`, the effect of the different `rate` levels can be examined in more detail using a one-way ANOVA to test the null-hypothesis that hemoglobin levels are the same for all levels.

```
model3 <- lm(hemoglobin~rate, data=hg_data); anova(model3)
```

```
## Analysis of Variance Table
##
## Response: hemoglobin
##           Df Sum Sq Mean Sq F value    Pr(>F)
## rate       3  90.56 30.1868  19.291 2.129e-09 ***
## Residuals 76 118.92  1.5648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model3)
```

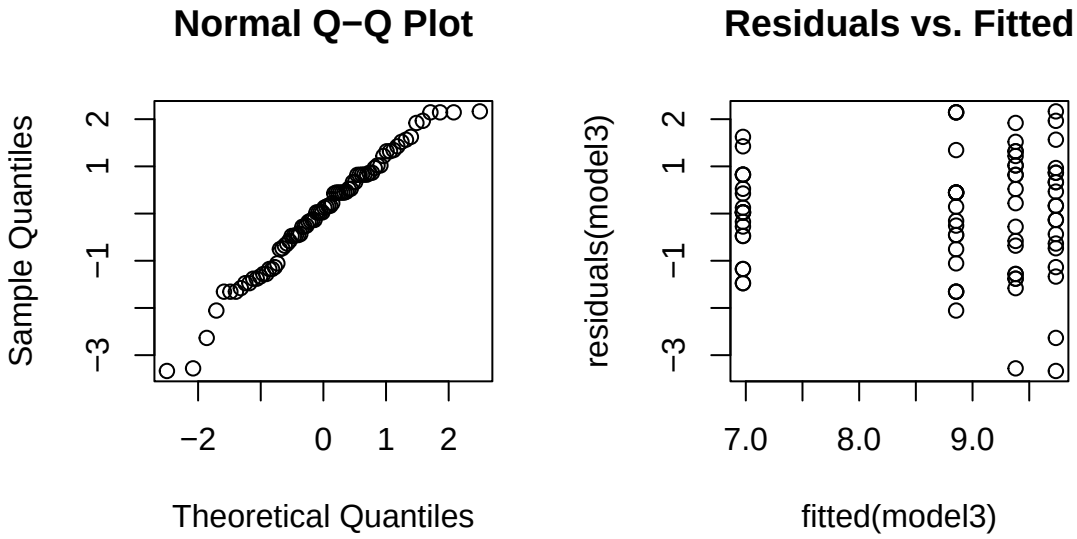
```
##
## Call:
## lm(formula = hemoglobin ~ rate, data = hg_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.335 -0.740  0.075  0.825  2.165
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.9750     0.2797  24.936  < 2e-16 ***
## rate2         2.7600     0.3956   6.977 9.69e-10 ***
## rate3         2.4050     0.3956   6.080 4.48e-08 ***
## rate4         1.8800     0.3956   4.753 9.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.251 on 76 degrees of freedom
```

```
## Multiple R-squared:  0.4323, Adjusted R-squared:  0.4099
## F-statistic: 19.29 on 3 and 76 DF,  p-value: 2.129e-09
```

Once again, the p-value of the `rate` factor is below the significance level, indicating a rejection of the null-hypothesis. The estimated hemoglobin values for each rate are obtained from the summary: for rate 1: 6.975, for rate 2: 11.49625, for rate 3: 11.14125 and for rate 4: 10.61625. Again, rate 2 is identified as the highest yielding one.

Whether this test is valid, remains to be seen by checking the assumptions of the ANOVA test. The residuals of the model are examined for normality using a Q-Q-plot and are plotted against the fitted model values.

```
par(mfrow=c(1,2))
qqnorm(residuals(model3)); plot(fitted(model3), residuals(model3)); title("Residuals vs. Fitted")
```



Even though a linear relationship between the theoretical and sample quantiles is visible, the tails show a more extreme deviation the QQ-plot of the two-way ANOVA, casting doubts on the normality assumption. This could be problematic for the correct interpretation of the test output.

e)

As an alternative, a Kruskal-Wallis test is performed to test the same null-hypothesis, as it operates with ranks and does not assume normality.

```
kruskal.test(hg_data$hemoglobin, hg_data$rate)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  hg_data$hemoglobin and hg_data$rate
## Kruskal-Wallis chi-squared = 34.224, df = 3, p-value = 1.777e-07
```

The resulting p-value is far below the significance level, indicating a rejection of the null-hypothesis and confirming the indication from the one-way ANOVA test that the hemoglobin levels differ across the rate levels. Given the questionable normality of the residuals, the Kruskal-Wallis test is a more robust alternative. However, it provides less detailed insight into the differences between the rate levels, as it uses a more generalised test statistic about the group medians and does not involve the more sophisticated model fitting of ANOVA.