Statistical Models – October 2016 — Assignment 01

Chouliaras Georgios Christos, Jiayang Zhuo
Group 15

I. Computational Problems

Problem 1

(i) In order to conduct the two sample t-test the following parameters were chosen for two data sets:

$$n_1 = 40$$
$$n_2 = 45$$
$$\mu_1 = 30$$
$$\mu_2 = 32$$
$$\sigma^2 = 3$$

The significance level for the hypotheses was chosen as $\alpha = 0.05$. The t-test results can be seen below:

$$t - value = -5.2131, \;\; df = 83, \;\; p - value = 1.336 \cdot 10^{-6}$$

The p-value is much less than the significance level 0.05 and thus, the hypothesis $H_0 : \mu_1 = \mu_2$ is rejected, meaning that the means of the two populatios is significantly different.

(ii) Next, an one-way ANOVA is conducted on the same two populations in order to compare the results of the two tests. The model was built using the `aov` function and the ANOVA table which produced with `anova` function (the complete code can be found in the appendix) is as follows:

Source	*DF*	*SS*	*MS*	*F value*	*P value*
Between groups	1	90.229	90.229	27.176	$1.336 \cdot 10^{-6}$
Within groups	83	275.572	3.320		

Table 1: ANOVA table of the two data sets.

It should be noted here that in this particular case the Sum of Squares (SS) and the Mean Square $MS = SS/DF$ are the same because the degree of freedom is 1.

(iii) From a comparison between the p - values and the test statistics for both tests it can be seen that they are identical. The degrees of freedom are 83 and the p values are $1.336 \cdot 10^{-6}$, so both tests reject the null hypothesis for equal means. This similarity stems from the fact that the two-sample t-test is just a special case of ANOVA when there are only two means or averages to be compared. It means that ANOVA is an extended version of t-test which is used to compare variation within and between $n > 2$ variables. Thus, it is reasonable for these two tests to produce identical results when there are only two groups, as in this case.

(iv) When R computes the t value for the t-test, it uses the pooled variance, as the real variance is considered unknown. If the variance is known then it can be used in the computation of t-value instead of the pooled variance:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

In this case the two variances are the same so the formula can be reformed to:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The t value calculated using this formula equals with $t = -5.313689$ which is very close to the t value from the t-test. Having the t value, the real p value calculated through the `pt` function in R as follows:

```
2*pt(-abs(T),df=n1+n2-2)
```

which is equivalent to:

$$2P_0(T > |t|)$$

This function resulted to a $p - value = 8.862702 \cdot 10^{-7}$ which is also close to the p value which calculated from the two tests, but it is more precise as it used the real variance of the data. However, it is much lower than $\alpha = 0.05$ so the there is not any difference in the conclusion of this analysis.

Problem 2

(i) The side by side boxplots of chick weights' distribution with respect to the feed supplement can be seen in the figure below:

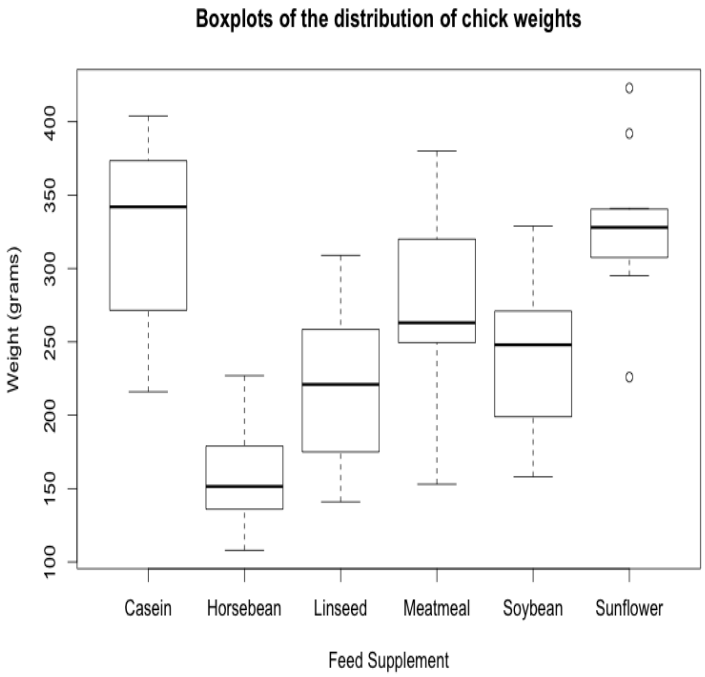


Figure 1: Side-by-side boxplots of the distribution of chick weights with respect to the feed supplement

From the boxplots it is easily observable that there is a significant difference between the ranges of chick weights for each different feed supplement. Chicks which were fed with casein seem to have the largest weights, while those who fed with horsebean seem to have the smallest. The median of the casein supplement is close to the median of the sunflower, however their distributions differ in a great extent. For the sunflower supplement there are also 2 outliers with larger weights and 1 outlier with lower weight. Moreover the box sizes differ largely, with casein having the tallest box with the biggest range of weights, while horsebean and sunflower having the shortest boxes. All the aforementioned lead to the hypothesis that the mean weights differ, however statistical analysis must be conducted in order for the hypothesis to be confirmed.

(ii) In order to conduct one-way ANOVA without using the built in R functions the following steps were followed:
Firstly, the total number of observations $N = 71$, the number of levels $I = 6$, as well as the *grand mean* $= 261.31$ were obtained. Then the response vector $\mathbf{Y}$ was created, along with the incidence matrix $\mathbb{X}$ in order to obtain the least squares estimator $\hat{\beta}$ using the following formula:

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-}\mathbb{X}^T Y$$

Before using the incidence matrix in this formula, the constraint $\mu = 0$ was added, removing the first column of 1's and resulting to a full rank matrix and a uniquely determined $\hat{\beta}$. Next, the estimator of the residuals $e_{ij} \overset{ind}{\sim} N(0, \sigma^2)$ was calculated by:

$$\hat{e_{ij}} = Y_{ij} - \bar{Y_{i.}} = \mathbf{Y} - \mathbb{X}\hat{\beta}$$

The residuals were used in the calculation of the *residual* or *within groups* sum of squares:

$$S_\Omega = (\mathbf{Y} - \mathbb{X}\hat{\beta})^T(\mathbf{Y} - \mathbb{X}\hat{\beta})$$

which led to the calculation of the unbiased variance estimator $\hat{\sigma}^2 = \frac{S_\Omega}{n-I}$. For the *between groups* sum of squares the formula is:

$$S_\omega - S_\Omega = \sum_{i=1}^{I} n_i(Y_i - \overline{Y}_{..})^2$$

However this formula applies to *balanced* design, while in this case the design is *unbalanced*, so the formula must be reformed. The formula for the *unbalanced* is as follows:

$$S_\omega - S_\Omega = n_1(Y_1 - \bar{Y_{..}})^2 + n_2(Y_2 - \bar{Y_{..}})^2 + n_3(Y_3 - \bar{Y_{..}})^2 + n_4(Y_4 - \bar{Y_{..}})^2 + n_5(Y_5 - \bar{Y_{..}})^2 + n_6(Y_6 - \bar{Y_{..}})^2$$

with $n_1, n_2, \ldots, n_6$ indicating the number of observations for each level, $Y_1, Y_2, \ldots, Y_6$ the vectors with observations of each level and $\bar{Y_{..}}$ the grand mean. Finally the *total* sum of squares was obtained by adding the *residual* and the *between groups* sums of squares. The aforementioned values were all used to calculate the value of F statistic:

$$\mathcal{F} = \frac{(S_\omega - S_\Omega)/(I-1)}{S_\Omega/(n-I)}$$

Using this F value, a p value was obtained in order to test the hypothesis. The calculated values are summarized in an ANOVA table:

Source	*DF*	*SS*	*MS*	*F value*	*P value*
Between groups	5	231129.162	46225.83	15.3648	$5.93642 \cdot 10^{-10}$
Within groups	65	195556	3008.554		
Total	70	426685.2			

Table 2: ANOVA table of the chick weight with respect to feed supplements.

(iii) The P value in the table is $P_{H_0}(\mathcal{F} \geq 15.3648) = 5.93642 \cdot 10^{-10} < 0.05$, so the null hypothesis for equality of the means H_0 is rejected and the alternative hypothesis H_1 holds. In order to check the validity of this result a similar analysis was conducted, but this time using the built in R functions. More specifically the functions which were used were:

`model = aov(weight~feed,data=data)` and `anova(model)`

These functions resulted to the following ANOVA table:

```
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
feed       5 231129   46226  15.365 5.936e-10 ***
Residuals 65 195556    3009
```

From a comparison of the two tables it can be derived that the values of both models are almost identical. The P value of the model created by the `aov` function is again $5.936 \cdot 10^{-10} < 0.05$, something that confirms the rejection of the null hypothesis H_0 for equality of the means. After this analysis it can be concluded that the feed supplement plays a significant role in the evolution of the chick weight and this is the reason for the differences which observed in the side by side boxplots of the chick weights.

(iv) In order for this analysis to be considered valid, there is the necessity for confirmation of the assumptions which were made during the procedure of the model building. One essential assumption which needs to be confirmed is the normality of the residuals. Hence, the residuals $\mathbf{Y} - \mathbb{X}\hat{\beta}$ were plotted against the fitted values $\mathbb{X}\hat{\beta}$ as it can be seen in the following figure:

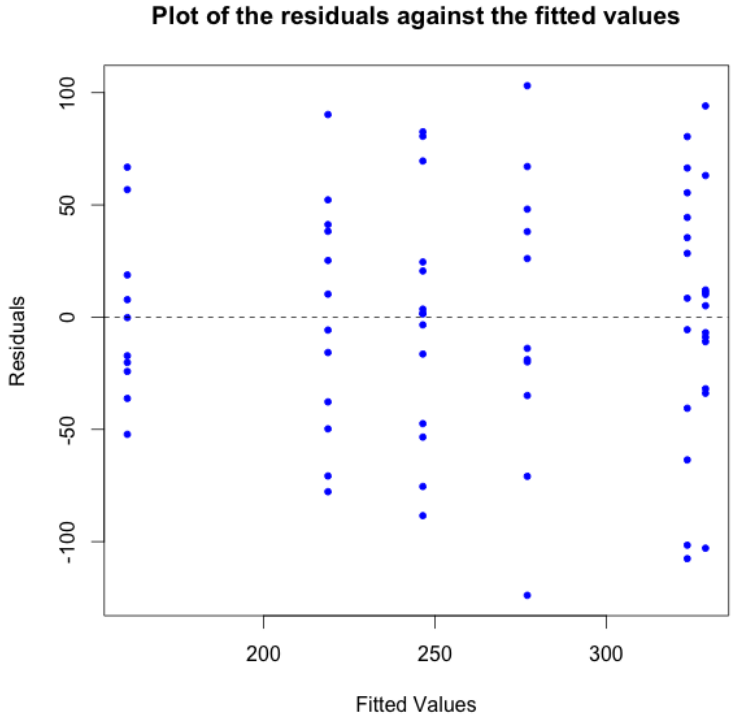


Figure 2: A plot of the residuals $\mathbf{Y} - \mathbb{X}\hat{\beta}$ *against the fitted values* $\mathbb{X}\hat{\beta}$

From an observation of figure 2 there is not any visible pattern in the residuals and they also seem to be spread in negative as well as positive values. Moreover, a normal quantile quantile plot of the residuals in figure 3 shows that the relationship between the sample and the theoretical quantiles is mainly linear, and thus the assumptions $\mathbb{E}(e_{ij}) = 0$ and $\mathbb{V}ar(e_{ij}) = \sigma^2$ are valid. This argument is also confirmed by the *Shapiro-Wilk* normality test on the residuals which produced a $p\ value = 0.6272 > 0.05$ and as a result the H_0 hypothesis for normality is not being rejected. Lastly, a *Bartlett test* for constancy of error variance returned a $p\ value = 0.66 > 0.05$, which does not reject the H_0 hypothesis for homoscedasticity and thus, the variances across groups are considered equal.

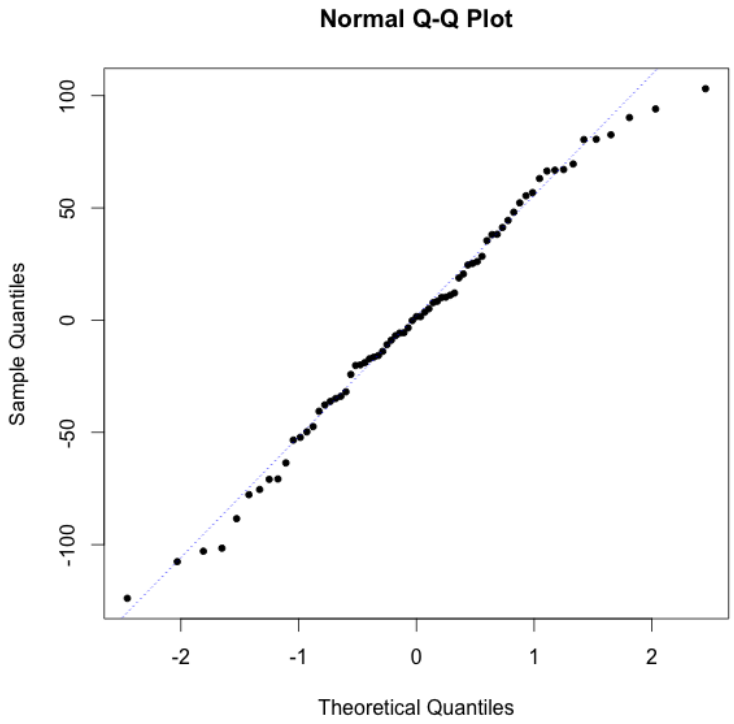


Figure 3: Normal Q-Q plot of the residuals for confirmation of normality

Problem 3

(i) In order to test the null hypothesis $H_{AB} : \gamma_{ij} = 0$ for interaction between the two factors, a two-way ANOVA model was created, with *hemoglobin* as the response variable, the *rates of sulfamerazine* as factor A with I=4 levels and *methods of administering* as factor B with J=2 levels. As ten fish were measured for each rate and each method, the total number of observations is $N = 10 \cdot I \cdot J = 10 \cdot 4 \cdot 2 = 80$. The two-way ANOVA model $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ was created using `aov` function:

```
model=aov(hemoglobin rate+method+rate:method,data)
```

and function `anova(model)` resulted to the ANOVA table with the following p values:

Source	*P value*
Rate	$2.404 \cdot 10^{-9}$
Method	0.2161
Interaction	0.3769

Table 3: P values of the full model, containing interaction.

From the table it is clear that the interaction has a p value larger than the significance level $\alpha = 0.05$ and thus, the null hypothesis $H_{AB} : \gamma_{ij} = 0$ is not rejected, leading to the conclusion that there is no interaction between the factors *rate* and *method*.

(ii) The parameters for the model were obtained using the `model.tables(model)` function, while the grand mean was obtained through `mean(data$hemoglobin)` and are summarized below:

μ	α_1	α_2	α_3	α_4
8.73625	-1.7613	0.9988	0.6438	0.1188

Table 4: Grand mean and the parameters for the rate factor under the full model

β_1	β_2
-0.17375	0.17375

Table 5: Parameters for the method factor under the full model

	method I	method II
rate 1	0.3987	-0.3987
rate 2	-0.2312	0.2312
rate 3	-0.1762	0.1762
rate 4	0.0088	-0.0088

Table 6: Parameters for the interaction between the two factors

(iii) Since the hypothesis H_{AB} was not rejected, the interaction can be removed from the model without affecting the results. This leads to an additive model $Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$, containing only the two factors. An ANOVA table of this model can produce evidence about the rejection or confirmation of the hypotheses $H_A : \alpha_1 = \ldots = \alpha_I = 0$ and $H_B : \beta_1 = \ldots = \beta_J = 0$. The additive model was created in a similar way:

```
additive=aov(hemoglobin~rate+method,data)
```

and the function `anova(additive)` resulted to the ANOVA table with the following p values:

Source	*P value*
Rate	$2.02 \cdot 10^{-9}$
Method	0.2163

Table 7: P values of the additive model, without interaction.

Also, the new estimates for the parameters of the additive model are as follows:

μ	α_1	α_2	α_3	α_4
8.73625	-1.7613	0.9988	0.6438	0.1188

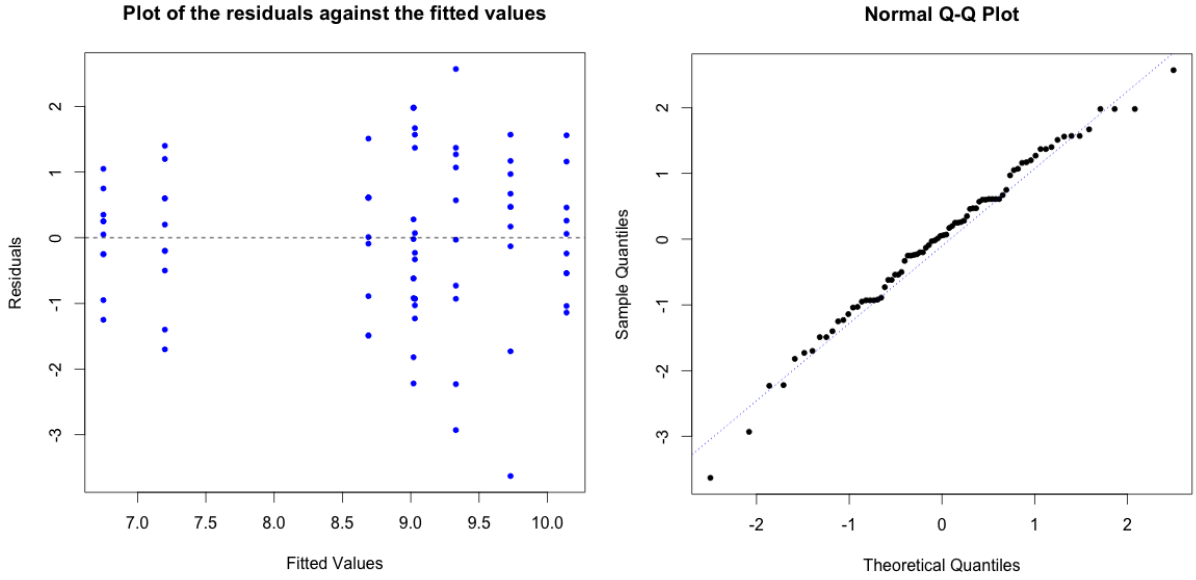
Table 8: Grand mean and the parameters for the rate factor under the additive model

β_1	β_2
-0.17375	0.17375

Table 9: Parameters for the method factor under the additive model

It can be noticed that the parameter estimates under the additive model are identical with those of the full model with only a slight, insignificant difference in the p values of the two factors. This result is reasonable, due to the fact that the H_{AB} hypothesis was not rejected, meaning that the interaction does not play a role to the model. Hence, the full model did not take the interaction into consideration, when it calculated its parameter estimates. As a result, the additive model which also did not take interaction into consideration as it was removed, led to exactly the same parameter estimations. Concerning the effects of the two factors, both models resulted to almost identical p values: $\approx 2 \cdot 10^{-9} < 0.05$ for the *rate* factor and $\approx 0.21 > 0.05$ for the *method* factor. As a consequence, the hypothesis $H_A : \alpha_1 = \ldots = \alpha_I = 0$ is rejected, while the $H_B : \beta_1 = \ldots = \beta_J = 0$ is not. This means that *hemoglobin* is being affected only by the four *rates of sulfamerazine*, and not from the two *methods of sulfamerazine's administering* or the *interaction* between them.

(iv) In order for the whole analysis to be considered valid, the assumptions for every model have to be verified. Regarding the full model, a plot of the residuals against the fitted values seems not to have patterns and a normal Q-Q plot shows linearity between sample and theoretical quantiles with only a few outliers. Also a *Shapiro-Wilk* normality test produced a p value = 0.526 and thus the null hypothesis for normality is not rejected. All the above confirm the normality assumptions for the full model.



*(a) Residuals $\boldsymbol{Y} - \mathbb{X}\hat{\beta}$ against the fitted values $\mathbb{X}\hat{\beta}$ for the **full** model*

(b) Normal Q-Q plot of the residuals for confirmation of normality

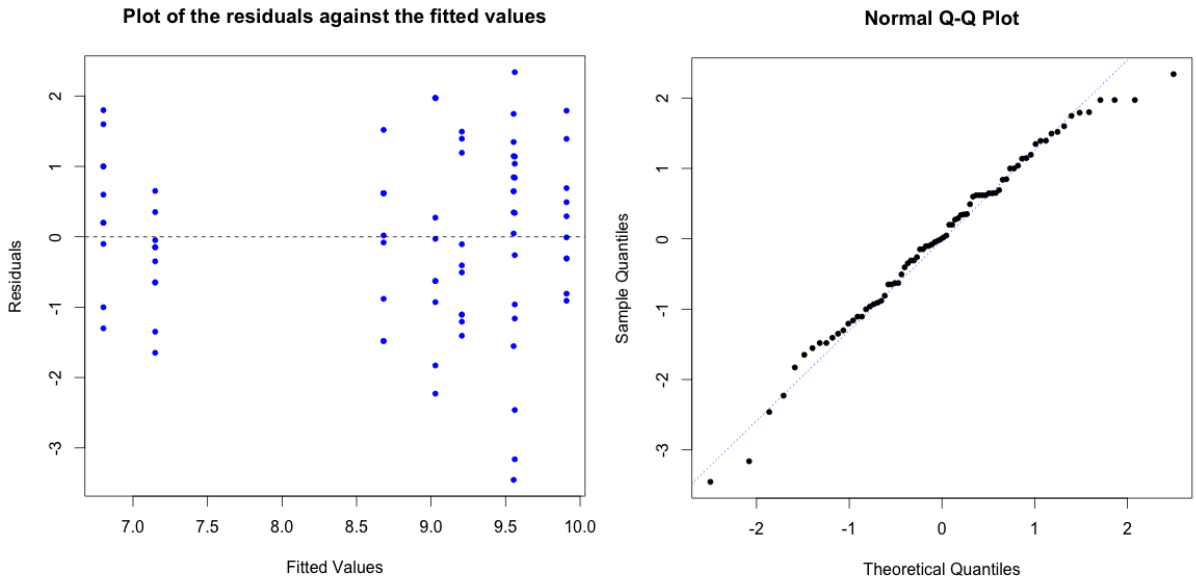
Figure 4: Check for validity of normality assumptions on the residuals of the full model

A similar analysis took place for the additive model. Again, the plot of the residuals against the fitted values as well as the normal Q-Q plot, show that the residuals can be assumed normal. The *Shapiro-Wilk* normality test had a p value = 0.3269, which is strong evidence that the residuals can be considered normal. In order to check for the equality of variances, three *Bartlett Tests* took place: one for the factor *rate*, one for the factor *method* and one for the *interaction*. The p values of the tests are summarized below:

Bartlett test	P value
factor *rate*	0.2177
factor *method*	0.2201
interaction	0.1291

Table 10: Bartlett tests for equality of variances.

All the p values are above the significance level, so the null hypothesis for homogeneity is not rejected in any of the tests.



*(a) Residuals $\boldsymbol{Y} - \mathbb{X}\hat{\beta}$ against the fitted values $\mathbb{X}\hat{\beta}$ for the **additive** model*

(b) Normal Q-Q plot of the residuals for confirmation of normality

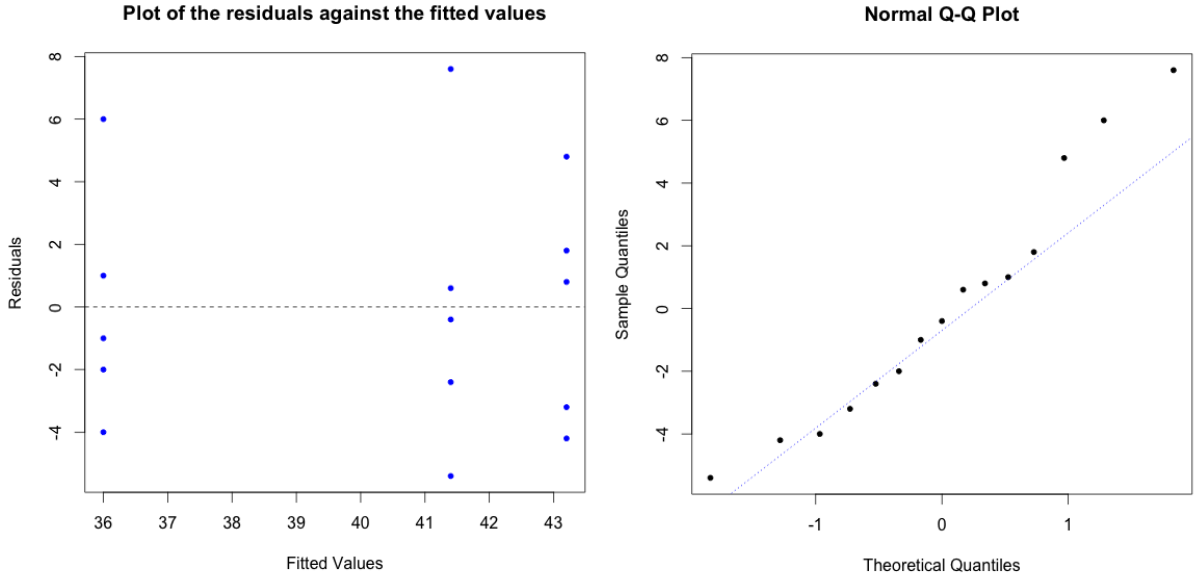
Figure 5: Check for validity of normality assumptions on the residuals of the additive model

Problem 4

(i) In order to investigate whether the factor *machine* influences the fiber *strength* or not, a one-way ANOVA model $Y_{ij} = \mu + \alpha_i + e_{ij}$ is constructed taking into account only the factor *machine*. The model which was built using `model1 = aov(strength~type,data = data)` produced the following p value:

$$P_{H_A}(\mathcal{F} \geq f) = 0.04423$$

The p value is smaller than the significance level 0.05 and thus, the null hypothesis $H_A : \alpha_1 = \ldots = \alpha_I = 0$ is being rejected, indicating that the type of the machine influences the fiber strength. However, this p value is only slightly smaller than the significance level and thus, the influence of the machine type to the fiber strength, is also small. At this point, it is essential for the normality assumptions to be checked for the residuals of this model. The plot of the residuals against the fitted values does not show any pattern which indicates that the assumption for normality of the residuals is valid. Furthermore, in a normal quantile quantile plot of the residuals, there is clearly a linear relationship between the sample and the theoretical quantiles, although there is some minor skewness. Lastly, a *Shapiro-Wilk* test for normality produces a p value 0.4991 which does not rejects the null hypothesis for normality and hence, the model can be considered valid.



(a) Residuals $\boldsymbol{Y} - \mathbb{X}\hat{\beta}$ against the fitted values $\mathbb{X}\hat{\beta}$ for the one way ANOVA model

(b) Normal Q-Q plot of the residuals for confirmation of normality

Figure 6: Check for validity of normality assumptions on the residuals of the one way ANOVA model

(ii) The next step is to check whether the parameter *machine* influences the fiber *strength*, including this time also the explanatory variable *thickness* as a covariate in the model. This test can be conducted by creating an ANCOVA model with one factor and one covariate $Y_{ij} = \mu + \alpha_i + \beta x_{ij} + e_{ij}$. It is essential that the factor is added second in the `aov` function:

```
model2 = aov(strength~thickness+type,data=data)
```

The p values for both the covariate and the factor can be obtained using the `drop1(model2,test="F")` which produces the correct values regardless of the factor's order, and they are summarized below:

Source	*P value*
Thickness	$4.264 \cdot 10^{-6}$
Machine Type	0.1181

Table 11: P values of the model containing both the covariate and the factor.

From the table it is indicated that the p value for the *machine type* is this time larger than the significance level and hence, the null hypothesis for the main effect of the factor $H_A : \alpha_1 = \ldots = \alpha_I = 0$ is not rejected. This can also be interpreted as different intercepts per group. Also, from the table can be derived that the the null hypothesis for presence of the covariate $H_\beta = 0$ is being rejected which is equivalent to saying that the regression lines are not horizontal and there is a dependence between the covariate and the response variable, since the p value is much smaller than the significance level α. The result for the hypothesis H_A contradicts with the findings from the one way ANOVA model, because the first time the *machine type* was (slightly) significant for the fiber *strength*, but after the inclusion of the covariate the factor is insignificant. This can be explained as follows: the ANCOVA model uses the covariate to reduce the variance in the error term in order to provide a more precise measurement of the effect. This means that the p value in the second case (0.1181) is more accurate than the one in the first case (0.04423), due to the fact that the covariate added explanatory value to the model.

(iii) The dependence of the fiber strength on thickness can be investigated visually by producing the regression lines of each *type* of machine in a plot between *strength* and *thickness*. The dependence of the strength on thickness is the same for all the machines only if these regression lines are parallel. From a visual inspection of figure 7 it seems that these regression lines are not parallel, however it is essential to investigate how significant is this difference in the slopes. This can be answered by testing the assumption for equal slopes in the groups: $H_{A\beta} = \beta_1 = \ldots = \beta_I$. The model which tests for interaction between the factor and the covariate can be built using:

```
aov(strength~thickness*type,data=data)
```

In the ANOVA table produced by `anova` function in this model, only the p value for the interaction between the factor and the covariate is valid. This value is :

$$P_{H_{A\beta}}(\mathcal{F} \geq f) = 0.6293$$

which indicates that the null hypothesis $H_{A\beta} = \beta_1 = \ldots = \beta_I$ for equality of slopes is not rejected. This means that the difference in the slopes which was observed visually, is not statistically significant and thus, they can be assumed equal. In other words, the dependence of the fiber strength on thickness is the same for all the types of machines.

The estimated strength for a fiber with the average thickness can be computed using the formulas:

$$Y_1 = \mu + \alpha_1 + \beta x_{average}$$
$$Y_2 = \mu + \alpha_2 + \beta x_{average}$$
$$Y_3 = \mu + \alpha_3 + \beta x_{average}$$

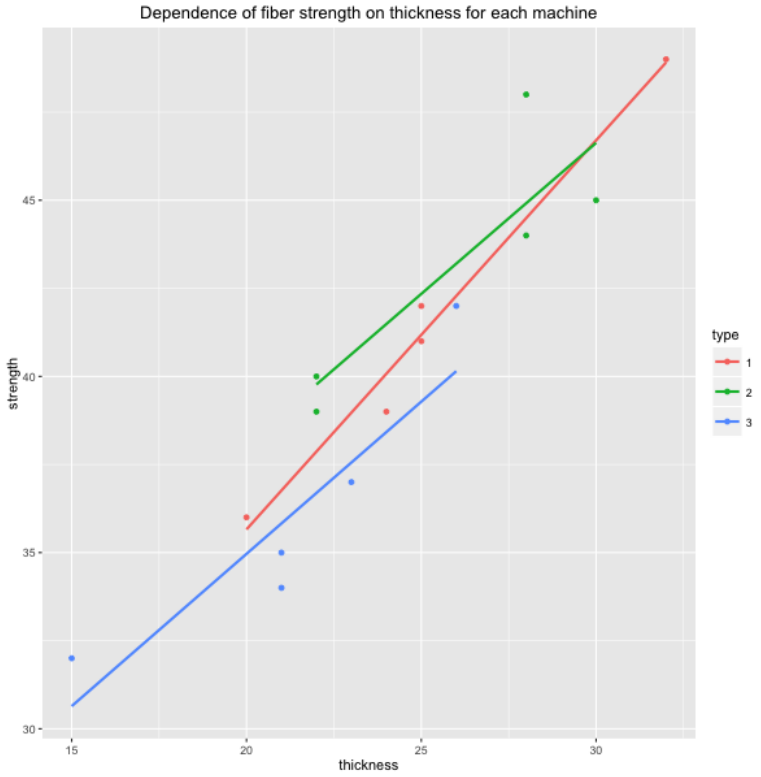


Figure 7: Regression lines for each type of machine

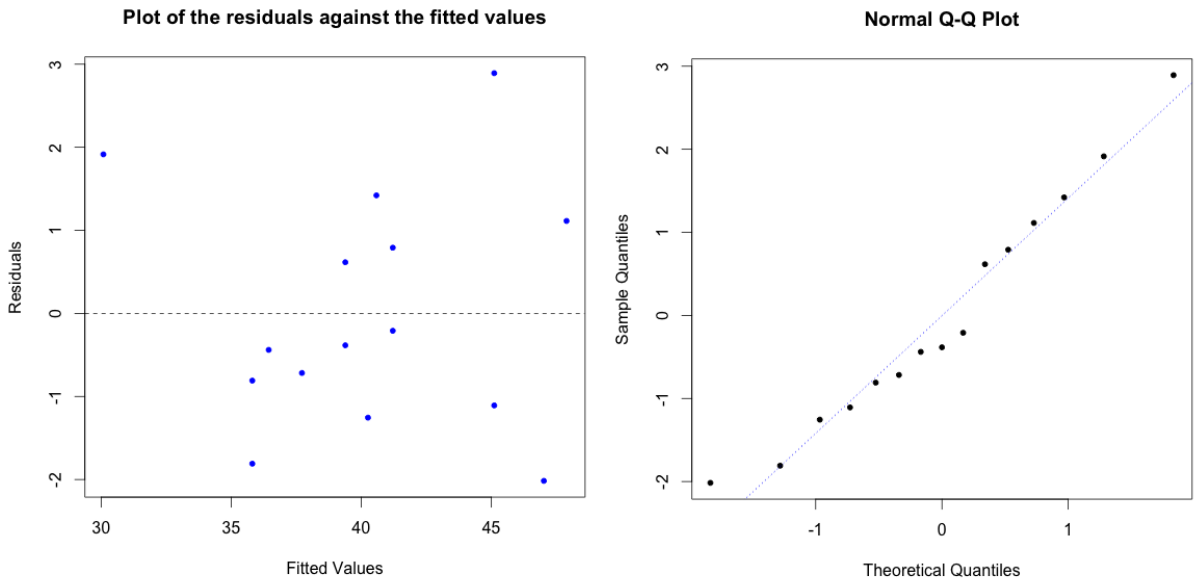
The parameters which obtained using the `model$coefficients` function led to the following estimated fiber strengths:

$$Y_1 = 40.38241 \quad Y_2 = 41.41922 \quad Y_3 = 38.79836$$

These values are very close to the grand mean $\mu = 40.2$ and this is another indication that the type of the machine is insignificant for the fiber strength.

(iv) Between the two models, the one which includes the covariate is more preferable, as it uses the covariate to reduce the error term's variance and thus, it provides more accurate estimations. The use of the model without the covariate, may mislead the analysis to false conclusions, as it assumes that the influence of the type of the machine is significant for the fiber strength. However, in order for the model which includes the covariate to be considered valid, there is the necessity for the normality assumptions to be confirmed.

From figure 8 it can be seen that the residuals against the fitted values are widely spread without any visible pattern and also the sample and theoretical quantiles have a linear relationship. Moreover a *Shapiro-Wilk* test produced a p value = 0.7201 which does not rejects the hypothesis for normality. Finally, the *Bartlett* test for homogeneity of variances resulted to a p value = 0.8521 and thus, the null hypothesis for homoscedasticity is not rejected. The aforementioned conclude that the model and also the analysis can be considered valid.



(a) Residuals $\boldsymbol{Y} - \mathbb{X}\hat{\beta}$ against the fitted values $\mathbb{X}\hat{\beta}$ for the ANCOVA model

(b) Normal Q-Q plot of the residuals for confirmation of normality

Figure 8: Check for validity of normality assumptions on the residuals of the ANCOVA model

II. Theoretical Problems

Problem 1

(i) In order for $\hat{\alpha_i}$ to be unbiased estimators of α_i for $i = 1, \ldots, I$ it needs to be shown that:

$$\mathbb{E}(\hat{\alpha_i}) = \alpha_i \tag{1}$$

Under the constraint $\mu = 0$:

$$\hat{\alpha_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \overline{Y}_{i.} \tag{2}$$

Hence:

$$\mathbb{E}(\hat{\alpha_i}) = \mathbb{E}(\frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}) = \frac{1}{n_i} \mathbb{E} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{E} Y_{ij} \tag{3}$$

But $\mathbb{E}Y_{ij} = \eta_i$ which under the constraint $\mu = 0$ equals with α_i, i.e. $\mathbb{E}Y_{ij} = \eta_i = \alpha_i$. So equation (3) becomes:

$$\mathbb{E}(\hat{\alpha_i}) = \frac{1}{n_i} \sum_{j=1}^{n_i} \alpha_i = \frac{1}{n_i} n_i \alpha_i = \alpha_i$$

□

(ii) It holds that:

$$\overline{Y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \Leftrightarrow n_i\overline{Y}_{i.} = \sum_{j=1}^{n_i} Y_{ij} \Leftrightarrow$$

$$\sum_{j=1}^{n_i} \overline{Y}_{i.} = \sum_{j=1}^{n_i} Y_{ij} \tag{4}$$

Also, $S_\omega = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2$ and $S_\Omega = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2$ so their difference:

$$S_\omega - S_\Omega = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2 - \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 \overset{\pm\overline{Y}_{i.}}{=}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}[(Y_{ij} - \overline{Y}_{i.}) + (\overline{Y}_{i.} - \overline{Y}_{..})]^2 - \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 =$$

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}[(Y_{ij} - \overline{Y}_{i.})^2 + 2(Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..}) + (\overline{Y}_{i.} - \overline{Y}_{..})^2] - \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 =$$

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}2(Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..}) + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{Y}_{i.} - \overline{Y}_{..})^2 - \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 =$$

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}2(Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..}) + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{Y}_{i.} - \overline{Y}_{..})^2 \overset{(4)}{=} \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{Y}_{i.} - \overline{Y}_{..})^2 \Leftrightarrow$$

$$S_\omega - S_\Omega = \sum_{i=1}^{I} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$$

□

(iii) It is true that:

$$\mathbb{E}(S_\omega + S_\Omega) = \mathbb{E}S_\omega + \mathbb{E}S_\Omega \tag{5}$$

It is known that the expected value for the χ^2_k equals to k. It is also known that $\frac{S_\Omega}{\sigma^2} \sim \chi^2_{n-I}$ and $\frac{S_\omega}{\sigma^2} \sim \chi^2_{n-1}$ so:

$$\mathbb{E}(\frac{S_\Omega}{\sigma^2}) = n - I \Leftrightarrow \mathbb{E}(S_\Omega) = \sigma^2(n - I)$$

$$\mathbb{E}(\frac{S_\omega}{\sigma^2}) = n - 1 \Leftrightarrow \mathbb{E}(S_\omega) = \sigma^2(n - 1)$$

Inserting these results in equation (5):

$$\mathbb{E}(S_\omega + S_\omega) = \sigma^2(n-I) + \sigma^2(n-1) = \sigma^2(2n-I-1)$$

(iv) The variance for the χ^2_k equals to 2k and also $Var(aX) = a^2Var(X)$.

$$Var(S_\Omega - S_\omega) = Var(-1(S_\omega - S_\Omega)) = (-1)^2 Var(S_\omega - S_\Omega) \Leftrightarrow$$

$$Var(S_\Omega - S_\omega) = Var(S_\omega - S_\Omega)$$

But it is known that $\frac{S_\omega - S_\Omega}{\sigma^2} \sim \chi^2_{I-1}$ and thus:

$$Var(\frac{S_\omega - S_\Omega}{\sigma^2}) = 2(I-1) \Leftrightarrow \frac{1}{(\sigma^2)^2} Var(S_\omega - S_\Omega) = 2(I-1) \Leftrightarrow$$

$$Var(S_\omega - S_\Omega) = 2\sigma^4(I-1)$$

Problem 2

(i) It needs to be shown that:

$$\mathbb{E}[\hat{\mu}] = \mu \quad and \quad \mathbb{E}[\hat{\alpha_i}] = \alpha_i$$

Under the constraint $\sum_{i=1}^{I} \alpha_i = 0$ the one way ANOVA model is $Y_{ij} = \mu + \alpha_i + e_{ij}$ and also $\mathbb{E}e_{ij} = 0$. For the estimator of the μ:

$$\hat{\mu} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} = \overline{Y}_{..} \Leftrightarrow$$

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} = \frac{1}{I}\mathbb{E}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} = \frac{1}{I}\mathbb{E}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i}(\mu + \alpha_i + e_{ij}) =$$

$$\frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i}(\mathbb{E}\mu + \mathbb{E}\alpha_i + \mathbb{E}e_{ij})$$

But $\mathbb{E}e_{ij} = 0,\ \mathbb{E}\alpha_i = \alpha_i,\ \mathbb{E}\mu = \mu$:

$$\mathbb{E}[\hat{\mu}] = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i}(\mu + \alpha_i) = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i}\mu + \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i}\alpha_i =$$

$$\frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}J_i\mu + \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}J_i\alpha_i = \frac{1}{I}\sum_{i=1}^{I}\mu + \frac{1}{I}\sum_{i=1}^{I}\alpha_i \overset{\sum_{i=1}^{I}\alpha_i = 0}{=\!=} \frac{1}{I}I\mu \Leftrightarrow$$

$$\mathbb{E}[\hat{\mu}] = \mu$$

□

For the estimator of α_i:

$$\hat{a}_i = \frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} - \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} \Leftrightarrow$$

$$\hat{a}_i = \frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} - \hat{\mu}$$

For the expected value of $\hat{a}_i$:

$$\mathbb{E}[\hat{a}_i] = \mathbb{E}\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} - \mathbb{E}[\hat{\mu}] \overset{\mathbb{E}[\hat{\mu}]=\mu}{=\!=} \frac{1}{J_i}\mathbb{E}[\sum_{j=1}^{J_i} \mu + \alpha_i + e_{ij}] - \mu =$$

$$\frac{1}{J_i}\sum_{j=1}^{J_i} \mathbb{E}\mu + \mathbb{E}\alpha_i + \mathbb{E}e_{ij} - \mu \overset{\mathbb{E}e_{ij}=0}{=\!=} \frac{1}{J_i} J_i(\mu + \alpha_i) - \mu \Leftrightarrow$$

$$\mathbb{E}[\hat{a}_i] = \alpha_i$$

□

(ii) For the variance of $\hat{\mu}$:

$$Var(\hat{\mu}) = Var(\frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij}) = \frac{1}{I^2}\sum_{i=1}^{I}\frac{1}{J_i^2}\sum_{j=1}^{J_i} Var(\mu + \alpha_i + e_{ij})$$

But μ and α_i are constants so $Var(\mu) = Var(\alpha_i) = 0$ and also, $Var(e_{ij}) = \sigma^2$ because of the normality assumption.

$$Var(\hat{\mu}) = \frac{1}{I^2}\sum_{i=1}^{I}\frac{1}{J_i^2}\sum_{j=1}^{J_i}\sigma^2 = \frac{1}{I^2}\sum_{i=1}^{I}\frac{1}{J_i^2} J_i\sigma^2 = \frac{1}{I^2}\sum_{i=1}^{I}\frac{1}{J_i}\sigma^2$$

Here $\sum_{i=1}^{I}\frac{1}{J_i}$ equals to $\frac{1}{n}$, where n is the total number of observations, hence:

$$Var(\hat{\mu}) = \frac{1}{I^2}\frac{\sigma^2}{n} = \frac{\sigma^2}{nI^2}$$

Similarly:

$$Var(\hat{\alpha}_i) = Var(\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij} - \hat{\mu}) = Var(\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij}) - Var(\hat{\mu}) =$$

$$\frac{1}{J_i^2}\sum_{j=1}^{J_i} Var(\mu + \alpha_i + e_{ij}) - \frac{\sigma^2}{nI^2} = \frac{1}{J_i^2}\sum_{j=1}^{J_i}\sigma^2 - \frac{\sigma^2}{nI^2} = \frac{\sigma^2}{J_i} - \frac{\sigma^2}{nI^2} \Leftrightarrow$$

$$Var(\hat{\alpha}_i) = \sigma^2(\frac{1}{J_i} - \frac{1}{nI^2})$$

For the covariate it holds that $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$ so in this case:

$$Var(\hat{\mu} + \hat{\alpha}_i) = Var(\hat{\mu}) + Var(\hat{\alpha}_i) + 2Cov(\hat{\mu}, \hat{\alpha}_i) \tag{6}$$

$$Var(\hat{\mu}+\hat{\alpha}_i) = Var(\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij}) = Var(\frac{1}{J_i}\sum_{j=1}^{J_i}(\mu+\alpha_i+e_{ij})) = \frac{1}{J_i^2}\sum_{j=1}^{J_i}\sigma^2 = \frac{1}{J_i^2}J_i\sigma^2 \Leftrightarrow$$

$$Var(\hat{\mu}+\hat{\alpha}_i) = \frac{\sigma^2}{J_i}$$

Equation (6) now becomes:

$$\frac{\sigma^2}{J_i} = \frac{\sigma^2}{nI^2} + \frac{\sigma^2}{J_i} - \frac{\sigma^2}{nI^2} + 2Cov(\hat{\mu},\hat{\alpha}_i) \Leftrightarrow$$

$$Cov(\hat{\mu},\hat{\alpha}_i) = 0$$

Problem 3

It has to be proved that:

$$\mathbb{E}[\hat{\gamma}_{ij}] = \gamma_{ij}$$

The two way ANOVA model is $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ with $\mathbb{E}e_{ijk} = 0$. In order for the model to be identifiable the additional conditions hold:

$$\sum_{i=1}^{I}\alpha_i = \sum_{j=1}^{J}\beta_j = 0$$

$$\sum_{i=1}^{I}\gamma_{ij} = \sum_{j=1}^{J}\gamma_{ij} = 0$$

The estimators of the parameters are:

$$\hat{\mu} = \overline{Y}_{...}$$
$$\hat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{...} = \overline{Y}_{i..} - \hat{\mu}$$
$$\hat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{...} = \overline{Y}_{.j.} - \hat{\mu}$$
$$\hat{\gamma}_{ij} = \overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}$$

The $\hat{\gamma}_{ij}$ can be written as:

$$\hat{\gamma}_{ij} = \underbrace{\frac{1}{K}\sum_{k=1}^{K}Y_{ijk}}_{\overline{Y}_{ij.}} - \underbrace{\frac{1}{J}\sum_{j=1}^{J}\frac{1}{K}\sum_{k=1}^{K}Y_{ijk}}_{\overline{Y}_{i..}} - \underbrace{\frac{1}{I}\sum_{i=1}^{I}\frac{1}{K}\sum_{k=1}^{K}Y_{ijk}}_{\overline{Y}_{.j.}} + \underbrace{\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{1}{K}\sum_{k=1}^{K}Y_{ijk}}_{\overline{Y}_{...}}$$

$$\mathbb{E}[\hat{\gamma}_{ij}] = \mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}(\mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijk})] - \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}\frac{1}{K}\sum_{k=1}^{K}(\mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijk})]-$$

$$\mathbb{E}[\frac{1}{I}\sum_{i=1}^{I}\frac{1}{K}\sum_{k=1}^{K}(\mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijk})] + \mathbb{E}[\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{1}{K}\sum_{k=1}^{K}(\mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijk})]$$

But $\mathbb{E}(\mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijk}) = \mu+\alpha_i+\beta_j+\gamma_{ij}$ and thus:

$$\mathbb{E}[\hat{\gamma}_{ij}] = \frac{1}{K}\sum_{k=1}^{K}(\mu+\alpha_i+\beta_j+\gamma_{ij}) - \frac{1}{J}\sum_{j=1}^{J}\frac{1}{K}\sum_{k=1}^{K}(\mu+\alpha_i+\beta_j+\gamma_{ij})-$$

$$\frac{1}{I}\sum_{i=1}^{I}\frac{1}{K}\sum_{k=1}^{K}(\mu+\alpha_i+\beta_j+\gamma_{ij}) + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{1}{K}\sum_{k=1}^{K}(\mu+\alpha_i+\beta_j+\gamma_{ij}) =$$

$$\frac{K}{K}(\mu+\alpha_i+\beta_j+\gamma_{ij}) - \frac{1}{J}\sum_{j=1}^{J}\frac{1}{K}K(\mu+\alpha_i+\beta_j+\gamma_{ij})-$$

$$\frac{1}{I}\sum_{i=1}^{I}\frac{1}{K}K(\mu+\alpha_i+\beta_j+\gamma_{ij}) + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{1}{K}K(\mu+\alpha_i+\beta_j+\gamma_{ij}) =$$

$$\mu+\alpha_i+\beta_j+\gamma_{ij}-\frac{1}{J}\sum_{j=1}^{J}(\mu+\alpha_i+\underbrace{\beta_j}_{=0}+\underbrace{\gamma_{ij}}_{=0})-\frac{1}{I}\sum_{i=1}^{I}(\mu+\underbrace{\alpha_i}_{=0}+\beta_j+\underbrace{\gamma_{ij}}_{=0})+\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(\mu+\alpha_i+\underbrace{\beta_j+\gamma_{ij}}_{=0}) =$$

$$\mu+\alpha_i+\beta_j+\gamma_{ij} - \frac{1}{J}\sum_{j=1}^{J}(\mu+\alpha_i) - \frac{1}{I}\sum_{i=1}^{I}(\mu+\beta_j) + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(\mu+\underbrace{\alpha_i}_{=0}) =$$

$$\mu+\alpha_i+\beta_j+\gamma_{ij} - \frac{J}{J}(\mu+\alpha_i) - \frac{I}{I}(\mu+\beta_j) + \frac{1}{IJ}J\sum_{i=1}^{I}\mu =$$

$$\mu+\alpha_i+\beta_j+\gamma_{ij} - \mu - \alpha_i - \mu - \beta_j + \frac{IJ}{IJ}\mu =$$

$$\mu + \alpha_i + \beta_j + \gamma_{ij} - \mu - \alpha_i - \mu - \beta_j + \mu \Leftrightarrow$$

$$\mathbb{E}[\hat{\gamma}_{ij}] = \gamma_{ij}$$

□

Problem 4

If the variance σ^2 is known, one can check for the presence of the main effects and interaction without calculating the F-statistic. More precisely, for *main effects* in the additive model:

$$Y_{ijk} = \mu + \alpha_i = \beta_j + e_{ijk}$$

The null hypothesis H_A can be checked by calculating the factor A sum of squares:

$$SS_A = S_{\omega A} - S_\Omega = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} (\overline{Y}_{i..} - \overline{Y}_{...})^2$$

It is known that SS_A/σ^2 (the variance is considered known) belongs to χ^2_{I-1} with I-1 degrees of freedom. Thus, one can determine if effect A is significant or not, by checking if the p value $Pr(X \geq \frac{SS_A}{\sigma^2} | H_A)$ is larger or smaller than the determined significance level α.

Similarly to test for the presence of *factor B*, it needs to be calculated the factor B sum of squares:

$$SS_B = S_{\omega B} - S_\Omega = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} (\overline{Y}_{.j.} - \overline{Y}_{...})^2$$

Again, SS_B/σ^2 belongs to χ^2_{J-1} with J-1 degrees of freedom, so H_B hypothesis can be tested by calculating the p value $Pr(X \geq \frac{SS_B}{\sigma^2} | H_B)$. If it is smaller than the significance level α the H_B is rejected, otherwise factor B can be removed from the model.

In the same fashion, the *interaction* can be tested, by considering the full model:

$$Y_{ijk} = \mu + \alpha_i = \beta_j + \gamma_{ij} + e_{ijk}$$

and calculating the interaction sum of squares:

$$SS_{AB} = S_{\omega AB} - S_\Omega = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2$$

Then the SS_{AB}/σ^2 belongs to $\chi^2_{(I-1)(J-1)}$ with (I-1)(J-1) degrees of freedom. Thus, H_{AB} can be tested by calculating the p value $Pr(X \geq \frac{SS_{AB}}{\sigma^2} | H_{AB})$. If it is smaller than the significance level α, then H_{AB} can be rejected. Rejection of H_{AB} implies significant evidence for the presence of the interaction effect.

Appendix

```
### Exercise 1 ###

#lengths of datasets

n1 = 40  #I use small n's because of t-distribution
n2 = 45

# the standard deviation (same for both datasets)

sd = sqrt(3)

# the two different means

mean1 = 30
mean2 = 32

# the 2 data sets generated from 2 normal distributions

set1 = rnorm(n1,mean1,sd)
set2 = rnorm(n2,mean2,sd)

# Perform t-test to check for equality of the population means
#H_0:means are the same, H_1: means are different

t.test(set1,set2, var.equal=TRUE, conf.level=0.95)

#Preparing the data for ANOVA

#vector with all the observations

y=c(set1,set2)

#vector with factor containing 2 levels corresponding to the two data sets

effect = factor(c(rep(1,n1),rep(2,n2)))

#Data frame with y and effect

data = data.frame(y,effect)

#Create aov model

aovmodel=aov(y~effect,data=data)

#Obtain anova table
```

```
anova(aovmodel)

#improve results by using the real variance in the calculation of T, instead of
#the pooled variance

T=(mean1-mean2)/(sd*(sqrt(1/n1+1/n2)))

#calculate new p value

p_value_real = 2*pt(-abs(T),df=n1+n2-2)
```

```
### Exercise 2 ###


data = chickwts

#Create the boxplot

boxplot(weight~feed,
names = c("Casein","Horsebean","Linseed","Meatmeal","Soybean","Sunflower"),
        main = "Boxplots of the distribution of chick weights",
        xlab = "Feed Supplement",ylab = "Weight (grams)",data = data)

#obtain the number of observations and the levels

n = length(data$weight)
i = length(levels(data$feed))
# obtain the grand mean
gm = mean(data$weight)

#Obtain Incidence Matrix and Response Variable

response = data$weight
feed = data$feed
X = model.matrix(~ feed - 1) # -1 indicates that we remove the intercept


#check the rank of the matrix to verify that it is of full rank
I = qr(X)$rank #Obtain the rank of matrix

X_T  = t(X) #X_Transpose

#betahat
betaH = solve(X_T %*% X) %*% X_T %*% response #inverse obtained by solve function

#residuals

e_hat = response  - X %*% betaH

#Residuals Sum of Squares

SSE = t(response - X %*% betaH) %*% (response - X %*% betaH)
cat("Residuals Sum of Squares: ",SSE )
wgDf = n - I
meanSSE = SSE/wgDf

cat("Within Groups DF: ", wgDf,"Mean Value: ", meanSSE)
```

```
#obtain unbiased estimator of variance

var = SSE / (n-I)

#Between groups sum of squares
#manually
n1=10
n2=12
n3=14
n4=12
n5=11
n6=12

y1=mean(data$weight[1:10])
y2=mean(data$weight[11:22])
y3=mean(data$weight[23:36])
y4=mean(data$weight[37:48])
y5=mean(data$weight[49:59])
y6=mean(data$weight[60:71])


bgSS = n1*(y1-gm)^2+n2*(y2-gm)^2+n3*(y3-gm)^2+n4*(y4-gm)^2+n5*(y5-gm)^2+n6*(y6-gm)^2


cat("Between Groups Sum of Squares: ", bgSS)

bgDf = I - 1
cat("Between Groups DF: ", bgDf)
meanbg = bgSS/bgDf

#total Sum of Squares

TSS = bgSS + SSE

cat("Total Sum of Squares: ", TSS)
#F statistic

f = meanbg/meanSSE
f
cat("F  value: ", f)

#obtain p value to determine influence of factor feed supplement
pv = pf(f, bgDf, wgDf, lower.tail = FALSE, log.p = FALSE)
pv
cat("P value:", pv)

#same analysis using anova function
```

```
model = aov(weight~feed,data=data)
anova(model)

#check model assumptions

#normality
fit = X %*% betaH

plot(fit,e_hat,xlab="Fitted Values",ylab="Residuals",
main= "Plot of the residuals against the fitted values", pch=20, cex=1, col="blue")
abline(a=0, b=0, lty= 2)

qqnorm(e_hat, cex = 1, pch= 20)
qqline(e_hat,lty=3,col="blue")

shapiro.test(e_hat)

## Bartlett's test of homogeneity (homoscedasticity) of variances
bartlett.test(weight~feed,data=data) # H0: homoscedasticity
```

```
### Exercise 3 ###

data=read.table("hemoglobin.txt",header=TRUE)

#Two way ANOVA for interaction hypothesis
hemoglobin=as.vector(unlist(data))

data$rate = as.factor(data$rate)
data$method = as.factor(data$method)

#contrasts in order to obtain correct parameters
contrasts(data$rate)=contr.sum
contrasts(data$method)=contr.sum

#full model
model=aov(hemoglobin~rate+method+rate:method,data)
anova(model)

#obtain model parameters
model$coef
summary(model)
model.tables(model)

gm = mean(data$hemoglobin)

#check residuals
plot(model$fitted,model$res,xlab="Fitted Values",ylab="Residuals",
main= "Plot of the residuals against the fitted values", pch=20, cex=1, col="blue")
abline(a=0, b=0, lty= 2)

qqnorm(model$res, cex = 1, pch= 20)
qqline(model$res,lty=3,col="blue")
shapiro.test(model$res)

#HAB has not been rejected so I create additive model to obtain new parameters

additive = aov(hemoglobin~rate+method,data=data)
anova(additive)

#obtain new parameters

additive$coef
summary(additive)
model.tables(additive,'means')
model.tables(additive)

#check assumptions
plot(additive$fitted,additive$res,xlab="Fitted Values",ylab="Residuals",
```

```
main= "Plot of the residuals against the fitted values", pch=20, cex=1, col="blue")
abline(a=0, b=0, lty= 2)

qqnorm(additive$res, cex = 1, pch= 20)
qqline(additive$res,lty=3,col="blue")
shapiro.test(additive$res)

#one bartlett test for factor A, one for factor B, and one for the interaction

bartlett.test(hemoglobin~rate,data=data) #bartlett for factor rate
bartlett.test(hemoglobin~method,data=data)#bartlett for factor method
bartlett.test(split(data$hemoglobin,list(data$rate,data$method)))#bartlett for interaction
```

```
### Exercise 4 ###

library(car)
library(ggplot2)


data=read.table("fiber.txt",header=TRUE)

#one-way ANOVA to check the machine influence, without taking thickness into account

strength=as.vector(unlist(data))
thickness=as.vector(data$thickness)
#contrasts in order to obtain correct parameters
data$type = as.factor(data$type)
contrasts(data$type)=contr.sum

model1 = aov(strength~type,data = data)

anova(model1)
summary(model1)
model.tables(model1)
model1$coef


#check residuals and model assumptions

plot(model1$fitted,model1$res,xlab="Fitted Values",
ylab="Residuals",
main= "Plot of the residuals against the fitted values", pch=20, cex=1, col="blue")
abline(a=0, b=0, lty= 2)

qqnorm(model1$res, cex = 1, pch= 20)
qqline(model1$res,lty=3,col="blue")
shapiro.test(model1$res)# H0:normality

bartlett.test(strength~type,data=data)# H0: homoscedasticity


#ANCOVA with one factor and one covariate(thickness)

model2 = aov(strength~thickness+type,data=data)#factor second!
anova(model2)
model.tables(model2,'means')
model.tables(model2)

#check the p values using drop1
drop1(model2,test="F")
```

```
#check residuals for model2

plot(model2$fitted,model2$res,xlab="Fitted Values",ylab="Residuals",
main= "Plot of the residuals against the fitted values", pch=20, cex=1, col="blue")
abline(a=0, b=0, lty= 2)

qqnorm(model2$res, cex = 1, pch= 20)
qqline(model2$res,lty=3,col="blue")
shapiro.test(model2$res)# H0:normality


#Compare slopes visually
ggplot(data, aes(thickness, strength,color=type))+
geom_point()+
stat_smooth(method="lm",se=FALSE)+
ggtitle('Dependence of fiber strength on thickness for each machine')

#check with model
model4 = aov(strength~thickness*type,data=data)
anova(model4)
model.tables(model4,'means')
model.tables(model4)

#purpose of model5 is to obtain the correct parameter estimates
#-1 to obtain correct coefficients

model5 = aov(strength~-1+thickness + type,data=data)

anova(model5)
model.tables(model5,'means')
model.tables(model5)
model5$coef

avg_thick = mean(data$thickness)
#estimated fiber strength for fiber with average thickness
y1 = model5$coefficients[2] + model5$coefficients[1] * (avg_thick)
y1
y2 = model5$coefficients[3] + model5$coefficients[1] * (avg_thick)
y2
y3 = model5$coefficients[4] + model5$coefficients[1] * (avg_thick)
y3
```