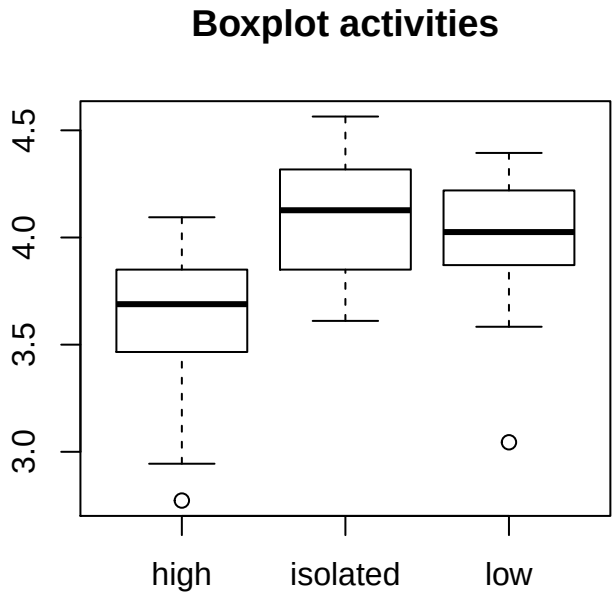
Assignment 3

Maud van den Berg, Mick IJzer, Tirza IJpma

March 15, 2020

Exercise 1

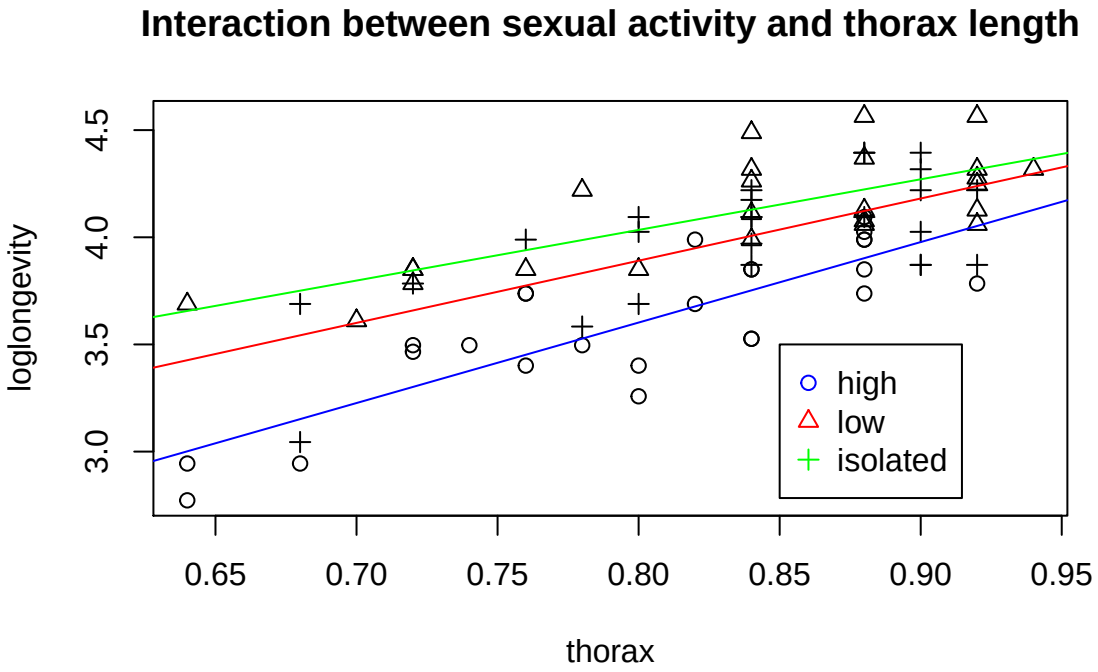
a)
To investigate whether sexual activity influences longevity an one-way ANOVA was carried out, where the activity groups are the independent variable and the log of the longevity was the dependent variable. The group averages are graphically shown in the boxplot below. Results show that activity has a significant effect $F(2, 72) = 19.42$, $p < .000$. on the longevity. Post-hoc tests (TukeyHSD) indicated that the fruitflies in the high sexual activity group lived significantly shorter ($M = 3.60$) than the fruitflies in the isolated ($M = 4.12$, $p < .000$) and low ($M = 4.00$, $p < .000$) activity groups. No group differences were found between the low activity and isolated groups, $p = .359$.



b)
Thorax length will be added as an explanatory variable in the previous ANOVA, thus making it an ANCOVA. Both thorax length, $F(1, 71) = 132.2$, $p < .000$, and the activity groups, $F(2, 71) = 25.7$, $p < .000$, have a significant effect on the log of the longevity. The effect of sexual activity is the same as in the previous question. The fruitflies in the high condition live significantly longer than the fruitflies in the other two

conditions. The estimated lifespan for a fruitfly with an average thorax for each group is 3.67 (high), 4.09 (isolated), and 3.96 (low).

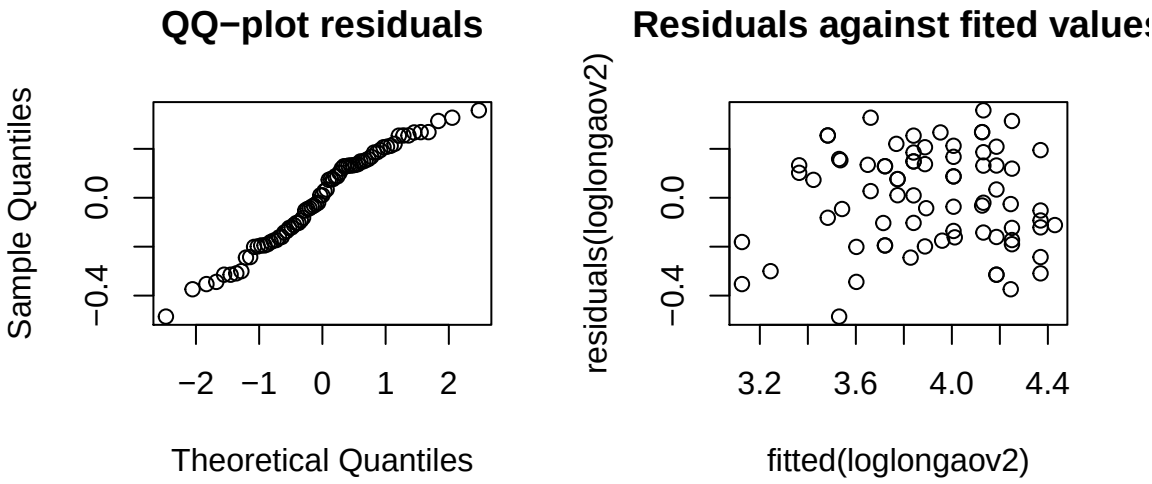
c)
To investigate the influence of thorax length on the longevity of the fruitflies, the previous ANCOVA was carried out once more. Only the interaction between thorax length and the sexual activity group was added as an explanatory variable. The results of the analysis show that the interaction is not significant, $F(2, 69) = 1.93$, $p = 0.154$. The graph below shows the estimates for each group depending on the thorax length. Since the interaction is not significant it can be concluded that the plotted lines are more or less parallel. Because of the insigificance of the interaction, the interaction term was removed from the model. This results in an exact copy of the ANCOVA in question b. Thorax length has a significant and positive relation with the log of longevity. This means that fruitflies with a longer thorax length live longer than fruitflies with a shorter thorax length.



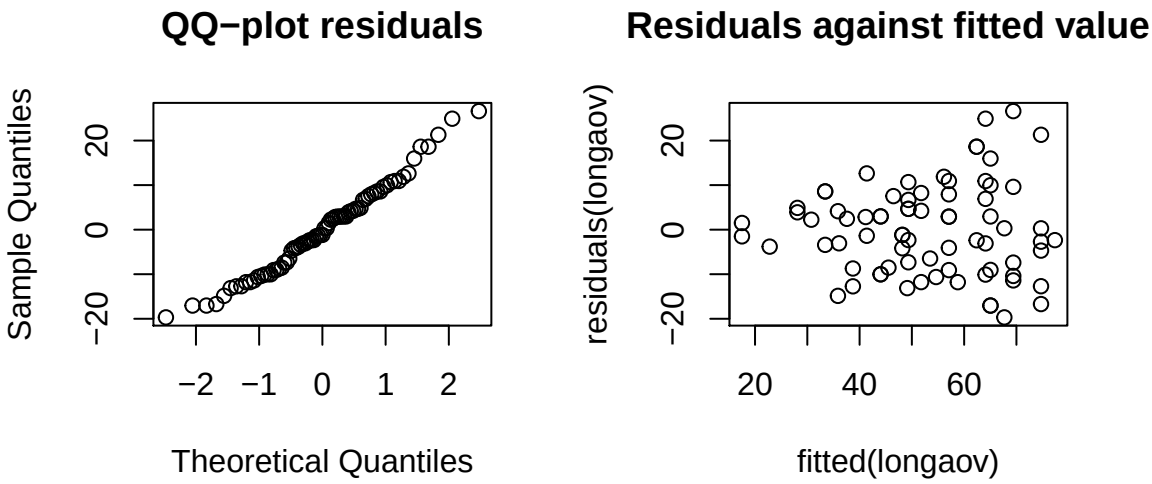
d)
To determine which of the two analysis is preferred the r-squared values are computed. The ANCOVA with thorax included in the model, has a explained variance of 70.9%. In contrast, the model without thorax has an explained variance of 33.2%. The addition of only thorax length almost doubles the explained variance of the model. Therefore the model with thorax length is preferred. In principle neither of the analysis is wrong. Since both are carried out according to the AN(C)OVA design. However, if beforehand it is known that thorax length has an influence on the lifespan of fruitflies it makes sense to incorporate that explanatory variable. Especially, to check whether the average thorax length is equal amongst groups. Otherwise the likelihood of a false positive result becomes larger.

e)
Normality and heteroscedasticity are checked by inspecting the normality of residuals (QQ-plot) and the relation between the estimates and residuals (fitted vs. residuals plot). These are computed for the ANCOVA with thorax length and the activity group as independent variables. The residuals look normally distributed. This is also confirmed by a Shapiro-Wilk normality test, $W = 0.97$, $p = .057$. The largest part of the range

of fitted values show no pattern with respect to the residuals. However, the lowest fitted values all seem to be underestimating the actual values. This is not a big problem since the lowest fitted values also have the lowest probability of occuring.



f)
An ANCOVA was carried out with thorax length and the activity group as explanatory variables. The dependent variable was the untransformed longevity measure. The effects found in the analysis are very similar to the results of the previous analysis. Next, the same QQ-plot and residuals plot are computed. The residuals (QQ-plot) are normally distributed. This is supported by a Sahpiro-Wilk test, W = 0.98, p = .318. However, the plot of the residuals against the fitted values shows a pattern of heteroscedasticity. The variance in residuals becomes increasingly larger as the estimates become larger. This means that the predictions from the model become more unreliable when the prediction becomes larger. Therefore, it was wise to use the logarithm of longevity as response variable instead of the original longevity measure.



Exercise 2

a)
The data for the personalized system of instruction is studied by looking at (1) the histogram of GPA, (2) the distribution of GPA for the groups that received psi or not, and (3) the distribution of GPA for the catagories pass or fail. Furthermore, a crosstab (table 1) is computed for the two factors. GPA seems to be normally distributed. The GPA of two groups of students is similar, although the group that received psi has a slightly higher GPA. The third plot indicates that the students who passed the assignment have a higher GPA than those who did not. The crosstabs show that there seems to be a dependence between psi and passing the assignment. This is because the majority of the students who did not receive psi did not pass the assignment, while on the other hand about half of the students passed the assignment when they did receive psi.

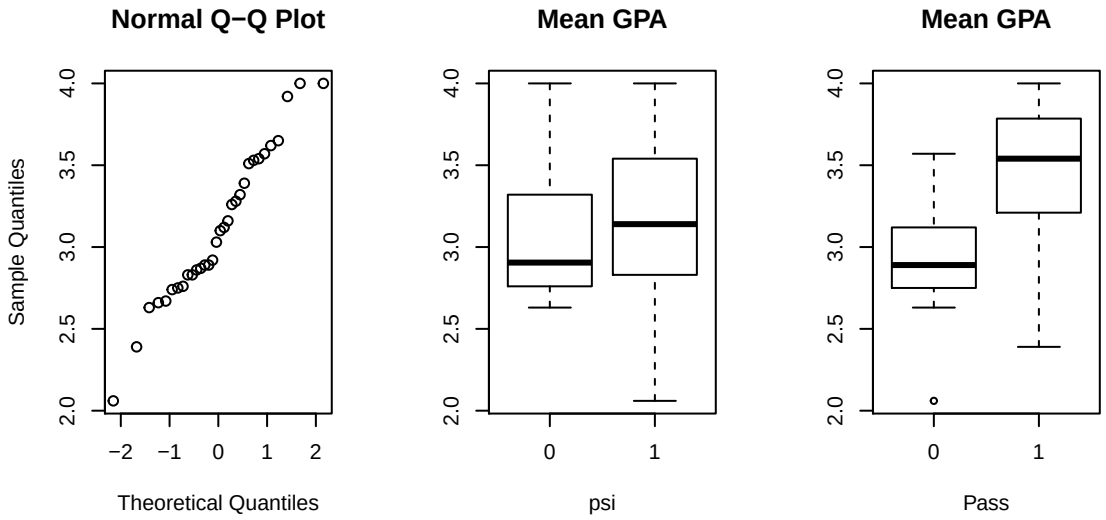


Table 1: Observed frequencies

	no psi	psi
fail	15	6
pass	3	8

b)
To investigate the effect of psi, a logistic regression is carried out with GPA and psi as predictors and pass/fail as outcome. Both GPA and psi have a significant effect. A higher GPA leads to a higher probability of passing the assignment, Z = 2.51, p = .012. Furthermore, receiving psi also increases the likelihood of passing the assignment, Z = -2.45, p = .024. It can be concluded that psi works.

c)
The probability of passing for a student with gpa 3 and no psi is 8.2%. For a student who did receive psi the probability is 48.2%. These probabilities are calculated by using the following formula, where the exponent is the result from the logistic regression:

$$\frac{1}{1+e^{-(-10.43+3.064*GPA+1.17*psi)}}$$

d)
Receiving psi increases the odds of passing the assignment by a factor $e^{2*1.17} = 10.36$. This is not dependent on gpa, since gpa and psi are independent of eachother.

e)
Table 2 shows the frequency of students who did or didn't receive psi in combination with whether they passed or not. The 15 are the number of students who did not receive psi and did not improve (or failed the assignment). The 8 are the number of students who did receive psi, but did not improve. A Fisher's exact test for 2x2-tables is carried out to assert the dependence of psi and the outcome of the assignment. The results reveal a significant dependence, p = 0.027. This means that the observed frequencies differ significantly from the expected frequenceis (table 3). When the contribution of each cell to the test statistic is inspected it becomes clear that the number of students that passed the assignment when receiving psi is much larger than expected. Thus it can be concluded that psi works.

Table 2: Observed frequencies

	no psi	psi
pass	3	8
fail	15	6

Table 3: Expected frequencies

	no psi	psi
pass	6	5
fail	12	9

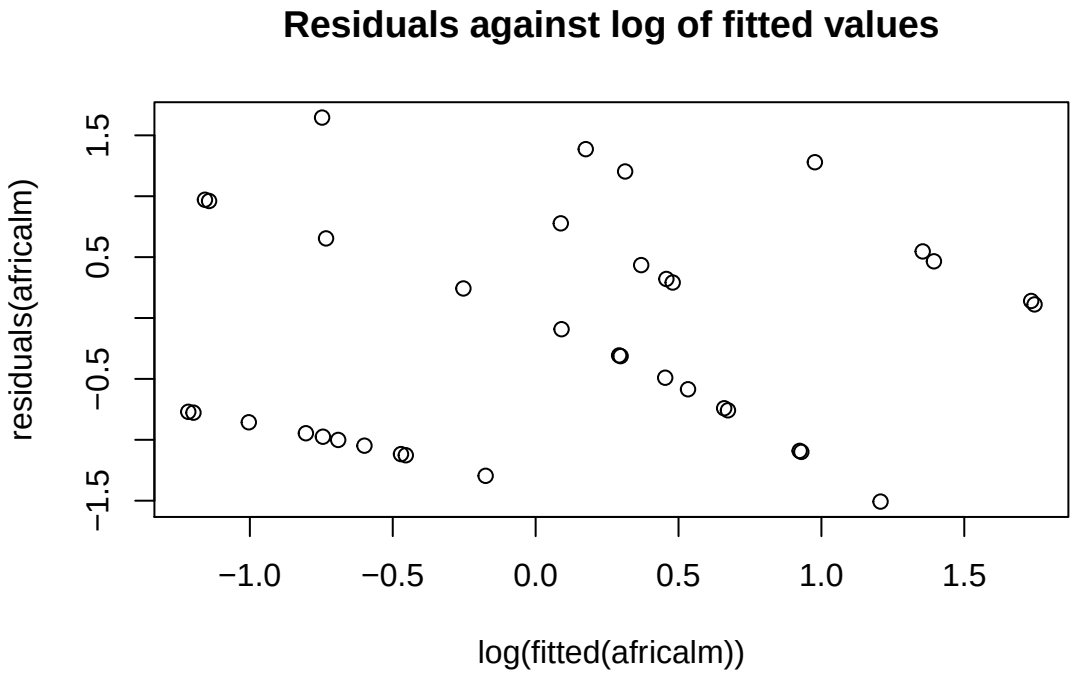
f)
Fisher's exact test for 2x2-tables is the wrong test to use given the experimental design. Contingency tables could be used when (a) a random sample is drawn from a population, (b) when a random sample is drawn for each level of the first factor, or (c) when a random sample is drawn for each level of the second factor. The current design doesn't include a random sample, since all students either do or do not receive psi. Therefore, a logistic regression is the correct analysis.

g)
With a logistic regression there is a lot of flexibility in choosing the predictors. In a Fisher's exact test for 2x2-tables the p-value can be computed exactly.

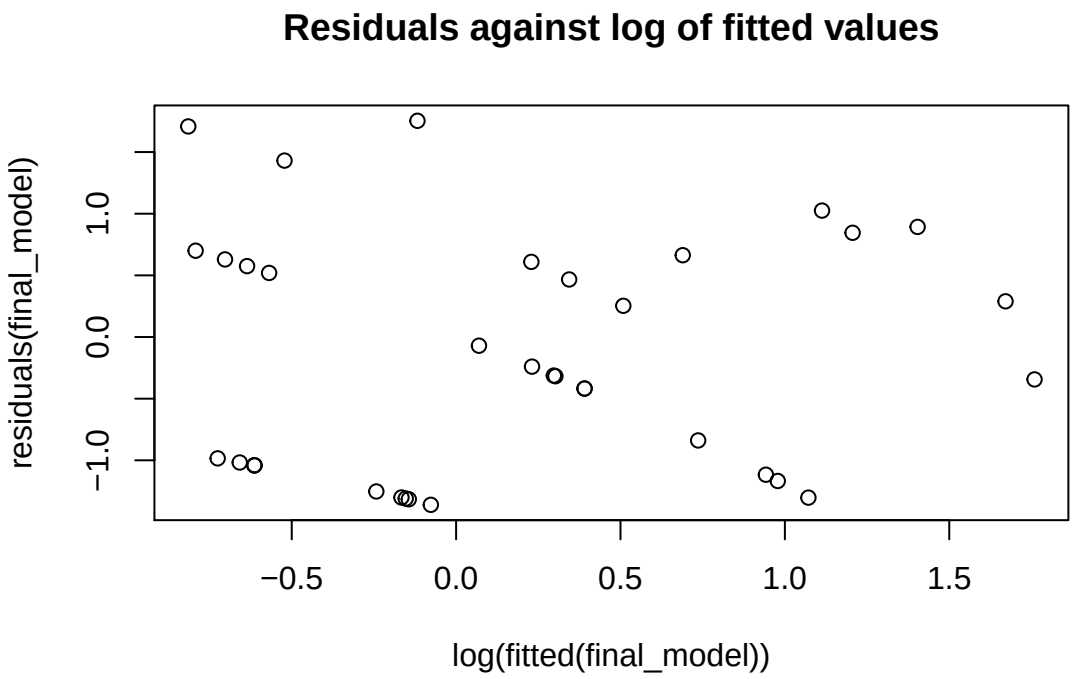
Exercise 3

a)
A poisson regression is performed, with the number of succesful military coups (miltcoup) as dependent variable. The other variables in the dataset are used as explanatory variables. However, it must be noted that the dependent variable does not seem to stem from a poisson distribution. This is because the mean (=1.58) is not similar to the variance (=3.11). The distribution of the used variables as well as the relation between independent variables are visually inspected. No indications for multicolinearity have been found. A few outliers are present for the variables parties, population size, and size. These outliers were not removed due influence the removal would have on the sample size. The rersults of the poisson regression show that the number of years the country was ruled by a military oligarchy (oligarchy), Z = 2.05, p = .040, the number of legal political parties (parties), z = 2.80, p = .005, and the political liberalization (pollib), have a significant effect on the number of military coups. Oligarchy and parties have a positive relation with miltcoup. If the political liberalization is that there are no civil rights for political expression the estimated number of

succesful military coups is also larger, Z = 2.17, p = .030. Furthermore, if there are full civil rights the estimated miltcoup is lower, Z = -2.69, p = .007.



b)
Since the number of independent variables is 8 and most of them are not significant the step down approach was used to reduce the number of predictors. First, the model with all predictors was inspected. The variable "numelec" was removed, because it was the least significant. This was repeated several times. In the end the model consisted of "oligarchy", "pollib", and "parties", which were the only significant predictors in the complete model aswell. The relation with the dependent variable remains the same as explained in question a. The residuals were plotted against the logarithm of the fitted values. No pattern was found in this plot. Lastly, the residuals of the model were plotted against all of the predictors, to investigate if any patterns emerge. In none of the plots a pattern was found that would indicate the need for a transformation (for the included variables) or inclusion (for the excluded variables).


Residuals against log of fitted values
residuals(final_model)
−1.0
0.0
1.0
−0.5
0.0
0.5
1.0
1.5
log(fitted(final_model))