Assignment 3

Experimental Design & Data Analysis

VU Amsterdam

S. Verhezen & L. Heek

March 22, 2020

Exercise 1: Fruit flies

a) The following code was used to read the data, add a column with the log function of the measured longevities, and perform an anova analysis without taking into account thorax length:

```
# read data
df = read.table("data/fruitflies.txt", header=TRUE)
df$log <- c(log(df$longevity))

# anova
a = lm(log~activity,data=df)
anova(a)
summary(a)
```

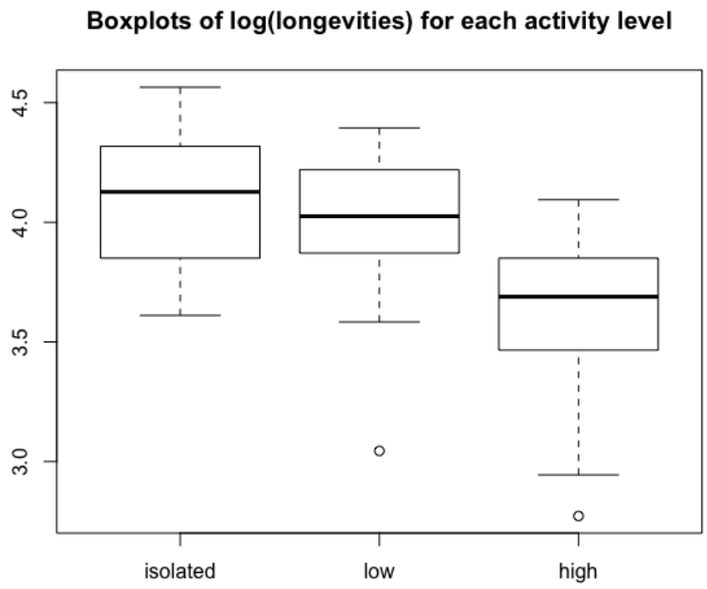

Figure 1: Boxplots of loglongevity for all three activity levels.

This analysis yielded the following results:

```
Analysis of Variance Table

Response: log
          Df Sum Sq Mean Sq F value Pr(>F)
activity  2 3.6665 1.8333 19.421 1.798e-07 ***
Residuals 72 6.7966 0.0944
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)     3.60212   0.06145 58.621 < 2e-16 ***
activityisolated 0.51722  0.08690  5.952 8.82e-08 ***
activitylow     0.39771   0.08690  4.577 1.93e-05 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```

From this output we can conclude that sexual activity does influence longevity: $F(2) = 19.421$ with $p = 1.798e-7$ is significant below the $\alpha = 0.01$ level. We accept the alternative hypothesis, that at least two of the groups with different activity levels have significantly different longevities. From the estimated longevities we can see that the high activity group lives longest, their estimated $log(longevity) = 3.6026.145e-2$. Followed by the isolated group with an estimated $log(longevity) = 5.172e-18.69e-2$. The low activity group has a lower log(longevity) than the isolated group, which indicates that log(longevity) is not a simple function of the amount of activity. However, since we don't know which group means are significantly different, it might be the case that there is no significant difference between the low activity and the isolated group, and that sexual activity only influences longevity from above a certain level.

b) To include thorax length as an explanatory variable, together with log(longevity), the following code was used:

```
# ancova
c = lm(log~thorax+activity,data=df)
anova(c)
summary(c)
```

This analysis yielded the following results:

```
Analysis of Variance Table

Response: log
          Df Sum Sq Mean Sq F value Pr(>F)
thorax     1 5.4322 5.4322 132.175 <2e-16 ***
activity   2 2.1129 1.0565 25.705 4e-09 ***
Residuals 71 2.9180 0.0411
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.21893   0.24865  4.902 5.79e-06 ***
thorax           2.97899   0.30665  9.715 1.14e-14 ***
activityisolated 0.40998   0.05839  7.021 1.07e-09 ***
activitylow      0.28570   0.05849  4.885 6.18e-06 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```

From the ancova analysis we see that both the effect of sexual activity and of thorax length have an significant effect for $\alpha = 0.01$ on longevity, with $F(2) = 25.705, p = 4e-09$ and $F(1) = 132.175, p < 2e-16$ respectively. Estimated log(longevity) for a fly with average thorax length is highest for flies with a high activity level, but isolated flies have higher estimated longevity than those with low activity (as can be seen in the coefficients table above). Therefore, we can not conclude whether sexual activity increases or decreases longevity.

c) As we can see from the estimated coëfficients in question b, the thorax length parameter is estimated at a value of $2.983.07e-1$. This indicates that thorax length has a positive influence on longevity: fruitflies with a larger thorax, live longer. To investigate whether this effect is the same for all three activity groups, a linear model was setup for each of the three groups separately using the following code:

```
activities = c("isolated", "low", "high")
for (activity in activities){
summary(lm(longevity~thorax,data=subset(df,activity==activity)))
}
```

Which resulted in the following estimates:

```
# isolated
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1443     0.3564   6.016 3.91e-06 ***
thorax       2.3625     0.4243   5.568 1.15e-05 ***
```

```
# low activity
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.5695     0.5479   2.865 0.008763 **
thorax       2.9016     0.6519   4.451 0.000183 ***

# high activity
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5978     0.4122   1.450    0.16
thorax       3.7554     0.5129   7.322 1.89e-07 ***
```

For all groups the effect of thorax length is significant below $\alpha = 0.01$. However, we see that the estimated coëfficient for thorax length increases with the amount of activity. That is, the effect of thorax length is larger for low activity flies than for isolated flies, and even larger for high activity flies than for low activity/isolated flies.

d) The preferred analysis, depends on the factors of interest of the researchers. In principle, you always want to include factors in your analysis that have a significant effect on the outcome variable. From this point of view, the ancova analysis in exercise 1b would be preferred compared to the anova analysis in ex. 1a. However, thorax length does not seem like a factor that can be changed or influenced by the researchers. In other words, we can not (simply) increase thorax length to increase longevity. Therefore, it could advocated that accounting for the effects of thorax length by making sure the sample of fruitflies is normally distributed for this variable, is sufficient. Then an anova analysis would be appropriate and sufficient to study the effects of sexual activity on longevity.

e) Normality and heteroscedasticity plots were generated using the following code:

```
plot(lm(log~thorax+activity,data=df))
```

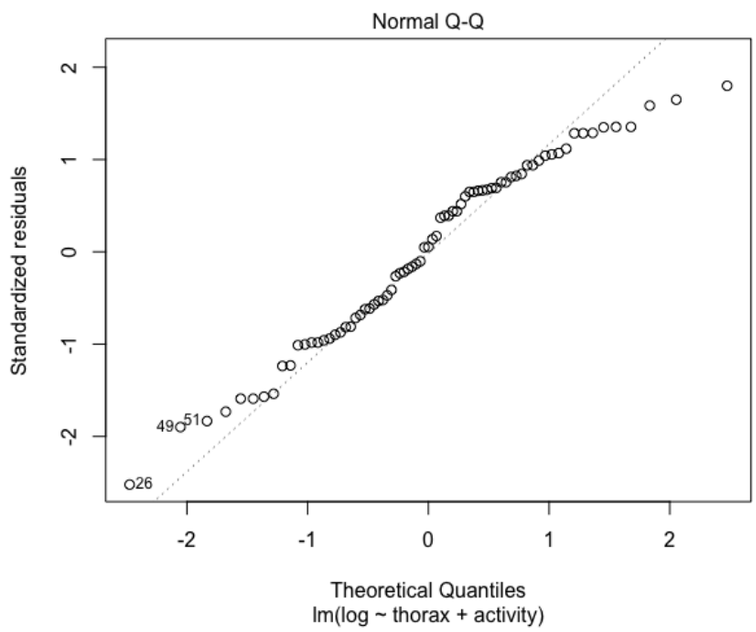

Figure 2: Normal QQ-plot of the linear model of log(longevity) with sexual activity and thorax length as explanatory variables.

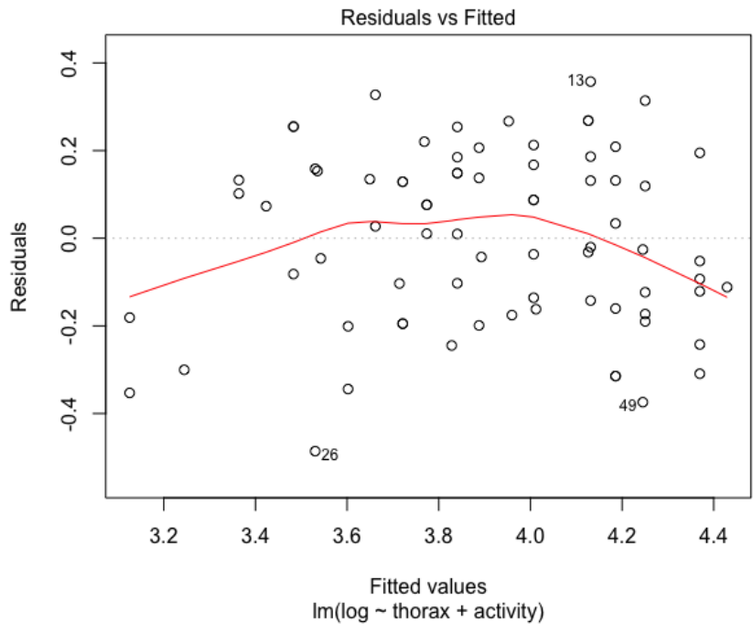

Figure 3: Residuals vs fitted plot of the linear model of log(longevity) with sexual activity and thorax length as explanatory variables.

The QQ-plot does not look completely satisfactory, the data points do not follow a straight line. However it is not a worrying outcome, this plot indicates that the data follows an approximately normal distribution. The residuals plot indicates heteroscedasticity.

f) The following code was used to perform an ancova analysis on the longevity in days, with activity level and thorax length as explanatory variables:

```
# ancova
d = lm(longevity~thorax+activity,data=df)
anova(d)
summary(d)
plot(d)
```

This analysis yielded the following outcomes:

```
Analysis of Variance Table

Response: longevity
          Df  Sum Sq Mean Sq F value  Pr(>F)
thorax     1 10959.3 10959.3 101.409 2.557e-15 ***
activity   2  4966.7 2483.4 22.979 2.016e-08 ***
Residuals 71 7673.0  108.1
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -67.375    12.750 -5.284 1.33e-06 ***
thorax           132.618    15.725  8.434 2.62e-12 ***
activityisolated 20.066     2.994  6.701 4.13e-09 ***
activitylow      13.054     2.999  4.352 4.43e-05 ***
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```

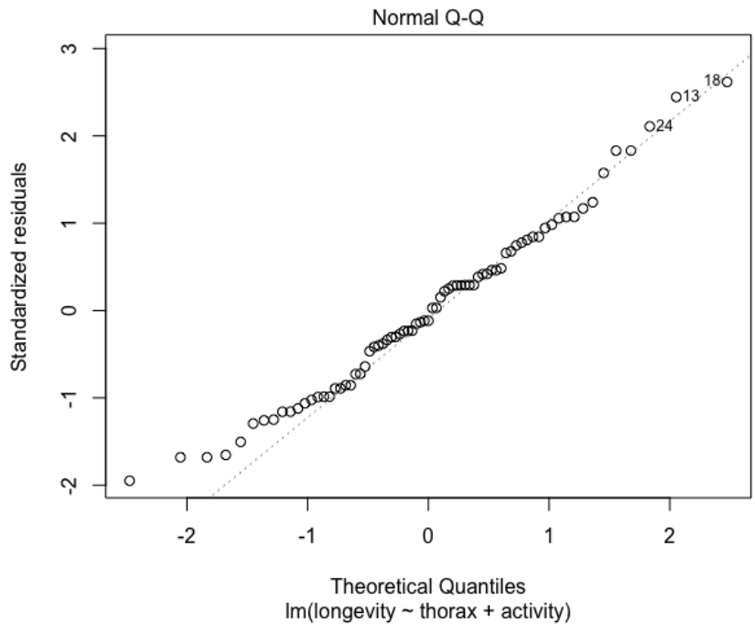

Figure 4: Normal QQ-plot of the linear model of longevity in days with sexual activity and thorax length as explanatory variables.

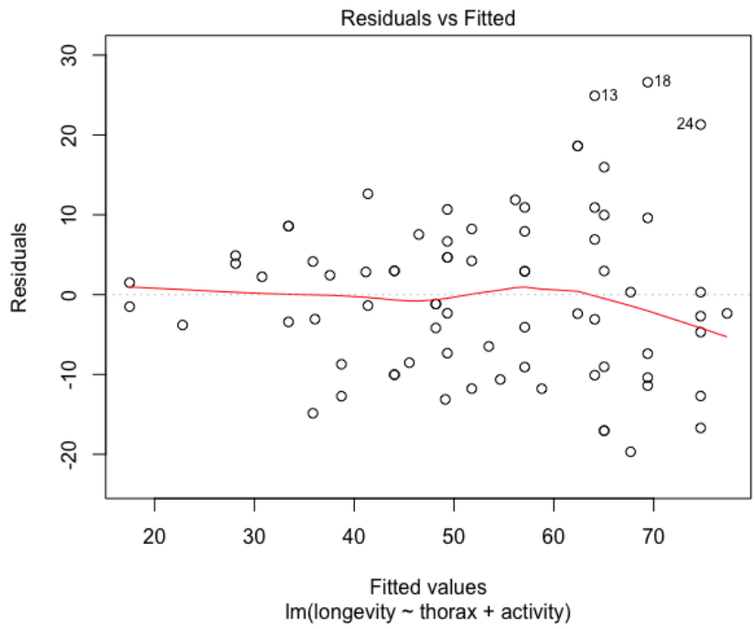

Figure 5: Residuals vs fitted plot of the linear model of longevity in days with sexual activity and thorax length as explanatory variables.

The QQ-plot of longevity in days satisfies the assumption of normality more than that of the log function of longevity. Furthermore, the residuals plot does not indicate homoscedasticity. The ancova analysis reaches significance below the $\alpha = 0.01$ level for both thorax length and activity level. However, from the estimates we can see that the estimated longevity for fruitflies in the high activity group is negative. For this data, where the outcome variable is longevity, we would not expect a negative estimate. This indicates that it was wise to use the logarithm as a response variable, because the resulting data was more suitable for the linear model on which the an(c)ova analysis is performed.

Exercise 2: Personalized system of instruction (PSI)

The effect of PSI is tested on 32 randomised students. The outcome Y is an assignment after the teaching period that they could pass (1) or not (0). Whether students received PSI (1) or not (0) is the factor explanatory variable. In addition, a numerical explanatory variable $(X_1, ..., X_p)$ is included, which is the average grade (GPA) of the students.

a) First, the data is summarised. The amount of students in each group is shown in the contingency table. Out of all 14 students who received PSI, eight passed the assignment (57.1%). From the 18 students who were taught with the standard method, three passed the assignment (16.7%).

	No PSI	PSI	Total
Pass	3	8	11
No pass	15	6	21
Total	18	14	32

The distributions of GPA scores of students in both teaching method groups are shown in Figure 6. The mean GPA for the standard teaching method is 3.10 (±0.42). For PSI, the mean GPA is 3.14 (±0.53).

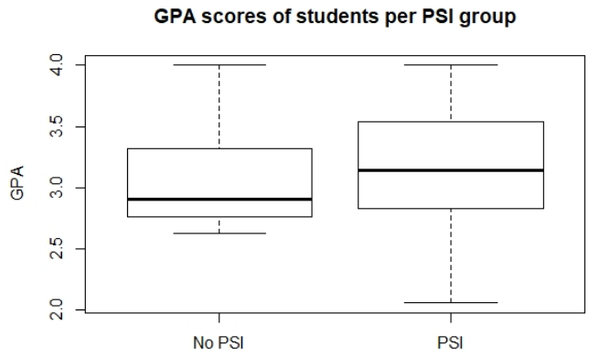

Figure 6: Boxplot of GPA scores of students who received the standard teaching method (no PSI), and the PSI method.

A histogram and QQ-plot of the numerical explanatory variable (GPA) are shown in Figure 7. The QQ-plot shows a linear pattern with some slight curves, indicating normal distribution of the GPA scores. A Shapiro-Wilk test confirms normality of the data (W = 0.96974, p-value = 0.4921 $>$ 0.05; accept H_0 that the data is normally distributed).

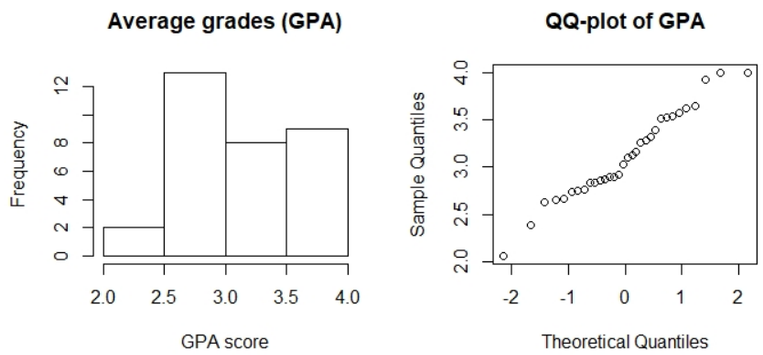

Figure 7: Histogram and QQ-plot of average grades (GPA) of students.

The amount of students to pass the test for different GPA and PSI groups is shown in Figure 8. For a better overview, the GPA scores were rounded to half a digit to generate the tables and plot (for Figure 8 only). In the upper table on the left, the total amount of passes in each group is shown. We see that out of all students who participated with a rounded GPA of 4, three passed the assignment. Two of them did not receive PSI and one did. In the lower table, the amount of passes as a ratio of the total amount of students are provided. The values of 1.0 for students with GPA of 4 denote that all participants in this group passed. In addition, the boxplot shows the ratio of students who passed over the total amount of students in each GPA group, disregarding

the teaching method. The code to obtain the data is provided as well. We see that one students with a GPA of 2 participated, received PSI, and did not pass the test.

```
> tot=xtabs(~gpa_50+psi,data=psi)
> tot
      psi
gpa_50 0 1
   2   0 1
   2.5 3 2
   3   8 5
   3.5 5 5
   4   2 1
> tot.c=xtabs(passed~gpa_50+psi,data=psi)
> round (tot.c/tot,2)
      psi
gpa_50   0   1
   2       0.0
   2.5 0.0 0.5
   3   0.0 0.4
   3.5 0.2 0.8
   4   1.0 1.0
```

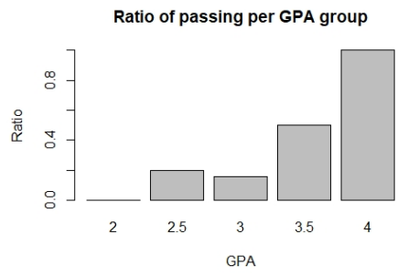

Figure 8: Tables and bar chart of individuals to pass test for combinations of GPA and PSI.

b) A logistic regression model is fitted with both explanatory variables using the function *glm*. *GPA* was taken as numeric, *PSI* and the outcome variable *passed* were taken as a factor. The following code is used.

```
# factorise PSI and passed
psi_data$passed = as.factor(psi_data$passed); psi_data$psi = as.factor(psi_data$psi)

# fit logistic regression model to data
psiglm=glm(passed~gpa+psi,data=psi,family=binomial)
summary(psiglm)
```

This results in the following coefficients (some output is deleted).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.602     4.213  -2.754  0.00589 **
gpa           3.063     1.223   2.505  0.01224 *
psi           2.338     1.041   2.246  0.02470 *
```

The results show that both explanatory variables (GPA and PSI) have a significant effect on the outcome (passing the test). GPA shows a p-value of 0.0112 (< 0.05; reject H0 that variable has no effect). PSI shows a p-value of 0.0247 (< 0.05; reject H0). PSI has an effect on the students' performance and its estimate has a positive sign. Positive signs of the parameter estimates mean that higher values of the variable result in higher probability of passing. Thus, PSI works.

c) We want to estimate the probability of a student with a GPA of 3 to pass the test, with and without receiving PSI. Here, we have a single factor with $I = 2$ levels (PSI) and a single numerical explanatory variable (GPA). The logistic regression model then assumes that output Y_{in} satisties

$$P(Y_{in} = 1) = \Psi(\mu + \alpha_i + \beta X_{in}), \qquad i = 1, 2; \qquad n = 1, ..., N$$

where Y_{in} is the outcome of a unit measured at level i of factor PSI that has explanatory variable X_{in}. This function is satisfied for $\mathrm{x} \mapsto \Psi(x) = 1/(1+e^{-x})$ (the logistic function). From subquestion 2.b), we obtained the following values.

Coeff.	Formula	Value
Intercept	μ	-11.602
gpa	β	3.063
psi	α	2.338

Inserting this in the formula for a GPA of 3 and PSI of 1 gives

$$P(Y_{in} = 1) = \Psi(-11.602 + 2.338 + 3.063 * 3) = \Psi(-0.075)$$

and by using the logistic function we obtain

$$P(Y_{in} = 1) = 1/(1 + e^{-(-0.075)}) = 0.48$$

The same is done for not receiving PSI (0). Here, α is not included in the formula. This renders

$$P(Y_{in} = 0) = \Psi(-11.602 + 3.063 * 3) = \Psi(-2.413)$$

$$P(Y_{in} = 0) = 1/(1 + e^{-(-2.413)}) = 0.08$$

So, a student with a GPA of 3 that receives PSI has a probability of 0.48 (48%) to pass the assignment. When not receiving PSI, the probability of passing is only 0.08 (8%).

These probabilities can also be computed with the *predict* function in R. The same results are obtained.

```
wo_psi = data.frame(psi=0,gpa=3)
w_psi = data.frame(psi=1,gpa=3)
predict(psiglm,wo_psi,type="response") # 0.08230274
predict(psiglm,w_psi,type="response") # 0.4815864
```

d) The relative change in odds for passing the assignment is estimated for (arbitrary) students that receive PSI, compared to the standard teaching method.

The relative change in odds equals $\frac{o(PSI=1)}{o(PSI=0)}$. The odds are obtained by taking the exponent of the coefficient. For the factor variable with I levels, we obtain $\frac{e^{\alpha_{i'}}}{e^{\alpha_i}} = e^{\alpha_{i'}-\alpha_i}$. In other words, given that a change of ν in the linear predictor $\mu + \alpha_i + \beta X_{in}$ multiplies the odds of passing by e^{ν}, a change from level i to level i' multiplies the odds by $e^{\alpha_{i'}-\alpha_i}$. For the binary factor variable PSI, we obtain a relative change in odds of $e^{2.34} = 10.36$. This indicates that students who receive PSI are 10.36 times more likely to pass the assignment than to fail it.

The variable GPA is not included in the calculation. However, it affects the probability of passing and the coefficient of PSI. Therefore, the value is dependent on GPA.

e) We consider the contingency table, where total values are added for the overview. Here, the number 15 and 6 are the amount of students that did not show improvement when they did and did not receive PSI, respectively.

	No PSI	PSI	Total
Improvement	3	8	11
No improvement	15	6	21
Total	18	14	32

To replicate the 2x2 contingency table the following code is used. Subsequently, the table is analysed using the Fisher's test.

```
x=matrix(c(3,15,8,6),2,2)
fisher.test(x)
```

The following output is obtained for the Fisher's test.

```
   Fisher's Exact Test for Count Data

data:  x
p-value = 0.0265
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.02016297 0.95505763
sample estimates:
odds ratio
 0.1605805
```

The test compares the success probabilities of not receiving PSI (p_1) and receiving PSI (p_2). The null hypothesis that they do not differ ($H_0 : p_1 = p_2$) is rejected (p-value $= 0.0265 < 0.5$). Thus, the probability of success is not equal for both sequences. Students who received PSI show a significantly higher probability of improvement (passing), compared to those who did not receive PSI.

f) This approach does not include the effect of the GPA of the students. Considering this effect was shown to be significant (see 2.b), it is better to include them. However, the students are considered as independent samples in this test. It is not wrong to study the effect of PSI on students using this method.

g) The Fisher's test can be used for 2x2 contingency tables, so only one explanatory variable can be taken into account. On the other hand, the logistic regression can be used with multiple explanatory variables.

An advantage of the Fisher's test is that the significance of the deviation from the null hypothesis (p-value) can is calculated exactly. Logistic regression relies on approximation and becomes exact in the limit, which is reached with the sample size growing to infinity.

Exercise 3: Military cops in Africa

a) A Poisson regression was performed on the full data set with the number of successful military coups from independence to 1989 (*miltcoup*). The variable *pollib* was factorised as it should not be treated as numerical. The code below was used.

```
# load data
africa = read.table("data/africa.txt", header=TRUE)

# factorise pollib
africa$pollib = as.factor(africa$pollib)

# perform Poisson regression
africaglm = glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,
          family=poisson,data=africa)
summary(africaglm)
```

The following coefficients were obtained using the *summary* function (some output was deleted). We see that variables *oligarchy*, *pollib* of 2 (full civil rights) and *parties* have a p-value smaller than 0.05. This means that there is a significant deviation from the null hypothesis that the variables have no effect on the outcome. Therefore, the test shows that the stated variables have an effect on *miltcoup*.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.2334274 0.9976112 -0.234 0.81500
oligarchy    0.0725658 0.0353457  2.053  0.04007 *
pollib1     -1.1032439 0.6558114 -1.682 0.09252 .
pollib2     -1.6903057 0.6766503 -2.498 0.01249 *
parties      0.0312212 0.0111663  2.796  0.00517 **
pctvote      0.0154413 0.0101027  1.528  0.12641
popn         0.0109586 0.0071490  1.533  0.12531
size        -0.0002651 0.0002690 -0.985 0.32444
numelec     -0.0296185 0.0696248 -0.425 0.67054
numregim     0.2109432 0.2339330  0.902  0.36720
```

b) Using the step-down method, the number of explanatory variables is reduced. This is done by consecutively selecting the variable with highest p-value and removing it until only significant variables remain.

The first step is performed by looking at the output of the full model in Section 3.a. We select variable *numelec* with the highest p-value of 0.806 and run the new model as follows. The p-values of the coefficients are evaluated.

```
> africaglm2 = glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numregim,
              family=poisson,data=africa)
> summary(africaglm2)

[ Some output deleted ]

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4577458 0.8602345 -0.532 0.59464
oligarchy    0.0812015 0.0288154  2.818  0.00483 **
pollib1     -0.9642976 0.5620939 -1.716 0.08625 .
pollib2     -1.5149509 0.5269441 -2.875 0.00404 **
parties      0.0293409 0.0103101  2.846  0.00443 **
pctvote      0.0139115 0.0094654  1.470  0.14164
popn         0.0099592 0.0067249  1.481  0.13862
size        -0.0002688 0.0002687 -1.000 0.31710
numregim     0.1804415 0.2241166  0.805  0.42075
```

Variable *numregim* is removed and the new model is evaluated.

```
> africaglm3 = glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size,
              family=poisson,data=africa)
> summary(africaglm3)

[ Some output deleted ]

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) 0.0419757 0.5774100  0.073 0.942048
oligarchy   0.0894951 0.0270440  3.309 0.000936 ***
pollib1    -0.9673253 0.5605601 -1.726 0.084412 .
pollib2    -1.5321126 0.5232779 -2.928 0.003412 **
parties     0.0288170 0.0102173  2.820 0.004796 **
pctvote     0.0149216 0.0093762  1.591 0.111513
popn        0.0071647 0.0056842  1.260 0.207510
size       -0.0002579 0.0002662 -0.969 0.332621
```

Variable *size* is removed.

```
> africaglm4 = glm(miltcoup~oligarchy+pollib+parties+pctvote+popn,
                   family=poisson,data=africa)
> summary(africaglm4)

[ Some output deleted ]

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.231435 0.528887 -0.438 0.66168
oligarchy   0.083468  0.025829  3.232  0.00123 **
pollib1    -0.683589  0.495822 -1.379  0.16799
pollib2    -1.320568  0.490268 -2.694  0.00707 **
parties     0.029770  0.010310  2.887  0.00388 **
pctvote     0.013925  0.009371  1.486  0.13728
popn        0.005659  0.005483  1.032  0.30204
```

Variable *popn* is removed.

```
> africaglm5 = glm(miltcoup~oligarchy+pollib+parties+pctvote,family=poisson,data=africa)
> summary(africaglm5)

[ Some output deleted ]

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.116499 0.513751 -0.227 0.82061
oligarchy   0.094712  0.023184  4.085 4.4e-05 ***
pollib1    -0.620756  0.487526 -1.273 0.20292
pollib2    -1.310374  0.489017 -2.680 0.00737 **
parties     0.025745  0.009552  2.695 0.00704 **
pctvote     0.012057  0.009072  1.329 0.18383
```

The variable with highest p-value is *pollib1*. However, *pollib* is a factorial variable and we cannot remove one factor. As *pollib2* is significant, the variable is kept and variable *pctvote* is removed in stead.

```
> africaglm6 = glm(miltcoup~oligarchy+pollib+parties,family=poisson,data=africa)
> summary(africaglm6)

[ Some output deleted ]

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.207981  0.445679  0.467  0.6407
oligarchy   0.091466  0.022563  4.054 5.04e-05 ***
pollib1    -0.495414  0.475645 -1.042  0.2976
pollib2    -1.112086  0.459492 -2.420  0.0155 *
parties     0.022358  0.009098  2.458  0.0140 *
```

The new model contains only variables that have a significant effect (p-value < 0.05) on the response variable *miltcoup*. Factor variable *pollib* is not significant for a factor of 1, but the effect of factor 2 is significant. Since the variable is considered as a whole, it should remain in the reduced model. The remaining variables are *oligarchy*, *pollib* and *parties*.