Assignment 3

Ella Smorenburg (2618639), Yoes Ywema (271544), Roel Rotteveel (271547)

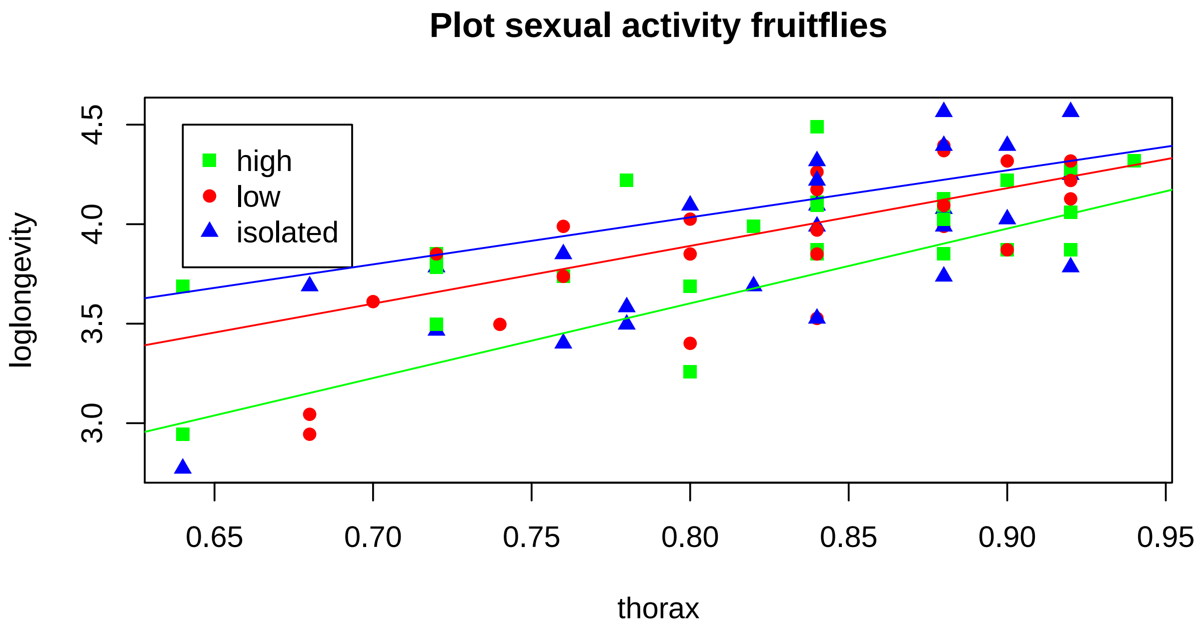
10-3-2021

EDDA Assignment 3

Exercise 1: Fruit Flies

A)

```
loglongevity <- log(longevity); flies$loglongevity <- loglongevity
activity = as.factor(activity)
plot(loglongevity~thorax, pch= c(15, 16, 17), col=c("green", "red", "blue"))
legend(0.64, 4.5, legend=c("high", "low", "isolated"), pch= c(15, 16, 17), col=c("green", "red", "blue"
title("Plot sexual activity fruitflies")
abline(lm(loglongevity~thorax,data=flies[activity=='low',]), col = "red")
abline(lm(loglongevity~thorax,data=flies[activity=="high",]), col = "green")
abline(lm(loglongevity~thorax,data=flies[activity=="isolated",]), col = "blue")
```



```
flies_wrong = lm(loglongevity~activity)
anova(flies_wrong)
```

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

```
summary(flies_wrong)
```

```
##
## Call:
## lm(formula = loglongevity ~ activity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95531 -0.13338  0.02552  0.20891  0.49222
##
## 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
##
## Residual standard error: 0.3072 on 72 degrees of freedom
## Multiple R-squared:  0.3504, Adjusted R-squared:  0.3324
## F-statistic: 19.42 on 2 and 72 DF,  p-value: 1.798e-07
```

According to an anova test, the influence of sexual activity on the loglongevity has a p-value of 1.798e-07, which is smaller than alpha and thus significant. So, we reject H0 (that loglongevity does not depend on the level of factor activity) and conclude that different levels of activity does influence the loglongevity. The estimated loglongevities are (in ascending order): high:3.60, low:3.99, isolated:4.11. Translating these to longevities can be done by using these as exponents. High: 36.598, Low: 54.055, Isolated: 60.947 All three have a p-value that is below 0.05, and thus significant. From this we can conclude that more sexual activity for fruit flies actually decreases the loglongevity, and thus decreases the longevity of the fruit fly.

B)

```
flies_correct = lm(loglongevity~thorax+activity)
drop1(flies_correct, test = "F")
```

```
## Single term deletions
##
## Model:
```

```
## loglongevity ~ thorax + activity
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                2.9180 -235.50
## thorax    1    3.8786 6.7966 -174.08  94.374 1.139e-14 ***
## activity  2    2.1129 5.0309 -198.64  25.705 4.000e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(flies_correct)
```

```
##
## Call:
## lm(formula = loglongevity ~ thorax + activity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4858 -0.1612  0.0104  0.1510  0.3574
##
## 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
##
## Residual standard error: 0.2027 on 71 degrees of freedom
## Multiple R-squared:  0.7211, Adjusted R-squared:  0.7093
## F-statistic:  61.2 on 3 and 71 DF,  p-value: < 2.2e-16
```

```
mean(thorax)
```

```
## [1] 0.8245333
```

When also including the thorax length as an explanatory variable, we now find that the activity significantly influences the loglongevity. In this case, loglongevity still increases the less sexual activity there is among fruit flies. When performing drop1 (anova for all versions of a model), we see that both activity and thorax are significant, as they have values lower than alpha, with 4.000e-19 for activity. When looking at the summary we see that the (log)longevity increases when activity decreases. More sexual activity thus decreases (log)longevity.

The average thorax length is 0.825, so this results in an average increase of loglongevity of (as added to the intercept): 2.4562766. The estimated loglongevity for the three levels are thus: high: 1.21+2.46=3.68. low: 3.68+0.28=3.97. isolated: 3.68+0.41 = 4.10. Again to obtain the longevities use these values as exponents: high: 39.646, low: 52.985 and isolated: 60.34

C)

```
flies2=lm(loglongevity~thorax,data=flies)
anova(flies2)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##           Df Sum Sq Mean Sq F value    Pr(>F)
## thorax     1 5.4322  5.4322  78.823 3.151e-13 ***
## Residuals 73 5.0309  0.0689
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
flies3=lm(loglongevity~thorax*activity,data=flies)
anova(flies3)
```

```
## Analysis of Variance Table
##
## Response: loglongevity
##                 Df Sum Sq Mean Sq  F value    Pr(>F)
## thorax           1 5.4322  5.4322 135.6195 < 2.2e-16 ***
## activity         2 2.1129  1.0565  26.3753 3.101e-09 ***
## thorax:activity  2 0.1542  0.0771   1.9251    0.1536
## Residuals       69 2.7638  0.0401
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

In the first anova test, we see that thorax length significantly influences the loglongevity, so we reject the null hypothesis of the thorax length being of no influence on the (log)longevity. From the plot in a) we can also see that the thorax length *positively* influences the (log)longevity.

From observing the plot in a) we see that within each level of the factor activity, the dependence of loglongevity on thorax is a straight line with approximately the same slope (the blue, red and green lines). Therefore the dependence seems similar under all three conditions.

Testing for the interaction between factor activity and predictor thorax is then done by including the interaction term activity*thorax in the model. The p-value for activity:thorax is higher than alpha and thus not significant. Therefore H0 (the interaction between thorax and activity is of no effect on the longevity) is not rejected, i.e. there is no significant interaction between factor activity and predictor thorax. So the dependence is similar under the three conditions.

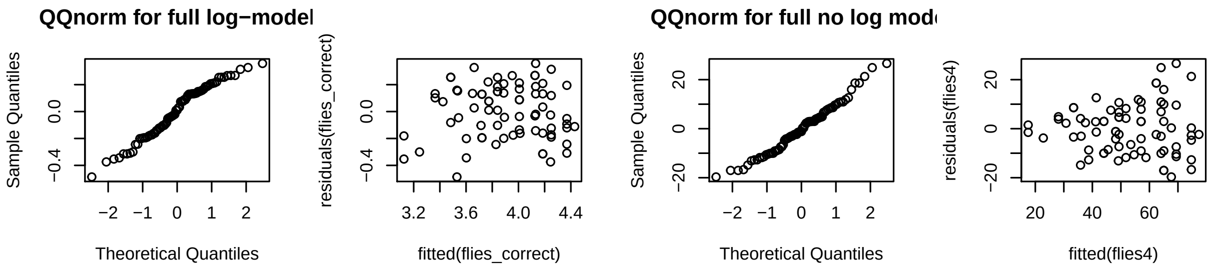
D)

We prefer the analysis with thorax since this is the most complete analysis & by observing the p-value for thorax in b) we can also clearly see this variable plays a role. Therefore excluding this variable from the analysis would be wrong.

E)

```
par(mfrow=c(1,4))
qqnorm(residuals(flies_correct), main="QQnorm for full log-model")
plot(fitted(flies_correct),residuals(flies_correct))

flies4=lm(longevity~thorax+activity,data=flies)
qqnorm(residuals(flies4), main="QQnorm for full no log model")
plot(fitted(flies4),residuals(flies4))
```



(2 left graphs) From the QQ-plot, we find that the residuals seem to be normally distributed, as the line is quite straight. Heteroscedasticity refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts is. If the scatterplot of the residuals versus the fitted model has a cone-like shape, this shows that the variability of the dependent variable widens or narrows as the value of the independent variable increases. The plot which shows the residuals versus the fitted model shows that there is no heteroscedasticity, as the plotted points appear very random and not in a cone-like pattern. Although the data is shifted slightly towards the right, we conclude that there is homoscedasticity.

F)

```
drop1(flies4,test="F")
```

```
## Single term deletions
##
## Model:
## longevity ~ thorax + activity
##          Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>                 7673 355.10
## thorax    1    7686.8 15360 405.15  71.127 2.624e-12 ***
## activity  2    4966.7 12640 388.53  22.979 2.016e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

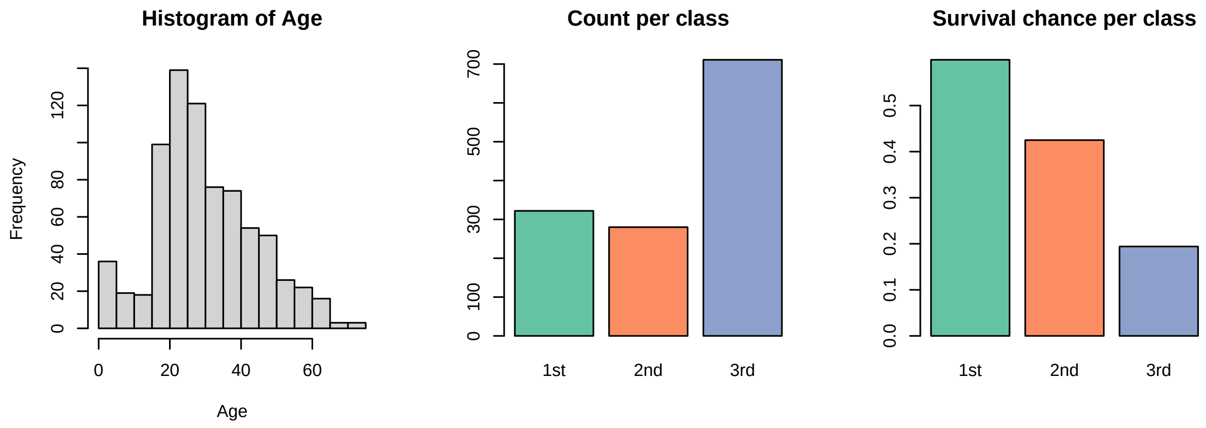
(2 right graphs) An ancova analysis with the longevity is performed above. We find that both thorax and activity have p-values lower than alpha. This means that we reject the null hypothesis of there being no effect for both thorax and activity on the longevity. The QQ-plot of this data shows normality through a straight, diagonal line. There is a difference from loglongevity in the plot of residuals vs the fitted model. Here we see a cone-shape, which indicates that there is heteroscedasticity instead of homoscedasticity. It was therefore wise to use log-longevity.

Exercise 2: Titanic

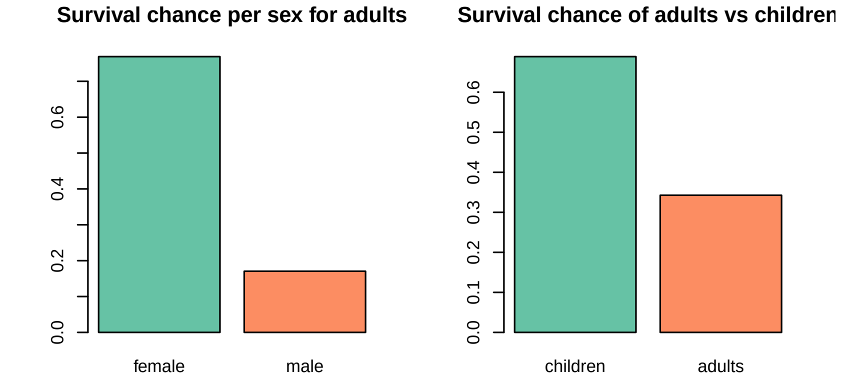
A)

```
par(mfrow=c(1,3))
library(RColorBrewer); coul <- brewer.pal(5, "Set2")
hist(Age) # NA automatically ommited from histogram
adultsurv = nrow(titanic[Survived=="1",])/nrow(titanic)
```

```
classtable = table(PClass) # all numbers from class
barplot(classtable, main = "Count per class", col=coul)
classsurv = xtabs(Survived~PClass) # all who survived from class
classsurvival = classsurv/classtable;
barplot(classsurvival, main="Survival chance per class", col=coul)
```



```
titanicAdult = titanic[Age>=18.00,]
sextable = table(titanicAdult$Sex)
sexsurv = xtabs(titanicAdult$Survived~titanicAdult$Sex)
sexsurvival = sexsurv/sextable;
barplot(sexsurvival, main = "Survival chance per sex for adults", col=coul)
titanicChild = titanic[Age<18.00,]
childsurv = nrow(titanicChild[Survived=="1",])/nrow(titanicChild)
barplot(c(childsurv, adultsurv), main = "Survival chance of adults vs children",
        names.arg = c("children", "adults"), col=coul)
```



In the age histogram we see that there are relatively a lot of 20-30 year old. We do however have a lot of missing data for this variable (42%, calculated as mean(is.na(Age))) so this histogram is far from truly representative. We extracted that 34% percent of the people survived, and the higher the class was you were in, the higher your chances of survival. Of the adults (18 or older) 77% of the woman survived, as opposed to 17% of the adult man. Of the children 69% survived, where of the adults 34% survived.

B)

```
PClass = as.factor(PClass)
Sex=as.factor(Sex)
titanicglm <- glm(Survived~PClass+Age+Sex, family = binomial)
summary(titanicglm)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Age + Sex, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7226  -0.7065  -0.3917   0.6495   2.5289
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.759662   0.397567   9.457  < 2e-16 ***
## PClass2nd   -1.291962   0.260076  -4.968 6.78e-07 ***
## PClass3rd   -2.521419   0.276657  -9.114  < 2e-16 ***
## Age         -0.039177   0.007616  -5.144 2.69e-07 ***
## Sexmale     -2.631357   0.201505 -13.058  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  695.14  on 751  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 705.14
##
## Number of Fisher Scoring iterations: 5
```

```
exp(3.76-2.631357*1)
```

```
## [1] 3.091459
```

```
1/0.071
```

```
## [1] 14.08451
```

Note: about 42% of the Age values is missing. This means that R in the default setting will omit rows that include NA values for modelling the generalized linear model. This could be overcome (imputation) by taking either the mean values for explanatory variables or the median values for categorical variables (factors). In this dataset the missing values are in the explanatory variable Age, therefore using the mean is most appropriate here. Given that it is not specified what to do with the missing values for this asssignment we decided to go with the standard procedure of GLM's: omitting the rows with NA data. However above we explained how these values can be used if desired.

The odds here are exp{3.76 -1.29*PClass2nd -2.52*PClass3rd -0.039*Age -2.63*Sexmale} (where PClass and Sexmale are one hot encodings, so 1 if you are in that class and 0 otherwise). This means that your odds change given your characteristics. The intercept is for a female child of age 0 in the 1st class. The more you deviate from this intercept the less likely you are to survive (estimates for all factors and explanatory variables are negative). For example if you are a male your odds decrease by exp{-2.63} = 0.07198072. So your odds to survive become 1/0.071 = 14.08 times worse than when you are a female.

C)

```
glm3=glm(Survived~Age*PClass,data=titanic,family=binomial)
anova(glm3, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         755    1025.57
## Age         1    2.849       754    1022.72  0.09141 .
## PClass      2  112.807       752     909.92  < 2e-16 ***
## Age:PClass  2    1.166       750     908.75  0.55816
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
glm4=glm(Survived~Age*Sex,data=titanic,family=binomial)
anova(glm4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      755    1025.57
## Age      1    2.849       754    1022.72   0.09141 .
## Sex      1  227.138       753     795.59 < 2.2e-16 ***
## Age:Sex  1   25.030       752     770.56 5.645e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the analysis above, we can conclude that Age and PClass do not interact, since the p-value is much higher than alpha and thus we cannot reject the H0 of B1=B2. We also find that Age and Sex do indeed interact, with a p-value lower than alpha, and thus rejecting H0 of B1=B2.

```r
titanicglm2=glm(Survived~PClass+Age*Sex,data=titanic,family=binomial)
summary(titanicglm2)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Age * Sex, family = binomial,
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4346  -0.6562  -0.3527   0.6964   2.7283
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.756563   0.437642   6.299 3.00e-10 ***
## PClass2nd   -1.543367   0.287358  -5.371 7.83e-08 ***
## PClass3rd   -2.653981   0.291423  -9.107  < 2e-16 ***
## Age          0.002443   0.011408   0.214    0.830
## Sexmale     -0.508187   0.442515  -1.148    0.251
## Age:Sexmale -0.075591   0.015009  -5.036 4.74e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  667.08  on 750  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 679.08
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(titanicglm2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      755    1025.57
## PClass   2   78.026       753     947.55 < 2.2e-16 ***
## Age      1   37.630       752     909.92 8.554e-10 ***
## Sex      1  214.776       751     695.14 < 2.2e-16 ***
## Age:Sex  1   28.064       750     667.08 1.174e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#exp{2.76 -1.54*PClass2nd -2.65*PClass3rd + 0.0024*Age -0.51*Sexmale -0.075591*Age:Sexmale}.
#exp(2.76 -1.54*0 -2.65*0 + 0.0024*53 -0.51*0 -0.075591*0) #Female 1st class
```

It is interesting to note that age is no longer negative, but the interaction between age:sexmale is. This means that a woman's odd increase (slightly) as she gets older, while a man's odds decrease.

This means that the estimates for probability of survival for the combinations of levels of factors PClass and Sex for a 53 year old are the following (calculated as follows; for male, class1: exp{2.76 -1.54*0 -2.65*0 + 0.0024*53 -0.51*1 -0.075591*53}):

Class \ Sex	Male	Female
1	0.196	17.943
2	0.042	3.847
3	0.014	1.268

D)

Fitting the model is similar to training a model in ML . After fitting we thus obtain theta-hat which we can use when predicting the success probability of new data. To generate a quality measure for our prediction we would need to split the data in training- and testing-data. The most common ratio for this is 80:20. In order to train a system well on training data, it is good to account for class imbalance (thus train as much on people that survived as on people that did not survive 50/50 instead of the 30/70 division there is in our dataset). Balanced classes can be obtained by oversampling on the survivors or undersampling on the non-survivors. On this data a model should then be fitted.

Then we could use this model (similar to above) to calculate the probability of success (P) for our testing data, and use this P in combination with a threshold value (p-0, e.g. 1/2) to generate outcomes of our model (Y-hat).

To test our model we could compare Y-hat with the true survival of rows in our testing data which we splitted from our original data in the beginning. Accuracy is not a very good metric to check whether our model performs well since there can be class imbalance in the test set as well, meaning that predicting everyone dies still gives an acuracy score of 70 % (given that 30% survived). Therefore using for instance precision and recall can give us more insight in the performance of our model.

E)

```
class_vs_survive = table(titanic$Survived, titanic$PClass)
class_vs_survive
```

```
##
##     1st 2nd 3rd
##   0 129 161 573
##   1 193 119 138
```

```
sex_vs_survive = table(titanic$Survived, titanic$Sex)
sex_vs_survive
```

```
##
##     female male
##   0    154  709
##   1    308  142
```

```
chisq.test(class_vs_survive)
```

```
##
##  Pearson's Chi-squared test
##
## data:  class_vs_survive
## X-squared = 172.3, df = 2, p-value < 2.2e-16
```

```
fisher.test(sex_vs_survive) # Fisher since 2x2
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  sex_vs_survive
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.07620521 0.13155709
## sample estimates:
## odds ratio
##  0.1003494
```

```
#class_vs_sex = xtabs(titanic$PClass, titanic) #<<< aanmaken
# total = xtabs(~Sex+PClass)
# total
# ifsurvived = xtabs(Survived ~Sex+PClass)
# ifsurvived
# percsurvived = round(ifsurvived/total, 3)
# percsurvived
#
# chisq.test(percsurvived)
```

To test this using contingency tables we have to create 2 tables: survive vs. class & survive vs. sex. It is also possible to summarize this info into one table by creating a table sex vs. class and having the percentages of people from the groups who survived in the table. The problem with this method is that it is no longer a contingency table, since a contingency table contains counts. We thus test with the aforementioned tables (fisher for survive vs. sex since this is a 2x2 table) and both results are significant with a p-value of 2.2e-16 (only the one of the fisher test is a true p-value however, the one of the chisq-test is an estimate by definition), meaning Sex and Class had a significant effect on survival.

F)

The test is not wrong, since it shows the dependencies between two variables. However, for this test only one variable is taken into account with respect to the survival rate per test. This is a disadvantage of the contingency test in comparison with the logistic regression model, because the latter method is not restricted by one variable while the contingency test is. An advantage of using contingency tables in comparison with glm's is the ease with which one can read the table. Counts are far more comprehensible than a linear model, meaning the data is easier to understand at first glance. Another disadvantage for the contingency test is that it cannot predict the odds for one person, while the glm can do this.

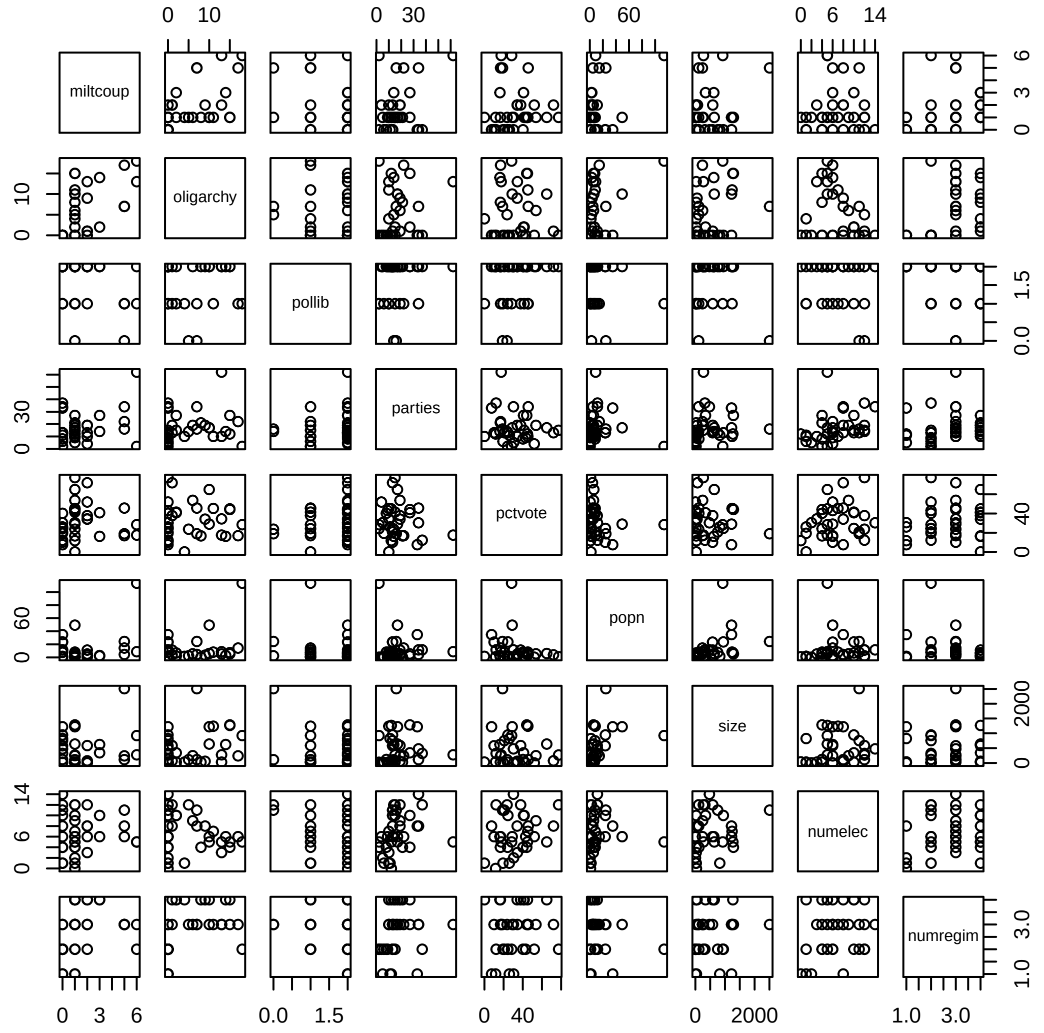
Exercise 3: Military coups in Africa

A)

```
pollib = as.factor(pollib)
africaglm=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,family=poisson,data=a
summary(africaglm)
```

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
##     popn + size + numelec + numregim, family = poisson, data = africa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3443  -0.9542  -0.2587   0.3905   1.6953
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5102693  0.9053301  -0.564  0.57301
## oligarchy    0.0730814  0.0345958   2.112  0.03465 *
## pollib      -0.7129779  0.2725635  -2.616  0.00890 **
## parties      0.0307739  0.0111873   2.751  0.00595 **
## pctvote      0.0138722  0.0097526   1.422  0.15491
## popn         0.0093429  0.0065950   1.417  0.15658
## size        -0.0001900  0.0002485  -0.765  0.44447
## numelec     -0.0160783  0.0654842  -0.246  0.80605
## numregim     0.1917349  0.2292890   0.836  0.40303
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 28.668  on 27  degrees of freedom
## AIC: 111.48
##
## Number of Fisher Scoring iterations: 6
```

```
pairs(africa) # = plot(africa)
```



We find that there are only three values that are significant out of 8: oligarchy, pollib, and parties. Furthermore, through the estimates, we find that pollitical liberization strongly decreases the number of successful military coups, while the number of regime types is the strongest variable that increases the number of successful coups. All other variables are relatively close to each other, which means that their influence is similar and around 0. In the figures, the figures with pollib, the figures with miltcoup, and the figures with numregim in them have separate rows of data, because these are either a factor (pollib) which has three options, or they have numeric values with a small variety (so, for instance, miltcoup has values only from 0 to 6, and numregim from 1 to 4). The variables we would then use are of course the ones that were a significant influence on miltcoup: oligarchy, pollib, and parties.

B)

```
# summary(glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,family=poisson,data=
```

While the analysis shows that there are many variables that have a p-value larger than alpha, we start with deleting the variable with the largest p-value: numelec

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

We now have a few variables with p-values larger than alpha, with the largest one being numregim This is the variable we will delete next.

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

The variable with the largest p-value is now size, so we delete this next.

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

The largest p-value is now popn, which we'll delete.

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

While pollib1 has the highest p-value here, pollib2 is still significant, meaning that we can't delete pollib. Pctvote is now the only variable with a p-value higher than alpha, so we'll delete it.

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

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
##     data = africa)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3583  -1.0424  -0.2863   0.6278   1.7517
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.251377   0.372689   0.674  0.50000
## oligarchy    0.092622   0.021779   4.253 2.11e-05 ***
## pollib      -0.574103   0.204383  -2.809  0.00497 **
## parties      0.022059   0.008955   2.463  0.01377 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 32.856  on 32  degrees of freedom
## AIC: 105.66
##
## Number of Fisher Scoring iterations: 5
```

Now we see that there are no p-values higher than alpha (other than pollib1, but again because of pollib2 being significant, we can't delete it), meaning that we're done with the step-down approach. This model uses the variables oligarchy, pollib and parties.

If we compare this model with the model in a, we find that both models use the same variables: oligarchy, pollib, and parties.

C)

Predicting within a poisson distribution is similar to predicting within a logistic regression, only now we are predicting the count and not the probability of success. The method (exp(mu-hat + alpha-i-hat + beta-hat * X-in)) is similar.

To predict the number of coups for a hypothetical country for all 3 levels of political liberalization and the averages of all other numerical characteristics, we first need the averages (all calculated as mean(x)):

```
mean(oligarchy)
```

```
## [1] 5.222222
```

Table 2: Averages values for numerical characteristics data africa

num. var:	*OLIGARCHY*	*PARTIES*	**pctvote**	**popn**	**size**	**numelec**	**numregim**
averages:	5.22	17.08	32.11	11.57	484.58	6.72	2.75

We can now fill in these values in the formula obtained from the summary of the model as obtained in b (miltcoup~oligarchy+pollib+parties) : 'exp(0.208 -0.495*pollib1 -1.112*pollib2 + 0.915*oligarchy + 0.022*parties)' (where level 0 of pollib is in mu) which yields:

Table 3: Predictions number of coups for 3 levels of political liberalization

levels pollib:	0	1	2
pred. number of coups:	2.890	1.762	0.313

What you see is that the model predicts that for an average Sub-Saharan African country (w.r.t. all numerical variables) the predicted number of successful military coups from independence to 1989 decreases as political liberalization (0 = no civil rights for political expression, 1 = limited civil rights for expression but right to form political parties, 2 = full civil rights) increases. In an average country the with full civil rights the predicted number of coups is even below 1. In countries with no civil rights for political expression however (level 0) the prediction is that 2.9 coups happened since the independence of the country until 1989, and 1.8 for an average country with limited civil rights (level 1).