Experimental Design and Data Analysis: Assignment 4

This assignment consists of 3 exercises. Throughout this assignment tests should be performed using a level of 0.05.

EXERCISE 1

To investigate the effect of sexual activity on longevity of fruit flies, 75 male fruit flies were divided randomly in three groups of 25. The fruit flies in the first group were kept solitary, those in the second were kept together with one virgin female fruit fly per day, and those in the third group were kept together with eight virgin female fruit flies a day. In the data-file `fruitflies.txt` the three groups are labelled `isolated`, `low` and `high`. The number of days until death (`longevity`) was measured for all flies, as was the length of their thorax.

1. Add a column `loglongevity` to the data-frame, containing the logarithm of the number of days until death. Use this as the outcome variable in the following.
2. Make an informative plot of the data.
3. Investigate whether sexual activity influences longevity by performing a statistical test, without taking the thorax length into account.
4. Does sexual activity increase or decrease longevity? What are the estimated longevities for the three conditions? To answer these questions, use the analysis as under 3), without taking thorax length into account.
5. Investigate whether sexual activity influences longevity by performing a statistical test, now including thorax length as an explanatory variable into the analysis.
6. Does sexual activity increase or decrease longevity? What are the estimated longevities for the three conditions, for a fly with average thorax length? And what are they for a typical fly as small as the smallest in the data set? To answer these questions, use the analysis as under 5), which includes thorax length.
7. How does thorax length influence longevity? Graphically investigate whether this dependence is similar under all three conditions of sexual activity.
8. Which of the two analyses, without or with thorax length, do you prefer? Is one of the analyses wrong? (To answer the last question, carefully (re)read the description of the design of the experiment.)
9. Verify normality and heteroscedasticity by making a normal QQ-plot of the residuals, and a residuals versus fitted plot, for the analysis that includes thorax length.
10. Perform the ancova analysis with the number of days as the outcome, rather than its logarithm. Verify normality and heteroscedasticity of the residuals of this analysis. Was it wise to use the logarithm as outcome?

EXERCISE 2

To study the effect of a new teaching method called "personalized system of instruction" (*psi*), 32 students were randomized to either receive *psi* or to be taught using the existing method. At the end of the teaching period the success of the teaching method was assessed by giving the students a difficult assignment, which they could pass or not. The average grade of the students (*gpa* on a scale of 0–4, with 4 being the best grade) were also available for analysis.

The data can be found in the file `psi.txt`.

1. Study the data and give a few (> 1) summaries (graphics or tables).
2. Fit a logistic regression model with both explanatory variables.
3. Does *psi* work?
4. Estimate the probability that a student with a `gpa` equal to 3 who receives *psi* passes the assignment. Estimate the same probability for a student who does not receive *psi*.
5. Estimate the relative change in *odds* of passing the assignment rendered by instructing students with *psi* rather than the standard method (for an arbitrary student). What is the interpretation of this number? Is it dependent on `gpa`?

Consider the following alternative method of analysis. Out of 18 students who did not receive *psi* 3 showed improvement, of the 14 remaining students 8 showed improvement. We perform a test for comparing two binomial proportions: we have two sequences of independent binary "experiments", of lengths 18 and 14. The experiments in the first sequence have success probability p_1, those in the second p_2. We wish to test the null hypothesis $H_0 : p_1 = p_2$ using the observed numbers of successes 3 and 8. You can find the test (Fisher's exact test or the chisquare test for a 2x2 table) in your introductory statistics book (or see e.g. Dalgaard, section 7.2). In `R` you can simply type:

```
> x=matrix(c(3,15,8,6),2,2)
> x
     [,1] [,2]
[1,]    3    8
[2,]   15    6
> fisher.test(x)
```

6. Do this. What are the numbers 15 and 6 in this table? What is the conclusion?
7. Given the way the experiment was conducted, is this second approach wrong? Why or why not?
8. Name both an advantage and a disadvantage of the two approaches, relative to each other.

EXERCISE 3
To study the influence of different political and geographical variables on the number of military coups, these data are collected for several Sub Saharan African countries in the file `africa.txt`. The meaning of the different variables is
`miltcoup` — number of successful military coups from independence to 1989
`oligarchy` — number years country ruled by military oligarchy from independence to 1989
`pollib` — 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)
`parties` — Number of legal political parties in 1993
`pctvote` — Percent voting in last election
`popn` — Population in millions in 1989
`size` — Area in 1000 square km
`numelec` — Total number of legislative and presidential elections
`numregim` — Number of regime types

In this exercise you will fit a regression model to these data using Poisson regression, but first Poisson distributions are studied.

1. Study (graphically) some different Poisson distributions using Poisson samples generated by `rpois(n,lambda)`, varying both `n` and (positive) `lambda`.

2. Two distributions are in the same location-scale family if there is a scale-and-shift transformation that maps one to the other. Investigate whether different Poisson distributions are in the same location-scale family, like all normal distribution. Clearly explain your approach to this question, and your answer.

3. Perform Poisson regression on the full `africa` data, taking `miltcoup` as response variable.

4. Use the step down approach (using output of the function `summary`) to reduce the number of explanatory variables.

5. Make one or more diagnostic plots of the deviance residuals for the model you have found in part (4). In case you see a certain pattern, check whether that is also present in the full model (using all explanatory variables), i.e. see whether is it due to deleting too many variables. Comment on the pattern (if possible).