

## EDDA: Assignment 3

Throughout this assignment tests should be performed using a level of 0.05, unless otherwise specified.

### Exercise 1. Fruit flies

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. Later, it was decided to measure also the length of their thorax. Add a column `loglongevity` to the data-frame, containing the logarithm of the number of days until death. Use this as the response variable in the following.

- Make an informative plot of the data. Investigate whether sexual activity influences longevity by performing a statistical test, without taking the thorax length into account. What are the estimated longevitys for the three conditions? Comment.
- Investigate whether sexual activity influences longevity by performing a statistical test, now including thorax length as an explanatory variable into the analysis. Does sexual activity increase or decrease longevity? What are the estimated longevitys for the three groups, for a fly with average thorax length?
- How does thorax length influence longevity? Investigate graphically and by using an appropriate test whether this dependence is similar under all three conditions of sexual activity.
- Which of the two analyses, without or with thorax length, do you prefer? Is one of the analyses wrong?
- 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.
- Perform the ancova analysis with the number of days as the response, rather than its logarithm. Verify normality and heteroscedasticity of the residuals of this analysis. Was it wise to use the logarithm as response?

### Exercise 2. Personalized system of instruction

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`.

- Study the data and give a few ( $> 1$ ) summaries (graphics or tables).
- Fit a logistic regression model with both explanatory variables. Does *psi* work?
- 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*.
- 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 apply Fisher’s exact test or the chisquare test for a 2x2 table. 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)
```

- e) Do this. What are the numbers 15 and 6 in this table? What is the conclusion?
- f) Given the way the experiment was conducted, is this second approach wrong?
- g) Name both an advantage and a disadvantage of the two approaches, relative to each other.

### Exercise 3. Military coups in Africa

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:

`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

- a) Perform Poisson regression on the full data set `africa`, taking `miltcoup` as response variable.
- b) Use the step down approach (using output of the function `summary`) to reduce the number of explanatory variables.