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.

a) 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 longevities for the three conditions? Comment.

b) 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 longevities for the three groups, for a fly with average thorax length?

c) 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.

d) Which of the two analyses, without or with thorax length, do you prefer? Is one of the analyses wrong?

e) 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.

f) Perform the ancova analysis with the number of days as the response, rather than its logarithm. Verify normality and homoscedasticity of the residuals of this analysis. Was it wise to use the logarithm as response?

Exercise 2. Titanic

On April 15, 1912, British passenger liner Titanic sank after colliding with an iceberg. There were not enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. The data file `titanic.txt` gives the survival status of passengers on the Titanic, together with their names, age, sex and passenger class. (About half of the ages for the 3rd class passengers are missing, although many of these could be filled in from the original source.) The columns: `Name` – name of passenger; `PClass` – passenger class (1st, 2nd or 3rd), `Age` – age in years, `Sex` – male or female, `Survived` – survival status (1=Yes or 0=No).

a) Study the data and give a few (>1) summaries (graphics or tables).

b) Fit a logistic regression model (no interactions yet) to investigate the association between the survival status and the predictors `PClass`, `Age` and `Sex`. Interpret the results in terms of odds, comment.

c) Investigate for interaction of predictor `Age` with factors `PClass` and `Sex`. From this and b), choose (and justify) a resulting model. For this model, report the estimate for the probability of survival for each combination of levels of the factors `PClass` and `Sex` for a person of age 53.

d) Propose a method to predict the survival status and a quality measure for your prediction and describe how you would implement that method (you do not need to implement it).

e) Another approach would be to apply a contingency table test to investigate whether factor passenger class (and gender) has an effect on the survival status. Implement the relevant test(s).

f) Is the second approach in e) wrong? Why or why not? Name both an advantage and a disadvantage of the two approaches, relative to each other.

Exercise 3. Military coups jn 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, Comment on your findings.
b) Use the step down approach (using output of the function `summary`) to reduce the number of explanatory variables. Compare the resulting model with your findings in a).
c) Predict the number of coups for a hypothetical country for all the three levels of political liberalization and the averages (over all the counties in the data) of all the other (numerical) characteristics. Comment on your findings.