02418, Assignment 2

Jan Kloppenborg Møller

October 2022

Introduction

During the semester you will analyze 3 different cases, in this second assignment you will continue working on the data from (two of the projects in) Assignments 1.

The data for the projects can be found together with the assignment on Learn.

It is important to describe your results and conclusions, not only numbers, but also in words interpreting your results.

Project 1: Wind power forecast

In this project you will be analyzing a data set from Tunø Knob wind power plant. Details on data may be found in Assignment 1, and you will continue the analysis from that assignment.

Regression models:

Here you should consider wind power as the response variable, and wind speed and wind direction as explanatory variables

1. Formulate an initial model,

$$\hat{y} = f(w_s) \tag{1}$$

   where y is observed power, and f, is a function of wind speed (e.g. $f = \beta_0 + \beta_1 w_s + \beta_2 ws^2$).

2. You might consider non-normal models and/or normal model with data transformation. Further you might consider including wind direction. You should devellop a suited model for prediction of daily power production.

3. Present the parameters of the final model, this include quantification of the unceartainty of the parameters.

4. Give an interpretation of the parameters in particular this should include presentation of any nonlinear functions (series expansions) of the explanatory variables. Evt. holde andre parms fast på MLE og plotte den non-linear function sammen med prædiktioner => fanger den adfærden?

5. Present the final model, e.g. some graphical presentation of predictions under different scenarios of wind speed and wind direction.

Project 2: Survival data

This project[1] treat binary and survival data. You will be analyzing two data sets studying the treatment of HIV patients.

The data can be found in the files `Logistic.txt` and `actg320.txt`.

A general description of the date is given in Assignment 1.

Analysis of the binary data

The first part of the assignment deals with the study of the effect of AZT on AIDS

- Read the data `Logistic.txt` into R.
- Fit a logistic regression model for the binary outcome AIDS="yes" versus AIDS="no" with the explanatory variable treatment with AZT (Yes, NO). Present the odds ratio for the effect of AZT on AIDS with 95% confidence interval and interpret the result in words.
- Test the hypothesis of no effect of AZT on AIDS using:
  - The likelihood ratio test
  - The Wald test
  - The score test (we will cover this in lecture 9). This test needs to be derived, so use the log-likelihood from the slides on logistic regression and apply it to this special case. Again we have that in our case the score test follows a χ^2 distribution with one degree of freedom.

    When we have more than one parameter in the model the general expression for the score test is

    $$t(S(\hat{\boldsymbol{\beta}}))V(S(\hat{\boldsymbol{\beta}}))^{-1}S(\hat{\boldsymbol{\beta}})$$

    Here $\hat{\boldsymbol{\beta}}$ is the estimate of the vector of parameters $\boldsymbol{\beta}$ under H_0 and $V(S(\hat{\boldsymbol{\beta}}))$ is the variance (matrix) of the score function (a vector) S evaluated in $\hat{\boldsymbol{\beta}}$.

[1] Originally this part of the assignment was created by Elisabeth Wreford Andersen

Analysis of the survival time data

In this part we look at the data-set `actg320.txt`. The main outcome of interest is the time variable. We want to see whether there is a difference for the two treatment groups (i.e. `tx=0` or `tx=1`). This type of data is called survival data and some data are censored, i.e. persons leave the study (or the study is terminated) without having developed AIDS, this imply that we only know the the time of event is longer than the reported time. The data is described in Assignment 1.

1 Analyses of the survival time data

The main outcome of interest is the time variable, where we want to see whether there is a difference for the two treatment groups.

1.1 Descriptive statistics

- Read the data `actg320.txt` into R. If you are using RStudio you can use the "Import Dataset" button.
- How many patients got AIDS or died in the two treatment groups? And how long was the total follow-up time in the two groups?
- Plot the survival functions in the two treatment groups, which group seems to be doing best?

Plot fits oven på kaplan meier og beregn Time Ratio for de to grupper da det er en nem måde at sammenligne på. (HUSK 95% CI for time ratios?).

- Plot the cumulative incidence functions for the two groups, which plot would you prefer?
- Compare the survival in the two treatment groups using a log-rank test.

1.2 Parametric survival models

- Fit parametric survival models containing treatment (tx) and CD4 count (cd4) as explanatory variables.

Har lavet noget lignende allerede tilføj nu bare CD4

  - Try using the exponential, Weibull and log-logistic models, which one gave the best fit (and why)?
- Using the survival model you chose, make a table of estimates and their 95% confidence intervals.

- Using your model compute the time ratio for the treatment effect. Similarly, compute the time ratio for the effect of increasing the CD4 count with 50. In both cases unceertainty evaluation (e.g. confidence intervals) should be included. Interpret the results in words.
- Assess the goodness of fit of this model using a plot based on the Cox Snell residuals.
- Give a graphical presentation of your model.