02418, Assignment 1

Jan Kloppenborg Møller

September 2020

Introduction

During the semester you will analyze 3 different cases, this first assignment will treat simple models for the data, and you will continue working on the data in the other two assignments.

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

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.

Motivation

The share of wind energy is increasing. In 2050 100% of the total electricity production in Denmark should come from renewable sources. The big and increasing proportion of wind power imply that accurate predictions of production is essential; E.g. to balance the grid and to decide optimal trading strategies on the Nordic power exchange (Nord Pool). In this project you will model the average daily wind power production for a wind power plant in Denmark.

Data

The data set consists average daily values for wind energy production for Tunø-Knob wind power plant, the plant is a small off shore wind power

Table 1: Overview of all variables and their meaning data set `tuno.txt` .

Variable	Meaning	Unit
`r.day:`	Days since 1/1 2003	days
`month:`	Month in year	
`day:`	Day in month	
`pow.obs:`	Average daily wind power production	kW
`ws30:`	Predicted wind speed 30 meters above ground level	m/s
`wd30:`	Predicted wind direction (0 north, $\pi/2$ east) 30 meters above ground level	rad

plant located north of Samsø. The installed capacity (maximum production) is 5.000kw.

Data prepossessing

This section gives a brief description of how data was treated before you use them for modeling of daily production. The original data set consists of average hourly production between midnight and 18.00, these values are in your data set translated into average production for each day. In addition to production the original data set contained corresponding meteorological predictions of a large number of variables, these are also converted to the daily averages (also based of the time period 00:00 to 18:00).

T he data set consisted of numbers between 0 and $5,000$ kW. Average daily values that are exactly zero are thought of as special phenomena where, for one reason or another the plant is down, these observations are removed from the data set.

The total data set can be divided into three categories; 3 variables to describe the time (days since 1/1-2003, month, and day of month), 1 variable describing the observed output (power) and 2 meteorological variable. The individual variables are described in Table 1.

Descriptive statistics

1. Read the data `tuno.txt` into R.
2. Make a graphical presentation of data or parts of the data, and present some summary statistics.

Before you start the model-ling it will make some of the questions simpler if you normalize (to numbers strictly between 0 and 1) the power production. Be careful to do this in a meaningful way (hint: relate to installed capacity) and describe precisely how it is done.

Simple models:

1. Fit different probability density models to wind power, wind speed and wind direction data. You might consider different models e.g. beta, gamma, log normal, and different transformations e.g. (for wind power)

$$y^{(\lambda)} = \frac{1}{\lambda} \log\left(\frac{y^\lambda}{1-y^\lambda}\right); \quad \lambda > 0 \tag{1}$$

$$y^{(\lambda)} = 2 \log\left(\frac{y^\lambda}{(1-y)^{1-\lambda}}\right); \quad \lambda \in (0,1) \tag{2}$$

   It is important that you consider if the distributions/transformations are reasonable for the data that you try to model.

2. Conclude on the most appropriate model for each variable, also report parameters including assessment of their uncertainty. For models that does not include a transformation you should also give an assessment of the uncertainty of the expected value in the model.

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`.

Introduction

There are two data sets for this assignment both dealing with the treatment of HIV patients. The first data set comes from a randomized study described in the New York Times (1991) looking at the effect of AZT (azidothymidine) on slowing the development of AIDS symptoms. The study included 338 individuals with HIV, who were randomized to either receive AZT immediately (AZT=yes) or only after severe immune weakness (AZT=no). At the end of the study the number of patients, who developed AIDS, was recorded for the two treatment groups. The data from this study are found in the file `Logistic.txt` .

The second data set for this assignment comes from a double-blind, placebo-controlled trial that compared a three-drug treatment with a standard two-drug treatment in HIV-infected patients (Hammer et al., 1997). Patients were eligible for the trial if they had no more than 200 CD4 cells per cubic millimeter and at least three months of prior zidovudine therapy. Randomization was stratified by CD4 cell count at the time of screening. The primary outcome measure was time to AIDS or death. Because efficacy results met a pre-specified level of significance at an interim analysis, the trial was stopped early.

The data come from Hosmer, D.W. and Lemeshow, S. and May, S. (2008) Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition, John Wiley and Sons Inc., New York, NY.

Table 2 shows the variables included in `actg320.txt`. In this project you should only consider the variables `time`, `event`, and `tx`.

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

Table 2: Variables in ”actg320.tx”

Variable	**Description**	**Values**
id	Identification Code	1-1156
time	Time to AIDS diagnosis or death	Days
event	Indicator for AIDS or death	1 = AIDS diagnosis or death, 0 = Otherwise
tx	Treatment indicator	1 = New treatment, 0 = Control treatment
sex	Sex	1 = Male, 2 = Female
raceth	Race/Ethnicity	1 = White Non-Hispanic, 2 = Black Non-Hispanic, 3 = Hispanic (regardless of race), 4 = Asian, Pacific Islander, 5 = American Indian, Alaskan Native, 6 = Other/unknown
karnof	Karnofsky Performance Scale	100 = Normal, 90 = Normal activity possible, 80 = Normal activity with effort, 70 = Normal activity not possible
cd4	Baseline CD4 count	Cells/milliliter
age	Age at Enrollment	Years

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 the Binomial distribution to the data (i.e. consider all data as coming from the same population).
- Fit the Binomial separately to the two distributions and test if there is a difference between the groups.
- Estimate parameters in the model (p_0 probability of AIDS in control group, p_1 probability of AIDS in treatment group)

$$p_0 = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{3}$$

$$p_1 = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}} \tag{4}$$

  and report a confidence interval for the parameter describing the difference, compare with the result above.

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 we will treat it in more detail later in the course. For now it suffice to know that some data are censored, i.e. persons leave the study (or the study is terminated) without having devellopedAIDS, this imply that we only know the the time of event is longer than the reported time.

Questions

- 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? What is the proportion of patients that got AIDS or died in the two group? Other relevant number that could be calculated?

- Fit an exponential distribution, using numerical methods, to the time of event (`time`) in the data set, remember to take into account that some of the data is censored (i.e. we only know that the time to the event is longer that the reported time).

  1: Using all data (i.e. ignore the treatment effect)

  2: Separately for the two treatments

- Compared the likelihood for the above models and conclude
- Formulate a model where one parameter indicate the treatment effect (e.g. $E[T] = e^{\beta_0}$ if control group and $E[T] = e^{\beta_0+\beta_1}$ if treatment group), find the MLE and compare with the result above.
- Find the Wald confidence interval for the treatment parameter in the model above.
- Derive the theoretical results for the models above, including the standard error estimates, use this to formulate and implement the profile likelihood function for the treatment parameter

Project 3: Financial data

Background

In this project, the weekly returns for an ETF is analyzed and modeled. An ETF (Exchange Traded Fund) can be described as a structured publicly traded pool of shares. ETFs are bought and sold in the same way as ordinary shares on a stock exchange.

An ETF is a pooled investment fund similar to a unit trust or mutual fund. For investors, ETFs blend the benefits of pooled funds and shares.

The available data is found in the file `finance_data.csv` and consists of 2 columns. The first column is a date column and the second column indicates the weekly returns (i.e. the ratio between the final and initial price for that week minus 1) of 1 ETF.

An important aspect of financial data is the volatility, i.e. standard error of the return. In this project you will explore properties of volatility and return in the given data, and a model for the time evolution of these.

1: Descriptive statistics and simple models

a) Present the data, estimate the parameters in a normal model, and asses if the normal model is appropriate.

b) Hypothesize a model that could fit the data better (Hint: consider tail probabilities), and compare with the normal model estimated above

c) Present the final model (i.e. relevant keynumbers for the estimates)