Assignment 2: Treatment of HIV patients

Jan Kloppenborg Møller, Elisabeth Wreford Andersen

October 2018

This assignment mainly deals with logistic regression and survival analysis and you will be analysing two data sets studying the treatment of HIV patients. It is important to describe your results and conclusions, not only in numbers but also in words interpreting your results.

You can find examples of R code for logistic regression and survival analysis in the scripts accompanying the lectures.

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

1 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` .

2 Analyses 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 (November 7)). 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 $\hat{\boldsymbol{\beta}}$.

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 1 shows the variables included in `actg320.txt`.

Table 1: 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

3 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.

3.1 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 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.

3.2 Parametric survival models

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