Introduction to Survival Analysis

Exercises

These exercises are part of a resource ‘Advanced Topics in Survival Analysis’ on <https://www.ncrm.ac.uk/resources/online/all/?id=20860>

Before you run the exercises

I have used R to create and illustrate these exercises. I have assumed that the readers have some familiarity with R: I will not explain in details R functions. If you are not familiar with R, there are several resources available. A good place to start is the UCLA Advanced Research Computing webpage:

<https://stats.oarc.ucla.edu/r/>

Together with these resources I have provided a document called “*Introduction to the Survival Library in R\_Intro to Survival Analysis.OPerra.docx*”. This provides a very short introduction to the library suite that is most commonly used for running Survival Analysis in R. If you have never used this library in R, please read the document indicated before running these exercises.

Together with the “survival” library, I will also use the following libraries:

library(survival)

library(survminer) #creates nicer plots and tables from survival analysis

library(muhaz) #estimates and plot kernel-smoothed hazards

library(psych) # useful for summary stats

library(kableExtra) # creates nice tables

library(dplyr) #for data manipulation

library(ggplot2) # to create plots

library("ggthemes") # to create plots with preset colours

library(ggpubr) #to combine ggplots

library(RColorBrewer) #to use preset colour combinations in the plots

library(tidyverse) # to manipulate dataset

While not necessary, using some R interface, e.g. *RStudio*, can greatly help you familiarise with and use R.

The Dataset

The dataset for these exercises is provided with the online material and it is called:

**df\_exercise.RData**

This is a simulated dataset I have created for these exercises. It includes 500 cases. In creating this dataset, I imagined a fictional setting of a study where children who had not yet uttered a sentence (subject – verb – object) are followed up for a year: the occurrence of their first sentence is recorded in days within the year. The main variables in the dataset are:

|  |
| --- |
| **ID**: The adolescent ID in the study. |
| **time**: The timing of first sentence utterance in days or the timing (in days) when participants left the study, if they are censored. |
| **censor**: A dummy variable to indicate whether participants are censored (censor==1) or not (censor==0) |
| **male**: A dummy variable for sex, where male==1 indicates male children and male==0 indicates female children (in this fictional example, only two options were provided) |
| **ses**: An indicator of Socio-Economic Status, ranging from 1 (High SES) to 3 (Deprived SES) |

Exercise Set #1

Use the dataset provided to complete the following tasks:

1. Create a survival object using the library “survival”.
2. Extract and plot the Kaplan-Meier **survivor function** with 95% Confidence Intervals (CIs), and estimate the sample **median survival time**.
3. Extract and plot the **cumulative hazard function** using Nelson-Aalen method; include 95% CIs.
4. Estimate and plot the **hazard function using kernel-smoothing**: I suggest you try three different bandwidths: 15, 30, and 90 days.

Exercise Set #2

Use the dataset provided to complete the following tasks:

1. Explore differences between males and females, firstly by plotting their respective Kaplan-Meier survivor functions, and their respective median survival time.
2. Plot sex differences in the log of the negative log of the survivor function.
3. Run a Cox regression model including child’s sex and SES as predictors. Since SES is a categorical variable, make sure you compare children from the most deprived background to other groups. Report and interpret the results.
4. Consider ties, and try another method such as Breslow in the Cox regression.
5. Recover “baseline” functions and create functions for other “prototypical” groups: Compare a group of female from high SES backgrounds, females from deprived SES backgrounds, as well as males from high or low SES backgrounds. Recover the survivor and cumulative hazard function of these groups, and plot them.