Introduction to Survival Analysis

Exercises

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 exercise.

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

library(survival)

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:

**sa\_exercise.csv**

This is a fictional dataset I have created for these exercises. It includes 500 cases. In creating this dataset, I imagined a fictional setting of a study where, after an intervention to prevent alcohol abuse in adolescents, the researcher measure whether and when adolescents report alcohol abuse for the first time (“relapse”). I imagined the study lasted for a year. The main variables in the dataset are:

|  |
| --- |
| **ID**: The adolescent ID in the study. |
| **time**: The timing of relapse in days or the timing when participants left the study, if they are censored (in days) |
| **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 adolescents and male==0 indicates female adolescents (in this fictional example, only two options were provided) |
| **ses**: An indicator of Socio-Economic Status, ranging from 1 (deprived) to 3 (affluent) |
| **time\_mo**: The timing of relapse (target event) or the censoring in monthly intervals |
| **timedis**: The timing of relapse (target event) or censoring in a discrete scale (bi-monthly periods, so that the time units range from 1 to 6) |

Exercise Set #1

Use the dataset provided to complete the following tasks:

1. Create a survival object through the R package “survival” and considering the discrete timing of the target event in the provided dataset (i.e.: variable “timedis”). Check that the survival object has been correctly specified.
2. Create a life table for the whole sample based on the discrete time variable. Ensure that the life table reports the risk set, the number of cases who experienced the event in each time interval, the number of cases censored in each time interval, the survival and the hazard function.
3. Plot the hazard function, preferably using the “ggplot” package.
4. Plot the survival function, preferably using the “ggplot” package.
5. Calculate the median lifetime of survival function using interpolation, and add this in the survival function plot.
6. Compare the two plots to draw some insights.

Exercise Set #2

Use the dataset provided to complete the following tasks:

1. Create life tables for females and males in the dataset considering the discrete time variable (“timedis”). Ensure the life tables report the risk set, the number of cases who experienced the event in each time interval, the number of cases censored in each time interval, the survival and the hazard function.
2. Plot the hazard functions for females and males in the same plot, preferably using the “ggplot” package.
3. Plot the survival function for females and male in the same plot, preferably using the “ggplot” package. Add the median lifetime for females and males using interpolation
4. Transform the hazard functions for females and males into odds and into logits, plot these transformed outcomes, and inspect the plots.
5. Use a the person-period dataset “sa\_exercise\_pp.csv” to fit a general specification model of the logit hazard functions for females and males over discrete time periods. Use a logistic regression to fit the model. Describe and interpret the results: in particular, how can you interpret the parameter of the covariate “male”?
6. Use the person-period dataset “sa\_exercise\_pp.csv” to fit a model assuming a linear change in the logit hazard function across time periods. Include the covariate “male”. Use a logistic regression to fit the model.