

Introducing survfuncs Package - Simple Survival Analysis with R

by Victor Wilson, Ashley Jacobson

Abstract This article introduces the R package `survfuncs`, which executes parametric survival analysis techniques similar to those in Minitab. The functions available in this package carry out basic survival analysis techniques. Among these are plotting hazard, cumulative hazard, and survival curves, based on specified parametric distributions, computing survival probabilities, and computing summary statistics. We describe appropriate usage of these functions, what the output means, and provide examples of how to utilize this functions in real-world datasets.

Introduction

Survival analysis is a branch of statistics that primarily deals with analyzing the time until an event of interest occurs. This event could be a variety of different things such as death, development of disease, or first score of a soccer game. Observations in survival analysis may also be described as censored, which occurs when an observation's survival time is incomplete. The most common way that this occurs is through right censoring, which occurs when a subject does not experience the event of interest within the duration of the study. Right censoring can also occur if a subject drops out before the end of the study and does not experience the event of interest. Due to the inherent issue of censoring that is typically found in datasets involving survival analysis, computations and analyses can be difficult to carry out with many standard functions available in R, as the majority of these do not account for censored data. The censored data here is of value and we cannot merely eliminate the observations which have censored data.

Some of the most popular techniques and statistics utilized when carrying out a survival analysis are computing what are known as the survival and hazard functions. The survival function is important because it gives the probability of surviving (also known as not experiencing the event of interest), for any given time. Similarly, the hazard function is also useful to compute because it gives the conditional probability that the subject will experience the event in the next instance of time, given that they have survived up until the specified point in time. Other popular statistics that are utilized are median survival time, mean survival time, and percentiles of survival time. In this package, all of the functions that we developed utilize parametric methods of survival analysis, which assumes that the distribution of the survival times follows a known probability distribution.

Currently, R does have many survival packages that have been developed that compute some of these statistics. However, we noticed that Minitab has very concise and easy to utilize functions for computing and displaying many of these survival statistics and plots, but this same output is not readily available in any single one package in R, or in some cases not available at all. Thus, we decided to develop a package that emulates the output found in Minitab for survival analysis, which contains all of these commonly utilized statistics and plots.

This paper describes the functions that this package contains, how the data is formatted in order to utilize these functions, and what the output of these functions represent. There are 3 major groups of functions that we created: fitting the censored data, displaying plots (hazard, cumulative hazard, and survival), and computing summary statistics (mean, median, survival probabilities). The majority of this paper will be organized following these groups of functions.

Fitting Right Censored Survival Data

As mentioned previously, this function is very similar to the function `fitdistcens` found in the `fitdistrplus` package. This package works to help fit a parametric distribution to data that is right censored. This function also computes the Maximum Likelihood Estimate (MLE).

Example

```
#> Fitting of the distribution 'logis' on censored data by maximum likelihood
#> Parameters:
#>           estimate
#> location 133.74054
#> scale    20.48286
```

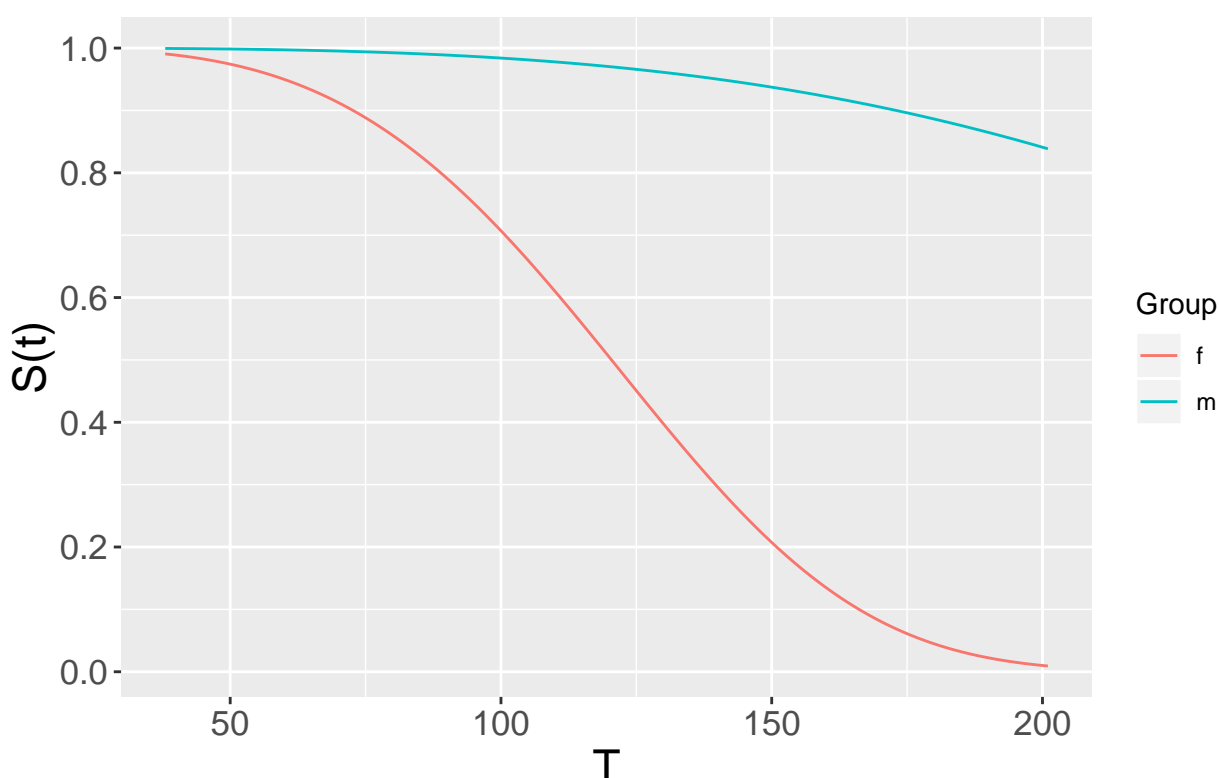
```
#> Fixed parameters:
#> data frame with 0 columns and 0 rows
```

Displaying Plots

Survival Plots

This section introduces an overview of the many types of plots that are available to be displayed via this package. Some of the most common plots used in Survival Analysis are survival plots, hazard plots, and cumulative hazard plots. Survival plots are used to estimate the proportion of subjects that survive beyond a specified time t . We were motivated to create the function `plot_surv` function in an attempt to mimic the hazard plots that are available in Minitab. This function plots the survival curve of right censored data given that it follows a specified parametric distribution. Some examples of the distributions that this function supports are the Weibull, Log-Normal, Exponential, Normal, and Logistic distributions. This function also provides the option to plot by a grouping variable, which if specified, displays separate survival plots for each group of the specified variable.

weibull survival function



In this example, we fit a Weibull distribution to the rats dataset available in the [survival](#) package, grouping by the “sex” variable. As seen in the plot above, two different survival curves were plotted. The blue line represents the estimated survival curve for male rats, while the red line represents the estimated survival curve for female rats.

Summary

This file is only a basic article template. For full details of *The R Journal* style and information on how to prepare your article for submission, see the [Instructions for Authors](#).

Victor Wilson

California Polytechnic State University, San Luis Obispo - Statistics Department

victorjw26@yahoo.com

Ashley Jacobson

California Polytechnic State University, San Luis Obispo - Statistics Department

ashleypjacobson@gmail.com