Introducing parmsurvfit Package - Simple Parametric Survival Analysis with R

by Victor Wilson, Ashley Jacobson

Abstract This article introduces the R package parmsurvfit, 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 probabilites, and computing summary statistics. We describe appropriate usage of these functions, what the output means, and provide examples of how to utilizie 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 desrcibed 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 avaiable 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 proability 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 survial time, mean survival time, and percentiles of survival time. In this package, all of the functions that we developed utilize parametric methods of survival anlaysis, 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 statistics (mean, median, survival probabilities). The majority of this paper will be organized following these groups of functions.

Fitting Right Censored Survival Data

As mentoined 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
```

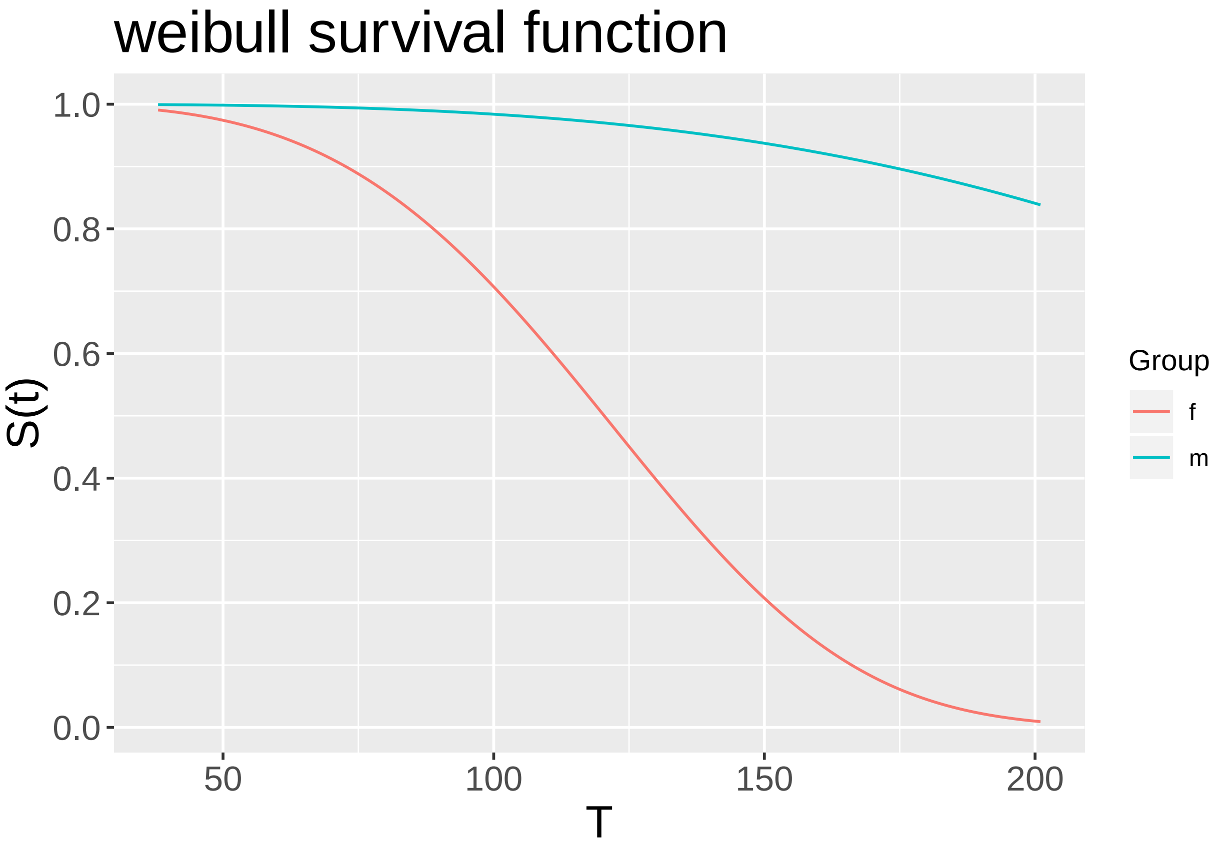
Displaying Plots

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, probability density plots, and cumulative hazard plots. We designed these functions with an intent to have the output displayed be very similar to the output that Minitab displays for these types of plots.

Survival 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 seperate survival plots for each group of the specified variable.

INSERT SURVIVAL FUNCTION EQUATION



In this example, we fit a Weibull distribution to the rats dataset available in the survival package, grouping by the "sex" variable. The rats dataset contains 300 observations, with 3 rats each being selected from 100 litters and 1 rat in each litter being adminstered a drug. The event of interest in this study was whether or not a rat developed a tumor following the beginning of the study. For each rat, the litter number (1-100), type of treatment received (coded as 1 = drug, 0 = control), time until development of tumor or last follow-up (measured in days), final event status (1 = tumor, 0 = censored), and sex were recorded. Below is an excerpt of the data.frame.

```
library(survival)
data(rats)
head(rats)
```

```
#>   litter rx time status sex
#> 1      1  1  101      0   f
#> 2      1  0   49      1   f
#> 3      1  0  104      0   f
#> 4      2  1   91      0   m
#> 5      2  0  104      0   m
#> 6      2  0  102      0   m
```

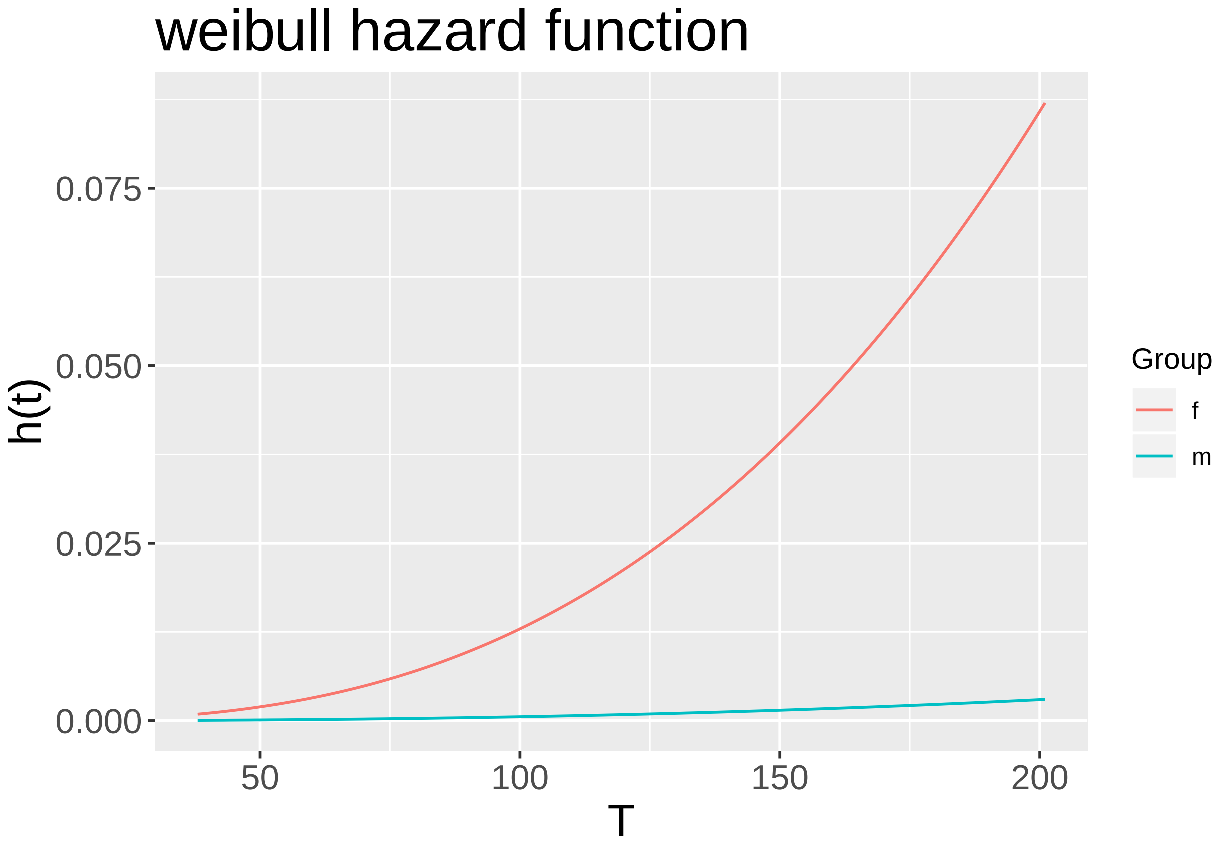
For example, the rat represented in line 2 came from Litter 1, did not receive the drug, experienced the development of a tumor at 49 days, and was a female rat.

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. From this plot, we see that the survival curve for male rats is consistently above the survival curve for female rats throughout all points in time. Due to this, we can conclude that female rats tend to experience the event of interest much more quickly than male rats. This can also be interpreted as male rats tend to survive longer than female rats in this study.

Hazard plots

Hazard plots, on the other hand, are used to display the conditional risk that a subject will experience the event of interest in the next instant of time, given that the subject has survived beyond a certain amount of time. Essentially, the hazard function attempts to assess the risk that an individual who has not yet experienced the event in the very next small amount of time. For example, if we observe that a rat has survived for 75 days already, the hazard function would estimate the risk that the rat will die in the next short instant of time, based on the fact that it has already survived 75 days. We created the "plot_haz" function in order to easily plot hazard functions given that it follows a specified parametric distribution, with the option to include a grouping variable.

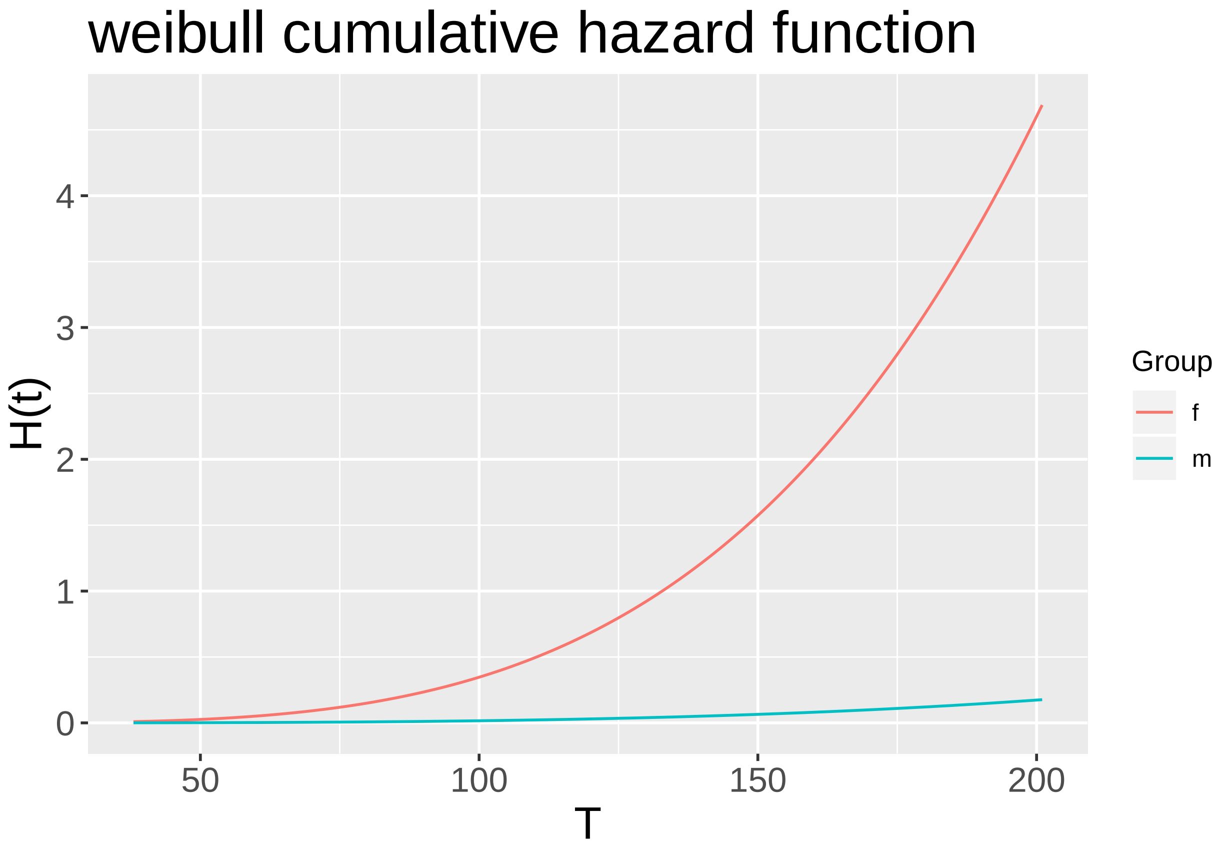
INSERT HAZARD FUNCTION FORMULA



From this plot above, also using the rats dataset available in the survival package, we can see that as female rats continue to survive, their risk of experiencing the event of interest in the next instant of time dramatically increases. Contrastly, male rats do not seem to have a greater risk of experiencing the event of interest as they survive longer. This is demonstrated by the blue line being mostly flat across all points of time.

Cumulative Hazard Plots

While hazard plots are usually useful in assessing a subject's risk of experiencing the event of interest in the next moment of time, these plots can be difficult to read and understand at times. Sometimes, the changes in hazard are very subtle, making it difficult to describe periods of increasing and decreasing risk. In order to accurately assess how hazard rates change over time, we investigate the accumulation of hazard rates over time, known as cumulative hazard. The cumulative hazard function, denoted **H(t)**, is the accumulated risk of experiencing an event up to time **t**. Since the cumulative hazard function is an accumulation of rates, it is important to note that this function is non-decreasing and is hardly ever remains constant by nature. We developed the function "plot_cumhaz" in order to easily display cumulative hazard plots, given that the data follows a specified parametric distribution. The functionality of this function is nearly identical to that of "plot_haz", with the only distinction being that it plots cumulative hazard curves instead of hazard curves.



Computing Survival Probabilities and Summary Statistics

While viewing plots such as those explained above are very useful in Survival Analysis, they only tell half of the story.

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