Introducing parmsurvfit Package - Simple Parametric Survival Analysis with R

by Victor Wilson, Ashley Jacobson, Shannon Pileggi

Abstract This article introduces the R package parmsurvfit, which executes basic parametric survival analysis techniques similar to those in Minitab. Among these are plotting hazard, cumulative hazard, survival, and density curves, computing survival probabilites, and computing summary statistics based on a specified parametric distribution. We describe appropriate usage of these functions, interpretation of output, and provide examples of how to utilize these 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 collected 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) beyond any given time t. 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 address non-parametric survival analysis, such as the **survival** package. Moreover, R does have some packages that aid in estimation for parametric survival analysis, including **fitdistrplus**. However, Minitab has very concise and easy to utilize functions for computing and displaying many parametric 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 parametric survival analysis, which contains all of these commonly utilized statistics and plots.

This paper describes the functions that the **parmsurvfit** package contains, how the data is formatted in order to utilize these functions, and what the output of these functions represent. There are four major groups of functions that we created: assessing fit, survival functions (density, 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.

Assessing fit

Since all of the functions available in this package assume that the survival data follows a known parametric distribution, it is important to have a method to analyze how well our assumed model fits the data. Utilizing such methods will allow us choose a distribution that adequately fits the data. Some common methods used to assess goodness of fit are viewing a histogram of the data, Q-Q (Quantile-Quantile) plots, the Anderson-Darling Test.

Fitting right censored survival data

As mentioned previously, this function is very similar to the function `fitdistcens` found in the fitdistrplus package, which computes the Maximum Likelihood Estimates (MLEs) for right-censored data. The `fitdistcens` function requires data to be organized into two columns, right and left. The

right column indicates a start time, and the left column indicates an end time. For example, if a time is right censored, then the left column would contain the time, and the right column would contain NA. This way of organizing survival data allows for different types of censoring to be unambiguous, but since the **parmsurvfit** package will only handle right censored data, the `fit_data` function makes it so that the user doesn't have to reorganize data.

This function takes in two required columns, a Time column and a Censor column. The time column contains the time-to-event variable, while the censor column indicates whether right censoring is present (0 corresponds to censored data and 1 corresponds to complete data). The function also takes in an optional grouping variable, which fits the data for each group individually. The function returns an object of class '`fitdistcens`', and if there's a grouping variable it returns a list of objects of class '`fitdistcens`'. Several of the other functions available in this package also utilize arguments that are similar to the ones found in this function.

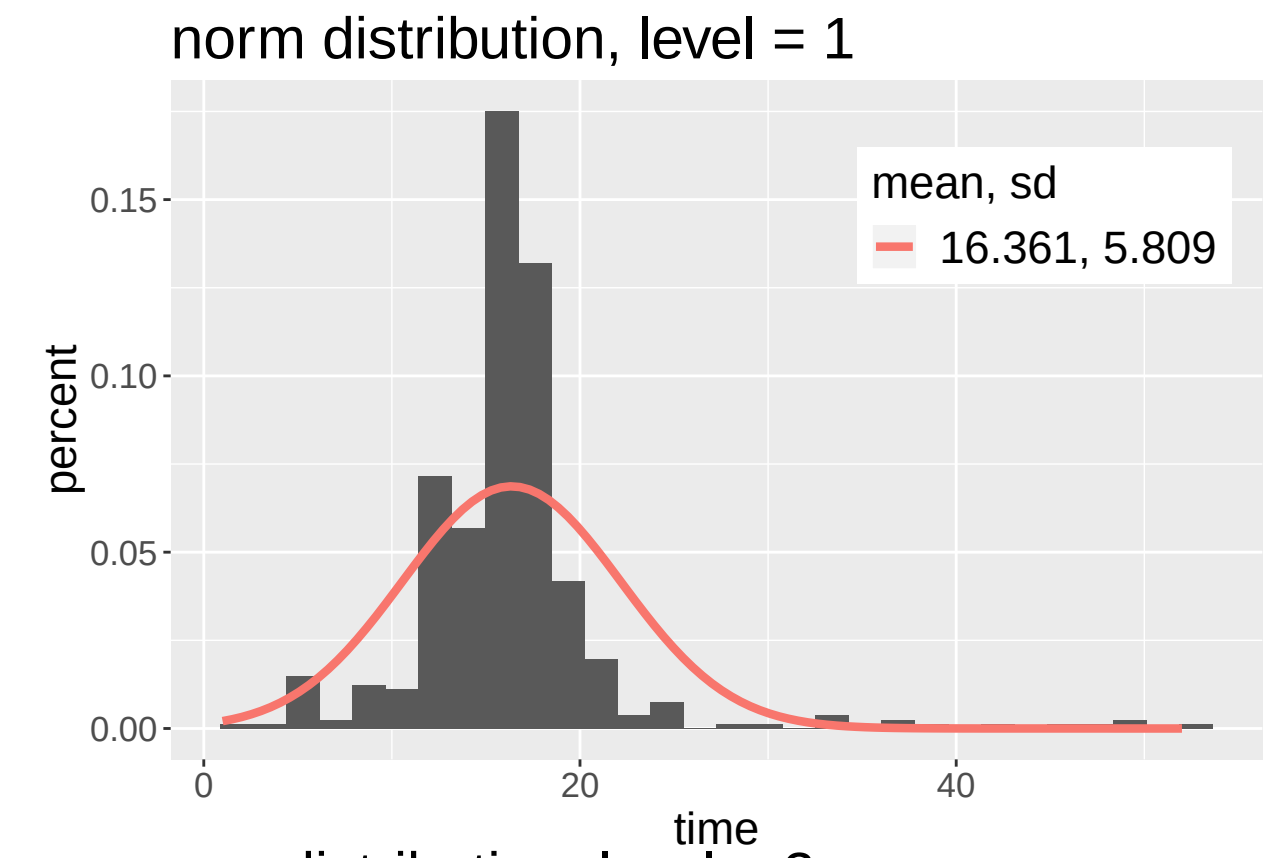
Example

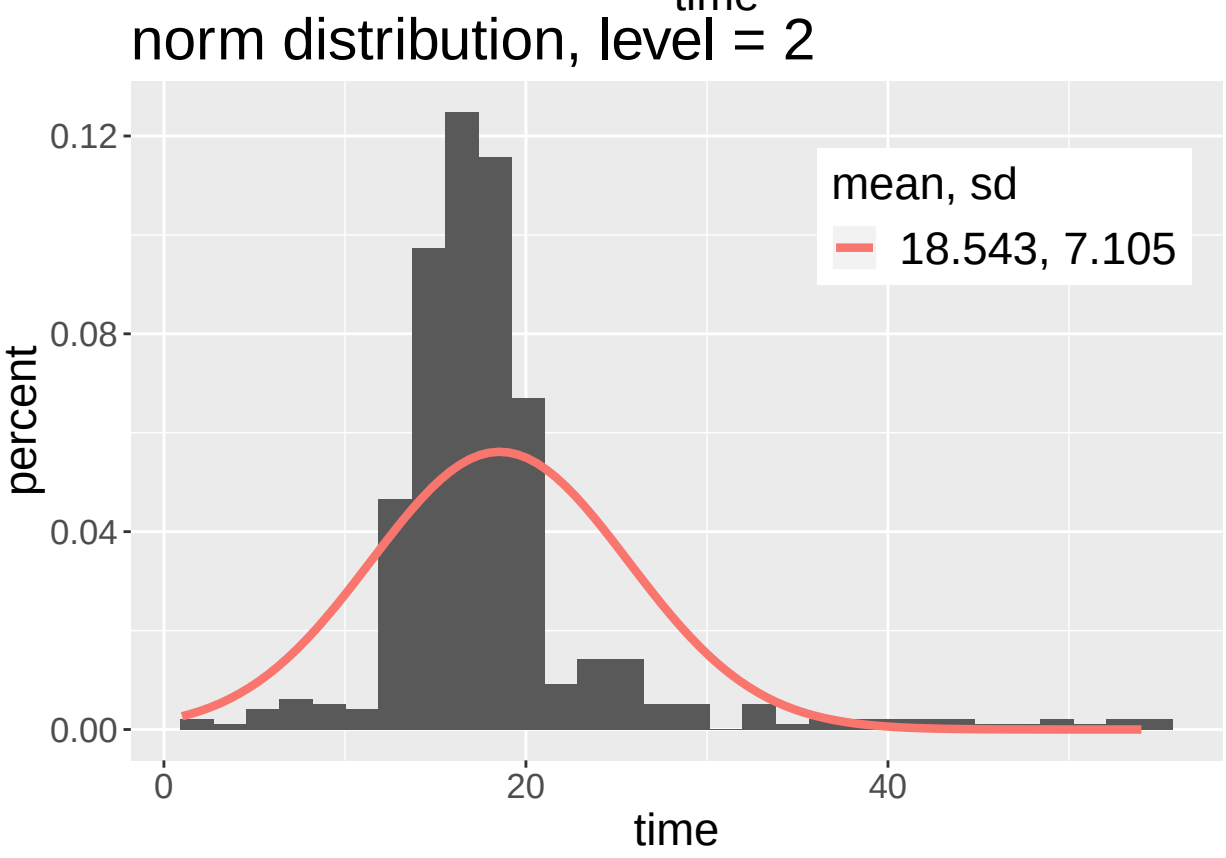
```
#> Fitting of the distribution ' logis ' on censored data by maximum likelihood
#> Parameters:
#>           estimate
#> location 16.741581
#> scale     2.798533
#> Fixed parameters:
#> data frame with 0 columns and 0 rows
```

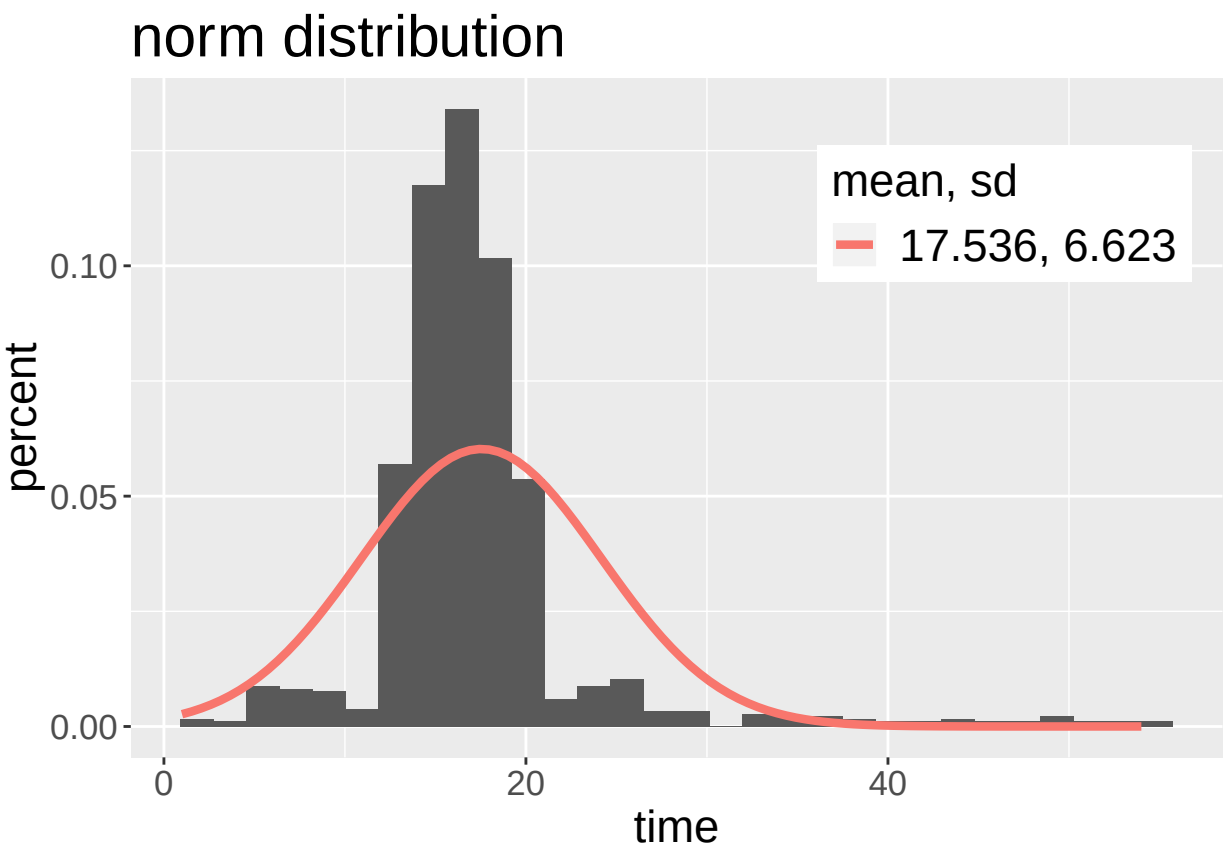
Density plots/histograms

The `plot_density` function creates a histogram of the data and overlays the density function of a fitted parametric distribution. Parameter estimates for the specified parametric distribution are provided as well. This function also supports the ability to plot separate histograms and density functions for each level of a grouping variable. An example of this function is shown below:

```
plot_density(Firstdrink, "norm", time = "Age", by = "Gender")
```


norm distribution, level = 1
mean, sd
16.361, 5.809
percent
0.15
0.10
0.05
0.00
0
20
40
time


norm distribution, level = 2
mean, sd
18.543, 7.105
percent
0.12
0.08
0.04
0.00
0
20
40
time

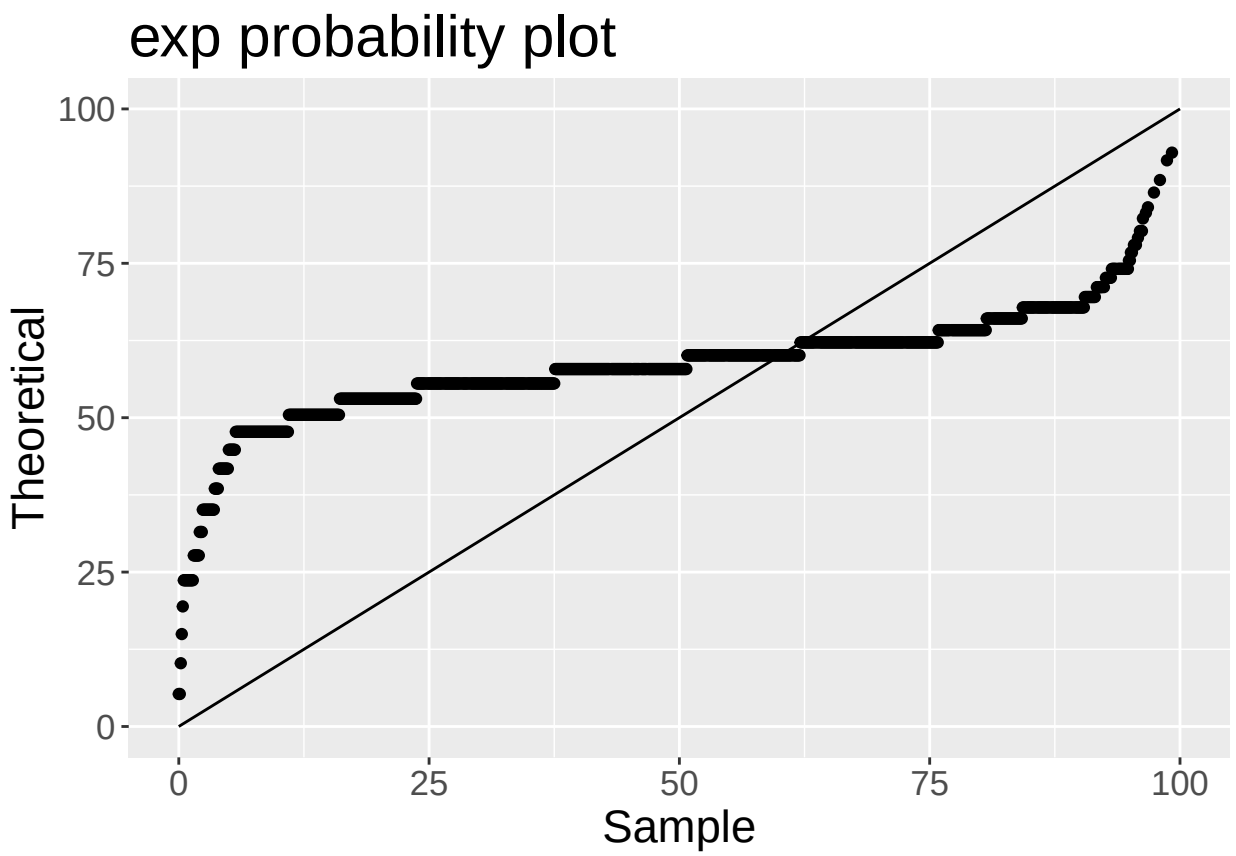


We ran the `plot_density` function, utilizing the 'Firstdrink.txt' dataset available in our package. This dataset contains data on the age of first consumption of alcoholic beverage for 1000 individuals. As seen above, a separate histogram and density plot was created for males and females.

Q-Q plots

As mentioned before, Q-Q plots are a very popular method used to evaluate the fit between two probability distributions. In these plots, the hypothesized quantiles are plotted on one axis and the observed quantiles are plotted on the other axis. A y=x line is typically included in these plots, because if the observed data fit the hypothesized distribution perfectly, all of the points would lie exactly on this line. We developed the `plot_qqsurv` function to create a quantile-quantile plot of right-censored data given that it follows a specified distribution.

```
plot_qqsurv(Firstdrink, "exp", time="Age")
```



We can create a Q-Q plot for the the Firstdrink data set to see how well an Exponential distribution fits the data. As seen in the plot above, there are some deviations from the provided $y=x$ line, indicates that an Exponential distribution may not be an ideal fit for the data.

Anderson-Darling Test

While creating Q-Q plots are a great way to visualize how a particular distribution may fit the data, it can be difficult at times to definitively decide how well the plot fits the data. The Anderson-Darling test provides a numerical test statistic that measures how well the data fits a particular distribution. **CITE MINITAB HELP PAGE**

Survival Functions

This section introduces an overview of the many types of survival functions that are able to be displayed via this package. Some of the most common functions used in Survival Analysis are the survival function, the hazard function, and the cumulative hazard function.

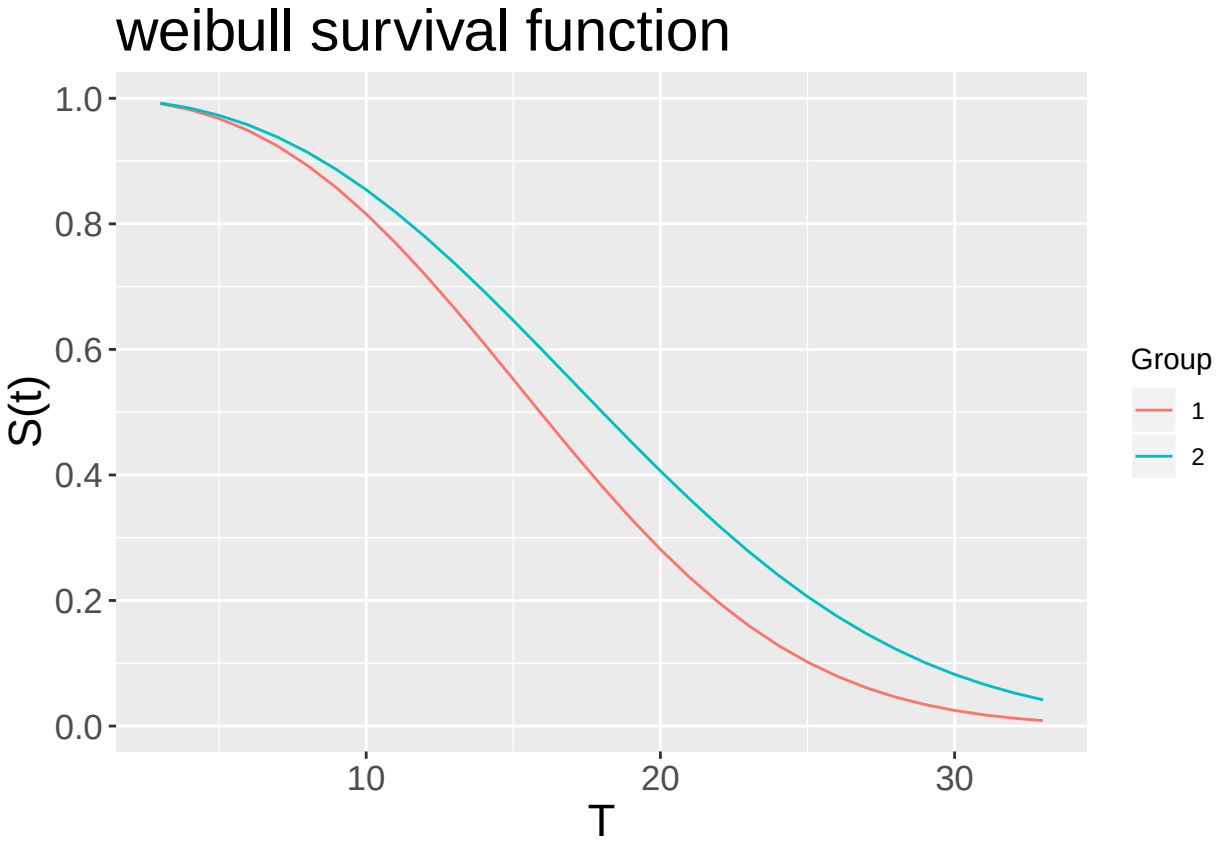
We designed functions within this package to display a plot for each of the aforementioned survival functions with an intent to have the output displayed be very easy to read and interpret. Below is a list of each function and it's relationship to other functions, as well as the formula used to compute each function.

Function	Relationships
PDF	$f(t) = \frac{d}{dt}F(t)$
CDF	$F(t) = \int_0^t f(y)dy$
Survival	$S(t) = 1 - F(t) = \exp[-H(t)] = \exp[-\int_0^t h(y)dy]$
Hazard	$h(t) = f(t)/S(t) = -\frac{d}{dt}\ln[S(t)]$
Cum. Haz.	$H(t) = \int_0^t h(y)dy = -\ln[S(t)]$

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` in an attempt to create hazard plots that are

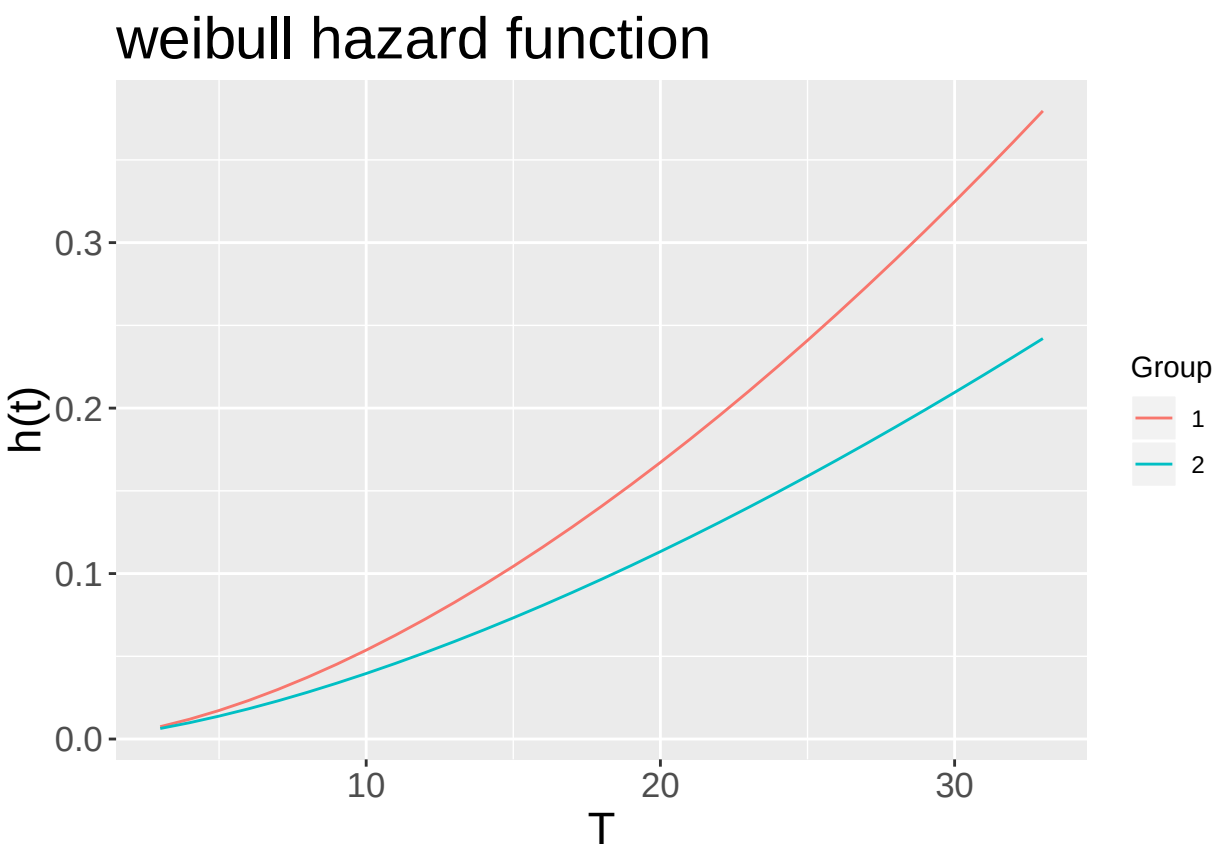
easy to produce, when dealing with data set up for Survival Analysis. This function plots the survival curve of right censored data given that it follows a specified parametric distribution. Some examples of the distributions that this function supports are the Weibull, Log-Normal, Exponential, Normal, and Logistic distributions. This function also provides the option to plot by a grouping variable, which if specified, displays separate curves for each group of the specified variable. In these plots, survival time is plotted on the x-axis, while survival probability is plotted on the y-axis.



In this example, we fit a Weibull distribution to the Firstdrink dataset, grouping by the Gender variable once again. As seen in the plot above, two different survival curves were plotted. The blue line represents the estimated survival curve for males, while the red line represents the estimated survival curve for females. From this plot, we see that the survival curve for female rats is consistently above the survival curve for male rats throughout all points in time. Due to this, we can conclude that males tend to experience their first drink of alcohol before males do.

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 person has survived for 17 years without first trying alcohol, the hazard function would estimate the risk that the person will experience their first drink of alcohol in the next short instant of time, based on the fact that it has already survived 17 years. 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.

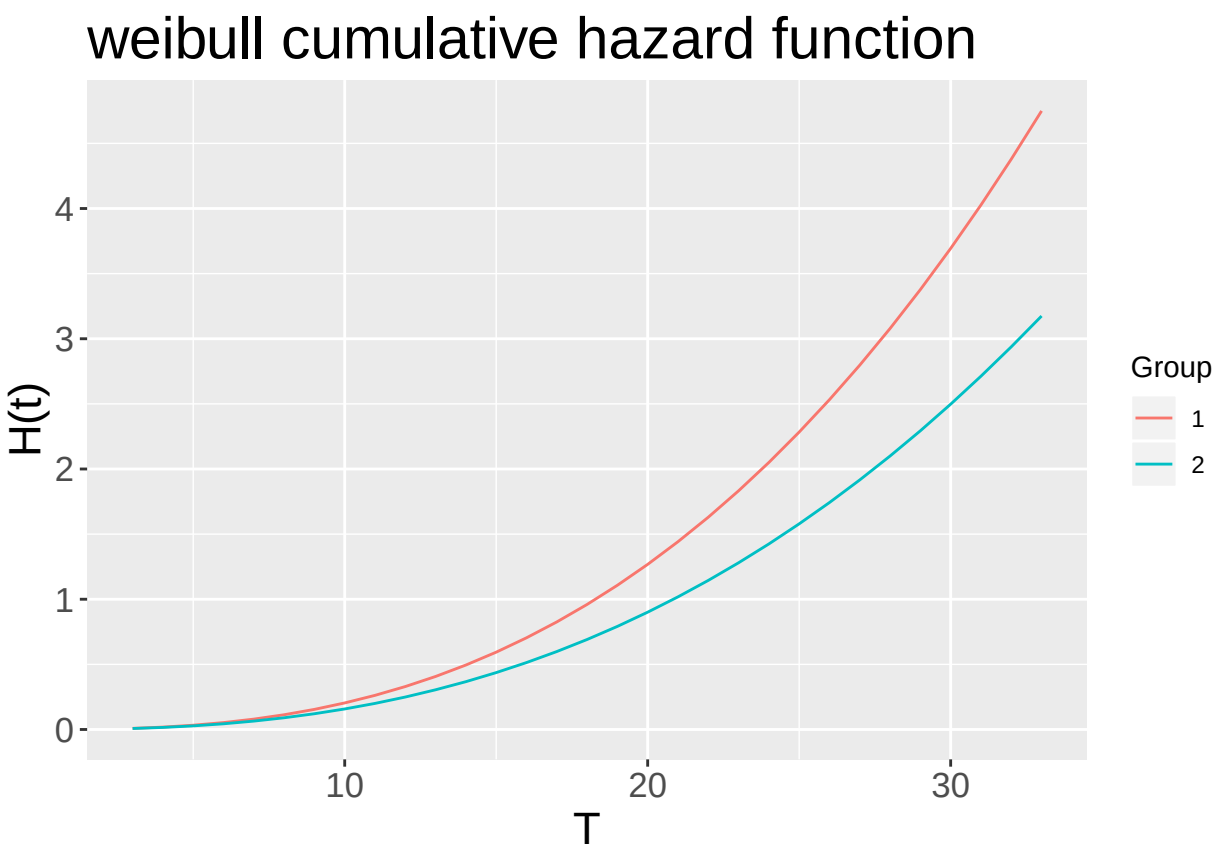


From this plot above, also using the Firstdrink dataset, we can see that as males continue to survive, their risk of experiencing the event of interest in the next instant of time dramatically increases. Similarly, females also seem to have a greater risk of experiencing the event of interest as they survive longer, but their risk is lower than that of males.

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.



As we can see, the cumulative hazard function is increasing for both males and females. This means that the risk for experiencing the first drink of alcohol for both males and females in the next instant of time is increasing over time, given that they have survived beyond time **t**.

Statistics/Computations

While viewing plots such as those explained above are very useful in survival analysis, they only tell half of the story. In order to carry out a complete survival analysis, we must also compute statistics in order to supplement our plots. Some of the most common statistics utilized in parametric survival analysis are survival probabilities and typical summary statistics such as the mean, median, standard deviation, and percentiles of survival time.

Computing survival probabilites

Being able to compute survival probabilities is especially of interest because it estimates the probability that a subject will not have experienced the event of interest beyond a specified time t. We developed the function `surv_prob` to compute probability of survival beyond time t, given that the data follows a specified parametric distribution.

```
library(parmsurvfit)
surv_prob(Firstdrink, "lnorm", 30, time="Age")

#> P(T > 30) = 0.05191602
```

As seen in the output from the function above, utilizing the Firstdrink data set and fitting a log-normal parametric distribution to the data, the estimated probability that a person survives for 30 years without having their first drink of alcohol is roughly 5%.

Computing summary statistics

Another useful form of output that we believed would be useful to also have in R is a table of summary statistics. Summary statistics that are typically included are the mean, standard deviation, median, and IQR. The surv_summary function that we developed estimates various summary statistics, including mean, median, standard deviation, and percentiles of survival time given that the data follows a

specified parametric distribution. This function also supports the option to provide separate summary statistics for each level of a grouping variable, if desired.

```
library(parmsurvfit)
surv_summary(Firstdrink, "lnorm", time="Age", by="Gender")
```

```
#>
#>
#> For level = 1
#> meanlog      2.743253
#> sdlog        0.3505774
#> Log Liklihood    -1370.309
#> AIC      2744.619
#> BIC      2752.885
#> Mean     16.52221
#> StDev        5.974932
#> First Quantile   12.26552
#> Median       15.53745
#> Third Quantile   19.68219
#>
#> For level = 2
#> meanlog      2.862891
#> sdlog        0.3674697
#> Log Liklihood    -1658.704
#> AIC      3321.408
#> BIC      3329.987
#> Mean     18.73528
#> StDev        7.123736
#> First Quantile   13.66772
#> Median       17.51209
#> Third Quantile   22.43778
```

As seen above, after specifying the grouping variable of sex, two separate tables were produced, one for males and one for females. We can see that the mean log survival time for males is smaller than the mean log survival time for females. The standard deviation of log survival time for males was also much smaller than that of females.

Conclusion

The R package **parmsurvfit** allows for parametric Survival Analysis methods involving right-censored data to be easily computed in R. The overall goal of developing this package was to provide a central package for R users to find typical methods used in Survival Analysis such as computing survival probabilities, create survival and hazard plots, and assess goodness of fit of a parametric distribution fit to a dataset.

Acknowledgments

This package would not have been possible without the generous donation of Bill and Linda Frost. We would also like to thank Dr. Shannon Pileggi for conceptualization of the package, as well as for guidance and assistance in issues that we encountered during the development of this package.

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

Shannon Pileggi
California Polytechnic State University, San Luis Obispo - Statistics Department

spileggi@calpoly.edu