Introducing parmsurvfit Package - Simple Parametric Models for Survival Analysis in 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 endpoints such as death, development of disease, or first score of a soccer game. Observations in survival analysis may also be subject to censoring, which occurs when an observation's actual time to event is unknown, but is known to be within some specified range. For example, if a subject does not experience the event of interest prior to study completion, then their "observed" event time (time until study completion) is less than their "actual" event time. This particular case is known as right-censoring, which is the most common form of censoring. Right censoring can also occur if a subject drops out before the end of the study and does not experience the event of interest.

Typically, survival analysis begins with computing what are known as the survival and hazard functions (Kleinbaum and Klein, 2012). The survival function estimates the probability of surviving (also known as not experiencing the event of interest) beyond time t. The hazard function computes the conditional risk 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. Summary statistics that are commonly reported include median survival time, mean survival time, and percentiles of survival time. These functions and statistics can be computed by either parametric or nonparametric techniques. In the **parmsurvfit** package, all of the functions that we developed utilize parametric methods for survival analysis, which assumes that the distribution of the survival times follows a known probability distribution.

R has many survival packages that address non-parametric survival analysis, such as the **survival** package. However, prior to the **parmsurvfit** package, parametric models for survival data and corresponding summary statistics was not easily attained. The **parmsurvfit** package attempts to emulate the ease and functionality of parametric survival analysis features available in Minitab.

This paper describes example survival data available in the **parmsurvfit** package, the functions that the package contains, how the data is formatted in order to utilize these functions, and what the output of these functions represent. There are three major groups of functions that we created: assessing fit, survival functions (density, survival, hazard, cumulative hazard), and computing statistics (mean, median, survival probabilities). Explanations are presented according to these groups of functions.

Data

The **parmsurvfit** package contains five data sets with observations subject to right-censoring (`aggressive`, `firstdrink`, `graduate`, `oscars`, and `rearrest`). Subsequent examples in this paper are based on the `firstdrink` data set, which contains 1,000 records from the 1990-1992 National Comorbidity Survey regarding age at first drink of alcohol (`age`). An observation is recorded as complete (`censor = 1`) if the age at first drink of alcohol is known. An observation is recorded as incomplete, or right-censored (`censor = 0`) if the subject had not yet consumed an alcoholic beverage at the time of the study execution. In this case, the subject's "actual" event time (age at first drink) is only known to exceed their "observed" event time (age at time of study). This data set also includes a `gender` variable such that `1` corresponds to males and `2` corresponds to females.

Fitting right-censored survival data

The fit_data function produces maximum likelihood estimates (MLE) for right censored data based on the input distribution. The fit_data function utilizes the fitdistcens function in the **fitdistrplus** package, but allows for more intuitive input of right-censored data than as specified with fitdistcens (which allows input of other types of censoring). The fit_data function in the **parmsurvfit** package inputs two variables: time and censor. The time variable contains the time-to-event variable, while the censor variable indicates whether right censoring is present (0 corresponds to censored data and 1 corresponds to complete data). Furthermore, the user specifies the desired parametric distribution in dist by inputting the base name of the distribution as a character string. For example, to utilize the normal distribution you would specify 'norm' as it is the base of dnorm, pnorm, etc. Commonly utilized distributions for survival analysis include Weibull ('weibull'), log-normal ('lnorm'), exponential ('exp'), and logistic ('logis') distributions. 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 is a grouping variable it returns a list of objects of class 'fitdistcens'.

Example

In this example, we fit the Weibull distribution to the 'firstdrink' data set where the time to event variable is age and the variable that indicates censoring status is censor. The maximum likelihood estimates of the location and scale parameters are returned.

```
library(parmsurvfit)
fit_data(data = firstdrink, dist = "weibull", time = "age", censor = "censor")

#> Fitting of the distribution ' weibull ' on censored data by maximum likelihood
#> Parameters:
#>        estimate
#> shape  2.536106
#> scale 19.684061
#> Fixed parameters:
#> data frame with 0 columns and 0 rows
```

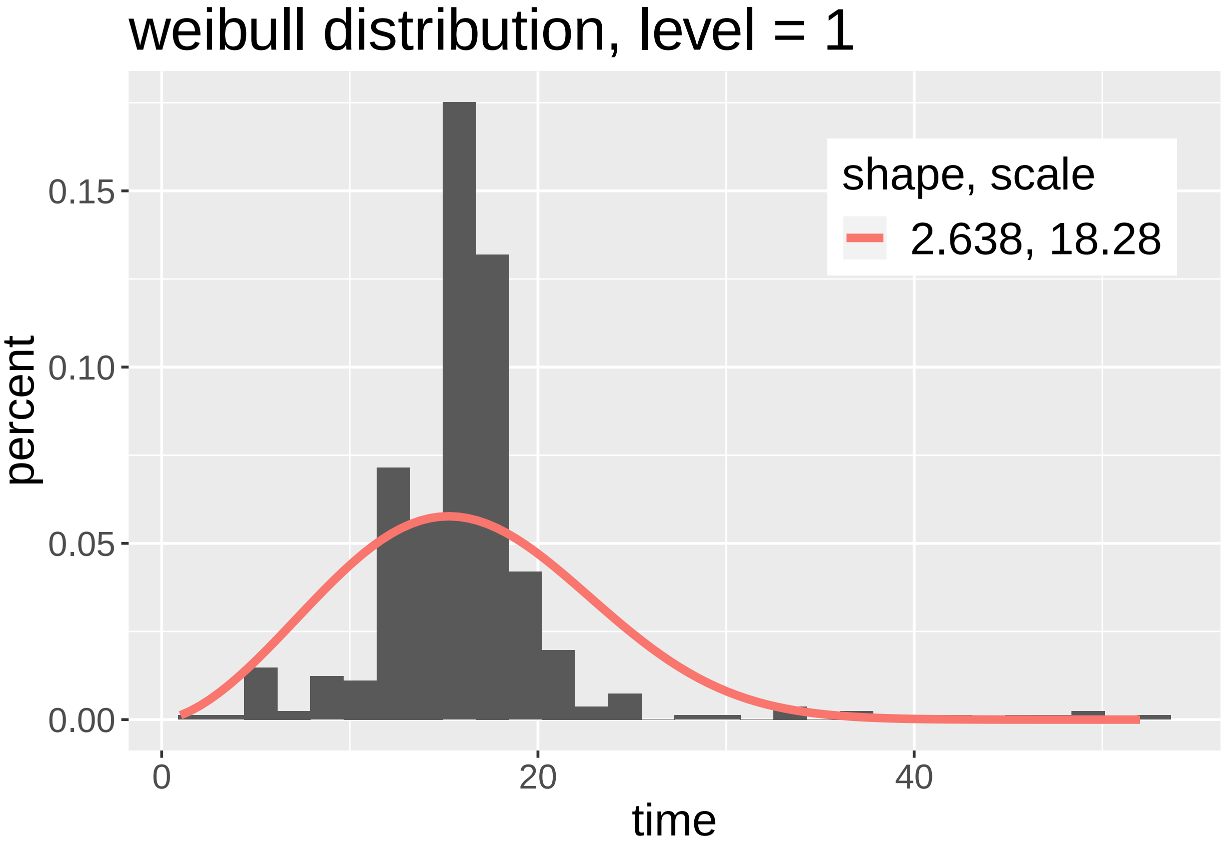
Assessing fit

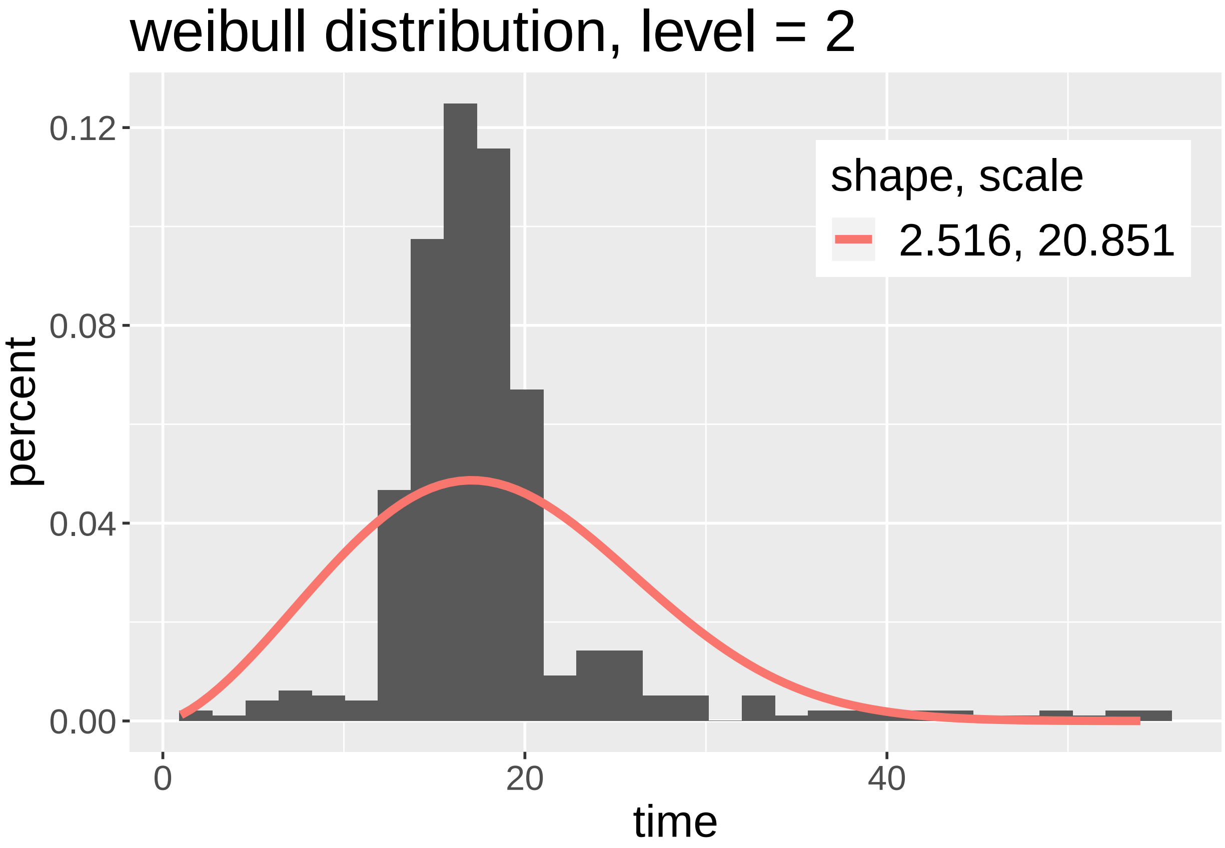
Since all of the functions available in this package assume that the survival data follows a known parametric distribution, we presents methods to evaluate how well the assumed model fits the data. Some common methods used to assess goodness of fit are viewing a histogram of the data, Q-Q (Quantile-Quantile) plots, the Anderson-Darling test statistic.

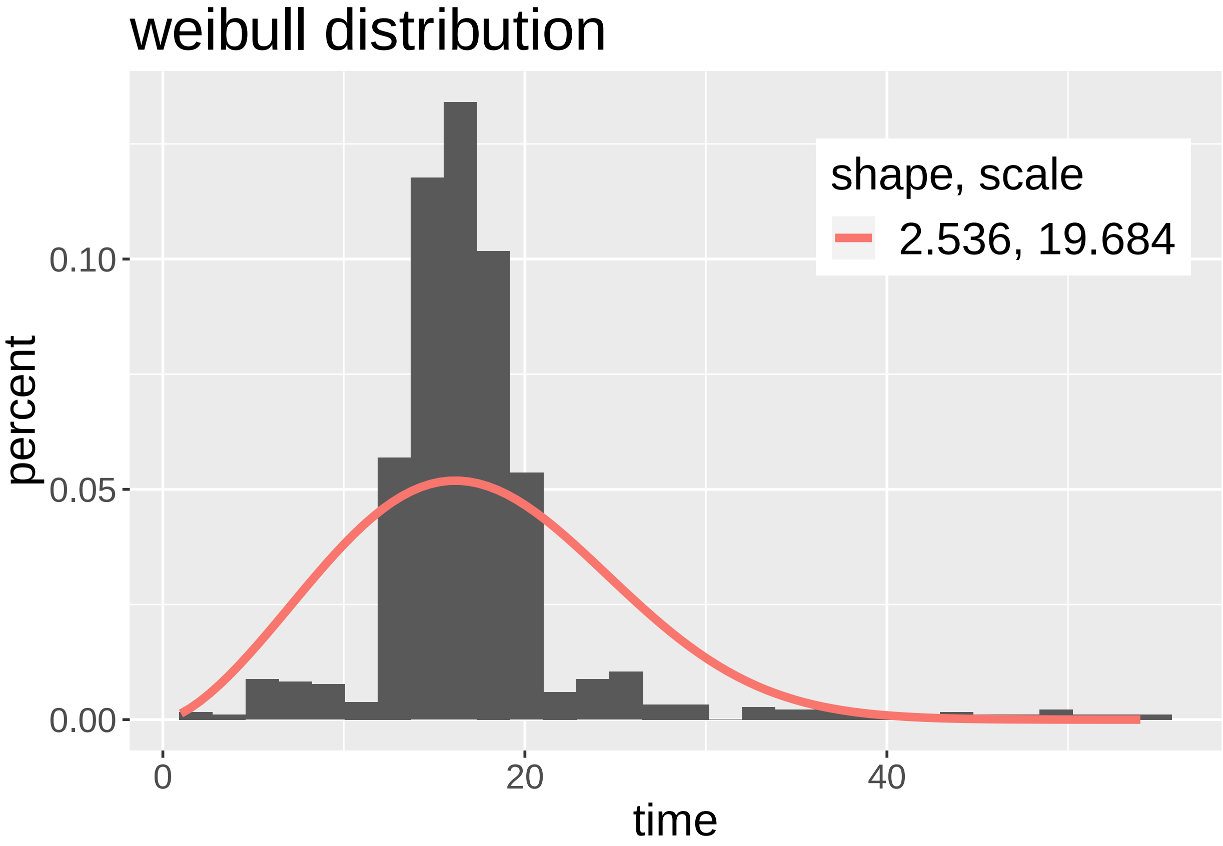
Histograms with density curves

The plot_density function produces histograms with overlayed density curves to allow a user to visually assess fit of a parametric distribution to data. 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. Below, we fit the weibull distribution to age until first drink by gender. Three plots are produced, each based on their respective MLE's: a plot for males (level = 1), females (level = 2), and overall. In these plots, all time to event data are plotted regardless of censoring status.

```
plot_density(data = firstdrink, dist = "weibull", time = "age", censor = "censor", by = "gender")
```


weibull distribution, level = 1
percent
0.15
0.10
0.05
0.00
0
20
40
time
shape, scale
2.638, 18.28


weibull distribution, level = 2
percent
0.12
0.08
0.04
0.00
0
20
40
time
shape, scale
2.516, 20.851



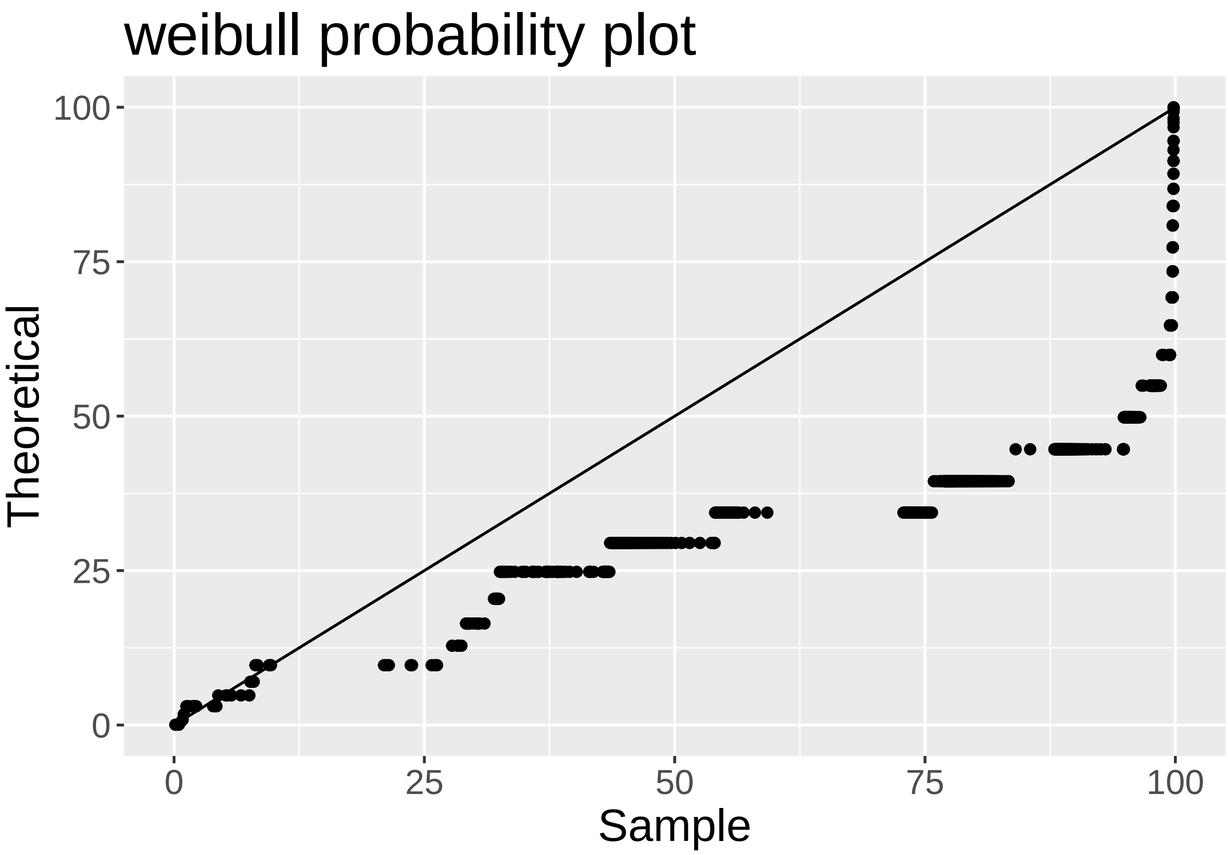
Here, we observe that the Weibull distribution could possibly be too flat for the distribution of age until first drink of alcohol.

Q-Q plots

Quantiles vs percentileS!!!

The `plot_qqsurv` function creates a quantile-quantile plot of right-censored data given that it follows a specified distribution. In typical Q-Q plots the hypothesized (theoretical) quantiles are plotted on the y-axis and the observed (sample) quantiles are plotted on the x-axis. A $y = x$ line is included in these plots, because if the observed data fit the hypothesized distribution perfectly, all of the points would lie exactly on this diagonal line. Here, the points are plotted according to the median rank method (Minitab, b) to accommodate the right-censoring features.

```
plot_qqsurv(data = firstdrink, dist = "weibull", time = "age", censor = "censor")
```



As seen in the Q-Q plot, there are some deviations from the provided $y = x$ line, indicating that a Weibull distribution may not be an ideal fit for the data.

Anderson-Darling test statistic

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 `compute_AD` function computes the Anderson-Darling (AD) test statistic, which provides a numerical test statistic that measures how well a particular distribution the data fits such that lower values indicate a better fit. Computation of the test statistic adhered to Minitab's documentation, utilizing the median rank plotting method (Minitab, a).

```
compute_AD(data = firstdrink, dist = "weibull", time = "age", censor = "censor")
```

```
#> [1] 315.5693
```

Here, the AD test statistic value is 315.5693; this can be used to judge fit relative to another computed AD test statistic.

Survival functions

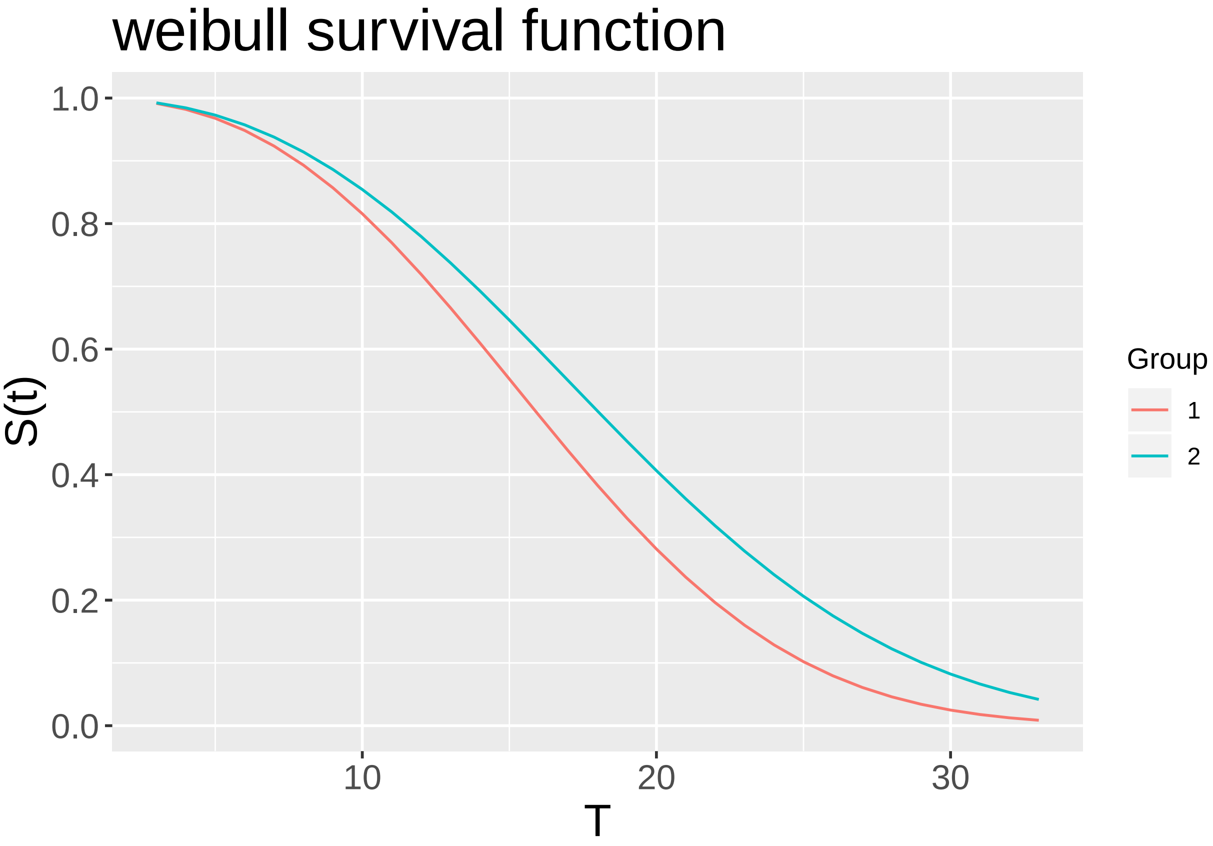
This section provides an overview of the survival functions available in this package. Some of the most common functions used in survival analysis are the survival function, the hazard function, and the cumulative hazard function. Table ?? lists each function and its relationship to the other functions.

Survival plots

The survival function $S(t)$ estimates the proportion of subjects that survive beyond a specified time t. The `plot_surv` function plots the survival curve of right censored data given a specified parametric distribution. This function also provides the option to plot by a grouping variable, which if specified, displays separate curves for each value of the specified grouping variable. In these plots, survival time is plotted on the x-axis, while survival probability is plotted on the y-axis.

Function	Relationship
PDF	$f(t) = \frac{d}{dt}F(t)$
CDF	$F(t) = \int_0^t f(y)dy$
Survival	$S(t) = 1 - F(t)$
Hazard	$h(t) = f(t)/S(t)$
Cumulative hazard	$H(t) = -\ln[S(t)]$

```
plot_surv(data = firstdrink, dist = "weibull", time = "age", censor = "censor", by = "gender")
```

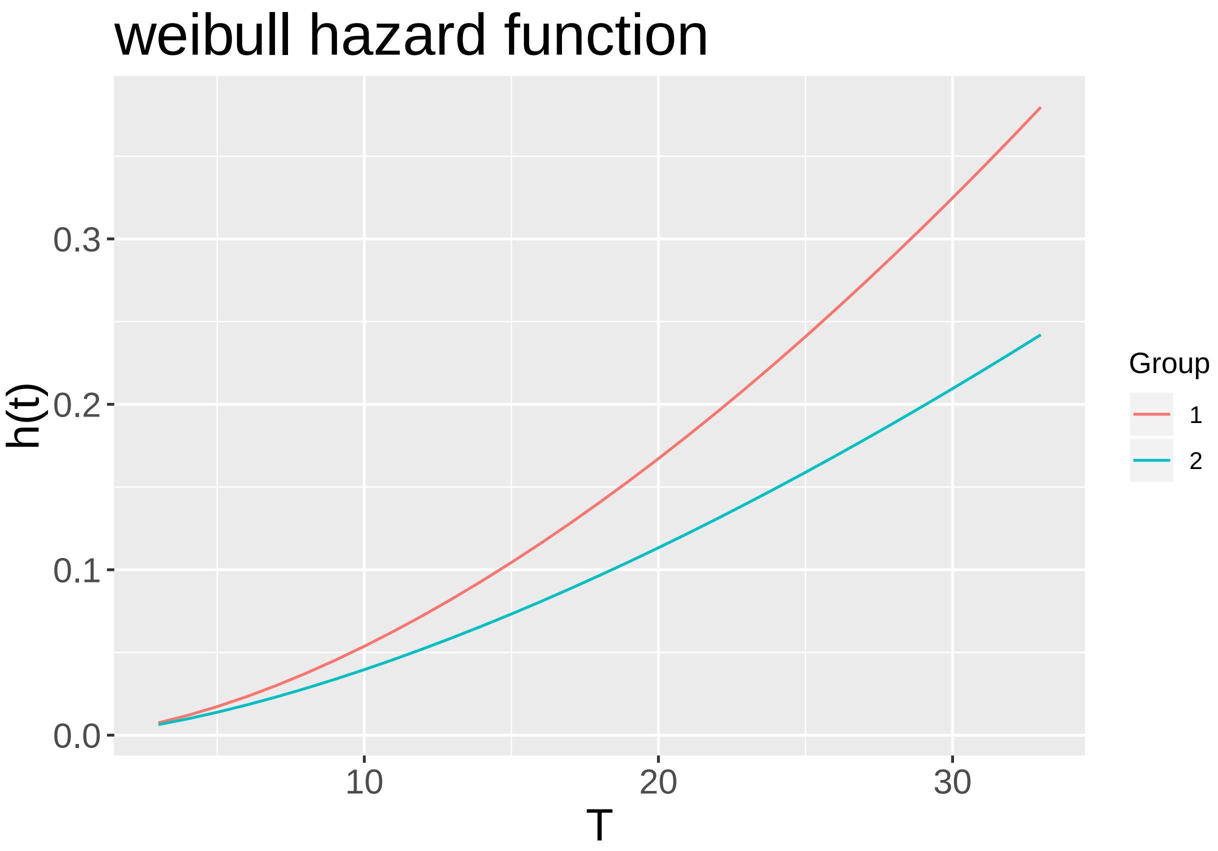


In this example, we fit a Weibull distribution to the `firstdrink` dataset, grouping by the `gender` variable. Here, the blue line represents the estimated survival curve for males (`group = 1`), while the red line represents the estimated survival curve for females (`group = 2`). From this plot, we see that the survival curve for females is consistently above the survival curve for males throughout all points in time. Due to this, we can conclude that males tend to experience their first drink of alcohol before females do.

Hazard function

The hazard function, denoted $h(t)$, esimates 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 time t. 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 the person has already survived 17 years alcohol free. However, hazard is a rate and not probability, and therefore can take values greater than one. Moreover, the hazard function can be both increasing or decreasing. The `plot_haz` function plots the hazard function based on specified parametric distribution, with the option to include a grouping variable.

```
plot_haz(data = firstdrink, dist = "weibull", time = "age", censor = "censor",  by = "gender")
```



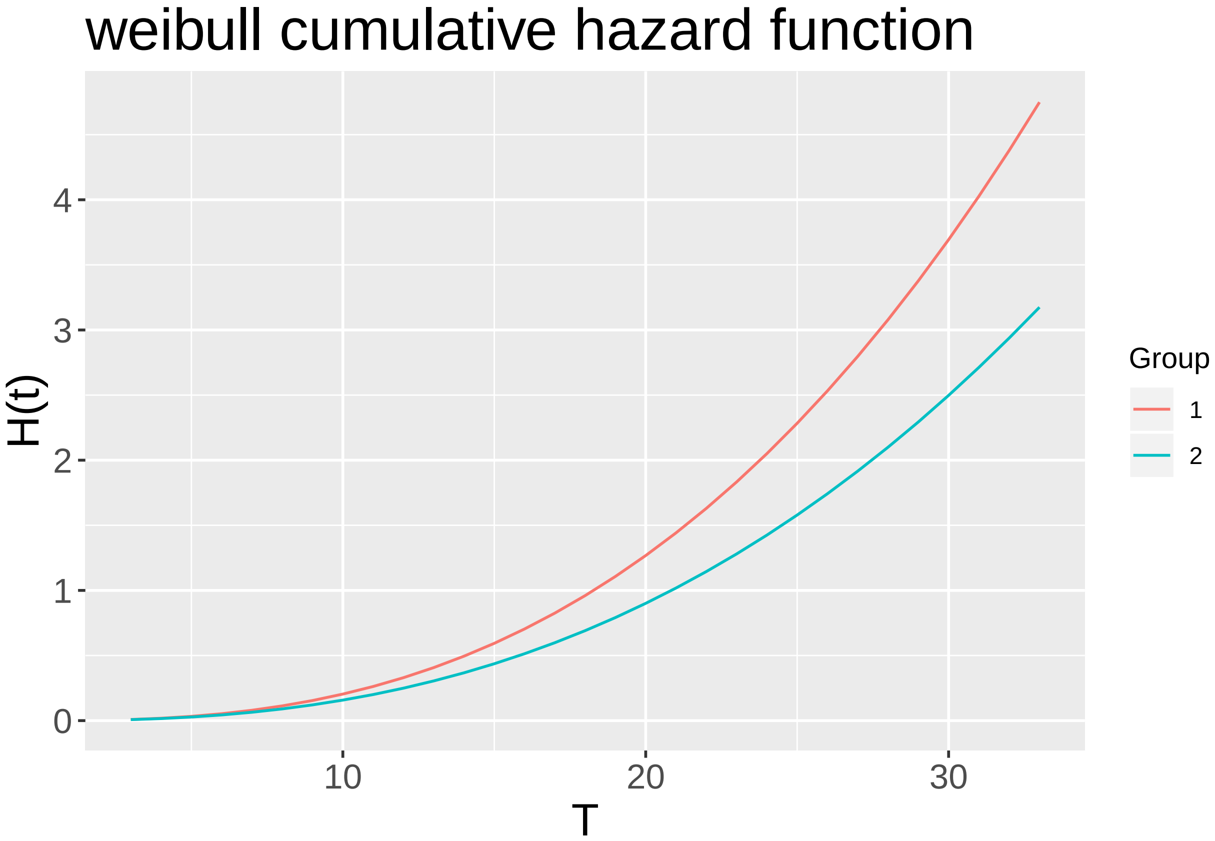
Based on the estimated hazard function, we see that as males and females continue to “survive” (not yet experience their first drink of alcohol), their risk of experiencing the event of interest (drinking) in the next instant of time increases.
This risk increases more dramatically for males relative to females.

Cumulative hazard function

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 total 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. The `plot_cumhaz` function displays cumulative hazard plots, given that the data follows a specified parametric distribution.

```
plot_cumhaz(data = firstdrink, dist = "weibull", time = "age", censor = "censor",  by = "gender")
```



As expected, the cumulative hazard function is increasing for both males and females. Here, the total accumulated risk of experiencing the first drink of alcohol is consistently greater for males compared to females.

Probabilities and Statistics

Survival analysis is typically accompanied by reporting summary statistics such as the mean, median, standard deviation, and percentiles of survival time. In addition, it is also common to compute survival probabilities.

Survival probability

A survival probability estimates the probability that a subject survives (does not experience the event of interest) beyond a specified time t. The function `surv_prob` computes the probability of survival beyond time t, given that the data follows a specified parametric distribution. The `num` argument specifies the time t of interest. The default for `surv_prob` is to compute an upper tail probability; this can be reversed to a lower tail probability using the `lower.tail` argument. The function can also optionally take a grouping variable with the by argument.

```
surv_prob(data = firstdrink, dist = "weibull", num = 30, lower.tail = F, time = "age", censor = "censor", by

#>
#> For level = 1
#> P(T > 30) = 0.02488195
#>
#> For level = 2
#> P(T > 30) = 0.08227309
#>
#> For all levels
#> P(T > 30) = 0.05439142
```

Given a Weibull distribution, the overall estimated probability that a person survives beyond 30 years without having their first drink of alcohol is approximately 0.05. However, males (`level = 1`) are less likely to survive beyond 30 years compared to females (`level = 2`).

Summary statistics

The `surv_summary` function 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 All summary statistics from the class '`fitdistcens`' are provided. If the distribution supplied is one of normal, lognormal, exponential, weibull, or logistic then the standard deviation reported is an exact computation from parameter estimates; however, if a user specifies a distribution other than that from this list, then the standard deviation is estimated from 1,000 randomly generated values from the distribution.

```
surv_summary(data = firstdrink, dist = "weibull", time = "age", censor = "censor", by = "gender")

#>
#>
#> For level = 1
#> shape        2.637645
#> scale        18.2804
#> Log Liklihood    -1425.271
#> AIC      2854.541
#> BIC      2862.808
#> Mean     16.24398
#> StDev        6.625303
#> First Quantile   11.39844
#> Median       15.90884
#> Third Quantile   20.6903
#>
#> For level = 2
#> shape        2.516025
#> scale        20.85053
#> Log Liklihood    -1730.273
#> AIC      3464.546
#> BIC      3473.126
#> Mean     18.50288
#> StDev        7.872356
#> First Quantile   12.70752
#> Median       18.02407
#> Third Quantile   23.74094
```

Utilizing the grouping variable of gender, produces two separate tables of summary statistics for males and females, respectively. The mean survival time for males (16.2 years) is less than the mean survival time for females (18.5 years).

Conclusion

The **parmsurvfit** package 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 implement typical methods used in parametric survival analysis such as computing survival probabilities, creating survival and hazard plots, and assessing goodness of fit of a parametric distribution to a dataset.

Acknowledgments

This package development was funded by the Bill and Linda Frost fund at California Polytechnic State University, San Luis Obispo. We would also like to thank Jeff Sklar for his input and guidance.

Bibliography

D. G. Kleinbaum and M. Klein. *Survival Analysis - A Self-Learning Text*. Springer, 2012. ISBN 1441966455. [p1]

Minitab. Methods and formulas for goodness-of-fit measures in Parametric Distribution Analysis (Right Censoring), a. URL https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/reliability/how-to/parametric-distribution-analysis-right-censoring/methods-and-formulas/goodness-of-fit-measures/. [p5]

Minitab. Methods and formulas for probability plot in Parametric Distribution Analysis (Right Censoring), b. URL https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/reliability/how-to/parametric-distribution-analysis-right-censoring/methods-and-formulas/probability-plot/. [p4]

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