SURVIVAL ANALYSIS: HOW TO?

A Technical Report

Presented to

The Department of Mathematics and Physics

by

Cherubin Makembele & Julie Binombo

STUDENT NO. 220418357, 219386773

As an Assessment for the Module

BIOSTATISTICS 4 **BMS470S**

Within the Qualification

ADVANCED DIPLOMA IN MATHEMATICAL SCIENCES

Lectured by: Dr Mozart Nsuami

Cape Peninsula University of Technology

November 2023

DECLARATION

We, Cherubin Makembele & Julie Binombo, declare that the contents of this research report represent our own work. We know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed and has been cited and referenced.

We have not allowed and will not allow anyone to copy my work to pass it off as his or her work.

Click on the box to confirm your agreement: ☒

Date of Declaration: 14 November 2023

CONTRIBUTION

Author 1: Cherubin Makembele

- Conceived and designed the analysis
- -Analysis tools (R programming)
- -Report Mounting
- -Analysis performance

Author 2: Julie Binombo

- Collected the data
- -Analysis Tools (R programming)
- -Analysis performance

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

- CDF: Cumulative Distribution Function
- $H(t)$: Cumulative Hazard
- δi: Event Indicator
- t: Time
- t: Increment of time
- u: Variable in the integral for cumulative hazard
- T: Increment of time for probability calculation
- $t \to 0$: As time increment approaches zero
- $H(t)$: Instantaneous Event Rate
- $S(t)$: Survival Function

GLOSSARY OF TERMS

- **Survival Time**: The duration until a specific event occurs.
- **Failure Event**: The occurrence of the event of interest, often referred to as an "event" or "nonevent" (failure).
- **Incomplete Response**: Survival time response for a subject that is not fully determined due to loss of follow-up, subject withdrawal, or subject death.
- **Cumulative Distribution Function (CDF):** Represents the probability that the event of interest occurs by a certain time.
- **Smooth Survival Function**: Theoretical representation of the survival function without discrete time intervals.
- **Non-Increasing:** Property of the survival function where the probability of survival does not increase over time.
- **Cumulative Hazard (H(t)):** The accumulated risk up to time t.
- Instantaneous Event Rate: Describes the probability that an event will occur in the next instant given no previous events.
- **Logarithmic Transformation:** Transformation used to derive the hazard function from the survival function.
- **Event Indicator**: Variable indicating whether an event was observed (1) or censored (0).
- **Risk in Analysis**: The potential for errors, biases, or uncertainties in survival analysis.
- **Parametric Estimation**: Estimation method assuming a specific distribution for survival times.
- **Non-Parametric** Estimation: Estimation method not assuming a specific distribution for survival times.
- **Follow-Up Loss**: Occurs when subjects are no longer being monitored during a study.
- **Withdrawal**: The voluntary act of a subject removing himsemf/herself from a study.

- **Censoring Assumption**: The assumption that censoring is non-informative about the event, not caused by the impending failure.
- **Risk Set:** The set of subjects at risk of experiencing an event at a specific time.
- **Survival Curve**: A graphical representation of the survival function over time.
- **Confounding**: A variable that is related to both the exposure and the outcome, potentially affecting the study results.
- **Cumulative incidence:** is a measure used in epidemiology to describe the probability of a specific event occurring within a specified period. It is often employed in the context of survival analysis, especially when dealing with the occurrence of diseases or other health-related events.
- **Longitudinal** often refers to studies or research that involves the repeated observation or examination of a set of subjects over time with respect to one or more study variables1. For example, a longitudinal study of heart transplant recipients over a five-year period1.
-

1. INTRODUCTION

In statistics, to answer a question, cohort studies of cohort analysis are of order. A cohort study is a type of longitudinal and observational study that follows a group of participants over a period examining a certain factor such as exposure to disease. There are two types of cohort studies:

- **Prospective cohort study:** It involves recruiting a group of participants and following them over time to gather new data.
- **Retrospective cohort study**: in the study, the scientists use data that are already available for a particular group.

To perform survival analysis, it is important to use a cohort study structure where it's possible to calculate the time spent for every individual until some event (event of interest) (Frost, 2023).

Survival analysis is a statistical approach utilized on time to event data. It is a popular method used in various fields such as engineering, social sciences, epidemiology, etc. to analyze the time before a certain event of interest happens. The event of interest is often referred to as "event" or "nonevent" (failure) and can range from anything to failure of a mechanical component, or the death of a living organism, time until a football player hits the bar when shooting a ball from a certain distance. Given all the information above, some instances of survival analysis are:

- Time until a person dies.
- Time until a person recovers from a disease following a treatment.
- Time until AIDS for HIV patients
- Time until a component fails in an engine, and many more (Rai et al., 2021).

The survival time response for a subject can be incompletely determined because of the loss of the subject to follow up, the withdrawal of the subject from the study or the death of the subject. Incompletely observed responses are called "**censored observations**" (Colombia University, 2004).

To conduct a survival analysis, it’s vital that individuals familiarize themselves with some concepts, functions such as:

1.1 Survival Function:

It is denoted S(t) is a fundamental concept in survival analysis. It represents the probability that a subject will survive past a specified point in time “t”. The survival function is smooth in theory. In actuality, events are observed throughout discrete time periods (days, weeks, etc.). Mathematically, a survival function is expressed as

$\mathrm{S(t)\ =\ P\,r(T\ >\ t)\ =\ 1\ -\ F(t)}$, where:

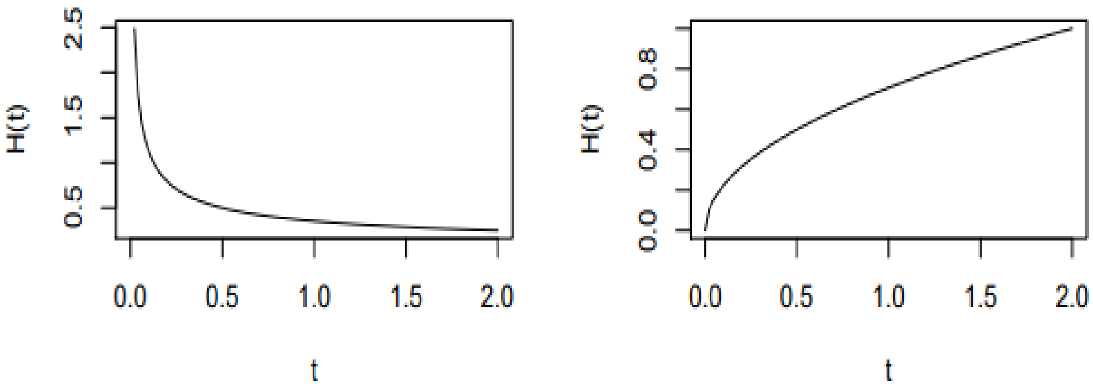
- S(t): This is the probability that a subject will survive beyond a specified time point, denoted as *t*. In the context of survival analysis, it is defined as $\mathrm{P\,r(T\ >\ t)}$
    - **As t ranges from 0 to ∞, the survival function has the following properties:**
        - It is non-increasing ∗ At time t = 0, S(t) = 1. In other words, the probability of surviving past time t=0 is 1.
        - At time t = ∞, S(t) = S(∞) = 0. As time goes to infinity, the survival curve goes to 0.
- T: Is a random variable representing the time until the event of interest occurs
- t: Specified time point,
- F(t): Represents the cumulative distribution function (CDF) of the event of interest, typically denoted by T, which represents the time until the event occurs (Colombia University, 2004).

1.2 Hazard Function:

The hazard function, h(t), describes the instantaneous event rate at a given time, representing the probability that an event will occur in the next instant, given no previous events (Colombia University, 2004).

$$\mathrm{h(t)}\ =\ \lim_{\Delta \mathrm{t}\to 0}\frac{\mathrm{P\,r(t\ <\ T\ \leq\ t\ +\ \Delta t|T\ >\ t)}}{\Delta \mathrm{t}}\ =\ \frac{f(t)}{S(t)}$$

The cumulative hazard describes the accumulated risk up to time t, $\mathrm{H(t)} = \int_0^t h(\mathrm{u})\mathrm{du}$.



Knowing S(t), H(t), or h(t), we can derive the other two functions:

$$h(t) = -\frac{\partial \log(S(t))}{\partial t}$$

$$H(t) = -log(S(t))$$

$$S(t) = e^{-H(t)}$$

1.3 Censoring:

In survival analysis, censoring occurs when the time to event is not observed for some subjects. This can happen when the study ends before an event occurs, the subject experiences event from a confounder variable or when a subject is lost to follow-up.

There are different types of censored data:

- **Right censoring**: a data point is above a certain value, but it is unknown by how much.
- **Left censoring:** a data point is below a certain value, but it is unknown by how much.

- **Fixed type I** censoring occurs when a study is designed to end after C years of follow-up. In this instance, censorship is applied to anyone who does not have an event observed during the study.
- **Random type I censoring**, the study is designed to end after C years, but censored subjects do not all have the same censoring time. This is the main type of right-censoring we will be concerned with.
- **Type II censoring**, a study ends when there is a pre-specified number of events.
- **Interval censoring:** when instead of knowing the exact time at which an event, we only know that the event occurred within a specified time range (Tripathy, 2020).

Regardless of the type of censoring, we must assume that it is non-informative about the event; that is, the censoring is caused by something other than the impending failure.

If we let δ_i denote the event indicator δ_i, then:

$$\delta\text{i} = \begin{cases} 1 \; if\ the\ event\ was\ observed \\ 0 \; if\ the\ event\ was\ censored \end{cases}$$

2. **Types of estimations**

2.1. Parametric Estimation

The **Kaplan-Meier estimator KM** is one of the main used estimators to estimate survival functions. The Kaplan-Meier model works with cumulative incidence and is bivariate with some confounding control, i.e.: one bivariate at a time.

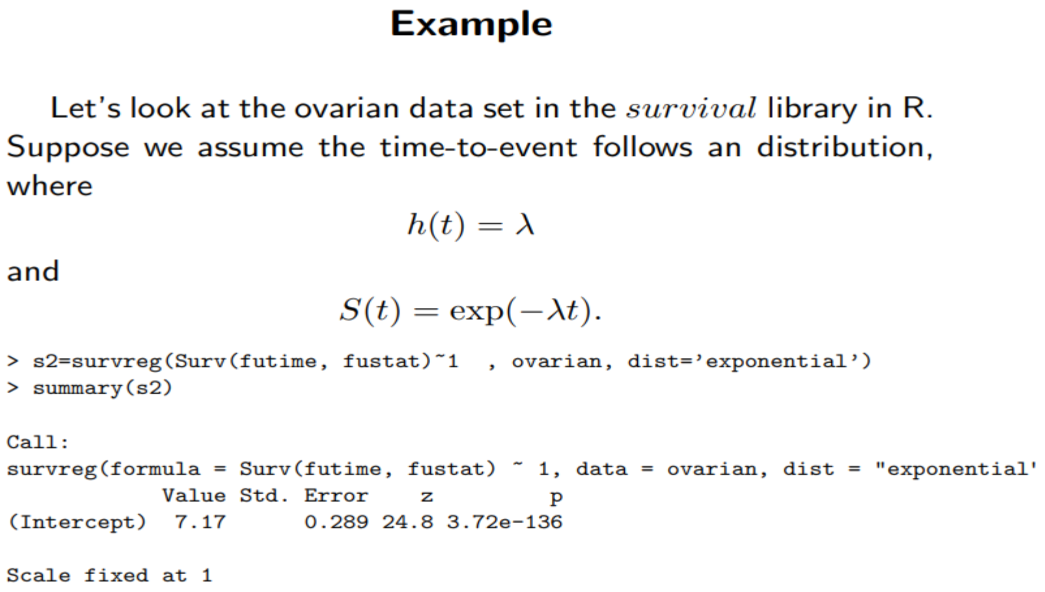
$$Cumulative\ Incidence = \frac{Number\ of\ New\ Cases\ of\ the\ Event}{Total\ Numbers\ at\ Risk} \times 100$$

The **Cox Proportional Hazard** is the major player used estimators to estimate survival functions as it works with incidence rates and can handle many confounders at a time.

$$Incidence\ Rate = \frac{Number\ of\ New\ Cases\ of\ the\ Event}{Total\ Person - TimeNumbers\ at\ Risk} \times Multiplier$$

- Where Multiplier is expressed per unit of time (e.g., per 1,000 person-years, per 100,000 person-months) for a more standardized rate.

- By specifying a parametric form for S(t), we can estimate it more precisely than KM. Some distributions for estimating curves are Exponential, Weibull, Log-normal and Log-logistic models (Colombia University, 2004).
- **For Exponential Distribution**: we can get the summary from the following example code and retrieve the value of lambda.



Example

Let's look at the ovarian data set in the *survival* library in R. Suppose we assume the time-to-event follows an distribution, where

$$h(t) = \lambda$$

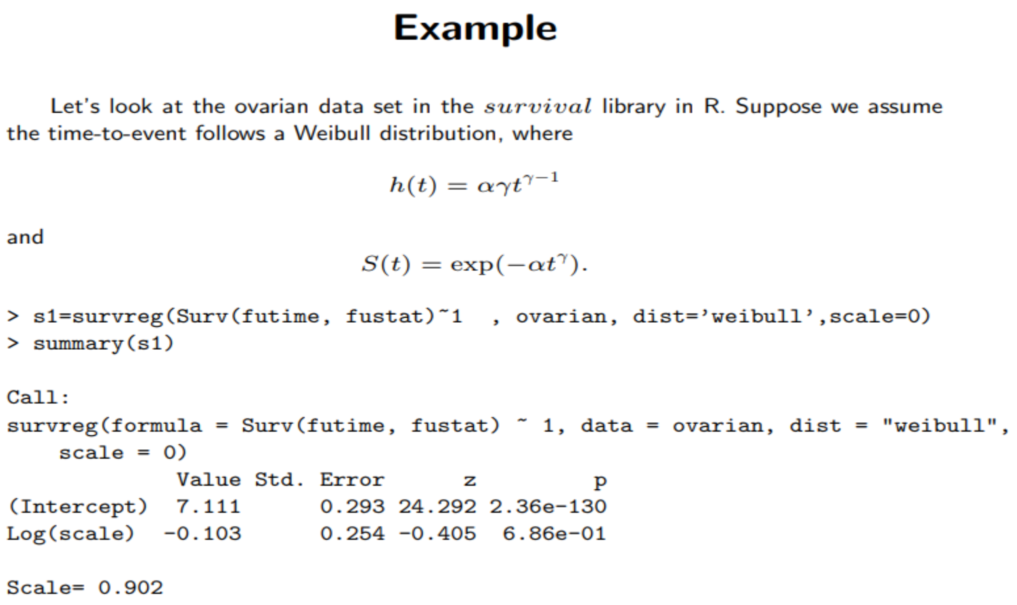
and

$$S(t) = \exp(-\lambda t).$$

```
> s2=survreg(Surv(futime, fustat)~1  , ovarian, dist='exponential')
> summary(s2)

Call:
survreg(formula = Surv(futime, fustat) ~ 1, data = ovarian, dist = "exponential'
            Value Std. Error    z         p
(Intercept)  7.17      0.289 24.8 3.72e-136

Scale fixed at 1
```

Figure 1 Example 1 exponential distr

  - $h(t) = \lambda$ and $S(t) = \ e^{-\lambda t}$
  - $\lambda = e^{-\ interceptt}$
- **For Weibull Distribution: w**e can get the summary from the following example code and retrieve the value of lambda.



Example

Let's look at the ovarian data set in the *survival* library in R. Suppose we assume the time-to-event follows a Weibull distribution, where

$$h(t) = \alpha\gamma t^{\gamma-1}$$

and

$$S(t) = \exp(-\alpha t^{\gamma}).$$

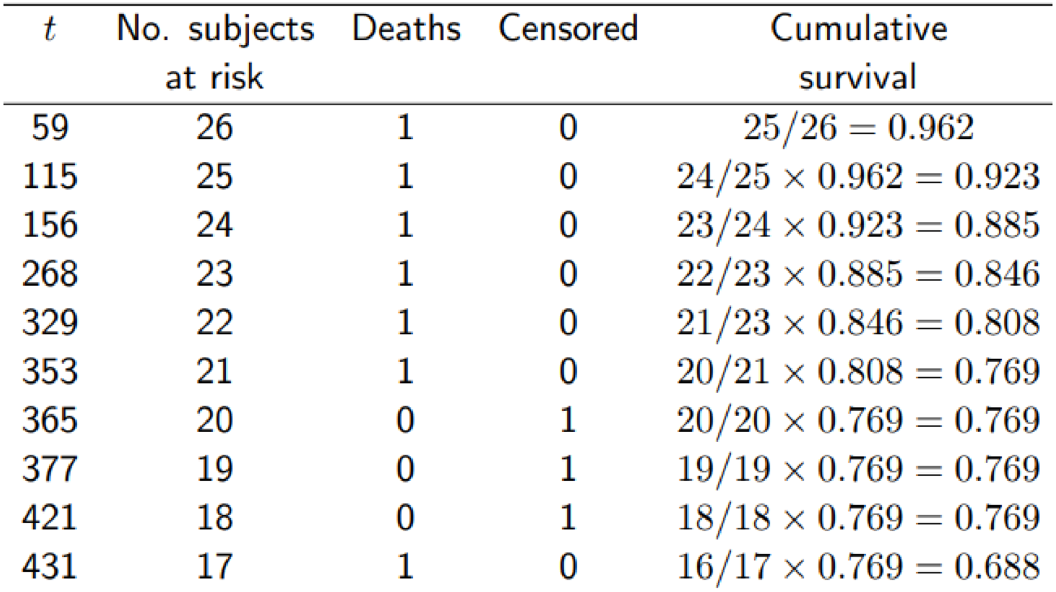
```
> s1=survreg(Surv(futime, fustat)~1  , ovarian, dist='weibull',scale=0)
> summary(s1)

Call:
survreg(formula = Surv(futime, fustat) ~ 1, data = ovarian, dist = "weibull",
    scale = 0)
              Value Std. Error      z         p
(Intercept)  7.111      0.293 24.292 2.36e-130
Log(scale)  -0.103      0.254 -0.405  6.86e-01

Scale= 0.902
```

Figure 2 Example 2 weibull distr

- ○ $h(t) = \alpha\gamma t^{\gamma-1}$ and $S(t) = \ e^{-\alpha t^{\gamma}}$
- ○ $\lambda = e^{-\ interceptt}$
- ○ $\gamma = \frac{1}{scale}$

2.2. Non-Parametric Estimation

- In the case when there are no censored observations, a parametric estimator of $S(t)$ is $1 - F(t)$, where F(t) is the empirical cumulative distribution function (Colombia University, 2004).
- On the other hand, if some observations are censored, we estimate S(t) using the Kaplan-Meier product-limit estimator as shown below for S(t) $= \prod_{i:t_{i\le t}} \frac{n_i-d_i}{n_i}$



t	No. subjects at risk	Deaths	Censored	Cumulative survival
59	26	1	0	$25/26 = 0.962$
115	25	1	0	$24/25 \times 0.962 = 0.923$
156	24	1	0	$23/24 \times 0.923 = 0.885$
268	23	1	0	$22/23 \times 0.885 = 0.846$
329	22	1	0	$21/23 \times 0.846 = 0.808$
353	21	1	0	$20/21 \times 0.808 = 0.769$
365	20	0	1	$20/20 \times 0.769 = 0.769$
377	19	0	1	$19/19 \times 0.769 = 0.769$
421	18	0	1	$18/18 \times 0.769 = 0.769$
431	17	1	0	$16/17 \times 0.769 = 0.688$

Figure 3 KM estimator example

3. Research Problem

This work aims to provide a comprehensive walkthrough of the implementation of Kaplan-Meier Survival Analysis for handling and analyzing survival data and to plot the curves along with their confidence interval.

4. Risk in analysis

Because of the way the cohort study is structured to get the data for analysis, researchers might be tempted to use standard regression procedures. However, the standard regressions could be not suitable for the following reasons:

- The distribution of time to an event is skewed and constrained to be positive.
- It's possible that the probability of living past a specific point in time will be more interesting than the anticipated timing of the event.
- Compared to linear regression, the hazard function, which is utilized for regression in survival analysis, can provide more information about the failure process (Colombia University, 2004).

5. METHODOLOGY

5.1. Data Source

The data was collected from the article “Survival Analysis; A Primer for Clinician Scientists”, which partly sourced this report.

5.2. Data Analysis

The programming language used was Rstudio, which is a powerful Integrated Development Environment (IDE), specifically among data scientists and statisticians. Rstudio hosts the programming language R (Posit, 2023).

Libraries are collections of similar functions for a certain task and the ones required for the task are the likes of:

- openxlsx
- Survival
- Autoplot
- Ggfortify
- stargazer

Data is presented as follows:

Table 1 Data with censored observations

Time, weeks	Number at risk	Number of deaths	Number censored
0	20	0	0

1	20	0	0
2	20	1	0
3	19	1	0
4	18	3	1
6	14	1	1
8	12	3	0
10	09	0	2
11	07	1	0
12	06	1	0
14	05	1	0
15	04	1	0
17	03	1	0
18	02	0	1
19	01	0	0

Since in RStudio, the number of deaths needed to be binary and a factor, data was further processed to expand where events were more than 1. This was fairly easy to do manually since to the data was small:

Table 2 Data with removed censored observations

time	Number at risk	Death
1	20	0
2	20	1
3	19	1
4	18	1
5	17	1
5	16	1
5	15	1
6	14	1
7	13	0
8	12	1
9	11	1
9	10	1
10	09	0
11	07	1
12	06	1
14	05	1
15	04	1

17	03	1
18	02	0
19	01	0

6. RESULTS AND DISCUSSION

From the data, the survival estimators and the confidence intervals type plain figure below. They are also plotted with the help of the library ggfortify:

Table 3 Surv Analysis results

time	n.risk	n.event	n.censor	surv	std.err	upper	lower
1	20	0	1	1.0000000	0.00000000	1.0000000	1.000000000
2	19	1	0	0.9473684	0.05407381	1.0000000	0.846963744
3	18	1	0	0.8947368	0.07868895	1.0000000	0.756743813
4	17	1	0	0.8421053	0.09933993	1.0000000	0.678145113
5	16	3	0	0.6842105	0.15585730	0.8932195	0.475201525
6	13	1	0	0.6315789	0.17521916	0.8484778	0.414680056
7	12	0	1	0.6315789	0.17521916	0.8484778	0.414680056
8	11	1	0	0.5741627	0.19948099	0.7986462	0.349679112

time	n.risk	n.event	n.censor	surv	std.err	upper	lower
9	10	2	0	0.4593301	0.25454403	0.6884886	0.230171652
10	8	0	1	0.4593301	0.25454403	0.6884886	0.230171652
11	7	1	0	0.3937116	0.29766120	0.6234049	0.164018176
12	6	1	0	0.3280930	0.34919267	0.5526414	0.103544479
14	5	1	0	0.2624744	0.41465108	0.4757876	0.049161137
15	4	1	0	0.1968558	0.50524138	0.3917932	0.001918377
17	3	1	0	0.1312372	0.64956564	0.2983186	0.000000000
18	2	0	1	0.1312372	0.64956564	0.2983186	0.000000000
19	1	0	1	0.1312372	0.64956564	0.2983186	0.000000000

Figure 4 Summary of analysis

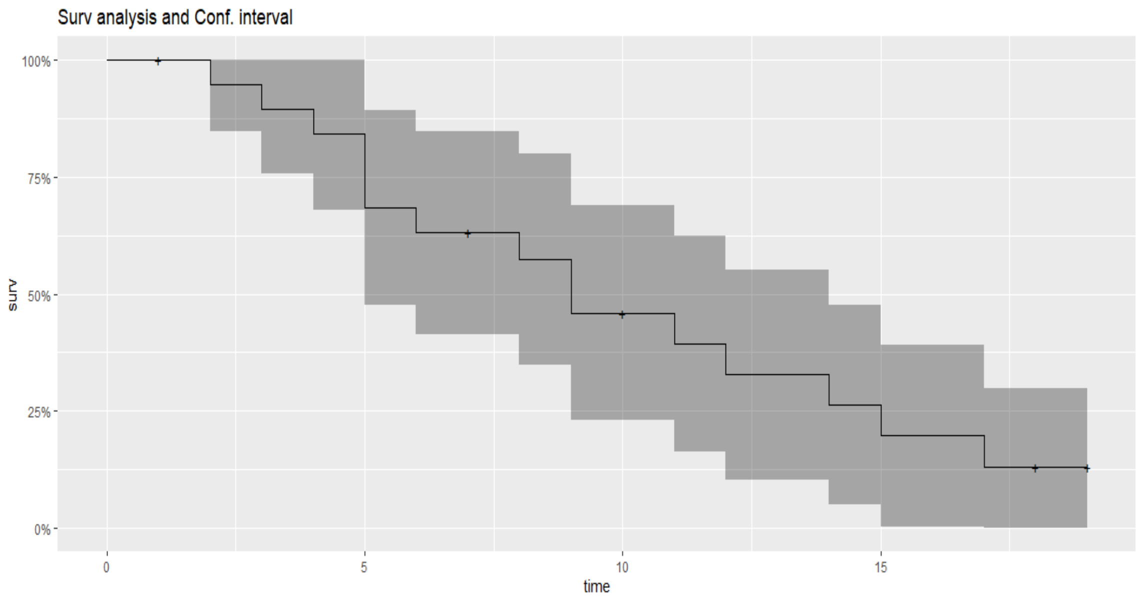


Figure 5 Plot of analysis

7. CONCLUSION AND RECOMMENDATIONS

To conduct a survival analysis, the data needs to be collected and some work has to be done on it. The work goes from expanding and coding the events related to the number of deaths into a binary outcome for a more streamlined analysis in RStudio. To enforce this report’s results perhaps a programming language where no encoding needs to be done should be used to do survival analysis so that better and faster algorithms are available for larger datasets.

8. APPENDIX A. Rstudio code

Click on here to see the code

9. REFERENCES

Columbia University. 2004. Lecture 15 introduction to survival analysis - department of statistics. https://www.stat.columbia.edu/~madigan/W2025/notes/survival.pdf (Accessed: 14 November 2023).

Frost, J. 2023. Cohort study: Definition, benefits & examples, Statistics By Jim. https://statisticsbyjim.com/basics/cohort-study/ [15 November 2023].

Posit. 2023. RStudio desktop. https://posit.co/download/rstudio-desktop/[14 November 2023].

Rai, S., Mishra, P. and Ghoshal, U.C. 2021. Survival analysis: A primer for the clinician scientists, Indian journal of gastroenterology : official journal of the Indian Society of Gastroenterology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8743691/ [14 November 2023].

Tripathy, A. 2020. Survival analysis: Censoring of Data, Medium. https://medium.com/@abhijittripathy99/survival-analysis-censoring-of-data-ea99928aa10b [14 November 2023].