

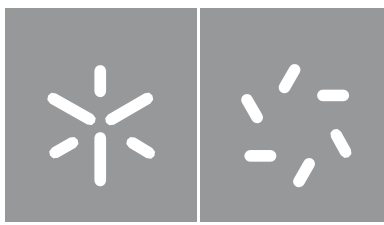
Universidade do Minho  
Escola de Ciências

Emanuel Vieira Monteiro da Silva

**Development of a Shiny application for  
survival data analysis**

Emanuel Vieira Monteiro da Silva  
**Development of a Shiny application for  
survival data analysis**





Universidade do Minho  
Escola de Ciências

Emanuel Vieira Monteiro da Silva

**Development of a Shiny application for  
survival data analysis**

Dissertação de Mestrado  
Mestrado em Estatística  
para Ciência de Dados

Trabalho efetuado sob a orientação da  
**Professor Doutor Luís Filipe Meira Machado**

## **COPYRIGHT AND CONDITIONS OF USE OF THE WORK BY THIRD PARTIES**

This is an academic work that can be used by third parties as long as the internationally accepted rules and good practices are respected, with regard to copyright and related rights.

If the user needs permission to be able to use the work under conditions not foreseen in the indicated license, he must contact the author, through the RepositóriUM of the University of Minho.

## **Acknowledgements**

I thank Professor Luís Filipe Meira Machado for his guidance during my dissertation. I also thank Professor Gustavo Soutinho for his suggestions for improving the dissertation.

## **STATEMENT OF INTEGRITY**

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

## **Abstract**

There is a great demand for graphical interfaces to perform statistical analysis by professionals who do not have good knowledge of programming. Therefore, the main objective of this project was to create an application that allows any user, regardless of their computational knowledge, to perform survival analyses. To this end, we have used the R software and the RStudio development environment and its packages, namely the shiny package, to develop an interactive web app that we called survapp. This application allows the use of different methodologies for the analysis of survival data. The survapp contains the core survival analysis models and techniques, including the Kaplan-Meier, log-rank test, Cox models, and parametric accelerated failure time models. The web application implements decision trees and random forests for survival data. An analysis of competitive risks is also possible, a particular case of multi-state models. A brief description of the mathematical background underlying survival analysis, focusing on the main methods and models and the development of the shiny application, is presented in this thesis. The Shiny app is available at the Shiny Apps repository: <https://emanuel-vieira.shinyapps.io/survapp/>. The application can be a very useful tool.

# Contents

- 1 Introduction** **1**
  
- 2 Basic concepts of Survival analysis** **2**
  - 2.1 Censorship . . . . . 2
  - 2.2 Mechanisms of censorship . . . . . 3
  - 2.3 Truncation . . . . . 3
  - 2.4 Probability density function . . . . . 3
  - 2.5 Survival function . . . . . 4
  - 2.6 Hazard function . . . . . 4
  
- 3 Nonparametric estimation** **5**
  - 3.1 The Kaplan-Meier estimator . . . . . 5
  - 3.2 Estimator for cumulative hazard function . . . . . 6
    - 3.2.1 The Nelson-Aalen estimator . . . . . 6
    - 3.2.2 The Fleming-Harrington estimator . . . . . 6
  - 3.3 The stratified Kaplan-Meier estimator . . . . . 7
    - 3.3.1 Tests to compare survival curves . . . . . 7
  - 3.4 Clusters of survival curves . . . . . 8
  
- 4 Cox regression model** **10**
  - 4.1 Formulation of the Cox Regression model . . . . . 10
  - 4.2 Estimation in the Cox Model . . . . . 10
  - 4.3 Selection of covariates in the Cox model . . . . . 11
  - 4.4 Selecting a suitable model . . . . . 11
  - 4.5 Validation of the proportional hazard assumption . . . . . 12
  
- 5 Univariate continuous distributions and some parametric survival models** **13**
  - 5.1 Exponential Distribution . . . . . 13



5.2	Weibull distribution . . . . .	13
5.3	Log-normal distribution . . . . .	14
5.4	Log-logistic distribution . . . . .	14
5.5	Gamma distribution . . . . .	15
5.6	The accelerated failure time model . . . . .	15
5.6.1	Formulation of the accelerated failure time model . . . . .	15
5.6.2	Construction of the likelihood function . . . . .	16
5.6.3	Akaike Information Criterion . . . . .	16
<b>6</b>	<b>Competing Risk Analysis</b>	<b>17</b>
6.1	Cause-specific functions and their estimators . . . . .	17
6.2	Competing risks regression . . . . .	18
<b>7</b>	<b>A brief introduction to Machine Learning Methods</b>	<b>19</b>
7.1	Classification and Regression Trees . . . . .	19
7.2	Decision tree . . . . .	20
7.2.1	Classification Trees . . . . .	22
7.2.2	Regression Trees . . . . .	23
7.3	Random Forests . . . . .	24
<b>8</b>	<b>Survapp</b>	<b>26</b>
8.1	Shiny . . . . .	26
8.1.1	UI Layout . . . . .	26
8.1.2	UI Inputs . . . . .	27
8.1.3	UI Outputs . . . . .	28
8.1.4	Interface builder functions . . . . .	29
8.1.5	Rendering functions . . . . .	29
8.1.6	Reactive programming . . . . .	29
8.2	Data.file . . . . .	30
8.3	Survival analysis . . . . .	31
8.3.1	Kaplan Meier . . . . .	32
8.3.2	Clusters of survival curves . . . . .	35
8.3.3	Cox PH Model . . . . .	36
8.3.4	AFT Model . . . . .	37
8.3.5	Regression Trees . . . . .	38
8.3.6	Classification Trees . . . . .	40

8.3.7	Random Forests . . . . .	42
8.3.8	Cumulative incidence . . . . .	43
8.3.9	Cumulative incidence between groups . . . . .	44
8.3.10	Competing risk regression . . . . .	46
8.3.11	Case Study Dataset IR_diabetes . . . . .	48

**9 Discussion and Future Work** **51**

# List of Figures

7.1	Decision Tree . . . . .	20
7.2	Decision tree with nodes . . . . .	21
7.3	Recursive partitioning of five regions in $\mathbb{R}^2$ , $R_1 - R_5$ , corresponding to the five terminal nodes. . . . .	21
8.1	Select Input for GBSG dataset. . . . .	27
8.2	File Input. . . . .	27
8.3	Numeric Input. . . . .	27
8.4	Action and radio buttons. . . . .	27
8.5	Check box group input for GBSG dataset. . . . .	28
8.6	Plot output from cumulative risk function. . . . .	28
8.7	Table and Text output from GBSG dataset. . . . .	29
8.8	Download button. . . . .	29
8.9	File input and change variable class. . . . .	31
8.10	Kaplan-Meier estimator for GBSG dataset. . . . .	33
8.11	Survival Curves by Kaplan-Meier estimator as a function of the hormon variable. . . . .	34
8.12	Survival Curves by Kaplan-Meier estimator as a function of the grade variable . . . . .	35
8.13	Clusters of survival curves . . . . .	36
8.14	Graphical Test of Proportional Hazards . . . . .	37
8.15	Regression tree for GBSG dataset . . . . .	39
8.16	Kaplan-Meier for different branches of the regression tree in Figure 8.15 . . . . .	40
8.17	Pruning of classification trees for GBSG dataset . . . . .	41
8.18	Classification tree for GBSG dataset . . . . .	41
8.19	20 survival curves of different decision trees for GBSG dataset . . . . .	42
8.20	Cumulative incidence . . . . .	44
8.21	Cumulative incidence between groups (sex variable) . . . . .	44
8.22	Cumulative incidence between groups (ulcer variable) . . . . .	46
8.23	Kaplan-Meier estimator for the IR_diabetes database . . . . .	49



# List of Tables

8.1	Survfit for gbsg . . . . .	33
8.2	Survfit for hormon variable . . . . .	34
8.3	Survfit for grade variable . . . . .	35
8.4	Cox Proporcional Hazard model for gbsg dataset . . . . .	37
8.5	Accelerated failure time model for gbsg dataset . . . . .	38
8.6	AFT interpretation . . . . .	38
8.7	Print of survfit with different branches . . . . .	40
8.8	Importance of variables for the GBSG database random forest model . . . . .	42
8.9	Estimates e variances for cumulative incidence . . . . .	43
8.10	Tests, estimates e variances for cumulative incidence between groups (sex variable) . . . . .	45
8.11	Tests, estimates e variances for cumulative incidence between groups (ulcer variable) . . . . .	47
8.12	Subdistribution hazard for dead from melanoma . . . . .	47
8.13	Subdistribution hazard for dead from other cause . . . . .	48
8.14	Cause-specific hazards for dead from melanoma . . . . .	48
8.15	Cause-specific hazards for dead from other cause . . . . .	48
8.16	Summary of the Kaplan-Meier estimator for the IR_diabetes database . . . . .	49
8.17	Summary of the Kaplan-Meier estimator for survival curves as a function of sex for the IR_diabetes database . . . . .	50
8.18	Parametric model for IR_diabetes . . . . .	50

# Chapter 1

## Introduction

The R software is one of the most commonly used for statistical analysis, but its graphical interface is still not user-friendly enough for those who do not have statistical and programming skills. The R software has a package library with loads of functions implemented for all phases of a data analysis project. The most relevant R package for this course and for the dissertation is the shiny one that allows R users to develop applications.

The main objective of the project was to create Survapp, an application that allows its users to carry out survival data analysis.

In the development of Survapp, some shiny applications already developed were tested. Some of them are MSM.app [24]. SmulTCan [20] and MEPHAS [30].

Survapp brings together a set of characteristics that are partly similar to existing applications, however, it stands out with the use of supervised algorithms applied to survival data. I consider it to be a more user-friendly technology that is accompanied by a statistical description that makes it stand out from other existing applications.

## Chapter 2

# Basic concepts of Survival analysis

Survival analysis is an area of statistics where the objective is to analyze and model the data where the result is the time until the occurrence of an event of interest. The Survival time refers to a variable which measures the time from a particular starting time (e.g., time initiated the treatment) to a particular endpoint of interest (time-to-event).

The main goals of survival analysis are the following: estimating time-to-event for a group of individuals, comparing time-to-event between two or more groups, or assessing the relationship of covariates to time-to-event.

These methods can also be applied to data from different areas: social sciences (time for doing some task); economics (time looking for employment) and engineering (time to a failure of some electronic component).

### 2.1 Censorship

The distinguishing feature of survival analysis is that it incorporates censoring. Censoring occurs when we have some information about individual survival time, but we don't know the time exactly.

The most frequent type of censorship is right censoring. In this type of censoring, the event of interest is not observed until the end of the study. Usually occurs when an individual leaves the study before the event occurs, or the study ends before the event has occurred. As an example, consider a clinical study in which the event of interest is the death of an individual after having been diagnosed with a certain malignant tumor, if the individual is alive at the end of the study, this is an observation censored right.

The lifetime of some subject is considered to be left censored if it is less than a censoring time. That is, the event of interest has already occurred for the individual before the observed time (not easy to deal with). The observed survival time is greater than or equal to true survival time.

Examples of left censoring: infection with a sexually-transmitted disease such as HIV/AIDS, onset of a pre-symptomatic illness such as cancer, time at which teenagers begin to drink alcohol, The age at which children are able to count from 1-10 at school. Some children are already able to count before joining School.

When the lifetime is only known to occur within an interval. Such interval censoring occurs when patients in a

clinical trial or longitudinal study have periodic follow-up and the patient's event time is only known to fall in some interval.

## 2.2 Mechanisms of censorship

- **Type I Censoring**

The event is observed only if it occurs prior to some specified time. Let  $t_c$  be some (preassigned) fixed number which we call the fixed censoring time. Instead of observing  $T_1, \dots, T_n$  (random variables of interest) we can only observe  $Y_1, \dots, Y_n$  where

$$Y_i = \begin{cases} T_i & T_i \leq t_c \\ t_c & t_c < T_i \end{cases}$$

- **Generalized Type I Censoring**

Also known as random Type I Censoring. When individuals enter the study at different times and the terminal point of the study is predetermined by the investigator, so that the censoring times are known when an individual is entered into the study.

- **Type II Censoring**

The study continues until the failure of the first  $r$  individuals, where  $r$  is some predetermined integer ( $r < n$ ). All subjects are put on test at the same time, and the test is terminated when  $r$  of the  $n$  subjects have "failed".

## 2.3 Truncation

When planning a study, it is of interest to design it to verify the event of interest. Thus, only individuals to whom the event of interest has occurred or will occur are included in the study. This mechanism, which consists of excluding individuals who are not relevant to the study in question, is called truncation.

## 2.4 Probability density function

Let  $T$  be a non-negative random variable representing the lifetime of a individual of a given homogeneous population. For the sake of simplicity in the following sections we will assume that  $T$  is continuous.

The probability of the failure time occurring at exactly time  $t$ ,  $f(t)$ .

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

The probability of the failure time (CDF) occur before or exactly at time  $t$ ,  $F(t)$ .



$$F(t) = P(T \leq t) = \int_0^t f(u) du$$

## 2.5 Survival function

The main goal of survival analysis is to estimate and compare survival experiences of different groups. Survival experience is described by the cumulative survival function.

$$S(t) = P(T > t) = 1 - F(t)$$

## 2.6 Hazard function

The hazard function is the probability that if you survive to  $t$ , you will experience the event in the next instant.

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Hazard from density and survival:

$$h(t) = \frac{f(t)}{S(t)}$$

In some cases it can be more interesting to present the cumulative hazard.

For continuous time

$$H(t) = \int_0^t h(u) du$$

For discrete time

$$H(t) = \sum_{t_i \leq t} h(t_i)$$

## Chapter 3

# Nonparametric estimation

The advantage of non-parametric estimation is that it is very flexible, as no assumptions are made about the distribution. The main disadvantage is that it is not easy to incorporate covariates.

### 3.1 The Kaplan-Meier estimator

In the presence of complete data, without censored observations in a sample of dimension  $n$ , the most natural estimator for survival is the empirical estimator, given by the survival function at time  $t$  is estimated by the proportion of individuals with failure times greater than  $t$ .

$$\hat{S}(t) = \frac{\text{Number of observations} > t}{n}, t \geq 0$$

Kaplan and Meier (1958) [14], obtained a nonparametric estimate of the survival function, when we are in the presence of a censored sample. This estimator is called the Kaplan-Meier (K-M) estimator or product-limit estimator, which is the generalization of the empirical estimator for censored data.

$$\hat{S}(t) = \prod_{j:t_i \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

Where  $t_1 < t_2 < \dots < t_k$  are the observed event times,  $d_j$  the number of events at time  $t_j$ ,  $n_j$  the number of individuals in risk at time  $t_j$ ,  $d_j/n_j$  the proportion that failed at the event time  $t_j$  and  $1 - d_j/n_j$  the proportion surviving the event time  $t_j$ .

$\hat{S}(t)$  represents estimated survival probability at time  $t$ :  $P(T > t)$ . Graphical representation of the estimator highlights all important aspects of the sampling distribution of survival time. The Kaplan-Meier estimate does not control for covariates or time-dependent variables.

The Greenwood variance estimate for a Kaplan-Meier curve is defined as:

$$Var(\hat{S}(t)) \approx \left[ \hat{S}(t) \right]^2 \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}, \text{ for } t_{(k)} \leq t < t_{(k+1)}$$

The confidence interval for the value of the survival function at a given time  $t$  can be obtained assuming that the estimator of the function at the instant  $t$  has a normal distribution with mean value  $S(t)$  and estimated variance  $Var(\hat{S}(t))$ . A confidence interval of  $100(1 - \alpha)\%$  is given per:

$$\left[ \hat{S}(t) - z_{\frac{\alpha}{2}} \sqrt{Var(\hat{S}(t))}, \hat{S}(t) + z_{\frac{\alpha}{2}} \sqrt{Var(\hat{S}(t))} \right]$$

where  $z_{\frac{\alpha}{2}}$  represents the probability quantile  $1 - \frac{\alpha}{2}$  of the centered and reduced normal distribution, that is, of the distribution  $N(0, 1)$ . In the case of the Kaplan-Meier estimator, where the standard deviation is  $SD\{\hat{S}(t)\} \approx \hat{S}(t) \sqrt{\sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}}$  a confidence interval of  $100(1 - \alpha)\%$  is given by:

$$\left[ \hat{S}_{KM}(t) - z_{\frac{\alpha}{2}} \hat{S}_{KM}(t) \sqrt{\sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}}, \hat{S}_{KM}(t) + z_{\frac{\alpha}{2}} \hat{S}_{KM}(t) \sqrt{\sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}} \right]$$

## 3.2 Estimator for cumulative hazard function

The estimator proposed by Kaplan and Meier (1958) is the most used estimator for the non-parametric estimation of the survival function. Often, it is also important to estimate the cumulative hazard function. A natural estimator of  $H(t) = -\log \hat{S}(t)$ , where  $\hat{S}(t)$  is the Kaplan and Meier estimator.

### 3.2.1 The Nelson-Aalen estimator

An alternative estimator, suggested by Nelson [19] and studied by Aalen [1], is another option that is becoming increasingly common.

Let  $t_{(1)} < \dots < t_{(r)}$  be the distinct moments of death in a sample of dimension  $n$  ( $r \leq n$ ),  $d_{(i)}$  the number of deaths occurred in  $t_{(i)}$  and  $n_{(i)}$  the number of individuals at risk in  $t_{(i)}$ . The Nelson-Aalen estimator is defined by:

$$\hat{H}_{NA}(t) = \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i}$$

### 3.2.2 The Fleming-Harrington estimator

The Fleming-Harrington estimator [9] is obtained through the Nelson-Aalen estimator.

$$\hat{S}_{FH}(t) = \exp\left(-\sum_{i:t_i \leq t} \left[1 - \frac{d_i}{n_i}\right]\right)$$

### 3.3 The stratified Kaplan-Meier estimator

The strategy used, based on the Kaplan-Meier estimator, to compare the different curves corresponding to the various groups, is stratification. This stratification consists of dividing the total set of observations into distinct groups, according to the covariates of interest, and estimating the survival functions separately for each of the groups.

#### 3.3.1 Tests to compare survival curves

The graphical representation of the Kaplan-Meier curves of survival functions, already allows to perceive that there are differences in the survival curves, in the respective groups. To assess whether there is a significant difference between the several curves, hypothesis tests must be used.

The hypotheses to be tested are:

$$H_0 : S_1(t) = S_2(t) \text{ vs } H_1 : S_1(t) \neq S_2(t)$$

#### a) The Log-Rank or Mantel-Haenszel and Gehan-Wilcoxon estimator

When the purpose of the analysis is to compare survival curves, it is necessary to determine whether there are statistically significant differences between the curves of two or more groups of individuals. It is possible to test the null hypothesis of no difference between survival curves of the two (or more) groups.

The log-rank test [17] is the most used nonparametric test and appropriate to use when the data are right skewed and censored. This test is the one with most power to test differences that fit the proportional hazards model so works well as a set-up for subsequent Cox regression. Gives equal weight to early and late failures. The Gehan-Wilcoxon estimator [10] weights strata by their size and is more sensitive to differences at earlier time points.

The survdiff function of the R survival package tests whether there are statistically significant differences between the curves of each group of individuals. A formula expression as for other survival models, of the form `Surv(time, status) ~ predictors`, can be used to implement two tests to compare survival functions. A scalar parameter named `rho` controls the type of test with default to the log-rank or Mantel-Haenszel test. In R, it is also possible to use the Peto & Peto modification of the Gehan-Wilcoxon test by using the argument `rho = 1`.

The function returns the number of subjects in each group, the weighted observed number of events in each group, the weighted expected number of events in each group, the Chi-square statistic for a test of equality and the variance matrix of the test.

### 3.4 Clusters of survival curves

When there are variables with a high number of levels, therefore a high number of survival curves, it may be relevant to understand if the curves can be grouped into clusters. Some authors have proposed different methods that can be used to compare estimates of nonparametric multi-sample functions. The null hypothesis is that all curves have identical functions,  $H_0 : F_1 = \dots = F_J$  with  $j = \{1, \dots, J\}$  as population [27].

If the null hypothesis is rejected, there is no available procedures that make it possible to determine groups between the curves, that is, to assess whether the levels  $1, \dots, J$  can be grouped into  $K$  groups  $(G_1, \dots, G_K)$  with  $K < J$ , so that  $F_i = F_j$  for all  $i, j \in G_k$ , for each  $k = 1, \dots, K$ . Note that  $(G_1, \dots, G_K)$  must be a partition of  $\{1, \dots, J\}$  and therefore must satisfy the following conditions:

$$G_1 \cup \dots \cup G_K = \{1, \dots, J\}$$

$$G_i \cap G_j = \emptyset, \forall i \neq j \in \{1, \dots, K\}$$

A procedure has been proposed to test, for a given number  $K$ , the null hypothesis  $H_0(K)$  that there is at least one partition  $(G_1, \dots, G_K)$  so that all the above conditions are met. The alternative hypothesis  $H_1(K)$  is that for any  $(G_1, \dots, G_K)$ , there is at least one group  $G_k$  in which  $F_i \neq F_j$  for some  $i, j \in G_k$ .

The procedure is based on the J-dimensional process:

$$\hat{U}(z) = (\hat{U}_1(z), \hat{U}_1(z), \dots, \hat{U}_J(z))$$

where, for  $j = 1, \dots, J$ ,

$$\hat{U}_j(z) = \sum_{k=1}^K [\hat{F}(z) - \hat{C}_k(z)] I_{j \in G_k}$$

and  $\hat{C}_k$  is the pooled nonparametric estimate based on the combined  $G_k$ -partition sample.

The following test stats were considered for testing  $H_0(K)$ : a Cramer-von Mises type statistic

$$D_{CM} = \min_{G_1, \dots, G_k} \sum_{j=1}^J \int_R \hat{U}_j^2(z) dz$$

and a modification of it based on the L1 norm proposed in the Kolmogorov-Smirnov test statistics

$$D_{KS} = \min_{G_1, \dots, G_k} \sum_{j=1}^J \int_R |\hat{U}_j(z)| dz$$

where R is the support of the lifetime distribution or the support of the independent variable in the case survival

or regression, respectively.

K-nonparametric curves algorithm:

1. With the original sample, for  $j = 1, \dots, J$  and  $i = 1, \dots, n_j$ , and using the nonparametric estimator obtain  $\hat{F}_j$ .
2. Initialize with  $K = 1$  and test  $H_0(K)$ :
  - (a) Obtain the “best” partition  $G_1, \dots, G_K$  by means of the k-means or k-medians algorithm.
  - (b) For  $k = 1, \dots, K$ , estimate  $\hat{C}_k$  and retrieve the test statistic  $D$ .
  - (c) Generate  $B$  bootstrap samples and calculate  $D^{*b}$ , for  $b = 1, \dots, B$ .
  - (d) **if**  $D > D^{*(1-\alpha)}$  **then**
    - reject  $H_0(K)$
    - $K = K + 1$
    - go back to (a)**else**
    - accept  $H_0(K)$**end**
3. The number  $K$  of groups of equal nonparametric curves is determined.

## Chapter 4

# Cox regression model

One of the aspects of the Cox proportional hazards model [6] is that it is formulated based on the relationship between the risk function and the covariates. This model allows estimating the relationship between the relative risk rate and the predictor variables. Although the effect of the covariates is modeled parametrically, this model is semi-parametric, given that the underlying risk function does not require the choice of a probabilistic model to represent the survival times of individuals, making the model more robust than parametric methods.

### 4.1 Formulation of the Cox Regression model

The hazard function is given by:

$$h_i(t; Z) = h_0(t) \exp(\beta Z_{i1} + \dots + \beta_k Z_{ik})$$

where  $Z = (Z_1, \dots, Z_k)$  is a vector of covariates,  $\beta = (\beta_1, \dots, \beta_k)$  a vector of regression parameters and  $h_0(t)$  the baseline hazard function.

### 4.2 Estimation in the Cox Model

Let  $t_{(1)} < \dots < t_{(r)}$  be the lifetimes and let  $R_j$  be the risk set at the instant  $t_{(j)}$ . The partial likelihood function proposed by David Cox (1972) [6] is defined by:

$$L(\beta) = \prod_{j=1}^r \frac{\exp(\beta Z_j)}{\sum_{l \in R_j} \exp(\beta Z_l)}$$

where  $Z_j$  is the vector of covariates.

### 4.3 Selection of covariates in the Cox model

We are interested in identifying which covariates have significant influences in the survival of individuals, among all that were registered.

Let  $T$  be a continuous random variable representing the lifetime and  $Z = (Z_1, \dots, Z_k)$  the vector of covariates associated with each individual.

The risk of "death" of an individual, at the instant  $t$  is given by  $h(t; Z) = h_0(t) \times g(\beta Z)$ , where  $\beta = \beta_1, \dots, \beta_k$  is the vector of the unknown regression parameters,  $h_0$  is the function of underlying hazard and  $g$  an arbitrary function of  $Z$  and  $\beta$ .

We intend to test the following hypothesis:

$$H_0 : \beta = 0 \text{ vs } H_1 : \beta \neq 0, j = 1, \dots, k$$

in the null hypothesis, we are testing whether the covariate  $Z_j$  does not influence survival in the presence of other covariates, against the alternative hypothesis, in which this covariate has a statistically significant influence on survival. If we reject the null hypothesis, it means that the variable in question influences survival. The inference methods are fundamentally based on the maximum likelihood method.

#### Likelihood ratio test

To assess whether or not the covariate in question influences survival, we will use the likelihood ratio test, which is based on comparing logarithm values of the maximized likelihood function under the null hypothesis validity.

The test statistic is given by:

$$LRT = -2 \log \left[ \frac{L(\text{reduced model})}{L(\text{full model})} \right] = -2 \log(Lr) + 2 \log(Ls)$$

Where  $Lr$  is the likelihood of the reduced model and  $Ls$  is the likelihood of the full model.

### 4.4 Selecting a suitable model

The objective is to find the most parsimonious model, that is, the model that involves the least possible parameters to be estimated and that explains the behavior of a response variable. We will place more emphasis on the Akaike information criterion (AIC) as a criterion for model selection.

The estimate of AIC for a given model is given by:  $AIC = -2L + 2r$  in that,  $L$  is the logarithm of the likelihood function with the parameters  $\theta$  and  $r$  is the number of model parameters.



## 4.5 Validation of the proportional hazard assumption

The Cox model is only valid if the proportional hazards assumption is verified, which can be verified through a graph  $\log(-\log(S(t)))$  against  $t$ . The proportional hazards assumption is verified if the curves obtained are parallel, if this is not the case, the proportional hazards assumption is not valid.

The proportional hazards assumption test tests the following hypotheses:

**H<sub>0</sub>**: proportional hazards assumption is satisfied

vs

**H<sub>0</sub>**: proportional hazards assumption is not satisfied

The null hypothesis is rejected if the global p-value is less than established significance level. Another way to validate the risk assumption proportional is through the graph of Schoenfeld residuals, proposed by Schoenfeld (1982) [23].

According to Schoenfeld, there is not just one residual for each individual, but several, as many as the covariates included in the model. Furthermore, it is not necessary to obtain an estimate of the cumulative risk function. For the  $i$ -th individual under study, the residual corresponding to the covariate  $Z_j, j = 1, \dots, p$  is given by:

$$r_{ji} = \delta_i (Z_{ji} - \alpha_{ji})$$

where  $\alpha = \frac{\sum_{l \in R_i} Z_{jl} \exp(Z_l \hat{\beta})}{\sum_{l \in R_i} \exp(Z_l \hat{\beta})}$ , where  $R_i$  is the set of individuals at risk at the instant  $t_i$ . These residuals can be interpreted as the difference between the observed values of the covariates of a given individual, whose death was observed in  $t_i$ , and is a weighted average of the values of these covariates, for all individuals at risk in  $t_i$ . The weights corresponding to these individuals are  $\exp(Z_l \hat{\beta})$ .

According to Schoenfeld, if the proportional hazards assumption is satisfied, there can be no trend in the graph, that is, the graphical representation should present the appearance of a cloud of points around zero.

## Chapter 5

# Univariate continuous distributions and some parametric survival models

If it is possible to admit a parametric model for the lifetime, there is the advantage of having of direct application inference methods [22]. However, the circumstances necessary for the application of a parametric model relevant to the data under study are not frequent. Next, some univariate continuous distributions and some of the parametric survival models are presented.

### 5.1 Exponential Distribution

Let  $T$  be a random variable with exponential distribution of parameter  $\lambda > 0$ , with given density function.

$$f(t) = \lambda \exp(-\lambda t), t \geq 0$$

The risk and survival functions are respectively  $h(t) = \lambda$  and  $S(t) = \exp(-\lambda t)$ . This distribution presents a constant risk function.

The exponential distribution is a reference in the analysis of survival data due to its mathematical simplicity and properties. The risk function being constant over time restricts the use of this distribution in many applications.

### 5.2 Weibull distribution

Let  $T$  be a random variable with a Weibull distribution with a scale parameter  $\lambda > 0$  of form  $\alpha > 0$ , with a density function given by:

$$f(t) = \lambda \alpha (\lambda t)^{\alpha-1} \exp^{-(\lambda t)^\alpha}, t \geq 0, \lambda, \alpha > 0$$

The risk and survival functions are respectively  $h(t) = \lambda \alpha (\lambda t)^{\alpha-1}$  and  $S(t) = \exp^{-(\lambda t)^\alpha}$ . The Weibull

distribution is a generalization of the exponential distribution having a much higher use than this one, since its hazard function can be either constant ( $\alpha = 1$ ), monotonous increasing ( $\alpha > 1$ ) or monotonous decreasing ( $0 < \alpha < 1$ ).

### 5.3 Log-normal distribution

Let be  $T$  a random variable with a log-normal distribution with parameters  $\mu \in (-\infty, +\infty)$  and  $\sigma > 0$  and a support  $t \in (0, +\infty)$ . The density function of probability is:

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp \left[ -\frac{1}{2} \left( \frac{\ln t - \mu}{\sigma} \right)^2 \right]$$

The survival function is:

$$S(t) = 1 - \Phi \left( \frac{\ln t - \mu}{\sigma} \right)$$

Where  $\Phi$  is the gaussian distribution function with mean 0 and standard deviation 1.

The hazard function is:

$$h(t) = \frac{\frac{1}{\sqrt{2\pi}\sigma t} \exp \left[ -\frac{1}{2} \left( \frac{\ln t - \mu}{\sigma} \right)^2 \right]}{1 - \Phi \left( \frac{\ln t - \mu}{\sigma} \right)}$$

The hazard function is increasing until it reaches a maximum value from which it becomes descending. This distribution is suitable when high values are of no interest.

### 5.4 Log-logistic distribution

The density function of probability of a random variable  $T$  with shape parameter  $\alpha > 0$  and scale  $\lambda > 0$  is:

$$f(t) = \lambda \alpha (\lambda t)^{\alpha-1} (1 + (\lambda t)^\alpha)^{-2}$$

The survival function is:

$$S(t) = \frac{1}{1 + (\lambda t)^\alpha}$$

The hazard function is:

$$h(t) = \frac{\lambda \alpha (\lambda t)^{\alpha-1}}{1 + (\lambda t)^\alpha}$$

It is used for events whose rate increases initially and decreases later.

## 5.5 Gamma distribution

The Gamma distribution is also a generalization of the exponential. Let  $T$  be a variable with Gamma distribution with scale parameters  $\theta$  and form  $k$ . The probability density, survival and hazard functions are given by:

$$f(t) = \frac{1}{\Gamma(k)} \theta^k t^{k-1} \exp^{-\theta t}$$

$$S(t) = \frac{1}{k\theta t^{k-1}}$$

$$h(t) = k\theta t^{k-1}$$

where  $t \geq 0, \theta > 0$  e  $k > 0$ .  $\Gamma(k)$  is the gamma function. The hazard function is constant when ( $k = 1$ ), monotonic increasing when ( $k > 1$ ) and monotonous decreasing ( $0 < k < 1$ ).

## 5.6 The accelerated failure time model

Accelerated failure time model (AFT model) is a parametric model that provides an alternative to the commonly used proportional hazards models. Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant [15].

Parametric multivariate regression techniques model the underlying risk/survival function. The time to event (dependent variable) has some known distribution, such as Weibull, exponential, lognormal, etc. Using parametric models to describe survival time becomes an advantageous methodology.

### 5.6.1 Formulation of the accelerated failure time model

The hazard function is:

$$h_i(t; Z) = h_0(t) * k(\beta Z)$$

where  $Z = (Z_1, \dots, Z_k)$  is a vector of covariates,  $\beta = (\beta_1, \dots, \beta_k)$  a vector of regression parameters,  $h_0(t)$  a baseline hazard function and  $k$  a specified link function.

## 5.6.2 Construction of the likelihood function

Let us consider that  $n$  individuals are being studied and that the data corresponding to the  $i - th$  individual are of the form  $(t_i, \delta_i, z_i)$ ,  $i = 1, \dots, n$ , where  $t_i$  is the lifetime ( $\delta_i = 1$ ) or censoring time ( $\delta_i = 0$ ) and  $z_i$  is a vector of fixed covariates. Let's assume that the distribution of the lifetime  $T$  given  $z$  is known to less than a vector of parameters  $\theta$ , on which we wish to perform inference and that the survival function for the  $i - th$  individual is  $S(t_i; z_i, \theta)$ , with the corresponding probability density function  $f(t_i; z_i, \theta)$  [22].

As individuals are subject to an independent censorship mechanism, tooth and non-informative, the likelihood function is given by:

$$L(\theta) = \prod_{i=1}^n f(t_i; z_i, \theta)^{\delta_i} S(t_i; z_i, \theta)^{1-\delta_i}$$

## 5.6.3 Akaike Information Criterion

As in the Cox Proportional Hazards model, in the parametric model the Akaike information criterion (AIC) can also be used as a criterion for selecting the best model.

$$AIC = -2 \log \hat{L} + 2(p + 1 + k),$$

where  $\hat{L}$  represents the maximized likelihood and  $p$  the number of regression parameters of the fitted model,  $k = 0$  for the exponential model and  $k = 1$  for the Weibull, log-logistic and log-normal models. The smaller the value of the AIC statistic, the better the model.

## Chapter 6

# Competing Risk Analysis

When the subjects under study have several possible events in a time-to-event setting, it is possible to perform a competitive risk analysis, a particular case of multi-state models. Multi-state models are very useful for describing complex event history data with multiple endpoints. These models may be considered a generalization of survival analysis where survival is the ultimate outcome of interest but where information is available about intermediate events which individuals may experience during the study period. For instance, in most biomedical applications, besides the 'healthy' initial state and the absorbing 'dead' state, one may observe intermediate (transient) states based on health conditions (e.g., diseased), disease stages (e.g., stages of cancer or HIV infection) [25]. Examples of events are recurrence, death from disease, death from other causes or treatment response.

The unobserved dependence between event times needs special consideration. For example, patients who recur are more likely to die, and therefore times to recurrence and times to death would not be independent events. For analysis in the presence of multiple potential outcomes there are two approaches, one in terms of the associated potential or latent lifetimes for each cause of death (a concept that was introduced only for mathematical convenience), the other commonly used is to describe the problem in terms of the cause-specific risk functions Rocha, Cristina, and Ana Luísa Papoila. (2009) [22].

### 6.1 Cause-specific functions and their estimators

Let's suppose that a population is subject to  $m$  causes of death. When a death occurs, we observe the lifetime  $T$  and cause of death  $J, J \in \{1, 2, \dots, m\}$ . The approach proposed by Prentice et al. (1978) [21] consists of describing the problem in terms of the cause-specific risk functions. The cause-specific risk function  $j(j = 1, \dots, m)$  is defined by:

$$h_j(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt, J = j | T \geq t)}{dt}$$

and describes the instantaneous probability of death due to cause  $j$  in the instant  $t$ , in the presence of the other

causes of death.

The lifetime survival function  $T$  can be represented per:

$$S(t) = \exp\left(-\sum_{j=1}^m \int_0^t h_j(u) du\right)$$

The cause-specific survival function and The cumulative incidence function of the cause  $j$  is also defined as being:

$$P(T > t, J = j) = \int_t^{\infty} h_j(u) S(u) du$$

$$I_j(t) = P(T \leq t, J = j) = \int_0^t h_j(u) S(u) du$$

## 6.2 Competing risks regression

Now suppose that each individual is associated with a vector of  $z$  covariates. The Cox model can be generalized to allow its application in problems of competitive risks, being the cause-specific risk function  $j$  given by:

$$h_j(t; z) = h_{0j}(t) \exp(\beta_j' z)$$

Subdistribution risks is the instantaneous rate of occurrence of a given type of event in individuals who have not yet experienced an event of that type and is estimated using Fine-Gray regression [8].

## Chapter 7

# A brief introduction to Machine Learning

## Methods

Machine learning is an interdisciplinary activity that mainly combines two major areas: computer science and statistics. Machine learning methods are increasingly being applied to decision-making, especially in medicine, biology, and economics.

### 7.1 Classification and Regression Trees

Decision tree methods are predictive machine learning techniques that can be used for classification and regression known as CART (Classification and Regression Trees). CART methods involve stratifying or segmenting the predictor space into several simple regions. The mean or mode of training observations is normally used to predict a given observation. These methods are known as decision tree methods because the set of division rules used to segment the predictor space can be summarized in a tree. Tree-based methods are simple and useful for interpretation. However, in terms of prediction accuracy, they are typically not competitive with the best supervised learning approaches. The decision tree model algorithm works by repeatedly splitting the data into multiple subspaces so that the results in each final subspace are as homogeneous as possible. This approach is called recursive partitioning. The result produced consists of a set of rules used to predict the outcome variable, regression tree for a continuous variable and classification trees for categorical variables.

Initially, tree-based methods were developed to model a categorical or continuous result using a set of covariates from a sample of data. They were introduced by Morgan and Sonquist (1963) [18], but became popular in the 1980's due in large part to the development of the CART paradigm by Breiman et al. (1984) [4]. The tree recursively splits the covariate space to form the nodes. For a categorical answer, the Gini and entropy measures of the impurity are popular, while the sum of squared deviations from the mean is more commonly used for a continuous answer. The basic approach focuses on binary splits using a single covariate. For a continuous or ordinal covariate  $X$ , a



division has the form  $X \leq c$  where  $c$  is a constant. For a categorical covariate  $X$ , a potential division has the form  $X \in c_1, \dots, c_k$  where  $c_1, \dots, c_k$  are possible values of  $X$ . The typical algorithm starts at the root node with all observations, searches through all potential binary splits with the covariates, and selects the best one according to a split criterion as a measure of impurity. In the CART approach, the process is repeated recursively until a stopping criterion is met. Thus, a large tree is produced that normally overfits the data. To find an appropriate subtree, a pruning and selection method is applied. In survival analysis the regression tree at each node uses the Kaplan-Meier estimate of the survival function.

## 7.2 Decision tree

There are several ways to represent a tree: hierarchical, inclusion diagram, bar diagram, numbering by levels, among others. In this work, the hierarchical form is used, represented in the figure 7.1

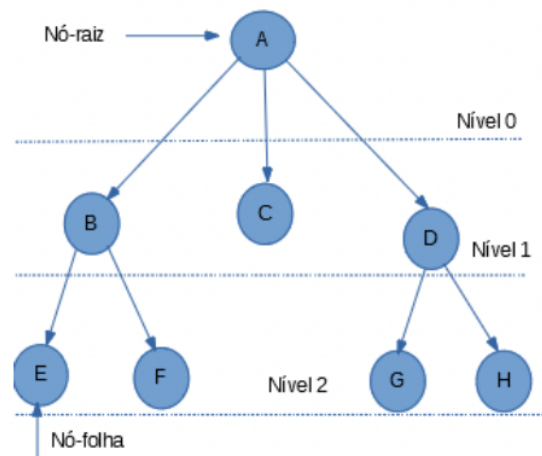


Figure 7.1: Decision Tree

The root node level is always 0. The height of a node is the length of the longest path between it and a leaf. Decision trees represent the conjunction and disjunction of attributes. Each path from the root of the tree to a leaf corresponds to a conjunction of attribute tests, and the tree itself to a disjunction of these conjunctions. A node can be called a decision node or a leaf node. A decision node can be split into two nodes (a binary split). This binary division is determined by a boolean condition that can be satisfied ("yes") or not satisfied ("no") by the observed value of this variable. Izenman (2008) [12] cites an example of a recursive partitioning involving two input variables,  $X_1$  and  $X_2$ , whose tree is represented in Figure 7.1. Figure 3 shows the resulting partition into 5 regions.

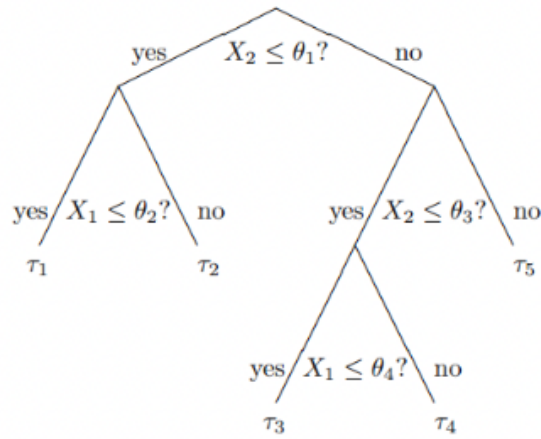


Figure 7.2: Decision tree with nodes

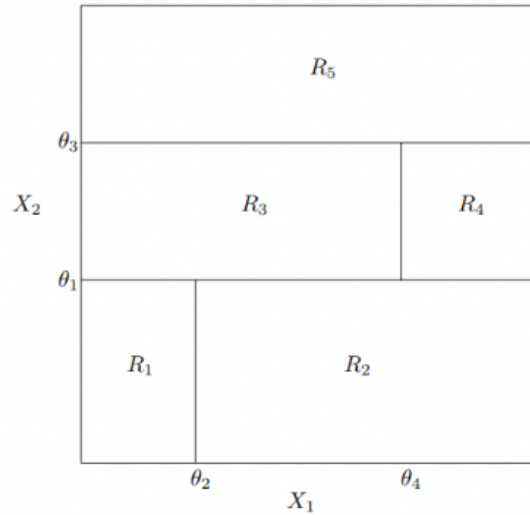


Figure 7.3: Recursive partitioning of five regions in  $\mathbb{R}^2$ ,  $R_1 - R_5$ , corresponding to the five terminal nodes.

The possible steps of this tree are as follows:

(1)  $X_2 \leq \theta_1$ ? If the answer is "yes", follow the left branch; otherwise, follow the right branch.

(2) If the answer to (1) is "yes", then the following question is asked:  $X_1 \leq \theta_2$ ? The answer "yes" produces the leaf node  $\tau_1$  with the corresponding region  $R_1 = X_1 \leq \theta_2, X_2 \leq \theta_1$ ; the answer "no" produces the leaf node  $\tau_2$  with the corresponding region  $R_2 = X_1 > \theta_2, X_2 \leq \theta_1$ .

(3) If the answer to (1) is "no", ask the next question:  $X_2 \leq \theta_3$ ? If the answer is "yes", then ask:  $X_1 \leq \theta_4$ ? If the answer is "yes", the leaf node  $\tau_3$  must be produced with the corresponding region  $R_3 = X_1 \leq \theta_4, \theta_1 < X_2 \leq \theta_3$ ; otherwise, the right branch to the leaf  $\tau_4$  node with the corresponding region  $R_4 = X_1 > \theta_4, \theta_1 < X_2 \leq \theta_3$ .

(4) If the answer to (3) is "no", the leaf node  $\tau_5$  is arrived at with the corresponding region  $R_5 = X_2 > \theta_3$ .

It is assumed that  $\theta_2 < \theta_4$  and  $\theta_1 < \theta_3$ .

In the regression tree, the predicted response for a given observation is given by the average of the training observations that belong to the same leaf node. In the classification tree, it can be predicted that each observation belongs to the most frequent training observation class in the region to which it belongs.

The induction algorithm must choose which predictive attribute will be used in each node of the tree. This choice can be based on different criteria, such as impurity, distance or dependence. Most algorithms try to split the data of a node in order to minimize the degree of impurity of the child nodes.

In classification, the sum of squares cannot be used as a criterion for binary divisions. An alternative to be used is the classification error rate. Misclassification is one of the common impurity measures used in binary decision trees.

### 7.2.1 Classification Trees

Classification aims to allocate objects from a population into one, two or more categories, based on a set of characteristics in each object. For example, classifying patients into low, medium and high risk groups.

In classification, as in regression, the data comprise  $n$  pairs  $(X_i, Y_i), i = 1, 2, \dots, n$ . It is important to use the data to define which components of the covariate vector  $X$  are needed to determine which category  $Y_i$ , the  $i$ -th observation belongs to. This information can be used to search for a function of explanatory variables that identify the class for a given  $X$  (James et al., 2013) [13].

The simplest classification problems separate a population into two classes, labeled 1 and 2. Binary classification problems can almost always be generalized to classification problems with multiple classes. The task is to find a decision function to discriminate between data from  $k$  different classes, where  $k \geq 2$ . The training set of a classifier  $f$  of samples  $(x_i, y_i)$  for  $i = 1, \dots, n$  where  $x_i \in \mathbb{R}^P$  are feature vectors and  $y_i \in \{1, \dots, K\}$  is the class label for the  $i$ -th sample. Based on the training set, the main objective is to learn the decision rule,

$$f(x) : \mathbb{R}^P \longrightarrow \{1, \dots, K\}$$

used to separate the  $K$  classes and predict the class label for a new entry  $x = x_{new}$ . Generally, a pre-trained multiclassifier is associated with a  $K$ -dimensional function

$$D(x) = (d_1(x), \dots, d_K(x))$$

where  $d_K(x)$  represents the strength of evidence that  $x$  belongs to class  $K$ . The classifier is induced from  $f$  and defined as

$$f(x) = \arg \max_{k=1, \dots, K} d_k(x)$$

The decision boundary between classes  $k$  and  $l$  is described by the set

$$\{x \in \mathbb{R}^P : d_k(x) = d_l(x)\} \forall k \neq l$$

If  $K$  is not too big, one way to simplify multiclass problems is to turn them into a series of binary problems. Each  $d_k(x)$  is trained to separate the class  $k$  from the rest. These  $K$  binary classifiers are then combined to give a final classification

$$\hat{f}(x) = \arg \max_{k=1, \dots, K} \hat{d}_k(x)$$

## 7.2.2 Regression Trees

To develop a regression tree, suppose that a dataset consists of  $p$  entries and a response, for each of the  $N$  observations. According to James et al.(2013) [13], there are two steps for building a regression tree:

Step 1: Split the predictor space, the set of possible values for  $X_1, X_2, \dots, X_p$  in  $J$  distinct, non-overlapping regions,  $R_1, R_2, \dots, R_J$ ;

Step 2: For each observation that occurs in the region  $R_j$ , the same prediction is made, which is the average of the response values for the training observations in  $R_j$ .

In step 1 two regions  $R_1$  and  $R_2$  are obtained and that the average response of the training observations in the first region is 10, while the average response of the observations of training in the second region is 20. Then, for a given observation  $X = x$ , if  $x \in R_1$  the value 10 will be predicted and if  $x \in R_2$  the value 20 will be predicted.

For step 1, in theory, the regions can have any shape. The predictor space is divided into rectangles or boxes, due to the simplicity and ease interpretation of the resulting predictive model. The objective is to find the boxes  $R_1, \dots, R_J$  that minimize the sum of squares.

$$\sum_{j=1}^J \sum_{i: x_i \in R_j} (y_i - \hat{y}_{R_j})^2$$

where  $\hat{y}_{R_j}$  represents the average response for the training observations inside the  $j$ -th box.

It is unfeasible to consider all possible partitions of the feature space into  $J$  boxes. Then we proceed with the recursive binary division. So, starting with all the data, consider a division variable  $j$  and division points  $s$ , and define the pair of "middle planes".

$$R_1(j, s) = \{X | X_j < s\} \text{ e } R_2(j, s) = \{X | X_j \geq s\}$$

where  $\{X | X_j < s\}$  means the region of the predictor space in which  $X_j$  takes on a value less than  $s$ .

One should look for the variable  $j$  and dot  $s$  that minimize the equation:

$$\sum_{x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

where  $\hat{y}_{R_1}$  is the average response for the training observations in  $R_1(j, s)$  and  $\hat{y}_{R_2}$  is the average response for the observations of training in  $R_2(j, s)$ . Once the best split is found, the data is split into the two resulting regions and the splitting process is repeated in each of the two regions. In order to minimize the sum of squares within each of the resulting regions. This process continues until a criterion is met.

The regions  $R_1, R_2, \dots, R_J$ , the response variable is predicted for a given test observation using the average of the training observations from the region that the test observation belongs to. When the resulting tree is very complex, good predictions can be generated in the training set, causing poor performance for the test set. In this way, a smaller tree with fewer divisions (fewer regions  $R_1, R_2, \dots, R_J$ ) can generate a smaller variance and better interpretation with a small bias.

So a better strategy is to grow a  $T_0$  tree a lot, stopping the splitting process only when some minimum number of nodes is reached. This large tree is pruned using cost complexity pruning. Cost complexity pruning provides a way to select a small set of these subtrees, because instead of considering each one, it considers a sequence of trees indexed by a non-negative adjustment parameter  $\alpha$ .

A subtree  $T \subset T_0$  is defined as any tree that can be obtained by pruning  $T_0$ . End nodes are indexed in  $m$ , with node  $m$  representing the region  $R_m$  (the rectangle or subset of the predictor space). The number of terminal nodes is denoted by  $|T|$ . Considering a sequence of trees indexed by a non-negative adjustment parameter  $\alpha$ , each value of  $\alpha$  corresponds to a sub-tree  $T \subset T_0$ , such that:

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

is as small as possible. The variable  $\hat{y}_{R_m}$  is the predicted response associated with  $R_m$ , that is, the average of the training observations in  $R_m$ . The  $\alpha$  fit parameter controls the trade-off between the size of the tree and the goodness of fit for the data.

### 7.3 Random Forests

The **Random Forests** (RF) technique is a powerful approach to data exploration, data analysis and predictive modeling. This technique was developed by Leo Breiman (creator of CART) at the University of California, Berkeley (Breiman, 2001) [3].

RF's are based on CART, Learning Ensembles, Committees of Experts and Bagging (Bootstrap Aggregation) (Dietterich, 2000) [7]. In Bagging you work with a sample of trees. However, the same complete set of predictors

is used to determine each split in the data. This results in a high correlation between the constructed trees, which are all very similar, resulting in little diversity (James et al., 2013) [13]. On the other hand, in RF's randomness is introduced by selecting random subsets of predictors for each division (Breiman, 2001) [3].

According to the website <https://dimensionless.in/introduction-to-random-forest/>, each of the decision trees results in a biased classifier (since it considers only one subset of the data). Each captures different trends in the data. In Classification problems, the result of most trees is used to classify a class. In the case of Regression problems, the arithmetic mean of the predictions obtained in all trees is used to describe the global prediction. It is also possible with the use of validation data to assign more weight to more decisive (important) trees in relation to others.

Random forests (RF) represent a collection of decision/classification trees (CART) that follow specific rules in tree growing determination, splitting, tree combination, self-testing and post-processing. In this process, a set containing  $B$  random vector samples of the predictors under study are randomly and independently selected. For each of these  $B$  samples, a tree is built. The selected trees describe a sample I.I.D. (independently and identically distributed) of trees from a given forest or population. The constructed trees are combined in order to obtain a joint prediction.

**Growth** - The growth (expansion) of trees occurs by binary partitioning (each node (parent) is divided into no more than two children). Each tree is grown/expanded at least partially at random. Randomness is achieved by expanding each tree based on a subsample chosen at random from the training data. Randomness is also obtained during the process of splitting the tree at each node.

**Division** - Assume that there are  $K$  predictor variables in the problem at hand. A small subset of these predictor variables is randomly selected. Usually  $\sqrt{K}$  is used. For example, if  $K = 500$ , you select about  $m_{try} = 23$  columns from the data matrix. Subsequently, each node is divided with the "best" of the 23 variables (not among the 500).

Each tree is expanded to its maximum size and no pruning is performed. It has been shown that pruning impairs the performance of these trees. The trees are deliberately overfitted in order to obtain predictors that resemble the nearest neighbor, a very robust non-parametric technique. Let  $N_{tree}$  be the number of trees to build. The Random Forests algorithm follows these steps for each of the  $N_{tree}$  iterations:

1. Select a new bootstrap sample (with replacement) from the training set.
2. Expand the tree: add more branches (no pruning).
3. At each internal node, randomly select  $m_{try}$  predictors and determine the best split using only these  $m_{try}$  predictors.
4. Without pruning, register (save) the tree, as it is, in a directory dedicated to these trees.

## Chapter 8

# Survapp

The main goal of Shiny applications is to allow the use of some methodologies that are implemented in R in a friendly environment. In our case, we intend to use the most common models and methods for survival analysis.

The different methodologies will be implemented in the different menus of the application. In this chapter, the constitution of the application is explained.

An analysis is presented for different databases that allows any user of the application to understand the interpretation of the different methodologies.

The source code of the application is available at:

<https://github.com/EmanuelVieira1111/Survapp/blob/main/app.R>.

### 8.1 Shiny

Shiny is a framework for creating web applications using R code. It is designed primarily with data scientists in mind, and to that end, you can create pretty complicated Shiny apps with no knowledge of HTML, CSS, or JavaScript. On the other hand, Shiny doesn't limit you to creating trivial or prefabricated apps: its user interface components can be easily customized or extended, and its server uses reactive programming to let you create any type of back end logic you want [28].

In this section, some of the Shiny package [5] functions used in the Survapp application are presented.

#### 8.1.1 UI Layout

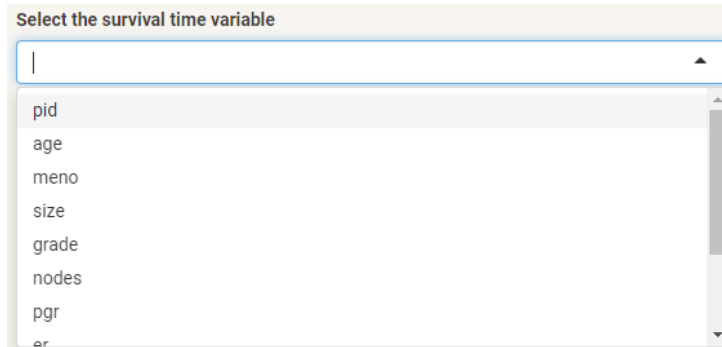
The user interface of Survapp is a navbar page with three tabpanels, About, Data.file and Survival analysis. The Survival analysis tabpanel is built by a navbar menu with several tabpanels representing the various methodologies for analyzing survival data. Each tabpanel has the constitution of a sidebarLayout with a sidebar panel and a main panel. The UI inputs shown below are the main elements in the sidebar panels of the entire application. UI Inputs are the application elements with which the user interactively interacts. In a reactive way, these inputs given by the

user are used by the server to generate the desired outputs in the main panels, be they graphs, tables or summaries.

## 8.1.2 UI Inputs

In this subsection are presented some of the Shiny UI Inputs functions used.

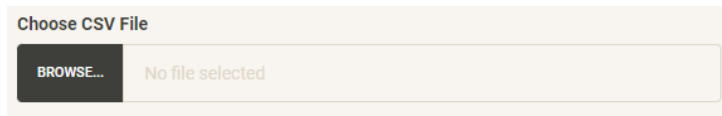
### Select input



The image shows a Shiny Select Input widget. The title is "Select the survival time variable". The dropdown menu is open, displaying a list of variables: pid, age, meno, size, grade, nodes, pgr, and er. The list is scrollable, and the current selection is empty.

Figure 8.1: Select Input for GBSG dataset.


### FileInput



The image shows a Shiny FileInput widget. The title is "Choose CSV File". There is a "BROWSE..." button and a text field displaying "No file selected".

Figure 8.2: File Input.

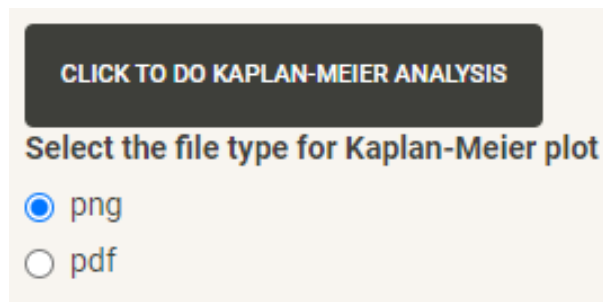
### Numeric input



The image shows a Shiny Numeric Input widget. The title is "Choose the estimation time". The text field contains the value "365".

Figure 8.3: Numeric Input.

### Action and radio buttons



The image shows a Shiny UI section. It contains an action button labeled "CLICK TO DO KAPLAN-MEIER ANALYSIS". Below it, the text "Select the file type for Kaplan-Meier plot" is followed by two radio buttons: "png" (selected) and "pdf".

Figure 8.4: Action and radio buttons.



## Check box group input

Select the covariates

- pid
- age
- meno
- size
- grade
- nodes
- pgr
- er
- hormon
- rfstime
- status

Figure 8.5: Check box group input for GBSG dataset.

### 8.1.3 UI Outputs

In this subsection are presented some of the Shiny UI Outputs functions used.

#### PlotOutput

Create an plot or image output element.

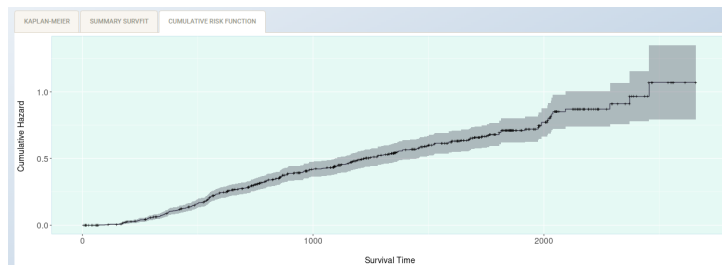


Figure 8.6: Plot output from cumulative risk function.

#### TableOutput

Create a table output element.

#### VerbatimTextOutput

Create a verbatim text output element.

DATA		DATA SUMMARY									
Show 10 entries		Search:									
	pid	age	meno	size	grade	nodes	pgr	er	hormon	rftime	status
1	132	49	0	18	2	2	0	0	0	1838	0
2	1575	55	1	20	3	16	0	0	0	403	1
3	1140	56	1	40	3	3	0	0	0	1603	0
4	769	45	0	25	3	1	0	4	0	177	0
5	130	65	1	30	2	5	0	36	1	1855	0
6	1642	48	0	52	2	11	0	0	0	842	1
7	475	48	0	21	3	8	0	0	0	293	1
8	973	37	0	20	2	9	0	0	1	42	0
9	569	67	1	20	2	1	0	0	1	564	1
10	1180	45	0	30	2	1	0	0	0	1093	1

```

Showing 1 to 10 of 686 entries
Previous 1 2 3 4 5 .. 69 Next

'data.frame':   686 obs. of  11 variables:
 $ pid      : int  132 1575 1140 769 130 1642 475 973 569 1180 ...
 $ age      : int  49 55 56 45 60 48 48 37 67 45 ...
 $ meno     : factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 1 ...
 $ size     : int  18 20 40 25 30 52 21 20 20 30 ...
 $ grade    : factor w/ 3 levels "1","2","3": 2 3 3 2 2 3 2 2 2 ...
 $ nodes    : int  2 16 3 1 5 11 8 9 1 1 ...
 $ pgr      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ er       : int  0 0 0 0 36 0 0 0 0 0 ...
 $ hormon   : factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 ...
 $ rftime   : int  1838 403 1603 177 1855 842 293 42 564 1093 ...
 $ status   : int  0 1 0 0 0 1 1 0 1 1 ...

```

Figure 8.7: Table and Text output from GBSG dataset.

### Download button

Create a download button or link.

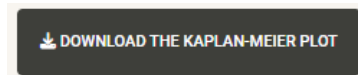


Figure 8.8: Download button.

## 8.1.4 Interface builder functions

HTML Builder Functions (a, br, code, div, em, h1, h2, h3, h4, h5, h6, hr, img, p, pre, span, strong, tags).

## 8.1.5 Rendering functions

Functions used in survapp server side code, assigning them to outputs that appear in user interface.

### renderPlot

Plot Output

### renderPrint

Printable Output.

### renderDataTable

Table output with the JavaScript library DataTables.

### downloadHandler

File Downloads.

## 8.1.6 Reactive programming

A sub-library that provides reactive programming facilities for R. In computing, reactive programming is a declarative programming paradigm concerned with data flows and the propagation of changes.

**reactive**

Create a reactive expression.

**observe**

Create a reactive observer.

**observeEvent**

Event handler (eventReactive).

**reactiveValues**

Create an object for storing reactive values.

## 8.2 Data.file

The data.file menu is the basis of the application, as it allows the user to import the desired database for analysis or use an example already contained in the application. Through a selectInput with the options "Load my own data" and "Example dataset" the user defines if he is going to do the analysis with his own data or with example data available as part of the application. Examples of data are GBSG (right censored), IR diabetes (interval censored) and Melanoma for competing risks analysis. There is no machine learning approach implemented for interval-censored data. The choice of one of these examples is allowed through a selectInput with their names as options, and this selectInput is only visible if the option "Example dataset" is selected in the previous selectInput. It is possible to download the sample databases through a downloadButton.

If the option selected in the first selectInput is "Load my own data", a file input appears and the user can choose the data to be imported in a csv format. The users can control some of the important parameters through radiobuttons. Selecting the header option places the first row of the database in the order corresponding to the column names. It is allowed to choose the separator type (comma, semicolon or tab) and the quote type (none, double quote or single quote).

In mainPanel, the data table, its summary, and its structure are given as output. Here it is possible to see if it is necessary to make any transformations in the classes of the variables for the future analysis.

In case it is necessary to transform the variable classes, there are two selectInputs and an actionButton. The first selectInput chooses the variable to transform; the second chooses the new class of the variable, having as options factor, numeric, integer, and character. By clicking on the action button, the transformation is performed.

Select an example dataset or upload your own

Load my CSV file

Choose CSV File

BROWSE... veteran.csv

Upload complete

Header

Separator

Comma

Semicolon

Tab

Quote

None

Double Quote

Single Quote

Choose column

Choose the class to change variable type

CHANGE CLASS

Figure 8.9: File input and change variable class.

### 8.3 Survival analysis

This section explains in detail each of the application's tabpanels referring to the methodologies for analyzing survival data. The R packages with greater relevance to the implemented methodologies and their respective functions are mentioned below.

#### **Package 'Survival'** [26]

Surv → Create a Survival Object.

survfit → Create survival curves.

survdif → Test Survival Curve Differences.

cox.zph → Test the Proportional Hazards Assumption of a Cox Regression.

coxph → Fit Proportional Hazards Regression Model.

survreg → Regression for a Parametric Survival Model

#### **Package 'clustcurv'** [27]

survclustcurves → Main function for determining groups of multiple survival curves and selecting automatically the optimal number of them.

**Package 'icenReg'** [2]

ic\_sp → Fits a semi-parametric model for interval censored data

ic\_par → Parametric Regression Models for Interval Censored Data

**Package 'rpart'** [26]

rpart → Recursive Partitioning and Regression Trees

**Package 'caret'** [16]

createDataPartition → Data Splitting functions

**Package 'ranger'** [29]

ranger → random forests

**Package 'cmprsk'** [11]

crr → Competing Risks Regression

cuminc → Cumulative Incidence Analysis

### 8.3.1 Kaplan Meier

Here, in case the data is right-censored, the user starts by selecting the variables related to the time and event of interest through two selectInputs. In the event that the data presents interval censoring, the variables referring to the left and right hand side of the interval are selected. After clicking the "Click to do full analysis" button, in mainPanel, automatically after selecting the variables, the graph of the Kaplan-Meier estimator, the print of the results, the summary, and the graph of the cumulative hazard function are generated. The user can, through a numericInput, select a specific time and obtain a print of the results of it. The option to download the generated graphics is given by clicking on the respective downloadsButtons with the possibility to choose the type of file, png or pdf.

Up to section 9 of this chapter we use data from a trial conducted by the German Breast Cancer Study Group (GBSG) in which a total of 720 women with primary node positive breast cancer is recruited in the period between July 1984 and December 1989. It retains the 686 patients with complete right censored data for the prognostic variables. Breast cancer is one of the most commonly occurring cancers in women. Fortunately, a large percentage of women survive their cancer for 1 year or more after diagnosis, but this prognosis depends on many things, including lifestyle factors, hormone levels, and some medical conditions. The database, available as part of the R survival package contains 11 variables. The rfstime and status variables are the most relevant for survival analysis. The rfstime indicates recurrence-free survival time, days to first recurrence, death, or last follow-up. Status indicates whether the patient is alive without recurrence (status=0) or if relapse or death has occurred (status=1). The remaining variables in the database are patient identifier, age in years, menopausal status, tumor size in millimeters, tumor grade, number of positive lymph nodes, progesterone receptors (fmol/ l), estrogen receptors (fmol/l), presence or absence of hormone therapy.

## a) Survival curve by the Kaplan-Meier estimator

Figure 8.10 shows the Kaplan–Meier curve of a study describing survivorship.

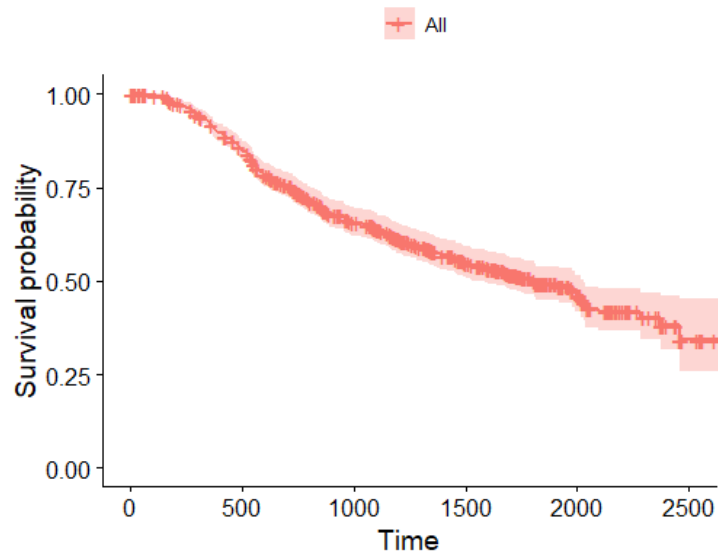


Figure 8.10: Kaplan-Meier estimator for GBSG dataset.

Table 8.1: Survfit for gbsg

```
Call: survfit(formula = Surv(rfstime, status) ~1, data = gbsg)
n      events  median  0.95LCL  0.95UCL
686    299    1807    1587    2030
```

In table 8.1 is the print of the survfit function. There are 686 observations and 299 events. The median survival is 1807 days, with the limits of the confidence interval being 1587 and 2030.

## b) Comparison of survival curves

Here, the user only needs to select the desired covariate and click on the button that allows executing the analysis. In the mainPanel are the following reactive outputs: graph with survival curves, Log-rank and Gehan-Wilcoxon tests (only for right censored data), print and summary of the Kaplan-Meier estimator for the different curves. Here, one can also download the generated graphic in PNG or PDF format.

In the German Breast Cancer data (gbsg database), there are two variables that show statistically significant differences between their levels; the hormone variable and the grade variable. The graphical differences are observed in Figure 8.11 and Figure 8.12. In both cases, the log-rank test obtained a probability value lower than 5%.

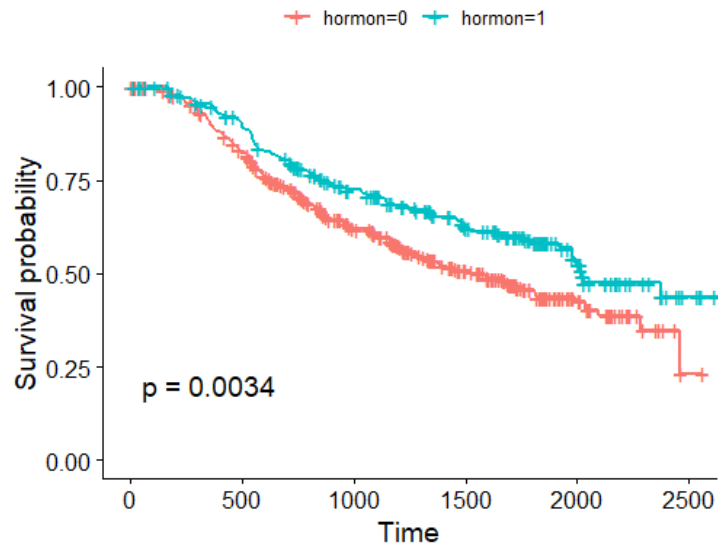


Figure 8.11: Survival Curves by Kaplan-Meier estimator as a function of the hormon variable.

Table 8.2: Survfit for hormon variable

Call: survfit(formula = Surv(rfstime, status) ~gbsg\$hormon, data = gbsg)					
	n	events	median	0.95LCL	0.95UCL
hormon=0	440	205	1528	1296	1814
hormon=1	246	94	2018	1918	NA

Table 8.2 shows the print obtained when using the survfit function with the hormon variable as a covariate. For the group with no therapy (hormone value 0), there are 440 observations and 205 events. The median survival time is 1528 days, with the limits of the confidence interval given by 1296 and 1814. For the group with therapy (hormone value 1), there are 246 observations and 94 events. The median survival time is 2018 days, with the lower limit of the confidence interval given as 1918 days.

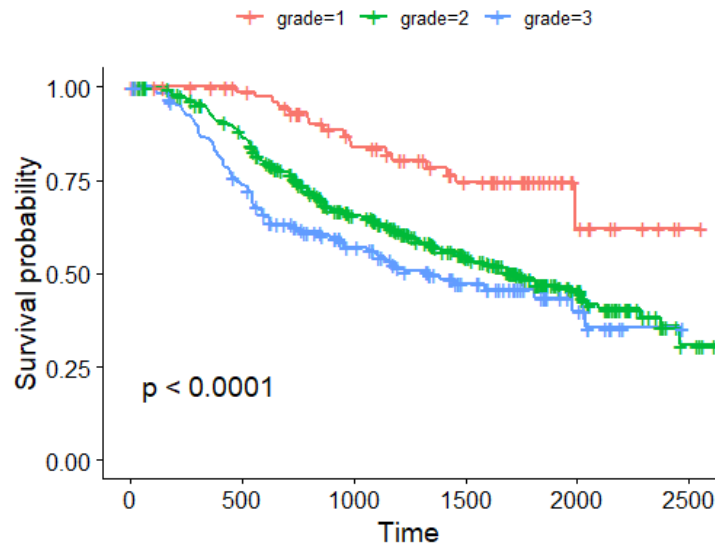


Figure 8.12: Survival Curves by Kaplan-Meier estimator as a function of the grade variable

Table 8.3: Survfit for grade variable

Call: survfit(formula = Surv(rfstime, status) ~grade, data = gbsg)					
	n	events	median	0.95LCL	0.95UCL
grade=1	81	18	NA	1990	NA
grade=2	444	202	1730	1493	2030
grade=3	161	79	1337	960	NA

Table 8.3 shows the print obtained when using the survfit function with the grade variable as a covariate. For the group with grade value 1, there are 81 observations and 18 events. The lower limit of the confidence interval is given by 1990 days. For the group with grade value 2, there are 444 observations and 202 events. The median survival time is 1730 days, with the limits of the confidence interval given by 1493 and 2030 days. For the group with grade value 3, there are 161 observations and 79 events, the median survival time is 1337 days, with the lower limit of the confidence interval given by 960 days.

### 8.3.2 Clusters of survival curves

In cases where the covariate used to compare survival curves has more than two levels, it is possible to try to find clusters of curves. However, this methodology is only possible for databases with the right censoring and with a binary event variable.

The mainPanel displays the summary and the graph with the identified clusters. The generated graph can be transferred in PNG and PDF format.

For this subsection the nodes variable is transformed into qualitative and node values greater than 13 are all stored with 14, so level 14 represents more than 13 nodes.



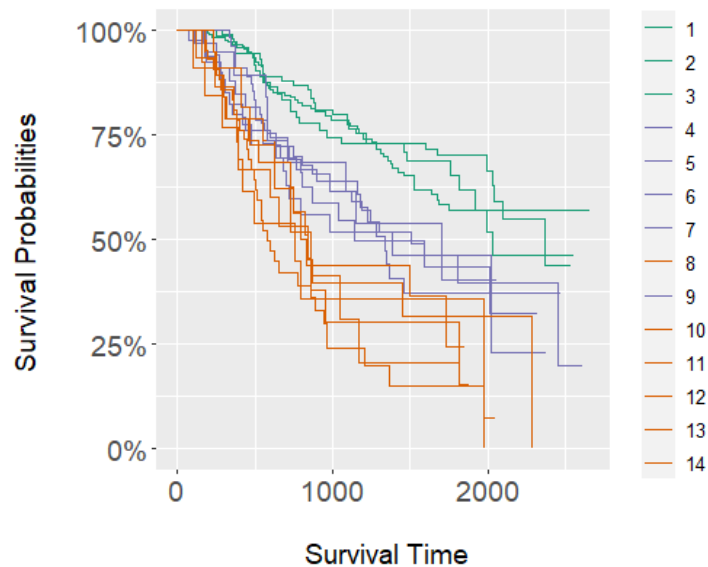


Figure 8.13: Clusters of survival curves

Three clusters are observed, represented by the green, blue, and orange colors. The first cluster in green is made up of three curves, in which individuals have 1, 2, or 3 nodes. The second in blue is made up of five curves, which represent individuals with 4, 5, 6, 7, and 9 nodes. The last orange cluster contains six individual curves, with 8, 10, 11, 12, 13 and more than 13 nodes.

### 8.3.3 Cox PH Model

In the Cox PH Model menu option contained in the sidebarPanel there is a checkboxGroupInput where the user can select the variables to be included in the model.

For right censored data in mainPanel, the print and graphs of the proportional hazards test of the model are reactively generated. It is possible to choose the combination of covariates that provides a better model since the Akaike Information Criterion (AIC) is also generated in the printout of the model.

For interval censoring, in mainPanel, the print of the model is generated.

Table 8.4 show the output for the Cox model with the lowest AIC, hence the best model according to the Akaike's Information Criterion. The probability value (p-value) for the likelihood ratio test (LRT) show that the model is clearly better than the null model. The LRT evaluate the null hypothesis that all of the betas ( $\beta$ ) are 0.

The variable nodes have a hazard ratio higher than 1 (HR = 1.0557), showing that for each additional nodule the hazard of recurrence or death increases by 5.5%. The p-value for pgr is 7.24e-05, with HR = 0.9977, showing that for each additional unit of progesterone receptors, the hazard of recurrence or death decreases by about 0.23%. The p-value for the variable hormone with hormone therapy (hormon = 1) is 0.01084, with HR = 0.7259, indicating a decrease of 28% in the hazard of recurrence or death for patients with hormone therapy. The level 2 of variable grade has a p-value of 0.00873 with a hazard ratio higher than 1 (HR = 1.92), indicating that the hazard of recurrence or

Table 8.4: Cox Proportional Hazard model for gbsg dataset

```
Call:
coxph(formula = Surv(rfstime, status) ~factor(grade) + nodes +
pgr + factor(hormon), data = gbsg)

            coef      exp(coef)    se(coef)      z      p
factor(grade)2  0.6523706    1.9200872   0.2487603   2.622  0.00873
factor(grade)3  0.8083537    2.2442103   0.2677781   3.019  0.00254
nodes           0.0542471    1.0557454   0.0067762   8.006  1.19e-15
pgr            -0.0022086    0.9977939   0.0005566  -3.968  7.24e-05
factor(hormon)1 -0.3202792    0.7259463   0.1257023  -2.548  0.01084
Likelihood ratio test=99.16 on 5 df, p=<2.2e-16
n= 686, number of events= 299
AIC value of the model:
[1] 3487.045
```

death increases 92% compared to those with grade = 1. The level 3 of variable grade has a p-value of 0.00254 with HR = 2.24, indicating that the hazard of recurrence or death increases by 124% compared to those with grade = 1, and 32% compared to those with grade = 2.

Looking at the Figure 8.14 the p-value for individual Schoenfeld test on the covariates grade is (0.0049) and pgr is (0.0398), so the null hypothesis, the assumption of proportional hazards for these covariates, is rejected. Graphically, it is observed that the residuals of these covariates are not parallel around 0. The global Schoenfeld test (p-value = 0.01418) rejects the proportional hazards assumption for this model.

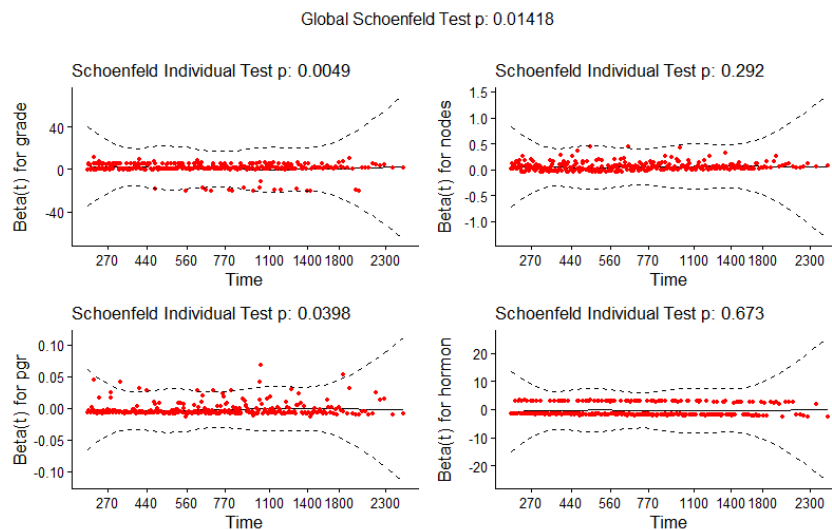


Figure 8.14: Graphical Test of Proportional Hazards

### 8.3.4 AFT Model

In the sidebarPanel of the AFT Model interface there are two checkboxGroupInputs, one to select the distribution to be used and the other to choose the covariates.

For right-censored data, the print of the parametric model is displayed on the mainPanel. It is possible to choose

the combination of covariates that provides a better model since the Akaike Information Criterion (AIC) is also given in the print of the model.

For interval censoring, the model print is generated in the mainPanel.

The accelerated failure time model can be considered as an alternative model to the cox model when the proportional hazards assumption does not hold.

Table 8.5: Accelerated failure time model for gbsg dataset

Call:				
survreg(formula = Surv(rfstime, status) ~factor(grade) + nodes +				
pgr + factor(hormon), data = gbsg, dist = "lognormal")				
	Value	Std. Error	z	p
(Intercept)	7.905705	0.174548	45.29	<2e-16
factor(grade)2	-0.499856	0.166663	-3.00	0.00271
factor(grade)3	-0.663241	0.184325	-3.60	0.00032
nodes	-0.054527	0.007690	-7.09	1.3e-12
pgr	0.001450	0.000352	4.12	3.7e-05
factor(hormon)1	0.309932	0.094693	3.27	0.00106
Log(scale)	-0.011888	0.044381	-0.27	0.78880
Scale= 0.988				
Log Normal distribution				
Loglik(model)= -2562.4 Loglik(intercept only)= -2618.9				
Chisq= 112.97 on 5 degrees of freedom, p= 9.6e-23				
Number of Newton-Raphson Iterations: 4				
n= 686				
AIC value of the model:				
[1] 5138.796				

Table 8.6: AFT interpretation

(Intercept)	grade2	grade3	nodes	pgr	hormon1
2712.7143	0.6066	0.5151	0.9469	1.001	1.3633

Table 8.5 show the results of the accelerated failure time model for the German breast cancer data. The AIC criterion was used to choose the covariates as well as the distribution of the model (log-normal). The p-value for the likelihood ratio test is significant, indicating that the model is significant better than the null model.

The level 2 from variable grade shorten survival time by  $\exp(-0.499856) = 0.6066$  times. The level 3 from variable grade shorten survival time by  $\exp(-0.6632) = 0.5151$  times. 1 unit change in nodes shorten survival time by  $\exp(0.0545) = 0.9469$ . 1 unit change in pgr extends survival time by  $\exp(0.0014) = 1.001$  times. The level 1 from variable hormon extends survival time by  $\exp(0.3099) = 1.3633$  times.

### 8.3.5 Regression Trees

To perform regression trees, in the sidebarPanel, there is a checkboxgroupInput where the user chooses the desired covariates. To perform the analysis just click on the actionButton. The mainPanel generates the decision

tree graphic, the print and summary of the Kaplan-Meier estimator for each node.

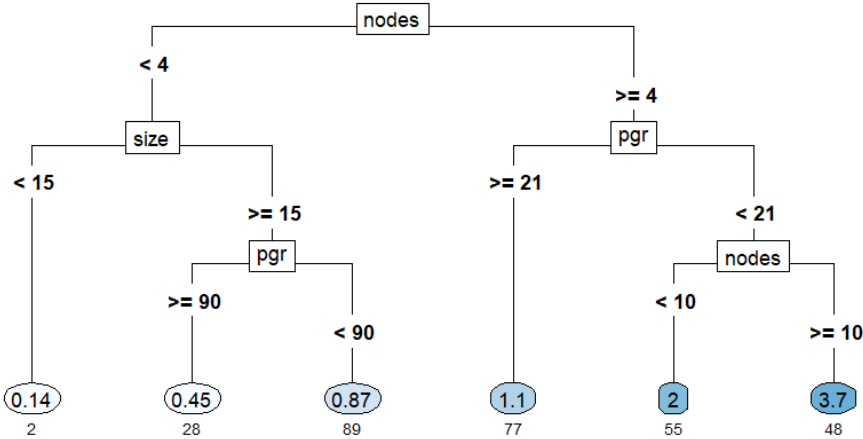


Figure 8.15: Regression tree for GBSG dataset

The decision tree for regression in the Figure 8.15 uses the variables nodes, size and pgr, resulting in 6 branches at the end.

The tree starts by dividing into 2, individuals that have less than 4 nodes (the ones with the lowest hazard of recurrence or death) and those that have 4 or more nodes.

Starting with individuals with less than 4 nodes, those with less than 15 millimeters of tumor have a mean hazard of recurrence or death 86% lower than all others, with 2 events being observed in these conditions (where = 3). When the tumor is 15 millimeters or more there is one more division, (where = 5) for individuals with 90 or more progesterone receptors where the mean hazard of recurrence or death is 55% lower compared to all others containing 28 events and (where = 6) for less than 90 progesterone receptors has a 13% lower mean hazard of recurrence or death than all others containing 89 events.

For individuals with 4 or more nodes those with 21 or more progesterone receptors (where = 8) have a 10% higher mean hazard of recurrence or death than all others containing 77 events. When there are less than 21 progesterone receptors there is another division, (where = 10) for individuals with less than 10 nodes with a 100% higher mean hazard of recurrence or death than all or others with 55 events and (where = 11) for individuals with 10 or more nodes, where there are 48 events with a 270% higher mean hazard of recurrence or death than all others.

In Table 8.7 is the print of the Kaplan-Meier estimator for each of the 6 branches of the decision tree. As already evidenced by the previous decision tree, nodes with (where = 3, 5 and 6) have a lower hazard of recurrence or death, as such, their median survival is higher than the other nodes.

In the Figure 8.16 is the plot of survival curves by the Kaplan-Meier estimator for each of the 6 branches of the decision tree.

Table 8.7: Print of survfit with different branches

Call: survfit(formula = Surv(rfstime, status) ~fit1\$where, data = gbsg)					
	n	events	median	0.95LCL	0.95UCL
fit1\$where=3	33	2	NA	NA	NA
fit1\$where=5	122	28	NA	NA	NA
fit1\$where=6	221	89	1989	1641	NA
fit1\$where=8	166	77	1701	1174	2018
fit1\$where=10	87	55	742	577	1366
fit1\$where=11	57	48	500	426	747

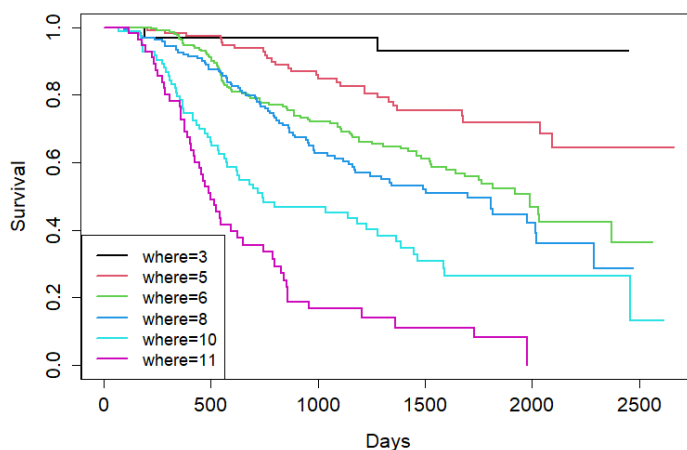


Figure 8.16: Kaplan-Meier for different branches of the regression tree in Figure 8.15

### 8.3.6 Classification Trees

Here, in the sidebarPanel, there is a checkboxgroupInput where the user chooses the desired covariates. To perform the analysis just click on the actionButton. The mainPanel generates the decision tree graphic and the model accuracy.

The best model is chosen taking into account pruning, which compares accuracy vs different values of complexity parameter. In the Figure 8.17 is the pruning chart. The best model has a complexity parameter of 0.02 and an accuracy of 73%.

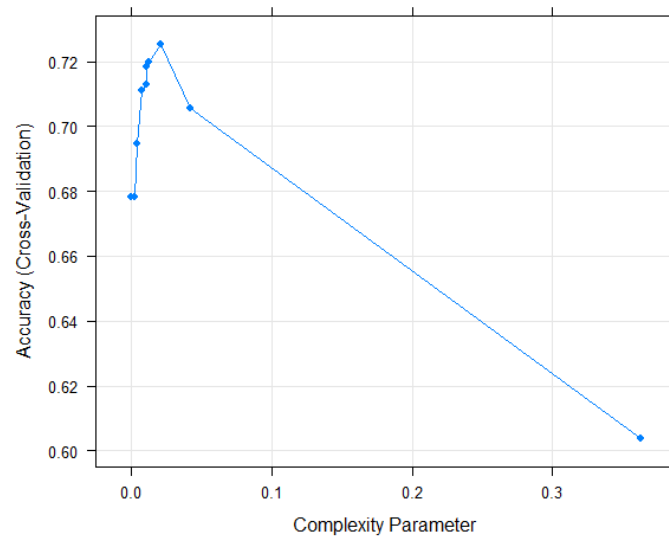


Figure 8.17: Pruning of classification trees for GBSG dataset

The decision tree for classification taking into account pruning is represented in Figure 8.18 and uses only the variable `rfstime` to predict one of two states of individuals, (0 = alive without recurrence) or (1 = recurrence or death).

According to the model, individuals with `rfstime` greater than or equal to 892 days remained alive without recurrence with a probability of 75% (236/315). Individuals with `rfstime` less than 85 days remained alive without recurrence with a probability of 100% (10/10). Individuals with `rfstime` between 85 and 892 days had recurrence or death with a probability of 72% (161/225).

The model accuracy is obtained by comparing the model predictions of the training data with the test data, in the calculated model it takes the value of 74%.

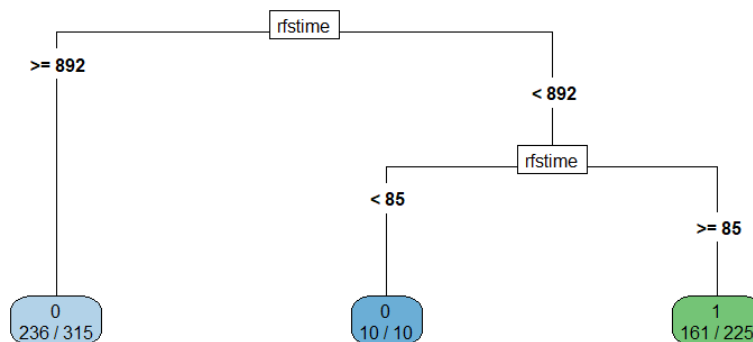


Figure 8.18: Classification tree for GBSG dataset

### 8.3.7 Random Forests

In the sidebarPanel of this interface, the user selects the covariates to be included in the model and then clicks on the action button to perform the analysis. The model graph, the prediction error and the importance of the variables are generated in the mainPanel.

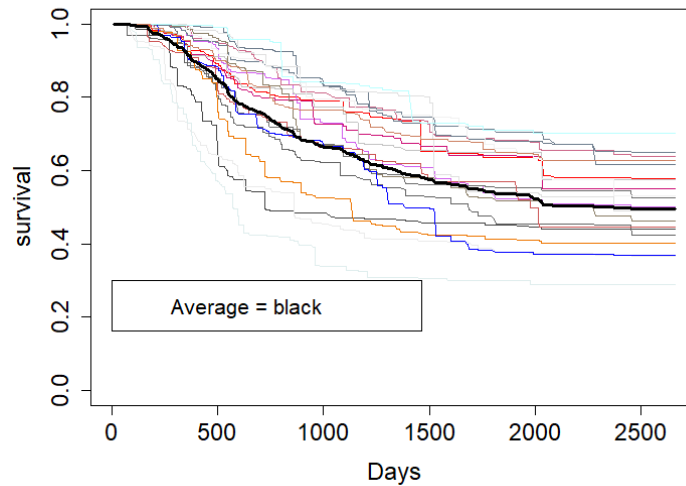


Figure 8.19: 20 survival curves of different decision trees for GBSG dataset

In the Figure 8.19, 20 survival curves of different decision trees are represented. In black is the curve of the average probability of survival over time with a median time to relapse or death of 1061 days and median survival of 66%.

The prediction error of this model is 0.32.

Table 8.8: Importance of variables for the GBSG database random forest model

	importance
nodes	0.0444
pgr	0.0144
hormon	0.0101
age	0.0099
grade	0.0087
er	0.0051
meno	0.0015
size	0.0007

In Table 8.8 the importance of each variable in the model is represented. The nodes variable is the most important, followed by pgr, hormon, age and grade. The remaining variables er, meno and size do not seem to have great relevance for the model.

### 8.3.8 Cumulative incidence

After having imported his data or chosen an example, the user selects, through two selectInputs, the variables referring to the time and the event of interest. It is also necessary to choose behind a numericInput the level of the variable of the event of interest that represents the censorship.

After the selected inputs, the graph and the print of the estimates and variances of the cumulative incidence are automatically generated. It is possible to download the graphic by clicking on the downloadButton and using the radioButtons to choose the file type PNG or PDF.

The dataset Melanoma will be used to illustrate a competitive risk analysis. The dataset consists of 205 measurements made on patients with malignant melanoma and is available as part of the boot R package. The variables present in the database are time (survival time in days), status (1 died from melanoma, 2 alive, 3 dead from other causes), sex (1 = male, 0 = female), age (in years), year of operation, thickness (tumor thickness in mm) and ulcer (1 = presence, 0 = absence).

Table 8.9: Estimates e variances for cumulative incidence

Estimates					
	1000	2000	3000	4000	5000
1	0.1274571	0.2301396	0.3096201	0.3387175	0.3387175
3	0.0342670	0.0504564	0.0581114	0.1059471	0.1059471
Variances					
	1000	2000	3000	4000	5000
1	0.0005481	0.0009001	0.0013789	0.0016907	0.0016907
3	0.0001628	0.0002451	0.0002998	0.0010401	0.0010401

Looking at the Figure 8.20 and Table 8.9 of the cumulative incidence functions, it is easy to see that at all times, the probability of a patient dying from melanoma is greater than the probability of dying from other causes. The variance of the cumulative incidence function for individuals who died of melanoma is always greater at all times than for individuals who died of other causes.



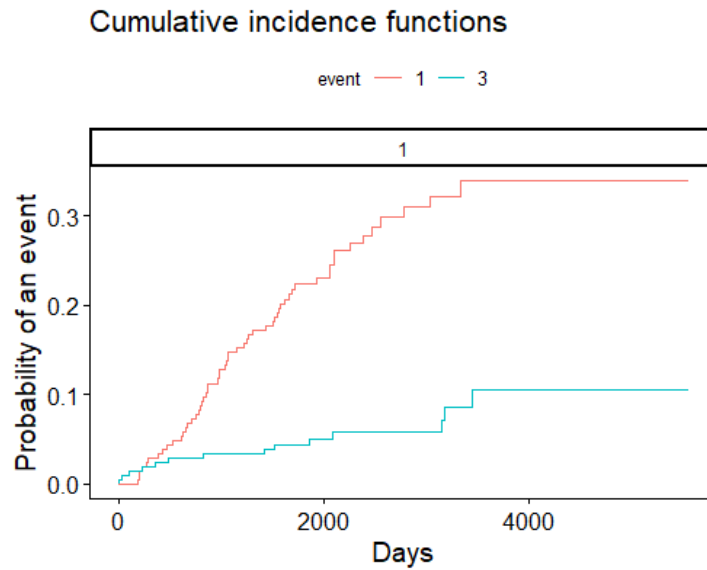


Figure 8.20: Cumulative incidence

### 8.3.9 Cumulative incidence between groups

In this menu, in the sidebarPanel, the user, through a selectInput, selects the desired covariate. The graphs, tests, estimates and variances of the cumulative incidence are automatically generated. It is also possible through a downloadButton to download the generated graphics, choosing one of the two radioButtons, with the options of the type of file to be transferred (png or pdf).

#### a) Sex variable

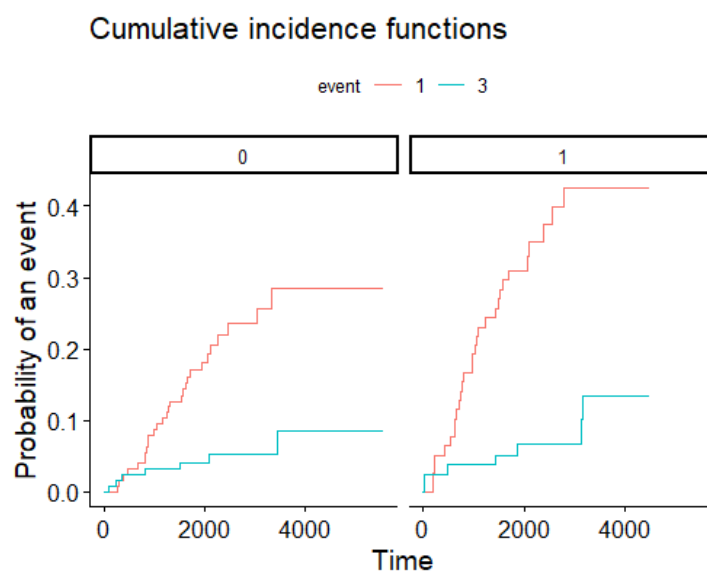


Figure 8.21: Cumulative incidence between groups (sex variable)

From Figure 8.21 we see that for death from other causes (in blue, status = 3), the cumulative incidence curves for the two genders are not very different. In Table 8.10 the p-value is 0.3553, so there is no statistically significant evidence to reject the equality of the curves, as such, it is concluded that death from other causes is not influenced by the individual's gender.

When the cumulative incidence functions of the sex variable are compared to death from melanoma (in red, status = 1), you can already see some difference looking at Figure 8.21. In the table 8.10 is the p-value of the test, which reveals statistical evidence to reject the equality of cumulative risk functions (0.0159), so it is concluded that the individual's gender influences death from melanoma. For all times, estimates of the probability of death from melanoma are always higher for men (sex = 1) than for women (sex = 0).

Table 8.10: Tests, estimates e variances for cumulative incidence between groups (sex variable)

Tests						
	stat	pv	df			
1	5.8140209	0.0158989	1			
3	0.8543656	0.3553203	1			
Estimates						
		1000	2000	3000	4000	5000
0	1	0.0873015	0.1807759	0.2356516	0.2842449	0.2842449
1	1	0.1923717	0.3100982	0.4245358	0.4245358	NA
0	3	0.0317460	0.0398351	0.0522064	0.0853838	0.0853838
1	3	0.0381412	0.0669394	0.0669394	0.1347427	NA
Variances						
		1000	2000	3000	4000	5000
0	1	0.0006378	0.0012450	0.0018102	0.0027555	0.0027555
1	1	0.0020223	0.0028196	0.0042695	0.0042695	NA
0	3	0.0002459	0.0003073	0.0004529	0.0015284	0.0015284
1	3	0.0004727	0.0008614	0.0008614	0.0029506	NA

## b) Ulcer variable

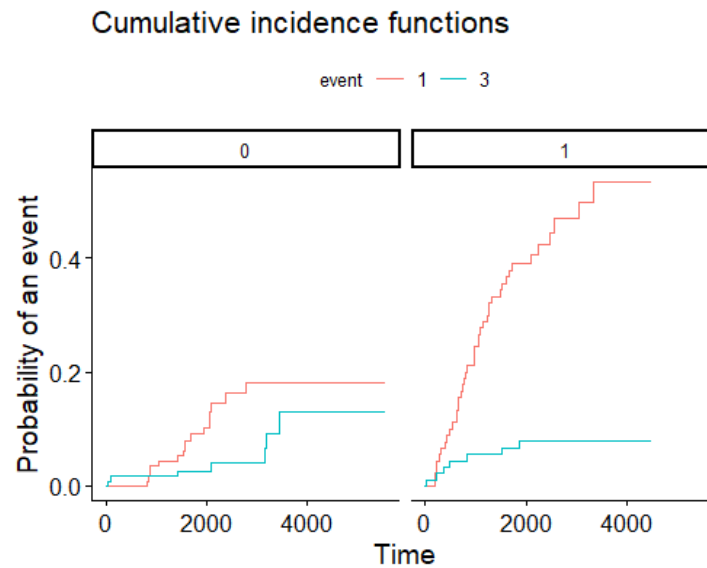


Figure 8.22: Cumulative incidence between groups (ulcer variable)

From Figure 8.22 we see that for death from other causes (in blue, status = 3), the cumulative incidence curves for the variable ulcer are not very different. In Table 8.11 the p-value is 6.903913e-01, so there is no statistically significant evidence to reject the equality of the curves, as such, it is concluded that death from other causes is not influenced by the presence or absence of ulcer.

When the cumulative incidence functions of the ulcer variable are compared to death from melanoma (in red, status = 1), you can already see some difference looking at the Figure 8.22. In Table 8.11 is the proof value of the test, which reveals the statistical evidence to reject the equality of cumulative risk functions (3.207240e-07), so it is concluded that the presence or absence of ulcer influences death from melanoma. For all times, estimates of the probability of death from melanoma are always higher for presence of ulcer (ulcer = 1) than for absence of ulcer (ulcer = 0).

### 8.3.10 Competing risk regression

In the competing risks regression menu, the user, through a selectInput, selects the desired approach, Subdistribution hazards or Cause-specific hazards.

#### a) Subdistribution hazard approach

Here, through a checkboxGroupInput, the user selects the covariates to be included in the model and chooses, through a numericInput, the level of the variable of the event of interest that represents the censorship. After the chosen inputs, the model is automatically generated in the mainPanel.

Table 8.11: Tests, estimates e variances for cumulative incidence between groups (ulcer variable)

Tests						
	stat	pv	df			
1	26.120719	3.207240e-07	1			
3	0.158662	6.903913e-01	1			
Estimates						
		1000	2000	3000	4000	5000
0	1	0.0350904	0.1032227	0.1816540	0.1816540	0.1816541
1	1	0.2444444	0.3897274	0.4697234	0.5330696	NA
0	3	0.0174682	0.0262408	0.0402817	0.1296081	0.1296081
1	3	0.0555555	0.0798143	0.0798143	0.0798143	NA
Variances						
		1000	2000	3000	4000	5000
0	1	0.0002997	0.0008952	0.0019180	0.0019180	0.0019180
1	1	0.0020796	0.0026929	0.0035308	0.0046320	NA
0	3	0.0001512	0.0002255	0.0004165	0.0029626	0.0029626
1	3	0.0005902	0.0008546	0.0008546	0.0008546	NA

For people who have not yet experienced such an event.

### Death from melanoma:

From the exponentials of the coefficients in Table 8.12 it is known that for every additional millimeter in tumor thickness, the hazard of dying from melanoma increases by 10%. Individuals with ulcers have a 223% higher hazard of dying from melanoma than individuals without ulcers.

Table 8.12: Subdistribution hazard for dead from melanoma

convergence:	TRUE
coefficients:	
thickness	ulcer1
0.09966	1.17300
standard errors:	
0.03579	0.30340
two-sided p-values:	
thickness	ulcer1
0.00540	0.00011

### Death from other causes:

From the exponentials of the coefficients in Table 8.13 it is known that for each additional year of age, the hazard of dying from other causes increases by 6%.

## b) Cause-specific hazards

Here, through a checkboxGroupInput, the user selects the covariates to be included in the model and chooses, through a numericInput, the level of the variable of the event of interest. After the chosen inputs, the model is automatically generated in the mainPanel.

Table 8.13: Subdistribution hazard for dead from other cause

```

convergence: TRUE
coefficients:
age
0.05868
standard errors:
0.01357
two-sided p-values:
age
1.5e-05
    
```

For individuals that are event free, didn't have any of the events

**Death from melanoma:**

Table 8.14: Cause-specific hazards for dead from melanoma

```

Call:
coxph(formula = formula, data = Melanoma)
      coef      exp(coef)  se(coef)  z      p
thickness 0.1140    1.1208   0.0361  3.158 0.00159
ulcer1    1.2180    3.3805   0.3091  3.941 8.12e-05
Likelihood ratio test=36.44 on 2 df, p=1.224e-08
n= 205, number of events= 57
    
```

From Table 8.14 it can be seen that for every additional millimeter in tumor thickness, the hazard of dying from melanoma increases by 12%. Individuals with ulcers have a 238% higher hazard of dying from melanoma than individuals without ulcers.

**Death from other causes:**

From Table 8.15 it can be seen that for each additional year of age, the hazard of dying from other causes increases by 8%.

Table 8.15: Cause-specific hazards for dead from other cause

```

Call:
coxph(formula = formula, data = Melanoma)
      coef      exp(coef)  se(coef)  z      p
age     0.07822    1.08136   0.02151  3.637 0.000276
Likelihood ratio test=15.77 on 1 df, p=7.153e-05
n= 205, number of events= 14
    
```

**8.3.11 Case Study Dataset IR\_diabetes**

In this section it is intended to show the operation and application for the analysis of interval censored data. We will use data from patients with Type 1 (insulin-dependent) diabetes. The data set is available as part of the icenReg R package. The data set contains interval censored survival time for the time from the onset of diabetes to diabetic

nephropathy. Accordingly, variables left and right correspond to the left and right sides of the observation interval. The gender of the subject is also available. This data set will be used to illustrate the application of the survival methods to interval censored data.

### a) Survival curves by the Turnbull estimator

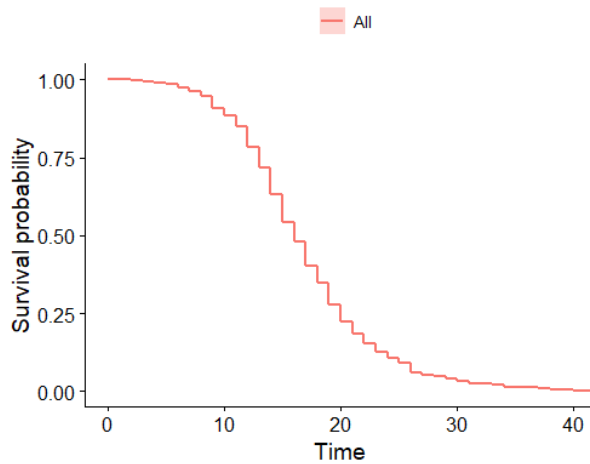


Figure 8.23: Kaplan-Meier estimator for the IR\_diabetes database

Table 8.16: Summary of the Kaplan-Meier estimator for the IR\_diabetes database

Call: <code>survfit(formula = Surv(left, right, type = "interval2") ~ 1, data = IR_diabetes)</code>				
n	events	median	0.95LCL	0.95UCL
731	731	16	15	17

There are 731 observations and 731 events. The median survival time is 16 days, with the confidence interval limits being 15 and 17 days.

#### Survival curves by gender

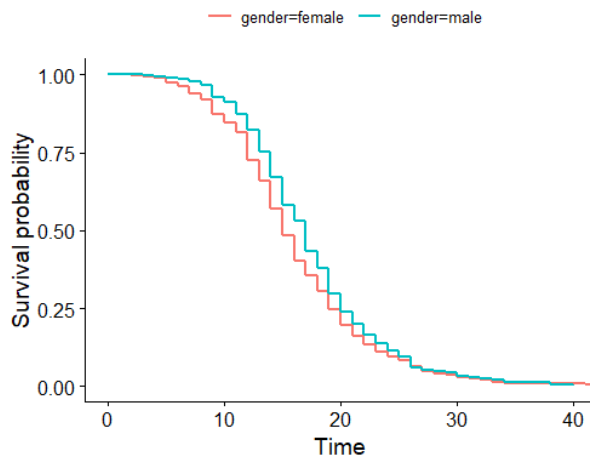


Figure 8.24: Survival curves of the Kaplan-Meier estimator as a function of gender for the IR\_diabetes database

Table 8.17: Summary of the Kaplan-Meier estimator for survival curves as a function of sex for the IR\_diabetes database

Call: survfit(formula = Surv(left, right, type = "interval2") ~gender, data = IR_diabetes)					
	n	events	median	0.95LCL	0.95UCL
gender=female	277	277	15	15	16
gender=male	454	454	17	16	18

For gender female there are 227 observations and 227 events. The median survival time is 15 days, with the confidence interval limits being 15 and 16 days. For gender male there are 454 observations and 454 events. The median survival time is 17 days, with the confidence interval limits being 16 and 18 days.

## b) AFT Model

The accelerated failure time model can be considered as an alternative model to the cox model when the proportional hazards assumption does not hold.

Table 8.18: Parametric model for IR\_diabetes

Call: ic_par(formula = cbind(left, right) ~gender, data = IR_diabetes, dist = "loglogistic")					
	Estimate	Exp(Est)	Std.Error	z-value	p
log_alpha	2.7730	16.0000	0.01379	201.000	0.0000
log_beta	1.5770	4.8420	0.03294	47.890	0.0000
gendermale	-0.1459	0.8643	0.07917	-1.843	0.0654
final llk	-2005.908				
Iterations	4				

As show in Table 8.18 it is estimated that the risk of diabetic nephropathy at any time will be approximately 14% times lower for men than for women.

## **Chapter 9**

# **Discussion and Future Work**

The developed application allows users to analyze survival data. For all the different methodologies, there are example databases contained in the application, but the main objective is for users to import their data and carry out their respective analyses.

The survapp can be found on [Shinyapps.io](https://shinyapps.io), an online service for hosting Shiny apps in the cloud. With a simple registration associating an email address, it is possible to access the application survapp. The initial objectives for the application were exceeded. However, in this type of application, there are always improvements to be made. A methodology in multi-state models, namely the illness-death model, is currently being implemented in the application.



# Bibliography

- [1] Odd O Aalen and Søren Johansen. “An empirical transition matrix for non-homogeneous Markov chains based on censored observations”. In: *Scandinavian Journal of Statistics* (1978), pp. 141–150.
- [2] Clifford Anderson-Bergman. “icenReg: regression models for interval censored data in R”. In: *Journal of Statistical Software* 81 (2017), pp. 1–23.
- [3] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.
- [4] Leo Breiman and Ross Ihaka. *Nonlinear discriminant analysis via scaling and ACE*. Department of Statistics, University of California Davis One Shields Avenue ..., 1984.
- [5] Winston Chang et al. “Package ‘shiny’”. In: See <http://citeseerx.ist.psu.edu/viewdoc/download> (2015).
- [6] David R Cox. “Regression models and life-tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.
- [7] Thomas G Dietterich. “Ensemble methods in machine learning”. In: *International workshop on multiple classifier systems*. Springer. 2000, pp. 1–15.
- [8] Jason P Fine and Robert J Gray. “A proportional hazards model for the subdistribution of a competing risk”. In: *Journal of the American statistical association* 94.446 (1999), pp. 496–509.
- [9] Thomas R Fleming, David P Harrington, and Margaret O’sullivan. “Supremum versions of the log-rank and generalized Wilcoxon statistics”. In: *Journal of the American Statistical Association* 82.397 (1987), pp. 312–320.
- [10] Edmund A Gehan. “A generalized Wilcoxon test for comparing arbitrarily singly-censored samples”. In: *Biometrika* 52.1-2 (1965), pp. 203–224.
- [11] Bob Gray and Maintainer Bob Gray. “Package ‘cmprsk’”. In: *Subdistribution analysis of competing risks. R package version 2* (2014), pp. 2–7.
- [12] Alan Julian Izenman. “Modern multivariate statistical techniques”. In: *Regression, classification and manifold learning* 10 (2008), pp. 978–.

- [13] Joby James, L Sandhya, and Ciza Thomas. "Detection of phishing URLs using machine learning techniques". In: *2013 International conference on control communication and computing (ICCC)*. IEEE. 2013, pp. 304–309.
- [14] Edward L Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.
- [15] Niels Keiding, Per Kragh Andersen, and John P Klein. "The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates". In: *Statistics in medicine* 16.2 (1997), pp. 215–224.
- [16] Max Kuhn et al. "Caret package". In: *Journal of statistical software* 28.5 (2008), pp. 1–26.
- [17] Nathan Mantel. "Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure". In: *Journal of the American Statistical Association* 58.303 (1963), pp. 690–700.
- [18] James N Morgan and John A Sonquist. "Problems in the analysis of survey data, and a proposal". In: *Journal of the American statistical association* 58.302 (1963), pp. 415–434.
- [19] Wayne Nelson. "Theory and applications of hazard plotting for censored failure data". In: *Technometrics* 14.4 (1972), pp. 945–966.
- [20] Ayse Ozhan, Melike Tombaz, and Ozlen Konu. "SmulTCan: A Shiny application for multivariable survival analysis of TCGA data with gene sets". In: *Computers in Biology and Medicine* 137 (2021), p. 104793.
- [21] Ross L Prentice et al. "The analysis of failure times in the presence of competing risks". In: *Biometrics* (1978), pp. 541–554.
- [22] Cristina Rocha and Ana Luísa Papoila. "Análise de sobrevivência". In: *XVII Congresso da Sociedade Portuguesa de Estatística. SPE*. 2009.
- [23] David Schoenfeld. "Partial residuals for the proportional hazards regression model". In: *Biometrika* 69.1 (1982), pp. 239–241.
- [24] Gustavo Soutinho and Luís Meira-Machado. "Analysis of Complex Survival Data: a tutorial using the Shiny MSM. app application". In: *arXiv preprint arXiv:2202.09160* (2022).
- [25] Gustavo Soutinho, Marta Sestelo, and Luís Meira-Machado. "survidm: An R package for Inference and Prediction in an Illness-Death Model." In: *R J*. 13.2 (2021), p. 22.
- [26] Terry M Therneau and Thomas Lumley. "Package 'survival'". In: *R Top Doc* 128.10 (2015), pp. 28–33.
- [27] Nora M Villanueva et al. "clustcurv: An R Package for Determining Groups in Multiple Curves." In: *R Journal* 13.1 (2021).
- [28] Hadley Wickham. *Mastering shiny*. " O'Reilly Media, Inc.", 2021.

- [29] Marvin N Wright and Andreas Ziegler. “ranger: A fast implementation of random forests for high dimensional data in C++ and R”. In: *arXiv preprint arXiv:1508.04409* (2015).
- [30] Yi Zhou et al. “MEPHAS: an interactive graphical user interface for medical and pharmaceutical statistical analysis with R and Shiny”. In: *BMC bioinformatics* 21.1 (2020), pp. 1–11.