

Extensions and Applications of Survival Trees in Medical Data

A thesis submitted

by

Alberto Alvarez Iglesias

to

School of Mathematics, Statistics and Applied Mathematics,
National University *of* Ireland, Galway

In conformity with the requirements for the degree of

PhD. in Statistics

September 2012

Supervisors: Dr. John Newell and Prof. John Hinde

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 2 |
| 1.1 | Datasets | 5 |
| 1.1.1 | Coronary Dataset | 6 |
| 1.1.2 | Breast cancer dataset | 6 |
| 1.2 | Structure of the thesis | 10 |
| 2 | Survival analysis | 14 |
| 2.1 | Introduction | 14 |
| 2.1.1 | The hazard function | 16 |
| 2.1.2 | Mean residual life function | 17 |
| 2.1.3 | Types of censoring | 20 |
| 2.1.4 | The likelihood of observed survival data | 22 |
| 2.2 | The Kaplan-Meier estimator of the survival function | 23 |
| 2.3 | The Logrank test | 25 |
| 2.4 | The Cox proportional hazards model | 31 |
| 2.5 | Applications of the Cox regression model | 33 |
| 2.6 | Chapter conclusion | 35 |
| 3 | Tree based methods | 37 |
| 3.1 | Introduction | 37 |
| 3.2 | Classification and regression trees | 39 |
| 3.2.1 | Continuous responses (regression tree) | 44 |
| 3.2.2 | Categorical responses (classification tree) | 47 |
| 3.2.3 | Extension of CART to Poisson regression | 48 |
| 3.3 | Survival trees | 49 |
| 3.4 | Conditional inference trees | 58 |
| 3.4.1 | Survival trees based on unbiased recursive partitioning | 61 |
| 3.5 | Random forest | 64 |
| 3.5.1 | Random survival forest | 65 |
| 3.6 | Chapter conclusion | 68 |

| | | |
|----------|--|------------|
| 4 | Trees based on node re-sampling | 71 |
| 4.1 | Introduction | 71 |
| 4.2 | Different trees, same structure | 73 |
| 4.2.1 | Synthetic example 1 | 75 |
| 4.3 | Node re-sampling: Continuous responses | 81 |
| 4.3.1 | Outliers in the response | 91 |
| 4.4 | Node re-sampling: survival responses | 93 |
| 4.4.1 | Synthetic example 2 | 93 |
| 4.4.2 | Interactions | 99 |
| 4.5 | Applications of the node re-sampling algorithm | 103 |
| 4.6 | Chapter conclusion | 109 |
| 5 | Estimating the mean residual life function | 111 |
| 5.1 | Introduction | 111 |
| 5.2 | The mean residual life function | 114 |
| 5.3 | Estimating MRL under non informative right censoring | 118 |
| 5.3.1 | Extrapolation of the survivor function using P-splines | 122 |
| 5.3.2 | A Parametric approach for estimating the MRL function | 124 |
| 5.4 | A semi-parametric approach based on extreme value theory | 128 |
| 5.4.1 | Description of the proposed method | 130 |
| 5.4.2 | Simulation study | 135 |
| 5.5 | Mean residual life trees based on node re-sampling | 142 |
| 5.6 | Chapter conclusion | 149 |
| 6 | Conclusions and future work | 151 |
| 6.1 | Future work | 155 |

List of Figures

| | | |
|-----|---|----|
| 1.1 | Example of a survival tree applied to a sample of women with breast cancer. | 3 |
| 2.1 | Examples of density, survival, hazard and MRL functions for different survival experiences. | 19 |
| 2.2 | Two different examples of the Kaplan-Meier estimate of the survival function. In (a) the probabilities of survival for women with breast cancer. In (b) the probability of disease free survival. | 25 |
| 2.3 | Kaplan-Meier estimate of the survival function applied to the coronary dataset. | 26 |
| 2.4 | Two examples of the Kaplan-Meier estimate of the survival function used to compare two different groups. In both cases the event of interest is the death of the patient. In (a) patients with cardiovascular disease with and without a previous myocardial infarction. In (b) women with breast cancer who tested positive and negative for HER2. | 30 |
| 3.1 | In (a) the cutpoint that minimizes the sum of squares of the left and right nodes. In (b) the sample (blue), regression line (black) and predicted outcomes generated by that cutpoint (red). | 40 |
| 3.2 | Graphical representation of the split generated in Figure 3.1. | 41 |
| 3.3 | Example of 7 splits using recursive partitioning. | 41 |
| 3.4 | Example of a regression tree with 8 terminal nodes and only one continuous predictor. | 42 |
| 3.5 | An example of overfitting. | 43 |
| 3.6 | Cross validated error as a function of the complexity parameter. | 46 |
| 3.7 | Optimal tree after pruning. On the left the structure of the tree. On the right the predicted values and the cut-points. | 46 |
| 3.8 | An example of a “pure” node (a) and the distance between the Kaplan-Meier estimate of the survival function (in a hypothetical node τ) and the survival function of the “pure” node (blue area in (b)). | 50 |

| | | |
|------|---|----|
| 3.9 | The cost-complexity plot applied to the coronary data using the method proposed by LeBlanc & Crowley (1992). | 55 |
| 3.10 | Tree obtained after applying the method proposed by Leblanc and Crowley (1992) to the coronary data. | 56 |
| 3.11 | Cost-complexity plot applied to the breast cancer dataset using the method proposed by Leblanc and Crowley (1992). | 57 |
| 3.12 | Unbiased recursive partitioning applied to the coronary data. . . . | 62 |
| 3.13 | Unbiased recursive partitioning applied to the breast cancer dataset. The event of interest is the recurrence of the disease. | 63 |
| 3.14 | Unbiased recursive partitioning applied to the breast cancer dataset. The event of interest is the death of the patient. | 64 |
| 3.15 | (a) Random survival forest applied to the coronary data. (b) Random survival forest applied to the breast cancer dataset (with recurrence as event). | 69 |
| 4.1 | An example in which 4 different structures generate virtually the same information. This highlights the point that there is no unique way of representing a tree. | 74 |
| 4.2 | Two representations of an interaction effect. | 75 |
| 4.3 | An example of a model represented in a tree fashion. In the terminal nodes the distribution of the response is assumed to be normal with means 60, 50, 70 and 30 and standard deviations of 20. | 76 |
| 4.4 | Output obtained using unbiased recursive partitioning after a random sample of 200 was drawn from the model presented in Figure 4.3 ($\alpha = 0.05$). | 77 |
| 4.5 | Output obtained using unbiased recursive partitioning after a random sample of 200 was drawn from the model presented in Figure 4.3 without noisy covariates ($\alpha = 0.10$). | 79 |
| 4.6 | Saturated tree based on CART. The data used to grow the tree was exactly the same than the data used to grow the tree in Figure 4.4. | 80 |
| 4.7 | Complexity parameter plot. The minimum is attained for the tree of size 7 but any other tree is within the 1SE of the achieved minimum and, therefore, the optimal tree is the tree with no splits. . . | 80 |
| 4.8 | An example of how recursive partitioning works at one particular node. For each predictor the algorithm records the maximum change in impurity (Y axis). | 82 |
| 4.9 | Bar-chart produced by the node re-sampling algorithm. | 83 |
| 4.10 | An example of the change in impurity plot for two continuous predictors. The plot on the left shows a meaningful split whereas the plot of the right shows a spurious split since the maximum values are attained in the boundaries of the possible splitting points. . . | 84 |

| | | |
|------|---|-----|
| 4.11 | Plot of the change in impurity of noise ₄ for 9 different bootstrap replicates. | 85 |
| 4.12 | In this example the 6 cutpoints leading to the maximum changes in impurity are considered. By doing this, splits that are spurious (like noise ₄ in this example) will produce a more meaningful value for the cutpoint. | 86 |
| 4.13 | An example of the relative importance plot produced by the node re-sampling algorithm. The value 1 represents equality for all the variables. In this example X_1 is clearly selected for the primary split. | 88 |
| 4.14 | Histogram of the bootstrap distribution of the cutpoints for X_1 . The red point is the median value. The two bumps observed in the histogram represent the two splitting points of this predictor in the underlying model. | 90 |
| 4.15 | The plot of the change of impurity for all the predictors after two outliers were introduced in the random sample (a). In (b), the plot of the change in impurity versus the cutpoints for the primary split noise ₁ | 92 |
| 4.16 | The relative importance plot after the node re-sampling algorithm was run with the two outliers (a). In (b), the histogram of the bootstrap distribution of possible cutpoints for X_1 | 92 |
| 4.17 | An example of a survival model represented in a tree fashion. The distribution of the response in each terminal node is assumed to be gamma with means and variances as specified in the squared boxes. | 94 |
| 4.18 | A snapshot of the graphical user interface created to accommodate the node re-sampling algorithm for survival responses. | 95 |
| 4.19 | Plot of the node re-sampling out of bag logrank for node 25 in Figure 4.18. | 95 |
| 4.20 | Survival tree based on the node re-sampling algorithm after all the irrelevant nodes have been eliminated. | 96 |
| 4.21 | Bootstrap distributions of the cutpoints for nodes 2 (a) and 3 (b) in Figure 4.20. | 97 |
| 4.22 | Survival tree using the the adapted version of CART for survival responses (LeBlanc & Crowley, 1992). The data used to generated the tree is identical to the one used to generate the tree in Figure 4.20. | 98 |
| 4.23 | Survival tree using conditional inference procedures (Hothorn <i>et al.</i> , 2006). The data used to generated the tree is identical to the one used to generate the tree in Figure 4.20. | 98 |
| 4.24 | An example of a model represented in a tree fashion with an interaction. The distribution of the response in each terminal node is assumed to be gamma with means and variances as specified in the squared boxes. | 100 |

| | | |
|------|--|-----|
| 4.25 | Snapshot of the optimal tree after running the node re-sampling algorithm with the underlying model containing an interaction term. | 100 |
| 4.26 | Plot of the out of bag values of the logrank statistic for each replicate and predictors in nodes 2 and 3 (left and right respectively). | 101 |
| 4.27 | Output obtained using unbiased recursive partitioning after a random sample was drawn from the underlying model with an interaction (Figure 4.24). | 102 |
| 4.28 | Plot of the complexity parameter versus the relative error for the tree based on CART where the underlying model has an interaction. | 102 |
| 4.29 | Survival tree based on node re-sampling for the coronary data. | 103 |
| 4.30 | Tree based on unbiased recursive partitioning (coronary data). | 104 |
| 4.31 | Bootstrap distribution of the cutpoints for Age in node 1 for the tree based on node re-sampling. | 105 |
| 4.32 | A snapshot of the graphical user interface after a survival tree has been grown for the breast cancer dataset (recurrence). | 106 |
| 4.33 | Distribution of “Size_rec” at node 4 in figure 4.32. | 106 |
| 4.34 | Survival tree based on unbiased recursive partitioning. Pathological variables, breast cancer dataset (recurrence). | 107 |
| 4.35 | Survival tree based on the node re-sampling algorithm. Biomarkers, breast cancer dataset (recurrence). | 108 |
| 4.36 | Survival tree based on the node re-sampling algorithm. Biomarkers, breast cancer dataset (recurrence). | 109 |
| 5.1 | Examples of different survival experiences. | 113 |
| 5.2 | Graphic representation of MRL at two different times. In (a) $t = 0$ and $MRL(0)$ is represented by the blue area. In (b) $t = 3$ and $MRL(3)$ is the blue area divided by $S(3) = 0.81$. | 115 |
| 5.3 | An example of the estimation of the MRL function when no censoring is present in the data. The blue line corresponds with the true MRL function. | 117 |
| 5.4 | An example of the estimate of the MRL at time 3 using the Kaplan-Meier estimate of the survivor function. The estimate is the red area divided by 0.79. | 119 |
| 5.5 | Estimates of the MRL function based on the Kaplan-Meier estimate of the survival function with corresponding 95% bootstrap confidence intervals. In (a) data were simulated using TSI whereas in (b) data were simulated using TSII (true MRL function in blue). | 120 |
| 5.6 | An example of the Kaplan-Meier estimate of $S(t)$ under TSI (left) and TSII (right). Whereas under TSI it is possible to get good estimates of MRL, under TSII it is not possible due to the fact that the gray area (right) cannot be estimated. | 121 |

| | | |
|------|--|-----|
| 5.7 | An example of the Kaplan-Meier estimate of the survivor function under TSI. The blue line is the true survivor function and the stepwise red line is the corresponding Kaplan-Meier estimate. | 123 |
| 5.8 | An example of extrapolation of the estimated survivor function using P-splines. Three fictitious points were located at $t = 20$, $t = 30$ and $t = 40$ (parts (a), (b) and (c) respectively). | 124 |
| 5.9 | An example of three different estimated survivor functions based on different choices of parametric shapes. The black stepwise curve is the KM estimator of the true survivor function (blue line). The dashed lines correspond to the estimated Weibull (black), log-normal (red) and log-logistic (green) distributions. | 126 |
| 5.10 | MRL functions corresponding to the estimated survivor functions in Figure 5.9. The blue curve is the true survivor function. The other curves correspond to the Weibull (black), log-normal (red) and log-logistic (green) distributions. | 126 |
| 5.11 | An example of the performance of the parametric approach for the estimation of the MRL function. The Y axis represents the deviations between the estimated and the theoretical values. . . . | 127 |
| 5.12 | Plot of the survival function of the GEV distribution for different values of the shape parameter. | 128 |
| 5.13 | Example of how to estimate the MRL function using the GPD distribution. The stepwise red line is the Kaplan-Meier estimate of the survival function. The dashed blue line is the estimated survival function using the GPD distribution. | 132 |
| 5.14 | The estimated MRL function using the GPD. In addition 95% bootstrap confidence intervals. The blue line is the true MRL function. | 133 |
| 5.15 | The estimated MRL function using the alternative approach based on the GPD. In addition 95% bootstrap confidence intervals. . . . | 134 |
| 5.16 | Areas where censoring occurs. The blue area corresponds to censoring due to lost to follow up or drop out. The red area corresponds to censoring due to the termination of the study. | 136 |
| 5.17 | Means of the absolute values of the differences between the estimated MRL function at time $t = 0$ and the theoretical value of the MRL function at time $t = 0$ | 139 |
| 5.18 | Standard deviations of the differences between the estimated MRL function at time $t = 0$ and the theoretical value of the MRL function at time $t = 0$ | 140 |
| 5.19 | Distribution of the differences between the estimated MRL function and the true MRL function both at time $t = 0$. The sample size is $n = 1000$ | 143 |

| | | |
|------|---|-----|
| 5.20 | Distribution of the differences between the estimated MRL function and the true MRL function both at time $t = 0$. The sample size is $n = 250$ | 144 |
| 5.21 | Distribution of the differences between the estimated MRL function and the true MRL function both at time $t = 0$. The sample size is $n = 50$ | 145 |
| 5.22 | Simulations based on method M1 for higher proportions of censoring (sample size $n = 1000$). | 146 |
| 5.23 | An example of a survival tree with the estimated survival functions in the terminal nodes. | 147 |
| 5.24 | An example of a survival tree with the estimated MRL functions in the terminal nodes. | 148 |
| 5.25 | Disease free survival by HER2. The blue lines correspond to HER2 negative patients whereas the red lines corresponds to HER2 positive patients. In (a), Kaplan-Meier estimates of the survival function. In (b), smooth estimates of the MRL functions. | 149 |
| 6.1 | Plot of the estimated MRL functions for HER2 positive (smooth red line, right node not shown here) and HER2 negative (smooth blue line, left node not shown here). The dots are the actual estimates of the MRL function. | 155 |

List of Tables

| | | |
|-----|--|----|
| 1.1 | Descriptive statistics for the coronary data covariates. | 7 |
| 1.2 | Descriptive statistics for the breast cancer data predictors. | 9 |
| 2.1 | Values of $Q(x_i, y_j)$ for the modified version of the Gehan statistic proposed by Efron (1967). | 27 |
| 2.2 | <i>Coronary dataset.</i> Logrank tests comparing the survival outcome of patients with and without previous history of myocardial infarction. | 30 |
| 2.3 | <i>Breast cancer dataset.</i> Logrank tests comparing the survival outcome of women who were HER2 positive versus women who were HER2 negative. | 30 |
| 2.4 | <i>Coronary data:</i> Cox proportional hazard model with Age, PreviousMI, Chol and ACE as covariates. Event of interest: death from any cause. | 33 |
| 2.5 | <i>Coronary data:</i> Cox proportional hazard model with Age, PreviousMI and ACE as covariates. Event of interest: death from any cause. | 33 |
| 2.6 | <i>Breast cancer dataset:</i> Cox proportional hazard model with Bcl2 as the only covariate. Event of interest: recurrence of the disease. | 34 |
| 2.7 | <i>Breast cancer dataset:</i> Cox proportional hazard model with HER2 as the only covariate. Event of interest: recurrence of the disease. | 34 |
| 2.8 | <i>Breast cancer dataset:</i> Cox proportional hazard model with Bcl2, Cdc7 and ER as covariates. Event of interest: death of the patient (overall survival). | 35 |
| 2.9 | <i>Breast cancer dataset:</i> Cox proportional hazard model with Bcl2 and ER as covariates. Event of interest: death of the patient (overall survival). | 35 |
| 4.1 | Distributions of the predictors involved in Example 1. Although it is not mentioned in the table, predictors X_1 and X_3 are correlated with correlation coefficient of 0.7. | 77 |
| 4.2 | Summary statistics of all 3 splits defined in Figure 4.3 after a random sample of size 200 was drawn from the underlying model. | 78 |

Abstract

Survival trees are a non-parametric modeling strategy that can be included in the area of statistical modeling. These type of models can be used as an alternative to Cox proportional hazards models for the analysis of survival data. Current methods for growing survival trees use estimates of the survival function in the terminal nodes of the tree. In this thesis, a new approach is proposed that incorporates estimates of the mean residual life function as the output of the model. The mean residual life function has been used traditionally in engineering and reliability but can also be used in the analysis of medical data. This function is easier to interpret (values are given in units of time) and summarizes the survival experience of the individuals under investigation in a very simple manner. In addition, this proposal is accompanied by two new methods for growing survival trees and for the estimation of the mean residual life function. The first method is based on a new algorithm called “node re-sampling” which uses bootstrapping to incorporate the sampling variability in the process of finding the optimal split in the nodes that are part of the tree. The second method is based on extreme value theory and aims to provide adequate estimates of the mean residual life function when the right tail of the underlying distribution is missing due to incomplete survival information due to termination of the study. Throughout the thesis, two datasets will be used to illustrate how the proposed methods can be used for the analysis of survival data. One of them consists of patients with cardiovascular disease and the other one consists of women diagnosed with breast cancer, both from the West of Ireland.

Acknowledgements

First of all I would like to thank Dr. John Newell who has been an excellent supervisor. Without his guidance and encouragement I would not have been able to complete this work. His vast experience as a lecturer in Statistics and as a statistical consultant have been an incredible asset for my PhD. His constant search to find new and simpler ways of explaining difficult statistical concepts motivated the initiation of this project and I have tried to follow his example during these years. But, above all, I would like to thank him for his friendship and support, especially in the difficult times, and I really hope that we can work together in future projects.

I would also like to thank Professor John Hinde, for his expertise and guidance. I really feel incredibly lucky to have worked with an internationally recognized statistician. Thanks to Dr. Carl Scarrott who spent six months in Galway during his sabbatical for his invaluable contribution to one of the Chapters. Thanks to Dr. Lyam Glynn and Prof. Grace Callagy who kindly made available the data that have been used in this thesis.

Many thanks to the Irish Research Council for Science, Engineering and Technology (IRCSET) for awarding me a postgraduate scholarship and for continuously funding projects like this.

Para concluir me gustaría agradecer también a mi familia, a mis padres Antonio y Carmina, los cuales me han apoyado siempre y me han animado a seguir estudiando y a aprender cosas nuevas. A mi hermano Carlos, que de alguna manera me inspiró a seguir el mismo camino iniciado por él. Por último me gustaría mencionar a mi querida hermana Zenaide que siempre ha estado conmigo cuando la he necesitado.

Chapter 1

Introduction

Survival trees is a generic term to refer to tree based methods applied to the analysis of survival data. These type of data arise when the outcome is the time until an event of interest occurs. Tree based methods are a non-parametric modeling strategy that can be included in the area of statistical modeling, and can be used as an alternative to generalized linear models or Cox proportional hazards models. A tree is basically a collection of binary partitions (nodes) that are defined (in a recursive manner) by the covariates included in the model. To illustrate some basic ideas related to these type of models, Figure 1.1 shows an example of a tree applied to survival data where a sample of women with breast cancer was used to build a survival tree (this dataset is described in more detail below).

In this example, the event of interest was the recurrence of the disease and the outcome was the time until the recurrence occurs (in months). A set of biomarkers (covariates) was considered for the construction of the model and the aim was to determine which one of the biomarkers (alone or in combination) had an effect on the time until recurrence. The three covariates presented in Figure 1.1 are HER2 (Human Epidermal Growth Factor Receptor 2), PR (Progesterone receptor) and ER (Estrogen Receptor). These are all proteins inside cells that, when over-expressed, have been associated with breast cancer. Each node in the tree represents a binary partition of the sample space. For instance, the split generated by HER2 divides the dataset into two groups, those who were HER2 positive and those who were HER2 negative. Among those who were HER2 positive, an additional partition was generated by ER, dividing the sample space into those who were ER positive and those who were ER negative. Moreover, women who were HER2 negative were further divided into two groups, PR positive and PR negative. Overall, this tree has partitioned the sample space into 4 parts which correspond to the four terminal nodes at the bottom of the tree. The information given in the terminal nodes (the output of the model) is related to the outcome

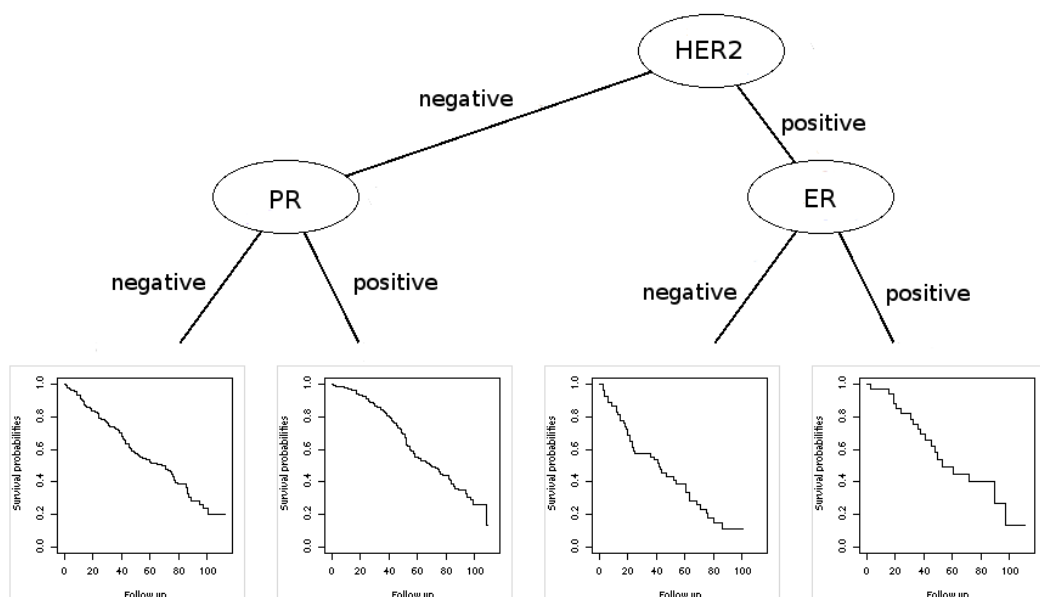


Figure 1.1: Example of a survival tree applied to a sample of women with breast cancer.

of interest and, generally, numerical or graphical summaries are provided. In this example, graphical summaries in the terminal nodes corresponds to the estimates of the survival functions of the time until recurrence. Without going into much detail, it seems that women who over-expressed HER2 (HER2 positive) and did not over-expressed ER (ER negative) have the worst prognosis. The best survival outcome seems to be for HER2 negative women and, among those, perhaps PR positive women have a better outcome than PR negative women.

It seems, however, that the interpretation of these plots is somehow complicated. In order to compare the survival outcome between the groups defined in the terminal nodes two strategies can be followed. One approach is to compare the medians, that can be approximated by identifying the values in the X-axis for which the survival function is 0.5. The other approach is to fix a time in the X-axis and to compare the probability that the survival outcome is greater than the selected time point. This approach is not very natural in the sense that the outcome being compared is given in terms of probability and not in units of time (the units in which the outcome of interest is measured). In this regard, the comparison of the medians is more natural but, in that case, a simple numerical summary would have sufficed as the outcome of the terminal nodes.

One of the main goals of this thesis is to incorporate a different and more interpretable function for the graphical summary in the terminal nodes of survival

trees. The aim is to use estimates of the mean residual life function as the output of the model. The mean residual life (MRL) function (Guess & Proschan, 1988) has been used traditionally in engineering and reliability (Watson & Wells, 1961; Kuo, 1984), although it can also be used in the analysis of survival data from the biomedical sciences (Gross & Clark, 1975). The MRL function at time t can be interpreted as the expected remaining lifetime of a patient given that the patient has survived up to time t (if the event of interest is the death of the patient). It also has a very nice interpretation in terms of summarizing the whole survival experience of the individuals under investigation. For instance, patients recovering from an operation or a transplant will display an increasing MRL function, whereas terminal patients with an incurable disease will display a decreasing MRL function. In this sense, the MRL function is very similar to the hazard function but with the advantage that the MRL function returns values in units of time rather than in units of risk.

It is the opinion of the author that this new approach will extend the capabilities of survival trees as a modeling tool and provide results that are easier to interpret. By using the MRL function as a graphical summary, the user will be able to obtain a model that describes the relationship between the predictors and the response in a very straightforward manner (as in the example above) and, at the same time, to obtain a model that returns values in units of time.

There are several challenges associated with the process of generating MRL trees. The first challenge is the generation of the survival tree itself. Although many different methods have been proposed in the literature, there are still some problems associated with the constructions of this type of model. One of these problems is related to the variable selection bias that has been identified by different authors such as White & Liu (1994), Shih (2004) and Kim & Yin Loh (2001). This is an inherent problem of the algorithm used to build this type of models which favors splits related to predictors with many different cutpoints, even if the corresponding predictor is not associated with the response. Other alternative methods, such as unbiased recursive partitioning proposed by Hothorn *et al.* (2006), are not affected by this problem, but are unable to identify interaction effects. Although little research has been done on this, it is a major drawback since the identification of interaction effects is one of the features that make tree based methods attractive. In this thesis, a novel method for growing survival trees will be explored that incorporates the sampling variability in the process of selecting the optimal split at any particular node. The proposed method aims not to be affected by any of the problems described above. The method involves the use of resampling procedures to generate the splits and it will be referred to as the **node re-sampling** algorithm.

The second challenge has to do with the estimation of the MRL function. It turns out that the MRL function is very dependent of the tail behavior of the un-

derlying distribution of the survival times. This fact makes the estimation of this function very complicated, especially with right censored data when the censoring is due to the study termination. In such a case, the right tail of the distribution is missing and some assumptions have to be made about this missing part of the distribution. The majority of methods for estimating the MRL function seem not to take this problem into account and fail to give adequate estimates when the right tail is missing. A novel method is proposed for the estimation of the MRL function that is based on some results from extreme value theory. This is based on the assumption that the conditional distribution of the exceedences from a chosen threshold u can be approximated by the generalized Pareto distribution. This new semi-parametric approach uses non-parametric methods for the estimation of the bulk of the distribution and this parametric assumption for the right tail of the distribution.

A third challenge is related to the development of software that is able to generate the survival trees applying these new methods. The methods developed here are aimed to be used by applied statisticians and clinicians. It would be of no practical use if this proposal were not be accompanied by the corresponding software able to fit the new models.

To summarize, the main goals of this thesis are:

- To incorporate the MRL function as a graphical summary in the terminal nodes of survival trees;
- To create a new and more robust algorithm for growing survival trees that is not affected by the variable selection bias and is able to identify interaction terms in the model.
- To develop a method for the estimation of the MRL function when the right tail of the underlying distribution is missing because of right censoring due to the end of the study.
- To design a statistical package in R that contains the necessary software to implement all the new methods proposed in this thesis.

1.1 Datasets

Two datasets will be used throughout the thesis to illustrate some of the methods and applications proposed in this work.

1.1.1 Coronary Dataset

This sample consisted of a cohort of 1609 patients with cardiovascular disease from the West of Ireland (Glynn *et al.*, 2008). These patients were identified from a stratified random sample of 35 general practices. To generate this sample, practices were randomly selected, after stratification by practice type (single-handed or group) and location (rural or urban), from the Health Services Executive, Western Area and asked to participate in the study. Patients were defined as having cardiovascular disease if they had a history of myocardial infarction, angina, or revascularisation by percutaneous coronary intervention, or coronary artery bypass grafting. Data at baseline were collected between 2000 and 2001 and, after a period of 5 years, patients who had not died or experienced a cardiovascular event were censored at that point. Individuals for whom follow-up data ceased to be available were also censored.

Outcome

The primary endpoint was death from any cause. The secondary endpoint was a composite endpoint that included death from a cardiovascular cause or any of the following cardiovascular events: myocardial infarction, heart failure, peripheral vascular disease, and stroke. In this thesis only the primary endpoint will be used for the examples.

Covariates

A set of predictors will be used for the generation of a prognostic model for the death of the patient from any cause. Predictors include age, gender, body mass index, smoking status, previous cardiovascular events, previous co-morbidity, baseline clinical status, co-morbidity and medication. Table 1.1 shows some descriptive statistics of each predictor at baseline.

Aims

The goal of the analysis of this dataset was to determine the prognostic factors that have a significant effect on the primary endpoint of interest, which is death from any cause. To answer this question, proportional hazards models and survival trees will be used to determine which of the predictors have an effect on the survival outcome, and to evaluate the nature and extent of such effects.

1.1.2 Breast cancer dataset

This sample consisted of a cohort of women with invasive breast cancer from Galway University Hospitals. A total of 666 invasive breast tumors had sufficient

| Covariates | Names | Summaries |
|-------------------------------------|---------------------------|--------------|
| Age, mean (SD) | Age | 70.9 (9.2) |
| Body mass index, mean (SD) | Bmi | 27.2 (4.5) |
| Gender, % | Gender Male | 65.2 |
| Smoking status, % | Smoking Current smoker | 30.6 |
| Previous CVD ¹ event, % | | |
| MI ¹ | PreviousMI (yes) | 46.1 |
| Angina | PreviousAngina (yes) | 85.3 |
| HF ¹ | PreviousHF (yes) | 6.1 |
| Previous Comorbidity, % | | |
| PVD ¹ | PreviousPVD (yes) | 5.5 |
| Stroke | PreviousStoke (yes) | 4.6 |
| Previous thrombo embolic events | PreviousTE (yes) | 10.9 |
| Baseline clinical status, mean (SD) | | |
| Systolic blood pressure, mmHg | Systolic | 138.9 (19.5) |
| Diastolic blood pressure, mmHg | Diastolic | 80.9 (9.5) |
| Total cholesterol, mmol/L | Chol | 5.4 (1.0) |
| Comorbidity, % | | |
| Diabetes | Diabetes (yes) | 11.3 |
| Medication, % | | |
| Aspirin | Aspirin (yes) | 75.0 |
| Beta blockers | BBlocker (yes) | 46.4 |
| Lipid-lowering agent | Lipids (yes) | 47.4 |
| ACE ¹ | ACE (yes) | 24.9 |

Table 1.1: Descriptive statistics for the coronary data covariates.

¹CVD = Cardiovascular Disease; MI = Myocardial Infarction; HF = Heart Failure; PVD = Peripheral Vascular Disease; ACE = Angiotensin Converting Enzyme Inhibitors

tissue available and were therefore eligible for inclusion in a West of Ireland breast cancer series. A database containing clinicopathological information was initially obtained from the Department of Surgery, National University of Ireland, Galway. This database contained over 140 columns of data and the relevant information was extracted and recorded. The collection of the data started in 1999 and patients were followed up for a period of 6 years. After that period, patients who did not experience the event of interest were censored at that point. Data on participants were also censored where follow-up data ceased to be available.

Outcomes

There were two different outcomes considered for the analysis of this dataset, disease-free survival (DFS) and overall survival (OS). Both outcomes were measured in months. For DFS the event of interest was the recurrence of the disease. Patients who were alive with locoregional disease or distant metastasis, or were dead with evidence of disease progression, were considered as having the event. Patients who were alive and well, or were dead with no evidence of disease progression, were censored. For overall survival (OS) the event of interest was death due to disease progression.

Covariates

Here two groups of predictors will be considered, pathologic variables and biomarkers. The evaluation of the biomarkers was carried out by immunohistochemistry (IHC) on the tissue microarray (TMA) cores available from formalin-fixed paraffin-embedded (FFPE) “donor” biopsies. Patients were considered to be positive for a particular biomarker if the score for percentage of positive cells, as per Allred scoring system, was greater than 10%. There were three novel biomarkers, Cdc7, tMcm2 and pMcm2 for which the values were analyzed in terms of the percentage of positive cells. Table 1.2 shows the descriptive statistics of each predictor at baseline.

Aims

For the purpose of this thesis the goal of the analysis of this dataset is to identify the set of biomarkers that can help to predict the two survival outcomes, DFS and OS. The traditional way of doing this was based on the use of pathological variables, such as lymph node status (LN_0_1 in the dataset, yes/no), lymphovascular invasion (LVI_0_1 in the dataset, yes/no), tumor size and tumor grade. The use of biomarkers can help to refine prognostication in breast cancer survival outcomes. The aim is to identify the biomarkers that have a significant effect on

| Covariates | Names | Summaries |
|-----------------------------------|---------------------------|-----------|
| Tumor size (mm), median(IQR) | Size | 24 (20) |
| Has the patient nodes positive, % | LN_0_1 (yes) | 50.7 |
| Vascular invasion, % | LVI_0_1 (yes) | 53.8 |
| Invasive tumor grade, % | Grade_MERGE (1,2,3) | |
| | 1 = grade 1 | 11.3 |
| | 2 = grade 2 | 57.6 |
| | 3 = grade 3 | 31.1 |
| Biomarkers % | | |
| ER | ER_TMA_0_1 (+) | 65.9 |
| PR | PR_TMA_0_1 (+) | 55.6 |
| HER2 | HER2_TMA_0_1 (+) | 14.1 |
| Ki67 | Ki67_TMA_0_1 (+) | 30.2 |
| Bcl2 | Bcl2_TMA_0_1 (+) | 54.3 |
| EGFR | EGFR_TMA_0_1 (+) | 14.6 |
| p53 | p53_TMA_0_1 (+) | 21.1 |
| CK5/6 | CK56_TMA_0_1 (+) | 9.1 |
| CK14 | CK14_TMA_0_1 (+) | 21.2 |
| tMcm2 score for % positive cells | tMcm2_TMA_P (0,1,2,3,4,5) | |
| | 0 = negative | 20.8 |
| | 1 = < 1 % | 1.9 |
| | 2 = 1 - 10 % | 18.3 |
| | 3 = 10 - 33 % | 28.0 |
| | 4 = 33 - 66 % | 15.4 |
| | 5 = > 66 % | 15.6 |
| pMcm2 % + cells, median(IQR) | pMcm2_TMA_P | 1.2 (9.2) |
| Cdc7 % + cells, median(IQR) | CDC7_TMA | 0.5 (2.5) |

Table 1.2: Descriptive statistics for the breast cancer data predictors.

both outcomes, the time to recurrence and the death of the patient. Cox proportional hazard models will be used along with survival trees for the construction of prediction models.

1.2 Structure of the thesis

This thesis consists of four main Chapters (along with this introduction and the conclusions and discussion in Chapter 6). In Chapter 2 an overview of survival analysis applied to the biomedical sciences is presented. Chapter 3 includes an extensive review of tree based methods with particular emphasis on survival trees. In Chapter 4 a new approach for generating survival trees is explored and some of its properties are examined. Finally, in Chapter 5 a novel method for estimating the mean residual life function is introduced and a simulation study is also presented. The following paragraphs contain a brief introduction and summary of each of these main Chapters.

Chapter 2: Survival analysis

Survival analysis is a set of statistical techniques to analyze data in which the outcome of interest is the time until an event occurs. Although the outcome is usually measured on a continuous scale, the usual methods for analyzing continuous responses cannot be applied for different reasons. One of the reasons is that time data are positive and standard normal methods generally do not apply. Another reason is the presence of censoring. This defining feature of survival data arises when some of the observations are incomplete due to causes that are not under the control of the investigator.

The Chapter begins with the definitions and interpretations of some of the functions that can be used to define the distribution of survival times, these include the survival, the hazard and the mean residual life functions. The different mechanisms and types of censoring will also be explained in detail in this part. The construction of the likelihood function, which incorporates all the information obtained from both censored and uncensored observations will also be described. The importance of this part will become clear in the context of the estimation of the mean residual life function in Chapter 5.

In the second part of this Chapter, the Kaplan-Meier estimate of the survival function (Kaplan & Meier, 1958) will be presented along with some examples given by the coronary and the breast cancer datasets. The Kaplan-Meier estimate of the survivor function is generally used to represent the survival distribution in the terminal nodes of survival trees.

In the third part of the Chapter, the topic of comparing two survival distributions will be examined in detail. The logrank test (Mantel, 1966) and many

of its variants (Harrington & Fleming, 1982) can be used to determine whether the differences between the survival experience of two cohorts of individuals are statistically significant or not.

Finally, the Cox proportional hazards model (Cox, 1972) will be described and fitted to the two datasets already introduced.

Chapter 3: Tree based methods

Chapter 3 is devoted to the explanation of tree based methods, in particular tree based methods applied to survival data commonly known as survival trees. Two methods will be explained in detail for their particular relevance: the CART algorithm (Classification and Regression Trees) by Breiman *et al.* (1984), and the unbiased recursive partitioning algorithm by Hothorn *et al.* (2006). For the sake of completeness a section of this Chapter is also devoted to the random forest approach by Breiman (2001).

In the first part of the Chapter, the recursive partitioning algorithm, along with the CART idea of pruning, is explained in detail for continuous responses. An example will be used to explain how CART obtains the optimal size of the tree. The extension of the same ideas to categorical and Poisson responses is also presented.

The second part of the Chapter is devoted to the description of tree based methods applied to survival data. A few methods have been proposed which are versions of the CART algorithm adapted to survival analysis (Gordon & Olshen, 1985; Davis & Anderson, 1989; LeBlanc & Crowley, 1992). These methods aim to use the same algorithm for pruning the tree as that used for categorical and continuous responses. Other methods developed their own algorithms for growing and pruning the tree. This is the case of Segal (1988) and LeBlanc & Crowley (1993) that use the logrank statistic as a measure of “between node separation”.

In the third part of the Chapter the unbiased recursive partitioning by Hothorn *et al.* (2006) is presented. This is a modified version of the recursive partitioning algorithm and it was developed to overcome the problem of variable selection bias that had been identified by different authors, including White & Liu (1994), Kim & Yin Loh (2001) and Shih (2004). The unbiased recursive partitioning algorithm is based on a general theory of permutation tests developed by Strasser & Weber (1999) and can deal with a wide range of different outcomes, including continuous, categorical and survival responses. One of the defining features of this method is the fact that no pruning is necessary, since the tree stops growing whenever the test performed at each split is not significant. This feature simplifies the selection of the optimal tree, but has some negative implications in relation to the identification of interaction effects, as will be demonstrated in Chapter 4.

Finally, the random forest approach by Breiman (2001) is introduced. This

method was extended by Ishwaran *et al.* (2004) to the case of survival responses (random survival forest) and this will also be examined. Although it is not the main topic of this thesis, some of the ideas of random survival forests are used for the development of the node re-sampling algorithm.

Throughout this Chapter the coronary and breast cancer datasets will be used to illustrate some of the methods.

Chapter 4: Trees Based On Node Re-Sampling

Chapter 4 presents a deeper study into tree based methods. The current methods for growing survival trees seem not to be completely satisfactory and some methods still have different problems, such as the variable selection bias or the incapacity to detect interaction effects. In this Chapter a new algorithm for growing survival trees is proposed and some of its properties are analyzed. This new method aims to incorporate the sampling variability in the process of finding the optimal split at any particular node giving some degree of robustness to the splitting procedure of the algorithm. This approach is based on an algorithm called the node re-sampling algorithm and it uses bootstrapping at a node level to generate the different splits of the tree.

The Chapter begins with an example which demonstrates that the structure of a tree is not unique and that many different representations are possible. This example, along with the fact that the structure of the tree is also subject to sampling variability aims to motivate the novel approach based on node re-sampling.

In the second part of the Chapter, a preliminary version of the node re-sampling algorithm is presented and applied to continuous responses. This new methodology incorporates the sampling variability at each split by bootstrapping the observations available at each node. The selection of the primary and surrogate splits is made using the so-called relative importance plot which is based on the out-of-bag values of the logrank statistic (the set of out-of-bag observations corresponds with the observations that are not included in each bootstrap replicate of the data). The use of the out-of-bag values of the logrank statistic is the key aspect to avoid the variable selection bias in node re-sampling.

In the third part of the Chapter the survival data version of the node re-sampling algorithm is introduced and the pruning mechanism is explained in detail. One of the novel features of this new method is the possibility of pruning a saturated tree interactively. To enable this feature, a new graphical user interface has also been developed and some of its functions are also presented in this section. The pruning of the tree is based on the idea of “irrelevant” nodes, which is also related to the out-of-bag values of the logrank statistics.

In the fourth part of the Chapter, an example is given which aims to illustrate how the method works in the presence of interactions. It is also demonstrated

that unbiased recursive partitioning is not able to detect such interactions. In fact, only methods that grow a saturated tree first and then prune back the tree are able to find interaction effects.

Finally, in the last section, the new method is applied to the coronary and breast cancer datasets.

Chapter 5: Estimating the mean residual life function

In this Chapter the problem of estimating the MRL function is studied. Under non-informative right censoring various methods have been proposed in the literature for the estimation of this function (Gill, 1983), (Chaubey & Sen, 2008) or (Zhou & Jeong, 2011). However, in many practical situations there is no information about the right tail of the underlying survival distribution because of censoring at the end of the study. In such cases, many of the existing methods seem to fail to give appropriate estimates of the MRL function. The new method proposed aims to address this problem and is based on results from extreme value theory.

In the first part of the Chapter a general overview of the MRL function is given, along with the description of some methods for the estimation of such a function when no censoring is present in the data. In addition, the choice of theoretical settings to accommodate different types of censoring is examined.

The second part is devoted to the problem of estimating the MRL function under non-informative right censoring. In particular, it will be shown how many of the current methods fail to give sensible estimates when no data are available for the right tail of the underlying survival distribution. Some novel solutions will be explored based on the extrapolation of the available data using smoothing techniques and a parametric approach. These solutions, however, are not entirely satisfactory, as will be demonstrated in this section.

In the third part of this Chapter a novel semi-parametric approach is introduced which is a combination of parametric and non-parametric estimation of the area under the survival function of the underlying distribution. The basic idea is to assume that the conditional distribution of the exceedences from a chosen threshold follows a generalized Pareto distribution. This result is based on extreme value theory and its adequacy depends upon the chosen threshold being sufficiently large. The method will be described in detail and the results of a simulation study will also be presented.

Finally, in the last part the MRL function will be incorporated as a graphical summary for the node re-sampling survival trees. The breast cancer dataset will be used to illustrate how the novel approach proposed in this thesis can be used to analyze survival data from the biomedical sciences.

Chapter 2

Survival analysis

In this Chapter an overview of survival analysis is presented and the most relevant aspects concerning the study of survival data are examined. This subject has been discussed by many authors and many excellent books have been written in the last few decades (see for instance Lawless, 2002; Kalbfleisch & Prentice, 2002; Kleinbaum & Klein, 2011). The Chapter begins with an introduction to the topic and a description of the most common functions used in survival analysis. In this part the definitions of the different types of censoring that can be encountered with survival data are also given along with the description of the likelihood function under different types of censoring. The second part explores the estimation of the distribution of the survival times and the problem of comparing two survival distributions. These two aspects of survival analysis are especially relevant in the context of this thesis, in particular the use of the logrank test to determine if there are significant differences between two groups. In the third part, the problem of modeling survival data is explored and an introduction to the Cox proportional hazards model is given. Finally, in the last part, the datasets described in Chapter 1 will be used to fit Cox proportional hazards models.

2.1 Introduction

Survival analysis is a set of statistical techniques concerning the analysis of data where the outcome of interest is the **time until an event occurs**. Survival data are found in many different disciplines such as medicine, biology, epidemiology, engineering, public health, economics, demography, etc. In general, the time until the event occurs is referred to as the survival time, time to event, event time, lifetime or failure time, whereas the event of interest is usually related to the death, recurrence or failure of the individual/unit under investigation. For instance, in medicine, the outcome of interest could be the time until a patient dies after diagnosis, the time to relapse from remission, or the time to recovery

after an operation or a transplant. In epidemiology one may be interested in the incubation times of different diseases such as AIDS, hepatitis B, SARS etc. In engineering the time of interest might be the time until a particular machine fails. Other examples are the life times of the elderly in particular social programmes, felons time to parole (criminology), duration of first marriage (sociology), or length of magazine subscription (marketing). Most of the examples and terminology presented in this work are related to the study of biomedical data and, therefore, the event of interest will be, mostly, the death of a patient and the time to event the survival time.

One of the characteristics that makes the analysis of survival data different from the analysis of other response variables is the presence of **censoring**. An observation is said to be censored if the time recorded does not correspond to the time at which the event of interest occurs (incomplete observation). This might happen, for instance, if a patient dies from a different cause than that under study or if a patient withdraws from the study and is lost to follow up. In both cases, the observed time is the time in which the patient was last seen. The key feature of censoring is that the event of interest occurs at some point in the future (*right censoring*) but the investigator does not know when. Due to this special feature, the observed outcome is comprised of a continuous measurement (the incomplete observed times) and a binary indicator that specifies if an observation has been censored or not. It is also possible to have *left censoring* (the event of interest occurs at some unknown time in the past) or *interval censoring* (the event occurs at an unknown time on an interval) although here only right censoring will be considered.

To accommodate censoring in the analysis of survival data, specific methods have been developed through the years which differ from those used with other type of responses. The notion of “*individuals at risk at time t* ” is a key feature in survival analysis and it is used by most of the methods for analyzing survival data. The reason is that the group of individuals at risk includes both, censored and uncensored observations and therefore it is not affected by the fact that the observed times are incomplete. Furthermore, it allows the estimation of the probability of survival beyond time t given the survival up to time t . This conditional argument is also used to determine the distribution of the survival times through the hazard and mean residual life functions as will be defined in this Chapter. Both of these functions, especially the hazard function, play an important role in the process of modeling survival data.

Throughout this thesis T will represent the random variable of the survival times of an individual under investigation. More specifically, T will be the time from a well defined starting point to the moment in which the event of interest occurs. Therefore, the distribution of T will be assumed to be continuous with support

on $[0, \infty)$. Furthermore, the probability density function (pdf) will be denoted by f and the cumulative distribution function (cdf) by F . At time t ,

$$F(t) = P(T \leq t) = \int_0^t f(t)dt.$$

In survival analysis, however, it is often more interesting to study the probability that an individual survived beyond time t which is given by the survival function

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - P(T \leq t).$$

2.1.1 The hazard function

The density function, the cumulative distribution function and the survival function characterize the distribution of T . Another way of representing the distribution of the survival times is by means of the hazard function

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (2.1)$$

Although the hazard function plays a very important role in the analysis and modeling of survival data its interpretation is not simple. There is a lot of confusion among non-statisticians about the meaning of the hazard function (see Spruance et al. 2004). A helpful analogy is based on the concept of instant velocity. To explain this analogy, the hazard function can be written as

$$\lambda(t) = \frac{1}{nP(T \geq t)} \lim_{\Delta t \rightarrow 0} \frac{nP(t \leq T < t + \Delta t)}{\Delta t}$$

where n is the number of individuals in the sample. Here, $nP(T \geq t)$ is the expected number of individuals at risk at time t and $nP(t \leq T < t + \Delta t)$ is the expected number of deaths between t and $t + \Delta t$. If there was a functional continuous relationship $N(t)$ between time and the number of deaths up to time t then the hazard function could be written as

$$\lambda(t) = \frac{1}{n - N(t)} \lim_{\Delta t \rightarrow 0} \frac{N(t + \Delta t) - N(t)}{\Delta t} = \frac{N'(t)}{n - N(t)}$$

Under this interpretation, $N'(t)$ represents the instantaneous rate of death. Therefore, the hazard function is the instantaneous rate of deaths at time t relative to the number of people at risk at time t . Given this interpretation, the hazard function is sometimes called mortality rate, conditional failure rate or instantaneous potential for death (Kleinbaum & Klein, 2011, pg. 10).

There is an easier way to write the hazard function in (2.1):

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\
&= \frac{1}{S(t)} \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} = \\
&= \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}.
\end{aligned} \tag{2.2}$$

Yet another way of expressing the hazard function, using (2.2), is:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} \implies F'(t) + \lambda(t)F(t) - \lambda(t) = 0$$

and one can obtain $F(t)$ as a solution of the differential equation with initial condition $F(0) = 1$. So it is clear that $\lambda(t)$ also completely characterizes the distribution of T .

Sometimes it is useful to work with the cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u) du$$

which can be interpreted as the expected number of events of $N(t)$ at time t , where $N(t)$ is the counting process defined by the distribution of T (Fleming & Harrington, 1991). It is immediate that $\lambda(t) = \Lambda'(t)$ and using (2.2) one can write the cumulative hazard as

$$\Lambda(t) = \int_0^t \lambda(u) du = \int_0^t \frac{f(u)}{S(u)} = \int_0^t \frac{-S'(u)}{S(u)} = [-\log S(u)]_0^t = -\log S(t). \tag{2.3}$$

Conversely, the survival function $S(t)$ can be expressed in terms of the cumulative hazard function $\Lambda(t)$ as

$$S(t) = e^{-\Lambda(t)}. \tag{2.4}$$

Based on these relationships the cumulative hazard function also characterizes the distribution of T .

2.1.2 Mean residual life function

The mean residual life (MRL) function at time $t \geq 0$ is defined as:

$$\text{MRL}(t) = E(T - t | T > t).$$

This can be interpreted as follows: given that an individual survived up to time t the MRL function is the expected value of the conditional distribution that

represents all the individuals who survived up to time t . Whereas the hazard function focuses on what happens in the very short term and gives an idea of the risk of instant death given the survival up to time t , the MRL function focuses on the long term survival experience of the individual and gives the mean value of the conditional distribution.

It is possible to write the MRL function in terms of the survival function by considering the conditional distribution of $T - t|T > t$. The cumulative distribution function of this conditional distribution is:

$$F_{T-t|T>t}(u) = P(T - t \leq u|T > t) = \frac{P(t < T \leq u + t)}{P(T > t)} = \frac{F(u + t) - F(t)}{S(t)}$$

where $u > 0$ ($F_{T-t|T>t}(u) = 0$ if $u \leq 0$). Thus, the survival function of the conditional distribution is:

$$S_{T-t|T>t}(u) = 1 - F_{T-t|T>t}(u) = 1 - \frac{F(u + t) - F(t)}{S(t)} = \frac{S(u + t)}{S(t)}$$

This conditional distribution has support on $(0, \infty)$ and therefore, the expected value¹ is:

$$\text{MRL}(t) = E(T - t|T > t) = \int_0^\infty \frac{S(u + t)}{S(t)} du = \frac{\int_t^\infty S(u) du}{S(t)} \quad (2.5)$$

The mean residual life function can then be related to the hazard function by simply taking the derivative of expression (2.5), i.e.

$$\text{MRL}'(t) = \frac{-S(t)^2 + (\int_t^\infty S(u) du) f(t)}{S(t)^2} = -1 + \lambda(t)\text{MRL}(t).$$

Therefore,

$$\lambda(t) = \frac{1 + \text{MRL}'(t)}{\text{MRL}(t)}, \quad (2.6)$$

so it is now apparent that the mean residual life function characterizes the distribution of T using the latter expression. For a review of all the properties of the MRL see Guess & Proschan (1988).

One of the reasons that makes the use of the hazard and mean residual life functions interesting in survival analysis is that they describe very nicely the survival

¹In general, if a random variable X is positive ($X > 0$), $E(X) = \int_0^\infty S(x) dx$ where $S(x)$ is the survival function of X . This is easy to prove by:

$$\int_0^\infty S(x) dx = \int_0^\infty \int_x^\infty f(u) du dx = \int_0^\infty f(u) \int_0^u dx du = \int_0^\infty u f(u) du = E(X)$$

where $f(x)$ is the density function of X .

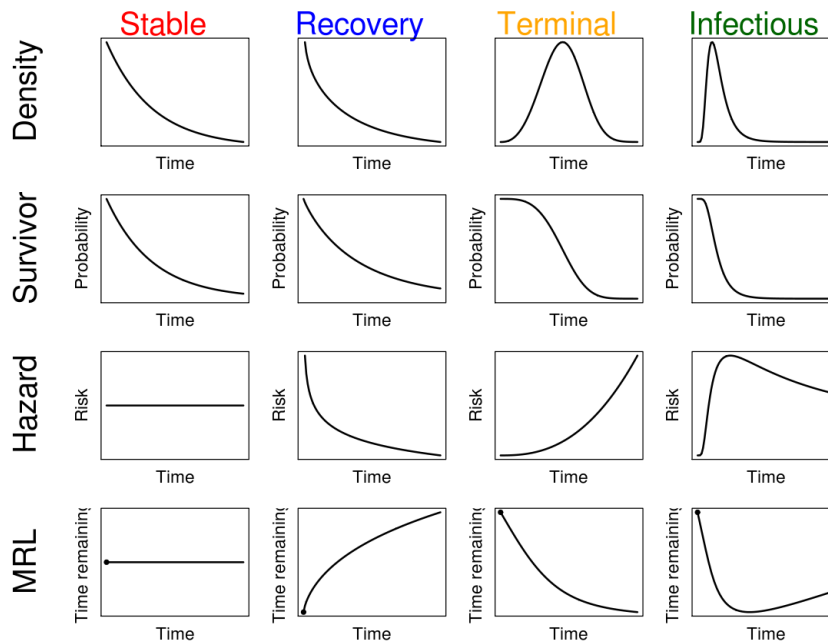


Figure 2.1: Examples of density, survival, hazard and MRL functions for different survival experiences.

experience of the individuals under investigation (see Figure 2.1). For instance, patients who have a stable condition in relation to a particular disease will display a constant MRL and hazard functions for a limited period of time. For this type of patients the exponential distribution adequately describes their survival experience since the risk of experiencing the event of interest (it could be death or recurrence of the disease) does not change over time. This is referred to as “Stable” patients in Figure 2.1. Patients who have undergone a surgical intervention (a serious operation or transplant) will have higher risk of death in the few hours following the operation and, if no major complications take place, the risk will decrease over time (decreasing hazard/increasing MRL). Terminal patients with a serious illness such as cancer will display increasing hazard (decreasing MRL) over time as the condition of the patient deteriorates due to the disease. Finally, patients with infectious diseases such as TB have a period of increasing hazard (decreasing MRL) until they receive treatment and, after that, the hazard decreases (MRL increases) over time.

2.1.3 Types of censoring

As stated in the introduction, survival data are very often incomplete due to censoring. An observation is said to be censored if the time recorded does not correspond with the time at which the event of interest occurs. This could happen in one of the following three situations: when the event occurs at an unknown time before the observed time (**left censoring**); when the event occurs at an unknown time after the observed time (**right censoring**); or when the event occurs at an unknown time between two observed times (**interval censoring**). By far the most common form of censoring assumption is right censoring, and all the methods described in this thesis are based on this assumption. In general, the different types of right censoring can be broadly classified as:

Type I The length of an experiment is set up in advance. Observations for which the event of interest did not occur before the end of the experiment are considered right censored and the censored times correspond to the length of the experiment.

Type II The number of observations is set up in advance. After a predetermined number of events occurs, the experiment stops, and observations for which the event did not occur are treated as right censored.

Random Each observation has associated a censoring time that is independent of the failure time (the time when the event of interest occurs). If the censoring time occurs before the failure time the observation is censored (right censored), if not the true event time is observed.

The first two types of censoring occur more generally in engineering, whereas the random censoring arises typically in the biomedical sciences. Note that the difference between type I and type II censoring is that, in the latter case, the number of events are fixed in advance, whereas in type I censoring the number of events is a random quantity. For instance, imagine that the lifetime of a set of n machines is going to be tested. In type I censoring, the machines are followed up for a predetermined period of time, and after that period, the machines that are still working are right censored (random number of failures and fixed length of the experiment). Now imagine that the machines are going to be followed up until $k < n$ machines fail. In that case, after the k th machine fails the experiment stops and the remaining $n - k$ machines are right censored (fixed number of failures and random length of the experiment). In both of these situations, only observations that exceed certain event time are censored. In random censoring, however, censored observations might be found at any time point. In this work only random right censoring will be considered because this type of censoring is the most common one in real data applications related to medicine. Some of the reasons for which data are collected with this type of censoring are:

- *Loss to follow up.* The individual under investigation disappears without any explanation. The clinicians never see him/her again.
- *Drop out.* The study has to finish prematurely for a particular individual due to some adverse effects of the treatment. Another cause is that the patient refuses to continue the treatment for some reason.
- *Competing risks.* The event could not be observed due to the occurrence of a competing event (for instance, the death of the patient due to other causes).
- *Termination of the study.* The individual did not experience the event of interest before the study ends.

The methods developed for analyzing survival data under the assumption of random right censoring generally adopt the following theoretical setting. Let T and C be the failure and censoring times respectively and assume that they are independent. Although interested in the distribution of T , we can only observe $X = \min(T, C)$ and

$$\delta = 1_{\{T \leq C\}} = \begin{cases} 1 & \text{if } T \leq C, \\ 0 & \text{if } T > C \end{cases}$$

which is called the censoring indicator that tells us whether the observed X is censored or not. Both X and δ are random variables, therefore, for a particular individual i the information that can be observed is the pair (X_i, δ_i) . Under this theoretical setting, a typical random sample of n individuals looks like $\{(x_1, \delta_1), (x_2, \delta_2), \dots, (x_n, \delta_n)\}$, where δ_i is 1 if the event of interest was observed at time x_i for individual i and it is 0 if x_i is the time in which individual i was last seen. The main problem in survival analysis is how to extract relevant information about the distribution of T using only the information obtained in (X, δ) . To analyze this problem it is convenient to write down the distribution function of X

$$\begin{aligned} F_X(t) &= P(X \leq t) = P(\min(T, C) \leq t) \\ &= 1 - P(\min(T, C) > t) \\ &= 1 - P(T > t \cap C > t) \\ &= 1 - P(T > t)P(C > t) \\ &= 1 - [(1 - F_T(t))(1 - F_C(t))]. \end{aligned}$$

Here F_T and F_C are the cumulative distribution functions of T and C respectively. One could be tempted to discard all the observations that are censored and to analyze only the true event times, even with the evident loss of statistical power.

However, this is not appropriate because the distribution function of the observed true events does not correspond with the distribution function of T . The reason is that

$$\begin{aligned}
P(X \leq t \cap \delta = 1) &= P(T \leq t \cap T \leq C) \\
&= \int_0^t \int_u^\infty f_C(c) f_T(u) dc du \\
&= \int_0^t f_T(u) S_C(u) du
\end{aligned} \tag{2.7}$$

is equal to $F_T(t)$ only if $S_C(u) = 1$ for all $u \in (0, t)$, i.e. if there is no censoring (f_T and f_C are the pdfs of T and C respectively and due to the independence of T and C , the joint density function $f_{T,C}(t, c) = f_T(t) f_C(c)$).

2.1.4 The likelihood of observed survival data

The likelihood function of the observed pair (X, δ) under this theoretical setting is

$$L(\mathbf{x}, \boldsymbol{\delta}; \boldsymbol{\theta}) = \prod_{i=1}^n f_{X,\delta}(x_i, \delta_i; \boldsymbol{\theta})$$

where $f_{X,\delta}$ is the joint density function of (X, δ) and $\boldsymbol{\theta}$ is the vector of the parameters. If $\delta = 1$ the cumulative distribution function of $(X, 1)$ is

$$F_{X,1}(t) = P(X \leq t \cap \delta = 1) = \int_0^t f_T(u) S_C(u) du$$

as shown in 2.7. On the other hand, if $\delta = 0$ the cumulative distribution function of $(X, 0)$ is

$$\begin{aligned}
F_{X,0}(t) &= P(X \leq t \cap \delta = 0) \\
&= P(C \leq t \cap T > C) \\
&= \int_0^t \int_c^\infty f_C(c) f_T(u) du dc \\
&= \int_0^t f_C(c) S_T(c) dc.
\end{aligned} \tag{2.8}$$

Thus, the density function of $(X, 1)$ is $f_{X,1}(t) = F'_{X,1}(t) = f_T(t) S_C(t)$ and the density function of $(X, 0)$ is $f_{X,0}(t) = F'_{X,0}(t, 0) = f_C(t) S_T(t)$. Therefore,

$$\begin{aligned}
f_{X,\delta}(t) &= \delta f_T(t) S_C(t) + (1 - \delta) f_C(t) S_T(t) \quad (\text{linear form}) \\
&= (f_T(t) S_C(t))^\delta (f_C(t) S_T(t))^{1-\delta} \quad (\text{multiplicative form}) \\
&= f_T(t)^\delta S_T(t)^{1-\delta} f_C(t)^{1-\delta} S_C(t)^\delta.
\end{aligned} \tag{2.9}$$

Considering all these results, the likelihood can be written now as

$$L(\mathbf{x}, \boldsymbol{\delta}; \boldsymbol{\theta}) = \prod_{i=1}^n f_T(x_i)^{\delta_i} S_T(x_i)^{1-\delta_i} f_C(x_i)^{1-\delta_i} S_C(x_i)^{\delta_i}. \quad (2.10)$$

As it has been shown, to get the latter expression the assumption that T and C are independent is fundamental. A further simplification can be made by assuming that the distribution of C does not include any of the parameters of the distribution of T . If this is the case, then the censoring is said to be non-informative. The reason why this simplifies the likelihood function is that the likelihood is defined up to a multiplicative constant, and if one is only interested in the estimation of the parameters of T (as in most cases) the terms $f_C(x_i)^{1-\delta_i} S_C(x_i)^{\delta_i}$ in (2.10) are absorbed into the constant of proportionality. Therefore, under **non-informative right censoring**, using (2.2) and (2.4), the likelihood function is

$$L(\mathbf{x}, \boldsymbol{\delta}; \boldsymbol{\theta}) = \prod_{i=1}^n f_T(x_i)^{\delta_i} S_T(x_i)^{1-\delta_i} = \prod_{i=1}^n \lambda_T(x_i)^{\delta_i} e^{-\Lambda_T(x_i)} \quad (2.11)$$

where $\boldsymbol{\theta}$ is now the vector of parameters of the distribution of T .

Both parametric and non-parametric methods can be used for the estimation of the survival function. The parametric approach assumes that the data have been generated from a parametric family of distributions and uses (2.11) in order to generate maximum likelihood estimates of the parameters of the underlying distribution. Based on the estimates of the parameters, survival and hazard functions can be easily obtained. On the other hand, non-parametric methods do not make any assumptions about the distribution of the survival times. Because of that, in general, they need more data to get reasonable estimates of the survival function although they are more robust to the selection of the wrong model. However, parametric and non-parametric methods might fail to give adequate estimates of the MRL function as will be shown in Chapter 5.

2.2 The Kaplan-Meier estimator of the survival function

The most common non-parametric method for the estimation of the survival function was developed by Kaplan & Meier (1958) and is based on the product limit formula. To explain this concept, suppose that $\{t_1, t_2, \dots, t_s\}$ is a set of s ordered

times smaller than t and that T is a positive random variable. Then

$$\begin{aligned}
P(T > t) &= P(T > t_1, T > t_2, \dots, T > t_s, T > t) = \\
&= \prod_{i=1}^{s+1} \left[P(T > t_i | T > t_{i-1}) \right] = \\
&= \prod_{i=1}^{s+1} \left[1 - P(T \leq t_i | T > t_{i-1}) \right]
\end{aligned} \tag{2.12}$$

where $t_0 = 0$ and $t_{s+1} = t$. To apply this formula to the problem of estimating the survival function suppose $\{(x_1, \delta_1), \dots, (x_n, \delta_n)\}$ is a random sample of T , C and $\{(x_{(1)}, \delta_{(1)}), \dots, (x_{(n)}, \delta_{(n)})\}$ the corresponding time ordered observations (the order is established for the values of (x_1, \dots, x_n) so $\delta_{(i)} = \delta_j$ iff $x_{(i)} = x_j$). If no ties are present (all the x_i s are different), then the Kaplan-Meier estimate of the survival function is

$$S_{KM}(t) = \prod_{i: x_{(i)} < t} \left[1 - \frac{\delta_{(i)}}{\sum_{j: x_j \geq x_{(i)}} 1} \right]. \tag{2.13}$$

If some observations are tied, let $(x_{(1)}, \dots, x_{(r)})$ the ordered distinct observed times corresponding only to uncensored observations, then, the Kaplan-Meier estimate is

$$S_{KM}(t) = \prod_{i: x_{(i)} < t} \left[1 - \frac{\sum_{j: x_{(i)} \leq x_j < x_{(i+1)}} \delta_j}{\sum_{j: x_j \geq x_{(i)}} 1} \right] = \prod_{i: x_{(i)} < t} \left[1 - \frac{d_i}{n_i} \right] \tag{2.14}$$

where d_i is the number of deaths between $x_{(i)}$ and $x_{(i+1)}$, and n_i the number of individuals at risk at time $x_{(i)}$. Therefore, the Kaplan-Meier estimate of the survival function is a non-increasing stepwise function changing only in the uncensored observed times (true failures). If the largest observed time $x_{(\max)}$ is censored then $S_{KM}(t) \neq 0$ for all $t > x_{(\max)}$ and in this situation it is not possible to estimate the mean $E[T]$ (as will be shown in Chapter 5).

Breslow & Crowley (1974) established the asymptotic normality of the Kaplan-Meier estimate of the survival function while Peterson (1977) proved the strong consistency. Confidence intervals can be obtained using Greenwood's formula (Greenwood 1926) which gives an estimate of the variance of $S_{KM}(t)$,

$$\widehat{\text{Var}}(S_{KM}(t)) = S_{KM}(t)^2 \sum_{i: x_{(i)} < t} \left[\frac{d_i}{n_i(n_i - d_i)} \right].$$

Figure 2.2 shows an example of the Kaplan-Meier estimate of the survival function with the corresponding 95% pointwise confidence intervals for the breast cancer dataset. In (a) the event of interest is death and the outcome is the time

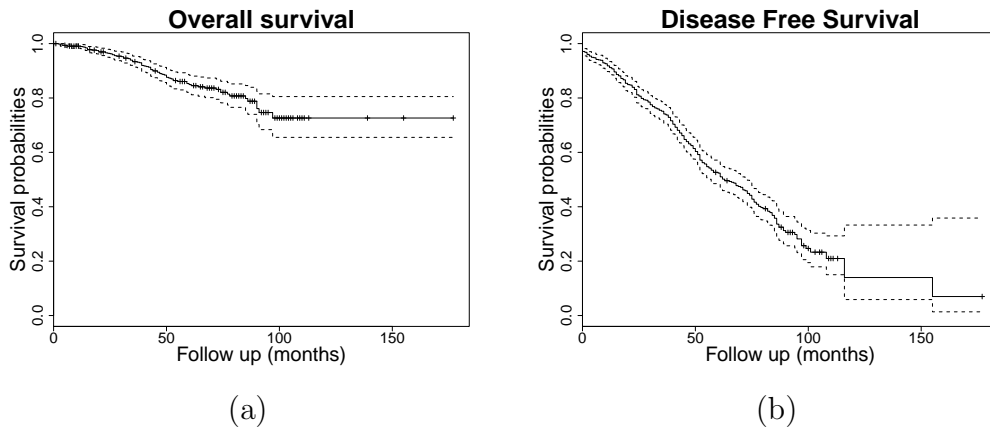


Figure 2.2: Two different examples of the Kaplan-Meier estimate of the survival function. In (a) the probabilities of survival for women with breast cancer. In (b) the probability of disease free survival.

from diagnosis until the event occurs. In (b) the event of interest is the recurrence of the disease and the outcome is the time until recurrence. The median time to recurrence is 63 months meaning that an estimated 50% of the women will have recurrence after 63 months. Also, there is a high probability that more than 70% of the women will get recurrence after 100 months. This plot suggests that there is a high probability of recurrence of the disease for the women from which this sample was taken. However, this fact does not translate into a higher incidence of mortality as Figure 2.2 (a) shows. In fact, more than 70% of the women were alive at the end of the study.

Another example is presented in Figure 2.3 where the coronary dataset was used to obtain Kaplan-Meier estimates of the survival function. As one can see in the plot the survival outcome (death from any cause) of the patients in this sample is positive in the sense that more than 70% of the patients were still alive after 1500 days (approximately 4 years) and more than 60% were alive after the 5 year period.

2.3 The Logrank test

Different methods have been developed in the literature to compare the distribution of the survival times in two or more different groups. In the context of this thesis, the comparison of two different survival functions is very relevant since in a survival tree, one of the ways of choosing the split at a node is by selecting the cutpoint that makes the survival distribution of the left and right nodes as different as possible. To measure these differences the values of the logrank test

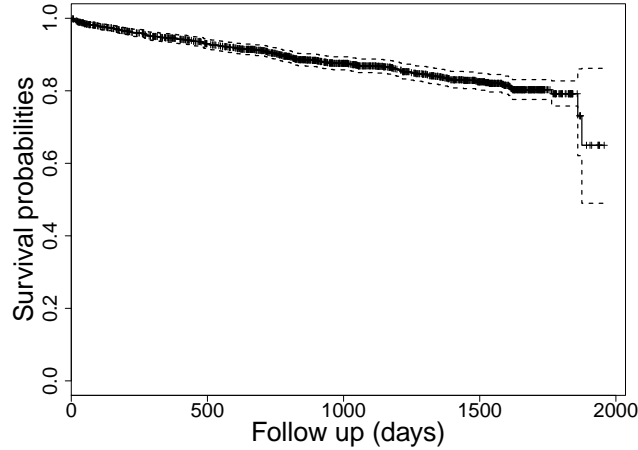


Figure 2.3: Kaplan-Meier estimate of the survival function applied to the coronary dataset.

statistic can be used. Thus, in this section an overview of some of the methods to compare the survival experience in two groups is given. The Coronary and the breast cancer datasets will be used to illustrate some of the results.

Because of the special characteristics of survival data, tests based on ranks are particularly appropriate. Gehan (1965) proposed an extension of the Wilcoxon statistic (Wilcoxon, 1945) to the case of censored data. To illustrate the method, let $\{(x_1, \delta_1), \dots, (x_n, \delta_n)\}$ and $\{(y_1, \epsilon_1), \dots, (y_n, \epsilon_n)\}$ be two random samples from two different populations A and B . Let $F_A(t)$ and $F_B(t)$ the corresponding distribution functions of the true survival times. Let $x'_i = x_i$ if $\delta_i = 0$ and $y'_j = y_j$ if $\epsilon_j = 0$. For a given pair (x_i, y_j) he defined the following score function

$$U_{(i,j)} = \begin{cases} 1 & \text{if } x_i < y_j \text{ or } x_i \leq y'_j, \\ 0 & \text{if } x_i = y_j \text{ or } x'_i = y'_j \text{ or } x'_i < y_j \text{ or } y'_j < x_i, \\ -1 & \text{if } x_i > y_j \text{ or } x'_i \geq y_j. \end{cases}$$

Based on these values he defined the test statistic $W = \sum_{(i,j)} U_{(i,j)}$ summing over all possible pairs (i, j) . He showed how this test is related to the Wilcoxon (1945) test statistic and proved that under the null hypothesis of no difference $H_0: F_A(t) = F_B(t)$, the test statistic is asymptotically normal.

Efron (1967) considered a modified version of the Gehan statistic in which the score function (which he called $Q(x_i, y_j)$) instead of giving a value 0 when $x_i = y_j$ or $\delta_i = 0$ or $\epsilon_j = 0$ gives values that depend on the Kaplan-Meier estimates of the survival functions of A and B (S_{KM}^A and S_{KM}^B). The actual values are shown in Table 2.1. The test statistic is defined as $\widehat{W} = \frac{1}{nm} \sum_{i,j} Q(x_i, y_j)$. For instance, if

| (δ_i, ϵ_j) | $x_i \geq y_j$ | $x_i < y_j$ |
|--------------------------|--|---|
| (1,1) | 1 | 0 |
| (0,1) | 1 | $\frac{S_{KM}^A(y_j)}{S_{KM}^A(x_i)}$ |
| (1,0) | $1 - \frac{S_{KM}^B(x_i)}{S_{KM}^B(y_j)}$ | 0 |
| (0,0) | $1 - \frac{S_{KM}^B(x_i)}{S_{KM}^B(y_j)} - \int_{x_i}^{\infty} \frac{S_{KM}^A(s)dS_{KM}^B(s)}{S_{KM}^A(x_i)S_{KM}^B(y_j)}$ | $-\int_{x_i}^{\infty} \frac{S_{KM}^A(s)dS_{KM}^B(s)}{S_{KM}^A(x_i)S_{KM}^B(y_j)}$ |

Table 2.1: Values of $Q(x_i, y_j)$ for the modified version of the Gehan statistic proposed by Efron (1967).

$y_j \leq x_i$ and $\epsilon = 0$ instead of scoring 0 as in the Gehan score function, $Q(x_i, y_j)$ will depend on how different the probabilities of survival are for x_i and y_j under the survival function defined by B .

Almost in parallel to the development of this family of statistics Mantel (1966) proposed the use of the Mantel-Haenszel procedure (Mantel & Haenszel, 1959), which combines the results of a series of 2×2 tables, to the particular case of comparing two different survival distributions. Let $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$ the combined ordered observed survival times of two groups A and B . For each observed time $x_{(i)}$ the following 2×2 table can be considered

| $x_{(i)}$ | Deaths | | At risk |
|-----------|-----------|---------------------|-----------|
| A | M_{0_i} | $N_{0_i} - M_{0_i}$ | N_{0_i} |
| B | M_{1_i} | $N_{1_i} - M_{1_i}$ | N_{1_i} |
| | M_i | $R_i - M_i$ | R_i |

where, at time $x_{(i)}$, M_{0_i} is the number of observed deaths in group A and N_{0_i} is the number of individual at risk in group A . M_{1_i} and N_{1_i} are the corresponding numbers in group B and M_i and R_i are the total number of deaths and the total individuals at risk at time $x_{(i)}$ respectively. If one considers the marginals of the table fixed, the range of possible values of M_{0_i} varies from $\max(0, M_i + N_{0_i} - R_i)$ to $\min(N_{0_i}, M_i)$ (notice that once the value of M_{0_i} is chosen the rest of the table can be completed). The probability of getting such a value, i.e. M_{0_i} , is given by the hypergeometric distribution

$$P(Y_i = M_{0_i}) = \frac{\binom{N_{0_i}}{M_{0_i}} \binom{N_{1_i}}{M_{1_i}}}{\binom{R_i}{M_i}}$$

where Y_i is the random variable that describes the number of deaths in group A at time $x_{(i)}$ and has support in $\{\max(0, M_i + N_{0_i} - R_i), \dots, \min(N_{0_i}, M_i)\}$. The

expected value is

$$E(Y_i) = \frac{N_{0i}}{R_i} M_i$$

which is the expected number of deaths in the 2×2 table under the hypothesis that there is no relationship between the number of deaths and the group. The variance of Y_i is

$$\text{Var}(Y_i) = M_i \frac{N_{0i}}{R_i} \frac{R_i - N_{0i}}{R_i} \frac{R_i - M_i}{R_i - 1}.$$

The proposed test statistic by Mantel (1966)² was

$$\chi_1^2 = \frac{[\sum_i (Y_i - E(Y_i))]^2}{\sum_i \text{Var}(Y_i)} \quad (2.15)$$

which it was shown by Crowley (1973) to converge in distribution to a chi-square distribution with 1 degree of freedom. This method would later be known as the logrank test. The observed value of the test statistic is

$$\chi_{1\text{obs}}^2 = \frac{\left[\sum_i (M_{0i} - \frac{N_{0i}}{R_i} M_i) \right]^2}{\sum_i M_i \frac{N_{0i}}{R_i} \frac{R_i - N_{0i}}{R_i} \frac{R_i - M_i}{R_i - 1}}$$

Cox (1972) showed that the results of the logrank test can be derived from a proportional hazard model in which the grouping factor is the only predictor in the model (see next section). Tarone & Ware (1977) studied the connections between the logrank test and the Gehan (1965) modified Wilcoxon test and proved that they only differ in the choice of weights, which are functions of the number of individuals at risk at any uncensored time. They showed that the Gehan modified Wilcoxon test is equivalent to the test

$$\chi_1^2 = \frac{[\sum_i w_i (O_i - E_i)]^2}{\sum_i w_i^2 \text{Var}(O_i)} \quad (2.16)$$

where $w_i = R_i$. Under the null hypothesis of no differences between the two survival functions, (2.16) follows an asymptotic chi-squared distribution with 1 degree of freedom. The logrank test statistic described in (2.15) is just the test statistic in (2.16) with weights $w_i = 1$ for all of the observed uncensored times. They proposed a new test statistic based on a different selection of the weights

$$w_i = \sqrt{R_i}.$$

²Note that the results shown here can be extended to a more general setting with more than 2 groups.

They concluded that the logrank test works extremely well under the proportional hazard assumption, but it loses power under deviations from this assumption in which case, the Gehan (1965) modified Wilcoxon test is more appropriate. In a separate study, Prentice & Marek (1979) found a large discrepancy between the p-values obtained by the logrank test and the Gehan modified Wilcoxon test when applied to the same data set. They argued that the reason is that the weights of the Gehan test, which are functions of the individuals at risk, included both, censored and uncensored observations. They proved that other generalizations of the Wilcoxon test by Peto & Peto (1972) and Prentice (1978) were also special cases of a weighted logrank test with weights

$$w_i = \prod_{j: x_{(j)} < x_{(i)}} \frac{N_j}{N_j + 1} \quad (2.17)$$

which are very similar to the Kaplan-Meier estimate of the survival function at any uncensored time i . They pointed out that because the weights of the Gehan's statistic depend on the censored and uncensored observations, the use of this test can lead to anomalous situations. For that reason they concluded that the Peto-Prentice generalized Wilcoxon statistic should always be used with censored data. Harrington & Fleming (1982) proposed a more general class of weights

$$w_i = [S_{KM}(x_{(i)})]^p .$$

where $p \geq 0$ and S_{KM} is the Kaplan-Meier estimate of the survival function. If $p = 0$ the test corresponds to the logrank test and if $p = 1$ the test is essentially equivalent to the Peto-Prentice generalized Wilcoxon statistic with weights given by (2.17).

An example of the use of the logrank test is illustrated in Figure 2.4. In (a) the two estimates of the survival function correspond to the group of patients with cardiovascular disease who had a previous myocardial infarction (red line) and the group of patients who did not have previous myocardial infarction (blue line). To test the hypothesis that the two groups have the the same distribution the logrank test can be used. Table 2.2 shows the values for the logrank test when different weights are considered. As one can see, although the three tests were significant, the Gehan modified Wilcoxon test gives the lowest values of the test statistic with the corresponding loss of power.

In Figure 2.4 (b) another example is given from the breast cancer dataset. The red line corresponds to the estimated survival function of women who tested positive for HER2 whereas the blue line corresponds to the estimated survival function of women who tested negative. The results of the different logrank tests are given in Table 2.3. Again, all three tests were significant with the Gehan modified Wilcoxon test giving the lowest values of the test statistic.

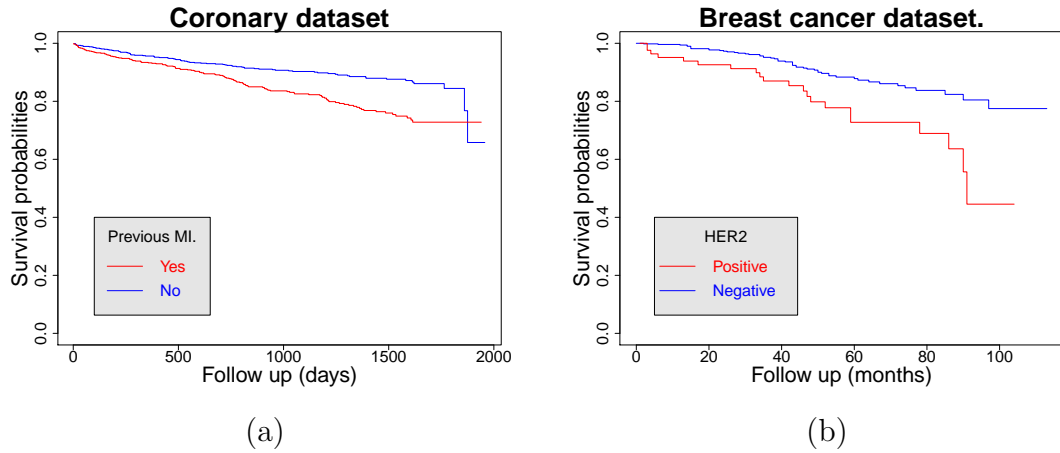


Figure 2.4: Two examples of the Kaplan-Meier estimate of the survival function used to compare two different groups. In both cases the event of interest is the death of the patient. In (a) patients with cardiovascular disease with and without a previous myocardial infarction. In (b) women with breast cancer who tested positive and negative for HER2.

| | Weights | χ_1^2 | p-value |
|---------------|-------------------------|------------|----------|
| Gehan | $w_i = R_i$ | 9.820557 | 0.001726 |
| Logrank | $w_i = 1$ | 12.267544 | 0.000461 |
| Peto-Prentice | $w_i = S_{KM}(x_{(i)})$ | 12.090948 | 0.000507 |

Table 2.2: *Coronary dataset*. Logrank tests comparing the survival outcome of patients with and without previous history of myocardial infarction.

| | Weights | χ_1^2 | p-value |
|---------------|-------------------------|------------|---------|
| Gehan | $w_i = R_i$ | 17.119425 | 3.5e-05 |
| Logrank | $w_i = 1$ | 22.184265 | 2e-06 |
| Peto-Prentice | $w_i = S_{KM}(x_{(i)})$ | 21.50661 | 4e-06 |

Table 2.3: *Breast cancer dataset*. Logrank tests comparing the survival outcome of women who were HER2 positive versus women who were HER2 negative.

2.4 The Cox proportional hazards model

In most of the practical situations in which one has to analyze survival data, the survival time is not the only information that is gathered about the patients. Although the outcome of interest is the time to event, one hopes that by knowing the values of other variables such as age, gender, weight, smoking status, etc, one can obtain a better understanding of the survival experience of the individuals under investigation. These other variables are usually referred to as covariates. Sometimes, the investigator is interested in the effect of one or more covariates on the survival times, treating other variables as *confounders*. For instance, one might want to know the effect of age in the survival times of individuals with cardiovascular disease while adjusting for gender, cholesterol level, blood pressure etc. Other times, the researcher might be interested in knowing which covariates have a significant effect on the survival times treating all the variables as variables of interest. All of the above can be handled in the framework of regression models.

The Cox proportional hazards model (Cox, 1972) is a regression model that allows the inclusion of the information provided by the covariates to analyze the survival times of the individuals. The need for a specific model to analyze survival data, different from those used with other responses, is due to the presence of censoring in the data. To introduce the model, let $\mathbf{Z} = (Z_1, \dots, Z_p)$ be the vector of covariates for any individual in the population. A random sample will be now of the form $\{(x_1, \delta_1, \mathbf{z}_1), \dots, (x_n, \delta_n, \mathbf{z}_n)\}$ where $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})$ is the vector of covariates of individual $i = 1, \dots, n$. The proportional hazards assumption states that

$$\lambda(t|\mathbf{z}_i) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{z}_i}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ is the vector of the parameters of interest and $\lambda_0(t)$ is the baseline hazard. The key aspect of this assumption is the separation of the shape of the hazard function and the multiplicative covariate effects. Cox (1972) proposed a semi-parametric model that does not make any assumption about the nature of the hazard function and it is based on the use of the partial likelihood. To illustrate the method, recall the expression of the likelihood function for non-informative right censoring in (2.11). Under the proportional hazards assumption, the likelihood can be written in terms of the baseline hazard function as

$$\prod_{i=1}^n [\lambda_0(x_i) \exp(\boldsymbol{\beta}'\mathbf{z}_i)]^{\delta_i} e^{-\Lambda_0(x_i) \exp(\boldsymbol{\beta}'\mathbf{z}_i)}. \quad (2.18)$$

Here the baseline hazard is still unknown at all the points x_i . If the cumulative baseline hazard function is estimated by (for simplicity it is assumed that there

are no ties and the times are ordered)

$$\Lambda_0(x_i) = \sum_{k=1}^i \lambda_0(x_k)$$

and this estimate is plugged in (2.18) one obtains

$$\prod_{i=1}^n [\lambda_0(x_i) \exp(\boldsymbol{\beta}' \mathbf{z}_i)]^{\delta_i} \exp\left(-\sum_{k=1}^i \lambda_0(x_k) \exp(\boldsymbol{\beta}' \mathbf{z}_k)\right). \quad (2.19)$$

The term $\prod_{i=1}^n \exp\left(-\sum_{k=1}^i \lambda_0(x_k) \exp(\boldsymbol{\beta}' \mathbf{z}_k)\right)$ in (2.19) can be written as

$$\begin{aligned} & e^{-\lambda_0(x_1)e^{(\boldsymbol{\beta}' \mathbf{z}_1)}} \\ & \cdot e^{-\lambda_0(x_1)e^{(\boldsymbol{\beta}' \mathbf{z}_2)}} \cdot e^{-\lambda_0(x_2)e^{(\boldsymbol{\beta}' \mathbf{z}_2)}} \\ & \quad \vdots \quad \quad \quad \vdots \\ & \cdot e^{-\lambda_0(x_1)e^{(\boldsymbol{\beta}' \mathbf{z}_n)}} \cdot e^{-\lambda_0(x_2)e^{(\boldsymbol{\beta}' \mathbf{z}_n)}} \quad \dots \quad e^{-\lambda_0(x_n)e^{(\boldsymbol{\beta}' \mathbf{z}_n)}} = \\ & = \prod_{i=1}^n e^{-\lambda_0(x_i) \sum_{k=i}^n e^{(\boldsymbol{\beta}' \mathbf{z}_k)}} \\ & = \prod_{i=1}^n \exp\left(-\lambda_0(x_i) \sum_{k=i}^n \exp(\boldsymbol{\beta}' \mathbf{z}_k)\right) \end{aligned}$$

and, therefore, equation (2.19) is now

$$\prod_{i=1}^n [\lambda_0(x_i) \exp(\boldsymbol{\beta}' \mathbf{z}_i)]^{\delta_i} \exp\left(-\lambda_0(x_i) \sum_{k=i}^n \exp(\boldsymbol{\beta}' \mathbf{z}_k)\right). \quad (2.20)$$

This function can be considered as a modified likelihood function that can be used to estimate $\boldsymbol{\beta}$. Essentially the modified likelihood can be profiled over the unknown values $\lambda_0(x_i)$ to be able to estimate $\boldsymbol{\beta}$. The maximum likelihood estimate of $\lambda_0(x_i)$ as a function of $\boldsymbol{\beta}$ is

$$\widehat{\lambda_0(x_i)} = \frac{\delta_i}{\sum_{k=i}^n \exp(\boldsymbol{\beta}' \mathbf{z}_k)} \quad (2.21)$$

for all $i = 1, \dots, n$. Again these estimates can be plugged in (2.20) resulting in the expression

$$\prod_{i=1}^n \left[\frac{\exp(\boldsymbol{\beta}' \mathbf{z}_i)}{\sum_{k=1}^n \exp(\boldsymbol{\beta}' \mathbf{z}_k)} \right]^{\delta_i} \delta_i^{\delta_i} e^{-\delta_i} \propto \prod_{i=1}^n \left[\frac{\exp(\boldsymbol{\beta}' \mathbf{z}_i)}{\sum_{k=1}^n \exp(\boldsymbol{\beta}' \mathbf{z}_k)} \right]^{\delta_i} \quad (2.22)$$

which is known as the *partial likelihood* for the estimation of the parameters $(\beta_1, \dots, \beta_n)$. This semi-parametric model is the most often applied one in survival analysis.

| | $\hat{\beta}_i$ | $\exp(\hat{\beta}_i)$ (95%CI) |
|------------------|-----------------|-------------------------------|
| Age | 0.06 | 1.06 (1.04,1.08) |
| PreviousMI (yes) | 0.57 | 1.77 (1.28,2.46) |
| Chol | 0.09 | 1.10 (0.95,1.27) |
| ACE (yes) | 0.64 | 1.89 (1.36,2.63) |

Table 2.4: *Coronary data*: Cox proportional hazard model with Age, PreviousMI, Chol and ACE as covariates. Event of interest: death from any cause.

| | $\hat{\beta}_i$ | $\exp(\hat{\beta}_i)$ (95%CI) |
|------------------|-----------------|-------------------------------|
| Age | 0.05 | 1.06 (1.04,1.08) |
| PreviousMI (yes) | 0.63 | 1.88 (1.42,2.48) |
| ACE (yes) | 0.50 | 1.66 (1.25,2.20) |

Table 2.5: *Coronary data*: Cox proportional hazard model with Age, PreviousMI and ACE as covariates. Event of interest: death from any cause.

2.5 Applications of the Cox regression model

To analyze the Coronary dataset, variable selection procedures³ (based on the Cox proportional hazard model) were applied for the selection of the relevant predictors of the survival outcome (death from any cause). The algorithm identified Age, PreviousMI, Chol and ACE as the relevant covariates to be included in the final model. Table 2.4 shows the estimated coefficients and the corresponding confidence intervals when the Cox proportional hazard model was fitted. ACE and PreviousMI were found to have a significant effect (CI does not contain 1) on the survival outcome with estimated adjusted hazard ratios of 1.89 and 1.77 respectively. The model also identified age as having a significant effect with an increase in risk as age increases. However, cholesterol level was found not to be significant given the presence of the other variables. This covariate was not significant in the univariate analysis either. Table 2.5 shows the model with Age, PreviousMI and ACE as the only covariates. As one can see all the three predictors are significant. The covariate with the largest effect is PreviousMI with an estimated adjusted hazard ratio of 1.88. Patients who had a previous myocardial infarction have 1.88 times more risk (hazard) of dying than patients who had no previous myocardial infarction, while adjusting for the other covariates. Patients who are under ACE medication have an increased hazard of 66% compared to patients who are not under this type of medication for any combination of levels of the other predictors. Age has also a significant effect while keeping the other covariates constant with an increase of 6% in the hazard of death for each unit

³Backwards elimination variable selection in SPSS.

| | $\hat{\beta}_i$ | $\exp(\hat{\beta}_i)$ (95%CI) |
|-----------------|-----------------|-------------------------------|
| Bcl2 (positive) | -0.37 | 0.69 (0.55,0.88) |

Table 2.6: *Breast cancer dataset*: Cox proportional hazard model with Bcl2 as the only covariate. Event of interest: recurrence of the disease.

| | $\hat{\beta}_i$ | $\exp(\hat{\beta}_i)$ (95%CI) |
|-----------------|-----------------|-------------------------------|
| HER2 (positive) | 0.58 | 1.78 (1.34,2.36) |

Table 2.7: *Breast cancer dataset*: Cox proportional hazard model with HER2 as the only covariate. Event of interest: recurrence of the disease.

increase in Age. Two-way interaction effects between the predictors in the Cox regression model were also examined and it was found that PreviousMI and Age had a significant effect when combined together in a multiplicative term in the equation. The adequacy of the proportional hazard assumption was investigated using the test based on weighted residuals proposed by Grambsch & Therneau (1994). It was found that the assumption was valid for all the covariates in the model.

Another example of a Cox proportional hazard model is given in Table 2.6 where the breast cancer dataset was used to established the set of biomarkers that had a significant effect on the survival outcome (recurrence of the disease). Variable selection procedures were run for this dataset and it was found that the final model only included Bcl2 as a significant covariate. The model suggests that Bcl2 positive patients have an approximately 31% reduction in the hazard of death when compared to Bcl2 positive patients. When assessing the adequacy of the proportional hazard assumption it was found that Bcl2 did not satisfy this assumption. Other biomarkers that did not satisfy the proportional hazard assumption (when doing univariate analyses) were ER and PR. From those that satisfied the proportional hazards assumption, only HER2, ki67, p53, CK14 and tMcm2 were found to be significant in univariate analysis with HER2 having the highest value of the likelihood ratio test. Table 2.7 shows the model with HER2. This model shows that women who test positive for HER2 have an estimating hazard ratio of 1.78 suggesting an increase of 78% in the hazard of death when compared to women who test negative.

Variable selection procedures were also run for this dataset when the event of interest was the death of the patient (overall survival). In this case, ER, Bcl2 and Cdc7 were the selected predictors for the final model. The estimates of the coefficients and the corresponding confidence intervals are given in Table 2.8. Both biomarkers, Bcl2 and ER, have a very similar effect and both were found to be significant. However, Cdc7 was not significant at a significance level of

| | $\hat{\beta}_i$ | $\exp(\hat{\beta}_i)$ (95%CI) |
|-----------------|-----------------|-------------------------------|
| Bcl2 (positive) | -1.05 | 0.35 (0.16,0.78) |
| Cdc7 | 0.03 | 1.03 (0.99,1.07) |
| ER (positive) | -1.02 | 0.36 (0.17,0.77) |

Table 2.8: *Breast cancer dataset*: Cox proportional hazard model with Bcl2, Cdc7 and ER as covariates. Event of interest: death of the patient (overall survival).

| | $\hat{\beta}_i$ | $\exp(\hat{\beta}_i)$ (95%CI) |
|-----------------|-----------------|-------------------------------|
| Bcl2 (positive) | -0.67 | 0.51 (0.29,0.90) |
| ER (positive) | -1.22 | 0.30 (0.16,0.53) |

Table 2.9: *Breast cancer dataset*: Cox proportional hazard model with Bcl2 and ER as covariates. Event of interest: death of the patient (overall survival).

0.05. In Table 2.9 the model obtained after considering Bcl2 and ER as the only predictors is given. In this final model, ER seems to have a higher effect when compared to Bcl2. Individuals who test ER positive have a better survival outcome with a reduction in the hazard of death of approximately 70% when compared to patients who tested negative while adjusting for Bcl2. On the other hand, women who test Bcl2 positive have a reduction of approximately 50% in the hazard of death compared to women who test negative while adjusting for ER. Two way interactions between the predictors included in the model were not significant and the proportional hazard assumption was found to be valid for both predictors.

2.6 Chapter conclusion

Different topics related to the analysis of survival data have been presented in this Chapter. The MRL function, the non-parametric estimate of the survival function based on the product-limit formula proposed by Kaplan & Meier (1958) and the use of the logrank statistic to compare the survival experience of two groups of individuals, are important topics in the context of this thesis. The MRL function will be used and studied again in Chapter 5 where the problem of estimating such a function will be examined in depth. The Kaplan-Meier estimate of the survival function is the most common method used for the representation of the distribution of survival times and it will be used many times in the following Chapters. An overview of different approaches to compare the distribution of survival times between two different groups was also given in this Chapter. Some of these methods will be applied in Chapters 3 and 4. In the last part of this Chapter, an introduction to the Cox proportional hazard model was given which

is the standard approach when modeling survival data. Based on this type of modeling it was found that Age, PreviousMI and ACE were significant predictors of the death from any cause of patients with cardiovascular disease included in the coronary dataset. Moreover, biomarker HER2 was found to have a significant effect in the time to recurrence for women with breast cancer included in the breast cancer dataset. When studying the overall survival of the same cohort of women, ER, Bcl2 and CDC7 had a significant effect in the survival outcome.

In the next Chapter, different types of models that can be used as an alternative to the Cox proportional hazard model are examined. These methods are known as tree based methods and can be used to model many different types of outcomes including continuous, categorical and survival responses. The advantages of this type of models are numerous, and to mention just a few of them, one could say that they do not rely in any assumptions (such as the proportional hazard assumption), there is no need to specify a functional form for the distribution of the response (like in parametric models) and these models are very robust to the presence of outliers. In addition to this, the models generated by these methods are very easy to interpret even for people with no statistical background. Although there are some drawbacks, some of which will be discussed in subsequent Chapters, the next Chapter will start with an extensive review of these type of models with special emphasis on the modeling of lifetime data. The coronary and the breast cancer datasets will be used for the application of these alternative models and the results will be compared to those obtained in this Chapter.

Chapter 3

Tree based methods

In this Chapter tree based methods are presented for modeling the relationship between a set of explanatory variables and a continuous, categorical, count or survival response. The Chapter begins with an introduction to the topic and some historical remarks about the evolution of this type of model. The second part of the chapter presents an overview of the general ideas surrounding recursive partitioning that is the algorithm generally used to build trees. The main focus will be on classification and regression trees (CART) and extensions of CART to count responses. In the third part an overview of the most relevant methods for growing survival trees is presented. Maybe due to the specific nature of survival data this particular field seems to have been a more active area of research and many different algorithms have been proposed. In the fourth part, unbiased recursive partitioning is introduced and some examples of survival trees constructed using this method are presented. Finally, a summary of the main ideas of ensemble methods in general, and random survival forests in particular, are presented in the final part of the Chapter.

3.1 Introduction

The representation of the results of a statistical analysis as a tree has been adopted by many different statistical techniques such as hierarchical cluster analysis, decision trees, phylogenetic trees, etc. The focus of this work is on what one broadly can call **tree based models**. They are used as an alternative to generalized linear models (including continuous, categorical, and count responses) and the Cox model in the case of survival responses. The history of tree based models goes back to the seminal work of Morgan & Sonquist (1963) and the development of the AID algorithm (Automatic Interaction Detection). In the context of social science they pointed out how, in the presence of many interaction effects, classical regression analysis was unable to account for such interactions. Algorithms for variable

selection such as forward selection, backwards elimination and stepwise variable selection do not take into account the presence of interactions. Morgan and Sonquist claimed that only after the dimension of the problem has been reduced can one handle interaction effects. They proposed an algorithm that searched over all the possible partitions of the sample space generated by each one of the predictors and selected the partition that lead to the highest reduction of the variance of the response. In their own words: ” *Considering all feasible divisions of the group of observations on the basis of each explanatory factor to be included (but not combinations of factors) find the division of the classes of any characteristic such that the partitioning of this group into two subgroups on this basis provides the largest reduction in the unexplained sum of squares*”. Such an algorithm can be repeated for each one of the subgroups generated and the process can be iterated many times. This method is commonly known as **recursive partitioning**. One of the early problems that this method had was the determination of the optimal size of the tree, which was usually generated based on some previously specified stopping criterion. One solution was given by Breiman *et al.* (1984) and the introduction of CART (Classification and Regression Trees) which introduced the idea of pruning a large tree based on a cross validated measure of the cost-complexity of a tree (i.e. a measure that evaluates the error of the tree taking into account the complexity of the model).

Another problem of recursive partitioning methods that has been identified by many authors (Doyle, 1973; White & Liu, 1994; Shih & Tsai, 2004; Kim & yin Loh, 2001) is the problem of variable selection bias. Predictors with many possible splits are more likely to be selected, even if they are not strongly associated with the response. Recently, Hothorn *et al.* (2006) proposed the use of unbiased recursive partitioning to grow the tree, based on a general theory of permutation tests developed by Strasser & Weber (1999). By using the results of an hypotheses test as stopping criteria they claim to avoid the problem of variable selection bias and to get the right size tree. This approach departs from the idea of pruning that dominated the use of recursive partitioning for many years. However, using the results of an hypothesis test as stopping criteria is not a new idea. The Chi-squared automated interaction detection algorithm (CHAID) (Kass, 1980) used the results of a chi-squared test to select the relevant predictors and to stop the growth of a binary partition tree, although this could only be implemented for nominal responses. In this regard, the work of Hothorn *et al.* comes as a generalization of the ideas proposed in the CHAID algorithm to any other type of response in a unified framework.

In addition to the research on recursive partitioning, in the last few years the use of ensemble methods applied to tree based methods has gained a good deal of popularity. At the core of this popularity is the random forest approach by Breiman (2001). Ensemble methods are mainly focused on prediction and the

basic idea is to use a set of models, that aim to give accurate predictions of the outcome of interest, to improve the prediction accuracy when new observations are considered. In general, the predictive value of the ensemble is the average of all the predictions over all the models. Random forests use bootstrap samples to generate a set of models which are each built using recursive partitioning. Although in terms of prediction random forests perform very well, they are sometimes criticized for the absence of an individual model that helps to understand the relationships between predictors and response. Therefore, ensemble methods are often referred to as “black boxes” for prediction.

In the following sections some of the topics described in the introduction will be explained in detail.

3.2 Classification and regression trees

One of the most popular algorithms for growing trees using recursive partitioning is CART (Classification and regression trees) by Breiman *et al.* (1984). This algorithm is able to model responses that are either continuous or categorical. The main feature of CART is the idea of pruning to avoid overfitting, after a large tree has been grown. The result is a model that can be represented as a tree and can be seen as a non-parametric alternative to the corresponding generalized linear model. The process of generating such a model has two key components, the

- splitting criterion and
- pruning procedure.

To illustrate some of the ideas related to the construction of a tree consider the following example.

Example Let Y be a continuous response and X a continuous predictor. Figure 3.1 (b) shows the relationship between Y and X and the fitted regression line (blue points and black line respectively). Let τ be the set of all the observations in the random sample $\tau = \{(x_i, y_i) : i = 1, \dots, n\}$. Any individual value x_i generates a binary partition of τ given by $\tau = \tau_{L_i} \cup \tau_{R_i}$ where $\tau_{L_i} = \{(x_j, y_j) : x_j \leq x_i\}$ is the set of observations for which $x_j \leq x_i$ (left node) and $\tau_{R_i} = \{(x_j, y_j) : x_j > x_i\}$ is the set of observations for which $x_j > x_i$ (right node). The sum of squares of the left and right nodes are $SS_{L_i} = \sum_{(x_j, y_j) \in \tau_{L_i}} (y_j - \bar{y}_{L_i})^2$ and $SS_{R_i} = \sum_{(x_j, y_j) \in \tau_{R_i}} (y_j - \bar{y}_{R_i})^2$ respectively where \bar{y}_{L_i} is the average of the values $\{y_j : (x_j, y_j) \in \tau_{L_i}\}$ and \bar{y}_{R_i} is the average of the values $\{y_j : (x_j, y_j) \in \tau_{R_i}\}$. Figure 3.1 (a) shows the values of $SS_{L_i} + SS_{R_i}$ as a function of x_i . One can see

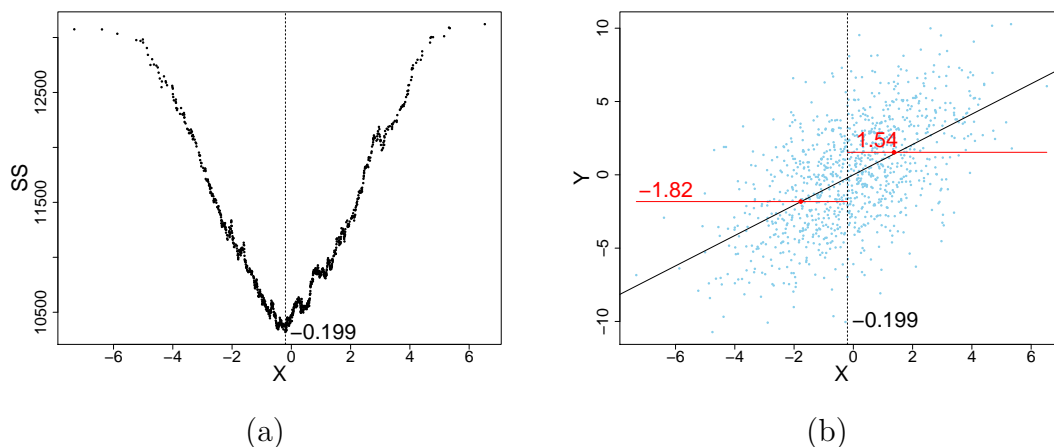


Figure 3.1: In (a) the cutpoint that minimizes the sum of squares of the left and right nodes. In (b) the sample (blue), regression line (black) and predicted outcomes generated by that cutpoint (red).

that this function attains its minimum at $x = -0.199$. At that point the right and the left nodes have the maximum level of homogeneity (minimum error of the left and right nodes). Figure 3.1 (b) shows the split $x = -0.199$ and the predicted values (**node summaries**) of the left and right nodes (-1.82 and 1.54 respectively). Notice that the selection of a constant (i.e. the mean) as the summary of the node is the simplest thing to do. One could think of more complex settings in which, for instance, a separate model could be fitted for the left and right nodes, although these will not be considered here. Another way of representing the split is displayed in Figure 3.2 where a parent node τ is partitioned into two offspring τ_L and τ_R . If $x \leq -0.199$, one descends to the left and the predicted outcome for the response is -1.82 and if $x > -0.199$, one descends to the right and the predicted outcome is 1.54.

The process just described can be repeated by simply considering the two daughter nodes as parents and repeating the same algorithm (recursive partitioning). At each stage, the set of splits of the corresponding parent node are considered, and the best split is appropriately selected. Figure 3.3 shows an example with 7 splits and the corresponding predicted values. As one can see, the structure of the data is better captured when more splits are considered. In addition, the within-node variability is also reduced with more terminal nodes. The resulting regression tree is presented in Figure 3.4.

One of the main questions regarding the use of recursive partitioning is when to stop growing the tree. One can keep generating splits until the tree is overfitted. An example is shown in Figure 3.5 where 30 splits were generated. As one can see, some of the predicted values are very local and clearly biased. Therefore,

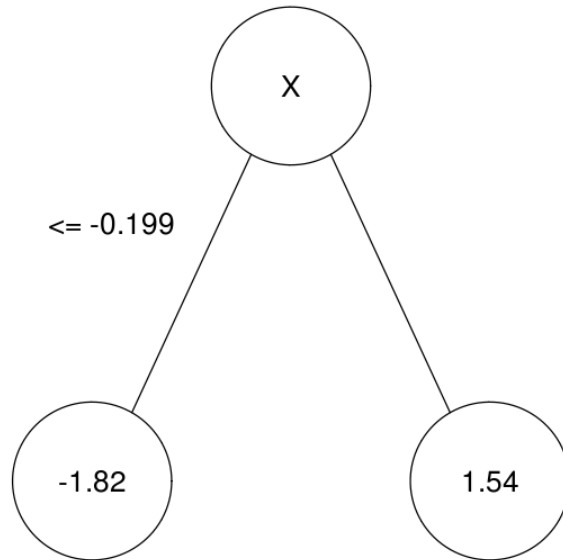


Figure 3.2: Graphical representation of the split generated in Figure 3.1.

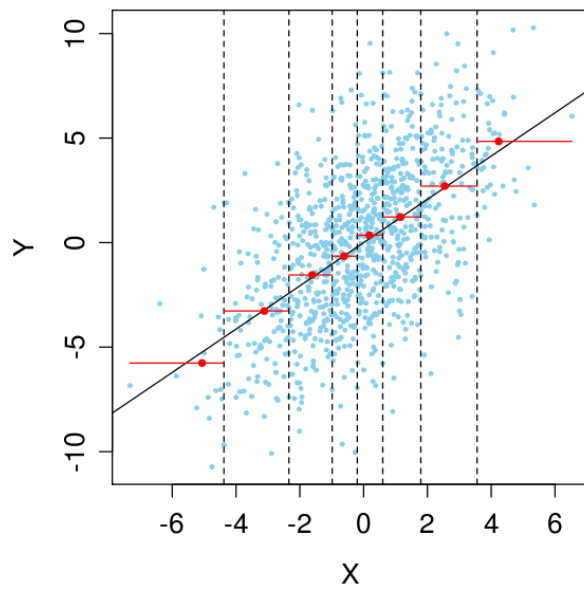


Figure 3.3: Example of 7 splits using recursive partitioning.

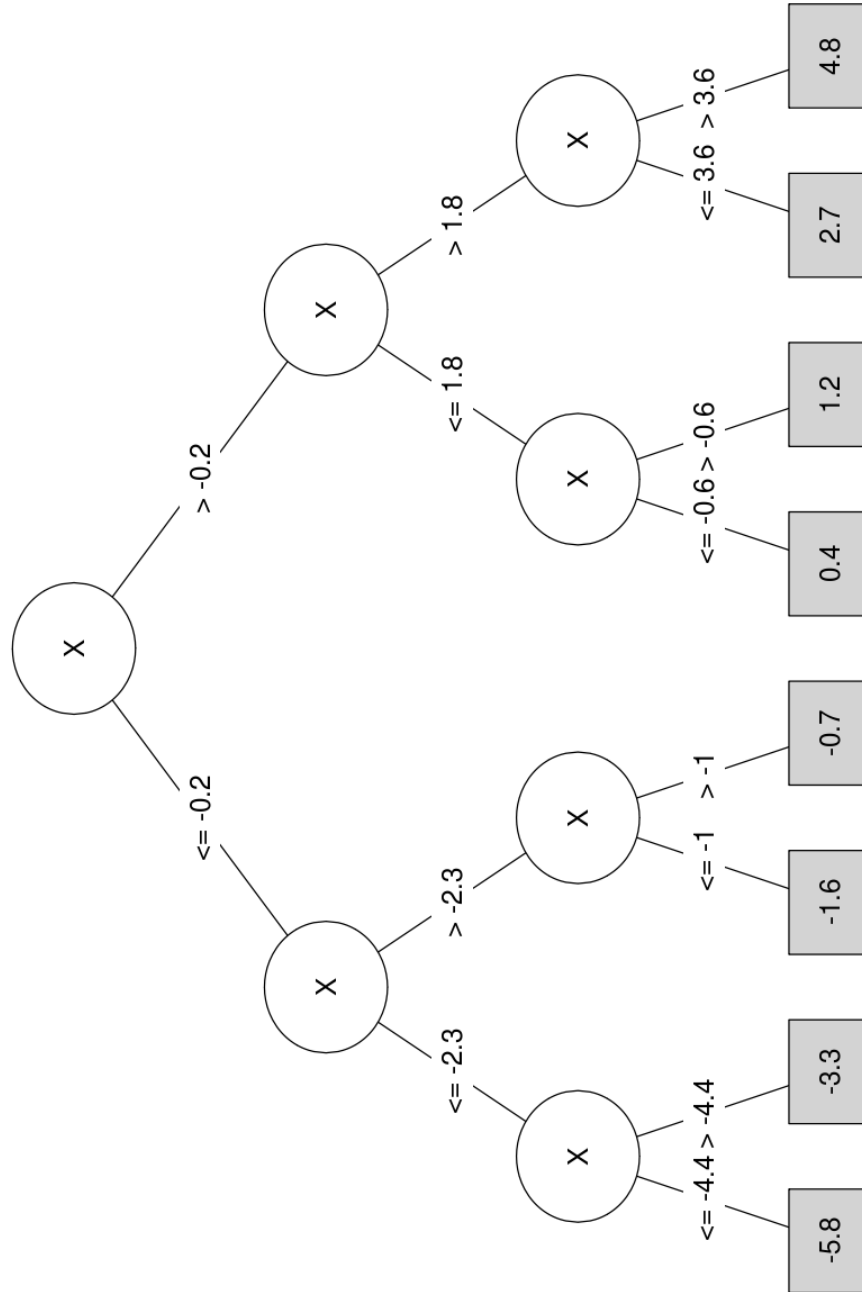


Figure 3.4: Example of a regression tree with 8 terminal nodes and only one continuous predictor.

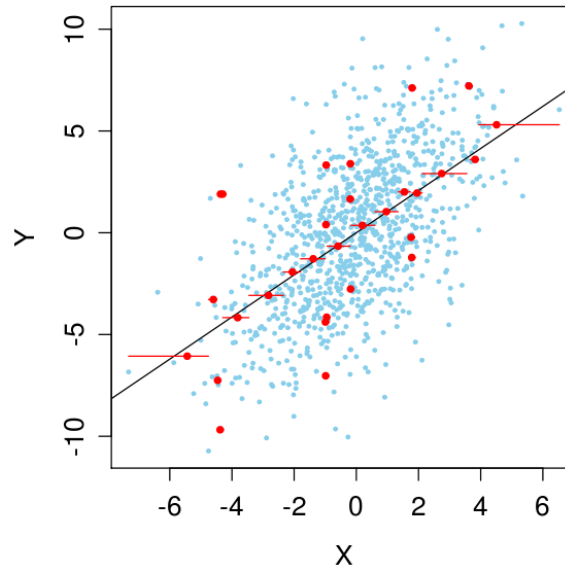


Figure 3.5: An example of overfitting.

if the number of nodes is too small, high variance dominates the predicted values. On the other hand, if the number of nodes is too high, high bias dominates the predicted values. In order to get the right size of the tree the usual bias-variance trade-off have to be considered (Hastie *et al.*, 2001).

Before going into more detail and a more formal description of the CART algorithm some remarks need to be made. The first one is that the example presented above has only one continuous predictor. However, if more predictors are present in the data the same algorithm can be used for each predictor. Each time a split has to be generated, the predictor is chosen such that the combined sum of squares of the two offspring for the optimal split is minimum compared to the combined sum of squares obtained by the other predictors. Another interesting remark is that the search for the optimal split given a particular predictor X depends on the type of data of the predictor. Although there are different ways in which this search can be performed, for the sake of simplicity only the following scheme will be considered:

- If the predictor is continuous, the search is done over all the observed values x_i (except the maximum and the minimum). Each one of those values generates the two offspring, i.e. the set of observations for which $X \leq x_i$ and the set of values for which $X > x_i$. The total number of possible splits is $n - 2$ where n is the number of observations in the node.

- If the predictor is categorical with only two levels there is only one possible split. If there are more than two levels the search is done over all the possible combinations of the levels. The number of splits is, in this case, $2^{k-1} - 1$ where k is the number of categories.
- If the predictor is measured on an ordinal scale, the search can be done in the same fashion as a continuous predictor by simply labeling the levels with the number corresponding to the ordered level of the factor. The number of splits here is k where k is the number of ordered categories.

3.2.1 Continuous responses (regression tree)

If the response is continuous the splitting criterion is based on the sum of squares as a measure of the homogeneity of the node. Let $\{(y_i, \mathbf{x}_i) : i = 1, \dots, n\}$ be a random sample where y_i is the value of the continuous response for individual i and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ is the vector of the corresponding values of p different predictors. For a given node τ , the impurity of the node can be defined as

$$I(\tau) = \sum_{i: (y_i, \mathbf{x}_i) \in \tau} (y_i - \bar{y}_\tau)^2$$

where \bar{y}_τ is the average of y_i for all the individuals in node τ . The split will be chosen such that the change of impurity

$$\Delta I(\tau) = I(\tau) - (I(\tau_L) + I(\tau_R))$$

is maximized. Therefore, the optimal split will produce two offspring that are as homogeneous as possible. This process can be repeated until a large tree is generated.

Once a large (or saturated) tree T_{\max} has been produced a measure of the relative error of the tree can be obtained by

$$R(T_{\max}) = \sum_{\tau \in \tilde{T}} R(\tau) \tag{3.1}$$

where $R(\tau) = I(\tau)/I(\tau_0)$ and \tilde{T} is the set of terminal nodes (note that τ_0 represents the parent node). This measure of relative error can be calculated for any tree. The process of pruning T_{\max} is based on the notion of the **cost-complexity measure of a tree** given by

$$R_\alpha(T) = R(T) + \alpha|T| \tag{3.2}$$

where T is any tree, α is the complexity parameter (tuning parameter) that regulates the bias-variance trade-off and $|T|$ is the size of the tree quantified by the number of terminal nodes. Based on these definitions, the algorithm to grow the optimal tree can be summarized in the following steps:

1. Using recursive partitioning grow a large (saturated) tree T_{\max} .
2. Prune the tree by snipping back splits of the saturated tree into smaller trees using the cost-complexity measure given by (3.2).
3. Obtain the optimal value of the complexity parameter α using cross validation and select the tree yielding the lowest cross-validated error rate.

To explain in more detail the last 2 steps of the algorithm, let T_{\max} be the saturated tree with $|\tilde{T}|$ terminal nodes. For a particular value of α one can calculate the relative error of any subtree¹ given by equation (3.1) and pick the subtree with the smallest error (optimal subtree). Thus, there is function between the value of the complexity parameter and the relative error of the optimal subtree. If the value of α is very small the cost of incorporating new splits is very small and larger subtrees would be chosen as optimal. On the other hand, if α is very large, each additional split would increase the value of the error very rapidly and small subtrees would be selected as optimal.

In order to obtain the optimal value of α CART uses cross validation (usually 10-fold cross validation). Basically, the procedure consists of dividing the full sample into $s = 10$ groups. Each of the groups is removed one at a time and a tree is grown without those observations. For each corresponding value of α the optimal subtree with the minimum error is considered and the individuals of the group that were left out are used to estimate the relative error. The process is repeated s times removing each group in turn. At the end, a sample of s relative errors is obtained for each value of α and their standard error can be calculated. In order to choose the best value of α the one standard error (1 SE) method outlined in Breiman et al (1984) is used. The tree with the minimum cross-validated error is considered and any other tree with cross-validated error within one standard error of the achieved minimum is marked as being equivalent to the minimum. The simplest tree is then chosen.

To illustrate this idea Figure 3.6 shows a plot of the relative error as a function of the complexity parameter for the example described in Figure 3.5. On the top axis of the plot there is the size of the optimal tree for each particular value of α . The minimum is attained at $\alpha = 0.0062$ which corresponds to the tree with 6 terminal nodes. However, trees with 3,4 and 5 terminal nodes could be considered equivalent as they are within the one standard error of the minimum. Figure 3.7 (left) shows the optimal tree with 3 terminal nodes.

Note that one of the advantages of using recursive partitioning is that the algorithm can deal very naturally with missing values. If one wants to obtain the predicted value of any observation for a particular tree it is enough to drop the

⁴A subtree is any tree created by snipping back splits of the saturated tree.

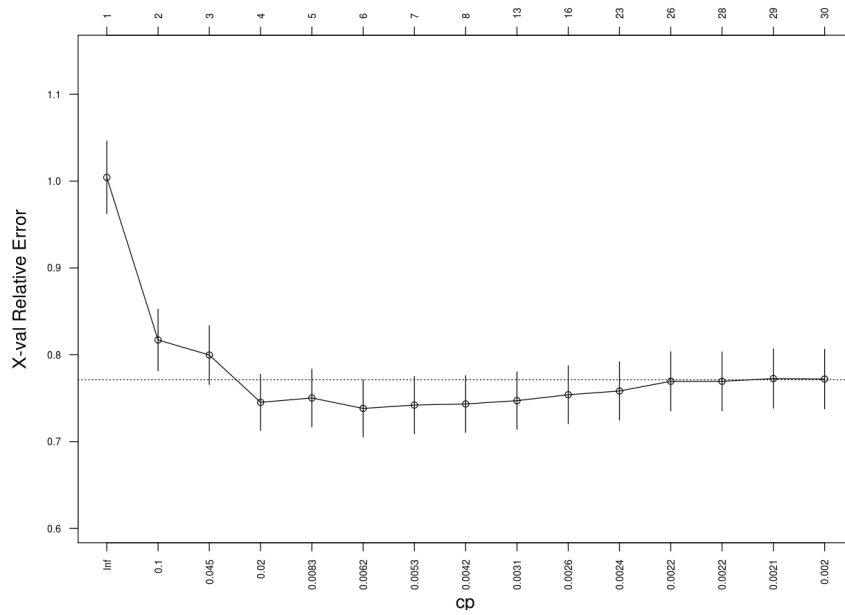


Figure 3.6: Cross validated error as a function of the complexity parameter.

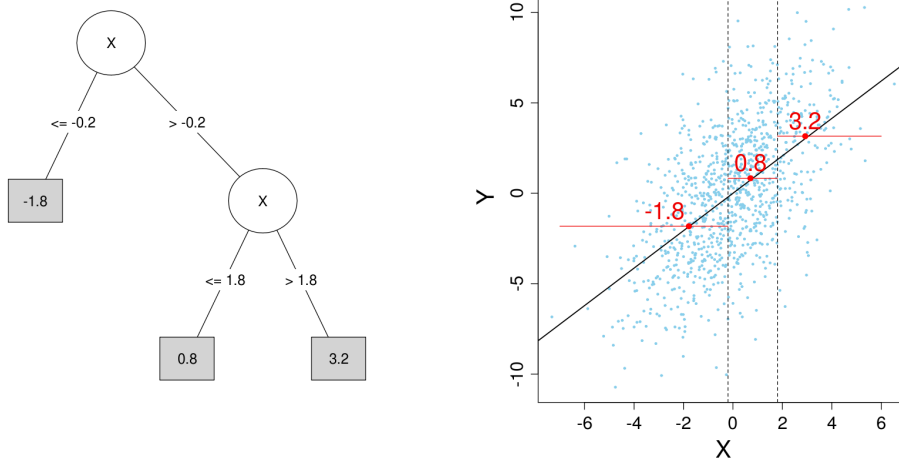


Figure 3.7: Optimal tree after pruning. On the left the structure of the tree. On the right the predicted values and the cut-points.

observation down the tree until a terminal node is reached (based on the values of the predictors). However, if the information is missing for some of the predictors it is possible to use surrogate splits which are the optimal splits based on the other predictors ordered in a decreasing order. The first surrogate is the second best predictor, the second surrogate is the third best predictor and so on. Therefore, unless the observation has missing values for all the predictors, one can always obtain the predicted value.

3.2.2 Categorical responses (classification tree)

If the response is categorical the splitting criterion is based on some value of the entropy of the distribution of the response as a measure of the homogeneity of the node. The most common measures used are the Gini diversity index (Gini, 1912), $I_{Gini}(\tau) = \sum_{i=1}^C p_{i|\tau}(1 - p_{i|\tau}) = 1 - \sum_{i=1}^C p_{i|\tau}^2$ and the information diversity index $I_{Info}(\tau) = \sum_{i=1}^C -p_{i|\tau} \log(p_{i|\tau})$ where τ is any given node and $p_{i|\tau}$ is the estimated probability, or proportion, of the i th level of the categorical response in node τ (it is assumed that there are C categories). These two functions attain their maximum when the distribution is uniform (maximum level of uncertainty) and play the same role as the sum of squares when the response is continuous. If the Gini diversity index is used, the impurity of a node τ can be measured by

$$I(\tau) = p_{\tau} I_{Gini}(\tau)$$

where p_{τ} is the proportion of individuals in node τ (in reality there is very little difference between I_{Gini} and I_{Info}). The split will be chosen such that the change of impurity

$$\Delta I(\tau) = I(\tau) - (I(\tau_L) + I(\tau_R))$$

is maximized. Thus, the optimal split will produce two offspring that are as pure as possible.

Once a large tree T_{\max} has been grown a measure of the error of the tree can be obtained by

$$R(T_{\max}) = \sum_{\tau \in \tilde{T}} p_{\tau} R(\tau)$$

where $R(\tau)$ is the proportion of misclassified² individuals in node τ and p_{τ} is the proportion of individuals in node τ relative to the total number of individuals. The process of pruning T_{\max} is analogous to the continuous case based on the cost-complexity measure given by (3.2).

²The predicted outcome for a categorical response in a particular node τ is the mode of the predicted probability mass function. Therefore the proportion of misclassified, or risk of the node, corresponds to the proportion of observations that fall in any of the remaining levels.

3.2.3 Extension of CART to Poisson regression

Although originally developed to undertake statistical analysis for continuous and categorical responses, the CART algorithm can be extended easily to deal with responses based on counts. Let $\{(c_i, t_i) : i = 1, \dots, n\}$ be a random sample where, for individual i , c_i is the number of times some event of interest occurs and t_i is the period of time in which the c_i events occurred. These data can be analyzed using a Poisson model where for individual i , the density function is

$$e^{-\lambda t_i} \frac{(\lambda t_i)^{c_i}}{c_i!}$$

where λ is the parameter of interest. The likelihood function for the entire dataset is

$$L(\lambda) = \prod_{i=1}^n e^{-\lambda t_i} \frac{(\lambda t_i)^{c_i}}{c_i!}$$

and the maximum likelihood estimate of λ is

$$\hat{\lambda}_{\text{mle}} = \frac{\sum_{i=1}^n c_i}{\sum_{i=1}^n t_i}.$$

To analyze these type of data from the tree based method perspective (Chambers & Hastie, 1992), the splitting criterion can be based on the deviance of the data as a measure of the homogeneity of the node. Given a node τ , the deviance is defined as

$$2[\log L_\tau(\text{saturated}) - \log L_\tau(\hat{\lambda}_\tau)] \quad (3.3)$$

where L_τ refers to the likelihood function of the individuals in τ , $\hat{\lambda}_\tau$ is the corresponding maximum likelihood estimate in τ and $L_\tau(\text{saturated})$ is the value of the likelihood function of the saturated model in which every individual i will have a different value of λ_i . Using the likelihood function, (3.3) can be written as

$$2 \sum_{(c_i, t_i) \in \tau} \left[c_i \log \frac{c_i}{t_i \hat{\lambda}_\tau} - (c_i - \hat{\lambda}_\tau t_i) \right].$$

The impurity of the node can be defined as

$$I(\tau) = \frac{1}{N} \sum_{(c_i, t_i) \in \tau} \left[c_i \log \frac{c_i}{t_i \hat{\lambda}_\tau} - (c_i - \hat{\lambda}_\tau t_i) \right]$$

and the split is chosen such that the change in impurity

$$\Delta I(\tau) = I(\tau) - (I(\tau_L) + I(\tau_R))$$

is maximized. Furthermore, given a tree T the error of the tree is

$$R(T) = \sum_{\tau \in \tilde{T}} I(\tau)$$

where \tilde{T} is the set of terminal nodes of tree T . The prediction error of a new observation is the deviance contribution of such an observation using the $\hat{\lambda}$ corresponding to the terminal node in which the new observation falls. Using this prediction error it is possible to obtain cross validated estimates of the cost-complexity measure of the error of the tree given by (3.2).

There is a problem with this method though. If none of the individuals in a terminal node τ experience an event, then $\hat{\lambda}_\tau = 0$, and therefore, the deviance contribution of other individual with at least one event will be ∞ . This could happen in the cross validation process where some of the observations are left out. Although some solutions have been suggested, such as the use of

$$\hat{\lambda} = \max \left(\hat{\lambda}_{\text{mle}}, \frac{0.5}{\sum t_i} \right)$$

as an estimate of λ , or the use of shrinkage estimates of λ (Therneau 1997), the cross-validation procedure tends to give very conservative results and very often chooses the tree with no splits as being the optimal tree.

3.3 Survival trees

Tree based methods for analyzing survival data have to be able to accommodate the presence of censoring. Unlike the methods presented above, where there is only one splitting criterion for continuous responses and one splitting criterion for categorical responses, in survival analysis there are many criteria to split the nodes and therefore to grow the tree. This section contains a review of such methods.

One of the first attempts to extend the use of recursive partitioning to the analysis of survival data was made by Gordon & Olshen (1985). They based their method on the idea of the impurity of a node similar to that given for categorical responses. A node would be “pure” if all the individuals in the node had the event of interest at the same time. This is illustrated in Figure 3.8 (a) where an example is given of the survival function³ of a distribution in which all the events occur at time 4.

Now let $S_{KM}(t)$ be the Kaplan-Meier estimate of the survival function of individuals belonging to a particular node τ in a tree T (black line in Figure 3.8 (b)).

³To simplify it will be assumed that there are no censored observations beyond the maximum observed time.

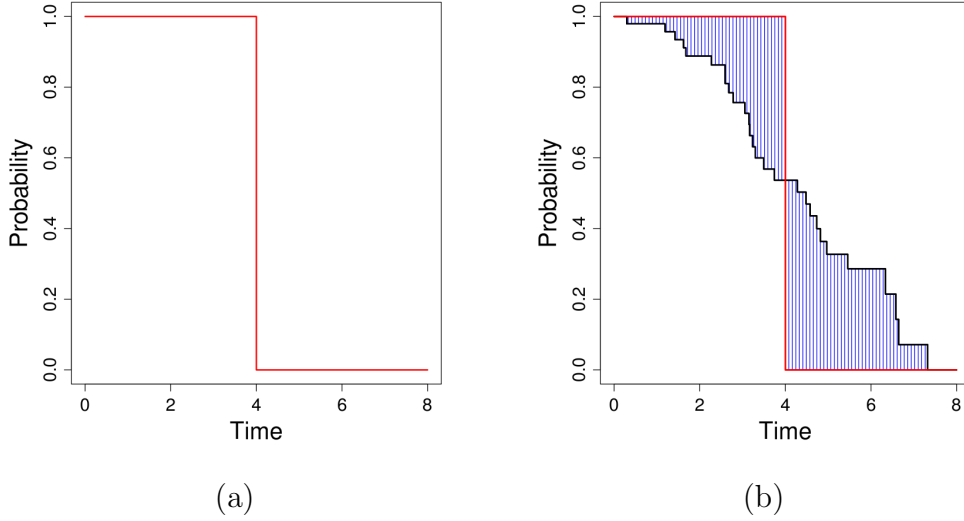


Figure 3.8: An example of a “pure” node (a) and the distance between the Kaplan-Meier estimate of the survival function (in a hypothetical node τ) and the survival function of the “pure” node (blue area in (b)).

The distance between S_{KM} and the survival function of the pure node at time 4, S_{P_4} , is represented as the blue area. This is the L^1 Wasserstein distance between the two survival functions. In general, the L^p Wasserstein distance between two distributions functions F_1 and F_2 is (full details are given in Zhang & Singer, 2010)

$$\left[\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p du \right]^{\frac{1}{p}}.$$

In order to measure the impurity of a node $\tau \in T$, Gordon & Olshen (1985) considered the following value

$$I(\tau) = p_\tau d_2(S_{KM}, S_{P_s})$$

where p_τ is the proportion of individuals at node τ (relative to number of individuals in the sample), d_2 is the L^2 Wasserstein distance and S_{P_s} is the “pure” survival function that is closest to S_{KM} . Therefore, $d_2(S_{KM}, S_{P_s})$ is the minimum L^2 Wasserstein distance between the Kaplan-Meier estimate of the survival function and the pure survival function S_{P_s} .

Based on this measure of impurity, the split will be chosen such that the change in impurity is maximized. Therefore the optimal split at node τ will generate the right (τ_R) and left (τ_L) nodes for which

$$\Delta I(\tau) = I(\tau) - (I(\tau_L) + I(\tau_R))$$

is maximum. Once the splitting rule has been defined, the rest of the process of growing the optimal tree is very similar to that of the classification trees seen above.

All of the splitting methods described so far, the one described in this section and the ones for continuous and categorical responses discussed in the previous section, are examples of splitting criteria based on **within node homogeneity**. In this type of methods, the optimal split is chosen such that the change in impurity is maximized. Therefore, once the impurity of a node has been defined the process of growing the optimal tree is very similar no matter what type of response is being analyzed. Segal (1988) proposed a different splitting criterion based on measurements of **between node separation** in the context of survival analysis. For regression problems with a continuous response, splitting criteria based on between node separation give analogous results to those based on within node homogeneity⁴. However, when analyzing survival data, because of the censoring, the homogeneity of a node cannot be evaluated using the sum of squares. Segal argued that the use of the logrank statistic (or one of its variants) as a measure of dissimilarity between nodes is particularly appropriate for survival trees. Some of the reasons for this are:

- due to the fact that the logrank statistic is part of the family of rank statistics, the results are invariant to monotone transformations of the response;
- therefore, the splits are insensitive to the presence of outliers in the response and not just to the presence of outliers in the predictors;
- the calculation of the logrank statistics is computationally feasible;
- censoring is easily accommodated.

The optimal splits based on measurements of between node separation are chosen such that the separation between the two offspring of a node is maximal. Segal (1988) proposed the use of any of the weighted logrank statistics seen in the previous Chapter and pointed out that the resulting trees are very similar for most of the weights, except for those that lead to the Gehan (1965) statistic.

Segal proposed a different method for pruning the tree since the minimal cost-complexity algorithm used by CART does not work here, due to the fact that the logrank statistic does not provide a measure of within node homogeneity. This new bottom-up approach is referred to as the subtree maximal statistic pruning procedure by Segal and can be summarized in the following steps:

1. initially grow a very large tree;

⁴This is due to the decomposition of the sum of squares as the sum of squares within plus the sum of squares between.

2. assign to each internal node the maximum logrank statistic contained in the subtree of which the node under consideration is the root;
3. locate the highest node in the subtree with the smallest maximum and remove all its descendants. The remaining tree is a maximal subtree;
4. repeat 2) and 3) until all that remains is the root node. This generates a sequence of maximal subtrees;
5. plot the maximal subtree split statistics against tree size and pick the tree corresponding to the characteristic “kink” in the curve.

This new method for pruning a tree does not need to use cross validation to obtain the optimal size of the tree.

LeBlanc & Crowley (1993) extended the use of survival trees based on logranks as splitting criterion by introducing the notion of the split-complexity measure of a tree, analogous to the cost-complexity measure used by CART. This measure can be used to prune a tree that has been grown using the logrank statistic as a measure of between node dissimilarity. For a given tree T the split-complexity is defined as

$$G_\alpha(T) = G(T) - \alpha|S|$$

where $|S|$ is the number of internal nodes of tree T , $\alpha \geq 0$ is the complexity parameter and $G(T)$ is the sum over the internal nodes of the standardized logrank statistics (for more details of how to pick the optimal tree see LeBlanc & Crowley (1993)).

Davis & Anderson (1989) proposed a completely different splitting criterion based on the likelihood function where an exponential model is assumed for the underlying distribution of the survival times. Under this model the log-likelihood function of a random sample of survival times $\{(x_1, \delta_1), \dots, (x_n, \delta_n)\}$ is

$$\log L(\lambda; \mathbf{x}, \boldsymbol{\delta}) = \sum_{i=1}^n \delta_i \log \lambda - \sum_{i=1}^n \lambda x_i$$

where λ is the parameter of interest. Thus, the maximum likelihood estimate of λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n x_i} = \frac{\text{total deaths}}{\text{total time on test}}.$$

The proposed splitting criterion was based on the value of

$$-\log L(\hat{\lambda}) = (1 - \log \hat{\lambda}) \sum_{i=1}^n \delta_i$$

which can be seen as a measure of the within node homogeneity. For a given node τ in a tree T the impurity is defined as $I(\tau) = -\log L(\hat{\lambda}_\tau)$ where $\hat{\lambda}_\tau$ is the maximum likelihood estimate of λ based on the observations belonging to node τ . Therefore the optimal split at node τ is obtained by maximizing $\Delta I(\tau) = I(\tau) - (I(\tau_L) + I(\tau_R))$ over all the predictors and all the possible cutpoints. The optimal size of the tree can be determined following the usual pruning procedure of the CART algorithm.

An extension of this approach was proposed by LeBlanc & Crowley (1992) which is based on the less restrictive assumption of the proportional hazard model. Let τ be a node in a tree T . Suppose that

$$\lambda_\tau(t) = \lambda_0(t)\theta_\tau$$

for all $\tau \in T$ where $\lambda_\tau(t)$ is the hazard function of individuals in node τ and $\lambda_0(t)$ is the baseline hazard. The full likelihood function of the tree T can be written as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{\tau \in \tilde{T}} L_\tau(\theta_\tau) \\ &= \prod_{\tau \in \tilde{T}} \prod_{(x_i, \delta_i) \in \tau} \lambda_\tau(x_i)^{\delta_i} e^{-\Lambda_\tau(x_i)} = \\ &= \prod_{\tau \in \tilde{T}} \prod_{(x_i, \delta_i) \in \tau} [\lambda_0(x_i)\theta_\tau]^{\delta_i} e^{-\Lambda_0(x_i)\theta_\tau} \end{aligned}$$

where \tilde{T} is the set of terminal nodes of tree T and $\boldsymbol{\theta} = \{\theta_\tau : \tau \in \tilde{T}\}$. The maximum likelihood estimate of θ_τ for any $\tau \in \tilde{T}$ is

$$\hat{\theta}_\tau = \frac{\sum_{(x_i, \delta_i) \in \tau} \delta_i}{\sum_{(x_i, \delta_i) \in \tau} \Lambda_0(x_i)}. \quad (3.4)$$

LeBlanc & Crowley (1992) proposed the use of the Nelson (1972) estimator of the cumulative baseline hazard function in (3.4) given by

$$\hat{\Lambda}_0(t) = \sum_{i: x_i \leq t} \frac{d_i}{n_i} \quad (3.5)$$

where d_i is the number of deaths at time x_i and n_i is the number of individuals at risk at time x_i . The value in (3.4) can be used as a meaningful summary of a terminal node that can be interpreted as the observed number of deaths divided by the expected number of deaths in node τ under the assumption of no structure in survival times.

The splitting criterion of this method is based on the deviance of any node τ and can be interpreted as a measure of within node homogeneity. The formula for the deviance is

$$2[\log L_\tau(\text{saturated}) - \log L_\tau(\hat{\theta}_\tau)] \quad (3.6)$$

where $L_\tau(\text{saturated})$ is the likelihood function that assumes a separate parameter for each observation. The maximum likelihood estimate of θ_τ is (3.4) and the maximum likelihood estimates of each θ_{τ_i} of the saturated model is

$$\hat{\theta}_{\tau_i} = \frac{\delta_i}{\Lambda_0(x_i)}$$

where $i = 1, \dots, n_\tau$ (note that the set of those indexes is equivalent to $\{i : (x_i, \delta_i) \in \tau\}$). Plugging these estimates into (3.6) one obtains

$$2 \left[\sum_{i=1}^{n_\tau} \delta_i \log \frac{\delta_i}{\Lambda_0(x_i)\hat{\theta}_\lambda} - (\delta_i - \Lambda_0(x_i)\hat{\theta}_\lambda) \right]$$

and therefore the deviance of a node τ can be written as $\sum_{i=1}^{n_\tau} d_i$ where

$$d_i = 2 \left[\delta_i \log \frac{\delta_i}{\Lambda_0(x_i)\hat{\theta}_\lambda} - (\delta_i - \Lambda_0(x_i)\hat{\theta}_\lambda) \right] \quad (3.7)$$

is the deviance residual for an observation i in node τ . These deviance residuals are equivalent to the deviance residuals one would obtain from a Poisson model with response δ_i and the mean parameter given by $\Lambda_0(x_i)\theta_\tau$ (see previous section 3.2.3).

The impurity of a node τ is define as

$$I(\tau) = \frac{1}{N} \left[\sum_{i=1}^{n_\tau} \delta_i \log \frac{\delta_i}{\hat{\Lambda}_0(x_i)\hat{\theta}_\lambda} - (\delta_i - \hat{\Lambda}_0(x_i)\hat{\theta}_\lambda) \right]$$

where N is the number of observations used to grow the tree. Thus, the optimal split is chosen such that the change in impurity

$$\Delta I(\tau) = I(\tau) - (I(\tau_L) + I(\tau_R))$$

is maximized. The process of pruning the tree after a large or saturated tree has been grown is analogous to the one used in CART. The cost-complexity of a tree T is defined as

$$R_\alpha(T) = \sum_{\tau \in \tilde{T}} R(\tau) + \alpha|\tilde{T}|$$

where $R(\tau)$ refers to the impurity the node τ .

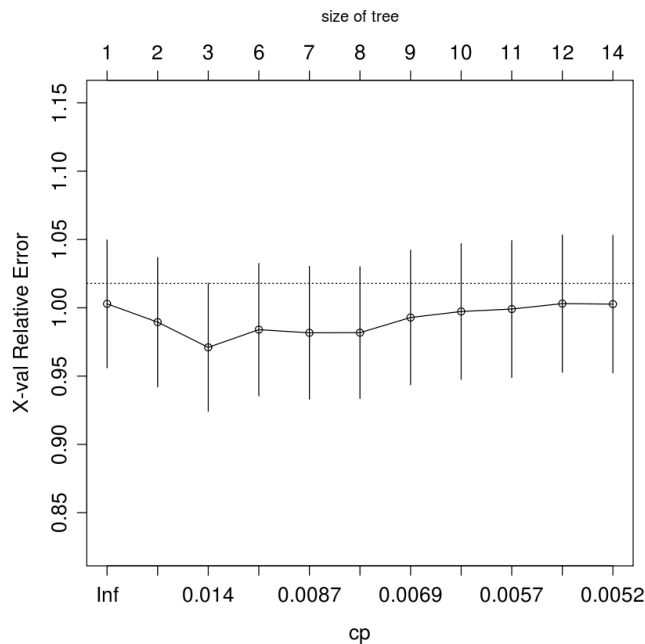


Figure 3.9: The cost-complexity plot applied to the coronary data using the method proposed by LeBlanc & Crowley (1992).

As mentioned in the previous section the problem with this method is where nodes have all observations censored, $\hat{\theta}_\tau = 0$, from (3.4). In that case, an uncensored observation that has been ‘left out’ when calculating the cross validated deviance residual will produce an infinite value in (3.7). An ad hoc solution proposed by LeBlanc & Crowley (1992) is to estimate θ_τ by

$$\hat{\theta}_\tau = \frac{1}{2 \sum_{(x_i, \delta_i) \in \tau} \Lambda_0(x_i)}$$

where again $\Lambda_0(x_i)$ can be estimated using (3.5). In this way, the estimate of θ_τ is never 0.

In Figure 3.9 the method just presented has been applied to the coronary data and the resulting cross-validated cost-complexity errors are displayed in the plot. As one can see, the minimum is attained for the tree with 3 terminal nodes but, following the 1 standard deviation rule proposed by Breiman *et al.* (1984), trees with 1 and 2 terminal nodes can be considered as equivalent. This would lead us to pick the tree with no splits as the optimal tree, although this might be the effect of the boundary problem mentioned previously. The tree with 3 terminal nodes is shown in Figure 3.10. The terminal nodes display the estimates of the hazard ratios ($\hat{\theta}$) and the number of uncensored events vs the total number of events.

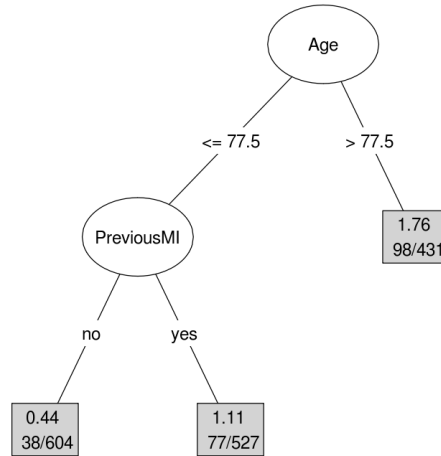


Figure 3.10: Tree obtained after applying the method proposed by Leblanc and Crowley (1992) to the coronary data.

For example, patients younger than 77.5 years old with no previous myocardial infarction have an estimated hazard ratio of 0.44 where the baseline hazard is the hazard of the whole data set. This means that this group of patients do better compared to the sample as a whole. On the other hand, the worse prognosis is for patients older than 77.5 years with an estimated hazard ratio of 1.76.

In Figure 3.11 the same method was applied to the breast cancer dataset and the corresponding cross validated cost-complexity errors are depicted in the plot. Once again, the boundary problem seems to have an effect in the cross-validated error rate based on the deviance residual.

An alternative approach was introduced by Zhang (1995) who suggested a different splitting criterion to grow survival trees. He based the notion of impurity on the idea that an homogeneous node should contain uncensored observations that are homogeneous (small variability) and observations should be mostly censored or mostly uncensored. To quantify the amount of homogeneity in any node τ the following value for the impurity is given

$$I(\tau) = w_1 I_x(\tau) + w_2 I_\delta(\tau)$$

where w_1 and w_2 are weights that have to be predefined, $I_x(\tau)$ is the impurity of

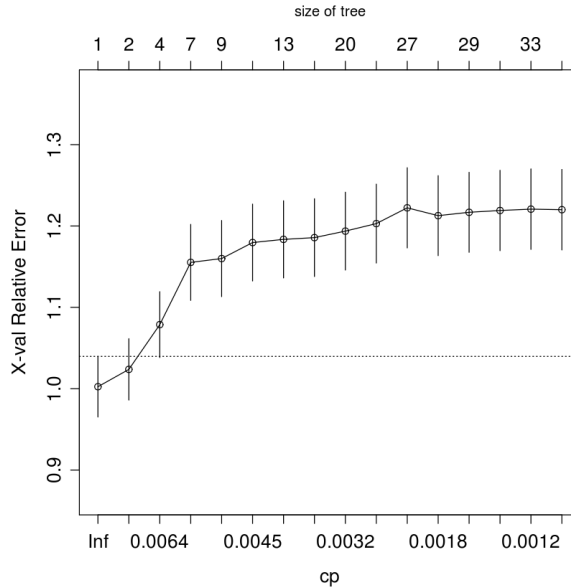


Figure 3.11: Cost-complexity plot applied to the breast cancer dataset using the method proposed by Leblanc and Crowley (1992).

the observed times given by

$$I_x(\tau) = \sum_{(x_i, \delta_i) \in \tau} \frac{[x_i - \bar{x}_\tau]^2}{\sum x_i^2}$$

where \bar{x}_τ is the average of the observed times in node τ , and $I_\delta(\tau)$ is the impurity of the censoring given by

$$I_\delta(\tau) = -p_\tau \log(p_\tau) - (1 - p_\tau) \log(1 - p_\tau).$$

where p_τ is the proportion of censoring in node τ . Notice that, in order to normalize the impurity of the observed times $I_x(\tau)$ and make it comparable with the other component of the impurity, the term $\sum x_i^2$ in the denominator is needed. If the summation is over the observations in τ Zhang called the criterion *adaptive normalization*, whereas if the summation is over the whole sample he called it *global normalization*. He showed in a simulation study that adaptive normalization seems to work better with a weight ratio of 1:2.

Although not discussed in Zhang (1995) he later suggested the use of the same bottom up pruning procedure suggested by Segal (1988) to simplify the tree.

3.4 Conditional inference trees

In this section a different algorithm for growing trees for continuous, categorical and time to event responses is presented. All the methods described above use recursive partitioning to generate a large or saturated tree, and these methods differ in the splitting criterion and the pruning procedure. However, one of the main drawbacks of the recursive partitioning algorithm (sometimes called the greedy search algorithm) is the problem of variable selection bias. This problem has been identified by many authors, including Breiman himself, and consists of that predictors with many possible splits are more likely to be chosen as a splitting variable. White & Liu (1994) showed in simulations that measures of information gain and gain ratio proposed by Quinlan (1986) for the classification problem (as an alternative to the Gini and information diversity indexes by Breiman) were affected by the variable selection bias. Shih (2004) produced some theoretical results to understand the selection bias of the greedy method when the response is two level categorical and the predictors are numerical with possibly missing values. Kim & yin Loh (2001) showed that the variable selection bias can be due also to the number of missing values of the predictor (predictors with many missing values have higher chance of being selected). Different solutions have been proposed by these and other authors for specific types of data. The most recent approach has been developed by Hothorn *et al.* (2006) who proposed the use of conditional inference procedures, based on a general theory of permutation tests developed by Strasser & Weber (1999), to grow the tree. In their own words, "*the separation of variable selection and splitting procedure into steps 1 and 2 of the algorithm is the key for the construction of interpretable tree structures not suffering a systematic tendency towards covariates with many possible splits or many missing values*". They called the algorithm **unbiased recursive partitioning**. The main strength of unbiased recursive partitioning is its flexibility, as it can handle any type of data within the same unified framework. Furthermore, due to the fact that the stopping criterion is based on the result of a statistical hypothesis test (i.e. a p-value), there is no need for pruning. The rest of the section explains the algorithm in detail.

The unbiased recursive partitioning algorithm

Let Y be a response (continuous, categorical or ordinal) and X_1, \dots, X_m the predictors, where any of the X_j can be continuous, categorical or ordinal. The algorithm to grow the tree can be summarized in the following steps:

1. At any node (starting with the parent node), test the global null hypothesis of independence between Y and any of the predictors X_j using the data belonging to that node. If the global test is not significant, the data is not

split any further. If the global test is significant, choose the predictor with the strongest association (smaller p-value).

2. Using the predictor selected in step 1, find the cut-point that maximizes the separation between the two children nodes. This can be done by searching over all the possible cut-points and obtaining the p-values of the test of independence between the left and the right part. The cut-point leading to the smallest p-value is selected for the split.
3. Repeat steps 1 and 2 until no further splits are possible.

To test the global null hypothesis of independence in step 1, multiple test procedures can be performed. The simplest way of doing this is to use a Bonferroni correction to achieve the desired global level of significance. More advanced methods for multiple testing include re-sampling-based adjustments such as the min-P-value re-sampling approach (Westfall & Young, 1993). There are other alternatives but multiple testing procedures are preferred when missing values are present in the data (for full details see Hothorn *et al.*, 2006). Therefore, the global test:

$$H_0 : D(Y|X_1 \dots X_m) = D(Y)$$

can be seen as a composite of the following individual tests:

$$H_0^j : D(Y|X_j) = D(Y)$$

where $j = 1, \dots, m$.

Test of independence between Y and X_j

Let A be a node and $P = \{(y_i, x_{1i}, \dots, x_{mi}) : i = 1, \dots, n_A\}$ the set of all the observations in node A . Although different tests could be used to test the null hypothesis of independence $H_0^j : D(Y|X_j) = D(Y)$ (depending on the nature of the response and the predictor), a general class of tests based on the theory of permutation tests by Strasser and Weber (1999) is used in the unbiased recursive partitioning algorithm. To measure the association between Y and X_j the following linear statistic is considered:

$$T_j = \text{vec} \left(\sum_{i=1}^{n_A} g_j(x_{ji}) h(y_i, (y_1, \dots, y_{n_A}))^T \right) \quad (3.8)$$

where $g_j(x_{ji}) \in \mathbb{R}^{p_j}$ and $h(y_i, (y_1, \dots, y_{n_A})) \in \mathbb{R}^q$. The function h is called the *influence function* and it depends on the vector (y_1, \dots, y_{n_A}) but not on the order in which the values of (y_1, \dots, y_{n_A}) are arranged. Furthermore, the term “vec”

in (3.8) implies that if $p_j > 1$ and $q > 1$ the linear statistic T_j will be considered a vector in $\mathbb{R}^{p_j q}$ and not a matrix.

One of the reasons why the linear statistic T_j is particularly appropriate is that by simply changing g and h the resulting test statistic can handle any combination of continuous, categorical and ordinal variables. In other words, the choice of the function g and the influence function h depends on the type of data of X_j and Y . For instance, if X_j is continuous g is the identity and therefore $g_j(x_{ji}) = x_{ji}$ for all $i = 1, \dots, n_A$. If X_j is categorical with, lets say, 3 levels a , b and c , then $g_j(x_{ji}) = (1, 0, 0)$ if $x_{ij} = "a"$, $g_j(x_{ji}) = (0, 1, 0)$ if $x_{ij} = "b"$ and $g_j(x_{ji}) = (0, 0, 1)$ if $x_{ij} = "c"$. The same can be extended to any number of categories. If X_j is ordinal $g_j(x_{ji}) = \text{rank}(x_{ji})$. For the response, if Y is continuous, the influence function is the identity $h(y_i) = y_i$. In some situations, if there are a few extreme values of the response one might considering using the ranks, i.e. $h(y_i) = \text{rank}(y_i)$. If the response is categorical, the same transformation as for the categorical predictor can be used.

In order to perform a test based on the linear statistic T_j , the values of X_j are considered fixed and the order in which the values of Y are arranged are random. Thus, the vector (y_1, \dots, y_{n_A}) is considered a random realization of the sample space of all the possible permutations of the values $\{y_1, \dots, y_{n_A}\}$. Therefore, for each permutation of $\{y_1, \dots, y_{n_A}\}$ the value of T_j is different under the null hypothesis H_0^j . Strasser & Weber (1999) gave a closed form of the expected value μ and covariance Σ of the distribution of T_j over the sample space of all the possible permutations of $\{y_1, \dots, y_{n_A}\}$:

$$\begin{aligned} \mu &= \text{vec} \left(\left(\sum_{i=1}^{n_A} g(x_{ji}) \right) \mathbb{E}(h)^T \right) \\ \Sigma &= \frac{n}{n-1} \mathbb{V}(h) \otimes \left(\sum_{i=1}^{n_A} g(x_{ji}) \otimes g(x_{ji})^T \right) - \\ &\quad - \frac{1}{n-1} \mathbb{V}(h) \otimes \left(\sum_{i=1}^{n_A} g(x_{ji}) \right) \otimes \left(\sum_{i=1}^{n_A} g(x_{ji}) \right)^T \end{aligned}$$

where $\mathbb{E}(h) = \frac{1}{n} \sum_{i=1}^{n_A} h(y_i)$, $\mathbb{V}(h) = \frac{1}{n} \sum_{i=1}^{n_A} (h(y_i) - \mathbb{E}(h))(h(y_i) - \mathbb{E}(h))^T$ and \otimes is the Kronecker product. Notice that the dimensions of μ and Σ depend on the dimension of g and the influence function h . In the continuous and ordinal cases the dimension is 1. When, either Y or X_j are categorical, the dimension is going to change according to the number of categories. In the general case $\mathbb{E}(h) \in \mathbb{R}^q$, $\mathbb{V}(h) \in \mathbb{R}^{q \times q}$, $\mu \in \mathbb{R}^{p_j q}$ and $\Sigma \in \mathbb{R}^{p_j q \times p_j q}$ where p_j is the number of categories of X_j and q is the number of categories of Y .

Strasser & Weber (1999) studied the asymptotic properties of the linear statistic T_j and concluded that, under H_0^j , as $n \rightarrow \infty$, T_j tends to a multivariate normal distribution with mean μ and covariance Σ . The asymptotic properties of T_j can be used to construct a test statistic for the test of independence. For instance, the quadratic form

$$c_{\text{quad}} = (T_j - \mu)^T \Sigma^+ (T_j - \mu)$$

where Σ^+ is the Moore-Penrose inverse of Σ , follows approximately a χ^2 distribution with degrees of freedom given by the rank of Σ . Notice that both T_j and μ are vectors in $\mathbb{R}^{p_j q}$, and Σ^+ is a matrix in $\mathbb{R}^{p_j q \times p_j q}$, thus c_{quad} is a number obtained from the sample that can be compared to the critical value. Another example refers to the test statistic

$$c_{\text{max}} = \max \left| \frac{T_j - \mu}{\text{diag}(\Sigma)^{1/2}} \right|$$

which follows a standard normal distribution in the univariate case ($p_j = q = 1$) and, in the multivariate case, its distribution can be computed by numerical algorithms. Notice that $(T_j - \mu)/\text{diag}(\Sigma)^{1/2}$ is a vector in $\mathbb{R}^{p_j q}$ and the maximization is obtained over all the values in the vector.

The p-values of the test can be approximated by either, using the asymptotic distribution of c_{quad} or c_{max} , or approximating the exact distribution of c_{quad} or c_{max} using re-sampling methods (getting a random sample of permutations of (y_1, \dots, y_{n_A})).

3.4.1 Survival trees based on unbiased recursive partitioning

The method just described can easily be extended to survival data by choosing the appropriate influence function h . In the case of survival data with right censoring h is chosen to be the logrank scores as in Hothorn & Lausen (2003). Let τ be any node in a tree T and let $\{(x_i, \delta_i) : i = 1, \dots, n_\tau\}$ be the survival response belonging to node τ . The logrank score for observation i is defined as

$$\delta_i - \sum_{(x_j, \delta_j): x_j \leq x_i} \frac{\delta_j}{n_\tau - \theta_j + 1}$$

where θ_j is the number of observations $x_l \leq x_j$.

Figure 3.12 shows the survival tree obtained using unbiased recursive partitioning when applied to the coronary data. This tree was grown using the contributed package in R *party* (Hothorn *et al.*, 2006). This model identified Age, PreviousMI and ACE as significant factors, which are three of the four predictors that the

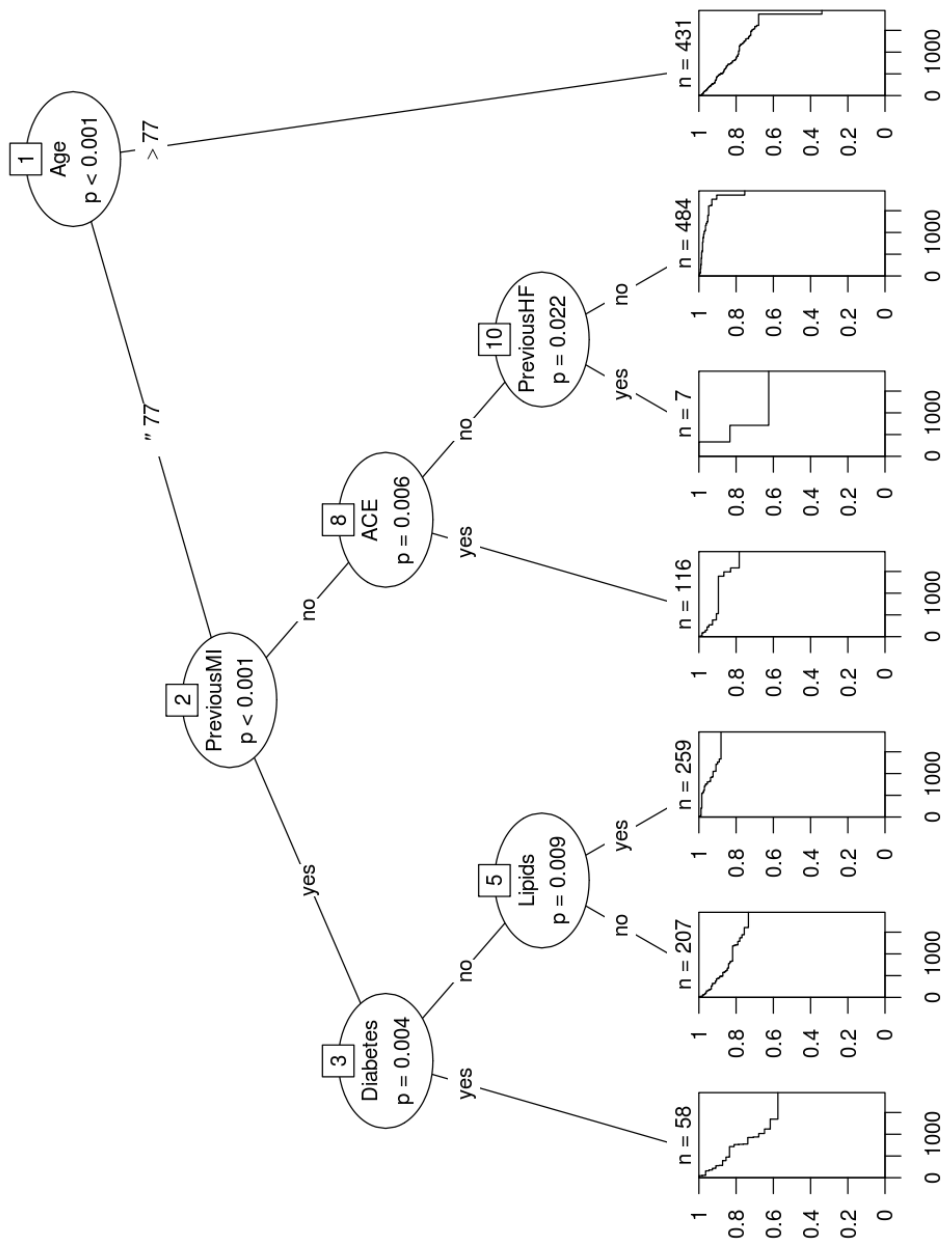


Figure 3.12: Unbiased recursive partitioning applied to the coronary data.

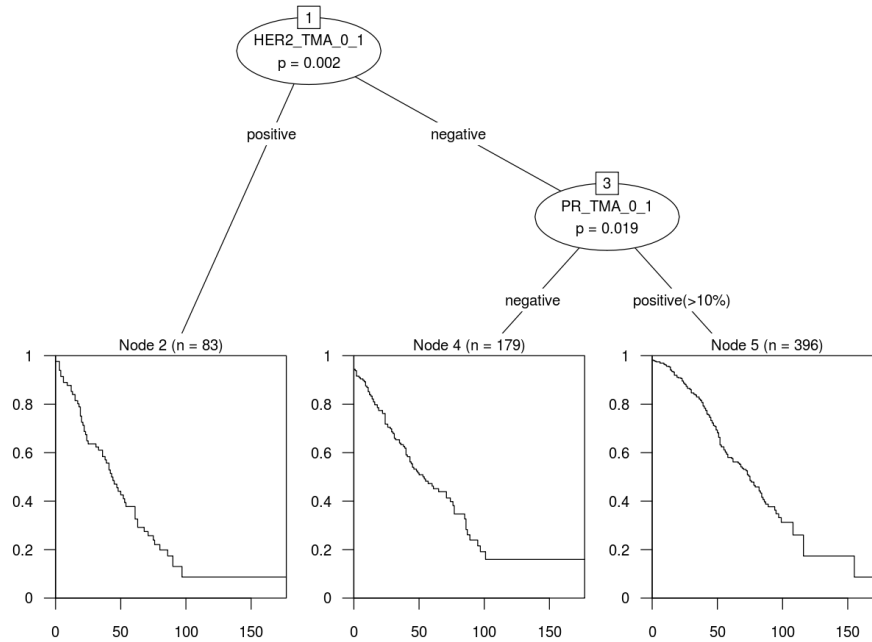


Figure 3.13: Unbiased recursive partitioning applied to the breast cancer dataset. The event of interest is the recurrence of the disease.

Cox proportional hazard model also identified as significant. The survival tree also found Diabetes, Lipids and PreviousHF as having an effect on the survival outcome. The worst prognostic groups correspond to the far left and far right terminal nodes: these are patients younger than 77 years old, with previous myocardial infarction and diabetes, and patients older than 77 years old respectively. The best prognostic cohort of patients are individuals younger than 77 years old with no previous myocardial infarction or heart failure and who are not on ACE medication (second terminal node from the right).

Another example of a survival tree using unbiased recursive partitioning is given in Figure 3.13 where the method was applied to the breast cancer dataset. The event of interest was the recurrence of the disease. The tree identified 2 of the 11 biomarkers used to generate the model as having a significant effect, i.e. HER2 and PR. These differ from those obtained in the previous Chapter, where variable selection identified Bcl2 as the only relevant predictor to be included in a proportional hazard model. The plot suggests that women who are HER2 positive have the worst prognosis (left terminal node) followed by women who are HER2 negative and PR negative (middle terminal node) and women who are HER2 negative and PR positive.

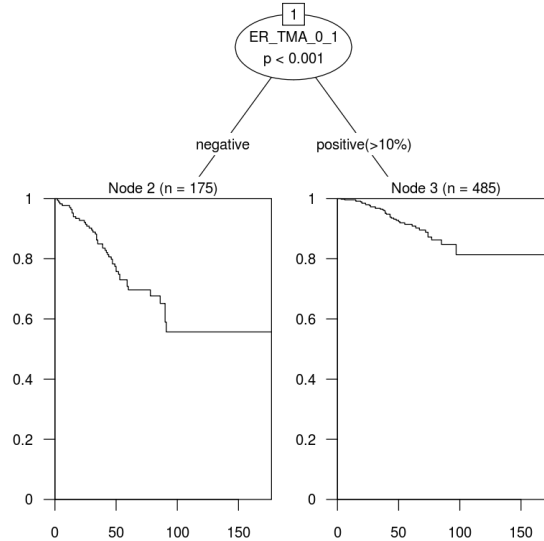


Figure 3.14: Unbiased recursive partitioning applied to the breast cancer dataset. The event of interest is the death of the patient.

Figure 3.14 depicts the corresponding survival tree when the event of interest is death. The tree only identified ER as having an effect in the survival outcome. This covariate was also identified when proportional hazard models were fitted. The tree shows that women who are ER negative have poorer prognosis than women who are ER positive. This apparent contradiction (one might expect that ER negative patients should have a better survival outcome) can be explained by the fact that women who test positive for ER are more likely to benefit from the established treatments and usually have a greater five-year overall survival than those who test negative.

3.5 Random forest

Although the main focus of this work is on individual trees, for the sake of completeness a brief description of ensemble methods based on trees is presented in this section. One of the motivations for the use of these types of methods is the fact that classification and regression trees, along with other model selection procedures such as subset selection in linear regression are unstable (as pointed out by Breiman, 1996b). This means that small changes in the learning set can result in large changes in the model used to predict the outcome of interest.

Breiman (1996a) proposed a method to improve the prediction error called **bagging** (**bootstrap aggregating**). The method is based on the use of bootstrapping to generate a set of prediction models. Let $\mathcal{L} = \{(y_i, \mathbf{x}_i) : i = 1, \dots, n\}$ be a random sample or training set where y is the outcome of interest and $\mathbf{x} = (x_1, \dots, x_m)$ is the set of covariates. Furthermore, let $\varphi(\mathbf{x}, \mathcal{L})$ be a predictor of \mathbf{x} based on the learning sample \mathcal{L} . In bagging, instead of using only one learning set \mathcal{L} for prediction, a set of learning sets $\{\mathcal{L}^{(B)}, B = 1, \dots, k\}$ is generated by taking repeated bootstrap samples from \mathcal{L} . The predicted outcome $\varphi_B(\mathbf{x})$ is the average of $\varphi(\mathbf{x}, \mathcal{L}^{(B)})$ over all the replicates B if the response is continuous and it is the mode over all the replicates B if the response is categorical.

Breiman applied this method to regression trees and found a reduction in the test set mean square error on datasets ranging from 21% to 46%. The same method was tested for classification trees and gave a reduction in test set misclassification rates ranging from 6% to 77%.

A slightly different version of bagging was proposed by Breiman (2001) and was called a **random forest**. The algorithm can be summarized in the following steps:

1. draw a bootstrap sample of the original dataset;
2. using recursive partitioning grow a large or saturated tree. At each node, consider only a random selection of predictors for the split;
3. repeat steps 1) and 2) until a forest of trees have been generated;
4. use this ensemble of trees for prediction in the same way as the bagging approach.

Breiman showed that the accuracy of a random forest depends on two key factors: the prediction strength of the individual trees and the correlation of the trees. He also proved that random forests do not over-fit the data even though the trees in the ensemble are not pruned. Note that the only difference between random forests and bagging is step 3) where only a subset of randomly selected predictors is considered for the splits. Therefore randomization is introduced in two forms, by selecting bootstrap replicates of the original data and by selecting at random a smaller set of covariates at each node.

3.5.1 Random survival forest

Extensions of the random forest approach to the study of survival data are problematic due to the fact that the predicted outcome in survival analysis is not as well defined as in regression and classification problems. One way of approaching this problem is by reformulating the survival tree in terms of a classification

problem by exploiting the equivalence between the likelihood of a tree for Poisson responses and the likelihood of a tree based on the proportional hazard assumption (Ishwaran *et al.*, 2004). The predicted outcome in that case is the estimated hazard ratio. Hothorn *et al.* (2006) proposed a different approach where log transformed survival times are used as the outcome in a weighted random forest regression analysis. The observations are weighted by, what they call, the inverse probability of censoring.

More recently, Ishwaran *et al.* (2008) has proposed a random survival forest specifically designed for survival data without the need of any transformation of the survival times. This approach is based on a new type of predicted outcome for survival data called ensemble mortality. In order to assess the prediction error, the concordance index (C-index) proposed by Harrell *et al.* (1982) is considered. The algorithm can be summarized in the following steps:

1. draw a bootstrap sample of the original data. Call **out-of-Bag (OOB)** the set of observations that are not included in the bootstrap replicate (on average 37% of the data will not be included in each replicate);
2. grow a saturated tree by choosing at each node only a random selection of predictors. The splitting criterion is based on some measure of between node separation (such as the logrank statistic);
3. repeat 1) and 2) until a forest of trees have been generated.
4. calculate the cumulative hazard function Λ_T for each tree T and average that over all the replicates to obtain the ensemble cumulative hazard function Λ_T^B ;
5. Calculate the prediction error for Λ_T^B using the OOB data;

To explain in more detail the elements of the algorithm, let τ be a node in any tree T . The cumulative hazard function at node τ is given by the Nelson-Aalen estimate (Nelson, 1972; Aalen, 1978)

$$\widehat{\Lambda}_\tau(t) = \sum_{i: x_i \leq t} \frac{d_i}{n_i}$$

where d_i is the number of deaths at time x_i , n_i is the number of individuals at risk and x_i represents the observed times. Let $\{T^{(B)} : B = 1, \dots, k\}$ be the forest of trees created in the ensemble. For any observation i , the cumulative hazard function of tree $T^{(B)}$ at any time t is

$$\Lambda^{(B)}(t|i) = \widehat{\Lambda}_{\tau_i^{(B)}}(t)$$

where $\tau_i^{(B)}$ is the terminal node in which observation i falls in tree B . The bootstrap cumulative hazard function for individual i at time t is defined as

$$\Lambda^*(t|i) = \sum_{B=1}^k \widehat{\Lambda}_{\tau_i^{(B)}}(t).$$

Alternatively, the OOB cumulative hazard function $\Lambda^{**}(t|i)$ is similar to $\Lambda^*(t|i)$ but in the former case the sum is only over the trees in the forest for which individual i is an OOB observation.

The predicted outcome for individual i is the **ensemble mortality** which is defined as

$$M_i^* = \sum_{j=1}^n \Lambda^*(x_j|i)$$

where $\{x_j : j = 1, \dots, n\}$ is the set of observed times in the random sample. Similarly, the **OOB ensemble mortality** is defined as

$$M_i^{**} = \sum_{j=1}^n \Lambda^{**}(x_j|i). \quad (3.9)$$

To estimate the prediction error, the C-index proposed by Harrell *et al.* (1982) is considered. Let $\{(x_i, \delta_i) : i = 1, \dots, n\}$ be the observed data where x is the observed time and δ is the censoring indicator. The concordance index is calculated as follows:

1. define the set of permissible pairs as the set of pairs (i, j) for which the shorter observed time is uncensored. If $x_i = x_j$ then (i, j) is a permissible pair only if, at least, one of them is uncensored;
2. For each permissible pair,
 - If $(x_i \neq x_j)$ count 1 if the shorter survival time has worse predicted outcome and count 0.5 if both have the same predicted outcome.
 - If $x_i = x_j$ and both are events count 1 if both have the same predicted outcome and count 0.5 otherwise.
 - if $x_i = x_j$ but not both are uncensored, count 1 if the uncensored observation has worse predicted outcome and 0.5 otherwise.
 - Let Concordance denote the sum over all the permissible pairs.
3. Calculate the concordance index as

$$C = \frac{\text{Concordance}}{\text{Permissible}}.$$

The predicted outcome is given by the OOB ensemble mortality defined in (3.9) and, therefore, an OOB estimate of the C-index C^{**} can be obtained. The OOB prediction error is calculated as $1 - C^{**}$.

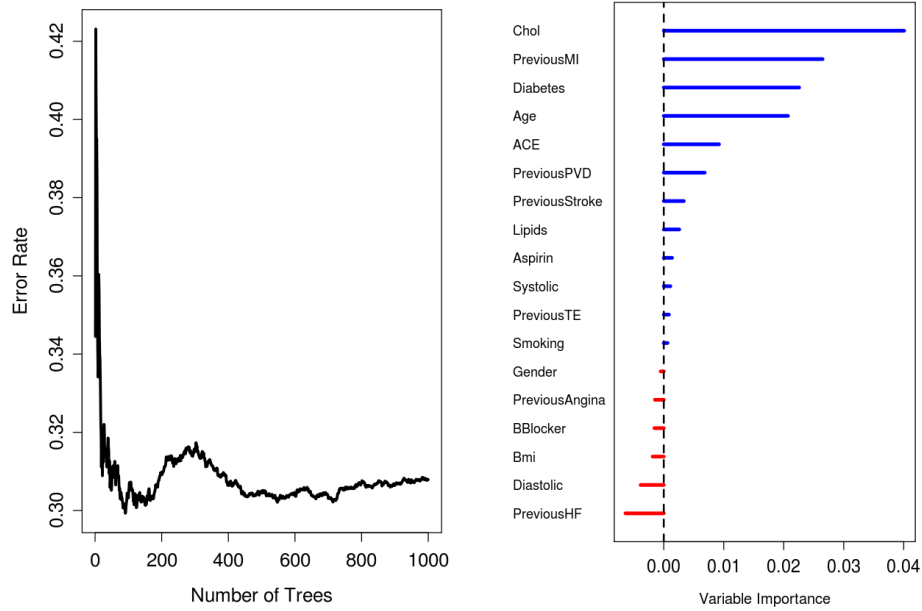
It is important to stress that random forests, in particular random survival forests, do not produce an individual tree as an outcome. One way of summarizing the combined information obtained by the forest of trees is to use a value called **variable importance** for each one of the predictors. To calculate the variable importance for a particular predictor x , the OOB ensemble mortality for an individual i is calculated in the exact same way described before, with the only difference that each time observation i encounters a split based on x the daughter node is assigned at random instead following the path described by the split. The variable importance is then calculated as the difference between the original prediction error and the prediction error calculated in this manner.

This method was applied to the coronary data and the results are shown in Figure 3.15 (a). On the left the OOB error rate as a function of the number of trees in the forest. On the right the ‘variable importance’ of each predictor. As one can see 4 out of 5 of the covariates ranked as having the highest values of ‘variable importance’ are identical to those obtained in the Cox proportional hazard model, i.e. Chol, PreviousMI, Age and ACE. When comparing the results of the random forest with the tree obtained using unbiased recursive partitioning, again one can see how most of the predictors identified by the tree have high values of ‘variable importance’ (Age, PreviousMI, ACE, Diabetes and Lipids) with the exception of PreviousHF which was ranked last.

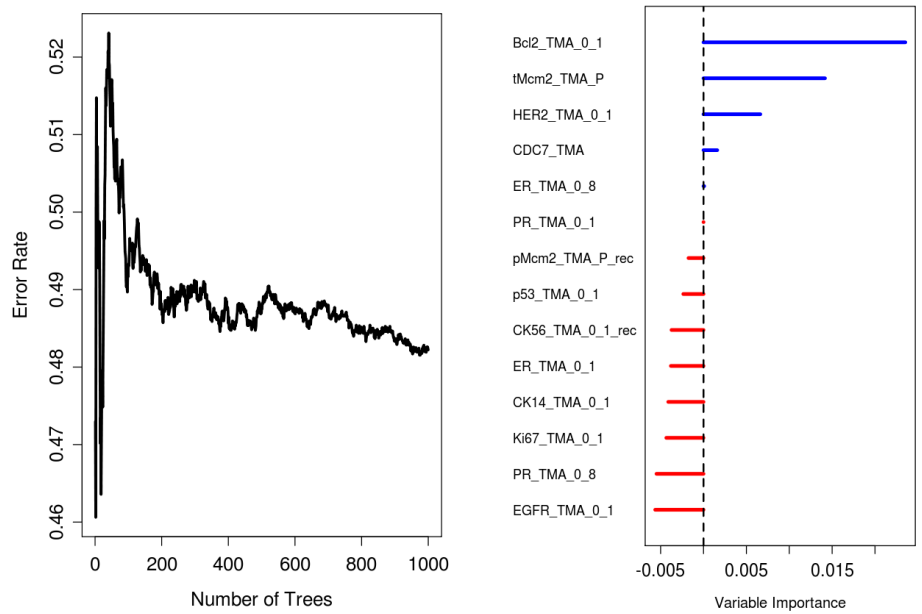
Another example is given in Figure 3.15 (b) where the random survival forest algorithm was applied to the breast cancer data when the event of interest is the recurrence of the disease. Bcl2, which was the only predictor identified by the Cox proportional hazard model, has the highest ‘variable importance’ value followed by tMcm2 and HER2. Recall that HER2 was one of the covariates identified by the tree based on unbiased recursive partitioning.

3.6 Chapter conclusion

In this chapter an extensive review of tree based methods has been presented. These methods can be used as an alternative to the generalized linear models for continuous, categorical, count and survival responses. The CART algorithm has been explained in detail and extensions to Poisson responses have also been presented. A whole section was dedicated to tree based methods applied survival responses. Some of these methods adopt the pruning procedure of the CART algorithm whereas others developed their own pruning strategies. However, trees that use recursive partitioning are affected by the variable selection bias. A possible



(a)



(b)

Figure 3.15: (a) Random survival forest applied to the coronary data. (b) Random survival forest applied to the breast cancer dataset (with recurrence as event).

solution for this problem is based on trees that use unbiased recursive partitioning. A section was dedicated to the explanation of these types of trees. Throughout the Chapter, the breast cancer dataset and the coronary dataset were used to illustrate some of the methods and to compare the results with those obtained in the previous Chapter. In the last part of this Chapter an overview of ensemble methods, in particular those applied to survival analysis, was presented.

The next Chapter includes a novel approach for growing survival trees that can be used as an alternative to trees based on unbiased recursive partitioning. The method is based on a novel algorithm called ‘node re-sampling’ which uses re-sampling procedures each time a split has to be performed. This new method addresses the problem of variable selection bias, it can identify interaction effects, and it takes into account the sampling variability in the process of generating the model. The aim of this new method is to provide a more robust and a more versatile tool to build survival trees.

Chapter 4

Trees based on node re-sampling

In this Chapter a novel algorithm for growing trees is introduced which uses re-sampling procedures at node level to obtain the optimal split at each node. The Chapter begins with an introduction that motivates the search for an alternative method to generate classification and regression trees and, particularly survival trees. In the second part of the Chapter, a brief section is presented which highlights the fact that the representation of a model in a tree fashion is not unique and, in reality, many different representations are valid. Taking this into account, the notion that the importance of a predictor is related to the position on the tree can be dismissed. In the third part a brief introduction to the node re-sampling algorithm is given. Initially, only continuous responses are considered and a preliminary version of the algorithm is studied in detail. The fourth part explains the final version of the algorithm and presents a graphical user interface for survival responses. In this section the whole process of growing the survival tree based on the node re-sampling algorithm is presented. One of the most interesting features of this new approach is the fact that interactions are easily detected. This will be demonstrated with a theoretical example and the results will be compared with unbiased recursive partitioning which fail to detect interaction effects. Finally, in the last part of the Chapter, the new method will be applied to the datasets presented in the introduction and the results will be discussed and compared to those obtained by other methods.

4.1 Introduction

There are many reasons why tree based methods have enjoyed enormous popularity since they first appeared. One of these reasons is the simplicity of the generated model and the possibility of representing the results of the statistical analysis in a tree fashion. This feature is particularly appealing for people with no statistical background. For instance, a researcher who has collected data and

has no experience in statistical modeling can easily run the recursive partitioning algorithm without the need of specifying a model for the response or transforming the scale of any of the variables involved in the study. Furthermore, the representation of the model in a tree fashion offers a very intuitive tool to interpret the results of the statistical analysis. Another reason for the popularity of tree based methods, which has more to do with statistical modeling, is that they can easily deal with interactions. This point is particularly true for tree based methods that use some form of pruning after a saturated tree has been grown (as will be demonstrated in this chapter). In this regard, interactions should naturally arise in the model without the need of specifying additional terms as in the case of regression models. Other reasons that make tree based methods interesting include the ease in which the model deals with missing values, through the use of surrogate splits, and the robustness of the method to the presence of outliers. In any case, tree based methods are accessible to a broader group of users and no particular expertise is necessary to understand and manage the statistical concepts involving the creation and interpretation of a tree.

Although all these appealing properties might appear to be an excellent reason to use this type of models, the recursive partitioning algorithm has some drawbacks. The most common criticism arising among statisticians has to do with the lack of any statistical inference of tree based methods. Apart from trees based on conditional inference procedures (described in the previous chapter) or any other trees that use a statistical test for the selection of predictors at each node, there is no information about how the sampling variability will affect the results of the obtained model. In this sense, tree based methods are seen as an exploratory tool which complements other algorithms for variable selection such as, backward elimination, forward selection or best subsets. It turns out that individual trees are very sensitive to sample variability as will be demonstrated in this chapter. The structure of the whole model depends dramatically on the data used to generate the tree and small changes in the data could lead to trees with a completely different appearance. This can be confusing for the inexperienced user and the current algorithms for growing and displaying trees might be playing a role by giving misleading information about the real nature of the model presented.

On the other hand, ensemble methods, in particular the random forest approach by Breiman (2001), have a stronger theoretical statistical basis but are criticized for the lack of an individual tree as an output that somehow diminishes the attractiveness of the method. The content of this chapter intends to fill this gap between the absence of the measurement of the sampling variability of individual trees and the lack of an individual tree as an outcome of ensemble methods. In order to do so, a novel method for growing trees is proposed which is based on re-sampling procedures performed at a node level. This new approach will allow the user to assess how small changes in the data could affect the structure

of the observed tree. This new method has been developed in parallel with the development of a new graphical user interface for growing survival trees. The new tool allows the user to generate the optimal tree by interactively eliminating nodes that are irrelevant for the model. The method presented here addresses some of the issues that are known to be problematic with the current algorithms for growing survival trees.

4.2 Different trees, same structure

One of the main misunderstandings when interpreting a tree is the notion that variables that are on top of the tree are somehow more important than variables that are at lower levels. To explain this point, suppose that the response is continuous and that there are 3 predictors: X_1 is continuous, X_2 is categorical with two levels ‘yes’, ‘no’ and X_3 is ordinal with levels ‘Low’, ‘Medium’ and ‘High’. Figure 4.1 shows 4 different representations of the same structure. The values of the terminal nodes corresponds to the predicted value of the response. All the trees displayed give the exact same information and any observation with particular values of X_1 , X_2 and X_3 will obtain the same predicted value no matter which one of the trees is chosen. As one can see, the root node in 3 of the 4 trees represented is different. The top left tree has X_1 in the root whereas the bottom right tree has X_3 in the root. The other two trees have X_2 in the root. Suppose that the top right tree is the tree obtained after running the recursive partitioning algorithm. That means that predictor X_2 produces the optimal split on top of the tree. However, other suboptimal splits could also be valid and, in fact, generate trees that give the same information. Moreover, some of these alternative trees might be simpler as in the case of the top right tree which has fewer terminal nodes. Because of the recursive nature of the algorithm, once a split has been generated, it cannot be modified and the split is only based on the information obtained in the splits generated before. Furthermore, there is no information about any future splits. As a result, the recursive partitioning algorithm will only create one of the many possible representations of the tree, without being necessarily the optimal one in terms of complexity. It is important to remark that these considerations are not related to the pruning procedure of CART (Breiman *et al.*, 1984) and they are valid for any procedure that uses recursive partitioning (as in the case of trees generated using unbiased recursive partitioning Hothorn *et al.*, 2006). The pruning of the tree is based on the existence of a saturated model which does not change apart from the elimination of some of the branches.

The important message is that the position of a variable in the tree cannot be a measurement of variable importance. The same applies to the use of p-values with the same objective. This point is easily illustrated when one considers an

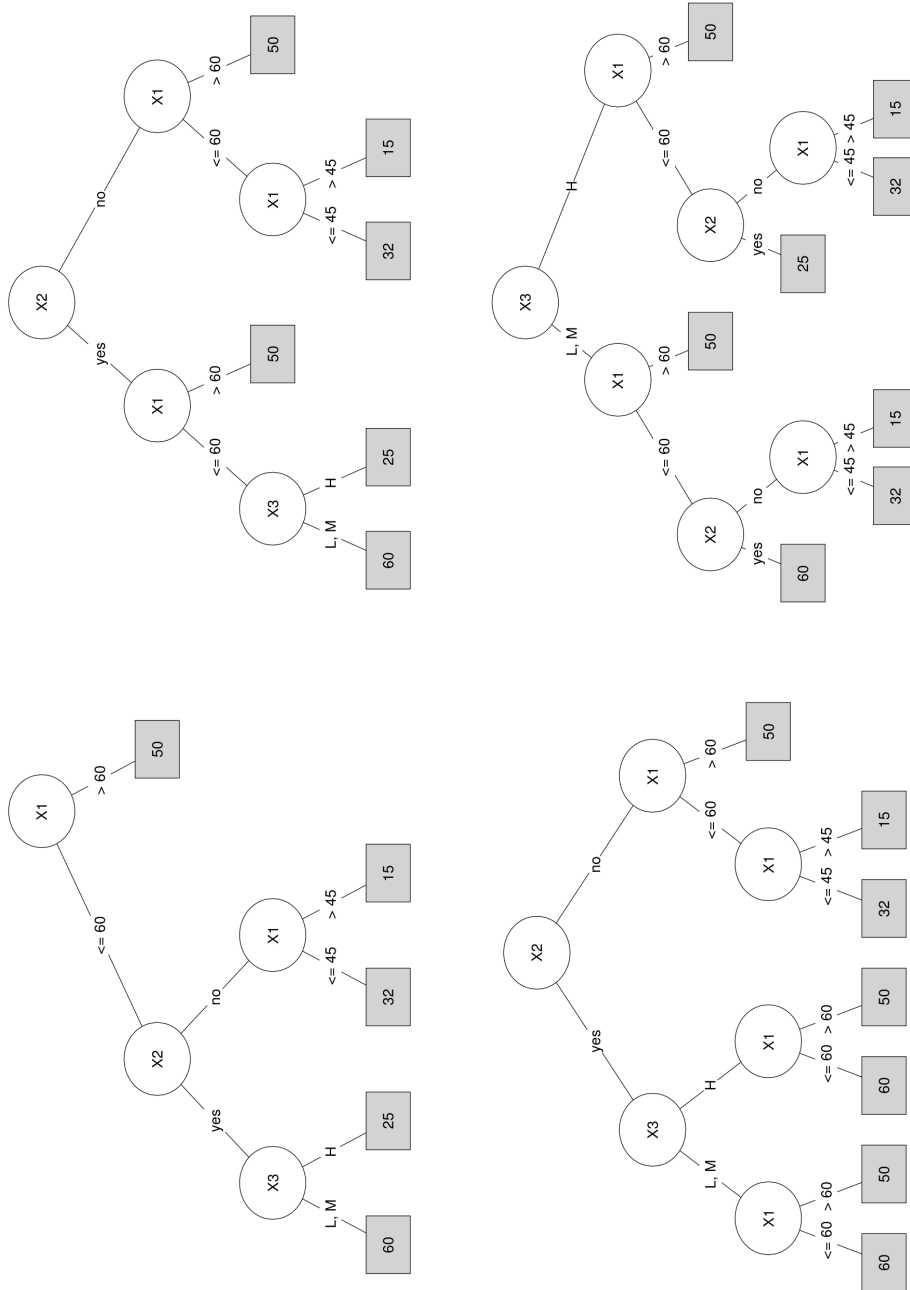


Figure 4.1: An example in which 4 different structures generate virtually the same information. This highlights the point that there is no unique way of representing a tree.

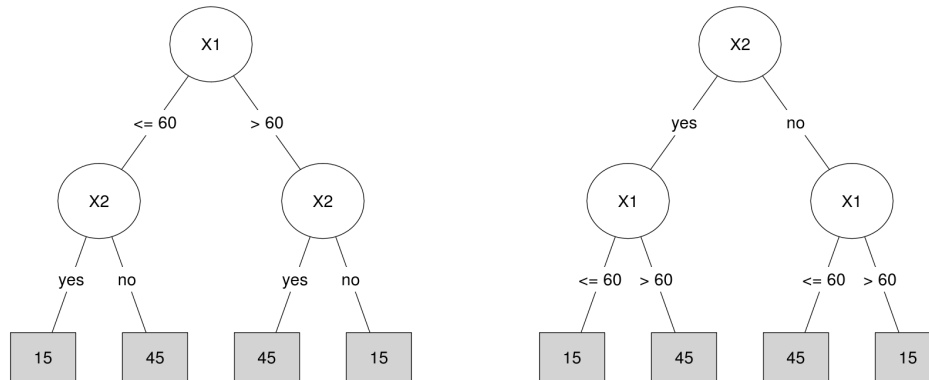


Figure 4.2: Two representations of an interaction effect.

interaction effect in the data. Figure 4.2 shows the two possible representations of an interaction between two predictors X_1 and X_2 . The terminal nodes contain the predictive value of a continuous response. In the left tree, the split in X_1 is not going to be significant since the two siblings will have the same distribution (with mean 30, assuming equal numbers in each node). However, X_2 will be significant conditioning to the split generated by X_1 . The tree on the right, which is equivalent to the tree on the left, will produce the exact opposite result. In that case, X_2 is not significant whereas the splits generated by X_1 are significant. Therefore, the use of p-values as a measurement of the importance of the predictor is relative and depends on the position occupied by that predictor in the tree. Since there are many different valid representations of the same tree, the use of p-values for such a purpose is ‘risky’ to say the least.

In summary, for a given dataset, the representation of the tree generated applying any of the methods that use recursive partition is not unique and many alternative representations are also valid. Before going into more detail about the node re-sampling algorithm an example is presented where simulated data are used to describe how the available methods perform when data are simulated from a given model.

4.2.1 Synthetic example 1

Suppose that the tree represented in Figure 4.3 defines the underlying model from which data can be simulated. In the terminal nodes, the n_s are the estimated number of observations included in the terminal nodes for a sample size of 200. Below n there are the mean and the standard deviation of the the distribution of

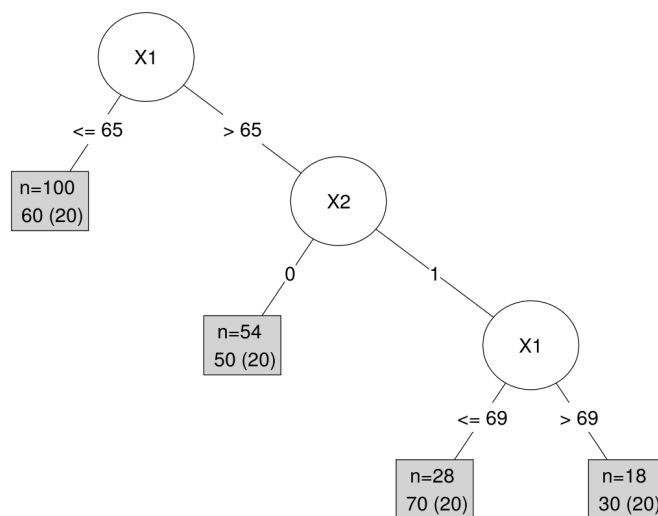


Figure 4.3: An example of a model represented in a tree fashion. In the terminal nodes the distribution of the response is assumed to be normal with means 60, 50, 70 and 30 and standard deviations of 20.

the response which is assumed to be normally distributed. In this example, data will be generated as follows. There will be 12 predictors in total with distributions given in Table 4.1. The first two predictors X_1 and X_2 will be related to the response in the way represented in Figure 4.3. The third predictor X_3 is related to X_1 with correlation coefficient $\rho = 0.7$ and, therefore, it will be related to the response indirectly. The rest of the predictors are pure noise with no relation whatsoever to the response.

The process of generating a random sample consists in, first obtaining the values of the predictors by drawing data from the distribution corresponding to each one of the predictors (as specified in Table 4.1). Once this is done, the rows corresponding to each one of the terminal nodes can be identified by dropping each observation down the tree in Figure 4.3. Finally, the values of the response are generated using the distribution specified in each one of the terminal nodes.

Once a random sample has been generated, the algorithms to grow the tree can be run and the results can be compared to the underlying model to see to what extent the estimated model reflects the underlying model. For instance, Figure 4.4 shows the results of running the unbiased recursive partitioning algorithm (Hothorn et al. 2006) after a random sample of size 200 was drawn from the model. As one can see, the optimal tree has no splits at all. Based on the

| Predictor | Distribution | Type |
|--------------------|--|-------------|
| X_1 | Normal ($\mu = 65, \sigma = 4.5$) | Continuous |
| X_2 | Bernoulli ($p = 0.45$) | Categorical |
| X_3 | Normal ($\mu = 70, \sigma = 4.5$) | Continuous |
| noise ₁ | Normal ($\mu = 30, \sigma = 10$) | Continuous |
| noise ₂ | {1, 2, 3, 4} | Categorical |
| | $p_1 = 0.20, p_2 = 0.40, p_3 = 0.10, p_4 = 0.30$ | |
| noise ₃ | Normal ($\mu = 70, \sigma = 10$) | Continuous |
| noise ₄ | Normal ($\mu = -2, \sigma = 10$) | Continuous |
| noise ₅ | Bernoulli ($p = 0.45$) | Categorical |
| noise ₆ | Bernoulli ($p = 0.45$) | Categorical |
| noise ₇ | {1, 2, 3} | Categorical |
| | $p_1 = 0.20, p_2 = 0.40, p_3 = 0.40$ | |
| noise ₈ | Bernoulli ($p = 0.45$) | Categorical |
| noise ₉ | Bernoulli ($p = 0.45$) | Categorical |

Table 4.1: Distributions of the predictors involved in Example 1. Although it is not mentioned in the table, predictors X_1 and X_3 are correlated with correlation coefficient of 0.7.

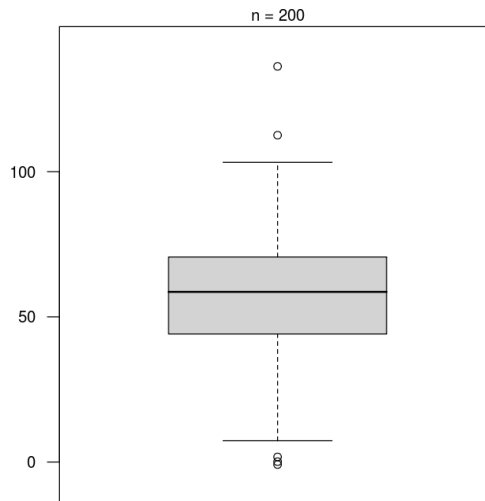


Figure 4.4: Output obtained using unbiased recursive partitioning after a random sample of 200 was drawn from the model presented in Figure 4.3 ($\alpha = 0.05$).

| Splits | \bar{y} | s | n | 95%CI dif | p-value |
|---------------|-----------|------|-----|--------------|---------|
| $X_1 \leq 65$ | 58.1 | 18.2 | 100 | (-13.3,5.8) | 0.216 |
| $X_1 > 65$ | 54.3 | 24.4 | 100 | | |
| $X_2 = 1$ | 58.4 | 26.0 | 85 | (-14.1,6.4) | 0.239 |
| $X_2 = 0$ | 54.6 | 17.5 | 115 | | |
| $X_1 \leq 69$ | 58.5 | 20.4 | 172 | (-32.4,-0.9) | 0.001 |
| $X_1 > 69$ | 41.9 | 23.5 | 28 | | |

Table 4.2: Summary statistics of all 3 splits defined in Figure 4.3 after a random sample of size 200 was drawn from the underlying model.

algorithm, there is no relationship between the predictors and the response. The global null hypothesis of independence is not significant and, therefore, the tree stops growing before generating any splits. In this regard, the method fails to provide a good representation of the underlying model. This could be due to the fact that the variability of the data does not allow us to detect any relationship. To find out if this is the reason, a simple t-test for each one of the possible splits, along with the corresponding summary statistics and p-values, are presented in Table 4.2. The Table shows that the only significant split corresponds to $X_1 \leq 69$. However, the global null hypothesis of independence was not significant when unbiased recursive partitioning was used. Another possible explanation is that there are too many predictors. Because the global test of independence is based on individual tests, some kind of correction must be introduced to the level of significance to achieve the desired global significance. As a result of this, the more predictors there are in the model the more conservative the test will be, and the harder will be for the test to detect any relationship.

To investigate if this is the case, the level of significance was increased to allow the method to obtain a larger tree. If the model was generated with all the predictors in the model, i.e. X_1 , X_2 , X_3 and the 9 covariates that are pure noise, the method was still unable to identify any splits for an α of 0.10. However, at the same level of significance, the model with only X_1 , X_2 and X_3 produced the output depicted in Figure 4.5. As one can see, the method returns the adequate model when less variables were used to grow the tree and, therefore, one can conclude that the number of covariates had an effect on the size of the tree.

To compare the results obtained with other methods, the recursive partitioning algorithm using CART (Breiman *et al.*, 1984) was run with the exact same random sample. Figure 4.6 shows part of the large tree generated. By looking at the complexity parameter plot (Figure 4.7) one can see that, again, the tree with no splits is the optimal tree. Therefore, based on the CART algorithm, there is no relationship between any of the predictors and the response.

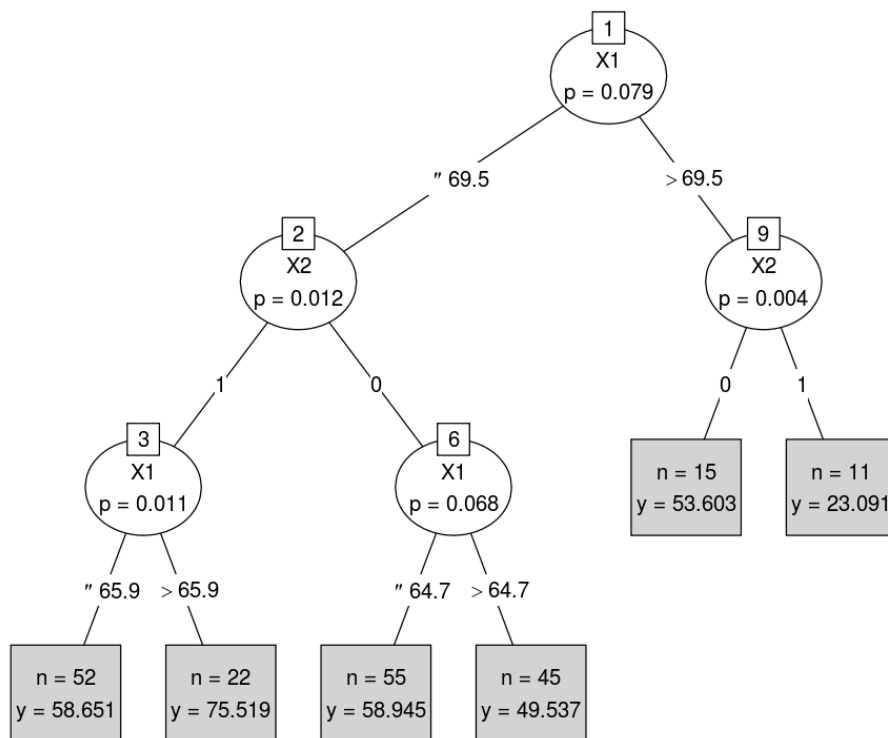


Figure 4.5: Output obtained using unbiased recursive partitioning after a random sample of 200 was drawn from the model presented in Figure 4.3 without noisy covariates ($\alpha = 0.10$).

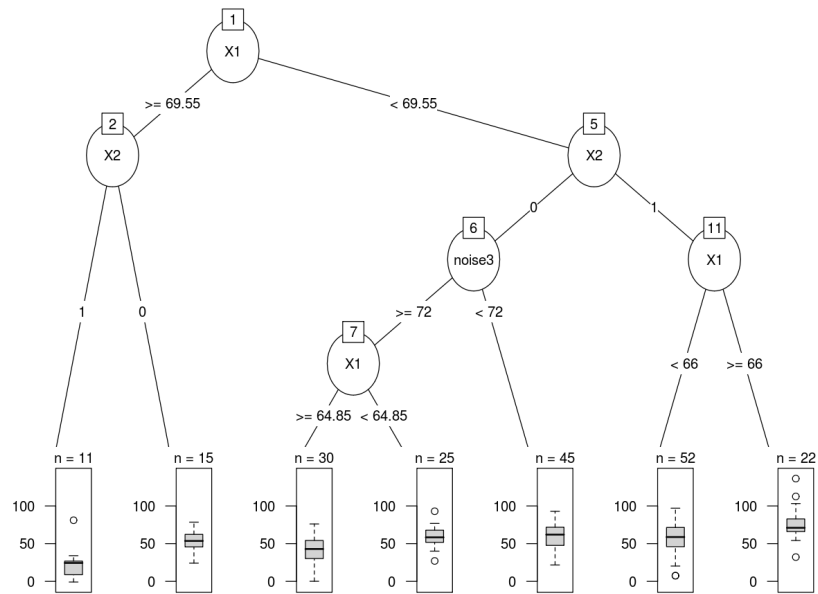


Figure 4.6: Saturated tree based on CART. The data used to grow the tree was exactly the same than the data used to grow the tree in Figure 4.4.

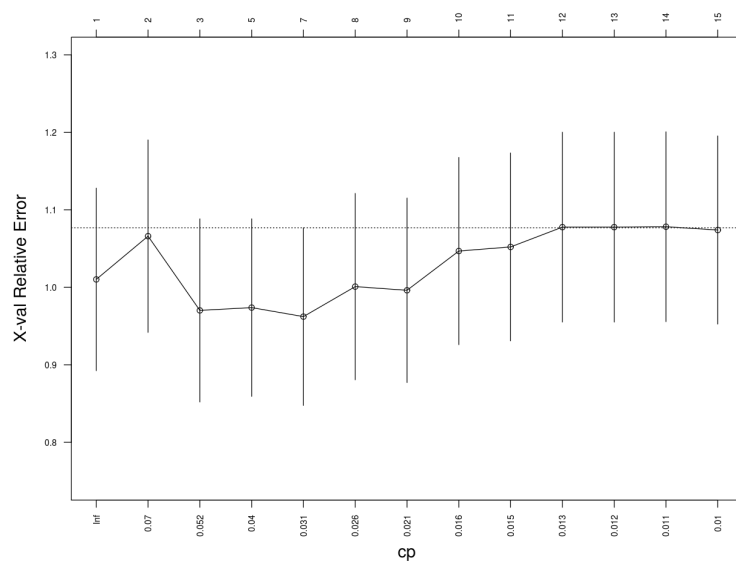


Figure 4.7: Complexity parameter plot. The minimum is attained for the tree of size 7 but any other tree is within the 1SE of the achieved minimum and, therefore, the optimal tree is the tree with no splits.

To summarize the results obtained with these simulated data, both methods, the unbiased recursive partitioning and the CART algorithm, seem to be conservative in this example and the models obtained are smaller than the model used to simulate the data. In both cases, the observed tree is based on a particular random sample of an underlying joint distribution. Different samples, however, might generate different trees (due to the sampling variability) which in turn will have many different representations. The next section aims to address the issue of how small changes of the training data will affect each one of the splits generated in each node.

4.3 Node re-sampling: Continuous responses

One of the most confusing features of tree based methods, especially for the non-experienced user, is the fact that small changes in the training data could result in major changes in the structure of the tree. These changes are generated at each node and each node represents a split that has been generated using the data from the random sample. The current methods for growing trees do not give any information about how the sampling variability might affect the selection of that particular split. The content of this section aims to provide a novel method to address this particular issue. Instead of making the decision of choosing a split based only on one dataset, the method will use the information obtained after many bootstrap replicates have been generated from the available data. By doing this, the selected split will inherently be more robust and information will be gained in terms of the variability of the cutpoints and the selection of meaningful surrogate splits.

To introduce the idea of node re-sampling, consider the same underlying model described in the example given in the previous section. To obtain the first node (on top of the tree) the recursive partitioning algorithm runs through all the predictors and all the possible cutpoints and chooses the predictor and the cutpoint that maximizes the change in impurity. Figure 4.8 shows the results in terms of the optimal change in impurity for all the predictors involved in the example. This plot is interesting for different reasons. First of all, it shows clearly that X_1 is the best predictor in terms of the change in impurity. Recall that the impurity is measured in regression problems using the sum of squares. It is interesting to see that the second best predictor (which will be considered the first surrogate) is noise_4 which has no relationship with the response. The same applies to the second surrogate noise_1 . The next surrogates are noise_7 , X_3 and noise_3 which are very close to each other. Another interesting comment that can be drawn from the plot is that one can see the results of variable selection bias. Predictors with many cutpoints are more likely to be selected for the split. This is very clear if

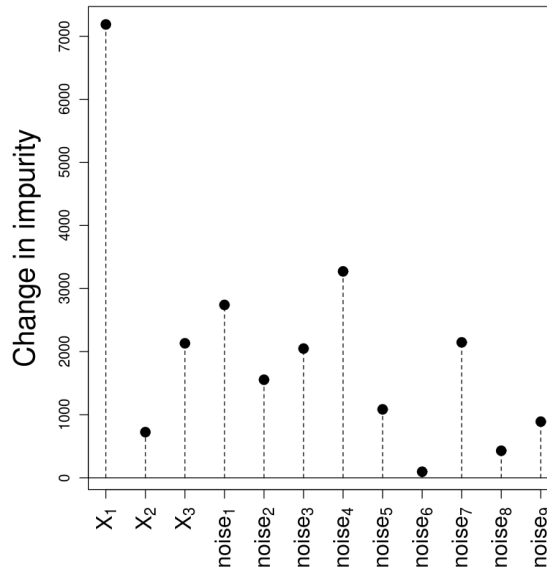


Figure 4.8: An example of how recursive partitioning works at one particular node. For each predictor the algorithm records the maximum change in impurity (Y axis).

one analyzes the set of variables representing pure noise. Those variables that are continuous produce higher values in terms of the change of impurity whereas factors with only two levels produce the lowest values.

From the point of view of making a decision about which split should be placed in the root node, the information derived from Figure 4.8 might not suffice, especially in relation to the surrogate splits. What would happen if some changes are introduced to the training set? What would happen if another random sample was drawn from the underlying model? Would one get the same predictors for the primary and surrogate splits? The obvious answer is that it is very likely that the results would vary. The idea of node re-sampling is to inject some variability in the training data and to base the decision about which split should be chosen for a particular node on the information obtained when this process is repeated many times. In the process of doing this one can also gather information about the variability of the splitting points and the candidates for surrogate splits.

The next step is to define the way in which the results obtained after bootstrapping the original data are going to be used to select the split in any particular node. One possible approach is to apply the following algorithm:

1. Draw N bootstrap replicates of the original data.

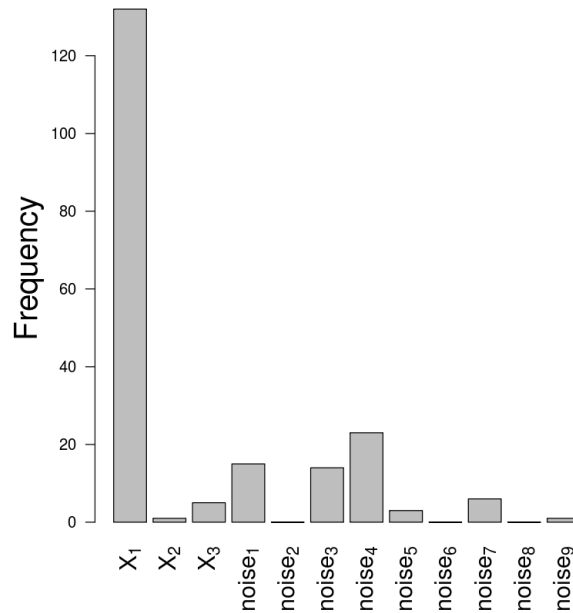


Figure 4.9: Bar-chart produced by the node re-sampling algorithm.

2. For each replicate, select the predictor for which the change in impurity is maximized and retain the information about the splitting points for all the predictors.
3. After all the replicates have been obtained, use the frequencies of the optimal predictors obtained in step 2) to determine the predictors for the primary and surrogate splits.
4. Select the typical value of the bootstrap distribution of the cutpoints of the primary split as the selected splitting point.

To illustrate the way this algorithm works suppose that the data used to generate Figure 4.8 is re-sampled with replacement 200 times. Figure 4.9 shows the bar-chart corresponding to the distribution of the predictors that were selected as the best candidates in each one of the replicates. As one can see, X_1 is the clear winner corroborating the impression that this predictor is related to the response and the best candidate for the primary split. However, in approximately 35% of the replicates other predictors were chosen. Using the information provided in the graph, one could consider noise_4 as the first surrogate, noise_1 as the second surrogate and noise_3 as the third surrogate. Surprisingly enough, these conclusions are identical to the ones obtained by simply looking at the change in impurity of

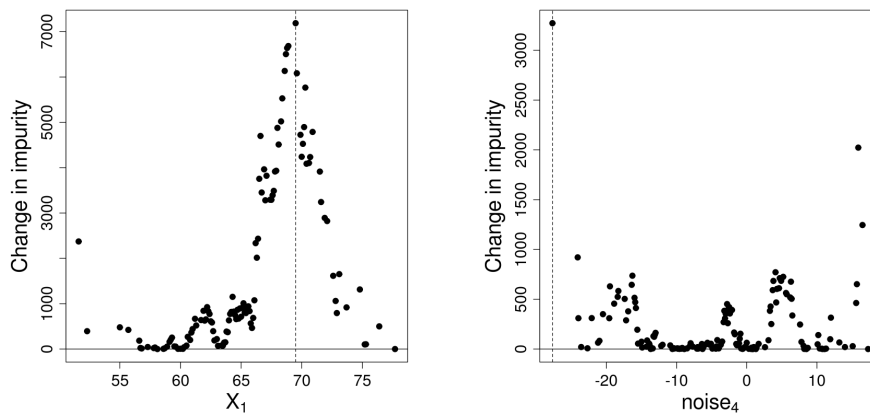


Figure 4.10: An example of the change in impurity plot for two continuous predictors. The plot on the left shows a meaningful split whereas the plot of the right shows a spurious split since the maximum values are attained in the boundaries of the possible splitting points.

the original dataset (Figure 4.8). Rather than being a coincidence, this seems to be the case when other data were simulated. After running all the replicates, the distribution of primary splits seems to provide a similar set of primary and surrogate splits than that one would obtain by looking simply at the change of impurity of each one of the predictors without doing any re-sampling.

Moreover, the fact that noise_4 (a predictor that has no relationship with the response) appears to be the second best split is somehow inadequate. To investigate in more detail why this is the case, Figure 4.10 shows the change in impurity versus the cutpoints for both X_1 and noise_4 . As one can see, the maximum change in impurity obtained for X_1 makes more sense than the maximum change in impurity obtained for noise_4 . In the latter case, the maximum is attained for the smallest value of the predictor. The reason is that the value of the response for that particular observation is an extreme value and this results in the change in impurity observed in the plot. Thus, the observed maximum could be considered as an outlier. To understand how this outlier will propagate when bootstrap replicates are generated Figure 4.11 shows the plot of noise_4 versus change in impurity for 9 such replicates. As one can see, the same pattern is observed for some of the replicates. The conclusion is that this outlier will, therefore, provide misleading information about the real nature of the split generated by this particular predictor. In fact, values of the predictor that are not close to the outlier give values for the change in impurity that are much smaller, as one could expect from a predictor that is not related to the response.

One way of addressing this problem is by considering not just the cutpoint

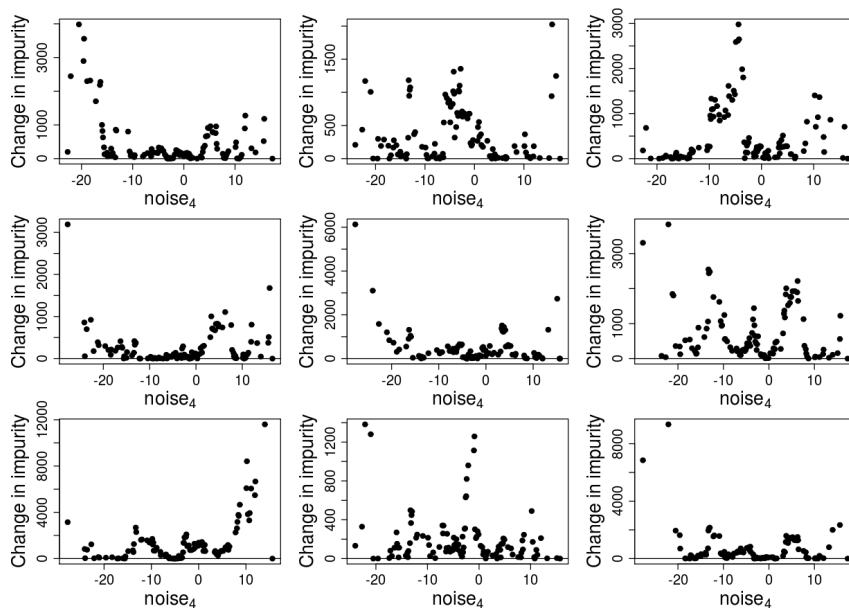


Figure 4.11: Plot of the change in impurity of noise_4 for 9 different bootstrap replicates.

for which the maximum change in impurity is attained but a few more of these points. Figure 4.12 shows the result of doing this when the 6 cutpoints with the highest reduction in the sum of squares were considered and averaged. If the split is meaningful, as in the case of X_1 , the result of applying this new procedure will not change significantly the selection of the cutpoint and the corresponding value of the change in impurity. However, if the split is spurious as in the case of noise_4 , the cutpoints associated with the highest values of the change in impurity will be scattered over different values of the predictor. By choosing the average of those values as the cutpoint, the corresponding change in impurity will be more meaningful since it will be a reflection of the change in impurity of, not just one point, but the range of the cutpoints for which the change in impurity is the highest.

Before implementing this new strategy for dealing with spurious splits, a remark about the algorithm presented previously is worth mentioning. At each replicate, the predictor with the maximum change in impurity is selected but the information about the other predictors and their respective change in impurity is lost. To explain in more detail this point, suppose that the change in impurity for one particular replicate is as follows:

| X_1 | X_2 | X_3 | N_1 | N_2 | N_3 | N_4 | N_5 | N_6 | N_7 | N_8 | N_9 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 718 | 72 | 213 | 274 | 155 | 204 | 327 | 108 | 9 | 214 | 42 | 88 |

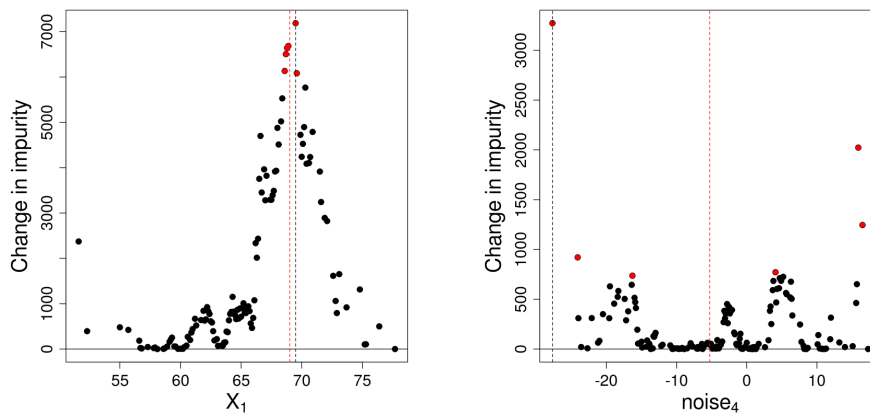


Figure 4.12: In this example the 6 cutpoints leading to the maximum changes in impurity are considered. By doing this, splits that are spurious (like noise_4 in this example) will produce a more meaningful value for the cutpoint.

The maximum is attained for X_1 but that does not tell one how important the change in impurity is when you compare it to the change in impurity of the other predictors. For instance, N_7 has a greater reduction than X_3 but both are very similar with values of 214 and 213. To assess the relative importance of each predictor one can divide each value of the change in impurity by the sum of all the values of the change in impurity for all the predictors. In the example being examined, the relative importance would be

| X_1 | X_2 | X_3 | N_1 | N_2 | N_3 | N_4 | N_5 | N_6 | N_7 | N_8 | N_9 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| .30 | .03 | .09 | .11 | .06 | .08 | .13 | .04 | .00 | .09 | .02 | .04 |

By doing this the information registered is not just about the best predictor but about how much better is the best predictor when compared to the others. To avoid making this value depend on the number of predictors one can multiply by the number of predictors. In this way, if all the predictors were very similar the relative importance would approximately be 1 for all the predictors. These values can be stored for all the bootstrap replicates and can be averaged in order to choose the variables for the primary split and the surrogate splits.

All of these considerations lead to a slightly more sophisticated version of the algorithm presented previously. In this version, the relative importance of each predictor and the optimal cutpoint is retained in each replicate. If the predictor is continuous, the optimal cutpoint is obtained by considering a small set of cutpoints leading to the highest values in the change in impurity (as explained in Figure 4.12). Typically the number of values can be 6 and the optimal cutpoint is just the average of those values. If the predictor is categorical, the cutpoint is

obtained in the usual way (by considering all possible combinations of the levels of the factor and choosing the one leading to highest change in impurity). Once the splitting point has been chosen, the change in impurity for each predictor is calculated using the data that were not used to obtain the cutpoint (this is called the Out of Bag data OOB). The reason for doing this is to obtain a more realistic estimate of the change in impurity and to avoid the variable selection bias. Although this might seem a little artificial, it is not when one thinks of the change of impurity from a different perspective. As defined in Chapter 3 the change in impurity for a continuous response in a given node τ is

$$\Delta I(\tau) = I(\tau) - (I(\tau_L) + I(\tau_R))$$

where $I(\tau) = \sum_{y_i \in \tau} (y_i - \bar{y}_\tau)^2$ and τ_L and τ_R are the corresponding left and right nodes generated by a particular split. It is easy to prove that

$$I(\tau) = I(\tau_L) + I(\tau_R) + n_L(\bar{y}_{\tau_L} - \bar{y}_\tau)^2 + n_R(\bar{y}_{\tau_R} - \bar{y}_\tau)^2$$

where n_L and n_R are the number of observations in the left and the right nodes respectively. To see this let

$$\begin{aligned} I(\tau) &= \sum_{y_i \in \tau} (y_i - \bar{y}_\tau)^2 = \sum_{y_i \in \tau_L} (y_i - \bar{y}_\tau)^2 + \sum_{y_j \in \tau_R} (y_j - \bar{y}_\tau)^2 \\ &= \sum_{y_i \in \tau_L} (y_i - \bar{y}_{\tau_L} + \bar{y}_{\tau_L} - \bar{y}_\tau)^2 + \sum_{y_j \in \tau_R} (y_j - \bar{y}_{\tau_R} + \bar{y}_{\tau_R} - \bar{y}_\tau)^2 = \\ &= \sum_{y_i \in \tau_L} (y_i - \bar{y}_{\tau_L})^2 + n_{\tau_L}(\bar{y}_{\tau_L} - \bar{y}_\tau)^2 + 2(\bar{y}_{\tau_L} - \bar{y}_\tau) \sum_{y_i \in \tau_L} (y_i - \bar{y}_{\tau_L}) + \\ &+ \sum_{y_j \in \tau_R} (y_j - \bar{y}_{\tau_R})^2 + n_{\tau_R}(\bar{y}_{\tau_R} - \bar{y}_\tau)^2 + 2(\bar{y}_{\tau_R} - \bar{y}_\tau) \sum_{y_j \in \tau_R} (y_j - \bar{y}_{\tau_R}). \end{aligned}$$

Given that $\sum_{y_i \in \tau_L} (y_i - \bar{y}_{\tau_L}) = 0$ and $\sum_{y_i \in \tau_R} (y_i - \bar{y}_{\tau_R}) = 0$ it follows that

$$I(\tau) = I(\tau_L) + I(\tau_R) + n_{\tau_L}(\bar{y}_{\tau_L} - \bar{y}_\tau)^2 + n_{\tau_R}(\bar{y}_{\tau_R} - \bar{y}_\tau)^2.$$

Therefore,

$$\Delta I(\tau) = n_L(\bar{y}_{\tau_L} - \bar{y}_\tau)^2 + n_R(\bar{y}_{\tau_R} - \bar{y}_\tau)^2.$$

Thus, the change in impurity is also measuring the effect that a particular split has on the response. The use of the OOB data to estimate the change in impurity indirectly provides an estimate of a measure of the effect of a particular predictor using the data that were not used to calculate the cutpoint. Furthermore, by using the OOB data to measure the change in impurity, the problem of variable selection bias is also eliminated as it will be shown in the example below.

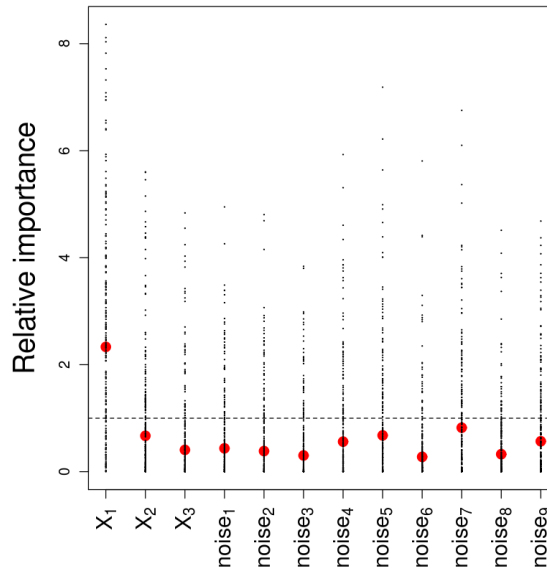


Figure 4.13: An example of the relative importance plot produced by the node re-sampling algorithm. The value 1 represents equality for all the variables. In this example X_1 is clearly selected for the primary split.

This last point is important as there are some similarities with the unbiased recursive partitioning by Hothorn *et al.* (2006). While the key feature to avoid variable selection bias in unbiased recursive partitioning is the separation of variable selection in one step followed by the selection of the splitting point in a second step, in the proposed algorithm, for each bootstrap replicate, the selection of the splitting point is made first and the selection of the covariate is made afterwards (using the OOB data).

Once the change in impurity is calculated for every predictor the final step is to calculate the relative importance of each predictor by dividing by the total amount of change in impurity. After repeating these steps for all the bootstrap replicates, the primary and surrogate splits can be identified as the predictors with the highest median relative importance. As for the splitting points, the typical values obtained from the bootstrap distribution of cutpoints can be considered. The pseudo-code of the algorithm is given in Algorithm 1. The input of the algorithm consists of the available data (in one particular node), the response variable, and the set of predictors.

An example is given in Figure 4.13 where the node re-sampling algorithm is applied to the data used to generate Figure 4.8. As one can see, the predictor chosen for the primary split is again X_1 , but the plot gives a more realistic picture of the nature of the split. The importance of the different predictors seem to

Algorithm 1 Node resampling pseudocode

$BS \leftarrow$ number of bootstrap replicates
for $j = 1$ to BS **do**
 $BSdata \leftarrow$ bootstrap replicate of the data
 $BSdataOOB \leftarrow$ Out of Bag data (OOB)
 $Y \leftarrow$ values of the response in $BSdata$
 $Y_{oob} \leftarrow$ values of the response in OOB
 $SS = \sum (Y_{oob} - \overline{Y_{oob}})^2$ sum of squares of Y_{oob}
 $VAR \leftarrow$ set of predictors
 for $i \in VAR$ **do**
 $X[i] \leftarrow$ values of predictor i in $BSdata$
 $splits \leftarrow$ set of all possible splitting points
 $SS \leftarrow$ change in impurity for each splitting point
 if i is continuous **then**
 $topSplits \leftarrow$ set of splits with the highest change in impurity
 ▷ Usually 6 values
 $spl[i][j] \leftarrow$ average of $topSplits$
 end if
 if i is categorical **then**
 $spl[i][j] \leftarrow$ split with highest change in impurity
 end if
 $Y_{left} \leftarrow$ values of Y_{oob} to the left of $spl[i][j]$
 $Y_{right} \leftarrow$ values of Y_{oob} to the right of $spl[i][j]$
 $SS_{left} = \sum (Y_{left} - \overline{Y_{left}})^2$
 $SS_{right} = \sum (Y_{right} - \overline{Y_{right}})^2$
 $Delta[i][j] = SS - (SS_{left} + SS_{right})$
 end for
 $sumDelta \leftarrow$ sum of all the values of $Delta$ for all predictors
 for $i \in VAR$ **do**
 $RDelta[i][j] \leftarrow Delta[i][j] / sumDelta * |VAR|$ ▷ relative importance
 end for
end for
for $i \in VAR$ **do**
 $aveRDelta[i] \leftarrow$ median of $RDelta[i][\cdot]$
 if i is continuous **then**
 $aveSpl[i] \leftarrow$ median of $spl[i][\cdot]$
 end if
 if i is categorical **then**
 $aveSpl[i] \leftarrow$ mode of $spl[i][\cdot]$
 end if
end for

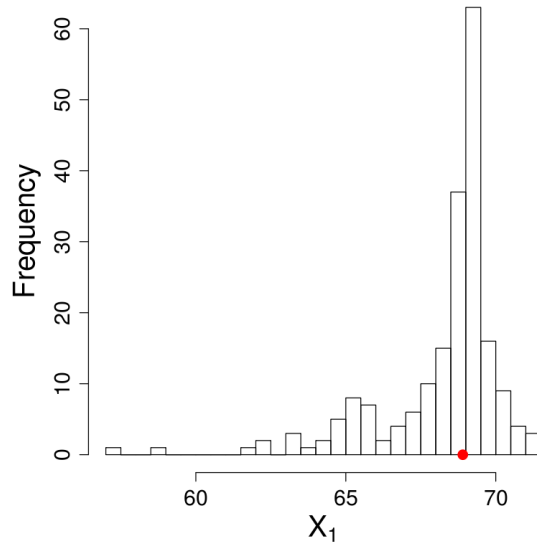


Figure 4.14: Histogram of the bootstrap distribution of the cutpoints for X_1 . The red point is the median value. The two bumps observed in the histogram represent the two splitting points of this predictor in the underlying model.

have changed when compared to the picture described in Figures 4.8 or 4.9. Now, noise_7 , noise_5 and X_2 are the surrogate predictors (all of them categorical). It seems, therefore, that the variable selection bias has been eliminated. The horizontal dashed line in the plot represents the threshold corresponding to equality of importance 1. The cutpoint for the primary split can be obtained from the bootstrap distribution of the cutpoints represented in Figure 4.14. The red point is the value of the median which is 68.9. It can also be seen in the histogram that there is another small bump around 65 which corresponds with the other splitting point for predictor X_1 . Based on the information obtained in Figures 4.13 and 4.14 the primary split suggested at node 1 (the root node) is $X_1 \leq 68.9$. To assess the validity of this split one can check if there are any significant differences between the two siblings. The results of the t test are:

```
data: Y by X1 <= 68.9
t = -3.5473, df = 33.952, p-value = 0.001161
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.205647 -7.115266
sample estimates:
mean in group X1 > 68.9  mean in group X1 <= 68.9
      41.87430             58.53475
```

The conclusion is that the differences based on that split are significant and,

therefore, it is worthwhile to include the split in the tree.

The algorithm presented here can be used every time a split has to be created. At each node, the information obtained consists of a summary of how the different predictors are related to the response through the use of the relative importance plot and an estimate of the distribution of the cutpoints. Of course, the question remains about how large the tree should be, or when to stop growing the tree. The node re-sampling algorithm does not answer that question. One possible approach is to use the result of the test to decide if the tree should grow any further at any particular node. This is the same approach that trees based on unbiased recursive partitioning use. However, as it will be shown in the next section, this is very problematic if interactions are present in the data. The other alternative is to grow a large tree and then to use some kind of pruning procedure to reduce the size of the tree. All these issues will be addressed in the next section where the node re-sampling algorithm is used in the survival context.

Before going to the next section, we comment briefly on how the algorithm deals with outliers in the response.

4.3.1 Outliers in the response

The recursive partitioning algorithm is immune to the presence of outliers in the continuous predictors since the role played by them is just to split the dataset in two groups. However, outliers in the response might have an effect in the creation of the split. To test if this is the case, 2 artificial outliers were introduced in the data used in the previous example. Two observations were chosen at random from the data and the values of the response were set to 240 and 250 respectively. Figure 4.15 (a) shows the optimal change in impurity of each one of the predictors. The change in impurity for each splitting point of the best predictor (in this case noise_1) is depicted in Figure 4.16 (b). Therefore, by simply incorporating two outliers, the split is modified notably. The node re-sampling algorithm was run for the exact same data resulting in Figure 4.16. The variable chosen for the optimal split is X_2 followed by X_1 although all the variables look very similar, as one can see in Figure 4.16 (a). The splitting points of X_1 are represented in Figure 4.16 (b). Although the distribution has varied, there is a peak remaining at around 69.

In summary, the presence of outliers in the response, when the response is continuous, might affect the selection of the splits dramatically. The node re-sampling algorithm was able to identify two of the predictors that are related to the response for the primary and secondary (first surrogate) splits, although the differences were very small when compared to the other predictors. The best approach for identifying outliers in the response is to plot the response first and to identify those observations that could create problems.

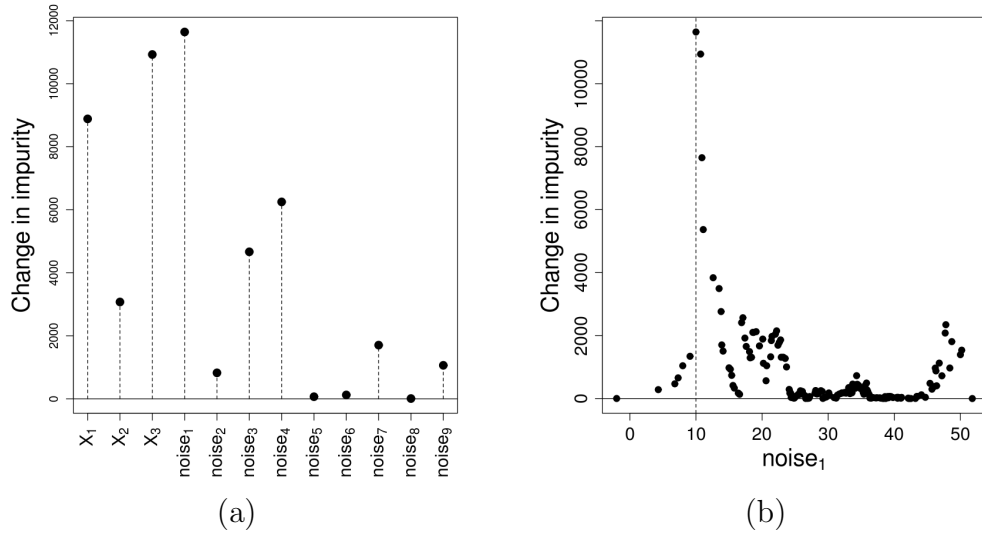


Figure 4.15: The plot of the change of impurity for all the predictors after two outliers were introduced in the random sample (a). In (b), the plot of the change in impurity versus the cutpoints for the primary split noise₁.

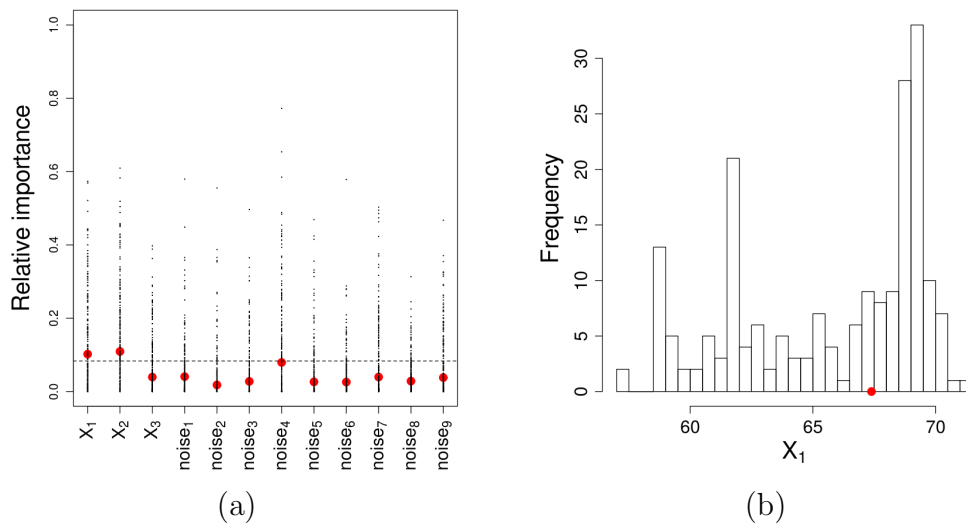


Figure 4.16: The relative importance plot after the node re-sampling algorithm was run with the two outliers (a). In (b), the histogram of the bootstrap distribution of possible cutpoints for X_1 .

4.4 Node re-sampling: survival responses

This section is dedicated to the extension of the ideas presented previously to the analysis of survival data. Although most of the points made in the previous section are valid, there are still a few issues that have to be addressed. One of these issues is how to obtain the optimal tree. This will be explained in this section in the survival context, although the method can also be applied to continuous and categorical responses. Another issue has to do with the use of a statistical test in the split. It was said that in order to assess the validity of the split after running the node re-sampling algorithm one could simply perform a test and check its significance in order to decide if the split is worth keeping. It will be shown here that this is dangerous for two reasons. One reason is that important interactions can be missed if this is used to stop growing the tree. The other reason is that some apparent significant splits might be spurious in the sense that the predictor involved in the split may not be related to the response. To illustrate all these considerations an example with survival responses is presented below.

4.4.1 Synthetic example 2

Suppose that the tree represented in Figure 4.17 defines the underlying model from which data can be simulated. The distribution of the response will be assumed to be gamma with means and variances as described in the terminal nodes. The set of predictors and their distributions are identical to those described in Table 4.1 and the response is simulated with 20% of the observations being right censored.

The first step will be to grow a large tree as in the CART algorithm. Because of the different nature of the survival responses, a different splitting criterion has to be considered. The logrank statistic as a measure of dissimilarity between the two daughter nodes is the most appropriate (Segal, 1988). The value of the logrank statistic is calculated as in the unbiased recursive partitioning (Hothorn *et al.*, 2006). Thus, the node re-sampling algorithm presented in the previous section can be run by simply changing the measure of the change in impurity to the value of the logrank statistic. Figure 4.18 shows a snapshot of the graphical user interface that has been developed to accommodate all the elements of the node re-sampling algorithm when applied to survival data. At each node, different pieces of information are visually available. The number on top of the node is the identification of the node. Below that, there are the size of the node, the name of the predictor and a p-value. The p-value corresponds with the logrank test once the predictor and the cutpoint have been selected for the split. To highlight the p-value the little colored circle on the right of each node will be displayed in green if the p-value is less than 0.01, in orange if the p-value is between 0.01 and 0.05, and in red if the p-value is > 0.05 .

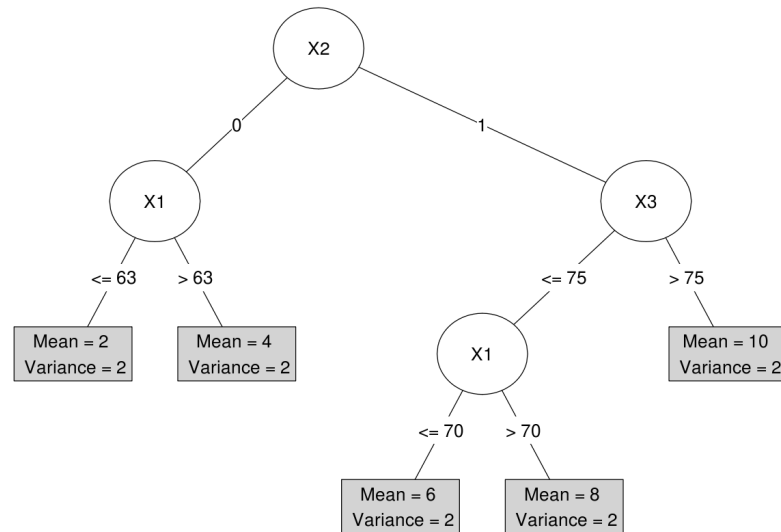


Figure 4.17: An example of a survival model represented in a tree fashion. The distribution of the response in each terminal node is assumed to be gamma with means and variances as specified in the squared boxes.

The right hand side of Figure 4.18 contains specific information about any chosen node. For instance, at node 1 (the root node) the relative importance plot, as described in the previous section¹, shows how X_2 is the most important predictor (primary split) followed by X_3 and X_1 . By looking at Figure 4.18 it is obvious that some of the splits are spurious. For instance, node 25 has noise_4 as the primary split and the split generated seems to be significant with a p-value of 0.0028. Although it is possible to have significant splits below other non significant splits (this will be considered later on when interactions are present in the data), noise_4 is not related to the response and should not be in the model.

In order to find which splits are relevant for the model and which ones are not, some additional information has to be analyzed. Figure 4.19 shows the plot with the node re-sampling OOB logranks in node 25. The plot shows the average value (red dots) of the logranks for all the bootstrap replicates and for each one of the predictors. Recall from the previous section that the value of the change of impurity (the logrank statistic in this case) is calculated using the OOB data in

¹The red points are the medians of the relative importance values for each replicate and the vertical lines go from the 0.05 to the 0.95 quantiles of such values. This way of displaying the plot was chosen instead of boxplots in order to make it simpler when many predictors are being analyzed.

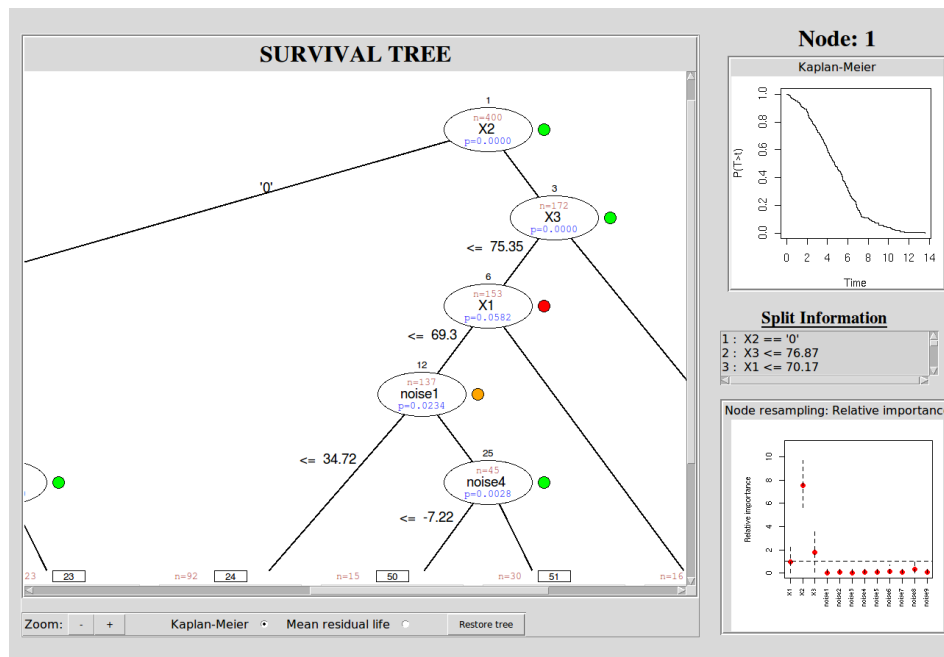


Figure 4.18: A snapshot of the graphical user interface created to accommodate the node re-sampling algorithm for survival responses.

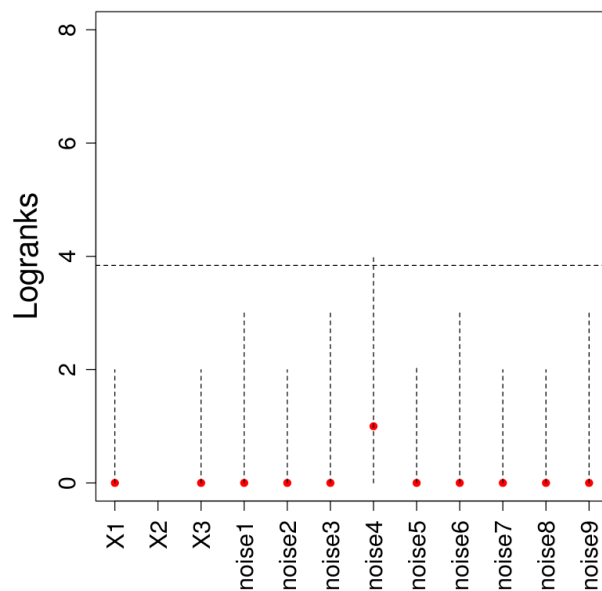


Figure 4.19: Plot of the node re-sampling out of bag logranks for node 25 in Figure 4.18.

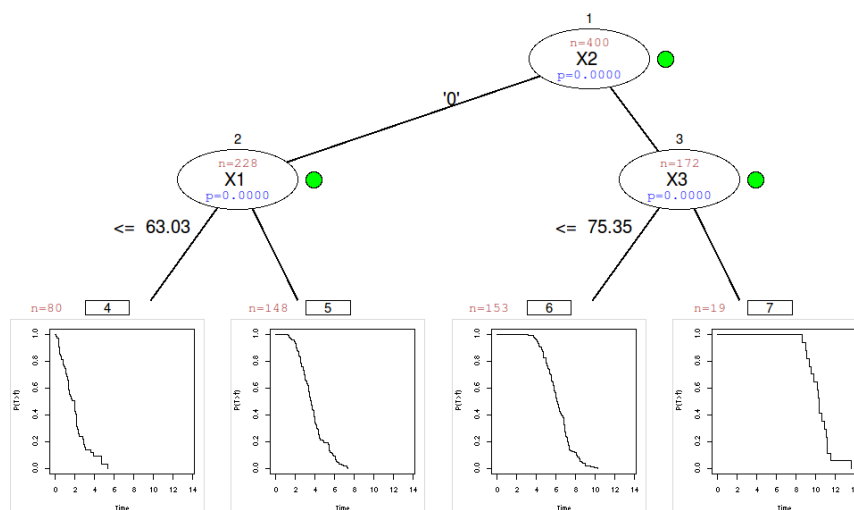


Figure 4.20: Survival tree based on the node re-sampling algorithm after all the irrelevant nodes have been eliminated.

each replicate. If the cutpoint for the split was fixed, and only one random sample was available, the criterion to determine if there was a significant difference between the two generated nodes would be to evaluate the logrank statistic and compare its value with the critical point 3.84 (0.95 quantile of a chi square distribution with 1 degree of freedom). If the split at the fixed cutpoint was significant (logrank statistic ≥ 3.84), one would expect that other random samples would also provide high values of the logrank statistic (unless of course there were not differences and the value obtained in the first place had less than 0.05 probability of being observed). The node re-sampling algorithm calculates the OOB values of the logrank statistic at the selected splitting points of each bootstrap replicate and obtains, at the end of this process, a global mean. The criterion to determine whether a split is relevant for the model is to compare this global average with the critical value 3.84. Based on this criterion, ‘irrelevant’ predictors would have an averaged value of the OOB logrank statistic (red dots in Figure 4.19) below the threshold 3.84 (horizontal line). As one can see, all the predictors in node 25 have values below the threshold and, therefore, the split generated at that particular node can also be considered ‘irrelevant’. This concept of ‘irrelevant’ splits provides a criterion to prune the large, or saturated, tree. Starting from the bottom of the tree, nodes that are ‘irrelevant’ are systematically deleted until a node is reached that is not ‘irrelevant’.

Figure 4.20 shows an example after all the ‘irrelevant’ splits have been deleted. The result is the optimal tree generated by the algorithm. This tree captures the structure of the underlying model with only one split missing in node 6.

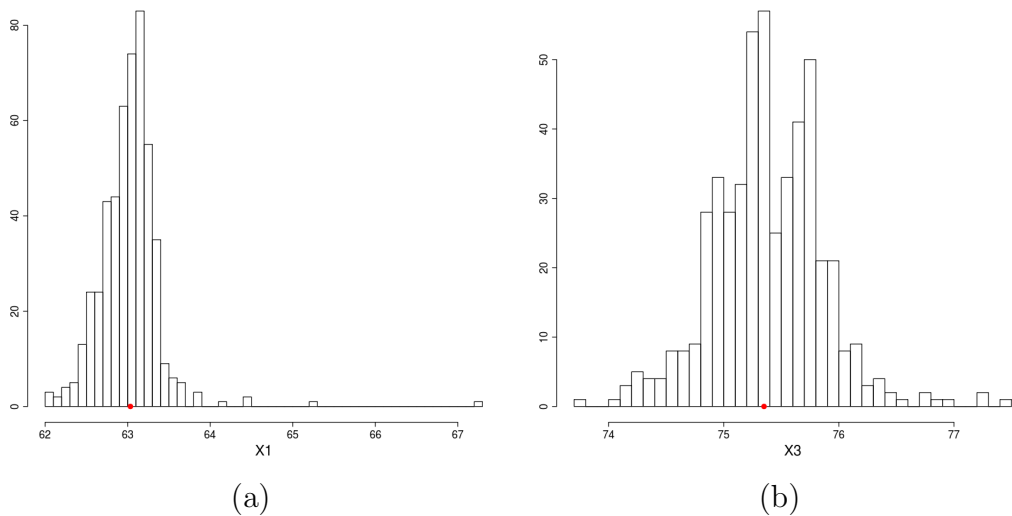


Figure 4.21: Bootstrap distributions of the cutpoints for nodes 2 (a) and 3 (b) in Figure 4.20.

Figure 4.21 shows the bootstrap distribution of the cutpoints for node 2 (left) and node 3 (right). This highlights the fact that all the cutpoints are random quantities.

To compare the results with other methods for growing survival trees Figures 4.23 and 4.22 shows the trees obtained using unbiased recursive partitioning and the version of CART adapted to survival responses, respectively.

The tree based on CART has captured only two of the 4 splits in the model. This tree has been selected looking at the complexity parameter plot (not shown here) and pruning a much larger tree. The use of the CART algorithm for survival responses based on deviance residuals (as proposed by LeBlanc & Crowley, 1992) can be problematic in the pruning step of the process (see Chapter 3). Another problem related to the use of CART is the problem of variable selection bias. Although this does not seem to be an issue in this particular example, there are no guarantees that predictors with many different cutpoints might be erroneously selected for the split.

On the other hand, the survival tree based on unbiased recursive partitioning (Hothorn *et al.*, 2006) has also captured 2 of the 4 splits. This method is not affected by the variable selection bias, since the selection of the split is divided into two different steps. However, there are two important problems with this method. One of them is that the global test of independence at each node is based on a set of individual tests, one for each one of the predictors. Because of that, the more predictors in the model, even if they are completely unrelated to the response, the more difficult will be for the test to find significant differences,

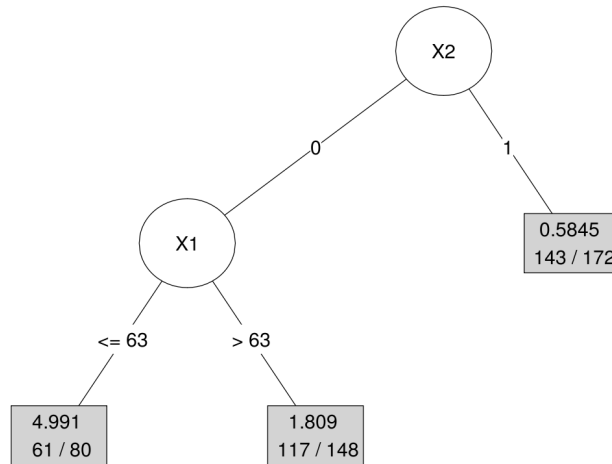


Figure 4.22: Survival tree using the the adapted version of CART for survival responses (LeBlanc & Crowley, 1992). The data used to generated the tree is identical to the one used to generate the tree in Figure 4.20.

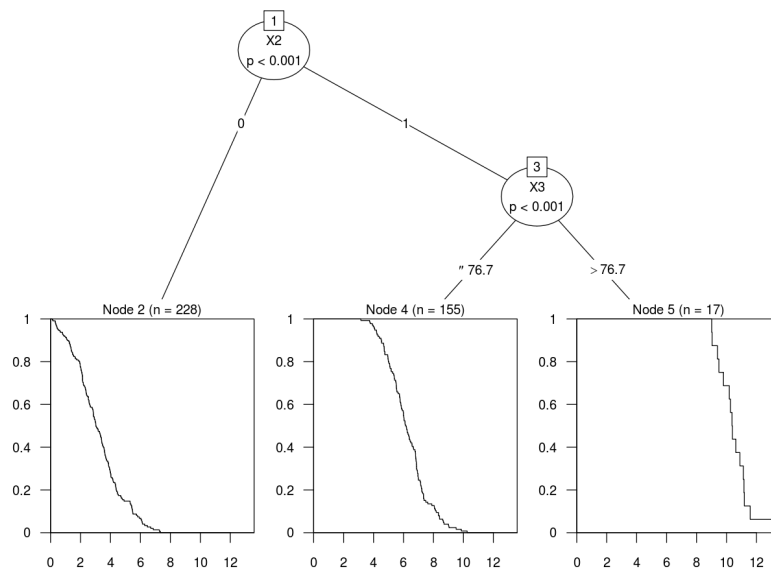


Figure 4.23: Survival tree using conditional inference procedures (Hothorn *et al.*, 2006). The data used to generated the tree is identical to the one used to generate the tree in Figure 4.20.

since the significance level has to be penalized by the number of predictors (see synthetic example 1). The second problem is that by stopping the growth of the tree one can miss important interactions in the model.

The tree based on node re-sampling is free from most of the problems presented above. It is not affected by the variable selection bias since the selection of the split is also divided into two parts. First of all, the cutpoints are identified and secondly the effect of the split generated by the cutpoint is evaluated using the OOB data. This procedure is repeated many times using bootstrap replicates of the original data and the results are averaged over all the replicates producing the, so-called, relative importance plot. The graphical user interface allows the growth of a large tree and the study of the information obtained by the algorithm for each one of the nodes in a very straight forward manner. Moreover, by analyzing the plot of the bootstrap OOB logranks, splits that are not relevant for the model are easily identified and eliminated. This provides the user with a standard method to prune the tree that can be used for any other type of response. Furthermore, the relative importance plot provides a tool for identifying the primary split and the subsequent surrogates. Because this is not based on any test, no penalization is required, and the same results should be obtained irrespective of the number of predictors involved in the study.

4.4.2 Interactions

Although the content of this section is related to survival responses, most of the considerations made here are also valid for categorical and continuous responses. Suppose one has an interaction in the model as represented in Figure 4.24. Data were simulated based on this model and the node re-sampling algorithm was run resulting in the tree displayed in Figure 4.25 (the tree has already been pruned and all the ‘irrelevant’ nodes eliminated). As one can see, the tree adequately captures the interaction in the model. At node 1 the plot of the bootstrap OOB logranks (bottom right in the plot) shows that the split should not be included in the model as all the red dots are below the threshold (‘irrelevant’ node). If no other splits below node 1 were found to be ‘relevant’ that node should be eliminated and the tree pruned. However, due to the interaction term, nodes 2 and 3 are ‘relevant’ nodes. This is shown in Figure 4.26 where the plot of the bootstrap OOB logranks is shown for nodes 2 and 3 respectively. As one can see, both nodes have mean logranks above the threshold and, as a result, they should be included in the model. Although not shown here, all the nodes below nodes 2 and 3 had values below the threshold and, therefore, were eliminated.

To see to what extent the other methods for growing survival trees are able to capture the interaction term, the tree based on unbiased recursive partitioning and the tree based on CART were generated using the exact same dataset. The results

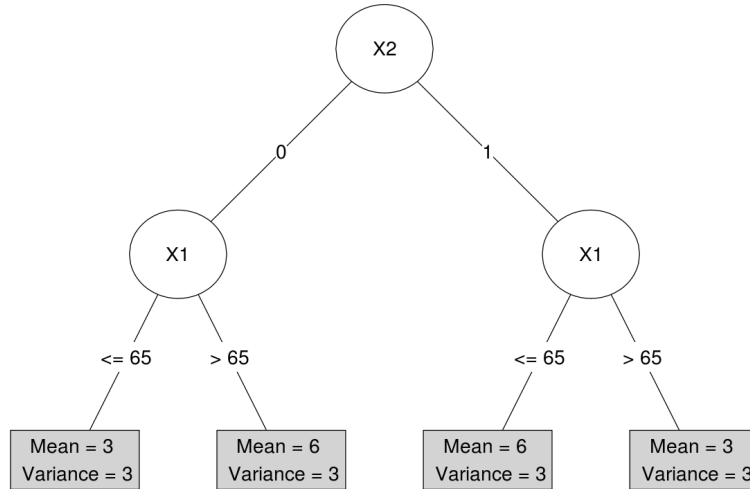


Figure 4.24: An example of a model represented in a tree fashion with an interaction. The distribution of the response in each terminal node is assumed to be gamma with means and variances as specified in the squared boxes.

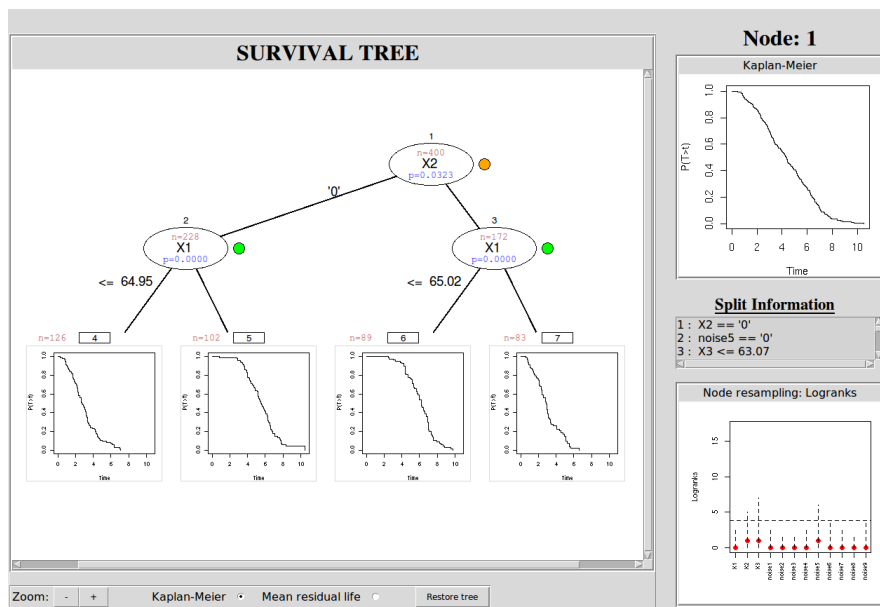


Figure 4.25: Snapshot of the optimal tree after running the node re-sampling algorithm with the underlying model containing an interaction term.

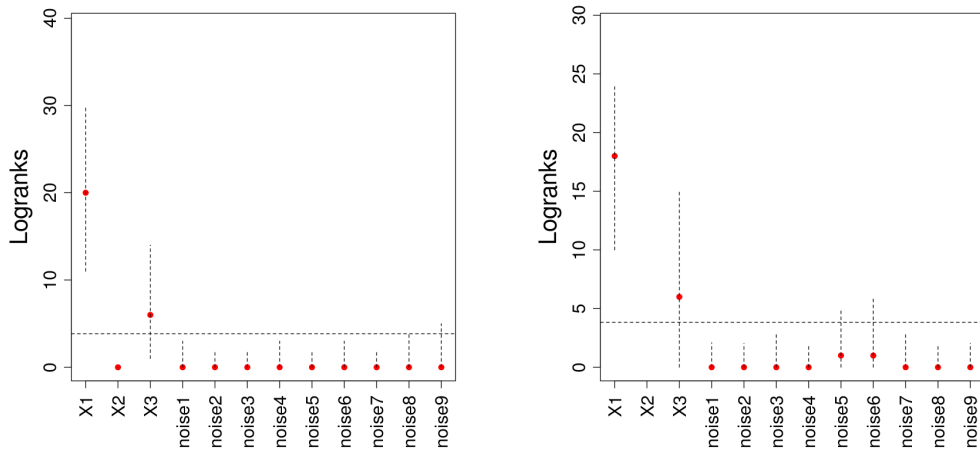


Figure 4.26: Plot of the out of bag values of the logrank statistic for each replicate and predictors in nodes 2 and 3 (left and right respectively).

of unbiased recursive partitioning are shown in Figure 4.27. As one can see, the algorithm was unable to even generate one node, as the growth of the tree stops after obtaining non-significant results. Due to the presence of the interaction, the non-significant results were obtained in the first node. This particular issue, somehow diminishes enormously the attractiveness of this method. The reason is that tree based methods using recursive partitioning were designed to deal with interactions without the need to look for them previous to the application of the method. In this sense, interactions should naturally arise in the model. The only way of achieving this goal is by growing a saturated tree first and by pruning the tree using some adequate pruning procedure. By stopping the growth of the tree, the algorithm does not allow the tree to find important splits that might be hiding in subsequent nodes.

The survival tree based on CART was also grown for the same dataset. Figure 4.28 shows the complexity parameter plot. Based on this plot, one can discard the whole tree and select as the optimal model the tree with no splits.

In summary, the proposed algorithm based on node re-sampling seems to handle the presence of interactions more adequately than the other two algorithms. The method based on unbiased recursive partitioning is inherently faulty in this regard and cannot perform well when interactions are present. One could increase the significance level in order to allow the tree to grow further but, in that case, many other irrelevant splits might be generated. Since no formal method for pruning the tree has been proposed using this method, this solution will not work either. This problem also applies for trees based on unbiased recursive partitioning with

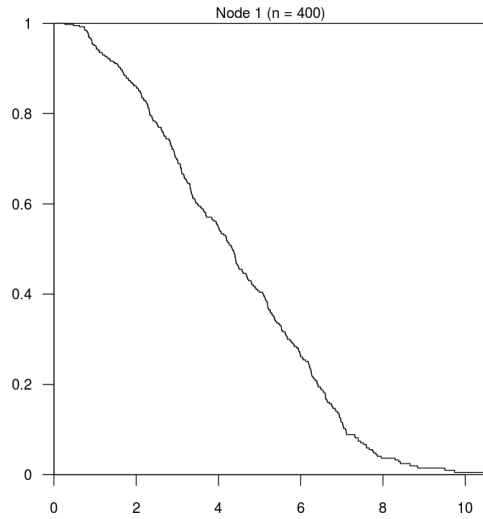


Figure 4.27: Output obtained using unbiased recursive partitioning after a random sample was drawn from the underlying model with an interaction (Figure 4.24).

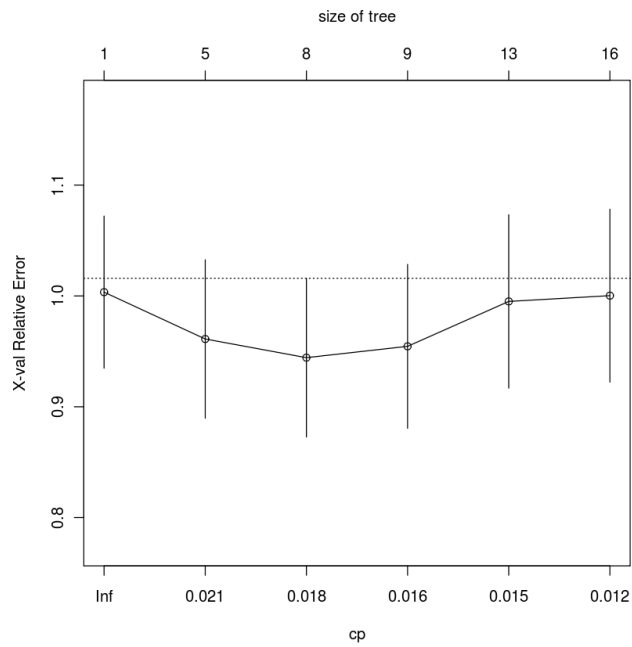


Figure 4.28: Plot of the complexity parameter versus the relative error for the tree based on CART where the underlying model has an interaction.

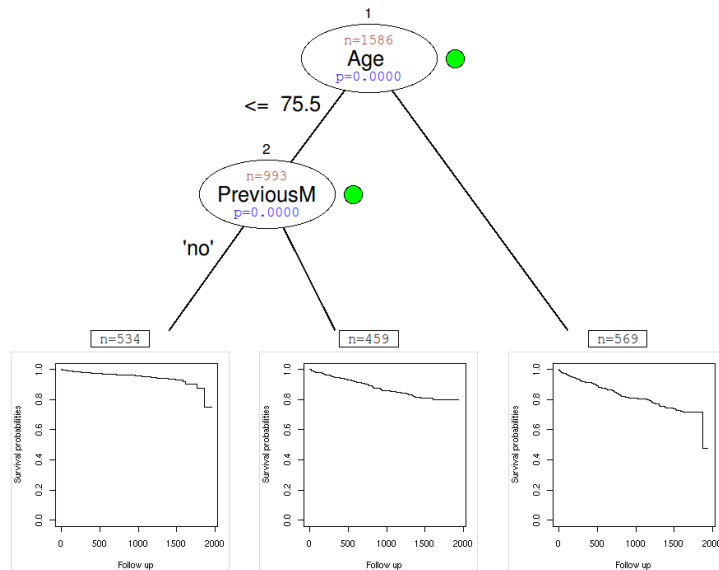


Figure 4.29: Survival tree based on node re-sampling for the coronary data.

any other type of responses. The method based on CART does allow for the incorporation of interactions since a large tree is generated first. However, the pruning procedure is rather problematic for survival responses, as was explained in previous Chapters. Moreover, the algorithm can be affected by the variable selection bias. The node re-sampling algorithm has been shown to be adequate even in the presence of interactions. It is, in fact, the only method that correctly identified the underlying model. In the last part of this chapter, the proposed method is applied to the study of the two datasets used in this thesis.

4.5 Applications of the node re-sampling algorithm

To examine how the novel method for growing survival trees performs when applied to real data, the two datasets presented in this thesis were used to grow survival trees based on node re-sampling. The first model is based on the coronary dataset where the event of interest was death from any cause. To analyze how the different predictors are related to the response, Figure 4.29 shows the result of applying the node re-sampling algorithm to this dataset. The tree depicted is already pruned in the manner described in the previous section. The two predictors present in the model are the age of the patient and the factor indicating if the patient had a previous myocardial infarction. In the terminal

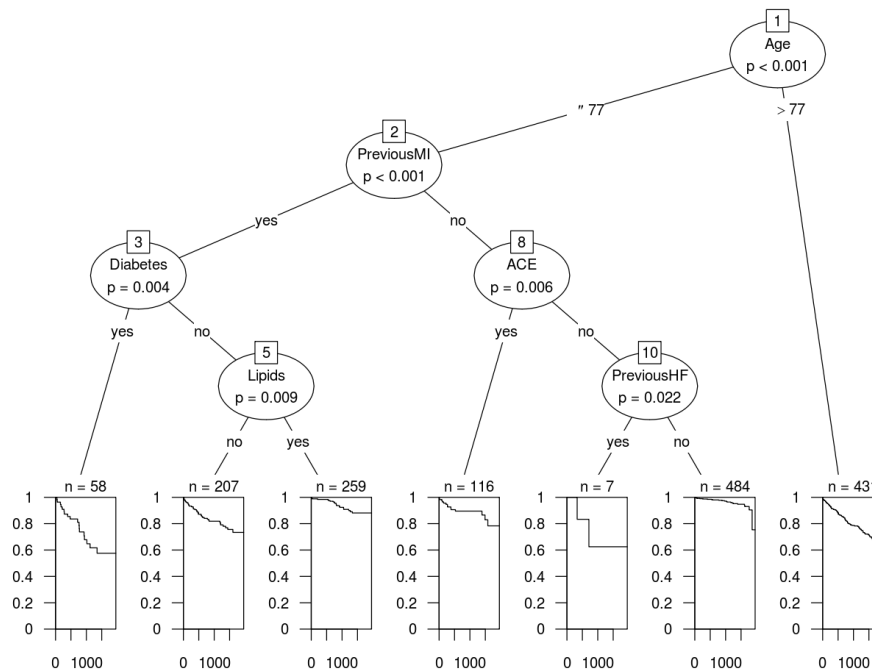


Figure 4.30: Tree based on unbiased recursive partitioning (coronary data).

nodes the Kaplan Meier estimates of the survival functions are displayed. As one can see older patients have worse prognosis, with 75 years old being the threshold that separates patients with higher risk. The ‘best’ group is formed by younger patients with no previous history of myocardial infarction. Patients younger than 75 years old with a previous history of myocardial infarction have an intermediate risk.

To compare the results obtained when the tree was grown using unbiased recursive partitioning Figure 4.30 shows the tree obtained after applying this method (this tree was previously shown in Chapter 4). One can see some similarities with the previous model. For instance, Age and PreviousMI appear in both models. However, other predictors such as Diabetes, Lipids, ACE or PreviousHF appear only in the latter model. Recall from Chapter 3 that the three covariates identified for the inclusion in the final model using Cox proportional hazard models were Age, PreviousMI and ACE. There is another difference between the node re-sampling and the unbiased recursive partitioning models which is the cutpoint for Age. Whereas the cutpoint is 77 years old for the latter model, the cutpoint is 75.5 years old for the node re-sampling model. To determine which one is more likely to be adequate, one can look at the distribution of the cutpoints generated by the node re-sampling algorithm. This is represented in Figure 4.31. The plot shows that the value of 75.5 is more adequate than the value of 77 based on

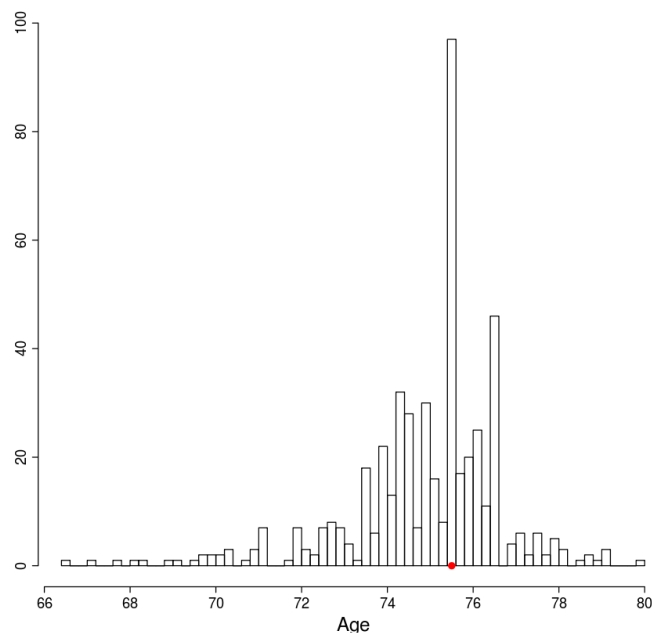


Figure 4.31: Bootstrap distribution of the cutpoints for Age in node 1 for the tree based on node re-sampling.

the bootstrap distribution of the cutpoints. Finally, the tree generated using the algorithm based on CART is not presented since the complexity parameter plot suggested the tree with no splits as the optimal tree.

The node re-sampling algorithm was also used for the breast cancer dataset. When only pathological predictors were considered for the analysis, Figure 4.32 shows the node re-sampling tree after pruning the irrelevant splits. The event of interest in this model is the recurrence of the disease. There are three variables identified by the model: LN_0_1, LVI_0_1 and Size_rec. The label ‘Size_rec’ refers to the covariate ‘Size’ categorized as: “0”, “ ≤ 20 mm”, “ > 20 to ≤ 50 mm” and “ > 50 mm”. The worse prognosis in terms of disease free survival corresponds to patients with ‘LN’ positive. It is interesting to see that even though LVI_0_1 does not generate a significant split (p-value=0.12), the plot of the bootstrap OOB logranks (bottom right in the picture) indicates that the node generates a ‘relevant’ split. This allows the split about size to be incorporated in the tree. Another interesting comment is that patients with positive size of the tumor seem to do better than patients with size 0 (node 4). The distribution of this covariate at that particular node is represented in Figure 4.33. The split at node 4 separates patients with no tumor from patients with a positive size of the tumor. The fact that patients with positive size of the tumor have lower probabilities of recurrence could be

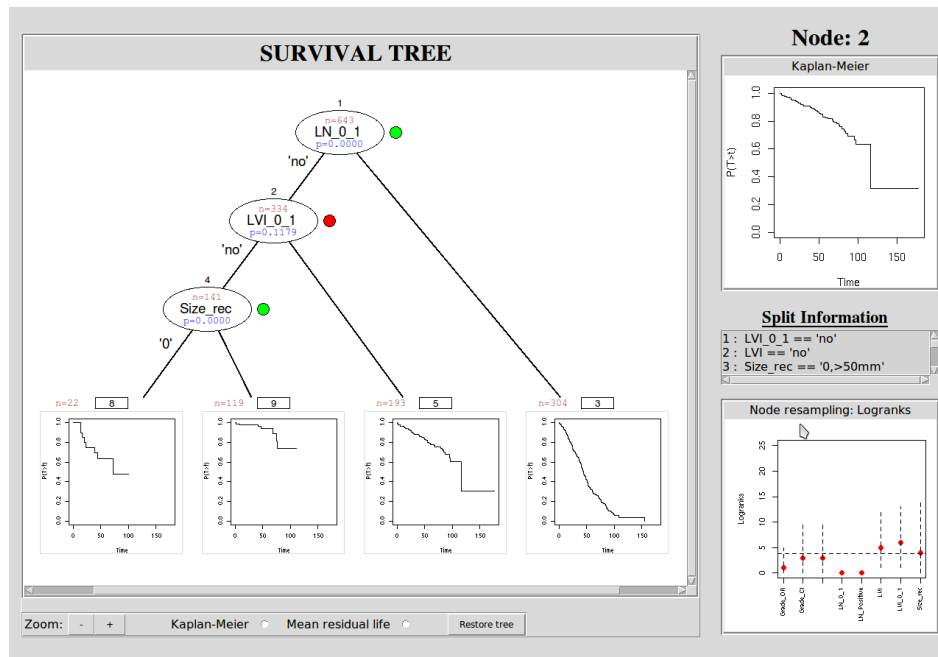


Figure 4.32: A snapshot of the graphical user interface after a survival tree has been grown for the breast cancer dataset (recurrence).

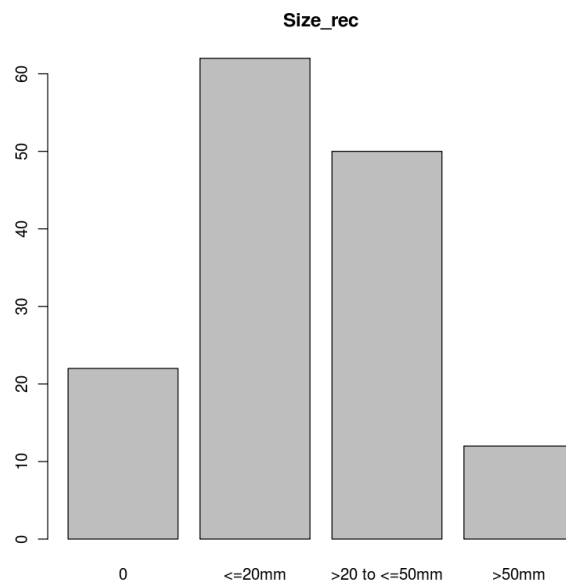


Figure 4.33: Distribution of “Size_rec” at node 4 in figure 4.32.

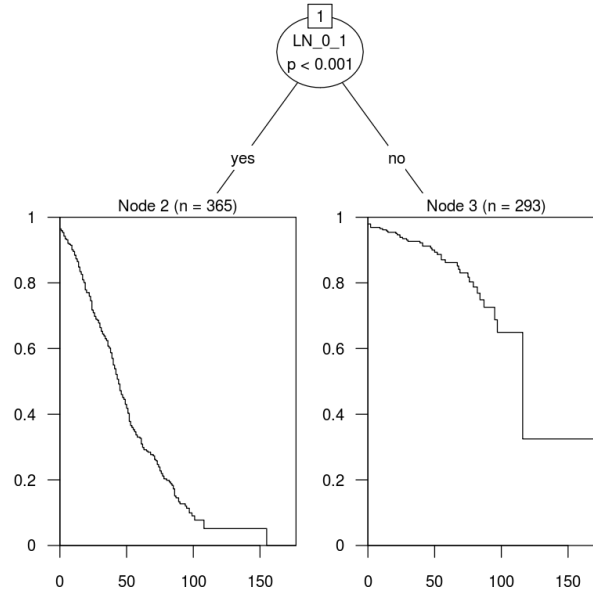


Figure 4.34: Survival tree based on unbiased recursive partitioning. Pathological variables, breast cancer dataset (recurrence).

explained by the fact that patients with no tumor might still have residual cancer cells that cannot be detected unless the appropriate tests are carried out. As a result of this, it is likely that patients in this group were not treated correctly, with the corresponding worsen of the survival outcome.

To compare the results of this analysis with those obtained using unbiased recursive partitioning, Figure 4.34 shows the survival tree applying this method. As one can see, only one split was included in the model. This highlights again the danger of using the results of the statistical test to stop growing the tree. The split based on `LVI_0_1` is not significant and, therefore, the algorithm stops growing the tree any further. However, as shown in the node re-sampling tree another ‘relevant’ split is identified below, which is also a significant split. The tree generated using CART was identical to that depicted in Figure 4.34.

To identify the relevant set of biomarkers, the node re-sampling algorithm was used to generate a survival tree. Figure 4.35 shows the tree after all the ‘irrelevant’ splits have been eliminated (the outcome of interest was the recurrence of the disease). As one can see, only `HER2` was identified by the model. This is the same result obtained in Chapter 2 when Cox regression models were fitted to analyze the same dataset. Recall from Chapter 3 that the tree based on unbiased recursive partitioning identified two of the biomarkers, `HER2` and `PR`.

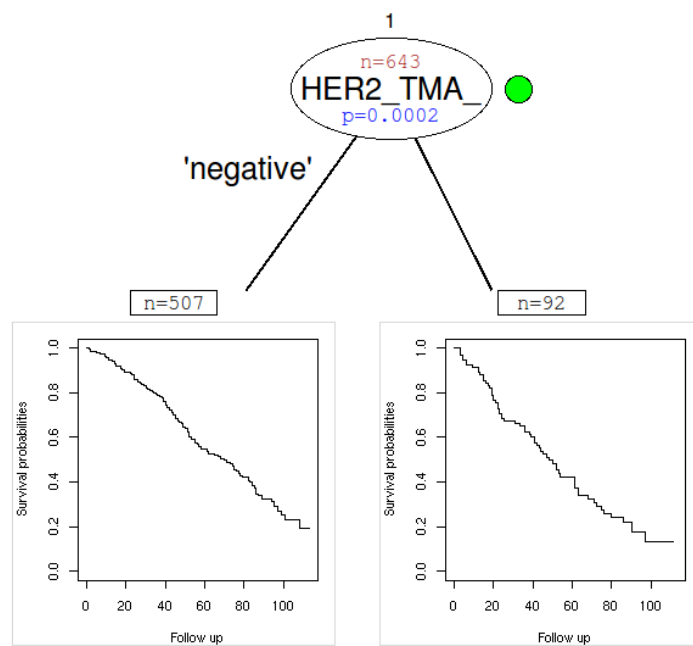


Figure 4.35: Survival tree based on the node re-sampling algorithm. Biomarkers, breast cancer dataset (recurrence).

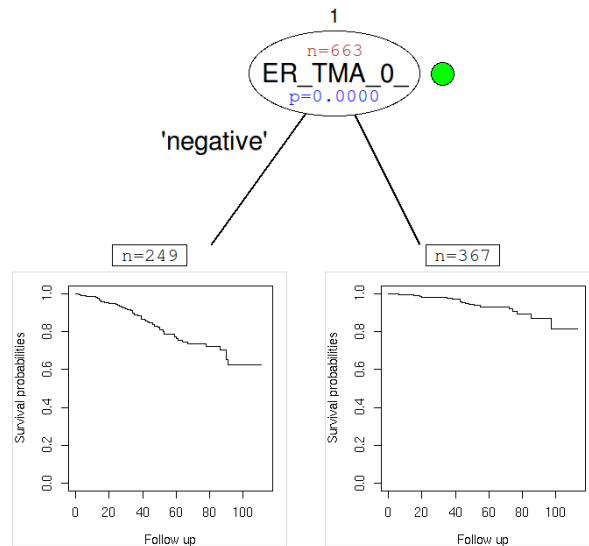


Figure 4.36: Survival tree based on the node re-sampling algorithm. Biomarkers, breast cancer dataset (recurrence).

Finally, the node re-sampling algorithm was also run for the same dataset when the outcome of interest was the death of the patient (overall survival). Figure 4.36 shows the pruned tree. The only covariate identified by this model was ER, which is exactly the same result obtained when applying the unbiased recursive partitioning algorithm (see results in Chapter 3). Recall from Chapter 2 that the Cox proportional hazard model also identified Bcl2 as having a significant effect in the survival outcome.

4.6 Chapter conclusion

A new method has been developed to grow and prune trees. This method uses re-sampling procedures at a node level and a graphical user interface to generate the optimal tree. The development of this method arises from the need for improvement of tree based methods based on the recursive partitioning algorithm. In particular, the vast majority of the algorithms to grow individual trees do not include the sampling variability in the process of generating the tree and are affected by the variable selection bias. On the other hand, methods based on some kind of hypothesis testing for generating the splits, such as trees based on unbiased recursive partitioning, do take into account the sampling variability but have two important fundamental problems. If many predictors are considered for the analysis, many simultaneous test must be performed, and the global significant

level must be corrected. As a result of this, the global test of independence tends to be conservative and some splits may be missed due to the reduction of the significance level. Another problem is related to the presence of interactions in the model. As has been demonstrated in this Chapter, these type of algorithms completely miss such interactions.

The presented method is immune to all these issues. The sampling variability is taken into account by using the information obtained after many bootstrap replicates of the original data have been drawn. This information is used to select the optimal split in each node. The basic idea is to use the recursive partitioning algorithm for each one of the replicates. The splitting criterion is based on the use of the logrank statistic as a measure of dissimilarity between the 2 daughter nodes. The two key components of the node re-sampling algorithm are; the use of the relative importance plot to select the predictors for the primary and surrogate splits and the use of the OOB logranks plot to determine if the split generated at any particular node is ‘relevant’ for the model. In addition, the cutpoints in each split are obtained using the bootstrap distribution of the splitting points. As has been demonstrated in this Chapter, this novel method is not affected by the variable selection bias and it is able to identify interaction effects when these are present in the theoretical model. The coronary and the breast cancer datasets were used to illustrate how the proposed method can be used for the analysis of survival data. When compared with the unbiased recursive partitioning, the method based on node re-sampling generated similar trees and, in some cases, more sensible results. That is the case of the breast cancer dataset when the pathological variables were used to grow the tree (disease free survival). The unbiased recursive partitioning failed to identify a significant split in “size” since a previously generated node gave non-significant results. Another example is the cutpoint of “age” for the coronary dataset. The cutpoint provided by the unbiased recursive partitioning algorithm was 77 years old, which, based on the bootstrap distribution of the cutpoints from the node re-sampling algorithm, is not very likely to be adequate.

Chapter 5

Estimating the mean residual life function

In this Chapter a novel approach for the estimation of the MRL function is presented. The proposed method can be used for the estimation of the MRL function in the terminal nodes of survival trees, which was one of the main goals of this thesis. The Chapter begins with an introduction in which the use of the proposed approach is justified. The second part of the Chapter is devoted to a general overview of the MRL function, its definition, some properties and a discussion about the theoretical settings when censoring is present in the data. In the third part, the problem of estimating the MRL function under non-informative right censoring is considered. The fourth part includes all the details of the novel approach which is based on some results from extreme value theory. Also included here are the results of a simulation study. Finally, in the last part of the Chapter, the application of this novel method to the estimation of the MRL function in the terminal nodes of survival trees based on node re-sampling is presented.

5.1 Introduction

One of the advantages of the use of tree based methods as a modeling tool for analyzing data is their flexibility in the selection of the splitting criterion and the summary statistics in the terminal nodes. Because tree based methods are based on an algorithm to obtain the optimal tree, different splitting criteria and numerical and graphical summaries can be easily incorporated into the model building process of the algorithm. Although there is no preferred option when analyzing survival data, the use of the logrank statistic as a splitting criterion has been adopted by many authors including Segal (1988), Hothorn *et al.* (2006) and Ishwaran *et al.* (2008). Segal (1988) mentions some reasons why the use of log-

ranks may be appropriate (see Chapter 3). Other authors have proposed splitting criterion based on likelihood such as Ciampi et al. (1986), Davis & Anderson (1989) and LeBlanc & Crowley (1992) (see Chapter 3).

Less attention has been paid to the summary statistics of a node, in particular to the graphical and numerical summaries of the terminal nodes. It seems that a wider choice of such summaries could enhance tremendously the functionality of trees as a tool to produce interpretable results. The range of possible choices depends to some extent on the splitting criterion. For instance, when log-ranks are used, median survival times are usually the default choice, as in Segal (1988) and Hothorn *et al.* (2006). Splitting criteria based on the likelihood tend to use relative risks measures as in LeBlanc & Crowley (1992) where proportional hazards are assumed and estimates of the hazard ratios are given as the node summaries. In recent years, the implementation of some of these methods in statistical packages such as R (R Core Team, 2012) has made it possible to display graphical summaries in the terminal nodes in addition to the corresponding numerical summaries. When analyzing survival data the preferred option almost unanimously is a plot of the estimated survival function.

Here we propose the use of the mean residual life (MRL) function as a terminal node graphical summary. This novel approach intends to extend the capabilities of survival trees as a modeling tool that is able to generate results that are very easy to interpret. Recall from Chapter 2 that the MRL function at any time t gives the expected remaining life time of a patient given that the patient survived up to time t . Furthermore, it completely characterizes the distribution and in that sense is equivalent to other functions such as the survival or hazard function. The mean residual life function has been used traditionally in engineering and reliability but the advantages of using it when analyzing survival data from medicine are clear. For example:

- The MRL function is the only function that summarizes survival in terms of time rather than as a probability or a hazard as in the survival or hazard functions respectively. This characteristic of the MRL function is particularly interesting when one tries to communicate the results of the statistical analysis to doctors and patients.
- The MRL function describes very nicely the survival experience of the individual under investigation (see Figure 5.1). For instance, patients with a stable condition will exhibit a constant MRL function. Increasing MRL corresponds to patients recovering from their condition (transplant, operation, etc) whereas decreasing MRL describes the survival experience of patients for which the condition is deteriorating. Finally, patients with infectious disease, such as tuberculosis, have a period of decreasing MRL until they

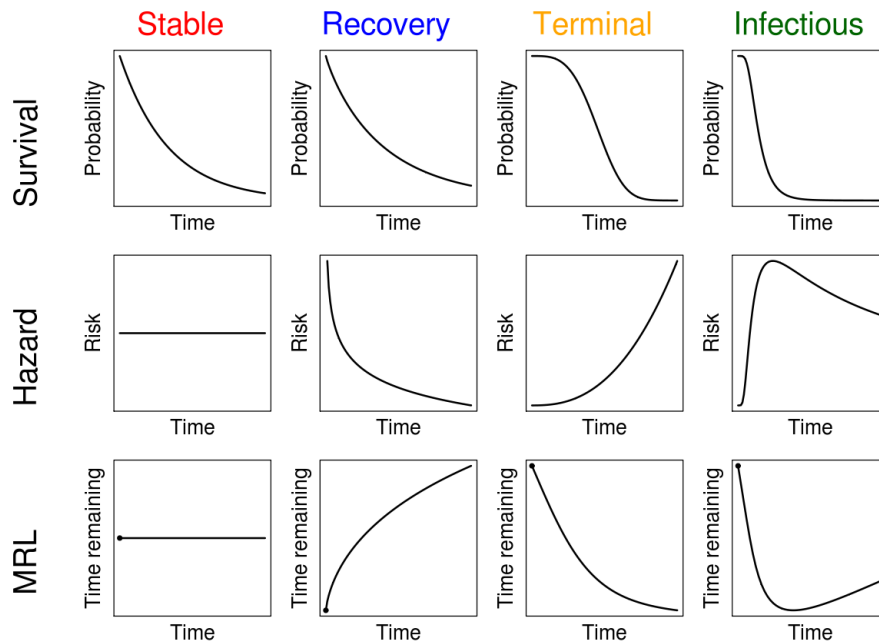


Figure 5.1: Examples of different survival experiences.

get treatment and after that, the expected remaining life time increases over time.

- Numerical summaries can be obtained straight from a plot of the MRL function. In this regard, the mean life time is just the value of the MRL at time 0. At any other time t the MRL is the expected remaining life time given that the patient is still event free at time t .

To illustrate in more detail the justification for this novel approach I will refer to a paper by Spruance *et al.* (2004) in which they express their concern about the misinterpretation of the estimates of the hazard ratios obtained from a Cox proportional hazard model. They state: “*We have observed substantial confusion among clinicians as to the meaning of hazard ratio. For many clinicians, hazard ratio is a relative speed. Words found in the literature that describe the effect of treatments on the resolution of viral diseases when the hazard ratio was significantly greater than one have included 'accelerated time [hazard ratio shown]', 'resolved [hazard ratio shown] times faster', 'hazard ratios indicate a 1.3 to 1.5-fold faster time', 'more than twice as fast [hazard ratio = 2.13]' and 'healing time was 15% shorter [hazard ratio = 0.85]'*”. They show different examples in which the ratio obtained by the estimated hazard ratio (\hat{HR}) may not necessarily correspond to the ratio of the typical values. In other words, if $\hat{HR} = 2$, that does not

mean that, on average, patients of one group are going to live two times longer than patients in the other group. Another interesting point is made in the last part of the paper where questions by patients and recommended responses by the physicians are discussed. One such question is “*Doctor, when will I heal if I use the new drug?* ”; the recommended answer is: “*The study showed that about half the people who used the new drug healed within 4 days, and 95% healed within 8 days. Your experience will vary, like those of the people in the study, because of the natural variation in severity characteristic of this illness.*”. These examples illustrate the difficulties in communicating the results of the statistical analysis first from the statistician to the the clinician and secondly from the clinician to the patients.

The use of survival trees with the MRL function in the terminal nodes can greatly help to overcome the difficulties described above. The possibility of plotting the model as a survival tree facilitate enormously the task of understanding the very often complicated relationships between the predictors and the response. In addition, the MRL function in the terminal nodes would easily answer the question of how long am I going to live? or when will I heal?

5.2 The mean residual life function

Let T be the random variable that represents the time to event of a particular individual. The **mean residual life (MRL)** at time t is defined as:

$$\text{MRL}(t) = E(T - t | T > t)$$

where $t \geq 0$. This can be interpreted as follows: given that an individual survived up to time t the MRL is the expected value of the conditional distribution that represents all the individuals who survived up to time t . The concept of using conditional distributions is used by the hazard function too. Whereas the hazard function focuses on what happens in the very short term and gives an idea of the instantaneous risk of death given the survival up to time t , the MRL function focuses on the long term survival experience of the individual and gives the mean value of the conditional distribution. Both, hazard and MRL functions completely characterize the distribution of T . For a review of all the properties of the MRL function see Guess & Proschan (1988).

The following characterization of the expected value of T is very convenient:

$$E[T] = \int_0^{\infty} S(t) dt \tag{5.1}$$

where $S(t)$ is the survival function of T . This characterization is easily proved by

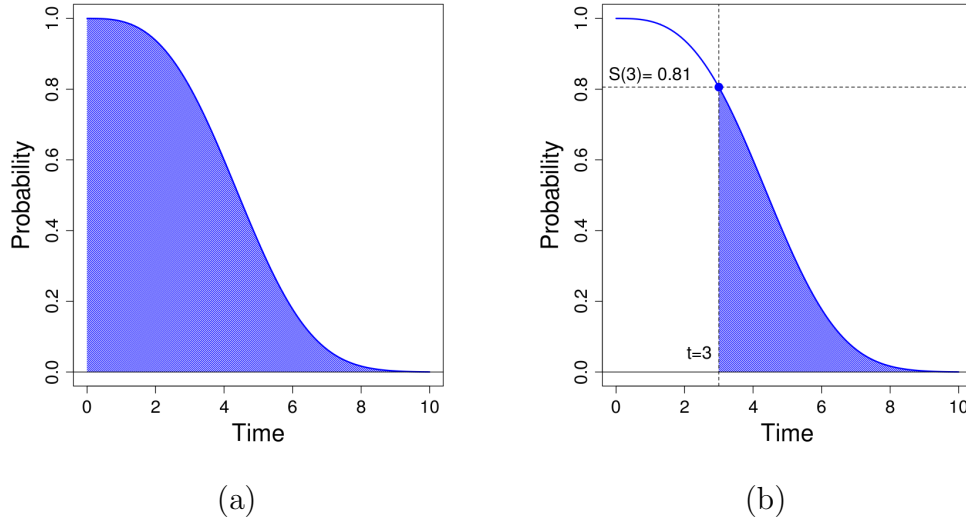


Figure 5.2: Graphic representation of MRL at two different times. In (a) $t = 0$ and $MRL(0)$ is represented by the blue area. In (b) $t = 3$ and $MRL(3)$ is the blue area divided by $S(3) = 0.81$.

interchanging the order of integration in the following expression:

$$\int_0^{\infty} S(t)dt = \int_0^{\infty} \left(\int_t^{\infty} f(x)dx \right) dt = \int_0^{\infty} f(x) \int_0^x 1dt dx = \int_0^{\infty} xf(x)dx = E[T]$$

where $f(t)$ is the density function of T . In graphical terms, $\int_0^{\infty} S(t)dt$ represents the area under the survival function as represented in Figure 5.2 (a).

This interpretation of the expected value of T is useful because it allows one to use the survivor function (or the cumulative distribution function) instead of the density function in order to calculate the expected value. One can extend this result to the MRL function by taking into account the fact that the survivor function of the conditional distribution $Y = T - t | T > t$ is:

$$S_Y(x) = \frac{S(t+x)}{S(t)} \quad x \geq 0$$

The MRL at time t is, therefore,

$$MRL(t) = \int_0^{\infty} \frac{S(t+x)}{S(t)} dx = \frac{1}{S(t)} \int_t^{\infty} S(u)du \quad (5.2)$$

An example is given in Figure 5.2 (b). At time $t = 3$ the $MRL(3)$ is the blue area divided by $S(3)$ which is 0.81 in this example.

Yang (1978) was the first to propose estimating the MRL function using the empirical survivor function in the complete data case (no censoring present). If (t_1, t_2, \dots, t_n) is a random sample of T the MRL function can be estimated by

$$\widehat{\text{MRL}}(t) = \frac{1}{S_n(t)} \int_t^\infty S_n(u) du 1_{\{t_{(n)}-t\}} \quad (5.3)$$

where $t_{(n)}$ is the maximum observed time, $S_n(t) = 1/n \sum_{i=1}^n 1_{\{t_i > t\}}$ is the empirical survivor function and $1_{\{\cdot\}}$ is the indicator function. Yang (1978) showed that (5.3) is asymptotically unbiased, uniformly strongly consistent and converges in distribution to a Gaussian process. It is easy to see that if $t < t_{(n)}$,

$$\begin{aligned} \widehat{\text{MRL}}(t) &= \frac{1}{\sum_{i=1}^n 1_{\{t_i > t\}}} \int_t^\infty \sum_{i=1}^n 1_{\{t_i > u\}} du = \\ &= \frac{1}{\sum_{i=1}^n 1_{\{t_i > t\}}} \sum_{i: t_i > t} \int_t^{t_i} 1_{\{t_i > u\}} du = \\ &= \frac{1}{\sum_{i=1}^n 1_{\{t_i > t\}}} \sum_{i: t_i > t} \int_t^{t_i} 1 du = \frac{\sum_{i: t_i > t} (t_i - t)}{\sum_{i=1}^n 1_{\{t_i > t\}}} \end{aligned}$$

which corresponds with the average of the excess times $(t_i - t)$ of those individuals for whom $t_i > t$. Confidence intervals can be calculated for the point estimates in the usual way by $\pm t_{\frac{\alpha}{2}, n_t-1} SE(\widehat{\text{MRL}}(t))$ where $SE(\widehat{\text{MRL}}(t))$ is the estimated standard error of the estimate. An example is given in Figure 5.3 where data were simulated from a gamma distribution with parameters $k = 2$ and $\theta = 3$ and sample size $n = 1000$. The blue curve represents the true MRL function and the black curve the estimated MRL function with the corresponding 95% point-wise confidence intervals.

Guillamón *et al.* (1998) proposed the use of a kernel smoothing based estimator of the MRL function. By noting that

$$\text{MRL}(t) = \frac{\int_t^\infty u f(u) du}{S(t)} - t$$

they proposed kernel estimators for both $\int_t^\infty u f(u) du$ and $S(t)$ in order to get a smooth estimate of the MRL function. They established the asymptotic normality of such an estimator and compared the results obtained with the empirical estimator.

Chaubey & Sen (1999) considered a different approach based on the classical Hille (1948) theorem (on uniform smoothing in real analysis) that can be used to obtain smooth estimators of $S(t)$ and $f(t)$. For a given random sample

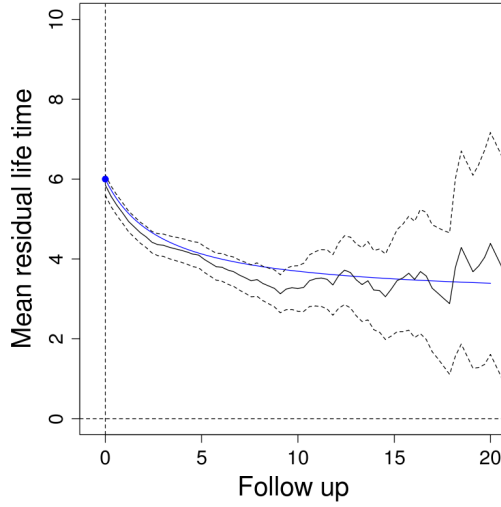


Figure 5.3: An example of the estimation of the MRL function when no censoring is present in the data. The blue line corresponds with the true MRL function.

(t_1, t_2, \dots, t_n) a smooth estimator of $S(t)$ can be obtained by

$$\hat{S}(t) = \sum_{k=0}^n w_{nk}(t\lambda_n) S_n\left(\frac{k}{\lambda_n}\right) \quad (5.4)$$

where $\lambda_n > 0$ is chosen so that $\lambda_n \rightarrow \infty$ but $n^{-1}\lambda_n \rightarrow 0$ as $n \rightarrow \infty$. The weights $w_{nk}(y)$ are defined as

$$w_{nk}(y) = \frac{\frac{y^k}{k!}}{\sum_{j=0}^n \frac{y^j}{j!}}$$

where $0 \leq k \leq n$. They showed that with these weights the integral in the numerator of (5.2) diverges when $S(t)$ is estimated by $\hat{S}(t)$. In order to get appropriate estimates of the MRL function using this approach they proposed different weights in (5.4):

$$w_{nk}(t\lambda_n) = e^{-t\lambda_n} \frac{(t\lambda_n)^k}{k!}$$

where $k = 0, 1, 2, \dots, \infty$.

All of the above methods assume that complete information about the failure time is obtained for all the individuals in the sample. However, in the biomedical sciences censored data are typical and, therefore, complete information is available

only for some of the individuals. Recall from Chapter 2 that random censoring is usually associated with medical studies and some of the reasons for which data of this type are collected with censoring are:

- *Loss to follow up.* The individual under investigation disappears without any explanation. The clinicians never see him/her again.
- *Drop out.* The study has to finish prematurely for a particular individual due to some adverse effects of the treatment. Another cause is that the patient refuses continuing the treatment for whatever the reasons.
- *Termination of the study.* The individual did not experience the event of interest before the study ends.

The rest of the Chapter is dedicated to the estimation of the MRL function under non-informative right censoring. As mentioned in Chapter 2, the theoretical setting of this type of censoring assumes that each individual has two associated times, the true time when the event of interest occurs (failure time) and the censoring time. More formally, let T and C be the failure and censoring times respectively and let them be independent of each other. The observed times are $X = \min(T, C)$ and the censoring indicator is defined as $\delta = 1_{\{T < C\}}$. For a particular individual i the information that can be observed is the tuple (X_i, δ_i) . Under this setting, a typical random sample of size n looks like $\{(x_1, \delta_1), (x_2, \delta_2), \dots, (x_n, \delta_n)\}$, where δ_i is 1 if the event of interest was observed at time x_i and 0 if x_i is the time in which individual i was last seen having not experience the event. This theoretical setting will be referred to as **Theoretical Setting I (TSI)**. However, this setting does not take into account the fact that studies are carried out only for a limited period of time. TSI accommodates censoring times due to lost to follow up and drop out, but not censoring times due to the study termination. To include these, a different theoretical setting has to be considered. This will be referred to as **Theoretical Setting II (TSII)**. Under this setting, the observed times are $X = \min(T, C, T_m)$ and the censor indicator is $\delta = 1_{\{(T < C) \cap (T < T_m)\}}$ where T_m represents the maximum observed time or length of the study. In this case, $\delta = 1$ if the failure time occurs before the censoring times and both occur before the end of the study T_m . If the failure time occurs after T_m the observation is censored and $x_i = T_m$.

5.3 Estimating MRL under non informative right censoring

The methods described above can be extended to the non informative right censoring case by means of replacing $S(t)$ by an appropriate estimate in equation (5.2).

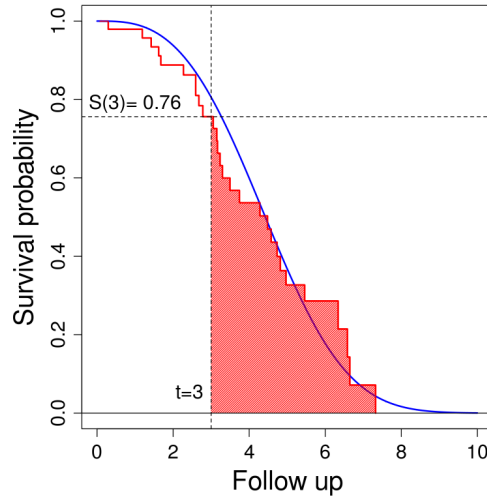


Figure 5.4: An example of the estimate of the MRL at time 3 using the Kaplan-Meier estimate of the survivor function. The estimate is the red area divided by 0.79.

One approach under TSI is to use the Kaplan-Meier estimator of $S(t)$. The estimate of the MRL function following this approach is thus,

$$\text{MRL}_{KM}(t) = \frac{\int_t^\infty S_{KM}(u) du}{S_{KM}(t)}$$

where KM stands for the Kaplan-Meier estimator. Figure 5.4 shows an example of such an estimate of the MRL function at time $t = 3$. The estimate would be the red area divided by $S_{KM}(3) = 0.76$. Gill (1983), among others, studied the asymptotic properties of this estimator.

More recently, Zhou & Jeong (2011) use the same estimator of the MRL function but derive confidence intervals obtained by an easy application of the general empirical likelihood ratio test. This approach has the advantage that there is no need to estimate the variance at all.

Chaubey & Sen (2008) have proposed an extension of their method (Chaubey & Sen, 1999) to the case of non-informative right censoring. They established strong uniform consistency and asymptotic normality and claim that such an estimator does not suffer from boundary bias as in the case of standard kernel smoothing. The estimator is the following:

$$\hat{\text{MRL}}(t) = \frac{\int_t^\infty \hat{S}(u) du}{\hat{S}(t)}$$

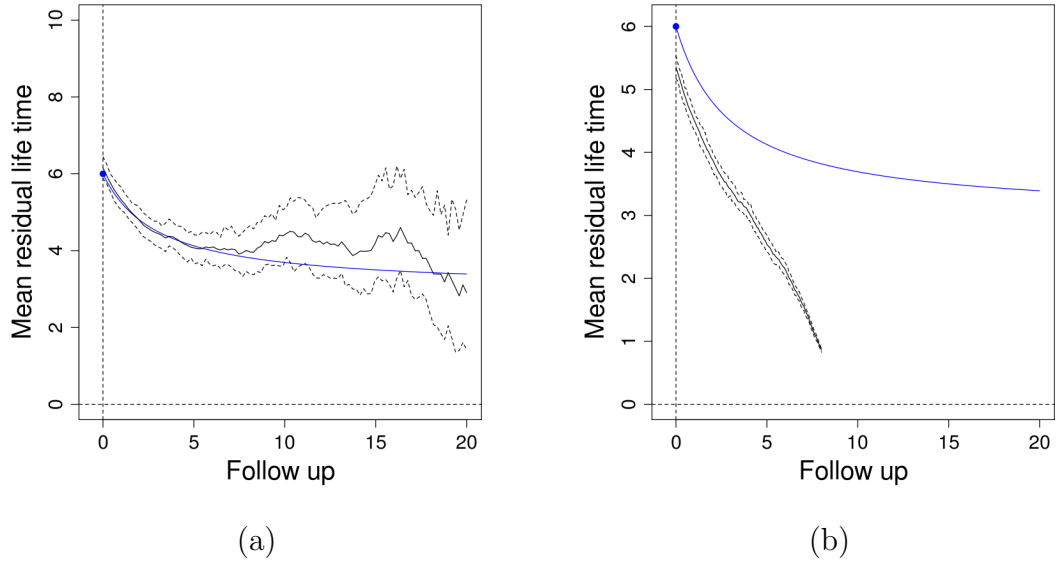


Figure 5.5: Estimates of the MRL function based on the Kaplan-Meier estimate of the survival function with corresponding 95% bootstrap confidence intervals. In (a) data were simulated using TSI whereas in (b) data were simulated using TSII (true MRL function in blue).

where

$$\hat{S}(t) = \sum_{k=0}^{\infty} S_{KM} \left(\frac{k}{\lambda_n} \right) p_k(\lambda_n t)$$

and

$$p_k(\lambda_n t) = e^{-\lambda_n t} \frac{(\lambda_n t)^k}{k!}, \quad k = 0, 1, 2, \dots, \infty$$

with λ_n being a constant to be chosen suitably. Under this setting $S_{KM}(t)$ is set to 0 for $t = t(n)$ where $t(n)$ is the maximum observed time.

All these methods use TSI to incorporate the censoring in the process of estimating the MRL function. However, under TSII, which assumes that the study has been carried out for a limited period of time, the estimates obtained using the above methods fail to give appropriate estimates. An example is given in Figure 5.5 where the Kaplan-Meier estimate of $S(t)$ is used for the estimation of the MRL function using the two theoretical settings. As one can see in (b) the method fails to give appropriate estimates.

To understand why the method fails, Figure 5.6 shows an example in which data are generated under TSI (left) and TSII (right). The problem under TSII is that the numerator in (5.2) cannot be estimated because, due to the termination

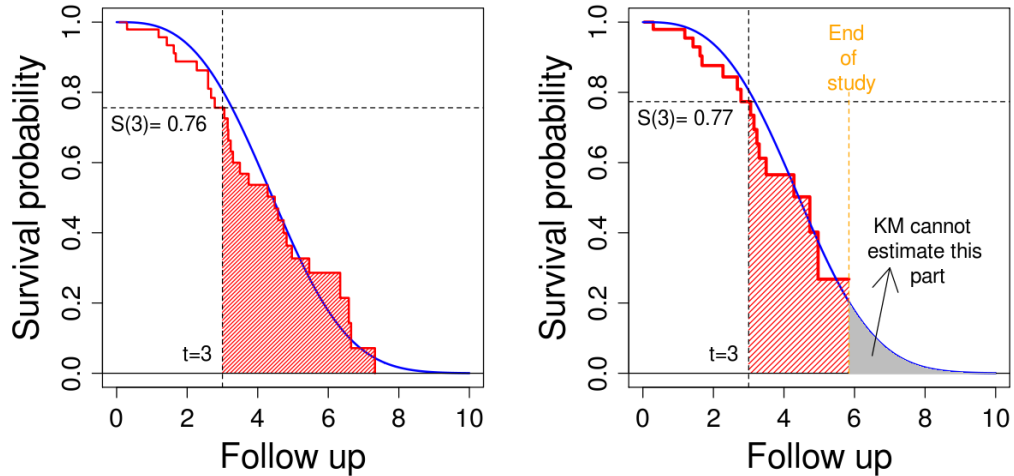


Figure 5.6: An example of the Kaplan-Meier estimate of $S(t)$ under TSI (left) and TSII (right). Whereas under TSI it is possible to get good estimates of MRL, under TSII it is not possible due to the fact that the gray area (right) cannot be estimated.

of the study, the Kaplan-Meier estimate of $S(t)$ does not give any information about the missing gray area (Figure 5.6 right).

As Shen *et al.* (2010) notice only a few papers have focused on the estimation of the MRL function when this type of censoring due to the termination of the study is also present in the data. Guess & Park (1991) proposed the use of conservative non-parametric confidence bounds for the MRL function based on the formula $MRL'(t) = MRL(t)\lambda(t) - 1$ where $\lambda(t)$ is the hazard function (this formula can be easily derived from (5.2)). The idea is that if $MRL'(t) > 0$ then $MRL(t) > 1/\lambda(t)$ and one can use $\lambda(t)$ to obtain lower bounds for the MRL function. This approach based on inverted confidence bounds on the hazard function is valid for both, increasing and decreasing MRL functions and can be used for right random censoring data even when the right tail of the survival function is missing. Shen *et al.* (2010) proposed a method for the estimation of the decreasing MRL function based on an initial estimation of $\mu = MRL(0)$. The method is based on the formula

$$MRL(t) = \frac{\mu - \int_0^t S(u)du}{S(t)} \quad (5.5)$$

which follows from (5.2). If μ is given, estimates of $MRL(t)$ are calculated using the Kaplan-Meier estimate of $S(t)$. An appropriate value for μ can be obtained by

comparing the Kaplan-Meier estimate of $S(t)$ and another estimate of $S(t)$ based on the formula

$$S(t) = \frac{\text{MRL}(0)}{\text{MRL}(t)} \exp\left(-\int_0^t \frac{1}{\text{MRL}(x)} dx\right) \quad (5.6)$$

where $\text{MRL}(t)$ is estimated as in (5.5). Formula (5.6) can be obtained solving the differential equation $W'(t) + \frac{1}{\text{MRL}(t)}W(t) = \frac{\text{MRL}(0)}{\text{MRL}(t)}$ which follows from (5.2) with $W(t) = \int_0^t S(u)du$.

Other approaches consist of the extrapolation of the survival function beyond the maximum observed event time T_m . Moeschberger & Klein (1985) proposed several methods (including the use of a parametric model) for completing the Kaplan-Meier estimator of the survivor function. Klein *et al.* (1990) suggested to treat non-parametrically uncensored observations and to use a parametric model for the censored observations. However, as pointed out by Su & Fang (2012) a potential problem of the use of a parametric approach is that the estimated function tends to fit poorly in the tail of the distribution (this will be demonstrated in the next section). They introduced a method that only uses observations that are near T_m to fit an exponential model. In order to identify the observations that are going to be used to fit the model, an algorithm was proposed that searches for a time point where the hazard rate changes significantly. This algorithm is based on a method proposed by Goodman *et al.* (2011) to detect multiple change points in survival functions and to approximating a survival function with piecewise exponential distributions. Gong & Fang (2012) extended the method proposed by Su & Fang (2012) studying the asymptotic properties of the estimate of the mean survival times providing a closed formula for the variance of the estimate.

The rest of the section is dedicated to describing several new attempts for the solution of the problem of estimating the MRL when the right tail of the estimated survival function is missing. In particular, three different approaches were considered. The first one is based on the use of P-splines (Eilers & Marx, 1996) to extrapolate the “trend” of the Kaplan-Meier estimate of the survivor function beyond time T_m . The second solution is based on the estimation of a parametric model given the sample data. The data available are used to get maximum likelihood estimates of the chosen distribution from which an estimate of the MRL function can be obtained. This approach is similar to that proposed by Moeschberger & Klein (1985) which only considered the Weibull distribution for the parametric model. The third solution is based on a new approach using extreme value theory.

5.3.1 Extrapolation of the survivor function using P-splines

Figure 5.7 depicts an example similar to that presented previously but with a smaller sample size of $n = 50$. In order to use (5.2) to estimate the MRL one

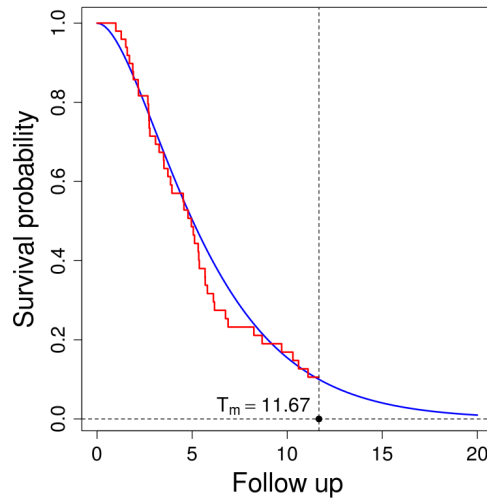


Figure 5.7: An example of the Kaplan-Meier estimate of the survivor function under TSI. The blue line is the true survivor function and the stepwise red line is the corresponding Kaplan-Meier estimate.

needs to extrapolate the Kaplan-Meier estimate of the survivor function beyond T_m . One possible way of doing this is by using some kind of smoothing that allocates basis functions that are situated to the right of T_m where no data are available. Although there are many ways of doing this, the results of the attempts using ordinary P-splines with shape constraints are presented. Other attempts involved the use of polynomials and P-splines with no constraints. Basically, the problem with all these methods is that in order to get accurate extrapolation of the survivor function one has to add a fictitious observation to the right of T_m . In a real life example, if time is measured in years and the event of interest is death, the fictitious observation could be at time $t = 110$ since it is very unlikely that anyone lives more than 110 years. By doing this, however, the extrapolated part of the survivor function depends very much on this fictitious point. Figure 5.8 shows three different possibilities where fictitious points are located at times $t = 20$, $t = 30$ and $t = 40$. In this example, the smoothed curve was constrained to be monotone non-increasing and positive (Bollaerts *et al.*, 2006). One can see how the estimated survivor function after T_m is very different depending on the fictitious point. If one tries to estimate $E[T]$ using the red curve in Figure 5.8 (b) and (c) an overestimate will arise. Any estimators based on (b) and (c) are going to be biased. A more sensible approach would be to assume some kind of parametric shape for the underlying distribution and, using the data available, to

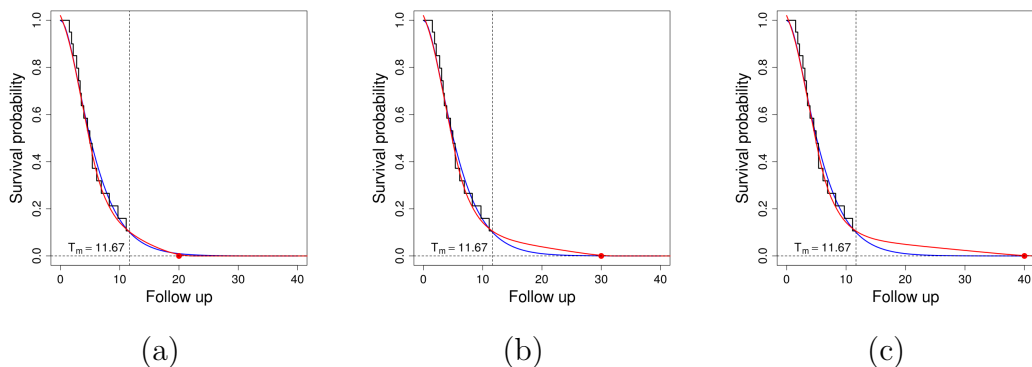


Figure 5.8: An example of extrapolation of the estimated survivor function using P-splines. Three fictitious points were located at $t = 20$, $t = 30$ and $t = 40$ (parts (a), (b) and (c) respectively).

estimate the survival function for the whole range of possible values of T .

5.3.2 A Parametric approach for estimating the MRL function

Assuming a parametric model for the distribution of T brings many advantages and simplifies many things if the distribution of T is, in fact, similar to the model one is assuming. Of course, the question is, what model should one choose? Graphical methods can be used to try to have an educated guess of what model could be more appropriate. Although all this is correct, it does not provide an automatic way of attaining the ultimate goal of this section which is to obtain an unbiased estimate of the MRL function. A solution would be to consider a set of distributions that are used more often in survival analysis and to pick the model that best fits the data. The results presented here include the Weibull, the log-normal and the log-logistic as the set of distributions to be used to estimate the MRL function. Of course this approach could be extended to a wider set of distributions without any significant modification of the method.

The proposed method is as follows. Let $f(t, \theta)$ be the density function of any of the three distributions mentioned before where θ is the vector of the corresponding parameters for each one of the models. Let $\{(x_1, \delta_1), (x_2, \delta_2), \dots, (x_n, \delta_n)\}$ be the observed times and censoring indicators of the random sample. Given the data, the likelihood function under non-informative right censoring can be written as:

$$L(\theta|(x, \delta)) = \prod_{i=1}^n f(x_i, \theta)^{\delta_i} S(x_i, \theta)^{1-\delta_i}.$$

as described in Chapter 2. Estimates of the parameters can be obtained by maximizing the likelihood function for each one of the three models. Let $\hat{\theta}_W$, $\hat{\theta}_{LN}$ and $\hat{\theta}_{LL}$ be the maximum likelihood estimates of the parameters for the Weibull, log-normal and log-logistic distributions respectively. Furthermore, let $\hat{S}_W(t)$, $\hat{S}_{LN}(t)$ and $\hat{S}_{LL}(t)$ be the corresponding estimated survival functions. In order to choose the model that best fit the data one can use ordinary least squares where the goodness of fit can be calculated as:

$$\text{gof}_M = \sum_{i=1}^n \delta_i (S_M(x_i) - S_{KM}(x_i))^2$$

Here the subscript M refers to any of the three models, i.e. $M \in \{W, LN, LL\}$, and S_{KM} is the Kaplan-Meier estimate of the survivor function. Notice that, for convenience, the sum of squares is only calculated for the values x_i that are uncensored (observations i for which $\delta_i = 1$). The model that leads to the smaller value of gof_M is selected for the analysis. Once the model has been chosen, one can get estimates of the MRL function using (5.2). Figure 5.9 illustrates the method. Data were simulated from a gamma distribution and the maximum likelihood estimates of the survivor function for the Weibull, log-normal and log-logistic distributions are represented in black, red and green respectively. By simulating the data using a different distribution, it is possible to evaluate how sensitive the method is to the election of the wrong model. The results obtained are summarized in the following output:

| name | mean | median | gof |
|-------------|------|--------|-------|
| Weibull | 5.75 | 5.07 | 0.111 |
| Lognormal | 5.98 | 5.98 | 0.035 |
| Loglogistic | 6.27 | 4.61 | 0.031 |

In this situation, the log-logistic distribution seems to give the smaller value of gof_M (0.031) and, therefore, will be chosen for the estimate of the mean residual life. Despite the apparent good fit of the three distributions it turns out that the same does not happen when estimating the MRL function as Figure 5.10 demonstrates. This gives an idea of the complexity of the task in hand. For instance, at time $t = 20$ the MRL function of the underlying distribution (gamma) is $\frac{0.0331}{0.0098} = 3.3913$ where $\int_{20}^{\infty} S_G(t)dt = 0.0331$ and $S_G(20) = 0.0098$ (G stands for gamma). The MRL function of the estimated log-logistic (green line) is $\frac{0.4294}{0.0292} = 14.6988$ which is a considerable over estimate.

To assess how the proposed method works in repeated sampling a small simulation was carried out in which 1000 random samples were obtained from the same underlying distribution (gamma with parameters $k = 2$ (shape) and $\theta = 3$ (scale)). The sample size was fixed at $n = 50$. The results are shown in Figure 5.11. The

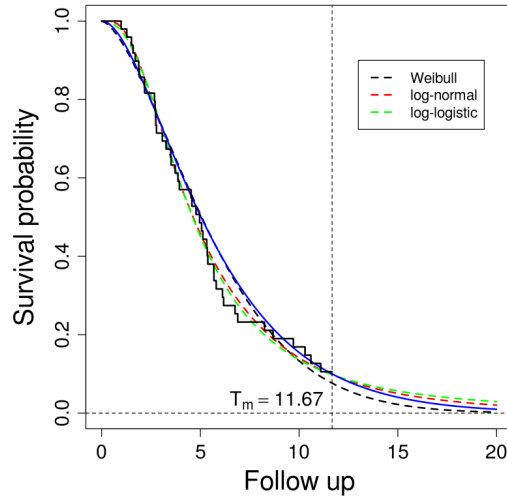


Figure 5.9: An example of three different estimated survivor functions based on different choices of parametric shapes. The black stepwise curve is the KM estimator of the true survivor function (blue line). The dashed lines correspond to the estimated Weibull (black), log-normal (red) and log-logistic (green) distributions.

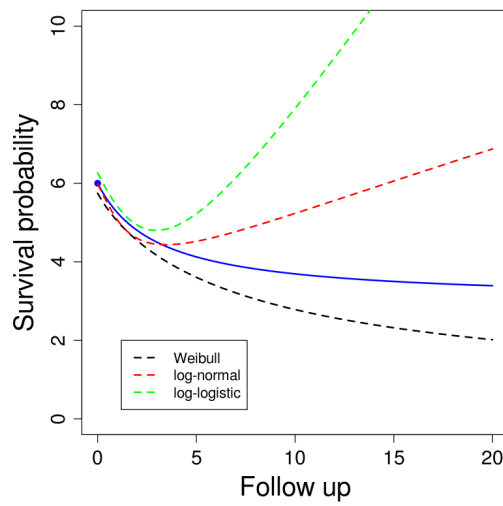


Figure 5.10: MRL functions corresponding to the estimated survivor functions in Figure 5.9. The blue curve is the true survivor function. The other curves correspond to the Weibull (black), log-normal (red) and log-logistic (green) distributions.

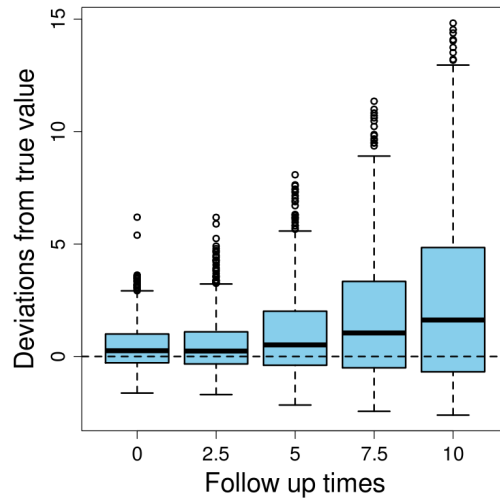


Figure 5.11: An example of the performance of the parametric approach for the estimation of the MRL function. The Y axis represents the deviations between the estimated and the theoretical values.

deviations from the true value of the MRL function and the estimated values using the method described above (Y axis) were determined at five different time points (X axis). As one can see, the method seems to give biased results in all five points and the variability increases as the follow up times increase.

To summarize the results obtained in this section, the two methods proposed for the estimation of the MRL function seem to fail to give appropriate estimates of such a function. The method based on P-splines is unable to determine what the behavior of the missing tail is, since, beyond the maximum observed time, no information related to the survival probabilities is available in the sample. As a result of this, the estimates of the MRL function will be affected. The method based on the parametric approach seems to provide good estimates of the survival function but fails to give good estimates of the MRL function as has been demonstrated in the example. In the next section a novel approach based on extreme value theory is presented.

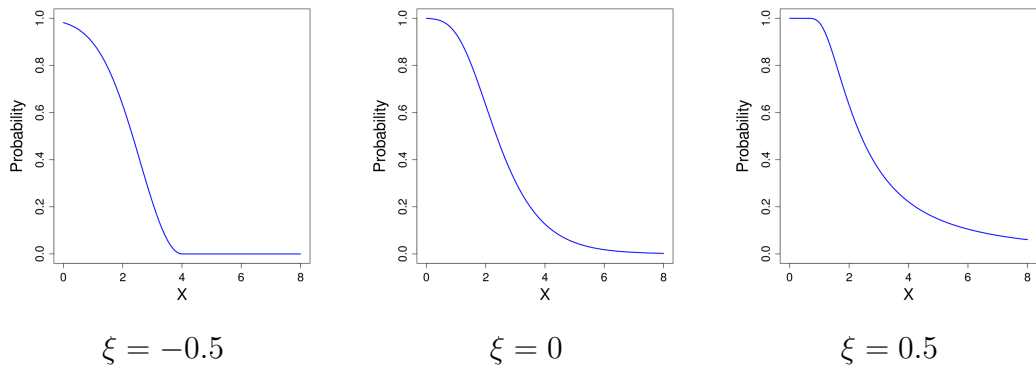


Figure 5.12: Plot of the survival function of the GEV distribution for different values of the shape parameter.

5.4 A semi-parametric approach based on extreme value theory

Extreme value theory is concerned with estimating the probability of highly unusual events. Because they are very unusual they are not likely to be captured in a particular random sample. A core result in extreme value theory, due to Leadbetter (1983), is that if there exists sequences of constants $a_n > 0$ and b_n , such that $\frac{M_n - b_n}{a_n}$ converges to $G(x)$ in distribution where $G(x)$ is non-degenerate and M_n is the sample maximum then $G(x)$ follows a generalized extreme value distribution (GEV). This distribution has 3 parameters μ (location), $\sigma > 0$ (scale) and ξ (shape) and its distribution function is:

$$G(x) = e^{-[1 + \xi(\frac{x - \mu}{\sigma})]_+^{-1/\xi}}$$

where $x > 0$ and $[u]_+ = \max(u, 0)$. The value of ξ is what determines the behavior of the tail of the distribution.

- If $\xi > 0$ heavy upper tail.
- if $\xi = 0$ exponential upper tail.
- if $\xi < 0$ short tail with a finite upper end-point.

Figure 5.12 shows the survival function for different values of the shape parameter (in this example $\mu = 2$ and $\sigma = 1$). These results provide an asymptotically justified model for the maximum of a random sample. The sampling distribution of the maxima defines somehow the behavior of the upper tail of the underlying distribution. In general, given a random sample (x_1, x_2, \dots, x_n) and a threshold

u , one can think of the observations $x_i \geq u$ as a sample of the maximum and use those observations that exceed the threshold to estimate the parameters of the GEV distribution.

In order to apply these ideas to the problem of estimating the MRL function, suppose one has a random sample of observed times (t_1, t_2, \dots, t_n) (to simplify the notation for now it is assumed that there is no censoring, although the ideas explained here will be extended when censoring is present). For a particular threshold u , consider the conditional distribution $T_u = T - u | T > u$. The survivor function of T_u is

$$S_{T_u}(t) = P(T - u > t | T > u) = P(T > u + t | T > u) = \frac{1 - F_T(u + t)}{1 - F_T(u)}$$

where $t \geq 0$ and $F_T(t)$ is the distribution function of T . If the threshold u is “large”, and there are m observations larger than u , one can approximate $F_T(t)^m$ (distribution function of the sample maximum) by the GEV distribution. Furthermore, due to the max-stability property (Leadbetter, 1983) $F_T(t)$ is still a GEV distribution. Thus,

$$S_{T_u}(t) \approx \frac{1 - G(u + t)}{1 - G(u)} = \frac{1 - e^{-[1 + \xi(\frac{u+t-\mu}{\sigma})]_+^{-1/\xi}}}{1 - e^{-[1 + \xi(\frac{u-\mu}{\sigma})]_+^{-1/\xi}}}$$

A further approximation can be obtained using the first order power series of the exponential function ($e^{f(x)} = 1 + f(x) + \dots$),

$$S_{T_u}(t) \approx \frac{[1 + \xi(\frac{u+t-\mu}{\sigma})]_+^{-1/\xi}}{[1 + \xi(\frac{u-\mu}{\sigma})]_+^{-1/\xi}} = \left[1 + \frac{\xi t}{\sigma + \xi u - \xi \mu} \right]_+^{-1/\xi}$$

Define $\sigma_u = \sigma + \xi(u - \mu)$, then

$$S_{T_u}(t) \approx \left[1 + \frac{\xi t}{\sigma_u} \right]_+^{-1/\xi}$$

which is the survivor function of the generalized Pareto distribution (GPD) with scale parameter σ_u (which depends on u) and shape parameter ξ .

Hence, the conditional distribution of $T - u | T > u$ can be approximated by the generalized Pareto distribution when the threshold u is “large”. The MRL function at time u can then be estimated by the expected value of the estimated conditional distribution. The expected value of the GPD as defined above is

$$\frac{\sigma_u}{1 - \xi}$$

In general, the r -th moment of the generalized Pareto distribution is

$$\frac{\sigma_u}{\xi} \sum_{j=0}^r \frac{(-1)^{j+1}}{1 - \xi j}$$

and only exists for $\xi < 1/r$. Of course the obvious question is: how large does u have to be for a good approximation? The answer lies in the usual trade-off between bias and variance. If u is too large, the approximation will be good but not many observations are available for the estimates of the parameters (high variance). On the other hand, if u is too small, more data are available but one might be estimating the wrong model (bias). There are some graphical methods available for choosing the optimal value of u but there is no automatic way of choosing u . Some simulations have shown, however, that for the task of estimating the MRL, a reasonable value for the threshold u might be a value such that between u and T_m there are 20% of the longest event times in the sample. In other words, the threshold u would be the 0.8 quantile of the uncensored observed times (see next section).

5.4.1 Description of the proposed method

Let $\{(x_1, \delta_1), (x_2, \delta_2), \dots, (x_n, \delta_n)\}$ be the observed times and censoring indicators in the random sample. The estimate of the MRL function at time t can be obtained following these steps:

1. Choose the threshold u to be the value such that between u and T_m there are 20% of the longest event times in the sample (0.8 quantile of the uncensored observed times).
2. Calculate the maximum likelihood estimates of the parameters ξ and σ_u of the generalized Pareto distribution (GPD) using only the observations between u and T_m . The survivor function of the GPD is

$$S_{GPD}(t) = \left[1 + \frac{\xi t}{\sigma_u} \right]_+^{-1/\xi},$$

and the density function is

$$f_{GPD}(t) = \frac{1}{\sigma_u} \left[1 + \frac{\xi t}{\sigma_u} \right]_+^{-\frac{1}{\xi}-1}.$$

Therefore, the likelihood function is:

$$L(\xi, \sigma_u; x_i, \delta_i) = \prod_{i=1}^n \left(\frac{1}{\sigma_u} \left[1 + \frac{\xi t}{\sigma_u} \right]_+^{-\frac{1}{\xi}-1} \right)^{\delta_i} \left(\left[1 + \frac{\xi t}{\sigma_u} \right]_+^{-1/\xi} \right)^{1-\delta_i}.$$

This function can be maximized using the usual optimization methods. Some restrictions are important though. For the GPD to have finite first and second moments, the shape parameter ξ must be < 0.5 . Some simulations have shown that for values of ξ being restricted to be ≤ 0.10 the estimates of the MRL function produce biased results. Values of ξ restricted to be ≤ 0.5 also produced very pronounced overestimates of the true values of the MRL function and considerable increase of the estimated standard errors. Restrictions of the shape parameter around 0.25 gave the best results (see next section for a detailed explanation of this point).

3. Estimate $MRL(t)$ using the KM estimate of $S(t)$ and setting $S_{KM}(T_m) = 0$. This estimate will be called $\hat{M}RL_{KM}(t)$.
4. Using the threshold stability property, estimate the mean residual life at time T_m as:

$$\hat{M}RL(T_m) = \frac{\hat{\sigma}_u + \hat{\xi}(T_m - u)}{1 - \hat{\xi}}$$

where $\hat{\xi}$ and $\hat{\sigma}_u$ are the maximum likelihood estimates from step 2.

The threshold stability property states that if a conditional random variable $T-u|T > u$ follows a GPD for some threshold u then for any higher threshold $v \geq u$ the conditional random variable $T-v|T > v$ follows a GPD with the same shape parameter ξ and $\sigma_v = \sigma_u + \xi(v - u)$. Furthermore, the mean residual life evaluated at time v is:

$$MRL(v) = E(T - v|T > v) = \frac{\sigma_u + \xi(v - u)}{1 - \xi}$$

5. Estimate the mean residual life at time t as:

$$\hat{M}RL(t) = \hat{M}RL_{KM}(t) + \frac{\hat{M}RL(T_m)S_{KM}(T_m)}{S_{KM}(t)}$$

Figure 5.13 provides a graphical explanation of step 5. Suppose one is interested in estimating the MRL function at time $t = 4$. The maximum observed time is in this example $T_m = 11.67$ (the border between the red and the blue areas). The red area in Figure 5.13 can be calculated as $\hat{M}RL_{KM}(4)S_{KM}(4)$. The estimate of the MRL at time T_m is $\hat{M}RL(T_m)$ which can be obtained in step 4. Thus, the blue area can be calculated by $\hat{M}RL(T_m)S_{KM}(T_m)$. The estimate of the MRL at time $t = 4$ is the area from 4 to ∞ (i.e. the red area plus the blue area) divided

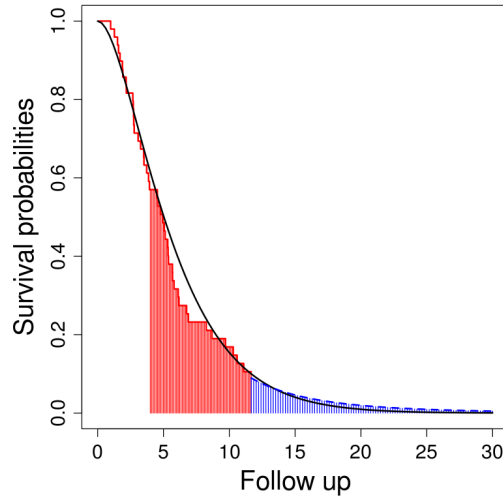


Figure 5.13: Example of how to estimate the MRL function using the GPD distribution. The stepwise red line is the Kaplan-Meier estimate of the survival function. The dashed blue line is the estimated survival function using the GPD distribution.

by $S_{KM}(4)$. This yields:

$$\begin{aligned} \hat{\text{MRL}}(4) &= \frac{\hat{\text{MRL}}_{KM}(4)S_{KM}(4) + \hat{\text{MRL}}(T_m)S_{KM}(T_m)}{S_{KM}(4)} = \\ &= \hat{\text{MRL}}_{KM}(4) + \frac{\hat{\text{MRL}}(T_m)S_{KM}(T_m)}{S_{KM}(4)} \end{aligned}$$

which is the formula used in step 5.

Figure 5.14 shows an example of the estimated MRL function using this method. As one can see, there is more variability than that obtained when the Kaplan-Meier estimate of the MRL function was used with data that had no censoring due to the study termination. The extra variability is due to the fact that, in this case, one has to estimate the missing upper tail of the distribution.

Alternative approach

An alternative method, which is very similar to the one described above, consists in using the maximum likelihood estimates $\hat{\sigma}$ and $\hat{\xi}$ to estimate the MRL function at time u rather than at time T_m . In this case there is no need to use the threshold

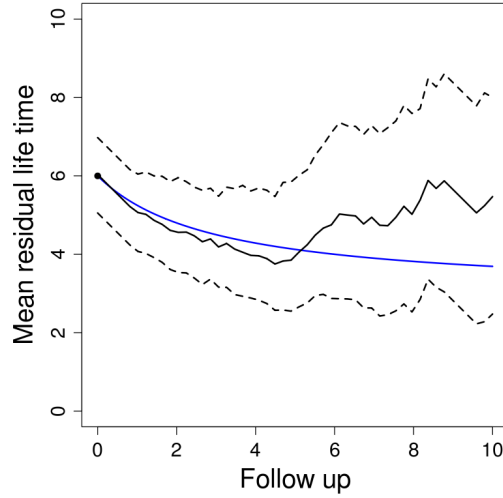


Figure 5.14: The estimated MRL function using the GPD. In addition 95% bootstrap confidence intervals. The blue line is the true MRL function.

stability property in step 4). The alternative proposed method can be summarized in the following steps:

1. Choose the threshold u in the same way.
2. Calculate the maximum likelihood estimates of the parameters ξ and σ_u in the same way.
3. Estimate $\text{MRL}(t)$ as the area from t to u of the KM estimate of the survival function. This estimate will be called $\hat{\text{MRL}}_{KM_u}(t)$.
4. Estimate the mean residual life at time u as:

$$\hat{\text{MRL}}(u) = \frac{\hat{\sigma}_u}{1 - \hat{\xi}}$$

where $\hat{\xi}$ and $\hat{\sigma}_u$ are the maximum likelihood estimates from step 2.

5. Estimate the mean residual life at time t as:

$$\hat{\text{MRL}}(t) = \hat{\text{MRL}}_{KM_u}(t) + \frac{\hat{\text{MRL}}(u)S_{KM}(u)}{S_{KM}(t)}$$

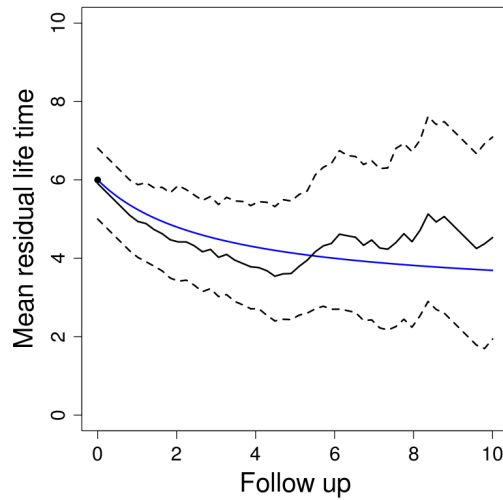


Figure 5.15: The estimated MRL function using the alternative approach based on the GPD. In addition 95% bootstrap confidence intervals.

An example is given in Figure 5.15 in which the alternative proposed method was applied to the same data used in Figure 5.14. The plot suggests perhaps a slight improvement in terms of the variability, especially for values of the MRL function near 10.

To summarize the content of this chapter so far, it has been shown that under TSI the Kaplan-Meier estimate of the MRL function does not give a good estimate of the true MRL function. The reason is that the method does not take into account the missing upper tail of the distribution. Trying to extrapolate that missing part using some kind of smoothing does not work either. A parametric approach seems to be a more sensible choice but, even though some distributions fit the data well, the estimated MRL function differs from that of the underlying distribution, especially for larger values of t . The novel method presented in this Chapter based on extreme value theory uses the generalized Pareto distribution to estimate the missing upper tail of the underlying distribution. Two approaches have been proposed and it seems that both are an improvement on the other methods discussed in terms of obtaining adequate results. To investigate whether this novel method is indeed an improvement, the results of a simulation study will now be presented.

5.4.2 Simulation study

In this section results from a simulation study are presented where the proposed methods are analyzed in terms of their adequacy. Several survival “experiences” are considered corresponding to the different cases described at the beginning of this Chapter. The survival distributions used for the simulations are:

- Patients with a stable condition: Exponential $\lambda = 3$.
- Recovery patients: Gamma $k = 0.7$, $\lambda = 3$.
- Terminal patients: Gamma $k = 2$, $\lambda = 3$.
- Infectious disease patient: Lognormal $\mu = 1$, $\sigma = 0.5$.

Let T be the true failure times and C the censoring times (T and C are assumed to be independent). In the simulations C is assumed to have a uniform distribution $(0, M)$. Furthermore, T_m will denote the termination or duration of the study (also M will be assumed to be greater than T_m). The simulations will be executed for different percentages of censoring. Two types of censoring will be taken into consideration; the censoring due to both lost to follow up and drop out, which will be represented by the region $(C < T) \cap (C < T_m)$ (blue area in Figure 5.16); and censoring due to the termination of the study which will be represented by the region $[(C < T) \cap (C > T_m)] \cup [(T < C) \cap (T > T_m)]$ (red area in Figure 5.16)

Simulation scheme

A formula is now derived in order to identify the choice of M (upper limit of censoring distribution) and T_m (study termination time) for a chosen proportion of censoring. Let A and B be the proportions in the red and blue areas respectively. Then

$$\begin{aligned}
 P(C < T \cap C < T_m) &= \int_0^{T_m} \int_c^\infty f_C(c) f_T(t) dt dc = \\
 &= \int_0^{T_m} f_C(c) S_T(c) dc & (5.7) \\
 &= \frac{1}{M} \int_0^{T_m} S_T(c) dc = B
 \end{aligned}$$

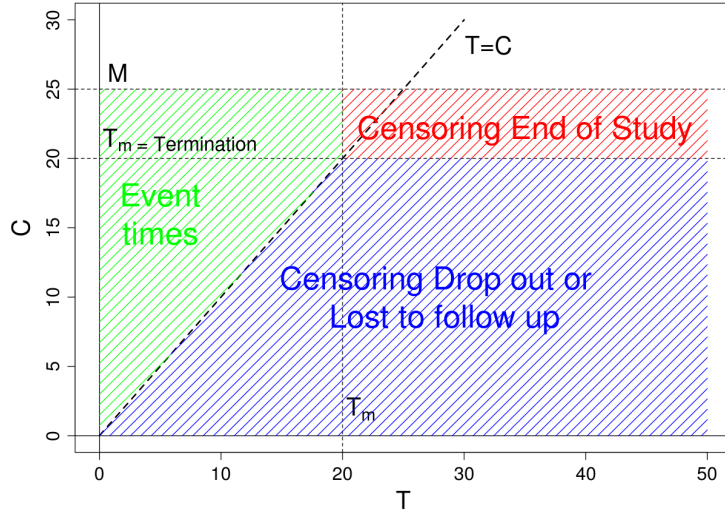


Figure 5.16: Areas where censoring occurs. The blue area corresponds to censoring due to lost to follow up or drop out. The red area corresponds to censoring due to the termination of the study.

On the other hand, $P((C < T \cap C > T_m) \cup (T < C \cap T > T_m))$ (red area in Figure 5.16) is

$$\begin{aligned}
 P((C < T \cap C > T_m) \cup (T < C \cap T > T_m)) &= \\
 &= \int_{T_m}^M \int_{T_m}^{\infty} f_C(c) f_T(t) dt dc = \\
 &= \int_{T_m}^M f_C(c) S_T(T_m) dc & (5.8) \\
 &= \frac{1}{M} \int_{T_m}^M S_T(T_m) dc = \\
 &= \frac{1}{M} S_T(T_m) (M - T_m) = A
 \end{aligned}$$

From (5.7)

$$M = \frac{1}{B} \int_0^{T_m} S_T(c) dc$$

and substituting for M in 5.8, one obtains

$$S_T(T_m) \left(\frac{1}{B} \int_0^{T_m} S_T(c) dc - T_m \right) = A \frac{1}{B} \int_0^{T_m} S_T(c) dc.$$

The latter expression is an implicit function of T_m and can be expressed as

$$f(T_m) = S_T(T_m) \left(\frac{1}{B} \int_0^{T_m} S_T(c) dc - T_m \right) - A \frac{1}{B} \int_0^{T_m} S_T(c) dc = 0$$

Therefore, the root of this implicit equation can be calculated applying Newton-Raphson. Starting at a value T_{m_0} iterate the expression

$$T_{m_{i+1}} = T_{m_i} - \frac{f(T_{m_i})}{f'(T_{m_i})}$$

until convergence.

Once T_m has been obtained, M can be calculated using (5.7).

Example: Suppose one desires to simulate data from a lognormal distribution $\mu = 3$, $\sigma = 2$ with a 10% of censoring due to lost to follow up and drop out and 20% of censoring due to the end of the study. In that case $B = 0.10$ and $A = 0.20$. Following the approach above, $M = 319.76$ and $T_m = 73.07$.

Methods

All the simulations will be executed as follows:

- Draw a random sample of size n from the theoretical distribution of T . This will give the vector (t_1, t_2, \dots, t_n) .
- Draw a random sample of size n from the uniform $(0, M)$. This will give the vector (c_1, c_2, \dots, c_n)
- Obtain the observed values as $x_i = \min(t_i, c_i, T_m)$ for $i = 1, 2, \dots, n$.
- Obtain the censored indicators as $\delta_i = 1_{\{(t_i \leq c_i) \cap (t_i \leq T_m)\}}$ for $i = 1, 2, \dots, n$.

Based on the simulated random sample, estimates of the MRL function will be obtained using the following three methods.

1. Method 1 (M1): Novel method proposed in this Chapter. The generalized Pareto distribution is used to estimate the MRL function at time T_m (termination of the study).
2. Method 2 (M2): Alternative novel method proposed in this Chapter. The generalized Pareto distribution is used to estimate the MRL function at time u (threshold).
3. Method 3 (M3): The MRL function is estimated only based on the Kaplan-Meier estimate of the survival function.

Objectives

There are two main goals in this simulation study. The first one is related to the selection of the threshold u and the restrictions in the maximum likelihood estimate of the shape parameter ξ of the GPD. The second objective is related to the assessment of the adequacy of the proposed methods in terms of unbiasedness and extent of variability of the estimates.

To explore the first goal Figures 5.17 and 5.18 show the results of simulating data under different values of the threshold and different restrictions of the shape parameter. Each point in the plots represents a simulation of 500 draws of sample size $n = 50$ from the underlying distribution. The amount of censoring is determined by the rows and the columns of the matrix of plots. The columns represent the percentage of censoring between 0 and the termination of the study T_m . The rows represent the percentage of censoring due to the termination of the study (which is labeled as “Proportion of upper tail missing”). The different colors represent the different distributions used for the simulations:

- **Red** Exponential $\lambda = 3$ (**Stable**).
- **Blue** Gamma $k = 0.7, \lambda = 3$ (**Recovery**).
- **Brown** Gamma $k = 2, \lambda = 3$ (**Terminal**).
- **Green** Lognormal $\mu = 1, \sigma = 0.5$ (**Infectious**).

Each letter represents a combination of a selected threshold u and a selected restriction of the shape parameter ξ . The correspondences are the following:

- A: Select u as the 0.9 quantile of the observed event times (uncensored observed times). $\xi < 0.10$.
- B: Select u as the 0.8 quantile of the observed event times. $\xi < 0.10$.
- C: Select u as the 0.6 quantile of the observed event times. $\xi < 0.10$.
- D: Select u as the 0.9 quantile of the observed event times. $\xi < 0.25$.
- E: Select u as the 0.8 quantile of the observed event times. $\xi < 0.25$.
- F: Select u as the 0.6 quantile of the observed event times. $\xi < 0.25$.
- G: Select u as the 0.9 quantile of the observed event times. $\xi < 0.50$.
- H: Select u as the 0.8 quantile of the observed event times. $\xi < 0.50$.
- I: Select u as the 0.6 quantile of the observed event times. $\xi < 0.50$.

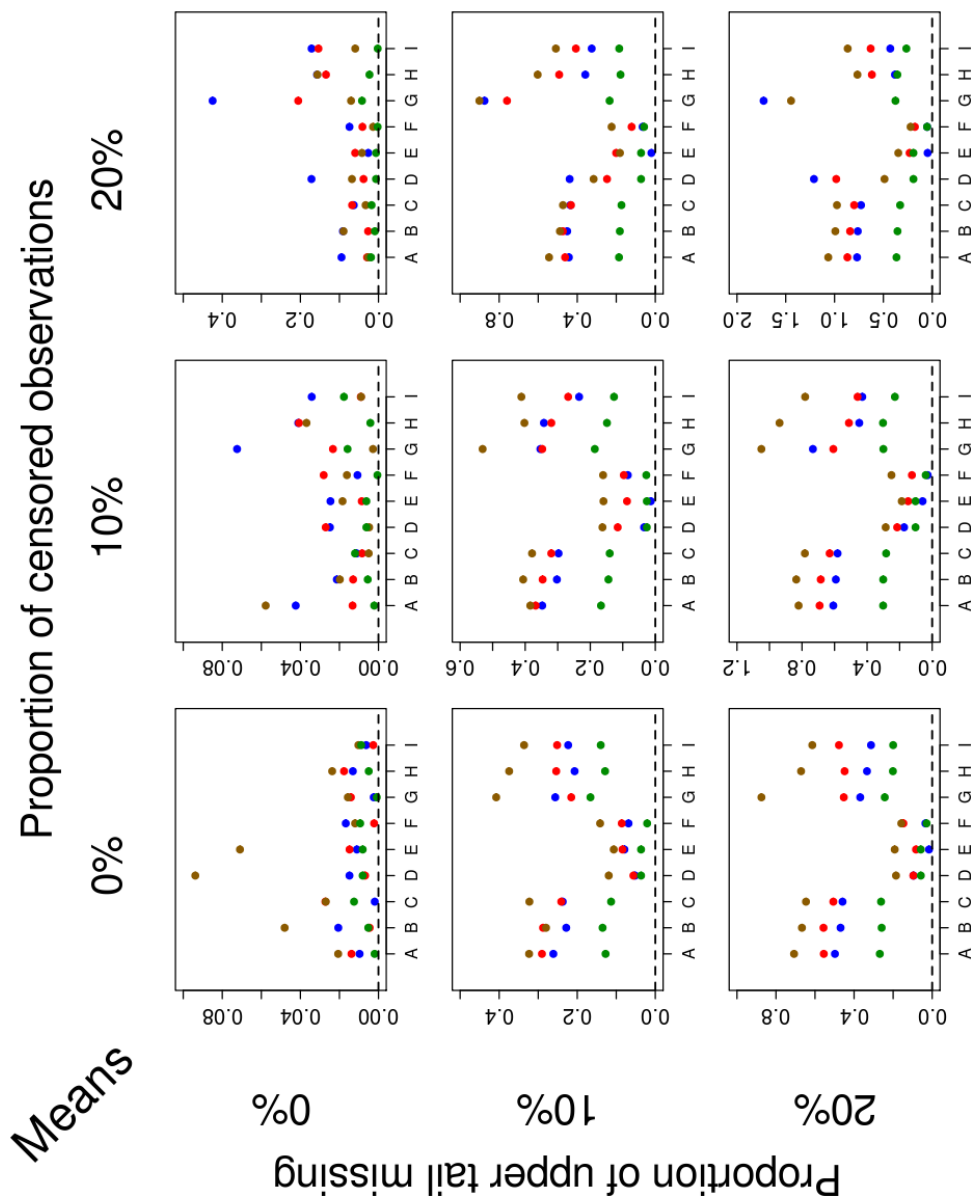


Figure 5.17: Means of the absolute values of the differences between the estimated MRL function at time $t = 0$ and the theoretical value of the MRL function at time $t = 0$.

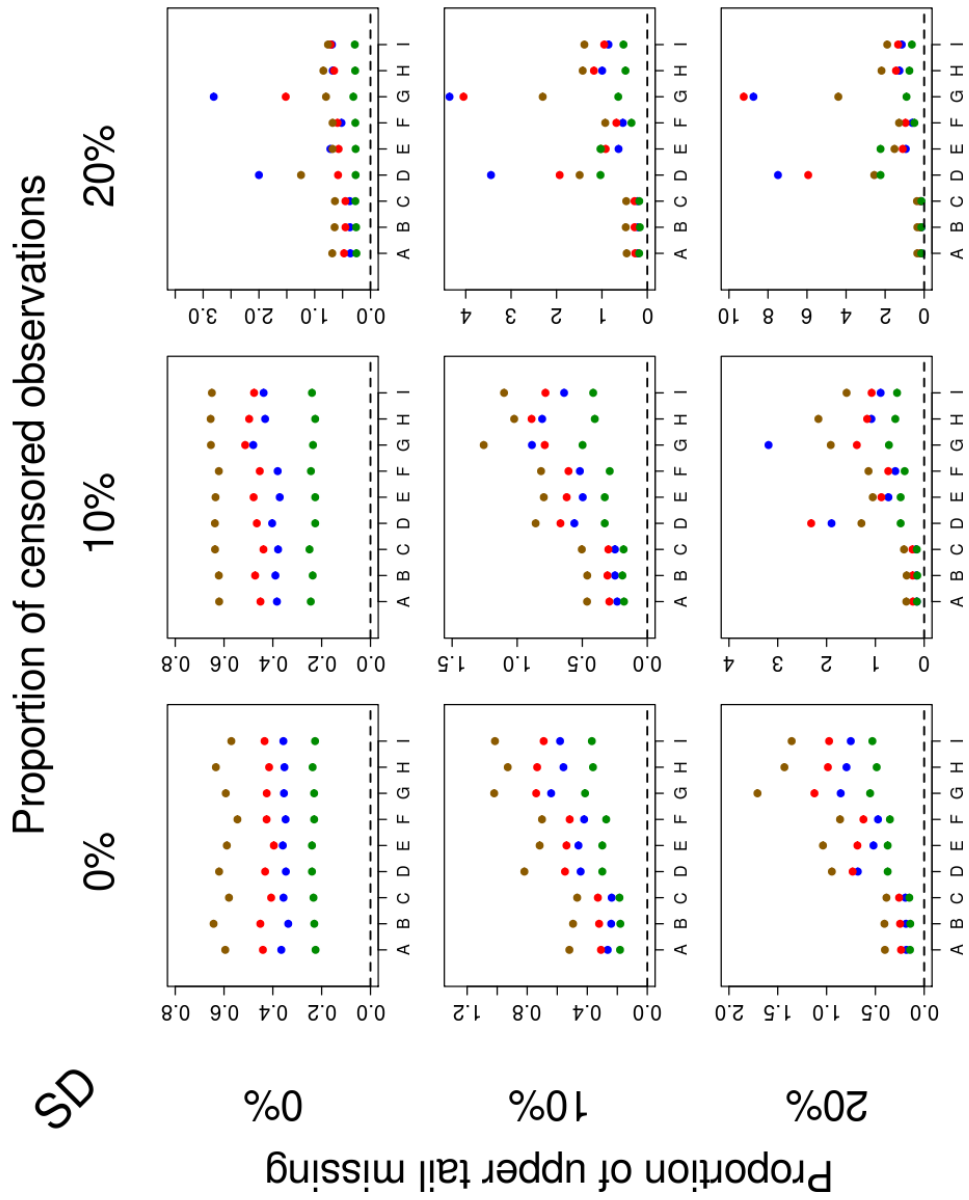


Figure 5.18: Standard deviations of the differences between the estimated MRL function at time $t = 0$ and the theoretical value of the MRL function at time $t = 0$.

For each simulation the differences between the estimated MRL function and the theoretical MRL function are evaluated. In Figures 5.17 and 5.18 the MRL function was only calculated at time $t = 0$. In Figure 5.17 the points are the means of the absolute value of the differences whereas in Figure 5.18 the points are the standard deviations of the differences.

As one can see A, B, C, G, H and I produce, in general, larger values of the mean absolute differences when compared to D, E or F (Figure 5.17). This means that the restriction of the shape parameter $\xi < 0.25$ seem to be the most appropriate. Although not shown here, simulations were also performed for the restriction $\xi < 1$ but the results obtained were even worse. The differences between D, E and F are due to the selection of the threshold u . D's tend to have more variability since less data are available for the estimation (Figure 5.18). On the other hand F's tend to have less variability but slightly higher absolute mean values suggesting higher bias compared to the E's. This is also corroborated by simulations performed for higher sample sizes.

In summary, the simulations presented here suggest that the best results are obtained when the threshold u is chosen such that between u and the termination of the study T_m there are 20% of the longest event times in the sample. The simulations also suggest that the most adequate results are obtained when the shape parameter of the GPD is restricted to be < 0.25 . The rest of the simulations are carried out using these restrictions.

The second goal of the simulations was to assess whether or not the proposed methods produce adequate estimates of the MRL function. Figures 5.19, 5.20 and 5.21 show the results of simulating samples under different sample sizes. Each box-plot in the matrix of plots represents the distribution of the differences between the estimated MRL function and the true MRL function both at time $t = 0$. This distribution is simulated by drawing 500 observations from the underlying distribution. The different distributions are now coded using the letters A, B, C and D:

- A: Exponential $\lambda = 3$ (Stable).
- B: Gamma $k = 0.7, \lambda = 3$ (Recovery).
- C: Gamma $k = 2, \lambda = 3$ (Terminal).
- D Log-normal $\mu = 1, \sigma = 0.5$ (Infectious).

The amount of censoring is determined in the same way explained above. The different colors represent the three methods used for the estimation of the MRL function:

- **Method 1 (M1)**: Novel method proposed in this Chapter.

- **Method 2 (M2)**: Alternative novel method.
- **Method 3 (M3)**: Method based on Kaplan-Meier.

The plots show that the proposed methods, M1 and M2, seem to perform well when compared to M3. The latter method only gives reasonable results when the percentage of censoring due to the termination of the study is 0 (first row in all the plots). Furthermore, methods M1 and M2 seem to provide very similar results across all the different distributions and all the different sample sizes. Another interesting feature revealed by the plots is the possible asymptotic normality of the method. Whereas for a sample size of $n = 50$ (Figure 5.21) the distribution of the differences seems to be skewed with very high estimates for some of the samples, for a sample size of $n = 1000$ (Figure 5.19) the distribution of the differences seems to be reasonably symmetric for all the possible combinations of censoring. Furthermore, both methods, M1 and M2, seem to perform well in terms of unbiasedness especially for larger sample sizes where the medians and means are similar since the distributions are more symmetric. The simulations were also performed at different time points. The time points were chosen to be $1/3$ and $2/3$ of T_m (termination of the study). The results obtained were similar to the ones obtained at time 0 with the corresponding increase of variability.

In the last part of the simulation study method M1 was tested for higher proportions of censoring. Figure 5.22 shows the results of simulating data from different distributions (A, B, C and D as before) with a sample size of $n = 1000$. As one can see, the method still performs reasonably well for proportions of censoring due to the termination of the study of 30% (fourth row), although with a considerable increase of variance especially for columns 4 and 5 (30% and 40% proportion of censored observations respectively). When the proportion of the upper tail missing is 40% the method starts to show clear signs of bias. These results show that one should be careful when applying the method in situations in which the percentage of upper tail missing is higher than 30% especially if the proportion of censored observations between 0 and T_m is 20% or higher.

5.5 Mean residual life trees based on node re-sampling

In the last part of this Chapter the estimated MRL function will be incorporated into a survival tree. As stated in the introduction, this is an attempt to extend the capabilities of the survival trees as a modeling tool that generate results that are very easy to interpret. Each terminal node in a survival tree represents a group of individuals with similar characteristics in relation to the values of the

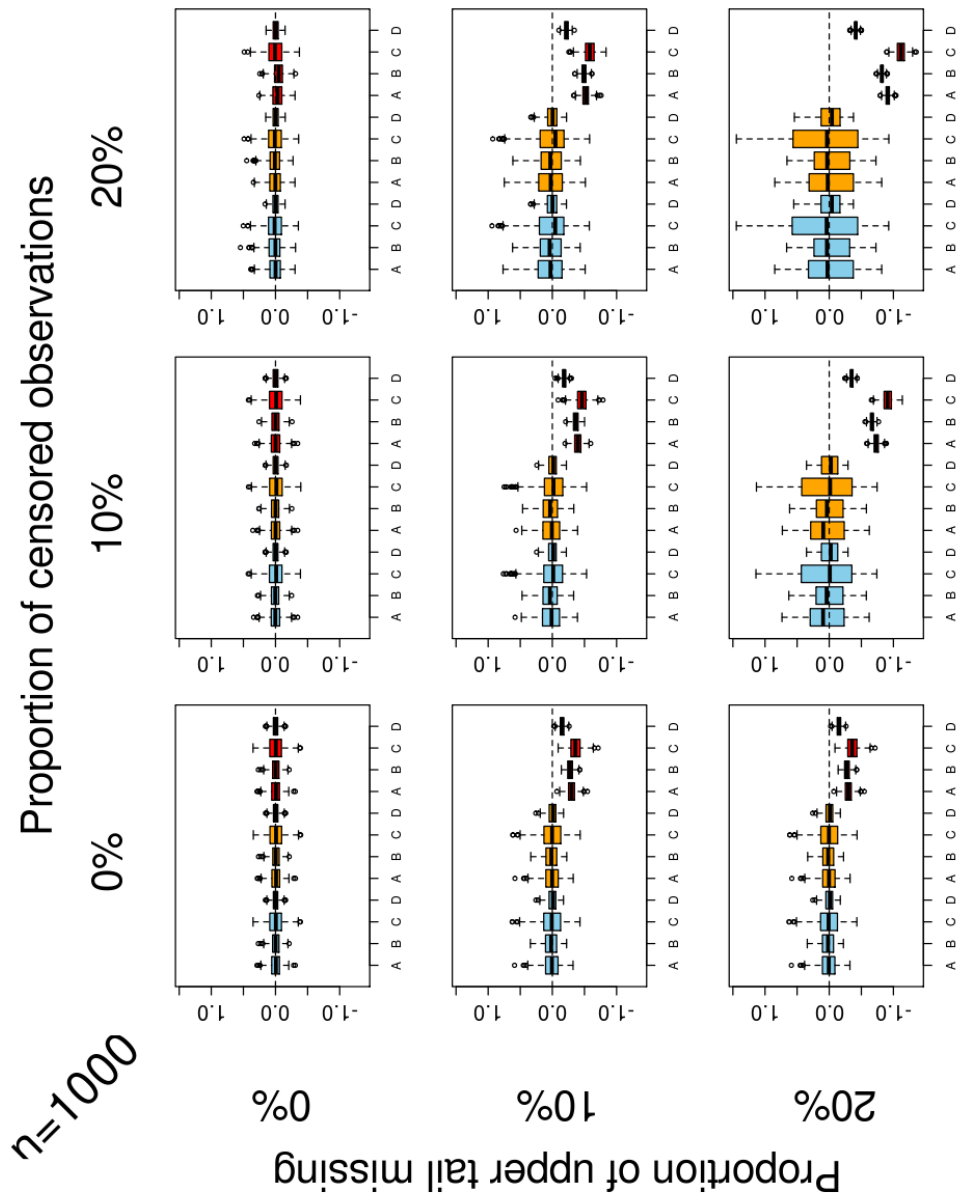


Figure 5.19: Distribution of the differences between the estimated MRL function and the true MRL function both at time $t = 0$. The sample size is $n = 1000$.

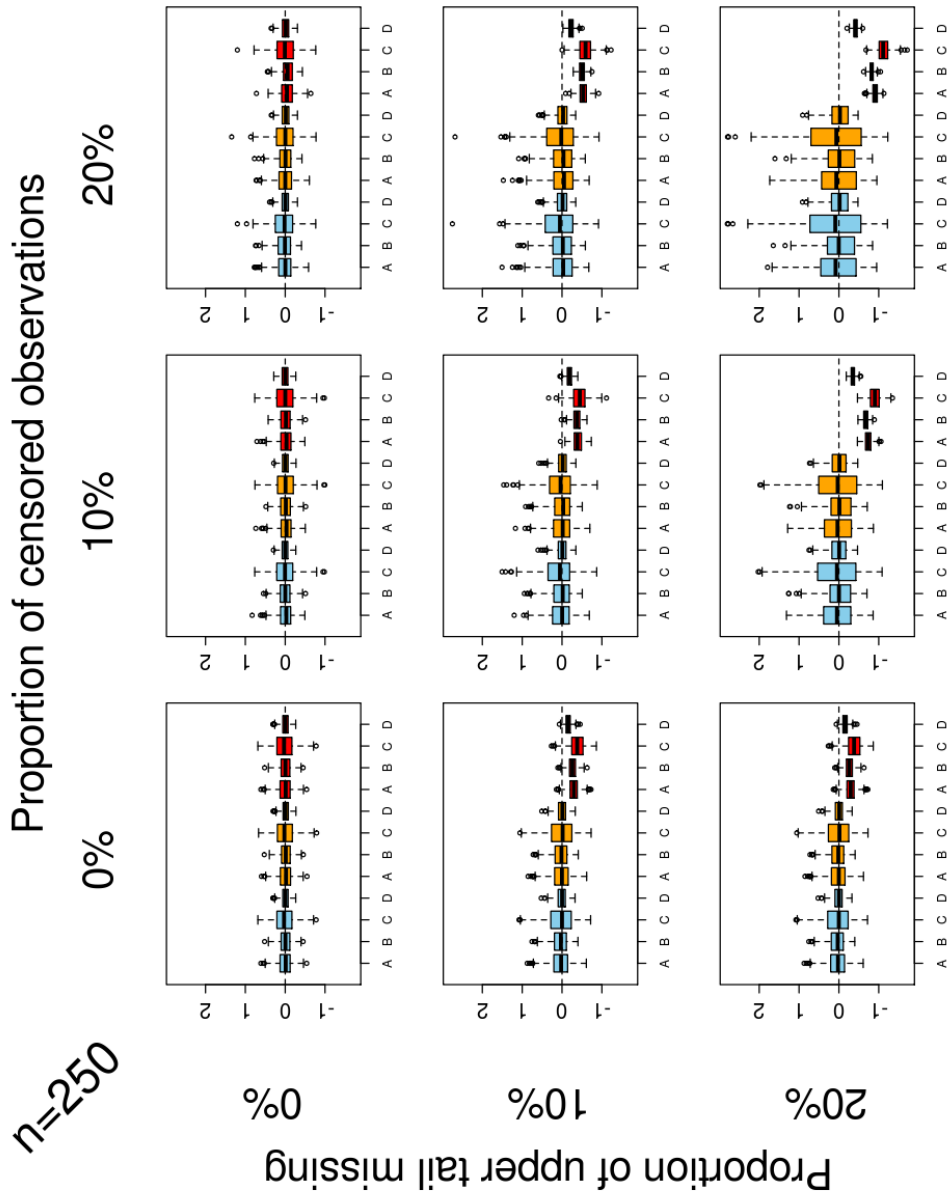


Figure 5.20: Distribution of the differences between the estimated MRL function and the true MRL function both at time $t = 0$. The sample size is $n = 250$.

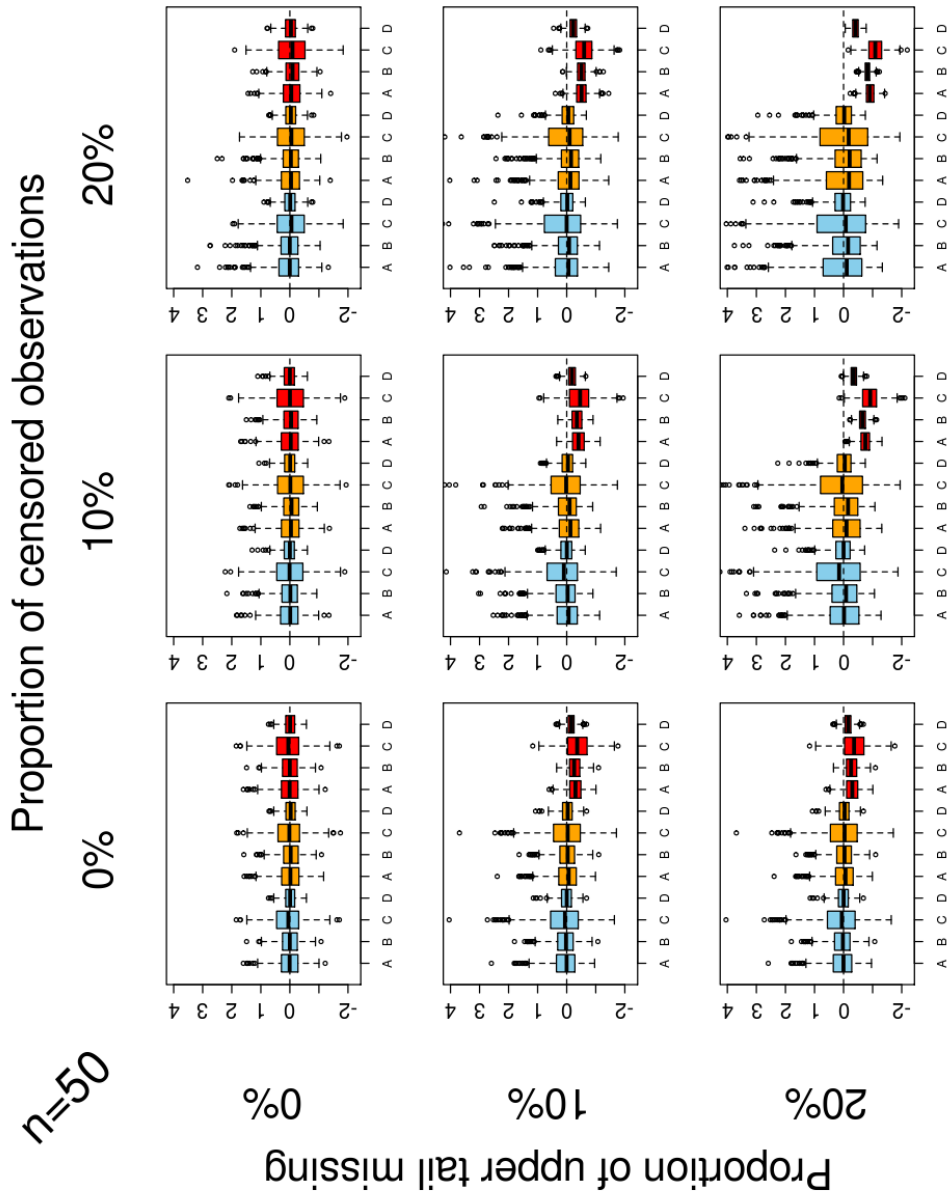


Figure 5.21: Distribution of the differences between the estimated MRL function and the true MRL function both at time $t = 0$. The sample size is $n = 50$.

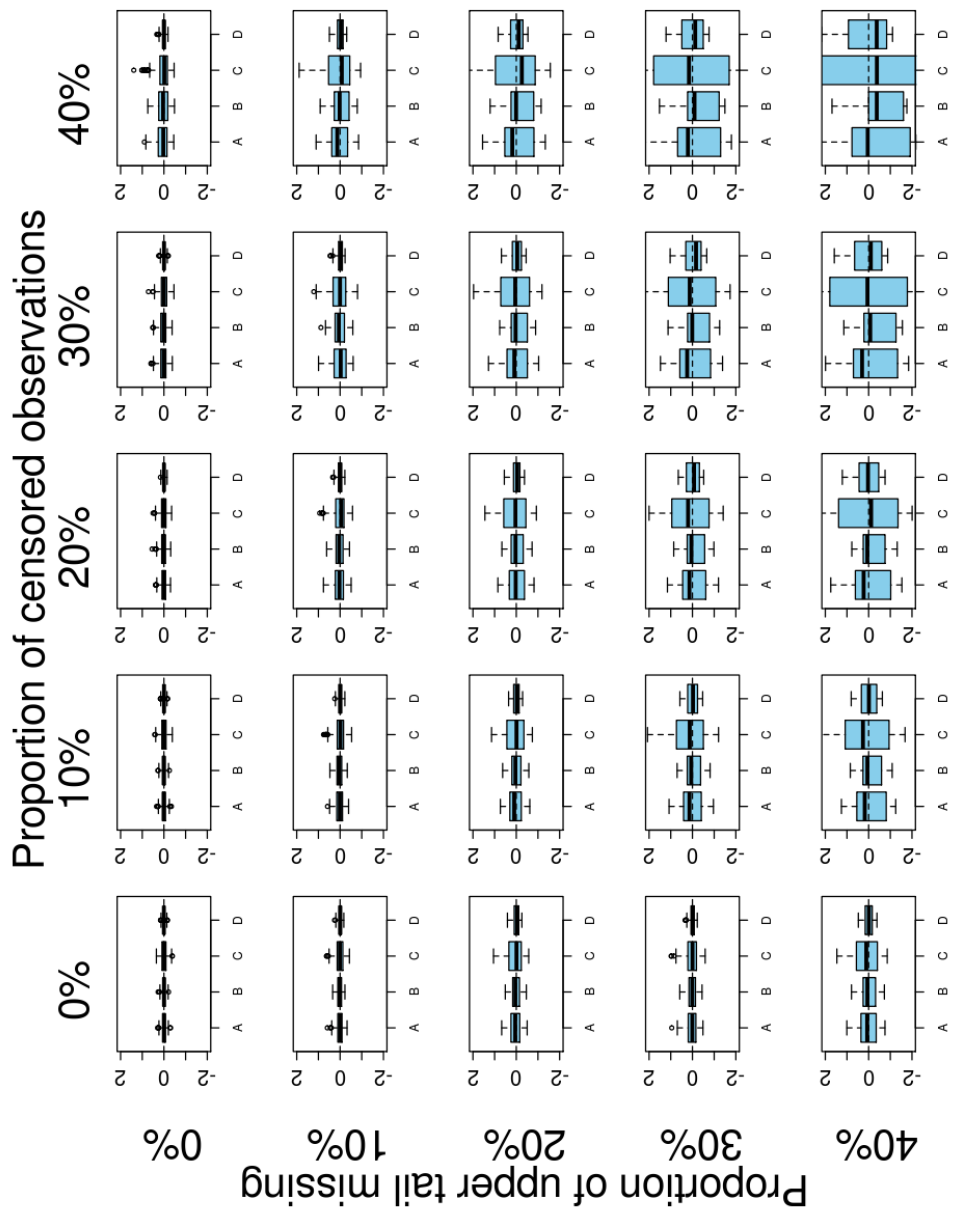


Figure 5.22: Simulations based on method M1 for higher proportions of censoring (sample size $n = 1000$).

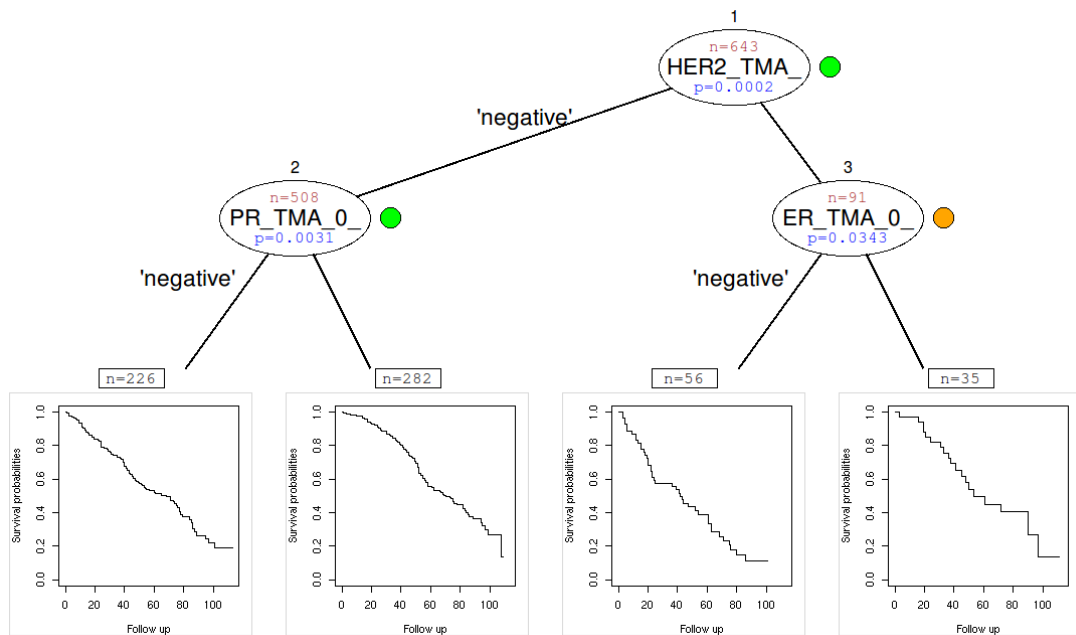


Figure 5.23: An example of a survival tree with the estimated survival functions in the terminal nodes.

predictors used to analyze the data. In Figure 5.23 a survival tree from the breast cancer dataset is displayed¹. The event of interest is the recurrence of the disease (in months) and, therefore, the analysis concerns disease free survival. The tree shows the three biomarkers that have historically been used to predict breast cancer outcomes. By looking at the tree, it is difficult to see at first glance which of the groups represented in the terminal nodes do better and which ones do worse. It seems that individuals in the third terminal node (from the left) are worse followed by those in the fourth terminal node. Individuals in the first and second terminal nodes seem to do better with better survival outcome perhaps for patients in the second terminal node.

The MRL plots in Figure 5.24, however, depict a much clear picture. Individuals in the group of HER2 positive seem to do worse than individuals in the group of HER2 negative. The average time to recurrence is around 70 months for HER2 negative patients and between 50 and 60 months for HER2 positive patients (MRL function at time $t = 0$ in the terminal nodes). Among those who are HER2 nega-

¹This tree was presented in the previous Chapter with only HER2 in the model. Although PR and ER were eliminated since they were found to be ‘irrelevant’, for the purpose of explaining the use of the MRL function in the terminal nodes, they are maintained in the tree. This is also the tree presented in the introduction of the thesis.

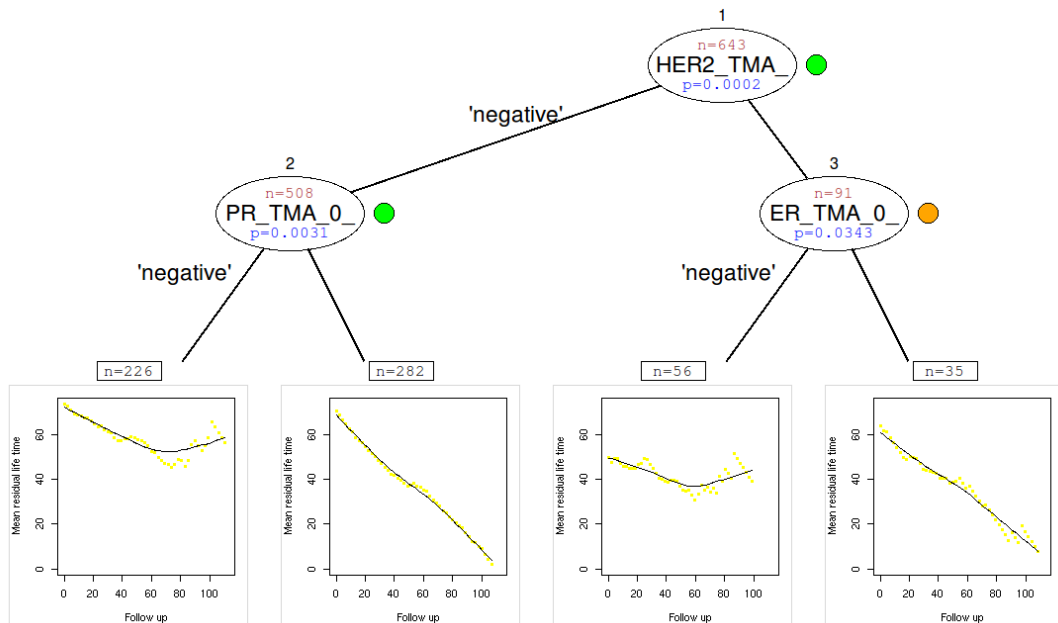


Figure 5.24: An example of a survival tree with the estimated MRL functions in the terminal nodes.

tive, the group corresponding to PR positive have worse prognosis when compared to the PR negative group. Whereas the latter group has a steady MRL function (around 60 month average for the recurrence of the disease) suggesting no deterioration of the condition, the former group displays a decreasing MRL function which indicates bad prognosis. For instance, patients who remain free of disease for 40 months have an estimated expected time for the recurrence of around 40 months (second terminal node from the left). However, patients who remain disease free for 80 months have only an estimated expected recurrence time of around 20 months. Among those who are HER2 positive, The group of ER negative have an estimated mean time to recurrence of 50 months, which is worse than the group of ER positives who has an estimated mean value of 60 months (third and fourth terminal nodes from the left). However, the former group displays a steady MRL function (around 40 months average) which again suggest no deterioration of the survival outcome. On the other hand, The group of ER positives displays a decreasing MRL function which suggests a negative survival outcome (fourth terminal node from the left).

Another interesting feature that can be investigated is the survival experience of individuals in each node of the survival tree. For instance, Figure 5.25 shows the disease free survival times of HER2 positives and HER2 negatives (root node

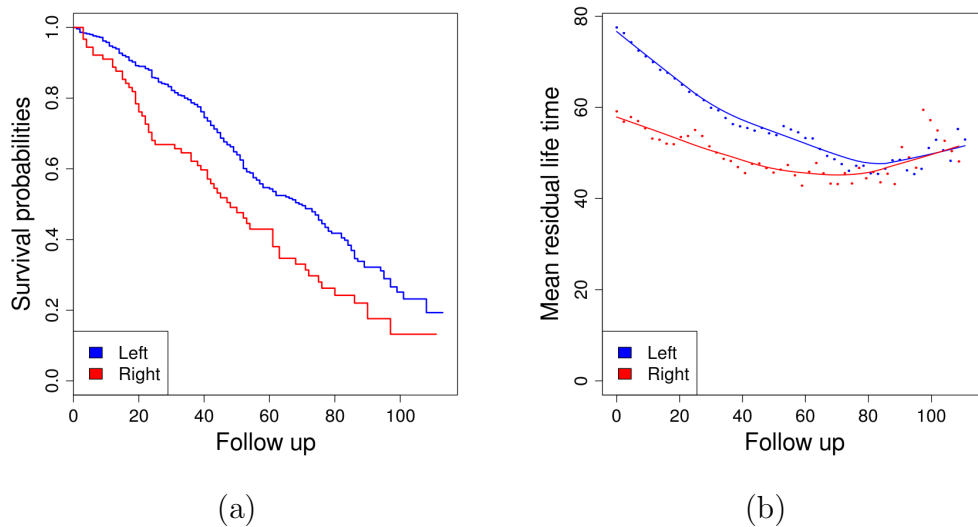


Figure 5.25: Disease free survival by HER2. The blue lines correspond to HER2 negative patients whereas the red lines corresponds to HER2 positive patients. In (a), Kaplan-Meier estimates of the survival function. In (b), smooth estimates of the MRL functions.

in Figure 5.24). Figure 5.25 (a) shows the Kaplan-Meier estimates of the survival function for both groups. It is clear that the group of HER2 positives (red line) have worse survival outcome than the group of HER2 negatives (blue line). The median time to recurrence is around 40 months for HER2 positive whereas the median time to recurrence is approximately 60 months for the group of HER2 negatives. Figure 5.25 (b) depicts the estimated MRL function. It is interesting to see that, although at time 0 the estimated differences of the expected time to recurrence for the two groups is approximately 20 months, after 80 months, if the patients remained free of the disease, the estimated time to recurrence is virtually the same for both groups.

5.6 Chapter conclusion

A new approach has been proposed for estimating the MRL function in survival problems. The use of the MRL function can help clinicians and nurses to interpret the results of the statistical analysis. The reason is that the output is given in units of time rather than in probabilities or hazards, which can be difficult concepts for individuals with no statistics training. The motivation behind the use of MRL functions is to convey results of an analysis in a more interpretable manner. The estimation of the MRL function under non-informative right censoring is

complicated when studies are carried out only for a limited period of time. The majority of existing methods do not take into account this fact in their theoretical settings and the estimates obtained are not adequate when analyzing real data. A novel semi-parametric approach has been presented which is based on extreme value theory. The behavior of the missing right tail of the distribution can be estimated using the generalized Pareto distribution. This method appears to perform well for simulated data across different distributions, different levels of censoring and different sample sizes. The new approach can be used in any survival analysis where the estimation of the survival distribution is required or where the survival experience of two or more cohorts of patients must be compared. In the field of statistical modeling the new method can be used to estimate MRL functions in the terminal nodes of survival trees. An example was presented in which this approach was used in a survival tree based on node re-sampling which was obtained using the breast cancer dataset.

The next Chapter contains a summary of all the content of this thesis and all the findings of this research. The incorporation of the MRL function in the terminal nodes of survival trees, which was the main goal of this thesis, has led to the development of two novel methods: one for the generation of survival trees (based on node re-sampling) and the other one for the estimation of the MRL function (presented in this Chapter). Two datasets have been used to illustrate these methods, one with patients with cardiovascular disease and the other one with breast cancer patients. A summary of these analyses is provided in the next Chapter. Finally, the Chapter concludes with suggestions for future work.

Chapter 6

Conclusions and future work

One of the main goals of this PhD thesis was the incorporation of the mean residual life function into tree based methods applied to survival data. The MRL function has been used traditionally in engineering and reliability but not so much in the biomedical sciences where the analysis is usually based on the survival and hazard functions. The reason for the use of survival and hazard functions is that survival functions can be easily estimated even in the presence of censoring, and hazard functions are a key element of the Cox proportional hazard model which has been used for analyzing survival data for decades. Although it is possible to use the MRL function in a proportional mean residual life model (Oakes & Dasu, 1990) this approach has not reached the degree of popularity that the proportional hazard model has enjoyed for many years. Yet, the use of the MRL function seems to be more natural in the sense that the values obtained by this function are given in terms of time rather than in terms of risk.

Tree based methods applied to survival analysis (commonly called survival trees) are a popular non-parametric alternative to the Cox proportional hazard model. One of the reasons for the popularity of these methods is that the results of the statistical analysis can be displayed in a tree fashion. This feature facilitates the interpretation of the model and makes the model particularly attractive for clinicians and physicians. In the current methods for growing survival trees, different graphical and numerical summaries are provided in the terminal nodes. These include, the median survival time, the estimated hazard ratio and the Kaplan-Meier estimate of the survival function.

The novel approach proposed in this work consists of the incorporation of the MRL function as the graphical summary in the terminal nodes. In the same way Oakes & Dasu (1990) extended the Cox proportional hazard model to the proportional mean residual life model, the work done in this thesis intends to emulate the same natural extension for survival trees. By incorporating the MRL function in the terminal nodes a non-parametric statistical model is generated that

produces as an outcome values in terms of time. This is important for different practical reasons which I will try to explain by providing some examples. Virtually any person could understand the model as no difficult statistical concepts are present in the output. By observing a survival tree with the MRL function in the terminal nodes, a clinician can easily identify cohorts of individuals with different survival experiences and the corresponding risk factors. The current methods mainly identify risk factors in terms of their effect in the hazard function. For instance, a Cox proportional hazard model might identify smoking as a significant factor in the survival times of patients with a particular disease. The estimate of the effect is given as a hazard ratio which will compare the hazard of dying for smokers and non-smokers over the period of time in which the study is carried out. The proposed approach will also identify smoking as a risk factor, but, unlike the Cox proportional hazard model, it will give information about the whole survival experience of both groups, smokers and non smokers, along with the estimates of their mean residual lifetimes at any time t . If a clinician says to a patient that smokers have twice as much risk of dying from the disease than non smokers, the patient will only get the message that smoking is ‘bad’. On the contrary, if the clinician says to the patient that smokers live 2 years on average whereas non-smokers survive for 8 years on average, the message is much clearer and the patient might decide to change the habit. Another example could be given in terms of the worth of some particular treatment or therapy. If a given treatment is known to have an effect by significantly reducing the risk of death it is worth asking how much the reduction represents in terms of time. It might be the case that such a reduction is of only a few months and therefore one could wonder if the treatment is really worth taking. By using the MRL function one can compare both survival experiences and estimate the differences in mean. Although the use of medians is also possible, the median values have a probabilistic origin, i.e. half of the population will survive longer than the estimated median. These are only a few examples of how the approach proposed in this thesis can be relevant in the area of modeling survival data. Of course the proposed method serves as a complement to the other traditional ways of modeling.

Challenges

Several challenges have been faced during the process of developing the new approach. These challenges have led to the creation of two new methods for growing survival trees and for estimating the mean residual life function. The MRL function has a serious disadvantage for statistical work since it is very dependent of the tail behavior of the survival function. This feature makes its estimation very complicated especially with right censoring data when the censoring is due to the termination of the study. The reason is that, in that case, the right tail of the

survival function is missing. To overcome this problem a new method has been developed that uses some results from the extreme value theory to estimate the behavior of the missing right tail of the survival function. The new approach produces similar results to the existing methods under the presence of right censoring and outperforms the results of the current methods in terms of unbiasedness if the right censoring is due to the end of the study. Although a more extensive theoretical study is needed, the results obtained in this thesis seem to be promising in terms of the adequacy of the method. Furthermore, this new approach for estimating the MRL function can be used not just in the context of survival trees but in any other context in which the estimation of the MRL function might be required.

The other challenge was related to the growth of survival trees. It has been shown in the literature that trees that are generated using the recursive partitioning algorithm are affected by the variable selection bias. Methods for growing survival trees based on conditional inference procedures (Hothorn *et al.*, 2006) were developed to eliminate this problem. This method introduced some notion of statistical significance by using the results of a statistical test in the splitting procedure. However, as has been demonstrated in this work, such a method is problematic for different reasons. The main drawback is the inability of detecting interaction effects in the model. This is a major drawback since the automatic detection of interactions was one of the main advantages of tree based methods. In this thesis, a new method for growing survival trees has been developed which aims to overcome all these difficulties. The method is based on the idea of node re-sampling. By basing the splitting procedure on the results obtained after bootstrapping the data available in each split the sampling variability is taken into account in the construction of the model. The new method is not affected by the variable selection bias and provides with a novel way of pruning the tree. Due to the fact that the optimal tree is obtained after a saturated tree has been pruned this method can also naturally identify interaction effects in the model.

All of these new ideas have been encapsulated in a new graphical user interface that is ready to be used by the scientific community, particularly by statisticians and clinicians. The new tool has been developed as a package to be used in the freely available language for statistical computing R (R Core Team, 2012). It is the intention of the author to make the package available from the web in the future. This new tool offers many advantages compared to the current packages for growing survival trees. One of these advantages is the possibility of scrolling out of the margins of the screen. This allows the user to visualize virtually any tree independently of its size. Thanks to this feature it is possible to grow a large tree and to prune it interactively. The new tool incorporates all the elements of the node re-sampling algorithm, including the bootstrap distribution of the cut-points,

the relative importance plot and the OOB values of the logrank statistic at each split. The process of pruning the tree is extremely easy and can be performed by any person, even those with non-statistical background. Moreover, it is possible to choose the graphical summaries for the terminal nodes including the Kaplan-Meier estimate of the survival function and the estimated MRL function using the novel method proposed in this thesis.

Datasets

Two datasets have been used to illustrate the new approaches proposed in this thesis. The Coronary data have been used to grow a survival tree based on node re-sampling (Chapter 4). When compared to the survival tree based on unbiased recursive partitioning, both trees identified ‘Age’ and ‘PreviousMI’ as relevant predictors. These two predictors were also identified as significant in the Cox proportional hazard model. The tree based on unbiased recursive partitioning also identified Diabetes, Lipids, ACE and PreviousHF of which only ACE was significant in the Cox proportional hazard model. The most relevant aspect of this example is the different cut-points for Age in both trees. The tree based on node re-sampling gave a value of 75.5 years whereas the tree based on unbiased recursive partitioning gave a value of 77 years. The figure of 75.5 years is based on the bootstrap distribution of the cut-points for Age and, therefore, it is the value obtained after the sampling variability was taken into account. From this point of view, 75.5 is a more robust value for the selection of the cut-point.

The breast cancer dataset was used to grow a node re-sampling survival tree. The tree identified three relevant splits (LN, LVI and Size) when the survival outcome corresponded to the recurrence of the disease. The split on Size was found to be significant, but it was not detected by the tree based on unbiased recursive partitioning, which only identified LN as the relevant predictor. The best prognosis, based on the node re-sampling tree, was for women with negative LN and negative LVI. Furthermore, among these patients, those who did not have the tumor (Size = 0) had worse prognosis than those who had a positive size of the tumor (probably due to some residual cancer cells that were not detected by simply measuring the size of the tumor).

In Chapter 5 an example was given of the use of the new approach proposed in this thesis where a tree based on node re-sampling was obtained with the MRL function in the terminal nodes. In this example only the biomarkers were considered to generate a survival tree able to model disease free survival. First of all, a tree with HER2, ER and PR was generated to illustrate how the novel approach of incorporating the estimated MRL function as a graphical summary in the terminal nodes can be used to interpret the tree. However, as shown in Chapter 4, ER and PR were considered as ‘irrelevant’ and only HER2 was identified as an

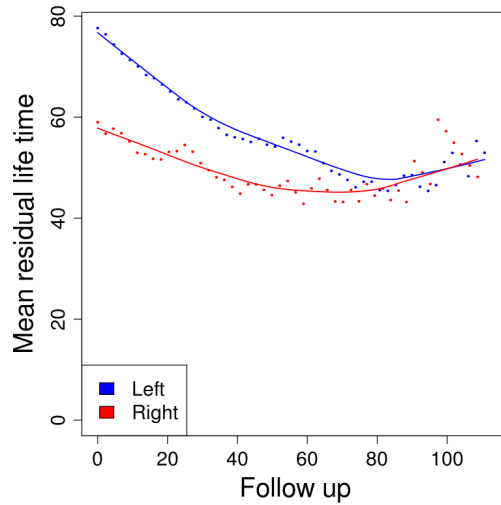


Figure 6.1: Plot of the estimated MRL functions for HER2 positive (smooth red line, right node not shown here) and HER2 negative (smooth blue line, left node not shown here). The dots are the actual estimates of the MRL function.

important predictor. Women who tested negative had a better prognostic as one can deduce by looking at the estimated medians of the Kaplan-Meier estimates of the survival function (approximately 60 months until recurrence for HER2 negative compared to approximately 40 months for HER2 positive. Plot shown in Chapter 5.). However, when one looks at the MRL function plot for both groups, the information one obtains is much richer (the plot is presented again in Figure 6.1). On average there is a difference of approximately 20 months until recurrence at time 0, meaning that women with HER2 negative have the recurrence 20 months later compared to the HER2 positive cohort. However, the most interesting aspect of the plot is the overall survival experience of both groups. As one can see in the picture, after approximately 80 months (6 and a half years) both lines merge indicating similar survival outcome for both groups. This feature is only revealed by the mean residual life function and, in some way, it diminishes the importance of HER2 as a predictor that identifies different survival outcomes.

6.1 Future work

The development of the new ideas proposed in this thesis opens up new opportunities for further analysis and research. The incorporation of the MRL function in the terminal nodes, the use of the novel method for growing and pruning survival

trees and the estimation of the MRL function based on the generalized Pareto distribution are still work in progress. I will explain the further work that these three topics might require separately.

The incorporation of the MRL function in the terminal nodes can be useful from the practical point of view by providing a new non-parametric statistical model that returns values in terms of time. It would be interesting to set up an experiment in which information about the survival outcome is given in terms of risk or probability and in terms of mean remaining time. This information would be given by a clinician to a patient showing the patient the survival tree with the Kaplan-Meier estimates of the survival function in the control group and the survival tree with estimates of the MRL function in the treatment group. The patients then could be followed up for a period of time and, after that period, one could determine if there are any differences in terms of the adherence of the patient to some specific treatment. Another experiment could be carried out by showing the two trees to different professionals who are involved in some particular study and to ask them which one do they understand. It is the hope of the author that the new way of presenting the information in the terminal nodes will become popular in the future. In the mean time, there are many studies that have already been carried out using other traditional techniques that constitute an excellent opportunity to compare the results with the new approach. For instance, if the effects of particular predictors are given in terms of hazard ratios it would be interesting to replicate the results in terms of remaining life time and see to what extent the new approach can be relevant.

The new method from growing and pruning survival trees can already be applied to any dataset with survival responses. Although more work is probably necessary to assess the validity of the method, the new approach, and all the examples in which it has been tested so far, have given excellent results. Even with the presence of interactions and many predictors adding noise to the model, the method seems to perform very well. A natural extension of the work in this thesis is to apply the same algorithm based on node re-sampling to the case of continuous and categorical responses. Although some work has already been shown in this thesis, the graphical user interface is a fundamental tool that has to be adapted for other types of responses.

But perhaps, the most exciting project that could be derived from this work is the creation of a predictive model based on the results obtained from the node re-sampling algorithm. This is one of the usual demands that clinicians make when they want to analyze statistically the data that they have collected. They want to predict the outcome for new observations. The random forest approach by Breiman (2001) is one of the standard tools that one can use to make predictions. But, as it has been said before, there is no model to look at. The

new predictive tool could be generated using the structure created by the node re-sampling algorithm. Once the predictors have been identified in the model, the data can be re-sampled again and the values of the cut-points, which can be considered the random part of the model, can be obtained by selecting at random a value from the bootstrap distribution of the cut-points. For each replicate and cut-points the new observation can be dropped down the tree and the predicted outcome recorded (if the terminal nodes contain the mean residual life function, the predicted outcome could be the mean life time at time 0 or the median). In that way different predicted outcomes would be obtained for each replicate and, therefore, a bootstrap distribution of predicted outcomes can be generated. This allows the user to obtain the predicted outcome (the mean of the bootstrap distribution) and some measure of the variability of such prediction. To assess the prediction error of the model the same method can be applied but, this time, for each replicate, the OOB data can be used to assess the error of the model in terms of the discrepancies between the observed and the predictive values. This error can be recorded for each replicate and an average can be calculated after all the replicates have been obtained. This new procedure could be used to provide tree based methods with an extra new functionality which could transform this type of methodology into a statistical tool comparable to any other modeling strategy.

Finally, the estimation of the MRL function based on the generalized Pareto distribution can be applied to any problem in which the estimation of the distribution of the survival times is necessary. The simulations presented in this thesis have given excellent results, although more theoretical work is probably necessary. Two key aspects of this novel method are the selection of the threshold u and the restrictions imposed to the shape parameter ξ in the maximum likelihood estimation. The selection of the threshold u is based on the results obtained from the simulations. However, it would be interesting to investigate if there is a way of obtaining the value of the threshold automatically, which would lead to the optimal value for the bias-variance trade-off. Another aspect related to the application of the proposed method is the generation of confidence intervals around the estimates. One obvious approach is the use of bootstrapping to obtain the standard errors of the estimates of the MRL function at each time t . In this way one can obtain piece-wise confidence bands for the MRL function. In any case, it would be interesting to study the variability of the estimates from the theoretical point of view rather than rely only on the results using bootstrap. Another application that can be derived from the use of the MRL function is the comparison of two survival distributions. Bootstrap confidence bands can be obtained for the differences between the two MRL functions at any time t . In this way, it would be possible to determine at which values of t the confidence bands produce significant results and to obtain the estimated values for the differences in the mean residual

life times.

The proposed method generates results that are equivalent to those obtained with the current methods when the termination of the study is not taken into account as has been shown in the simulations. However, when the termination of the study is embedded in the theoretical setting, the method presented in this thesis seems to work much better even for percentages of censoring due to the termination of the study up to 30% where the current methods fail to give sensible results. From that point of view, the proposed approach represents a step forward in the estimation of the mean residual life function.

To conclude, other extensions of the work done here could include comparisons between the novel approach presented in this thesis and other alternative methods such as probabilistic graphical models, particularly Bayesian networks. This type of models represent conditional dependencies between random variables with a directed acyclic graph.

Final remarks

To summarize in a few lines the content and results of this thesis, a new approach has been introduced which, in the opinion of the author, will enhance the attractiveness of tree based methods for modeling survival data. This approach consists of the incorporation of the MRL function in the terminal nodes of survival trees and aims to facilitate the task of communicating the results of the statistical analysis to any person involved in any particular study (including statisticians, clinicians and patients). In addition, a novel method has also been developed for growing survival trees based on the idea of node re-sampling. This new method addresses some of the problems that current methods have, such as the variable selection bias and the detection of interaction effects in the model. Furthermore, a new approach has been proposed for the estimation of the MRL function. This approach uses the generalized Pareto distribution to estimate the MRL function at values in which the estimated survival function does not provide any information due to censoring due to the termination of the study. Finally, a new graphical user interface has been developed that is ready to be used for any survival data with all the new ideas and methods.

Bibliography

- Aalen, Odd. 1978. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*, **6**(4), pp. 701–726.
- Bollaerts, Kaatje, Eilers, Paul H C, & van Mechelen, Iven. 2006. Simple and multiple P-splines regression with shape constraints. *Br J Math Stat Psychol*, **59**(Pt 2), 451–69.
- Breiman, Leo. 1996a. Bagging Predictors. *Machine Learning*, **24**(2), 123–140.
- Breiman, Leo. 1996b. Heuristics of instability and stabilization in model selection. *Ann. Stat.*, **24**(6), 2350–2383.
- Breiman, Leo. 2001. Random Forests. *Machine Learning*, **45**, 5–32.
- Breiman, Leo, Friedman, Jerome H., Stone, Charles J., & Olshen, Richard A. 1984. *Classification and regression trees*. Boca Raton, Florida: CHAPMAN & HALL/CRC.
- Breslow, N., & Crowley, J. 1974. A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *The Annals of Statistics*, **2**(3), pp. 437–453.
- Chambers, John M., & Hastie, Trevor. 1992. *Statistical models in S*. Advanced Books & Software. Wadsworth & Brooks/Cole.
- Chaubey, Yogendra P., & Sen, Arusharka. 2008. Smooth estimation of mean residual life under random censoring.
- Chaubey, Yogendra P., & Sen, Pranab K. 1999. On smooth estimation of mean residual life. *Journal of Statistical Planning and Inference*, **75**(2), 223–236.
- Cox, David R. 1972. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, **B**(34), 187–220.

- Crowley, John James. 1973. *Non-parametric Analysis Censored of Survival Data, with Distribution Theory for the k-sample Generalized Savage Statistic*. Unpublished Ph.D. dissertation.
- Davis, R B, & Anderson, J R. 1989. Exponential survival trees. *Stat Med*, **8**(8), 947–61.
- Doyle, Peter. 1973. The Use of Automatic Interaction Detector and Similar Search Procedures. *Operational Research Quarterly (1970-1977)*, **24**(3), pp. 465–467.
- Efron, Bradley. 1967. The two sample problem with censored data. *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, **4**, 831–853.
- Eilers, Paul H.C., & Marx, Brian D. 1996. Flexible smoothing with *B*-splines and penalties. *Stat. Sci.*, **11**(2), 89–121.
- Fleming, Thomas R., & Harrington, David P. 1991. *Counting processes and survival analysis*. Wiley series in probability and mathematical statistics. New York [u.a.]: Wiley.
- Gehan, Edmund A. 1965. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika*, **52**(1/2), pp. 203–223.
- Gill, Richard. 1983. Large Sample Behaviour of the Product-Limit Estimator on the Whole Line. *The Annals of Statistics*, **11**(1), pp. 49–58.
- Gini, C. 1912. *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini.
- Glynn, Liam G, Buckley, Brian, Reddan, Donal, Newell, John, Hinde, John, Dinneen, Sean F, & Murphy, Andrew W. 2008. Multimorbidity and risk among patients with established cardiovascular disease: a cohort study. *Br J Gen Pract*, **58**(552), 488–94.
- Gong, Qi, & Fang, Liang. 2012. Asymptotic properties of mean survival estimate based on the Kaplan–Meier curve with an extrapolated tail. *Pharmaceutical Statistics*, **11**, 135–140.
- Goodman, Melody S., Li, Yi, & Tiwari, Ram C. 2011. Detecting multiple change points in piecewise constant hazard functions. *Journal of Applied Statistics*, **38**(11), 2523–2532.
- Gordon, L, & Olshen, R A. 1985. Tree-structured survival analysis. *Cancer Treat Rep*, **69**(10), 1065–9.

- Grambsch, Patricia M., & Therneau, Terry M. 1994. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika*, **81**(3), pp. 515–526.
- Gross, A. J., & Clark, V. A. 1975. *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wiley.
- Guess, Frank, & Park, Dong Ho. 1991. Nonparametric Confidence Bounds, Using Censored Data, on the Mean Residual Life. *IEEE Transactions on Reliability*, **40**(1), 78–80.
- Guess, Frank, & Proschan, Frank. 1988. Mean residual life: Theory and applications. *Pages 215–224 of: Krishnaiah, P.R., & Rao, C.R. (eds), Quality Control and Reliability*. Handbook of Statistics, vol. 7. Elsevier.
- Guillamón, A., Navarro, J., & Ruiz, J. 1998. Nonparametric estimator for mean residual life and vitality function. *Statistical Papers*, **39**, 263–276.
- Harrell, F E, Califf, R M, Pryor, D B, Lee, K L, & Rosati, R A. 1982. Evaluating the yield of medical tests. *JAMA*, **247**(18), 2543–6.
- Harrington, David P., & Fleming, Thomas R. 1982. A Class of Rank Test Procedures for Censored Survival Data. *Biometrika*, **69**(3), pp. 553–566.
- Hastie, T., Tibshirani, R., & Friedman, J. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- Hille, E. 1948. Functional Analysis and Semigroups. *In: American Mathematical Society*, vol. 31. Am. Math. Soc. Colloq. Pub.
- Hothorn, Torsten, & Lausen, Berthold. 2003. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, **43**(2), 121–137.
- Hothorn, Torsten, Hornik, Kurt, & Zeileis, Achim. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.
- Ishwaran, Hemant, Blackstone, Eugene H., Pothier, Claire E., & Lauer, Michael S. 2004. Relative Risk Forests for Exercise Heart Rate Recovery as a Predictor of Mortality. *Journal of the American Statistical Association*, **99**, 591–600.
- Ishwaran, Hemant, Kogalur, Udaya B., Blackstone, Eugene H., & Lauer, Michael S. 2008. Random survival forests. *Ann. Appl. Stat.*, **2**(3), 841–860.

- Kalbfleisch, John D., & Prentice, Ross L. 2002. *The Statistical Analysis of Failure Time Data (Wiley Series in Probability and Statistics)*. 2 edn. Wiley-Interscience.
- Kaplan, E. L., & Meier, Paul. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**(282), 457–481.
- Kass, G. V. 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **29**(2), pp. 119–127.
- Kim, Hyunjoong, & yin Loh, Wei. 2001. Classification trees with unbiased multi-way splits. *Journal of the American Statistical Association*, **96**, 589–604.
- Klein, John P., Lee, Shih-Chang, & Moeschberger, M. L. 1990. A Partially Parametric Estimator of Survival in the Presence of Randomly Censored Data. *Biometrics*, **46**(3), pp. 795–811.
- Kleinbaum, David G., & Klein, Mitchel. 2011. *Survival Analysis: A Self-Learning Text, Third Edition (Statistics for Biology and Health)*. 3rd ed. 2011 edn. Springer.
- Kuo, Way. 1984. Reliability Enhancement Through Optimal Burn-In. *Reliability, IEEE Transactions on*, **R-33**(2), 145–156.
- Lawless, Jerald F. 2002. *Statistical Models and Methods for Lifetime Data (Wiley Series in Probability and Statistics)*. 2 edn. Wiley-Interscience.
- Leadbetter, M. R. 1983. Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields*, **65**, 291–306.
- LeBlanc, Michael, & Crowley, John. 1992. Relative Risk Trees for Censored Survival Data. *Biometrics*, **48**(2), pp. 411–425.
- LeBlanc, Michael, & Crowley, John. 1993. Survival Trees by Goodness of Split. *Journal of the American Statistical Association*, **88**(422), pp. 457–467.
- Mantel, N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, **50**(3), 163–70.
- Mantel, N., & Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Cancer Inst*, **22**, 719–748.

- Moeschberger, M. L., & Klein, John P. 1985. A Comparison of Several Methods of Estimating the Survival Function When There is Extreme Right Censoring. *Biometrics*, **41**(1), pp. 253–259.
- Morgan, J.N., & Sonquist, J.A. 1963. Problems in the analysis of survey data, and a proposal. *J. Am. Stat. Assoc.*, **58**, 415–434.
- Nelson, Wayne. 1972. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, **14**(4), pp. 945–966.
- Oakes, David, & Dasu, Tamraparni. 1990. A note on residual life. *Biometrika*, **77**(2), 409–410.
- Peterson, Arthur V. Jr. 1977. Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions. *Journal of the American Statistical Association*, **72**(360), pp. 854–858.
- Peto, Richard, & Peto, Julian. 1972. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A (General)*, **135**(2), pp. 185–207.
- Prentice, R. L. 1978. Linear Rank Tests with Right Censored Data. *Biometrika*, **65**(1), pp. 167–179.
- Prentice, R. L., & Marek, P. 1979. A Qualitative Discrepancy between Censored Data Rank Tests. *Biometrics*, **35**(4), pp. 861–867.
- Quinlan, John Ross. 1986. Induction of Decision Trees. *Machine Learning*, **1**(1), 81–106.
- R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Segal, Mark Robert. 1988. Regression Trees for Censored Data. *Biometrics*, **44**(1), pp. 35–47.
- Shen, Yan, Xie, Min, & Tang, Loon Ching. 2010. Nonparametric Estimation of Decreasing Mean Residual Life With Type II Censored Data. *IEEE Transactions on Reliability*, **59**(1), 38–44.
- Shih, Y.-S. 2004. A note on split selection bias in classification trees. *Computational Statistics & Data Analysis*, **45**(3), 457–466.
- Shih, Yu-Shan, & Tsai, Hsin-Wen. 2004. Variable selection bias in regression trees with constant fits. *Computational Statistics & Data Analysis*, **45**(3), 595–607.

- Spruance, Spotswood L., Reid, Julia E., Grace, Michael, & Samore, Matthew. 2004. Hazard Ratio in Clinical Trials. *Antimicrobial Agents and Chemotherapy*, **48**(8), 2787–2792.
- Strasser, H., & Weber, Ch. 1999. The asymptotic theory of permutation statistics. *Math. Methods Stat.*, **8**(2), 220–250.
- Su, Zheng, & Fang, Liang. 2012. A Novel Method to Calculate Mean Survival Time for Time-to-Event Data. *Communications in Statistics - Simulation and Computation*, **41**(5), 611–620.
- Tarone, Robert E., & Ware, James. 1977. On distribution-free tests for equality of survival distributions. *Biometrika*, **64**(1), 156–160.
- Watson, G. S., & Wells, W. T. 1961. On the Possibility of Improving the Mean Useful Life of Items by Eliminating Those with Short Lives. *Technometrics*, **3**(2), pp. 281–298.
- Westfall, Peter H., & Young, S. Stanley. 1993. *Resampling-based Multiple Testing*. John Wiley & Sons.
- White, Allan P., & Liu, Wei Zhong. 1994. Technical Note: Bias in Information-Based Measures in Decision Tree Induction. *Machine Learning*, **15**, 321–329.
- Wilcoxon, Frank. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, **1**(6), pp. 80–83.
- Yang, Grace L. 1978. Estimation of a Biometric Function. *The Annals of Statistics*, **6**(1), pp. 112–116.
- Zhang, Heping, & Singer, Burton H. 2010. *Recursive Partitioning and Applications*. Springer Series in Statistics. Springer.
- Zhang, H.P. 1995. Splitting criteria in survival trees. *Pages 305–314 of: Proceedings of the 10-th international workshop on statistical modeling*.
- Zhou, Mai, & Jeong, Jong-Hyeon. 2011. Empirical likelihood ratio test for median and mean residual lifetime. *Stat Med*, **30**(2), 152–9.