POLITECNICO DI TORINO

Master's Degree
in Mathematical Engineering

Master's Degree Thesis

Event time prediction based on inversion of the partial likelihood equations in survival analysis

Supervisors
prof. Mauro Gasparini
Dietmar Trummer

Candidate
Meike Adani

Academic Year 2021-2022

To my lovely family

Summary

Clinical trials often collect and assess data of survival (or time-to-event) with the objective of comparing different treatments or identifying risk factors that are linked to individuals risk rate of experiencing an event, that can be death, tumor progression or any other meaningful clinical outcome. When dealing with these types of survival data, the most popular method is the Cox proportional hazards regression model, used to explore the relationship between survival experience and characteristics of patients. The standard outcome of the Cox model is a semiparametric estimate of the hazard ratio, a relative measure that informs on the rank of patients' risk among others, but does not meaningfully inform on individual patients. However, especially in the context of personalized medicine, it is of interest to identify an accurate model for lifetime prediction on an individual level.

The aim of this work is to inspect if it is possible to obtain an estimate of survival time for a new patient, starting from a validated Cox model and known regression coefficients.

First of all two functions of the `survival` package in the software R are examined, namely `predict.coxph()` and `survfit.coxph()`, in order to identify which methods can be used for the purpose of individual survival time prediction. The first function turns out to provide risk predictions, that are relative measures not adequate to describe individual survival times; the second one provides instead individualized survival curves, that are absolute measures and from which two different values can be extrapolated: median survival time and restricted mean survival time. These can be used for prediction but rely on an estimate of the baseline survival function and are thus classified as parametric approaches. Since the Cox proportional hazards model is classified as a semiparametric model, not requiring the hazard and the survival function to be specified, the aim of this work is to preserve this feature and find a non-parametric approach for the estimation of survival times.

In particular, it is questioned whether it is possible to invert the equation characterizing the Cox model and solve it for the survival time of a single individual, once the estimated coefficients are known. At this scope, a kind of inversion of the model is proposed, where the partial likelihood used in the estimation of the regression coefficients is exploited. It is shown that, if the regression coefficients are known, the model can be inverted somehow and a range of survival times can be obtained for a new patient; practically, there are some issues when the real values of the regression coefficients are not known and the main limitations are underlined. Finally, the results obtained with the parametric approaches of median and restricted mean survival time are analyzed over two example data sets and compared with the novel algorithmic approach in terms of predictive performance.

Acknowledgements

I would like to sincerely thank my supervisor Mauro Gasparini for trusting me from the beginning and giving me the opportunity to make a great experience from which this thesis originates.

I really would like to thank Lidia who made everything possible and has been a loyal companion and colleague during my entire journey at Bayer. Of course a special thanks goes to Dietmar, who gave me the chance to join his team and from whom I learned a lot about clinical development. I would like to mention the entire team, thanking everyone who shared his or her knowledge and experience with me; among them, a particular thanks goes to Anke, who proposed me this interesting and challenging topic, and to Florian for spending his time with me.

Thanks to all my friends that support me in everything I do, I am really grateful for having you all by my side. Thanks to my Ballotta for being my certainty whenever I came back home. Thank you Agne and Ire for the big support in all this long journey and thank you Meg for being from the very beginning a person I can count on. Thanks to the new born "Asse Trapani-Beaulard", my second new family in Turin: you made this five years special and I will always have great memories. Thanks to David for the friendship that has grown constantly and to Davide who is the best student rep. ever.

Thanks to Michele for always believing in me, for his constant support and for sharing with me all my pains and successes.

Last but not least, I really would like to thank all my family from all my heart. Thanks to my grandmother who has always a special words to spend and thanks to my sisters without whom I were lost: thanks to my older sister for her recommendations and being my example and thanks to my twin for simply being my second half and being always with me whatever it happens. Finally, thanks Mum and Dad for being such great parents: thank you for giving me the possibility to realize all my desires and for always let me feel that you are proud of me.

This thesis would not have been here without you all.

Contents

Chapter 1

Introduction

The work of this thesis originates from an experience in the pharmaceutical company Bayer AG as an intern in clinical statistics in the Oncology SBU department. Among all the different tasks and roles that characterize the team, the biomarker team stands out for its fast growing and its increasingly importance in many aspects of pharmaceutical discovery and development.

Biomarkers act as a predictor in a model for different clinical outcomes, be these in terms of disease prognosis, treatment response or occurrence of toxicities, with the aim of explaining the variation in the responses among individuals and being potentially valuable for individual patient management and personalized healthcare.

It is in this setting of increasing interest in personalized and precision medicine that some interesting questions may arise. In the field of survival analysis, fundamental questions regard the possibility of predicting patients' lifetime: " What can we say about survival times of individuals? Can we predict how long patients with a particular disease are expected to live?".

Starting from this background, this work aims at finding an answer to the previous questions under specific assumptions and conditions, with the objective of predicting lifetime of patients that could potentially impact the field of personalized medicine and improve the decision-making process at an individual level.

The thesis is organized as follows. Chapter 2 introduces the concept of survival analysis in the particular context of medical research giving the basis for understanding what is the final objective. Chapter 3 presents the popular Cox model used for analyzing survival data, together with its main characteristics concerning partial likelihood estimation (Section 3.2), survivor function estimate (Section 3.3) and prediction (Section 3.4). Chapter 4 concerns survival data and prediction using the `survival` package in R; it introduces an example of survival data (Section 4.1) and the function `coxph()` that implements the Cox model (Section 4.2). Finally, it analyzes the functions `predict.coxph()` (Section 4.3) and `survfit.coxph()` (Section 4.4) used for predictions. Chapter 5 contains a novel approach for prediction of individual lifetimes starting from the Cox partial likelihood; in particular, Section 5.1 contains the theoretical formulation and considerations while Section 5.2 shows its application to an example data set. Chapter 6 presents a performance evaluation metric (Section 6.1) and compares, graphically and through this measure, the presented prediction models (Section 6.2); Section 6.3 presents another example of survival data in order to make further comparisons and assess quality of predictions. To conclude, chapter 7 summarizes the results obtained, highlighting the main issues and limitations that are encountered when dealing with individual lifetime prediction.

Chapter 2

Survival Analysis

2.1 Introduction to Survival Analysis

The term *survival analysis* describes a statistical methodology used for the analysis of data in the form of *times until an event occurs.* In particular, the times refer to the time elapsed from a well-defined time origin until the occurrence of some particular event of interest and they will generally be defined as *survival times.* Although the methods that characterize survival analysis can be used in many areas, the focus here is on the application of survival analysis in medical research, where the time origin will often be the recruitment of an individual into an experimental study, and possible end-points will be disease remission, progression or death of the patient.

2.1.1 Main Features of Survival Analysis

In clinical trial setting it may be not possible to observe the end-point for every patient: this may be because the data from the study have to be analyzed at a specific point in time (e.g. end-of-study) when some individuals are still alive, or because the individual has been withdrawn from the study or *lost to follow up.* These incomplete observations are called *censored* survival times and for these patients the only available information is the last date in which it is known they were alive.

There are in general various categories of censoring, such as right censoring, left censoring, and interval censoring but throughout this work only the most common *right-censoring* will be considered, where the right-censored survival time is smaller than the actual, but unknown, survival time. Moreover, an important assumption that is made in modelling survival data, is that the actual survival time of an individual does not depend on any mechanism that causes that individual's survival time to be censored; censoring that meets this requirement is called *non-informative* [2]. Censoring is one of the main reasons why, to analyze survival data, special tools are required.

2.1.2 Summarising Survival Data

There are three basic concepts that pervade the whole theory of survival analysis and are used in summarising survival data, namely the survivor function, the hazard function and the cumulative hazard function [2].

Survivor function

The *survivor function* $S(t)$ gives the expected proportion of individuals for which the event has not yet happened at time t and, in other terms, gives the unconditional probability that the individual will survive beyond time t (experiencing the event after time t). If the survival time of an individual t is regarded as an observation of a random variable T that has probability density function $f(t)$, formally the survivor function is given by:

$$S(t) = \mathrm{P}(T \geq t) = \int_t^{\infty} f(u)\, du. \tag{2.1}$$

The survivor function is a monotone, non-increasing function that is equal to one at origin and zero as the time approaches infinity; its rate of decline varies according to the risk of experiencing the event at time t, that is given by the hazard function.

Hazard function

The *hazard function* $h(t)$, on the contrary, is defined by means of conditional probability: it represents the probability that an individual experiences the event at time t, conditional on he or she having survived to that time. Also called a *hazard rate*, it expresses the instantaneous risk of an event occurring and is formally defined as:

$$h(t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \mathrm{P}(t \leq T < t + \delta t \,|\, T \geq t). \tag{2.2}$$

The hazard rate can take many different shapes and the only restriction on $h(t)$ is that it is non-negative, i.e. $h(t) \geq 0$.

Cumulative hazard function

A quantity related to the hazard rate is the *cumulative hazard function* $H(t)$, that summarizes the cumulative risk of an event occurring by time t. It is derived from the previous two functions as follows:

$$H(t) = \int_0^t h(u)\, \mathrm{d}u\,, \tag{2.3}$$

and

$$H(t) = -\log S(t). \tag{2.4}$$

In the analysis of survival data, estimates of the survivor function, hazard function and cumulative hazard function just mentioned can be obtained from the observed survival times and be used to draw on the survival experience of patients.

2.2 Objective of the thesis

The objective of survival analysis in clinical setting is often to assess the effect of a treatment over placebo on the survival experience for a population in the study, but this is not its only application. Survival analysis is also widely used to assess the effect of one or more predictors, usually termed *covariates* or *explanatory variables*, on the survival outcome of the individuals in the study. When the objective is to find a relationship between survival experience of patients and a number of explanatory variables, regression models are called for. The most popular and widely used regression model is the *Cox regression model* that can handle the peculiarity of censored-data and is presented in Chapter 3 with more details .

With the objective of assessing the effect of certain variables on survival, statistical models, considering the increasing demand for accuracy in practical applications, play also an important role in the *prediction* of survival times for new patients. With the ever increasing attention paid to personalized medicine, it would be of interest to predict what is the expected time of experiencing the event of interest for a new patient.

The aim of this work is to identify a method that can be used to predict survival times of new individuals in the setting of a Cox regression model. With this objective, the `survival` package in software R is firstly inspected to analyze if there are already available options for prediction and to explore what are the main limitations. Secondly, possible improvements are proposed in order to provide a patient with an estimation of his or her *expected survival time.*

Chapter 3

Cox Proportional Hazard Model

3.1 Introduction and Notation

The *Cox proportional hazard model* [3] is a regression model that is widely used in survival analysis to investigate the relationship between the survival experience of a patient and some explanatory variables.

For an individual with covariates $\mathbf{x}_i = (x_{1i}, ..., x_{pi})$, the hazard function takes the form of

$$h_i(t) = h_0(t) \exp{(\beta_1 x_{1i} + ... + \beta_p x_{pi})}, \tag{3.1}$$

where $h_0(t)$ is a beseline hazard that describes the shape of the hazard as a function of time and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a p x 1 column vector of coefficients. Model 3.1 is often expressed in terms of the *hazard ratio* and the model, which is multiplicative for the hazard, is then linear for the log-hazard ratio:

$$\log\left[\frac{h_i(t)}{h_0(t)}\right] = \beta_1 x_{1i} + ... + \beta_p x_{pi}. \tag{3.2}$$

One important aspect of the model is that it relies on the assumption of *proportional hazards (PH)*, which means that:

- Covariates have a multiplicative effect on the hazard.
- The effect of the covariates on the hazard function doesn't change with time, letting the hazard ratio between two subjects be constant in time

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_j)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\exp(\boldsymbol{\beta}'\mathbf{x}_j)}. \tag{3.3}$$

It is important to assess these assumptions when fitting a Cox model, otherwise violating the PH assumption can seriously invalidate the model.

The exponentiated coefficients $\exp(\beta_j)$ take the name of *hazard ratios (HR)* and are interpreted as the multiplicative effects on the hazard:

- HR = 1: the covariate has no effect.
- HR < 1: the effect of the covariate is to reduce the hazard and the risk of experiencing the event of interest.
- HR > 1: the effect of the covariate is to increase the risk of experiencing the event.

The coefficient β_j of x_{ij}, that represent the logarithm of the hazard ratio, can be interpreted as the change in the logarithm of the hazard ratio when the value of x_{ij} is increased by one unit and the other covariates are kept constant.

The vector of regression coefficients $\hat{\boldsymbol{\beta}}$ can be estimated for the model in Equation 3.1, and the next section (Section 3.2) will give an idea on how it is estimated.

Since no assumptions are made about the actual form of the baseline hazard function $h_0(t)$ in Equation 3.1, the Cox model is called a *semi-parametric* model where no particular form of probability distribution is assumed for survival times. This is one of the main reasons for which this method is widely used, since no extra assumptions are required for the survival times that may be not appropriate.

However, the simplifying aspects of the Cox model that make is so useful are exactly those that should be verified to determine whether a fitted Cox regression model adequately describes the data, in order to make future inferences. Mainly four underlying assumptions should be considered: proportional hazards, additivity, linearity, and lack of any high leverage point [6]. Different types of *residuals* are used for this purpose, see for example [2, Chapter 4].

3.2 Partial Likelihood

Estimation of the vector of regression coefficients $\boldsymbol{\beta}$ is based on the method of *maximum likelihood*. If the individual experiences the event at time t_i, his or her contribution to the likelihood would be the density $f(t_i)$, given by the product of the survivor function $S(t_i)$ and the hazard function $h(t_i)$:

$$L_i(\boldsymbol{\beta}) = S(t_i)h(t_i).$$

If the unit is still alive at t_i, meaning that the time is censored, the contribution to the likelihood would then be given by the only survivor function:

$$L_i(\boldsymbol{\beta}) = S(t_i).$$

The full likelihood function can thus be written including both contributions as

$$L(\boldsymbol{\beta}) = \prod_i L_i(\boldsymbol{\beta}) = \prod_i h(t_i)^{\delta_i} S(t_i). \tag{3.4}$$

Since the likelihood in equation 3.4 requires the knowledge of the shape of the hazard function, a *partial likelihood* is considered for the Cox model instead. For a study with n patients, the partial likelihood for the model in Equation 3.1 is defined as

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right]^{\delta_i}, \tag{3.5}$$

or, in the form of partial log-likelihood, as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \Big[\boldsymbol{\beta}'\mathbf{x}_i - \log \big[\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l) \big] \Big], \tag{3.6}$$

where $R(t_i)$, called the *risk set*, indicates the set of individuals who are at risk at time t_i and δ_i is an event indicator being equal to zero if the time t_i is right-censored and equal to unity otherwise. Although the partial-likelihood is not, in general, a likelihood in the sense of being proportional to the probability of an observed dataset, nonetheless it can be treated as a likelihood for purposes of asymptotic inference [8]. Estimates of the β-parameters can thus be found by maximising the partial log-likelihood in Equation 3.6. Taking the first derivative with respect to $\boldsymbol{\beta}$ gives the *score vector* $U(\boldsymbol{\beta})$:

$$U_j(\boldsymbol{\beta}) = \frac{d\ell}{d\beta_j} = \sum_{i=1}^{n} \delta_i \Big[x_{ji} - \frac{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l) x_{jl}}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \Big], \qquad j = 1 \ldots p. \tag{3.7}$$

The *information matrix* is the negative of the matrix of second derivatives of the log-likelihood and is given by $I(\boldsymbol{\beta}) = [I_{zj}(\boldsymbol{\beta})]_{pxp}$ with the (z, j)-th element given by

$$I_{zj}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l) x_{jl} x_{zl}}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} - \sum_{i=1}^{n} \frac{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l) x_{zl}}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \Big[\frac{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l) x_{jl}}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \Big]. \tag{3.8}$$

The (partial) maximum likelihood estimates are found by solving the set of p nonlinear equations $U_j(\boldsymbol{\beta}) = 0$ in Equation 3.7 for $j = 1, \ldots, p$. Since they cannot be solved analytically, numerical methods like the *Newton-Raphson algorithm* will be used to find the estimates.

It is worth noting that the partial likelihood in Equation 3.5, used in a Cox model, depends only on the ranking of the event-times, thus inferences about the effect of explanatory variables on the hazard function depend only on the rank order of the survival times and not on their absolute values [2].

Tied Events

The Cox regression model and its partial likelihood hold under the assumption of continuous data but, in real situations, survival times are usually recorded to the nearest day, month, year etc, and so it is possible to have two or more events happening at the same time, as a consequence of this rounding process. These events take the name of *tied events* and variants of the partial likelihood are needed to address this situation. The simplest approximation for Equation 3.5 is due to Breslow [1] and is given by:

$$\prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_j)}{\big[\sum_{l \in R(t_j)} \exp(\boldsymbol{\beta}'\mathbf{x}_l) \big]^{d_j}}, \tag{3.9}$$

where d_j is the number of events at time $t_{(j)}$ and s_j is the vector of sums of each of the p covariates for those individuals who experience the event at the j^{th} event time $t_{(j)}$. This approximation is adequate especially when the number of tied observations at any one event time is not too large.

Another approximation is due to Efron [4] who proposed

$$\prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_j)}{\prod_{k=1}^{d_j}\left[\sum_{l\in R(t_{(j)})}\exp(\boldsymbol{\beta}'\mathbf{x}_l) - (k-1)d_j^{-1}\sum_{l\in D(t_{(j)})}\exp(\boldsymbol{\beta}'\mathbf{x}_l)\right]}, \tag{3.10}$$

where $D(t_{(j)})$ is the set of all the individuals who experience the event at time $t_{(j)}$. This approximation is closer to the appropriate likelihood function than the one due to Breslow, although in practical situations they have very similar performance.

Cox's approximation treats the underlying time scale as discrete rather than continuous and is takes the name of exact partial likelihood:

$$\prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_j)}{\sum_{l\in R(t_{(j)};d_j)}\exp(\boldsymbol{\beta}'\mathbf{s}_l)}, \tag{3.11}$$

where $R(t_{(j)}; d_j)$ denotes a set of d_j individuals drawn from the risk set $R(t_{(j)})$ at time $t_{(j)}$.

Because of the superiority of the Efron approximation (Eq 3.10), the Cox model implemented in software R uses this method as default for handling tied event time, while other Cox regression programs use Breslow approximation (Eq 3.9) for its simplicity in computation. If there are few tied event times than all the approximations in Equations 3.9,3.10,3.11 will be nearly equivalent and when there are no ties, they all reduce to the same form [2].

3.3 Estimation of Hazard and Survivor Functions

In order to make inferences about the effect of explanatory variables on the hazard function, all what is needed is the estimate of the vector $\boldsymbol{\beta}$ of the regression coefficients. But, once these estimates are obtained, one could further summarize survival experience of the patients in the study computing estimates for the hazard and survivor functions. For these purposes, an estimate of $h_0(t)$ itself is also needed. These estimates should match the way in which ties are treated in the likelihood for the Cox model, as discussed in Section 3.2 [8].

One popular estimator of the cumulative baseline hazard, due to its simplicity, is given by:

$$\hat{H}_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)} \qquad \text{for} \quad t_{(k)} \le t < t_{(k+1)}, \tag{3.12}$$

where $k = 1, .., r-1$ and r is the total number of different event-times observed in the study. The corresponding estimator of the survivor function, derived from the exponential of the cumulative hazard, is then

$$\hat{S}_0(t) = \prod_{j=1}^{k} \exp\left[\frac{-d_j}{\sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)}\right] \qquad \text{for} \quad t_{(k)} \le t < t_{(k+1)}. \tag{3.13}$$

The estimate of Equation 3.12 is often referred to as the *Nelson-Aalen* estimate and the derived one in Equation 3.13 as the *Breslow* estimate.

Finally, from the estimates of the baseline cumulative hazard and survivor functions, the corresponding estimates can be obtained for an individual characterized by a vector of explanatory variables $\mathbf{x}_i$:

$$\hat{S}_i(t) = \{\hat{S}_0(t)\}^{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)} \qquad \text{for} \quad t_{(k)} \leq t < t_{(k+1)}. \tag{3.14}$$

From this definition, the estimated survivor function for an individual is a piecewise constant function that is defined only until the last observed event-time. Equation 3.14 shows that a *survival curve* for each patient in the study can be obtained from the estimated regression coefficients and this is already an important input to clinical decision making. Furthermore, the entire distribution can be summarized in single values and used to compute estimates of other measures of interest like, for example, the *p-year* probability of survival for an individual or the *median* and *mean* survival times.

Confidence intervals can also be computed for the estimated survivor function $\hat{S}_0(t)$, once its variance has been computed. One common estimator is due to Greenwood:

$$v\hat{a}r(\hat{S}_0(t)) = \hat{\sigma}^2(\hat{S}_0(t)) = \hat{S}_0(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{r_i(r_i - d_i)}, \tag{3.15}$$

where r_i is the number at risk right before t and d_i is the number of deaths occurred at time t_i. The standard deviation is then indicated as $\hat{\sigma} = se[\hat{S}_0(t)]$, where se refers to the standard error.

In the same manner, 95% confidence intervals for $\hat{S}_i(t)$ can be computed on the *plain scale*

$$\hat{S}_i(t) \pm 1.96\,\mathrm{se}[\hat{S}_i(t)], \tag{3.16}$$

or on the cumulative hazard (or *log survival*) scale

$$\exp\{\,\log S_i(t) \pm 1.96\,\mathrm{se}[\log \hat{S}_i(t)]\,\}. \tag{3.17}$$

Also *log-log* and *logit* transformations of the survivor function can be used, and they are obtained simply substituting $\hat{S}_i(t)$ with the corresponding transformation in the confidence interval in Equation 3.16.

Many authors have investigated the behavior of transformed intervals like that in Equation 3.17, and a general conclusion is that the direct intervals (Eq.3.16) do not behave well, particularly near 0 and 1, while all the others are acceptable. For this reason in the software R they are computed with the log-transformation as default [6].

3.4 Predictions from a Cox Model

Once a suitable Cox model for a set of survival data has been identified and estimates for the vector of coefficients $\hat{\boldsymbol{\beta}}$, the hazard $\hat{h}_0$ and the survivor function $\hat{S}_0$ have been computed, it would be of interest to *predict* a survival time for a new patient that is added to the study. In order to compute estimates for new individuals, the predictive ability of the Cox model should be assessed and there are essentially two aspects that are considered regarding the evaluation of the predictive performance: *discrimination* and *calibration*.

- Discrimination measures how well model-estimated risks translate to patient outcomes: patients predicted to be at higher risk (greater hazard ratio) should have experienced the event before those deemed at lower risk in the observed dataset. C-index, that is a measure of *concordance*, is a common statistic used to assess discrimination ability of a Cox model; values of concordance equal to one represent perfect discrimination.

- Calibration reflects prediction accuracy, in particular the accuracy of survival probabilities at any time after the time origin. The *Brier Score* is a common statistic used to assess calibration performance of a Cox model, measuring the distance between the predicted survival probability and the actual outcome at that time. The best possible value is zero, coinciding with perfect accuracy.

For a new patient, it is supposed that all the values corresponding to the explanatory variables used in the Cox model for the original set of patients are known, so that he or she is characterized by a set of covariates $\mathbf{x}_{new} = (x_{1,new}, ..., x_{p,new})$. Furthermore, for the proportional hazards assumption to still hold, the new patient is assumed to be taken from a group that is qualitatively the same as the original one used to derive the Cox proportional hazards model.

Since the Cox PH model is a *risk prediction model*, for each new patient it predicts a *risk score* (also called *linear predictor* or *prognostic index*), that is a relative measure for the probability of occurrence of the event of interest. The risk score corresponds to the linear component of the model and is given by a linear combination of the values of the p explanatory variables in $\mathbf{x}_{new}$:

$$\hat{\eta}_{new} = \log\left\{\frac{h_{new}(t)}{h_0(t)}\right\} = \hat{\boldsymbol{\beta}}'\mathbf{x}_{new} = \hat{\beta}_1 x_{1,new} + \hat{\beta}_2 x_{2,new} + \cdots + \hat{\beta}_p x_{p,new}\,. \tag{3.18}$$

The predicted risk score informs on the rank of the patient's risk among the others, but does not meaningfully inform on an individual level, since it cannot directly give an individual survival time and not even a survival probability. To achieve this last objective the Cox model has to be combined with an estimator of the baseline hazard function as discussed in Section 3.3, from which other estimates and information can be obtained. However, it is not straightforward to directly estimate survival times for new individuals.

Based on this consideration, the purpose of the next chapter is to analyze what are the estimates that can be computed for a new patient taking advantage of the `survival` package included in the software R and to highlight what still remains a challenge in the context of survival times prediction.

Chapter 4

Survival Analysis and Prediction in R

The survival package included in software R ($\geq$ 3.5) is concerned with time-to-event analysis that, as already stated, refers to a special type of outcome which arises very often in the analysis of medical data: time to progression for a tumor or death itself are only two examples. A key common principle for such outcomes is that "it takes time to observe time" and this is is what led to the creation of an appropriate package in order to manage the resulting challenges [6].

Survival data is often represented as a pair (t_i, δ_i) where t_i is the time until end-point or last follow-up, and δ_i is a binary variable equal to 0 if the subject i was censored at time t_i and equal to 1 if the subject i had an event at time t_i. In R code survival data take the form of `Surv(time, status)` and are called *survival objects*.

In the survival package a function called `coxph()` is used to fit a Cox model to a given dataset of patients and other two R functions, `predict()` and `survfit()`, if applied to the result of a fitted Cox model, can give further information on the survival experience of a new patient.

A dataset from the survival library will be used throughout this chapter to illustrate the response of these functions and analyze the different options regarding the input arguments.

4.1 MGUS Data Set

The library survival contains different examples of survival data that can be used for survival analysis. The dataset `mgus` has been selected and is here used for illustrative purposes. The dataset contains the natural history of 241 subjects with monoclonal gammopathy of undetermined significance (MGUS), that is a condition characterized by the presence of a monoclonal paraprotein in the blood that can eventually progress to a plasma cell malignancy. The dataset contains 12 variables for each subject in the study:

- id: subject id;

- age: age in years at the detection of MGUS;
- sex: male or female;
- dxyr: year of diagnosis;
- pcdx: for subjects who progress to a plasma cell malignancy the subtype of malignancy (multiple myeloma (MM), amyloidosis (AM), macroglobulinemia (MA), lymphprolifative disorders (LP));
- pctime: days from MGUS until diagnosis of a plasma cell malignancy;
- futime: days from diagnosis to last follow-up;
- death: 1= follow-up is until death;
- alb: albumin level at MGUS diagnosis;
- creat: creatinine at MGUS diagnosis;
- hgb: hemoglobin at MGUS diagnosis.
- mspike: size of the monoclonal protein spike at diagnosis;

The variables include some demographic aspects, like age or sex, and other clinical values reported at the time of diagnosis. For subjects who progressed to a plasma cell malignancy also the subtype of malignancy is reported together with their time until progression. As very common in cancer treatment studies, overall survival (OS) is here considered as the primary end-point. It is defined as the duration from the date of diagnosis to death, with no restriction on the cause of death that can be for cancer or not. With this choice of end-point the variable reporting the time until progression will not be considered but the information on the experience of patients progression is still included in the variable `pcdx`.

The response variables for each subject are then the time from diagnosis of MGUS to the last follow up (`futime`) and the opposite of the censoring state (`death`), defining if death has been observed or not.

After the removal of variable `pctime`, the modification of variable `pcdx` to let it include patients who did not experience progression, and the removal of subjects for which some values were unknown, a subset of observations from the new dataset `"mgus_os"` takes the form shown below.

```
> head(mgus_os)
  id    sex dxyr pcdx futime death alb creat  hgb mspike
1  1 female   68    0    748     1 2.8   1.2 11.5    2.0
3  3   male   68    0    277     1 2.2   1.1 11.2    1.3
4  4   male   69    0   1815     1 2.8   1.3 15.3    1.8
5  5 female   68    0   2587     1 3.0   0.8  9.8    1.4
6  6   male   68    0    563     1 2.9   0.9 11.5    1.8
7  7 female   68    0   1135     1 3.0   0.8 13.5    1.3
```

This data set is then used to fit a Cox model that can later be used for prediction purposes.

4.2 Function "coxph"

This section shows the main features of the function `coxph()` using as an example the previously presented data set `mgus_os`.

The first step is to fit a Cox regression model to the data in order to get an estimate of the vector of coefficients $\hat{\boldsymbol{\beta}}$ and this is done using the function `coxph()` of the survival package. Three continuous variables, namely the age of the individuals and their initial values of creatinine and hemoglobin, showed to be statistically significant fitting a Cox model with all the variables and will thus be used for the analysis of overall survival.

```
> coxfit<-coxph(Surv(futime,death)~age+creat+hgb,data=mgus_os)
> coxfit
Call:
coxph(formula = Surv(futime,death)~age+creat+hgb,data=mgus_os)

           coef exp(coef)  se(coef)      z       p
age    0.072865  1.075585  0.008446  8.627 < 2e-16
creat  0.421832  1.524752  0.138054  3.056 0.00225
hgb   -0.117225  0.889385  0.056031 -2.092 0.03643

Likelihood ratio test=94.1  on 3 df, p=< 2.2e-16
n= 176, number of events= 165
```

Other arguments can be specified inside the `coxph` function (see [7] for more details) and, in particular, `method=c("efron","breslow","exact")` allows to choose the way in which tied event times should be handled in the computation of the partial likelihood (see Section 3.2). The default option is to use Efron approximation.

The main terms of the resulting `coxph` object are briefly described in order to have a clear understanding of what information could eventually be used to make future predictions, that is the final goal:

- **coefficients:** the vector of the estimated coefficients $\hat{\boldsymbol{\beta}}$.

- **concordance:** the concordance statistic for the model, that is used to measure the discriminative power of a risk prediction model. In survival analysis, a pair of patients is called *concordant* if the risk of the event predicted by a model is lower for the patient who experiences the event at a later timepoint. The concordance probability (C-index) is the frequency of concordant pairs among all pairs of subjects.

- **means:** vector $\bar{\mathbf{x}}$ of values used as the "reference" for each covariate. This is not statistically ideal since it could be seen as the representation of an “average” patient, which is not because it just represents a patient with the mean value for the covariates [5].

- **linear.predictors:** the linear predictors for each observation, centered with respect to the "reference" value: $\hat{\eta}_i = \hat{\boldsymbol{\beta}}'(\mathbf{x}_i - \bar{\mathbf{x}})$, $i = 1, \ldots, n$.

- **logLik:** a vector containing the partial log-likelihood (Eq.3.6) computed with the initial values and the final values of the coefficients.

Finally, once the model has been fitted to the observed survival data, the interest is on predicting a survival estimate for an additional patient. In order to do this, the functions `predict.coxph()` and `survfit.coxph()` are analyzed and discussed.

4.3 Function "`predict.coxph`"

The aim of this section is to analyze which kind of prediction can be obtained from the `predict()` function if applied to the result of a fitted Cox model. The interest would be on having predictions for the survival time of a new patient, but it will be clear at the end of this section that such a direct prediction cannot be obtained and only other types of predictions are provided.

The function `predict.coxph()` produces various types of predicted values from a Cox model that will be examined one by one, but firstly the main arguments of the function are presented:

- **object:** an object of class `coxph` that is the result of a fitted Cox model.
- **newdata:** optionally, a new dataset with values of the covariates that characterize new individuals for which predictions are desired. If this is absent, predictions are for all the observations used to fit the model.
- **type:** the type of predicted value. Choices are `"lp"`, `"risk"`, `"expected"`, `"terms"` or `"survival"`, that will be discussed later on in more details.
- **se.fit:** whether or not to return pointwise standard errors of the predictions.
- **na.action:** how to handle missing values if there is new data.
- **terms:** the terms that are desired if `type="terms"` is chosen.
- **collapse:** an optional vector of subject identifiers, over which to sum or 'collapse' the results of the prediction.
- **reference:** the reference context for centering the results obtained with choices of type `"lp","risk"` and `"terms"`. If not defined, this is given by $\bar{\mathbf{x}}$ obtained from the fitted Cox model with `coxfit$means`. The option `"zero"` causes no centering to be done.

[5].

Since the aim is to make predictions for new patients, the `newdata` argument will here be used to describe a patient with covariates $\mathbf{x}_{new}$. Note that the function requires this argument to be a dataset, thus a new patient could be represented as follows:

```
new_patient<-data.frame(age=74,creat=0.8,hgb=9.8).
```

Now the different types of predicted values are discussed in more details.

- `type="lp"`

  The default predicted value for a Cox PH model is the linear predictor, that corresponds to the predicted risk score of Equation 3.18 centered with respect to the mean values of the variables, that is indicated as $\bar{\mathbf{x}}$. The mean is used because it is more practical, since it is just needed to get $\hat{\boldsymbol{\beta}}'\mathbf{x}$ in the neighborhood of zero. It represents the log-hazard ratio centered with respect to the reference value; in formula, the linear predictor is given by:

  $$\hat{\eta}_{new} = \hat{\boldsymbol{\beta}}'(\mathbf{x}_{new} - \bar{\mathbf{x}}) = \hat{\beta}_1(x_{1,new} - \bar{x}_1) + \cdots + \hat{\beta}_p(x_{p,new} - \bar{x}_p). \tag{4.1}$$

  This is a *relative* measure, meaning that it is relative to the sample used to fit the model. This value gives a measure of the risk of the new patient relative to the risk of a "reference" patient who has the mean values for all the covariates.

```
> pred_lp<-predict(coxfit, newdata = new_patient,type="lp")
> pred_lp
       1
1.052579
```

- `type="risk"`

  If this is the choice for the type of prediction, the outcome corresponds to the exponential of the linear predictor, again centered with respect to a "reference" set of covariates, which represents the hazard ratio. In formula:

  $$e^{\hat{\eta}_{new}} = e^{(\hat{\boldsymbol{\beta}}'(\mathbf{x}_{new} - \bar{\mathbf{x}}))} = e^{\hat{\beta}_1(x_{1,new} - \bar{x}_1) + \cdots + \hat{\beta}_p(x_{p,new} - \bar{x}_p)}. \tag{4.2}$$

  Since it comes directly from the previous linear predictor, this is also a *relative* measure, and not an absolute one, for the risk of a new subject.

```
> pred_risk<-predict(cox,newdata = new_patient,type="risk")
> pred_risk
       1
2.865029
```

  The interpretation of this value is that the new patient has a risk of death that is 2.87 times the risk of a "reference" patient from the original study sample.

- `type="expected"`

  The description of the function refers this value as "the expected number of events given the covariates and follow-up time". If the argument `newdata` is used, the output is a single value that corresponds to the *cumulative hazard function* at a specific time, that depends on the follow-up time for the future subject as well as on his or her covariates. This type of prediction is generally meaningful when an observation can have multiple events as it gives an estimate of how many event-times are expected over the predefined follow-up time. In formula, this type of prediction gives the cumulative hazard:

  $$\hat{H}_{new}(t) = \int_0^t \hat{h}_0(u) \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new}) du. \tag{4.3}$$

Note that this predicted value is a function of time and *does* require an estimate of the baseline hazard that needs therefore to be computed. In this case, the argument `newdata` requires both the right and the left hand side of the formula that defines the Cox model (Section 4.2): the variable `status` will not be used, but is required since the underlying code needs to reconstruct the entire formula, while the variable `time` represents the time at which it is of interest to compute the cumulative hazard function. Lets consider again an example. If the follow-up time for the new individual is 365 days, than the new data, that represent the subject of the previous examples, should now be defined to consider also this follow-up time:

```
new_patient_time<-data.frame(futime=365,death=1,age=74,
                             creat=0.8,hgb=9.8)
```

The predicted cumulative hazard at that time is:

```
> pred_expected<-predict(coxfit,newdata = new_patient_time,
                         type="expected")
> pred_expected
[1] 0.09601368
```

Since predictions of type `expected` incorporate an estimate of the baseline hazard, they are *absolute* measures instead of relative ones.

- `type="terms"`

  With this option, the terms of the linear predictor (relative to the "reference") are produced, one for each explanatory variable in the model:

  $$\{\,\hat{\beta}_j(x_{j,new}-\bar{x}_j)\,\}_{j=1}^{p}\,. \tag{4.4}$$

  ```
  > pred_terms<-predict(coxfit,newdata=new_patient,type="terms")
  > pred_terms
          age      creat       hgb
  1 0.7882628 -0.1279875 0.3923033
  attr(,"constant")
  [1] 3.528067
  ```

  The term `attr(,"constant")` refers to the linear predictor for the "reference" patient, whose covariates are given by `coxfit$means`. In formula, this means:

  $$\hat{\bar{\eta}} = \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}} = \hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_p\bar{x}_p. \tag{4.5}$$

  As for the predictions of type `lp` and `risk`, it is worth underlying that predictions of type `terms` are relative to the original sample of patients, thus not absolute measures.

- `type="survival"`

  This option produces a predicted survival probability for a new patient at a specific time and it is directly computed from the prediction of type `"expected"` as exp(-expected):

  $$\hat{S}_{new}(t) = \exp(-\hat{H}_{new}(t)). \tag{4.6}$$

Again note that it is a function of time and so the follow-up time for the future subject has to be provided to the argument `newdata`. The prediction is then a single value that represents the probability of survival until a specific time point.

```
> pred_survival<-predict(coxfit,newdata = new_patient_time,
                         type="survival")
> pred_survival
[1] 0.9084516
```

As the prediction of type `expected`, this type of prediction is absolute rather then relative.

It can be seen that the contribution of the function `predict.coxph()` to the prediction of the survival time of a new patient is only indirect. Predictions of the prognostic index give a measure of the relative risk of that new patient, but since this is a relative measure, it is not adequate to provide an estimate of the survival time. On the contrary, predicted probabilities of survival are absolute measures and can inform an individual on what is his or her probability to survive until a specific time point, but cannot directly provide estimates of survival times.

To proceed with the goal of prediction, there is another function in R that is used to predict quantities from a Cox model and this is the `survfit()` function. The next section will analyze this function and will show which types of predictions can be obtained.

4.4 Function "survfit.coxph"

The function `survfit.coxph()` of the survival package computes the predicted survivor function for a Cox PH model. The default procedure used by the function is to estimate the baseline survivor function as the exponential of the cumulative hazard, better known as the Breslow estimate (Eq.3.13) presented in Section 3.3. The option `newdata` allows to produce a subject-specific survival curve that is representative for an individual whose covariates $\mathbf{x}_{new}$ correspond to the values in `newdata`. The arguments of the function are:

- **formula:** an object of class `coxph` that is the result of a fitted Cox model.
- **newdata:** contains the data values of the new individuals for which curves should be predicted. If it is not present, the value of `coxfit$means` from a fitted Cox model is used as the default covariate set.
- **se.fit:** logical value true/false that indicates if pointwise standard errors of the survival curve should be computed.
- **conf.int:** the level of the two-sided confidence interval on the survival curve(s). Default is 95%.
- **stype:** states if the survival curve should be computed directly (=1) or as exponential of the cumulative hazard; since the direct estimate of survival can be very difficult to compute, the default procedure is to compute an estimate of the cumulative hazard function and use the relation of Equation 2.4.

- **ctype:** option to include correction for ties in the computation of the cumulative hazard (Sect. 3.2), where 1=no, 2=yes.
- **conf.type:** which transformation to use in the computation of the confidence intervals for the survival curve(s); default is the log-transformation in Equation 3.17 (`"log"`). Other options are `"none"`,`"plain"`,`"log-log"` or `"logit"`.
- **censor:** if false, any times that have no events are removed from the output.
- **id:** optional variable name of subject identifiers.
- **start.time:** a single numeric value that gives an optional starting time. If present, the result is a *conditional* survival curve that contains survival after time `start.time` conditional on surviving to that time-point.
- **influence:** option to return the influence values in case of multi-state data. Since it is not the case here, comments on its meaning are left to [6].
- **na.action:** the action to be used for new data if there are missing values

[7].

Using the `survfit()` function on a previously fitted Cox model and the explanatory variables that characterize a new individual, the outcome is the following.

```
> surv<-survfit(coxfit,newdata=new_patient)
> surv
Call: survfit(formula = coxfit, newdata = new_patient)

       n events median 0.95LCL 0.95UCL
[1,] 176    165   2587    2016    3855
```

From the response of the function a summary measure, the *median survival time*, is given together with its 95% confidence interval, for which more details are given in Section 4.4.1. In addition, combining the result with the `plot()` function, the *predicted individual survival curve* of Figure 4.1 is obtained.

```
> plot(surv,main="Predicted Survival Curve",xlab="Time",
       ylab="Survival Probability")
> segments(median_time,0,median_time,surv_prob,lty=2,col="red")
> segments(0,surv_prob,median_time,surv_prob,lty=2,col="red")
```

Other values can then be extrapolated from a 'survfit' object:

```
> str(surv)
 $ n          : int 176
 $ time       : num [1:173] 6 7 31 32 39 61 277 362 370  ...
 $ n.risk     : num [1:173] 176 175 174 173 172 171 169 168 167 ...
 $ n.event    : num [1:173] 1 1 1 1 1 2 1 1 1  ...
 $ n.censor   : num [1:173] 0 0 0 0 0 0 0 0 0 ...
 $ surv       : num [1:173] 0.99 0.98 0.97 0.959  ...
 $ cumhaz     : num [1:173] 0.0101 0.0204 0.0308 0.0414  ...
 $ std.err    : num [1:173] 0.0103 0.015 0.0189 0.0224  ...
```

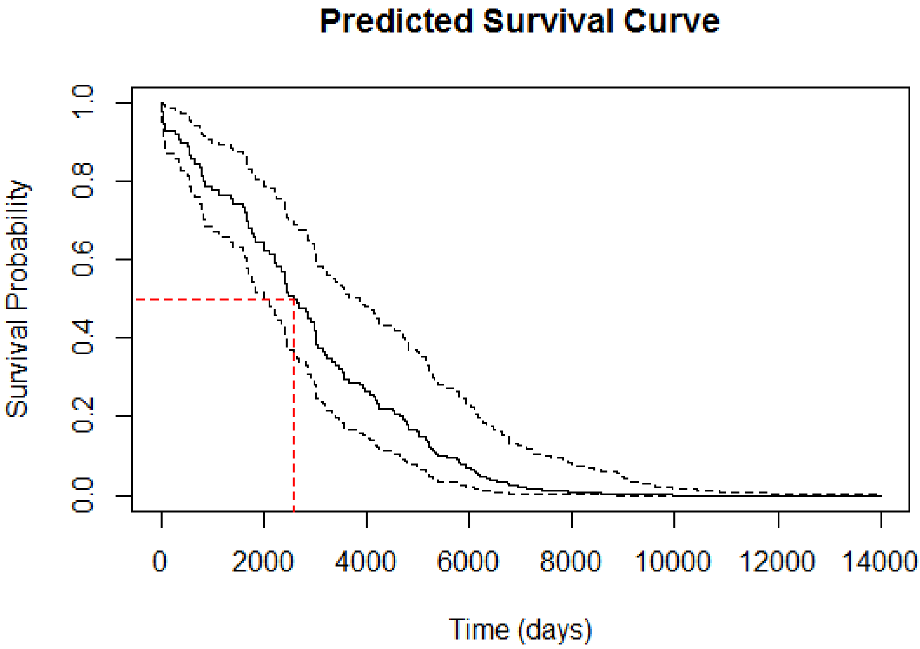


Figure 4.1. Predicted individual survival curve for a new individual with his or her predicted median survival time (red). Broken lines represent a 95% confidence interval.

```
$ logse    : logi TRUE
$ std.chaz : num [1:173] 0.0103 0.015 0.0189 0.0224  ...
$ lower    : num [1:173] 0.97 0.951 0.934 0.918 0.903 ...
$ upper    : num [1:173] 1 1 1 1 0.998 ...
$ conf.type: chr "log"
$ conf.int : num 0.95
$ call     : language survfit(formula=coxfit,newdata=new_patient)
- attr(*, "class")= chr [1:2] "survfitcox" "survfit"
```

It can be seen that standard errors and confidence intervals are computed for each event time together with the estimated survival probability (`surv$surv`).

An estimate of the cumulative hazard function is also provided and it is given by `surv$cumhaz`. Note that the same cumulative hazard can be obtained from another function `basehaz()` in R but, since this latter does the actual work of `survfit()` and has less options, there are not more advantages in using it. If a time point at which to compute the cumulative hazard is specified, then this result will also be the same as the prediction of type `"expected"` of the previously discussed `predict.coxph()` function. The `survfit` object has to be combined with the `summary()` function, specifying the time at which the hazard should be computed with the argument `time=y`. Assuming again that $y = 365$ days,

```
> s_probability<-summary(survfit(coxfit,newdata=new_patient),
                         time=365)
> s_probability$cumhaz
[1] 0.09601368
```

it can be seen that the results of the two functions coincide.

Finally, estimates of the probability of survival until a specific time y for the new patient can also be obtained from $\hat{S}_i(t)$:

$$\hat{S}_i(y) = \{\hat{S}_0(y)\}^{exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)}. \tag{4.7}$$

Referring again to the example, the predicted probability of survival until time $y = 365$ days is:

```
> s_probability<-summary(survfit(coxfit,newdata=new_patient),
                         time=365)
> s_probability$surv
[1] 0.9084516
```

This predicted probability coincides with the prediction of type `"survival"` in the function `predict.coxph`.

To summarize, the function `survfit.coxph()` produces the entire estimated survival curve, rather then predicted probabilities of survival for predefined follow-up times as done by the `predict()` function. Next, having the entire estimated individual curve, quantities that refer to specific time points can also be extracted. Furthermore, the survivor distribution can be summarized in single values that could be used to obtain predictions for new individuals that are median and restricted mean survival time and are described in the next two sections 4.4.1 and 4.4.2 respectively.

4.4.1 Median Survival Time

The first quantity that can be derived from the estimated survival curve of a new patient with covariates $\mathbf{x}_{new}$ is his or her *median survival time* that is directly shown in the output of the function. It corresponds to the smallest time for which the value of the estimated survivor function $S_{new}(t)$ is smaller than 0.5, that is

$$\begin{aligned}\hat{t}_m &= \min\{t_{new} \mid \hat{S}_{new}(t_{new}) < 0.5\} \\ &= \min\{t_{new} \mid \hat{S}_0(t_{new})^{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})} < 0.5\}.\end{aligned} \tag{4.8}$$

This is visualized by graphing the estimated survival function and drawing a horizontal line at probability value of 0.5: the estimated median time equals the time where the function and line intersect (Figure 4.1). The 95% confidence interval for the median $\hat{t}_m$ is then given by the points at which this horizontal line crosses over the pointwise confidence intervals of $\hat{S}_{new}(t)$. The median survival time and its 95% confidence interval are given directly by the `survfit()` function and they can be extracted as follows.

```
> summary(surv)$table['median']
median
```

```
  2587
> cbind(as.numeric(summary(surv)$table[’0.95LCL’]),
      as.numeric(summary(surv)$table[’0.95UCL’]))
     [,1] [,2]
[1,] 2016 3855
```

Estimates of other *percentiles* can also be obtained from the estimated survival curve $\hat{S}_{new}(t)$ with the same definition. The estimated p^{th} percentile is in fact defined as the smallest observed time $\hat{t}(p)$ for which

$$\hat{S}_{new}\{\hat{t}(p)\} < 1 - (p/100).$$

In R the percentiles, together with their confidence intervals, can be obtained from the function `quantile()` applied to a `survfit` object. Hereafter it is shown, as an example, the 25th percentile with its 95% confidence interval.

```
> quantile(surv,probs=0.25)$quantile
  25
1392
> cbind(quantile(surv,probs=0.25)$lower,
        quantile(surv,probs=0.25)$upper)
   [,1] [,2]
25  748 2339
```

Median survival time provides a summary of a predicted survival curve that, on a population level has a clear useful meaning, but on an individual level the question is whether is good or not to predict a patient's survival time whit that time-point for which he or she has 50% estimated probability of survival if a group of identical individuals is observed. It could be questioned how accurate would be such a prediction. In addition, one drawback of using the median survival time as an estimate for the survival time is that it is not always defined: if the estimated survivor function of an individual stays above 0.5 for the entire follow-up period, it can only be said that the median time is greater than the last observed time, but it remains undefined. For this reason it would be of interest to find another value that could be used for prediction of survival time, that has not this issue and is always defined.

Before introducing this alternative measure, an application to the study sample in `mgus_os` is shown. Exploiting the leave-one-out procedure of cross validation, one by one each subject is considered to be a new patient and a prediction of his or her survival time is made. The Cox regression model is fitted on the training set of all the other individuals while the selected patient is used as a single-item test set.

Results are shown graphically in Figure 4.2 in which predicted survival time is compared to the actual one for non-censored patients. This choice is done for comparative purposes; in fact, for censored patients their actual time is unknown and the comparison with the predicted time is not immediate. Survival time is plotted against the predicted prognostic index (Eq.4.1 obtained with `predict.coxph()`)in order to visualize the relation among the predicted risk of patients and their survival times. As expected, patients with higher predicted score are predicted to have smaller survival time.

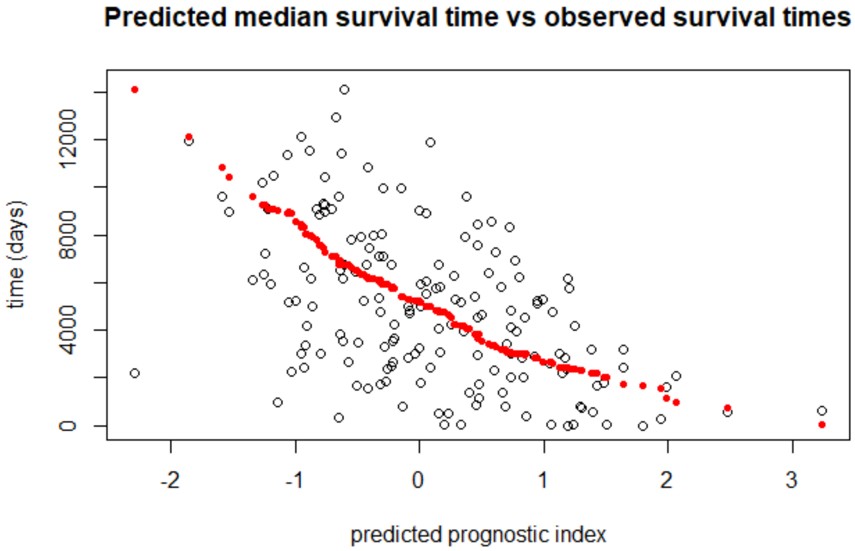


Figure 4.2. Predicted median survival time (red) and observed event time (black) for each non-censored patient of the original study sample.

Predicted median survival time is, in this example, defined for each patient. However it can happen that it is not defined for some individuals and a general measure that does not suffer from this issue is preferable.

4.4.2 Restricted Mean Survival Time

In addition to median survival time, a second quantity can be derived from the `survfit()` function and this is called *restricted mean survival time (RMST)*. Its definition is firstly presented and then it is shown how to obtain this value in R.

The idea is to estimate survival time with its expected value under a certain probability distribution and one way to compute this expected value is to go from an estimate of the hazard function (Section 3.3) to an estimate of the distribution function $F(t)$ of survival times. The actual survival time of an individual t can be regarded as the observed value of a *random variable* T that has a probability distribution. Assuming that $f(t)$ is its probability density function, the distribution function of T is then given by

$$F(t) = Prob(T < t) = \int_0^t f(u)\,du, \tag{4.9}$$

that is linked to the survivor function through the following relation:

$$F(t) = Prob(T < t) = 1 - Prob(T \geq t) = 1 - S(t). \tag{4.10}$$

The *expected* survival time is then defined as:

$$E[T] = \int_0^\infty u\, dF(u) = \int_0^\infty u\, f(u)\, du\,. \tag{4.11}$$

Integrating by parts and using the relation in Equation 4.10, the integral becomes

$$\begin{aligned}\int_0^\infty u\, f(u)\, du &= uF(u)|_0^\infty - \int_0^\infty F(u)\, du \\ &= u(1 - S(u))|_0^\infty - \int_0^\infty (1 - S(u))\, du \\ &= \int_0^\infty S(u)\, du,\end{aligned}$$

so that finally

$$E[T] = \int_0^\infty S(u)\, du. \tag{4.12}$$

Since an estimate of $\hat{S}(t)$ is available using Equation 3.14, it could be put into Equation 4.12 in order to obtain the estimated expectation of survival time. However, the integral in Equation 4.12 will diverge if $\hat{S}(t)$ does not converge to zero and the survivor function is only defined until the last observed time t_{max}. This property creates a challenge for most data, and one resolution is to use a finite value τ as the bound for the integral, where τ may be a predetermined time point (smaller than the maximum observed time t_{max}) or the maximum observed time t_{max} itself. Restricting the computation of the expected value to this time leads to the restricted mean survival time:

$$\mu_\tau = \int_0^\tau S(t)\, dt \tag{4.13}$$

[2]. The RMST is then defined as the area under the survival curve up to a time τ and its interpretation on a population level would be “when patients are followed-up for τ, patients will survive for μ_τ on average”, which is quite a straightforward and clinically meaningful summary of the censored survival data. For a single individual, RMST is interpreted as the time he or she is expected to survive if followed for a time period of τ

From Equation 4.13 a natural *estimate* of the restricted mean survival time for an individual with covariates $\mathbf{x}_i$ is given by:

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}_i(t)\, dt. = \int_0^\tau \hat{S}_0(t)^{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)}\, dt. \tag{4.14}$$

Equation 4.14 can then be used to predict the time for a new patient, substituting $\mathbf{x}_i$ with the new covariates $\mathbf{x}_{new}$.

In R such a prediction may be obtained using the function `survfit()` in conjunction with the function `print()`, that gives the restricted mean survival time with its standard error: the `print.rmean=TRUE` argument is used to obtain them.

```
> print(surv,print.rmean = TRUE)
Call: survfit(formula = coxfit, newdata = new_patient)
```

```
       n events rmean se(rmean) median 0.95LCL 0.95UCL
[1,] 176    165  2833      86.9   2587    2016    3855
    * restricted mean with upper limit =  14325

> summary(surv)$table['rmean']
   rmean
2833.478
```

The upper bound τ is automatically set as the largest observed or censored time (in the example 14325 days). If one is interested in computing the mean time until a precise time point, a different τ may be specified using the `rmean` argument

As done with the median survival time, restricted mean survival time is used to predict time-to-event for each non-censored patient in the original data set `mgus_os` using technique of leave-one-out cross validation. For each patient the chosen upper bound τ is taken to be the last observed time among the others. Results are shown in Figure 4.3 where predicted restricted mean survival time is compared with the observed one.

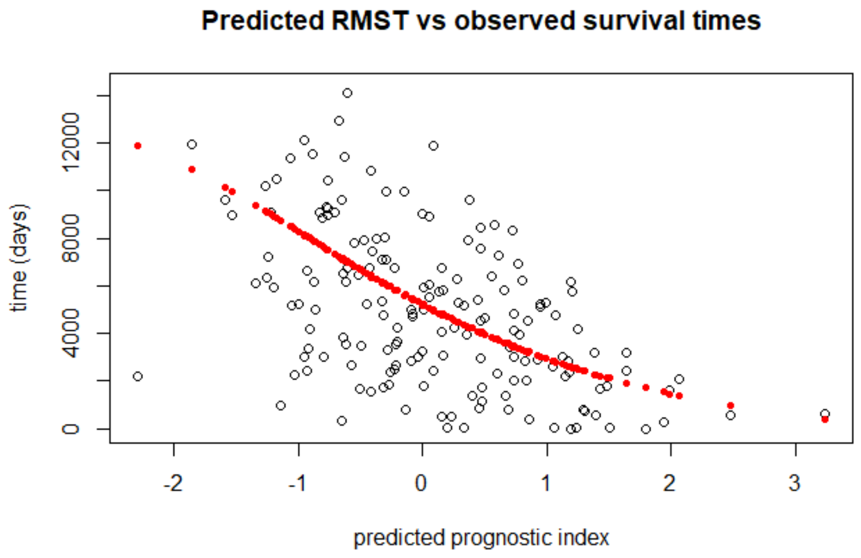


Figure 4.3. Predicted restricted mean survival time (red) and observed event time (black) for each non-censored patient of the original study sample.

In conclusion, the main contribution of the function `survfit.coxph()` to the prediction of survival times, is to provide estimates of subject-specific survival curves. This latter is advantageous from a visualization point of view and, although it does not directly estimate the survival time of the new individual, it can be used to extrapolate some values that can provide a kind of estimate for this time, like the median survival time or the mean restricted survival time. Due to the median survival time not always been defined, however, it seems more reasonable to take the restricted mean survival time for this purpose.

The reason why the survival package in R doesn't contain a function that directly predicts survival times from a Cox model is because this latter has not been developed for this purpose. The Cox model is a *relative risk* model that can explain the impact of several explanatory variables on survival times of individuals but this relationship is modelled only through the hazard function. In order to have estimates of time it is necessary to go through an estimation of the hazard function and also in R the informative measures for times prediction (like median survival time or restricted mean survival time) are computed only after an estimation of the survivor function.

A first objective should be to assess if these estimated quantities could be employed to adequately predict survival times for new patients; secondly, it may be worth considering if there are alternatives that could give direct estimates of time, without having to estimate the hazard function. The next chapter aims to discuss if there are reasonable alternatives or improvements in the direction of non-parametric prediction.

Chapter 5

Non Parametric Survival Times Predictions

The aim of this chapter is to describe a possible candidate method, that is different from those already available in the survival package in R, to be used for the estimation of times-to-event for new individuals that enter into a study. Having a Cox PH model that has been fitted and validated for a certain population, the objective is to determine a value that can provide a prediction for a survival time when a new patient is added in order to answer the question of "how long will he or she survive?". At this scope, the definition of a novel *Cox prediction time* is provided.

5.1 "Inversion" of Partial Likelihood

Until now it has not been shown any direct method that allows to compute estimates of survival times without needing an estimation of hazard and survivor functions. While the Cox model is a semi parametric model with unknown shape of the hazard function, the estimates of the previous chapter (median and restricted mean survival time) are computed only after an estimation of the survivor function and are thus classified as parametric approaches. Preserving the strength of the Cox model, the desire would be to develop a non-parametric method to predict survival times.

An attempt in this direction is to use the partial likelihood involved in the Cox model (Eq.3.5): as shown in Section 3.2, maximization of the partial likelihood is used to estimate the regression coefficients of the Cox model in Equation 3.1. For its computation, the *ranking* of the observed survival times is exploited, rather than the exact survival times themselves. In simple words, this process takes as input the covariates of the patients together with their survival times and gives as an output the vector of estimated regression coefficients $\hat{\boldsymbol{\beta}}$, solving Equation 3.7 for each coefficient. For a sample of n patients, each characterized by p explanatory variables, the input-output process is summarized as:

$$(\mathbf{x}_1, \ldots, \mathbf{x}_n, t_1, ..., t_n) \rightarrow (\hat{\beta}_1, \ldots, \hat{\beta}_p). \tag{5.1}$$

The idea is then to "invert" this process: if the regression coefficients are known from a previous fitted Cox model and the covariates of the patients are also available, the goal is to obtain the estimated survival times:

$$(\mathbf{x}_1, \dots, \mathbf{x}_n, \hat{\beta}_1, \dots, \hat{\beta}_p) \to (t_1, ..., t_n). \tag{5.2}$$

In order to obtain the survival time for each patient in the study, a number of equations that equals the number of unknown variables should be needed. However, most of the times, there will be more individuals than predictor variables (with relative coefficients) and consequently more unknowns than equations. If the aim is to estimate the time of a single individual, only one equation is necessary if the survival times of the other individuals are known. Suppose then that the objective is to estimate the time of the k^{th} individual, the "inverse" process would be summarized as:

$$(\{\mathbf{x}_i\}_{i \neq k}, x_k, \{t_i\}_{i \neq k}, \{\hat{\beta}_j\}_{j=1,\dots,p}) \to t_k. \tag{5.3}$$

The "inverse" process is defined through the same equation used to find the coefficients $\hat{\boldsymbol{\beta}}$, but this time solved for the unknown time t_k. First of all, the derivative of the partial log-likelihood of Equation 3.7 can be rewritten so to highlight the contribution of survival times:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i \left[x_{ji} - \frac{\sum_{l=1}^{n} \mathbb{1}(t_l \geq t_i) \exp(\boldsymbol{\beta}' \mathbf{x}_l) x_{jl}}{\sum_{l=1}^{n} \mathbb{1}(t_l \geq t_i) \exp(\boldsymbol{\beta}' \mathbf{x}_l)} \right] = 0, \qquad j = 1 \dots p, \tag{5.4}$$

where the characteristic function $\mathbb{1}(t_l \geq t_i)$ is used to indicate the risk set $R(t_i)$: it is equal to one if t_l is greater than t_i and zero otherwise.

Next, suppose that the survival time of the k^{th} individual is unknown, while the regression coefficients and the survival times of all the other patients in the study are given. Then from Equation 5.4 it would be possible to obtain the time t_k of the k^{th} individual: the derivative of the partial log-likelihood with respect to each β_j is theoretically equal to zero if computed at the estimated coefficients $\hat{\boldsymbol{\beta}}$:

$$\left. \frac{\partial \ell}{\partial \beta_j} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = 0, \qquad j = 1 \dots p. \tag{5.5}$$

In order to solve the equation for the unknown t_k, only one of the p equations in 5.5 has to be considered. The coefficient β_j to be used in the computation of the derivative of the partial likelihood is then arbitrary, resulting in the same solution t_k; here a generic β_j is considered. Equation 5.4 can then be rewritten highlighting the time t_k and substituting the vector of coefficients $\boldsymbol{\beta}$ with its estimate $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} \left. \frac{\partial \ell}{\partial \beta_j} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} = & \sum_{i \neq k} \left[x_{ji} - \frac{\sum_{l=1, l \neq k}^{n} \mathbb{1}(t_l \geq t_i) \exp{(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} x_{jl} + \mathbb{1}(t_k \geq t_i) \exp{(\hat{\boldsymbol{\beta}}' \mathbf{x}_k)} x_{jk}}{\sum_{l=1, l \neq k}^{n} \mathbb{1}(t_l \geq t_i) \exp{(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} + \mathbb{1}(t_k \geq t_i) \exp{(\hat{\boldsymbol{\beta}}' \mathbf{x}_k)}} \right] \\ & + x_{jk} - \frac{\sum_{l=1, l \neq k}^{n} \mathbb{1}(t_k \leq t_l) \exp{(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)} x_{jl} + \exp{(\hat{\boldsymbol{\beta}}' \mathbf{x}_k)} x_{jk}}{\sum_{l=1, l \neq k}^{n} \mathbb{1}(t_k \leq t_l) exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l) + \exp{(\hat{\boldsymbol{\beta}}' x_{jk})}} = 0. \end{aligned} \tag{5.6}$$

The only unknown of equation 5.6 is the time t_k and if this equation could be solved for this variable, then it would be possible to obtain the survival time for the k^{th} individual.

However, if a solution exists, it will not be a single value but a range of times that lie in-between two observed survival times; in fact, as already stated, the partial likelihood does not use the exact values of the survival times, but only their relative ranking. This means that, "inverting" the process, the outcome will be a position of the k^{th} patient in the ranking of the already known survival times of the others.

The "inverse" process can be summarized as follows:

- Consider an estimate of the vector of regression coefficients $\hat{\boldsymbol{\beta}}$ obtained from a Cox regression model using the method of maximum likelihood estimation.
- Consider the k^{th} patient with known covariates $\mathbf{x}_k$ but unknown survival time t_k.
- "Solve" Equation 5.6 for the unknown t_k using $\hat{\boldsymbol{\beta}}$ and the known survival times of the other patients t_i, $i \neq k$.
- Obtain the range of survival times for the k^{th} individual that corresponds to its relative ranking among the survival times of the other patients.

In using Equation 5.6 to obtain ranges of survival times some practical considerations need to be done:

- First of all, Equation 5.6 has been obtained from the partial likelihood (Eq. 3.5) of the Cox model that, as discussed in section 3.2, does not consider the possibility of two events happening at the same time point. In order to handle tied-observations, approximations of the partial likelihood are necessary and thus other equations need to be used for the "inverse" process, computing derivatives of these approximated versions of the partial likelihood. Equation 5.6 could still be used in case of ties, adding a small error to the time points that coincide, so to avoid the problem and have all distinct event times.
- Equation 3.7 is not solved exactly for the regression coefficients β_j but only through iterative methods, like the Newton Raphson method. The result is that the derivative of the partial log-likelihood in Equation 5.4, computed at the estimated values $\hat{\beta}_j$, that theoretically should be equal to zero, will in practice only be approximately so:

$$\left.\frac{\partial \ell}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \approx 0 \quad \text{for} \quad j = 1, \dots, p.$$

5.1.1 "Inverse" approach for prediction

When a new patient with covariates $\mathbf{x}_{new}$ is added to the study, the approach summarized by Equation 5.6 could still be used to provide an interval of his or her survival time, that again corresponds to a ranking among known times. Replacing the unknown t_k with the unknown t_{new} leads to the following:

$$\begin{aligned}\left.\frac{d\ell}{d\beta_j}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = &\sum_{i=1}^{n}\left[x_{ji} - \frac{\sum_{l=1}^{n}\mathbb{1}(t_l \geq t_i)\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)x_{jl} + \mathbb{1}(t_{new} \geq t_i))\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})x_{j,new}}{\sum_{l=1}^{n}\mathbb{1}(t_l \geq t_i)\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) + \mathbb{1}(t_{new} \geq t_i)\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})}\right] \\ &+ x_{j,new} - \frac{\sum_{l=1}^{n}\mathbb{1}(t_{new} \leq t_l)\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)x_{jl} + \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})x_{j,new}}{\sum_{l=1}^{n}\mathbb{1}(t_{new} \leq t_l)\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) + \exp(\hat{\boldsymbol{\beta}}'x_{j,new})} = 0,\end{aligned} \tag{5.7}$$

for $j = 1, \dots, p$. In this case the resulting interval for t_{new}, if it exists, is interpreted as the survival time that would give the same estimated coefficients, if patient with covariates $\mathbf{x}_{new}$ would belong to the study sample used to fit the model. Practically, if a Cox model were fit including the new subject with covariates x_{new}, what is asked with the "inverse" method is that the difference in the estimated coefficients in $\hat{\boldsymbol{\beta}}$ should be zero. In general, the difference $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(k)}$ in the estimates of the regression coefficients including and not including the k^{th} individual (respectively given by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(k)}$) is called *delta-beta* and can be approximated by:

$$\Delta_k \hat{\boldsymbol{\beta}} \approx \mathbf{r}'_{Uk} var(\hat{\boldsymbol{\beta}}), \tag{5.8}$$

where $\mathbf{r}'_{Uk}$ represents the score residual for the k^{th} individual (see [2] for its definition). Even if the mean of these delta-beta residuals is zero, in real examples it is rare to find single observations that have $\Delta\hat{\boldsymbol{\beta}}$ equal to zero and estimating the survival time of a new individual asking for this difference to be zero seems not so reasonable. It sounds rather more appropriate to consider the difference in the coefficients given by a new subject inside the algorithm, using $\hat{\boldsymbol{\beta}} + \Delta_{new}\hat{\boldsymbol{\beta}}$ instead of $\hat{\boldsymbol{\beta}}$ in solving Equation 5.7. If the value of $\Delta_{new}\hat{\boldsymbol{\beta}}$ is known for a new patient or there is a way to estimate it, then an accurate predicted interval can be found solving the following equation for t_{new}, using the same approach described earlier for the k^{th} patient in the data set:

$$\left.\frac{d\ell(t_{new})}{d\beta_j}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}+\Delta_{new}\hat{\boldsymbol{\beta}}} = 0. \tag{5.9}$$

It is worth noting again that the choice of the coefficient β_j to use in Equation 5.9 is arbitrary since the vector $\hat{\boldsymbol{\beta}}$ has been estimated satisfying all the equations.

However, when the value of the delta-beta residual for the new patient is not known, the solution t_{new} of Equation 5.7 may be different for the different β_j. As previously mentioned, this is a consequence of the fact that in practice new individuals will almost never leave the estimates unchanged. For this reason a new global measure needs to be used that takes into account derivative of partial log-likelihood with respect to all p coefficients and is given by the $l2$-norm of the gradient of the partial log-likelihood:

$$\|\nabla\ell\| = \left\|\left(\frac{\partial\ell}{\partial\beta_1}, \dots, \frac{\partial\ell}{\partial\beta_p}\right)\right\|. \tag{5.10}$$

Substituting the coefficients β_j with their estimates in $\hat{\boldsymbol{\beta}}$, the objective would be to have the norm of the gradient equal to zero. However, the nature of this method implies that a solution (convergence) cannot always be achieved and the aim is then to search for the time that provides the smallest $l2$-norm.

Definition 1. *The event time prediction, or* Cox prediction time*, based on the minimization of the l2-norm of the partial log-likelihood gradient, is defined as:*

$$\begin{aligned}\hat{t}_{new} &= \arg\min_{t_{new}} \left\|\nabla\ell(t_{new})\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right\| \\ &= \arg\min_{t_{new}} \sqrt{\left(\left.\frac{\partial\ell(t_{new})}{\partial\beta_1}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right)^2 + \dots + \left(\left.\frac{\partial\ell(t_{new})}{\partial\beta_p}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right)^2},\end{aligned} \tag{5.11}$$

where the derivatives of the partial log-likelihood, that constitute the components of the gradient, are those in Equation 5.7.

This definition for the Cox prediction time is just a point definition, representing the best time point corresponding to the minimum $l2$-norm in Definition 1. The solution of this minimization problem is usually not a single one but it turns out to be an interval of survival times. The reason of this claim is the nature of the Cox regression model itself, which only works with ranks of survival rather than absolute values of observed times.

Theorem 1. *Let consider a sample of n individuals, each characterized by a set of predictors $\boldsymbol{x}_i$, an observed survival time t_i and its censoring state δ_i. Let define the order statistics of the time intervals, i.e the ordered observed survival times $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$.*

For a new patient with known set of predictors $\boldsymbol{x}_{new}$, the Cox prediction time t_{new} is the solution of the minimization problem in Definition 1. This solution is invariant to the choice of t_{new} in any of the intervals defined by the order statistics of the time intervals. It turns out that:

- *if t_{new} is a solution such that $t_{(i)} < t_{new} < t_{(i+1)}$, then the entire interval $(t_{(i)}, t_{(i+1)})$ is a solution;*
- *the search of the minimizer in Equation 5.11 reduces to the search of the best over $n+1$ possible intervals.*

Proof. The two propositions exposed in Theorem 1 follow directly from the definition of the partial likelihood given for a Cox regression model. The latter works only with the rank of the survival times and their contribution to the partial likelihood is observed only in the risk-sets. Considering the order statistics of the time intervals, the contribution of time t_{new} to the partial log-likelihood derivative can be highlighted. Suppose here that $\mathbf{x}_i$ represents the covariates for the patient with time $t_{(i)}$.

$$\begin{aligned}\left.\frac{d\ell}{d\beta_j}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} =& \sum_{i=1}^{n}\left[x_{ji} - \frac{\sum_{l=i}^{n}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)x_{jl} + \mathbb{1}(t_{new} \geq t_{(i)})\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})x_{j,new}}{\sum_{l=i}^{n}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) + \mathbb{1}(t_{new} \geq t_{(i)})\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})}\right] \\ &+ x_{j,new} - \frac{\sum_{l=1}^{n}\mathbb{1}(t_{new} \leq t_{(l)})\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)x_{jl} + \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})x_{j,new}}{\sum_{l=1}^{n}\mathbb{1}(t_{new} \leq t_{(l)})\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) + \exp(\hat{\boldsymbol{\beta}}' x_{j,new})}.\end{aligned} \tag{5.12}$$

If $t_{(k)} < t_{new} < t_{(k+1)}$, then:

$$\begin{aligned}\left.\frac{d\ell}{d\beta_j}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} =& \sum_{i=1}^{k}\left[x_{ji} - \frac{\sum_{l=i}^{n}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)x_{jl} + \mathbf{1}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})x_{j,new}}{\sum_{l=i}^{n}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) + \mathbf{1}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})}\right] \\ &+ \sum_{i=k+1}^{n}\left[x_{ji} - \frac{\sum_{l=i}^{n}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)x_{jl}}{\sum_{l=i}^{n}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)}\right] \\ &+ x_{j,new} - \frac{\sum_{l=k+1}^{n}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)x_{jl} + \exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_{new})x_{j,new}}{\sum_{l=k+1}^{n}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l) + \exp(\hat{\boldsymbol{\beta}}' x_{j,new})}.\end{aligned} \tag{5.13}$$

In Equation 5.13 the time t_{new} is not explicit, highlighting that the position in the ordered observed times $t_{(k)}$, $k = 1, \ldots, n$, is the only information used about the time. If the derivative of the partial log-likelihood with respect to each coefficient depends only on the $k-th$ interval for the new patient, the same will be for the norm of the gradient. As a consequence, an entire interval $(t_{(k)}, t_{(k+1)})$ will be the solution for the $l2$-norm minimization in Definition 1.

Since it has been shown that values of $\|\nabla\ell(t_{new})\|$ change only in correspondence of different positions k in the order of the survival times, it is sufficient to consider a finite number of possible values for its minimization. This finite number equals the number of intervals defined by the order statistics $t_{(1)}, \ldots, t_{(n)}$: it is equal to $n-1$ in-between intervals plus two intervals (one for those times smaller than $t_{(1)}$ and one for those greater than $t_{(n)}$). In conclusion, the minimization problem reduces to the comparison of only $n+1$ values. □

Considerations and future work

The result of the "inverse" method, in either one of the two forms presented throughout this section, is a point definition, corresponding in particular to a range of times. A comparison with the linear regression setting allows to support the "inverse" approach by one side, and to think on possible improvements by the other. If survival times would be related to individual predictors through a simple linear regression, then prediction of individual survival time for a new patient would simply be obtained as:

$$t_{new} = \hat{\beta} x_{new}. \tag{5.14}$$

In this context, regression coefficients are estimated minimizing the residual sum of squares and they are obtained imposing the derivative of the latter to be equal to zero. The question is now whether inverting this process, as done for the Cox model, would give the same predictions. In practice it can be seen that the same predictions of Equation 5.14 can be obtained following an "inverse" approach asking for a time prediction that gives the same estimated coefficients (or the smallest variation). This result supports the idea of inverting the equations of the partial log-likelihood also in the context of a more complex Cox regression model, as proposed at the beginning of this section.

In analogy of linear regression, the desire would also to obtain real prediction intervals for the Cox prediction times. As already mentioned, the definition of a Cox prediction time is a point prediction; however it would be of interest to find a prediction interval in which a future observation will fall, with a certain probability. Until now it has not identified a clear method to obtain such intervals but further investigation could be done to improve available information on individual times prediction.

The next section provides practical examples on how to implement this non-parametric approach for survival times prediction, highlighting in particular the main issues and difficulties that are encountered.

5.2 Application

In practical applications, it is worth distinguishing if the new patient for whom an estimate of the survival time is wanted, has also been used to fit the original Cox model or not. In the first case, the interval obtained from the "inverse" process (summarized in Equation 5.6) will contain the exact observed time for uncensored patients; indeed, it represents the time that would give those estimated coefficients that are supposed to be already known. In the second case, instead, the computed interval will not necessarily contain the actual time if the same method is followed. Since the difference in the estimates of the regression coefficients is rarely zero, as discussed in the theoretical section, knowing its value would lead to predicted survival intervals that are close to the actual observed times. For example, accurate survival time prediction would be obtained if values of the delta-beta residuals would be known. Otherwise, the request for the time of the new patient would be to be to minimize the variation in the estimates of the coefficients, knowing that this latter will almost never be zero.

The "inverse" non-parametric approach has been implemented in R and it is hereafter shown for both cases previously mentioned.

5.2.1 Survival time for the k-th patient in the study sample

In this section it is assumed that the regression coefficients are computed fitting a model over the entire data set of individuals and that in a second moment the interest is on obtaining the time of the k^{th} individual.

The computation of the derivative of partial log-likelihood in Equation 5.6 is done by the function `derivative_pll()`, that depends on the data of the individuals in the sample excluding the k^{th} patient (`data_small`), the covariates that characterize the new individual (`xnew`) and the estimated coefficients from a previous Cox model (`coeff`). With specific values of the just mentioned parameters, the function is a function of the only unknown `tnew`. As already stated, the choice of the coefficient β_j with respect to which compute the derivative is arbitrary; here the coefficient β_1 relative to the first variable `x_1` is used. Suppose that variables `time` and `status` represent the observed event-time and the censoring status of the subjects respectively and that three covariates characterize each patient. The function is implemented as follows.

```
derivative_pll<-function(tnew,data_small,xnew,coeff){
 ns=nrow(data_small)
 d_i=0*c(1:ns)

 for (i in (1:ns)){
   d1=as.matrix(data_small[data_small$time>=data_small$time[i],
                 c("x_1","x_2","x_3")])
   d2=as.matrix(data_small[data_small$time>=data_small$time[i],
                "x_1"])
   m1=sum(exp(d1%*%coeff)*d2)
   m2=sum(exp(d1%*%coeff))

   #contribution of the i-th individual to the pll derivative
   d_i[i]= data_small$status[i]*(data_small$x_1[i]-((m1+
          ifelse(tnew>=data_small$time[i],1,0)*
          exp(as.vector(xnew)%*%coeff)*xnew[1])/
          (m2+ifelse(tnew>=data_small$time[i],1,0)*
          exp(as.vector(xnew)%*%coeff))) )
   }

 d3=as.matrix(data_small[data_small$time>=tnew,
              c("x_1","x_2","x_3")])
 d4=as.matrix(data_small[data_small$time>=tnew,"x_1"])

 m3=sum(exp(d3%*%coeff)*d4)
 m4=sum(exp(d3%*%coeff))

 #contribution of the new individual to the pll derivative
 d_new= xnew[1]-( (m3+exp(as.vector(xnew)%*%coeff)*xnew[1])/
 (m4+exp(as.vector(xnew)%*%coeff)) )

 d_pll=sum(d_i)+d_new
 return(d_pll)
}
```

The "inverse" process is now summarized for the situation in which the actual estimated vector $\hat{\boldsymbol{\beta}}$ is known and the actual observed event time for the k^{th} subject can be used to compare the results of the method. Suppose to have an original study sample made of n patients and p predictors, then it can be proceeded as follows:

1. Remove tied-observations (if any) so to use Equation 3.5 for the computation of the partial likelihood.

2. Fit a Cox regression model on the data of the n individuals.

3. Obtain the estimated vector of regression coefficients $\hat{\boldsymbol{\beta}}$ maximizing the partial log-likelihood of the model and using iterative methods to solve the equations.

4. Take the k^{th} patient with covariates $\mathbf{x}_k$ out of the n ones and suppose now that his or her time t_k is unknown.

5. Choose one j out of the p coefficients and compute the derivative of the partial log-likelihood with respect to that β_j.

6. Insert the vector of estimated coefficients $\hat{\boldsymbol{\beta}}$ and the survival times of the $n-1$ individuals in Equation 5.6.

7. "Solve" the equation for the time t_k.

8. Obtain the relative ranking for the survival time of the k^{th} patient.

One way to solve Equation 5.6 in point 6. is to use grid search and choose the time for which the computed derivative of the partial log-likelihood, that will never be exactly zero due to approximation, is smallest and approximately zero. Since the function does not need an exact time but just the ranking of all survival times of the subjects, it is sufficient to define a vector (`time_grid`) that includes the information on the ranking: it will contain a time point between every two consecutive observed times and this will be used to obtain the lower and upper bounds of the interval, that correspond to those two consecutive times.

```
time_grid=0*c(1:(nrow(data_small)+1)  #time points between two
                                       #consecutive observed times
for (k in (2:nrow(data_small))){
    time_grid[k]=(data_small_ordered$time[k]+
                  data_small_ordered$time[k-1])/2
  }
time_grid[1]=data_small_ordred[1]-1/2
time_grid[n+1]=data_small_ordered$time[nrow(data_small)]+1/2
```

Again the dataset mgus_os with its relative notation is used for illustrative purposes; the 8 points that characterize the method are now shown more in detail using this data set as an example.

1. Adjust for tied-observations: a small error of 0.01 has been added for two equal event-times and a new data set mgus_noties is obtained, where only one event occurs at each observed time.

2. Fit a Cox model:

   ```
   cox_noties<-coxph(Surv(futime,death)~age+creat+hgb,
                     data=mgus_noties).
   ```

3. Get the estimated regression coefficients:

   ```
   beta_noties<-as.vector(cox_noties$coefficient).
   ```

4. Choose the k^{th} patient: suppose, as an example, to be interested in obtaining the survival time of the first patient. The explanatory variables that characterize this patient are:

   ```
   > mgus_noties[1,]
     id age    sex dxyr pcdx futime death alb creat  hgb mspike
   1  1  78 female   68    0    748     1 2.8   1.2 11.5      2
   ```

   where it can be seen that the observed event time is equal to 748 days; this value is used for comparison with the result of the "inverse" process. Since only the three variables age, creat and hgb are used in the Cox regression model, the covariates for the new subject are

   ```
   x_k<-as.matrix(mgus_noties[1,c("age","creat","hgb")]).
   ```

5. Define the coefficient to use for the computation of function derivative_pll(). If the first explanatory variable age is used, Equation 5.6 implemented in the function refers to $\frac{\partial \ell}{\partial \beta_{age}} \approx 0$.

6. Define the derivative with respect to β_{age} as a function of time tnew:

   ```
   derivative_pll(tnew,mgus_train,x_k,beta_noties),
   ```

   where mgus_train refers to the data set obtained excluding the first patient (mgus_train<-mgus_noties[-1,]).

7. As already said, the equation is here simply solved with grid search, using the vector `time_grid` that identifies the possible new intervals. For each value in the vector, it is saved the value of the computed partial log-likelihood derivative and the interval for which the value is approximately zero (the smaller in absolute value in this case) is taken.

```
for ( j in (1:length(time_grid))){
    d_pll[j]=abs(derivative_pll(time_grid[j],mgus_train,x_k,
                 beta_train))
  }
i_opt<-which(d_pll==min(d_pll))
if (length(i_opt)==1){
    i1<-i_opt
    i2<-i_opt
}else{
    i1<-i_opt[which(i_opt==min(i_opt))]
    i2<-i_opt[which(i_opt==max(i_opt))]
}
if(i1==1){
    opt_u<-mgus_train_o$futime[i2]
    opt_l=0
}else if (i2==(nrow(mgus_train)+1)){
    opt_u<-max(mgus_train_o$futime)
    opt_l<-mgus_train_o$futime[i1-1]
}else{
    opt_l<-mgus_train_o$futime[i1-1]
    opt_u<-mgus_train_o$futime[i2]
}
interval_k<-cbind(opt_l,opt_u)
```

8. Finally the bounds of the new interval are obtained from the observed times of the individuals and the resulting survival interval for the first patient is:

```
 > interval_k
[1] 652 779
```

   that contains the actual time of 748 days.

This process can be repeated for each individual in the study and the result is an interval for each of them that contains the actual observed time. Note that this is true for the subjects who were not censored, but the interval for the patients whose survival time was censored does not necessarily contain the original censored survival time, since it is supposed to contain the actual event-time that is not known in this case.

For each patient it can can be computed the predicted score (Eq.3.18) using the function `predict.coxph` of Chapter 4 and plot the survival time against this value to compare the observed time with the predicted one. The resulting interval from the "inverse" approach is shown in red in Figure 5.1 where it is compared to the observed survival time (black points) for those subjects who were uncensored. From Figure 5.1 it is evident that the obtained intervals cover exactly the observed times for each patient whose actual

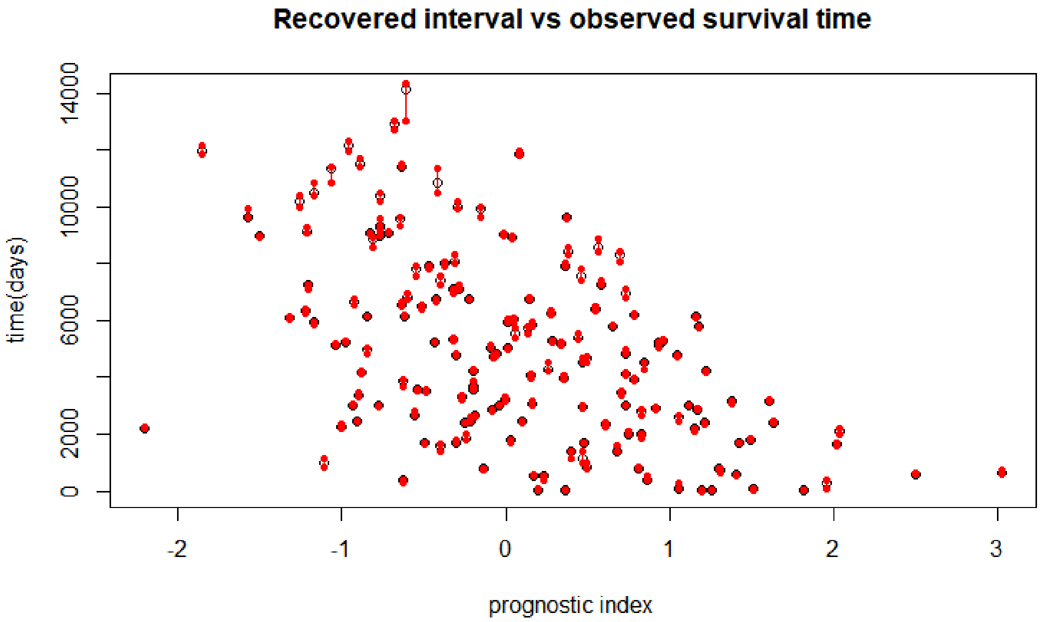


Figure 5.1. Predicted survival interval (red) and observed event time (black) for each non-censored patient of the original study sample in `mgus_noties` knowing the exact value for $\hat{\beta}$.

time was known. The size of the predicted intervals varies among the patients and this is because it depends on the *known* observed survival times of the other individuals in the study and, in particular, on their relative distance. This aspect underlies the strong dependence of this methodology on the observed data.

If a single time-point is preferred rather than an interval of times, then the median time of the predicted interval could be taken as the survival time prediction for the k^{th} individual:

```
t_k=(interval_k[1]+interval_k[2])/2
> t_k
[1] 715.5
```

Otherwise it could be thought of an exponential distribution for the survival function between the two bounds of the interval (`t1` and `t2`). Computing the empirical survival probability for these two time-points (`s1` and `s2`) allows to compute an exponential function for the considered interval and to take the time that corresponds to the mean of the survival probabilities of the two bounds.

```
b=(log(s2)-log(s1))/(t2-t1)
a= s1*exp(-b*t1)
t_k<-1/b*( log(((s1+s2)/2)/a))
>t_k
```

```
[1] 715.4005
```

5.2.2 Survival time for a new patient

In the second case, the new patient is supposed not to be in the original study used to fit the model. As for the parametric cases, leave-one-out cross validation is used to validate the prediction. The new patient is taken again as one of the original n subjects for whom the actual survival time is known but the model is then fitted over the remaining $n-1$ individuals.

If the values of the delta-beta residuals are known, the function `derivative_pll()` can be used to get a prediction of the intervals, including this further information. If for example, the first patient is considered to be the new one, his or her predictors are given by:

```
x_new<-as.matrix(mgus_noties[1,c("age","creat","hgb")])
```

Using the function `residuals()` on a fitted Cox model over the full data set, allows to obtain the delta-beta residuals $\Delta_1\hat{\beta}$ that are shown to be different from zero.

```
> res_beta<-residuals(cox_noties,type="dfbeta")
> res_beta[1,]
[1]   0.0004202891  0.0017042735 -0.0028970961
```

If this information were known for the new subject, than it could be included in the model in order to obtain times that are nearer to those actually observed. In fact, adding the approximated difference on the estimated coefficients when computing the partial likelihood derivative,

```
derivative_pll(tnew,mgus_train,x_new,beta_train+res_beta_[1,])
```

allows to obtain an interval that is close to the observed time of death (748 days):

```
> interval_new[1,]
[1] 779 791
```

If this process is repeated for each non-censored patient with the leave-one-out approach, the obtained results are the ones in Figure 5.2.

As more realistically happens, when exact values or estimates of the difference in the regression coefficients are not given, Definition 1 is used to obtain a time prediction. Rather than computing the partial log-likelihood derivative with respect to one coefficient, now a global measure is needed and the $l2$-norm of the gradient of the partial log-likelihood is used instead, as discussed in Section 5.1. This latter is implemented in the function `gradient_norm` that again, with specific values of `data_small`, `xnew` and `coeff`, is a function of the only unknown `tnew`.

```
gradient_norm<-function(tnew,data_small,xnew,coeff){

  ns=nrow(data_small)

  dl1_i=0*c(1:ns)
  dl2_i=0*c(1:ns)
  dl3_i=0*c(1:ns)
```

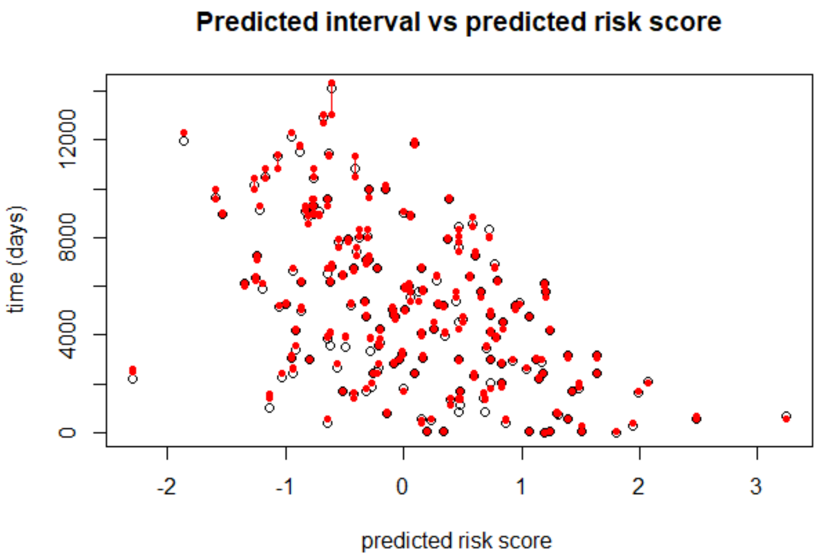


Figure 5.2. Observed and predicted survival time vs predicted risk score for each non–censored patient of the original study sample `mgus_noties` using the "inverse" process and supposing to know values of the difference in β ($\Delta_{new}\hat{\boldsymbol{\beta}} \approx$ delta-beta residuals).

```
for (i in (1:ns)){
  x=as.matrix(data_small[data_small$futime>=
              data_small$futime[i], c("age","creat","hgb")])
  x_1=as.matrix(data_small[data_small$futime>=
                data_small$futime[i],"age"])
  x_2=as.matrix(data_small[data_small$futime>=
                data_small$futime[i],"creat"])
  x_3=as.matrix(data_small[data_small$futime>=
                data_small$futime[i],"hgb"])

  num_1=sum(exp(x%*%coeff)*x_1)
  num_2=sum(exp(x%*%coeff)*x_2)
  num_3=sum(exp(x%*%coeff)*x_3)

  den=sum(exp(x%*%coeff))

  dl1_i[i]= data_small$death[i]*( data_small$age[i]-((num_1 +
            ifelse(tnew>=data_small$futime[i],1,0)*
            exp(xnew%*%coeff)*xnew[1])/
            (den+ ifelse(tnew>=data_small$futime[i],1,0)*
            exp(xnew%*%coeff))) )
```

```
    dl2_i[i]= data_small$death[i]*( data_small$creat[i]-((num_2+
              ifelse(tnew>=data_small$futime[i],1,0)*
              exp(xnew%*%coeff)*xnew[2])/
              (den+ifelse(tnew>=data_small$futime[i],1,0)*
              exp(xnew%*%coeff))) )

    dl3_i[i]= data_small$death[i]*( data_small$hgb[i]-((num_3 +
              ifelse(tnew>=data_small$futime[i],1,0)
              *exp(xnew%*%coeff)*xnew[3]) /
              (den+ ifelse(tnew>=data_small$futime[i],1,0)*
              exp(xnew%*%coeff))) )
  }

  xstar=as.matrix(data_small[data_small$futime>=tnew,
                  c("age","creat","hgb")])
  xstar_1=as.matrix(data_small[data_small$futime>=tnew,"age"])
  xstar_2=as.matrix(data_small[data_small$futime>=tnew,"creat"])
  xstar_3=as.matrix(data_small[data_small$futime>=tnew,"hgb"])

  num_star1=sum(exp(xstar%*%coeff)*xstar_1)
  num_star2=sum(exp(xstar%*%coeff)*xstar_2)
  num_star3=sum(exp(xstar%*%coeff)*xstar_3)

  den_star=sum(exp(xstar%*%coeff))

  dl1_new= xnew[1]-((num_star1+exp(as.vector(xnew)%*%coeff)*
           xnew[1])/ (den_star+exp(as.vector(xnew)%*%coeff)))
  dl2_new= xnew[2]-((num_star2+exp(as.vector(xnew)%*%coeff)*
           xnew[2])/ (den_star+exp(as.vector(xnew)%*%coeff)))
  dl3_new= xnew[3]-((num_star3+exp(as.vector(xnew)%*%coeff)*
           xnew[3])/ (den_star+exp(as.vector(xnew)%*%coeff)))

  dl1=sum(dl1_i)+dl1_new
  dl2=sum(dl2_i)+dl2_new
  dl3=sum(dl3_i)+dl3_new

  norm=sqrt(dl1^2+dl2^2+dl3^2)
  return(norm)
}
```

The idea would be to use the same process of the previous case but using the estimated coefficients of a Cox model fitted over the sample of the remaining $n-1$ subjects instead of using the full sample of n patients and using the function `gradient_norm()` instead of `derivative_pll()` which computes instead the derivative with respect to only one chosen coefficient.

Except for these changes, the methodology used in practice looks very similar to the previous one:

1. Remove tied-observations and obtain `mgus_noties` so to use Equation 3.5 for the partial likelihood.

2. Take the k^{th} patient with covariates $\mathbf{x}_k$ out of the n ones: he or she is treated as the new patient with covariates $\mathbf{x}_{new} = \mathbf{x}_k$ and survival time t_{new}. If $k = 1$:

   ```
   x_new<-as.matrix(mgus_noties[1,c("age","creat","hgb")]).
   ```

3. Fit a Cox regression model on the data of the other $n-1$ individuals.

   ```
   mgus_train<-mgus_noties[-1,]
   cox_train<-coxph(Surv(futime,death)~age+creat+hgb,
                    data=mgus_train)
   ```

4. Obtain the estimated vector of regression coefficients $\hat{\boldsymbol{\beta}}$ maximizing the partial log-likelihood of the model and using iterative methods to solve the equations.

   ```
   beta_train<-as.vector(cox_train$coefficients)
   ```

5. Compute the gradient for the partial log-likelihood together with its norm as done in Equation 5.10 and insert the vector of estimated coefficients $\hat{\boldsymbol{\beta}}$ and the survival times of the $n-1$ individuals:

   ```
   gradient_norm(tnew,mgus_train,x_new,beta_train)
   ```

6. Minimize the equation for the time t_{new}. As for the previous case, grid search is used to find the time that satisfies the request in Definition 1.

   ```
   for ( j in (1:length(time_grid))){
       grad_norm[j]=gradient_norm(time_grid[j],mgus_train,x_new,
       beta_train)
     }
   i_opt<-which(grad_norm==min(grad_norm))
   if (length(i_opt)==1){
       i1<-i_opt
       i2<-i_opt
   }else{
       i1<-i_opt[which(i_opt==min(i_opt))]
       i2<-i_opt[which(i_opt==max(i_opt))]
   }
   if(i1==1){
       opt_u<-mgus_train_o$futime[i2]
       opt_l=0
   }else if (i2==(nrow(mgus_train)+1)){
       opt_u<-max(mgus_train_o$futime)
       opt_l<-mgus_train_o$futime[i1-1]
   }else{
       opt_l<-mgus_train_o$futime[i1-1]
       opt_u<-mgus_train_o$futime[i2]
   }
   interval_new<-cbind(opt_l,opt_u)
   ```

8. Obtain the relative ranking for the survival time of the new individual among the others:

```
> interval_new
[1]  3027  3062
> t_pred=(interval_new[1]+interval_new[2])/2
> t_pred
[1]  3044.5
```

The observed event-time for patient 1 in this case was 748 days, quite far from the predicted one. Exploiting leave-one-out cross validation, results of the Cox prediction time are given by Figure 5.3 that shows the Cox prediction times (intervals in red) compared to the observed death-times (black points).

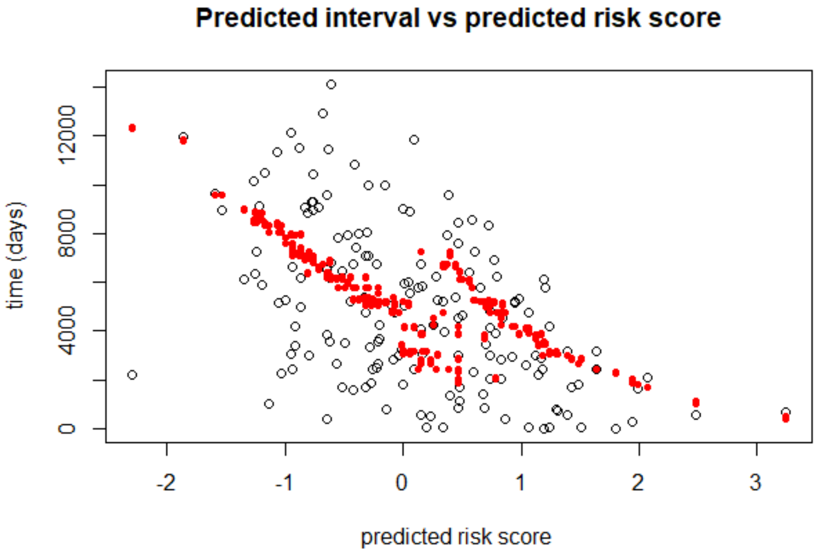


Figure 5.3. Observed and predicted survival time vs predicted risk score for each non–censored patient of the original study sample `mgus_noties` using the "inverse" process asking for the smallest difference in the coefficients.

Summarising all the results obtained, it has been shown that theoretically it would be possible to "invert" the process of maximum likelihood estimation solving Equation 5.6 and obtaining an interval of survival times for a certain patient. This is verified in the case in which the right values of the coefficients $\hat{\beta}$ are known and the time of the k^{th} individual can in this way be "recovered". If only the coefficients computed over a certain population not including the new patient are available, as more realistically would be, then practical issues arise in solving Equation 5.7. If estimates of the difference in the regression coefficients $\Delta\hat{\beta}$ can be computed, than again it is possible to make the "inversion" including the latter information and get prediction times that are close to the real death-times. In the general case, when such values are not known, practical issues are solved proposing a new definition (Def 1) based on the minimization of the $l2$-norm of the partial log-likelihood gradient. The function `gradient_norm()` is minimized through grid search, exploiting the nature of the Cox model which is only based on the ranking of the survival times. Results seem not to overcome the limitations of the parametric approaches of Section 4 and in the next Chapter a comparison is done through visualization and some performance evaluation metrics.

Chapter 6

Results and Comparisons

In this chapter possible performance metrics for prediction evaluation are defined and are then used to compare the three methods (median survival time, restricted mean survival time and cox prediction time obtained from the "inverse" approach) for survival time prediction.

6.1 Performance Evaluation Metrics

Once a method for prediction of survival times has been identified, its performance should be evaluated to assess whether the predictions are accurate or not. However, performance evaluation is not straightforward in survival analysis: the censoring problem of survival data is the main reason why survival models are difficult to evaluate, since the actual survival time of certain subjects is not known. One way could be to not consider the information of the censored data and construct a measure that is used only over the uncensored sample.

A first measure that will be used to compare the performance of the different methods proposed, is the *root mean square error (RMSE)* for the uncensored observations:

$$RMSE = \sqrt{\frac{1}{n_{event}} \sum_{i=1}^{n} [\delta_i (t_{pred,i} - t_{obs,i})]^2}, \tag{6.1}$$

where n is the observed sample size and n_{events} is the event sample size that corresponds to individuals who showed genuine death. The RMSE measures the difference between the predicted survival time $t_{pred,i}$ and the observed uncensored survival time for each individual. Note that RMSE defined in this way can be used to compare performances of different prediction models over the same data but is not appropriate to compare performances over different data sets since it depends on the scale of times, that can vary a lot from one data set to the other.

For this scope, a very intuitive way to evaluate performance starts from a definition of *"serious error"* given by Parkes [1]: the error in prediction is defined to be serious if the predicted outcome differs from the observed event-time by a multiplicative factor of two, that is, if the predicted time is less than half the observed time ($t_{pred} \leq 0.5\, t_{obs}$) or more than twice the observed time ($t_{pred} \geq 2\, t_{obs}$). With this definition, the number of serious errors can be computed for different models in different data sets and used to compare the different prediction performances.

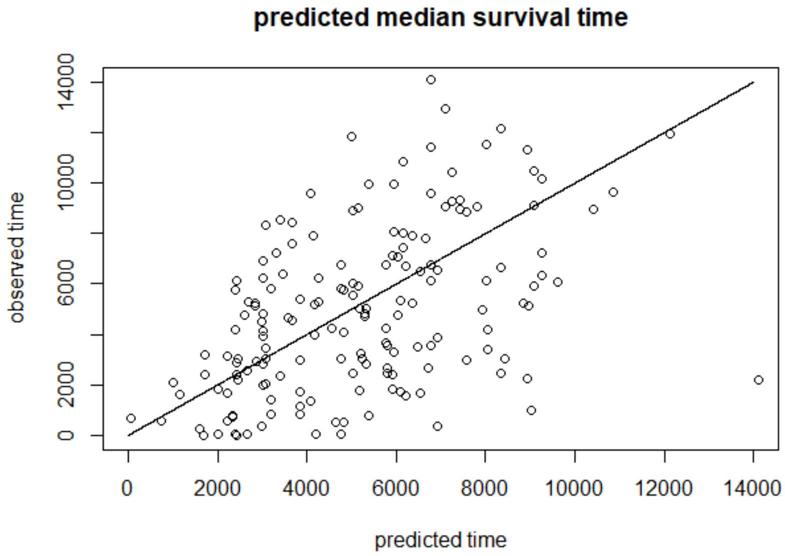
6.2 Models Comparison

In this section a comparison between the performance of the three different prediction models, namely median survival time, restricted mean survival time and the new proposed cox prediction time obtained from the "inverse" approach, is made in terms of the root mean square error for uncensored observations for the example data set `mgus_os`.

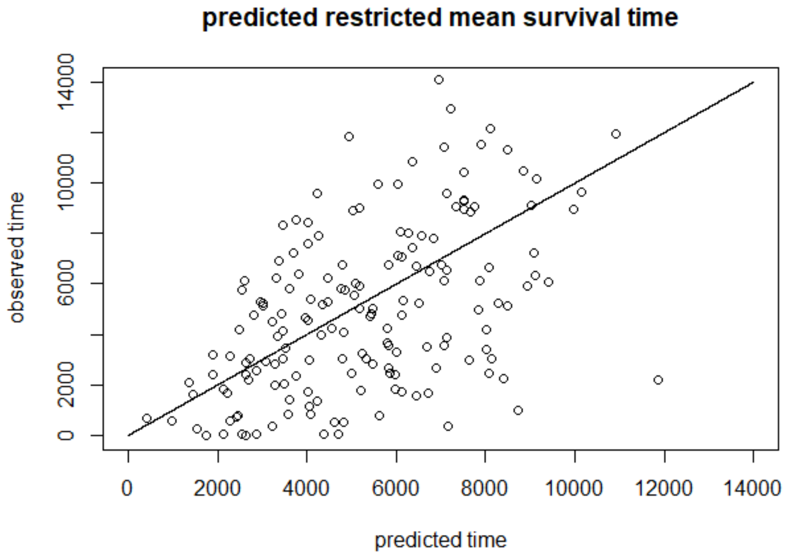
For each model a graphical representation shows the difference between the observed survival time and the prediction obtained from the model: a diagonal line would mean a perfect prediction, where predicted survival times for the uncensored patients coincide exactly with the actual ones and deviations from this line are measured by the root mean square error. Figure 6.1 shows the errors for the median survival time (a) and the restricted mean survival time (b) with their values for RMSE; restricted mean survival time turns out to be more accurate than median survival time in predicting individual survival times. In addition, although it is not the case of this example, median survival time could be undefined for certain values of the explanatory variables and using restricted mean survival time instead is even more appropriate because it does not suffer from this limitation. Concerning predictions obtained with the "inverse" algorithm, as discussed in Chapter 5, it is worth making some considerations. If the survival time is obtained from Definition 1, this should represent the time which gives the smaller variation in the estimates of the regression coefficients. The result is in general an entire interval of times, but here the mean time of each interval is taken in order to make single time points comparisons. The predicted times so obtained are shown in Figure 6.2(a). It can be seen that some predicted times are quite far from the diagonal representing perfect prediction and the RMSE is slightly higher than the two parametric methods.

If for each subject a predefined value of the difference in the β-coefficients is available, coinciding for example with the value of the delta-beta residuals, then predicted times will move close to diagonal (Figure 6.2(b)) improving prediction accuracy. However, since the beta-residuals will not be available for a new individual, another reasonable value should be found for this difference in order to obtain such accurate predictions.

[1] Parkes CM. *Accuracy of predictions of survival in later stages of cancer*, British Medical Journal 1972; 2: 29 –31.

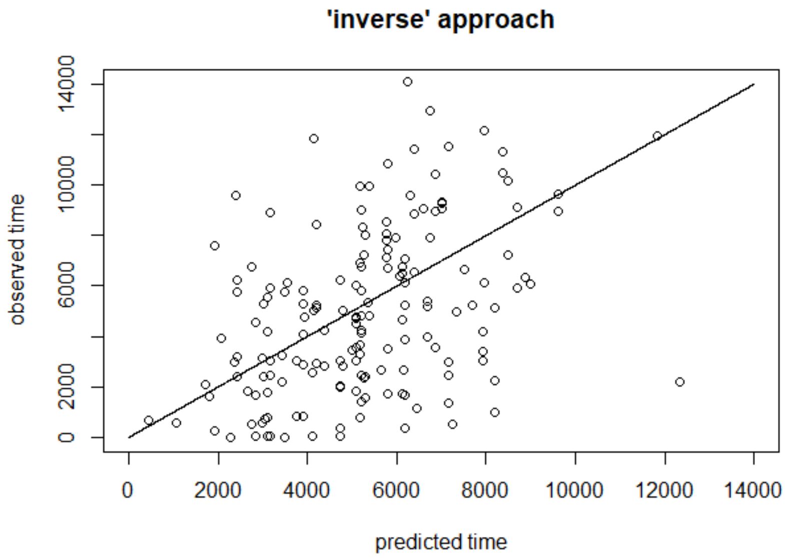


(a) RMSE: 3019.585 .

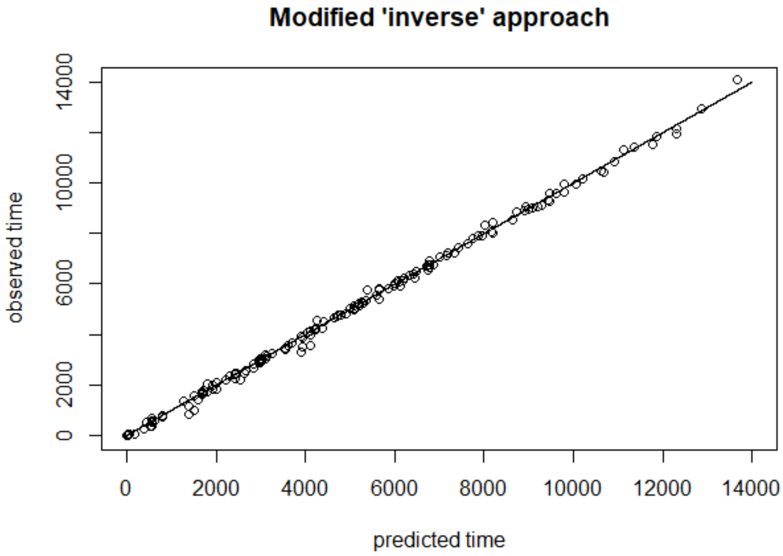


(b) RMSE: 2920.068 .

Figure 6.1. Observed vs predicted survival time for uncensored patients in the `mgus_os` data set computed with median survival time (a) and restricted mean survival time (b).



(a) RMSE: 3048.771 .



(b) $\Delta_{new}\hat{\beta} \approx$ delta-beta residuals.

Figure 6.2. Observed vs predicted survival times for uncensored patients in `mgus_os` computed with the "inverse" approach minimizing the variation in the coefficients (a) and supposing to know the exact values of the delta-beta residuals (b) .

6.3 PBC Data Set

In Section 4.1 the data set `mgus` has been presented and then used throughout this work as an example to firstly visualize the responses of the functions `predict.coxph()` and `survfit.coxph()` in R and then to show how the proposed non-parametric "inverse" method works. Here another example data set from the library `survival` in R is reported and used for comparative purposes, assessing the impact of some sample features as, for example, the number of patients and the percentage of censoring.

The data set `pbc` from the survival package collects the data from the Mayo Clinic trial in "primary biliary cirrhosis (PBC)" between 1974 and 1984 and contains a total number of 418 patients (more details on the data set in [7, Pag.76]). Primary sclerosing cholangitis is an autoimmune disease leading to destruction of the small bile ducts in the liver; progression is slow but inexhortable, eventually leading to cirrhosis and liver decompensation. This condition takes the name of "primary biliary cirrhosis".

Taking advantage of the variables already selected to be significant for overall survival in [8] and successively selecting those that meet the proportional hazards assumption, finally each patient can be described by:

- **age:** age in years.
- **edema:** categorical variable that indicates the presence of an edema (0: no edema, 0.5: untreated or successfully treated, 1: edema despite diuretic therapy).
- **albumin:** serum albumin (g/dl).
- **bili:** serum bilirunbin (mg/dl).
- **status:** status at endpoint (0: censored, 1: transplant, 2: dead). For the analysis of overall survival, liver transplantation is treated as censoring so that the event is represented by "status"=2.
- **time:** number of days between registration and the earlier of death, transplantion, or study end (July 1986).

Since in [8] it has be shown that the relationship between survival and the continuous variables `albumin` and `bili` is logarithmic rather than linear, a log-transformation is considered also in this case. A Cox regression model can then be fitted on that data.

```
cox_pbc<-coxph(Surv(time,status==2)~age+edema+log(bili)
               +log(albumin),data=pbc)
```

This dataset is here used as a second example of survival data for which prediction times are desired and can be compared with the firstly described dataset `mgus_os` in terms of prediction performances.

Two features differentiate mainly the two data sets and those are given by sample size and censoring: `mgus_os` contains 11 censored observations out of 176, while death was not observed for 257 over 418 patients in `pbc`. Since the first dataset has a low percentage of censoring (6.25%) and the second has a very high censoring (61.48%), a comparison can be done to assess how much censoring can affect prediction of survival times. The discriminative power, measured by the C-index, of the Cox model on the two different data sets is slightly different,

```
> c_index
             C-index mgus C-index pbc
concordance     0.7069157   0.8339209
```

showing higher discriminative power for the pbc data set. However, as already claimed in Section 3.4, high discriminative predictive models can show low calibration (thus accuracy) that is however fundamental in predicting individual survival times. Unfortunately this is one of these cases, in which survival times predictions are not accurate at an individual level. Hereafter the restricted mean survival time is preferred to the median survival time in order to make predictions and compare the results of the proposed algorithm, since median survival time is not defined for some individuals in the original sample. Using restricted mean survival time to predict survival times for individuals in the pbc data set, leads then to results in Figure 6.3 for non-censored patients. The reason for such a poor

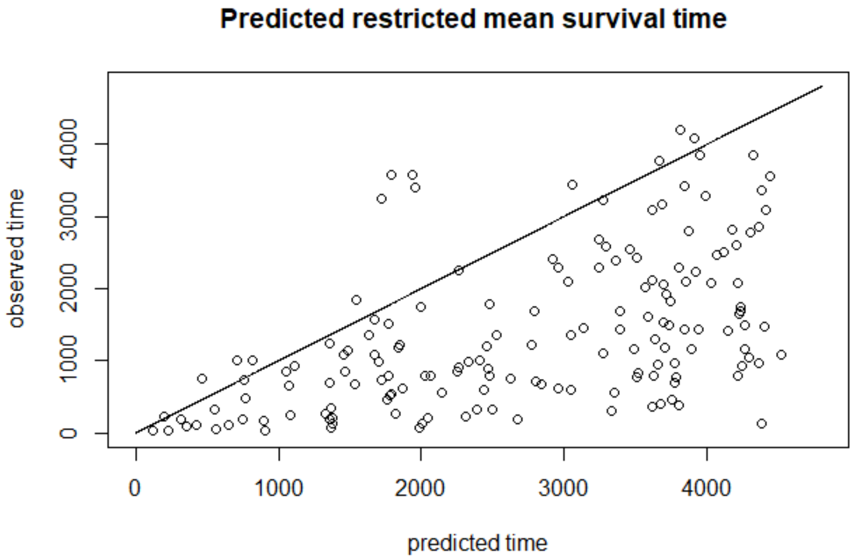


Figure 6.3. Observed vs predicted survival times for uncensored patients in pbc using restricted mean survival time (rmst) for prediction.

result is the high number of censored observations, that leads to predicted survival times that are almost always greater than the actual observed ones. Applying the new "inverse" method, results appear slightly better having a smaller mean square error and smaller percentage of "serious" errors as defined in Section 6.1.

```
> RMSE_error_pbc
         rmst   'inverse'
[1,] 1742.968    1648.604
```

```
> serious_error_pbc
          rmst   'inverse'
[1,] 0.5652174   0.5403727
```

However, Figure 6.4 shows that again the impact of censored observations is very high, affecting the quality of Cox prediction times obtained with the "inverse" algorithm. By modifying the algorithm adding the information of a known difference in the estimated values of the coefficients, i.e delta-beta residuals, high censoring would not affect much the performance as it can be seen in Figure 6.5.

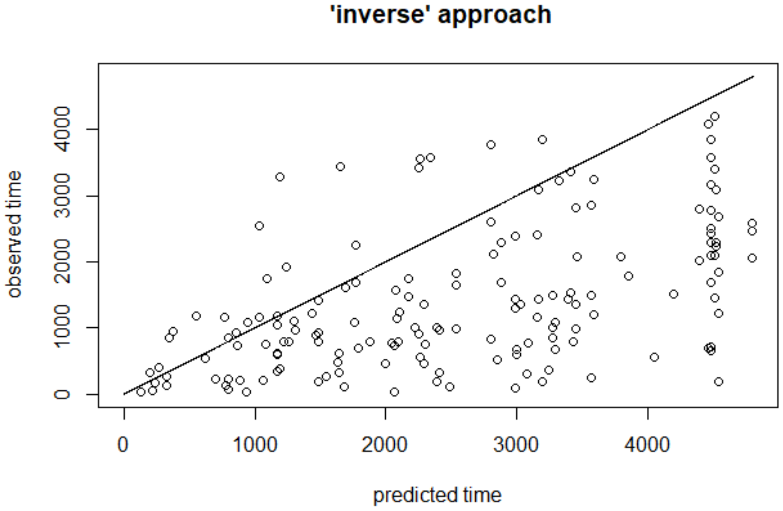


Figure 6.4. Observed vs predicted survival times for uncensored patients in `pbc` computed with the "inverse" approach.

If the number of serious errors is used as a measure for qualitatively comparing performance over the two different data sets, it can be seen that for samples with high number of censored patients, the number of errors increases. Indeed, for the `mgus_os` data set the percentage of serious errors decreases, containing very few censored observations.

```
> serious_error_mgus
          rmst   'inverse'
[1,] 0.2909091   0.2969697
```

In conclusion, the use of `mgus_os` and `pbc` data sets to assess performances of the illustrated predictive methods together with their comparison, allows to draw some considerations. First of all, it can be seen that good discriminative power is not enough for a model to make accurate predictions at the individual level. Even if the Cox model fitted over the `pbc` data set has a high value of concordance, greater than the first example proposed, the high number of censored observation affects the accuracy of individual time prediction when using the restricted mean survival time. Unfortunately the same impact of censored observations can be seen also in the "inverse" method that, even if it has a

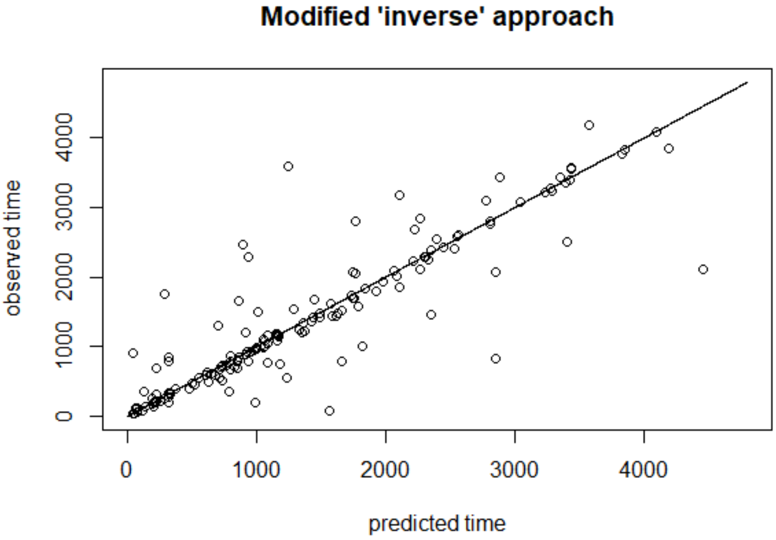


Figure 6.5. Observed vs predicted survival times for uncensored patients in `pbc` computed with the modified "inverse" approach supposing to know the exact values of the delta-beta residuals.

slightly smaller root mean square error, cannot overcome this dependence. Also in the first example of `mgus_os` the performance looks quite similar but the "inverse" method in the proposed form for a new patient, in this case, fails also in the attempt to improve the accuracy in prediction.

The dependence on censoring could be overcome if the exact values of the estimated coefficients $\hat{\boldsymbol{\beta}}$ or estimates for the delta-beta residuals $\Delta_{new}\hat{\boldsymbol{\beta}}$ were known, making possible to "invert" the equation used in the maximization of the partial log-likelihood and obtain accurate predictions on an individual level.

Chapter 7

Conclusions

Individual lifetime prediction is a challenging and interesting goal of survival analysis in clinical research since it could improve the decision-making process for the single patients moving forward into the field of personalized medicine.

The Cox proportional hazards regression model is the most popular method when dealing with survival data, explaining relationships between individual characteristics and survival experience. The standard output of this model is a relative measure of the risk of the individuals among each other, and not an absolute estimate of their survival times; indeed, for this latter objective, the model has to be combined with a parametric estimate of the hazard and survival function.

The `survival` package in R can be used to predict some useful quantities from a fitted Cox model. Function `predict.coxph()` allows to predict the relative risk score for a new patient, together with its survival probability at a specific time. Function `survfit.coxph()` instead provides a predicted individualized survival curve that gives information at an individual level: survival probabilities at each time can be extrapolated for a new patient along with its median and restricted mean survival time. Restricted mean survival time in particular turns out to be a better value for survival time prediction, even if it is strongly affected by censoring and by the assumptions of linearity of the underlying Cox model.

A novel approach that aims to somehow invert the Cox model has been proposed, starting from the partial likelihood involved in the estimation of the regression coefficients. It has been shown that, if the coefficients are known, it is possible to invert the model and obtain a range of survival times for each individual, that correspond to its ranking among the known survival times of the rest of the individuals in the study. This algorithmic approach does not require any estimation of the hazard or survival function, reflecting the semi-parametric nature of the Cox model. Inversion is possible if the right coefficients are supposed to be known but, if the same algorithm is used for prediction, practical issues coming from approximations may arise. However, if approximated values of delta-beta residuals would be available, then it would be possible to invert the model again and obtain predictions of survival times that does not suffer too much from the model assumptions. With these assumptions, it has been shown that accurate predictions can be obtained for individual patients, improving the parametric approaches of median and mean survival time.

In more realistic conditions, when such information is not available, a novel definition of a Cox prediction time is provided, based on the minimization of the $l2$-norm of the partial-likelihood gradient. The result of the approach is an interval of times rather than a single value, that aims at minimizing the variation in the estimates of the regression coefficients.

Finally, conclusions on the restrictive assumptions of a Cox model are made in order to improve in the field of individual lifetime prediction.

Bibliography

[1] N. E. Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 30:89–99, 1974.

[2] D. Collett. *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 3 edition, 2015. ISBN 9781498731690. URL https://books.google.it/books?id=Okf7CAAAQBAJ.

[3] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246. URL http://www.jstor.org/stable/2985181.

[4] Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977. ISSN 01621459. URL http://www.jstor.org/stable/2286217.

[5] Terry Therneau. *Survival Package Functions*, 2014.

[6] T.M. Therneau. *A Package for Survival Analysis in R*, 2022. URL https://CRAN.R-project.org/package=survival. R package version 3.4-0.

[7] T.M. Therneau. *Package 'survival'*, 2022. URL https://CRAN.R-project.org/package=survival. R package version 3.4-0.

[8] T.M. Therneau and P.M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.