

**PONTIFICIA UNIVERSIDAD
CATÓLICA DEL PERÚ**

Escuela de Posgrado



M-Splines baseline hazard approximation for the proportional hazard model with right censored data

Tesis para obtener el grado académico de Magíster en Estadística
que presenta:

Omar Alejandro Juárez García

Asesor:

Víctor Giancarlo Sal y Rosas Celi

Lima, 2023

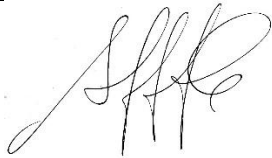
Informe de Similitud

Yo, Victor Giancarlo Sal y Rosas Celi, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis de investigación titulado M-splines baseline Hazard approximation for the proportional Hazard model with right censored data, del autor Omar Juárez García, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 21%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 26/06/2023.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio. Las similitudes se dan en palabras frecuentemente usadas en estadística, referencias y en todos los casos son menores a 1%.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

Lima 26 de junio del 2023

Apellidos y nombres del asesor: <u>Sal y Rosas Celi, Victor Giancarlo</u>	
DNI: 40361284	
ORCID: 0000-0001-8636-7142	

Abstract

The proportional hazard model plays a fundamental role in the analysis of time-to-event data. In this thesis, we conduct a simulation study to evaluate the performance of M-splines to estimate the baseline cumulative hazard function for the proportional hazard model. We assess the effect of sample size and number of knots in the estimation process. Finally, we apply this method to a sample of students from a university where the event of interest is the payment on time of the last tuition fee.

Key words: Cox model, M-splines, right censored data.

Contents

1	Introduction	1
2	Theoretical Background	3
2.1	Preliminary concepts	3
2.2	Regression Models	6
2.2.1	Accelerated failure time	6
2.2.2	Proportional hazards model	9
2.3	Introduction to Splines	10
2.3.1	Definition of a Spline	10
2.3.2	M-splines and I-splines	12
3	Estimation of the baseline hazard function via M-splines	14
3.1	M-splines approximation of the baseline hazard function	14
3.2	Likelihood Estimation	15
4	Simulation Analysis	18
4.1	Results	20
4.2	Model comparison	23
5	Analysis of Students Data	26
5.1	Data Description	26
5.2	Model Fitting	27
6	Conclusions	30
6.1	Suggestions for further analysis	31
A	R code	32
	Bibliography	38

Chapter 1

Introduction

Motivation. The area of survival analysis studies the time to occurrence of certain event. It appears in medicine to study the time to reinfection of diseases such as chlamydia (Brunham et al., 2005), in the banking sector when trying to model the time to reach default (Bellotti and Crook, 2009), in education to study the time until a drop-out from university (Ameri et al., 2016), in marketing when analysing the time until a consumer purchases a product (Ihwah, 2015), among other areas.

In order to implement a regression model to assess the effect of several variables on the time to the event of interest, we could go for either a parametric approach using known distributions via the accelerated time failure model (AFT) or a semi-parametric proportional hazard model (PH) developed by (Cox, 1972, 1975). The PH model assumes that the hazard function, given the covariates, is equal to a baseline hazard function (that does not depend on the covariate) times the exponential of a linear predictor. This is a popular choice because of its flexibility in not specifying the baseline hazard function. However, we can still use the PH model and maintain some flexibility if we approximate the baseline hazard function by a linear combination of spline functions.

Splines are widely used in different fields. For example, natural cubic splines have been used for medical trials in patients suffering from Alzheimer (Donohue et al., 2022), and to describe soil water retention (Kastanek and Nielsen, 2001). In survival analysis, natural cubic splines have been applied to approximate the stepwise cumulative hazard function in the context of the PH model (Bantis et al., 2020), and in the estimation of the hazard function using B-splines (Rosenberg, 1995).

Since the baseline hazard is a non-negative function we will approximate it via cubic M-splines, measuring its performance and complementing previous analyses made by Angelos et al. (1991), Rosenberg (1995), Bantis et al. (2020), Herndon and Harrell Jr (1995), Shih and Emura (2021), and Etezadi-Amoli and Ciampi (1987). More specifically, Etezadi-Amoli and Ciampi (1987) developed an extension for the hazard regression model, named Extended Hazard Regression (EHR), which has the AFT and PH models as special cases. They also approximate the baseline hazard function via quadratic splines for the Weibull, Log-Normal, and Generalised Gamma models.

Objectives. In this thesis we will carry out a deeper analysis of the work developed by Etezadi-Amoli and Ciampi (1987) by performing multiple simulations and evaluating the performance of cubic

M-splines to approximate the baseline hazard using different sample sizes and censoring levels of 30%, 40% and 50% approximately. Furthermore, we will assess the effect of the sample size and number of knots in the estimation process, apply the method against a real-world dataset, and compare its results with the Cox model implemented in R.

Document structure. This thesis is structured as follows. Chapter 2 presents the theoretical background. In this chapter, we cover topics such as survival function, hazard function, cumulative hazard function, censoring, distributions, regression models, and an introduction to splines: their definition, continuity constraints, and a specific type called M-splines. In chapter 3, we derive the proposed model, present the log-likelihood function, and briefly describe the estimation process. The simulation study and its performance are presented in chapter 4. In chapter 5, we apply the method against a dataset from an educational institution and compare these results with the ones provided by the Cox model. Finally, chapter 6 concludes by gathering all the findings and giving a brief discussion for future analyses.

Chapter 2

Theoretical Background

2.1 Preliminary concepts

Cumulative distribution function. Let T be a non-negative random variable specifying the failure time of a subject and t a value within its range. The cumulative distribution function is defined as:

$$F(t) = P(T \leq t), t > 0. \quad (2.1)$$

This function represents the probability that T will take values less or equal to t . Its derivative (when exists) is the probability density function defined as $F'(t) = f(t)$. This function follows the properties of i) $f_T(t) \geq 0$ for all t and

$$\int_{-\infty}^{\infty} f_T(t) = 1$$

Survival function. The survival function is defined as the complement of the cumulative distribution function. This is defined as $S(t) = 1 - F(t) = P(T > t)$, $t > 0$. This function is a non-increasing function and has the following property:

$$\begin{aligned} \lim_{t \rightarrow 0} S(t) &= 1 \\ \lim_{t \rightarrow \infty} S(t) &= 0 \end{aligned}$$

Hazard function. The hazard function is defined as:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}. \quad (2.2)$$

Since we know that the cumulative probability function is $F(t)$ and if its derivative can be computed, then

$$\lambda(t) = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}. \quad (2.3)$$

The hazard function can vary from zero (no risk) to infinity (certainty of failure at that instant), i.e. is a positive function. Over time, the hazard rate does not have a pre-defined shape: it can increase, decrease, remain constant, or take some more complex shapes.

Cumulative hazard function. The cumulative hazard function is defined as:

$$\Lambda(t) = \int_0^t \lambda(u) du. \quad (2.4)$$

The cumulative hazard determines the total amount of risk that has been accumulated up to time t . This can be easily understood under the count-data interpretation if only the failure event were repeatable (Cleves et al., 2010). For example, if we get a cumulative hazard of 30 over 2 months, this means we expect a failure 30 times.

All these equations presented before are linked as well. More specifically:

$$\Lambda(t) = -\log S(t)$$

$$S(t) = \exp(-\Lambda(t))$$

$$F(t) = 1 - S(t)$$

$$f(t) = -S'(t)$$

Details are given in Kalbfleisch and Prentice (2002) Cleves et al. (2010) or Reid and Cox (2018).

Censoring. When the event happens and the subject is not under observation, then we have censoring observations. We now present 3 types of censoring: i) right censoring, ii) interval censoring, and iii) left censoring.

Right censoring. Let $T \sim F(\cdot)$ and $Y \sim G(\cdot)$, where the first represents the failure time and the latter the censoring time, both following some distribution. A right-censored observation is when the failure time is observed before or at the censoring time. The censoring indicator is defined as:

$$\delta = I(T \leq Y),$$

where $I(\cdot)$ denotes the indicator function. This returns a 1 if the observation is censored and a 0 if not. In an observed sample, we can identify the observed time and censoring indicator as the pair:

$$(\tilde{T}, \delta) = (\min(T, Y), \delta).$$

This is the most common case of censoring. For example, this happens when the failure has not yet occurred but the study was already over. Another cause might be if the subject leaves or withdraws before the end of the study or if the subject is lost to follow-up meaning their disappearance from the study (e.g. relocation to a different city).

Interval censoring. Occurs when the investigator knows that the event happened but he can not track when the event took place. For example, in a clinical study, a patient has to visit the clinic monthly for 12 months. In the 8-month the patient tested negative but the following month he tests

positive. This means the event happened between the 8-month and 9-month but the clinician can not know for sure when exactly.

Left censoring. This type of censoring occurs when the event happens and the subject was not under observation. For example, a client that paid before the issue date of his credit card invoice.

Distributions. Common distributions used in survival analysis are part of the Generalised Gamma (GG) family. Depending on the values of the parameters one can compute the Gamma, Weibull, Log-Normal, and Exponential distribution. The Generalized Gamma is also part of a bigger family called the generalised-F distribution (GF) which includes the Log-Logistic (Cox, 2008).

We will present briefly the probability density function, the hazard function, and the survival function for three common distributions from the GG family as well as the Gompertz distribution.

Generalised Gamma distribution. T follows a Generalised Gamma distribution if its probability density function is

$$f(t) = \frac{pb(bt)^{pa-1}\exp(-(bt)^p)}{\Gamma(a)}, \quad p > 0, b > 0, a > 0, \quad (2.5)$$

where $\Gamma(a)$ is the gamma function.

Gamma distribution. If $p = 1$ in (2.5) then $T \sim \text{Gamma}(a, b)$. For this specific case, the density, survival, and hazard functions are given by:

$$f(t) = \frac{1}{\Gamma(a)} b^a t^{a-1} \exp(-bt). \quad (2.6)$$

$$S(t) = \frac{\Gamma(a) - \gamma(a, bt)}{\Gamma(a)}. \quad (2.7)$$

$$\lambda(t) = \frac{b^a t^{a-1} \exp(-bt)}{\Gamma(a) - \gamma(a, bt)}, \quad (2.8)$$

where

$$\Gamma(a) = \int_{bt}^{\infty} u^{a-1} e^{-u} du$$

$$\gamma(a, bt) = \int_0^{bt} u^{a-1} e^{-u} du$$

Weibull distribution. If $a = 1$ in (2.5), then $T \sim \text{Weibull}(p, b')$ with $b' = b^p$. The density, survival, and hazard functions are given by:

$$f(t) = pb't^{p-1}\exp(-b't^p). \quad (2.9)$$

$$S(t) = \exp(-b't^p). \quad (2.10)$$

$$\lambda(t) = pb't^{p-1}. \quad (2.11)$$

Exponential distribution. If $a = 1$ and $p = 1$ in (2.5) then $T \sim \text{Exponential}(b)$. The density, survival, and hazard functions are given by:

$$f(t) = b \exp(-bt). \quad (2.12)$$

$$S(t) = \exp(-bt). \quad (2.13)$$

$$\lambda(t) = b. \quad (2.14)$$

Gompertz distribution. The Gompertz distribution does not belong to the previous family, however, is often used in survival analysis. T follows the Gompertz distribution if its probability density function is

$$f(t) = be^{\eta t} \exp\left(-\frac{b}{\eta}(e^{\eta t} - 1)\right), \quad (2.15)$$

and

$$S(t) = \exp\left(-\frac{b}{\eta}(e^{\eta t} - 1)\right) \quad (2.16)$$

$$\lambda(t) = be^{\eta t} \quad (2.17)$$

2.2 Regression Models

2.2.1 Accelerated failure time

Let T be the time elapsed until a certain event and Z be a vector of covariates that can affect T . Then, the accelerated failure time model (AFT) is defined as:

$$\log(T) = Z\beta + \sigma W, \quad (2.18)$$

where the error term, W , follows certain distribution while σ is a scale parameter. This model assumes that the covariates can accelerate or decelerate the time-to-event.

Its hazard notation has the form of:

$$\lambda(t|Z) = e^{-\beta Z} \lambda_0(te^{-\beta Z}) = \theta \lambda_0(t\theta), \quad (2.19)$$

where $\theta = e^{-\beta Z}$. As seen, the covariates affect time by multiplying it by θ giving us the acceleration or deceleration of an event.

In order to get from (2.18) to (2.19) we can start by taking the exponential of (2.18)

$$T = e^{Z\beta} e^{\sigma W}. \quad (2.20)$$

Since the term $e^{Z\beta}$ is affected by covariates we could define $T_0 = e^{\sigma W}$ as a reference subject when the covariates are 0. The probability of survival for the reference subject after time t is:

$$S_0(t) = P(T_0 > t) = P(e^{\sigma W} > t) = P\left(W > \frac{\log(t)}{\sigma}\right). \quad (2.21)$$

Considering the effect of the covariates, the probability that a subject with covariates values Z to be alive after time t is:

$$S(t|Z) = P(T > t) = P(T_0 e^{Z\beta} > t) = P(T_0 > t e^{-\beta Z}) = S_0(t e^{-\beta Z}). \quad (2.22)$$

Recall that (2.21) represents the probability of a reference subject to survive after t , now (2.22) follows the same logic representing the survival function for the subject affected by covariates. As seen, this is no other than the reference subject survival function evaluated at time $t e^{-\beta Z}$ instead of t .

Since the probability density function is the negative of the derivative of the survival function, then:

$$f(t|Z) = S'_0(t e^{-\beta Z}) = e^{-\beta Z} \underbrace{-S'_0(t e^{-\beta Z})}_{f_0(t e^{-\beta Z})} = f_0(t e^{-\beta Z}) e^{-\beta Z}, \quad (2.23)$$

where $f_0(\cdot)$ is the probability density function for the reference subject.

Using the definition of the hazard function and (2.3) we have:

$$\lambda(t|Z) = \frac{f_0(t e^{-\beta Z})}{S_0(t e^{-\beta Z})} \cdot e^{-\beta Z}. \quad (2.24)$$

The ratio $f_0(\cdot)/S_0(\cdot)$ represents the baseline hazard $\lambda_0(\cdot) = f_0(\cdot)/S_0(\cdot)$ and reordering terms we get:

$$\lambda(t|Z) = e^{-\beta Z} \lambda_0(t e^{-\beta Z}) = \theta \lambda_0(t\theta), \quad (2.25)$$

where $\theta = e^{-\beta Z}$ is a common way of representing the covariates and their effects. And as we can see, the covariates affect time by multiplying it by θ giving us the acceleration or deceleration of an event.

Example. Suppose we have a covariate that follows a binomial distribution that takes the value of 1 when the covariate is active and 0 if not. Also, we have a regression coefficient $\beta = -0.5$. In this context, imagine we have 2 groups A and B, where the former receives a new treatment and the latter receives a standard treatment. If we say that the probability of survival in 12 months is 0.5 for group B then, using (2.21), the survival for group A is:

$$S_A = S_B(e^{-(-0.5)(1)} \times 12) = S_B(19.785).$$

As we see, to get the probability of survival for A we need to evaluate a higher time (19.785 vs 12 months) in the survival function of B, in other words, we have multiplied the time of B by $e^{-(-0.5)} = 1.649$ increasing it by 65%. The same analysis can be done for the hazard via (2.24) resulting in $\lambda_A = 1.649 \times \lambda_B(19.785)$.

Mean survival time. Since we have defined the probability density function in (2.23) we can compute the mean as

$$\begin{aligned} E(T) = \mu_T &= \int_0^\infty f_0(t e^{-\beta Z}) e^{-\beta Z} dt \\ &= e^{-\beta Z} \underbrace{\int_0^\infty f_0(t e^{-\beta Z}) dt}_{\mu_0} \\ &= e^{-\beta Z} \mu_0. \end{aligned} \quad (2.26)$$

This result implies that the expected time for the subject with covariates Z is $e^{-\beta Z}$ times greater than the mean for the reference subject. Furthermore, if β is small we can obtain the relative change between means from (2.26) as $\mu_T/\mu_0 - 1 = e^{-\beta Z} - 1$.

Example: Exponential Regression. Let us assume that W follows an extreme value distribution. More specifically:

$$f(w) = \exp(w - \exp(w)), \quad w > 0, \quad (2.27)$$

then it is easy to see that the hazard and survival functions are expressed by:

$$\lambda(t|Z) = \exp(-Z\beta). \quad (2.28)$$

$$S(t|Z) = \exp(-t \exp(-Z\beta)). \quad (2.29)$$

Example: Weibull Regression. If W follows an extreme value distribution with a probability density function of the form:

$$f(w) = \frac{1}{\sigma} \exp\left(\frac{w}{\sigma} - \exp\left(\frac{w}{\sigma}\right)\right), \quad (2.30)$$

then the hazard and survival functions can be expressed as:

$$\lambda(t|Z) = \sigma^{-1} \exp(Z\beta) t^{\sigma^{-1}-1}. \quad (2.31)$$

$$S(t|Z) = \exp\left(-t^{\sigma^{-1}} \exp\left[\frac{(-1/\sigma)(Z\beta)}{t^{\sigma^{-1}}}\right]\right). \quad (2.32)$$

Example: Log-Logistic Regression. If W follows an extreme value distribution with the following probability density function:

$$f(w) = \frac{e^w}{(1 + e^w)^2}, \quad (2.33)$$

then the hazard and survival functions can be expressed as:

$$\lambda(t|Z) = \frac{\sigma^{-1} \exp(-Z\beta) (t \exp(-Z\beta))^{\sigma^{-1}-1}}{[1 + (t \exp(-Z\beta))^{\sigma^{-1}}]}. \quad (2.34)$$

$$S(t|Z) = \frac{1}{1 + (t \exp(-Z\beta))^{\sigma^{-1}}}. \quad (2.35)$$

Estimation: Full Likelihood. Considering a sample of size N and

$$(t_i, \delta_i, Z_i), \quad i = 1, \dots, N, \quad (2.36)$$

where t refers to the observed time, δ is the censoring indicator (1 if time is observed and 0 if not), Z is a covariate vector and i is the index for each observation. The likelihood function is

$$\begin{aligned}
\mathcal{L}(\beta) &= \prod_{i:\delta_i=1}^N f(t_i|Z_i) \cdot \prod_{i:\delta_i=0}^N S(t_i|Z_i) \\
&= \prod_{i=1}^N f(t_i|Z_i)^{\delta_i} \cdot S(t_i|Z_i)^{1-\delta_i} \\
&= \prod_{i=1}^N \lambda(t_i|Z_i)^{\delta_i} \cdot S(t_i|Z_i)^{\delta_i} \cdot S(t_i|Z_i)^{1-\delta_i} \\
&= \prod_{i=1}^N \lambda(t_i|Z_i)^{\delta} \cdot S(t_i|Z_i)
\end{aligned}$$

By replacing $\lambda(t_i|Z_i)$ and $S(t_i|Z_i)$ for their respective forms based on the selected model (e.g. Exponential, Weibull, Log-Logistic) we can get the estimates for the regression coefficients.

2.2.2 Proportional hazards model

The proportional hazard model is defined as:

$$\lambda(t|Z) = \lambda_0(t) \exp(Z\beta), \quad (2.37)$$

where $\lambda_0(t)$ is left unspecified which corresponds to the hazard at time t when the vector of covariates is equal to zero, $Z = 0$, for all t . To understand why this is called a proportional hazard model, let us assume a case in which Z is an indicator covariate ($z = 1$ for the treatment group and $z = 0$ for the control group). Then the hazard ratio at time t between the two groups is:

$$\frac{\lambda(t|z_1 = 1)}{\lambda(t|z_0 = 0)} = \frac{\lambda_0(t)\exp(\beta)}{\lambda_0(t)} = \exp(\beta).$$

As we see the hazard in the two groups are proportional and β measures the effect of the treatment which remains the same even at different time points. Similarly, if we had a continuous covariate,

$$\frac{\lambda(t|Z = z_1)}{\lambda(t|Z = z_0)} = \frac{\lambda_0(t)\exp(\beta z_1)}{\lambda_0(t)\exp(\beta z_2)} = \exp((z_1 - z_2)\beta).$$

Estimation: Partial Likelihood. This estimation method was proposed by Cox (1972) and developed further in Cox (1975). The partial likelihood is expressed as:

$$\mathcal{L}^p(\beta) = \prod_{j=1}^m \frac{\exp(Z_j\beta)}{\sum_{\ell \in R(t_j)} \exp(Z_\ell\beta)}, \quad (2.38)$$

where t_i are the times where at least an event is observed, $R(t_j)$ represents all individuals at risk at time t_i , and m are the number of observed events.

The partial likelihood focuses on the estimation of the regression parameters without the need of specifying the baseline hazard. The baseline hazard is cancelled out in the derivation of the expression. For a deep understanding, we can redirect to Cox (1975), Kalbfleisch and Prentice (2002), and Wong (1986) for its asymptotic theory.

It is important to point out that the partial likelihood method works well in the majority of cases, nevertheless is important to state that, when faced with outliers, the partial likelihood could lead to

non-robust estimates (Ghosh and Basu 2019; Hutton and Monaghan 2002).

2.3 Introduction to Splines

Let us start by using a simple regression model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon$. In this, we see a polynomial of degree 2, which means that we have the following functions: i) $f_0(x_i) = 1$ (the intercept represented by a constant function), ii) $f_1(x_i) = x_i$ (linear function), and iii) $f_2(x_i) = x_i^2$ (quadratic function). These three functions are called basis functions. Now in order to fit the data we need to find the optimal weights that, in combination with the basis functions, allow us to model the response y_i . Figure 2.1 shows this fitting process using the three basis functions to fit the data represented by the black dots.

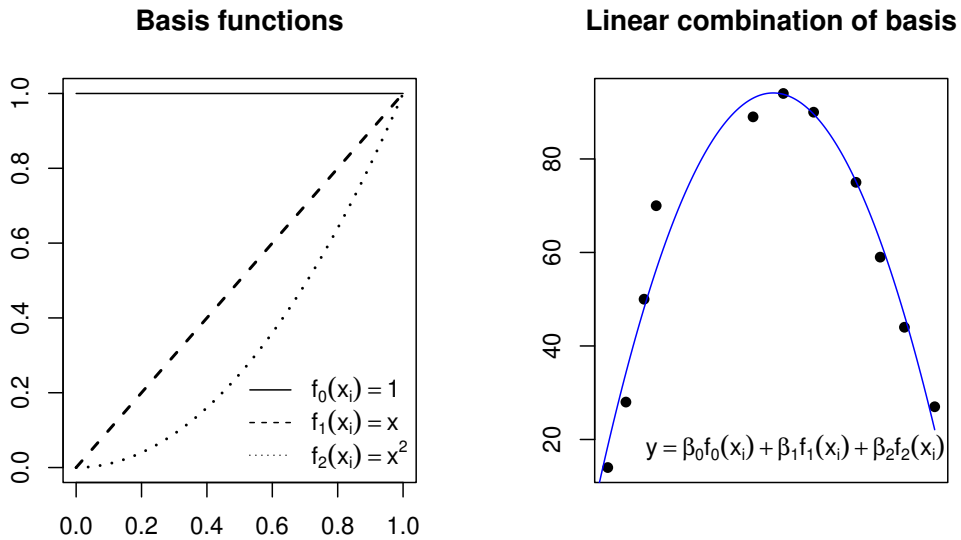


Figure 2.1: Basis functions and Linear combination

As seen, if we combine the optimal weights, that for this example were $\beta_0 = -18.25$, $\beta_1 = 6.74$, and $\beta_2 = -0.10$, with the basis functions we end up with the blue solid line from the right panel representing the linear combination.

2.3.1 Definition of a Spline

Following Agarwal (1989), if we have a sequence of real numbers, t_1, t_2, \dots, t_m , a spline function $S(t)$ of degree $k - 1$ with internal knots an interval $[a, b]$ having these 2 properties:

- In each interval $[t_i, t_{i+1}]$ for $i = 0, 1, \dots, m$, where $t_0 = a$ and $t_{m+1} = b$, $S(t)$ is given by some polynomial of degree $k - 1$ or less.
- $S(t)$ and its derivatives of order $1, 2, \dots, k - 2$ are continuous everywhere.

To get a better understanding of this we could use a function S that takes values from an interval $[a, b]$ and maps them to the set of real numbers \mathbb{R} and make it piecewise-defined. For this, let the interval $[a, b]$ be covered by m ordered disjoint intervals such as,

$$[t_i, t_{i+1}], \quad i = 0, \dots, k - 1$$

$$[a, b] = [t_0, t_1) \cup [t_1, t_2) \cup \dots \cup [t_{m-1}, t_m] \cup [t_m]$$

$$a = t_0 \leq t_1 \leq \dots \leq t_{k-1} \leq t_k = b$$

On each of these m pieces of $[a, b]$ we define a polynomial P_i .

$$P_i : [t_i, t_{i+1}] \rightarrow \mathbb{R}$$

On the i -th subinterval of $[a, b]$, S is defined by P_i as follows:

$$\begin{aligned} S(t) &= P_0(t) & , & & t_0 \leq t < t_1 \\ S(t) &= P_1(t) & , & & t_1 \leq t < t_2 \\ & \vdots \\ S(t) &= P_{m-1}(t) & , & & t_{m-1} \leq t < t_m \end{aligned}$$

For the second property, related to the smoothness, the two polynomial pieces P_{i-1} and P_i need to share a common derivative at t_i ,

$$\begin{aligned} P_{i-1}^0(t) &= P_i^0(t) \\ P_{i-1}^1(t) &= P_i^1(t) \\ & \vdots \\ P_{i-1}^{r_i}(t) &= P_i^{r_i}(t), \end{aligned}$$

where the superscript represents the order of the derivative.

Example. Suppose we have the interval $[a, b] = [0, 3]$ and the subintervals are $[0, 1]$, $[0, 2]$ and $[2, 3]$ and the polynomial pieces to be a quadratic polynomial. The pieces on $[0, 1]$ and $[1, 2]$ must join in value and first derivative at $t = 1$ while the pieces on $[1, 2]$ and $[2, 3]$ must join in value and first derivative at $t = 2$. Based on this, we could define a spline $S(t)$ as:

$$\begin{aligned} S(t) &= P_0(t) = -1 + 4t - t^2 & , & & 0 \leq t < 1 \\ S(t) &= P_1(t) = 2t & , & & 1 \leq t < 2 \\ S(t) &= P_2(t) = 2 - t + t^2 & , & & 2 \leq t \leq 3 \end{aligned}$$

Then, it is easy to see that i) follows the condition of being polynomials defined by pieces on subintervals of degree 2 or less and ii) $S(t)$ and its derivatives are continuous everywhere since at $t = 1 \rightarrow P_0(t) = P_1(t)$ as well as $P_0'(t) = P_1'(t)$ and at $t = 2 \rightarrow P_1(t) = P_2(t)$ as so their derivative $P_1'(t) = P_2'(t)$; so we can say $S(t)$ is a spline function of degree 2.

2.3.2 M-splines and I-splines

M-splines. Since our objective is to estimate a non-negative function (i.e. the hazard), it would be reasonable to think about a family of splines that share the same characteristic (being non-negative). The M-spline family is one type that satisfies this condition.

This spline, is positive inside the internal knots $[t_i, t_{i+k}]$ and zero elsewhere and has the normalisation $\int M_i(t)dt = 1$ (De Boor, 1978). Based on this, we can see that each M_i of the family has the properties of a probability density function in the interval $[t_i, t_{i+k}]$. As we can recall, a function $f_X(x)$ is a probability density function of some random variable X if and only if $f_X(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f_X(x) = 1$ which holds for M_i .

As a recap, we list the following properties for M_i

1. M_i is non-negative.
2. M_i is zero unless $t_i \leq t < t_{i+k}$.
3. M_i has $k - 2$ continuous derivatives at interior knots, k being the order of the spline.
4. M_i integrates to 1.

Additional to the previous properties, in order to assure non-negativity, the associate coefficients must be positive as well (i.e. $\gamma_i \geq 0$). Moreover, the total degrees of freedom (df), which represents the number of independent variables in a spline (number of parameters), is given by the degree and the number of knots, $df = k + m$. If we estimate an intercept then $df = k + m + 1$.

The computation can be defined by a recursion method since is more appropriate from a computational perspective (Ramsay, 1988). More specifically, for M_i of degree 1 ($k = 1$) we have

$$M_i(t) = \begin{cases} 1/(t_{i+1} - t_i) & \text{if } t_i \leq t < t_{i+1} \\ 0, & \text{otherwise.} \end{cases} \quad (2.39)$$

For a higher degree ($k > 1$) and $p_0 = (t - t_i)$, $p_1 = (t_{i+k} - t)$, $p_2 = (t_{i+k} - t_i)$,

$$M_i(t) = \begin{cases} k[p_0 M_i(t) + p_1 M_{i+1}(t)]/(k-1)p_2, & \text{if } t_i \leq t < t_{i+1} \\ 0, & \text{otherwise.} \end{cases} \quad (2.40)$$

I-splines. The cumulative hazard is the integral of the hazard. It is a non-decreasing function representing the accumulation of risks until a certain time. To obtain this we can employ the integral of M-splines, called I-splines, which are monotonically non-decreasing functions of the form

$$I_i(t) = \int_{t_{\min}}^t M_i(u)du, \quad (2.41)$$

where t_{\min} is the lower limit of the interval of the splines. The computation for this family at each interior boundary $t_j \leq t_{j+1}$ has the form

$$I_i(t) = \begin{cases} 0, & i > j \\ \sum_{m=i}^j (t_{m+k+1} - t_m) M_m^{k+1} / (k+1), & j - k + 1 \leq i \leq j \\ 1, & i < j - k + 1. \end{cases} \quad (2.42)$$

Finally, we present in Figure 2.2 a family of cubic M-splines ($k = 3$) with 3 knots ($m = 3$) and their integral with the resultant curve in a red dashed line.

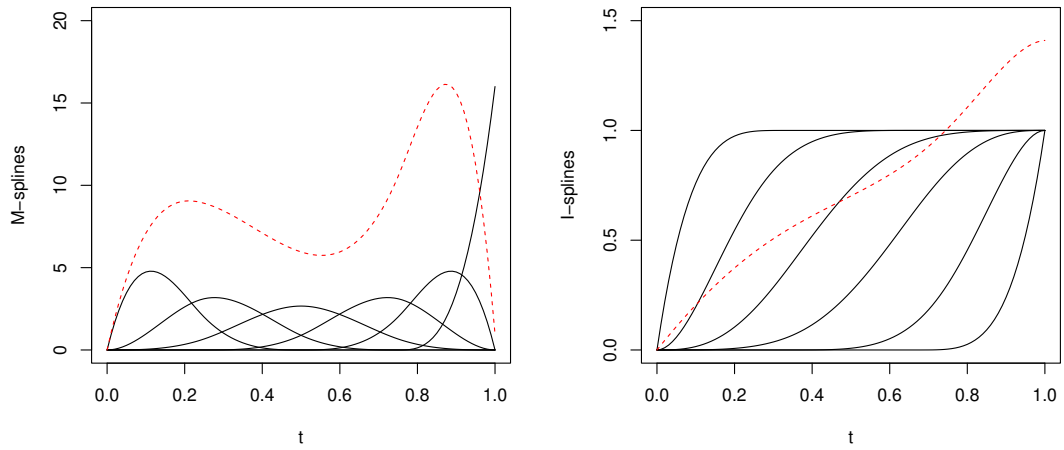


Figure 2.2: Cubic M_i and I_i splines with 3 knots ($m = 3$). The red curve was generated using equation $1.1M_1 + 1.9M_2 + 1.2M_3 + 1.2M_4 + 3M_5 + 0.06M_6$

Chapter 3

Estimation of the baseline hazard function via M-splines

3.1 M-splines approximation of the baseline hazard function

Angelos et al. (1991), Rosenberg (1995), Herndon and Harrell Jr (1995) and Etezadi-Amoli and Ciampi (1987) use splines to approximate the baseline hazard function. They use quadratic, cubic, and B-Splines, respectively. However, the use of M-splines is not covered. Knowing that M-splines share the same property of non-negativity with the baseline hazard function it would seem valid to propose this approximation. We will evaluate this approximation under the proportional hazard model using simulations and a real-world data application.

Etezadi-Amoli and Ciampi (1987) defined the approximation of the baseline hazard using quadratic splines of the form $\lambda_0(t) \approx sp_{2,m}(t)$, where $sp_{2,m}(t)$ is defined as:

$$sp_{2,m}(t) = \sum_{k=0}^2 \gamma_{0,k} t^k + \sum_{n=1}^m \gamma_{n2} (t - \tau_n)_+^2, \quad (3.1)$$

being $(a)_+ = a$ if $a > 0$ and 0 otherwise, m the number of knots, and τ_1, \dots, τ_m the knots.

For example, if we have one knot $t = \tau_1$ we can rewrite (3.1) as

$$sp_{2,1}(t) = \gamma_{00} + \gamma_{01}t + \gamma_{02}t^2 + \gamma_{10} + \gamma_{11}(t - \tau_1)_+ + \gamma_{12}(t - \tau_1)_+^2.$$

Similarly, we propose the following approximation via cubic M-splines. We select cubic splines since these are C^2 continuous, meaning that their second derivatives are the same at the joints where they meet and giving us a higher level of smoothness than linear or quadratic splines.

What we propose is to approximate the baseline hazard function by an M-spline of order three. More specifically: $\lambda_0(t) \approx M_{i,m}^3(t)$ of the form

$$M_{i,m}^3(t) = \sum_{k=0}^3 \gamma_k t^k + \sum_{i=1}^m \phi_i M_i^3, \quad (3.2)$$

where m represents the number of knots, k is the degree of the splines, γ indicates the coefficients for the polynomial part at the left side of the equation and ϕ_i are the coefficients for the M-splines associated with the number of knots.

Example. If we set $m = 0$, there are no internal knots, so the condition $t_i \leq t < t_{i+1}$ does not hold since we are not dividing our data points, then $M_i = 0$ and

$$M_{i,m}^3 = \sum_{k=0}^3 \gamma_0 t^k + \sum_{\substack{i=1 \\ j \neq 1}}^m \phi_i M_i^3 = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3. \quad (3.3)$$

The estimation of four parameters in this case comes from an ordinal polynomial approximation. If we use one internal knot ($m = 1$), then we would have to estimate five parameters,

$$M_{i,m=1}^3(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3 + \phi_1 M_1^3. \quad (3.4)$$

For I-splines with one knot, we have:

$$I_{i,m=1}^3(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3 + \phi_1 \int_{t_i}^t M_1^3(u) du. \quad (3.5)$$

3.2 Likelihood Estimation

The proposed extended hazard regression (EHR) from Etezadi-Amoli and Ciampi (1987) will serve us as a guide to later estimate the baseline hazard with M-splines. This general form for the hazard entails both models, the proportional hazard and the accelerated failure time. The EHR has the form

$$\lambda(t|Z) = \lambda_0[g(Z\alpha)t] g(Z\beta), \quad (3.6)$$

where $g(x) = \exp(x)$, $\lambda_0(t)$ is the baseline hazard and α and β are regression parameters associated to the vector of covariates Z .

Depending on the values of α and β , the previous expression could yield different results. For example, if we define $\beta = \alpha$ then (3.6) turns to the accelerated failure time model,

$$\lambda(t|Z) = \lambda_0[\exp(Z\beta)t] \exp(Z\beta). \quad (3.7)$$

On the other hand, if we set $\alpha = 0$ then the EHR yields a proportional hazard model similar to (2.41).

$$\lambda(t|Z) = \lambda_0(t) \exp(Z\beta). \quad (3.8)$$

Using (3.8) with a sample size of N observations with times (t_i) , the censoring indicator (δ_i) , and a vector of covariates z_i , we have: (t_i, δ_i, z_i) , $i = 1, \dots, N$

We can write the likelihood function for the N observations as

$$\mathcal{L} = \prod_{i=1}^N f(t_i, z_i)^{\delta_i} \cdot S(t_i, z_i)^{1-\delta_i} \quad (3.9)$$

$$\begin{aligned} &= \prod_{i=1}^N \lambda(t_i, z_i)^{\delta_i} \cdot S(t_i, z_i)^{\delta_i} \cdot S(t_i, z_i)^{1-\delta_i} \\ &= \prod_{i=1}^N \lambda(t_i, z_i)^{\delta_i} \cdot S(t_i, z_i). \end{aligned} \quad (3.10)$$

Since $S(t) = \exp(-\Lambda(t))$, then

$$\mathcal{L} = \prod_{i=1}^N \lambda(t_i, z_i)^{\delta_i} \cdot \exp[-\Lambda(t_i, z_i)]. \quad (3.11)$$

Taking the logarithmic value of the last expression, the log likelihood function is defined as:

$$l = \sum_{i=1}^N \left[\delta_i [\log \lambda(t_i, z_i)] - \Lambda(t_i, z_i) \right] \quad (3.12)$$

$$= \sum_{i=1}^N \left[\delta_i [\log \lambda(t_i, z_i)] - \int_0^t \lambda(t_i, z_i) dt \right] \quad (3.13)$$

Replacing (3.6) in (3.13)

$$\begin{aligned} l &= \sum_{i=1}^N \left[\delta_i [\log \{g(\beta \cdot z_i) \lambda_0[g(\alpha \cdot z_i)t]\}] - \int_0^t g(\beta \cdot z_i) \lambda_0[g(\alpha \cdot z_i)t] dt \right] \\ &= \sum_{i=1}^N \left[\delta_i [\log \{g(\beta \cdot z_i) \lambda_0[g(\alpha \cdot z_i)t]\}] - g(\beta \cdot z_i) \int_0^t \lambda_0[g(\alpha \cdot z_i)t] dt \right]. \end{aligned} \quad (3.14)$$

If we define $u = g(\alpha \cdot z)t$, then $du = g(\alpha \cdot z)dt$ and replacing this in (3.14),

$$\begin{aligned} l &= \sum_{i=1}^N \left[\delta_i [\log \{g(\beta \cdot z_i) \lambda_0(u_i)\}] - g(\beta \cdot z_i) \int_0^{(\alpha \cdot z_i)t} \frac{\lambda_0(u_i)}{g(\alpha \cdot z_i)} du \right] \\ &= \sum_{i=1}^N \left[\delta_i [\log \{g(\beta \cdot z_i) \lambda_0(u_i)\}] - \frac{g(\beta \cdot z_i)}{g(\alpha \cdot z_i)} \int_0^{g(\alpha \cdot z_i)t} \lambda_0(u_i) du \right]. \end{aligned} \quad (3.15)$$

Given that the integral of λ_0 is Λ_0 we can rewrite (3.15) as

$$\begin{aligned} l &= \sum_{i=1}^N \left[\delta_i [\log g(\beta \cdot z_i) + \log \lambda_0(u_i, \xi_i)] - \frac{g(\beta \cdot z_i)}{g(\alpha \cdot z_i)} \Lambda_0(u_i, \xi_i) \right] \\ &= \sum_{i=1}^N \delta_i [\log g(\beta \cdot z_i) + \log \lambda_0(u_i, \xi_i)] - \sum_{i=1}^N \frac{g(\beta \cdot z_i)}{g(\alpha \cdot z_i)} \Lambda_0(u_i, \xi_i), \end{aligned} \quad (3.16)$$

where $\xi_i = (\gamma_{0,k}, \phi_i)$ are the coefficients for the M and I-splines, similar to (3.4) and (3.5). Then our approximations are $\lambda_0(u_i, \xi_i) \approx M_m^3(t)$ and $\Lambda_0(u_i, \xi_i) \approx \int_{t_i}^t M_m^3(t) = I_m^3(t)$.

However, since our work focuses on the proportional hazard model, then $\alpha = 0$ and the general form given by (3.16) reduces to:

$$l = \sum_{i=1}^N \delta_i [\log g(\beta \cdot z_i) + \log \lambda_0(\xi_i)] - \sum_{i=1}^N g(\beta \cdot z_i) \Lambda_0(\xi_i). \quad (3.17)$$

To optimize this function we will use the *optimx* package from R. This package, as documented is “a wrapper function that calls other R tools for optimization including the existing *optim*”. Since the baseline hazard is a non-negative function we need to ensure that the coefficients of the M-splines are positive. For this, we use $\exp(\xi)$ instead of ξ . This allows us to solve an unconstrained optimization that can be solved by using a quasi-Newton method like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. This is an iterative method that also computes an approximation of the Hessian matrix which also allows us to compute the covariance matrix, hence finding the estimators of the standard errors. An example of this and its code can be found in Appendix A.

Chapter 4

Simulation Analysis

In this chapter, we focus on a simulation study to evaluate the performance of the proportional hazard model via M-splines. For this, we generate data sets of size 250, 500, and 1,000 and we set the number of simulations to 1,000 datasets replication.

Time-to-event of interest will be simulated considering three parametric models: Exponential, Weibull, and Gompertz. The formulas for each model are presented in Table 4.1. This is a similar setting as the one described in Austin (2012).

Table 4.1: From Austin (2012)

Characteristic	Exponential	Weibull	Gompertz
Scale(Rate)	$b > 0$	$b > 0$	$b > 0$
Shape		$a > 0$	$-\infty < \eta < \infty$
Baseline hazard	$\lambda_0(t) = b$	$\lambda_0(t) = bat^{b-1}$	$\lambda_0(t) = be^{\eta t}$
Cumulative baseline	$\Lambda_0(t) = bt$	$\Lambda_0(t) = bt^a$	$\Lambda_0(t) = \frac{b}{\eta}(e^{\eta t} - 1)$
Prob.Density Function	$b \exp(-bt)$	$bat^{a-1} \exp(-bt^a)$	$b \exp(\frac{b}{\eta}(1 - \exp(\eta t)))$
Time $u \sim U(0, 1)$	$T = -\frac{\log(u)}{b \exp(\beta'z)}$	$T = \left(-\frac{\log(u)}{b \exp(\beta'z)}\right)^{1/a}$	$T = \frac{1}{\eta} \log\left(1 - \frac{\eta \log(u)}{b \exp(\beta'z)}\right)$

For the censoring times, we assume a uniform distribution,

$$Y_i \sim U(0, \rho), i = 1, \dots, n$$

where ρ defines the censoring level. For this, we test 30%, 40%, and 50% of censoring. To define ρ , since we are using the same combination of covariates, we need to define $P(Y < T)$ which is the probability that the censored times beat the observed times in order to obtain a censored observation,

$$P(Y < T) = P(Y < T \cap T < \rho) + P(T \geq \rho) \quad (4.1)$$

The first part of the equation relates to events that happened before the observed time and the right side represents the events that happened after.

The two terms of (4.1) are defined as

$$P(Y < T \cap T < \rho) = \int_0^\rho \int_0^t \frac{1}{\rho} f(t) dt. \quad (4.2)$$

$$P(T \geq \rho) = \int_\rho^\infty f(t) dt. \quad (4.3)$$

Replacing (4.2) and (4.3) in (4.1) and solving for ρ we get the value needed for a particular level of censoring

$$\begin{aligned} P(Y < T) &= \int_0^\rho \int_0^t \frac{1}{\rho} f(t) dt + \int_\rho^\infty f(t) dt \\ &= \frac{1}{\rho} \int_0^\rho t f(t) dt + \int_\rho^\infty f(t) dt. \end{aligned} \quad (4.4)$$

Once we obtain the times T from Table 4.1 and the censoring times Y based on the value of ρ , we identify the observed and unobserved times based on the censoring indicator δ_i ,

$$(\tilde{T}_i, \delta_i) = (\min(Y_i, T_i), I(T_i \leq Y_i)),$$

where \tilde{T} is the observed time. Then, we set the regression coefficients to $\beta_1 = 0.5$, $\beta_2 = 0.7$, and the scale and shape parameters to $b = 2$, $a = \eta = 3$. We also use two covariates: $Z_1 \sim \text{Bernoulli}(0.5)$ and $Z_2 \sim N(0, 1)$.

To evaluate our approximation we compute the relative bias and the coverage for the regression parameters,

$$\text{Relative Bias} = \frac{1}{1000} \sum_{i=1}^{1000} \left(\frac{\hat{\beta}_i - \beta}{\beta} \right).$$

$$\text{Coverage} = \frac{1}{1000} \sum_{i=1}^{1000} I(\beta \in [LI_i, LS_i]).$$

where LI_i and LS_i are the lower and upper bound for the 95% confidence interval for each simulation.

4.1 Results

Polynomial Approximation: No knots. As a first exercise, we start with zero knots (i.e. $m = 0$), by doing this we are approximating the baseline hazard using a similar polynomial to (3.3). The results are presented in Table 4.2.

Table 4.2: Bias and coverage for regression parameters with zero knots ($m = 0$).

	<i>Exponential</i>				<i>Weibull</i>				<i>Gompertz</i>			
	<i>Bias</i>		<i>Coverage</i>		<i>Bias</i>		<i>Coverage</i>		<i>Bias</i>		<i>Coverage</i>	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
<i>Censoring ~ 30%</i>												
$N = 250$	0.13	0.14	0.90	0.74	0.03	0.03	0.94	0.93	0.02	0.02	0.94	0.95
$N = 500$	0.16	0.16	0.85	0.47	0.01	0.02	0.95	0.94	0.01	0.01	0.94	0.94
$N = 1000$	0.19	0.18	0.73	0.11	0.01	0.02	0.94	0.94	0.00	0.00	0.93	0.94
<i>Censoring ~ 40%</i>												
$N = 250$	0.25	0.26	0.85	0.46	0.05	0.04	0.94	0.93	0.00	0.02	0.95	0.95
$N = 500$	0.28	0.27	0.72	0.16	0.04	0.03	0.95	0.94	0.02	0.02	0.95	0.95
$N = 1000$	0.29	0.29	0.56	0.01	0.03	0.03	0.95	0.93	0.02	0.02	0.94	0.95
<i>Censoring ~ 50%</i>												
$N = 250$	0.35	0.36	0.81	0.27	0.03	0.05	0.95	0.92	0.01	0.01	0.95	0.96
$N = 500$	0.39	0.37	0.65	0.05	0.04	0.05	0.95	0.92	0.00	0.01	0.95	0.95
$N = 1000$	0.41	0.39	0.41	0.00	0.05	0.06	0.93	0.87	0.00	-0.01	0.96	0.95

We can notice that the exponential model presents difficulties in the recovery of the parameters. As the level of censoring increases, the bias also increases and the coverage decreases. For scenario I, with 30 percent of censoring, we get 0.13 as the lowest bias for β_1 and 0.14 for β_2 . In terms of coverage, the highest is 0.9 for β_1 and 0.74 for β_2 with a sample size of 250 while the lowest coverage is 0.73 for β_1 and only 0.11 for β_2 for the largest sample size. If we increase the level of censoring to 40%, the bias increases by nearly 10 percentage points for a sample size of 1,000 and the coverage decreases by nearly 20 percentage points for β_1 and 10 percentage points for β_2 returning a coverage of 1%. This pattern repeats when we look at scenario III (Censoring ~ 50%) where the bias for both parameters is no less than 35% and the coverage decreases to 0% for β_2 , results that prevent us to make any inference about the parameters.

However, the opposite happens for the Weibull and Gompertz models. Focusing on the Weibull model we can see that among all three scenarios the highest bias is 0.05 (5%) for β_1 when the level of censoring is 40% and 50%. The lowest coverage is for β_2 when the censoring grasps 50% returning 0.87 (87%). For the rest of the values, the bias is very low reaching 0.01 and the coverage reaches 0.93 or higher among the studied scenarios. The results for the Gompertz model are slightly better than the Weibull. As we see, the bias in every case is less than 5% or even less than 1% regardless of the censoring level being as high as 50% for both parameters. The coverage for every scenario, except for the value of 0.93, goes from 0.94 to 0.96.

Setting knots. Based on previous results, we noticed that, at least, for the Exponential model an

approximation using only polynomials (i.e. $m = 0$) is not enough to get the proper level of coverage and bias. Etezadi-Amoli and Ciampi (1987) study was conducted on three models that did not include the Exponential (e.g. Weibull, Generalised Gamma, and Log-normal). In their study, they conclude that $m = 0$ suffices to retrieve the regression parameters and the results were acceptable when compared to $m = 1$ or $m = 2$. However, they also mention that inference on the regression parameters could depend on the number of knots and one should observe this by testing if the addition of extra knots does not change the log-likelihood substantially. Given this statement and knowing the limitations of polynomial approximations (De Boor, 1978), we run another simulation to evaluate how the bias and the coverage may vary with additional knots. We test $m = 3$ and $m = 5$ for all 3 models from Table 4.2. Just to keep in mind that, in doing this, we estimate 7 parameters when $m = 3$ and 9 when $m = 5$, this is 4 γ 's parameters for the polynomial part of (3.2) plus the associated parameters for the knots ϕ_i . For the placement of the knots, they were placed evenly in the range of T which is the default method in the *splines2* package from R.

The results for the exponential model are shown in Table 4.3.

Table 4.3: Bias and coverage for regression parameters for the Exponential model with zero, three, and five knots.

	<i>Exponential (m = 0)</i>				<i>Exponential (m = 3)</i>				<i>Exponential (m = 5)</i>			
	<i>Bias</i>		<i>Coverage</i>		<i>Bias</i>		<i>Coverage</i>		<i>Bias</i>		<i>Coverage</i>	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
<i>Censoring ~ 30%</i>												
$N = 250$	0.13	0.14	0.90	0.74	0.02	0.02	0.96	0.94	0.04	0.03	0.95	0.96
$N = 500$	0.16	0.16	0.85	0.47	0.00	0.01	0.93	0.94	0.03	0.02	0.94	0.95
$N = 1000$	0.19	0.18	0.73	0.11	0.00	0.00	0.96	0.95	0.03	0.03	0.94	0.94
<i>Censoring ~ 40%</i>												
$N = 250$	0.25	0.26	0.85	0.46	0.00	0.02	0.95	0.94	0.02	0.02	0.95	0.95
$N = 500$	0.28	0.27	0.72	0.16	0.04	0.03	0.95	0.94	0.02	0.02	0.95	0.95
$N = 1000$	0.29	0.29	0.56	0.01	0.03	0.03	0.95	0.93	0.02	0.02	0.94	0.95
<i>Censoring ~ 50%</i>												
$N = 250$	0.35	0.36	0.81	0.27	0.03	0.05	0.95	0.92	0.01	0.01	0.95	0.96
$N = 500$	0.39	0.37	0.65	0.05	0.02	0.01	0.94	0.95	0.01	0.00	0.95	0.96
$N = 1000$	0.41	0.39	0.41	0.00	0.00	0.00	0.96	0.95	0.00	0.00	0.95	0.95

As we can see there is a substantial change from $m = 0$ to $m = 3$. The bias decreases to less than 4% in each scenario while the coverage is no lower than 0.93. If we compare one of the poorest results with $m = 0$ (with a censoring level of 50% with a sample size of 1,000) with $m = 3$, we get a steep decrease in bias to less than 1% in comparison of what we had with $m = 0$, while the coverage for β_1 increases from 0.41 to 0.96 and β_2 goes from basically no coverage at all to 0.95 (95%). These same results can be seen in each scenario for $m = 3$ in which the lowest coverage now is 0.93 instead of 0 when the approximation was made with no knots. These results also hold if we increase two more knots to $m = 5$. As we can also see, there is a marginal improvement in the coverage of the regression parameters and the bias remain low, but the change is minimal in comparison to what $m = 0$ vs

$m = 3$ showed.

We do the same for the Weibull model. In this case, the results obtained with $m = 0$ were satisfactory in most cases except for, perhaps, the coverage of β_2 for a sample size of 1,000 with $\sim 50\%$ of censoring where the coverage was less than 0.9 (90%) and a bias of 6%. Table 4.4 shows the results with $m = 3$ and $m = 5$ for the Weibull model. And similar to what we expressed already, we can see that with $m = 3$ there is no bias greater than 0.05, as a matter of fact, the bias for each scenario decreases which also leads to an increase in the coverage of the parameters. As we see, the lowest coverage is no longer 0.87 for a censoring level of 50%, instead increases to 0.96. Furthermore, if we increase two more knots ($m = 5$) there is a marginal improvement in terms of bias and coverage.

Table 4.4: Bias and coverage for regression parameters for the Weibull model with zero, three, and five knots $m = 0$, $m = 2$ and $m = 3$.

	<i>Weibull</i> ($m = 0$)				<i>Weibull</i> ($m = 3$)				<i>Weibull</i> ($m = 5$)			
	<i>Bias</i>		<i>Coverage</i>		<i>Bias</i>		<i>Coverage</i>		<i>Bias</i>		<i>Coverage</i>	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
<i>Censoring</i> $\sim 30\%$												
$N = 250$	0.03	0.03	0.94	0.93	0.02	0.02	0.96	0.94	0.03	0.02	0.95	0.95
$N = 500$	0.01	0.02	0.95	0.94	0.00	0.01	0.93	0.94	0.00	0.01	0.95	0.95
$N = 1000$	0.01	0.02	0.94	0.94	0.00	0.00	0.96	0.95	0.00	0.00	0.94	0.95
<i>Censoring</i> $\sim 40\%$												
$N = 250$	0.05	0.04	0.94	0.93	0.00	0.02	0.95	0.94	0.01	0.02	0.95	0.95
$N = 500$	0.04	0.03	0.95	0.94	0.02	0.01	0.95	0.95	0.01	0.00	0.94	0.95
$N = 1000$	0.03	0.03	0.95	0.93	0.00	0.00	0.96	0.95	0.00	0.01	0.94	0.95
<i>Censoring</i> $\sim 50\%$												
$N = 250$	0.03	0.05	0.95	0.92	0.00	0.03	0.95	0.94	0.03	0.01	0.95	0.94
$N = 500$	0.04	0.05	0.95	0.92	0.02	0.01	0.94	0.95	0.01	0.00	0.96	0.94
$N = 1000$	0.05	0.06	0.93	0.87	0.00	0.00	0.96	0.95	0.01	0.00	0.95	0.96

Finally, we can do the same for the Gompertz model and the results are presented in Table 4.5. This model yielded more than decent results with $m = 0$. Its bias was less than 0.05 and the coverage for the regression parameters was no lower than 0.93 among all scenarios. If we increase the number of knots we see a decrease in bias (however, the bias was really small from the beginning) and the coverage increases a bit even reaching a maximum value of 0.97 with $m = 5$. At least for this model, we can see more clearly that an increase in the level of complexity of the model (reflected in the number of parameters to estimate) does not vary much from the results obtained with $m = 0$.

Table 4.5: Bias and coverage for regression parameters for the Gompertz model with zero, three and five knots.

	<i>Gompertz (m = 0)</i>				<i>Gompertz (m = 3)</i>				<i>Gompertz (m = 5)</i>			
	<i>Bias</i>		<i>Coverage</i>		<i>Bias</i>		<i>Coverage</i>		<i>Bias</i>		<i>Coverage</i>	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
<i>Censoring ~ 30%</i>												
$N = 250$	0.02	0.02	0.94	0.95	0.01	0.01	0.94	0.95	0.01	0.02	0.93	0.94
$N = 500$	0.01	0.01	0.94	0.94	0.00	0.01	0.95	0.95	0.00	0.01	0.95	0.95
$N = 1000$	0.00	0.00	0.93	0.94	0.00	0.00	0.95	0.95	0.00	0.00	0.95	0.94
<i>Censoring ~ 40%</i>												
$N = 250$	0.00	0.02	0.95	0.95	0.01	0.02	0.95	0.96	0.02	0.01	0.95	0.96
$N = 500$	0.02	0.02	0.95	0.95	0.00	0.01	0.96	0.95	0.00	0.02	0.96	0.95
$N = 1000$	0.02	0.02	0.94	0.95	0.00	0.01	0.95	0.96	0.00	0.01	0.95	0.97
<i>Censoring ~ 50%</i>												
$N = 250$	0.01	0.01	0.95	0.96	0.03	0.03	0.94	0.93	0.01	0.02	0.95	0.95
$N = 500$	0.00	0.01	0.95	0.95	0.01	0.02	0.96	0.95	0.00	0.01	0.93	0.94
$N = 1000$	0.00	-0.01	0.96	0.95	0.03	0.02	0.95	0.93	0.00	0.01	0.95	0.94

4.2 Model comparison

In the previous section, we described that the number of knots could lead to an improvement in bias and coverage hence in the inference on the regression parameters. The results showed a substantial improvement for $m = 3$ for the Exponential model and a slight improvement for $m = 5$ not only in the former but in the Weibull and Gompertz as well. Being aware that model complexity increases when we set a higher number of knots, we use the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Likelihood Ratio Test (LRT) which are described below to compare different models:

$$\text{AIC} = 2k - 2\log(\hat{\mathcal{L}}). \quad (4.5)$$

$$\text{BIC} = k\log(N) - 2\log(\hat{\mathcal{L}}). \quad (4.6)$$

$$\lambda_{\text{LR}} = -2 \left[\ell(\theta_0) - \ell(\hat{\theta}) \right], \quad (4.7)$$

where \mathcal{L} refers to the Likelihood evaluated in the parameters and ℓ the log-likelihood. The likelihood ratio, under regularity conditions and with a large sample size, follows a chi-squared distribution with parameter r , which is the difference between the dimensions of θ_0 and $\hat{\theta}$ for the reduced and full model, respectively. We reject the null hypothesis if $\lambda_{\text{LR}} > \chi_{1-\alpha}^2$ where $\chi_{1-\alpha}^2$ is the $100(1-\alpha)$ percentile of the chi-squared distribution with r degrees of freedom. In our case, the null hypothesis is the approximation of the baseline hazard with no knots (i.e. reduced model) and how this compares with more complex models using three or five knots.

For this, we use one artificial data set for each sample size with an approximate censoring level of 30% and we estimate the parameters defining $m = 0$, $m = 3$, and $m = 5$ and compute their AIC and BIC. The results are presented in Table 4.6

Table 4.6: AIC and BIC for $m = 0$, $m = 3$ and $m = 5$

	$N = 250$			$N = 500$			$N = 1000$		
	AIC	BIC	ℓ	AIC	BIC	ℓ	AIC	BIC	ℓ
<i>Exponential</i> $\sim 30\%$									
$m = 0$	86.01	107.14	-37.01	28.16	53.45	-8.08	107.65	137.1	-47.82
$m = 3$	84.23	115.92	-33.11	27.05	64.99	-4.53	71.65	115.82	-26.83
$m = 5$	87.8	126.54	-32.9	16.93	63.29	2.53	52.43	106.42	-15.22
<i>Weibull</i> $\sim 30\%$									
$m = 0$	146.76	167.89	-67.38	360.16	385.45	-174.08	563.01	592.45	-275.5
$m = 3$	149.49	181.18	-65.75	364.5	402.43	-173.25	561.89	606.06	-271.95
$m = 5$	150.97	189.7	-64.48	367.91	414.27	-172.95	564.53	618.52	-271.27
<i>Gompertz</i> $\sim 30\%$									
$m = 0$	-110.5	-89.37	61.25	-230.37	-205.08	121.18	-475.93	-446.48	243.96
$m = 3$	-108.78	-77.08	63.39	-226.12	-188.19	122.06	-473.16	-428.99	245.58
$m = 5$	-105.02	-66.29	63.51	-223.78	-177.42	122.89	-469.18	-415.19	245.59

Focusing on the Exponential model with $N = 250$ the lowest AIC and BIC are obtained with $m = 3$ having a value of 84.23 and 115.92, respectively while the log-likelihood gets a higher value with $m = 3$ vs $m = 0$. This pattern also repeats itself for larger sample sizes. In the case of $N = 500$, going from $m = 0$ to $m = 3$ increases the likelihood in 44% (from -8.08 to -4.53), and the same happens with the larger sample size ($N = 1000$). Now, for $m = 5$, the change in likelihood is not as high as in going from 0 to 3, while the AIC and BIC tend to increase for a sample size of 250 but gets a lower value for $N = 500$ and $N = 1000$.

For the Weibull model, the lowest AIC and BIC are the ones obtained with $m = 0$ with the exception of $N = 1,000$. However, if we look at the log-likelihood values, the best fit occurs when we use $m = 5$ with an improvement of 2% approximately.

A similar behavior occurs for the Gompertz model. In this model, we see that $m = 0$ suffices in terms of AIC and BIC while the change in the log-likelihood is also small. For example, not being greater than 3% when we compare $m = 0$ vs $m = 3$ for $N = 250$.

Now we present the likelihood ratio test for each model presented in Table 4.7 where we reject the null hypothesis for $N = 500$ and $N = 1,000$ for the Exponential model (a result that was expected since we saw the Exponential model had a poor performance with $m = 0$). Also, we are available to reject the null hypothesis when comparing $m = 3$ vs $m = 5$ and $m = 0$ vs $m = 5$. The only contradiction is with $N = 250$, when we fail to reject the null hypothesis for all scenarios, however since we are using only one data set for this example, we are inclined to believe that in a repeated experiment we would end up rejecting the null hypothesis multiple times.

For the other 2 models, fail to reject the null hypothesis in every scenario so we could use a polynomial approximation with no knots for these. However, taking into consideration the results for the Exponential model with $m = 5$ and also the improvement for the Weibull model when going from $m = 3$ to $m = 5$, we conclude that 5 knots could be suitable since this number of knots suffices and passes the tests for all models.

Table 4.7: Likelihood Ratio Test for 3 data sets with a censoring level of $\sim 30\%$

	χ_r			LR		
	<i>Parameters</i>	<i>r</i>	χ_r^2	<i>Exponential</i>	<i>Weibull</i>	<i>Gompertz</i>
<i>N = 250</i>						
<i>m = 0 vs m = 3</i>	6 vs 9	3	7.81	7.79	3.27	4.28
<i>m = 3 vs m = 5</i>	9 vs 11	2	5.99	0.43	2.52	0.25
<i>m = 0 vs m = 5</i>	6 vs 11	5	11.07	8.21	5.79	4.53
<i>N = 500</i>						
<i>m = 0 vs m = 3</i>	6 vs 9	3	7.81	7.10	1.66	1.75
<i>m = 3 vs m = 5</i>	9 vs 11	2	5.99	14.12	0.59	1.66
<i>m = 0 vs m = 5</i>	6 vs 11	5	11.07	21.22	2.26	3.41
<i>N = 1000</i>						
<i>m = 0 vs m = 3</i>	6 vs 9	3	7.81	42.00	7.11	3.24
<i>m = 3 vs m = 5</i>	9 vs 11	2	5.99	23.22	1.36	0.02
<i>m = 0 vs m = 5</i>	6 vs 11	5	11.07	65.21	8.47	3.25

Finally, while performing the simulations the algorithm encountered a few numerical problems in finding the standard errors (SEs) even though convergence was reached. For example, with $m = 0$ the polynomial approximation is unable to compute the SE for 3 simulations for the exponential and 7 for the Weibull. When we increase the number of knots to $m = 3$ we notice that the Exponential model with 50% of censoring for a sample size of 250 has the most difficulty with 18 observations where the SEs were not computed followed by the Weibull model with 17 observations, however, both of them are less than 2% of the cases. With $m = 5$ the behaviour repeats itself, even reaching 43% for only one case of the exponential model with a sample size of 500 and censoring of 50%. Nevertheless, we found that by changing the initial parameters to the Cox regression estimates this issue either decreases totally or reduces significantly. For example, we found that for the Exponential model, we reduce the NAs from 43% to 22% for $m = 5$. So, in order, to apply this method we suggest feeding the algorithm with the Cox regression estimates as initial values to tackle this inconvenience and if the problem persists, we could suggest using Bootstrap to find the SEs given the asymptotic normality of the maximum likelihood estimator.

Chapter 5

Analysis of Students Data

5.1 Data Description

In this chapter, we analyse the payment behaviour of students from an educational Institution. We will apply the standard proportional hazards model implemented in R and compare it with the proposed methodology to assess the effect of different factors on the risk of payment.

Data description. The data contains a cohort of 50,146 students with a censoring level of 27%. For this study we use seven covariates: i) tuition fee, ii) cumulative credits, iii) absence to classes, iv) current grade point average (GPA), and payment behaviour from the first three out of a total of four academic fees: v) payment behaviour of the first fee, vi) payment behaviour of the second fee, and vii) payment behaviour of the penultimate fee. The explanations of these seven covariates and the response are as follow:

- Tuition fee: Students have a tuition fee that is calculated at the beginning of the enrolment stage. This can vary according to the number of credits, the faculty, and the campus.
- Cumulative credits: Researches like McGrath and Braunstein (1997), Anderson (1981), Pantages and Creedon (1978) show that students from the first o second year are more likely to drop out from college. Based on this, we are interested in how the level of seniority could have an impact on the risk of payment.

- Absence to classes until week 12: Absences from scheduled classes can affect grades (Jones, 1984) and also can have an impact on attrition (Whannell, 2013) and based on this we are interested to see if there is an impact on the risk of payment.
- Grade point average (GPA): GPA is an indicator of how a student is performing in university and we would like to know how this covariate can affect the risk of payment on time.
- Payment behaviour of the first 3 fees: Financial variables tend to have an autoregressive component. For example, if we want to forecast credit growth (Dinh, 2020) or study non-performing loans (Radivojevic et al., 2017) we would have to dive into their past behaviour. In relation to that, we want to test if the past behaviour can shed some light on the payment of the last fee.
- Days until payment (days): This is our response variable. Is a vector containing 50,146 times of which 27% of them are censored.

5.2 Model Fitting

After fitting the PH model via M-splines we find a negative relationship between the amount of the tuition fee (tfee) and the absence from classes until week 12 (abs), and the response variable, whereas we find a positive relationship between cumulative credits, GPA, and their payment behaviour of the previous fees. In Table 5.1 we can see that all covariates are statistically significant at 5% since their confidence intervals do not contain 0, yielding similar results to the Cox model.

Table 5.1: Factors associated with time to payment of fee 4 using a proportional hazard model.

	PH via M-splines				Cox PH			
	coef	HR	Lower	Upper	coef	HR	Lower	Upper
Tuition fee	-0.062 (0.012)	0.940	0.918	0.963	-0.060 (0.012)	0.942	0.919	0.964
Cumulative credits	0.078 (0.013)	1.081	1.054	1.109	0.079 (0.013)	1.082	1.055	1.110
Absences	-0.504 (0.048)	0.604	0.550	0.664	-0.505 (0.048)	0.604	0.549	0.663
GPA	0.039 (0.005)	1.040	1.030	1.050	0.041 (0.005)	1.041	1.032	1.051
Fee 1	0.420 (0.012)	1.522	1.488	1.558	0.425 (0.012)	1.530	1.495	1.566
Fee 2	0.475 (0.014)	1.607	1.564	1.652	0.466 (0.014)	1.594	1.550	1.638
Fee 3	0.849 (0.013)	2.338	2.278	2.401	0.842 (0.013)	2.320	2.260	2.382

Values inside parenthesis represent the standard errors of the regression coefficients.

In the following paragraphs, we will interpret the regression coefficients.

For tuition fees, holding the other covariates constant, if the tuition fee is greater than 3,000 then there is a decrease in the relative risk in comparison with students with tuition fees less than 3,000. The estimated relative risk is $\exp(-0.062) = 0.940$, with a confidence interval of $\exp(-0.062 \pm 1.96 \times 0.012) = (0.918, 0.963)$.

For cumulative credits, Holding the rest of covariates constant, if the cumulative credits increase in 1 unit then there is an increase in the relative risk of payment on time. The estimated relative risk is $\exp(0.078) = 1.081$ with a confidence interval of $\exp(0.078 \pm 1.96 \times 0.013) = (1.054, 1.109)$.

Focusing on absence to classes until week 12, a student with one additional percentage point in absence of classes reduces the relative risk of payment on time while keeping the rest of the variables constant. The estimated relative risk is $\exp(-0.504) = 0.604$ with a confidence interval of $\exp(-0.504 \pm 1.96 \times 0.048) = (0.550, 0.664)$.

For GPA, controlling for the rest of covariates, students with 1 more point in their GPA increases the relative risk of payment on time. For this the estimated relative risk is $\exp(0.039) = 1.040$ with a confidence interval of $\exp(0.039 \pm 1.96 \times 0.005) = (1.030, 1.050)$.

For the case of payment behaviour for the first 3 fees, we find that the penultimate fee has the strongest effect. Holding all other covariates constant, if a student has paid on time the third fee, then there is an increase in the relative risk of payment. Furthermore, the estimated relative risk is $\exp(0.849) = 2.338$ with a confidence interval of $\exp(0.849 \pm 1.96 \times 0.013) = (2.278, 2.401)$. The second strongest effect comes from the payment behaviour of the second fee with the estimated relative risk being $\exp(0.475) = 1.607$ with a confidence interval of $\exp(0.475 \pm 1.96 \times 0.014) = (1.564, 1.652)$ while for the first fee, we have an estimated relative risk of $\exp(0.420) = 1.522$ with a confidence interval of $\exp(0.420 \pm 1.96 \times 0.012) = (1.488, 1.558)$.

Furthermore, Figure 5.1 shows the baseline curves thanks to the estimated coefficients associated with M-splines.

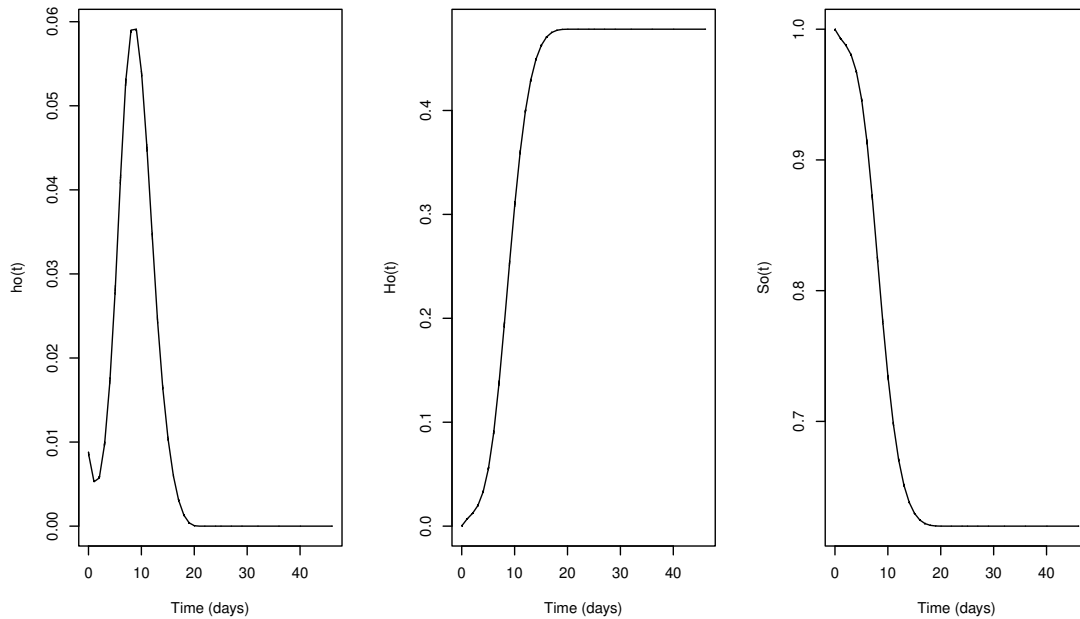


Figure 5.1: From left to right one has the baseline hazard, cumulative hazard, and the baseline survival function.

Focusing on the left side panel, we see that the hazard decreases at the start but then increases abruptly since more students start paying their fees approximately on day 10. After that, we can see a decrease in the risk of payment. Also, we need to consider that normally students have some more days after the due date to be considered on time (otherwise they could be susceptible to additional charges) which is why there are still payments but every single day that passes the risk of payment on time decreases. This is also reflected in the cumulative hazard from the centre panel. As we see, there is an increase until almost day 20, then apparently reaches a plateau. While in the right-side panel, we can see that the survival function decreases rapidly since the majority of payments occur in the first 20 days approximately.

Chapter 6

Conclusions

This thesis has developed a methodology to estimate the baseline hazard function via M-splines by maximising the log-likelihood. The contributions of this work are: i) to have a broader explanation of survival analysis and splines as well as the extended hazard regression model (EHR) proposed by Etezadi-Amoli and Ciampi (1987), ii) understand via simulation study why the number of knots is important in the inference of the parameters, even for models that are not complex like the Exponential model, iii) evidence the difficulties this method has as well as their strengths, since as stated in Shih and Emura (2021), Bryson and Johnson (1981) and Wichitsa-Nguan et al. (2016) the partial likelihood might not always have a maximum and a spline approximation could be useful in these cases, and iv) serve as a starting point for further studies.

In our simulation study, we tested three models: i) Exponential, ii) Weibull, and iii) Gompertz. The first had difficulties when we used only a polynomial approximation (i.e. no knots), while the others had a decent level of bias and coverage for the regression parameters with $m = 0$. Then, we decided to add more knots equally spaced in the range of the time variable. Using three and five knots we saw an improvement in terms of bias and coverage enabling us to make inferences about the regression parameters. The greater impact was seen in the Exponential model that went from basically no coverage at all to 95% when $m = 5$. We also saw marginal improvements in the other models as well (for example, a 0.93 coverage increases to 0.95 for the Weibull and Gompertz model). These results lead us to conclude that 5 knots might be suitable for this method.

Apart from this we also identified that the computational time to get the estimates is higher than the Cox proportional model, mainly because of the number of parameters the M-splines method needs to compute. Furthermore, we encountered negative values in the inverse of the Hessian matrix which caused problems with the computation of the standard errors. This was present in only a few of the simulated samples so we do not consider this a major issue. One approach to deal with this is using the estimates of the Cox regression as initial values for the algorithm and, if the problem persists, one could use Bootstrap methods to approximate the standard errors.

6.1 Suggestions for further analysis

As potential lines of research, we could: i) study a way to implement a penalised maximization if we would want to use more than 5 knots, ii) test this method with more complex models from the generalised F distribution, iii) evaluate its performance for the Accelerated Failure Time model, iv) evaluate how this method could deal with repeated times (ties), v) evaluate the trade-off between this method and the Cox regression from a computational perspective, and vi) test this implementation against the *splineCox.reg()* function from the *joint.Cox* package in R developed by Emura et al. (2017).

Appendix A

R code

Packages used: *survival*, *splines2* y *optimx*.

llph: Negative of the log-likelihood function

Description: Computes the negative of the log likelihood function

```
llph <- function(dat, time_col, cen_col, covar_cols, df, ms_degree, par){

  To <- dat[,time_col]
  d <- dat[,cen_col]
  Z <- as.matrix(dat[,covar_cols])
  n_reg_par <- ncol(Z)
  betas <- as.matrix(par[1:n_reg_par])
  xi <- as.matrix(par[(n_reg_par + 1):length(par)])

  # Baseline hazard
  ho <- mSpline(To, intercept = TRUE,
               df = df, degree = ms_degree) %%% exp(xi)
  ht <- exp(Z %%% betas) * ho # Hazard function

  # Cumulative baseline hazard
  Ho <- mSpline(To, intercept = TRUE, df = df,
               degree = ms_degree, integral = TRUE) %%% exp(xi)

  # Log-Likelihood
  val <- sum(d*log(ht)) - sum((exp(Z %%% betas))*Ho)
  return(-val)
}
```

```
wrap_llph <- function(par,...){
  # wrapper for llph that prints optimisation
  cat(par,"\n")
  llph(par,...)
}
```

Arguments:

1. dat: Data.
2. time_col: Column index for time.
3. cen_col: Column index for the censoring indicator.
4. covar_cols: Index for the columns containing the covariates.
5. df: Degree of freedom regarding the number of knots. For more information check the splines2 documentation.
6. ms.degree: M-spline degree
7. par: Initial parameters.

sim_df: Data simulation

Description: Simulates data for a proportional hazard model for the Exponential, Weibull, and Gompertz distribution with two covariates. For this, the function identifies the value of ρ that meets the desired level of censoring. The covariates are: $z_1 \sim \text{Binomial}(0.5)$ and $z_2 \sim N(0, 1)$. Finally, the scale and shape parameters were fixed to $b = 2$, $a = \eta = 3$ in the simulation process.

Note: For the Gompertz distribution, the function works with a transformation of the probability density function in order to avoid numerical difficulties. It uses $\exp(\log(f(t)))$ instead of $f(t)$.

```
dataSim <- function(dist,N,a,b,eta,betasTrue,targetCensProb){

  z1 <- rbinom(N,1,0.5)
  z2 <- rnorm(N)
  Z <- cbind(z1,z2)

  # Exponential distribution
  if(dist == "exp"){
    f1 <- function(t){t * b*exp(-b*t)}
    f2 <- function(t){b*exp(-b*t)}
    # Generating times (similar for weibull and gompertz)
    tvals <- (-log(runif(N))/(b*exp(Z %*% betasTrue)))
  }
}
```

```

# Weibull distribution
if(dist == "wei"){
  f1 <- function(t){t*b*a*(t**(a-1))*exp(-b*t**a)}
  f2 <- function(t){b*a*(t**(a-1))*exp(-b*t**a)}
  tvals <- tvals <- (-log(runif(N))/(b*exp(Z %>% betasTrue)))*(1/a)
}

# Gompertz distribution
if(dist == "gomp"){
  f1 <- function(t){t * exp(log(b) + eta*t + b/eta*(1-exp(eta*t)))}
  f2 <- function(t){exp(log(b) + eta*t + b/eta*(1-exp(eta*t)))}
  tvals <- (1/eta)*log(1-((eta*log(runif(N)))/(b*exp(Z %>% betasTrue))))
}

# Finding rho for desired level of censoring
f <- function(rho){
  int1 <- integrate(f1,lower = 0,upper = rho)[["value"]]
  int2 <- integrate(f2,lower = rho,upper = Inf)[["value"]]
  res <- targetCensProb - (1/rho)*int1 - int2
}

resTmp <- uniroot(f,lower = 0.001,upper = 5)
rho <- resTmp$root
Y <- runif(N,0,rho) # Censoring times
To <- pmin(tvals,Y) # Observed times
d <- as.numeric(tvals < Y) # delta indicator (censoring indicator)
simData <- as.data.frame(cbind(tvals,d,Z))
colnames(simData)[1] <- "t"
return(simData)
}

```

Arguments:

1. dist: Distribution. For Exponential = 'exp', for Weibull = 'wei', for Gompertz = 'gomp'.
2. N: Sample size.
3. a: Shape for the Weibull distribution.
4. b: Scale for the Exponential and Weibull distribution.
5. eta: Shape for the Gompertz distribution.
6. betasTrue: True values of the regression parameters to recover.
7. targetCensProb: Desired level of censoring (i.e. 30%, 40%, 50%).

Description: Returns the estimates of the regression parameters, the coefficients for the M-splines, the standard errors, the lower and upper bound, and other relevant information of the optimisation.

```
llphOpt <- function(par,dat,time_col,cen_col,
                   covar_cols,ms_degree,df,printOpt=1,...){
  fun <- ifelse(printOpt == 0,llph,wrap_llph)
  opt <- optimx(fun,
               par = par,
               method = "BFGS",
               dat = dat,
               time_col = time_col,
               cen_col = cen_col,
               covar_cols = covar_cols,
               ms_degree = ms_degree,
               df = df)

  # 1. M-splines and regression coefficients
  # 1.1 Estimates for regression parameters (betas)
  b_est <- coef(opt)[1:length(covar_cols)]
  # 1.2 Estimates for M-spline coefficients
  m_est <- coef(opt)[(length(covar_cols)+1):length(coef(opt))]
  # 1.2.1 Estimates for polynomial part
  gamma_est <- m_est[1:(ms_degree+1)]
  # 1.2.2 Estimates for coefficients related to knots
  phi_est <- if(df > 0){
    m_est[((ms_degree+1)+1):length(m_est)]
  } else{NULL}

  # 2. Hessian Matrix and Inverse Hessian to compute SE
  temp_hessian <- attributes(opt)$details["BFGS", ][[ "nhatend" ]]
  temp_inv_hessian <- tryCatch(solve(temp_hessian),
                              error = function(e) NULL)
  sd_b_est <- sqrt(diag(temp_inv_hessian))[1:length(b_est)]

  # 3. Confidence Interval (95%)
  sig_level <- 0.05
  conf_level <- 1-(sig_level)/2
  lb <- b_est - qnorm(conf_level)*sd_b_est
  ub <- b_est + qnorm(conf_level)*sd_b_est
  est_results <- c(opt$value,b_est,sd_b_est,lb,ub,m_est,
                  opt$convcode,opt$kkt1,opt$kkt2,opt$xtime)
```

```

vecNames <- if(df > 0){
  c("negLogLikeVal",
    paste("estimate",1:length(b_est),sep = ""),
    paste("SE",1:length(b_est),sep = ""),
    paste("Lower",1:length(b_est),sep = ""),
    paste("Upper",1:length(b_est),sep = ""),
    paste("gam", 0:(length(gamma_est)-1), sep = ""),
    paste("phi", 1:length(phi_est),sep = ""),
    "convcode","kkt1", "kkt2","xtimes")
}else{
  c("negLogLikeVal",
    paste("estimate",1:length(b_est),sep = ""),
    paste("SE",1:length(b_est),sep = ""),
    paste("Lower",1:length(b_est),sep = ""),
    paste("Upper",1:length(b_est),sep = ""),
    paste("gam", 0:(length(gamma_est)-1), sep = ""),
    "convcode","kkt1", "kkt2","xtimes")
  }
  names(est_results) <- vecNames
  return(est_results)
}

```

Arguments:

1. par: Initial values.
2. dat: Data.
3. time_col: Column index for time.
4. cen_col: Column index for the censoring indicator.
5. covar_cols: Column index for the censoring indicator.
6. ms_degree: M-spline degree.
7. df: Degrees of freedom. For more information check the splines2 documentation
8. printOpt: Indicator if optimisation should be printed. Default = 1 (prints optimisation) if 0 does not.

Example: Optimisation of an artificial dataset of 250 observations from a Weibull distribution with an approximate censoring level of 40% with coefficients values of 0.5 and 0.7. Optimisation using 7 degrees of freedom resulting in 3 knots.

```
library(survival)
library(splines2)
library(optimx)

set.seed(123456)
tmpData <- dataSim(dist = "wei",N = 250, a = 3, b = 2,
                   betasTrue = c(0.5,0.7),targetCensProb = 0.4)
opt <- llphOpt(par = c(0,0,rep(0.01,7)),dat = tmpData,
              time_col = 1,cen_col = 2,covar_cols = c(3,4),
              ms_degree = 3,df = 7,printOpt=1)
print(opt)
```

#negLogLikeVal	estimate1	estimate2	SE1	SE2
# 65.31887104	0.38404277	0.73870963	0.15934918	0.09053048
# Lower1	Lower2	Upper1	Upper2	gam0
# 0.07172412	0.56127315	0.69636142	0.91614611	-4.67555640
# gam1	gam2	gam3	phi1	phi2
# -3.16898221	-2.16743105	-0.28314826	0.20925607	-2.15265130
# phi3	convcode	kkt1	kkt2	xtimes
# -0.71285462	0.00000000	1.00000000	1.00000000	0.14000000

Bibliography

- Agarwal, G. (1989). Splines in statistics, *Bulletin of the Allahabad Mathematical Society* **4**: 1–55.
- Ameri, S., Fard, M. J., Chinnam, R. B. and Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts, *Proceedings of the 25th ACM international on conference on information and knowledge management*, pp. 903–912.
- Anderson, K. L. (1981). Post-high school experiences and college attrition, *Sociology of Education* pp. 1–15.
- Angelos, J., Lee, C. and Singh, K. (1991). B-spline approximation for the baseline hazard function, *Environmetrics* **2**(3): 323–339.
- Austin, P. (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates, *Statistics in Medicine* **31**(29): 3946–3958.
URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5452>
- Bantis, L. E., Tsimikas, J. V. and Georgiou, S. D. (2020). Survival estimation through the cumulative hazard with monotone natural cubic splines using convex optimization-the hcns approach, *Computer methods and programs in biomedicine* **190**: 105357.
- Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis, *Journal of the Operational Research Society* **60**(12): 1699–1707.
- Brunham, R. C., Pourbohloul, B., Mak, S., White, R. and Rekart, M. L. (2005). The unexpected impact of a chlamydia trachomatis infection control program on susceptibility to reinfection, *The Journal of infectious diseases* **192**(10): 1836–1844.
- Bryson, M. C. and Johnson, M. E. (1981). The incidence of monotone likelihood in the cox model, *Technometrics* **23**(4): 381–383.
- Cleves, M., Gould, W., Gutierrez, R. and Marchenko, Y. (2010). *An Introduction to Survival Analysis Using Stata*, 3rd edn, Stata Press, Texas.
- Cox, C. (2008). The generalized f distribution: an umbrella for parametric survival analysis, *Statistics in medicine* **27**(21): 4301–4312.
- Cox, D. (1972). Regression models and life-tables, *Royal Statistical Society* **34**(2): 187–220.
URL: <https://www.jstor.org/stable/2985181>

- Cox, D. (1975). Partial likelihood, *Biometrika* **62**(2): 269–276.
URL: <https://www.jstor.org/stable/2335362>
- De Boor, C. (1978). *A practical guide to splines*, Vol. 27, springer-verlag New York.
- Dinh, D. v. (2020). Forecasting domestic credit growth based on arima model: Evidence from vietnam and china, *Management Science Letters* **10**(5): 1001–1010.
- Donohue, M. C., Langford, O., Insel, P. S., van Dyck, C. H., Petersen, R. C., Craft, S., Sethuraman, G., Raman, R., Aisen, P. S. and Initiative, A. D. N. (2022). Natural cubic splines for the analysis of alzheimer’s clinical trials, *Pharmaceutical Statistics* .
- Emura, T., Nakatochi, M., Murotani, K. and Rondeau, V. (2017). A joint frailty-copula model between tumour progression and death for meta-analysis, *Statistical methods in medical research* **26**(6): 2649–2666.
- Etezadi-Amoli, J. and Ciampi, A. (1987). Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function, *Biometrics* **43**(1): 181–192.
URL: <https://www.jstor.org/stable/2531958>
- Ghosh, A. and Basu, A. (2019). Robust and efficient estimation in the parametric proportional hazards model under random censoring, *Statistics in Medicine* **38**(27): 5283–5299.
- Herndon, J. E. and Harrell Jr, F. E. (1995). The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables, *Statistics in medicine* **14**(19): 2119–2129.
- Hutton, J. L. and Monaghan, P. (2002). Choice of parametric accelerated life and proportional hazards models for survival data: asymptotic results, *Lifetime data analysis* **8**: 375–393.
- Ihwah, A. (2015). The use of cox regression model to analyze the factors that influence consumer purchase decision on a product, *Agriculture and Agricultural Science Procedia* **3**: 78–83.
- Jones, C. H. (1984). Interaction of absences and grades in a college course, *The Journal of Psychology* **116**(1): 133–136.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*, John Wiley & Sons.
- Kastanek, F. J. and Nielsen, D. R. (2001). Description of soil water characteristics using cubic spline interpolation, *Soil Science Society of America Journal* **65**(2): 279–283.
- McGrath, M. and Braunstein, A. (1997). The prediction of freshmen attrition: An examination of the importance of certain demographic, academic, financial and social factors., *College student journal* .
- Pantages, T. J. and Creedon, C. F. (1978). Studies of college attrition: 1950—1975, *Review of educational research* **48**(1): 49–101.
- Radivojevic, N., Jovovic, J. et al. (2017). Examining of determinants of non-performing loans, *Prague Economic Papers* **26**(3): 300–316.
- Ramsay, J. (1988). Monotone regression splines in action, *Statistical Science* **3**(4): 425–441.
URL: <https://www.jstor.org/stable/2245395>

- Reid, N. and Cox, D. (2018). *Analysis of survival data*, Chapman and Hall/CRC.
- Rosenberg, P. (1995). Hazard function estimation using b-splines, *Biometrics* **51**(3): 874–887.
- Shih, J.-H. and Emura, T. (2021). Penalized cox regression with a five-parameter spline model, *Communications in Statistics-Theory and Methods* **50**(16): 3749–3768.
- Whannell, R. (2013). Predictors of attrition and achievement in a tertiary bridging program., *Australian Journal of Adult Learning* **53**(2): 280–301.
- Wichitsa-Nguan, K., Läuter, H. and Liero, H. (2016). Estimability in cox models, *Statistical Papers* **57**: 1121–1140.
- Wong, W. H. (1986). Theory of partial likelihood, *The Annals of statistics* pp. 88–123.