Stockholms universitet

Generalized survival models applied to interval censored data

Albin Niva Printz

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:10
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2018:10**
http://www.math.su.se

Generalized survival models applied to interval censored data

Albin Niva Printz*

June 2018

Abstract

The generalized survival model (GSM) is a parametric spline based survival model, whose model scope includes many widely used parametric survival models. The GSM is implemented in the R package *rstpm2*. Performance checks of the package has previously been made on right censored data. In this thesis we aim to assess the performance GSM implementation on interval censored data. We apply the implementation to simulated proportional hazards data from the Weibull and mixture Weibull distributions, and to the Signal Tandmobiel data set. We then compare the GSM to a Weibull proportional hazards (PH) model, implemented in the *survreg* function of the *survival* package, which provides the core survival analysis routines in R.

Applied to Weibull data, we find that the special case of an proportional hazards GSM wih one spline term yields identical estimators as the standard PH model. In the case of a mixture of Weibull distributions, the GSM successfully captures the more complex distribution, outperforming the PH model. We find that in the case of very coarse censoring, the GSM fails to adequately capture the data. Finally we apply a lognormal accelerated failure time (AFT) model, and a probit GSM to the Signal Tandmobiel data set. The GSM does seem to capture more detailed features of the data than the AFT, suggesting that it is a better fit.

We conclude that the GSM implementation performs well on interval censored data, given the types of data tested and a reasonable resolution of the censoring.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: albin.printz@gmail.com. Supervisor: Mark Clements, Kristoffer Lindensjö and Felix Wahl.

Acknowledgements

I would like to express my gratitude to my supervisors Felix Wahl and Kristoffer Lindensjö from Stockholm University for their help and suggestions, and Mark Clements from Karolinska Institutet for his invaluable guidance and support throughout the project.

I would also like to sincerely thank Emmanuel Lesaffre for allowing me to use the Signal-Tandmobiel data set.

Finally, I would like to thank my friends and family for their unending support.

Contents

1 Introduction and background

1.1 Objectives and background

In the field of time-to-event data and survival analysis, one attempts to analyse the time to some event occurring. The theory is widely used in many fields, including medicine and engineering [11]. Two common model families are the parametric and non-parametric models. In this thesis, we will focus on the former.

In 2009, Younes and Lachin [18] introduced a parametric "link-based model", with a more flexible nature than previous models. Their model scope included many other popular parametric models, such as the proportional hazards model and the proportional odds model, making it a generalization of these. The model has been developed by Royston and colleagues under the name *flexible parametric models* to include general link functions for natural splines of log time. In 2018, Liu, Pawitan, and Clements [10] generalized these models to allow for a broader range of parametric and penalized smooth functions and used the name *generalized survival models* (GSM). Liu and colleagues also presented an implementation of GSMs in the R package *rstpm2*[1], and applied the model to right censored data.

We will use the R package *survival* [14] created by Therneau, as a comparison tool. The *survival* package contains many of the common survival analysis tools, and is considered a canonical package in survival analysis. In particular, it contains an accelerated failure time model in the *survreg* function.

In this thesis, our aims are:

1. To assess the statistical properties of the GSM implementation on simulated interval censored data. We will do this by considering specific cases when the GSM is equivalent to an accelerated failure time and proportional hazards model, as well as simulating from a mixed Weibull proportional hazards model.

2. We will demonstrate the flexibility of the GSM by applying it to the Signal-Tandmobiel® data set.

1.2 Data set

The Signal-Tandmobiel® study [16] is a large longitudinal study of oral hygiene for children conducted in Flanders, Belgium during 1996–2001. There were over 6000 seven year old children who participated in the study, making up approximately 7% of the target group. As a part of the study, an oral health education program was conducted. The goals of the study were to determine the oral health condition of Flemish school children in the ages 7–12, to educate the children in this age group about oral health, as well as to measure the effect of the health education program. The children were divided in three groups:

Group A: The 4468 children in this group were followed during the six years the study was performed. They were examined annually, and participated in the health education program.

Group B: A control group of 520 children not participating in the study. These children were examined in the first and last year of the study, allowing a

comparison with the children in group A. The children in this group did not participate in the health education program.

Group C: The children in this group were selected anew each year, to act as a cross-sectional control group. Each year, the group comprised approximately 500 children. The children in this group did not participate in the health education program.

Since the children were only visited on a yearly basis and not continuously, this study yielded what is called censored data, which will be introduced in Section 2.2.

2 Concepts for survival analysis

We will introduce some concepts that are common in the analysis of time-to-event data. This mainly consists of defining what we actually mean by censored data, and some useful functions. The definitions presented in this section are adapted from Cox and Oakes [2].

2.1 The survival, hazard and cumulative hazard functions

Throughout this paper, we are going to refer to the survival, hazard, and cumulative hazard functions. There are several different ways of representing these in terms of each other, some of which are presented in Section 2.1.3. All of these functions refer to a non-negative, continuous random variable T. We will regularly omit subscripting functions of random variables when the underlying random variable is obvious from context, i.e. write $f(t)$ instead of $f_T(t)$.

2.1.1 Survival function

We begin by defining the survival function.

Definition 1 (Survival function, p. 13 in [2])**.** *The survival function — denoted $S_T(t)$ — of a random variable T, is defined as*

$$S_T(t) = P_T(T > t) = \int_t^\infty f_T(x)dx. \tag{2.1}$$

That is, $S(t)$ is the probability that the event happens after time t — or, conversely, that the event does not happen by time t. The connection between the survival function and the cumulative distribution function $F(t)$ is simply that

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

As a consequence of this relation, the survival function is monotonically decreasing, and satisfies $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$. This follows directly from the corresponding properties of the cumulative distribution function. Note that improper distributions may also be possible, such as due to statistical cure, where $\lim_{t\to\infty} S(t) > 0$.

Example 2.1. The Weibull distribution is a common choice for parametric survival analysis models. We will use it as an example throughout this paper to illustrate certain concepts. A Weibull distributed random variable T with scale parameter λ and shape parameter k has density function

$$f_T(t) = \frac{kt^{k-1}}{\lambda^k} e^{-(t/\lambda)^k}, \quad t, \lambda, k \geq 0.$$

Using Equation 2.1, we can calculate the survival function of T:

$$S_T(t) = \int_t^\infty \frac{kx^{k-1}}{\lambda^k} e^{-(x/\lambda)^k} \, dx = e^{-(t/\lambda)^k}.$$

2.1.2 Hazard function

Another function that appears in time-to-event analysis is the hazard function. The interpretation of the hazard function is less intuitive than the survival function.

Definition 2 (Hazard function, p. 14 in [2])**.** *The hazard function of a random variable T is defined as the instantaneous rate of event occurrence:*

$$h_T(t) = \lim_{\Delta t \to 0} \frac{P_T(T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.2}$$

It is the conditional limiting probability of the event occurring in the interval $[t, t + \Delta t)$, divided by the length of the interval, as the length of the interval goes to zero. The hazard function is closely related to the density function $f(t)$, as will be presented shortly.

Sometimes one also has use of the cumulative hazard function $H(t)$, defined by

$$H(t) = \int_0^t h(t')dt', \tag{2.3}$$

which, as we shall see, has a close connection to the survival function.

2.1.3 Connecting $S(t)$ and $h(t)$

Expanding the conditional probability in the numerator of Definition 2 using the definition of conditional probability, the following expression for the hazard function is obtained:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)}. \tag{2.4}$$

which, if one notices that the derivative of $S(t)$ with regards to t is $-f(t)$, then the hazard function can be re-expressed as:

$$h(t) = -\frac{d}{dt} \log S(t).$$

Using this formula to rewrite the survival function in terms of the hazard function, we obtain

$$S(t) = \exp\left\{-\int_0^t h(t')dt'\right\} = \exp\left\{-H(t)\right\}. \tag{2.5}$$

Example 2.2. Continuing with the Weibull distribution, we can use Equation (2.4) to calculate the hazard of $T \sim \text{Weibull}(k, \lambda)$:

$$h(t) = \frac{f(t)}{S(t)} = \frac{kt^{k-1}}{\lambda^k}. \tag{2.6}$$

The cumulative hazard function can either be calculated by integrating the hazard, like in Equation (2.3), or using Equation (2.5). Either way, we will arrive at the cumulative hazard function:

$$H(t) = -\log(S(t)) = \int_0^t h(t')dt' = \left(\frac{t}{\lambda}\right)^k.$$

2.2 Censored data

In survival analysis one attempts to model the time to some event from a longitudinal study, which gives rise to a special kind of imprecision in the data. Consider a hypothetical study on new born children investigating the time until the child learns to walk. Usually parents notice when their toddler starts wandering away, but for the sake of this example let us assume the children in the study have especially inattentive parents. A natural way to collect these data would be to select some infants, observing their walking skills once a month until they can successfully take a few steps. However, the time when a child actually learned to walk could be at any point in time in between the last test before the first successful test, and the first successful test. This is the idea of *interval censored data*.

Let us put this in more mathematical terms. The participants in the study will throughout the paper be referred to as *units*. When talking about a *failure time*, we mean the occurrence of the event of interest. The times the units are observed will be referred to as *examinations*. In the previous example, the time the ith child learned to walk would be referred to as the failure time of the ith unit. Denote the failure time of the ith unit by T_i, the last examination before T_i by L_i, and the first examination after T_i by R_i. The concept is visualized in Figure 1.

Other types of censoring are left- and right-censored data. A right censored observation is only known to have happened after some point in time, and a left censored observation is only known to have happened before some point in time. Right and left censored data occur as special cases of interval censored data. With right censored data we get the interval $[L_i, \infty)$, and with left censored data we get the interval $(0, R_i]$.

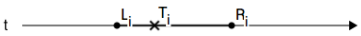


Figure 1: A visualization of censored data, displaying the end points of the censored interval L_i and R_i, as well as the actual failure time T_i.

3 Likelihood theory

3.1 Non-informative censoring

In the case of exact data, the contribution of an observation t to the likelihood $L(\boldsymbol{\theta})$ for parameters $\boldsymbol{\theta}$ is simply $f_T(t|\boldsymbol{\theta})$. It would seem reasonable to assume that the contribution of observing an observation censored on $[L, R]$ is simply $P(T \in [L, R])$, and under certain circumstances this is indeed the case. It turns out [12] that this holds under the assumption of *non-informative censoring*. In this paper we will not go into very much mathematical detail of this assumption, but merely try to give the reader some intuition.

We will use the informal definition of non-informativity presented on page 42 in [7]. The assumption states that the distribution of failure times yields no information about the distribution of censoring intervals, and vice versa.

Now, one might wonder if there are any cases of *informative censoring* in practice. The answer is yes, and this will hopefully be clarified in the following two examples.

Let us revisit the fictional infant study previously discussed in Section 2.2. Consider a situation where the event of a child learning to walk made the parents of said child lose interest in the study, and drop out before the study was over. Then the distribution of the failure times heavily influences the distribution of the censoring intervals, since no examinations would be visited after the event occurred. This is a case of informative censoring, since the failure times influence the distribution of the censoring intervals.

Another possible scenario would be if the examination process of the children had a significant impact on their walking performance. Then the children attending more examinations would learn to walk faster than the other children. Or in other words, the distribution of censoring intervals yields information about the failure times, and thus the censoring is informative.

A mathematical definition of the non-informativity assumption can be found in [12].

There are two other assumptions usually made about the censoring, *random censoring* and *independent censoring* [7]. These treat the randomness of the censoring process, and basically assume that the hazards of censored units are the same as those of the non censored units over the whole data set, as well as conditioned on covariates. These assumptions will not be discussed in more detail, since these assumptions are automatically fulfilled if the failure times and censoring times are simulated independently.

3.2 Likelihood theory for censored data

When we have exact observations, the contribution of an observation to the likelihood function is $f_T(t|\boldsymbol{\theta})$. Under the assumption of non-informative censoring, the contribution of an observation censored on the interval $[L, R]$ to the likelihood is $P_{T|\boldsymbol{\theta}}(L \leq T < R)$. Expressing this probability in terms of the survival function, we get that

$$P_{T|\boldsymbol{\theta}}(L \leq T < R) = S(L|\boldsymbol{\theta}) - S(R|\boldsymbol{\theta}). \tag{3.1}$$

In the special cases right and left censoring of interval censoring, (3.1) simplifies to $S(L|\boldsymbol{\theta})$ and $1 - S(R|\boldsymbol{\theta}) = F(R|\boldsymbol{\theta})$ respectively, recalling that $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$. The censoring type of an observation and the likelihood contribution is summarized in Table 1.

Censoring type	Likelihood contribution
Exact, $L = R = t$	$f(t\|\boldsymbol{\theta})$
Interval censored, $[L, R]$	$S(L\|\boldsymbol{\theta}) - S(R\|\boldsymbol{\theta})$
Left censored, $[0, R]$	$1 - S(R\|\boldsymbol{\theta}) = F(R\|\boldsymbol{\theta})$
Right censored, $[L, \infty)$	$S(L\|\boldsymbol{\theta})$

Table 1: The different types of censored observations, and their likelihood contributions.

Let $\boldsymbol{\theta}$ be a parameter vector, E be the set of exact observations, L the set of left censored observations, R the set of right censored observations, and I be the set of censored observations. The likelihood and log likelihood function becomes

$$L(\boldsymbol{\theta}|E, L, R, I) = \prod_{t\in E} f(t|\boldsymbol{\theta}) \prod_{r\in L} (1 - S(r|\boldsymbol{\theta})) \prod_{l\in R} S(l|\boldsymbol{\theta}) \prod_{(l,r)\in I} (S(l|\boldsymbol{\theta}) - S(r|\boldsymbol{\theta})), \tag{3.2}$$

$$\begin{aligned} l(\boldsymbol{\theta}|E, L, R, I) = &\sum_{t\in E} \log f(t|\boldsymbol{\theta}) + \sum_{r\in L} \log\{1 - S(r|\boldsymbol{\theta})\} \\ &+ \sum_{l\in R} \log\{S(l|\boldsymbol{\theta})\} \\ &+ \sum_{(l,r)\in I} \log\{S(L|\boldsymbol{\theta}) - S(R|\boldsymbol{\theta})\}. \end{aligned} \tag{3.3}$$

The parameter vector $\hat{\boldsymbol{\theta}}$ which maximizes the likelihood function (or equivalently the log likelihood function) is called the *maximum likelihood (ML) estimator* of $\boldsymbol{\theta}$ [4]. The ML-estimator can be calculated with the use of the derivative and second derivative of the (log) likelihood function. As per usual, the *score vector* is obtained by taking the gradient of the log likelihood

$$\nabla l(\boldsymbol{\theta}|E, L, R, I) = \left(\frac{\partial l(\boldsymbol{\theta}|E, L, R, I)}{\partial \theta_i}\right)_i.$$

The *Fisher information matrix* is obtained by taking the negative hessian of the log likelihood

$$I(\boldsymbol{\theta}|E, L, R, I) = -\left(\frac{\partial^2 l(\boldsymbol{\theta}|E, L, R, I)}{\partial \theta_i \partial \theta_j}\right)_{i,j}.$$

Example 3.1. For this example, we will consider exponentially distributed failure times $T_i \sim \text{Exp}(\lambda)$ with mean λ^{-1}. Given n_E exact observations, and n_I interval censored observations we get explicit expressions of the likelihood and log likelihood using Equation (3.2) and (3.3).

$$L(\lambda|E,I) = \lambda^{n_E} \prod_{t\in \boldsymbol{E}} e^{-\lambda t} \prod_{(L,R)\in \boldsymbol{I}} \left(e^{-\lambda L} - e^{-\lambda R}\right),$$
$$l(\lambda|E,I) = n_E \log\lambda - n_E\lambda\bar{t} + \sum_{(L,R)\in I} \log\left(e^{-\lambda L} - e^{-\lambda R}\right),$$

where $\bar{t} = \frac{1}{n_E}\sum_{t\in E} t$. As we shall now see, the last sum does not simplify. The score function is calculated by differentiating the log-likelihood with respect to λ,

$$\nabla l(\lambda|E,I) = \frac{n_E}{\lambda} - n_E\bar{t} - \sum_{(L,R)\in I} \frac{Le^{-\lambda L} - Re^{-\lambda R}}{e^{-\lambda L} - e^{-\lambda R}}.$$

Without the censored observations the solution to the score equation can be expressed analytically, but the censoring clearly complicates the expression. The Fisher information becomes

$$\begin{aligned} I(\lambda|E,I) &= -\frac{d}{dt}\left(\nabla l(\lambda|E,I)\right) \\ &= \frac{2n_E}{\lambda} - \sum_{(L,R)\in I} \frac{\left(L^2e^{-\lambda L} - R^2e^{-\lambda R}\right)\left(e^{-\lambda L} - e^{-\lambda R}\right)}{\left(e^{-\lambda L} - e^{-\lambda R}\right)^2} \\ &\quad - \sum_{(L,R)\in I} \frac{\left(Le^{-\lambda L} - Re^{-\lambda R}\right)^2}{\left(e^{-\lambda L} - e^{-\lambda R}\right)^2}. \end{aligned}$$

The expressions for the score and information are not presented here in expectation that the reader will have any use of these expressions, but merely to illustrate the fact that they will require a numerical solution. This is also the reason for going with the exponential function for this example, the expression for the Fisher information matrix in the case of Weibull failure times is more complex.

In practice the score and Fisher information can also be calculated using finite differences in the log likelihood. For example, in one variable we have

$$\nabla l(\theta) \approx \frac{l(\theta+h) + l(\theta-h)}{2h},$$
$$I(\theta) \approx \frac{l(\theta+h) - 2l(\theta) + l(\theta-h)}{h^2},$$

for sufficiently small h.

4 Survival analysis models

In this section we present three parametric survival analysis models. In Section 11.1.2 of the appendix we describe the non-parametric Kaplan-Meier estimator of the survival function.

4.1 Proportional hazards model

Before we can look at a more general model, we will introduce the proportional hazards (PH) model. In this model, one assumes that the hazard function $h_x(t)$ of t under the effect of a covariate x is proportional to the baseline hazard $h_0(t)$. Let us clearly define what we mean by this.

Definition 3 (Proportional hazards model, p.70 in [2])**.** *Let $\boldsymbol{x}$ be a vector of covariates, $h_T(t;\boldsymbol{x})$ be the hazard function of a random variable T given $\boldsymbol{x}$, and $h_T(t;\mathbf{0})$ be the baseline hazard of t. By employing a* proportional hazards model*, we assume the effect of covariates satisfies the relation*

$$h_T(t;\boldsymbol{x}) = h_T(t;\mathbf{0})\psi(\boldsymbol{x}), \quad \psi(\mathbf{0}) = 1.$$

The function ψ is a link function between the covariates and their effect. In this thesis, we will always put $\psi(\boldsymbol{x}) = e^{\boldsymbol{\beta}^T\boldsymbol{x}}$ for some parameter vector $\boldsymbol{\beta}$ and covariate vector $\boldsymbol{x}$. We see that for any covariate vector $\boldsymbol{x}$, $h(t;\boldsymbol{x}) \propto h(x;\mathbf{0})$, which is the source of the name proportional hazards model.

This model yields a simple interpretation of covariate effects. Given a binary covariate and a parameter $\beta = \log(2)$, under the assumption of the PH we can state that the hazard of those under covariate effect is always twice that of those not under covariate effect.

Example 4.1. Recall that the hazard function of a Weibull-distribution with scale parameter λ and shape parameter k is

$$h_0(t) = \frac{kt^{k-1}}{\lambda^k}.$$

Under the PH model we find, using Definition 3, the hazard of t under covariate effect from $\boldsymbol{x}$ to be

$$h_x(t) = \frac{kt^{k-1}}{\lambda^k}e^{\boldsymbol{\beta}^T\boldsymbol{x}}. \tag{4.1}$$

In Figure 2 the two hazard functions $h_0(t)$ and $h_1(t)$ are illustrated for the special case when $\boldsymbol{x}$ is a binary covariate, $\beta = 1$, $\lambda = 1$, and $k = 1.4$.

The expressions for the cumulative hazard, log cumulative hazard, and survival function can be derived from Equation (4.1) and the re-

lations in Section 2.1.3.

$$
\begin{aligned}
H(t|\boldsymbol{x},\boldsymbol{\beta}) &= \int_0^t h(t'|\boldsymbol{x},\boldsymbol{\beta})dt' = \frac{t^k}{\lambda^k} e^{\boldsymbol{\beta^T x}} \\
\log(H(t|\boldsymbol{x},\boldsymbol{\theta})) &= k\log(t) - k\log(\lambda) + \boldsymbol{\beta^T x} \qquad (4.2) \\
S(t|\boldsymbol{x},\boldsymbol{\theta}) &= \exp\{-H(t|\boldsymbol{x},\boldsymbol{\beta})\} = \exp\left\{-\frac{t^k}{\lambda^k}\exp\{\boldsymbol{\beta}^T\boldsymbol{x}\}\right\}.
\end{aligned}
$$

Taking one extra look at Equation (4.2), we notice that the effect of covariates is linear on the log cumulative hazard scale.

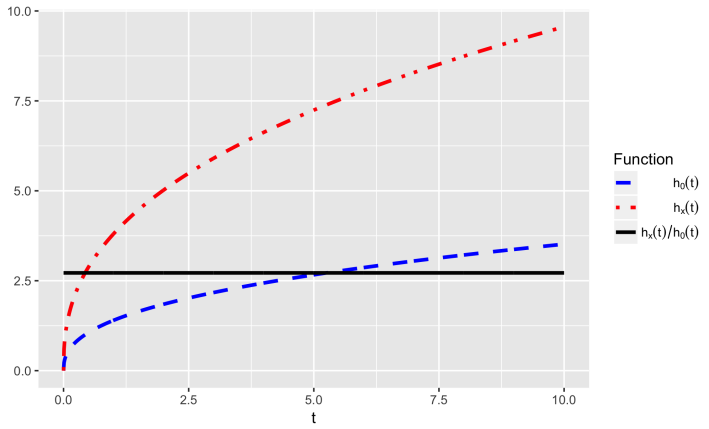


Figure 2: Illustrated are the hazard functions $h_0(t)$ and $h_1(t)$ from Example 4.1, as well as the ratio $\frac{h_1(t)}{h_0(t)}$ between them. Notice how because of the proportional hazards effect the ratio between the two hazard functions is constant.

4.2 Accelerated time model

Another common parametric survival model is the accelerated failure time model (AFT). The idea behind the AFT is that the time to failure of a unit under effect of covariates $\boldsymbol{x}$ is accelerated by some constant amount $\psi(\boldsymbol{x})$. Let us define what we mean by this.

Definition 4 (Accelerated failure time model, p. 64 in [2])**.** *Denote the survival function of a random variable T affected by covariates in $\boldsymbol{x}$ by $S_T(t;\boldsymbol{x})$. By employing the* accelerated time model *we assume that the effect of covariates can be described by the relation*

$$S_T(t;\boldsymbol{x}) = S_T\left(\psi(\boldsymbol{x})\, t; \mathbf{0}\right).$$

Just like with the PH model, we will always let $\psi(\boldsymbol{x}) = e^{\boldsymbol{\beta^T x}}$ in this paper.

Expressions for the density, hazard, and cumulative hazard of a unit under

covariate influence can be derived from their definitions:

$$f_T(t;\boldsymbol{x}) = \psi(\boldsymbol{x}) f_T\left(\psi(\boldsymbol{x})t;\mathbf{0}\right) \tag{4.3}$$
$$h_T\left(t;\boldsymbol{x}\right) = \psi(\boldsymbol{x}) h_T\left(\psi(\boldsymbol{x})t;\mathbf{0}\right) \tag{4.4}$$
$$H_T\left(t;\boldsymbol{x}\right) = H_T(\psi(\boldsymbol{x})t;\mathbf{0}) \tag{4.5}$$

One way to think of the AFT assumption is that the underlying random variable (time to failure) of a unit under covariate effect $T_{\boldsymbol{x}}$ is identically distributed to $T_{\mathbf{0}}/\psi(\boldsymbol{x})$. This also yields the interpretation of covariate effects. An example is if we have a binary covariate, with parameter value $\beta = \log 2$. Then, a unit under the influence of the covariate cuts the expected time to failure in half compared to if it was not under covariate effect.

Example 4.2. Let us assume that the AFT model holds for some independent and identically distributed failure times T_i under the effect of covariates $\boldsymbol{x}$, and assume a baseline Weibull distribution. That is,

$$S_{\mathbf{0}}(t) = \exp\left\{-\left(t/\lambda\right)^k\right\}.$$

Under the AFT, the survival, density, hazard, and cumulative hazard functions conditioned on a covariate vector $\boldsymbol{x}$ is found to be

$$\begin{aligned}
S_{\boldsymbol{x}}(t) &= \exp\left\{-\left(\frac{t}{\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}}\right)^k\right\} \\
f_{\boldsymbol{x}}(t) &= \frac{k}{\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{t}}\boldsymbol{x}}}\left(\frac{t}{\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}}\right)^{k-1}\exp\left\{-\left(\frac{t}{\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}}\right)^k\right\} \\
h_{\boldsymbol{x}}(t) &= \frac{k}{\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{t}}\boldsymbol{x}}}\left(\frac{t}{\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}}\right)^{k-1} \\
H_{\boldsymbol{x}}(t) &= \left(\frac{t}{\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}}\right)^k,
\end{aligned}$$

using Definition 4.2 and Equations (4.3), (4.4), and (4.5). Writing this is terms of the log cumulative hazard and reparametrizing $\tilde{\boldsymbol{\beta}} = k\boldsymbol{\beta}$ we obtain

$$\log\left(H_{\boldsymbol{x}}(t)\right) = k\log t - k\log\lambda + \tilde{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x}$$

Notice that this is exactly the same as the PH model when assuming a baseline Weibull distribution derived in Example 4.1. In general, this is not the case. Under the AFT model, and Weibull baseline distributed failure times $T_0 \sim \text{Weib}\left(k,\lambda\right)$, covariate affected failure times are distributed according to $T_{\boldsymbol{x}} \sim \text{Weib}\left(k, \lambda e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}\right)$.

4.3 Generalized survival model

As the name suggests, the generalized survival model (GSM) is a generalization of parametric survival models. It is more general in the sense that it allows for a much wider variety of distributions of the failure times than, for example, the proportional hazard model can offer.

Definition 5 (Generalized survival model, section 2.1 in [10])**.** *Let g be a link function, t be a realization of a failure time (positive random variable) T, $\boldsymbol{x}$ a covariate vector, and $\boldsymbol{\theta}$ a parameter vector. The generalized survival model takes the form*

$$g(S(t|\boldsymbol{x};\boldsymbol{\theta})) = X(t,\boldsymbol{x})\boldsymbol{\theta} = \eta(t,\boldsymbol{x};\boldsymbol{\theta}) \tag{4.6}$$

where X is the model matrix, and η is a linear predictor.

We could express the linear predictor as $\eta(t,\boldsymbol{x};\boldsymbol{\theta}) = g(S_0(t)) + \boldsymbol{\beta^T x}$, which expresses the covariate effects as being *additive*. Note that the model matrix X is time dependent, where the linear predictor is a function of time. Hopefully the example in the end of this section will help to illustrate this.

What assumptions are we putting on the link function? Well, as the name suggests, the purpose of a link function is to act as a *link* between the linear predictor (whose image can be all of $\mathbb{R}$) and in this case, the survival function (whose image is the interval $(0,1)$). Then it seems like a necessity that such a function is invertible, since we want to be able to retrieve the survival function given the linear predictor, just as well as we can obtain the linear predictor given the survival function. We will use the same assumptions on the link function made by Younes and Lachin [18] when they first presented the link-based model that the GSM stems from. Thus, we assume that the link-function $g : (0,1) \to \mathbb{R}$ is a bijective, strictly monotone function and known. This yields that g is invertible. We also make another assumption about g that is not made in [18], and that is that the inverse link function is differentiable with respect to t. The reason for this will become apparent when we write the hazard function under the GSM.

The linearity of η refers to the fact that it is linear with respect to the parameter vector $\boldsymbol{\theta}$. From the model matrix representation of the GSM, the linearity is captured by the matrix multiplication of X with $\boldsymbol{\theta}$, while the way in which the model matrix depends on time may be complex. This is hopefully clarified in the example at the end of this section.

One assumption that we make about η is that it is twice differentiable with respect to time. The need for the once differentiable assumption will (just as with the link function) become apparent when we try to write the hazard function under the GSM. In biological settings, it is common to assume that the underlying hazard "changes smoothly with time" [18]. Taking smooth to mean differentiable, we will see that this means exactly that the linear predictor should be twice differentiable in the next paragraph.

Let $G = g^{-1}$ be the inverse link function. Using the Equation (4.6), and the connections between the survival, hazard, and cumulative hazard found in Section 2.1.3, we find explicit expressions for them:

$$\begin{aligned} S(t|\boldsymbol{x};\boldsymbol{\theta}) &= G(\eta(t,\boldsymbol{x};\boldsymbol{\theta})) \\ H(t|\boldsymbol{x};\boldsymbol{\theta}) &= -\log\left(G(\eta(t,\boldsymbol{x};\boldsymbol{\theta}))\right) \\ h(t|\boldsymbol{x};\boldsymbol{\theta}) &= -\frac{G'(\eta(t,\boldsymbol{x};\boldsymbol{\theta}))}{G(\eta(t,\boldsymbol{x};\boldsymbol{\theta}))}\,\eta'_t(t,\boldsymbol{x};\boldsymbol{\theta}). \end{aligned} \tag{4.7}$$

Here we see the reason for differentiability of the inverse link function and the twice differentiability of the linear predictor. If they were not, the hazard function would not be defined and differentiable under the GSM.

In *stpm2*, the default is that the time effects in the linear predictor are modelled as natural splines for log time [18]. That is, we assume that the true baseline hazard can be expressed as

$$g(S_0(t)) = s_K^d(\log t)$$

where $s_K^d(\log t)$ is a spline of degree d over knots in the set K. It is not necessary to model on the log time scale, but it has been suggested a good choice for the most common link functions[10]. A natural spline is a picewise polynomial function, with a zero second derivative in the left- and rightmost knots. The values of t where the polynomial pieces meet are called knots, and the spline is required to be continuous and twice differentiable at these knots (recall the remark made earlier in this section about the underlying linear predictor changing smoothly with respect to time). Given a degree d, and $|K|$ knots in the set K, there exists a *basis spline* having the property that all splines $s_K^d(t)$ can be written as a linear combination

$$s_K^d(\log t) = \sum_{i=1}^{|K|-1} \theta_i B_i^d(\log t)$$

where $B_i^d(\log t)$ is a basis function with natural spline properties. In R, the natural splines are calculated by the `splines::ns()` function using a matrix projection of B-splines. An algorithm for constructing the B-splines is presented in [18]. Having found a basis, we fit the spline by adjusting the parameters θ_i. In order to enforce positivity of the hazard, we use a quadratic penalty to ensure that $\eta_t'(t, \boldsymbol{x}; \boldsymbol{\theta}) > 0$ [10]. One of the advantages of the GSMs is that the calculation of survival does not require integration, while the calculation of the hazard requires differentiation of the linear predictor with respect to time.

Wold [17] proposed that knot number and placement should not be considered ordinary parameters of the curve fitting process, but instead thought of as deciding the underlying functional form. Instead, Wold suggested that knots should be placed in areas where data density is high, or places where it attains a maximum or minimum. In *stpm2* the knots are placed on the observed failure time quantiles, a strategy which has been suggested when nothing is known about the underlying distribution the data [18]. In the case of censored observations the knots are placed on the quantiles of the left censoring interval boundary if the unit is interval or right censored, and on the right censoring interval boundary if the unit is left censored. The number of knots to use can be determined using forward selection on AIC [18].

Example 4.3. Consider this special case of the GSM:
Let $\boldsymbol{x}$ be a covariate vector of length n, and

$$\begin{aligned} \boldsymbol{\theta} &= (k, \alpha, \beta_1, \ldots, \beta_n)^T \\ g(S(t)) &= \log(-\log(S(t))) \\ \eta(t, \boldsymbol{x}; \boldsymbol{\theta}) &= k \log t + \alpha + \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{x}. \end{aligned}$$

Notice that the linear predictor is differentiable with respect to time, $\eta_t'(t, \boldsymbol{x}; \boldsymbol{\theta}) = \frac{k}{t}$. Also, the link function has a differentiable inverse, $G(\eta(t, \boldsymbol{x}; \boldsymbol{\theta})) = \exp\{-\exp\{\eta(t, \boldsymbol{x}; \boldsymbol{\theta})\}\}$.

Then, the model matrix and parameter vector for n covariates and m units (denoting the jth covariate of the ith unit by x_{ij}) is simply

$$X(t, \boldsymbol{x}) = \begin{pmatrix} \log t & 1 & x_{11} & \dots & x_{1n} \\ \log t & 1 & x_{21} & \dots & x_{2n} \\ \log t & 1 & x_{31} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \log t & 1 & x_{m1} & \dots & x_{mn} \end{pmatrix}$$

and

$$\begin{pmatrix} k \\ \alpha \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}.$$

Now, using equation (4.7) we can find an expression for the hazard function:

$$h(t|\boldsymbol{x}, \boldsymbol{\theta}) = kt^{k-1}e^{\alpha}e^{\boldsymbol{\beta}^T\boldsymbol{x}}.$$

Reparametrizing $\alpha = -k \log \lambda$ we get the hazard function of the Weibull distribution under PH (or equivalently AFT) effects, presented in Examples 4.2 and 4.1. That is, this special case of the GSM is equivalent to the AFT and PH models. This fact will be used to compare the GSM implemented in the R package *rstpm2* to the AFT in the standard R package *survival* in Section 7.

5 Statistical comparison methods of models

Following Liu et al [10], we will primarily use two methods of comparing different models. The first one is a quantitative measure, simply measuring the area between the estimated baseline hazard curve, and the real baseline hazard, between the first and last observed failure time. That is:

$$A(\hat{h}, h) = \int_{t_{\min}}^{t_{\max}} |\hat{h}_0(t') - h_0(t')|dt'$$

In this paper we will estimate the hazard functions and use the average area between the true hazard and the estimated hazards. We will refer to this measure as the area difference of a model. This is a very simple way to compare models, where the more an estimated hazard diverges from the true hazard, then the larger the area between them becomes.

Another method of comparison that we are going to use is a graphical one. Recall the definition of the emprical mean square error:

$$\text{EMSE}(\hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - Y_i)^2$$

In this case though we are not comparing a estimated parameter and a true parameter value, or predictions with observed values, but an estimated hazard function with the true underlying hazard function, depending on time. That is, the EMSE of the hazard function is itself going to be a function, dependent on time. Given $i = 1, \cdots, N$ estimated baseline hazard functions, the ith of which is denoted $\hat{h}_{0,i}(t)$, the EMSE is:

$$\text{EMSE}(\hat{h}_0(t)) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{h}_{0,i}(t) - h_0(t) \right)^2.$$

In this thesis, we compare the EMSE of different models over the region of time where events can be observed.

6 Simulation methods

We will simulate interval censored data according to prescription 1, p.675 in [8], also discussed in [3]. The idea is to generate examination times in some manner, and then independently of that simulate failure times from the desired distribution. Examination times are the moments when we observe the unit. If a failure occurs in between two examinations, it is censored on that interval. Different interval lengths are obtained by each unit missing out on each examination with some predestined probability p. In the next paragraph the simulation method is described in more detail.

Let N be the total number of observations we want to simulate, indexed by i.

1. Simulate N failure times $t_1, t_2, \ldots, t_N$ from a distribution with some known parameter θ. This step is described specifically for the Weibull distribution and mixed Weibull distribution in Sections 7.1 and 8.2 respectively.

2. Generate M potential examination times C_j, $j = 1, 2, \ldots, M$. In Section 7 we simply use $C_j = j$, while in Section 8 we consider more frequent examinations as well. These potential examination times are the same for all units in the sample.

3. For every unit, each potential examination is visited with some probability p. The largest measure time smaller than t_i becomes L_i, the lower bound of the censoring interval. The smallest measure time larger than t_i becomes R_i, the upper bound of the censoring interval.

This method can produce intervals of different lengths by adjusting the visiting probability p, and produces left, right, and interval censored data. If there is no visited measure time before t_i, then the observation is left censored. Similarly, if there is no visited measure time after t_i, the observation becomes right censored. If none of the measure times are visited, the observation is discarded.

Recall the earlier discussion on non-informativity in Section 3.1. The simulation method presented in this section satisfies the non-informativity assumption, since the measure times and failure times are simulated independently. An example of a simulation method subject to informative censoring could look something like this:

Simulate failure times T_i and generate potential measure times $C_{ij},\ j = 1, \ldots, M$ exactly like in the previous proposed method. Now, for the ith unit, the jth measure time C_{ij} is visited with the probability p if $C_{ij} < T_i$, and zero otherwise. With this simulation method the censoring intervals are clearly dependent on the failure times, and therefore informative.

7 Weibull simulation study

In this section, we will continue the analysis from Example 4.3. We will simulate from a Weibull distribution with a binary and a continuous covariate, and fit both the GSM and AFT models to the data. We will only consider the special case of the GSM discussed in the example, when it is equivalent to the AFT model. This will allow us to make a simple performance check of the GSM implemented in *stpm2* compared to the AFT implemented in *survreg*.

7.1 Data simulation

The data were simulated according to the procedure described in Section 6. Failure times were simulated from the Weibull distribution by the following procedure.

1. Simulate n uniformly distributed numbers $u_i,\ i = 1, \cdots, n$, between 0 and 1.

2. Using the quantile function, $q_W(u)$ of the Weibull distribution with shape k and scale $\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}$, the values $q_W(u_i)$ are observations from the Weibull distribution.

We used $M = 10$ potential examinations throughout this simulation. The probability with which each examination is visited was put to 0.9 with one exception (which will be explained shortly).

We use a total of four different parameter scenarios in this study. We consider three different distributional forms of the Weibull distribution, with shapes equal to 0.5, 1, and 1.5. From there, we adjust the scale parameter such that about 90% of the simulated failure times would be expected to occur on the interval $(0, M] = (0, 10]$. For some shape k, percentage q, random variable T, and M, this is done by solving the equation

$$F_T\left(M; \lambda, k\right) = q,$$

for the scale λ. These three scenarios will be referred to as scenarios 1, 2 and 3, respectively.

In the fourth scenario we simulate from the same distribution as in scenario 2, but we decrease the probability by which each measuring time is visited from 0.9 to 0.6. This will result in wider censoring intervals, as well as a larger proportion of right and left censored observations.

With each scenario we also attach two covariates: one binary, and one continuous. This is done by simulating covariate values for the binary covariate from the bernoulli distribution (with parameter 0.5) and for the continuous from the standard normal distribution. These will have the same coefficients for all scenarios.

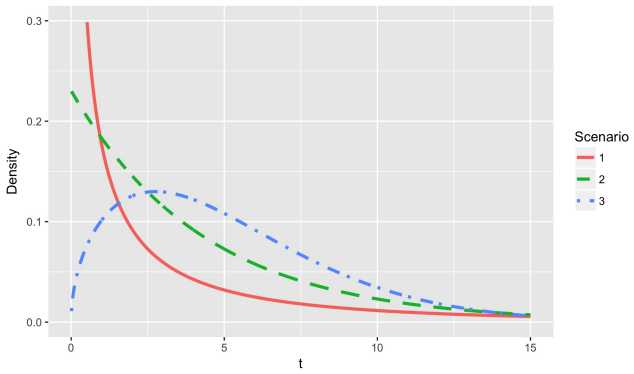


Figure 3: The three different distributional shapes used in the four different parameter scenarios of Section 7.

Scenario	Shape	Scale	Binary covariate	Continuous covariate
1	0.5	1.886	0.25	0.25
2	1.0	4.343	0.25	0.25
3	1.5	5.735	0.25	0.25
4	1.0	4.343	0.25	0.25

Table 2: Parameter values for the four different simulation scenarios from Section 7. The first three are different distributional shapes, and in the fourth we simulate from the same distribution as in the second but with a lower probability of each examination being visited, resulting in wider censoring intervals.

The three distributions used are visualized in Figure 3, and the parameter values can be found in Table 2.

The simulation procedure is summarized as follows:

1. Simulate n observations from the Weibull distribution with shape k and scale $\lambda e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}$ (recall the remark in the end of Example 4.2).
2. Apply the AFT model using *survreg* from the package *survival* and the GSM model using *stpm2* function from the *rstpm2* package.
3. Reparametrize both models using the multivariate delta method (see the appendix Section 11.1.3 for a short description of the multivariate delta method), to get estimates and confidence intervals for the shapes, scales, and covariate effects.
4. Repeat steps 1-3 N times to get N different estimates for each parameter.
5. Repeat steps 1-4 for each parameter scenario.

In this study, we used $n = 500$ and $N = 10^4$. These values were chosen based on an compromise between available computing power, and obtaining an acceptable level of accuracy.

7.2 Analysis and results

The two functions produce extremely similar results. The kernel density estimator (for a short introduction, see Section 11.1.4 of the appendix) of the binary covariate parameter estimator from scenario four can be observed in Figure 4, and plots for all parameters and scenarios can be found in Figure 14 in the appendix. Simply by looking at the distributions of the estimators, one notices how similar they are.

This is further reinforced when inspecting the distribution details of all the parameters in Table 9 in the appendix. The sample mean, sample standard deviations, and 95% confidence interval coverages are identical for the two models and all parameters.

In Table 3, we find the average area between the estimated hazard and true hazard, Akaike information criterion (AIC), covariate parameter estimates, and coverage proportions of the 95% confidence intervals of the covariate parameters, for both models. Again, the two models produce strikingly similar results. One observation that can be made here is that for scenario 2, the area $A(\hat{h}, h)$ is about 70% to the area of scenario 4. This is not very surprising, since even though the two scenarios use the same parameter estimates, scenario four will generally have wider censoring intervals because of the lower examination probability. This results in less accurate estimators, and thus generally a larger deviance from the true hazard.

The time dependent EMSE plots look identical for all parameter scenarios. In Figure 5 the time dependent EMSE's of the second parameter scenario is illustrated, while plots for all of the parameter scenarios can be found in Figure 15 in the appendix.

The confidence interval coverage proportions show no indication of conservative or anti-conservative confidence intervals. In Table 3 we observe that the confidence intervals of the two model have the same average coverage proportions. In fact, we found that both implementations produce equal confidence intervals.

If we want to make inference about the biases of the covariate parameter estimators, the sixteen different scenario and parameter combinations unfortunately makes us subject to the multiple comparisons problem. Taking this into consideration, we use the Holm-Bonferroni method [5] to adjust the p-values to counteract the issue. That is, we create a family of m hypotheses $\{H_i\}$ and their corresponding p-values p_i. Then we sort the hypotheses and p-values in ascending order according to the p-values. We then compare the ith hypothesis to $\frac{\alpha}{m-i}$, $i = 0, \cdots, m-1$, rejecting any null hypotheses with $p_i \leq \frac{\alpha}{m-i}$. In this case, there are 16 parameter-scenario combinations so we have $m = 16$ hypotheses, and we are testing on the level $\alpha = 0.05$.

In order to investigate bias in the parameters we create a family of hypotheses, one for each parameter and scenario. For the ith parameter β_i with true parameter value β_i^{true}, the hypothesis becomes

$$H_0 : E(\hat{\beta}_i) = \beta_i$$
$$H_a : E(\hat{\beta}_i) \neq \beta_i.$$

Creating t statistics for all scenarios and parameters using the sample means and standard errors from Table 9 in the appendix, and testing these using the

Scenario	Implementation	$A(\hat{h}, h)$	df	AIC	$\hat{E}\left[\hat{\beta}_{\text{bin}}\right]$ (95% CI)	$\hat{E}\left[\hat{\beta}_{\text{cont}}\right]$ (95% CI)	Binary CP	Continous CP
1	stpm2	0.74	4	1511.71	0.25 (0.246, 0.254)	0.25 (0.248, 0.252)	0.95	0.94
	survreg	0.74	4	1511.71	0.25 (0.246, 0.254)	0.25 (0.248, 0.252)	0.95	0.94
2	stpm2	0.68	4	2058.07	0.25 (0.248, 0.252)	0.25 (0.249, 0.251)	0.95	0.94
	survreg	0.68	4	2058.07	0.25 (0.248, 0.252)	0.25 (0.249, 0.251)	0.95	0.94
3	stpm2	0.78	4	2118.19	0.25 (0.249, 0.251)	0.25 (0.249, 0.251)	0.95	0.96
	survreg	0.78	4	2118.19	0.25 (0.249, 0.251)	0.25 (0.249, 0.251)	0.95	0.96
4	stpm2	1	4	810.75	0.25 (0.248, 0.252)	0.25 (0.249, 0.251)	0.95	0.95
	survreg	1	4	810.75	0.25 (0.248, 0.252)	0.25 (0.249, 0.251)	0.95	0.95

Table 3: A summary of the results from the simulation study of *stpm2* and *survreg* on Weib(k, λ) distributed censored data. Visualized in the table is the area difference, degrees of freedom, AIC, average parameter estimates and confidence intervals, and 95% confidence interval coverage proportions for each covariate.

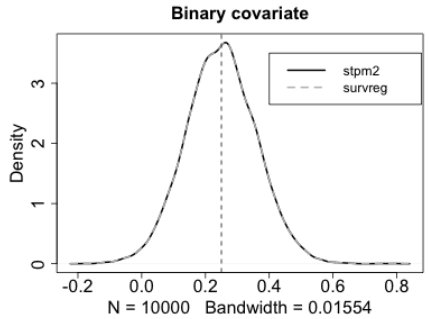


Figure 4: The kernel density estimates of the binary covariate parameter estimator for both the GSM and AFT models from parameter scenario 2 of Section 7. The dotted line represents the true parameter value.

previously discussed Holm-Bonferroni method, we do not obtain any significant indication of bias in the estimators for a sample size of 500.

8 Mixed Weibull simulation study

In the previous section we examined the performance of the two implementations *stpm2* and *survreg* of the AFT model, in the case of an underlying Weibull distribution. This is of course an interesting and important study to make, but it does not showcase the possible need for the GSM. If all failure times followed a standard distribution, then the AFT and PH models would always suffice. In many practical settings in biostatistics, this is unlikely to be the case. In this section, we are going to assess the performance of the Weibull PH implemented in *survreg*, compared to the GSM in *stpm2*.

We would expect that the GSM would outperform the PH, since it defaults back to the Weibull PH when using zero knots. This is an ideal case though, and will depend on the implementations of the two models.

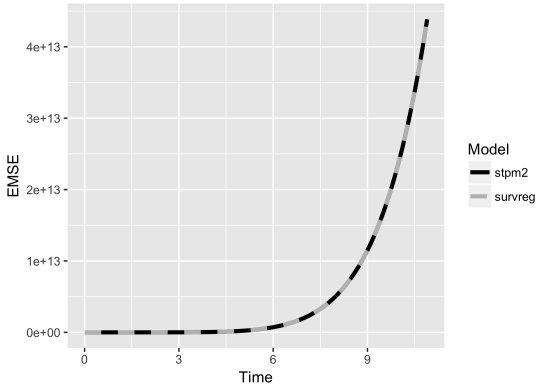


Figure 5: The time dependent empirical square mean error (EMSE) plots for parameter scenario 2. The two implementations produce identical EMSE for all examined values of t.

8.1 Mixed Weibull distribution

The underlying distribution of the failure times in this section will be a mixture of Weibull distributions. This distribution has been used in previous simulation studies [13] to evaluate the performance of spline based models.

The idea of the mixture distribution is simple. Let $m_i \in [0,1],\ \sum_{i=1}^{n} m_i = 1$ be n *mixing parameters*, $k_1, \ldots, k_n$ shape parameters, and $\lambda_1, \ldots, \lambda_n$ scale parameters. Also let $f_i(t)$ be the density of a Weibull distribution with parameters k_i and λ_i. We say that a random variable T is mixture of Weibull distributions with parameters $m_i,\ k_i,\ \lambda_i,\ i = 1, \ldots, n$ if it has the density function

$$f_T(t) = \sum_{i=1}^{n} m_i f_i(t), \tag{8.1}$$

for $t > 0$, and 0 for $t \leq 0$. In this section, we will refer to the $f_i(t)$s as *component distributions*.

Expressions for the survival function and hazard function of T can easily be calculated from the density:

$$S_T(t) = \sum_{i=1}^{n} m_i e^{-\left(\frac{t}{\lambda_i}\right)^{k_i}} \tag{8.2}$$

$$h_T(t) = \frac{\sum_{i=1}^{n} m_i \frac{k_i}{\lambda_i} \left(\frac{t}{\lambda_i}\right)^{k_i - 1} e^{-\left(\frac{t}{\lambda_i}\right)^{k_i}}}{\sum_{i=1}^{n} m_i e^{-\left(\frac{t}{\lambda_i}\right)^{k_i}}} \tag{8.3}$$

The nice properties of the Weibull distribution transfer naturally to the mixed Weibull distribution. For example, since all the m_is sum to one, and the density function of any Weibull distribution integrate to one from 0 to ∞, we have that

$$\int_0^\infty f_T(s)ds = 1 \tag{8.4}$$

for a mixed Weibull distributed random variable T.

8.2 Simulation of mixed Weibull distributed data

Using the definition of the PH model (Definition 3), and the equations presented in Section 2.1.3 we can rewrite the PH assumption to

$$S\left(t|\boldsymbol{x}\right) = S\left(t|\mathbf{0}\right)^{e^{\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}}.$$

From here, we can use the inverse of the previous stated function to generate the failure times. Below follows a more detailed description of the method.

1. Simulate n uniformly distributed numbers $u_j,\ i = 1, \ldots, n$ between 0 and 1.

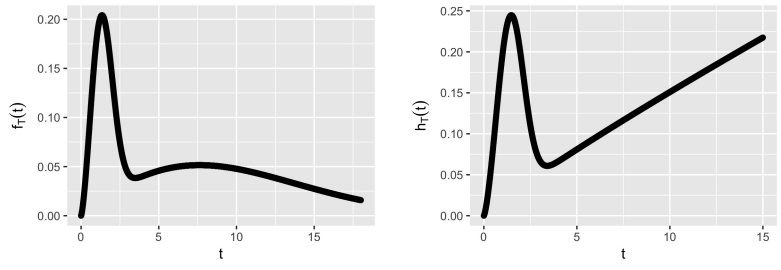
2. Adjust the uniformly simulated values according to their corresponding covariate values $\tilde{u}_j = u_j^{e^{-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}}}$.

3. Let W be of the baseline mixed Weibull distribution, and q_W its quantile function. The values $q_W(\tilde{u}_j)$ are now observations from $W|X = x$.

The quantile function of the mixed Weibull distribution can not be easily calculated. Instead we use the root finding function *uniroot* in R to find the value of $S(t|0)$ which equals $\tilde{u}_j$.

In this study we will consider failure times T with the mixture of two Weibulls as baseline distribution, more specifically $T_0 \sim \text{MixWei}(m = 0.7, k_1 = 1.9, k_2 = 2.5, \lambda_1 = 11.29, \lambda_2 = 1.62)$. This is the same distribution used in Scenario 4 in [13]. This distribution is unimodal with a right shoulder, and the density and hazard are visualized in Figure 6.

In order to capture the more complicated features of the mixed Weibull distribution, we will use more detailed potential examination times $C_i = i \times 10^{-1},\ i = 1, \ldots, 160$. Define the step length of a censoring technique to be the shortest distance between any pair of possible examination times. In this case, the step length is 10^{-1}. To compensate for the increase in potential examination times, we will use a lower probability of each potential examination actually being visited to $p = 0.5$. As in the previous section the units will be under



(a) Density (b) Hazard

Figure 6: The density and hazard of the mixed Weibull distribution considered in Section 8, $T \sim \text{MixWei}(m = 0.7, k_1 = 1.9, k_2 = 2.5, \lambda_1 = 11.29, \lambda_2 = 1.62)$.

the effect of two covariates, one binary and one continuous. We use different

values of the covariate parameters though. In this section we use 0.5 for both parameters.

500 simulated values from the baseline distribution are visualized in Figure 7.

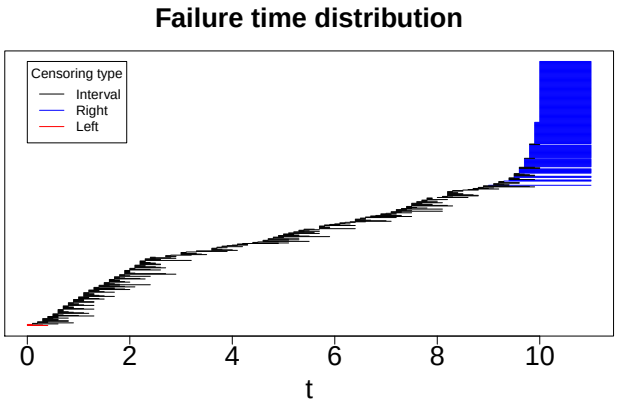


Figure 7: Visualizing the failure times from the mixed Weibull simulation study. Each line segment represents the interval an observation in censored on. The observations are sorted on their left interval limits. Notice the group of right censored observations in the top right corner.

8.3 Method

In order to fit a GSM allowing for more than zero knots we need a way to determine the amount of knots to use. We will use AIC as a knot count selection tool. More precisely, we will fit the GSM with $0, \ldots, 8$ knots and then choose the model with the highest AIC. This is not a very fast procedure to repeat a lot of times. A faster but less fair approach would be to test the optimal knot count on a small number of simulated data sets, and then simply test for the two most common knot count, or even just using the most common knot count (see the `logspline` package on CRAN for an approach that smooths on the log hazard scale with knot selection).

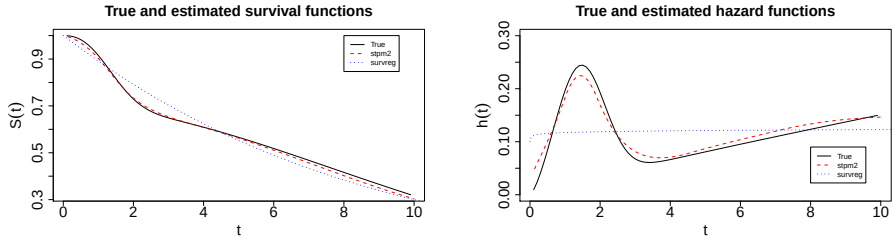
Due to numerical instability, we did not calculate the area between the estimated and true hazard between the minimum and maximum of the simulated failure times. The two models yield numerical overflow for estimated hazards that are outside the possible censoring interval. In order to compensate for this, we instead measure the area between the first simulated failure time and the last possible examination time. The area will in this section then measure the performance of the two models up to the last possible examination time.

8.4 Performance results

Unlike in the last section, the two models do not display the same behavior when applied to the mixed Weibull data. The PH model implemented in *survreg* misses the complex features of the mixed Weibull distribution. The GSM implemented in *stpm2* is often able to capture these features, placing knots as

	$A\,(t_{\min}, C_{\max})$	Area SD	df	AIC	$\hat{\mathrm{E}}[\hat{\beta}_{\mathrm{bin}}]$ (95% CI) ($\beta_{\mathrm{bin}} = 0.5$)	$\hat{\mathrm{E}}[\hat{\beta}_{\mathrm{cont}}]$ (95% CI) ($\beta_{\mathrm{cont}} = 0.5$)
stpm2	0.260	0.391	7.779	2728.848	0.495 (0.493,0.497)	0.495 (0.494,0.496)
survreg	0.387	0.026	4.000	2764.068	0.498 (0.496,0.500)	0.498 (0.497,0.499)

Table 4: Summarising the results of the simulation study. The mean area between the estimated hazard and the true hazard, degrees of freedom, AIC, and means and standard errors of the covariate parameters are displayed. Besides the covariate headers, their true values are displayed.



(a) The survival function.

(b) The hazard function.

Figure 8: Illustrated here are examples of the estimated and the true hazard and survival functions for the two models. The examples were chosen because they have area differences close to the observed mean area difference and were fitted on the same data set.

discussed in Section 4.3 and fitting natural splines to the observed values of $-\log(-\log(S(t)))$.

The simulation results are summarised in Table 4. We note that the GSM produces a lower mean area and AIC than the Weibull PH, indicating a better fit. Also, the mean covariate parameter estimates of the GSM are slightly closer to the true values compared with those from the PH model. The covariate parameter kernel density estimates of the two models can be observed in Figure 16 in the appendix. The two models seem to produce similar, but slightly different covariate estimates. In Table 4 we see that both models produce slightly biased covariate estimates for this sample size, except for *survreg* and the binary covariate.

In Figure 8 representatives of the estimated survival and hazard functions are drawn together with the true curves. The representatives were chosen at random, but were applied to the same simulated data set. Observe how in Figure 8b the Weibull PH model simply cuts through the peak of the true hazard function, while the GSM does a better job of capturing this feature.

In Figure 9, the time dependent EMSE is plotted for both models. The figure also indicates that the GSM in *stpm2* does a better job of capturing the more complex features of the distribution.

8.5 Varying the censoring parameters

In the mixed Weibull simulation study the possible examinations were more frequent than in the ordinary Weibull case, in order to capture the more advanced

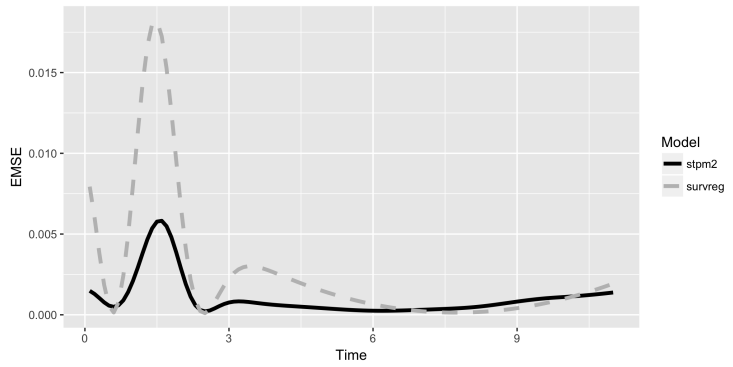


Figure 9: The time dependent empirical mean square error (EMSE) of the two models implemented on the simulated mixed Weibull data. The GSM (*stpm2*) seems to perform much better than the Weibull PH (*survreg*).

features of the more complex distributional shape. In actual studies, control of examination frequency is often limited due to financial or practical aspects. In this section we examine the behavior of *stpm2* when we change the censoring resolution. We consider three cases:

1. High resolution: step length: 10^{-2}, probability of examination visit: 0.5.
2. Medium resolution: step length: 10^{-1}, probability of examination visit: 0.5.
3. Low resolution: step length: 1, probability of examination visit: 0.5.

In each of the 500 simulation iterations we simulate 500 observations. The failure times for one of the simulation iterations of the three censoring cases are visualized in Figure 10.

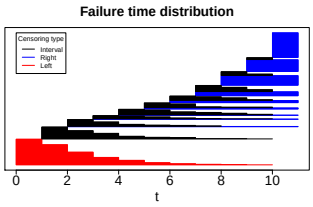
Intuitively, the more exact the data is, the more complicated behavior can be captured by *stpm2*. This is visualized in Figure 11, where we see how the EMSE of *stpm2* decrease drastically with the higher censoring resolution while the performance of *survreg* barely changes. The application of *stpm2* on the step length 1 data stands out as much worse than the other two with regards to EMSE. This suggests a limit at which any efforts to further shorten the censoring intervals leaves one subject to diminishing returns. The length of the censoring intervals determines the magnitude of failure time distribution characteristics one is able to detect.

The covariate estimates continue to exhibit slight bias, except for *survreg* with step length 0.1 and the binary covariate.

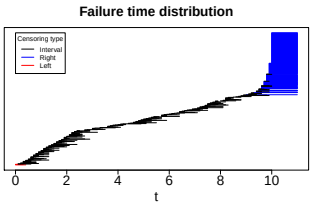
9 The Signal-Tandmobiel®study

9.1 Application

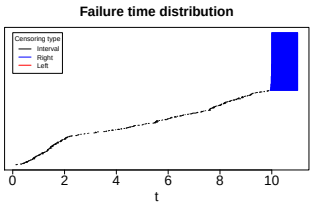
As has been done in previous demonstrations using the data set [3], we will use the emergence of permanent tooth 24 as our event. Due to the yearly



(a) Step length 1.



(b) Step length 10^{-1}.



(c) Step length 10^{-2}.

Figure 10: Failure times for three different censoring resolution scenarios of Section 8.5. One thing to note is that the proportion of left censored observations decreases with increasing censoring resolution.

Step length	Implementation	$A(t_-, C_+)$	df	AIC	$\hat{\mathrm{E}}[\hat{\beta}_{\mathrm{bin}}]$ (95% CI) ($\beta_{\mathrm{bin}} = 0.5$)	$\mathrm{sd}(\hat{\beta}_{\mathrm{bin}})$	$\hat{\mathrm{E}}[\hat{\beta}_{\mathrm{cont}}]$ (95% CI) ($\beta_{\mathrm{cont}} = 0.5$)	$\mathrm{sd}(\hat{\beta}_{\mathrm{cont}})$
1	stpm2	0.365	5.237	1381.552	0.504 (0.502,0.506)	0.127	0.505 (0.504,0.506)	0.069
	survreg	0.388	4.000	1383.185	0.505 (0.503,0.507)	0.127	0.505 (0.504,0.506)	0.069
0.1	stpm2	0.260	7.779	2728.848	0.495 (0.493,0.497)	0.123	0.495 (0.494,0.496)	0.067
	survreg	0.387	4.000	2764.068	0.498 (0.496,0.500)	0.124	0.498 (0.497,0.499)	0.067
0.001	stpm2	0.330	7.811	4111.811	0.489 (0.487,0.491)	0.122	0.493 (0.492,0.494)	0.068
	survreg	0.387	4.000	4146.836	0.492 (0.490, 0.494)	0.123	0.496 (0.495, 0.497)	0.068

Table 5: AIC, area difference, and covariate estimates and confidence intervals for the three different censoring resolution cases of Section 8.5.

examinations, the events are either left, right, or interval censored. We will consider the following two covariates:

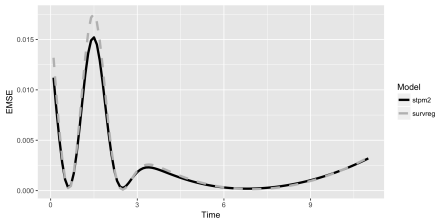
Sex — Binary variable, indicating the gender of the child: 0 for boy, 1 for girl.

DMF — Binary variable, indicating whether or not deciduous tooth 64 was decayed, missing or filled at the time of the last examination before the emergence of permanent tooth 24.

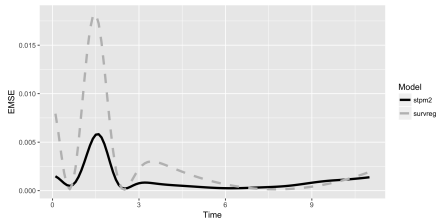
The baseline distribution of the AFT, and the link function of the GSM was chosen based on AIC scores. The available distributions and link functions are presented in Table 6. As before, we select the knot count for the GSM based on AIC as well.

The assumptions of non-informative, independent and random censoring introduced in Section 3.1 are simply assumed to hold, allowing the use of the simplified likelihood function (3.2).

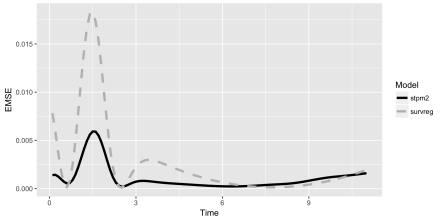
The failure times are visualized in Figure 12.



(a) Step length 1.



(b) Step length 10^{-1}.



(c) Step length 10^{-2}.

Figure 11: EMSE for the three different censoring resolution scenarios of Section 8.5. From Subfigure 11a to Subfigure 11b the EMSE decreases significantly. This is not the case when increasing the resolution further.

AFT distributions	GSM
Weibull	Proportional hazard
Exponential	Proportional odds
Gaussian	Probit
Logistic	Additive hazard
Log-normal	
Log-logistic	

Table 6: The considered distributions for the AFT, and the link functions for the GSM. The lognormal distribution and the probit link functions produced the lowest AIC values.

9.2 Results

For the AFT a baseline log-normal distribution provided the best AIC, and for the GSM the probit link function with five knots performed the best. The probit link function is $g(\cdot) = -\Phi^{-1}(\cdot)$, with inverse $-\Phi(\cdot)$. Results from both models are summarized in Table 7. First, we note that all covariate effects are significant with $p < 0.05$. We interpret the covariate estimates as follows:

For the AFT assuming a baseline log-normal distribution, the distribution of failure times for girls is log-normally distributed with a mean that is $\exp(-0.03663493) = 0.964028$ times their male counterparts. The other covariate effect estimate is interpreted in the same manner.

The effects with a negative probit link is interpreted as the change for a $-\Phi^{-1}(\cdot)$ transformation, which is the *negative* change in Z scores under a standard normal distribution. This indicates that the Z scores for girls are 0.3116609 *lower* than for boys, which is the same direction as the estimate from the AFT.

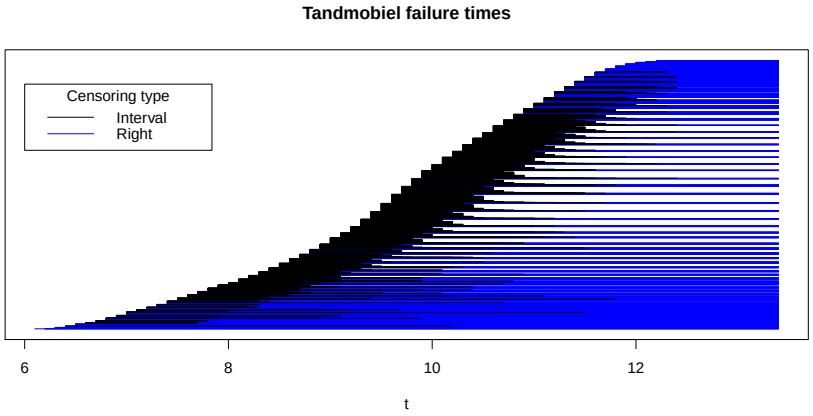


Figure 12: The failure times of the Tandmobiel data set visualized. Each horizontal bar represents a censoring interval.

	Link/Distribution	df	AIC	Sex (95% CI)	DMF (95% CI)
stpm2	probit	9.0000	10946.9103	0.3117 (0.2631, 0.3602)	0.3952 (0.3453, 0.4452)
survreg	lognormal	5.0000	10966.5792	-0.0366 (-0.0447, -0.0286)	-0.0468 (-0.0549, -0.0388)

Table 7: Summary of the results of the Signal Tandmobiel data set analysis. We note that the GSM scores a lower AIC value, indicating a better fit. All covariate effects are significant with $p < 0.05$.

That is,

$$\Phi^{-1}(S(t|\boldsymbol{x}, \text{girl})) = \Phi^{-1}(S(t|\boldsymbol{x}, \text{boy})) - 0.3116609.$$

More intuitive interpretations can be made by calculating marginal effects, such as the expected survival or expected hazard. If one wanted to calculate the estimated effect of the ith covariate one can calculate the partial derivative of the estimated survival function with respect to the covariate in question,

$$\frac{\partial S_{\boldsymbol{x}}(t)}{\partial x_i} = -\frac{\partial \eta(t, \boldsymbol{x}; \boldsymbol{\theta})}{\partial x_i} \varphi\left(-\eta(t, \boldsymbol{x}; \boldsymbol{\theta})\right),$$

which under an additive effects assumption simplifies to $-\beta_i \varphi\left(\Phi(S_0(t)) - \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{x}\right)$.

The GSM AIC is appreciably lower than AFT AIC, indicating that there are some features of the data that are not well captured by the AFT model. In Figure 13 the estimated cumulative hazard function for girls is plotted together with the non parametric Turnbull estimate (the Turnbull estimator is described in short in Section 11.1.2 of the appendix), as well as the estimated hazards for both boys and girls. In both figures, the DMF covariate is set to zero. Comparing the estimated curves with the Turnbull curve, it is hard to distinguish differences between the two models in the lower failure time regions, but it seems like the GSM provides a slightly better fit on the later failure times. Taking a look at the estimated hazards, the GSM seems to capture more detailed changes than the AFT model.

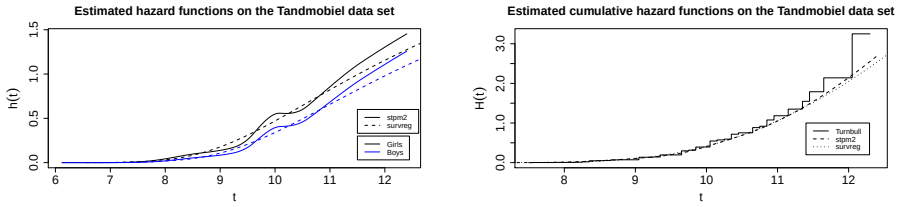


Figure 13: Estimated cumulative hazard functions of *stpm2* and *survreg* applied on the tandmobiel data set for girls with healthy teeth, and the estimated hazard functions for both boys and girls (in both figures, DMF = 0). The left figure agrees with our interpretations of the effect of gender.

10 Discussion

To summarize, in Section 7 we showed that the GSM implementation displays similar performance to the AFT in *survreg* in the special case of the GSM being equivalent to a Weibull PH. We tested this by applying the implementations on three different parameter scenarios of the Weibull distribution, as well as an additional scenario with wider censoring intervals. We continued by applying the GSM to a mixture of two Weibull distributions, and found that the GSM more closely captured the features of the more complex distributional shape. Moreover, we noticed that for very coarse censoring none of the discussed models managed to capture the data adequately. Finally we applied a lognormal AFT and a GSM with the probit link to the Signal Tandmobiel data set, and showed that the GSM seemed to capture the property of the data more closely than the more standard AFT model. The covariate interpretations we made were similar to previously published work [3][9].

The GSM implementation could be improved in a few ways. Currently, the spline knots of the GSM are placed on the quantlies of the left censoring interval limit (unless the data is left censored, in which case the right limit is used instead. There are other knot placement methods that could be used. A simple example would be to place the knots based on censoring interval mid-points. As previously mentioned, the GSM did not perform well when the censoring was of very low resolution. It would be useful to develop a diagnostic to detect these kinds of problems.

For further studies, there are a couple of different topics one could discuss. The scope of the GSM includes models with time varying effects. These types of models should be tested on interval censored data. More ambitiously, the *rstpm2* package recently implemented generalized AFTs for left truncated and right censored data. This could be extended to include interval censored data, which would allow for more direct comparison against to the AFT used in this paper. Included in the *rstpm2* package is a penalized GSM, which removes the need for knot count selection based on an information criterion. The package aslo includes support for normal random effect models for interval censored data. These two models could also be tested in a simulation study similar to this one.

11 Appendix

11.1 Theory

11.1.1 Censoring granularity

Here follows a short discussion about the effect of censoring granularity on the *stpm2* knot placement.

As discussed in Section 4.3, the knots of *stpm2* are placed on the quantiles of the observed left censoring limit in the case of interval and right censored data, and on the right censoring limit in the case of left censored data. For one simulated data set, the knot placements for one, up to and including five knots, for different censoring interval placement resolutions can be found in Table 8. In the study we used the step length 0.1 and examination probability 0.5. This yields (approximately) an expected interval length (given that the observation is interval censored) of 0.3. Employing a step length of for example, one, with the same probability of examination will yield significantly longer censoring intervals. Similarly using a shorter step length between each possible examination will yield much shorter censoring intervals.

Censoring parameters	knots	Knot placement				
$M = 10$	1	4				
$p = 0.9$	2	2	6			
$s = 1$	3	2	4	7		
	4	2	3	5	8	
	5	1	2	4	6	9
$M = 10^2$	1	4.2				
$p = 0.5$	2	2.5	6.8			
$s = 0.1$	3	1.9	4.2	8.7		
	4	1.5	3.2	5.6	9.7	
	5	1.3	2.5	4.2	6.8	9.8
$M = 10^3$	1	4.325				
$p = 0.06$	2	2.58	6.80			
$s = 0.01$	3	1.96	4.325	8.72		
	4	1.64	3.19	5.66	9.94	
	5	1.43	2.58	4.325	6.8	9.98

Table 8: Knot placement for different censoring parameters and different number of knots. Notice that the expected interval censoring length is different for the three scenarios, but yet the knot locations are fairly similar.

11.1.2 Kaplan-Meier estimator

In this paper our focus is directed at parametric survival models. A common non-parametric estimator of the survival function is the Kaplan-Meier (KM) estimator[6]. We use the KM estimator in Section 9 to visualize the Tandmobiel data. The KM estimator will not be described in detail, since it is not central to the results of this thesis. In the case of non-censored failure times the KM-estimator is calculated as follows.

The failure times are divided into n intervals $t_i = [a_i, b_i),\ i = 1, \ldots, n$, such that at least one event occurs at a_i, the beginning of each interval. The KM-estimate of the survival curve then takes the form

$$\hat{S}(t) = \prod_{i;\ \min t_i < t} \left(1 - \frac{e_i}{r_i}\right),$$

where e_i are the number of events in time interval t_i, and r_i are the number of units at risk in the beginning of time interval t_i. In the case of interval censored observations, Turnbull presented an iterative algorithm which converges to the ML-estimate [15][14]. The KM and Turnbull estimators are implemented in `survfit.formula::survival` in R.

11.1.3 Multivariate delta method

The multivariate delta method[4] is a method of calculating the distribution of transformations of the mean of i.i.d, p-dimensional random variables. Let $\bar{\boldsymbol{Y}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$ be the mean of n i.i.d random variables $\boldsymbol{X}_i$ with expectation $\boldsymbol{\mu} < \infty$ and covariance matrix $\boldsymbol{\Sigma}$. Furthermore, let f be a function from $\mathbb{R}^p$ to $\mathbb{R}^q$ with $q \leq p$ which is continuously differentiable in a neighborhood of $\boldsymbol{\mu}$, with $p \times q$ Jacobian matrix $\boldsymbol{J}$. Then

$$\sqrt{n}\left(f(\boldsymbol{Y}_n) - f(\boldsymbol{\mu})\right) \overset{\text{apprx}}{\sim} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{J}^{\boldsymbol{t}}\boldsymbol{\Sigma}\boldsymbol{J}\right), \quad n \to \infty.$$

In this thesis we use the multivariate delta method to transform parameter estimates produced by different models to the same parameterization.

11.1.4 Kernel density estimator

The kernel density estimator is a non-parametric estimator of the density function some random variable. We introduce the concept of a kernel and take a look at the definition. A kernel function K is simply any non-negative function which satisfies

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

Given a kernel function K, n observations $x_i,\ i = 1, \ldots, n$ of some random variable X, and a bandwidth parameter $h > 0$, the kernel density estimator of the density f of X is

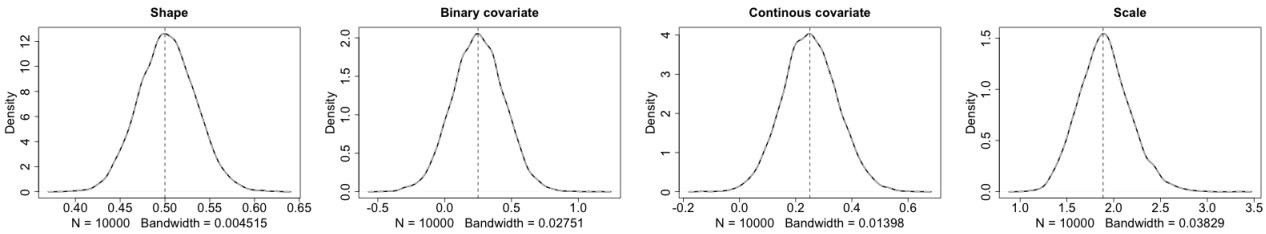
$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

There are many possible choices of kernel. In this paper, we chose the gaussian kernel ϕ.

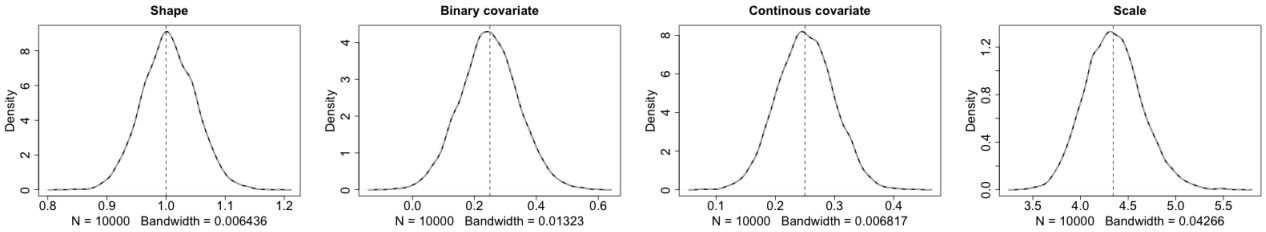
11.2 Plots and tables

		Shape			Binary Covariate			Continous Covariate			Scale		
Scenario	Implementation	Mean	SE	Coverage	Mean	SE	Coverage	Mean	SE	Coverage	Mean	SE	Coverage
1	stpm2	0.50	0.03	0.95	0.25	0.20	0.95	0.25	0.10	0.94	1.91	0.28	0.94
	survreg	0.50	0.03	0.95	0.25	0.20	0.95	0.25	0.10	0.94	1.91	0.28	0.94
2	stpm2	1.00	0.04	0.95	0.25	0.09	0.95	0.25	0.05	0.94	4.36	0.30	0.95
	survreg	1.00	0.04	0.95	0.25	0.09	0.95	0.25	0.05	0.94	4.36	0.30	0.95
3	stpm2	1.51	0.06	0.96	0.25	0.06	0.95	0.25	0.03	0.96	5.74	0.26	0.95
	survreg	1.51	0.06	0.96	0.25	0.06	0.95	0.25	0.03	0.96	5.74	0.26	0.95
4	stpm2	1.01	0.07	0.94	0.25	0.11	0.95	0.25	0.06	0.95	4.35	0.34	0.96
	survreg	1.01	0.07	0.94	0.25	0.11	0.95	0.25	0.06	0.95	4.35	0.34	0.96

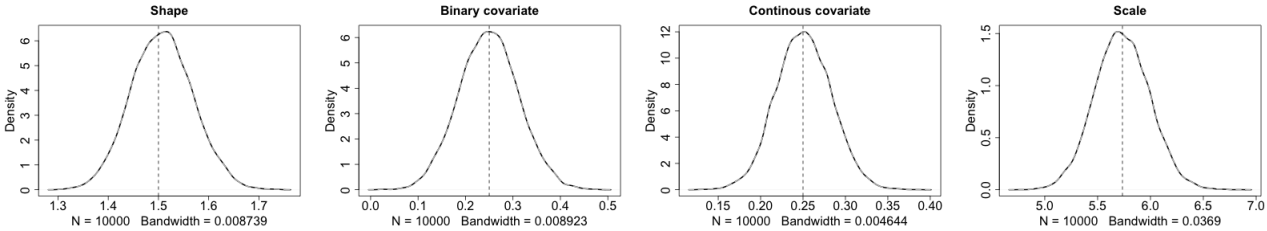
Table 9: Details on the distribution of the estimators of *stpm2* and *survreg* from the analysis of Section 7. Again, we notice that the two implementations produce identical results.



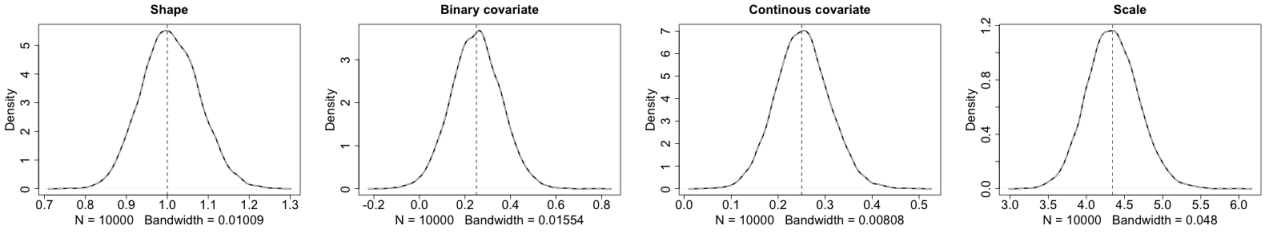
(a) Scenario 1: Shape 0.5 and probability 0.9



(b) Scenario 2: Shape 1 and probability 0.9

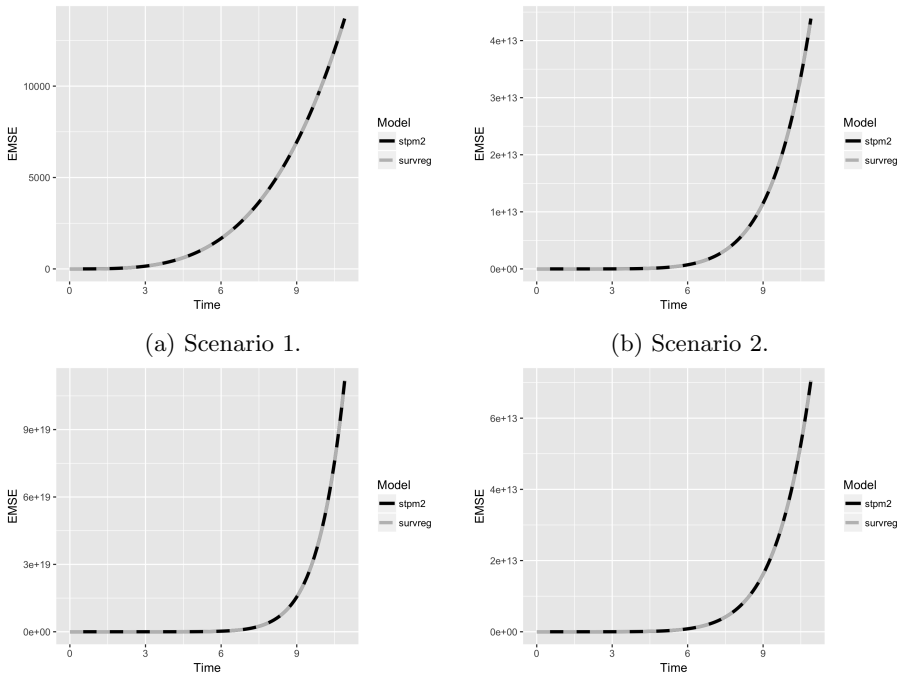


(c) Scenario 3: Shape 1.5 and probability 0.9



(d) Scenario 4: Shape 1 and probability 0.6

Figure 14: The kernel density estimations for scenarios and both models from section 7. Notice that they look very similar between the models. This is a good thing since it indicates that the models are implemented in a satisfactory manner. Theoretically, they should be the same.



(a) Scenario 1.

(b) Scenario 2.

(c) Scenario 3.

(d) Scenario 4.

Figure 15: The time dependent empirical square mean error (EMSE) plots for the four different parameter scenarios of Section 7.

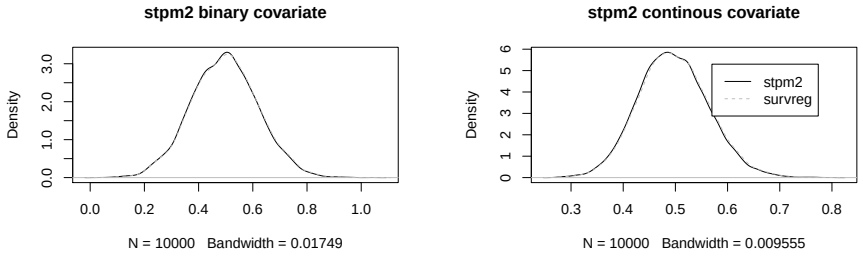


Figure 16: The kernel density estimates of the covariate parameter distributions of the two models from Section 8.

References

[1] Mark Clements and Xing-Rong Liu. *Generalized Survival Models*, 2017. version 1.4.1.

[2] D.R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1984.

[3] Guadalupe Gómez, M Luz Calle, Ramon Oller, and Klaus Langohr. Tutorial on methods for interval-censored data and their implementation in r. *Statistical Modelling*, 9(4):259–297, 2009.

[4] L. Held and D.S. Bove. *Applied Statistical Inference: Likelihood and Bayes*. Springer, 2013.

[5] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[6] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[7] David G Kleinbaum and Mitchel Klein. *Survival analysis*, volume 3. Springer, 2010.

[8] JF Lawless and Denise Babineau. Models for interval censoring and simulation-based inference for lifetime distributions. *Biometrika*, 93(3):671–686, 2006.

[9] Emmanuel Lesaffre, Arnošt Komárek, and Dominique Declerck. An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research*, 14(6):539–552, 2005.

[10] Xing-Rong Liu, Yudi Pawitan, and Mark Clements. Parametric and penalized generalized survival models. *Statistical methods in medical research*, page 0962280216664760, 2016.

[11] Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[12] Ramon Oller, Guadalupe Gómez, and M Luz Calle. Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics*, 32(3):315–326, 2004.

[13] Mark J Rutherford, Michael J Crowther, and Paul C Lambert. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation*, 85(4):777–793, 2015.

[14] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.

[15] Bruce W Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American statistical association*, 69(345):169–173, 1974.

[16] Jacques Vanobbergen, Luc Martens, Emmanuel Lesaffre, and Dominique Declerck. The signal-tandmobiel project a longitudinal intervention health promotion study in flanders (belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, 2:87–96, 2000.

[17] Svante Wold. Spline functions in data analysis. *Technometrics*, 16(1):1–11, 1974.

[18] Naji Younes and John Lachin. Link-based models for survival data with interval and continuous time censoring. *Biometrics*, pages 1199–1211, 1997.