Local Linear Smoothing to Estimate Accelerated Lifetime Model with Censoring and Truncation

Antonio Jesús López-Montoya
University of Granada
María Luz Gámiz-Pérez[1]
University of Granada
María Dolores Martínez-Miranda
University of Granada
Cass Business School, City University London

Abstract

A nonparametric Accelerated Failure Time (AFT) model is considered to evaluate the dependability measures of a water supply system laid in a city of the Mediterranean Sea. To do it, a set of breakdown data of sections of pipe of the system is available.

Unlike the usual methods that assume a parametric family for the underlying lifetime, we propose an AFT model with an unspecified distribution for the underlying lifetime.

To carry out the model fitting we suggest a two-stage procedure. Firstly, we estimate the influence of certain factors over the lifetime of the system. Secondly, we propose modern and flexible statistical tools based on counting processes to construct a smooth estimate the reliability measure of a specific system. We prove the good asymptotic properties of the estimator.

Simulation experiments show that the proposed methods have a good performance for finite samples and improve on the performance of other semi-parametric methods commonly used in practice.

Keywords: Semi-parametric; nonparametric; Survival Function; Local Linear; Water Supply System; Bootstrap.

1. Introduction

In Reliability Analysis, right-censoring and left-truncation are common features arising in lifetime datasets. Consequently, specific models and statistical procedures to analyze failure times data have been developed in the specialized literature. The proportional hazards (PH) model (Cox [1]) and the accelerated failure time (AFT) model (Lawless [2] and Nelson [3])

[1]Corresponding author. Address: Department of Statistics and O.R., Faculty of Sciences, University of Granada. 18071. Granada. Spain. E-mail: mgamiz@ugr.es. Tel.: +34958243156. Fax: +34958243267.

are frequently used in applied studies (see for example Srivastava and Mittal [4], Leea et al. [5] and Ismail [6]).

The main advantage of the PH model is that no additional assumptions about the baseline hazard function are needed to evaluate the effect of the covariates on the lifetime distribution. However, the basic PH assumption may not hold in many practical cases. In such situations, the AFT model has proven to be a convenient and intuitive alternative.

The AFT model establishes a direct relationship between the failure time and the covariates. The estimation of the model is usually carried out by assuming a parametric distribution for the lifetime. In this paper we refer to this model as parametric AFT. Several approaches have been proposed for the estimation of a parametric AFT model, see Lawless [2], Kalbfleisch and Prentice [7]. However, such parametric AFT model is very restrictive in most cases. As an alternative, semi-parametric models, where no assumptions are specified for the underlying survival distribution, can be more convenient in practice. Ritov [8] studied the general linear square estimation method, rank-based methods for censored data have been proposed by Tsiatis [9], Lai and Ying [10] or Jin et al. [11], and least squares based methods for censored data have been explored, for example, by Miller [12], Buckley and James [13] or Stute [14]. Stute [15, 16] considered a semi-parametric AFT model and introduced a procedure to estimate the regression coefficients under random censorship. Gross and Lai [17] developed a regression analysis under the presence of left-truncation in addition to right-censoring. Their approach is based on the estimation of a trimmed functional of the survival distribution, given that, with this sampling scheme, the lifetime is only observable within the range between the lower boundary of the support of the truncation variable and the upper boundary of the support of the censoring variable. It leads to estimators of the regression parameters which are relatively simple to obtain and useful to explore the relationship between the response variable and the covariate vector.

The interest in this paper goes further the estimation of the regression parameters since our practical motivation is to evaluate the dependability measures of a water supply system laid in a city of the Mediterranean Sea. So we aim to provide a reliable estimation of the probability of survival of a particular pipe beyond a specific period of time. To this goal we propose the use of nonparametric tools for evaluating the risk of failure in the water supply system. With the term "nonparametric" we mean that we do not consider any particular family of distributions for the underlying lifetime. In other words, a semi-parametric AFT that directly links the failure time of a pipe to its particular characteristics has been taken into account.

We follow a sequential procedure. Firstly, we consider the methods suggested by Gross and Lai [17] to estimate the parameters involved in the regression problem. These estimates are used to transform the data into the baseline scale of time. Then we conduct a nonparametric procedure to estimate the baseline survival function. Finally, a back transformation provides the estimator of the survival function for a specific subject.

The baseline survival function is estimated using a weighted-least-squares minimization approach that provides a local linear estimator. The estimator is closely related to the hazard estimator suggested by Nielsen and Tanggaard [18] and the density estimator by

Nielsen et al. [19]. In these papers, the authors consider a counting process formulation to represent survival data. The advantage of this formulation is that complicated truncation and/or censoring schemes may be easily incorporated to the model. We adopt this point of view in this paper since the dataset that we analyze consists of breakdown data of sections of pipe of the system, where left-truncation and right-censoring are present.

The rest of the paper is structured as follows. In Section 2 it is described the model under a counting process formulation which comprises the important cases of left-truncation or/and right-censoring. Our proposal is fully described in Section 3. The sequential procedure to derive the semi-parametric estimator of the survival function is introduced in Section 3.1. Section 3.2 describes the estimation of the regression coefficients in the semi-parametric model. Section 3.3 describes the nonparametric local linear estimator of the baseline survival. The asymptotic properties of the local linear estimator are derived in the Appendix. Section 4 shows a simulation study to report the performance of the proposed model and methods. The analysis of the water supply system dataset is described in Section 5. Section 6 concludes the paper.

2. The model

For $i = 1, \ldots, n$, let T_i denote the event time for the i-th subject. We assume that the subjects are independent. Define $N_i(t)$ as the number of events that have occurred on the i-th subject by time t in the absence of filtering. That is, $N_i(t) = I(T_i \leq t)$, where $I(\cdot)$ is the indicator function. Also define $Y_i(t)$ as the indicator function that takes the value 1 if the i-th individual is at risk at time t, which means that it has not failed and it is under observation. So, $Y^{(n)}(t) = \sum_{i=1}^{n} Y_i(t)$ counts the number of individuals at risk at time t and $\{Y^{(n)}(t), t \geq 0\}$ is called the *risk process*.

Suppose that the mean function of the counting process $N_i(t)$ associated with a k-vector of covariates X_i takes the form

$$E\left\{N_i(t)|X_i\right\} = Pr\left\{T_i \leq t|X_i\right\} = \Phi_0\left(t \exp\left(-\beta' X_i\right)\right), \tag{1}$$

where β is a k-vector of unknown regression parameters, and Φ_0 is an unspecified continuous function. If we write $T_{0,i} = T_i \exp\left(-\beta' X_i\right)$ and define $N_{0,i}(t) = I(T_{0,i} \leq t)$, then, clearly $N_{0,i}(t) = N_i\left(t \exp\left(\beta' X_i\right)\right)$, and also, under (1) we have that

$$E\left\{N_{0,i}(t)\right\} = Pr\left\{T_{0,i} \leq t\right\} = \Phi_0(t).$$

It means that the probability of failure by time t at the level $X_i = x$ equals the probability of failure by time $t \exp\left(-\beta' x\right)$ at the level $X_i = 0$. In other words, the set of covariates X_i affects the probability of occurrence of the failure by expanding or contracting the time scale on which this event occurs by a multiplicative factor $\exp\left(-\beta' x\right)$ relative to that of a zero-valued covariate vector, that is we have the following direct and intuitive relationship

$$T_i = \exp\left(\beta' X_i\right) T_{0,i},$$

or, equivalently

$$\log T_i = \beta' X_i + \epsilon_i, \tag{2}$$

where the error terms $\epsilon_i = \log T_{0,i}$, for $i = 1, \ldots, n$, have a common distribution with survival function S_0. This function is called the baseline survival and it represents the survival function of a subject at the zero-valued level of covariates.

The important situation of right-censoring and/or left-truncation can be described from the general counting process formulation given above just as particular cases. Many papers in the literature which are relevant to our purposes are described just in the case of right-censoring or both right-censoring and left-truncation (e.g. Stute [15, 16] and Gross and Lai [17]). To facilitate the reference to these previous papers we conclude this section by describing the data formulation commonly used in the case of right-censoring and left-truncation, and how it is actually incorporated by the more general counting process formulation provided above.

Let F denote the distribution function of the failure time T and let assume that model (2) holds. Right-censoring is one of the most common features of survival data, which means that the actual failure time T is not always observable and instead it is observed the random variable $Z = \min(T, C)$, where C denotes the random censoring variable (assumed to be independent of T). The variable $\delta = I(T \leq C) = I(Z = T)$ is an indicator of whether T has been observed or not. Additionally, in many prospective and retrospective survival studies, where the subjects under study are time followed-up, left-truncation in addition to right-censoring can arise. Left-truncation may occur if the time origin of the lifetime precedes the time origin of the observation period. Only those individuals that fail after the start of the study are being followed, otherwise they are left-truncated.

In the empirical analysis described in Section 5 some of the subjects that are followed in the study are susceptible to be right-censored during the follow-up period and left-truncation is present. In consequence, we consider a left-truncated and right-censored ($LTRC$) model. To represent this kind of data, the common formulation defines triplets of random variables (L, T, C), where T is the lifetime, L is the random left-truncation time and C is the random right-censoring time. Also it is usual to assume that L, T, and C are mutually independent. In the conditions of the $LTRC$ the data consist of a set of i.i.d. triplets $\{(L_1, Z_1, \delta_1), \ldots, (L_n, Z_n, \delta_n)\}$, with $L_i \leq Z_i$, for all $i = 1, \ldots, n$, as in case that $L > Z$ there is no observation. Let $\beta = Pr\{L > Z\}$ defined as the probability of being left-truncated, it is obvious that $\beta < 1$ in order to have non-empty sample sets.

Note that the $LTRC$ formulation actually corresponds to a particular case of the failure and risk processes introduced above. In fact, for each individual $i = 1, \ldots, n$, the corresponding failure process is given by $N_i(t) = I(Z_i \leq t)\,\delta_i$, and the risk process by $Y_i(t) = I(L_i \leq t \leq Z_i)$, for $t \geq 0$. So the counting process formulation assumed in this paper is indeed a more general formulation which incorporates the $LTRC$ case usually assumed in the literature of survival analysis.

3. The proposed methods

3.1. Sequential estimation

Let assume the counting process formulation described in Section 2. The equation

$$T_X = \exp(X\beta)\, T_0$$

defines the standard AFT model. From this equation we can assess the effect that the covariates have on the length of the lifetime of a particular subject. However, our interest goes further the assessment of these effects and we are concerned about the probability that a particular subject survives beyond a specified time t. In other words, we are interested in estimating the following function:

$$\begin{aligned} S_X(t) &= Pr\{T_X > t\} = Pr\{T_0 \exp(X\beta) > t\} = Pr\{T_0 > t\exp(-X\beta)\} \\ &= S_0(t\exp(-X\beta)), \end{aligned} \tag{3}$$

where S_0 is the baseline survival function. Note that $S_X(\cdot)$ is the survival function for a subject associated to level $X = x$, while $S_0(\cdot)$ represents the survival function of an individual with covariate level equal to 0.

To estimate the survival function in (3) we propose a sequential procedure. Firstly, we estimate the vector of regression coefficients β in the model (2) using the semi-parametric approach by Stute [15, 16] and Gross and Lai [17]. Let denote by $\widehat{\beta}$ such an estimator. This estimator is used to transform the data into the baseline scale of time as it is described in Section 2. Then, from the transformed data, we estimate the baseline survival function by a nonparametric estimator $\widehat{S}_0(\cdot)$. Specifically we consider the local linear approach by Nielsen and Tanggaard [18] and Nielsen et al. [19]. Finally, the survival function for a subject associated to level $X = x$ is estimated by

$$\widehat{S}_x(t) = \widehat{S}_0\left(t\exp(-x\widehat{\beta})\right). \tag{4}$$

A detailed description of these steps is provided in the following sections.

3.2. Estimation of the regression coefficients

In this section the influence of a set of covariates on the failure time is analyzed. To this goal we consider the regression model specified in (2) and follow the approach by Stute [15, 16] and Gross and Lai [17]. These papers assume the $LTRC$ formulation described in Section 2. Using the notation introduced in that section, let us consider a random sample of size n of the form $\{(L_i, Z_i, \delta_i, X_i);\ i = 1, \ldots, n\}$, with $L_i \leq Z_i$, $Z_i = \min\{T_i, C_i\}$ and $\delta_i = I(Z_i = T_i)$. Here $X_i = (X_{i1}, \ldots, X_{ik})'$ is a k-dimensional vector of covariates describing the i-th subject.

As we showed in Section 2 the above data formulation is a particular case of the general counting process model where the failure process is $N_i(t) = I(Z_i \leq t)\,\delta_i$, and the risk process

is $Y_i(t) = I\left(L_i \leq t \leq Z_i\right)$, for $t \geq 0$. Then, the cumulative hazard function can be estimated by

$$\widehat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{Y^{(n)}(s)} ds,$$

where $Y^{(n)}(s) = \sum_{i=1}^{n} Y_i(s)$, and thus the corresponding estimator for the survival function is obtained as the following product-limit estimator

$$\widehat{S}(t) = \prod_{s \leq t} \left(1 - d\widehat{\Lambda}(s)\right).$$

Computationally it results more tractable the following expression

$$\widehat{S}(t) = \prod_{T_i \leq t} \left(1 - \frac{\delta_i}{Y^{(n)}(T_i)}\right).$$

The last expression gives a step function that only has jumps at the failure times T_i whose sizes are given by

$$W_i = \widehat{S}(T_{i-1}) - \widehat{S}(T_i) = \prod_{j=1}^{i-1} \left(1 - \frac{\delta_j}{Y^{(n)}(T_j)}\right) \frac{\delta_i}{Y^{(n)}(T_i)}. \tag{5}$$

To derive the estimator of the vector of parameters β for right-censored data, Stute [15, 16] proposed a method that requires very general hypotheses and the estimation procedure is developed using weighted-least-squares. Under the model (2), the estimator of β can be obtained by minimizing

$$\sum_{i=1}^{n} W_i \left(\log Z_{(i)} - X_{[i]}\beta\right)^2, \tag{6}$$

where $\log Z_{(i)}$ is the i-th ordered value of the observed response variable $\log Z$, $X_{[i]}$ is the covariate associated with $\log Z_{(i)}$. When only right-censoring is considered in the sample, W_i are the Kaplan-Meier weights, which are obtained as the successive increments of the Kaplan-Meier estimator.

The minimization of (6) leads to the following estimator of β given by

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W} \log Z, \tag{7}$$

where $\log Z = \left(\log Z_{(1)}, \ldots, \log Z_{(n)}\right)'$, $\mathbf{W}$ is a diagonal matrix with the Kaplan-Meier weights, see Kaplan and Meier [20], and $\mathbf{X}$ is the $(n \times k)$-matrix with rows X_i $(i = 1, \ldots, n)$. Stute [15, 16] studied the consistency of this estimator and its asymptotic normal distribution and also proposed the use of a simpler jackknife estimator to calculate the asymptotic variance.

In a similar way, Gross and Lai [17] consider the regression problem when truncation is also present in the dataset. In this case, the weights are obtained as in the expression

(5). These authors proved that, under certain regularity conditions, the solution $\widehat{\beta}$ of the equation (6), given in (7), is strongly consistent and asymptotically normal. The authors also suggested a simple bootstrap method to estimate the standard error of the estimator $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)'$. The bootstrap algorithm consists of the following steps:

1. Generate B random samples of size n with replacement from the n observed quadruples $\{(L_i, Y_i, \delta_i, X_i);\ i = 1, \ldots, n\}$.
2. From each bootstrap sample ($b = 1, \ldots, B$) calculate a bootstrap estimator of the regression parameters β using the expression (7). Let denote such estimator by $\widehat{\beta}^{(b)}$.
3. Estimate the standard error $\widehat{\beta}_j$ ($j = 1, \ldots, k$) by:

$$s.e.(\widehat{\beta}_j) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\beta}_j^{(b)} - \overline{\beta}_j\right)^2},\ \ \forall\, j = 1, \ldots, k,$$

   with $\overline{\beta}_j = \dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\widehat{\beta}_j^{(b)}$.
4. Calculate the $(1-\alpha)\times 100\%$ confidence interval for $\widehat{\beta}_j$ ($j = 1, \ldots, k$) by:

$$\left(\widehat{\beta}_j \pm z_{\alpha/2}(n-1) \times s.e.(\widehat{\beta}_j)\right),$$

   where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

3.3. Nonparametric estimation of the baseline survival function

In the previous section we have described estimators for the coefficients of the *AFT* regression model. Now, to obtain the survival function for a subject associated to level $X = x$, defined in (4), the only thing that remains is to estimate the baseline survival function $S_0(\cdot)$. In the following we describe a nonparametric local linear estimator of such function. Without any loss of generality we can assume that numeric components of the covariate vector are centered at 0, that is $E[X] = 0$.

3.3.1. Transformation of the data

For simplicity let start by considering the *LTRC* data formulation. Thus, we consider a random sample of size n of the form $\{(L_i, Z_i, \delta_i, X_i);\ i = 1, \ldots, n\}$. In the previous section we have derived an estimator $\widehat{\beta}$ of the coefficients vector β in the *AFT* model given in (2). Using such estimator let now consider the transformation of the y axis given by $u = \exp(-\widehat{\beta}x)y$. So, we construct a new transformed dataset as follows:

$$\begin{cases} L_{0,i} = \exp(-\widehat{\beta}X_i)L_i \\ Z_{0,i} = \exp(-\widehat{\beta}X_i)Z_i \\ \delta_{0,i} = \delta_i \end{cases}$$

where $L_{0,i} \leq Z_{0,i}$ and $\exp(-\widehat{\beta}X_i) \geq 0$. Note that the dataset $\{(L_{0,i}, Z_{0,i}, \delta_{0,i});\ i = 1, \ldots, n\}$ is a *LTRC* sample from the baseline population which can be used to estimate $S_0(t) = Pr\{T_0 > t\}$.

Let now simplify the notation by using the general counting process formulation described in Section 2. For each subject $i = 1, \ldots, n$, let define the failure process by

$$N_{0,i}(t) = \begin{cases} 1; & \text{if } T_{0,i} \leq t \\ 0; & \text{otherwise} \end{cases},$$

and the risk process by

$$Y_{0,i}(t) = \begin{cases} 1; & \text{if } L_{0,i} \leq t \leq T_{0,i} \\ 0; & \text{otherwise} \end{cases}.$$

Note that the information in the dataset $\{(L_{0,i}, Z_{0,i}, \delta_{0,i});\ i = 1, \ldots, n\}$ is equivalent to the one contained in the dataset $\{(N_{0,i}(t), Y_{0,i}(t));\ t \geq 0, i = 1, \ldots, n\}$.

Using this transformed dataset, we derive in the next section a nonparametric estimator of the survival function S_0 using the local linear approach (Nielsen and Tanggaard [18], Nielsen et al. [19]).

3.3.2. The local linear estimator

Let $\widetilde{S}_0(t)$ be an empirical estimator of the survival function S_0, (e.g. extended Nelson-Aalen estimator) and K a probability density function. Let define $K_b(\cdot) = b^{-1}K(\cdot/b)$, with b being the bandwidth parameter ($b > 0$). The local linear approach by Nielsen and Tanggaard [18] and Nielsen et al. [19] provides an estimator of $S_0(t)$, as a result of the following weighted-least-squares minimization problem:

$$\underset{\theta_0,\theta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \int_0^\infty \left(\widetilde{S}_0(s) - \theta_0 - \theta_1(t-s)\right)^2 K_b(t-s)W(s)Y_i(s)ds.$$

Here $W(\cdot)$ denotes a general weighting function. Following recommendations by the authors we consider the natural weighting $W(s) = 1$. Let denote by $\widehat{\theta}_0$ and $\widehat{\theta}_1$ the minimizers of the above problem for $W(s) = 1$. Then, $S_0(t)$ is estimated by $\widehat{S}_0(t) = \widehat{\theta}_0$.

An explicit expression of $\widehat{\theta}_0$ can be derived by solving the following equations:

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n} \int_0^\infty \left(\widetilde{S}_0(s) - \theta_0 - \theta_1(t-s)\right) K_b(t-s)Y_i(s)ds \\ &= \int_0^\infty \widetilde{S}_0(s)K_b(t-s)Y^{(n)}(s)ds - \theta_0 \int_0^\infty K_b(t-s)Y^{(n)}(s)ds \\ &\quad -\theta_1 \int_0^\infty (t-s)K_b(t-s)Y^{(n)}(s)ds, \end{aligned}$$

$$
\begin{aligned}
0 &= -2\sum_{i=1}^{n}\int_0^\infty \left(\widetilde{S}(s)-\theta_0-\theta_1(t-s)\right)(t-s)K_b(t-s)Y_i(s)ds \\
&= \int_0^\infty \widetilde{S}(s)(t-s)K_b(t-s)Y^{(n)}(s)ds - \theta_0\int_0^\infty (t-s)K_b(t-s)Y^{(n)}(s)ds \\
&\quad -\theta_1\int_0^\infty (t-s)^2K_b(t-s)Y^{(n)}(s)ds.
\end{aligned}
$$

Let us consider the following definitions:

$$
a_j(t)=\int_0^\infty K_b(t-s)(t-s)^jY^{(n)}(s)ds, \quad \text{for } j=0,1,2
$$

and

$$
G_j(t)=\int_0^\infty \widetilde{S}_0(s)K_b(t-s)(t-s)^jY^{(n)}(s)ds, \quad \text{for } j=0,1.
$$

With these definitions, the above equations can be written as:

$$
\begin{aligned}
G_0(t) &= \theta_0a_0(t)+\theta_1a_1(t) \\
G_1(t) &= \theta_0a_1(t)+\theta_1a_2(t).
\end{aligned}
$$

By simple calculations we get the following expression for the solution in θ_0:

$$
\widehat{\theta}_0=\frac{a_2(t)G_0(t)-a_1(t)G_1(t)}{a_0(t)a_2(t)-a_1^2(t)}.
$$

Thus, the local linear survival estimator can be explicitly written by:

$$
\widehat{S}_0(t)=\sum_{i=1}^{n}\int_0^\infty \left(\frac{a_2(t)-a_1(t)(t-s)}{a_2(t)a_0(t)-a_1^2(t)}\right)K_b(t-s)\widetilde{S}_0(s)Y_i(s)ds,
$$

or equivalently by:

$$
\widehat{S}_0(t)=\sum_{i=1}^{n}\int_0^\infty \overline{K}_{b,t}(t-s)\widetilde{S}_0(s)Y_i(s)ds, \tag{8}
$$

using the local linear kernel function given by

$$
\overline{K}_{b,t}(t-s)=\frac{a_2(t)-a_1(t)(t-s)}{a_2(t)a_0(t)-a_1^2(t)}K_b(t-s).
$$

Notice that

$$\int_0^\infty \overline{K}_{t,b}(t-s)Y^{(n)}(s)ds = 1, \qquad \int_0^\infty \overline{K}_{t,b}(t-s)(t-s)Y^{(n)}(s)ds = 0,$$
$$\int_0^\infty \overline{K}_{t,b}(t-s)(t-s)^2Y^{(n)}(s)ds > 0.$$

In the appendix we prove a result which states the asymptotic pointwise normality of the local linear estimator (8).

4. Simulation study

In this section we evaluate the finite sample properties of the estimation procedure proposed in Section 3. The purpose is twofold. On the one hand, we are interested in the parametric part of the model, that is, we evaluate the accuracy of our estimates of the regression parameters and compare them with alternative methods. On the other hand, we focus on assessing the accuracy of the local linear estimator of the underlying distribution of the lifetime.

To achieve our first goal we compare the AFT model and the PH model in a scenario where the assumptions of both methods hold and the estimations obtained for the regression parameters are directly comparable. So, in our first comparison study, we do not pay attention to the underlying distribution family but we focus on the estimators of the regression parameters, which describe the effect of the covariates on the lifetime for the AFT case, and on the hazard function for the PH case. Besides, we measure the estimation error obtained with the two semi-parametric fits and compare these errors with the error obtained under a parametric approach. The computations have been performed using the statistical software environment R [21]. To implement the parametric AFT model we used the packages *survival* and *eha*. In particular, the parametric fit has been implemented by using maximum likelihood estimation procedures developed by the *aftreg* function included in the aforementioned *eha* package, see Broström [22].

Since the Weibull parametric regression model can be seen as a PH model as well as an AFT model, we have decided to use this family of distributions to compare the performance of the two approaches in a fair scenario. So, we have generated a Weibull regression model under different sampling schemes. Specifically we consider samples that present different rates of censoring and truncation.

To be more precise, let us define the PH model according to the following expression

$$\lambda(t;X) = \lambda_0(t)\exp\left(\gamma_1 X_1 + \gamma_2 X_2\right), \tag{9}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\gamma' = (\gamma_1, \gamma_2)$ is a vector of regression coefficients associated to the covariate vector $X' = (X_1, X_2)$. In terms of the AFT model we have that the logarithm of the lifetime of a subject with covariate vector $X' = (X_1, X_2)$ can be written as

$$\log T = \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{10}$$

where $\epsilon = \log T_0$, with T_0 being the lifetime of a subject at a *zero* level of covariates, that is, the baseline population, with $\beta_1 = -\lambda_1$, and $\beta_2 = -\lambda_2$.

To carry out the simulations we consider the same specifications as given in Orbe et al. [23]. So, we assume that X_1 has uniform distribution $U[0, 2]$, X_2 has distribution $U[3, 9]$, and, $\beta' = (\beta_1, \beta_2) = (1, 3)$.

For the baseline population, T_0, we consider a Weibull distribution with scale parameter $sc = 1$ and three different choices for shape parameter in order to account for decreasing hazard rate $sh = 0.5$, increasing hazard $sh = 5$, and constant hazard function $sh = 1$, which is an Exponential distribution. In this situation, both models PH and AFT are directly comparable (i.e. $\gamma = -\beta$).

The samples have been generated according to the following algorithm:

1. Generate a sample of failure times $\{T_1, \ldots, T_n\}$ from the Exponential regression model with $\beta' = (\beta_1, \beta_2) = (1, 3)$, $X_1 \in U[0, 2]$, $X_2 \in U[3, 9]$, and ϵ a standard minimum extreme value distribution.
2. Generate a sample of censoring times independently of the lifetimes, $\{C_1, \ldots, C_n\}$, from a Uniform distribution $U[0, \tau]$, where τ is properly chosen to get a desirable censoring rate $\alpha_1\%$ (i.e. 10%, 30% and 45%).
3. Define the censoring indicator by $\delta_i = I(T_i \leq C_i)$. Then, represent the i-th observation by the pair (Z_i, δ_i), with $Z_i = \min(T_i, C_i)$, for $i = 1, \ldots, n$.
4. Generate n left-truncation times $\{L_1, \ldots, L_n\}$ independently of censoring and lifetimes, from a Uniform distribution $U[0, \eta]$, where η is chosen to get a desirable truncation rate, α_2= 10%. Only those subjects for which $L_i \leq Z_i$ are kept in the sample whereas the others are discarded.

Each simulated sample consists of a sequence of triplets $\{(L_i, Z_i, \delta_i); i = 1, 2 \ldots, n^*\}$, where $L_i \leq Z_i$ for each $i = 1, \ldots, n^* \leq n$. And we have repeated the procedure up to $R = 2000$ times with $n = 50$, $n = 100$ and $n = 200$, for each combination of censoring and truncation levels.

Given a particular simulated sample, on the one hand we estimate the bias and standard error of the estimators of the regression parameters using the semi-parametric AFT model specified in (10) and compare them with the results obtained by using the PH model given in (9). On the other hand, we provide a measure of the error of the local linear estimation of the baseline survival distribution. In other words, for the estimations of the regression parameters ($\widehat{\beta}$ or $\widehat{\gamma}$) we have computed the value of the bias ($Bias$) by means of the expression $Bias(\widehat{\beta}) = E(\widehat{\beta}) - \beta$; and, the standard error (SE) by using the equation $SE(\widehat{\beta}) = \sqrt{E\left[\left(\widehat{\beta} - E(\widehat{\beta})\right)^2\right]}$. Finally, as a summary measure of the error estimation, we provide the mean square error (MSE), which is usually split into a bias term and a variance term as follows:

$$MSE(\widehat{\beta}) = SE^2(\widehat{\beta}) + Bias^2(\widehat{\beta}).$$

In order to assess the performance of the local linear estimator of the baseline survival, we consider the average square error (ASE). That is, we first estimate the regression parameters as it is explained in Section 3.2 and, second we obtain the survival function evaluated at the observed data times transformed to the baseline time-scale (see Section 3.3). To calculate the local linear estimator we have considered the Epanechnikov kernel, $K(t) = \frac{3}{4}(1-t^2)I(|t| \leq 1)$, and the bandwidth parameter has been chosen using the reference rule (see Gámiz et al. [28], pp. 44). The ASE for a particular sample of size n is thus defined by

$$ASE = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{S}_0\left(\exp\left(-\widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 x_{2,i}\right) t_i\right) - S_0\left(\exp\left(-\widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 x_{2,i}\right) t_i\right)\right]^2.$$

The simulation results are summarized in Tables A.1-A.3 and Figs. A.1-A.3. We deduce that higher levels of censoring tend to increase the variance of estimators, therefore the estimates lose precision as the censoring rate increases. In terms of the MSE, the AFT method provides more precise estimates than the PH model in almost all cases, according to Tables A.1-A.3. For the PH model we see from Table A.2 that the summary measure of the error gives values that are significantly higher than the ones reported by the AFT model. We have inspected the simulated samples and we have noticed that for a few simulated samples, the value of the estimated bias is extremely large which causes this high error measure. It could be that the use of another measure to summarize the error could result more convenient in this case to have a more representative picture of the actual situation (i.e. the median or perhaps a trimmed mean).

Besides, semi-parametric models are, as expected, less accurate when we estimate the parametric method, which uses the true probability distribution from which we have generated the data. However, this loss of accuracy is slight, in the sense of justifying the use of the semi-parametric AFT model when the probability distribution is unknown.

The ASE values for the proposed method are significantly smaller than the ones reported by the PH method in all cases considered. Notice that, as expected, the ASE values decrease with increasing sample sizes, see Figs. A.1-A.3.

5. Analysis of the breakdown dataset of a water supply network

5.1. The data and the analysis objectives

In this section we analyze failure data registered in a water supply network of a medium-sized city of the Spanish Mediterranean coast. The aim is to evaluate the lifetime of pipes using the semi-parametric approach described above. We consider the same dataset used by Carrión et al. [27] who evaluated the pipes failure probability using a Cox proportional hazards model. The full dataset comprises 26034 registers of pipes sections with information about the installation year of the pipes, the section length, the section diameter, the pipe material, the traffic conditions as well as the data of failure. According to the dataset, four types of different *material* have been employed: ductile cast iron, gray cast iron, polyethylene and asbestos cement. Also it has been considered three types of *traffic conditions* of the installation area of the pipes: under sidewalk, under normal traffic and heavy traffic.

In our analysis we do not consider the fact that a pipe section can fail more than once and, following the discussion by Carrión et al. [27], we have restricted the observation window to the interval of time between years 2000-2005. Note that this sampling scheme induces left-truncation and right-censoring and it is remarkable that the data present a heavy censoring rate, up to 98%.

We define the random variable $L = \max\{0,$ 2000-*installation date*$\}$ to represent the left-truncation time. It takes value 0 if the corresponding item is not left-truncated. The failure time is represented by the random variable $T =$ *registered failure date - installation date*. And the censoring time by the random variable $C =$ 2006 - *installation date*. All dates are given in years. We assume that L, T and C are mutually independent, non negative and the censoring scheme is non informative, see Fleming and Harrington [24]. Let G, F and H denote the cumulative distribution functions (c.d.f.) of L, T and C, respectively.

Under the above sampling specifications our objective is to estimate the lifetime of the water main. To this goal we start by fitting a Cox PH model and checking whether the PH assumption holds. Then we calculate the semi-parametric AFT approach described in Section 3.

5.2. Checking the proportional hazards assumption

As a first approach the PH model is proposed to analyze the time period until failure occurs in a particular pipe in the network. It is identified in terms of a vector of covariates that consists of the following:

Ductile CI: Type of material (1=ductile cast iron, 0=others).

Gray CI: Type of material (1=gray cast iron, 0=others).

Polyethylene: Type of material (1=polyethylene, 0=others).

Sidewalk: Level of stress of road traffic (1=sidewalk, 0=others).

Normal: Level of stress of road traffic (1=normal, 0=others).

Length: The length of the pipe in (m).

Diameter: The diameter of the section pipe in (mm).

The main hypothesis in the PH model is proportionality of the hazard rates. If the proportional hazards assumption holds, the logarithms of the cumulative hazard functions are expected to describe parallel curves. To fit this model to the data we have used the *cox.zph* function, included in the R-package *survival*, see Therneau and Lumley [25]. Tests and graphical diagnostics of the PH model are based on the scaled Schoendfeld residuals, see Therneau and Grambsch [26].

A visual inspection reveals that the corresponding curves provided by the *material*, *traffic* and *diameter* covariates do not seem to be parallel, as can be deduced from Fig. A.4. In consequence, the PH model is questionable in this study. In addition, we use the Schoenfeld

residual test to verify the *PH* assumption. The results from the test are displayed in Table A.4. If the p-value reported by the Schoenfeld residual test is below a significant level 0.05, we conclude that *PH* assumption does not hold. The results suggest that the *PH* assumption is not acceptable for the covariates *material* and *traffic*. Besides, the global test for *PH* assumption leads to similar conclusions, as the p-value is 0.0005. The graphical diagnostics displayed in Fig. A.5 show that the covariates *material* and *traffic* do not hold the *PH* assumption.

In this context, one way of accommodating non-proportional hazards may be to consider time-dependent covariates in the formulation of the regression model. As an alternative, it can also result adequate to consider a stratified Cox model according to the levels of the covariates *material* and/or *traffic*. This strategy has been carried out by Carrión et al. [27]. However, this extended Cox model has several disadvantages as the loss of the ability to quantify the effects of these stratified covariates, which in some cases might be of direct interest, as it is in our dataset the case of the *material* and *traffic* covariates.

An additional issue that arises is the non-linearity of the model. That is, an incorrectly specified functional form in the parametric part of the model. To detect non-linearity the martingale residuals may be plotted against covariates. To this goal, we examine the plots of the martingale residuals against *Length* and *Diameter* since non-linearity is not an issue concerning dichotomous covariates. The residual plots are presented in Fig. A.6. As in the plots of Schoenfeld residuals, smoothing is also helpful to get a clear idea of the picture. The smoothed curves displayed in Fig. A.6 are produced by local linear regression (using the *lowess* function of R software). The graphs in the figures suggest that the hypothesis of linearity is reasonable.

5.3. The proposed approach

According to the study in the previous section, *PH* is not adequate for our data so we propose an alternative procedure. In this section, we describe the results of the semi-parametric AFT model. The estimation is performed in two steps as it is described in Section 3.1. In the first step we calculate the estimates of the regression coefficients in the model (2) without assuming any parametric family for the underlying distribution. The second step consists of estimating nonparametrically the baseline survival function.

The estimated coefficients and the corresponding standard errors are presented in Table A.5. The reported standard errors have been calculated using the bootstrap method suggested by Gross and Lai [17]. A full description of the bootstrap algorithm is provided in Section 3.2. The number of simulations considered in the algorithm is $B = 10000$.

The regression coefficients of the semi-parametric AFT model are given by $\exp(\beta)$. Note that the categories considered as reference levels, i.e. *asbestos cement* and *heavy traffic*, for the qualitative covariates *material* and *traffic*, respectively, are not displayed in Table A.5.

The coefficients can be interpreted for each covariate as follows:

Ductile CI: the failure time corresponding to a pipe made of ductile cast iron is expected about 29.75% shorter than others.

Gray CI: the failure time corresponding to a pipe made of gray cast iron is expected about 51.20% longer than others.

Polyethylene: the failure time corresponding to a pipe made of polyethylene is expected about 52.53% shorter than others.

Sidewalk: under a sidewalk traffic the failure time increases by 245.29%.

Normal: under a normal traffic the failure time increases by 366.90%.

Length: the increase in failure time for an increase of 1 m is 29.62%.

Diameter: the decrease in failure time for an increase of 1 mm is 23.86%.

We consider that the reference (baseline) population is determined by the category *asbestos cement* for the covariate *material* and *heavy traffic* for the covariate *traffic*. For the numeric covariates, which are *Length* and *Diameter*, we take the mean values as reference levels. To estimate the baseline survival function we have considered the local linear estimator proposed in Section 3.3, equation (8). As in Section 4 we have calculated the estimator using the Epanechnikov kernel and the bandwidth parameter has been chosen using the reference rule (following the suggestions by Gámiz et al. [28], pp. 44), which provides the value $b = 0.0707$. The resulting local linear estimator of the baseline survival function is shown in Fig. A.7.

5.4. Discussion

When the *PH* assumption does not hold, the results obtained from the Cox *PH* model lead us to wrong conclusions. On the contrary, our semi-parametric *AFT* approach provides more reliable results.

Inadequate parametric assumptions can lead to wrong conclusions derived from wrong specifications of the probability distribution of the failure time. Thus, if the probability distribution is unknown, that is, the most usual situation in practice, it seems to be more convenient to consider semi-parametric approaches.

The methods described in the paper allow us to evaluate the impact of different covariates on the lifetime of pipes of the analyzed water supply system. In fact, the analysis described above shows that pipes which were more prone to failure had the following characteristics: large length, small diameter, with ductile cast iron or polyethylene material and placed under heavy traffic. The results agree with the derived by previous works that analyze similar datasets (see for example the recent paper by Carrión et al. [27]).

6. Conclusions

We propose a practical strategy that allows us to estimate an *AFT* model under a nonparametric approach. If the *PH* assumption of Cox's model holds, the semi-parametric *AFT* approach presented here performs better in terms of standard errors and subsequently in their total mean squared errors. In the case that the *PH* assumption cannot be hold, the

semi-parametric AFT is a very convenient alternative to the traditionally used Cox model, as it is the case of the practical application we present in Section 6. Thus, we use our procedure to understand the influence of some variables on the reliability of pipes in a water supply network.

Through simulation experiments, the semi-parametric AFT proposed has been also compared with the parametric AFT model based on the correct probability distribution of errors. As expected, the semi-parametric proposed is less accurate, although this loss of accuracy is slight. Besides, this is compensated since in the semi-parametric approach is not necessary to consider any parametric assumptions about the underlying failure time.

Furthermore, we have compared the baseline survival functions for both models and we confirm the greater effectiveness of the semi-parametric AFT approach in contrast to the PH model.

The insights provided in this paper can be very valuable when making decisions concerning the design and construction of water supply networks. The conclusions from the performed data analysis are valid not only for the particular dataset studied here but also (to some extent) for any other water supply systems.

Appendix A. Asymptotic theory

In this section we derive the asymptotic properties of the local linear estimator presented in Section 3.3.2. For simplicity in the notation we write the baseline survival function $S_0(\cdot) = S(\cdot)$ and we denote the local linear estimator by $\widehat{S}(t)$. This estimator can be written as

$$\widehat{S}(t) = \int_0^\infty \overline{K}_{b,t}(t-s)\widetilde{S}(s)Y^{(n)}(s)ds,$$

just by substituting $Y^{(n)}(s) = \sum_{i=1}^n Y_i(s)$.

Following Nielsen and Tanggaard [18], the error $\widehat{S}(t) - S(t)$ can be split in two terms as

$$\widehat{S}(t) - S(t) = \widehat{S}(t) - S^*(t) + S^*(t) - S(t) = V(t) + B(t),$$

where we define

$$S^*(t) = \int_0^\infty \overline{K}_{b,t}(t-s)S(s)Y^{(n)}(s)ds,$$

so that $V(t)$ is the variable part converging to a normal distribution, and $B(t)$ is the stable part converging in probability to a constant.

Firstly, we study the stable part of the error term, $B(t)$:

$$B(t) = S^*(t) - S(t) = \int_0^\infty \overline{K}_{b,t}(t-s)S(s)Y^{(n)}(s)ds - S(t),$$

given that $\int_0^\infty \overline{K}_{t,b}(t-s)Y^{(n)}(s)ds = 1$ we express the expression above of the form

$$B(t) = \int_0^\infty \overline{K}_{t,b}(t-s)\left[S(s) - S(t)\right]Y^{(n)}(s)ds.$$

We assume that $S \in C^2(0,\infty)$, and via Taylor expansion we have:

$$B(t) = \int_0^\infty \overline{K}_{t,b}(t-s)\left[S(t) + S^{'}(t)(s-t) + S^{''}(t)\frac{(s-t)^2}{2} + O_P\left((s-t)^2\right) - S(t)\right] Y^{(n)}(s)ds,$$

as $\int_0^\infty \overline{K}_{t,b}(t-s)(t-s)Y^{(n)}(s)ds = 0$, since $\overline{K}_{t,b}$ can be interpreted as a second-order kernel with respect to the measure μ, where $d\mu(s) = Y^{(n)}(s)ds$ (see Nielsen et al. [19]). Therefore,

$$B(t) = \int_0^\infty \overline{K}_{t,b}(t-s)\left[S^{''}(t)\frac{(s-t)^2}{2} + O_P\left((s-t)^2\right)\right] Y^{(n)}(s)ds.$$

As proven in Nielsen and Tanggaard [18], the stochastic local linear kernel $\overline{K}_{t,b}(t-s)$ is asymptotically equivalent to the kernel $K_b(t-s)\{Y^{(n)}(s)\}^{-1}$, where $K_b(t-s) = b^{-1}K\left((t-s)b^{-1}\right)$. Defining the second moment $\mu_2(K) = \int u^2K(u)du$, we conclude that

$$B(t) = b^2\frac{1}{2}S^{''}(t)\mu_2(K) + O_P(b^2).$$

Secondly, we study the variable part of the error term $V(t)$:

$$V(t) = \widehat{S}(t) - S^*(t) = \int_0^\infty \overline{K}_{t,b}(t-s)[\widetilde{S}(s) - S(s)]Y^{(n)}(s)ds.$$

Furthermore we use the following approximation

$$\begin{aligned}
\widetilde{S}(s) &= \exp\left(-\int_0^s \frac{dN^{(n)}(u)}{Y^{(n)}(u)}\right) = \exp\left(-\widetilde{\Lambda}(s)\right) \approx 1 - \widetilde{\Lambda}(s),\\
S(s) &= \exp\left(-\int_0^s \alpha(u)du\right) = \exp\left(-\Lambda(s)\right) \approx 1 - \Lambda(s).
\end{aligned}$$

Here $\alpha(\cdot)$ and $\Lambda(\cdot)$ denote the hazard and the cumulative hazard functions, respectively. Replacing $\widetilde{S}(s)$ and $S(s)$ into the expression $V(t)$ we have

$$\begin{aligned}
V(t) &= \int_0^\infty \overline{K}_{t,b}(t-s)\left[1 - \widetilde{\Lambda}(s) - (1-\Lambda(s))\right] Y^{(n)}(s)ds\\
&= \int_0^\infty \overline{K}_{t,b}(t-s)\left[-\widetilde{\Lambda}(s) + \Lambda(s)\right] Y^{(n)}(s)ds\\
&= -\int_0^\infty \overline{K}_{t,b}(t-s)\left[\int_0^s \left(\frac{dN^{(n)}(u)}{Y^{(n)}(u)} - \alpha(u)du\right)\right] Y^{(n)}(s)ds\\
&= -\int_0^\infty K_b(t-s)\left(\int_0^s \frac{dN^{(n)}(u) - \alpha(u)Y^{(u)}(u)du}{Y^{(u)}(u)}\right) ds\\
&= -\int_0^\infty K_b(t-s)\left(\int_0^s \frac{dM^{(n)}(u)}{Y^{(n)}(u)}\right) ds,
\end{aligned}$$

where $M^{(n)}(t) = N^{(n)}(t) - \alpha(t)Y^{(n)}(t)$ is a squared integrable local martingale, and again we use the asymptotic equivalence between the kernels. We can exchange the order of the integrals in the above expression and, after a convenient change of variable, we get

$$\begin{aligned} V(t) &= -\int_0^\infty \left(\int_u^\infty K_b(t-s)ds \right) \frac{dM^{(n)}(u)}{Y^{(n)}(u)} \\ &= \int_0^\infty \left(-\mathbf{K}\left(\frac{t-u}{b} \right) \right) \frac{dM^{(n)}(u)}{Y^{(n)}(u)}, \end{aligned}$$

where $\mathbf{K}(x) = \int_{-\infty}^{x} K(v)dv$, and hence we can proceed following from standard martingale theory (see Nielsen and Tanggaard [18]). To obtain the limiting distribution we only need to prove, for some sequence $a_n \to \infty$ and some positive function $C(t)$ (see Ramlau-Hansen [29]), that

$$a_n \langle V \rangle (t) \xrightarrow{P} C(t),$$

for $n \to \infty$, where

$$\langle V \rangle (t) = \int_0^\infty \left(\mathbf{K}\left(\frac{t-u}{b} \right) \right)^2 \frac{1}{Y^{(n)}(u)} \alpha(u)du.$$

Under the usual assumption $\left| Y^{(n)}(t)/n - \gamma(t) \right| \xrightarrow{P} 0$, where γ is a positive function such that $\gamma \in C^1(0,\infty)$, we obtain that the previous integral can be approximated, after some easy computations, by the expression

$$\frac{b}{n\gamma(t)} \alpha(t) \left(\int_{-\infty}^{t/b} \left\{ \mathbf{K}(v) \right\}^2 dv \right), \tag{A.1}$$

where we have used the continuity of functions γ and α.

In most cases the kernel function K is a symmetric density function with domain the interval $[-1,1]$, so that $\mathbf{K}$ denotes the corresponding cumulative distribution function, and then $(A.1)$ can be written

$$\frac{b}{n\gamma(t)} \alpha(t) \left(\int_{-1}^{1} \left(\mathbf{K}(v) \right)^2 dv + \int_1^{t/b} dv \right) = \frac{b}{n\gamma(t)} \alpha(t) \left(R\left(\mathbf{K} \right) + \frac{t}{b} - 1 \right), \tag{A.2}$$

where $R\left(\mathbf{K} \right) = \int_{-1}^{1} \left(\mathbf{K}(v) \right)^2 dv$ is a constant that only depends on the kernel. Equation $(A.2)$ can be written as

$$\frac{\alpha(t)}{\gamma(t)} \left(\frac{R\left(\mathbf{K} \right) - 1}{nb^{-1}} + \frac{t}{n} \right). \tag{A.3}$$

The leading term of the expression $(A.3)$ is of order $O_P\left(n^{-1} \right)$, so we can approximate it by

$$\frac{\alpha(t)}{\gamma(t)} \frac{t}{n} + O_P\left(n^{-1} \right).$$

All the above arguments prove the asymptotic pointwise normality of the local linear estimator. The following theorem summarizes the above derivations.

Theorem 1. *Let assume the following conditions:*

(A.1) The bandwidth $b \to 0$ *and* $nb \to \infty$.

(A.2) There exist a positive function γ *such that* $\left|Y^{(n)}(t)/n - \gamma(t)\right| \xrightarrow{P} 0$.

(A.3) $S \in C^2(0, \infty)$.

(A.4) α, $\gamma \in C^1(0, \infty)$.

Then,

$$B(t) = b^2 \frac{1}{2} S''(t) \mu_2(K) + O_P(b^2),$$

and

$$n^{-1/2} V(t) \xrightarrow{D} N\left(0, \sigma(t)\right),$$

where $\sigma(t) = [t\alpha(t)]/\gamma(t)$.

Note that this result is in accordance to the discussion in Chapter 1 of Gámiz *et al.* [28]. It states that the order of convergence of variance of the local linear estimator agrees with the one of the empirical cumulative distribution function, that is, $O_P\left(n^{-1}\right)$. However, from $(A.3)$, we also get here that the variance has an additional term of a lower order, i.e. $O_p(bn^{-1})$, and this term, which depends only on the kernel, can be negative. In particular for the most usual kernel function, the Epanechnikov kernel, we obtain that $R\left(\mathbf{K}\right) - 1 = -0.2571429$. This means that local linear estimator presented here provides an improvement in efficiency estimation.

Acknowledgments

We sincerely thanks Dr. A. Carrión, (Universidad Politécnica de Valencia, Spain) for providing the dataset. This work has been financial supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) by the grants MTM2008-03010 and MTM2013-41383-P; and "Junta de Andalucía" by the grant TIC-06902.

References

[1] D.R. Cox, Regression models and life-tables (with discusion), J. Roy. Stat. Soc. B. 34 (1972) 187 – 220.

[2] J.F. Lawless, Statistical Models and Methods for Lifetime Data Analysis, John Wiley & Sons, New York, 1982.

[3] W. Nelson, Accelerated Testing: Statistical Models, Test Plans and Data Analyses, John Wiley & Sons, New York, 1990.

[4] P.W. Srivastava, N. Mittal, Optimum step-stress partially accelerated life tests for the truncated logistic distribution with censoring, Appl. Math. Mod. 34 (2010) 3166 – 3178.

[5] P.S. Leea, C.C. Torngb, Y.C. Linb, Determination of the optimal accelerated burn-in time under Arrhenius-Lognormal distribution assumption, Appl. Math. Mod. 35 (2011) 4023 – 4030.

[6] A.A. Ismail, Estimating the parameters of Weibull distribution and the acceleration factor from hybrid partially accelerated life test, Appl. Math. Mod. 36 (2012) 2920 – 2925.

[7] J.D. Kalbfleisch, R.L. Prentice, The Statistical Analysis of Failure Time Data, John Wiley & Sons, New York, 2002.

[8] Y. Ritov, Estimation in a linear regression model with censored data, Ann. Stat. 18 (1990) 303 – 328.

[9] A.A. Tsiatis, Estimating regression parameters using linear rank tests for censored data, Ann. Stat. 18 (1990) 354 – 372.

[10] T.L. Lai, Z. Ying, Rank Regression Methods for Left-Truncated and Right-Censored Data, Ann. Stat. 19 (1991) 505 – 1108.

[11] Z. Jin, D. Lin, L.J. Wei, Z. Ying, Rank-based inference for the accelerated failure time model, Biometrika. 90 (2003) 341 – 353.

[12] R.G. Miller, Least squares regression with censored data, Biometrika. 63 (1976) 449 – 464.

[13] J.J. Buckley, I.R. James, Linear regression with censored data, Biometrika. 66 (1979) 429 – 436.

[14] W. Stute, Consistent estimation under random censorship when covariables are present, J. Multivariate. Anal. 45 (1993) 89 – 103.

[15] W. Stute, Distributional convergence under random censorship when covariables are present, Scand. J. Stat. 23 (1996) 461 – 471.

[16] W. Stute, The jack-knife estimate of variance of a Kaplan-Meier integral. Ann. Stat. 24 (1996) 2679 – 2704.

[17] S.T. Gross, T.L. Lai, Bootstrap Methods for Truncated and Censored Data, Stat. Sinica. 6 (1996) 509 – 530.

[18] J.P. Nielsen, C. Tanggaard, Boundary and bias correction in kernel hazard estimation, Scand. J. Stat. 28 (2001) 675 – 698.

[19] J.P Nielsen, C. Tanggaard, M.C. Jones, Local linear density estimation for filtered survival data, Statistics. 43 (2009) 176 – 186.

[20] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Am. Stat. Assoc. 53 (1958) 457 – 481.

[21] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2011. URL http://www.R-project.org.

[22] G. Broström, Event History Analysis with R, Chapman & Hall/CRC The R Series, United States, 2012.

[23] J. Orbe, E. Ferreira, V. Núñez-Antón, Comparing proportional hazards and accelerated failure time models for survival analysis, Stat. Med. 21 (2002) 3493 – 3510.

[24] T. Fleming, D. Harriton, Counting Processes and Survival Analysis, John Wiley & Sons, New York, 2002.

[25] T.M. Therneau, T. Lumley, Survival analysis, including penalised likelihood, R package version 2.34-1, 2008.

[26] T.M. Therneau, P.M. Grambsch, Modeling Survival Data: Extending the Cox Model, Springer-Verlag, New York, 2000.

[27] A. Carrión, H. Solano, M.L. Gámiz, A. Debón, Evaluation of the reliability of a water supply network from right-censored and left-truncated break data, Water Resour. Manag. 24 (2010) 2917 – 2935.

[28] M.L. Gámiz, K.B. Kulasekera, N. Limnios, B.H. Lindquist, Applied Nonparametric Statistic in Reliability, Springer Series in Reliability Engineering, London, New York, 2011.

[29] H. Ramlau-Hansen, Smoothing counting process intensities by means of kernel functions, Ann. Stat. 11 (1983) 453 – 466.

Table A.1: MSE for PH, AFT and Exponential regression models.

Sample	Censorship 10%			Censorship 30%			Censorship 45%		
size	PH	AFT	Exp	PH	AFT	Exp	PH	AFT	Exp
$n = 50$	0.4750	0.1368	0.0975	0.7336	0.1763	0.1335	1.2125	0.2458	0.2122
$n = 100$	0.1715	0.0770	0.0461	0.2085	0.0772	0.0545	0.3116	0.1197	0.1330
$n = 200$	0.0689	0.0444	0.0205	0.0932	0.0503	0.0277	0.1403	0.0671	0.0393

Table A.2: MSE for AFT and Weibull(1, 5) regression models.

Sample	Censorship 10%			Censorship 30%			Censorship 45%		
size	PH	AFT	Weib	PH	AFT	Weib	PH	AFT	Weib
$n = 50$	0.6472	0.0048	0.0036	0.9302	0.0072	0.0057	1.3221	0.0108	0.0082
$n = 100$	0.2014	0.0028	0.0018	0.2282	0.0033	0.0022	0.3523	0.0048	0.0032
$n = 200$	0.0656	0.0016	0.0008	0.0852	0.0018	0.0010	0.1379	0.0027	0.0014

Table A.3: MSE for AFT and Weibull(1, 0.5) regression models.

Sample	Censorship 10%			Censorship 30%			Censorship 45%		
size	PH	AFT	Weib	PH	AFT	Weib	PH	AFT	Weib
$n = 50$	0.7062	0.5983	0.3682	1.1566	0.8357	0.5627	1.6963	0.9948	0.8012
$n = 100$	0.3223	0.3712	0.1828	0.4365	0.3959	0.2509	0.5625	0.5001	0.3247
$n = 200$	0.1384	0.2166	0.0840	0.1909	0.2309	0.1149	0.2676	0.3046	0.1595

Table A.4: Testing results about the proportional hazards assumption.

Covariate	ρ	χ^2	$p-value$
Ductile CI	0.1110	10.5	0.0012
Gray CI	0.0007	0.0003	0.9859
Polyethylene	0.0396	0.983	0.3214
Sidewalk	0.1206	8.86	0.0029
Normal	0.1388	11.8	0.0005
Length	0.0395	0.529	0.4670
Diameter	0.0583	1.88	0.1707
Global		25.7	0.0005

Table A.5: Estimated regression coefficients using a semi-parametric AFT model.

Covariate	$\widehat{\beta}$	$\exp(\widehat{\beta})$	$se(\widehat{\beta})$	$95\%\, bootstrap\, CI$
Ductile CI	-0.3530	0.7025	0.0974	(-0.5439, -0.1621)
Gray CI	0.4134	1.5120	0.1058	(0.2059, 0.6209)
Polyethylene	-0.7449	0.4747	0.1838	(-1.1052, -0.3845)
Sidewalk	1.2392	3.4529	0.5466	(0.1677, 2.3107)
Normal	1.5409	4.6690	0.5058	(0.5495, 2.5323)
Length	0.2594	1.2962	0.0727	(0.1169, 0.4019)
Diameter	-0.2725	0.7614	0.0619	(-0.3939, -0.1511)

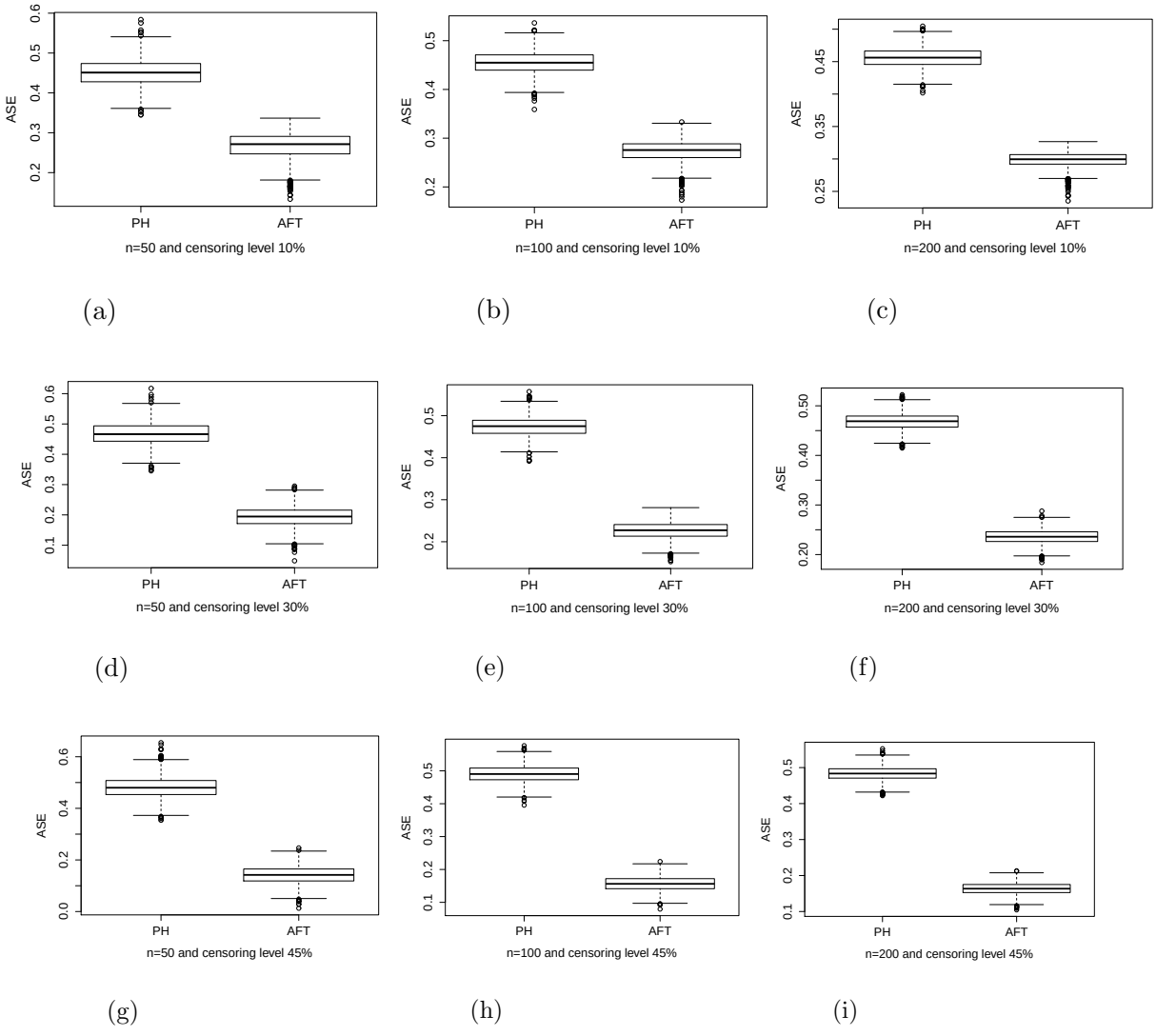


Figure A.1: **Case 1. Exponential Model.** Box-plots of the ASE estimations for the *PH* and *AFT* models $n = 50$, 100 and 200 with truncation level $\alpha_2 = 10\%$ and a censoring level $\alpha_1 = 10\%$, 30% and 45%.

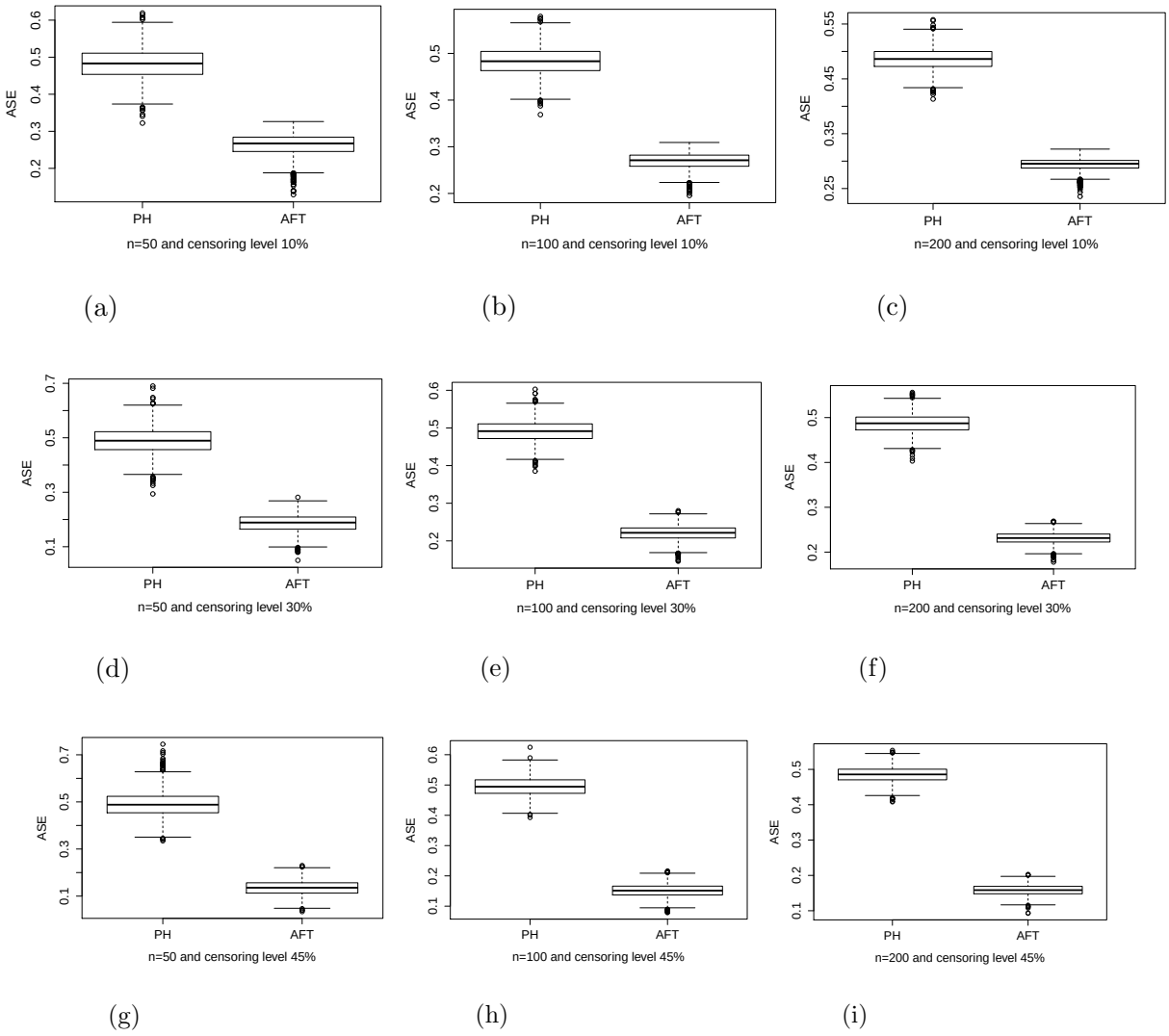


Figure A.2: **Case 2. Model Weibull with shape parameter 5.** Box-plots of the ASE estimations for the PH and AFT models $n = 50$, 100 and 200 with truncation level $\alpha_2 = 10\%$ and a censoring level $\alpha_1 = 10\%$, 30% and 45%.

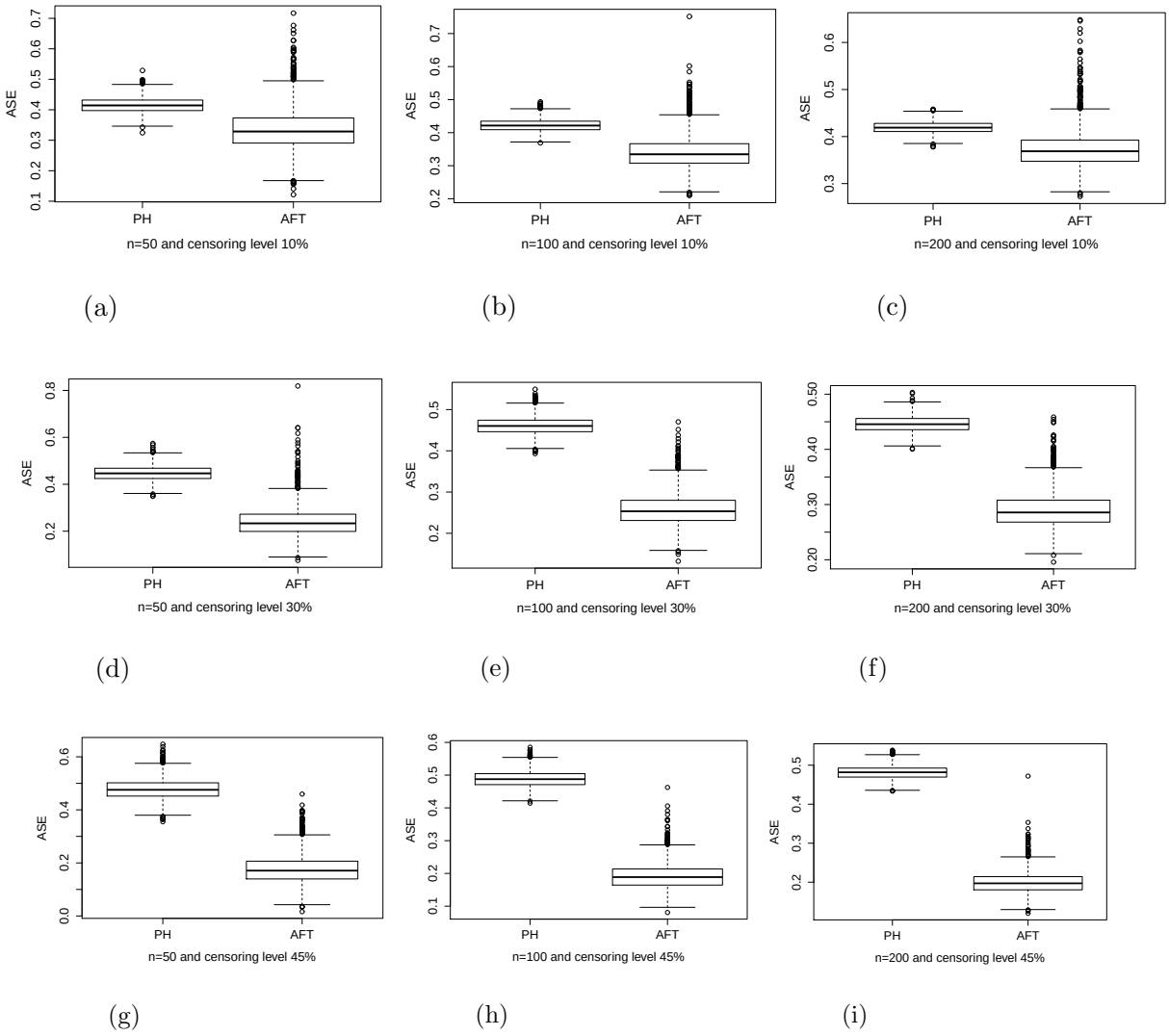
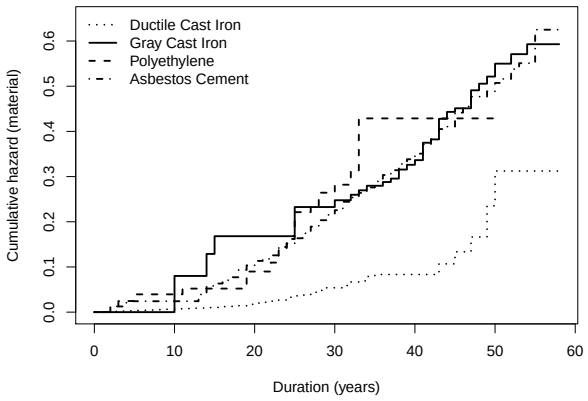
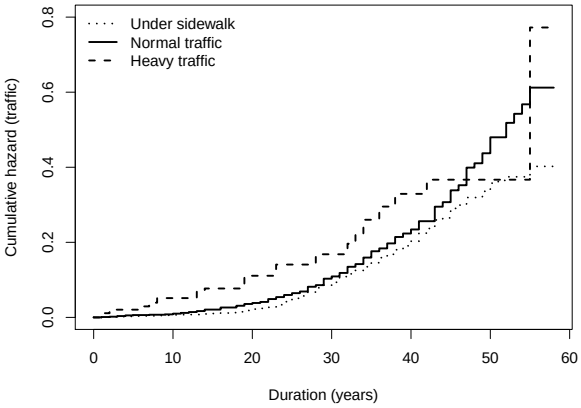


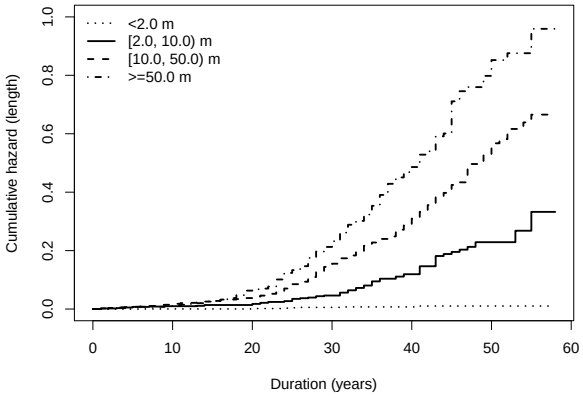
Figure A.3: **Case 3. Model Weibull with shape parameter 0.5.** Box-plots of the ASE estimations for the *PH* and *AFT* models $n = 50$, 100 and 200 with truncation level $\alpha_2 = 10\%$ and a censoring level $\alpha_1 = 10\%$, 30% and 45%.



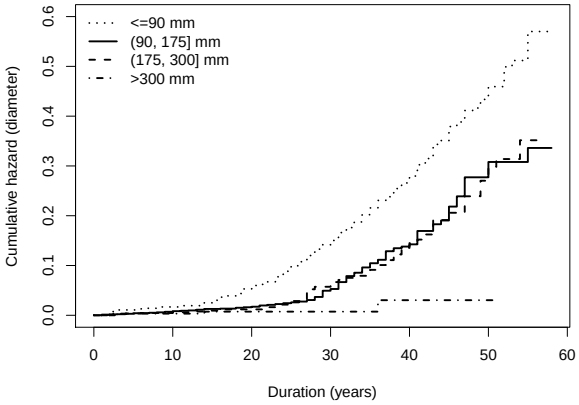
(a)



(b)

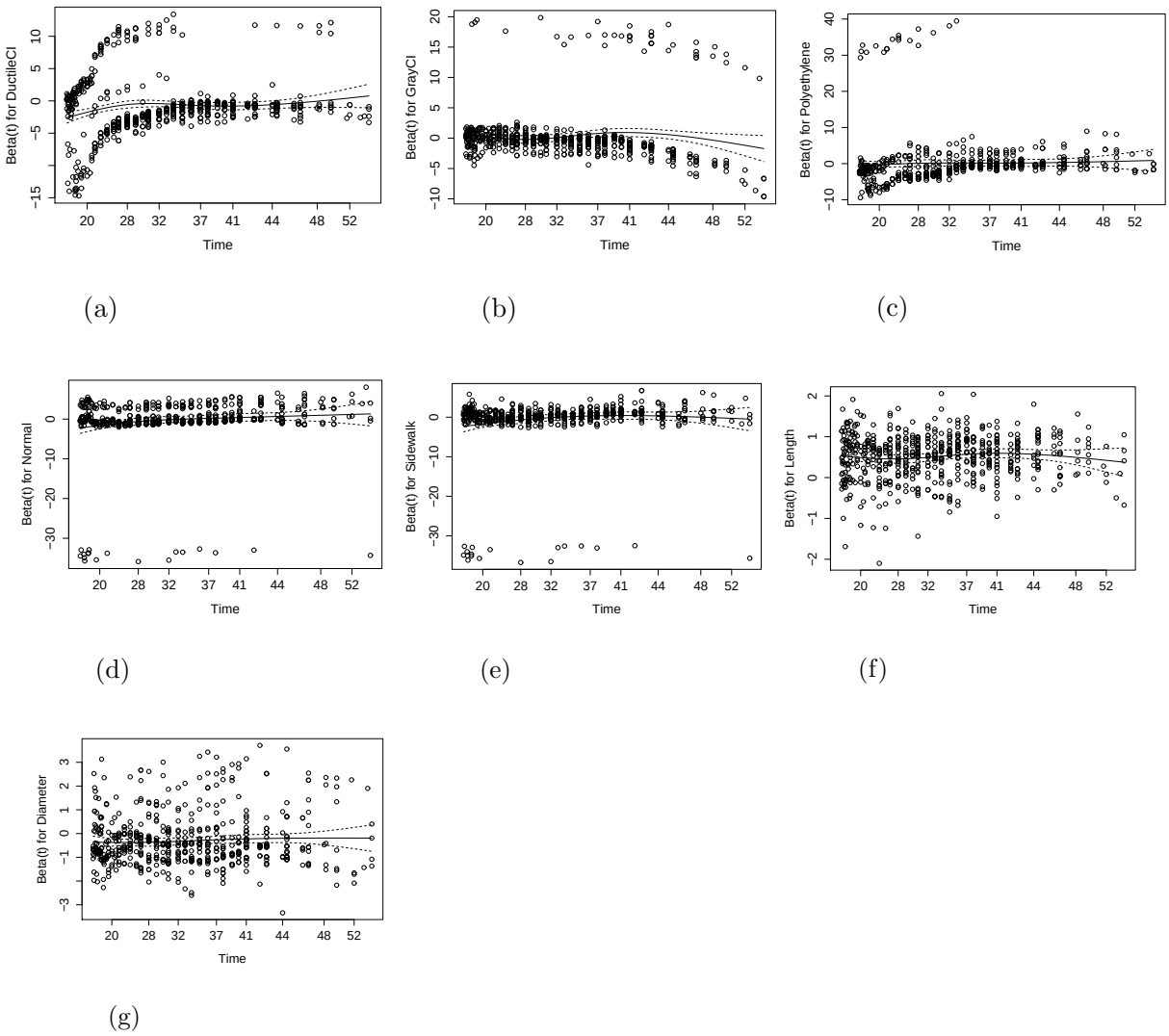


(c)



(d)

Figure A.4: Cumulative hazard functions stratifying by material, traffic, length and diameter, respectively

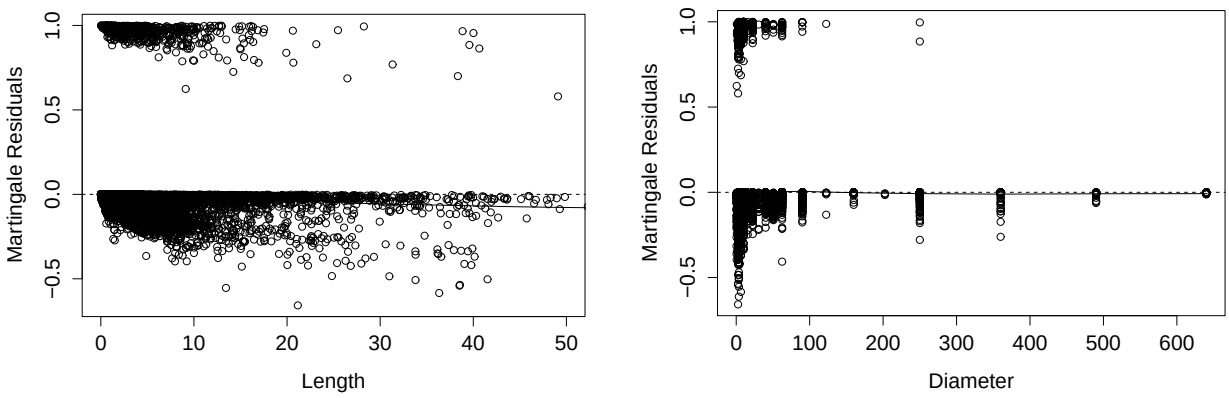


(a) (b) (c)

(d) (e) (f)

(g)

Figure A.5: Scaled Schoenfeld residuals plots for covariates material, traffic, length and diameter, respectively.



(a) (b)

Figure A.6: Martingale residual plots for the covariates length and diameter.

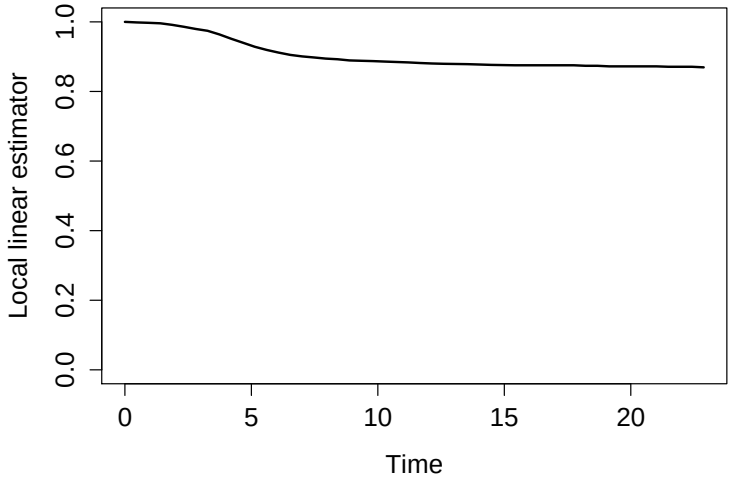


Figure A.7: Local linear estimator of the baseline survival function.