THE NONPARAMETRIC LOCATION-SCALE MIXTURE CURE MODEL

Justin Chown, Cédric Heuchenne, Ingrid Van Keilegom

REPRINT | 2021 / 34

UCLouvain

ISBA | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

THE NONPARAMETRIC LOCATION-SCALE MIXTURE CURE MODEL

JUSTIN CHOWN[1*], CÉDRIC HEUCHENNE[2], AND INGRID VAN KEILEGOM[3]

Abstract. We propose completely nonparametric methodology to investigate location-scale modelling of two-component mixture cure models that is similar in spirit to accelerated failure time models, where the responses of interest are only indirectly observable due to the presence of censoring and the presence of *long-term survivors* that are always censored. We use nonparametric estimators of the location-scale model components that depend on a bandwidth sequence to propose an estimator of the error distribution function that has not been considered before in the literature. When this bandwidth belongs to a certain range of undersmoothing bandwidths the proposed estimator of the error distribution function is root-n consistent. A simulation study investigates the finite sample properties of our approach, and the methodology is illustrated using data obtained to study the behavior of distant metastasis in lymph-node-negative breast cancer patients.

Keywords: censored data, cure model, error distribution function, nonparametric regression

2010 AMS Subject Classifications: Primary: 62G08, 62N01; Secondary: 62G05, 62N02.

1. Introduction

A common problem faced in medical studies is for some subjects to never experience the event of interest during the study period. Consider a follow-up study examining harmful side-effects of a pharmaceutical product. Since side-effects are commonly rare, it is expected that many subjects involved in the study will not experience the harmful effect of the treatment by the end of the study. These subjects will then be censored at the conclusion of the study, and are called *long-term survivors*. The first known study involving statistical analysis of data containing long-term survivors dates back to Boag (1949), and this author coined the term *cure model* to indicate the data contained a non-trivial proportion of long-term survivors.

We consider the case of observing responses that are right censored, and only a minimum of two variables is recorded in the data. In many practical situations additional information of a covariate is available, and a typical statistical analysis of the data is then to assess the dependence of the response on that covariate. Since the response is usually a positive-valued observational time, we will assume that some fixed monotone transformation of the response variable is available such that a location-scale model of the transformed data is appropriate similar to *accelerated failure time* models used in survival analysis. A log-transformation of a time-based response like *time of death*, *time of relapse*, or *time of dropout* (in the case of censoring) is a popular choice.

Throughout this article we will refer to Y and C as the post-transformed response and censoring variables, respectively. The data therefore consists of n independent and identically distributed copies of the triple (X, Z, δ), where X is the covariate of interest, $Z = Y \wedge C$ is the minimum between Y and C, and $\delta = \mathbf{1}[Y \leq C]$ is the right-censoring indicator. Here an observation from the subpopulation of uncured individuals is denoted by Y_u, which is only *indirectly observed* due to censoring. We study the heteroscedastic nonparametric regression of

[1]Fakultät für Mathematik, Ruhr-Universität Bochum, Germany
[2]HEC-Management School of ULg, University of Liège, Belgium
[2]ISBA, Université catholique de Louvain, Belgium
[3]ORSTAT, KU Leuven, Belgium
E-mail addresses: justin.chown@rub.de, c.heuchenne@uliege.be, ingrid.vankeilegom@kuleuven.be.
*Corresponding author.

Y_u given the covariate X, that is,

$$Y_u = m(X) + s(X)\varepsilon, \tag{1.1}$$

where m is the regression function and s is the scale function (bounded away from zero), which are both assumed to be smooth. We assume that the error ε is a continuous random variable that is independent of the covariate X and has distribution function F_e.

Many approaches to statistical inference rely on quantities calculated from F_e. For example, one may be interested in estimating survival probabilities. In many parametric and semiparametric models of data the regression errors are assumed to have a specified distribution, e.g. normally distributed errors. For accelerated failure time models, popular choices of F_e include the normal distribution and the logistic distribution. However, from a nonparametric perspective, F_e is unknown and a convenient form for it should be verified from the data. The goodness-of-fit of a convenient choice of distribution can be verified using, for example, a Kolmogorov-Smirnov or Cramér-von Mises statistic. González-Manteiga and Crujeiras (2013) provide an overview of goodness-of-fit testing for regression models.

A completely nonparametric statistical methodology for examining location-scale modelling of data containing long-term survivors has not been previously studied. Since the appearance of this type of data, the literature on this subject has been divided into two distinct categories. The original cure model proposed by Boag (1949) is now known as the two-component mixture cure model [see, for example, Taylor (1995)]. We will refer to the second category as simply the non-mixture cure model [see, for example, Haybittle (1959, 1965)]. Farewell (1986) observes that cure models should not be used without clear empirical or biological need (see Section 2 for further discussion). Results on censored data models should be used in these cases, and studies involving subjects with censored responses are common. Well known methods can be employed to study these data [see Lawless (1982), Aitkin et al. (1989), Harris and Albert (1990), Collett (1994), among others].

Several advancements have been made in the literature when these models are assumed to have a parametric or a semiparametric form. Kuk and Chen (1992) investigate combining logistic regression methods (used to estimate the unknown proportion of cured cases) with proportional hazards techniques to obtain estimators of their model parameters. Taylor (1995) works with a similar model as that of Kuk and Chen (1992) but proposes an Expectation-Maximization algorithm for simultaneously fitting the model parameters and estimating the baseline hazard function, and this author also uses simulations to conjecture that a crucial assumption is required for identifying components of these models (see Section 2 for further discussion). Sy and Taylor (2000) also consider a similar model as that of Kuk and Chen (1992) but investigate an estimation technique based on the Expectation-Maximization algorithm.

More recently, Lu (2008) considers the proportional hazards mixture cure model and proposes a semiparametric estimator of the unknown parameters of that model by maximizing a so-called nonparametric likelihood function. The estimator is shown to have minimum asymptotic variance. Lu (2010), the same author as before, investigates an accelerated failure time cure model (a special case of the two-component mixture cure model) and proposes an Expectation-Maximization algorithm for fitting the unknown parameters of that model as well as obtaining an estimator of the unknown error density function using a kernel-smoothed conditional profile likelihood technique. Xu and Peng (2014) and López-Cheda et al. (2017) consider estimating the unknown cure fraction in a completely nonparametric setting. Patilea and Van Keilegom (2017) consider a general approach to modelling the conditional survival function of a subject given that the subject is not cured by proposing so-called inversion formulae that allows one to express the conditional survival function of the uncured subjects in terms of the proportion of cured subjects and the subdistributions of the response and the censoring variable.

A particularly interesting model belonging to the non-mixture case is proposed by Yakovlev and Tsodikov (1996). Earlier, in a back-and-forth exchange over letters to the editors of *Statistics in Medicine*, Andrej Yakovlev very clearly details shortcomings of the two-component

mixture cure model. Specifically, he argues that the two-component mixture cure model implicitly assumes that there is only a single *risk operative* in the population, i.e. a single latent variable that determines whether or not subjects are cured / uncured. Yakovlev argues that, in general, populations have multiple risk operatives and he proposes what we are referring to as the non-mixture cure model in his letter to the editors. The exchange between him and the authors Alan Cantor and Jonathan Shuster can be found in Yakovlev et al. (1994).

Sometimes this model is called a promotion time cure model, and it has gained popularity due to motivations from Biology (in particular cancer research). Several results on parametric and semiparametric estimation of the unknown parameters in these models are available in the literature. However, this model is not in the scope of the current article, and we only mention a few of the notable works in this area. Tsodikov (1998) compares likelihood-based fitting techniques for the non-mixture cure model. Later, Tsodikov et al. (2003) survey the literature and report that, in less than a decade from its introduction, the non-mixture cure model is already in popular use, and these authors promote use of the non-mixture cure model in semiparametric and Bayesian settings. Zeng et al. (2006) propose a recursive algorithm for obtaining maximum likelihood estimates in the non-mixture cure model. Additionally, these authors show their regression parameter estimators have minimum asymptotic variance. Portier et al. (2017) consider an extension of the non-mixture cure model.

A popular theme in both modelling categories is to use proportional hazards methods, with the baseline hazard function estimated by nonparametric methods. In this case, the proportional hazards model is made additionally flexible through a nonparametric estimator of the baseline hazard function. Recently, efforts have been made to unify the two seemingly distinct categories of cure models. Sinha et al. (2003) discuss the benefits and disadvantages of the mixture and non-mixture cure models. Yin and Ibrahim (2005) propose transforming the unknown population survival function in a manner that is analogous to the Box-Cox transformation for non-normally distributed random variables.

The article is organized as follows. We further discuss model (1.1) in Section 2, and motivate our nonparametric function estimators in Section 3, and we give asymptotic results on these estimators in Section 4. Automated bandwidth selection for the regression and scale estimators, and the incidence rate estimator, are discussed in Section 5. Section 6 investigates the finite sample behavior of the error distribution estimator and the regression function estimator proposed in Section 3. In Section 7, we examine a set of data collected to study distant metastasis in lymph-node-negative breast cancer patients. The proofs of these results and further supporting technical results are given in the supplemental material.

2. Two-component mixture cure models

In this section we discuss identifiability of the cure model parameters. Let the support of X be $[0,\,1]$, and we will write F for the conditional distribution function of the responses Y given X and F_u for the conditional distribution function of Y_u given X. In addition, we will write F_c for the conditional distribution function of C given X. We will assume that the conditional survival function of C given X is proper in the sense that $1-F_c(t\,|\,X)\to 0$ as $t\to\infty$, and that Y and C are conditionally independent given X. For both cure models and censored response models, it is important that we place conditions on the distribution function F (and therefore F_u) so that we may identify and estimate the regression model components, m and s, and the error distribution function F_e. As previously discussed, a cure model should only be imposed in certain situations depending on empirical or biological need.

Here, empirical or biological need means that the support of the censoring variable C is never contained in the support of the uncured response variable Y_u; that is, we require that

$$\tau_0 = \sup_{x\in[0,\,1]} \tau_{F_u}(x) < \inf_{x\in[0,\,1]} \tau_{F_c}(x), \tag{2.1}$$

where $\tau_{F_u}(x) = \inf\{t \in \mathbb{R} : 1 - F_u(t \,|\, x) = 0\}$ and $\tau_{F_c}(x) = \inf\{t \in \mathbb{R} : 1 - F_c(t \,|\, x) = 0\}$. Taylor (1995) uses simulation evidence to conjecture the necessity of (2.1). Xu and Peng (2014) and López-Cheda et al. (2017) observe that (2.1) leads to identifiability of the cure model components [Lemma 1 of López-Cheda et al. (2017)]. Specifically, this assumption is required to identify the conditional proportion $\pi(X)$ of cured cases given X as well as the conditional distribution function F_u of Y_u given X.

We assume that the remaining mass in the conditional distribution of Y given X beyond $\tau_{F_u}(X)$ occurs at $Y = \infty$, i.e. $1 - F(\tau_{F_u}(X) \,|\, X) = 1 - F(\infty \,|\, X) = \pi(X)$, where $0 < \pi_l \leq \pi(X) \leq \pi_u < 1$, for some constants $0 < \pi_l \leq \pi_u < 1$. This justifies writing

$$1 - F(t \,|\, X) = \pi(X) + (1 - \pi(X))(1 - F_u(t \,|\, X)), \qquad t \in \mathbb{R}. \tag{2.2}$$

Further, we can make use of the fact that the conditional boundary $\tau_{F_u}(X)$ in the definition of $\pi(X)$ above may be replaced by the unconditional boundary $\tau_0 = \sup_{x \in [0,\,1]} \tau_{F_u}(x)$, because $1 - F(\tau_{F_u}(X) \,|\, X) = \pi(X) = 1 - F(\tau_0 \,|\, X)$. This means that $\pi(X)$ depends on X solely through the conditional distribution function F of Y given X.

We can see that identification of π in (2.2) allows for identification of F_u as well, since (2.2) implies the equivalence

$$F(t \,|\, X) = (1 - \pi(X))F_u(t \,|\, X), \quad t \leq \tau_0, \tag{2.3}$$

fundamentally relating the conditional distribution function F of Y given X to F_u of Y_u given X. The model equation (1.1) implies structural constraints on F_u that depend on the quantities m, s, and F_e in the sense that $F_u(t \,|\, X) = F_e((t - m(X))/s(X))$, for $t \leq \tau_{F_u}(X)$ and every X. Further, the conditional boundary $\tau_{F_u}(X)$ of Y_u given X can be standardized to obtain the unconditional boundary τ_{F_e} of the model error ε, with $\tau_{F_e} = \inf\{t \in \mathbb{R} : 1 - F_e(t) = 0\}$ for the largest observable standardized error. Here, standardization implies that $1 = F_u(\tau_{F_u}(X) \,|\, X) = F_e((\tau_{F_u}(X) - m(X))/s(X)) = F_e(\tau_{F_e})$, and, therefore, $\tau_{F_e} = (\tau_{F_u}(X) - m(X))/s(X)$.

The data we work with have responses that are right-censored, and we will assume specialized forms of m and s as in Dabrowska (1987), who, along with Beran (1981), recommend working with L–type regression functionals when the data contain censored values. The reasoning for using specialized forms of m and s comes from the fact that the censored values in the data will affect estimators of the classical mean and scale functionals. To explain the idea, we introduce the score function J and define the conditional quantiles of Y_u given X as $\xi_u(p \,|\, X) = \inf\{y \in (-\infty, \tau_0] : F_u(y \,|\, X) \geq p\}$, for $p \in [0, 1]$. Here, the score function J must be nonnegative and satisfy $\int_0^1 J(p)\, dp = 1$. The regression components m and s are then defined as

$$m(x) = \int_0^1 \xi_u(p \,|\, x) J(p)\, dp \quad \text{and} \quad v(x) = \int_0^1 \xi_u^2(p \,|\, x) J(p)\, dp - m^2(x),$$

where $s(x) = v^{1/2}(x)$ and $x \in [0, 1]$. Note that different score functions J lead to different model definitions of m and s, and so the choice of score function should include practical considerations. An example of a score function commonly used in practice is $J(p) = \mathbf{1}[\alpha < p < 1 - \alpha]/(1 - 2\alpha)$, which results in an α-trimmed mean m and an α-trimmed scale s. If we worked with $J(p) = \mathbf{1}[0 < p < 1]$, then we recover the classical definitions of m and s, where now m is a mean function and s is a scale function.

Using these specialized forms of m and s, we will now restrict the error distribution F_e so that we can identify and consistently estimate m and s. If F_e satisfies

$$\int_0^1 \xi_{F_e}(p) J(p)\, dp = 0 \quad \text{and} \quad \int_0^1 \xi_{F_e}^2(p) J(p)\, dp = 1, \tag{2.4}$$

where ξ_{F_e} is the quantile function of F_e (i.e. $\xi_{F_e}(p) = \inf\{t \in (-\infty, \tau_{F_e}] : F_e(t) \geq p\}$, for $p \in [0, 1]$), then m and s are identifiable. Note that these conditions are analogous to the typical mean 0 and variance 1 identification requirements imposed in classical regression models, which, as before, can be observed by considering $J(p) = \mathbf{1}[0 < p < 1]$.

3. Estimation of the model parameters

In this section we define the estimators of F_u, π, m, and s that lead to a consistent estimator of F_e. Write H for the conditional distribution function of Z given X and H^1 for the conditional subdistribution function of both Z and $\delta = 1$ given X. Here, we will take advantage of the fact that F and F_u are consistently estimated, when $-\infty < t \leq \tau_0$ and $x \in [0,\, 1]$, cf. Van Keilegom and Akritas (1999).

We begin by defining estimators of H and H^1. For this purpose, we introduce the Nadaraya-Watson weights depending on $x \in [0,\, 1]$ as

$$W_j(x) = K\bigg(\frac{x - X_j}{a_n}\bigg) \bigg/ \bigg\{ \sum_{k=1}^{n} K\bigg(\frac{x - X_k}{a_n}\bigg)\bigg\}, \qquad j = 1, \ldots, n,$$

where K is a given kernel function and $\{a_n\}_{n\geq 1}$ is a sequence of bandwidth parameters. Later, we will specify conditions on choosing K and $\{a_n\}_{n\geq 1}$. Estimators of H and H^1 follow from the classical approach of Stone (1977), that is, for $-\infty < t \leq \tau_0$,

$$\hat{H}(t \,|\, X) = \sum_{j=1}^{n} W_j(X) \mathbf{1}[Z_j \leq t] \quad \text{and} \quad \hat{H}^1(t \,|\, X) = \sum_{j=1}^{n} W_j(X) \delta_j \mathbf{1}[Z_j \leq t]. \tag{3.1}$$

The conditional cumulative hazard function Λ of Y given X may be written as

$$\Lambda(t \,|\, X) = \int_{-\infty}^{t} \frac{F(ds \,|\, X)}{1 - F(s- \,|\, X)} = \int_{-\infty}^{t} \frac{H^1(ds \,|\, X)}{1 - H(s- \,|\, X)}, \qquad -\infty < t \leq \tau_0. \tag{3.2}$$

Substituting (3.1) into the far right-hand side of (3.2) leads to an estimator of $F(t \,|\, X) = 1 - \exp(-\Lambda(t \,|\, X))$ in the classical approach of Beran (1981), that is,

$$\hat{\mathbb{F}}(t \,|\, X) = 1 - \prod_{Z_{(j)} \leq t} \bigg\{ 1 - \frac{W_{(j)}(X)}{\sum_{k=j}^{n} W_{(k)}(X)} \bigg\}^{\delta_{(j)}}, \qquad -\infty < t \leq \tau_0. \tag{3.3}$$

Here, $Z_{(1)} \leq \ldots \leq Z_{(n)}$ is the ascending ordering of $Z_1, \ldots, Z_n$, and both of $\delta_{(1)}, \ldots, \delta_{(n)}$ and $W_{(1)}(X), \ldots, W_{(n)}(X)$ are ordered according to $Z_{(1)}, \ldots, Z_{(n)}$. For simplicity, we will assume that the data contain no tied responses, which is reasonable, because our assumptions imply that the responses Z_j are continuous random variables. Otherwise, the ordering of the variables indicated above is not unique, and the estimator $\hat{\mathbb{F}}$ is affected.

Xu and Peng (2014) propose estimating τ_0 by the largest uncensored response $Z^1_{\max}$, and then combining this estimator with (3.3) forms an estimator the unknown proportion π of cured cases, that is,

$$\hat{\pi}(x) = 1 - \hat{\mathbb{F}}(Z^1_{\max} \,|\, x), \qquad x \in [0,\, 1]. \tag{3.4}$$

The estimator $\hat{\pi}$ is shown to be consistent and asymptotically normally distributed. López-Cheda et al. (2017) further study this estimator, and these authors propose bootstrap-based methods for selecting the bandwidth parameter.

Estimators of m and s require an estimator of the unknown conditional quantile function ξ_u. It is easy to show that $\xi_u(p \,|\, X) = \xi((1 - \pi(X))p \,|\, X)$, using (2.3), with $p \in [0,\, 1]$. Consistent estimators of $\xi((1-\pi(X))p \,|\, X)$ are given by $\hat{\xi}((1-\hat{\pi}(X))p \,|\, X) = \inf\{y \in (-\infty,\, \tau_0] : \hat{\mathbb{F}}(y \,|\, X) \geq \{1 - \hat{\pi}(X)\}p\}$, for $p \in [0,\, 1]$. Estimators of m and s can then be defined analogously to those of Van Keilegom and Akritas (1999), that is,

$$\hat{m}(x) = \int_0^1 \hat{\xi}\big(\big(1 - \hat{\pi}(x)\big)p \,|\, x\big) J(p)\, dp \quad \text{and} \quad \hat{v}(x) = \int_0^1 \hat{\xi}^2\big(\big(1 - \hat{\pi}(x)\big)p \,|\, x\big) J(p)\, dp - \hat{m}^2(x),$$

with $\hat{s}(x) = \hat{v}^{1/2}(x)$ and $x \in [0,\, 1]$.

Using the relation (2.3) and the model equation (1.1), we can write the error distribution function F_e as an expectation with respect to X, that is,

$$F_e(t) = E\left[\frac{F(ts(X) + m(X) \,|\, X)}{1 - \pi(X)}\right], \qquad -\infty < t \leq \tau_{F_e}.$$

Here, we transfer the support of ε, which is the interval $(-\infty, \tau_{F_e})$, into the support of Y_u, which is the interval $(-\infty, \tau_{F_u}(X)) = (-\infty, \tau_{F_e} s(X) + m(X))$, because F, π, m, and s can each be consistently estimated from the data. This is analogous to the transfer of tail information studied in Van Keilegom and Akritas (1999). However, in order to consistently estimate F_e from the data, we transfer information from F_e to F_u. This is in contrast to the work of the previously mentioned authors, who transfer tail information from the responses to the errors in order to consistently estimate the conditional distribution of response given covariate in their model, which does not include cured cases. We arrive at the proposed estimator of F_e:

$$\hat{\mathbb{F}}_e(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{\hat{\mathbb{F}}(t\hat{s}(X_j) + \hat{m}(X_j) \,|\, X_j)}{1 - \hat{\pi}(X_j)}, \qquad -\infty < t \leq \tau_{F_e}. \tag{3.5}$$

Note that this estimator is averaging over the local model estimators of F_e at each covariate X_j that are consistent with rates slower than root-n and depend on the chosen bandwidth a_n. However, an appropriate choice of bandwidth a_n leads to root-n consistency of the estimator $\hat{\mathbb{F}}_e$ and a functional central limit theorem, which is described in the next section.

4. Large sample properties of the proposed estimators

In order to state our asymptotic results for the estimators introduced in the previous section, we require the following assumptions:

(A1) Y and C are conditionally independent given X, and C is a continuous random variable with proper survival function such that (2.1) is satisfied.

(A2) The bandwidth a_n satisfies $(na_n^2)^{-1}\log\log(n) = O(1)$ and $na_n^5 \log^{-1}(n) = O(1)$.

(A3) There are real numbers $0 < \pi_l \leq \pi_u < 1$ satisfying $\pi_l \leq \pi(X) \leq \pi_u$ for every X.

(A4) (i) The kernel function K is a symmetric probability density function with support $[-1, 1]$.

(ii) K has bounded second derivative.

(A5) (i) The support of the covariate X is $[0, 1]$.

(ii) The density function g of X has two bounded derivatives.

(A6) (i) There is a continuous nondecreasing function L_1, satisfying $L_1(-\infty) = 0$ and $L_1(\tau_0) < \infty$, such that

$$H(t \,|\, x) - H(s \,|\, x) \leq L_1(t) - L_1(s), \qquad x \in [0, 1],\ -\infty < s < t \leq \tau_0.$$

(ii) The conditional (sub)distribution functions H and H^1 have continuous partial derivatives, with respect to x, $\dot{H}$ and $\dot{H}^1$, respectively, bounded in $(-\infty, \tau_{F_c}] \times [0, 1]$.

(iii) There are continuous nondecreasing functions L_2 and L_3; satisfying $L_2(\tau_0) < \infty$, $L_3(\tau_0) < \infty$, and $L_2(-\infty) = L_3(-\infty) = 0$; such that

$$\dot{H}(t \,|\, x) - \dot{H}(s \,|\, x) \leq L_2(t) - L_2(s), \quad x \in [0, 1], \quad -\infty < s < t \leq \tau_0,$$

and

$$\dot{H}^1(t \,|\, x) - \dot{H}^1(s \,|\, x) \leq L_3(t) - L_3(s), \quad x \in [0, 1], \quad -\infty < s < t \leq \tau_0.$$

(iv) The conditional (sub)distribution functions H and H^1 have second partial derivatives, with respect to x, bounded in $(-\infty, \tau_{F_c}] \times [0, 1]$.

(A7) The conditional (sub)distribution functions H and H^1 admit bounded Lebesgue density functions.

(A8) The (conditional) distribution functions $P(Z \le t)$ and $P(Z \le t \,|\, \delta = 1)$ are twice continuously differentiable, and the derivatives are bounded away from zero in absolute value on any compact interval in the region $(-\infty, \tau_0]$, with the density function of $P(Z \le t \,|\, \delta = 1)$ also bounded away from zero at $t = \tau_0$.

(A9) (i) The function $L = L_1 + L_2 + L_3$; with L_1, L_2, and L_3 given by (A6); is differentiable, with bounded derivative L', such that L' is uniformly continuous, the function $t \mapsto tL'(t)$ is uniformly continuous, and $\sup_{-\infty < t \le \tau_0} |tL'(t)| < \infty$.

(ii) The error distribution function F_e admits a bounded Lebesgue density function f_e satisfying $\sup_{-\infty < t \le \tau_{F_e}} |tf_e(t)| < \infty$. Further, the density function f_e is differentiable, with bounded derivative f_e', satisfying $\sup_{-\infty < t \le \tau_{F_e}} |t^2 f_e'(t)| < \infty$.

(A10) (i) The score function J is bounded and nonnegative, and there are constants $0 < p_l < p_u \le 1$ such that J is bounded away from zero on (p_l, p_u) and equal to zero on $[0, p_l] \cup [p_u, 1]$.

(ii) J is continuously differentiable with bounded derivative J'.

Assumptions (A4) and (A5) are common assumptions taken for nonparametric regression models, which guarantee good performance of nonparametric function estimators. Assumption (A5) guarantees that the density function g of X is bounded and bounded away from 0 on $[0, 1]$. Hence, as n increases more data local to any $x \in [0, 1]$ become available for use in estimation. A referee has kindly pointed out that in practice the covariate X rarely has support $[0, 1]$. The results of this paper continue to hold with the interval $[0, 1]$ replaced by an arbitrary interval I of finite length, and X is supported on I, which is more realistic in light of applications.

Note that Assumptions (A6) (i) and (iii) are satisfied for many distributions. Suppose that H is the logistic distribution function with a positive, bounded mean function m and constant scale function $s \equiv 1$. Write $l_m = \inf_{x \in [0,1]} m(x)$ and $u_m = \sup_{x \in [0,1]} m(x)$. Then Assumption (A6) (i) is satisfied by choosing $L_1(t) = \int_{-\infty}^{t} \exp(u_m - s)\{1 + \exp(l_m - s)\}^{-2}\, ds$. When m is also differentiable, with a bounded derivative, then bounding functions L_2 and L_3 (that are similar to L_1) can be chosen to satisfy Assumption (A6) (iii) as well. Assumptions (A6) (ii) and (iv) and (A7) imply that the conditional distribution functions F_u and F_c are smooth, and that π must be smooth as well, with two bounded derivatives. This is due to the conditional independence of Y and C given X, where we can write

$$1 - H(t \,|\, X) = \big(\pi(X) + (1 - \pi(X))(1 - F_u(t \,|\, X))\big)\big(1 - F_c(t \,|\, X)\big).$$

Finally, m and s defined in Section 3 are truncated, weighted averages, which means that the integrals in the definitions of these functions are restricted to compact subsets of $(-\infty, \tau_0]$, which is due to the facts that J has support contained in $[p_l, 1]$ and that $\tau_0 < \infty$. Combining the Leibniz integral rule for differentiation with Assumptions (A6) (ii) and (iv) yields that both m and s are twice differentiable with bounded derivatives. Assumption (A8) is a technical assumption required for the consistency of $Z^1_{\max}$ for τ_0, which many probability distributions satisfy.

For a simpler exposition, we define the functions

$$\zeta\big(x, Z_j, \delta_j, t\big) = \frac{\delta_j \mathbf{1}[Z_j \le t]}{1 - H(Z_j - \,|\, x)} - \int_{-\infty}^{t} \frac{\mathbf{1}[Z_j > s]}{(1 - H(s - \,|\, x))^2} H^1(ds \,|\, x),$$

$$\eta_m(x, Z_j, \delta_j) = \int_{-\infty}^{\tau_0} \zeta(x, Z_j, \delta_j, y) \frac{1 - F(y \,|\, x)}{1 - \pi(x)} J\big(F_u(y \,|\, x)\big)\, dy,$$

$$\xi_m(x, Z_j, \delta_j) = \eta_m(x, Z_j, \delta_j) - \frac{\pi(x) C_m(x)}{1 - \pi(x)} \zeta(x, Z_j, \delta_j, \tau_0),$$

$$\eta_s(x, Z_j, \delta_j) = \int_{-\infty}^{\tau_0} \zeta(x, Z_j, \delta_j, y) \frac{1 - F(y \,|\, x)}{1 - \pi(x)} \frac{y - m(x)}{s(x)} J\big(F_u(y \,|\, x)\big)\, dy,$$

and

$$\xi_s(x, Z_j, \delta_j) = \eta_s(x, Z_j, \delta_j) - \frac{\pi(x)C_s(x)}{1-\pi(x)}\zeta(x, Z_j, \delta_j, \tau_0),$$

with

$$C_m(x) = \int_{-\infty}^{\tau_0} F_u(y \,|\, x) J\big(F_u(y \,|\, x)\big)\, dy$$

and

$$C_s(x) = \int_{-\infty}^{\tau_0} \frac{y - m(x)}{s(x)} F_u(y \,|\, x) J\big(F_u(y \,|\, x)\big)\, dy.$$

Our first result specifies the consistency and an expansion of the estimator $\hat{\pi}$. Similar properties of this estimator are given in Theorem 3 of López-Cheda et al. (2017).

Proposition 1. *Let Assumptions* (A1) – (A9) *hold. Then,*

$$\sup_{x\in[0,\,1]} \Big|\hat{\pi}(x) - \pi(x)\Big| = O\big((na_n)^{-1/2}\log^{1/2}(n)\big), \qquad a.s.$$

Additionally,

$$\hat{\pi}(x) - \pi(x) = -\frac{\pi(x)}{g(x)}\frac{1}{na_n}\sum_{j=1}^{n} K\left(\frac{x - X_j}{a_n}\right)\zeta\big(x, Z_j, \delta_j, \tau_0\big) + R_{1,n}(x),$$

where $\sup_{x\in[0,\,1]} |R_{1,n}(x)| = O((na_n)^{-3/4}\log^{3/4}(n))$, *almost surely.*

The next two results specify the consistency and expansions of the estimators $\hat{m}$ and $\hat{s}$.

Proposition 2. *Let Assumptions* (A1) – (A9) *and* (A10) *(i) hold. Then,*

$$\sup_{x\in[0,\,1]} \Big|\hat{m}(x) - m(x)\Big| = O\big((na_n)^{-1/2}\log^{1/2}(n)\big), \quad a.s.$$

Additionally, if Assumption (A10) *(ii) holds,*

$$\hat{m}(x) - m(x) = -\frac{1}{g(x)na_n}\sum_{j=1}^{n} K\left(\frac{x - X_j}{a_n}\right)\xi_m(x, Z_j, \delta_j) + R_{2,n}(x),$$

where $\sup_{x\in[0,\,1]} |R_{2,n}(x)| = O((na_n)^{-3/4}\log^{3/4}(n))$, *almost surely.*

Proposition 3. *Let Assumptions* (A1) – (A9) *and* (A10) *(i) hold. Then,*

$$\sup_{x\in[0,\,1]} \Big|\hat{s}(x) - s(x)\Big| = O\big((na_n)^{-1/2}\log^{1/2}(n)\big), \quad a.s.$$

Additionally, if Assumption (A10) *(ii) holds,*

$$\hat{s}(x) - s(x) = -\frac{1}{g(x)na_n}\sum_{j=1}^{n} K\left(\frac{x - X_j}{a_n}\right)\xi_s(x, Z_j, \delta_j) + R_{3,n}(x),$$

where $\sup_{x\in[0,\,1]} |R_{3,n}(x)| = O((na_n)^{-3/4}\log^{3/4}(n))$, *almost surely.*

We are now prepared to state our main result: the weak convergence of $n^{1/2}(\hat{\mathbb{F}}_e - F_e)$.

Theorem 1. *Let Assumptions* (A1) – (A10) *hold, but choose the bandwidth sequence* a_n *such that* $a_n^2 = o(n^{-1/2})$ *and* $(na_n)^{-3/4}\log^{3/4}(n) = o(n^{-1/2})$. *Then,* $n^{1/2}(\hat{\mathbb{F}}_e - F_e)$ *is asymptotically linear, that is,*

$$n^{1/2}\big(\hat{\mathbb{F}}_e(t) - F_e(t)\big) = n^{-1/2}\sum_{j=1}^{n}\frac{1}{2}b_t(X_j, Z_j, \delta_j) + R_{4,n}(t),$$

where $\sup_{-\infty<t\le\tau_F} |R_{4,n}(t)| = o_P(1)$ *and the influence function is* $(1/2)b_t(X_j, Z_j, \delta_j)$, *with*

$$b_t(X_j, Z_j, \delta_j) = \frac{1 - F(ts(X_j) + m(X_j) \,|\, X_j)}{1 - \pi(X_j)}\zeta\big(X_j, Z_j, \delta_j, ts(X_j) + m(X_j)\big)$$

$$- f_e(t)\big(\eta_m(X_j, Z_j, \delta_j) + t\eta_s(X_j, Z_j, \delta_j)\big)$$
$$- \frac{\pi(X_j)}{1-\pi(X_j)}\big(F_e(t) - f_e(t)(C_m(X_j) + tC_s(X_j))\big)\zeta\big(X_j, Z_j, \delta_j, \tau_0\big),$$

with $-\infty < t \leq \tau_{F_e}$. *Consequently, the process* $\{n^{1/2}(\hat{\mathbb{F}}_e(t) - F_e(t))\}_{-\infty<t\leq\tau_{F_e}}$ *weakly converges in the space* $\ell^\infty([-\infty, \tau_{F_e}])$ *to a mean zero Gaussian process* $\{Z(t)\}_{-\infty<t\leq\tau_{F_e}}$, *with covariance function* $\Sigma(u, v) = (1/4)E[b_u(X, Z, \delta)b_v(X, Z, \delta)]$, *for* $-\infty < u, v \leq \tau_{F_e}$.

Remark 1. The results of Theorem 1 require that the bandwidth a_n should always *undersmooth* the estimators $\hat{\mathbb{F}}$, $\hat{m}$, $\hat{s}$, and $\hat{\pi}$. To gain a better perspective on the proper amount of undersmoothing, consider a bandwidth sequence of the form $a_n = O(n^{-1/4-\gamma}\log^{1/4+\gamma}(n))$, which satisfies $a_n^2 = o(n^{-1/2})$ and $(na_n)^{-3/4}\log^{3/4}(n) = o(n^{-1/2})$, for $0 < \gamma < 1/12$. When $\gamma = 1/12$ we have $a_n = O(n^{-1/3}\log^{1/3}(n))$, but this choice undersmoothes by too much and does not lead to root-n consistency of $\hat{\mathbb{F}}_e$. Similarly, when $\gamma = 0$ we have $a_n = O(n^{-1/4}\log^{1/4}(n))$, but this choice does not undersmooth by enough and also does not lead to root-n consistency of $\hat{\mathbb{F}}_e$. Interestingly, when we have that $a_n^2 = O((na_n)^{-3/4}\log^{3/4}(n))$ it follows that $a_n = O(n^{-3/11}\log^{3/11}(n))$, which is a sequence of the previously discussed form for $\gamma = 1/44$, and this choice of bandwidth allows $\hat{\mathbb{F}}_e$ to be a root-n consistent estimator of F_e.

Remark 2. Many statistical inference procedures rely on a correct specification of the statistical model underlying the data. In the context of accelerated failure time models of survival data, where a log-transformation of the observed response times is commonly used, one needs to verify the distribution of the unmoderated lifetimes, which may be interpreted as errors in a regression model. The null hypothesis is then $H_0 : F_e = F_{e,0}$, for a specified $F_{e,0}$. One may test the goodness-of-fit of $F_{e,0}$ using (say) a Kolmogorov-Smirnov statistic, which is calculated under H_0 as $T_n = n^{1/2}\sup_{-\infty<t\leq\tau_{F_{e,0}}}|\hat{\mathbb{F}}_e(t) - F_{e,0}(t)|$. Theorem 1 implies the limiting distribution of T_n is that of $\sup_{-\infty<t\leq\tau_{F_{e,0}}}|Z(t)|$, where the Gaussian process Z is given in the statement of the theorem. Critical values from this distribution may be difficult to calculate. One can proceed by using bootstrap-based critical values instead by considering the distribution of bootstrap test statistics T_n^*, which are constructed similarly to T_n, calculated from bootstrap samples of data. Here the bootstrap samples of data are constructed under H_0.

5. Choices of bandwidth

In this section we discuss a bootstrap option for selecting bandwidths used in estimators of F, π, m and s. Here, we will focus our discussion on selecting a bandwidth for $\hat{m}$ that is in the same spirit as the bandwidth selection recommendation of López-Cheda et al. (2017) for the estimator $\hat{\pi}$. Our discussion follows the perspective that statistical analyses of data should be objective oriented. For example, when characterization of the mean m is desired, the choice of bandwidth is made so that the estimator $\hat{m}$ provides the best characterization of m with respect to alternative bandwidth choices, and, hence, relies on an application-dependent optimality criterion.

Many nonparametric estimators require the selection of the bandwidth parameter. It is customary to select the bandwidth with respect to an optimality criterion, and, consequently, the resulting estimators (usually) obtain desirable properties like optimal rates of decay in certain performance metrics. Traditionally, the mean squared error or the integrated mean squared error (IMSE) play central roles in bandwidth selection, because these criteria are popular and useful in regression problems. Optimally choosing the bandwidth in these cases means that the squared bias of a nonparametric estimator is made proportional to its variance. This optimality constraint requires a certain choice of bandwidth, which typically does not yield an undersmoothing bandwidth.

In cases where undersmoothing is desired, fully automated selection may not be possible, because the estimator may need to satisfy certain properties like having a squared bias that is

decaying faster than its variance. Rather, a partially automated choice can be made as follows. For $\hat{\mathbb{F}}_e$ to be a root-n consistent estimator of F_e, the bandwidth used in the nonparametric estimators of F, π, m, and s is required to undersmooth (see Remark 1). We can, however, automate the selection of the bandwidth parameter, and optimize the IMSE as in the case of $\hat{m}$ above. Here the bandwidth will be of the form $C_{\text{opt}} n^{-1/4}$, for some constant $C_{\text{opt}} > 0$, and we can scale this bandwidth by $\log^{-1/4}(n)$ to obtain a root-n consistent estimator of F_e.

For the purpose of selecting an appropriate bandwidth for the estimator $\hat{m}$, we consider a simple weighted bootstrap, without resampling the covariate X, inspired by López-Cheda et al. (2017), which is similar to the weighted bootstrap proposed by Li and Datta (2001). For each $j = 1, \ldots, n$, we set $X_j^* = X_j$ and generate a pair (Z_j^*, δ_j^*) from an estimator $\hat{\mathbb{F}}_b$ of the conditional distribution function of (Z_j, δ_j) given X_j, that is,

$$\hat{\mathbb{F}}_b(u, v \,|\, X_j) = \sum_{i=1}^{n} W_{i,b}(X_j) \mathbf{1}[Z_i \le u,\, \delta_i = v], \quad u \in \mathbb{R},\ v = 0, 1,$$

where $W_{i,b}$ is the Nadaraya-Watson weight W_j previously introduced, but now with $j = i$ and pilot bandwidth b in place of the bandwidth a_n. The bootstrap sample of data is then $(X_1,\, Z_1^*,\, \delta_1^*), \ldots, (X_n,\, Z_n^*,\, \delta_n^*)$. In the following, we write E^* for conditional expectation given the original sample of data $(X_1,\, Z_1,\, \delta_1), \ldots, (X_n,\, Z_n,\, \delta_n)$.

The bandwidth for $\hat{m}$ is chosen as the minimizer of a bootstrap version of the IMSE, that is,

$$a_n^* = \underset{a \in [a_l,\, a_u]}{\arg\min}\, \text{IMSE}^*(a) = \underset{a \in [a_l,\, a_u]}{\arg\min} \int_{[0,\,1]} E^*\Big[\big(\hat{m}_a^*(x) - \hat{m}_b(x)\big)^2\Big]\, dx, \tag{5.1}$$

with $\hat{m}_a^*$ given as the estimator $\hat{m}$, but using bandwidth a in place of a_n and bootstrap data $(X_j,\, Z_j^*,\, \delta_j^*)$ replacing the original data $(X_j,\, Z_j,\, \delta_j)$, $j = 1, \ldots, n$. Here, $\hat{m}_b$ is the estimator $\hat{m}$ calculated with the original data, but now with pilot bandwidth b in place of a_n. The window $[a_l,\, a_u]$ is chosen so that its length is $a_u - a_l = O(n^{-1/5} \log^{1/5}(n))$ and it captures the true IMSE optimal bandwidth, that is

$$a_{n,\text{opt}} = \underset{a \in [a_l,\, a_u]}{\arg\min}\, \text{IMSE}(a) = \underset{a \in [a_l,\, a_u]}{\arg\min} \int_{[0,\,1]} E\Big[\big(\hat{m}_a(x) - m(x)\big)^2\Big]\, dx.$$

Note that the same technique to bandwidth selection may be applied to choose bandwidths for the estimators of F, π, and s. Further, the pilot bandwidth b could be optimally chosen so that (5.1) is minimized over choices a and b, for each set of data. In this case, the optimal pilot bandwidth typically satisfies $b^* = O(n^{-1/9})$ and is larger than the desired a_n^*. Simulation evidence shows that the choice of pilot bandwidth has little effect on the optimal bandwidth choice a_n^* [pages 149-150 of López-Cheda et al. (2017)]. The rule-of-thumb given in López-Cheda et al. (2017),

$$b = (X_{(n)} - X_{(1)}) 10^{-7/9} n^{-1/9},$$

where $X_{(1)}$ and $X_{(n)}$ are the respective minimum and maximum of the observed covariates, is a practical choice of pilot bandwidth.

6. Numerical studies of the previous results

In this section we investigate the finite sample properties of the proposed estimators $\hat{\mathbb{F}}_e$ and $\hat{m}$ using simulation studies. For the estimator $\hat{\mathbb{F}}_e$, we consider a bandwidth chosen by bootstrap as detailed in the previous section, and scale this data-driven bandwidth by $\log^{-1/4}(n)$, with sample size n. We measure the resulting estimation performance of $\hat{\mathbb{F}}_e$ using this scaled bandwidth choice. For the estimator $\hat{m}$, we consider two cases where a parametric specification of the mean function m by m_θ, with parameter θ, is appropriate and inappropriate. Here, we also use an automated bandwidth selector as discussed in the previous section, and compare estimation performance of the nonparametric estimator $\hat{m}$ to two semiparametric estimators.

n	-2	-1	0	1	2
100	0.066	0.101	0.097	0.076	0.060
200	0.095	0.091	0.071	0.063	0.072
300	0.110	0.082	0.049	0.045	0.086
500	0.121	0.071	0.053	0.042	0.091

TABLE 1. Simulated asymptotic mean squared error values of $\hat{\mathbb{F}}_e$ at the points -2, -1, 0, 1 and 2, that is, the simulated mean squared error values of $\hat{\mathbb{F}}_e$ multiplied by the sample size.

6.1. **Estimating the error distribution function F_e.** To study the finite sample performance of the estimator $\hat{\mathbb{F}}_e$, we conducted simulations of 300 runs using sample sizes 100, 200, 300 and 500 under the following data generation scheme. The covariates X are uniformly distributed on the interval $[-1, 1]$, and the location and scale functions are chosen as

$$m(x) = 1 - 2x + 1.25\cos\left(\pi x^2\right) \quad \text{and} \quad s(x) = 1 + 0.5\cos\left(\pi x\right), \qquad x \in [-1, 1]. \tag{6.1}$$

Here, the score function J is chosen as $J(p) = \mathbf{1}[0.05 < p < 0.95]/0.9$, for $0 \leq p \leq 1$. This choice results in m and s being the 5%-trimmed mean and scale functions, respectively. The error distribution is an appropriately centered and scaled standard normal distribution initially truncated at 2. An initial set of responses Y are then obtained using (1.1).

We work with a cure proportion function given by the logistic distribution function with standard scaling that has been centered at 7/4, which gives about 16% cured cases on average. Cure indicators are randomly generated based on this probability function, and whenever a cure indicator is equal to one we replace the corresponding value of Y with ∞. Finally, the response values Z are taken as the minimum of each Y and a censoring variable C, and a censoring indicator δ is set equal to one, if $Y \leq C$, or zero, otherwise.

To generate censoring variables C, we use a mixture distribution that ensures large valued censoring variables appear across the support of X as follows. The first component distribution is a normal distribution centered at 10, with variation 1/2. The second component distribution is a shifted version of (1.1), where m and s are as in (6.1), but now m is shifted up by 1/2 and the model errors are standard normally distributed (no truncation). The two component distributions in the mixture are assigned equal mixing probabilities. These choices result in about 18% censored values for the uncured cases. When we combine the censoring from both cases (cured and uncured) we expect a typical dataset generated in our simulations to present with about 31% of censored response values. This choice of censoring distribution ensures that the cure proportion function π can be well estimated from the data, which is crucial for the estimator $\hat{\mathbb{F}}_e$.

We select the bandwidth parameter for $\hat{\mathbb{F}}_e$ using the bootstrap method discussed in the previous section, and we numerically measure the performance of the estimator $\hat{\mathbb{F}}_e$ in two ways. Firstly, the asymptotic mean squared errors at t-values -2, -1, 0, 1 and 2 are simulated, where this performance metric is calculated by first obtaining the simulated mean squared errors and then multiplying these by the corresponding sample size. Secondly, the asymptotic integrated mean squared error is simulated, where this quantity is calculated similarly to the asymptotic mean squared errors, but now we integrate over t. These performance metrics are predicted to be stable from Theorem 1, which means that they should not fluctuate much with increasing sample size.

The results of the simulated asymptotic mean squared errors are displayed in Table 1, and the results of the simulated asymptotic mean integrated squared errors are given in Table 2. The values in Table 1 show the estimator $\hat{\mathbb{F}}_e$ has asymptotically stable pointwise mean squared errors (at the t-values -2, -1, 0, 1 and 2), and this metric clearly shows the estimator $\hat{\mathbb{F}}_e$ to have good performance on samples as small as 100. The values in Table 2 show a strong

100	200	300	500
0.443	0.432	0.436	0.446

TABLE 2. Simulated asymptotic integrated mean squared error values of $\hat{\mathbb{F}}_e$, that is, the simulated integrated mean squared error values of $\hat{\mathbb{F}}_e$ multiplied by the sample size.

mirroring of the conclusions drawn from Table 1, which indicate that $\hat{\mathbb{F}}_e$ is a good estimator of F_e even for samples as small as 100.

6.2. **Estimating the mean function** m**.** In this section, we investigate the performance of the estimator $\hat{m}$ introduced in Section 3 against two semiparametric estimators $\check{m}$ and $\tilde{m}$, which are defined as follows. The estimator $\check{m}(x) = (1, x)\check{\theta}$, for $x \in [-1, 1]$, is obtained by using the *smcure* package for the *R* computing language to form the estimate $\check{\theta}$, see Cai et al. (2012) for details. Here, $\mathbf{v}^T$ denotes the transpose of the vector $\mathbf{v}$.

To define $\tilde{m}$, write $\mathbf{x}_j = (1, X_j)^T$, for $j = 1, \ldots, n$ and define $U_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ as the matrix of covariates. Additionally, define the diagonal weight matrix

$$\hat{W}_n = \operatorname{diag}\big(\delta_1/(1 - \hat{\mathbb{F}}_c(\delta_1 Z_1 \,|\, X_1)), \ldots, \delta_n/(1 - \hat{\mathbb{F}}_c(\delta_n Z_n \,|\, X_n))\big)$$

and the vector of responses $\mathbf{z}_n = (Z_1, \ldots, Z_n)^T$. The estimator $\tilde{m}$ is then $\tilde{m}(x) = (1, x)\tilde{\theta}$, for $x \in [-1, 1]$, where $\tilde{\theta}$ is defined as the solution to the normal equations

$$U_n^T \hat{W}_n U_n \tilde{\theta} = U_n^T \hat{W}_n \mathbf{z}_n.$$

Note that $\tilde{\theta}$ is an estimator of the solution θ_0 to the least squares problem: find θ, such that

$$E\Big[\big(Y_u - m_\theta(X)\big)^2\Big] = E\left[\frac{\delta}{1 - F_c(\delta Z \,|\, X)}\big(Z - m_\theta(X)\big)^2\right]$$

has minimum value. In order to avoid the situation of $0/0$ in the formula above, we use as denominator $1 - F_c(\delta Z \,|\, X)$ rather than the more traditional $1 - F_c(Z \,|\, X)$, which does not affect θ_0. Further, our assumptions guarantee that there is at least one data point (X_j, Z_j, δ_j) for which $1 - \hat{\mathbb{F}}_c(Z_j \,|\, X_j) = 0$, with high probability, and we can assume without loss of generality that $\tau_{F_c}(\cdot) > 0$. Here, $m_\theta(x) = (1, x)\theta$, for $x \in [-1, 1]$, is a linear regression model with parameter θ. The estimator $\hat{\mathbb{F}}_c$ is defined as in (3.3), but with $1 - \delta_{(j)}$ in place of $\delta_{(j)}$, and consistently estimates the conditional distribution function F_c of the censoring variable C given X. Note that this estimator requires a bandwidth c_n to be chosen. Here we simply work with the undersmoothing choice $c_n = 2(n \log(n))^{-1/4}$, and we expect that $\tilde{\theta}$ is a root-n consistent estimator of θ_0.

We will consider two scenarios for m in (1.1), and s is taken as a constant function equal to 1. The first scenario is $m_1(x) = 1 - 2x$ and the second scenario is $m_2(x) = 1 - 2x + 1.25\cos(\pi x^2)$, for $x \in [-1, 1]$, as in the previous simulation study. We measure the performance of the estimators $\hat{m}$, $\check{m}$, and $\tilde{m}$ using the IMSE. In the case of m_1, we expect that $\check{m}$ and $\tilde{m}$ will perform better than $\hat{m}$ in the sense that the IMSE of these estimators will be smaller than that of $\hat{m}$. However, in the case of m_2, we expect that $\hat{m}$ will perform better than $\check{m}$ and $\tilde{m}$.

The simulation consists of 300 runs with samples of sizes 100, 200, and 300. The simulation data is generated as in the previous example, with m_1 and m_2 in place of m and 1 in place of s, for each scenario. The bandwidth used in the estimator $\hat{m}$ is selected by bootstrap as in Section 5. The results of this study are summarized in Table 3.

We can see that the linear model m_1 is well estimated by $\tilde{m}$, and that this estimator appears to be root-n consistent. The nonparametric estimator $\hat{m}$ does not perform as well as the semiparametric estimator $\tilde{m}$ with respect to the IMSE. Further, as expected, the conclusions change when we consider the nonlinear model m_2. Here, the nonparametric estimator $\hat{m}$ outperforms the semiparametric estimator $\tilde{m}$ in this case, which is explained by the fact that

Est \ n	100	200	300
$\hat{m}$	0.172	0.104	0.065
$\check{m}$	0.196	0.160	0.150
$\tilde{m}$	0.056	0.029	0.018
$\hat{m}$	0.837	0.491	0.282
$\check{m}$	1.820	1.730	1.713
$\tilde{m}$	1.611	1.570	1.548

TABLE 3. Simulated integrated mean squared error values of $\hat{m}$, $\check{m}$, and $\tilde{m}$. The values in the first three rows correspond to the regression m_1 and the remaining values correspond to the regression m_2.

$\tilde{m}$ is simply an inconsistent estimator of m_2 but that $\hat{m}$ is a consistent estimator of m_2. Unfortunately, the semiparametric estimator $\check{m}$, based on estimator $\check{\theta}$ obtained from the *smcure* package, performed the worst in both scenarios. We are at a loss to explain this behaviour.

7. Analysis of breast cancer data

In this section we illustrate the estimators of the components from model (1.1), i.e. π, m, s, and F_e, using a set of data obtained from frozen samples of tumour tissue stored at the Erasmus Medical Center (Rotterdam, Netherlands) of patients who were treated for lymph-node negative breast cancer during 1980–95. It has been observed that about 60%–70% of patients treated are cured [page 671 of Wang et al. (2005)]. These data were collected to study distant metastasis in lymph-node-negative breast cancer sufferers, where it is desirable to identify medical treatments that increase the amount of time before metastasis appear. See Wang et al. (2005) for a complete description of these data.

Our analysis considers three variables measured: the number of days before metastasis were detected (observed or censored), a censoring indicator, and the patient's age in years. There are 286 data points, and about 63% of the reported time lengths are censored at large values. The ages of the patients range between 26 and 83 years, with a median age of 52 years. The oldest patient with an uncensored response is 78 years. Since there were only two (censored) observations for patients older than 80, these cases were removed because the data was too sparse in this region to obtain good model estimates, and our analysis considers the remaining 284 patients. A possible cure effect is reasonable in this case, because 2,768 days (or just over 7.5 years) elapsed between the last observed case of detectable metastasis and the last (censored) observation time. We are interested in a nonparametric location-scale modelling of the log-transformed *time before detectable metastasis appear* by the patient's age that accounts for both the presence of censoring and an apparent cure effect.

For each nonparametric estimator, we use automated bandwidth selection by bootstrap as previously discussed to choose appropriate bandwidths. For the estimator $\hat{\mathbb{F}}_e$, the data-driven bandwidth is scaled by $\log^{-1/4}(n)$, with $n = 284$. Note that the different bandwidths for the estimators of π, m, s, and F_e are taken so that separate analysis of these quantities can be pursued. The score function J is chosen similarly to the previous section, but now is a smooth approximation of the indicator function $\mathbf{1}[0.05 < p < 0.95]$, for $0 \leq p \leq 1$, so that m and s represent approximate 5%-trimmed mean and scale functions, respectively. This choice allows the nonparametric estimators $\hat{m}$ and $\hat{s}$ to have smooth graphs, which does not occur when the score function is simply an indicator. As before, we work with the cosine smoothing kernel.

To construct pointwise confidence intervals for π, we use the *npcure* package for the *R* computing language [see López-Cheda et al. (2017) for further information].

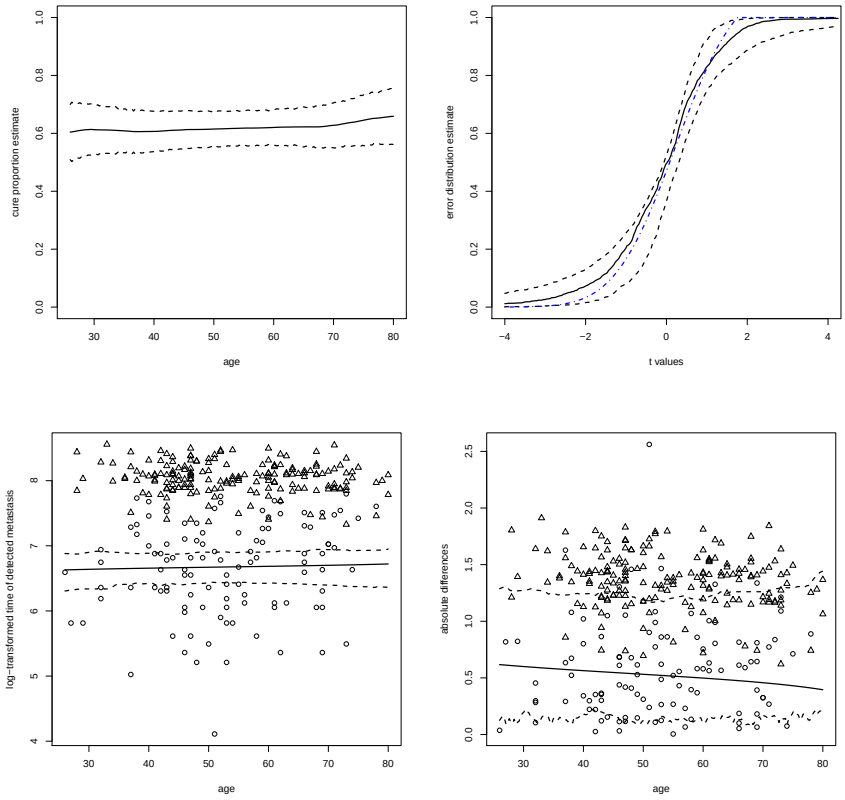

FIGURE 1. Non-model based CI's! Left to right, top to bottom: a plot of the cure proportion estimate (solid curve) overlaid by approximate 95% pointwise confidence intervals (dashed curves); a plot of the error distribution function estimate (solid curve) overlaid by approximate 95% pointwise confidence intervals (dashed curves) and a plausible truncated normal error distribution (blue dot-dashed curve); a plot of the observed responses (dots) and censored responses (triangles) overlaid by the regression estimates (solid curve) and approximate 95% pointwise confidence intervals (dashed curves); a plot of the absolute differences between the responses and corresponding regression estimates (dots indicate observed and triangles indicate censored response values) overlaid by scale estimates (solid curve) and approximate 95% confidence intervals (dashed curves).

Pointwise confidence intervals for m, s and F_e are constructed using the bootstrap discussed in Section 5 to approximate the sampling distributions of $\hat{m}$, $\hat{s}$, and $\hat{\mathbb{F}}_e$. The resulting confidence intervals are simply taken as the 2.5^{th} and 97.5^{th} percentiles from these distributions.

Pointwise confidence intervals for m, s, and F_e are built using a model-based bootstrap as follows. Here, bootstrap uncured responses are constructed using model (1.1), where m is replaced by the estimator $\hat{m}$, s is replaced by the estimator $\hat{s}$, and the model errors are sampled independently from $\hat{\mathbb{F}}_e$. The remaining procedures for generating the bootstrap data are as in Section 5. The sampling distributions of $\hat{m}$, $\hat{s}$, and $\hat{\mathbb{F}}_e$ are approximated using 300 bootstrap data sets. The resulting confidence intervals are simply taken as the 2.5^{th} and 97.5^{th} percentiles from these distributions.

Figure 2 shows plots of the cure proportion estimates, error distribution function estimates, regression estimates, and scale estimates. The cure proportion estimates appear to fluctuate between 0.55 and 0.65 as age increases, and it appears that the $60\% - 70\%$ cure proportion reported in Wang et al. (2005) is accurate. Inspecting the shape of the estimate of the cure

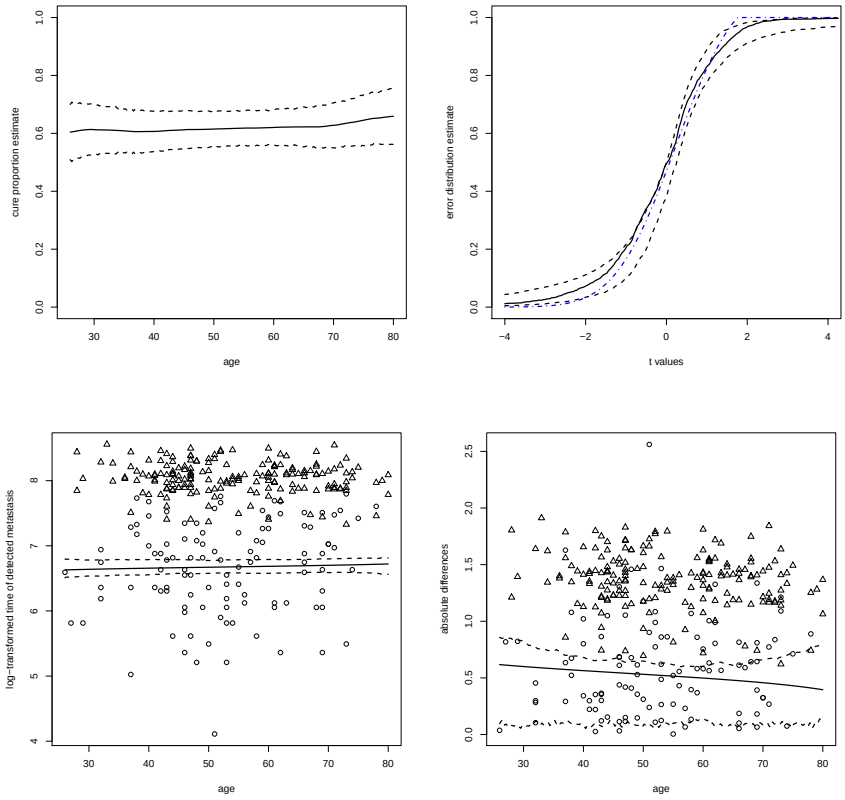

FIGURE 2. Model based CI's! Left to right, top to bottom: a plot of the cure proportion estimate (solid curve) overlaid by approximate 95% pointwise confidence intervals (dashed curves); a plot of the error distribution function estimate (solid curve) overlaid by approximate 95% pointwise confidence intervals (dashed curves) and a plausible truncated normal error distribution (blue dot-dashed curve); a plot of the observed responses (dots) and censored responses (triangles) overlaid by the regression estimates (solid curve) and approximate 95% pointwise confidence intervals (dashed curves); a plot of the absolute differences between the responses and corresponding regression estimates (dots indicate observed and triangles indicate censored response values) overlaid by scale estimates (solid curve) and approximate 95% confidence intervals (dashed curves).

proportion curve and the confidence intervals, it appears that the cure proportion curve is well described by a linear function. Similarly, the regression and scale curves appear to be well described by linear functions. Finally. the error distribution function from this regression model appears to be well-described by a normal distribution truncated near 1.75 (comparing the solid black curve with the blue dot-dashed curve). We performed a bootstrap based Kolmogorov-Smirnov goodness-of-fit test as in Remark 2, for the null hypothesis $H_0 : F_e = F_{e,0}$, where $F_{e,0}$ is the distribution function given by centering and scaling a standard normal distribution initially truncated at 1.3 so that (2.4) is satisfied and $\tau_{F_{e,0}} \approx 1.76$ after centering and scaling. The bootstrap based p-value of the test is 0.517, which indicates that assuming a truncated normal error distribution is reasonable in light of the data.

ACKNOWLEDGEMENTS

The authors would like to thank and acknowledge the following sources of financial support. This research has been supported by the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409), the IAP research network grant nr. P7/06 of the Belgian government (Belgian science policy), the Collaborative Research Center "Statistical modelling

of nonlinear dynamic processes" (SFB 823, Project C1) of the German Research Foundation, and the Bundesministerium für Bildung und Forschung (project "MED4D: Dynamic medical imaging: Modeling and analysis of medical data for improved diagnosis, supervision, and drug development").

References

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1989). *Statistical Modelling in GLIM*. Clarendon Press, New York.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical Report*.

Boag, J. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 11(1):15–44.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An r-package for estimating semiparametric mixture cure models. *Comput. Methods Programs Biomed.*, 108(3):1255–1260.

Collett, D. (1994). *Modelling Survival Data in Medical Research*. CRC Monographs on Statistics & Applied Probability. CRC Press, Boca Raton.

Dabrowska, D. (1987). Non-parametric regression with censored survival time data. *Scand. J. Statist.*, 14(3):181–197.

Farewell, V. (1986). Mixture models in survival analysis: Are they worth the risk? *Canad. J. Statist.*, 14(3):257–262.

González-Manteiga, W. and Crujeiras, R. (2013). An updated review of goodness-of-fit tests for regression models. *TEST*, 22(3):361–411.

Harris, E. and Albert, A. (1990). *Survivorship Analysis for Clinical Studies*. Statistics: A Series of Textbooks and Monographs. CRC Press, Boca Raton.

Haybittle, J. (1959). The estimation of the proportion of patients cured after treatment for cancer of the breast. *Br. J. Radiol.*, 32(383):725–733.

Haybittle, J. (1965). A two-parameter model for the survival curve of treated cancer patients. *J. Amer. Statist. Assoc.*, 60(309):16–26.

Kuk, A. and Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541.

Lawless, J. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, New York.

Li, G. and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Ann. Inst. Statist. Math.*, 53(4):708–729.

López-Cheda, A., Cao, R., Jácome, M., and Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput. Statist. Data Anal.*, 105:144–165.

Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Ann. Inst. Statist. Math.*, 60(3):545–574.

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Stat. Sin.*, 20(2):661–674.

Patilea, V. and Van Keilegom, I. (2017). A general approach for cure models in survival analysis. *Ann. Statist.*, page (under revision).

Portier, F., Van Keilegom, I., and El Ghouch, A. (2017). On an extension of the promotion time cure model. *Ann. Statist.*, page (under revision).

Sinha, D., Chen, M., and Ibrahim, J. (2003). Bayesian inference for survival data with a surviving fraction. In *Crossing Boundaries: Statistical Essays in Honor of Jack Hall*, Lecture Notes–Monograph Series, pages 117–138. Institute of Mathematical Statistics, Beachwood, OH.

Stone, C. (1977). Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–620.

Sy, J. and Taylor, J. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Taylor, J. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, 51(3):899–907.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, 54(4):1508–1516.

Tsodikov, A., Ibrahim, J., and Yakovlev, A. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *J. Am. Stat. Assoc.*, 98(464):1063–1078.

Van Keilegom, I. and Akritas, M. (1999). Transfer of tail information in censored regression models. *Ann. Statist.*, 27(5):1745–1784.

Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M., Yu, J., Jatkoe, T., Berns, E., Atkins, D., and Foekens, J. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679.

Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canad. J. Statist.*, 42(1):1–17.

Yakovlev, A., Cantor, A., and Shuster, J. (1994). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Stat. Med.*, 13(9):983–986.

Yakovlev, A. and Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Series in Mathematical Biology and Medicine. World Scientific, Singapore.

Yin, G. and Ibrahim, J. (2005). Cure rate models: A unified approach. *Canad. J. Statist.*, 33(4):559–570.

Zeng, D., Yin, G., and Ibrahim, J. (2006). Semiparametric transformation models for survival data with a cure fraction. *J. Amer. Statist. Assoc.*, 101(474):670–684.