Goodness-of-fit tests for the cure rate in a mixture cure model

Müller U, Van Keilegom I.

Goodness-of-fit tests for the cure rate in a mixture cure model

Ursula U. Müller and Ingrid Van Keilegom

ABSTRACT. We consider models for time-to-event data that allow that an event, e.g. a relapse of a disease, never occurs for a certain percentage p of the population, called the “cure rate”. We suppose that these data are subject to random right censoring and we model the data using a mixture cure model, in which the survival function of the uncured subjects is left unspecified. The aim is to test whether the cure rate p (as a function of the covariates) satisfies a certain parametric model. To do so, we propose a test statistic that is inspired by a goodness-of-fit test for a regression function by Härdle and Mammen. We show that the statistic is asymptotically normally distributed under the null hypothesis that the model is correctly specified and under local alternatives. A bootstrap procedure is proposed to implement the test. The good performance of the approach is confirmed with simulations: we test whether the logistic curve is an adequate model for p , and also whether a cure proportion p exists at all. For an illustration we apply the test to data on the times between first and second birth.

Key words: Asymptotic normality, Beran estimator, bootstrap, censoring, existence of cure fraction, Kaplan-Meier, logistic model, kernel estimator.

1. Introduction

The classical approach to time-to-event data assumes that an event will occur eventually. This is appropriate in many situations but not all, for example the event that a person marries or becomes a parent never occurs for a certain percentage of the population. The same is true of the event that an unemployed person finds a job, since some people always stay unemployed. The proportion of subjects for which the event does not happen is often called “cure rate” and is of particular interest in health sciences, especially if there are covariates that are likely to have an effect on the cure rate. In this case it would be useful to have a parametric model for the cure rate as a function of the covariates that is easy to interpret. We want to find out whether such a model fits the data well.

Let T denote the time until an event of interest occurs. The random variable T is typically called “survival time” or “failure time”. Write $S(t|x) = P(T > t|X = x)$ ($t \geq 0$) for the conditional survival function given that the covariate X equals x . We set $T = \infty$

if a subject is cured. Then $S_0(t|x) = P(T > t|T < \infty, X = x)$ is the conditional survival function given the covariate x and given that the subject is not cured.

Typical for survival time data is the presence of censoring, e.g. when patients leave a study before an event occurs. This makes estimation of the cure rate difficult because it is not clear whether they were cured or not when they left. In this paper we consider cure models for survival time data with right censoring. There are two types of cure models that are typically considered in the literature. The first one is the so-called *mixture cure model*,

$$(1.1) \quad S(t|x) = P(T > t|X = x) = p(x) + (1 - p(x))S_0(t|x), \quad t \geq 0,$$

where $p(\cdot)$ denotes the cure rate, which is the type of model we are interested in. The second type is the *promotion time cure model*,

$$S(t|x) = \exp\{-\theta(x)F_0(t)\}, \quad t \geq 0,$$

where F_0 is a baseline distribution function, and $\theta(x)$ is usually modelled as $\theta(x) = \exp(\beta_0 + \beta_1 x)$. Parametric estimation in the latter model is, for example, discussed in Yakolev, Cantor and Shuster (1994); for semiparametric methods see e.g. Tsodikov (1998, 2003).

There are various articles studying the estimation of the mixture cure model (1.1). Most papers assume a logistic model for the cure rate $p(x)$, while various model assumptions have been made on the survival function $S_0(t|x)$. Parametric estimators are, for example, given in Boag (1949) and Farewell (1982). A semiparametric approach based on a proportional hazards model is provided by Kuk and Chen (1992), Sy and Taylor (2000), and Lu (2008), just to name a few, whereas a completely nonparametric estimation approach for $S_0(t|x)$ is given in Patilea and Van Keilegom (2017+). The literature on cure models that do not assume a logistic model for $p(x)$ is much more sparse. Maller and Zhou (1992) propose a nonparametric estimator of the cure rate in a model that does not involve covariates; Laska and Meisner (1992) discuss nonparametric inference with discrete covariates, whereas Xu and Peng (2014) propose a conditional Kaplan-Meier estimator which involves one continuous covariate, and which we will use here to construct goodness-of-fit tests about the form of the cure rate.

As is clear from the above discussion, most papers in the literature assume either that a cure rate does not exist, or that it does exist and follows a logistic model. However, in practice it is possible that neither of these two model assumptions is correct. First, the possible existence of a cure rate has been discussed in detail in the literature on cure models (see e.g. the above mentioned papers). Second, there is no reason to believe that the cure rate is always a monotone function of x , let alone that it satisfies a logistic model. There are many other (monotone or non-monotone) models for $p(x)$ that could fit the data much better than a logistic model. All this motivates us to consider in this paper tests for the parametric form of the cure rate function. Tests for the logistic model and for the existence of a cure rate will be notable special cases. Tests for the existence of a cure rate have been proposed by Maller and Zhou (1994) and by Li, Tiwari and Guha (2007). However, neither approach incorporates covariates. To the best of our knowledge, the literature does not

contain any test when covariates are present. Hence, our test is the first to allow the cure rate to depend on a (one-dimensional) covariate. The study of our test is also a necessary first hurdle to be taken before the case of multi-dimensional covariates can be considered. The latter case requires, however, a semi- or nonparametric estimator of $p(x)$ when X is multi-dimensional, which is yet to be developed in the literature.

This paper is organised as follows. In Section 2 we describe the model and the test statistic, which is an adaptation of the statistic proposed by Härdle and Mammen (1993), who test for the parametric form of a regression function. Section 2 also contains our asymptotic result, Theorem 2.1, which states the limiting normality of the test statistic, including assumptions. For the implementation of the test we recommend a bootstrap procedure. This is explained in Section 3, where we also state the consistency of the bootstrap in Theorem 3.1. In Section 4 we present the results of a simulation study. We test whether the logistic model is appropriate for p , and we also consider the important problem whether a cure fraction p exists or not. The results are quite encouraging. In Section 5 we apply our method to data on the time between first and second births in Norway. Some concluding remarks are in Section 6. The proofs of Theorems 2.1 and 3.1 are given at the end of the paper in Sections 7 and 8.

2. Model and theoretical results

Let T be a survival time and C a censoring time. The observations are independent copies (Y_i, δ_i, X_i) , $i = 1, \dots, n$, of

$$Y = \min(T, C), \quad \delta = 1(T \leq C), \quad X,$$

where X is a covariate. We consider the mixture cure model (1.1), i.e.

$$S(t|x) = p(x) + (1 - p(x))S_0(t|x), \quad t \geq 0.$$

The usual model for the cure rate p is the logistic model, i.e. $p(x) = p_\theta(x) = 1/(1 + \exp(\theta_0 + \theta_1 x))$, with parameter vector $\theta = (\theta_0, \theta_1)$. We want to test whether the logistic model, or any other parametric model, is appropriate. The hypotheses are

$$H_0 : \exists \theta \in \Theta : p = p_\theta \quad \text{vs.} \quad H_1 : p \neq p_\theta \quad \forall \theta \in \Theta,$$

where Θ is a finite-dimensional parameter space and p_θ a function that is known up to a parameter vector $\theta \in \Theta$. The test statistic we propose is a weighted L_2 distance between a nonparametric estimator of $p(x)$ and a parametric estimator obtained under H_0 . This is in the same spirit as the test statistic proposed by Härdle and Mammen (1993) in the context of tests for the parametric form of a regression function. More precisely, we define

$$(2.1) \quad \mathcal{T}_n = n\sqrt{h} \int \{\hat{p}(x) - p_{\hat{\theta}}(x)\}^2 \pi(x) dx,$$

where $\hat{p}$ is a nonparametric estimator with bandwidth h , $\hat{\theta}$ estimates θ , and $\pi(\cdot)$ is a given density function. For practical purposes we recommend an empirical version $\tilde{\mathcal{T}}_n$ of the special

case of $\mathcal{T}_n$ where $\pi(\cdot)$ equals the covariate density $f(\cdot)$ (cf. Cao and González-Manteiga, 2008),

$$(2.2) \quad \tilde{\mathcal{T}}_n = n\sqrt{h} \frac{1}{n} \sum_{i=1}^n \{\hat{p}(X_i) - p_{\hat{\theta}}(X_i)\}^2.$$

This statistic naturally puts more weight on covariates that are more likely.

Both statistics, $\mathcal{T}_n$ and $\tilde{\mathcal{T}}_n$, require a nonparametric estimator $\hat{p}$ and a parametric estimator $p_{\hat{\theta}}$ of the cure rate. We use the nonparametric estimator $\hat{p}$ introduced in Xu and Peng (2014), which is described below. In order to handle the finite-sample bias we choose, similarly to Härdle and Mammen (1993) and Cao and González-Manteiga (2008), the parametric estimator $p_{\hat{\theta}}$ as follows. First we create a new data set $(X_1, \hat{p}(X_1)), \dots, (X_n, \hat{p}(X_n))$, where $\hat{p}$ is again the Xu-Peng estimator. Then we use a parametric estimator for θ to obtain $p_{\hat{\theta}}$ from the new data set; see Remark 1 for more details.

The nonparametric cure rate estimator by Xu and Peng is the conditional Kaplan-Meier estimator $\hat{S}$ proposed by Beran (1981) evaluated at the largest uncensored survival time, i.e.

$$\hat{p}(x) = \hat{S}(Y_{(n)}^1 | x) \quad \text{with } Y_{(n)}^1 = \max_{i: \delta_i=1} Y_i$$

and

$$\hat{S}(t|x) = 1(t \leq Y_{(n)}) \prod_{j=1}^n \left(\frac{1 - \sum_{i=1}^n 1(Y_i \leq Y_j) w_{hi}(x)}{1 - \sum_{i=1}^n 1(Y_i < Y_j) w_{hi}(x)} \right)^{1(Y_j \leq t; \delta_j=1)}.$$

Here $w_{h1}(x), \dots, w_{hn}(x)$ are non-negative weights which add up to one, and $Y_{(i)}$ is the i -th order statistic. For simplicity we use the familiar Nadaraya-Watson weights,

$$w_{hi}(x) = \frac{K_h(x - X_i)}{\sum_{j=1}^n K_h(x - X_j)}, \quad K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right),$$

where h is a bandwidth and K a kernel function. Note that $\hat{p}(x) = \hat{S}(\infty|x)$ since $\hat{S}(\cdot|x)$ is constant for values to the right of $Y_{(n)}^1$.

To ensure that $\hat{p}(x)$ is consistent for $p(x)$, we assume that

$$(2.3) \quad \tau_0(x) = \inf\{t : S_0(t|x) = 0\} < \inf\{t : G(t|x) = 0\} \text{ for every } x,$$

which is slightly weaker than the corresponding condition by Xu and Peng (who consider the supremum with respect to x on the left-hand side). Here G is the survival function of the censoring time C . This implies that subjects censored after the time point $\tau_0(x)$ are known to be cured. Since $Y_{(n)}^1$ converges to $\sup_x \tau_0(x)$ in probability (see Xu and Peng, 2014), the assumption therefore guarantees, asymptotically, that all subjects censored after the largest observed survival time are cured.

Before we state our main result, the limiting normality of $\mathcal{T}_n$, we provide some notation and a list of technical assumptions.

In the following we will write

$$H(t|x) = P(Y \leq t|X = x), \quad H_1(t|x) = P(Y \leq t, \delta = 1|X = x)$$

and

$$(2.4) \quad \zeta_i(t|x) = \frac{1(Y_i \leq t, \delta_i = 1)}{1 - H(Y_i|x)} - \int_0^t \frac{1(Y_i \geq s)}{\{1 - H(s|x)\}^2} dH_1(s|x), \quad i = 1, \dots, n.$$

Let $\zeta(t|x)$ denote the version of $\zeta_i(t|x)$ that involves the base observations (Y, δ) instead of (Y_i, δ_i) . Further set

$$\mu_{xz}(t) = E\{\zeta(\tau_0(x)|x)\zeta(\tau_0(z)|z)|X = t\}$$

and note that $\mu_{xx}(t) = E\{\zeta^2(\tau_0(x)|x)|X = t\}$.

To prove the limit theorem we will need the following assumptions.

- (A1) X is quasi-uniform, i.e. X has a density f with support $[0, 1]$ and f is bounded and bounded away from zero on $[0, 1]$. The distribution function F has bounded second and third derivatives f' and f'' .
- (A2) The kernel function K is a symmetric probability density with compact support and two bounded derivatives.
- (A3) (i) There is a continuous nondecreasing and bounded function L_1 with $L_1(0) = 0$ and $|H(t_1|x) - H(t_2|x)| \leq |L_1(t_1) - L_1(t_2)|$ for all x, t_1, t_2 .
(ii) The first two derivatives of $H(t|x)$ and $H_1(t|x)$ with respect to x exist and are bounded uniformly for all x and t .
(iii) There are continuous nondecreasing and bounded functions L_2 and L_3 with $L_2(0) = L_3(0) = 0$ and

$$\begin{aligned} \left| \frac{\partial H(t_1|x)}{\partial x} - \frac{\partial H(t_2|x)}{\partial x} \right| &\leq |L_2(t_1) - L_2(t_2)| \text{ for all } x, t_1, t_2, \\ \left| \frac{\partial H_1(t_1|x)}{\partial x} - \frac{\partial H_1(t_2|x)}{\partial x} \right| &\leq |L_3(t_1) - L_3(t_2)| \text{ for all } x, t_1, t_2. \end{aligned}$$

- (A4) The functions $H(t|x)$, $S_0(t|x)$ and $G(t|x)$ are continuously differentiable in t , for any value of x , and have bounded second derivatives with respect to x , for any given value t .
- (A5) The function $\mu_{xz}(z)$ is continuous in z for any given value of x .
- (A6) The cure rate $p(\cdot)$ is continuously differentiable; the parametric model $p_\theta(\cdot)$ is Lipschitz in θ uniformly in x , i.e. for all $\theta, \theta' \in \Theta$,

$$|p_\theta(x) - p_{\theta'}(x)| \leq L(x)|\theta - \theta'| \quad \text{with } \sup_x L(x) < \infty.$$

- (A7) The estimator $\hat{\theta}$ is $\sqrt{n}$ -consistent, i.e. $\hat{\theta} - \theta = O_p(n^{-1/2})$.

Note that assumption (A3)(i) (and likewise (A3)(iii)) is verified for a wide class of distributions. For instance, if $H(\cdot|x)$ is an exponential distribution with mean a_x^{-1} , then (A3)(i) is satisfied for $L(t) = 1 - \exp(-mt)$ with $m = \inf_x a_x$, provided $m > 0$ and $\sup_x a_x < \infty$.

REMARK 1 (ESTIMATION OF θ). Theorem 2.1 only requires that there is some $\sqrt{n}$ -consistent estimator for the parameter θ in the cure rate model p_θ . The most natural estimator is obtained by maximising the following logistic log-likelihood, based on the artificial data $(X_i, \hat{p}(X_i))$, $i = 1, \dots, n$:

$$\hat{\theta} = \operatorname{argmax}_\theta \sum_{i=1}^n \left\{ \hat{p}(X_i) \log p_\theta(X_i) + (1 - \hat{p}(X_i)) \log(1 - p_\theta(X_i)) \right\}.$$

The root- n consistency of $\hat{\theta}$ follows from Theorems 1 and 2 in Chen, Linton and Van Keilegom (2003).

We now state our main result, the limiting normality of the test statistic $\mathcal{T}_n = n\sqrt{h} \int \{\hat{p}(x) - p_{\hat{\theta}}(x)\}^2 \pi(x) dx$; see equation (2.1). The proof is given in Section 7. Recommendations on bandwidth selection are provided in Remark 3 in the next section.

THEOREM 2.1. *Suppose assumptions (A1)-(A7) and (2.3) hold true and that the bandwidth $h = h_n$ satisfies $nh^3 \rightarrow \infty$ and $nh^5/(\log n) = O(1)$ as $n \rightarrow \infty$. Then, under the null hypothesis, the statistic $\mathcal{T}_n$ from equation (2.1) is asymptotically normally distributed,*

$$\mathcal{T}_n - b_h \xrightarrow{d} N(0, V).$$

The asymptotic bias is $b_h = h^{-1/2} R(K) \int_0^1 \frac{p^2(x)\pi(x)}{f(x)} \mu_{xx}(x) dx = O(h^{-1/2})$, with $R(g) = \int g^2(t) dt$. The asymptotic variance is

$$V = 2K^{(4)}(0) \int \left(\frac{p^2(x)\pi(x)\mu_{xx}(x)}{f(x)} \right)^2 dx,$$

where $K^{(4)}$ denotes the fourth convolution product of K .

Under local alternatives of the form $p(x) = p_\theta(x) + n^{-1/2}h^{-1/4}\Delta_n(x)$, with $\Delta_n(x)$ bounded uniformly in x and n , $\mathcal{T}_n$ is asymptotically normally distributed,

$$\mathcal{T}_n - b_h - \int_0^1 \Delta_n(x)^2 \pi(x) dx \xrightarrow{d} N(0, V)$$

(which reduces to the above statement if H_0 is true and $\Delta_n \equiv 0$).

Note that the asymptotic limit is degenerate under the hypothesis that there is no cure fraction ($p \equiv 0$). It should also be noted that the test can only detect deviations from the null hypothesis of order larger than $n^{-1/2}$: the assumptions on the bandwidths, $nh^3 \rightarrow \infty$ and $nh^5/(\log n) = O(1)$, are, for example, satisfied if $h = n^{-1/5}$. For this choice the deviation from H_0 has order $n^{-1/2}h^{-1/4} = n^{-9/20}$, i.e. larger than $n^{-1/2}$.

REMARK 2 (ASYMPTOTIC BEHAVIOUR OF $\tilde{\mathcal{T}}_n$). The second test statistic $\tilde{\mathcal{T}}_n$ from equation (2.2) can be regarded an empirical version of $\mathcal{T}_n$ for the special case $\pi(\cdot) = f(\cdot)$, i.e.

with $\pi(x)$ replaced by the empirical density $\hat{f}(x) = 1/n$. We expect that $\mathcal{T}_n$ and $\tilde{\mathcal{T}}_n$ are asymptotically equivalent, i.e. with $\hat{F}$ denoting the empirical distribution function,

$$\tilde{\mathcal{T}}_n - \mathcal{T}_n = n\sqrt{h} \int \{\hat{p}(x) - p_{\hat{\theta}}(x)\}^2 d\{\hat{F}(x) - F(x)\} = o_p(1).$$

Then, applying Theorem 2.1 to the case $\pi(\cdot) = f(\cdot)$, we obtain that $\tilde{\mathcal{T}}_n$ is asymptotically normally distributed under H_0 ,

$$\tilde{\mathcal{T}}_n - \tilde{b}_h \xrightarrow{d} N(0, \tilde{V}),$$

with asymptotic bias $\tilde{b}_h = h^{-1/2}R(K) \int_0^1 p^2(x)\mu_{xx}(x) dx = O(h^{-1/2})$ and asymptotic variance

$$\tilde{V} = 2K^{(4)}(0) \int \{p^2(x)\mu_{xx}(x)\}^2 dx.$$

3. Bootstrap procedure

Härdle and Mammen (1993) noticed that a test based on the asymptotic result (using normal quantiles) cannot be recommended due to the slow convergence rate and since the approximation is only of first order. Instead they suggest a bootstrap approach, which is noted for its higher order accuracy. For the same reason we will not use the normal approximation to implement the test. Instead we propose the following bootstrap procedure to approximate the critical value of the test statistic:

- (1) Use the data and a pilot bandwidth $h = h_0$ to calculate, for $i = 1, \dots, n$, the nonparametric estimator $\hat{p}_{h_0}(X_i)$.
- (2) Fit the parametric model p_{θ} for p into the new ‘data’ that were obtained in the previous step, i.e. use the $\hat{p}_{h_0}(X_i)$ ’s to calculate the parametric estimators $p_{h_0, \hat{\theta}}(X_i)$, $i = 1, \dots, n$.
- (3) For $b = 1, \dots, B$ we proceed as follows.
 - Generate binary bootstrap data $Z_{i,b}^*$, $i = 1, \dots, n$, such that $P^*(Z_{i,b}^* = 0) = p_{h_0, \hat{\theta}}(X_i)$.
 - Write $\hat{S}_{h_0}(\cdot|x)$ for $\hat{S}(\cdot|x)$ with bandwidth $h = h_0$. For $i = 1, \dots, n$ set $T_{i,b}^* = \infty$ (or a large number) if $Z_{i,b}^* = 0$; generate

$$T_{i,b}^* \sim \frac{\hat{S}_{h_0}(\cdot|X_i) - \hat{p}_{h_0}(X_i)}{1 - \hat{p}_{h_0}(X_i)} \quad \text{if } Z_{i,b}^* = 1.$$

- For $i = 1, \dots, n$ generate $C_{i,b}^* \sim \hat{G}_{h_0}(\cdot|X_i)$, where $\hat{G}_{h_0}(\cdot|x)$ is given by

$$\hat{G}_{h_0}(t|x) = 1(t \leq Y_{(n)}) \prod_{j=1}^n \left(\frac{1 - \sum_{i=1}^n 1(Y_i \leq Y_j) w_{h_0 i}(x)}{1 - \sum_{i=1}^n 1(Y_i < Y_j) w_{h_0 i}(x)} \right)^{1(Y_j \leq t; \delta_j=0)}.$$

- The bootstrap data are $(Y_{i,b}^*, \delta_{i,b}^*, X_i)$, $i = 1, \dots, n$, with

$$Y_{i,b}^* = \min(T_{i,b}^*, C_{i,b}^*), \quad \delta_{i,b}^* = 1(T_{i,b}^* \leq C_{i,b}^*).$$

Use the bootstrap data and a bandwidth h to calculate the nonparametric bootstrap estimator $\hat{p}_{hh_0,b}^*(X_i)$ for $p(X_i)$, $i = 1, \dots, n$.

- Analogously to step (2), use the nonparametric bootstrap estimates $\hat{p}_{hh_0,b}^*(X_i)$ to obtain the parametric estimates $p_{hh_0,\hat{\theta}^*,b}(X_i)$, $i = 1, \dots, n$.

(4) Order $\mathcal{T}_{n,1}^*, \dots, \mathcal{T}_{n,B}^*$, where

$$\mathcal{T}_{n,b}^* = n\sqrt{h} \frac{1}{n} \sum_{i=1}^n \{\hat{p}_{hh_0,b}^*(X_i) - p_{hh_0,\hat{\theta}^*,b}(X_i)\}^2, \quad b = 1, \dots, B,$$

and select the $[(1 - \alpha)B]$ -th order statistic as the critical value of the test.

For the following theorem, which shows the consistency of the above bootstrap procedure, we need to introduce additional assumptions. In what follows P^* denotes the probability with respect to the bootstrap data and conditionally on the original data.

(A8) The estimator $\hat{\theta}^*$ is $\sqrt{n}$ -consistent, i.e. $\hat{\theta}^* - \hat{\theta} = O_{p^*}(n^{-1/2})$ a.s.

(A9) The derivatives $\frac{\partial^2}{\partial x^2} H(t|x)$, $\frac{\partial^2}{\partial t^2} H(t|x)$, $\frac{\partial^2}{\partial x \partial t} H(t|x)$, $\frac{\partial^2}{\partial x^2} H_1(t|x)$, $\frac{\partial^2}{\partial t^2} H_1(t|x)$ and $\frac{\partial^2}{\partial x \partial t} H_1(t|x)$ exist and are continuous in (x, t) .

THEOREM 3.1. *Suppose assumptions (A1)-(A9) and (2.3) hold true and that the bandwidths $h = h_n$ and $h_0 = h_{0n}$ satisfy $h = Cn^{-1/5}$ for some $C > 0$, $h_0 \rightarrow 0$, $nh_0^5/\log n \rightarrow \infty$ and $(nh_0^5/\log n)(h/h_0) = O(1)$ as $n \rightarrow \infty$. Then,*

$$\sup_t \left| P^*(\mathcal{T}_n^* \leq t) - \Phi\left(\frac{t - b_h}{V}\right) \right| = o(1) \text{ a.s.},$$

where Φ is the distribution function of a standard normal random variable. In particular we have under H_0 ,

$$\sup_t \left| P^*(\mathcal{T}_n^* \leq t) - P(\mathcal{T}_n \leq t) \right| = o(1) \text{ a.s.}$$

The proof of this result is given in Section 8.

REMARK 3 (BANDWIDTH SELECTION). The bandwidths h and h_0 can be chosen as $h = c_h n^{-.20}$ and $h_0 = c_h n^{-.11}$ (i.e. $h_0 = h \times n^{.09}$) with, for example, $c_h = 1, 1.5$ or 2 . The choice for h is motivated by the conditions stated in Theorem 2.1, whereas the choice of h_0 comes from Li and Datta (2001) (see their Remark 2.1, p. 714). For practical applications we recommend selecting the bandwidth h of the nonparametric estimator $\hat{p}(x)$ via cross-validation, which is data-driven and straightforward to apply. For the pilot bandwidth one can simply take $h_0 = h \times n^{.09}$, in accordance with Li and Datta (2001).

4. Simulations

We now check the performance of our test procedure with simulations, using R and the following model. Let X be a uniform random variable on the interval $[-1, 1]$ and, for a given value x of X , generate a survival time T from model (1.1) with conditional survival function

$$S_0(t|x) = \begin{cases} 1 - 0.98 \frac{F(t|x)}{F(\tau|x)} - 0.02I(t = \tau) & 0 \leq t \leq \tau \\ 0 & t > \tau, \end{cases}$$

where $F(t|x) = 1 - \exp(-(t/b_x)^{a_x})$ ($t \geq 0$) is a Weibull distribution function with shape parameter $a_x = (1+x)^{c_T}$ (for some $c_T \geq 0$) and scale parameter $b_x = \exp(-(1+x)/(2a_x))$. Further $\tau = F^{-1}(0.9)$, with $F(t) = 1 - \exp(-(t/b)^a)$, $a = E(a_X)$ and $b = E(b_X)$, which means that $\tau = b(\log 10)^{1/a}$. We truncate at τ to make sure that condition (2.3) is satisfied. In addition, we generate C independently of X from an exponential distribution with mean 1.5, i.e. $G(t|x) = G(t) = \exp(-2t/3)$ for $t \geq 0$.

Standard logistic vs. extended logistic. For our first illustration we assume a (standard) logistic model under the null hypothesis, i.e.

$$H_0 : \quad \exists \theta \in \Theta : \quad p(x) = p_\theta(x) = \frac{1}{1 + \exp(\theta_0 + \theta_1 x)} \quad \forall x,$$

and study the performance of our test if in fact an extended logistic model is true, i.e. if

$$p(x) = \frac{1}{1 + \exp(\theta_0 + \theta_1 x - \theta_2 x^2)} \neq p_\theta(x),$$

for appropriate values of θ_0, θ_1 and $\theta_2 \neq 0$. Note that $\theta_2 = 0$ yields the standard logistic model (H_0). The proportion of censoring $p_{cens} = P(C < T)$ and the average cure proportion $p_{cure} = E[p(X)]$ are given in Table 1 for several values of c_T , θ_0 , θ_1 and θ_2 . Table 1 also shows the rejection proportions of the test statistic $\tilde{T}_n$ defined in (2.2) for each of these settings. We work with $n = 50, 100$ and 200 , and with bandwidths h and h_0 as stated in Remark 3 above. We used the R package *np*, which provides the cross-validation bandwidth for estimating a conditional distribution. For the kernel function K we use the Epanechnikov kernel on $[-1, 1]$.

The figures in Table 1 are based on 500 simulation runs, where for each sample the bootstrap critical value was obtained from 500 bootstrap samples. The nominal level is $\alpha = 0.05$. The table rows corresponding to $\theta_2 = 0$ show that under the null hypothesis that the logistic model is true the level is well respected, even for sample size $n = 50$. Table 1 also shows that the rejection proportion increases with the sample size and with the value of θ_2 , as is to be expected. The constant c_h , which determines the value of the bandwidths h and h_0 , does not seem to have an important impact on the rejection proportion. The data-driven cross-validation method yields very similar results, so both approaches to bandwidth selection appear to be acceptable. These findings are true for both values of c_T and for both choices of the vector (θ_0, θ_1) considered in Table 1.

c_T	θ_0	θ_1	θ_2	p_{cens}	p_{cure}	$n = 50$				$n = 100$				$n = 200$			
						c_h			CV	c_h			CV	c_h			CV
						1	1.5	2		1	1.5	2		1	1.5	2	
0	1	1	0	.46	.28	.06	.05	.06	.06	.05	.04	.05	.05	.05	.04	.05	.04
			1	.51	.35	.08	.08	.08	.08	.10	.10	.11	.10	.16	.22	.21	.18
			2	.56	.42	.19	.20	.20	.19	.36	.39	.40	.41	.64	.66	.69	.67
			4	.66	.54	.59	.65	.65	.64	.90	.92	.93	.92	1	1	1	1
	.5	.5	0	.54	.38	.05	.06	.04	.05	.06	.06	.06	.05	.05	.05	.05	.05
			1	.60	.46	.08	.08	.08	.07	.15	.17	.17	.17	.26	.28	.29	.29
			2	.65	.53	.22	.25	.27	.26	.48	.51	.53	.53	.73	.76	.79	.78
			4	.74	.64	.52	.63	.63	.61	.89	.92	.94	.92	1	1	1	1
	1	1	0	.45	.28	.05	.05	.06	.04	.05	.04	.05	.04	.04	.05	.05	.04
			1	.51	.35	.07	.07	.08	.08	.10	.10	.11	.10	.19	.19	.20	.19
			2	.56	.42	.19	.21	.19	.21	.37	.39	.41	.39	.64	.71	.72	.67
			4	.66	.54	.61	.65	.66	.65	.91	.93	.94	.93	1	1	1	1
	.5	.5	0	.52	.38	.05	.05	.04	.04	.04	.04	.05	.05	.05	.05	.06	.05
			1	.59	.46	.05	.07	.07	.07	.14	.18	.18	.16	.27	.32	.33	.30
			2	.65	.53	.22	.24	.24	.24	.50	.53	.55	.52	.74	.79	.82	.78
			4	.73	.64	.53	.62	.61	.61	.90	.92	.95	.94	.99	1	1	1

TABLE 1. *Proportion of censoring $p_{cens} = P(C < T)$, proportion of cure $p_{cure} = E[p(X)]$ and rejection proportion of $\tilde{T}_n$ for several values of n , c_h , c_T , θ_0 , θ_1 and θ_2 . The null hypothesis corresponds to $\theta_2 = 0$; the significance level is $\alpha = 0.05$.*

A degenerate special case is given if $p \equiv 0$, which can be regarded as a standard logistic model with $\theta_1 = 0$ and a very large θ_0 . We therefore expect that in this case the test will have difficulties rejecting the null hypothesis of a logistic cure rate, $p(x) = 1/[1 + \exp(\theta_0 + \theta_1 x)]$. This is indeed confirmed by a small simulation study: see Table 2 (with $c_T = 1$ and cross-validation bandwidth as before), which shows power figures close to the nominal level $\alpha = 0.05$.

$n = 50$	$n = 100$	$n = 200$
.108	.054	.054

TABLE 2. *Power of the test for a logistic model, when in reality there is no cure fraction.*

Logistic vs. probit. We also carried out a small simulation study to explore what happens if we test again for a logistic model (H_0), but in fact the probit model is more appropriate for the cure rate. We therefore generated data from a probit model

$$1 - p(x) = \Phi(\theta_0 + \theta_1 x)$$

with $\theta_0 = \theta_1 = 0.5$ (Φ denotes the standard normal distribution function). The model for T, C and X is as in Table 1, with $c_T = 1$ and a bandwidth chosen by cross-validation. The results are in Table 3. Obviously the power figures are close to the nominal level $\alpha = 0.05$. This is, however, not surprising: the probit and the logistic model are very close to each other (except in the tails), so that large samples are necessary “for even modest sensitivity”, as laid out in Chambers and Cox (1967).

$n = 50$	$n = 100$	$n = 200$
.054	.048	.056

TABLE 3. *Power of the test for a logistic model, when the true cure rate is a probit model.*

Logistic vs. mixture model. The previous scenario presents, of course, a special case. Further simulations, with models for the alternative that are clearly different from the logistic curve, yielded much better results. We generated, for example, data from a mixture of a logistic model and a polynomial of order three:

$$1 - p(x) = (1 - \alpha) \frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)} + \alpha \{0.9(\theta_0 + \theta_1 x) + (\theta_0 + \theta_1 x)^2 - (\theta_0 + \theta_1 x)^3\},$$

for $-1 \leq x \leq 1$, $0 \leq \alpha \leq 1$ and $\theta_0 = \theta_1 = 0.75$. Hence $0 \leq \theta_0 + \theta_1 x \leq 1.5$ for all $-1 \leq x \leq 1$. Note that on the interval $[0, 1.5]$ the function $u \rightarrow 0.9u + u^2 - u^3$ takes values between 0 and 1 and is not monotone. We use the cross-validation bandwidth as before, and the same scenario as for Table 1, with $c_T = 1$. Table 4 below shows that the power of this test increases with sample size and with the value of α , as can be expected.

α	$n = 50$	$n = 100$	$n = 200$
0	.048	.062	.056
0.25	.054	.096	.138
0.50	.122	.280	.508
0.75	.274	.572	.836
1	.458	.780	.986

TABLE 4. *Power of the test for a logistic model, when the true cure rate is a mixture of a logistic model and a polynomial model.*

Testing for a cure rate. Finally we used simulations to test whether $p \equiv 0$ (H_0), i.e. whether a cure rate exists, which is an important question in applications. Although the asymptotic result shows that the limit is degenerate (so the rate of convergence is faster than the rate $n\sqrt{h}$ in Theorem 2.1), the bootstrap is capable of detecting deviations from the null hypothesis.

c_T	p	p_{cens}	$n = 50$				$n = 100$				$n = 200$			
			c_h			CV	c_h			CV	c_h			CV
			1	1.5	2		1	1.5	2		1	1.5	2	
0	0	0.25	.11	.09	.07	.08	.06	.05	.04	.05	.05	.05	.04	.04
	0.05	0.28	.40	.46	.52	.50	.57	.65	.73	.69	.84	.90	.92	.91
	0.15	0.36	.76	.83	.86	.83	.97	.99	.99	.99	1	1	1	1
	0.25	0.44	.94	.96	.97	.95	1	1	1	1	1	1	1	1
1	0	0.22	.11	.10	.10	.10	.07	.05	.04	.05	.05	.04	.03	.04
	0.05	0.26	.47	.49	.53	.51	.63	.69	.73	.67	.88	.94	.95	.92
	0.15	0.34	.79	.81	.85	.80	.98	.98	.99	.98	1	1	1	1
	0.25	0.42	.95	.97	.97	.97	1	1	1	1	1	1	1	1

TABLE 5. Proportion of censoring p_{cens} and rejection proportion of $\tilde{T}_n$ for several values of n , c_h and c_T . The null hypothesis corresponds to $p \equiv 0$, the alternatives are $p \equiv 0.05, 0.15, 0.25$; the significance level is $\alpha = 0.05$.

c_T	$p(0)$	p_{cens}	$n = 50$				$n = 100$				$n = 200$			
			c_h			CV	c_h			CV	c_h			CV
			1	1.5	2		1	1.5	2		1	1.5	2	
0	0.05	0.28	.39	.47	.52	.51	.58	.66	.72	.68	.88	.91	.95	.93
	0.15	0.36	.80	.84	.88	.85	.97	.98	.99	.99	1	1	1	1
	0.25	0.44	.94	.95	.96	.95	1	1	1	1	1	1	1	1
1	0.05	0.26	.47	.53	.58	.52	.66	.74	.78	.72	.88	.93	.96	.92
	0.15	0.34	.80	.83	.86	.83	.97	.98	.99	.98	1	1	1	1
	0.25	0.42	.93	.95	.96	.95	1	1	1	1	1	1	1	1

TABLE 6. Proportion of censoring p_{cens} and rejection proportion of $\tilde{T}_n$ as in Table 5, now with alternatives $p(x) = 1/\{1 + \exp(\theta(1 + 0.2x))\}$, with θ chosen such that $p(0) = 0.05, 0.15, 0.25$ (which is also equal to $E[p(X)]$).

We considered again the scenario from Table 1, in particular the same survival function S_0 . Only the probability $p(\cdot)$ differs: we consider the null hypothesis, $p \equiv 0$, and alternatives

$p \equiv 0.05, 0.15, 0.25$ in Table 5, as well as $p(x) = 1/\{1 + \exp(\theta(1 + 0.2x))\}$ in Table 6, with θ chosen such that $p(0) = 0.05, 0.15, 0.25$. Apart from the case $n = 50$, the test again respects the level $\alpha = 5\%$. It also appears to be very powerful for all alternatives and sample sizes.

5. Application

As an illustration of our method we analysed data from by the Medical Birth Registry of Norway (see <http://folk.uio.no/borgan/abg-2008/data/data.html>). The data contain information on births in Norway since 1967, related to a total of 53,558 women. We are interested in the time between the births of the first and second children, and how this is affected if the first child died within one year. The covariate of interest is *age* (X), which is the age of the mother at the birth of the first child. The censoring indicator equals 1 if the mother had a second birth, and 0 if the observation is censored. We want to test whether the logistic model is a suitable model for the “cure rate” p , i.e. the fraction of women who gave birth only once.

FIGURE 1. *Left: Raw (univariate) data (case 1); circles are uncensored, triangles are censored data. Right: Kaplan-Meier estimator of all data; the plateau (containing 7 observations) suggests that a cure fraction is present (sufficiently long follow-up).*

We consider two cases (subpopulations). Case 1 is concerned with the subset of observations for which the first child has died within one year ($n = 262$). For case 2 we consider a random subset of size $n = 300$ of the entire data set. Our main reason for working with a sample and not with the complete data set is feasibility. The subset is obtained using the random number generator in R. We expect that the cure rate, i.e. the probability of no second birth (as a function of age), is different in the two groups. This is confirmed by our analysis; see Figures 1 and 2 for case 1 and Figures 3 and 4 for case 2. For case 1 (first child did not survive the first 12 months) the logistic model seems to be an appropriate model

(the null hypothesis is not rejected for all values of the bandwidth h), whereas for case 2 the logistic model is not appropriate (the p -value is close to zero for all values of h).

For our analysis we chose $h_0 = h \times n^{.09}$, which was motivated by the simulations. The number of bootstrap samples is 500, the time between the first and second births is expressed in years. Further comments are provided in the figure captions.

FIGURE 2. *Left: Nonparametric estimator of $p(x)$ for case 1 with $h = 3$ (solid curve); logistic estimator of $p(x)$ (dashed curve). Right: Significance trace (plot of p -value as a function of h), clearly showing that the null hypothesis of logistic cure rate is not rejected for all bandwidths in a wide range of possible bandwidths; the dashed line is the 5% level.*

FIGURE 3. *Raw data (case 2) and Kaplan-Meier estimator as in Figure 1; again there is a plateau indicating that a cure fraction is present.*

FIGURE 4. *Left: Nonparametric estimator of $p(x)$ for case 2 with $h = 3$ (solid curve); logistic estimator of $p(x)$ (dashed curve). Right: Significance trace as in Figure 2, this time showing that the null hypothesis of logistic cure rate is rejected for all bandwidths in a wide range; the dashed line is again the 5% level.*

6. Concluding remarks

We have proposed a test for a parametric form of the cure rate p that was inspired by Härdle and Mammen (1993), who test parametric models for a regression function. Analogously to that paper we have established asymptotic normality, but we cannot recommend using the asymptotic result in practice. Instead we propose a bootstrap procedure that is easy to implement. Simulations confirm that the bootstrap method works very well. Our application to two related real data sets further reassures that the method is powerful and can be recommended: in one data set the null hypothesis of a logistic form for p is clearly rejected, in the second data set clearly not.

Our proposed test is based on the nonparametric estimator $\hat{p}(X)$ from Xu and Peng (2014), where X is a one-dimensional covariate. A reasonable next step would be to target models with more than one covariate, since most data sets contain information in the form of many relevant covariates. In order to avoid the curse of dimensionality, it would certainly make sense to work with dimension reducing techniques. For example, one could replace the Xu-Peng estimator $\hat{p}(X)$ by a single index model estimator $\hat{p}(S)$ based on an estimator of the index $S = \beta^\top X$, now with X multi-dimensional (containing various types of covariates), and a parameter vector β of matching dimension. The estimation of such a model (or any other semiparametric model for the cure rate $p(x)$ that allows for multi-dimensional covariates) has not been studied so far in the literature. So the next step will be to tackle estimation of $p(x)$ in such models. The implementation of the test is then straightforward using our proposed approach with the new cure rate estimator $\hat{p}(X)$ in place of the Xu-Peng estimator.

7. Proof of Theorem 2.1

To derive the desired result for the Härdle-Mammen type test statistic $\mathcal{T}_n$ from equation (2.1) we write

$$\mathcal{T}_n = n\sqrt{h} \int_0^1 [\{\hat{p}(x) - p(x)\}^2 + \{p(x) - p_{\hat{\theta}}(x)\}^2 + 2\{p(x) - p_{\hat{\theta}}(x)\}\{\hat{p}(x) - p(x)\}]\pi(x) dx.$$

Consider local alternatives $p(x) = p_{\theta}(x) + n^{-1/2}h^{-1/4}\Delta_n(x)$, where θ denotes the true parameter. This covers the null hypothesis as a special case with $\Delta_n = 0$. We write $\mathcal{T}_{0n}$ for the first term of $\mathcal{T}_n$. We will show that $\mathcal{T}_{0n}$ determines the distribution of the test statistic under the null hypothesis and that

$$(7.1) \quad \mathcal{T}_{0n} - b_h = n\sqrt{h} \int_0^1 \{\hat{p}(x) - p(x)\}^2 \pi(x) dx - b_h \xrightarrow{d} N(0, V),$$

i.e. $\mathcal{T}_n$ is under H_0 approximately $N(b_h, V)$ distributed, with $b_h = O(h^{-1/2})$ tending to infinity.

Before proving (7.1) we consider the second term of $\mathcal{T}_n$ and, writing $p(x) = p_{\theta}(x) + n^{-1/2}h^{-1/4}\Delta_n(x)$, we can split it into three parts:

$$\begin{aligned} & n\sqrt{h} \int_0^1 \{p_{\theta}(x) - p_{\hat{\theta}}(x) + n^{-1/2}h^{-1/4}\Delta_n(x)\}^2 \pi(x) dx \\ &= n\sqrt{h} \int_0^1 \{p_{\theta}(x) - p_{\hat{\theta}}(x)\}^2 \pi(x) dx + n\sqrt{h} \int_0^1 \{n^{-1/2}h^{-1/4}\Delta_n(x)\}^2 \pi(x) dx \\ & \quad + 2n\sqrt{h} \int_0^1 \{p_{\theta}(x) - p_{\hat{\theta}}(x)\}n^{-1/2}h^{-1/4}\Delta_n(x) \pi(x) dx \\ &= \int_0^1 \Delta_n(x)^2 \pi(x) dx + o_p(1). \end{aligned}$$

The resulting integral (which only plays a role if $\Delta_n \neq 0$) comes from the middle part. To see that the first part involving $(p_{\theta} - p_{\hat{\theta}})^2$ is asymptotically negligible, use assumptions (A6) and (A7), which ensure that p_{θ} is Lipschitz in θ , uniformly in x , and that $\hat{\theta}$ is $\sqrt{n}$ -consistent. This gives the order $O_p(h^{1/2}) = o_p(1)$ as $h \rightarrow 0$. The third part in the above display vanishes with rate $O_p(h^{1/4}) = o_p(1)$, which follows by the same arguments, combined with the assumption that $\Delta_n(\cdot)$ is uniformly bounded.

Now consider the (mixed) third term of $\mathcal{T}_n$. Applying the Cauchy-Schwarz inequality yields the order $O_p(h^{-1/4})$, which is smaller than the asymptotic order of the leading term $\mathcal{T}_{0n} = O_p(h^{-1/2})$. These arguments combined with (7.1) prove the claim of the theorem for local alternatives, namely that $\mathcal{T}_n$ is approximately $N(b_h + \int_0^1 \Delta_n(x)^2 \pi(x) dx, V)$ distributed.

It remains to verify (7.1) for $\mathcal{T}_{0n}$, the first term of $\mathcal{T}_n$. Let Λ denote the cumulative conditional hazard function $\Lambda(t|x) = -\log\{S(t|x)\}$ and let $\hat{\Lambda}(t|x)$ be the nonparametric

estimator provided in Xu and Peng (2014). These authors show that

$$(7.2) \quad \sqrt{nh}\{\hat{p}(x) - p(x)\} = S(Y_{(n)}^1|x)\sqrt{nh}\{\hat{\Lambda}(Y_{(n)}^1|x) - \Lambda(Y_{(n)}^1|x)\} + o_p(1).$$

They also show that $Y_{(n)}^1$ is weakly consistent for $\sup_x \tau_0(x)$, and that $S(Y_{(n)}^1|x) \rightarrow S(\tau_0(x)|x) = p(x)$ in probability. This yields

$$\sqrt{nh}\{\hat{p}(x) - p(x)\} = p(x)\sqrt{nh}\{\hat{\Lambda}(Y_{(n)}^1|x) - \Lambda(Y_{(n)}^1|x)\} + o_p(1).$$

In order to derive asymptotic normality of $\mathcal{T}_n$, we will need an approximation of $\hat{\Lambda} - \Lambda$, which is, for example, provided in Du and Akritas (2002). They show that, uniformly in x ,

$$\hat{\Lambda}(t|x) - \Lambda(t|x) = \sum_{i=1}^n w_{hi}(x)\zeta_i(t|x) + r(t, x)$$

with $\sup_{x, t \leq t_0} |r(t, x)| = O\{(nh)^{-3/4}(\log n)^{3/4}\}$ a.s., where t_0 satisfies $\sup_x H(t_0|x) < 1$, with w_{hi} as specified above, and with $\zeta_i(t|x)$ defined in (2.4). It follows from arguments in Xu and Peng that $\hat{\Lambda}(Y_{(n)}^1|x) - \Lambda(Y_{(n)}^1|x)$ is asymptotically equivalent to $\sum_{i=1}^n w_{hi}(x)\zeta_i(\tau_0(x)|x)$. Then,

$$\begin{aligned} (7.3) \quad \mathcal{T}_{0n} &= n\sqrt{h} \int_0^1 p^2(x) \left\{ \sum_{i=1}^n w_{hi}(x)\zeta_i(\tau_0(x)|x) \right\}^2 \pi(x) dx + o_p(1) \\ &= n\sqrt{h} \int_0^1 p^2(x) \left\{ n^{-1} \sum_{i=1}^n \frac{K_h(x - X_i)}{f(x)} \zeta_i(\tau_0(x)|x) \right\}^2 \pi(x) dx + o_p(1) \\ &= \mathcal{T}_{1n} + \mathcal{T}_{2n} + o_p(1), \end{aligned}$$

where

$$\begin{aligned} \mathcal{T}_{1n} &= \frac{\sqrt{h}}{n} \int_0^1 p^2(x) \sum_{i=1}^n \frac{K_h^2(x - X_i)}{f^2(x)} \zeta_i^2(\tau_0(x)|x) \pi(x) dx, \\ \mathcal{T}_{2n} &= \frac{\sqrt{h}}{n} \int_0^1 p^2(x) \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \frac{K_h(x - X_i)K_h(x - X_j)}{f^2(x)} \zeta_i(\tau_0(x)|x)\zeta_j(\tau_0(x)|x) \pi(x) dx. \end{aligned}$$

Analogously as in the proof of Proposition 1 in Härdle and Mammen (1993), but now with ζ_j in place of heteroscedastic errors ε_j , we obtain the bias term as follows:

$$\begin{aligned} E(\mathcal{T}_{1n}) &= \sqrt{h} \int_0^1 \frac{p^2(x)}{f^2(x)} E\{K_h^2(x - X)\zeta^2(\tau_0(x)|x)\} \pi(x) dx \\ &= \sqrt{h} \int_0^1 \frac{p^2(x)}{f^2(x)} E\left[\frac{1}{h^2} K^2\left(\frac{x - X}{h}\right) E\{\zeta^2(\tau_0(x)|x)|X\}\right] \pi(x) dx \\ &= \sqrt{h} \int_0^1 \frac{p^2(x)}{f^2(x)} \int \frac{1}{h} K^2(v) f(x + vh) \mu_{xx}(x + vh) dv \pi(x) dx \\ &= b_h + o(1), \end{aligned}$$

with, using assumptions (A1) and (A5),

$$b_h = h^{-1/2} R(K) \int_0^1 \frac{p^2(x)\pi(x)}{f(x)} \mu_{xx}(x) dx = O(h^{-1/2}).$$

We have $nh^3 \rightarrow \infty$, thanks to our assumptions on the bandwidth, and therefore $\text{var}(\mathcal{T}_{1n}|X_1, \dots, X_n) = O_p(n^{-1}h^{-3}) = o_p(1)$,

$$\mathcal{T}_{1n} = b_h + o_p(1).$$

The desired normal approximation (7.1), $\mathcal{T}_{0n} - b_h \xrightarrow{d} N(0, V)$, now follows from this statement, combined with the asymptotic normality of $\mathcal{T}_{2n}$,

$$\mathcal{T}_{2n} \xrightarrow{d} N(0, V),$$

which follows analogously to the arguments outlined in Härdle and Mammen (1993), using Theorem 2.1 by de Jong (1987): write

$$\begin{aligned} \mathcal{T}_{2n} &= \frac{\sqrt{h}}{n} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n V_{ij}, \\ V_{ij} &= \int_0^1 p^2(x) \frac{K_h(x - X_i)K_h(x - X_j)}{f^2(x)} \zeta_i(\tau_0(x)|x) \zeta_j(\tau_0(x)|x) \pi(x) dx. \end{aligned}$$

De Jong's theorem requires that $\mathcal{T}_{2n}$ is clean in the sense that the conditional expectations of the V_{ij} given X_i ($i = 1, \dots, n$) vanish. Using similar arguments as in Cao and González-Manteiga (2008; pages 183-184), we see that $\zeta_i(\tau_0(x)|x)$ and $\zeta_j(\tau_0(x)|x)$ in the expression of V_{ij} can be replaced by $\zeta_i(\tau_0(X_i)|X_i)$ and $\zeta_j(\tau_0(X_j)|X_j)$. Since $E\{\zeta(\tau_0(x)|x)|X = x\} = 0$, it follows that $\mathcal{T}_{2n}$ is indeed clean up to an asymptotically negligible term (and the asymptotic mean is zero).

In order to obtain the asymptotic variance, it suffices to calculate the second moment of V_{ij} , which is obtained as follows. We have

$$\begin{aligned} E(V_{ij}^2) &= E\left(\int_0^1 \int_0^1 \frac{p^2(x)}{f^2(x)} K_h(x - X_i) K_h(x - X_j) \zeta_i(\tau_0(x)|x) \zeta_j(\tau_0(x)|x) \pi(x) \right. \\ &\quad \times \left. \frac{p^2(z)}{f^2(z)} K_h(z - X_i) K_h(z - X_j) \zeta_i(\tau_0(z)|z) \zeta_j(\tau_0(z)|z) \pi(z) dx dz \right) \\ &= \int_0^1 \int_0^1 \frac{p^2(x)p^2(z)\pi(x)\pi(z)}{f^2(x)f^2(z)} E\{K_h(x - X_i) K_h(z - X_i) \zeta_i(\tau_0(x)|x) \zeta_i(\tau_0(z)|z)\} \\ &\quad \times E\{K_h(x - X_j) K_h(z - X_j) \zeta_j(\tau_0(x)|x) \zeta_j(\tau_0(z)|z)\} dx dz \end{aligned}$$

$$\begin{aligned}
&= \int_0^1 \int_0^1 \frac{p^2(x)p^2(z)\pi(x)\pi(z)}{f^2(x)f^2(z)} \\
&\quad \times \left(\int K_h(x-u)K_h(z-u)E\{\zeta(\tau_0(x)|x)\zeta(\tau_0(z)|z)|X=u\}f(u)du \right)^2 dx dz \\
&= \int_0^1 \int_0^1 \frac{p^2(x)p^2(z)\pi(x)\pi(z)}{f(x)f(z)} \mu_{xz}(x)\mu_{xz}(z) \left\{ \frac{1}{h} K * K \left(\frac{z-x}{h} \right) \right\}^2 dx dz + o(1)
\end{aligned}$$

as $h \rightarrow 0$. Hence

$$\text{var } \mathcal{T}_{2n} = 2 \sum_{i \neq j} \text{var} \left(\frac{\sqrt{h}}{n} V_{ij} \right) = 2hE(V_{ij}^2) + o(1) = V + o(1),$$

with

$$\begin{aligned}
V &= 2h \int_0^1 \int_0^1 \frac{p^2(x)p^2(z)\pi(x)\pi(z)}{f(x)f(z)} \left\{ \frac{1}{h} K * K \left(\frac{z-x}{h} \right) \right\}^2 \\
&\quad \times E\{\zeta(\tau_0(x)|x)\zeta(\tau_0(z)|z)|X=x\}E\{\zeta(\tau_0(x)|x)\zeta(\tau_0(z)|z)|X=z\} dx dz.
\end{aligned}$$

Since $p(\cdot)$ is continuous and $\mu_{xz}(z)$ is continuous in z by assumption (A5), V simplifies, as on page 1931 of Härdle and Mammen (1993), and we obtain the desired formula

$$V = 2K^{(4)}(0) \int \left(\frac{p^2(x)\pi(x)E\{\zeta^2(\tau_0(x)|x)|X=x\}}{f(x)} \right)^2 dx.$$

□

8. Proof of Theorem 3.1 (sketch)

We use a decomposition (and notation) as in the proof of Theorem 2.1:

$$\begin{aligned}
\mathcal{T}_n^* &= n\sqrt{h} \int_0^1 [\hat{p}_{hh_0}^*(x) - p_{hh_0, \hat{\theta}^*}(x)]^2 \pi(x) dx \\
&= n\sqrt{h} \int_0^1 \left[\{\hat{p}_{hh_0}^*(x) - p_{h_0, \hat{\theta}}(x)\}^2 + \{p_{h_0, \hat{\theta}}(x) - p_{hh_0, \hat{\theta}^*}(x)\}^2 \right. \\
&\quad \left. + 2\{\hat{p}_{hh_0}^*(x) - p_{h_0, \hat{\theta}}(x)\}\{p_{h_0, \hat{\theta}}(x) - p_{hh_0, \hat{\theta}^*}(x)\} \right] \pi(x) dx.
\end{aligned}$$

In a similar way as was done for $\mathcal{T}_n$ and by using assumption (A8) we can show that the second and third term above vanish asymptotically.

Let us now focus on $\hat{p}_{hh_0}^*(x) - p_{h_0, \hat{\theta}}(x)$. We know that $\hat{p}_{hh_0}^*(x) = \hat{S}_{hh_0}^*(Y_{(n)}^{*1}|x)$ and that $\hat{S}_{hh_0}^*(Y_{(n)}^{*1}|x) - \hat{S}_{hh_0}^*(Y_{(n)}^1|x) = o_{p^*}(1)$ a.s., where $\hat{S}_{hh_0}^*(\cdot|x)$ is the Beran estimator based on the bootstrap data $(Y_1^*, \delta_1^*), \dots, (Y_n^*, \delta_n^*)$, and where $Y_{(n)}^{*1} = \max_{i: \delta_i^*=1} Y_i^*$. We also know that for given x , the variable T^* is drawn from the survival function

$$P^*(T^* > t|x) = \frac{\hat{S}_{h_0}(t|x) - \hat{p}_{h_0}(x)}{1 - \hat{p}_{h_0}(x)}(1 - p_{h_0, \hat{\theta}}(x)) + p_{h_0, \hat{\theta}}(x).$$

Hence, $P^*(T^* > Y_{(n)}^1 | x) = p_{h_0, \hat{\theta}}(x)$, since $\hat{p}_{h_0}(x) = \hat{S}_{h_0}(Y_{(n)}^1 | x)$. This shows that $p_{h_0, \hat{\theta}}(x)$ is the correct centering term for $\hat{p}_{hh_0}^*(x)$.

As our bootstrap procedure is similar to the bootstrap procedure studied by Van Keilegom and Veraverbeke (1997) in the context of nonparametric regression with right censored data without cure fraction, it can be shown, similarly as in their paper that

$$\begin{aligned} & \hat{p}_{hh_0}^*(x) - p_{h_0, \hat{\theta}}(x) \\ &= p(x) \left[\sum_{i=1}^n w_{hi}(x) \zeta_i^*(\tau_0(x) | x) - \sum_{i=1}^n w_{h_0i}(x) \zeta_i(\tau_0(x) | x) \right] + O_{p^*}((nh)^{-3/4}(\log n)^{3/4}), \end{aligned}$$

a.s. and uniformly in x , where $\zeta_i^*(\cdot | x)$ is as $\zeta_i(\cdot | x)$ except that (Y_i, δ_i) is replaced by (Y_i^*, δ_i^*) . Defining $\eta_i^*(\cdot | x) = \zeta_i^*(\cdot | x) - \sum_{j=1}^n w_{h_0j}(x) \zeta_j(\tau_0(x) | x)$, the latter can be written as

$$p(x) \sum_{i=1}^n w_{hi}(x) \eta_i^*(\tau_0(x) | x) + O_{p^*}((nh)^{-3/4}(\log n)^{3/4}).$$

Now, we follow the same arguments as in the proof of Theorem 2.1, and write

$$\mathcal{T}_n^* = n\sqrt{h} \int_0^1 \{ \hat{p}_{hh_0}^*(x) - p_{h_0, \hat{\theta}}(x) \}^2 \pi(x) dx + o_{p^*}(1) = \mathcal{T}_{1n}^* + \mathcal{T}_{2n}^* + o_{p^*}(1) \text{ a.s.},$$

and for $\mathcal{T}_{1n}^*$ we find that

$$E^*(\mathcal{T}_{1n}^*) = h^{-1/2} R(K) \int_0^1 \frac{p^2(x) \pi(x)}{f(x)} \mu_{xx}^*(x) dx,$$

where

$$\begin{aligned} \mu_{xx}^*(x) &= E^* \{ \eta^*(\tau_0(x) | x)^2 | X = x \} = \text{Var}^* \{ \zeta^*(\tau_0(x) | x)^2 | X = x \} \\ &= \text{Var} \{ \zeta(\tau_0(x) | x)^2 | X = x \} + o(1) \text{ a.s.}, \end{aligned}$$

using arguments similar to those in the proof of Lemma 8(b) in Van Keilegom and Veraverbeke (1997), and the latter equals $\mu_{xx}(x) + o(1)$ a.s. Hence, $\mathcal{T}_{1n}^* = E^*(\mathcal{T}_{1n}^*) + o_{p^*}(1) = b_h + o_{p^*}(1)$ a.s. Using similar arguments we obtain that $P^*(\mathcal{T}_{2n}^* \leq t) - \Phi(t/V) \rightarrow 0$ a.s., which shows the result. $\square$

Acknowledgment. The authors would like to thank the Referees and the Associate Editor for helpful comments that greatly improved the paper. I. Van Keilegom acknowledges financial support from the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409), and from the IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy).

References

- [1] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Tech. Rep., University of California, Berkeley.
- [2] Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 11, 15-53.

- [3] Cao, R. and González-Manteiga, W. (2008). Goodness-of-fit tests for conditional models under censoring and truncation. *J. Econometrics*, 143, 166-190.
- [4] Chambers E.A. and Cox D.R. (1967). Discrimination between alternative binary response models, *Biometrika* **54**, 573-578.
- [5] Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71**, 1591-1608.
- [6] de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probab. Theory Related Fields.*, 75, 261-277.
- [7] Du, Y. and Akritas, M.G. (2002). Uniform strong representation of the conditional Kaplan-Meier process. *Math. Methods Statist.*, 11, 152-182.
- [8] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38, 1041-1046.
- [9] Härdle, W. and Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21, 1926-1947.
- [10] Kuk, A.Y.C. and Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79, 531-541.
- [11] Laska, E.M. and Meisner, M.J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, 48, 1223-1234.
- [12] Li, G. and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Ann. Inst. Statist. Math.*, 53, 708-729.
- [13] Li, Y., Tiwari, R.C. and Guha, S. (2007). Mixture cure survival models with dependent censoring. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69, 285-306.
- [14] Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Ann. Inst. Statist. Math.*, 60, 545-574.
- [15] Maller, R.A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, 79, 731-739.
- [16] Maller, R.A. and Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data. *J. Amer. Statist. Assoc.*, 89, 1499-1506.
- [17] Patilea, V. and Van Keilegom, I. (2017+). A general approach for cure models in survival analysis. Preprint.
- [18] Sy, J.P. and Taylor, J.M.G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56, 227-236.
- [19] Tsodikov, A. (1998). Asymptotic efficiency of a proportional hazards model with cure. *Statist. Probab. Lett.*, 39, 237-244.
- [20] Tsodikov, A. (2003). Semiparametric models: A generalized self-consistency approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 65, 759-774.
- [21] Van Keilegom, I. and Veraverbeke, N. (1997). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Ann. Inst. Statist. Math.*, 49, 467-491.
- [22] Yakovlev, A.Y., Cantor, A.B. and Shuster, J.J. (1994). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Stat. Med.*, 13, 983-986.
- [23] Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canad. J. Statist.*, 42, 1-17.

URSULA U. MÜLLER, DEPARTMENT OF STATISTICS, TEXAS A&M UNIVERSITY, COLLEGE STATION, TX 77843-3143, USA, E-MAIL: USCHI@STAT.TAMU.EDU

INGRID VAN KEILEGOM, ORSTAT, KU LEUVEN, NAAMSESTRAAT 69, 3000 LEUVEN, BELGIUM, E-MAIL: INGRID.VANKEILEGOM@KULEUVEN.BE

FACULTY OF ECONOMICS AND BUSINESS

Naamsestraat 69 bus 3500

3000 LEUVEN, BELGIË

tel. + 32 16 32 66 12

fax + 32 16 32 67 91

info@econ.kuleuven.be

www.econ.kuleuven.be