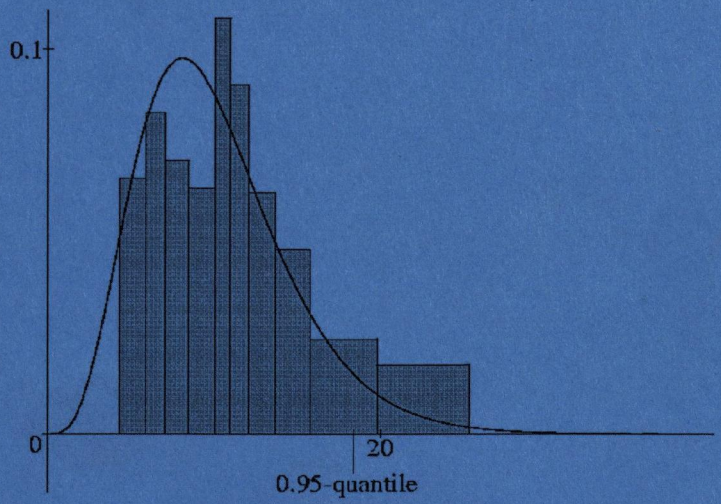
UNIVERSITÄT POTSDAM
Institut für Mathematik

Estimation and Testing the Effect of Covariates in Accelerated Life Time Models under Censoring

Hannelore Liero



Mathematische Statistik und
Wahrscheinlichkeitstheorie

Universität Potsdam – Institut für Mathematik

Mathematische Statistik und Wahrscheinlichkeitstheorie

Estimation and Testing the Effect of Covariates in Accelerated Life Time Models under Censoring

Hannelore Liero

Institute of Mathematics, University of Potsdam

e-mail: liero@uni-potsdam.de

Preprint 2010/02

January 2010

Impressum

Herausgeber: Mathematische Statistik und Wahrscheinlichkeitstheorie
am Institut für Mathematik

Adresse: Universität Potsdam
Am Neuen Palais 10
14469 Potsdam
Telefon:
Fax: +49-331-977 1500
E-mail: +49-331-977 1578
neisse@math.uni-potsdam.de

ISSN 1613-3307

Estimation and Testing the Effect of Covariates in Accelerated Life Time Models under Censoring Preliminary Version

Hannelore Liero
Institute of Mathematics, University of Potsdam
e-mail: liero@uni-potsdam.de

Abstract

The accelerated lifetime model is considered. To test the influence of the covariate we transform the model in a regression model. Since censoring is allowed this approach leads to a goodness-of-fit problem for regression functions under censoring. So nonparametric estimation of regression functions under censoring is investigated, a limit theorem for a L_2-distance is stated and a test procedure is formulated. Finally a Monte Carlo procedure is proposed.

Key words: Accelerated life time model; censoring; goodness-of-fit testing; nonparametric regression estimation; Monte Carlo testing
AMS subject classification: Primary 62G10; Secondary 62N05

1 Introduction

We consider a life time model which describes the following situation: By some covariate X the time to failure may be accelerated or retarded relative to some baseline. The speeding up or slowing down is accomplished by some positive function ψ, and we may write

$$T = \frac{T_0}{\psi(X)},$$

where T_0 is the so-called baseline life time and T is the observable life time. We will assume that T is an absolute continuous random variable (r.v.) and that the covariate X does not depend on the time. For simplicity of presentation let X be one-dimensional.

For statistical application a suitable choice of the function ψ is important and the problem of testing ψ arises. A survey of test procedures for testing ψ under different model assumptions is given in Liero H. and Liero M. (2008). The aim of the present paper is to propose a test procedures for testing whether the function ψ belongs to a pre-specified parametric class of functions

$$\mathcal{F} = \{\psi \,|\, \psi(\cdot) = \psi(\cdot;\beta), \;\; \beta \in \mathbb{R}^d\}. \tag{1}$$

For data without censoring this problem was already considered in Liero (2008). In this paper we assume that the independent and identically distributed life times T_i are subject to random right censoring, i.e. the observations are

$$V_i = \min(T_i, C_i), \;\; \Delta_i = 1(T_i \leq C_i) \;\; \text{and} \;\; X_i, \;\; i = 1, \ldots, n$$

where the C_i's are independent and identically distributed censoring times with distribution function G. Furthermore we assume that the T_i's and the C_i's are conditionally independent given the X_i's.

The inference is based on the log transformation of the lifetime model to a regression model: The conditional expectation of $Y = \log T$ given the covariate X has the form

$$\mathsf{E}\,(Y|X = x) = \mu - \log\psi(x) \quad \text{with} \quad \mu = -\int \log z \, \mathrm{d}S_0(z) = \mathsf{E}\,(\log T_0),$$

where S_0 is the survival function of the baseline life time T_0, and we can translate the considered problem into a problem of testing the regression function in a nonparametric regression model

$$Y = \log T = m(X) + \varepsilon,$$

where $m(x) = \mu - \log\psi(x)$, and with $\varepsilon = \log(T_0) - \mathsf{E}(\log(T_0))$

$$\mathsf{E}\,(\varepsilon|X = x) = 0, \;\; \text{and} \;\; \mathsf{E}\,(\varepsilon^2|X = x) = \sigma^2$$

for some $\sigma^2 > 0$. For identifiability we assume $\psi(0) = 1$.
Test problem (1) is translated into the following problem

$$H: \;\; m \in \mathcal{M} \;\; \text{versus} \;\; K: \;\; m \notin \mathcal{M}$$

where

$$\mathcal{M} = \{m \,|\, m(\cdot;\beta,\mu) = -\log\psi(\cdot;\beta) + \mu, \;\; \beta \in \mathbb{R}^d, \mu \in \mathbb{R}\},$$

that is we have to check whether the regression function has a parametric form or alternatively that this regression is nonparametric.
As test statistic a weighted L_2-distance between a parametric and the nonparametric regression estimator is proposed. To formulate the corresponding test procedure one has to investigate the properties of nonparametric estimators for regression functions under censoring. Therefore in Section 2 the nonparametric estimation of the regression function under censoring is considered. In Section 3 asymptotic properties of the nonparametric regression estimator are presented; the main result is the asymptotic normality of the weighted L_2-distance of the estimator. This limit theorem is based on a so-called asymptotic (conditional) i.i.d. representation of the difference between the estimator and the regression function. The test procedure is given in Section 4.

2 Nonparametric estimation of the regression function under censoring

We start with the a nonparametric estimator for the conditional distribution function of the transformed r.v. $Y = \log T$. Such an estimator was introduced by Beran (1981). On one hand the Beran estimator can be regarded as an extension of the well-known Kaplan-Meier estimator proposed for models with censored data without covariates, on the other hand it is an extension of nonparametric estimators for conditional distributions functions studied for data sets without censoring. To define the Beran estimator it is useful to introduce the following functions and their estimators: The conditional distribution function of the r.v. $Z = \log V$ with $V = \min(T, C)$ is given by $H(z|x) = \mathsf{P}(Z \leq z|X = x)$ and estimated by the kernel estimator

$$\hat{H}_n(z|x) = \sum_i W_{b_n i}(x, X_1, \ldots, X_n) 1(Z_i \leq z) \tag{2}$$

where $W_{b_n i}$ are the kernel weights defined by

$$W_{b_n i}(x, X_1, \ldots, X_n) = \frac{\frac{1}{b_n} K\left(\frac{x - X_i}{b_n}\right)}{\frac{1}{b_n} \sum_{j=1}^n K\left(\frac{x - X_j}{b_n}\right)}.$$

Here $K : \mathbb{R} \to \mathbb{R}$ is a kernel function, and b_n is a sequence of bandwidths tending to zero as $n \to \infty$. The symbol "1" denotes the indicator function.

The estimator of the conditional subdistribution function $H^U(z|x) = \mathsf{P}(Z \leq z, \Delta = 1|X = x)$ is given by

$$\hat{H}_n^U(z|x) = \sum_i W_{b_n i}(x, X_1, \ldots, X_n) 1(Z_i \leq z, \Delta_i = 1). \tag{3}$$

For the conditional cumulative hazard function Λ we have for $y \leq \tau$

$$\Lambda(y|x) = \int_{-\infty}^{y} \frac{\mathrm{d}F(s|x)}{1 - F(s_-|x)} = \int_{-\infty}^{y} \frac{\mathrm{d}H^U(s|x)}{1 - H(s_-|x)}$$

where F denotes the conditional cdf of the transformed $Y = \log T$ and $H(s_-|x) = \lim_{t \uparrow s} H(t|x)$, $\tau_x = \inf\{y|H(y|x) = 1\}$ is the upper bound of the support of $H(\cdot|x)$. Replacing H^U and H by their estimators (2) and (3) leads to the weighted Nelson-Aalen type estimator for the conditional cumulative hazard function:

$$\hat{\Lambda}_n(y|x) = \int_{-\infty}^{y} \frac{\mathrm{d}\hat{H}_n^U(s|x)}{1 - \hat{H}_n(s_-|x)}. \tag{4}$$

Now from the well-known relation between the cumulative hazard function and the survival function we obtain as estimator for $S_Y(y|x) = 1 - F(y|x) = \mathsf{P}(Y > y|X = x)$

$$\hat{S}_{Yn}(y|x) = \prod_{t \leq y} \left(1 - \Delta\hat{\Lambda}_n(t|x)\right) \tag{5}$$

where $\Delta\hat{\Lambda}_n(t|x) = \hat{\Lambda}_n(t|x) - \hat{\Lambda}_n(t - |x)$ is the jump of $\hat{\Lambda}_n(\cdot|x)$ at t. An equivalent form of (5) is

$$\hat{F}_n(y|x) = 1 - \prod_{\substack{Z_i \leq y \\ \Delta_i = 1}} \left\{ 1 - \frac{W_{b_n i}(x, \mathbb{X})}{\sum_j 1(Z_j \geq Z_i) W_{b_n j}(x, \mathbb{X})} \right\} \tag{6}$$

1

Note that for weights $W_{b_n i} = \frac{1}{n}$ the estimator $\hat{F}_{Yn}$ is the classical Kaplan–Meier estimator; for $\Delta_i = 1$ for all i the estimator $\hat{F}_n$ is the estimator of the conditional distribution function, and for $W_{b_n i} = \frac{1}{n}$ and $\Delta_i = 1$ for all i the estimator $\hat{F}_n$ is simply the empirical distribution function.
Several authors considered the asymptotic behavior of $\hat{F}_n$. Consistency of $\hat{F}_n(y|x)$ is proven for $y \leq \tau_x$.

[1] For simplicity we write $\mathbb{X} = (X_1, \ldots, X_n)$.

The regression function $m(x) = \mathsf{E}(Y|X = x)$ is defined by $\int y \, \mathrm{d}F(y|x)$. However, for the estimation of m and the investigation of the properties of the resulting estimator the following identities are useful:

$$\begin{aligned} m(x) &= \mathsf{E}(Y|X = x) \\ &= \int y \, \mathrm{d}F(y|x) \qquad (7) \\ &= \mathsf{E}\left(\frac{1 - F(Z_-|X)}{1 - H(Z_-|X)} Z\Delta | X = x\right) \qquad (8) \end{aligned}$$

and

$$m(x) = \int_0^1 F^{-1}(u|x) \, \mathrm{d}u \qquad (9)$$

where $F^{-1}(u|x) = \inf\{y|F(y|x) \geq u\}$.
To estimate m we replace $F(y|x)$ in (7) by the Beran estimator and obtain as nonparametric estimator for m

$$\hat{m}_n(x) = \int y \, \mathrm{d}\hat{F}_n(y|x). \qquad (10)$$

One can show that for this estimator the empirical versions of (8) and (9) hold, i.e.:

$$\hat{m}_n(x) = \sum_{i=1}^{n} W_{bni}(x, \mathbb{X}) \frac{1 - \hat{F}_n(Z_{i-}|x)}{1 - \hat{H}_n(Z_{i-}|x)} Z_i \Delta_i \qquad (11)$$

and

$$\hat{m}_n(x) = \int_0^1 \hat{F}_n^{-1}(u|x) \, \mathrm{d}u \qquad (12)$$

where $\hat{F}_n^{-1}(u|x) = \inf\{y|\hat{F}_n(y|x) \geq u\}$. We see that as in the case without censoring the regression estimator is a weighted average; now, in the case with censoring an average of the uncensored observations. The weights depend on the kernel and on the ratio of the Kaplan- Meier estimator and the empirical df of the observations.

3 Properties of the nonparametric regression estimator

Györfi et al (2002) showed that an estimator of this type [2] is L_2-consistent if the right endpoint of the support of F is smaller than that of G. We follow another approach than those authors. We will use a conditional i.i.d. presentation of the difference between estimator and regression function; such a presentation is based on the corresponding result for the estimator $\hat{F}_n$ which is derived by Akritas and Du (2002). Since this presentation holds only for $y \leq y^*$, where $y^* < \sup_x \tau_x$ we will truncate the estimator (due to the right censoring):
Instead of $m(x) = \int_{-\infty}^{\infty} y \, \mathrm{d}F(y|x)$ we estimate the function

$$m^*(x) = \int_{-\infty}^{y^*} y \, \mathrm{d}F(y|x). \tag{13}$$

The function m^* is estimated by

$$\hat{m}_n^*(x) = \int_{-\infty}^{y^*} y \, \mathrm{d}\hat{F}_n(y|x). \tag{14}$$

Before we state the results let us formulate the assumptions

A1 The marginal density g of X is bounded on $\mathbb{R}$ and twice continuously differentiable in a neighborhood of a set $\mathcal{I}$ and $g(x) \geq c > 0$ for some c and all $x \in \mathcal{I}$.

A2 The kernel K is a symmetric density with compact support; furthermore, it is twice continuously differentiable.

A3 We will need typical smoothness conditions on the functions $H(\cdot|\cdot)$ and $H^U(\cdot|\cdot)$. We formulate here for a general (sub)distribution function L:

The derivatives

$$\ddot{L}(y|x) = \frac{\partial^2 L(y|x)}{\partial^2 x}, \quad L''(y|x) = \frac{\partial^2 L(y|x)}{\partial^2 y}, \quad \dot{L}'(y|x) = \frac{\partial^2 L(y|x)}{\partial y \partial x}$$

exist and are continuous for all y, and all x in a neighborhood of $\mathcal{M}$.

[2] Instead of kernel weights considered here they used nearest neighbor weights.

Lemma 3.1 *Suppose that A1 and A2 hold and that A3 is satisfied by H and H^U. then*

$$\hat{m}_n^*(x) - m^*(x) = \sum_{i=1}^{n} W_{b_n i}(x, \mathbb{X})\, \eta(Z_i, \Delta_i | x) + R_n(x) \tag{15}$$

with

$$\begin{aligned}\eta(Z_i, \Delta_i | x) =& y^*(1 - F(y^*|x))\xi(Z_i, \Delta_i, y^*|x) \\ &- \int_{-\infty}^{y^*} (1 - F(s|x))\xi(Z_i, \Delta_i, s|x)\, \mathrm{d}s\end{aligned}$$

where

$$\xi(Z_i, \Delta_i, s|x) = \frac{1(Z_i \le s, \Delta_i = 1)}{(1 - H(Z_i|x))} - \int_{-\infty}^{s} \frac{1(Z_i \ge w)\, \mathrm{d}H^U(w|x)}{(1 - H(w|x))^2},$$

and where

$$\sup_{x \in \mathcal{I}} R_n(x) = O_{\mathsf{P}}\left((nb_n)^{-\frac{3}{4}} (\log n)^{\frac{3}{4}}\right) \qquad as\ n \to \infty.$$

Based on the presentation given in Lemma 3.1 we will prove the asymptotic normality of $\hat{m}_n(x)$ at an arbitrary fixed point x and a limit theorem for a weighted integrated squared error. Let us consider the conditional i.i.d. presentation as process and set

$$\mathcal{A}_n(x) = \sum_{i=1}^{n} W_{b_n i}(x, \mathbb{X})\, \eta(Z_i, \Delta_i | x). \tag{16}$$

In a first step we will split $\mathcal{A}_n$ in a stochastic and in a systematic part:

$$\begin{aligned}\mathcal{A}_n(x) =& \sum_{i=1}^{n} W_{b_n i}(x, \mathbb{X})\, (\eta(Z_i, \Delta_i | x) - \mathsf{E}[\eta(Z_i, \Delta_i | x) | X_i]) \\ &+ \sum_{i=1}^{n} W_{b_n i}(x, \mathbb{X}) \mathsf{E}[\eta(Z_i, \Delta_i | x) | X_i].\end{aligned} \tag{17}$$

Note that

$$W_{b_n i}(x, \mathbb{X}) = \frac{\frac{1}{n} K_{b_n}(x - X_i)}{\hat{g}_n(x)}$$

where

$$\hat{g}_n(x) = \frac{1}{n}\sum_{j=1}^{n} K_{b_n}(x - X_j)$$

is the estimator for the marginal density g of the covariate X.
The first part, the stochastic one, is approximated by

$$\mathcal{A}_{n1}(x) = \frac{1}{\mathsf{E}\hat{g}_n(x)}\frac{1}{n}\sum_{i=1}^{n} K_{b_n}(x - X_i)\ (\eta(Z_i, \Delta_i|x) - \mathsf{E}[\eta(Z_i, \Delta_i|x)|X_i])\,. \quad (18)$$

Using the well-known asymptotic properties of a nonparametric density estimator it is shown that the stochastic part of $\mathcal{A}_n$ and the statistic $\mathcal{A}_{n1}$ have the same asymptotic behavior. The stochastic behavior of $\mathcal{A}_{n1}$ is characterized by the covariance function

$$\mathcal{C}_n(x, y) = \mathsf{Cov}(\mathcal{A}_{n1}(x), \mathcal{A}_{n1}(y)). \quad (19)$$

Since this function plays a key role in proving limit theorems and deriving the corresponding standardizing terms an asymptotic expression for $\mathcal{C}_n(x, y)$ is presented in the following lemma:

Lemma 3.2 *Suppose that A1 and A2 hold, and H and H^U are Lipschitz continuous with respect to x. Set*

$$\beta_x(v) = \int_{-\infty}^{v} \frac{\mathrm{d}H^U(w|x)}{(1 - H(w|x))^2}$$

and

$$\begin{aligned}\gamma_{xy}(s,t) &= \int_{-\infty}^{\infty} \beta_x(s \wedge z)\beta_y(t \wedge z)\,\mathrm{d}H(z|x) \\ &- \int_{-\infty}^{s} \beta_y(t \wedge z)\,\mathrm{d}\Lambda(w|x) - \int_{-\infty}^{t} \frac{1 - H(z|x)}{1 - H(z|y)}\beta_x(s \wedge z)\,\mathrm{d}\Lambda(z|x) \\ &+ \int_{-\infty}^{s\wedge t} \frac{\mathrm{d}H^U(z|x)}{(1 - H(z)|x)(1 - H(z|y))}.\end{aligned}$$

Then

$$\mathcal{C}_n(x, y) = \frac{1}{nb}(g(x))^{-1}(K * K)\left(\frac{y - x}{b}\right) \times$$

$$
\begin{aligned}
&\Big(y^{*2}(1-F(y^*|x))(1-F(y^*|y))\gamma_{xy}(y^*,y^*) \\
&- y^*(1-F(y^*|x))\int_{-\infty}^{y^*}(1-F(t|y))\gamma_{xy}(y^*,t)\,\mathrm{d}t \\
&- y^*(1-F(y^*|y))\int_{-\infty}^{y^*}(1-F(s|x))\gamma_{xy}(s,y^*)\,\mathrm{d}s \qquad (20)\\
&+ \int_{-\infty}^{y^*}\int_{-\infty}^{y^*}(1-F(s|x))(1-F(t|y))\gamma_{xy}(s,t)\,\mathrm{d}s\,\mathrm{d}t\Big) + o\left(n^{-1}\right)
\end{aligned}
$$

where $K * K$ denotes the convolution.

The approximating statistic $\mathcal{A}_{n1}(x)$ is a sum of i.i.d. r.v.'s. Applying the central limit theorem we obtain immediately the asymptotic normality at a fixed point x:

$$\frac{\mathcal{A}_{n1}(x)}{\mathcal{C}_n(x,x)} \xrightarrow{\mathcal{D}} \mathsf{N}(0,1).$$

After some transformations we obtain from Lemma 3.2 for $x = y$

$$
\begin{aligned}
\mathcal{C}_n(x,x) =& \frac{1}{nb}(g(x))^{-1}(K*K)(0) \\
&\times \int_{-\infty}^{y^*}\left(y^*(1-F(y^*)) - A_x(s;y^*)\right)^2 \frac{\mathrm{d}H_x^U(s)}{(1-H_x(s))^2} + o(n^{-1}) \\
=& \frac{1}{nb}\kappa^2\rho^2(x) + o(n^{-1}). \qquad (21)
\end{aligned}
$$

with $A_x(s;y^*) = \int_s^{y^*}(1-F(t|x))\,\mathrm{d}t$ and $\kappa^2 = (K*K)(0)$.
Hence

$$\sqrt{nb}\,\mathcal{A}_{n1}(x) \xrightarrow{\mathcal{D}} \mathsf{N}(0,\rho^2(x)\kappa^2). \qquad (22)$$

Since $\hat{g}_n(x)$ is consistent,

$$\mathcal{A}_{n1}(x) = O_{\mathsf{P}}((nb)^{-1/2})$$

and

$$\hat{g}_n(x) - \mathsf{E}\hat{g}_n(x) = O_{\mathsf{P}}((nb)^{-1/2})$$

we obtain

$$(\mathcal{A}_n(x) - \sum_{i=1}^{n} W_{bi}(x,\mathbb{X})\mathsf{E}(\eta_x(Z_i,\Delta_i)|X_i)) \; - \; \mathcal{A}_{n1}(x)$$
$$= \mathcal{A}_{n1}(x)\frac{\mathsf{E}\hat{g}_n(x) - \hat{g}_n(x)}{\hat{g}_n(x)} \; = \; O_{\mathsf{P}}((nb)^{-1}).$$

Hence

$$\sqrt{nb}\left(\mathcal{A}_n(x) \; - \sum_{i=1}^{n} W_{bi}(x,\mathbb{X})\mathsf{E}(\eta_x(Z_i,\Delta_i)|X_i)\right) \xrightarrow{\mathcal{D}} \mathsf{N}(0,\rho^2(x)\kappa^2). \quad (23)$$

Now, to characterize the systematic part of the deviation define

$$B_1(s,x) \; = \; \int_{-\infty}^{s} \frac{\mathrm{d}\dot{H}^U(t|x)}{1-H(t|x)} + \int_{-\infty}^{s} \frac{\dot{H}(t|x)\,\mathrm{d}H^U(t|x)}{(1-H(t|x))^2}$$

and

$$B_2(s,x) \; = \; \int_{-\infty}^{s} \frac{\mathrm{d}\ddot{H}^U(t|x)}{1-H(t|x)} + \int_{-\infty}^{s} \frac{\ddot{H}(t|x)\,\mathrm{d}H^U(t|x)}{(1-H(t|x))^2}.$$

Using standard techniques for the investigation of a bias we obtain the following asymptotic expansion for the term $\sum_{i=1}^{n} W_{bi}(x,\mathbb{X})\mathsf{E}(\eta_x(Z_i,\Delta_i)|X_i)$:

$$\sum_{i=1}^{n} W_{bi}(x,\mathbb{X})\mathsf{E}(\eta_x(Z_i,\Delta_i)|X_i) \; = \; b_n^2\, B(x)\,\mu_2(K) \; + \; o_{\mathsf{P}}(b_n^2), \quad (24)$$

where

$$B(x) \; = \; \frac{g'(x)}{g(x)}\left(y^*(1-F(y^*))B_1(y^*,x) - \int_{-\infty}^{y^*}(1-F(s))B_1(s,x)\,\mathrm{d}s\right) +$$
$$\frac{1}{2}\left(y^*(1-F(y^*))B_2(y^*,x) - \int_{-\infty}^{y^*}(1-F(s))B_2(s,x)\,\mathrm{d}s\right)$$

and $\mu_2(K) = \int u^2 K(u)\,\mathrm{d}u$.
If $nb_n^5 \to 0$

$$\sqrt{nb_n}\sum_{i=1}^{n} W_{bi}(x,\mathbb{X})\mathsf{E}(\eta_x(Z_i,\Delta_i)|X_i) \; = \; o_{\mathsf{P}}\left((nb_n^5)^{1/2}\right) \; = \; o_{\mathsf{P}}(1),$$

in other words, the systematic part is asymptotically negligible.

If $nb_n^5 \to c > 0$ we have

$$\sqrt{nb_n} \sum_{i=1}^{n} W_{bi}(x, \mathbb{X}) \mathsf{E}(\eta_x(Z_i, \Delta_i)|X_i) \to \sqrt{c}\, B(x)$$

and

$$\sqrt{nb}\, \mathcal{A}_n(x) \xrightarrow{\mathcal{D}} \mathsf{N}(\sqrt{c} B(x) \mu_2(K), \rho^2(x) \kappa^2). \tag{25}$$

By Lemma 3.1 we conclude from the asymptotic behavior of $\mathcal{A}_n(x)$ to that of $\hat{m}_n^*(x) - m^*(x)$ and formulate the following theorem:

Theorem 1 (Asymptotic normality at a fixed point) *Under the assumptions given above and* $b_n \to 0$ *and* $nb_n \to \infty$

(i) *If* $nb_n^5 \to 0$ *then*

$$\sqrt{nb_n}\, (\hat{m}_n^*(x) - m^*(x)) \xrightarrow{\mathcal{D}} \mathsf{N}(0, \kappa^2\, \rho^2(x)) \tag{26}$$

with

$$\rho^2(x) \;=\; (g(x))^{-1} \int_{-\infty}^{y^*} (y^*(1 - F(y^*|x)) - A_x(s; y^*))^2 \, \frac{\mathrm{d}H^U(s|x)}{(1 - H(s|x))^2}.$$

(ii) *If* $nb_n^5 \to c > 0$ *then*

$$\sqrt{nb_n}\, (\hat{m}_n^*(x) - m^*(x)) \xrightarrow{\mathcal{D}} \mathsf{N}(\sqrt{c} B(x) \mu_2(K), \rho^2(x) \kappa^2). \tag{27}$$

Remark: Consider the case without censoring. Using integration by parts we obtain for $y^* = \infty$, $H = H^U = F$

$$\int_{-\infty}^{\infty} A_x^2(s; \infty) \frac{\mathrm{d}F(s|x)}{(1 - F(s|x))^2} \;=\; \mathsf{Var}\,(Y|X = x) \;=\; \sigma^2.$$

Thus, in this case Theorem 1 coincides with the well-known limit theorem stating asymptotic normality of nonparametric kernel regression estimators.

The asymptotic normality at a fixed point characterizes the local behavior. For testing the formulated hypothesis it seems to be better to use a global

deviation measure. So, let us consider the integrated squared difference, weighted by a known function a with $a(x) = 0$ for $x \notin \mathcal{I}$:

$$Q_n = \int (\hat{m}_n^*(x) - m^*(x))^2 a(x)\,dx.$$

Heuristically speaking this is an infinite sum of squares of asymptotically normally distributed r.v.'s. which are asymptotically independent as b_n tends to zero. Thus Q_n, properly standardized, converges in distribution to the standard normal distribution.
Using the method proposed by Hall (1981) for proving the asymptotic normality of the integrated squared error of kernel density estimators one can show the following limit theorem

Theorem 2 (Asymptotic normality of the ISE) *Under the assumptions formulated above and* $nb_n \to \infty$ *and* $n^{\frac{2}{9}} b_n \to 0$

$$Q_n = \int (\hat{m}_n^*(x) - m^*(x))^2 a(x)\,dx$$

$$nb_n^{1/2}\,(Q_n - e_n) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \nu^2)$$

with

$$e_n = e_n(g, H, H^U; K, a, b_n) = (nb_n)^{-1}\kappa^2 \int \rho^2(x) a(x)\,dx$$

$$\nu^2 = \nu^2(g, H, H^U; K, a) = 2\,\kappa_1 \int \rho^4(x)\,a^2(x)\,dx$$

with $\kappa_1 = \int (K * K)^2(x)\,dx$.

4 Formulation of the test procedure

Let us apply the limit theorem for the L_2-type distance of the truncated estimator from the truncated regression function to formulate a test procedure for testing the hypothesis $m \in \mathcal{M}$.
The alternative is characterized by the nonparametric estimator $\hat{m}_n^*$. Suppose the null hypothesis is true. The hypothetical function $m^*(\cdot;\vartheta)$ is unknown. Firstly, one has to estimate the unknown parameter ϑ. There

are several proposals in the literature to do this; we refer to Tsiatis (1990), Ritov (1990) or Bagdonavičius/Nikulin (2001). The basic idea is to replace the unknown cumulative hazard function by an efficient estimator depending on ϑ and to estimate this unknown parameter then by the maximum likelihood method. The authors show that under suitable assumptions the resulting estimator is $\sqrt{n}$-consistent, i.e.

$$\sqrt{n}\left(\hat{\vartheta}_n - \vartheta\right) = O_P(1) \qquad \text{as } n \to \infty. \tag{28}$$

The next step is to determine $m^*(\cdot;\hat{\vartheta})$. With the estimator $\hat{\beta}$ the Breslow estimator for the cumulative hazard function of the unobservable r.v. T_0 is constructed as follows:

$$\hat{\Lambda}_{n0}(t;\hat{\beta}) = \int_0^t \frac{\mathrm{d}\hat{H}^U_{n0}(s;\hat{\beta})}{1-\hat{H}_{n0}(s_-;\hat{\beta})}$$

where

$$\hat{H}_{n0}(t;\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{V}_{0i} \le t)$$

is the empirical distribution function of the estimated hypothetical baseline observations $\hat{V}_{0i} = V_i\psi(X_i,\hat{\beta})$, and

$$\hat{H}^U_{n0}(t;\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{V}_{0i} \le t, \Delta_i = 1)$$

is the corresponding estimator of the subdistribution of the uncensored observations. Then the baseline survival function is estimated by

$$\hat{S}_0(t;\hat{\beta}) = \prod_{s\le t}(1 - \Delta\hat{\Lambda}_0(s;\hat{\beta}))$$

and the hypothetical truncated regression function by

$$\begin{aligned}
\tilde{m}^*(x;\hat{\vartheta}) &= -\int_{-\infty}^{y^*} y\,\mathrm{d}\hat{S}_0(\mathrm{e}^y\psi(x;\hat{\beta})) \\
&= -\int_0^{\mathrm{e}^{y^*}\psi(x;\hat{\beta})} \log\frac{z}{\psi(x;\hat{\beta})}\,\mathrm{d}\hat{S}_0(z) \\
&= -\int_0^{\mathrm{e}^{y^*}\psi(x;\hat{\beta})} \log z\,\mathrm{d}\hat{S}_0(z) + \log\psi(x;\hat{\beta})\left(\hat{S}_0(\mathrm{e}^{y^*}\psi(x;\hat{\beta})) - 1\right).
\end{aligned}$$

We see, for $y^* \to \infty$ the function $\tilde{m}^*(x;\hat{\vartheta})$ converges to

$$-\int_0^\infty \log z \, \mathrm{d}\hat{S}_0(z) - \log\psi(x;\hat{\beta}) = \hat{\mu} - \log\psi(x;\hat{\beta}) = m(x;\hat{\vartheta}).$$

The test procedure has the following form: The hypothesis $m \in \mathcal{M}$, that is $\psi \in \mathcal{F}$ is rejected if the estimated L_2-distance

$$\hat{\mathrm{Q}}_n = \int (\hat{m}_n^*(x) - m^*(x;\hat{\vartheta}))^2 a(x)\,\mathrm{d}x$$

satisfies the inequality

$$\hat{\mathrm{Q}}_n \geq z_{1-\alpha}\frac{\hat{\nu}}{nb_n^{1/2}} + \hat{e}_n$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the limiting distribution and $\hat{e}_n = e_n(\hat{g}_n, \hat{H}_n, \hat{H}^U, K, a, b_n)$ and $\hat{\nu}^2 = \nu^2(\hat{g}_n, \hat{H}_n, \hat{H}^U, K, a)$ are the estimated standardizing terms.

4.1 A proposal for a Monte Carlo procedure

Finally a Monte Carlo method for determining empirical p-values of the test procedure is proposed. The aim of this method is to generate data

$$(V_{ir}^*, X_{ir}^*, \Delta_{ir}^*), \qquad r = 1,\ldots,R, \quad i = 1,\ldots,n$$

according to the hypothetical model. Based on these data the test statistics $\hat{\mathrm{Q}}_{n1},\ldots,\hat{\mathrm{Q}}_{nr},\ldots,\hat{\mathrm{Q}}_{nR}$ are computed and from their empirical distribution the p-value is determined.
The data can be constructed as follows:

1. Let $\hat{\beta}$ the estimator for β based on the original data. Construct the Breslow estimator $\hat{\Lambda}_0(\cdot;\hat{\beta})$ for the cumulative baseline hazard function. Then

$$\hat{S}_0(t;\hat{\beta}) = \prod_{s\leq t}(1 - \Delta\hat{\Lambda}_0(s;\hat{\beta})).$$

2. Generate data T_{0ir}^* from the estimated survival function $\hat{S}_0(t;\hat{\beta})$ and set

$$T_{ir}^* = \frac{T_{0ir}^*}{\psi(X_i;\hat{\beta})}$$

3. Estimate the distribution function of the C_i by the weighted Kaplan-Meier estimator $\hat{G}_n$ and generate censoring variables C_{ir}^* from the estimated survival function $\hat{G}_n$.

4. Finally set

$$V_{ir}^* = \min(T_{ir}^*, C_{ir}^*), \quad \Delta_{ir}^* = 1(T_{ir}^* \leq C_{ir}^*), \quad X_{ir}^* = X_i.$$

As of yet this MC procedure is only a proposal and further investigations should be pursued.

References

[1] M. G. Akritas and Y. Du. IID representation of the conditional Kaplan-Meier process for arbitrary distributions. *Mathematical Methods in Statistics*, 11:152–182, 2002.

[2] V. Bagdonavičius and M. Nikulin. *Accelerated Life Models.* Springer Series in Statistics. Chapman and Hall, 2001.

[3] R. Beran. Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley, 1981.

[4] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer, 2002.

[5] P. Hall. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Analysis*, 14:1–16, 1984.

[6] H. Liero. Testing in nonparametric accelerated life time models. *Austrian Journal of Statistics*, 37(1), 2008.

[7] H. Liero and M. Liero. Testing the influence function in accelerated life time models. In F. Vonta, M. Nikulin, N. Limnios, and C. Huber, editors, *Statistical models and methods for biomedical and technical systems.* Birkhäuser, 2008.

[8] Y. Ritov. Estimation in a linear regression model with censored data. *Ann. Statist.*, 18:303–328, 1990.

[9] A. A. Tsiatis. Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.*, 18:354–372, 1990.