TECHNICAL

REPORT

0245

COX'S REGRESSION MODEL UNDER PARTIALLY INFORMATIVE CENSORING

R. BRAEKERS and N. VERAVERBEKE

IAP STATISTICS NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

COX'S REGRESSION MODEL UNDER PARTIALLY INFORMATIVE CENSORING

Roel BRAEKERS, Noël VERAVERBEKE
Limburgs Universitair Centrum
Universitaire Campus, B-3590 Diepenbeek, Belgium
roel.braekers@luc.ac.be
noel.veraverbeke@luc.ac.be

ABSTRACT

We extend Cox's classical regression model to accomodate partially informative censored data. In this type of data, each observation is the minimum of one lifetime and two censoring times. The survival function of one of these censoring times is a power of the survival function of the lifetime. We call this the informative censoring time. The distribution of the other censoring time has no relation with the distribution of the lifetime. It is called the non-informative censoring time. In this model we specify a semiparametric relation between the lifetime and a covariate where we take into account that also informatively censored observations contribute to this relation. We introduce an estimator for the cumulative baseline hazard function and use maximum likelihood techniques for the estimation of the parameters in the model. Our main results are strong consistency and asymptotic normality of these estimators. The proof uses the general theory of Murphy and van der Vaart (2000) on profile likelihoods. Finally the method is applied to a real data example on survival with malignant melanoma.

AMS 1991 Subject classifications. Primary: 62G05; secondary: 62E20.
Key words and phrases: asymptotic normality, censoring, consistency, Cox's regression model, hazard function, partially informative censoring, profile likelihood inference
Running head: Partially informative censoring
Corresponding author: Noël Veraverbeke

1 Introduction

In the original regression model of Cox (1972), the relationship between a lifetime Y and a covariate X is modelled via the conditional hazard rate function of Y given $X = x$, defined as

$$\lambda(t \mid x) = \lim_{\substack{h \to 0 \\ >}} \frac{1}{h} P(Y < t + h \mid Y \geq t; X = x)$$

Cox's proportional hazards model specifies that $\lambda(t \mid x)$ has the form

$$\lambda(t \mid x) = \lambda_0(t) e^{\beta_0 x}$$

where $\lambda_0(t)$ is an unspecified baseline hazard rate function (the hazard for an individual with $x = 0$) and β_0 is an unknown regression parameter. For simplicity we assume here that the covariate X is one-dimensional, but generalization to vector valued X and β_0 is possible.

In survival analysis applications it typically occurs that independent observations $Y_1, \ldots, Y_n$ on Y are not fully observed. Here we consider the following right censorship pattern: each Y_i may be censored by the minimum of two non-negative variables C_i and D_i and the observed random variables are (Z_i, δ_i) $(i = 1, \ldots, n)$, where $Z_i = Y_i \wedge C_i \wedge D_i$ and $\delta_i = 1$ if $Y_i \leq C_i \wedge D_i$, $\delta_i = 0$ of $C_i \leq Y_i \wedge D_i$ and $\delta_i = -1$ if $D_i \leq Y_i \wedge C_i$. In the absence of regression, this model for censoring has been introduced by Gather and Pawlitschko (1998). They assumed the C_i to be informative censoring times (in the sense defined below), while the D_i were arbitrary non-informative censoring times. A typical example in a clinical study on survival of patients after a treatment could be that C_i describes survival time of the patient till death from other causes while D_i represents survival time of the patient when alive at the end of the study. This partially informative censoring pattern has also been studied nonparametrically in the fixed design regression case by Braekers and Veraverbeke (2001).

We use the following notations for the conditional distribution functions $F(t \mid x) = P(Y \leq t \mid X = x)$, $G_1(t \mid x) = P(C \leq t \mid X = x)$, $G_2(t \mid x) = P(D \leq t \mid X = x)$

and denote their corresponding conditional densities by $f(t \mid x)$, $g_1(t \mid x)$, $g_2(t \mid x)$. We assume that the covariate random variable X has density function $f(x)$.

For our analysis we consider the observed data (X_i, Z_i, δ_i) $(i = 1, \ldots, n)$ as an iid sample from (X, Z, δ), where $Z = Y \wedge C \wedge D$ and $\delta = 1$, 0 or -1 according to $Z = Y$, C or D. Throughout, we also assume that:

(a) Y, C and D are conditionally independent given X (independent censoring)

(b) The conditional distribution function of C given $X = x$ satisfies

$$1 - G_1(t \mid x) = (1 - F(t \mid x))^{\beta_x}$$

for some constant $\beta_x > 0$, depending on the covariate value x (Koziol-Green assumption)

(c) The conditional distribution function of D given $X = x$ does not involve the parameters of interest (non-informative censoring)

(d) The conditional hazard function of Y given $X = x$ has the form

$$\lambda(t \mid x) = \lambda_0(t) e^{\beta_0 x}$$

(proportional hazards assumption)

(e) The parameter β_x in (b) satisfies a model

$$\beta_x = \varphi(x, \boldsymbol{\beta}^{(0)})$$

with φ some strictly positive function and $\boldsymbol{\beta}^{(0)} = (\beta_1, \ldots, \beta_p)$ a vector of p unknown parameters. We assume that φ has partial derivatives of first and second order in a neighborhood of $\boldsymbol{\beta}^{(0)}$. These will be denoted by $\dot{\varphi}_j = \dfrac{\partial \varphi}{\partial \beta_j}$, $\ddot{\varphi}_{ij} = \dfrac{\partial^2 \varphi}{\partial \beta_i \partial \beta_j}$ $(i, j = 1, \ldots, p)$.

(f) $\log \varphi(x, \boldsymbol{\beta}^{(0)})$ is concave and $\log(\varphi(x, \boldsymbol{\beta}^{(0)}) + 1)$ is convex in $\boldsymbol{\beta}^{(0)}$.

(g) The non-informative censoring time D is bounded by some $T_0 > 0$ and $P(Z \geq T_0) > 0$. The value of T_0 is the prespecified time at which the study is terminated.

Remarks

(1) The condition in (b) on the censoring time C reflects a simple model of informative censoring. It is originally due to Koziol and Green (1976) in the case without covariates. In the fixed design regression case it has been studied by Veraverbeke and Cadarso Suárez (2000). This condition is often called 'simple proportional hazards model', but because of confusion with Cox's proportional hazards model, we prefer to call it Koziol-Green assumption.

(2) By direct calculation it is easily seen that the parameter β_x in (b) has the following interpretation:

$$\beta_x = \frac{P(\delta = 0 \mid X = x)}{P(\delta = 1 \mid X = x)}.$$

An important example in (e) is the loglinear model $\log \beta_x = a + bx$, but it is clear that other modelling could be proposed.

In this paper we develop maximum likelihood techniques for joint estimation of the $p + 1$ parameters in this model. There are the p parameters $\beta_1, \dots, \beta_p$ for the modelling of the exponent β_x and the other one is the regression parameter β_0. The likelihood and likelihood equations are established in Sections 2 and 3. We prove consistency and asymptotic normality in Sections 4 and 5 respectively. Finally, in Section 6, the method is implemented in the analysis of a real data example.

2 The likelihood

We begin by calculating the likelihood contribution of an item i with $X = x_i$, $Z = z_i$ and $\delta = d_i$. Under assumption (a) it is given by

$$\lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}P(z_i-\varepsilon\le Z\le z_i+\varepsilon,\delta=d_i\mid X=x_i)$$
$$=\begin{cases} f(z_i\mid x_i)(1-G_1(z_i\mid x_i))(1-G_2(z_i\mid x_i))\dots & \text{if } d_i=1\\ g_1(z_i\mid x_i)(1-F(z_i\mid x_i))(1-G_2(z_i\mid x_i))\dots & \text{if } d_i=0\\ g_2(z_i\mid x_i)(1-F(z_i\mid x_i))(1-G_1(z_i\mid x_i))\dots & \text{if } d_i=-1.\end{cases}$$

Using assumptions (b) - (e), we obtain for the likelihood, after removing non-important factors:

$$\prod_{\delta_i=1} f(Z_i\mid X_i)(1-F(Z_i\mid X_i))^{\beta_{X_i}}\prod_{\delta_i=0}\beta_{X_i}f(Z_i\mid X_i)(1-F(Z_i\mid X_i))^{\beta_{X_i}}$$
$$\cdot\prod_{\delta_i=-1}(1-F(Z_i\mid X_i))^{\beta_{X_i}+1}$$
$$=\prod_{\delta_i=0}\beta_{X_i}\prod_{\delta_i\neq-1}\lambda_0(Z_i)e^{\beta_0X_i}\prod_{i=1}^{n}(1-F(Z_i\mid X_i))^{\beta_{X_i}+1}$$

and, by taking logarithms, we obtain the loglikelihood

$$\sum_{\delta_i=0}\log\varphi(X_i,\boldsymbol{\beta}^{(0)})+\sum_{\delta_i\neq-1}[\log\lambda_0(Z_i)+\beta_0X_i]-\sum_{i=1}^{n}\left(\varphi(X_i,\boldsymbol{\beta}^{(0)})+1\right)e^{\beta_0X_i}\Lambda_0(Z_i)$$

where $\Lambda_0(t)=\int_0^t\lambda_0(u)du$ is the cumulative baseline hazard function.

We want to obtain estimators $\widehat{\beta}_0,\widehat{\beta}_1,\dots,\widehat{\beta}_p$ that maximize this expression. In the ordinary Cox proportional hazards model, the standard inference for the regression parameter β_0 can be based on the partial likelihood, an expression which does not depend on the infinite dimensional nuisance parameter $\lambda_0(t)$ (Cox 1972, 1975). It is well known (see for example the explanation in Fan and Gijbels (1996)) that this gives exactly the same estimator as using the full likelihood in which $\Lambda_0(t)$ is replaced by a 'least informative' nonparametric estimator. In our situation the partial likelihood analysis is not possible, due to the presence of the unknown parameters $\beta_1,\dots,\beta_p$ in the modelling of β_x. Our approach will be a profile likelihood technique in which $\Lambda_0(t)$ is replaced in the full likelihood by a nonparametric maximum likelihood estimator, given $\beta_0,\beta_1,\dots,\beta_p$.

The least informative nonparametric modelling for $\Lambda_0(t)$ is given by

$$\widehat{\Lambda}_0(t)=\sum_{j=1}^{N}\lambda_jI(Y_j^0\le t)$$

where $Y_1^0 < Y_2^0 < \ldots < Y_N^0$ are the N ordered times for which $\delta_i \neq -1$. This is a step function with jumps at any observation which is uncensored or informatively censored. The motivation for this choice is that the nonparametric approach in Gather and Pawlitschko (1998) turns out to be precisely of this type. We then have that

$$\widehat{\Lambda}_0(Z_i) = \sum_{j=1}^{N} \lambda_j I(Y_j^0 \leq Z_i) = \sum_{j=1}^{N} \lambda_j I\ (i \in \mathcal{R}_j)$$

where $\mathcal{R}_j = \{i : Z_i \geq Y_j^0\}$ is the risk set at time Y_j^0-.

With this the loglikelihood becomes

$$\sum_{\delta_i=0} \log \varphi(X_i, \boldsymbol{\beta}^{(0)}) + \sum_{j=1}^{N} [\log \lambda_j + \beta_0 X_{(j)}] - \sum_{i=1}^{n} (\varphi(X_i, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_i} \sum_{j=1}^{N} \lambda_j I(i \in \mathcal{R}_j) \quad (1)$$

where $X_{(1)}, \ldots, X_{(N)}$ are the covariates associated with the ordered $Y_1^0 < Y_2^0 < \ldots < Y_N^0$. Maximization with respect to λ_j gives

$$\widehat{\lambda}_j = \frac{1}{\sum_{i \in \mathcal{R}_j} (\varphi(X_i, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_i}}$$

and substituting this into (1) leads to the profile loglikelihood

$$\sum_{\delta_i=0} \log \varphi(X_i, \boldsymbol{\beta}^{(0)}) + \sum_{j=1}^{N} [- \log \sum_{i \in \mathcal{R}_j} (\varphi(X_i, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_i} + \beta_0 X_{(j)}] - N$$

which has to be maximized with respect to $\beta_0, \beta_1, \ldots, \beta_p$. This is of course equivalent to maximizing

$$\widehat{H}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{\delta_i=0} \log \varphi(X_i, \boldsymbol{\beta}^{(0)}) - \frac{1}{n} \sum_{\delta_i \neq -1} \log \left(\frac{1}{n} \sum_{k \in \mathcal{R}_i} (\varphi(X_k, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_k} \right) + \frac{\beta_0}{n} \sum_{\delta_i \neq -1} X_i. \quad (2)$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}^{(0)}) = (\beta_0, \beta_1, \ldots, \beta_p)$.

3 The likelihood equations

The estimators $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ are solutions to the equations $\frac{\partial \widehat{H}}{\partial \beta_0} = \ldots = \frac{\partial \widehat{H}}{\partial \beta_p} = 0$, that is

$$\sum_{\delta_i \neq -1} X_i - \sum_{\delta_i \neq -1} \frac{\sum_{k \in \mathcal{R}_i} X_k (\varphi(X_k, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_k}}{\sum_{k \in \mathcal{R}_i} (\varphi(X_k, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_k}} = 0 \quad (3)$$

$$\sum_{\delta_i=0} \frac{\dot{\varphi}_j(X_i, \boldsymbol{\beta}^{(0)})}{\varphi(X_i, \boldsymbol{\beta}^{(0)})} - \sum_{\delta_i \neq -1} \frac{\sum\limits_{k \in R_i} \dot{\varphi}_j(X_k, \boldsymbol{\beta}^{(0)}) e^{\beta_0 X_k}}{\sum\limits_{k \in R_i} (\varphi(X_k, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_k}} = 0 \qquad (j = 1 \ldots p). \tag{4}$$

It will be convenient to introduce the following shorthand notations. For any continuous function g, we put:

$$\begin{aligned}
E(g(x), t) &= \int g(x) P(Z \geq t \mid X = x) f(x) dx \\
E^0(g(x), t) &= \int g(x) P(Z \geq t, \delta = 0 \mid X = x) f(x) dx \\
E^1(g(x), t) &= \int g(x) P(Z \geq t, \delta = 1 \mid X = x) f(x) dx \\
E^{0,1}(g(x), t) &= \int g(x) P(Z \geq t, \delta \neq -1 \mid X = x) f(x) dx
\end{aligned}$$

where $f(x)$ is the density function of the covariate X.

The empirical versions will be denoted by $\widehat{E}(g(x), t)$, $\widehat{E}^0(g(x), t)$, etc. For example,

$$\widehat{E}^{0,1}(g(x), t) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) I(Z_i \geq t, \delta_i \neq -1).$$

Further abbreviations will be

$$\begin{aligned}
Q(t) &= P(Z \geq t, \delta \neq -1) \\
\widehat{Q}(t) &= \frac{1}{n} \sum_{i=1}^{n} I(Z_i \geq t, \delta_i \neq -1).
\end{aligned}$$

With this, $\widehat{H}$ in (2) can be rewritten as

$$\widehat{H}(\boldsymbol{\beta}) = \widehat{E}^0(\log \varphi(x, \boldsymbol{\beta}^{(0)}), 0) + \int_0^{T_0} \log \widehat{E}((\varphi(x, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 x}, t) d\widehat{Q}(t) + \widehat{E}^{0,1}(\beta_0 x, 0). \tag{5}$$

Consider the ‘population version’ of (5):

$$H(\boldsymbol{\beta}) = E^0(\log \varphi(x, \boldsymbol{\beta}^{(0)}), 0) + \int_0^{T_0} \log E((\varphi(x, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 x}, t) dQ(t) + E^{0,1}(\beta_0 x, 0). \tag{6}$$

In the following Lemma 1, we show that this function H is concave.

Lemma 1. Under assumption (f), we have that the functions $\widehat{H}$ and H are concave. Furthermore, the first order partial derivatives of the function H are zero at $\boldsymbol{\beta}$:

$$\frac{\partial H}{\partial \beta_0} = E^{0,1}(x,0) + \int_0^{T_0} \frac{E(x(\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},t)}{E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},t)} dQ(t) = 0.$$

$$\frac{\partial H}{\partial \beta_j} = E^0\left(\frac{\dot{\varphi}_j(x,\boldsymbol{\beta}^{(0)})}{\varphi(x,\boldsymbol{\beta}^{(0)})},0\right) + \int_0^{T_0} \frac{E(\dot{\varphi}_j(x,\boldsymbol{\beta}^{(0)})e^{\beta_0 x},t)}{E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},t)} dQ(t) = 0 \qquad (j = 1,\ldots,p).$$

Proof. Take $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}^{(0)})$, $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\beta}^{*(0)})$ and $0 < \lambda < 1$. From (f) we have, by taking expectations, that

$$\lambda E^0(\log(\varphi(x,\boldsymbol{\beta}^{(0)})),0) + (1-\lambda)E^0(\log(\varphi(x,\boldsymbol{\beta}^{*(0)})),0) \leq E^0(\log\varphi(x,\lambda\boldsymbol{\beta}^{(0)} + (1-\lambda)\boldsymbol{\beta}^{*(0)}),0)$$

Furthermore we find:

$$\begin{aligned}&\left[\varphi(x,\lambda\boldsymbol{\beta}^{(0)} + (1-\lambda)\boldsymbol{\beta}^{*(0)}) + 1\right] e^{\left[\lambda\beta_0 + (1-\lambda)\beta_0^*\right]x} \\ &\leq \left[(\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x}\right]^{\lambda} \left[(\varphi(x,\boldsymbol{\beta}^{*(0)})+1)e^{\beta_0^* x}\right]^{1-\lambda}\end{aligned}$$

and hence

$$\begin{aligned}&E((\varphi(x,\lambda\boldsymbol{\beta}^{(0)} + (1-\lambda)\boldsymbol{\beta}^{*(0)}) + 1)e^{(\lambda\beta_0 + (1-\lambda)\beta_0^*)x}, t) \\ &\leq \int \left[(\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x}\right]^{\lambda} \left[(\varphi(x,\boldsymbol{\beta}^{*(0)})+1)e^{\beta_0^* x}\right]^{1-\lambda} P(Z \geq t | X = x) f(x) dx \\ &\leq E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},t)^{\lambda} E((\varphi(x,\boldsymbol{\beta}^{*(0)})+1)e^{\beta_0^* x},t)^{1-\lambda}\end{aligned}$$

where we used Hölder's inequality. Taking logarithms and expectations with respect to the (decreasing) function $Q(t)$ gives:

$$\begin{aligned}&\lambda \int_0^{T_0} \log E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},t) dQ(t) + (1-\lambda)\int_0^{T_0} \log E((\varphi(x,\boldsymbol{\beta}^{*(0)})+1)e^{\beta_0^* x},t) dQ(t) \\ &\leq \int_0^{T_0} \log E\left[(\varphi(x,\lambda\boldsymbol{\beta}^{(0)} + (1-\lambda)\boldsymbol{\beta}^{*(0)}) + 1)e^{(\lambda\beta_0 + (1-\lambda)\beta_0^*)x}, t\right] dQ(t).\end{aligned}$$

So we get : $\lambda H(\boldsymbol{\beta}) + (1-\lambda)H(\boldsymbol{\beta}^*) \leq H(\lambda\boldsymbol{\beta} + (1-\lambda)\boldsymbol{\beta}^*)$.

This means that H is concave. By changing integrals into sums, we get that $\widehat{H}$ is also concave.

That the partial derivatives of H vanish at $\boldsymbol{\beta}$ can be seen through the following two relations. For any continuous function g we have that:

$$\frac{E(g(x)\beta_x e^{\beta_0 x}, t)}{E((\beta_x + 1)e^{\beta_0 x}, t)} dQ(t) = dE^0(g(x), t) \tag{7}$$

and

$$\frac{E(g(x)(\beta_x + 1)e^{\beta_0 x}, t)}{E((\beta_x + 1)e^{\beta_0 x}, t)} dQ(t) = dE^{0,1}(g(x), t). \tag{8}$$

We only show the derivation of (7). We have

$$\begin{aligned} P(Z \geq t, \delta \neq -1 \mid X = x) &= (1 + \beta_x) \int_t^\infty (1 - G_2(u \mid x))(1 - F(u \mid x))^{\beta_x} dF(u \mid x) \\ &= (1 + \beta_x) \int_t^\infty (1 - G_2(u \mid x))(1 - F(u \mid x))^{\beta_x + 1} \lambda(u \mid x) du \\ &= (1 + \beta_x) e^{\beta_0 x} \int_t^\infty P(Z \geq u \mid X = x) \lambda_0(u) du \end{aligned}$$

and hence

$$dQ(t) = -\lambda_0(t) E((1 + \beta_x)e^{\beta_0 x}, t) dt. \tag{9}$$

Also, similarly,

$$P(Z \geq t, \delta = 0 \mid X = x) = \beta_x e^{\beta_0 x} \int_t^\infty P(Z \geq u \mid X = x) \lambda_0(u) du$$

and hence

$$dE^0(g(x), t) = -\lambda_0(t) E(g(x)\beta_x e^{\beta_0 x}, t) dt. \tag{10}$$

The relation (7) now follows from (9) and (10).

4 Strong consistency

In Theorem 1 of this section we establish the existence of a strongly consistent solution to the likelihood equations. We also prove Lemma 2 which will be used in the proof of the next section. It concerns the consistency of an estimator for the cumulative hazard function $\Lambda_0(t)$.

Theorem 1. Assume that $E|X| < \infty$ and that $E|\log \varphi(X, \boldsymbol{\beta}^{(0)})|$ and $E[((\varphi(X, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 X})^2]$ are bounded uniformly in a neighborhood of $\boldsymbol{\beta}$. There exists a sequence of solutions $\widehat{\boldsymbol{\beta}}$ of the equations (3) - (4) such that

$$\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$$

a.s. as $n \to \infty$.

Proof. Lemma 1 implies that the function H has a local maximum at $\boldsymbol{\beta}$. Hence for $\boldsymbol{\beta}^*$ in a δ-neighborhood of $\boldsymbol{\beta}$ ($\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| \leq \delta$, with $\| \ \|$ Euclidean distance) we have that

$$H(\boldsymbol{\beta}) - H(\boldsymbol{\beta}^*) \geq 0 \tag{11}$$

with strict inequality if $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = \delta$. From the strong law of large numbers together with Lemmas A1 and A2 in Tsiatis (1981), it follows that

$$\widehat{H}(\boldsymbol{\beta}) - \widehat{H}(\boldsymbol{\beta}^*) \to H(\boldsymbol{\beta}) - H(\boldsymbol{\beta}^*). \tag{12}$$

Relations (11) and (12) entail that, on a set of probability one, there exists an n_0 such that for all $n \geq n_0$:

$$\widehat{H}(\boldsymbol{\beta}) - \widehat{H}(\boldsymbol{\beta}^*) > 0 \text{ for } \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = \delta. \tag{13}$$

In Lemma 1 we saw that $\widehat{H}$ is concave. So $\widehat{H}$ has a local maximum on $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| \leq \delta$. This maximum cannot be on the boundary ($\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = \delta$) since (13). A consequence of this is that the first derivatives vanish somewhere on $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| < \delta$. The value where $\dfrac{\partial \widehat{H}}{\partial \beta_0} = \ldots = \dfrac{\partial \widehat{H}}{\partial \beta_p} = 0$ is the ML-estimate $\widehat{\boldsymbol{\beta}}$ which was discussed in Section 3. We can now repeat this argument for δ decreasing with n. In this way, we get a sequence $\widehat{\boldsymbol{\beta}}_n$ with $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}$ a.s. as $n \to \infty$.

In the next section we will need the estimator for $\Lambda_0(t)$ which is obtained by maximizing the likelihood for a fixed value of $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}^{(0)})$. It is given by

$$\widehat{\Lambda}_{\boldsymbol{\beta}}(t) = \sum_{j=1}^{n} \frac{I(Z_j \leq t, \delta_j \neq -1)}{\sum\limits_{i \in \mathcal{R}_j} (\varphi(X_i, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_i}}. \tag{14}$$

We have the following consistency result.

Lemma 2. Assume that $E[((\varphi(X, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 X})^2]$ is bounded uniformly in a neighborhood of $\boldsymbol{\beta}$. If $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^{(0)})$ is any random sequence with $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ as $n \to \infty$, then

$$\sup_{0 \leq t \leq T_0} |\widehat{\Lambda}_{\widehat{\boldsymbol{\beta}}}(t) - \Lambda_0(t)| \xrightarrow{P} 0.$$

Proof. From (9) it follows that

$$\Lambda_0(t) = \int_0^t \frac{-dQ(s)}{E((\varphi(x, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 x}, s)}.$$

Also $\widehat{\Lambda}_{\widehat{\boldsymbol{\beta}}}(t)$ can be rewritten as

$$\widehat{\Lambda}_{\widehat{\boldsymbol{\beta}}}(t) = \int_0^t \frac{-d\widehat{Q}(s)}{\widehat{E}((\varphi(x, \widehat{\boldsymbol{\beta}}^{(0)}) + 1)e^{\widehat{\beta}_0 x}, s)}.$$

We have that

$$\begin{aligned}
&\sup_{0 \leq t \leq T_0} \left| \widehat{E}((\varphi(x, \widehat{\boldsymbol{\beta}}^{(0)}) + 1)e^{\widehat{\beta}_0 x}, t) - E((\varphi(x, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 x}, t) \right| \\
\leq &\sup_{0 \leq t \leq T_0} \left| \widehat{E}((\varphi(x, \widehat{\boldsymbol{\beta}}^{(0)}) + 1)e^{\widehat{\beta}_0 x}, t) - E((\varphi(x, \widehat{\boldsymbol{\beta}}^{(0)}) + 1)e^{\widehat{\beta}_0 x}, t) \right| \\
+ &\sup_{0 \leq t \leq T_0} \left| E((\varphi(x, \widehat{\boldsymbol{\beta}}^{(0)}) + 1)e^{\widehat{\beta}_0 x}, t) - E((\varphi(x, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 x}, t) \right|.
\end{aligned}$$

The first term tends to zero a.s. by Lemma A1 in Tsiatis(1981). The second term tends to zero in probability since $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ and since the function $\sup\limits_{0 \leq t \leq T_0} \Big| E((\varphi(x, \widetilde{\boldsymbol{\beta}}^{(0)}) + 1)e^{\widetilde{\beta}_0 x}, t)$ $-E((\varphi(x, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 x}, t) \Big|$ is continuous in $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \widetilde{\boldsymbol{\beta}}^{(0)})$.

This leads to

$$\begin{aligned}
&\sup_{0\le t\le T_0} |\widehat{\Lambda}_{\widehat{\boldsymbol{\beta}}}(t)-\Lambda_0(t)| \\
\le\ &\sup_{0\le t\le T_0}\left|\int_0^t \frac{-d\widehat{Q}(s)}{\widehat{E}((\varphi(x,\widehat{\boldsymbol{\beta}}^{(0)})+1)e^{\widehat{\beta}_0 x},s)}-\int_0^t \frac{-dQ(s)}{\widehat{E}((\varphi(x,\widehat{\boldsymbol{\beta}}^{(0)})+1)e^{\widehat{\beta}_0 x},s)}\right| \\
+\ &\sup_{0\le t\le T_0}\left|\int_0^t \frac{-dQ(s)}{\widehat{E}((\varphi(x,\widehat{\boldsymbol{\beta}}^{(0)})+1)e^{\widehat{\beta}_0 x},s)}-\int_0^t \frac{-dQ(s)}{E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},s)}\right| \\
\le\ &\frac{\sup_{0\le t\le T_0}|\widehat{Q}(t)-Q(t)|}{\widehat{E}((\varphi(x,\widehat{\boldsymbol{\beta}}^{(0)})+1)e^{\widehat{\beta}_0 x},T_0)} \\
+\ &\frac{\sup_{0\le t\le T_0}\left|\widehat{E}((\varphi(x,\widehat{\boldsymbol{\beta}}^{(0)})+1)e^{\widehat{\beta}_0 x},t)-E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},t)\right|}{\widehat{E}((\varphi(x,\widehat{\boldsymbol{\beta}}^{(0)})+1)e^{\widehat{\beta}_0 x},T_0)E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},T_0)}
\end{aligned}$$

which finishes the proof of the lemma.

5 Asymptotic normality

Theorem 2. Assume that $E|\log\varphi(X,\boldsymbol{\beta}^{(0)})|$, $E(|X|^5(\varphi(X,\boldsymbol{\beta}^{(0)})+1)^2e^{2\beta_0 X})$, $E(X^2|\dot{\varphi}_j(X,\boldsymbol{\beta}^{(0)})|e^{\beta_0 X})$, $E\left(\frac{\dot{\varphi}_j(X,\boldsymbol{\beta}^{(0)})\dot{\varphi}_{j'}(X,\boldsymbol{\beta}^{(0)})e^{2\beta_0 X}}{\varphi^2(X,\boldsymbol{\beta}^{(0)})}\right)$ and $E\left(\frac{X^2|\ddot{\varphi}_{jj'}(X,\boldsymbol{\beta}^{(0)})|}{\varphi(X,\boldsymbol{\beta}^{(0)})}\right)$ for all $j,j'=1,\dots,p$ are bounded uniformly in a neighborhood of $\boldsymbol{\beta}$. Then the solution $\widehat{\boldsymbol{\beta}}$ given in Theorem 1 is asymptotically normal as $n\to\infty$:

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N\left(\mathbf{0}\ ;I^{-1}\right)$$

where $\mathbf{0}=(0,\dots,0)$ and I is the information matrix of the function H

$$I=I(\boldsymbol{\beta})=\left(-\frac{\partial^2 H}{\partial\beta_i\partial\beta_j}\right)\qquad(i,j=0,1,\dots,p).$$

Proof. We follow the general approach of Murphy and van der Vaart (2000) for verifying the validity of the profile likelihood method. More in particular we will check the conditions of their Theorem 1, which guarantees that the profile likelihood allows an asymptotic expansion, which then leads to the asymptotic normality of the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$.

We had as loglikelihood in Section 2

$$\begin{aligned}&\sum_{\delta_i=0} \log \varphi(X_i, \boldsymbol{\beta}^{(0)}) + \sum_{\delta_i \neq -1} [\log \lambda_0(Z_i) + \beta_0 X_i] - \sum_{i=1}^{n} (\varphi(X_i, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 X_i} \Lambda_0(Z_i)\\ &= \sum_{i=1}^{n} \log L(\boldsymbol{\beta}, \Lambda_0)(X_i, \delta_i, Z_i)\end{aligned}$$

where $\log L(\boldsymbol{\beta}, \Lambda_0)(x, \delta, z)$

$$= I(\delta = 0) \log \varphi(x, \boldsymbol{\beta}^{(0)}) + I(\delta \neq -1) [\log \lambda_0(z) + \beta_0 x] - (\varphi(x, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 x} \Lambda_0(z)$$

is the contribution of the datapoint (x, δ, z).

We start by calculating the score functions for $\boldsymbol{\beta}$ and Λ_0. The parameter $\boldsymbol{\beta}$ is finite dimensional, so the score function is the vector $S(\boldsymbol{\beta}, \Lambda_0)(x, \delta, z)$ of partial derivatives of $\log L(\boldsymbol{\beta}, \Lambda_0)(x, \delta, z)$ with respect to β_j $(j = 0, 1, \ldots, p)$:

$$S(\boldsymbol{\beta}, \Lambda_0)(x, \delta, z) = \begin{pmatrix} S_0(\boldsymbol{\beta}, \Lambda_0)(x, \delta, z) \\ S_1(\boldsymbol{\beta}, \Lambda_0)(x, \delta, z) \\ \cdots \\ S_p(\boldsymbol{\beta}, \Lambda_0)(x, \delta, z) \end{pmatrix} = \begin{pmatrix} I(\delta \neq -1) x - (\varphi(x, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 x} \Lambda_0(z) \\ I(\delta = 0) \dfrac{\dot{\varphi}_1(x, \boldsymbol{\beta}^{(0)})}{\varphi(x, \boldsymbol{\beta}^{(0)})} - \dot{\varphi}_1(x, \boldsymbol{\beta}^{(0)}) e^{\beta_0 x} \Lambda_0(z) \\ \cdots \\ I(\delta = 0) \dfrac{\dot{\varphi}_p(x, \boldsymbol{\beta}^{(0)})}{\varphi(x, \boldsymbol{\beta}^{(0)})} - \dot{\varphi}_p(x, \boldsymbol{\beta}^{(0)}) e^{\beta_0 x} \Lambda_0(z) \end{pmatrix}.$$

For the score function of the infinite dimensional nuisance parameter Λ_0, we use $\frac{\partial}{\partial t} \log L(\boldsymbol{\beta}, \Lambda_t)(x, \delta, z)|_{t=0}$, where $\Lambda_t(z) = \int\limits_0^z (1 + th(s)) d\Lambda_0(s)$ with $h : I\!\!R \to I\!\!R$ some bounded function. The boundedness of h entails that Λ_t is an absolutely continuous cumulative hazard function for $|t|$ small. This gives

$$\frac{\partial}{\partial t} \log L(\boldsymbol{\beta}, \Lambda_t)(x, \delta, z)|_{t=0} = I(\delta \neq -1) h(z) - (\varphi(x, \boldsymbol{\beta}^{(0)}) + 1) e^{\beta_0 x} \int\limits_0^z h(s) d\Lambda_0(s) := Ah(z)$$

where $A : L_2(\Lambda_0) \to L_2(\boldsymbol{\beta}, \Lambda_0)$ is a bounded linear operator between the Hilbert spaces $L_2(\Lambda_0)$ and $L_2(\boldsymbol{\beta}, \Lambda_0)$ with in-products given by, respectively $< f, g >_{\Lambda_0} = \int\limits_0^{T_0} f(s) g(s) d\Lambda_0(s)$ and $< f, g >_{\boldsymbol{\beta}, \Lambda_0} = \int f g dP(x, \delta, z)$.

The score function depends on the infinite dimensional nuisance parameter Λ_0 and therefore we calculate the efficient score function for $\boldsymbol{\beta}$, i.e. the original score function minus its original projection onto the score function of the nuisance parameter Λ_0. Since A is a linear operator, this efficient score function is given by

$$\widetilde{S}(\boldsymbol{\beta},\Lambda_0)(x,\delta,z)=S(\boldsymbol{\beta},\Lambda_0)(x,\delta,z)-A(A^*A)^{-}A^*S(\boldsymbol{\beta},\Lambda_0)(x,\delta,z) \tag{15}$$

where A^* is the adjoint operator and $(A^*A)^-$ is a generalized inverse.
The identity $< Ah, g >=< h, A^*g >$, for every $h \in L_2(\Lambda_0)$ and $g \in L_2(\boldsymbol{\beta},\Lambda_0)$ can be used to find expressions for A^*A and A^*. Direct calculations give

$$\begin{aligned}(A^*A)^-g(z) &= \frac{g(z)}{E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},z)}\\ A^*S_0(\boldsymbol{\beta},\Lambda_0)(x,\delta,z) &= E(x(\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},z)\\ A^*S_j(\boldsymbol{\beta},\Lambda_0)(x,\delta,z) &= E(\dot{\varphi}_j(x,\boldsymbol{\beta}^{(0)})e^{\beta_0 x},z) \quad (j=1,\dots,p).\end{aligned}$$

Hence the efficient score function for $\boldsymbol{\beta}$ has components

$$\begin{aligned}\widetilde{S}_0(\boldsymbol{\beta},\Lambda_0)(x,\delta,z) = I(\delta\neq -1)[x-\frac{E(x(\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},z)}{E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},z)}]\\ -(\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x}\int_0^z[x-\frac{E(x(\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},s)}{E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},s)}d\Lambda_0(s)]\end{aligned}$$

and, for $j=1,\dots,p$,

$$\begin{aligned}\widetilde{S}_j(\boldsymbol{\beta},\Lambda_0)(x,\delta,z) = \frac{I(\delta=0)\dot{\varphi}_j(x,\boldsymbol{\beta}^{(0)})}{\varphi(x,\boldsymbol{\beta}^{(0)})}-\frac{I(\delta\neq -1)E(\dot{\varphi}_j(x,\boldsymbol{\beta}^{(0)})e^{\beta_0 x},z)}{E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},z)}\\ -\int_0^z[\dot{\varphi}_j(x,\boldsymbol{\beta}^{(0)})e^{\beta_0 x}-\frac{(\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x}E(\dot{\varphi}_j(x,\boldsymbol{\beta}^{(0)})e^{\beta_0 x},s)}{E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},s)}]d\Lambda_0(s).\end{aligned}$$

For the covariance matrix $I=(I_{ij})$ of this efficient score function we obtain, after long but straightforward calculations, that for $i,j=0,1,\dots,p$,

$$I_{ij}=E(\widetilde{S}_i(\boldsymbol{\beta},\Lambda_0)(X,\delta,Z).\widetilde{S}_j(\boldsymbol{\beta},\Lambda_0)(X,\delta,Z))=-\frac{\partial^2 H}{\partial\beta_i\partial\beta_j}.$$

Since, by Lemma 1, H is a concave function, we have that I is a positive definite matrix.

In the remaining part of the proof we have to define an approximately least favorable submodel and verify the conditions of Theorem 1 in Murphy and van der Vaart (2000).

For any $(\widetilde{\boldsymbol{\beta}}, \Lambda)$ and $\boldsymbol{t} = (t_0, t_1, \ldots, t_p) = (t_0, \boldsymbol{t}^{(0)})$, we define the approximately least favorable submodel by

$$\Lambda_{\boldsymbol{t}}(\widetilde{\boldsymbol{\beta}}, \Lambda)(z) = \int_0^z [1 + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{t}) h_0(s)] d\Lambda(s)$$

where h_0 is the least favorable direction given by

$$h_0(z) = \begin{pmatrix} h_{00}(z) \\ h_{01}(z) \\ \cdots \\ h_{0p}(z) \end{pmatrix} = \frac{1}{E((\varphi(x, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 x}, z)} \begin{pmatrix} E(x(\varphi(x, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 x}, z) \\ E(\dot{\varphi}_1(x, \boldsymbol{\beta}^{(0)})e^{\beta_0 x}, z) \\ \cdots \\ E(\dot{\varphi}_p(x, \boldsymbol{\beta}^{(0)})e^{\beta_0 x}, z) \end{pmatrix}.$$

We see that at $\boldsymbol{t} = \widetilde{\boldsymbol{\beta}}$, $\Lambda_{\widetilde{\boldsymbol{\beta}}}(\widetilde{\boldsymbol{\beta}}, \Lambda)(z) = \Lambda(z)$, which is condition (8) in Murphy and van der Vaart (2000). Next we define the function $l(\boldsymbol{t}, \widetilde{\boldsymbol{\beta}}, \Lambda)$ as

$$l(\boldsymbol{t}, \widetilde{\boldsymbol{\beta}}, \Lambda)(x, \delta, z) = \log L(\boldsymbol{t}, \Lambda_{\boldsymbol{t}}(\widetilde{\boldsymbol{\beta}}, \Lambda))(x, \delta, z).$$

The remaining conditions in Murphy and van der Vaart (2000) are on the vector $\dot{l}$ of first order partial derivatives and the matrix $\ddot{l}$ of second order partial derivatives of the function l. By way of example we only calculate $\dfrac{\partial l}{\partial t_0}$ and $\dfrac{\partial^2 l}{\partial t_0^2}$:

$$\begin{aligned} \frac{\partial l}{\partial t_0}(\boldsymbol{t}, \widetilde{\boldsymbol{\beta}}, \Lambda)(x, \delta, z) &= I(\delta \neq -1)[x - h_{00}(z)] \\ &- (\varphi(x, \boldsymbol{t}^{(0)}) + 1)e^{t_0 x} \int_0^z [x - h_{00}(s)] d\Lambda_{\boldsymbol{t}}(\widetilde{\boldsymbol{\beta}}, \Lambda)(s) \\ \frac{\partial^2 l}{\partial t_0^2}(\boldsymbol{t}, \widetilde{\boldsymbol{\beta}}, \Lambda)(x, \delta, z) &= (\varphi(x, \boldsymbol{t}^{(0)}) + 1)e^{t_0 x} \int_0^z [h_{00}^2(s) - x^2] d\Lambda_{\boldsymbol{t}}(\widetilde{\boldsymbol{\beta}}, \Lambda)(s). \end{aligned}$$

We have that the functions $\dot{l}(\boldsymbol{t}, \widetilde{\boldsymbol{\beta}}, \Lambda)$ and $\ddot{l}(\boldsymbol{t}, \widetilde{\boldsymbol{\beta}}, \Lambda)$ are continuous at $(\boldsymbol{\beta}, \boldsymbol{\beta}, \Lambda_0)$. If we evaluate the vector $\dot{l}$ at the true parameter, we find

$$\dot{l}(\boldsymbol{\beta}, \boldsymbol{\beta}, \Lambda_0) = \widetilde{S}(\boldsymbol{\beta}, \Lambda_0)$$

where $\widetilde{S}(\boldsymbol{\beta},\Lambda_0)$ is the vector of efficient scores given in (15).

The next conditions (10) and (11) in Murphy and van der Vaart (2000) require that for any random sequence $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$, we should have that

$$\widehat{\Lambda}_{\widehat{\boldsymbol{\beta}}}(t) \xrightarrow{P} \Lambda_0(t) \tag{16}$$

and

$$E(\dot{l}(\boldsymbol{\beta},\widehat{\boldsymbol{\beta}},\Lambda_{\widehat{\boldsymbol{\beta}}})) = o_P(||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}||+n^{-1/2}) \tag{17}$$

where $\widehat{\Lambda}_{\widehat{\boldsymbol{\beta}}}(t)$ is the cumulative hazard estimator in (14). Condition (16) follows from our Lemma 2 above, while (17) is according to Murphy and van der Vaart (2000) equivalent to

$$E(\dot{l}(\boldsymbol{\beta},\boldsymbol{\beta},\Lambda_{\widehat{\boldsymbol{\beta}}})) = o_P(||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}||+n^{-1/2}). \tag{18}$$

But this is trivially true in our situation, since some easy calculations show that the left hand side in (18) is equal to zero, independent of the function $\Lambda_{\widehat{\boldsymbol{\beta}}}$. For any cumulative hazard function Λ, we have that

$$E(\dot{l}(\boldsymbol{\beta},\boldsymbol{\beta},\Lambda)) = 0.$$

The last condition requires that the class of functions $\{\dot{l}(\boldsymbol{t},\widetilde{\boldsymbol{\beta}},\Lambda)|(\boldsymbol{t},\widetilde{\boldsymbol{\beta}},\Lambda)\in V\}$ is Donsker and that the class $\{\ddot{l}(\boldsymbol{t},\widetilde{\boldsymbol{\beta}},\Lambda)|(\boldsymbol{t},\widetilde{\boldsymbol{\beta}},\Lambda)\in V\}$ is Glivenko-Cantelli in a neighborhood V of the true parameter $(\boldsymbol{\beta},\boldsymbol{\beta},\Lambda_0)$. From page 270 in van der Vaart (1998) it suffices to verify these properties componentwise. We use the bound on the bracketing number in Corollary 2.7.4. of van der Vaart and Wellner (1996). In this corollary, we divide $I\!R\times[0,T_0]$ into a partition of bounded, convex sets. By assumption (g), the functions $\dot{l}(\boldsymbol{t},\widetilde{\boldsymbol{\beta}},\Lambda)$ and $\ddot{l}(\boldsymbol{t},\widetilde{\boldsymbol{\beta}},\Lambda)$ are uniformly bounded as a function of z. This is not the case when we look at these as functions of the covariate x. Therefore we take a partition $I\!R\times[0,T_0]=\bigcup_{j\in\mathbb{Z}}]j-1/2,j+1/2]\times[0,T_0]$. As explained in van der Vaart and Wellner (1996, page 159), checking the Donsker (or Glivenko-Cantelli) property can be done by establishing the convergence of the series $\sum_j M_jP^{1/2}(I_j)$ (or $\sum_j \widetilde{M}_jP(I_j)$) where M_j (or

$\widetilde{M}_j$) is the maximum of a component of $\dot{l}$ (or $\ddot{l}$) in each set I_j of the partition and $P(I_j)$ is the probability of this set. The moment conditions allow to use a Markov bound for $P(I_j)$, which makes the above series convergent and hence the Donsker and Glivenko-Cantelli properties are proved.

All the conditions of Theorem 1 of Murphy and van der Vaart (2000) are satisfied. This, together with Corollary 1 of the same paper, proves the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ and finishes our proof.

6 Example: survival with malignant melanoma

In this section we illustrate our estimation method with the analysis of clinical trial data on malignant melanoma (skin cancer) of the Department of Plastic Surgery, University Hospital of Odense, Denmark. See example I.3.1 in Anderson *et al* (1993). This study took place in the period 1962-77 and looked at the survival of 225 patients after their tumor was completely removed. Along with the survival time, several covariates, like sex, age, ... were recorded. Twenty patients were left out of the study due to missing values in their covariates. For each of the remaining 205 patients, they recorded the cause of death or whether the patient was alive at the end of the study.

As an example, we study survival time till death from malignant melanoma versus sex of the patient as a covariate. (It is obvious that our results above also cover the discrete covariate case). As informative censoring variable we take the survival time till death from other causes. There are several reasons for this. A first reason is that we actually observe the death of an individual within the time interval under study. A second reason is that this cause of death is presumably an indirect consequence of malignant melanoma. In this way we also have an explanation for the difference we make between informative and non-informative censoring in this model. As non-informative censoring variable, we use the survival time of the patient when alive at the end of the study. For such individuals we do not know whether they will ever experience a death caused by malignant melanoma or a death which is an indirect consequence of malignant melanoma.

To study whether survival time till death from malignant melanoma is different for the sexes, we recall the two basic equations of our model:

$$\lambda_F(t \mid x) = \lambda_0(t) e^{\beta_0 x} \tag{19}$$

$$\beta_x = \frac{P(\delta = 0 \mid X = x)}{P(\delta = 1 \mid X = x)} = \varphi(x; \beta_1, \ldots, \beta_p). \tag{20}$$

The equation (19) expresses the hazard function of the uncensored observations as a function of the covariate. Equation (20) models the ratio of the uncensored and informatively censored observations as a function of the covariate. In this example we take this function as $\varphi(x; \beta_1, \beta_2) = e^{\beta_1 + \beta_2 x}$. There are several reasons for this choice. A first reason is that this ratio of probabilities reminds of the generalized logit model. In this case, our model contains an important submodel. If we take $\beta_1 = \beta_2 = 0$ then this model reduces to the ordinary Cox-model where we compare the group of uncensored and informatively censored observations with the group of non-informatively censored observation. A second reason for this choice of the function φ is that it simplifies the calculations in such a way that we can use existing statistical software to compute the estimate for the different parameters. This is not possible for other choices of φ like for example: $\varphi(x, b) = \dfrac{e^{-bx}}{1 + e^{bx}}$.
Table 1 shows the number of the different type of observations for each value of the covariate sex.

Table 1

Sex	uncens.($\delta = 1$)	infor.cens.($\delta = 0$)	non-infor.cens($\delta = -1$)	Total
Female (0)	28	7	91	126
Male (1)	29	7	43	79
Total	57	14	134	205

Note that the percentage of censored observations in this dataset is high (75%) and that the major part of this censoring is non-informative.
To estimate jointly the parameters in (19) and (20), we were able to use two different methods. For arbitrary choices of φ, we can use a multivariate Newton-Raphson method

to solve the maximum likelihood equations (3) - (4). The second method, which is only valid for this specific choice of φ is the so called data duplication method as described in Lunn and McNeil (1995). The numerical values of the estimators are given in Table 2, together with asymptotic standard errors obtained by inverting the information matrix. The complete asymptotic variance-covariance matrix is given in Table 3 and clearly shows the interaction between the two parts of the model given in (19) and (20). The last two columns in Table 2 are the Wald chisquare statistic and its asymptotic P-value, based on a chisquare distribution with 1 degree of freedom.

Table 2

Coef	Estimate	ASE	Wald chisq	P-value
β_0	0.6630029	0.265151	6.252365	0.01240276
β_1	-1.3862944	0.422577	10.762148	0.00103597
β_2	-0.0350913	0.596583	0.003460	0.95309507

Table 3

	β_0	β_1	β_2
β_0	0.070305052	0.035714286	-0.070197044
β_1	0.035714286	0.17857143	-0.17857143
β_2	-0.070197044	-0.17857143	0.35591133

It is seen that the parameters β_0 and β_1 are significantly different from zero and that this is not the case for the parameter β_2. From (20) it follows that for these data, β_x does not depend on x. Hence the ratio of the conditional probabilities of being informatively censored and being uncensored does not change with the covariate. By integrating out it is also seen that the ratio of marginal probabilities has the same value.

The estimate for β_0 shows a significant effect of the covariate on the survival time till death from melanoma. From Table 2 it follows that the hazard rate for males is 1.94 times the hazard rate for females. The estimates for the cumulative hazard functions for males and females are shown in Figure 1.

The cumulative hazard function for females is also the baseline cumulative hazard. Note that both curves reflect an increase in hazard for larger survival times (which are mostly non-informatively censored).

[Place Figure 1 about here]

We conclude with some further comments on the model. As already said above, if β_1 and β_2 are equal to zero, then the model reduces to a Cox regression model where we treat the uncensored and informatively censored observations as one group versus the non-informatively censored observations. For the present example, the Wald test for the null hypothesis that $\beta_1 = \beta_2 = 0$ results in a value of 22.1546 with an asymptotic P-value of 0.00002. This shows that there is a difference between our model and the Cox regression model with uncensored and informatively censored versus non-informatively censored. An other extreme case of this model is when the estimates for β_1 or β_2 are infinity and no proper fit for β_0 can be obtained. A way out is to interchange the role of informative and non-informative or to consider the classical Cox regression model.

Acknowledgement

The authors acknowledge partial support by the IAP Research Network $n°$ P5/24 of the Belgian State (Federal Office for Scientific Technical and Cultural Affairs).

References

P.K. Anderson, Ø. Borgan, R.D. Gill, N. Keiding, "Statistical Models based on Counting Processes", Springer, New York, 1993.

R. Braekers, N. Veraverbeke, The partial Koziol-Green model with covariates, *J. Statist. Planning Inf.* **92** (2001), 55-71.

D.R. Cox, Regression models and life tables, *J. Roy. Statist. Soc. Ser. B* **34** (1972), 187-220.

D.R. Cox, Partial likelihood, *Biometrika* **62** (1975), 269-276.

J. Fan, I. Gijbels, "Local Polynomial Modelling and its Applications", Chapman and Hall, London, 1996.

U. Gather, J. Pawlitschko, Estimating the survival function under a generalized Koziol-Green model with partially informative censoring, *Metrika* **48** (1998), 189-209.

J.A. Koziol, S.B. Green, A Cramér-von Mises statistic for randomly censored data, *Biometrika* **63** (1976), 465-474.

E.L. Lehmann, "Theory of Point Estimation", Wiley, New York, 1983.

M. Lunn and D. McNeil, Applying Cox regression to competing risks, *Biometrics* **51** (1995), 524-532.

S.A. Murphy, A.W. van der Vaart, On profile likelihood, *J. Amer. Statist. Soc.* **95** (2000), 449-485.

A.A. Tsiatis, A.A. A large sample study of Cox's regression model, *Ann. Statist.* **9** (1981), 93-108.

A.W. van der Vaart, J.A. Wellner, "Weak Convergence and Empirical Processes with Applications to Statistics", Springer, New York, 1996.

A.W. van der Vaart, A.W. "Asymptotic Statistics", Cambridge University Press, 1998.

N. Veraverbeke, C. Cadarso Suárez, Estimation of the conditional distribution in a conditional Koziol-Green model, *Test* **9** 2000, 97-122.

Figure 1: The estimates for the cumulative hazard functions for males and females.

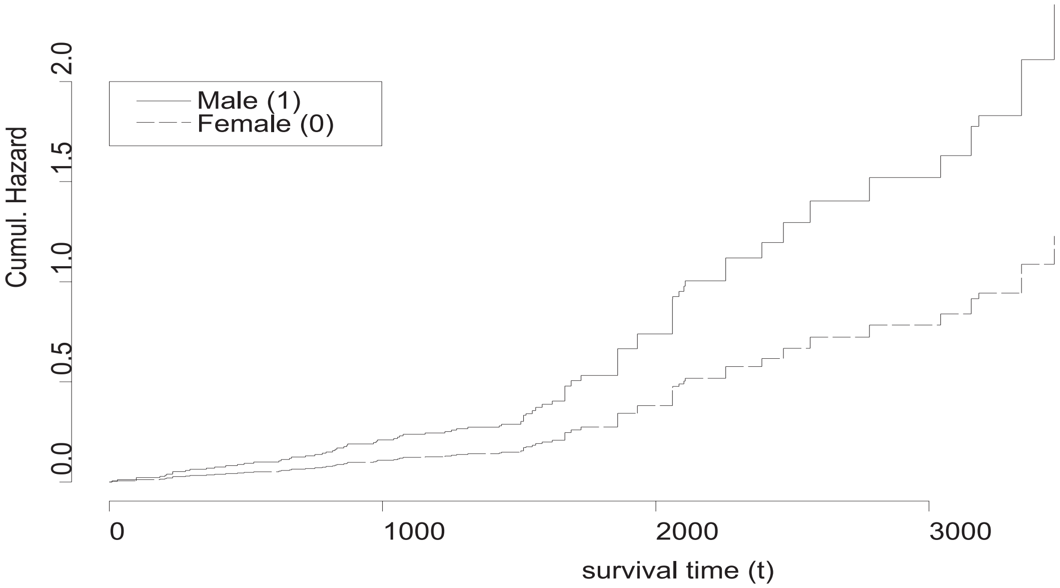

Figure 1: Cumulative hazard function