Made available by Hasselt University Library in https://documentserver.uhasselt.be

Cox's regression model under partially informative censoring

Non Peer-reviewed author version

BRAEKERS, Roel & VERAVERBEKE, Noel (2005) Cox's regression model under partially informative censoring. In: COMMUNICATIONS IN STATISTICS-THEORY AND METHODS, 34(8). p. 1793-1811.

DOI: 10.1081/STA-200066346
Handle: http://hdl.handle.net/1942/2067

COX'S REGRESSION MODEL UNDER PARTIALLY INFORMATIVE CENSORING

Roel BRAEKERS, Noël VERAVERBEKE*

Limburgs Universitair Centrum, Belgium

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 simulated data and to real data examples on survival with malignant melanoma and survival after bone marrow transplantation.

Key Words: Asymptotic normality, Censoring, Consistency, Cox's regression model, Hazard function, Partially informative censoring, Profile likelihood inference

*Correspondence: N. Veraverbeke, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium; E-mail: noel.veraverbeke@luc.ac.be.

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

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

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 known function and $\boldsymbol{\beta}^{(0)} = (\beta_1, \ldots, \beta_p)$ a vector of p unknown parameters. We assume that φ is strictly positive in a neighborhood of $\boldsymbol{\beta}^{(0)}$ and has partial derivatives of first and second order in this neighborhood. 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) The non-informative censoring time D is bounded above by some bound $M < \infty$ and for a prespecified time $T_0 \leq M$ we assume that $P(Z \geq T_0) > 0$.

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

$$\begin{aligned} P(\delta = 1|X = x) &= \int_0^\infty (1 - G_1(t|x))(1 - G_2(t|x))dF(t|x) \\ &= \int_0^\infty (1 - F(t|x))^{\beta_x}(1 - G_2(t|x))dF(t|x) \\ P(\delta = 0|X = x) &= \int_0^\infty (1 - F(t|x))(1 - G_2(t|x))dG_1(t|x) \\ &= \beta_x \int_0^\infty (1 - F(t|x))^{\beta_x}(1 - G_2(t|x))dF(t|x). \end{aligned}$$

Hence 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. We discuss other examples of φ functions in Section 7.

(3) The value of T_0 is typically some prespecified time which in most cases denotes the end of the study. The condition $P(Z \geq T_0) > 0$ is needed in the proof of Lemma 1 below and a possible weakening to $P(Z \geq T_0) = 0$ is discussed on p. 101 of Tsiatis (1981).

In this paper we develop maximum likelihood techniques for joint estimation of the $p+1$ parameters in this model. There are the p parameters $\beta_1, \ldots, \beta_p$ 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. In Section 6, a simulation study is set up to explore the effect of misspecification of β_x. Finally, in Section 7, the method is implemented in the analysis of real data examples.

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

$$\lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P(z_i - \varepsilon \leq Z \leq z_i + \varepsilon, \delta = d_i \mid X = x_i)$$

$$= \begin{cases} f(z_i \mid x_i)(1 - G_1(z_i \mid x_i))(1 - G_2(z_i \mid x_i)) \ldots & \text{if } d_i = 1 \\ g_1(z_i \mid x_i)(1 - F(z_i \mid x_i))(1 - G_2(z_i \mid x_i)) \ldots & \text{if } d_i = 0 \\ g_2(z_i \mid x_i)(1 - F(z_i \mid x_i))(1 - G_1(z_i \mid x_i)) \ldots & \text{if } d_i = -1. \end{cases}$$

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

$$\begin{aligned} &\prod_{\delta_i=1} f(Z_i \mid X_i)(1 - F(Z_i \mid X_i))^{\beta_{X_i}} \prod_{\delta_i=0} \beta_{X_i} f(Z_i \mid X_i)(1 - F(Z_i \mid X_i))^{\beta_{X_i}} \\ &\times \prod_{\delta_i=-1} (1 - F(Z_i \mid X_i))^{\beta_{X_i}+1} \\ &= \prod_{\delta_i=0} \beta_{X_i} \prod_{\delta_i \neq -1} \lambda_0(Z_i) e^{\beta_0 X_i} \prod_{i=1}^{n} (1 - F(Z_i \mid X_i))^{\beta_{X_i}+1} \end{aligned}$$

and, by taking logarithms, we obtain the loglikelihood

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

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

We want to obtain estimators $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ 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 $\beta_1, \ldots, \beta_p$ in the modelling of β_x. Our approach will be a profile likelihood technique. As in the classical Cox model (see Johansen (1983), Anderson et al (1993), Murphy and van der Vaart (2000)), maximization of the full likelihood over arbitrary Λ_0 leads to maximization over $\widehat{\Lambda}_0$ which is a piecewise constant function with jumps at the observed deaths Y_j^0 only. This least informative nonparametric estimator is given by

$$\widehat{\Lambda}_0(t) = \sum_{j=1}^{N} \lambda_j I(Y_j^0 \leq t)$$

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 is

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

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

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

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 $\dfrac{\partial \widehat{H}}{\partial \beta_0} = \ldots = \dfrac{\partial \widehat{H}}{\partial \beta_p} = 0$, that is

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

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

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

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

Consider the 'population version' of (5):

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

and denote the information matrix of the function H as

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

Furthermore, the first order partial derivatives of the function H are zero at $\boldsymbol{\beta}$:

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

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

This can be seen through the following two relations. For any continuous function g we have that:

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

and

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

We only show the derivation of (7) since (8) can be found in a similar fashion. 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 1 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 I is positive definite at $\boldsymbol{\beta}$. 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. The positive definiteness of I at $\boldsymbol{\beta}$ 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 $n \geq n_0$:

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

Since $\widehat{H}$ is continuous and differentiable at $\boldsymbol{\beta}$, we get that $\widehat{H}$ 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 1. 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

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

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

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

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 $\sup_{0\le t\le T_0}\left|E((\varphi(x,\widetilde{\boldsymbol{\beta}}^{(0)})+1)e^{\widetilde{\beta}_0 x},t)\right.$ $\left.-E((\varphi(x,\boldsymbol{\beta}^{(0)})+1)e^{\beta_0 x},t)\right|$ is continuous in $\widetilde{\boldsymbol{\beta}}=(\widetilde{\beta}_0,\widetilde{\boldsymbol{\beta}}^{(0)})$.

This leads to

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

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

which finishes the proof of the lemma.

5 ASYMPTOTIC NORMALITY

Theorem 2. Assume that I is positive definite at $\boldsymbol{\beta}$. Assume that $E|\log\varphi(X,\boldsymbol{\beta}^{(0)})|$, $E(|X|^5(\varphi(X,\boldsymbol{\beta}^{(0)})+1)^2 e^{2\beta_0 X})$, $E(X^2|\dot{\varphi}_j(X,\boldsymbol{\beta}^{(0)})|e^{\beta_0 X})$, $E\left(\frac{\dot{\varphi}_j(X,\boldsymbol{\beta}^{(0)})\dot{\varphi}_{j'}(X,\boldsymbol{\beta}^{(0)})e^{2\beta_0 X}}{\varphi^2(X,\boldsymbol{\beta}^{(0)})}\right)$ and $E\left(\frac{X^2|\ddot{\varphi}_{jj'}(X,\boldsymbol{\beta}^{(0)})|}{\varphi(X,\boldsymbol{\beta}^{(0)})}\right)$ for all $j,j'=1,\ldots,p$ 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 $\mathbf{0} = (0,\ldots,0)$ and I is the information matrix of the function H.

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

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

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 $\dfrac{\partial}{\partial t} \log L(\boldsymbol{\beta}, \Lambda_t)(x, \delta, z)|_{t=0}$, where $\Lambda_t(z) = \int_0^z (1 + th(s))d\Lambda_0(s)$ 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

$$\frac{\partial}{\partial t} \log L(\boldsymbol{\beta}, \Lambda_t)(x, \delta, z)|_{t=0} = I(\delta \neq -1)h(z) - (\varphi(x, \boldsymbol{\beta}^{(0)}) + 1)e^{\beta_0 x} \int_0^z h(s)d\Lambda_0(s) := Ah(z)$$

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 $< f, g >_{\Lambda_0} = \int_0^{T_0} f(s)g(s)d\Lambda_0(s)$ and $< f, g >_{\boldsymbol{\beta}, \Lambda_0} = \int fgdP(x, \delta, z)$.

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 orthogonal 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

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

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

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

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

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

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

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

and I is a positive definite matrix by assumption.

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

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

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 1 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 $\mathbb{R} \times [0, T_0]$ into a partition of bounded, convex sets. By assumption (f), 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 $\mathbb{R} \times [0, T_0] = \bigcup_{j \in \mathbb{Z}}]j - 1/2, j + 1/2] \times [0, T_0]$. 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 $\sum_j M_j P^{1/2}(I_j)$ (or $\sum_j \widetilde{M_j} P(I_j)$) 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.

Remark.

An estimator for the asymptotic variance-covariance matrix can be based on the matrix of minus the second derivatives of the profile loglikelihood.

6 SIMULATION STUDY

In this simulation study we compare the performance of the estimators in the partially informative Cox model. The main interest is on the estimation of the β_0 parameter. If we take β_x equal to 1, we see that the partially informative model reduces to an ordinary Cox model where the uncensored and informatively censored observations are treated as one group versus the group of non-informatively censored observations. In this situation we lose some information of the data because we neglect that uncensored and informatively censored observations have a different distribution. The advantage of this sub-model is that we are able to use standard software for the Cox model to estimate β_0. In this simulation study we find out whether the loss of information is substantial and therefore influences the estimation of β_0. Our simulation scheme is as follows:

$$\begin{aligned} X &\sim U[0,1] \\ Y|X &\sim \text{Weibull}(e^{\beta_0 X}, \tilde{b}) \end{aligned}$$

where β_0, $\tilde{b}$ are constants and $\tilde{b} > 0$. We see that the conditional distribution function is given by $F(t|x) = 1 - \exp(-e^{\beta_0 x} t^{\tilde{b}})$ and the hazard function is $\lambda(t|x) = \tilde{b} t^{\tilde{b}-1} e^{\beta_0 x}$. By the simple proportional hazard assumption we get that $C|X \sim \text{Weibull}(\beta_x e^{\beta_0 x}, \tilde{b})$. In this setting we will work with $\beta_x = e^{a+bx}$. For the non-informative censoring times we assume that

$$D|X \sim \text{Weibull}(e^{\beta_1 X}, \tilde{b})$$

with β_1, $\tilde{b}$ constants and $\tilde{b} > 0$. The probabilities of an uncensored, informatively censored or non-informatively observation are easily calculated:

$$\begin{aligned} P(\delta = 1|X = x) &= e^{\beta_0 x}(e^{a+(b+\beta_0)x} + e^{\beta_0 x} + e^{\beta_1 x})^{-1} \\ P(\delta = 0|X = x) &= e^{a+(b+\beta_0)x}(e^{a+(b+\beta_0)x} + e^{\beta_0 x} + e^{\beta_1 x})^{-1} \\ P(\delta = -1|X = x) &= e^{\beta_1 x}(e^{a+(b+\beta_0)x} + e^{\beta_0 x} + e^{\beta_1 x})^{-1} \end{aligned}$$

We consider 4 different situations expressed by different choices of the parameters β_0, a, b, β_1 and $\tilde{b}$.

[Place Figure 1 about here]

In (a), we take $\beta_0 = 0.5$, $a = b = 0$, $\beta_1 = 1$ and $\tilde{b} = 11$. Consequently β_x is now equal to 1 such that the partially informative Cox model and the reduced Cox model should give the same estimate for β_0. In (b), we take $\beta_0 = 0.5$, $a = -0.5$, $b = -0.4$, $\beta_1 = 1$ and $\tilde{b} = 11$. The probability of a non-informatively censored observation is larger than the other probabilities. Furthermore the informatively censored probability becomes smaller when the covariate x increases. In (c), we reverse the situation of (b). By choosing $\beta_0 = 0.5$, $a = 0.5$, $b = 1$, $\beta_1 = 1$ and $\tilde{b} = 11$, we get that the probability of an informatively censored observation is larger than the other probabilities. The probability of an uncensored observation is the smallest probability here. In (d), we take $\beta_0 = 0.5$, $a = -0.5$, $b = -1$, $\beta_1 = 0.45$ and $\tilde{b} = 11$. The probabilities of an uncensored observation or non-informatively censored observation are now almost equal and the probability of an informatively censored observation is smaller than the other two.

For each situation, we take 5000 samples of two different sample sizes (50 and 150). By comparing the results for these sample sizes, we get an idea about whether the sample size influences the loss of information between both models. In the following tables we show the results of the simulation study. For each situation, we give the average estimate for the parameters a, b and β_0 in both models. Furthermore we also show between brackets, the standard deviation of these estimates by calculating the standard deviation of the 5000 estimates.

Table 1 : Simulation results for $n = 50$

		(a)	(b)	(c)	(d)
	β_0	0.5124 (0.0143)	0.5118 (0.0133)	0.4328 (0.0194)	0.5349 (0.0119)
partial	a	-0.0231 (0.0114)	-0.5703 (0.0139)	0.5196 (0.0116)	-0.5469 (0.0143)
	b	0.0317 (0.0201)	-0.4489 (0.0277)	1.124 (0.0216)	-1.1834 (0.0279)
reduced	β_0	0.5316 (0.0099)	0.3815 (0.0107)	1.2637 (0.0093)	0.2408 (0.0100)

Table 2 : Simulation results for $n = 150$

		(a)	(b)	(c)	(d)
	β_0	0.4988 (0.0076)	0.5018 (0.0071)	0.4738 (0.0096)	0.5132 (0.0064)
partial	a	0.0044 (0.0060)	-0.5222 (0.0067)	0.5038 (0.006)	-0.5111 (0.0067)
	b	-0.0091 (0.0106)	-0.4041 (0.0122)	1.0398 (0.0111)	-1.0306 (0.0127)
reduced	β_0	0.4958 (0.0054)	0.3705 (0.0058)	1.2358 (0.0050)	0.2375 (0.0054)

Except for situation (a), we note that the estimate of β_0 in the reduced Cox model is largely different from the estimate of β_0 in the partially informative Cox model. As we noted before, we treat in the reduced Cox model the uncensored and the informatively censored observations as one group and neglect that the distribution of these types of observations are different. These results show that when we estimate the parameter β_0 in a partially informative Cox model by the estimator of the reduced Cox model, the estimate can be very different of what we would expect because we do not use all the information in the data. This simulation study also shows that a misspecification of β_x can have serious consequences for the estimate of β_0. In this example we note that the reduced Cox model can also be seen as a partial informative Cox model with $\beta_x = 1$. Comparing the results of Table 1 with those of Table 2 we see that the estimates for the reduced Cox model and the partial informative Cox model are not much different in each situation. Therefore we note that the sample size has no influence on the difference in the estimate of β_0 in the reduced Cox model and in the partial informative Cox model.

The estimates in Table 1 and 2 are obtained via the CoxPh software in Splus. For the reduced Cox model we treated the uncensored and the informatively censored observations as uncensored in the calculations. For the partially informative Cox model, we used the duplication method for competing risk models of Lunn and McNeil (1995). Via this data augmentation method we are able to deal with the three types of observations in the calculations.

7 EXAMPLES WITH REAL DATA

7.1 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 because 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. 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 mainly 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. Due to the limitations of this software, we cannot use it for other choices of φ, like for example: $\varphi(x, \beta_1, \beta_2) = 1+(\beta_1+\beta_2 x)$, $\varphi(x, \beta_1) = \dfrac{e^{-\beta_1 x}}{1+e^{\beta_1 x}}$, $\varphi(x, \beta_1) = \dfrac{e^{\beta_1 x}}{1+e^{\beta_1 x}}$, $\varphi(x, \beta_1, \beta_2, \beta_3) = [1 + \beta_3(\beta_1 + \beta_2 x)]^{1/\beta_3}$, $\varphi(x, \beta_1, \beta_2, \beta_3) = e^{\beta_1 - \beta_2 e^{-\beta_3 x}}$, ... In a practical situation it is not always clear which is the best choice of φ. This will for example depend on how the parameters of φ will be interpreted in this situation. We use in this example several choices of φ for illustrating purpose and believe that the optimal choice of φ in a practical situation is a subject for future research.

To estimate jointly the parameters in (19) and (20), we are able to use two different methods. For arbitrary choices of φ, we can use a numerical Newton-type optimizer to find estimates for the parameters. The second method, which is only valid for $\varphi(x, \beta_1, \beta_2) = e^{\beta_1+\beta_2 x}$ is the data duplication method as described in Lunn and McNeil (1995). The numerical values of the estimators for this choice φ are given in Table 3, together with asymptotic standard errors obtained by inverting the information matrix. The complete asymptotic variance-covariance matrix is given in Table 4 and clearly shows the interaction between the two parts of the model given in (19) and (20). The last two columns in

Table 3 are the Wald chisquare statistic and its asymptotic P-value, based on a chisquare distribution with 1 degree of freedom.

Table 3

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 4

	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 3 it follows that the hazard rate for males is 1.94 times the hazard rate for females.

We conclude with some further comments on the model. As already said above, if β_1 and β_2 are equal to zero, this 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 significant 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. Another remark on the partially informative Cox model with $\varphi(x;\beta_1,\beta_2) = e^{\beta_1+\beta_2 x}$, is that we can rewrite it as the following competing risk model with cause-specific hazards

$$\begin{aligned} \lambda_{\text{Melanoma}}(t|x) &= \lambda_0(t)e^{\beta_0 x} \\ \lambda_{\text{Other}}(t|x) &= \lambda_0(t)\beta_0 e^{\beta_0 x} = \lambda_0(t)e^{a+(b+\beta_0)x} \end{aligned}$$

and an independent right censoring process. A slightly different model was considered for this data set on p. 493 of Anderson *et al.* (1993) with hazards $\lambda_{\text{Melanoma}}(t|x) = \lambda_0(t)e^{\beta_0 x}$ and $\lambda_{\text{Other}}(t|x) = \lambda_0(t)e^{\beta_1 x}$. However they found the same value for β_0 as we did.
We finish this section by showing some results for other choices of φ. We took examples of φ functions for which the estimate of the variance-covariance matrix is positive definite (for this particular data set). In Table 5, we give for different φ the estimates for the parameters and their standard errors.

Table 5

	Model 1	Model 2	Model 3
	$\varphi(x,\beta_1,\beta_2,\beta_3) = [1+\beta_3(\beta_1+\beta_2 x)]^{1/\beta_3}$	$\varphi(x,\beta_1,\beta_2) = 1+(\beta_1+\beta_2 x)$	$\varphi(x,\beta_1) = \frac{e^{\beta_1 x}}{1+e^{\beta_1 x}}$
β_0	0.6628 (0.2652)	0.6628 (0.2652)	0.8451 (0.2513)
β_1	-1.1957 (10.0999)	-0.7500 (0.1057)	-1.1451 (0.5551)
β_2	-0.0258 (0.6389)	-0.0086 (0.1467)	
β_3	0.2189 (12.8271)		

We see that the estimates of β_0 in the first two models are identical and very close to the estimate in Table 3. An explanation for this is that these models are related to the model in Table 3. $\varphi(x,\beta_1,\beta_2) = e^{\beta_1+\beta_2 x}$ is the limit of β_3 to zero for the first model and the first order approximation of this φ gives the second model. We also note that for a totally different choice of φ as in model 3, we get a different estimate for β_0.

7.2 BONE MARROW TRANSPLANTATION

As second real data set we take the bone marrow transplantation data of Klein and Moeschberger (1997). In this data set, 137 patients with acute leukemia were followed after their leukemia was treated by a bone marrow transplantation. For each patient the disease free survival time and the reason why the patient left the study, were recorded. As an example we want to show how the time until relapse is influenced by the age of the patient. In this setting we take the time until death as informative censoring time and patients still alive at the end of the study are considered as non-informatively censored. We give the results of the partially informative Cox model where we used several different choices of φ in Table 6. These choices are selected such that the estimate for the variance-covariance matrix is positive definite.

Table 6

	β_0	β_1	β_2
Model 1 $\varphi(x,\beta_1,\beta_2) = e^{\beta_1+\beta_2 x}$	0.0085 (0.0165)	-1.1814 (0.7359)	0.0054 (0.0234)
Model 2 $\varphi(x,\beta_1,\beta_2) = 1+(\beta_1+\beta_2 x)$	0.0096 (0.0148)	-0.1172 (0.5518)	0.0032 (0.0182)
Model 3 $\varphi(x,\beta_1) = \frac{e^{\beta_1 x}}{1+e^{\beta_1 x}}$	0.0112 (0.0117)	0.7989 (13.4556)	
Model 4 $\varphi(x,\beta_1,\beta_2) = e^{(\frac{\beta_1 x e^{-\beta_1 x}}{\beta_2})}$	0.0113 (0.0125)	-0.0002 (0.0074)	1.000 (42.6874)

As in the previous example, we see that in models with related choices of φ such as model 1 and 2, the estimates of β_0 are almost the same. However when these choices have different properties, for example model 1 versus model 3, the estimates are also different. In each model we note that the estimate for β_0 is not significant and so the age of the patient does not influence the time until relapse of acute leukemia.

ACKNOWLEDGEMENT

The authors acknowledge partial support by the IAP Research Network n° P5/24 of the Belgian Government (Belgian Science Policy).

REFERENCES

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

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

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

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

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

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

Johansen, S. An extension of Cox's regression model. *International Statistical Review* **1983**, 51, 165-174.

Klein, J.P.; Moeschberger, M.L. *Survival analysis: Techniques for censored and truncated data*; Springer-Verlag: New York, 1997.

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

Lehmann, E.L. *Theory of Point Estimation*; Wiley: New York, 1983.

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

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

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

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

van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge,1998.

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

Figure 1: The probabilities of an uncensored, informatively censored and non-informatively censored observation.

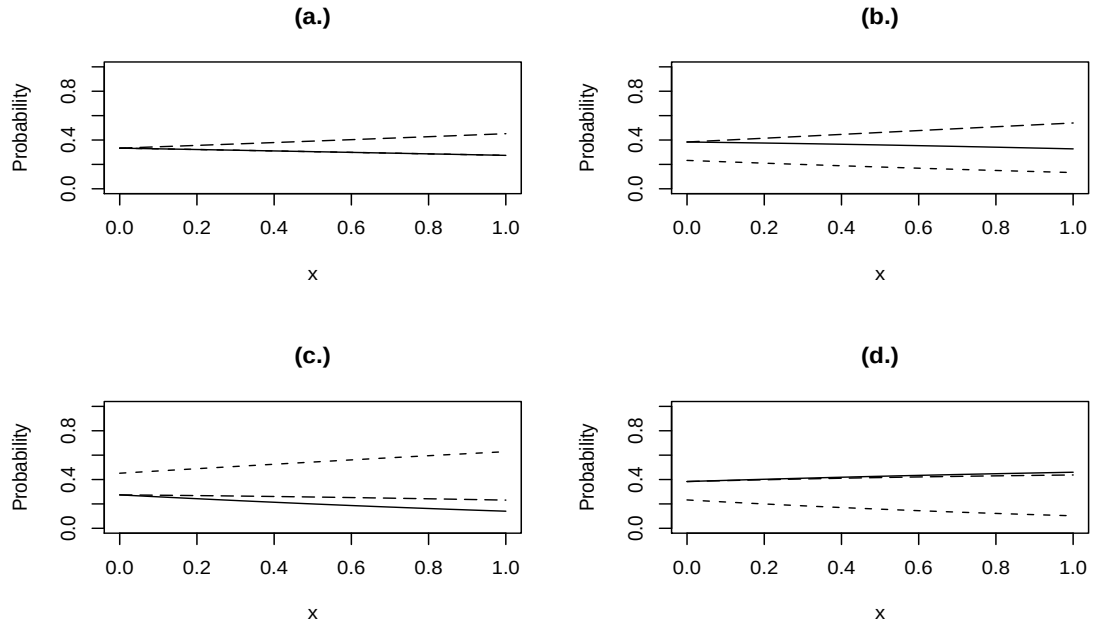

Figure 1 : The probabilities of an uncensored observation $P(\delta = 1|X = x)$ (solid line), informatively censored observation $P(\delta = 0|X = x)$ (dashed line) and non-informatively censored observation $P(\delta = -1|X = x)$ (longdashed line) in 4 different situations. In (a), $\beta_0 = 0.5$, $a = b = 0$, $\beta_1 = 1$ and $\tilde{b} = 11$. In (b), $\beta_0 = 0.5$, $a = -0.5$, $b = -0.4$, $\beta_1 = 1$ and $\tilde{b} = 11$. In (c), $\beta_0 = 0.5$, $a = 0.5$, $b = 1$, $\beta_1 = 1$ and $\tilde{b} = 11$. In (d), $\beta_0 = 0.5$, $a = -0.5$, $b = -1$, $\beta_1 = 0.45$ and $\tilde{b} = 11$.