Lifetime Data Anal (2016) 22:570–588
DOI 10.1007/s10985-015-9351-y

Cox regression with missing covariate data using a modified partial likelihood method

Torben Martinussen[1] · **Klaus K. Holst**[1] · **Thomas H. Scheike**[1]

Received: 18 March 2015 / Accepted: 5 October 2015 / Published online: 22 October 2015

Abstract Missing covariate values is a common problem in survival analysis. In this paper we propose a novel method for the Cox regression model that is close to maximum likelihood but avoids the use of the EM-algorithm. It exploits that the observed hazard function is multiplicative in the baseline hazard function with the idea being to profile out this function before carrying out the estimation of the parameter of interest. In this step one uses a Breslow type estimator to estimate the cumulative baseline hazard function. We focus on the situation where the observed covariates are categorical which allows us to calculate estimators without having to assume anything about the distribution of the covariates. We show that the proposed estimator is consistent and asymptotically normal, and derive a consistent estimator of the variance–covariance matrix that does not involve any choice of a perturbation parameter. Moderate sample size performance of the estimators is investigated via simulation and by application to a real data example.

Keywords Cox model · Missing covariate data · Recursive estimation · Survival data

✉ Torben Martinussen
tma@sund.ku.dk

Klaus K. Holst
kkho@sund.ku.dk

Thomas H. Scheike
ts@sund.ku.dk

[1] Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5B, 1014 Copenhagen K, Denmark

1 Introduction

The possible effect of prognostic factors X on a censored time-to-event outcome T is very often modelled using the Cox-model (Cox 1972), specified by the conditional hazard function

$$\lambda(t|X = x) = \lambda_0(t) \exp\left(\beta^T x\right), \tag{1}$$

where β denotes the regression parameters of interest, and $\lambda_0(t)$ is the baseline hazard function that is not further specified. Let $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$ denote the integrated baseline hazard function. We will work under right-censoring so we only observe $\tilde{T} = \min(T, U)$ and $D = (T \leq U)$, where U denotes the potential censoring time. Also, the covariates are assumed to be constant with time. Statistical inference for this model is well known (Andersen and Gill 1982). However, in practice, it is very common that some prognostic factors are unobserved for some subjects. Interestingly, the complete case (CC) estimator that uses only the subjects with complete covariate values gives unbiased estimates if the missing data mechanism does not depend on the outcome. However, it loses efficiency since it discards incomplete observations. Also, if the missing data mechanism does depend on the follow-up time, then the CC-estimator will be biased. Therefore, several other approaches have been suggested to tackle both the efficiency and the bias problem of the CC-estimator. These approaches basically split up in two directions. One that is likelihood based such as Chen and Little (1999), Martinussen (1999) that use a full likelihood approach, see also Herring and Ibrahim (2001). These methods need a specification of the covariate distribution, however. This is alleviated to some extent by the double-semiparametric likelihood method of Chen (2002). All these methods use the EM-algorithm that may have slow convergence and the variance-covariance estimator of Chen (2002) also involves the choice of a perturbation parameter. Related to the full likelihood approach is multiple imputation that seems to be widely used in practice being available in many standard software packages such as R and SPSS. It is well known for this to work properly that one needs to include the outcome variable in the imputation model, see Sterne et al. (2009) for further comments on this, and it is not clear when dealing with a (possibly censored) follow-up time what to use as the outcome variable. This was investigated under the Cox model by White and Royston (2009) and they found that one should probably use $(\Lambda_0(\tilde{T}), D)$ as the "outcome-variable" rather than for instance $(\tilde{T}, D)$. However, this result could only be established in some very simple settings. Failing to include the outcome in the imputation model will bias the outcome-covariate association towards the null when using the imputed data (Sterne et al. 2009). Another direction has been to use inverse probability weighted (IPW) estimators. IPW estimators can correct bias from the CC-estimator in the case where the missing data mechanism depends on the follow-up time and/or the failure status. Such an estimator was suggested by Pugh et al. (1994) requiring correct specification of the missing data mechanism. This demand is alleviated using the augmented inverse probability weighting (AIPW) concept of Robins et al. (1994) and was applied for the Cox-model by Wang and Chen (2001). These estimators are unbiased as long as either the conditional distribution of the missing covariate given observed data or the missing data mechanism is modelled correctly and the AIPW estimator thus possess a double robustness

property. When selection probabilities are small for some individuals, the variance of the IPW estimators may be inflated because of heavy weights. This led Xu et al. (2009) to consider re-weighted estimators. Simple IPW-estimators may not always have good efficiency as demonstrated by Qi et al. (2005), which led these authors to suggest non-parametric estimation of the selection probabilities to gain efficiency. This, however, requires some kind of smoothing involving the choice of a bandwidth parameter. This kernel assisted estimator is as efficient as the fully augmented one, see Qi et al. (2005).

In this paper we will focus on the situation where the selection probabilities only depend on the always observed covariates and thus does not depend on the follow-up time. In this case it is possible to construct an estimation method that is close to the maximum likelihood estimator thus having high efficiency and at the same time avoiding the use of the EM-algorithm. It further gives rise to a direct estimator of the variance-covariance matrix and it is thus not necessary to choose a perturbation parameter as in Chen (2002). Instead of maximising the full likelihood directly, we try to profile out the non-parametric cumulative baseline hazard function exploiting that the observed hazard function is still multiplicative in the baseline hazard function. This approach has been suggested before in other settings such as when doing estimation within the transformation model, see Bagdonavicius and Nikulin (1999) and Martinussen and Scheike (2006, Chap. 8), and has been termed a modified partial likelihood method. This gives rise to an estimating function for the regression parameters. This depends, however, on the cumulative baseline hazard function but it can be estimated recursively leaving us with an unbiased estimating function for the regression parameters. The practical calculations involve calculating (ratios) of certain conditional means that could be carried out upon imposing structure on the joint distribution of the prognostic factors. However, when all observed covariates are categorical, as we will assume, this can be carried out without having to model the joint distribution of the covariates. We show that the proposed estimator is consistent and asymptotically normal, and provide also a consistent estimator of the variance-covariance matrix.

The rest of the paper is organised as follows. In Sect. 2, we derive the suggested estimator, and give the large sample results in Sect. 3. Moderate sample size performance of the estimators is examined via simulations in Sect. 4, and we apply the methods to data on Kidney cancer data in Sect. 5. We conclude in Sect. 6. Technical results are deferred to the Appendix.

2 Methods

We focus on the situation where there is a set of prognostic factors X (p-dimensional) so that some X's may be missing for some individuals. We will impose the following restrictions (and discuss later how to alleviate these): we can write $X = (X^c, X^d, W)$ so that W is always observed and is assumed to be discrete. The part (X^c, X^d) consist of continuous and discrete covariates that may be missing but we will not allow that X^d is missing if X^c (or parts of it) is observed. We allow, however, X^c to be empty, that is, we may only have categorical covariates. We use P_g, $g = 1, \ldots, G$ to denote the different missing data patterns with P_1 corresponding to no missing data. Let R^g

be the indicator of pattern g and X^g and Z^g be the missing and observed covariates under pattern g, respectively. We let β_g be the entries of vector β corresponding to Z^g and similarly let β_{g^c} be the entries corresponding to X^g. As an example, consider the situation with three covariates, $X = (X_1, X_2, X_3)$, such that X_1 is continuous, and X_2 and X_3 are categorical. In this example we assume that X_3 is always observed, so that $W = X_3$, but there may be missing values in X_1 and X_2. The missing data patterns that are allowed, with "·" denoting missing information, are P_1: (X_1, X_2, X_3) (all covariates observed); P_2: $(\cdot, X_2, X_3)$ and P_3: $(\cdot, \cdot, X_3)$. We assume that data are missing at random so that the missing data pattern only depends on the part of X that is always observed, that is $P(R^g = 1|\text{obs data}) = P(R^g = 1|W)$. Assume that the conditional distribution of T given X is governed by the Cox model, that is, model (1) is assumed to hold. Since T may be censored by U we only observe the minimum of the two $\tilde{T} = \min(T, U)$ and the event indicator $D = I(T \leq U)$. We assume that T and U are independent given X, and that the censoring distribution only depends on the always observed covariates W. Based on the assumed missing data generating mechanism it is easy to see that

$$P(T > t|R^g = 1, Z^g = z) = P(T > t|Z^g = z).$$

The observed hazard function, for a subject under missing data pattern, P_g, can then be found by calculating the derivative of $-\log\{P(T > t|Z^g = z)\}$ and is therefore given by

$$\lambda_g(t) = \lambda(t|Z^g = z) = \lambda_0(t)\phi_g\{\Lambda_0(t), \beta, z\}\exp\left(\beta_g^T z\right), \tag{2}$$

where

$$\phi_g\{\Lambda_0(t), \beta, z\} = \frac{E\left[\exp\left(\beta_{g^c}^T X^g\right)\exp\left\{-\Lambda_0(t)\exp\left(\beta^T X\right)\right\}|Z^g = z\right]}{E\left[\exp\left\{-\Lambda_0(t)\exp\left(\beta^T X\right)\right\}|Z^g = z\right]}. \tag{3}$$

The key observation is that $\lambda_g(t)$ is still multiplicative in $\lambda_0(t)$, and the idea is to profile this function out. For the missing data pattern P_1 we have all covariates observed, and then $Z^1 = X$ and the corresponding ϕ_1 is simply 1. We let $\phi = (\phi_1, \ldots, \phi_G)$.
We assume that we have n iid replicates from this model. Denote the ith counting process by $N_i(t) = I(\tilde{T}_i \leq t, D_i = 1)$, and assume that we observe in the time interval $[0, \tau]$ with $\tau < \infty$. The increment of the compensator for this counting process is then given by

$$\sum_{g=1}^{G} Y_i(t)R_i^g\phi_g\{\Lambda_0(t), \beta, Z_i^g\}\exp\left(\beta_g^T Z_i^g\right)d\Lambda_0(t)$$

where $Y_i(t)$ is the at risk indicator, and R_i^g is one if missing data pattern P_g is seen for individual i, and zero otherwise.

Counting process theory (see Martinussen and Scheike 2006, Chap. 3) suggest the following estimating function

$$\tilde{U}(\beta, \phi, \Lambda_0) = \sum_i \int_0^\tau \left[\left\{ \frac{d\lambda^i(t)}{d\beta} \right\}^T / \lambda^i(t) \right] \{dN_i(t) - Y_i(t)\lambda^i(t)dt\},$$

where $\lambda^i(t)$ denotes the ith observed hazard function. Using the specific expression for the observed hazard function leads to the following estimating function for β:

$$\tilde{U}(\beta, \phi, \Lambda_0) = \sum_{i=1}^n \int_0^\tau \left\{ \tilde{X}_i(t) - \frac{S_1(t, \beta)}{S_0(t, \beta)} \right\} dN_i(t), \tag{4}$$

where

$$\tilde{X}_i(t) = \sum_g \left\{ R_i^g \frac{D_\beta \phi_g\{\Lambda_0(t), \beta, Z_i^g\}}{\phi_g\{\Lambda_0(t), \beta, Z_i^g\}} + R_i^g Z_i^g \right\}$$

$$S_0(t, \beta, \Lambda_0, \phi) = \sum_{i=1}^n \sum_{g=1}^G Y_i(t) R_i^g \phi_g\{\Lambda_0(t), \beta, Z_i^g\} \exp\left(\beta_g^T Z_i^g\right)$$

and $S_1(t, \beta, \Lambda_0, \phi) = D_\beta S_0(t, \beta, \Lambda_0, \phi)$. Here D_β means the derivative with respect to β. In the definition of $\tilde{X}_i(t)$, if $R_i^g = 1$, then the $R_i^g Z_i^g$ should be understood as the p-vector with the observed covariate values at the relevant places, and zero entries corresponding to the covariates that are missing for that individual. We can not use $\tilde{U}(\beta, \phi, \Lambda_0)$ directly for estimation of β as ϕ and Λ_0 are unknown. As is seen from display (3), ϕ is not only depending on β and Λ_0 but is also a functional of the covariate distribution. Below we outline how to estimate it without any assumptions about the covariate distribution. Let R_i be equal to R_i^g with $g = 1$ (all covariates observed). If $\Lambda_0(t)$ was known then we suggest estimating $\phi_g\{\Lambda_0(t), \beta, z^*\}$ non-parametrically by

$$\hat{\phi}_g\{\Lambda_0(t), \beta, z^*\} = \frac{\sum_{i=1}^n R_i \exp\left(\beta_{g^c}^T X_i^g\right) \exp\left\{-\Lambda_0(t) \exp\left(\beta^T X_i\right)\right\} I\left(Z_i^g = z^*\right)}{\sum_{i=1}^n R_i \exp\left\{-\Lambda_0(t) \exp\left(\beta^T X_i\right)\right\} I\left(Z_i^g = z^*\right)},$$

if $\sum_i R_i I(Z_i^g = z^*) > 0$, and set it to 1 otherwise. We estimate the needed derivatives in the estimating function in a similar way. We shall require that $P(R = 1|W = w^*) > 0$ and $P(Z^g = z^*) > 0$ for all z^*. This will ensure that $P\{RI(Z^g = z^*) = 1\} > 0$.

Exploiting that (2) is multiplicative in the $\lambda_0(t)$ we estimate the increments of $\Lambda_0(t)$ by the following Breslow type estimator

$$d\hat{\Lambda}_0(t, \beta) = \frac{dN_\cdot(t)}{S_0(t, \beta, \hat{\Lambda}_0(t-, \beta), \hat{\phi}_{\hat{\Lambda}_0})}, \tag{5}$$

where $N_\cdot(t) = \sum_i N_i(t)$ and $\hat{\phi}_{\hat{\Lambda}_0}$ is short for $\hat{\phi}$ evaluated at $\hat{\Lambda}_0(t-)$. Note that (5) needs only to be calculated at the jump times of $N_\cdot$, and for two such consecutive jumping times, τ_j and τ_{j+1}, we can calculate (5) at $t = \tau_{j+1}$ if $\hat{\Lambda}_0(\tau_j, \beta)$ is known.

The calculation is started by putting $\hat{\Lambda}_0(0, \beta) = 0$. Finally, we estimate β by $\hat{\beta}$ that is the solution to $U(\beta) = 0$, where

$$U(\beta) = \tilde{U}\big\{\beta, \hat{\phi}, \hat{\Lambda}_0(\cdot, \beta)\big\}.$$

Finally, we estimate $\Lambda_0(t)$ by

$$\hat{\Lambda}_0(t, \hat{\beta}) = \int_0^t \frac{1}{S_0\big(s, \hat{\beta}, \hat{\Lambda}_0\big(s-, \hat{\beta}\big), \hat{\phi}_{\hat{\Lambda}_0}\big)} dN_\cdot(s).$$

3 Large sample properties

In the Appendix we show that the suggested estimators are consistent. Further, we show that

$$n^{-1/2} U(\beta_0) = n^{-1/2} \sum_{i=1}^{n} \epsilon_i^U + o_p(1),$$

where ϵ_i^U are zero-mean iid terms. Hence, $n^{-1/2} U(\beta_0)$ converges in distribution to a normal variate with variance Σ_U that can be estimated consistently by

$$\hat{\Sigma}_U = n^{-1} \sum_{i=1}^{n} \epsilon_i^U \big(\epsilon_i^U\big)^T.$$

It therefore follows that $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to a normal variate with a variance Σ_β that is consistently estimated by

$$\hat{\Sigma}_\beta = (n^{-1} D_\beta U)^{-1} \hat{\Sigma}_U (n^{-1} D_\beta U)^{-1},$$

where $D_\beta U$ denotes the derivative of U with respect to β, and evaluated at $\hat{\beta}$. The analytical expression for $D_\beta U$ is complicated but it is easily calculated numerically using complex step derivatives, see for instance Sherman (2006).

We also show that $n^{1/2}\{\hat{\Lambda}_0(t, \beta_0) - \Lambda_0(t)\}$ can be written essentially as a sum of independent and identically distributed mean-zero random variables. This result, along with the expansion

$$\begin{aligned} n^{1/2}\big\{\hat{\Lambda}_0\big(t, \hat{\beta}\big) - \Lambda_0(t)\big\} &= n^{1/2}\big\{\hat{\Lambda}_0(t, \beta_0) - \Lambda_0(t)\big\} + n^{1/2}\big\{\hat{\Lambda}_0\big(t, \hat{\beta}\big) - \hat{\Lambda}_0(t, \beta_0)\big\} \\ &= n^{1/2}\big\{\hat{\Lambda}_0(t, \beta_0) - \Lambda_0(t)\big\} \\ &\quad + D_\beta\big(\hat{\Lambda}_0(t, \beta)\big)_{|\hat{\beta}} n^{1/2}\big(\hat{\beta} - \beta_0\big) + o_p(1), \end{aligned}$$

allows us to write $n^{1/2}\{\hat{\Lambda}_0(t, \hat{\beta}) - \Lambda_0(t)\}$ in terms of a sum of independent and identically distributed mean-zero random variables: $n^{1/2}\{\hat{\Lambda}_0(t, \hat{\beta}) - \Lambda_0(t)\} =$

$n^{-1/2}\sum_{i=1}^{n}\epsilon_i^{\Lambda_0}(t,\beta_0)+o_p(1)$, It thus follows that $n^{1/2}\{\hat{\Lambda}_0(t,\hat{\beta})-\Lambda_0(t)\}$ converges to a zero-mean Gaussian process with a variance that is consistently estimated by $n^{-1}\sum_k \hat{\epsilon}_i^{\Lambda_0}(t,\hat{\beta})\hat{\epsilon}_i^{\Lambda_0}(t,\hat{\beta})^T$.

4 Simulation studies

In this section we investigate the moderate sample performance of the suggested estimator, and compare it to the simple IPW estimator (that uses the correct selection probability model) and to the kernel assisted IPW estimator suggested by Qi et al. (2005) that they found to be as efficient as the fully augmented weighted estimator. We use simulation scenario 1 of Chen (2002) so that we can also compare to the double-semiparametric estimator suggested by Chen (2002). We hence consider the situation where W is Bernoulli with success probability 0.5 and the continuous covariate X^c is uniformly on [0, 1] or normally distributed $N(0,1)$. The survival time follows a Cox regression model with baseline hazard rate of 1, and (β_1,β_2) are either set to $(0,0)$ or to $(1,1)$. Censoring was induced using an exponential censoring distribution. The parameter of the censoring distribution was adjusted to obtain censoring rates of 30 and 70 %. Two missing-data mechanism were used to obtain 50 % missingness in X^c. The first mechanism randomly deleted the continuous covariate X^c of one half of the subjects (MCAR) and the other mechanism deleted X^c according to the probability $1/(1+e^{-0.92+1.85W})$ (MAR) also resulting in 50 % missingness. The simple IPW-estimator uses the weights based on the inverse of the estimates of $\pi(w)=P(R_i=1|W_i=w)$, which is just $\hat{\pi}(w)=n^{-1}\sum_j R_j I(W_j=w)$, $w=0,1$. For the kernel assisted IPW estimator we let $\pi()$ depend on $(\tilde{T},D,W)$ as $\pi_{kl}(\tilde{T})$ if $D=k$, $W=l$, and here we used a Nadaraya–Watson estimator applying the R-function `ksmooth` with a normal kernel and with bandwidth $h=6n_{kl}^{-1/3}$, where $n_{kl}=\sum_i(D_i=k,W_i=l)$. The results for the situation with X^c being uniformly distributed on [0, 1] are given in Table 1. It is seen that all methods are unbiased as expected. We also see that the proposed method has the same efficiency as the complete case estimator for the covariate that has missing values but that it is much more efficient with regard to the covariate that has complete information. The mean-squared error (MSE) of the proposed estimator on that component compared to the MSE for the full data estimator varies from 1.01 to 1.20 so in many of the considered scenarios the proposed estimator is almost as efficient as the full data estimator with regard to the covariate with complete information.

These results compares well to the results of the double-semiparametric estimator reported in Chen (2002). The simple IPW estimator that uses the correct missing data model is consistently less efficient than the complete case estimator except under the MCAR where the estimated weights tend to be close to 0.5 for all individuals and hence cancels out, and therefore being very close to the complete case estimator. But when the missing data are generated under the MAR, the simple IPW estimator is less efficient than the CC estimator. For instance, under MAR with $n=400$ and censoring percent equal to 30, we see that the MSE for the CC and simple IPW estimator with respect to the full data estimator for β_1 are 2.03 and 2.51, respectively. The kernel assisted IPW estimator does a better job. In the MCAR situation it is as efficient as the

Table 1 Continuous covariate distributed as U(0,1)

n	Censoring	Method	MCAR				MAR			
			β_1		β_2		β_1		β_2	
			Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
True value of $(\beta_1, \beta_2) = (0, 0)$										
200	30 %	Full	−0.00	0.09	0.00	0.03	−0.00	0.09	−0.00	0.03
		CC	0.00	2.15	−0.00	2.20	−0.00	2.24	−0.00	2.68
		PP	0.00	2.13	0.00	1.03	−0.00	2.23	−0.00	1.04
		IPW1	0.00	2.17	−0.00	2.20	0.00	2.78	0.00	2.71
		IPW2	−0.00	2.13	0.00	1.32	0.00	2.61	0.00	1.66
	70 %	Full	−0.00	0.21	−0.00	0.07	0.01	0.22	−0.00	0.08
		CC	−0.01	2.31	0.00	2.20	0.02	2.24	−0.03	4.96
		PP	−0.01	2.25	−0.00	1.03	0.03	2.16	−0.00	1.03
		IPW1	−0.01	2.31	0.00	2.20	0.02	2.84	−0.03	5.42
		IPW2	−0.01	2.30	0.00	1.30	0.02	2.92	0.01	4.07
400	30 %	Full	−0.01	0.04	0.00	0.02	−0.00	0.04	0.00	0.02
		CC	−0.00	2.05	0.00	2.02	−0.00	2.03	0.01	2.51
		PP	−0.00	2.04	0.00	1.01	−0.00	1.99	0.00	1.01
		IPW1	−0.00	2.06	0.00	2.02	−0.00	2.51	0.01	2.53
		IPW2	−0.00	2.00	0.01	1.20	−0.00	2.39	0.05	1.55
	70 %	Full	0.01	0.11	0.00	0.03	0.01	0.01	0.00	0.03
		CC	−0.00	2.11	−0.01	2.20	0.00	2.15	−0.00	2.75
		PP	−0.00	2.07	−0.01	1.02	0.00	2.10	0.00	1.01
		IPW1	−0.00	2.11	−0.01	2.20	0.01	2.61	−0.00	2.76
		IPW2	−0.01	2.10	−0.01	1.26	0.01	2.62	0.02	1.37
True value of $(\beta_1, \beta_2) = (1, 1)$										
200	30 %	Full	0.00	0.10	0.00	0.03	0.01	0.10	0.02	0.04
		CC	0.02	2.16	0.02	2.17	0.02	2.18	0.03	2.49
		PP	0.03	2.19	0.01	1.12	0.03	2.23	0.02	1.20
		IPW1	0.02	2.17	0.02	2.17	0.02	2.70	0.04	2.59
		IPW2	0.01	2.14	0.01	1.66	0.02	2.73	0.05	2.11
	70 %	Full	0.03	0.22	0.02	0.08	0.03	0.22	0.02	0.09
		CC	0.03	2.29	0.04	2.37	0.07	2.80	0.03	2.23
		PP	0.03	2.28	0.02	1.07	0.06	2.90	0.03	1.14
		IPW1	0.03	2.31	0.04	2.37	0.07	3.49	0.03	2.29
		IPW2	0.03	2.31	0.03	1.49	0.07	3.51	0.03	1.60
400	30 %	Full	0.01	0.05	0.01	0.02	0.00	0.05	0.01	0.02
		CC	0.02	2.01	0.01	2.02	0.00	2.32	0.02	2.37
		PP	0.02	1.97	0.01	1.10	0.01	2.20	0.01	1.16
		IPW1	0.01	2.02	0.01	2.02	0.00	2.76	0.02	2.44
		IPW2	0.01	2.00	0.01	1.49	−0.00	2.68	0.01	1.90
	70 %	Full	0.01	0.11	0.01	0.04	0.01	0.11	0.00	0.04
		CC	0.00	2.03	0.02	1.99	0.00	2.51	0.01	2.07

Table 1 continued

n	Censoring	Method	MCAR				MAR			
			β_1		β_2		β_1		β_2	
			Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
		PP	0.00	2.00	0.01	1.04	0.00	2.49	0.01	1.05
		IPW1	0.00	2.03	0.02	1.99	0.00	3.02	0.01	2.10
		IPW2	0.00	2.03	0.01	1.28	0.00	3.03	0.01	1.45

Simulation results from 2000 replications. Bias and MSE for the β_1 and β_2 estimators under the considered MCAR and MAR with a missing percentage on X^c equal to 30 and 70 %. Results are given for the full data (Full) estimator, the complete case (CC) estimator, the proposed estimator (PP), the simple IPW estimator (IPW1) and the kernel assisted estimator (IPW2). The MSE for the CC, PP, IPW1 and the IPW2 estimators are taken relative to the MSE for the full data estimator

CC estimator with regard to the effect of the covariate that has missing information, and more efficient than the CC estimator with regard to the effect of the covariate with complete information. The latter also holds in the MAR situation, but for the first covariate effect it can be less efficient than the CC estimator. Furthermore, it is consistently less efficient than the proposed estimator.

When going to the situation where the continuous covariate is normally distributed $N(0, 1)$ the picture is more or less the same, although the proposed estimator looses some efficiency under MAR scenario when the true coefficients are $(1, 1)$. Still, it is more efficient than the kernel assisted IPW estimator except for the effect of the covariate with missing information where it can be slightly less efficient under MCAR. Under MAR, however, the kernel assisted IPW estimator is less efficient than the CC estimator with regard to the effect of the covariate with missing information (Table 2).

Table 3 investigates the performance of the variance estimator for the proposed estimator in the case where the continuous covariate X^c is uniformly on $[0, 1]$ (results for the normal covariate case were similar and are not shown). Here we also considered sample size 800. For sample size 200 it is seen that the estimated standard errors tend to be a bit too small thus resulting in a bit too small coverage probabilities, but as sample size increases the estimated standard errors are close to the observed standard deviation and coverage probabilities are then also close to the nominal 95 %.

As suggested by a referee, we also considered the setup of the kidney cancer data application described in the next section where there is one continuous covariate approximately normal with mean 4.5 and sd equal to 1.44. This covariate is missing for half the patients under a MCAR mechanism. The other covariate, treatment, is categorical with two levels and is observed for all patients. Censoring percent is 37 % with censoring taking place after 1 year. Since the Breslow estimator applied to the data (based on complete cases) is approximately a straight line with a slope of approximately 0.003 (time scale in days), we took for this simulation $\lambda_0(t) = 0.003$. The two regression coefficients were set to the complete case estimates $(\hat{\beta}_X, \hat{\beta}_Z) = (0.45, -0.36)$. Sample size is $n = 347$. We ran 2000 simulations and calculated the estimator based on full data, the complete case estimator and the proposed estimator. Results are shown in Table 4, and are similar to the other simulation results.

Table 2 Continuous covariate distributed as N(0,1)

n	Censoring	Method	MCAR				MAR			
			β_1		β_2		β_1		β_2	
			Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
True value of $(\beta_1, \beta_2) = (0, 0)$										
200	30%	Full	−0.00	0.01	0.00	0.03	−0.00	0.01	−0.00	0.03
		CC	−0.01	2.07	0.00	2.10	0.00	2.16	−0.01	2.58
		PP	−0.01	2.06	0.00	1.03	0.00	2.15	−0.00	1.04
		IPW1	−0.01	2.08	0.00	2.10	0.00	2.76	−0.01	2.63
		IPW2	−0.01	2.03	−0.00	1.35	0.00	2.65	0.06	1.59
	70%	Full	0.00	0.02	−0.00	0.07	0.00	0.019	0.01	0.07
		CC	0.00	2.15	−0.02	2.16	0.00	2.23	−0.02	5.54
		PP	0.00	2.07	0.00	1.04	0.00	2.15	0.01	1.03
		IPW1	0.00	2.18	−0.02	2.17	0.002	2.73	−0.02	5.77
		IPW2	0.00	2.18	−0.00	1.30	0.00	2.82	0.03	4.34
400	30%	Full	0.00	0.004	−0.00	0.015	−0.00	0.004	0.00	0.014
		CC	0.00	2.05	−0.01	2.01	0.00	2.27	0.01	2.54
		PP	0.00	2.05	−0.00	1.02	0.00	2.27	0.00	1.01
		IPW1	0.00	2.06	−0.01	2.01	0.00	2.76	0.01	2.57
		IPW2	0.00	2.01	−0.00	1.20	0.00	2.66	0.05	1.57
	70%	Full	0.00	0.004	0.00	0.01	−0.00	0.009	0.00	0.036
		CC	0.00	2.06	0.00	2.13	−0.00	2.69	0.01	2.69
		PP	−0.00	2.06	−0.00	1.02	−0.00	2.07	0.00	1.02
		IPW1	0.00	2.06	0.00	2.08	−0.00	2.62	0.01	2.69
		IPW2	0.00	2.02	0.00	1.28	−0.00	2.64	0.02	1.36
True value of $(\beta_1, \beta_2) = (1, 1)$										
200	30%	Full	0.02	0.01	0.00	0.04	0.01	0.01	0.01	0.04
		CC	0.03	2.19	0.02	2.14	0.03	2.22	0.02	2.52
		PP	0.03	2.17	0.02	1.67	0.03	2.45	0.01	1.95
		IPW1	0.03	2.20	0.02	2.14	0.03	2.74	0.02	2.60
		IPW2	0.01	2.08	0.01	1.71	0.01	2.60	0.05	2.28
	70%	Full	0.01	0.02	0.01	0.08	0.01	0.02	0.01	0.06
		CC	0.03	2.43	0.03	2.14	0.03	2.48	0.02	2.60
		PP	0.03	2.38	0.03	1.43	0.04	2.66	0.03	1.81
		IPW1	0.03	2.46	0.03	2.14	0.04	3.13	0.03	2.67
		IPW2	0.03	2.38	0.03	1.61	0.03	3.09	0.03	2.16
400	30%	Full	0.00	0.01	0.00	0.02	0.01	0.006	0.01	0.02
		CC	0.01	2.13	0.01	2.12	0.02	2.34	0.02	2.50
		PP	0.01	2.12	0.01	1.67	0.02	2.38	0.01	1.86
		IPW1	0.01	2.13	0.01	2.12	0.01	2.83	0.02	2.58
		IPW2	0.00	2.03	0.01	1.69	−0.00	2.61	0.04	2.17
	70%	Full	0.01	0.01	0.01	0.04	0.01	0.01	0.01	0.04

Table 2 continued

n	Censoring	Method	MCAR				MAR			
			β_1		β_2		β_1		β_2	
			Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
		CC	0.01	2.14	0.01	2.13	0.02	2.43	0.01	2.33
		PP	0.01	2.15	0.02	1.41	0.02	2.41	0.02	1.55
		IPW1	0.01	2.15	0.01	2.13	0.03	3.03	0.01	2.37
		IPW2	0.01	2.12	0.01	1.58	0.03	2.97	0.02	1.85

Simulation results from 2000 replications. Bias and MSE for the β_1 and β_2 estimators under the considered MCAR and MAR with a missing percentage on X^c equal to 30 and 70 %. Results are given for the full data (Full) estimator, the complete case (CC) estimator, the proposed estimator (PP), the simple IPW estimator (IPW1) and the kernel assisted estimator (IPW2). The MSE for the CC, PP, IPW1 and the IPW2 estimators are taken relative to the MSE for the full data estimator

Table 3 Continuous covariate distributed as U(0,1)

n	Censoring	MCAR						MAR					
		β_1			β_2			β_1			β_2		
		SE	SEE	CP	SE	SEE	CP	SE	SEE	CP	SE	SEE	CP
True value of $(\beta_1, \beta_2) = (0, 0)$													
200	30 %	0.45	0.43	93.5	0.18	0.17	94.8	0.45	0.43	93.5	0.18	0.17	93.9
	70 %	0.69	0.66	95.2	0.28	0.27	94.3	0.68	0.65	93.6	0.28	0.27	94.3
400	30 %	0.30	0.30	94.3	0.12	0.12	94.4	0.30	0.30	94.6	0.12	0.12	95.2
	70 %	0.47	0.45	94.6	0.19	0.19	95.2	0.47	0.46	94.7	0.19	0.19	94.5
800	30 %	0.21	0.21	94.6	0.086	0.085	94.9	0.21	0.21	94.5	0.085	0.085	95.0
	70 %	0.32	0.32	94.9	0.13	0.13	95.0	0.32	0.32	95.1	0.13	0.13	95.2
True value of $(\beta_1, \beta_2) = (1, 1)$													
200	30 %	0.47	0.43	93.9	0.19	0.19	95.4	0.48	0.45	94.0	0.21	0.20	94.3
	70 %	0.71	0.68	94.8	0.29	0.30	96.0	0.79	0.75	94.9	0.31	0.30	94.7
400	30 %	0.31	0.30	94.5	0.14	0.13	94.2	0.32	0.31	94.6	0.14	0.14	95.0
	70 %	0.48	0.47	95.1	0.20	0.20	96.0	0.53	0.51	95.3	0.21	0.21	94.5
800	30 %	0.21	0.21	95.1	0.097	0.095	94.5	0.22	0.22	93.7	0.10	0.10	95.1
	70 %	0.34	0.33	94.5	0.14	0.14	94.4	0.36	0.35	94.9	0.14	0.14	95.6

Summary statistics for the β_1 and β_2 estimators under the considered MCAR and MAR with a missing percentage on X^c equal to 30 and 70 %. SE corresponds to the standard error of the estimator; SEE, to the mean of the standard error estimator; and CP, to the coverage probability of the 95 % confidence interval

The proposed estimator is unbiased and much more efficient than the CC estimator with regard to the covariate that is fully observed. The means of the estimated standard error of the proposed estimator are 0.071 and 0.153 for the two covariates, and the 95 % coverage probabilities are 94.7 and 93.6 showing reasonable performance.

Table 4 Simulation setup corresponding to the Kidney cancer data

Method	β_1			β_2		
	Bias	SE	MSE	Bias	SE	MSE
Full	0.004	0.053	0.003	−0.001	0.142	0.02
CC	0.009	0.075	2.01	−0.006	0.203	2.05
PP	0.008	0.074	1.92	−0.004	0.157	1.24

Results from 2000 replications. Bias, SE (standard error of the estimator) and MSE for the β_1 and β_2 estimators. Results are given for the full data (Full) estimator, the complete case (CC) estimator, and the proposed estimator (PP). The MSE for the CC and PP estimators are taken relative to the MSE for the full data estimator

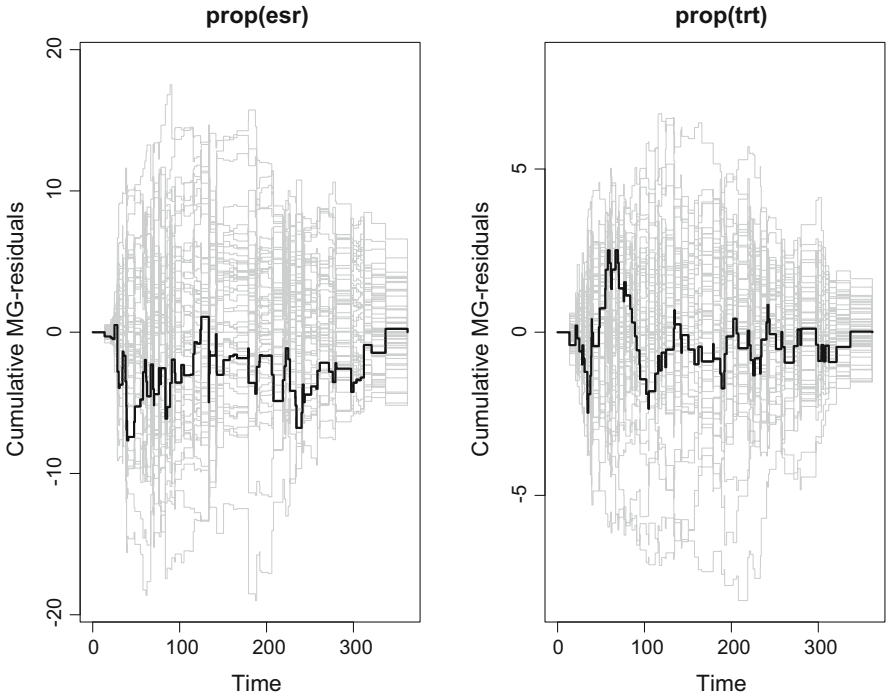

Fig. 1 Kidney cancer data. Score process plots

5 Application kidney cancer data

To illustrate the suggested method we use the kidney cancer data described in White and Royston (2009). These data are from the MRC RE01 study that was a randomised controlled trial comparing treatment with interferon-α (IFN) with the best supportive care and hormone treatment with medroxyprogesterone acetate (control) in patients with metastatic renal carcinoma. We use the same 347 patients as in White and Royston (2009). In this illustrative analysis we will only use the treatment variable and the erythrocyte sedimentation rate (*esr*), a variable that was only collected for half of the

Table 5 Kidney cancer data
Results from complete case (CC) analysis and from proposed (PP) with the two explanatory variables erythrocyte sedimentation rate (*esr*) and treatment (*trt*) with the control treatment as the reference level

Method	Variable	Estimate	s.e.	p value
CC	*esr*	0.45	0.076	2.9e-09
	trt	−0.37	0.201	0.07
PP	*esr*	0.42	0.073	8.5e-09
	trt	−0.34	0.148	0.02

patients, and it is argued in White and Royston (2009) that the missing data generating mechanism appears to be approximately MCAR. Further, we consider survival only in the first year after treatment, the censoring percent being 37 % with all except one censored after one year. The complete case analysis uses only 169 patients. The *esr* factor was power transformed to the power 0.4 to meet model requirements. The score process plots of Lin et al. (1993) was calculated using the R-package `timereg` (see, Martinussen and Scheike 2006, Chap. 6) and are shown in Fig. 1. It is seen that proportional hazards assumption seems reasonable with the supremum test resulting in the p values 0.74 and 0.92 for the two covariates.

Results from the complete case (CC) analysis and the proposed method (PP) are given in Table 5 where it is seen that point estimates are not very different. The estimated standard error concerning the treatment effect is smaller, however, for the proposed method compared to the complete case analysis (Table 5).

6 Concluding remarks

We have proposed a novel estimator for the Cox regression model under certain missing covariate scenarios which commonly arises in registry studies. We have shown consistency and asymptotic normality of the estimator, and provide explicit expressions for the variance-covariance matrix of the parameter estimates. The estimator is obtained as a modification of the observed partial likelihood and as a consequence is close to the maximum likelihood estimator which would typically have to be obtained via an EM-algorithm. Multiple imputation is conceptually related to the latter and is widely adopted in applied statistics, but unfortunately few general theoretical results exists for this approach in the survival analysis framework. We give such results for the proposed estimator. Our suggested inferential technique, however, is quite different compared to EM algorithm, and by exploiting a recursive structure offers great computational advantages with faster convergence and direct consistent estimates of the asymptotic covariance matrix.

In a simulation study we have examined the properties of the estimator in moderate sample sizes, and the results shows that it generally outperforms IPW estimators, and with efficiency for the covariates that are never missing which is close to the efficiency of the estimator with full data.

We have considered the situation where the missing data mechanism does not depend on the time to event outcome. While this may be too restrictive in some case, this is certainly a reasonable assumption in many prospective cohort studies and the

proposed estimator therefore provides an attractive alternative to multiple imputation which is routinely used in registry studies.

A limitation in our approach is that we only consider the case where the covariates which are always observed are discrete. We hereby avoid any need for explicitly modelling the joint distribution of the covariates which would otherwise be necessary. In practice this may not be a severe limitation, as flexible models can still be obtained by categorizing the relevant variables provided we have a sufficiently large dataset. In principle a flexible parametric model for the distribution of the covariates could be specified and we could reuse the estimation technique described in this paper. The main problem in this approach will be need of an approximation of (3), but this may be achieved with sufficiently precision with a computationally very appealing fully exponential approximation, see Tierney et al. (1989). Close-formed expressions may also be obtained in this case if we instead consider additive hazard models and limit our attention to for example mixtures of normal distributions for the covariates. The extension to these cases will be presented in a future paper.

Appendix

Consistency

The following shows that $\hat{\phi}_g\{\Lambda_0(t), \beta, z^*\}$ is a consistent estimator of $\phi_g\{\Lambda_0(t), \beta, z^*\}$ under the assumed MAR-assumption. The numerator (normed with n) of $\hat{\phi}_g$ converges in probability to

$$
\begin{aligned}
&E\big[R\exp\big(\beta_{g^c}^T X^g\big)\exp\big\{-\Lambda_0(t)\exp\big(\beta^T X\big)\big\}I(Z^g = z^*)\big] \\
&\quad = E\big[\exp\big(\beta_{g^c}^T X^g\big)\exp\big\{-\Lambda_0(t)\exp\big(\beta^T X\big)\big\}I(Z^g = z^*)P(R=1|X)\big] \\
&\quad = E[\exp\big(\beta_{g^c}^T X^g\big)\exp\{-\Lambda_0(t)\exp\big(\beta^T X\big)\}I(Z^g = z^*)P(R=1|W)] \\
&\quad = E\big[\exp\big(\beta_{g^c}^T X^g\big)\exp\big\{-\Lambda_0(t)\exp\big(\beta^T X\big)\big\}|Z^g = z^*\big]f(Z^g = z^*)P(R=1|w^*)
\end{aligned}
$$

and likewise with the denominator. Therefore, $\hat{\phi}_g$ converges in probability to

$$
\begin{aligned}
&\frac{E\big[\exp\big(\beta_{g^c}^T X^g\big)\exp\big\{-\Lambda_0(t)\exp\big(\beta^T X\big)\big\}|Z^g = z^*\big]f(Z^g = z^*)P(R=1|w^*)}{E\big[\exp\big\{-\Lambda_0(t)\exp\big(\beta^T X\big)\big\}|Z^g = z^*\big]f(Z^g = z^*)P(R=1|w^*)} \\
&\quad = \frac{E\big[\exp\big(\beta_{g^c}^T X^g\big)\exp\big\{-\Lambda_0(t)\exp\big(\beta^T X\big)\big\}|Z^g = z^*\big]}{E\big[\exp\big\{-\Lambda_0(t)\exp\big(\beta^T X\big)\big\}|Z^g = z^*\big]}
\end{aligned}
$$

as wanted. One can now prove that $\hat{\Lambda}_0(t, \beta)$ converges in probability to some $\Lambda_0(t, \beta)$ along the lines of the proof given in Zucker (2005). To do so we need the assumptions stated in the Appendix of Zucker (2005) adapted to the situation considered here. We list the relevant ones here for completeness.

I The function $\Lambda_0(t)$ is strictly increasing with derivative $\lambda_0(t)$.

II The covariates are bounded and the parameter β lies in a compact set B of $\Re^p$ that includes and open neighborhood of the true parameter β_0.
III $P(Y_i(\tau) = 1) > 0$.
IV The limiting value $v(\beta_0, \Lambda_0)$ of $D_\beta U(\beta)$ evaluated at β_0 and Λ_0 is positive definite.
V The baseline hazard function λ_0 is bounded over $[0, \tau]$ by some constant $\lambda_{\max}$.
VI The censoring distribution has at most a finite number of jumps on $[0, \tau]$.

It is not easy to give conditions ensuring the positive definiteness in condition **IV**, see Asgharian (2014) and references therein for results about singularities in the information matrix in general.

The consistency proof now goes almost unchanged compared to that in Zucker (2005), but there are some differences that we focus on next. Define

$$A_0(\beta, t, c) = n^{-1} \sum_{i=1}^{n} \sum_{g=1}^{G} Y_i(t) R_i^g \hat{\phi}_g\{c, \beta, Z_i^g\} \exp\left(\beta_g^T Z_i^g\right)$$

and

$$Q_n^g(t, z^g) = n^{-1} \sum_{i=1}^{n} Y_i(t) R_i^g I\left(Z_i^g = z^g\right).$$

Then $A_0(\beta, t, c)$ can be rewritten as

$$A_0(\beta, t, c) = \sum_{g=1}^{G} \sum_{z^g} \hat{\phi}_g\{c, \beta, z^g\} \exp\left(\beta_g^T z^g\right) n^{-1} Q_n^g(t, z^g)$$

that converges in probability to

$$a_0(\beta, t, c) = \sum_{g=1}^{G} \sum_{z^g} \phi_g\{c, \beta, z^g\} \exp\left(\beta_g^T z^g\right) q^g(t, z^g),$$

where $q^g(t, z^g)$ is the limit in probability of $Q_n^g(t, z^g)$.

Let, for a general function Λ

$$\Upsilon_n(t, \beta, \Lambda) = \int_0^t \frac{n^{-1} dN_\cdot(s)}{A_0(\beta, s, \Lambda(s-))}$$

and

$$\begin{aligned} \Upsilon(t, \beta, \Lambda) &= \int_0^t \frac{\sum_g E\left[Y_i(s) R_i^g \phi_g\{\Lambda_0(s), \beta, Z_i^g\} \exp\left(\beta^T Z_i^g\right)\right] \lambda_0(s) ds}{a_0(\beta, s, \Lambda(s-))} \\ &= \int_0^t \frac{\sum_g \sum_{z^g} \phi_g\{\Lambda_0(s), \beta, z^g\} \exp\left(\beta_g^T z^g\right) q^g(t, z^g) \lambda_0(s) ds}{a_0(\beta, s, \Lambda(s-))}. \end{aligned}$$

Then $\hat{\Lambda}_0(t, \beta)$ is the solution to the equation $\Lambda(t) = \Upsilon_n(t, \beta, \Lambda)$ subject to $\Lambda(0) = 0$. We further assume that

$$\inf_{t\in[0,\tau],c,\beta} a_0(\beta, t, c) > 0.$$

Under this assumption we can now conclude, as in Zucker (2005), that there exists a unique solution $\Lambda_0(t, \beta)$ to the functional equation $\Lambda = \Upsilon(t, \beta, \Lambda)$ subject to $\Lambda(0) = 0$. Similarly, one can also show that $\hat{\Lambda}_0(t, \beta)$ converges uniformly a.s. to $\Lambda_0(t, \beta)$, and that $\Lambda_0(t, \beta_0) = \Lambda_0(t)$. To the latter point note that $\Upsilon(t, \beta_0, \Lambda_0) = \Lambda_0(t)$, and hence $\Lambda_0(t, \beta_0) = \Lambda_0(t)$. Also, along the lines of the proof in Zucker (2005), one may show that $\tilde{U}(\beta, \hat{\phi}, \Lambda_0(\cdot, \beta))$ converges almost surely to a limit $\tilde{u}(\beta, \Lambda_0(\cdot, \beta))$, and that $\tilde{u}(\beta_0, \Lambda_0(\cdot)) = 0$ with the latter point following since in $\tilde{U}(\beta_0, \phi, \Lambda_0)$ we can replace the counting process increment $dN_i(t)$ in that expression with the martingale increment $dM_i(t)$ since the compensator of (4) is zero. It hence follows as in Zucker (2005) that $\hat{\beta}$ is consistent.

Asymptotic normality

The key to derive the large sample properties of the suggested estimators is to look at the following difference

$$n^{-1/2}\{S_0(t, \beta, \hat{\phi}, \hat{\Lambda}_0(\cdot, \beta)) - S_0(t, \beta, \phi, \Lambda_0)\} = \sum_{g=1}^{G}\sum_{z^g} n^{1/2}h_n(t, z^g)\exp(\beta_g^T z^g)\}Q_n^g(t, z^g),$$

where

$$h_n(t, z^g) = \hat{\phi}_g\{\hat{\Lambda}_0(t-, \beta), \beta, z^g\} - \phi_g\{\Lambda_0(t), \beta, z^g\}.$$

By a Taylor expansion of ϕ_g (in the first argument) we get

$$n^{1/2}h_n(t, z^g) = n^{1/2}[\hat{\phi}_g\{\Lambda_0(t), \beta, z^g\} - \phi_g\{\Lambda_0(t), \beta, z^g\}] + \hat{\phi}_g'(\Lambda_0, t, z^g)n^{1/2}\{\hat{\Lambda}_0(t-) - \Lambda_0(t-)\}$$

where $\hat{\phi}_g'$ denote the derivative of $\hat{\phi}_g$ with respect to first argument. It is easy to see that

$$n^{1/2}[\hat{\phi}_g\{\Lambda_0(t), \beta, z^g\} - \phi_g\{\Lambda_0(t), \beta, z^g\}] = n^{-1/2}\sum_{i=1}^{n}\epsilon_i^{\phi_g}(t, z^g) + o_p(1),$$

where

$$\epsilon_i^{\phi_g}(t,z^*)=\frac{R_i\exp\{-\Lambda_0(t)\exp\left(\beta^T X_i\right)\}I(Z_i^g=z^*)\{\exp\left(\beta_{g^c}^T X_i^g\right)-\phi_g\{\Lambda_0(t),\beta,z^*\}\}}{E\{R_i\exp\{-\Lambda_0(t)\exp\left(\beta^T X_i\right)\}I(Z_i^g=z^*)\}},$$

$i=1,\ldots,n$, are zero-mean iid variables. We also have that $\hat{\phi}_g'(\Lambda_0,t,z^g)$ converges in probability, denote this limit by $\mu(t,z^g)$. Let also $s_j(t)$ be short for $s_j(t,\beta,\Lambda_0,\phi)$ that are defined as the limits in probability of $n^{-1}S_j(t,\beta,\Lambda_0,\phi)$, j=0,1.

We now study the properties of $\hat{\Lambda}_0(t,\hat{\phi}_{\hat{\Lambda}_0})$ evaluated at the true β_0. Let

$$M_i(t)=N_i(t)-\int_0^t\sum_{g=1}^G Y_i(s)R_i^g\phi_g\{\Lambda_0(s),\beta,Z^g\}\exp\left(\beta_g^T Z^g\right)d\Lambda_0(s)$$

be the ith counting process martingale. We therefore obtain

$$\begin{aligned}n^{1/2}\left(\hat{\Lambda}_0\left(t,\hat{\phi}_{\hat{\Lambda}_0}\right)-\Lambda_0(t)\right)&=n^{1/2}\int_0^t S_0\left(u,\hat{\phi}_{\hat{\Lambda}_0}\right)^{-1}dM_\cdot(t)\\&\quad-\int_0^t\frac{n^{-1/2}\left\{S_0\left(u,\hat{\phi}_{\hat{\Lambda}_0}\right)-S_0(u,\phi_{\Lambda_0})\right\}}{n^{-1}S_0\left(u,\hat{\phi}_{\hat{\Lambda}_0}\right)}d\Lambda_0(u)\end{aligned}$$

using here an abbreviated notation to keep the expressions simple. From the calculations above it follows that $B_n(t)=n^{1/2}(\hat{\Lambda}_0(t,\hat{\phi}_{\hat{\Lambda}_0})-\Lambda_0(t))$ solves the following Volterra equation (see Andersen et al. (1993), p. 91)

$$B_n(t)=W_n(t)+\int_0^t B_n(s-)dV_n(s),\tag{6}$$

where

$$\begin{aligned}W_n(t)=&n^{1/2}\int_0^t S_0\left(s,\hat{\phi}_{\hat{\Lambda}_0}\right)^{-1}dM_\cdot(s)\\&+n^{-1/2}\sum_{i=1}^n\int_0^t\frac{\sum_{g=1}^G\sum_{z^g}\epsilon_i^{\phi_g}(s,z^g)\exp\left(\beta_g^T z^g\right)\}Q_n^g(s,z^g),}{n^{-1}S_0\left(s,\hat{\phi}_{\hat{\Lambda}_0}\right)}d\Lambda_0(s),\\V_n(t)=&\int_0^t\frac{\sum_{g=1}^G\sum_{z^g}\hat{\phi}_g'(\Lambda_0,s,z^g)\exp\left(\beta_g^T z^g\right)\}Q_n^g(s,z^g),}{n^{-1}S_0\left(s,\hat{\phi}_{\hat{\Lambda}_0}\right)}d\Lambda_0(s).\end{aligned}$$

The Volterra equation (6) has the solution (see Andersen 1993, p. 91)

$$B_n(t)=\int_0^t\prod_{(s,t]}\{1+dV_n(u)\}dW_n(s)\tag{7}$$

and we therefore have

$$n^{1/2}\{\hat{\Lambda}_0(t,\beta_0)-\Lambda_0(t)\}=n^{-1/2}\sum_{i=1}^{n}\epsilon_i^{\Lambda_0}(t)+o_p(1),$$

where

$$\epsilon_i^{\Lambda_0}(t)=\int_0^t\prod_{(s,t]}\{1+d\nu(u)\}s_0^{-1}(s)\times\left\{dM_i(s)+\sum_{g=1}^{G}\sum_{z^g}\epsilon_i^{\phi_g}(s,z^g)\exp(\beta_g^Tz^g),\}q^g(s,z^g)d\Lambda_0(s)\right\},$$

$i=1,\ldots,n$, are zero-mean iid terms. In the last display $\nu(t)$ is the limit in probability of $V_n(t)$.

We now look at the asymptotic distribution of $n^{-1/2}U(\beta_0)$. Let $\hat{X}_i(t)$ be equal to $\tilde{X}_i(t)$ with the ϕ_g's and $\Lambda_0(t)$ replaced with their estimates (evaluated at β_0), and similarly with $\hat{S}_j(t,\beta_0)$, $j=0,1$. We can decompose $n^{-1/2}U(\beta_0)$ in the following way:

$$\begin{aligned}n^{-1/2}U(\beta_0)=&n^{-1/2}\sum_{i=1}^{n}\int_0^\tau\left\{\hat{X}_i(t)-\frac{\hat{S}_1(t,\beta)}{\hat{S}_0(t,\beta)}\right\}dM_i(t)\\&-\int_0^\tau\sum_{g=1}^{G}\sum_{z^g}n^{1/2}h_n(t,z^g)\left\{\frac{D_\beta\phi_g\{\Lambda_0(t),\beta,z^g\}}{\phi_g\{\Lambda_0(t),\beta,z^g\}}+z^g\right\}\\&\times\exp\left(\beta_g^Tz^g\right)\Big\}Q_n^g(t,z^g)d\Lambda_0(t)\\&+\int_0^\tau\frac{\hat{S}_1(t,\beta)}{\hat{S}_0(t,\beta)}\sum_{g=1}^{G}\sum_{z^g}n^{1/2}h_n(t,z^g)\exp\left(\beta_g^Tz^g\right)\Big\}Q_n^g(t,z^g)d\Lambda_0(t),\end{aligned}$$

and it therefore follows that

$$n^{-1/2}U(\beta_0)=n^{-1/2}\sum_i\epsilon_i^U+o_p(1),$$

where

$$\begin{aligned}\epsilon_i^U=&\int_0^\tau\left\{\tilde{X}_i(t)-\frac{s_1(t)}{s_0(t)}\right\}dM_i(t)\\&-\int_0^\tau\sum_{g=1}^{G}\sum_{z^g}\{\epsilon_i^{\phi_g}(t,z^g)\\&+\mu(t,z^g)\epsilon_i^{\Lambda_0}(t)\}\left\{\frac{D_\beta\phi_g\{\Lambda_0(t),\beta,z^g\}}{\phi_g\{\Lambda_0(t),\beta,z^g\}}+z^g\right\}d\Lambda_0(t,\beta,z^g)\\&+\int_0^\tau\frac{s_1(t)}{s_0(t)}\sum_{g=1}^{G}\sum_{z^g}\left\{\epsilon_i^{\phi_g}(t,z^g)+\mu(t,z^g)\epsilon_i^{\Lambda_0}(t)\right\}d\Lambda_0(t,\beta,z^g),\end{aligned}$$

$i = 1, \ldots, n$, are zero-mean iid terms. In the latter display, $d\Lambda_0(t, \beta, z^g)$ is short for $\exp(\beta_g^T z^g)\}q^g(t, z^g)d\Lambda_0(t)$. It thus follows that $n^{-1/2}U(\beta_0)$ converges in distribution towards a normal variate, and therefore also that $n^{1/2}(\hat{\beta}_0 - \beta)$ converges in distribution towards a normal variate. The variance of the latter, Σ, is estimated consistently by

$$\hat{\Sigma} = n\{D_\beta U(\hat{\beta})\}^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i^U \{\hat{\epsilon}_i^U\}^T \{D_\beta U(\hat{\beta})\}^{-1}.$$

References

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer-Verlag, New York

Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. Ann Stat 10:1100–1120

Asgharian M (2014) On the singularities of the information matrix and multipath change-point problems. Theory Probab Appl 58:546–561

Bagdonavicius V, Nikulin M (1999) Generalised proportional hazards model based on modified parital likelihood. Lifetime Data Anal 5:329–350

Chen H (2002) Double-semiparametric method for missing covariates in Cox regression models. J Am Stat Assoc 97:565–576

Chen H, Little R (1999) Proportional hazards regression with missing covariates. J Am Stat Assoc 94:896–908

Cox DR (1972) Regression models and life-tables (with discussion). J R Stat Soc: Ser B 34:187–220

Herring AH, Ibrahim JG (2001) Likelihood-based methods for missing covariates in the Cox proportional hazards model. J Am Stat Assoc 96:292–302

Lin DY, Wei LJ, Ying Z (1993) Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 80:557–572

Martinussen T (1999) Cox regression with incomplete covariate measurements using the EM-algorithm. Scand J Stat 26:479–491

Martinussen T, Scheike TH (2006) Dynamic regression models for survival data. Springer-Verlag, New York

Pugh M, Robins J, Lipsitz S, Harrington D (1994) Inference in the Cox proportional hazards model with missing covariate data. Technical report, Harvard School og Public Health, Dept. of Biostatistics

Qi L, Wang CY, Prentice RL (2005) Weighted estimators for proportional hazards regression with missing covariates. J Am Stat Assoc 100:1250–1263

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89:846–866

Sherman M (2006) Complex step derivatives: how did i miss this? Biomed Comput Rev 2(3):27

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338:157–160

Tierney L, Kass RE, Kadane JB (1989) Fully exponential laplace approximations to expectations and variances of nonpositive functions. J Am Stat Assoc 84:710–716

Wang CY, Chen HY (2001) Augmented inverse probability weighted estimator for Cox missing covariate regression. Biometrics 57:414–419

White IR, Royston P (2009) Imputing missing covariate values for the Cox model. Stat Med 28:1982–98

Xu Q, Paik MC, Luo X, Tsai W-Y (2009) Reweighting estimators for Cox regression with missing covariates. J Am Stat Assoc 104:1155–1167

Zucker D (2005) A pseudo partial likelihood method for semi-parametric survival regression with covariate errors. J Am Stat Assoc 100:1264–1277