Lifetime Data Analysis, 7, 207–224, 2001


# A Profile Conditional Likelihood Approach for the Semiparametric Transformation Regression Model with Missing Covariates

HUA YUN CHEN hychen@uic.edu
*Division of Epidemiology and Biostatistics, School of Public Health, UIC 2121 West Taylor Street, Chicago, IL 60612*

RODERICK J. LITTLE rlittle@umich.edu
*Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109*

*Received March 16, 2000; Revised November 28, 2000; Accepted December 6, 2000*

**Abstract.** We propose a profile conditional likelihood approach to handle missing covariates in the general semiparametric transformation regression model. The method estimates the marginal survival function by the Kaplan-Meier estimator, and then estimates the parameters of the survival model and the covariate distribution from a conditional likelihood, substituting the Kaplan-Meier estimator for the marginal survival function in the conditional likelihood. This method is simpler than full maximum likelihood approaches, and yields consistent and asymptotically normally distributed estimator of the regression parameter when censoring is independent of the covariates. The estimator demonstrates very high relative efficiency in simulations. When compared with complete-case analysis, the proposed estimator can be more efficient when the missing data are missing completely at random and can correct bias when the missing data are missing at random. The potential application of the proposed method to the generalized probit model with missing continuous covariates is also outlined.

**Keywords:** Cox model, gamma odds model, missing pattern

## 1. Introduction

In survival analysis, covariate adjustments are often conveniently modeled by the Cox proportional-hazards model, which assumes that the hazard rate of a subject with covariate $\mathbf{Z}$ is $\lambda(t)\exp(\beta^T\mathbf{Z})$, where $\lambda$ is the baseline hazard rate and $\beta$ is a vector of parameters of interest and $\mathbf{Z}$ is a p-dimensional vector. The Cox model can also be interpreted in terms of a linear regression model between the transformed survival time and the covariate. Specifically, let $T$ be the survival time. Then the Cox regression model can be equivalently written as

$$\log\{\Lambda(T)\} = -\beta^T\mathbf{Z} + \epsilon, \tag{1}$$

where $\Lambda$ is the cumulative baseline hazard and $\epsilon$ has the extreme value distribution with density $\exp(\epsilon - e^{\epsilon})$, or equivalently $\log\epsilon$ is an exponential random variable with unit mean (Bennett, 1983; Cheng, Wei, and Ying, 1995; Murphy, Rossini, and Van der Vaart, 1997).

Alternatives to the Cox model are of interest when it does not produce a good fit to the data. One way of generating alternative models is to replace $\log\{\Lambda(\cdot)\}$ by an unspecified

monotonic transformation function of the survival time and consider error distributions other than the extreme value distribution for $\epsilon$ in Equation (1). For example, parallel to the simple linear regression model, the choice of the standard normal distribution yields the generalized probit model. Another choice is the family of Pareto distributions (Clayton and Cuzick, 1985; Dabrowska and Doksum, 1988) with distribution function

$$F(\epsilon) = 1 - (1 + \gamma e^{\epsilon})^{-\frac{1}{\gamma}}, \gamma > 0,$$

which generates the proportional gamma odds model. This family of models has an interesting interpretation, namely, the gamma odds defined as $[1-\{\text{pr}(T>t \mid \mathbf{Z})\}^{\gamma}]/\{\text{pr}(\boldsymbol{T}>t \mid \mathbf{Z})\}^{\gamma}$ are proportional across subjects with the proportional rates depending on covariates only. An alternative interpretation of this family of models is the Cox regression model with heterogeneity modeled by including unobserved gamma-distributed random variables with mean 1 and variance $\gamma$. The model reduces to the Cox regression model when $\gamma$ tends to zero and to the proportional odds model when $\gamma$ equals 1.

It is well known that maximizing the Cox partial likelihood (Cox, 1972, 1975) in the proportional-hazards regression model will produce efficient estimates of the regression parameters (Andersen and Gill, 1982; Begun et al, 1983). For censored data with fully-observed covariates, the analogous partial likelihood exists for the semiparametric transformation regression model (Dabrowska and Doksum, 1988), but it is much more complicated to analyze in general. Methods for estimation of the regression parameters include Monte-Carlo simulation to maximize the partial likelihood (Dabrowska and Doksum, 1988), methods based on marginal rank (Pettitt, 1984; Clayton and Cuzick, 1985; Cuzick, 1988), the estimating equation method (Cheng, Wei, and Ying, 1995), and the nonparametric maximum likelihood method (Bennett, 1983; Murphy, Rossini, and Van der Vaart, 1997). We consider a modification of these methods when covariates are not fully observed.

For the special case of the Cox regression model, methods for handling incomplete covariates have been proposed by Lin and Ying (1993), Robins, Rotnitzky, and Zhao (1994), Zhou and Pepe (1995), Wang et al (1997), Paik and Tsai (1997) based on the modification of the partial likelihood, and Lipsitz and Ibrahim (1998), Chen and Little (1999) based on the nonparametric likelihood. Because no closed-form partial likelihood is available for the general semiparametric transformation model, methods based on the direct modification of the partial likelihood are not straightforward to generalize. Methods based on the nonparametric likelihood can in principle be generalized to the semiparametric transformation regression model. However, the asymptotic properties of the estimates have not been established in the missing-data setting and no closed form of the variance estimate of the regression parameter exists in general. We propose a new profile conditional likelihood approach to incorporate cases with incomplete covariates into the analysis when censoring is independent of the covariates. In handling the functional parameter in the conditional likelihood, the approach uses a strategy similar to the pseudo-likelihood method proposed by Gong and Samaniego (1981) for the elimination of nuisance parameters in a parametric model. This approach yields consistent estimates of the regression parameters when the missing data are missing at random in the sense of Rubin (1976), see also Little and Rubin (1987). When data are complete, the proposed method involves loss of

efficiency when compared with some known fully efficient methods. However, the simulation study reported in §3 suggests that the loss of efficiency is often minor. When there are missing covariates, the proposed method can be more efficient and less biased than fully-efficient methods applied to the subset of complete cases, particularly when the fraction of incomplete cases is large.

The remainder of this paper is organized as follows. The basic idea of the maximum profile conditional likelihood method is introduced in §2.1. The asymptotic theory of the proposed estimator is established in §2.2. Extensive simulations of the efficiency of the proposed methods are reported in §3. §4 discusses the special case of the normal transformation regression model, which allows relatively easy handling of missing continuous covariates.

## 2. The Profile Conditional Likelihood Approach

Under the model (1), the survival time distribution conditional on the covariate can be expressed as

$$\mathrm{pr}(T > t \mid \mathbf{Z}) = \mathrm{pr}\{\epsilon > \log\Lambda(t) + \beta^T\mathbf{Z} \mid \mathbf{Z}\} = 1 - F\{\log\Lambda(t) + \beta^T\mathbf{Z}\},$$

where $F$ is the distribution function of $\epsilon$ and $\Lambda$ is an unspecified monotonic transformation function. The survival function depends on the unspecified monotonic transformation function $\Lambda$, the regression parameter $\beta$, and the error distribution function $F$ known either completely (e.g., in the Cox model) or up to a unknown parameter (e.g., in the gamma odds model). For the remainder of this paper, we use $\theta$ to denote $\beta$ when $F$ is completely known or both $\beta$ and the unknown parameter associated with $F$ when $F$ is known up to an unknown parameter. For notational simplicity, we rewrite the survival function in the following general form.

$$\mathrm{pr}(T > t \mid \mathbf{Z}) = \phi\{\mathbf{Z}, \theta, \Lambda(t)\}, \tag{2}$$

where the functional form of $\phi$ is assumed known. It is easy to see that model (1) is a special case of model (2). Theory development in the remainder of this section is directed to model (2).

Let $C$ denote the censoring time whose distribution is denoted by $G$. Let $X = \min(T, C)$ and $\delta$ be the censoring indicator taking 1 when $T \leq C$, 0 otherwise. We assume that the distribution of $C$ is independent of $T$ and $\mathbf{Z}$. Let $\varphi$ denote the derivative of $-\phi$ with respect to the argument $\Lambda$. Suppose that the covariates $\mathbf{Z}$ follow a parametric distribution denoted by $H_\gamma(\mathbf{z})$, where $\gamma$ is a parameter of $q$-dimension. Covariates are subject to missing values. Let $\mathbf{e}=(e^1,\cdots, e^p)$ be any vector in $R^p$ with components being 0 or 1. $\mathbf{e}$ also defines a mapping from $R^p$ to subsets of $R^p$ in the following way,

$$\mathbf{e}(\mathbf{Z}) = \{\mathbf{z} = (z_1,\cdots,z_p) : z_i = Z_i \text{ if } e^i = 1;\ \text{arbitrary if } e^i = 0\}$$

Let $\mathbf{M}$ denote the missing data indicator with the $i$th component taking the value 1 if the corresponding component of $\mathbf{Z}$ is observed, 0 otherwise. Using the notation defined above,

we can express the observed data in the form $(\mathbf{M}_i, X_i, \delta_i, \mathbf{M}_i(\mathbf{Z}_i))$, $i = 1, \cdots, n$. Let $\mathbf{e}_1, \cdots, \mathbf{e}_K$ denote all possible missing data patterns. The likelihood for the observed data is

$$\prod_{k=1}^{K}\prod_{i=1}^{n}\{\int_{\mathbf{z}\in\mathbf{e}_k(\mathbf{Z}_i)} p(\mathbf{M}=\mathbf{e}_k \mid X_i, \delta_i, \mathbf{Z}=\mathbf{z})p(X_i, \delta_i \mid \mathbf{Z}=\mathbf{z}, \theta, \Lambda)dH_\gamma(\mathbf{z})\}^{1_{\{\mathbf{M}_i=\mathbf{e}_k\}}},$$

where $p(x, \delta \mid \mathbf{z})$ is the density of the conditional distribution of $(X, \delta)$ given $\mathbf{Z}=\mathbf{z}$. Suppose further that missing data are missing at random, which means that

$$p(\mathbf{M}=\mathbf{e}_k \mid X_i, \delta_i, \mathbf{Z}_i) = p\{\mathbf{M}=\mathbf{e}_k \mid X_i, \delta_i, \mathbf{e}_k(\mathbf{Z}_i)\}, \text{ for } k = 1, \cdots, K.$$

As a result, the parts of likelihood describing the missing data mechanism and the censoring mechanism can be dropped from the full likelihood. The remaining part can be written as

$$\prod_{k=1}^{K}\prod_{i=1}^{n}(\int_{\mathbf{z}\in\mathbf{e}_k(\mathbf{Z}_i)} [\varphi\{\mathbf{z}, \theta, \Lambda(X_i)\}]^{\delta_i}[\phi\{\mathbf{z}, \theta, \Lambda(X_i)\}]^{1-\delta_i}dH_\gamma(\mathbf{z})\{d\Lambda(X_i)\}^{\delta_i})^{1_{\{\mathbf{M}_i=\mathbf{e}_k\}}} \quad (3)$$

For the proportional-hazards regression model, maximizing the likelihood (3) with respect to $(\theta, \Lambda)$ was proposed by Chen and Little (1999). We take a different approach here. The basic idea is first to transform the parameters $(\theta, \gamma, \Lambda)$ into $(\theta, \gamma, R)$, where $R$ denotes the marginal survival function without covariate. The likelihood for the new set of parameters can be expressed as the product of the conditional likelihood of $\mathbf{e}_k(\mathbf{Z})$ given $(X, \delta)$, which contains all the information about $(\theta, \gamma)$ and a small amount of information about $R$, and the marginal likelihood of $(X, \delta)$, which contains most information about $R$. We then estimate $R$ by the Kaplan-Meier estimator based on the marginal survival data $(X_i, \delta_i)$, $i = 1, \cdots, n$, i.e., ignore the part of information about $R$ contained in the conditional likelihood, and estimate $(\theta, \gamma)$ by maximizing the conditional likelihood of $\mathbf{e}_k(\mathbf{Z})$ given $(X, \delta)$ with $R$ replaced by the Kaplan-Meier estimate. We call the latter procedure the profile conditional likelihood method.

Specifically, consider the following parameter transformation from $(\theta, \gamma, \Lambda)$ to $(\tilde{\theta}, \tilde{\gamma}, R)$ such that

$$\begin{cases} \tilde{\theta} = \theta \\ \tilde{\gamma} = \gamma \\ R(t) = \int \phi\{\mathbf{z}, \theta, \Lambda(t)\}dH_\gamma(\mathbf{z}) \end{cases}$$

Under weak conditions, the inverse transformation

$$\Lambda(t) = v\{\tilde{\theta}, \tilde{\gamma}, R(t)\}$$

exists and we have

$$\frac{dR}{d\Lambda}(t) = -\int \varphi\{\mathbf{z}, \theta, \Lambda(t)\}dH_\gamma(\mathbf{z}).$$

Rewriting the likelihood (3) under the new set of parameters (with ˜ dropped from the transformed parameters for notational simplicity):

$$L(\theta, \gamma, R) = L_1(\theta, \gamma, R) L_2(R)$$

where $L_1(\theta, \gamma, R) = \Pi_{k=1}^{K} \Pi_{i=1}^{n} [p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \gamma, \theta, R\}]^{1_{\{\mathbf{M}_i = \mathbf{e}_k\}}}$ and $p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \gamma, \theta, R\}$ is

$$\left(\frac{\int_{\mathbf{z} \in \mathbf{e}_k(\mathbf{Z}_i)} \varphi[\mathbf{z}, \theta, v\{\theta, \gamma, R(X_i)\}] dH_\gamma(\mathbf{z})}{\int \varphi[\mathbf{z}, \theta, v\{\theta, \gamma, R(X_i)\}] dH_\gamma(\mathbf{z})}\right)^{\delta_i} \left(\frac{\int_{\mathbf{z} \in \mathbf{e}_k(\mathbf{Z}_i)} \phi[\mathbf{z}, \theta, v\{\theta, \gamma, R(X_i)\}] dH_\gamma(\mathbf{z})}{R(X_i)}\right)^{1-\delta_i},$$

and

$$L_2(R) = \prod_{i=1}^{n} R^{1-\delta_i}(X_i) \{dR(X_i)\}^{\delta_i}.$$

The likelihood $L$ consists of two parts. The first part $L_1$ is the conditional likelihood of observed covariates given the follow-up time and the censoring indicator. The second part $L_2$ is the marginal likelihood of $R$ given $(X, \delta)$. Maximizing this second part with respect to $R$ gives us the Kaplan-Meier estimator $\hat{R}$ based on data $(X, \delta)$. The properties of the Kaplan-Meier estimator have been extensively studied. We will concentrate on studying the estimate of $(\theta, \gamma)$ obtained by maximizing $L_1(\theta, \gamma, \hat{R})$.

Maximizing $L_1$ with respect to $(\theta, \gamma)$ is a routine calculation. In particular, the EM algorithm (Dempster, Laird, and Rubin, 1977) can be applied to simplify the computations. Note that

$$\begin{aligned} \log L_1(\theta, \gamma, R) \;=\; & \sum_{k=1}^{K} \sum_{i=1}^{n} 1_{\{\mathbf{M}_i = \mathbf{e}_k\}} (E\{\log p(\mathbf{Z}_i \mid X_i, \delta_i, \theta, \gamma, R) | \mathbf{e}_k(\mathbf{Z}_i), X_i, \delta_i, \theta^*, \gamma^*, R^*\} \\ & - E[\log p\{\mathbf{Z}_i | \mathbf{e}_k(\mathbf{Z}_i), X_i, \delta_i, \theta, \gamma, R\} | \mathbf{e}_k(\mathbf{Z}_i), X_i, \delta_i, \theta^*, \gamma^*, R^*\}]). \end{aligned}$$

Given $\theta^*$ and $\gamma^*$, by fixing both $R$ and $R^*$ at $\hat{R}$, we can iteratively maximize the first part of the right-hand side of the equation to obtain the maximum profile conditional likelihood estimator. In order to actually carry out the maximization, we need to find (a) the conditional likelihood score equations for $(\theta, \gamma, \Lambda)$ under full data and their expectations conditional on the observed data, (b) the derivatives of $v(\theta, \gamma, R)$ with respect to $(\theta, \gamma)$. For (a), analytical integration is preferred when it is possible. Otherwise, numerical integration can be used. For (b), we can obtain the derivatives analytically by the chain rule for an implicit function. $v$ in the derivative expression can then be replaced by the closed-form solution when it exists; see next section for an example. When no closed-form solution is available, we can solve for $v$ from the transformation expression with $(\theta, \gamma)$ fixed at current estimates.

The large sample behavior of the proposed estimator can be characterized in the following theorems. To avoid technical complications in the discussion of the asymptotic properties of the estimator, we maximize

$$\prod_{k=1}^{K} \prod_{i=1}^{n} [p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \theta, \gamma, \hat{R}\}]^{1_{\{a \le X_i \le b, \mathbf{M}_i = \mathbf{e}_k\}}}$$

instead of $L_1(\theta, \gamma, \hat{R})$, where $1 > R_0(a) > R_0(b) > 0$. $R_0$ is the true marginal survival function of $T$. Because of the conditional likelihood we use, this modification represents little change of the profile conditional likelihood. Proofs of the theorems are postponed to the appendix.

**Theorem 1** *Under conditions specified in the appendix, the maximum profile conditional likelihood estimate $\hat{\alpha}$ of $\alpha = (\theta, \gamma)$ is strongly consistent and asymptotically normally distributed, that is,*

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to N(0, \Sigma)$$

*where $\alpha_0$ is the true parameter and*

$$\Sigma = A^{-1} + A^{-1}BA^{-1},$$

*where*

$$A = -E\{\sum_{k=1}^{K} 1_{\{a \le X \le b, \mathbf{M}=\mathbf{e}_k\}} \frac{\partial^2}{\partial \alpha^2} \log\, p(\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha_0, R_0)\},$$

$$B = \int_0^{\tau} \{\frac{1}{R_0(u)\bar{G}(u)} \int_u^{\tau} \omega(X, \delta) R_0(X) dP(X, \delta)\}^{\otimes 2} \bar{G}(u) d\bar{R}_0(u),$$

$$\omega(X, \delta) = E[\sum_{k=1}^{K} 1_{\{a \le X \le b, \mathbf{M}=\mathbf{e}_k\}} \frac{\partial^2}{\partial \alpha \partial R} \log\, p\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha_0, R_0\} \mid X, \delta],$$

*$P(x, \delta)$ is the distribution function of $(X, \delta)$. $\tau \ge b$ such that $pr(X > \tau) > 0$.*

**Theorem 2** *The asymptotic variance of the maximum profile likelihood estimate can be consistently estimated by $\hat{A}^{-1} + \hat{A}^{-1}\hat{B}\hat{A}^{-1}$, where*

$$\hat{A} = -\sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} 1_{\{a \le X_i \le b, \mathbf{M}_i=\mathbf{e}_k\}} \frac{\partial^2}{\partial \alpha^2} \log\, p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \hat{\alpha}, \hat{R}\}$$

*and*

$$\hat{B} = \frac{1}{n} \sum_{i=1}^{n} \delta_i 1_{\{X_i \le \tau\}} [\frac{\hat{k}(X_i)}{\frac{1}{n}\sum_{j=1}^{n} 1_{\{X_j \ge X_i\}}}]^{\otimes 2},$$

*with*

$$\hat{k}(u) = \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \ge u\}} \hat{R}(X_i) \frac{\partial^2}{\partial R \partial \alpha} \log\, p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \hat{\alpha}, \hat{R}1\}_{\{\mathbf{M}_i=\mathbf{e}_k, a \le X_i \le b\}}.$$

## 3. Simulation Study of the Properties of the Proposed Estimator

Note that the proposed estimator is in general not fully efficient for estimating ($\alpha$, $R$) since information about $R$ in $L_1(\alpha, R)$ is not used when estimating $R$. However, we expect the loss of efficiency to be limited because the parameter estimation for $\alpha$ is based on the conditional likelihood $L_1$ which would generate the efficient estimating score for estimating $\alpha$ had we used an efficient estimate of $R$ when we profile $R$ out in the conditional likelihood. The structure of the variance of the proposed estimate gives a lower bound of the relative efficiency of the proposed estimator to the fully efficient estimator. To see this, note that the variance of the proposed estimate divides into two pieces. The first piece corresponds to the variance of the efficient estimate when the monotonic transformation is assumed to be known. Therefore, the variance of the efficient estimate in the semiparametric transformation regression model cannot be smaller than that. If the second piece turns out to be very small compared with the first piece, we can conclude that the proposed estimator should be close to fully efficient. However, it is worth noting that even when the second piece is large compared with the first piece, it is still possible that the proposed estimator does not involve a substantial loss of efficiency, since the monotonic transformation is not assumed to be known, and the regression parameter cannot be estimated as well as when the transformation is known. This bound can serve as a crude measure of magnitude of the loss of efficiency when no fully-efficient estimator is available for comparison.

Our study of the efficiency of the proposed estimator involves two parts. The first part is the efficiency loss of the proposed estimator compared with available fully-efficient estimators such as the Cox partial likelihood method in the Cox model and the nonparametric maximum likelihood method in the proportional-odds model, when covariates are fully observed. The second part is the efficiency gain of the proposed method over complete-case analysis when covariates are missing. In the simulation study, we consider primarily two semiparametric transformation models, namely, the Cox regression model and the proportional-odds model. Two independent binary covariates, with respective success probabilities 0.3 and 0.5, are used in this study. All the simulation results are based on 1000 repetitions of a sample size 200 except for the first set of simulations, which has sample size 100.

The first set of simulations was designed to examine the loss of efficiency of the proposed method when covariates are fully observed. The censoring distribution is uniform $U(0, a)$. Different values of $a$ are used to obtain approximately 0%, 40% , and 80% of censoring in each problem. Table 1 lists the ratios of mean squared errors between the efficient estimates and the maximum conditional profile estimates. In most of the simulated situations, the efficiency loss of the proposed estimate is less than 5%.

To investigate the potential gain of efficiency of the proposed method using all the observed data over the efficient estimate using only complete cases when the covariates have missing values, both the Cox regression model and the proportional odds model are simulated. The basic models are the same as in the first set of simulations except that half of the second covariate values are missing. Two missing data mechanisms are simulated. The first missing-data mechanism is a missing completely at random mechanism in which the second covariate values are randomly deleted. The second missing-data mechanism is a missing at random mechanism in which the second covariate value of a subject with

*Table 1.* Mean squared error comparison between the efficient estimate and the proposed profile conditional likelihood estimate when covariates are fully observed.

| Censoring(%) | $\beta_1 = 0$ | $\beta_2 = 0$ | $\beta_1 = 1$ | $\beta_2 = 1$ |
|---|---|---|---|---|
| *Cox regression model. Sample size = 100* | | | | |
| 0 | 105 | 105 | 99 | 100 |
| 40 | 100 | 100 | 100 | 101 |
| 80 | 98 | 98 | 98 | 98 |
| *Proportional odds model. Sample size = 100* | | | | |
| 0 | 98 | 97 | 97 | 97 |
| 40 | 97 | 96 | 96 | 96 |
| 80 | 95 | 95 | 95 | 95 |

*Table 2.* Comparison of parameter estimates between the complete case analysis using efficient estimates and the proposed method using all the observed data. 50% of the second covariate values are randomly deleted.

| | | Complete Case Analysis | | | | Profile Conditional Likelihood | | | |
|---|---|---|---|---|---|---|---|---|---|
| True ($\beta_1$, $\beta_2$) | Censoring (%) | $\beta_1$ bias | $\beta_1$ var | $\beta_2$ bias | $\beta_2$ var | $\beta_1$ bias | $\beta_1$ var | $\beta_2$ bias | $\beta_2$ var |
| *Cox regression model* | | | | | | | | | |
| (0,0) | 0 | 0.00 | 0.0458 | 0.01 | 0.0541 | 0.00 | 0.0214 | 0.01 | 0.0543 |
| | 40 | −0.02 | 0.0708 | 0.01 | 0.0861 | −0.01 | 0.0338 | 0.01 | 0.0848 |
| | 80 | −0.02 | 0.2408 | −0.05 | 0.2995 | −0.00 | 0.1096 | −0.05 | 0.2970 |
| (1,1) | 0 | 0.02 | 0.0568 | 0.03 | 0.0613 | 0.00 | 0.0306 | 0.03 | 0.0573 |
| | 40 | 0.02 | 0.0877 | 0.02 | 0.0810 | 0.02 | 0.0456 | 0.02 | 0.0765 |
| | 80 | 0.04 | 0.2671 | 0.04 | 0.2305 | 0.04 | 0.1369 | 0.04 | 0.2226 |
| *Proportional odds model* | | | | | | | | | |
| (0,0) | 0 | 0.00 | 0.1207 | −0.01 | 0.1475 | 0.00 | 0.05750 | −0.01 | 0.1496 |
| | 40 | 0.01 | 0.1725 | 0.01 | 0.1625 | −0.01 | 0.0707 | 0.01 | 0.1610 |
| | 80 | 0.01 | 0.3069 | −0.02 | 0.2959 | 0.01 | 0.1335 | −0.02 | 0.3008 |
| (1,1) | 0 | −0.01 | 0.1366 | 0.08 | 0.1621 | 0.02 | 0.0674 | 0.09 | 0.1687 |
| | 40 | 0.02 | 0.1476 | 0.02 | 0.1822 | 0.03 | 0.0799 | 0.03 | 0.1824 |
| | 80 | 0.07 | 0.2816 | −0.04 | 0.3248 | 0.05 | 0.1429 | −0.03 | 0.3368 |

follow-up time greater than the median follow-up time is not observed. When missing data are missing completely at random, simulation results in Table 2 suggest that by including the incomplete cases in the analysis, the regression coefficient estimate of the completely-observed covariate has a substantial gain of efficiency over the maximum likelihood estimator applied to the complete cases. A nearly 50% reduction in variance is achieved in all the simulated situations. The relative precision of the two estimates of the coefficient of the incompletely observed covariate is not significantly different. When missing data are missing at random, complete-case estimates are seriously biased. As indicated in the

*Table 3.* Comparison of parameter estimates between the complete case analysis using efficient estimates and the proposed method using all the observed data. Subjects with follow-up times greater than median follow-up time are missing their second covariate values.

| | | Complete Case Analysis | | | | Profile Conditional Likelihood | | | |
|---|---|---|---|---|---|---|---|---|---|
| rue ($\beta_1$, $\beta_2$) | Censoring (%) | $\beta_1$ bias | $\beta_1$ var | $\beta_2$ bias | $\beta_2$ var | $\beta_1$ bias | $\beta_1$ var | $\beta_2$ bias | $\beta_2$ var |
| *Cox regression model* | | | | | | | | | |
| (0,0) | 0 | 0.00 | 0.0435 | 0.01 | 0.0544 | −0.00 | 0.0248 | −0.09 | 0.8684 |
| | 40 | 0.00 | 0.0650 | 0.00 | 0.0793 | −0.01 | 0.0344 | 0.00 | 0.2196 |
| | 80 | −0.01 | 0.1757 | −0.07 | 0.1942 | 0.01 | 0.1038 | −0.07 | 0.2437 |
| (1,1) | 0 | −0.75 | 0.0491 | −0.71 | 0.0496 | −0.01 | 0.0355 | −0.03 | 0.2628 |
| | 40 | −0.54 | 0.0760 | −0.55 | 0.0680 | 0.01 | 0.0494 | −0.00 | 0.1823 |
| | 80 | −0.17 | 0.2085 | −0.23 | 0.1690 | 0.07 | 0.1413 | 0.03 | 0.2439 |
| *Proportional odds model* | | | | | | | | | |
| (0,0) | 0 | 0.00 | 0.1255 | −0.01 | 0.1516 | −0.01 | 0.0672 | −0.06 | 0.5812 |
| | 40 | −0.01 | 0.1308 | −0.02 | 0.1716 | −0.01 | 0.0729 | −0.01 | 0.3099 |
| | 80 | 0.01 | 0.2132 | −0.02 | 0.2779 | 0.01 | 0.1249 | −0.02 | 0.2797 |
| (1,1) | 0 | −0.42 | 0.1671 | −0.40 | 0.1432 | 0.03 | 0.0800 | 0.02 | 0.3379 |
| | 40 | −0.33 | 0.1553 | −0.32 | 0.1538 | 0.02 | 0.0730 | 0.03 | 0.2651 |
| | 80 | −0.08 | 0.2440 | −0.14 | 0.2204 | 0.04 | 0.1348 | 0.04 | 0.2600 |

*Table 4.* Bias of the estimator under covariate dependent censoring distribution $U(0, a(1+bz_2))$. $a$ is selected to obtain approximate 50% censoring in each situation. Parameters $(\beta_1, \beta_2)=(1,1)$.

| | $\beta_1$ | $\beta_1$ | $\beta_1$ | $\beta_2$ | $\beta_2$ | $\beta_2$ |
|---|---|---|---|---|---|---|
| b | bias | var | est var | bias | var | est var |
| *Cox regression model* | | | | | | |
| 0.5 | −0.02 | 0.0445 | 0.0437 | −0.03 | 0.0461 | 0.0419 |
| 1.0 | −0.03 | 0.0408 | 0.0419 | −0.07 | 0.0429 | 0.0406 |
| 2.0 | −0.02 | 0.0476 | 0.0467 | −0.17 | 0.0366 | 0.0465 |
| 10.0 | −0.05 | 0.0417 | 0.0429 | −0.27 | 0.0294 | 0.0430 |
| *proportional odds model* | | | | | | |
| 0.5 | −0.01 | 0.0861 | 0.0833 | 0.010 | 0.0903 | 0.0899 |
| 1.0 | 0.00 | 0.0780 | 0.0816 | −0.03 | 0.0798 | 0.0875 |
| 2.0 | 0.00 | 0.0981 | 0.0846 | −0.02 | 0.0936 | 0.0886 |
| 10.0 | −0.02 | 0.0830 | 0.0820 | −0.08 | 0.0764 | 0.0857 |

simulation results in Table 3, the conditional profile-likelihood estimate corrects bias at the cost of increase variance of the estimates.

The proposed method assumes that the censoring is independent of the covariates. We also did some simulations to evaluate the robustness of the proposed estimator when this

assumption is violated. Complete data are generated with censoring distribution uniform $U(0,a(1+bZ_2))$, where $b=0.5$, 1, 2 and 10. $a$ is adjusted to obtain approximately 50% censoring cases. Results are listed in Table 4. As expected, the estimates are biased when censoring depends on covariates. For small to moderate association between the censoring variable and covariates, the bias is not large in the simulated scenarios.

## 4. Some Applications

In this section, we discuss some specific applications of the proposed method.

### *4.1. The Generalized Probit Model with Missing Continuous Covariates*

The proposed method can in principle handle both discrete and continuous covariates. However, when continuous covariates are involved, the application of this method, like other likelihood-based methods, needs to resolve the problem of the possibly intractable integrals. With the generalized probit model, however, the computation can be substantially simplified when the continuous covariates are normally distributed with mean $\mu$ and variance $\Sigma$. Denote the normal density by $N(\mathbf{Z} \mid \mu, \Sigma)$. The joint density of $(T, \mathbf{Z})$ is

$$N\{\Lambda(t) \mid \beta^T\mathbf{Z}, 1\}N(\mathbf{Z} \mid \mu, \Sigma),$$

which can be rewritten as

$$N\{\mathbf{Z} \mid Z_0(t), (\Sigma^{-1} + \beta\beta^T)^{-1}\}N\{\Lambda(t) \mid \beta^T\mu, 1 + \beta^T\Sigma\beta\},$$

where $Z_0(t)=\mu+\{\Lambda(t)-\beta^T\mu\}\Sigma\beta/(1+\beta^T\Sigma\beta)$. Hence, the inverse transformation for $\Lambda$, as a function of $\beta$, $\mu$ and $R$, can be obtained explicitly from $\Phi[\{\Lambda(t)-\beta^T\mu\}/(1+\beta^T\Sigma\beta)^{1/2}] = 1-R(t)$ as

$$\Lambda(t) = \beta^T\mu + (1 + \beta^T\Sigma\beta)^{1/2}\Phi^{-1}\{1 - R(t)\},$$

where $\Phi$ denotes the standard normal distribution function. Hence, given the survival time $t$, the covariate distribution is

$$N[\mathbf{Z}|\mu + \Phi^{-1}\{1 - R(t)\}\frac{\Sigma\beta}{(1 + \beta^T\Sigma\beta)^{1/2}}, \Sigma - \frac{\Sigma\beta\beta^T\Sigma}{1 + \beta^T\Sigma\beta}].$$

Given that the survival time is greater than $t$, the distribution of the covariates is

$$\Phi[\beta^T(\mu - \mathbf{Z}) + (1 + \beta^T\Sigma\beta)^{1/2}\Phi^{-1}\{1 - R(t)\}]N(\mathbf{Z}|\mu, \Sigma)/R(t).$$

*Table 5.* Analysis of the mouse leukemia data.

| Complete case | | Proposed | | NPMLE | |
|---|---|---|---|---|---|
| $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ |
| *Proportional odds model* | | | | | |
| −1.44 | 1.44 | −1.59 | 1.68 | −1.49 | 1.67 |
| 0.3598 | 0.2096 | 0.3984 | 0.2601 | 0.3481 | 0.1879 |
| 0.2096 | 0.5194 | 0.2601 | 0.3943 | 0.1879 | 0.2981 |
| *Proportional odds model* | | | | | |
| −1.61 | 1.65 | −1.94 | 2.19 | | |
| 0.3562 | 0.1933 | 0.6009 | 0.3502 | | |
| 0.1933 | 0.5062 | 0.3502 | 0.5265 | | |

The first entry is parameter estimates, the next two entries are covariance estimates.

When no covariate is missing, the profile conditional likelihood,

$$(N[\mathbf{Z}|\mu + \Phi^{-1}\{1 - R(t)\}\frac{\Sigma\beta}{(1+\beta^T\Sigma\beta)^{1/2}}, \Sigma - \frac{\Sigma\beta\beta^T\Sigma}{1+\beta^T\Sigma\beta}])^{\delta}$$

$$\times(\Phi[\beta^T(\mu - \mathbf{Z}) + (1+\beta^T\Sigma\beta)^{1/2}\Phi^{-1}\{1 - R(t)\}]N(\mathbf{Z}|\mu, \Sigma))^{1-\delta},$$

can be maximized directly except that evaluations of the univariate normal distribution function and its inverse are needed. When some of the covariates are missing, the EM type algorithms can be applied. One additional complication is the need to evaluate an integral involving the ratio of normal density to its distribution function. The computation of the M-step can be simplified by using an ECM algorithm (Meng and Rubin, 1993).

### *4.2. An Example with Missing Discrete Covariates*

The data (Kalbfleisch and Prentice, 1980, Appendix 1) were collected from the study examining the viral and genetic effects on the incidence of spontaneous mouse leukemia. Several covariates were recorded in the original dataset. Preliminary exploration of the data suggested that the viral level and the phenotype of the gene Gpd-1 were the two important covariates. For simplicity, as in Chen and Little (1999), we include only these two covariates in the regression model and dichotomize the viral level according to whether it is greater than $10^4$ or not. One problem with this dataset is that the Gpd-1 phenotype is subject to heavy missing values. Among 204 mice included in the study, only 100 of the mice have the Gpd-1 phenotype information. The other covariate also has missing values.

We fit a Cox regression model to this dataset, using our proposed method and the nonparametric maximum likelihood method in Chen and Little (1999). Estimates given by the two methods in Table 5 are very close. As expected, the nonparametric maximum

likelihood method is more efficient and gives a smaller variance estimate. The use of Cox regression model in this example is largely based on the convenience and availability of methods of estimation. We also fit a proportional-odds model to the dataset by the proposed method. The difference between the complete-case estimate and the proposed estimate is larger than the corresponding comparison when a Cox model is fitted. This suggests that including incomplete cases into the analysis helps correct bias when the proportional-odds model is fitted to the data in the example.

## 5. Discussion

We have proposed a method of estimating the regression parameter in the general semiparametric regression model with incomplete covariate information. Although it is not fully efficient, theory suggests and simulations show that the proposed estimator has very high relative efficiency. The method does have the limitation that it requires censoring independent of the covariates, whereas the nonparametric maximum likelihood method only requires censoring to be independent of the covariates that have missing values. When compared with the full likelihood method, the proposed method has at least two advantages. First, the estimate of the infinite dimensional parameter $R$ has closed-form expression and iterations are needed only in solving for $(\theta, \gamma)$. Generally, the additional computation in solving for $v$ with $(\theta, \gamma)$ fixed is much simpler and stabler than in maximizing the likelihood with respect to $\Lambda$ with $(\theta, \gamma)$ fixed even when no closed-form solution for $v$ is available. This is because the latter requires solving a system of $N$ nonlinear equations simultaneously while the former requires solving $N$ independent nonlinear equations which can be solved separately, where $N$ is the number of observed events. In addition, the proposed method also gives an explicit expression for the variance estimate of the regression parameter without involving high dimension matrix operations. Second, the proposed method can be easily adapted to incorporate additional information about the marginal survival function $R$ from other sources. This can often happen in practice. In the extreme case where $R$ is known a priori, we need only to maximize likelihood $L_1$ for $(\theta, \gamma)$.

## Appendix A. Proofs

Let $h_\gamma(\mathbf{z})$ denote the density function of $H_\gamma(\mathbf{z})$. The following lemmas are needed in checking the conditions of the theorem or proving the theorems.

**Lemma 1** *(Existence of the inverse transformation): Suppose that,*

1. *for any $\theta$, $\gamma$, and $v \in (v_L, v_U)$, $\int \phi(\theta, \mathbf{z}, v) dH_\gamma(\mathbf{z})$ is continuous and strictly monotonic in $v$,*
2. *$\lim_{v \to v_L} \int \phi(\theta, \mathbf{z}, v) dH_\gamma(\mathbf{z}) = 1$ and $\lim_{v \to v_U} \int \phi(\theta, \mathbf{z}, v) dH_\gamma(\mathbf{z}) = 0$.*

*Then for any* $u \in (0,1)$, $\int \phi(\theta, \mathbf{z}, v) dH_\gamma(\mathbf{z}) = u$ *has a unique solution* $v = v(\theta, \gamma, u) \in (v_L, v_U)$ *which is strictly monotonic and continuous with respect to* $u$.

**Proof:** The existence and uniqueness are easy to see. To prove the continuity of $v$ with respect to $u$, let $u$, $u + \Delta u \in (0,1)$,

$$u + \Delta u = \int \phi(\theta, \mathrm{z}, v + \Delta v) dH_\gamma(\mathbf{z}) \text{ and } \mathrm{u} = \int \phi(\theta, \mathbf{z}, \mathrm{v}) \mathrm{dH}_\gamma(\mathbf{z}).$$

If $\lim_{\Delta u \to 0} \Delta v = 0$ does not hold, then there is a subsequence $\Delta v_n$ such that $\lim_{n \to \infty} \Delta v_n = v^* \neq 0$ and $v + v^* \in [v_L, v_U]$. As $\Delta u$ goes to zero, we have

$$0 = \int \phi(\theta, \mathbf{z}, \mathrm{v} + \mathrm{v}*) \mathrm{dH}_\gamma(\mathbf{z}) - \int \phi(\theta, \mathbf{z}, \mathrm{v}) \mathrm{dH}_\gamma(\mathbf{z})$$

which contradicts the strict monotonicity assumption.

**Lemma 2** *Suppose that* $c \leq i$, $d \leq i + j + k$, *and* $c + d = i + j + k$ *and that*

1. $\partial^{i+j+k} \phi(\theta, \mathbf{z}, v) / \partial \theta^c \partial v^d$ *and* $\partial^k h_\gamma(\mathbf{z}) / \partial \gamma^k$ *exist and are continuous;*
2. $$\int \sup_{|v - v'| < \epsilon, |\theta - \theta'| < \epsilon} \left| \frac{\partial \phi^{i+j+k-l}}{\partial \theta^c \partial v^{d-l}} (\theta', \mathbf{z}, v') \right| \sup_{|\gamma - \gamma'| < \epsilon} \left| \frac{\partial^l \log h_{\gamma'}(\mathbf{z})}{\partial \gamma^l} \right| dH_\gamma(\mathbf{z}) < +\infty \tag{4}$$

   *for any* $\theta$, $\gamma$, *and some* $\epsilon > 0$;
3. $\int \varphi(\theta, \mathbf{z}, v) dH_\gamma(\mathbf{z}) > 0$;

*then* $\partial^{i+j+k} v / \partial \theta^i \partial u^j \partial \gamma^k$ *exist and are continuous with respect to* $\theta$, $\gamma$, $u$.

**Proof:** We first show that $v$ is continuous with respect to $(\theta, \gamma, u)$.

$$\begin{aligned}
\Delta u &= \int \phi(\theta + \Delta\theta, \mathbf{z}, v + \Delta v) dH_{\gamma + \Delta\gamma}(\mathbf{z}) - \int \phi(\theta, \mathbf{z}, v) dH_\gamma(\mathbf{z}) \\
&= \int \{\phi(\theta + \Delta\theta, \mathbf{z}, v + \Delta v) - \phi(\theta, z, v + \Delta v)\} dH_{\gamma + \Delta\gamma}(\mathbf{z}) \\
&\quad + \int \phi(\theta, \mathbf{z}, v + \Delta v)) \{h_{\gamma + \Delta\gamma}(\mathbf{z}) - h_\gamma(\mathbf{z})\} d\mathbf{z} \\
&\quad + \int \{\phi(\theta, \mathbf{z}, v + \Delta v) - \phi(\theta, \mathbf{z}, v)\} dH_\gamma(\mathbf{z})
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\left| \int \{\phi(\theta, \mathbf{z}, v + \Delta v) - \phi(\theta, \mathbf{z}, v)\} dH_\gamma(\mathbf{z}) \right| \\
&\quad \le |\Delta u| + \int \phi(\theta, \mathbf{z}, v + \Delta v) \left| h_{\gamma+\Delta\gamma}(\mathbf{z}) - h_\gamma(\mathbf{z}) \right| d\mathbf{z} \\
&\quad\quad + \int \left| \phi(\theta + \Delta\theta, \mathbf{z}, v + \Delta v) - \phi(\theta, \mathbf{z}, v + \Delta v) \right| dH_{\gamma+\Delta\gamma}(\mathbf{z})
\end{aligned}$$

As $\Delta u$, $\Delta\theta$, $\Delta\gamma$ go to zero, The right side of the above inequality goes to zero. The second term goes to zero because of $0 \le \phi \le 1$ and Sheffé's theorem, and the third term goes to zero due to assumption 2. Therefore,

$$\lim_{\Delta u, \Delta\theta, \Delta\gamma \to 0} \int \phi(\theta, \mathbf{z}, v + \Delta v) dH_\gamma(\mathbf{z}) = \int \phi(\theta, \mathbf{z}, v) dH_\gamma(\mathbf{z}).$$

By the same argument as in *Lemma 1*, we have the continuity of $v$.

As an illustration, we prove the existence of derivatives $\partial v/\partial\theta$.

$$\begin{aligned}
&\int \frac{\phi(\theta + \Delta\theta, \mathbf{z}, v + \Delta v) - \phi(\theta + \Delta\theta, \mathbf{z}, v)}{\Delta v} dH_\gamma(\mathbf{z}) \frac{\Delta v}{\Delta\theta} \\
&\quad = -\int \frac{\phi(\theta + \Delta\theta, \mathbf{z}, v) - \phi(\theta, \mathbf{z}, v)}{\Delta\theta} dH_\gamma(\mathbf{z})
\end{aligned}$$

Since we have

$$\left| \frac{\phi(\theta + \Delta\theta, \mathbf{z}, v + \Delta v) - \phi(\theta + \Delta\theta, \mathbf{z}, v)}{\Delta v} - \frac{\partial\phi}{\partial v}(\theta, \mathbf{z}, v) \right| \le 2 \sup_{|\theta - \theta'| < \epsilon, |v - v'| < \epsilon} \left| \frac{\partial\phi}{\partial v}(\theta, \mathbf{z}, v) \right|$$

By the dominant convergence theorem, we have $\partial v/\partial\theta$ exists and equals

$$\frac{\int \partial\phi(\theta, \mathbf{z}, v)/\partial\theta dH_\gamma(\mathbf{z})}{\int \varphi(\theta, \mathbf{z}, v) dH_\gamma(\mathbf{z})}.$$

Let $\hat{R}$ be the Kaplan-Meier estimator of the marginal survival distribution based on the marginal survival times $(t, \delta)$. Let $R_0(t)$ be the true marginal survival distribution. Let $\tau > 0$ such that $R_0(\tau)\bar{G}(\tau) > 0$. We assume that $R_0$ is continuous. The following result is adapted from Lo and Singh (1986).

**Lemma 3** *(Uniform convergence of Kaplan-Meier estimator) suppose that censoring is independent of covariates. Then*

$$\sup_{t \le \tau} \left| \hat{R}(t) - R_0(t) + \frac{R_0(t)}{n} \sum_{i=1}^{n} \eta_i(t) \right| = O\{n^{-3/4} (\log n)^{3/4}\} \text{ a.s.}$$

*where*

$$\eta_i(t) = \frac{\delta_i 1_{\{x_i \le t\}}}{R_0(x_i)\bar{G}\,(x_i-)} - \int_0^{\min(t,x_i)} \frac{d\bar{R}_0(u)}{\{R_0(u)\}^2 \bar{G}\,(u-)}.$$

From now on, we will denote $(\theta, \gamma)$ by $\alpha$ and $DD'$ by $D^{\otimes 2}\}$ for any matrix $D$. Let $\alpha_0$ be the true parameter. Let

$$W_k\{\mathbf{e}_k(\mathbf{z}) \mid x, \delta, \alpha, \rho\} = 1_{\{a \le X \le b\}} \sup_{|\alpha - \alpha'| < \rho} p\{\mathbf{e}_k(\mathbf{z}) \mid x, \delta, \alpha', R_0\}, \text{ for } k = 1, \cdots, K.$$

The following assumptions are used in the proof of the consistency of the maximum profile conditional likelihood estimates of $\alpha$.

**A 1** *(Compactification): The parameter space $\alpha \in \Omega$ is compact under the same metric used in defining $\omega$. For simplicity we assume $\Omega$ is a closed bounded subset of a Euclidean space.*

**A 2** *(Continuity): For $k = 1, \cdots, K$,*

$$\lim_{\rho \to 0} W_k\{\mathbf{e}_k(\mathbf{z}) \mid x, \delta, \alpha, \rho\} = p\{\mathbf{e}_k(\mathbf{z}) \mid x, \delta, \alpha, R_0\} 1_{\{a \le X \le b\}}$$

**A 3** *(Identifiability): If $\alpha \neq \alpha_0$, then*

$$E[\sum_{k=1}^{K} \int \mid p\{\mathbf{e}_k(\mathbf{z}) \mid X, \delta, \alpha, R_0\} - p\{\mathbf{e}_k(\mathbf{z}) \mid X, \delta, \alpha_0, R_0\} \mid dz \; 1_{\{a \le X \le b, \mathbf{M} = \mathbf{e}_k\}}] \neq 0.$$

**A 4** *(measurability): For each $k$, $W_k\{\mathbf{e}_k(\mathbf{z}) \mid x, \delta, \alpha, \rho\}$ is a measurable function of $(\mathbf{e}_k(\mathbf{z}), x, \delta)$.*

**A 5** *(integrability):*

$$E(\sum_{k=1}^{K} 1_{\{a \le X \le b, \mathbf{M} = \mathbf{e}_k\}} [\log \frac{W_k\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha, \rho\}}{p\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha_0, R_0\}}]^+) < +\infty$$

**A 6**

$$\lim_{n \to +\infty} \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} 1_{\{a \le X \le b, \mathbf{M}_i = \mathbf{e}_k\}} \log \frac{p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \alpha, \hat{R}\}}{p\{\mathbf{e}_k(\mathbf{Z}_i \mid X_i, \delta_i, \alpha, R_0\}} = 0 \text{ a.s.}$$

*uniformly in $\alpha$.*

**A 7**

$$\lim_{n \to +\infty} \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} 1_{\{a \le X_i \le b, \mathbf{M}_i = \mathbf{e}_k\}} \frac{\partial^2}{\partial \alpha^2} \log p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \alpha, \hat{R}\}$$

$$= E[\sum_{k=1}^{K} 1_{\{a \le X_i \le b, \mathbf{M}_i = \mathbf{e}_k\}} \frac{\partial^2}{\partial \alpha^2} \log p\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha, R_0\}] \text{ a.s.}$$

*uniformly in a neighborhood of $\alpha_0$, and the right-hand side is a negative definite matrix.*

**A 8** *(equicontinuity) For $k = 1, \cdots, K$,*

$$\sum_{i=1}^{n} [1_{\{a \le X_i \le b, \mathbf{M}_i = \mathbf{e}_k\}} \frac{\partial}{\partial \alpha} \log \frac{p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \alpha, \hat{R}\}}{p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \alpha, R_0\}}$$

$$-E 1_{\{a \le X \le b, \mathbf{M} = \mathbf{e}_k\}} \frac{\partial}{\partial \alpha} \log \frac{p\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha, \hat{R}\}}{p\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha, R_0\}}] = o_p(n^{\frac{1}{2}})$$

**A 9** *$E[\sum_{k=1}^{K} 1_{\{a \le X \le b, \mathbf{M} = \mathbf{e}_k\}} \frac{\partial}{\partial \alpha} \log p\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha_0, R\} \mid X = x, \delta]$ is compact differentiable at $R_0$ and the derivative is a bounded linear operator in $L^2(P)$.*

Conditions A1-A6 are used to prove the consistency in addition to conditionsfor *Lemma 3*. Conditions A1–A5 are analogous to those of Kiefer and Wolfowitz (1956). Conditions A7–A9 are needed for establishing normality of the estimator of $\alpha$. Conditions A1–A5 can be checked for specific examples. Conditions A6–A8 are generally more difficult to check. However, when $\phi$ are continuous differentiable with respect to $(\theta, \nu)$ and have bounded derivatives up to the fifth order, $h_\gamma$ has continuous and bounded derivatives up to the third order, and $Z$ is bounded, A6–A9 are true (see Theorem 2.7.5 on page 159 and Theorem 2.10.6 on page 192, Van der Vaart and Wellner, 1996).

**Proof of Theorem 1:** The consistency follows from Wald (1996) and Kiefer and Wolfowitz (1956) arguments. To prove the normality, expanding the likelihood scores around $(\alpha_0, \hat{R})$ and applying conditions A6–A8, we have

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = -[\sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} 1_{\{a \le X_i \le b, \mathbf{M}_i = \mathbf{e}_k\}} \frac{\partial^2}{\partial \alpha^2} \log p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \alpha_0, R_0\} + o_p(1)]^{-1}$$

$$(\sum_{k=1}^{K} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1_{\{a \le X_i \le b, \mathbf{M}_i = \mathbf{e}_k\}} \frac{\partial}{\partial \alpha} \log p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \alpha_0, R_0\}$$

$$+\sqrt{n} E[\sum_{k=1}^{K} 1_{\{a \le X \le b, \mathbf{M} = \mathbf{e}_k\}} \frac{\partial}{\partial \alpha} \log \frac{p\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha_0, \hat{R}\}}{p\{\mathbf{e}_k(\mathbf{Z}) \mid X, \delta, \alpha_0, R_0\}}] + o_p(1)).$$

We have

$$\sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} 1_{\{a \le X_i \le b, \mathbf{M}_i = \mathbf{e}_k\}} \frac{\partial^2}{\partial \alpha^2} \log p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \alpha_0, R_0\} + o_p(1) \to A$$

$$\sum_{k=1}^{K} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1_{\{a \le X_i \le b, \mathbf{M}_i = \mathbf{e}_k\}} \frac{\partial}{\partial \alpha} \log p\{\mathbf{e}_k(\mathbf{Z}_i) \mid X_i, \delta_i, \alpha_0, R_0\} \to N(0, A)$$

By assumption A9, the second term in the expansion without the $o_p(1)$ term is $\int 1_{\{a\le x\le b\}}\omega(x,\delta)\sqrt{n}(\hat{R}-R_0)(x)dP(x,\delta)$, where

$$w(x,\delta)=E[\sum_{k=1}^{K}1_{\{a\le X\le b,\mathbf{M}=\mathbf{e}_k\}}\frac{\partial^2}{\partial\alpha\partial R}\log p\{\mathbf{e}_k(\mathbf{Z})\mid X,\delta,\alpha_0,R_0)\mid X=x,\delta\}].$$

Since $\sqrt{n}(R-R_0)(t)$ converges weakly to a Gaussian process on $t\in[a,\ b]$ with covariance function

$$v(t,s)=R_0(t)R_0(s)\int_0^{\min(t,s)}\frac{d\bar{R}_0(u)}{R_0(u)^2\bar{G}(u-)},$$

by the continuous mapping theorem, the second term in the Taylor expansion converges in distribution to a normal random variable with variance

$$\int\int\omega(t,\delta_1)\omega(s,\delta_2)v(t,s)dp(t,\delta_1)dp(s,\delta_2).$$

Separate the expression in two terms with regard to $t>s$ and $t\le s$, interchange the order of the integral. It is then easy to see the above expression is the same as $B$.

Note that the first term is asymptotically independent of the second term in the expansion. Hence, the asymptotic result about $\alpha$ follows.

**Proof of Theorem 2:** Note that

$$\hat{B}=\int_0^{\tau}\{\frac{\hat{k}(u)}{\frac{1}{n}\sum_{i=1}^{n}1_{\{X_i\ge u\}}}\}^{\otimes 2}d\frac{1}{n}\sum_{i=1}^{n}\delta_i 1_{\{X_i\le u\}}$$

Applying the uniform law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i 1_{\{X_i\le u\}}\to \mathrm{pr}(\delta=1)+\int_u^{\tau}G(s)dR_0(s)$$

$$\hat{k}(u)\to k(u)$$

$$\frac{1}{n}\sum_{i=1}^{n}1_{\{X_i\ge u\}}\to\bar{G}(u)R_0(u)$$

uniformly on $[0,\ \tau]$. The result follows.

## References

P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: A large sample study," *Annals of Statistics* vol. 10 pp. 1100–1120, 1982.

J. B. Begun, W. J. Hall, W. M. Huang, and J. A. Wellner, "Information and asymptotic efficiency in parametric-nonparametric models," *Annals of Statistics* vol. 11 pp. 432–452, 1983.

S. Bennet, "Analysis of survival data by the proportional odds model," *Statistics in Medicine* vol. 2 pp. 273–277, 1983

H. Y. Chen and R. J. Little, "Proportional hazards regression with missing covariates," *Journal of American Statistical Association* vol. 94 pp. 896–908, 1999.

S. C. Cheng, L. J. Wei, and Z. Ying, "Analysis of Transformation models with censored data," *Biometrika* vol. 82 pp. 835–45, 1995.

D. Clayton and J. Cuzick, "Multivariate generalization of the proportional hazards model," *Journal of Royal Statistical Society, Series A* vol. 148 pp. 82–117, 1985.

D. R. Cox, "Regression models and life-tables (with discussion)," *Journal Royal Statistical Society, Series B* vol. 34 pp. 187–220, 1972.

D. R. Cox, "Partial likelihood," *Biometrika* vol. 62 pp. 269–279, 1975.

J. Cuzick, "Rank regression," *Annals of Statistics* vol. 16 pp. 1369–1389, 1988.

D. D. Dabrowska and K. A. Doksum, "Partial likelihood in transformation models with censored data," *Scandinavian Journal of Statistics* vol. 15 pp. 1–23, 1988.

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)," *Journal of Royal Statistical Society, Series B* vol. 39, pp. 1–39, 1977.

G. Gong and F. J. Samaniego, "Pseudo maximum likelihood estimation: Theory and applications," *Annals of Statistics* vol. 9 pp. 861–869, 1981.

J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, Wiley: New York, 1980.

J. Kiefer and K. Wolfowitz, "Consistency of the maximum likelihood estimator in the presence of finitely many incidental parameters," *Annals of Mathematical Statistics* vol. 27 pp. 887–906, 1956.

D. Y. Lin and Z. Ying, "Cox regression with incomplete covariate measurements," *Journal of American Statistical Association* vol. 88 pp. 1341–1349, 1993.

S. R. Lipsitz and J. G. Ibrahim, "Estimating equations with Incomplete categorical covariates in the Cox model," *Biometrics* vol. 54 pp. 1002–1013, 1998.

R. J. Little and D. B. Rubin, *Statistical Analysis of Missing Data*, Wiley: New York, 1987.

S. H. Lo and K. Singh, "The product-limit estimator and the Bootstrap: Some asymptotic representations," *Probability Theory and Related Fields* vol. 71 pp. 455–65, 1986.

X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: a general framework," *Biometrika* vol. 80 pp. 267–278, 1993.

S. A. Murphy, A. J. Rossini, and A. W. Van der Vaart, "Maximum likelihood estimation in the proportional odds model," *Journal of American Statistical Association* vol. 92 pp. 968–976, 1997.

M. C. Paik and W. Y. Tsai, "On using the Cox proportional hazards model with missing covariates," *Biometrika* vol. 84 pp. 579–593, 1997.

A. N. Pettitt, "Proportional odds model for survival data and estimating using ranks," *Applied Statistics* vol. 33 pp. 169–75, 1984.

J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of American Statistical Association* vol. 89 pp. 846–866, 1994.

D. B. Rubin, "Inference and missing data," *Biometrika* vol. 63 pp. 581–592, 1976.

A. W. Van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer–Verlag: New York, 1996.

A. Wald, "Note on the consistency of the maximum likelihood estimate," *Annals of Mathematical Statistics* vol. 20 pp. 595–601, 1996.

C. Y. Wang, L. Hsu, Z. D. Feng, and R. L. Prentice, "Regression calibration in failure time regression," *Biometrics* vol. 53 pp. 131–145, 1997.

H. B. Zhou and M. S. Pepe, "Auxiliary covariate data in failure time regression," *Biometrika* vol. 82 pp. 139–149, 1995.