Metrika (2010) 72:199–217
DOI 10.1007/s00184-009-0248-1

Semiparametric maximum likelihood estimation in Cox proportional hazards model with covariate measurement errors

Chi-Chung Wen

Received: 27 August 2008 / Published online: 12 March 2009

Abstract This paper studies semiparametric maximum likelihood estimators in the Cox proportional hazards model with covariate error, assuming that the conditional distribution of the true covariate given the surrogate is known. We show that the estimator of the regression coefficient is asymptotically normal and efficient, its covariance matrix can be estimated consistently by differentiation of the profile likelihood, and the likelihood ratio test is asymptotically chi-squared. We also provide efficient algorithms for the computations of the semiparametric maximum likelihood estimate and the profile likelihood. The performance of this method is successfully demonstrated in simulation studies.

Keywords Covariate measurement error · Cox model · Semiparametric maximum likelihood estimate · Profile likelihood

1 Introduction

The proportional hazards model (Cox 1972) is commonly used to characterize the relationship between failure time and covariates. In many circumstances, because of the measuring mechanism or the nature of the environment, the covariates are often measured with error. Starting from work of Prentice (1982), various methods have been proposed to deal with measurement error survival data in the proportional hazards model. The regression calibration method (e.g., Prentice 1982; Wang et al. 1997; Xie et al. 2001) is a frequently used approach which gives a approximately consistent estimator for the regression parameter. The estimated partial likelihood method (e.g., Zhou and Pepe 1995; Zhou and Wang 2000) estimates the induced relative hazard

C.-C. Wen (✉)
Department of Mathematics, Tamkang University, Taipei, Taiwan
e-mail: ccwen@mail.tku.edu.tw

rate and presents consistent estimation. Other general consistent approaches include the conditional score method (Tsiatis and Davidian 2001), the corrected score method (e.g., Nakamura 1992; Kong and Gu 1999; Huang and Wang 2000; Hu and Lin 2002; Song and Huang 2005) and the generalized estimating equations based method (Cheng and Wang 2001).

In this paper, we describe a semiparametric maximum likelihood (SPML) method for the proportional hazards model with measurement error covariates. Let T^0, C and X denote the failure time, the censoring time, and the d-vector of true covariate respectively. The Cox model assumes the conditional hazard of T^0 at time t given $X = x$ is

$$e^{\beta^T x}\lambda(t), \tag{1}$$

where β is the regression parameter and λ is the baseline hazard function. For simplicity, we consider the case where X is a completely unobserved error-prone covariate vector and let Z is the surrogate measure of X. The observable survival data consist of the follow-up time $T = \min\{T^0, C\}$, the failure indicator $\delta = I(T^0 \leq C)$, and error-prone surrogate Z. Denote $w(x|z)$ the conditional density function of X given Z relative to a dominating measure m. Assume T^0 and (C, Z) are conditional independent given X, and the joint distribution of (C, Z) has nothing to do with parameters. Then the full likelihood for (T, δ, Z) is

$$L(\beta, \Lambda) = \int \left(\lambda(T)e^{\beta^T x}\right)^{\delta} \exp\left(-\Lambda(T)e^{\beta^T x}\right) w(x|Z)m(dx),$$

where $\Lambda(t) = \int_0^t \lambda(u)du$.

We will study the semiparametric maximum likelihood estimate (SPMLE) of (β, Λ) based on the n i.i.d. copies of (T, δ, Z), assuming that the measurement error distribution $w(x|z)$ is known. The assumption that $w(x|z)$ is known may be weakened if validation data or replicate measurement data are available, but we do not pursue it in this paper. We note that the pseudo-partial likelihood method of Zucker (2005) and the imputed partial likelihood score based method of Li and Ryan (2006) likewise make use of the conditional distribution of X given Z, but are otherwise different from our full likelihood-based approach. Full likelihood-based approaches in survival measurement error literature include the methods of Zhong et al. (1996) and Hu et al. (1998), but they, unlike ours, resort to the use of the conditional distribution of Z given X. Besides, the properties of their full likelihood-based estimators are mostly unknown and the computations are burdensome when the number of nuisance parameters to be estimated is large. In this paper, we will develop a profile likelihood theory for our likelihood-based estimator and propose an efficient numerical method for maximum likelihood estimation.

Joint analyses of failure time and longitudinal data have generated considerable recent interest (e.g., Henderson et al. 2000; Song et al. 2002; Zeng and Lin 2007), in which the failure time and repeated measures are modelled to depend on a common random effect, the asymptotic properties of the SPMLE are derived, and the EM is employed to compute the estimate. However, the model we assume is a measurement error model and replication data in our method is not necessary. Our method works with the conditional density $w(x|z)$ of the true covariate X given the surrogate Z, and the distribution of Z is arbitrary or unspecified. The conditional density $w(x|z)$ can be estimated by available validation or replication data parametrically or nonparametrically in practice. Besides, we do not pursue the EM algorithm to compute the SPMLE but propose a hybrid algorithm to compute the SPMLE. We found the hybrid algorithm works efficiently and is easy to implement.

The SPMLE has asymptotic properties analogous to those of parametric likelihood estimates. We establish the consistency of the SPMLE for (β, Λ), following the approaches and techniques developed in Murphy (1994), Parner (1998), and Kosorok et al. (2004), among others. With the consistency, we then treat the SPMLE of (β, Λ) as a solution to certain estimating equations and prove its asymptotic normality by Theorem 19.26 of van der Vaart (1998). In addition, we follow Murphy and van der Vaart (2000) to develop a profile likelihood theory for the regression parameter β. We show that the SPMLE of β is asymptotically normal and efficient, its covariance matrix can be estimated consistently by means of the profile likelihood, and the likelihood ratio test is asymptotically chi-squared. We would like to point out that, to the best of our knowledge, there exists no published work on the likelihood ratio test for the Cox regression in the presence of covariate measurement error.

For computing the SPMLE, we present an efficient hybrid algorithm, which uses a composite algorithmic mapping combining the Newton–Raphson method based algorithm and the self-consistency equation based algorithm. The self-consistency equation is an integral characterization of the score function. Simulation studies use the hybrid algorithm and the profile likelihood theory. They indicate that the proposed SPML method is quite satisfactory and that the normal approximation of SPMLE and the chi-squared approximation of profile likelihood statistic are valid with reasonable sample sizes.

This paper is organized as follows. Section 2 contains the model assumptions and asymptotic results. Section 3 provides the algorithms for computing the SPMLE and profile likelihood. Section 4 illustrates the method in simulation studies and Sect. 5 gives concluding remarks. The Appendix contains the proofs of model identifiability and asymptotic properties.

Throughout, let P_n, P_0, and $P_{(\beta,\Lambda)}$ be the expectations taken under the empirical distribution, the true underlying distribution, and a given model, respectively. We use the notations $o_P(1)$ and $O_P(1)$, respectively, for a sequence of random vectors converging to zero in probability and being uniformly tight.

2 Semiparametric maximum likelihood estimator

We make the following assumptions throughout. We assume that the data are observed on the time interval $[0, \tau]$ for some $\tau > 0$, and that $P(C \geq \tau) = P(C = \tau) > 0$ and $P(T^0 > \tau) > 0$. The last condition is satisfied if the survival study ends at some time τ at which a positive fraction of subjects is still at risk. For β to be identifiable, we assume that the support of Z is bounded and non-degenerate and that if

$$P\left(\int e^{c_1^T x} w(x|Z)m(dx) \Big/ \int e^{c_2^T x} w(x|Z)m(dx) = c_0\right) = 1 \qquad (2)$$

for some constant c_0, then $c_1 = c_2$. It is easy to show that $w(x|z)$ under the classical independent additive error model satisfies this identifiability condition. The identifiability of the model is given by Lemma A.1 in Appendix. The parameter space of (β, Λ) is $\mathcal{B} \times \mathcal{L}$, where $\mathcal{B}$ is a compact subset of $\Re^d$ with nonempty interior and $\mathcal{L} = \{\Lambda : [0, \tau] \to [0, \infty)|\Lambda(0) = 0, \Lambda$ is non-decreasing and right continuous$\}$. Denote (β_0, Λ_0) the true parameter. We assume β_0 is an interior point of $\mathcal{B}$ and Λ_0 has positive derivative on $[0, \tau]$.

Because L could become arbitrarily large within the class of absolutely continuous Λ, we extend the parameter space of Λ to a class of right continuous functions to allow for a discrete estimator and replace $\lambda(t)$ in the likelihood by $\Lambda\{t\}$, the jump of size of Λ at time t. The resulting likelihood function for the data $\{T_i, C_i, Z_i|i = 1, \ldots, n\}$ is

$$L_n(\beta, \Lambda) = \prod_{i=1}^{n} \int \left(\Lambda\{T_i\}e^{\beta^T x}\right)^{\delta_i} \exp\left(-\Lambda(T_i)e^{\beta^T x}\right) w(x|Z_i)m(dx). \tag{3}$$

The maximizer of (3), denoted by $(\hat{\beta}_n, \hat{\Lambda}_n)$, is referred to as the semiparametric maximum likelihood estimator of the parameter (β, Λ). We note that $\hat{\Lambda}_n$ is discrete with positive jumps at the observed failure times only.

The existence and asymptotic properties of the SPMLE are described by the following theorems, and the proofs of which are given in Appendix.

Theorem 1 (Existence) *The SPMLE* $(\hat{\beta}_n, \hat{\Lambda}_n)$ *exists and satisfies the equations*

$$\hat{\Lambda}_n(t) = \int_0^t \frac{1}{W_n(\hat{\beta}_n, \hat{\Lambda}_n; u)} dG_n(u), \tag{4}$$

where W_n *and* G_n *are monotone functions in* u *defined by*

$$W_n(\beta, \Lambda; u) = P_n\left[\frac{\int e^{\beta^T x} e^{\beta^T x \delta} \exp\left(-\Lambda(T)e^{\beta^T x}\right) w(x|Z)m(dx)}{\int e^{\beta^T x \delta} \exp\left(-\Lambda(T)e^{\beta^T x}\right) w(x|Z)m(dx)} I(u \leq T)\right],$$

and

$$G_n(u) = P_n[\delta I(T \leq u)].$$

We note that the integral equation (4) is called the self-consistency equation and is the basis of algorithm for computing the SPMLE.

Theorem 2 (Consistency) *The SPMLE* $(\hat{\beta}_n, \hat{\Lambda}_n)$ *is consistent*; *that is*, $\|\hat{\beta}_n - \beta_0\|$ *and* $\|\hat{\Lambda}_n - \Lambda_0\|_\infty$ *converge to* 0 *almost surly, where* $\|\cdot\|$ *is the Euclidean norm and* $\|\cdot\|_\infty$ *is the supremum norm on* $[0, \tau]$.

Let the profile likelihood for β be denoted by $pL_n(\beta)$, which is equal to $sup_\Lambda L_n(\beta, \Lambda)$. The following Theorems 3–5 are a consequence of the profile likelihood theory.

Theorem 3 (Asymptotic normality) *The SPMLE $\hat{\beta}_n$ is asymptotically normal and efficient at* (β_0, Λ_0)*; that is,*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Sigma^{-1}).$$

Here the variance Σ^{-1} in estimating β_0 is the inverse of the efficient Fisher information Σ, which is given in Appendix A.3.

Theorem 4 (Observed information estimate) *For all sequences $v_n \xrightarrow{p} v \in \Re^d$ and $\rho_n \xrightarrow{p} 0$ such that $(\sqrt{n}\rho_n)^{-1} = O_p(1)$,*

$$-2\frac{\log pL_n(\hat{\beta}_n + \rho_n v_n) - \log pL_n(\hat{\beta}_n)}{n\rho_n^2} \xrightarrow{p} v^T \Sigma v.$$

Let e_i be the d-dimensional unit vector with the ith component being 1. Using Theorem 4, we can show that

$$(\hat{\Sigma}_n)_{ij} \equiv -\Big[\log pL_n(\hat{\beta}_n + \rho_n e_i + \rho_n e_j) - \log pL_n(\hat{\beta}_n + \rho_n e_i) - \log pL_n(\hat{\beta}_n + \rho_n e_j) + \log pL_n(\hat{\beta}_n)\Big] \Big/ (n\rho_n^2) \qquad (5)$$

converges in probability to the (i, j)-entry of Σ.

Theorem 5 (Likelihood ratio inference) *Under the null hypothesis* $\mathbf{H}_0 : \beta = \beta_0$*, the profile likelihood ratio statistic,*

$$lrt_n(\beta_0) \equiv 2 \log \frac{pL_n(\hat{\beta}_n)}{pL_n(\beta_0)}$$

is asymptotically chi-squared with d degrees of freedom.

3 Algorithm

This section contains the methods for the computations of the SPMLE, variance estimate and profile likelihood ratio statistic. Making use of Newton–Raphson method and the self-consistency equation (4), we first present the following algorithm for computing SPMLE.

Algorithm for computing SPMLE

Step 1. Choose an initial value $(\beta^{(1)}, \Lambda^{(1)})$. One example might be $\beta^{(1)} = 0$ and $\Lambda^{(1)}(t) = t$.

Step 2. Update each current estimate $(\beta^{(k)}, \Lambda^{(k)})$, $k \geq 1$, by the following substeps.

Step 2.1. Update $\beta^{(k)}$ to $\beta^{(k+1)}$ by

$$\beta^{(k+1)} = \beta^{(k)} - L_{\beta\beta}^{-1}(\beta^{(k)}, \Lambda^{(k)}) L_\beta(\beta^{(k)}, \Lambda^{(k)}),$$

where L_β and $L_{\beta\beta}$ are the first and second derivatives of the log-likelihood function relative to β, respectively.

Step 2.2. Update $\Lambda^{(k)}$ to $\Lambda^{(k+1)}$ by

$$\Lambda^{(k+1)}(t) = \int_0^t \frac{1}{W_n(\beta^{(k+1)}, \Lambda^{(k)}; u)} dG_n(u),$$

where W_n and G_n are given in Theorem 1.

Step 3. If the updated estimate $(\beta^{(k+1)}, \Lambda^{(k+1)})$ is close to $(\beta^{(k)}, \Lambda^{(k)})$, then stop the procedure; otherwise, return to step 2.

The algorithm for computing SPMLE is based on an iterative procedure. The iterations of the algorithm are generated by a composite algorithmic mapping alternating steps of a Newton–Raphson method based algorithm and a self-consistency equation based algorithm. A minor variation of the algorithm is to inverse the order of steps 2.1 and 2.2. Step 2.1 is eventually the Newton–Raphson method for maximizing the log-likelihood function with Λ-parameter being fixed. The basis of step 2.2 is the self-consistency equation obtained from the integral characterization (4) of Λ. We note form (4) that the estimate of Λ given by the algorithm is a step function with positive jumps at the observed failure times only.

It is obvious from Theorem 4 and Theorem 5 that we need only to compute the profile likelihoods to obtain the variance estimate $\hat{\Sigma}_n$ and the profile likelihood ratio statistic $lrt_n(\beta_0)$. We now describe the method for the computation of profile likelihood. For every fixed β, let $\hat{\Lambda}_{n,\beta}$ be a random element at which the supremum in the definition of profile likelihood $pL_n(\beta)$ is taken. Notice that $\hat{\Lambda}_{n,\beta}$ can be calculated from the preceding algorithm by fixing β-parameter at β and omitting step 2.1. The profile likelihood $pL_n(\beta)$ is then computed by $L_n(\beta, \hat{\Lambda}_{n,\beta})$. That these algorithms are fast is seen clearly in the simulation studies.

4 Simulation studies

The purposes of this section is to assess the numerical performance of the proposed semiparametric maximum likelihood estimator and examine the normal approximation and chi-squared approximation which are claimed in the profile likelihood theory. All the computations are done on an ordinary PC using *Matlab* written by the author.

The simulation studies are presented for a setup with a signal error-prone continuous covariate X. We assume X is $N(0, 1)$ distributed and the measured value Z is given by $Z = X + e$, where the error term e is assumed to be independent of X and $N(0, \sigma_e^2)$ distributed. In this case, $w(x|z)$ is a normal density with mean $z/(1 + \sigma_e^2)$ and variance $\sigma_e^2/(1 + \sigma_e^2)$, and satisfies the identifiability condition (2).

Table 1 Simulation study with sample size 150

True β_0	Error var. σ_e^2	Mean	SD	MSE	SD$^{\text{prof}}$	CP (%)
0	0.25	−0.0080 (−0.0064)	0.1172 (0.0937)	0.0138 (0.0088)	0.1173 (0.0933)	94.7 (94.3)
	0.5	−0.0092 (−0.0061)	0.1278 (0.0850)	0.0164 (0.0073)	0.1290 (0.0852)	95.4 (94.8)
	1	−0.0110 (−0.0054)	0.1475 (0.0735)	0.0219 (0.0054)	0.1488 (0.0737)	96.2 (95.0)
0.5	0.25	0.4994 (0.3912)	0.1365 (0.1029)	0.0186 (0.0224)	0.1359 (0.0979)	95.4 (76.6)
	0.5	0.4980 (0.3208)	0.1535 (0.0929)	0.0236 (0.0407)	0.1566 (0.0886)	95.8 (46.5)
	1	0.4958 (0.2360)	0.1814 (0.0792)	0.0329 (0.0760)	0.1911 (0.0759)	96.1 (9.0)
1	0.25	1.0024 (0.7412)	0.1745 (0.1148)	0.0305 (0.0801)	0.1840 (0.1095)	95.5 (34.9)
	0.5	0.9979 (0.5890)	0.2055 (0.1021)	0.0422 (0.1793)	0.2235 (0.0967)	95.6 (2.6)
	1	0.9875 (0.4193)	0.2500 (0.0850)	0.0626 (0.3445)	0.2837 (0.0808)	95.1 (0)

Numbers in brackets are obtained by the naive method, others are obtained by our method

Table 2 Simulation study with sample size 300

True β_0	Error var. σ_e^2	Mean	SD	MSE	SD$^{\text{prof}}$	CP (%)
0	0.25	−0.0020 (−0.0016)	0.0852 (0.0681)	0.0073 (0.0046)	0.0822 (0.0656)	93.9 (93.9)
	0.5	−0.0021 (−0.0014)	0.0929 (0.0619)	0.0086 (0.0038)	0.0903 (0.0599)	94.7 (94.5)
	1	−0.0021 (−0.0010)	0.1066 (0.0532)	0.0114 (0.0028)	0.1047 (0.0518)	94.8 (94.0)
0.5	0.25	0.5019 (0.3928)	0.0974 (0.0731)	0.0095 (0.0168)	0.0943 (0.0686)	95.3 (62.5)
	0.50	0.5021 (0.3230)	0.1097 (0.0658)	0.0120 (0.0357)	0.1082 (0.0621)	95.2 (21.1)
	1	0.5023 (0.2383)	0.1303 (0.0558)	0.0170 (0.0716)	0.1317 (0.0533)	95.3 (2)
1	0.25	1.0063 (0.7418)	0.1263 (0.0815)	0.0160 (0.0733)	0.1272 (0.0764)	95.3 (10.2)
	0.5	1.0057 (0.5901)	0.1510 (0.0722)	0.0228 (0.1733)	0.1544 (0.0675)	95.5 (1)
	1	1.0025 (0.4204)	0.1882 (0.0599)	0.0354 (0.3395)	0.1976 (0.0564)	95.8 (0)

Numbers in brackets are obtained by the naive method, others are obtained by our method

We set $\sigma_e^2 = 0.25, 0.5$ or 1; $\beta_0 = 0, 0.5$ or 1; $\lambda_0(t) = 1$; common censoring at time 1. There are 1,000 replicates in each combination study and each replicate is a random sample of size $n = 150$ or 300. For each sample, we base on the algorithm in Sect. 3 and the profile likelihood theory to calculate the $\hat{\beta}_n$, $\hat{\Sigma}_n$, and $lrt_n(\beta_0)$. The ρ_n in (5) is set to be $n^{-1/2}$; the various integrals involved in the SPML estimation procedure were evaluated by 20-point Gauss-Hermite quadrature. The number of iteration used in the algorithm is set at 300, and the starting values are set as $\beta^{(1)} = 0$ and $\Lambda^{(1)}(t) = t$. Average computing times needed for one replicate to calculate $(\hat{\beta}_n, \hat{\Lambda}_n)$, $\hat{\Sigma}_n$, and $lrt_n(\beta_0)$ with sample size $n = 150$ and 300 are 7.5 and 40 seconds respectively, which indicate that the computation methods are efficient. We note that the algorithm usually converges within 100 iterations.

Tables 1 and 2 summarize the results of the simulation studies. Table 1 gives the results with sample size 150 and Table 2 gives the results with sample size 300. The first and second columns of Tables 1 and 2 list, respectively, the true value of regression parameter β_0 and error variance σ_e^2. The third, fourth, and fifth columnsreport respectively the sample mean (mean), sample standard deviation (SD) and sample mean-squared error (MSE) of the 1,000 estimates of β for each simulation scenario. The sixth column reports the average of the 1,000 standard deviations computed by profile likelihood (SD$^{\text{prof}}$); the final column gives the 95% coverage probability (CP) based on the normal approximation. For comparison, the simulated data are also analyzed by the naive method, which directly use the observed values of the covariate Z in the Cox model. The numbers in the brackets in the tables are the corresponding results obtained using naive method.

In terms of the sample mean (mean), mean-squared error (MSE) or 95% coverage probability (CP), Tables 1 and 2 indicate clearly that when no covariate effect presents ($\beta_0 = 0$), both naive method and ours work nicely; when covariate effect presents ($\beta_0 \neq 0$), our method still works well but the naive method may fail. The latter is more prominent when the covariate effect β_0 or the measurement error variance σ_e^2 increases. Further, when there is a covariate effect in the study, for larger sample size ($n = 300$), our method has more accurate 95% coverage probability but the naive method provides worse coverage probability. Among other things, the averages of the standard deviations obtained by profile likelihood (SD$^{\text{prof}}$) are quite close to the sample standard deviations of the 1,000 estimates (SD) for the naive method and our method. For our method, both SD and SD$^{\text{prof}}$ increase when the covariate effect or measurement error variance increases; for the naive method, they increase when the covariate effect increases or measurement error variance decreases.

Figures 1 and 2 are based on the simulation scenario with $\beta_0 = 0.5, \sigma_e^2 = 0.5$, and $n = 150$. Figure 1 is the Q–Q plot of standardized estimates $\sqrt{n}\hat{\Sigma}_n^{1/2}(\hat{\beta}_n - \beta_0)$ versus standard normal and Fig. 2 is the Q–Q plot of likelihood ratio statistics $lrt_n(\beta_0)$ versus chi-squared distribution with degree of freedom 1. They indicate that the normal approximation and chi-squared approximation claimed in the profile likelihood theory are quite satisfactory.

Remark Here is a remark concerning the model identifiability when the error variance σ_e^2 is treated as an unknown parameter in the above simulation example. If we treat σ_e^2 as an unknown parameter under independent additive error model, in which that Z is $N(0, 1 + \sigma_e^2)$ distributed is assumed, then it can be proved that (the proof is not presented) the parameter $(\beta, \Lambda, \sigma_e^2)$ is identifiable without validation or replication data in the study. If we assume $w(x|z)$ is $N(z/(1+\sigma_e^2), \sigma_e^2/(1+\sigma_e^2))$ density but the distribution of Z is free of the parameter $(\beta, \Lambda, \sigma_e^2)$, then, when there are no validation or replication samples within the study, the identifiability of the parameter $(\beta, \Lambda, \sigma_e^2)$ is doubted. In fact, we found from a simulation study (not presented) that, as sample size increases, the bias of SPMLE for β (or σ_e^2) decrease but the variance of SPMLE for β (or σ_e^2) does not.

5 Concluding remarks

In this paper, we have presented the semiparametric maximum likelihood approach to Cox regression with covarite measurement errors, assuming that the conditional distribution $w(x|z)$ of the true covariate X given the surrogate Z is known. As mentioned in

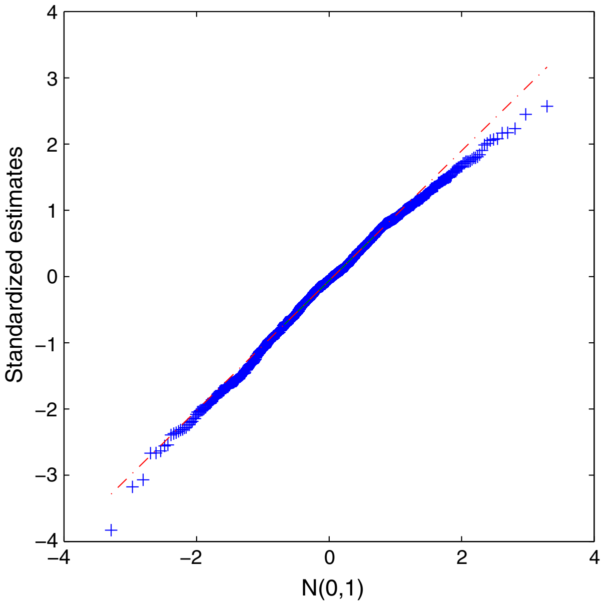

Fig. 1 Q–Q plot of standardized estimates versus standard normal for the study scenario with $\beta_0 = 0.5, \sigma_e^2 = 0.5$ and $n = 150$

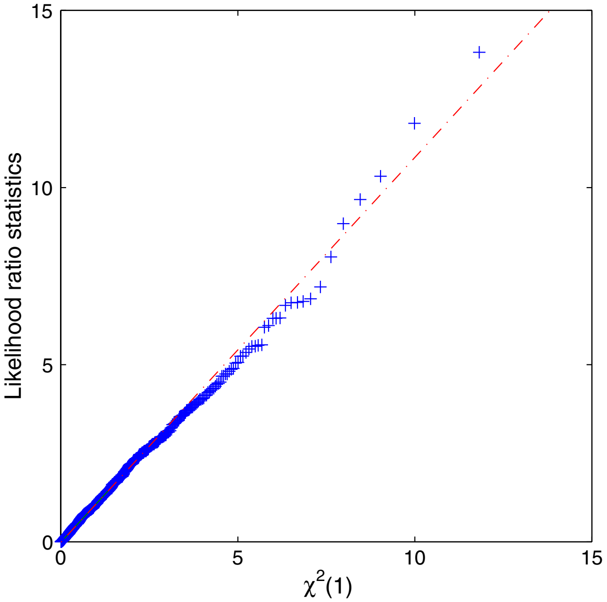

Fig. 2 Q–Q plot of likelihood ratio statistics versus chi-squared distribution with degree of freedom 1 for the study scenario with $\beta_0 = 0.5, \sigma_e^2 = 0.5$ and $n = 150$

Sect. 1, theoretical justification of asymptotic properties are deficient and computations are burdensome in the full likelihood-based methods of Zhong et al. (1996) and Hu et al. (1998). We have provided a profile likelihood theory with rigorous proofs and efficient computation methods for our full likelihood-based estimator. Our simulation studies indicate that the numerical performance of SPMLE is excellent, the profile likelihood-based normal and chi-squared approximations are valid, and the computation methods are feasible and efficient.

We assume that the covariate error distribution $w(x|z)$ is known in the method. The information about $w(x|z)$ may be provided by manufacturers of the measuring instrument or external studies. Besides, we may estimate $w(x|z)$ by using available validation data or replicate data, if any, and consider the estimated $w(x|z)$ to be known in practice. However, when validation data or replicate data are available, one might attempt to model the covariate error distribution $w(x|z)$ parametrically or nonparametrically and use the maximum likelihood principle to estimate β, Λ and $w(x|z)$ simultaneously. This is a potential topic for future work.

Appendix

Lemma A.1 (Identifiability) *If Λ is absolutely continuous relative to Λ_0, then $L(\beta, \Lambda) = L(\beta_0, \Lambda_0)$ a.s. implies that $(\beta, \Lambda) = (\beta_0, \Lambda_0)$.*

Proof The condition $L(\beta, \Lambda) = L(\beta_0, \Lambda_0)$ a.s. implies that

$$\left(\frac{d\Lambda}{d\Lambda_0}(T)\right)^{\delta} \frac{\int e^{\beta^T x\delta} \exp(-\Lambda(T)e^{\beta^T x})w(x|Z)m(dx)}{\int e^{\beta_0^T x\delta} \exp(-\Lambda_0(T)e^{\beta_0^T x})w(x|Z)m(dx)} = 1, a.s, \qquad \text{(A.1)}$$

where $d\Lambda/d\Lambda_0$ is the Radon–Nikodym derivative of Λ with respect to Λ_0. By considering (A.1) for T near 0 and $\delta = 1$, we have

$$\frac{d\Lambda}{d\Lambda_0}(0) = \frac{\int e^{\beta_0^T x} w(x|Z)m(dx)}{\int e^{\beta^T x} w(x|Z)m(dx)},$$

for almost all Z. Assumption that $P(\int e^{c_1^T x} w(x|Z)m(dx)/\int e^{c_2^T x} w(x|Z)m(dx) = c_0) = 1$ for some constant c_0 implies $c_1 = c_2$ gives us $\beta = \beta_0$. Because Λ_0 has positive derivative on $[0, \tau]$, we can set $\beta = \beta_0$, $\delta = 1$ in $L(\beta, \Lambda) = L(\beta_0, \Lambda_0)$ and integrate likelihoods with respect to T over $(0, t)$ to obtain

$$\int \exp\left(-\Lambda(t)e^{\beta_0^T x}\right) w(x|Z)m(dx) = \int \exp\left(-\Lambda_0(t)e^{\beta_0^T x}\right) w(x|Z)m(dx).$$

Hence $\Lambda = \Lambda_0$ on $[0, \tau]$. This completes the proof. □

A.1: Existence and consistency of SPMLE

Let $h_1 \in \Re^d$ and $h_2 \in BV$, the set of all functions of bounded variation on $[0, \tau]$. By considering the submodels specified by $(\beta + \epsilon h_1, \int_0^{\cdot}(1 + \epsilon h_2)d\Lambda)$ with $\mathbf{h} = (h_1, h_2) \in \Re^d \times BV$ and ϵ near 0, the score operator for (β, Λ) takes the form $\ell(\beta, \Lambda)[\mathbf{h}] = h_1^T \ell_1(\beta, \Lambda) + \ell_2(\beta, \Lambda)[h_2]$, where

$$\ell_1(\beta, \Lambda)(T, \delta, Z) = \frac{\int x[\delta - \Lambda(T)e^{\beta^T x}]e^{\beta^T x\delta} \exp(-\Lambda(T)e^{\beta^T x})w(x|Z)m(dx)}{\int e^{\beta^T x\delta} \exp(-\Lambda(T)e^{\beta^T x})w(x|Z)m(dx)},$$

$$\ell_2(\beta, \Lambda)[h_2](T, \delta, Z) = \delta h_2(T) - \int_0^T h_2 d\Lambda \left[\frac{\int e^{\beta^T x\delta} \exp(-\Lambda(T)e^{\beta^T x})e^{\beta^T x} w(x|Z)m(dx)}{\int e^{\beta^T x\delta} \exp(-\Lambda(T)e^{\beta^T x})w(x|Z)m(dx)}\right].$$

Proof of Theorem 1 Given observed data $\{(T_i, \delta_i, Z_i)|i = 1, \ldots, n\}$. Using the fact that $\lim_{y\to\infty} ye^{-y} = 0$, we conclude from

$$|L_n(\beta, \Lambda)| \leq \prod_{i=1}^n \int \left(\Lambda(T_i)e^{\beta^T x}\right)^{\delta_i} \exp\left(-\Lambda(T_i)e^{\beta^T x}\right) w(x|Z_i)m(dx)$$

that there exists $M_0 > 0$ such that $\max_{\Lambda(T_{(n)})\leq M_0} L_n(\beta, \Lambda) > \sup_{\Lambda(T_{(n)})>M_0} L_n(\beta, \Lambda)$, where $T_{(n)}$ is the largest observed follow-up time. Because L_n is a continuous function, it has a maximizer on any compact set. Consequently, the existence of SMPLE is obtained. This implies that a necessary condition for $(\hat\beta_n, \hat\Lambda_n)$ to be the SPMLE is $P_n\ell_2(\hat\beta_n, \hat\Lambda_n)[h_2] = 0$. Setting $h_2(t) = I(t \leq u)$ in $P_n\ell_2(\hat\beta_n, \hat\Lambda_n)[h_2] = 0$ and changing the order of integration result in (4). This completes the proof. □

Further, we can show that $P_n\ell(\hat\beta_n, \hat\Lambda_n)[\mathbf{h}] = 0$ for all large n and that

$$\Lambda_0(t) = \int_0^t \frac{1}{W(\beta_0, \Lambda_0; u)} dG(u) \tag{A.2}$$

as follows. Here $W(\beta, \Lambda; u) = P_0 W_1(\beta, \Lambda; u)$ and $G(u) = P_0 G_1(u)$. By Lemma A.1 and the Kullback–Leibler divergence theorem (see, for example, Lemma 5.35 of van der Vaart 1998), the function $\epsilon \mapsto P_0 \log[L(\beta_0 + \epsilon h_1, \int_0^{\cdot}(1 + \epsilon h_2)d\Lambda_0)/L(\beta_0, \Lambda_0)]$ attains its maximum at 0, which implies $P_0\ell(\beta_0, \Lambda_0)[\mathbf{h}] = 0$ by the fact that (β_0, Λ_0) is an interior point. Because $(\hat\beta_n, \hat\Lambda_n)$ asymptotically is an interior point by Theorem 2, we can similarly obtain $P_n\ell(\hat\beta_n, \hat\Lambda_n)[\mathbf{h}] = 0$ for all large n. Using the fact that $P_0\ell(\beta_0, \Lambda_0)[\mathbf{h}] = 0$ and a similar argument in proving (4), we obtain (A.2).

We will repeatedly use the following lemma, taken from section 19.4 of van der Vaart (1998), to prove Theorem 2.

Lemma A.2 *Suppose that $\mathcal{F}$ is a P-Glivenko–Cantelli class of P-measurable functions and $\hat f_n$ is a sequence of random functions that are contained in $\mathcal{F}$. If $\hat f_n$ converges* *almost surely to a function f_0 and the sequence is dominated such that $P(\hat{f}_n - f_0)$ converges to 0 almost surely, then $P_n(\hat{f}_n - f_0)$ converges to 0 almost surely.*

Proof of Theorem 2 Before the proof, one particular remark is that all the arguments in this proof are made and hold for a fixed sample point ω in the probability space except a set with zero probability. From the definition of W_n, there exists c_1 and c_2 such that

$$c_1 P_n[I(\tau \le T)] \le W_n(\beta, \Lambda; u) \le c_2 P_n[I(u \le T)] \le c_2,$$

which implies that W_n is uniformly bound away from 0 and infinity for large n by the condition that $P(T^0 \ge \tau) > 0$. Thus, it follows from (4) that the sequence $\hat{\Lambda}_n(\tau)$ is bounded. With this, we now claim the existence of a uniformly convergent subsequence of $(\hat{\beta}_n, \hat{\Lambda}_n)$. Because $\mathcal{B}$ is compact, it suffices to show $\{\hat{\Lambda}_n\}$ has a uniformly convergent subsequence. Let E be a countable dense subset of $[0, \tau]$. By Theorem 7.23 of Rudin (1976), $\{\hat{\Lambda}_n\}$ has a subsequence, say $\{\hat{\Lambda}_n\}$ again to simplify the notation, that converges on E. Let $\varepsilon > 0$. By the continuity of G and the fact that

$$|\hat{\Lambda}_n(s) - \hat{\Lambda}_n(t)| \le O(1)|G_n(s) - G_n(t)| = O(1)|G(s) - G(t)| + o(1)$$

for $s, t \in [0, \tau]$, we can choose $\delta > 0$ and N_1 such that $|\hat{\Lambda}_n(s) - \hat{\Lambda}_n(t)| < \varepsilon$ for all $n > N_1$ and $|s - t| < \delta$. Let $V(s, \delta) = \{t \in [0, \tau] \,|\, |s - t| < \delta\}$. Since E is a countable dense subset of $[0, \tau]$, there exists $s_1, \ldots, s_K$ in E such that $[0, \tau] \subset \bigcup_{k=1}^{K} V(s_k, \delta)$. Because $\hat{\Lambda}_n(s)$ converges for every $s \in E$, we can choose $N > N_1$ such that $|\hat{\Lambda}_m(s_k) - \hat{\Lambda}_n(s_k)| < \varepsilon$ whenever $m, n \ge N$, and $k = 1, \ldots, K$. If $s \in [0, \tau]$, then $s \in V(s_k, \delta)$ for some k, so that $|\hat{\Lambda}_m(s) - \hat{\Lambda}_m(s_k)| < \varepsilon$ for every $m \ge N$. Consequently, for $m, n \ge N$, we have

$$\begin{aligned}|\hat{\Lambda}_m(s) - \hat{\Lambda}_n(s)| &\le |\hat{\Lambda}_m(s) - \hat{\Lambda}_m(s_k)| + |\hat{\Lambda}_m(s_k) - \hat{\Lambda}_n(s_k)| + |\hat{\Lambda}_n(s_k) - \hat{\Lambda}_n(s)| \\ &< 3\varepsilon.\end{aligned}$$

This indicates $\hat{\Lambda}_n$ converges uniformly on $[0, \tau]$, and hence the claim is verified.

Now, denote the uniformly convergent subsequence of $(\hat{\beta}_n, \hat{\Lambda}_n)$ still by the original sequence $(\hat{\beta}_n, \hat{\Lambda}_n)$ for simplicity and suppose its limit is (β^*, Λ^*). We will show that $(\beta^*, \Lambda^*) = (\beta_0, \Lambda_0)$.

Define $\tilde{\Lambda}_n(t) = \int_0^t W_n(\beta_0, \Lambda_0; u)^{-1} dG_n(u)$. It is known that the class of monotone functions with a common upper bound is a Donsker class. Using the Donsker class argument (Theorem 2.10.6 of van der Vaart and Wellner 1996), we can show that the class $\{W_1(\beta, \Lambda; u) | \beta \in \mathcal{B}, \Lambda \in \mathcal{L}_c, u \in [0, \tau]\}$ is a Donsker class, where $\mathcal{L}_c = \{\Lambda \in \mathcal{L} | \Lambda(\tau) \le c\}$ for some $c \in (0, \infty)$. Hence, two applications of Lemma A.2 yield, uniformly in t,

$$\hat{\Lambda}_n(t) = \int_0^t \frac{1}{W_n(\hat{\beta}_n, \hat{\Lambda}_n; u)} dG_n(u) \to \int_0^t \frac{1}{W(\beta^*, \Lambda^*; u)} dG(u), \tag{A.3}$$

$$\tilde{\Lambda}_n(t) = \int_0^t \frac{1}{W_n(\beta_0, \Lambda_0; u)} dG_n(u) \to \int_0^t \frac{1}{W(\beta_0, \Lambda_0; u)} dG(u). \tag{A.4}$$

(A.3) indicates that $\Lambda^*(t) = \int_0^t W(\beta^*, \Lambda^*; u)^{-1} dG(u)$; (A.4), together with (A.2), indicates that $\tilde{\Lambda}_n$ converges uniformly to Λ_0. Consequently, we have, uniformly in t,

$$\frac{d\hat{\Lambda}_n(t)}{d\tilde{\Lambda}_n(t)} = \frac{W_n(\beta_0, \Lambda_0; t)}{W_n(\hat{\beta}_n, \hat{\Lambda}_n; t)} = \frac{W(\beta_0, \Lambda_0; t)}{W(\hat{\beta}_n, \hat{\Lambda}_n; t)} + o(1) = \frac{W(\beta_0, \Lambda_0; t)}{W(\beta^*, \Lambda^*; t)} + o(1),$$

where the limit($= d\Lambda^*(t)/d\Lambda_0(t)$) is bounded and bounded away from 0. The second equality of the preceding display follows from the fact that the class $\{W_1(\beta, \Lambda; t) | \beta \in \mathcal{B}, \Lambda \in \mathcal{L}_c, t \in [0, \tau]\}$ is a Donsker class and the last equality follows by the dominated convergence theorem. We note the expectation in the preceding display is taken only over (T, δ, Z) with parameter estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$ substituted after taking the expectation. The expectations in (A.3), (A.4) and the displays below in this proof are to be understood in the same manner.

Because the class of uniformly bounded functions of bounded variations and the class of functions $g(\beta, \Lambda) \equiv \log \int e^{\delta\beta^T x} \exp(-\Lambda(T) e^{\beta^T x}) w(x|Z) m(dx)$ indexed by $(\beta, \Lambda) \in \mathcal{B} \times \mathcal{L}_c$ are Glivenko–Cantelli, we can apply Lemma A.2 again to show that,

$$P_n \left[\log \frac{d\hat{\Lambda}_n(T)}{d\tilde{\Lambda}_n(T)} \right]^{\delta} \to P_0 \left[\log \frac{d\Lambda^*(T)}{d\Lambda_0(T)} \right]^{\delta},$$

and

$$P_n \left[g(\hat{\beta}_n, \hat{\Lambda}_n) - g(\beta_0, \tilde{\Lambda}_n) \right] \to P_0 \left[g(\beta^*, \Lambda^*) - g(\beta_0, \Lambda_0) \right].$$

Hence

$$\begin{aligned} \frac{1}{n} \log \frac{L_n(\hat{\beta}_n, \hat{\Lambda}_n)}{L_n(\beta_0, \tilde{\Lambda}_n)} &= P_n \left[\log \frac{d\hat{\Lambda}_n(T)}{d\tilde{\Lambda}_n(T)} \right]^{\delta} + P_n \left[g(\hat{\beta}_n, \hat{\Lambda}_n) - g(\beta_0, \tilde{\Lambda}_n) \right] \\ &\to P_0 \left[\log \frac{L(\beta^*, \Lambda^*)}{L(\beta_0, \Lambda_0)} \right]. \end{aligned}$$

The fact that $(\hat{\beta}_n, \hat{\Lambda}_n)$ maximizes $L_n(\beta, \Lambda)$ entails $P_0 \log[L(\beta^*, \Lambda^*)/L(\beta_0, \Lambda_0)] \geq 0$. Using Jensen's inequality and the identifiability of the model, we get $(\beta^*, \Lambda^*) = (\beta_0, \Lambda_0)$. This completes the proof. □

A.2: Asymptotic normality of SPMLE

We shall prove the asymptotic normality of the SPMLE (Theorem A.1) by verifying the conditions in Theorem 19.26 of van der Vaart (1998). The proof of Theorem A.1 concerns many useful results for the profile likelihood theory, discussed in Murphy and van der Vaart (2000).

Let $H_p = \{\mathbf{h} = (h_1, h_2) \in \Re^d \times BV \mid \|h_1\| + \|h_2\|_v \leq p\}$, where $\|\cdot\|_v$ is the variation norm. If $p = \infty$, then the previous inequality is strict. We consider the parameter (β, Λ) as a functional on H_p given by $(\beta, \Lambda)(\mathbf{h}) = h_1^T\beta + \int_0^\tau h_2 d\Lambda$ and the parameter space of (β, Λ) as a subset of $\ell^\infty(H_p)$, the space of all bounded functionals on H_p, equipped with the supremum norm $||(\beta, \Lambda)||_p = \sup_{\mathbf{h}\in H_p} |(\beta, \Lambda)(\mathbf{h})|$. We note that

$$(p/\sqrt{d})(\|\beta\| \vee \|\Lambda\|_*) \leq \|(\beta, \Lambda)\|_p \leq 2p(\|\beta\| \vee \|\Lambda\|_*), \tag{A.5}$$

where $\|\Lambda\|_* = \sup_{\|h\|_v \leq 1} |\int_0^\tau h\, d\Lambda|$.

Theorem A.1 $\sqrt{n}((\hat{\beta}_n, \hat{\Lambda}_n) - (\beta_0, \Lambda_0))$ *converges weakly to a tight Gaussian process* $\mathcal{G}$ *on* $\ell^\infty(H_p)$ *with mean zero and covariance process*

$$Cov(\mathcal{G}(\mathbf{h}), \mathcal{G}(\mathbf{h}')) = h_1^T\tilde{\sigma}_1(\mathbf{h}') + \int_0^\tau h_2\tilde{\sigma}_2(\mathbf{h}')d\Lambda_0, \tag{A.6}$$

where $(\tilde{\sigma}_1, \tilde{\sigma}_2) = \tilde{\sigma}$ *is the inverse of some continuously invertible linear operator* σ *from* H_∞ *onto* H_∞.

Proof It is known that the class of uniformly bounded functions of bounded variations is Donsker. Applying Theorem 2.10.6 of van der Vaart and Wellner (1996), we can prove that the class of functions

$$\left\{\ell(\beta, \Lambda)[\mathbf{h}] \;:\; \|(\beta, \Lambda) - (\beta_0, \Lambda_0)\|_p < \delta, \mathbf{h} \in H_p\right\} \tag{A.7}$$

is uniformly bounded Donsker for some $\delta > 0$. In particular, we have $\sqrt{n}(P_n - P_0)\ell(\beta_0, \Lambda_0)$ converges weakly to a tight mean zero Gaussian process $\mathcal{W}$ on $\ell^\infty(H_p)$ for every $0 < p < \infty$. Applying the first order Taylor expansion of $\ell(\beta, \Lambda)[\mathbf{h}]$ (T, δ, Z), as a function of $(\beta, \Lambda(T), \int_0^T h_2 d\Lambda)$ and using (A.5), we have

$$\begin{aligned}
&\ell(\beta, \Lambda)[\mathbf{h}](T, \delta, Z) - \ell(\beta_0, \Lambda_0)[\mathbf{h}](T, \delta, Z)\\
&\quad = \eta\left(\beta - \beta_0, h_1, (\Lambda - \Lambda_0)(T), \int_0^T h_2 d(\Lambda - \Lambda_0)\right)\\
&\qquad + O\left(\|\beta - \beta_0\|^2 + |(\Lambda - \Lambda_0)(T)|^2 + |\int_0^T h_2 d(\Lambda - \Lambda_0)|^2\right)
\end{aligned}$$

$$= \eta\left(\beta-\beta_0, h_1, (\Lambda-\Lambda_0)(T), \int_0^T h_2 d(\Lambda-\Lambda_0)\right)$$
$$+o\left(||(\beta-\beta_0, \Lambda-\Lambda_0)||_p\right), \tag{A.8}$$

for some multi-linear function η. This gives

$$\sup_{\mathbf{h}\in H_p} P_0(\ell(\beta,\Lambda)[\mathbf{h}]-\ell(\beta_0,\Lambda_0)[\mathbf{h}])^2 \to 0, \quad \text{as } (\beta,\Lambda)\to(\beta_0,\Lambda_0). \tag{A.9}$$

Again by (A.8), we have

$$P_0\left(\ell(\beta_0+\beta,\Lambda_0+\Lambda)[\mathbf{h}]-\ell(\beta_0,\Lambda_0)[\mathbf{h}]\right) = -\left[\sigma_1(\mathbf{h})^T\beta+\int_0^\tau \sigma_2(\mathbf{h})d\Lambda\right]$$
$$+R(\beta,\Lambda)(\mathbf{h}),$$

for some continuous linear operator $\sigma=(\sigma_1,\sigma_2)$ form H_∞ to H_∞; and the remainder term $R(\beta,\Lambda)$ satisfying $\|R(\beta,\Lambda)\|_p = o(\|(\beta,\Lambda)\|_p)$. This indicates that $(\beta,\Lambda)\mapsto P_0\ell(\beta,\Lambda)$ is Fréchet differentiable at (β_0,Λ_0) with derivative $V : lin(\mathcal{B}\times\mathcal{L})\mapsto \ell^\infty(H_p)$ given by

$$V(\beta,\Lambda)(\mathbf{h}) = -\left[\sigma_1(\mathbf{h})^T\beta+\int_0^\tau \sigma_2(\mathbf{h})d\Lambda\right]. \tag{A.10}$$

The operator σ is called the information operator. By considering (A.10) and the negative second directional derivative of the log-likelihood specified by the parametric submodel $(\epsilon_1,\epsilon_2)\mapsto(\beta_0+\epsilon_1 h_1+\epsilon_2 h_1', \Lambda_0+\epsilon_1\int_0^\cdot h_2 d\Lambda_0+\epsilon_2\int_0^\cdot h_2' d\Lambda_0)$ for ϵ_1,ϵ_2 near 0, we have

$$P_0\left(\ell(\beta_0,\Lambda_0)[\mathbf{h}]\ell(\beta_0,\Lambda_0)[\mathbf{h}']\right) = h_1'^T\sigma_1(\mathbf{h})+\int_0^\tau \sigma_2(\mathbf{h})h_2' d\Lambda_0, \tag{A.11}$$

for $\mathbf{h}=(h_1,h_2)$ and $\mathbf{h}'=(h_1',h_2')$ in H_∞. We note that (A.11) is the covariance process of $\mathcal{W}$.

The continuous invertibility of V for some p is equivalent to the fact that

$$\inf_{(\beta,\Lambda)\in lin(\mathcal{B}\times\mathcal{L})} \frac{\|V(\beta,\Lambda)\|_p}{\|(\beta,\Lambda)\|_p} > \epsilon, \tag{A.12}$$

for some $\epsilon>0$ (see, for example, Bickel et al. 1993, p. 418). To prove (A.12), we shall show that $\sigma : H_\infty \to H_\infty$ is continuously invertible, which implies that for all $p>0$ there exists $r>0$ such that $\sigma^{-1}(H_r)\subset H_p$. Thus, it follows from (A.10) that $\|V(\beta,\Lambda)\|_p \geq \sup_{\mathbf{h}\in\sigma^{-1}(H_r)} |\sigma_1(\mathbf{h})^T\beta + \int_0^\tau \sigma_2(\mathbf{h})d\Lambda| = \|(\beta,\Lambda)\|_r$, which gives (A.12) with $\epsilon = r/(2p\sqrt{d})$ by (A.5).

By Theorem 4.25 in Rudin (1973), we now show that σ is one-to-one and can be written as the sum of a continuously invertible operator and a compact operator to prove that σ is continuously invertible.

Suppose that $\sigma(\mathbf{h}) = 0$, then $\ell(\beta_0,\Lambda_0)[\mathbf{h}]$=0, $a.s.$ by (A.11). A similar argument as in the proof of Lemma A.1 gives h_1=0 and h_2=0 $a.s.$ $[\Lambda_0]$. Putting h_1=0 and h_2=0 $a.s.$ $[\Lambda_0]$ in $\sigma_2(\mathbf{h})$=0, we obtain from (A.13) below that $h_2(u)W(\beta_0,\Lambda_0;u) = 0$ for $u \in [0,\tau]$. Since $W(\beta_0,\Lambda_0;\cdot)$ is uniformly bounded away from zero, $h_2 = 0$ identically. Thus, σ is one to one.

Write $\sigma = A + K$ where $A(h_1,h_2) = (h_1, h_2(\cdot)W(\beta_0,\Lambda_0;\cdot))$. Since W is bounded away from 0, A is continuously invertible. Let $A = (A_1,A_2)$ and $K = (K_1,K_2)$. Since a bounded linear operator with finite-dimensional range is compact, we need only show that K_2 is compact on $\{(0,h_2)|h_2 \in BV\}$ to obtain the compactness of K. Observing from (A.10) that $-\int_0^\tau \sigma_2(0,h_2)d\Lambda = V(0,\Lambda)(0,h_2) = \frac{d}{d\varepsilon}\big|_{\varepsilon=0} P_0\ell(\beta_0,\Lambda_0 + \varepsilon\Lambda)[(0,h_2)]$, we can show that

$$\begin{aligned}\sigma_2(0,h_2)(u) = h_2(u)W(\beta_0,\Lambda_0;u) - P_0\Bigg\{&\left[I(u\leq T)\int_0^\tau h_2 d\Lambda_0\right]\\ &\times\left[\frac{\int e^{\beta_0^T x(\delta+2)}\exp(-\Lambda_0(T)e^{\beta_0^T x})w(x|Z)m(dx)}{\int e^{\beta_0^T x\delta}\exp(-\Lambda_0(T)e^{\beta_0^T x})w(x|Z)m(dx)}\right.\\ &\left.-\frac{\left(\int e^{\beta_0^T x(\delta+1)}\exp(-\Lambda_0(T)e^{\beta_0^T x})w(x|Z)m(dx)\right)^2}{\left(\int e^{\beta_0^T x\delta}\exp(-\Lambda_0(T)e^{\beta_0^T x})w(x|Z)m(dx)\right)^2}\right]\Bigg\},\end{aligned} \tag{A.13}$$

which implies $\|K_2(0,h_2)\|_v = \|(\sigma_2 - A_2)(0,h_2)\|_v \leq c\int_0^\tau |h_2|d\Lambda_0$ for some $c > 0$. By Helly's lemma, K_2 is compact on $\{(0,h_2)|h_2 \in BV\}$.

Since (A.7), (A.9), (A.10), (A.12), and Theorem 2 combined indicate that all the conditions in the Theorem 19.26 in van der Vaart (1998) are satisfied, we obtain

$$V\sqrt{n}\left((\hat{\beta}_n,\hat{\Lambda}_n) - (\beta_0,\Lambda_0)\right) = -\sqrt{n}P_n\ell(\beta_0,\Lambda_0) + o_P(1), \tag{A.14}$$

which implies that $\sqrt{n}((\hat{\beta}_n,\hat{\Lambda}_n) - (\beta_0,\Lambda_0))$ converges weakly to a tight mean zero Gaussian process $\mathcal{G} = -V^{-1}\mathcal{W}$ on $\ell^\infty(H_p)$. To calculate the asymptotic variance, we use (A.10) and (A.14) to obtain

$$\sigma_1(\mathbf{h})^T\left(\sqrt{n}(\hat{\beta}_n - \beta_0)\right) + \int_0^\tau \sigma_2(\mathbf{h})d\left(\sqrt{n}(\hat{\Lambda}_n - \Lambda_0)\right) = \sqrt{n}P_n\ell(\beta_0,\Lambda_0)[\mathbf{h}] + o_{P^*}(1).$$

Setting $\mathbf{g} = \sigma(\mathbf{h})$ and using the weak convergence of $\sqrt{n}P_n\ell(\beta_0, \Lambda_0)$, we know from (A.11) that $g_1^T(\sqrt{n}(\hat{\beta}_n - \beta_0)) + \int_0^\tau g_2 d(\sqrt{n}(\hat{\Lambda}_n - \Lambda_0))$ has asymptotic variance $g_1^T\tilde{\sigma}_1(\mathbf{g}) + \int_0^\tau g_2\tilde{\sigma}_2(\mathbf{g})d\Lambda_0$, which gives (A.6) immediately. This completes the proof. □

A.3: Profile likelihood theory

In this subsection, we focus our attention on the estimation of β and present the efficient score function, the least favorable submodel and finally the profile likelihood theory. Theorems 3–5 can be obtained immediately from the profile likelihood theory.

Define the $d \times d$ matrix $\Sigma^{-1} = (\tilde{\sigma}_1(e_1, 0), \ldots, \tilde{\sigma}_1(e_d, 0))^T$, where e_i is the d-dimensional unit vector with the ith component is 1. The matrix is symmetric and positive definite. Let $\tilde{l}_0 = \ell_1(\beta_0, \Lambda_0) - \ell_2(\beta_0, \Lambda_0)[g^*]$, where

$$g^* = -\Sigma \begin{pmatrix} \tilde{\sigma}_2(e_1, 0) \\ \vdots \\ \tilde{\sigma}_2(e_d, 0) \end{pmatrix}$$

It can be shown that $\tilde{\ell_0}$ is the efficient score function for estimating β and Σ is equal to the covariance matrix of $\tilde{\ell_0}$. For fixed (β, Λ) and $\gamma \in \Re^d$, define

$$\Lambda_\gamma(\beta, \Lambda)(t) = \int_0^t \left[1 + (\beta - \gamma)^T g^*\right] d\Lambda. \tag{A.15}$$

We introduce the least favorable submodel specified by the log-likelihood $\gamma \mapsto l(\gamma, \beta, \Lambda) \equiv \log L(\gamma, \Lambda_\gamma(\beta, \Lambda))$. Denote by $\dot{l}$ and $\ddot{l}$ the first and second derivatives of l in γ, respectively. Noting that $\dot{l}(\gamma, \beta, \Lambda) = \ell_1(\gamma, \Lambda_\gamma(\beta, \Lambda)) - \ell_2(\gamma, \Lambda_\gamma(\beta, \Lambda))[g^*]$, we have

$$\dot{l}(\beta_0, \beta_0, \Lambda_0) = \tilde{l}_0. \tag{A.16}$$

Furthermore, noting that $\gamma \mapsto l(\gamma, \beta_0, \Lambda_0)$ is a smooth parametric submodel, its derivatives at β_0 satisfies $P_0\ddot{l}(\beta_0, \beta_0, \Lambda_0) = -P_0\dot{l}^T(\beta_0, \beta_0, \Lambda_0)\dot{l}(\beta_0, \beta_0, \Lambda_0) = -\Sigma$.

We shall establish the asymptotic expansion of the profile likelihood by verifying the conditions in Theorem 1 of Murphy and van der Vaart (2000).

Theorem A.2 *For every random sequence β_n that converges to β_0 in probability,*

$$\log pL_n(\beta_n) - \log pL_n(\beta_0) = n(\beta_n - \beta_0)^T P_n\tilde{l}_0 - \frac{1}{2}n(\beta_n - \beta_0)^T \Sigma(\beta_n - \beta_0) + o_{P_0}\left(\sqrt{n}\|\beta_n - \beta_0\| + 1\right)^2.$$

Proof Using the similar argument as in the proof of Theorem 2, we can show that

$$\hat{\Lambda}_{n,\beta_n} \xrightarrow{P} \Lambda_0. \tag{A.17}$$

We note that the "no-bias condition"

$$P_0 \dot{l}\left(\beta_0, \beta_n, \hat{\Lambda}_{n,\beta_n}\right) = o_p\left(\|\beta_n - \beta_0\| + n^{-1/2}\right), \tag{A.18}$$

is certainly satisfied if

$$\|\hat{\Lambda}_{n,\beta_n} - \Lambda_0\|_\infty = O_p\left(\|\beta_n - \beta_0\| + n^{-1/2}\right), \tag{A.19}$$

(see Murphy and van der Vaart 2000, p. 457–458). Using (A.7), (A.10), (A.12), and (A.17), we have (A.19) by Theorem 3.1, on rate of convergence, in Murphy and van der Vaart (1999).

It is clear that the functions $(\gamma, \beta, \Lambda) \mapsto \dot{l}(\gamma, \beta, \Lambda)$ and $(\gamma, \beta, \Lambda) \mapsto \ddot{l}(\gamma, \beta, \Lambda)$ are continuous at $(\beta_0, \beta_0, \Lambda)$ almost surely. By the Donsker class argument, we can show that there exists a neighborhood $\mathcal{V}$ of $(\beta_0, \beta_0, \Lambda)$ such that $\{\dot{l}(\gamma, \beta, \Lambda) | (\gamma, \beta, \Lambda) \in \mathcal{V}\}$ is a uniformly bounded Donsker class, and $\{\ddot{l}(\gamma, \beta, \Lambda) | (\gamma, \beta, \Lambda) \in \mathcal{V}\}$ is a uniformly bounded Glivenko–Cantelli class. In view of (A.15)–(A.18), we know all the conditions in Theorem 1 of Murphy and van der Vaart (2000) are satisfied. Thus, the proof is complete. □

Using the consistency of $\hat{\beta}_n$ given in Theorem 2, the invertibility of the efficient Fisher information matrix Σ, and the second order expansion of the profile likelihood, we obtain Theorems 3–5 immediately from the profile likelihood theory of Murphy and van der Vaart (2000).

Acknowledgments The author gratefully acknowledge the editor and two referees for their insightful comments and suggestions that substaintially improved the paper.

References

Bickel PJ, Klaassen C, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. The Johns Hopkins University Press, Baltimore

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

Cheng SC, Wang N (2001) Linear transformation models for failure time data with covariate measurement error. J Am Stat Assoc 96:706–716

Henderson R, Diggle P, Dobson A (2000) Joint modelling of longitudinal measurements and event time data. Biostatistics 4:465–480

Hu C, Lin DY (2002) Cox regression with covariate measurement error. Scand J Stat 29:637–656

Hu P, Tsiatis AA, Davidian M (1998) Estimating the parameters in the Cox model when covariate variables are measured with error. Biometrics 54:1407–1419

Huang Y, Wang CY (2000) Cox regression with accurate covariate unascertainable: a semiparametric correction approach. J Am Stat Assoc 95:1209–1219

Kong FH, Gu M (1999) Consistent estimation in Cox proportional hazards model with covariates measurement errors. Stat Sin 9:953–969

Kosorok M, Lee B, Fine J (2004) Robust inference for proportional hazards univariate frailty regression models. Ann Stat 32:1448–1491

Li Y, Ryan L (2006) Inference on survival data with covariate measure error—an imputation-based approach. Scand J Stat 33:169–190

Murphy SA (1994) Consistency in a proportional hazards model incorporating a random effect. Ann Stat 22:712–731

Murphy SA, van der Vaart AW (1999) Observed information in semi-parametric models. Bernoulli 5:381–412

Murphy SA, van der Vaart AW (2000) On profile likelihood. J Am Stat Assoc 95:449–465

Nakamura T (1992) Proportional hazards model with covariates subject to measurement error. Biometrics 48:829–838

Parner E (1998) Asymptotic theory for the correlated gamma-frailty model. Ann Stat 26:183–214

Prentice RL (1982) Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika 69:331–342

Rudin W (1973) Functional analysis. McGraw-Hill, New York

Rudin W (1976) Principles of mathematical analysis. McGraw-Hill, New York

Song X, Davidian M, Tsiatis AA (2002) An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. Biostatistics 58:742–753

Song X, Huang Y (2005) On corrected score approach for proportional hazards model with covariate measurement error. Biometrics 61:702–714

Tsiatis AA, Davidian M (2001) A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. Biometrika 88:447–458

van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge

van der Vaart AW, Wellner JA (1996) Weak convergence and empirical processes. Springer, New York

Wang CY, Hsu L, Feng ZD, Prentice RL (1997) Regression calibration in failure time regression. Biometrics 53:131–145

Xie SX, Wang CY, Prentice RL (2001) A risk set calibration method for failure time regression by using a covariate reliability sample. J Roy Stat Soc Ser B 63:855–870

Zeng D, Lin DY (2007) Maximum likelihood estimation in semiparametric regression models with censored data. J Roy Stat Soc Ser B 69:507–564

Zhong M, Sen PK, Cai J (1996) Cox regression model with mismeasured covariates or missing covariate. In: ASA proceedings of the biometrics section, pp 323–328

Zhou H, Pepe MS (1995) Auxiliary covariate data in failure time regression analysis. Biometrika 82: 130–149

Zhou H, Wang CY (2000) Failure time regression analysis with a continuous auxiliary variable. J Roy Stat Soc Ser B 62:657–665

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