Failure Time Regression Analysis With A Continuous Auxiliary Variable

Haibo Zhou

Department of Biostatistics,

University of North Carolina at Chapel Hill,

Chapel Hill, NC 27599-7400, U.S.A.

zhou@bios.unc.edu

C. Y. Wang

Fred Hutchinson Cancer Research Center, Seattle, U.S.A.

SUMMARY

We consider failure time regression analysis with an auxiliary variable in the presence of a validation sample. We extend the non-parametric inference procedure of Zhou and Pepe (1995) to handle a continuous auxiliary or proxy covariate. We estimate the induced relative risk function with a kernel smoother and allow the selection probability of the validation set to depend on the observed covariates. We present some asymptotic properties for the kernel estimator and provide some simulation results. The proposed method is illustrated with a data set from an on going epidemiologic study.

Keyword: Estimated partial likelihood; Kernel smoother; Martingale; Surrogate/mismeasured covariate; Survival; Validation sample.

1. Introduction

The problem of auxiliary information arises frequently in epidemiologic studies in which the exposure assessment often relies on the study subjects' responses to some questionnaires. The value of exposure collected using the questionnaire method (W) is known to be subject to substantial error and, whenever affordable, investigators often measure some biological markers (X) in addition to conduct a questionnaire survey. However, due to the cost of the measuring these biological markers, it is usually only a subset of the study cohort has the X values measured. X is sometimes referred to as the "gold standard" for the exposure measurement and W as an auxiliary or proxy to X. The subsample where $\{X, W\}$ is measured is referred to as the validation set or the calibration sub-study. In a study on iron intake in relation to preterm delivery (Savitz et al. 1999), X is the Serum ferritin concentration, a biological marker that reflects iron stores measured using an immunoratiometric assay, W is a crude score on iron intake measured from self-administered food frequency questionnaires. The food frequency questionnaire is conducted on all subjects but the serum assay is only conducted for a random subsample of the study subjects. Statistical methods that achieve greater efficiency by utilizing the available auxiliary information are important to the success of such epidemiologic studies.

For censored failure time data with measurement error in covariates, Prentice (1982) introduced a partial likelihood estimator based on the induced relative risk function after correcting for covariate measurement error under a "rare disease" assumption. This approach is further developed by Pepe, Self and Prentice (1989) using a parametric modeling. Lin and Ying (1993) proposed an approach in the pure missing data context. Hughes (1993) showed that ignoring measurement error may cause serious bias problems. Lipsitz and Ibrahim (1996) considered an EM approach for incomplete categorical covariates. Zhou and Pepe (1995) proposed an estimated partial likelihood approach for discrete auxiliary covariates which does not require parametric assumptions on either the frequency of events or the underlying distributions of covariates. In this article, we extend the approach of Zhou and Pepe (1995) using a kernel smoother based estimated partial likelihood to handle a continuous auxiliary variable.

To fix notation, suppose there are n independent individuals in a study cohort. For individual i, let $\{X_i(t), Z_i(t)\}$ denote the covariate vector at time t. Assume that $X_i(t)$ is the covariate that is only observed in the validation set and $Z_i(t)$ is the remaining covariate vector that is always being observed. Let W denote the auxiliary variable for X. In the study on iron intake in relation to preterm delivery (Savitz et al. 1997), X is the serum ferritin, W is the iron intake score measured from the self reported diet information, and Z is the remaining covariate vector that includes demographic information and other potential confounding factors. Let η_i be an indicator variable with $\eta_i = 1$ indicating that the ith individual is in the validation set. Then the observed data for individual i is $\{S_i, \delta_i, X_i(\cdot), Z_i(\cdot), W_i(\cdot)\}$, if $\eta_i = 1$, and $\{S_i, \delta_i, Z_i(\cdot), W_i(\cdot)\}$, if $\eta_i = 0$, where S_i is the observed event time, which is the minimum of the potential failure time T_i and the censoring time C_i, and δ_i is the indicator of failure. The validation set V is the set of indexes such that $V := \{i : \eta_i = 1\}$. Following the proportional hazards regression model (Cox, 1972), we postulate the hazard function of failure as

$$\begin{aligned}\lambda\{t; X_i(t), Z_i(t), W_i(t)\} &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathcal{P}\{t \le T_i < t + \Delta t | T_i \ge t, X_i(t), Z_i(t), W_i(t)\} \\ &= \lambda_0(t) \exp\{\beta_1 X_i(t) + \beta_2' Z_i(t)\}, \qquad (1)\end{aligned}$$

where $\lambda_0(\cdot) \ge 0$ is the unspecified baseline hazard, and $\beta' = (\beta_1, \beta_2')$ is the relative risk parameter vector to be estimated. It is assumed that the validation set is chosen at the baseline under the ignorable missing value mechanism (Rubin, 1976). This allows the observation of X_i to depend on the observed baseline covariates $\{Z_i(0), W_i(0)\}$ but independent of $\{S_i, X_i(\cdot)\}$, *i.e.*, $\Pr(\eta_i = 1) = \pi(Z_i(0), W_i(0))$.

The rest of this paper is organized as follows. In Section 2, we propose a kernel smoother estimator for the induced relative risk and extend the estimated partial likelihood approach to continuous W. We propose a variance estimator for the proposed estimator based on the asymptotic results. We present some simulation results in Section 3 and demonstrate the proposed method in Section 4 with the iron intake data set. Concluding remarks are given in Section 5.

2. Estimated Partial Likelihood with Kernel Smoother

We first derive the induced hazard function for an individual i with only $\{Z_i(t), W_i(t)\}$ observed. Conditional on (1) and assuming that conditional on $X_i(\cdot)$, $W_i(\cdot)$ provides no additional information about the hazard of failure, we have

$$
\begin{aligned}
\lambda\{t; \lambda_0(t), Z_i(t), W_i(t)\} &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathcal{P}\{t \le T_i < t + \Delta t | T_i \ge t, Z_i(t), W_i(t)\} \\
&= \lambda_0(t) \exp\{\beta_2' Z_i(t)\} E[\exp\{\beta_1 X_i(t)\} | T_i \ge t, Z_i(t), W_i(t)]. \quad (2)
\end{aligned}
$$

where $\lambda_0(\cdot)$ above reflects the fact that the condition of $\{T_i \ge t\}$ in the expectation usually imply that λ will depend on the baseline hazard. Equations (1) and (2) imply that the relative risk functions for an individual with complete and incomplete data are $\exp\{\beta_1 X_i(t) + \beta_2' Z_i(t)\}$ and $\exp\{\beta_2' Z_i(t)\} E[\exp\{\beta_1 X_i(t)\} | T_i \ge t, Z_i(t), W_i(t)]$, respectively. Under the independent censoring assumption (equation (5), Prentice, 1982) that $\lambda\{t; Z_i(t), W_i(t), \text{no censorship in } [0,t)\}$ $= \lambda\{t; Z_i(t), W_i(t)\}$ and notice that the event of {no censorship in [0,t)} is $\{C \ge t\}$, we rewrite the induced relative risk for an individual in our setup as

$$
r_i(\beta, t, \lambda_0(t)) = \exp\{\beta_1 X_i(t) + \beta_2' Z_i(t)\}\eta_i + \exp\{\beta_2' Z_i(t)\} E[\exp\{\beta_1 X_i(t)\} | S_i \ge t, Z_i(t), W_i(t)](1 - \eta_i). \quad (3)
$$

In addition to the unspecified λ_0, the expectation above also depends the underlying distributions of $X_i(t)$ and $Z_i(t)$. If $f(X_i(t) | S_i \ge t, Z_i(t), W_i(t))$ is a known function up to a parameter θ, then the inference about β and θ can be drawn from the usual partial likelihood based on the general relative function $r_i(\beta, \theta, t)$ (Andersen and Gill, 1982; Prentice and Self, 1983). To avoid making the undesirable parametric assumptions on λ_0 and the distributions of covariates, Zhou and Pepe used an empirical relative risk estimator for those in the non-validation set when W is discrete. We extend their approach to continuous W by estimating the induced relative risk function with a kernel smoother and then estimate β from the resulting estimated partial likelihood function.

2.1. Kernel Estimation of the Relative Risk Function

Kernel smoothing is a useful nonparametric tool to estimate a regression function, *i.e.* conditional expectation of the outcome variable given a covariate variable. Assume (X_1, Y_1), $\ldots, (X_n, Y_n)$ are n pairs of observations for an explanatory variable X and an outcome variable Y, and $E(Y_i|X_i = x) = m(x)$ where $m(\cdot)$ is a smooth function. Then the kernel smoother of $m(x)$ based on Nadaraya (1964) and Watson (1964) can be expressed as $\sum_{i=1}^n Y_i K_h(X_i - x)/\{\sum_{i=1}^n K_h(X_i - x)\}$, where $K_h(\cdot) = K(\cdot/h)$ and $h > 0$ is the bandwidth. The function $K(\cdot)$ is called the kernel function which is a piecewise smooth function satisfying $\int K(u)du = 1$. The basic idea of kernel estimation is that it estimates $m(x)$ with a weighted mean in a neighborhood around x that is defined by the bandwidth h. We say that a kernel function $K(\cdot)$ has an order of α if $\int u^q K(u)du = 0$ for $q = 1, \ldots, (\alpha - 1)$, $\int u^\alpha K(u)du \neq 0$ and $\int K^2(u)du < \infty$. Further details on kernel smoother can be found in Eubank (1988, Chapter 4).

For simplicity, we denote $\Gamma(\cdot) = (Z'(\cdot), W'(\cdot))'$, d be the dimension of Γ, $\gamma_i(\beta, t) = \exp\{\beta_1 X_i(t) + \beta_2' Z_i(t)\}$ and $\phi_i(\beta, t) = \exp\{\beta_2' Z_i(t)\} E[\exp\{\beta_1 X_i(t)\}|S_i \geq t, Z_i(t), W_i(t)]$. Hence $r_i(\beta, t)$ in (3) can be written as $\gamma_i(\beta, t)\eta_i + \phi_i(\beta, t)(1 - \eta_i)$. Let K be a kernel function of order α. Then the kernel estimator for $\phi(\beta, t)$ based on Nadaraya (1964) and Watson (1964) is defined as

$$\hat{\phi}_i(\beta, t) = \frac{\sum_{j=1}^{n_V} I_{[S_j \geq t]} K_h(\Gamma_i(t) - \Gamma_j(t))\gamma_j(\beta, t)/n_V}{\sum_{j=1}^{n_V} I_{[S_j \geq t]} K_h(\Gamma_i(t) - \Gamma_j(t))/n_V}, \tag{4}$$

where $n_V = \sum_{i=1}^n \eta_i$ is the sample size of the validation set and h is the bandwidth. When the validation sample is a simple random sample of the full cohort, the denominator in (4) can be further improved with the one that is based on the full cohort data. We estimate β with $\hat{\beta}_{EPL}$ which is the maximizer of the following estimated partial likelihood function:

$$EPL(\beta) = \prod_{i=1}^n \left[\frac{\hat{r}_i(\beta, S_i)}{\sum_{j \in \mathcal{R}(S_i)} \hat{r}_j(\beta, S_i)}\right]^{\delta_i}, \tag{5}$$

where $\hat{r}_i(\beta, t) = \gamma_i(\beta, t)\eta_i + \hat{\phi}_i(\beta, t)(1 - \eta_i)$ is the estimated relative risk function for individual i and $\mathcal{R}(S_i)$ is the risk set at time $S_i - 0$.

In small sample situation, the kernel relative risk estimator $\hat{r}_i(\beta, t)$ is not defined when the denominator in (4) is 0, which occurs when the risk set at time t in the validation set is the null set. This could happen in the later part of the time interval being studied. In this case, we could either (i) do estimation without using these points; or (ii) use imputation of the relative risk at these points by interpolation based on neighboring points of $\Gamma_i(\cdot)$. The efficiency loss associated with (i) will be minimal if these points consist of a small portion of the data. Method (ii) is based on the idea that the induced relative risk is a continuous function of $\Gamma_i(\cdot)$. We find that either methods worked reasonably well in terms of mean square error in our simulation studies.

2.2. Asymptotic Results And the Proposed Variance Estimator

The large sample theory for kernel smoother based partial likelihood estimator can be established following closely to the argument of Zhou and Pepe (1995). We present the main asymptotic result and the full proofs under the kernel smoother setting can be found in Zhou and Wang (1998). Recall that $\gamma(\beta,t) = \exp\{\beta_1 X(t) + \beta_2' Z(t)\}$, $\phi(\beta,t) = \exp\{\beta_2' Z(t)\} E[\exp\{\beta_1 X_i(t)\} | S_i \geq t, Z_i(t), W_i(t)]$ and $r_i(\beta,t) = \gamma(\beta,t)\eta_i + \phi(\beta,t)(1-\eta_i)$. Define $s^{(0)}(\beta,t) = E\{Y_i(t) r_i(\beta,t)\}$, $s^{(1)}(\beta,t) = (\partial/\partial\beta) s^{(0)}(\beta,t)$, $(\partial/\partial\beta) s^{(1)}(\beta,t)$, $s^{(2)}(\beta,t) = E[Y_i(t)\{r_i^{(1)}(\beta,t) r_i(\beta,t)^{-1}\}^{\otimes 2} r_i(\beta_0,t)]$, where $Y_i(t) = I_{[S_i \geq t]}$ is the at risk indicator and $a^{\otimes 2}$ denotes matrix aa'. Define $N_i(t) = I_{[S_i<t,\delta_i=1]}$ be a counting process and $M_i(t) = N_i(t) - \int_0^t Y_i(\omega) r(\beta,\omega)\lambda_0(\omega) d\omega$ be a martingale (Fleming and Harrington, 1991). We only consider a finite time interval and assume t is in $[0,1]$. The following theorem summarizes the asymptotic properties of the estimator $\hat\beta_{EPL}$ that maximizes (5).

Theorem 2.1 Under the bandwidth condition that $nh^{4\alpha} \to 0$ and $nh^{2d}\{\log(n)\}^{-2} \to \infty$, $\hat\beta_{EPL}$ is a consistent estimator of β, and $\sqrt{n}(\hat\beta_{EPL} - \beta)$ is asymptotically normally distributed with mean 0 and asymptotic covariance $\Sigma_{EPL}(\beta)$, where $\Sigma_{EPL}(\beta) = I^{-1}(\beta)\{(1-\rho)\Sigma_1(\beta) + \rho\Sigma_2(\beta)\}(I^{-1}(\beta))'$, $\rho = \lim_{n\to\infty} n_V/n$, and $I(\beta) = \int_0^1 \{s^{(2)}(\beta,\omega)/s^{(0)}(\beta,\omega) - [s^{(1)}(\beta,\omega)/s^{(0)}(\beta,\omega)]^{\otimes 2}\}$ $s^{(0)}(\beta,\omega)\lambda_0(\omega)d\omega$, $\Sigma_1(\beta) = \int_0^1 \{E[Y(\omega)\phi(\beta,\omega)^{\otimes 2}/\phi(\beta,\omega)] - s^{(1)}(\beta,\omega)^{\otimes 2}/s^{(0)}(\beta,\omega)\}\lambda_0(\omega)d\omega$, and $\Sigma_2(\beta) = E\{\int_0^1 [\gamma^{(1)}(\beta,\omega)/\gamma(\beta,\omega) - s^{(1)}(\beta,\omega)/s^{(0)}(\beta,\omega)] dM(\omega) - (1-\rho)/\rho [\int_0^1 (\phi^{(1)}(\beta,\omega)/\phi(\beta,\omega)$ $- s^{(1)}(\beta,\omega)/s^{(0)}(\beta,\omega)) Y(\omega)(\gamma(\beta,\omega) - \phi(\beta,\omega))\lambda_0(\omega)d\omega]\}^{\otimes 2}$.

A consistent variance estimator for $\hat\beta_{EPL}$ can be found by replacing the population quantities in the asymptotic covariance matrix $\Sigma_{EPL}(\beta)$ with their corresponding sample quantities.

For the cumulative hazard $\Lambda_0(t) = \int_0^t \lambda_0(\omega)d\omega$, we follow the idea of Breslow (1974) and use

$$d\tilde{\Lambda}_0(t) = \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n Y_i(t)\hat{r}_i(\hat{\beta}_{EPL}, t)} = \frac{1}{\hat{S}^{(0)}(\hat{\beta}_{EPL}, t)} \frac{1}{n} \sum_{i=1}^n dN_i(t).$$

3. Simulation studies

We examine the small sample behavior of $\hat{\beta}_{EPL}$ using Monte-Carlo studies. The covariate X is assumed to follow a uniform distribution on [0,2]. The failure time is generated from a Weibull distribution with hazard function given by $\lambda(t; X) = \kappa p t^{p-1} exp[\beta X]$. The Weibull distribution reduces to the exponential distribution when $p = 1$. The auxiliary variable W is defined as $W = X + e$, where $e \sim N(0, \sigma^2)$ is a random error and σ^2 is the parameter that controls the strength of the association between X and W. In the motivating epidemiologic study on iron intake and preterm delivery, σ^2 is the measurement error associated with the self reported dietary information. We assume an uniform distributed censoring variable that is independent of the failure time variable. The selection probability for the validation set follows the model

$$\text{pr}(\eta_i = 1|W_i) = \{1 + \exp(1 - W_i)\}^{-1},$$

which satisfies the ignorable missing value mechanism (Rubin, 1976). The validation fraction is chosen at 50 %. We apply the Epanechnikov kernel function (Eubank, 1988, Chapter 4) with the bandwidth $h = 2\hat{\sigma}_W n^{-1/3}$, where $\hat{\sigma}_W$ is the sample standard deviation of W, which satisfies the bandwidth conditions in Theorem 2.1. A boundary kernel is used to avoid possible boundary bias problems. We use method (ii) described in Section 2 for finite sample correction.

The parameter values in our studies are $p = 1$ or 2, $\kappa = 1$, $\beta = ln(2)$, and $\sigma = 0.2$ or 0.8. The sample sizes studied are $n = 100$ and 300. Since the results for $p = 1$, which corresponds to the exponentially distributed failure times, is similar to that for $p = 2$, we only report the results for the latter case. Table 1 summarizes the results of simulation studies. The methods studied are the proposed estimated partial likelihood ($\hat{\beta}_{EPL}$), the complete-case Cox regression analysis ($\hat{\beta}_{CC}$), the Cox regression analysis with W substituted for the missing X

($\hat{\beta}_N$), and the full data Cox regression analysis ($\hat{\beta}_F$). The full data Cox regression analysis is the standard proportional hazard regression assuming X is available for all n individuals in the study. The complete-case Cox regression is data restricted to the validation set only. $\hat{\beta}_N$ is an naive estimator that uses the auxiliary W directly when X is not observed. For a given sample size, the estimated means, medians, standard errors and the 95% confidence intervals of an estimator are obtained from 500 independent runs. Under the model studied, we observe that: (i) Both the complete-case Cox regression analysis and the proposed estimated partial likelihood method are valid. The naive estimator $\hat{\beta}_N$ is biased. There is no observable bias in the means or medians of $\hat{\beta}_{EPL}$ and $\hat{\beta}_{CC}$. (ii) $\hat{\beta}_{EPL}$ is more efficient than $\hat{\beta}_{CC}$ and the efficiency gain is higher when W is more informative about X (smaller σ). For $\sigma = 0.2$ and 50% censoring, the ratios of variances of $\hat{\beta}_{EPL}$ relative to $\hat{\beta}_{CC}$ at $n = 100$ and $n = 300$ are 47% and 53% respectively. These ratios increase to 65% and 67% respectively when $\sigma = 1$. The results also show a higher efficiency gain by $\hat{\beta}_{EPL}$ relative to $\hat{\beta}_{CC}$ when the sample sizes are small. (iii) The proposed variance estimator provides good estimation of $\Sigma_{EPL}(\beta)$. The 95% confidence intervals based on the proposed standard errors provide good coverages for the cases studied. Clearly, $\widehat{\beta}_F$ is not obtainable for practical studies because of the missing X. Nevertheless, we observe in this case that when σ is small, the proposed method recovered most of the information from the non-validation set.

4. Iron Intake In Relation To Preterm Delivery Study

We apply the proposed method to the ongoing epidemiologic study of iron intake and preterm delivery (Savitz et al., 1999) conducted at the UNC hospitals at Chapel Hill. Serum ferritin concentration reflects iron stores, which are required to sustain the production of essential iron compounds. Published epidemiologic studies based on examination of low ferritin concentrations and preterm delivery through logistic regression had generated both positive association (Ulmer and Goepel, 1988) and no association (Scholl et al., 1992). We consider gestation age, the time from pregnancy to delivery, as the failure time variable and evaluate iron intake in relation to time to delivery. The full data includes 1362 individuals among which a subset of 287 individuals had their serum ferritin concentration (FERRITIN) measured using an immunoratiometric assay. For all individuals, a crude score for dietary iron intake (I-DIET) is being assessed using a quantitative food frequency questionnaire. Potential confounders are the pregravid weight status category of women (BMI) and poverty index (POVCAT). Using the notation of the proposed method, W is I-DIET, X is FERRTIN and Z includes BMI and POVERTY.

Table 3 shows the results of failure time regression analyses for the iron intake in relation to time to delivery data. The proposed estimated partial likelihood method which uses all 1362 observations is contrasted with the complete-case Cox regression analysis which uses only 272 observations. Both analyses confirm that the serum ferritin concentration is significantly related to time to delivery. The variance estimate for $\hat{\beta}_{EPL}$ is calculated based on the proposed estimator $\hat{\Sigma}_{EPL}$. For estimating the effect of ferritin concentration, there is substantial efficiency gain by using the proposed partial likelihood approach over the complete-case Cox regression analysis. The ratio of the estimated standard error for $\hat{\beta}_{EPL}$ relative to $\hat{\beta}_{CC}$ is 0.431. The 95% confidence intervals for the regression parameter of ferritin is (0.102, 0.395) for the complete-case Cox regression analysis versus that of (0.040, 0.223) for the estimated partial likelihood method. Assuming all other factors being equal, the estimated hazard ratios for ferritin concentration are 1.282 for the complete-case analysis and 1.141 for the proposed method.

5. Concluding remarks

It is well known that the kernel smoother used in the proposed method is not practical if the dimension d of $\Gamma = (Z', W')'$ is large. This is the so called curse of dimensionality. Our estimator would be useful in the case where the dimension of the vector that is correlated to the missing covariate is at most two. This would be satisfied in many epidemiologic studies where the interested exposure variable and its auxiliary variable are scalar variables. For high dimensional cases, Carroll, Knickerbocker and Wang (1995) consider the problem in the special situation that the distribution of X given Γ depends only on lower dimensional linear combinations of Γ. Although using the auxiliary information is intuitive and our simulation results show efficiency gain by the EPL method, it is possible that in some extreme cases, the EPL estimator may not improve over that from the complete-case Cox regression analysis. In practice, we recommend the use of the proposed method when the validation fraction is not too small. In some extreme situations where there is extremely small validation fraction and very large relative risk parameter, one should also check the results from the Cox regression analysis with data restricted to the validation set.

Acknowledgement

The research of the second author was supported by the US National Institutes of Health grants (AG 15026, CA 53996). The authors would like to thank Dr. D. Savitz for providing the data set. We also thank Drs. Gregg Dinse, David Dunson, Margaret Pepe, Jamie Robins, Anna Maria Siega-Riz and Michael Schell for helpful discussion and comments.

References

Andersen, P.K. and Gill, R.D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100-1120.

Breslow, N. E. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89-99.

Carroll, R. J., Knickerbocker, R. K. and Wang, C. Y. (1995), "Dimension reduction in a semiparametric regression model with errors in covariates," *The Annals of Statistics*, 90, 157-169.

Cox, D.R. (1972), "Regression Models and Life Tables (With Discussion)," *J. R. Statist. Soc., B*, 34, 187-220.

Eubank, R. L. (1988), *Smoothing Spline and Nonparametric Regression*, New York and Basel: Marcel Dekker.

Fleming, T.R. and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, Wiley, New York.

Goepel E, Ulmer HU, and Neth RD. Premature labor contractions and the value of serum ferritin during pregnancy. Gynecol Obstet Invest 1988;26:265-73.

Hughes, M. D. (1993), "Regression dilution in the proportional hazards model," *Biometrics*, 49, 1056-1066.

Kalbfleisch, J. D. and Prentice, R. L. (1980), "The statistical analysis of failure time data", New York: John Wiley.

Nadaraya, E. A. (1964), "On estimating regression," *Theory Probab. Appl.* , 10, 186–190.

Lin, D. Y. and Ying, Z. (1993), "Cox regression with incomplete covariate measurements", *Journal of the American Statistical Association*, 88, 1341-1349.

Lipsitz, S. and Ibrahim, J. G. (1996). Using the E-M algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, 2, 5-14.

Pepe, M. S., Self, S. G. and Prentice, R. L. (1989), "Further Results on Covariate Measurement Errors in Cohort Studies with Time to Response Data," *Statistics in Medicine*, 8, 1167-1178.

Prentice, R. L. (1982), "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model," *Biometrika*, 69, 331-342.

Prentice, R. L. and Self, S. G. (1983), "Asymptotic Distribution Theory for Cox-Type Regression Models with General Relative Risk Form," *The Annals of Statistics*, 11, 804-813.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63: 581-592.

Savitz DA, Dole N, Williams J, Thorp JM, McDonald T, Carter AC, Euker B. Determinants of participation in an epidemiological study of preterm delivery. Paediatric and Perinatal Epi 1999;13:114-25.

Scholl T, Hediger ML, Fischer RL, and Shearer JW. Anemia vs iron deficiency: Increased risk of preterm delivery in a prospective study. Am J Clin Nutr 1992;55:985-88.

Watson, G. S. (1964), "Smooth regression analysis," *Sankhya, Ser. A*, 26, 359-372.

Zhou, H and Pepe, M.S. (1995) "Auxiliary Covariate Data in Failure Time Regression Analysis," *Biometrika*, 82, 139-149.

Zhou, H and Wang, C. Y. (1998) "Some asymptotic results for using kernel smoother with covariate measurement error problem in survival analysis". Technical Report, Department of Biostatistics, University of North Carolina at Chapel Hill.

Table 1. Simulation results comparing $\hat{\beta}_{EPL}$, $\hat{\beta}_{CC}$, $\hat{\beta}_N$ and $\hat{\beta}_F$ under the Weibull model with the validation set selection probability $\pi(W) = \{1 + \exp(1 - W)\}^{-1}$.

			σ=.2			$\sigma = .8$		
n		$\hat{\beta}_F$	$\hat{\beta}_{CC}$	$\hat{\beta}_N$	$\hat{\beta}_{EPL}$	$\hat{\beta}_{CC}$	$\hat{\beta}_N$	$\hat{\beta}_{EPL}$
	50 % censoring							
n=100	mean	.688	.703	.652	.691	.702	.542	.640
	median	.690	.698	.642	.682	.703	.562	.642
	s.e.	.260	.401	.258	.275	.389	.364	.313
	mean $(\widehat{\text{s.e.}})$	.256	.387	.249	.266	.384	.252	.310
	95 % cov. prob.	.950	.946	.938	.934	.950	.667	.940
n=300	mean	.699	.706	.664	.700	.699	.380	.662
	median	.697	.698	.660	.691	.685	.379	.657
	s.e.	.151	.212	.147	.154	.214	.113	.175
	mean $(\widehat{\text{s.e.}})$	.147	.217	.143	.152	.216	.109	.178
	95 % cov. prob.	.940	.948	.940	.950	.958	.212	.946
	80 % censoring							
n=100	mean	.703	.791	.665	.708	.761	.381	.659
	median	.689	.726	.646	.690	.683	.367	.642
	s.e.	.435	.688	.424	.454	.671	.317	.526
	mean $(\widehat{\text{s.e.}})$	.413	.645	.400	.422	.638	.302	.492
	95 % cov. prob.	.938	.960	.944	.936	.956	.786	.930
n=300	mean	.703	.710	.671	.709	.700	.391	.670
	median	.699	.693	.668	.708	.684	.393	.651
	s.e.	.229	.348	.220	.236	.350	.177	.282
	mean $(\widehat{\text{s.e.}})$	.234	.349	.228	.241	.346	.173	.282
	95 % cov. prob.	.960	.944	.968	.954	.950	.582	.946

NOTE: The true parameter $\beta_0 = \log(2) \doteq 0.6931$, $X \sim Unif[0, 2]$, $W_i = X_i + e_i$, where $e_i \sim N(0, \sigma^2), i = 1, \ldots, n$. The bandwidth h is chosen as $h = 2\widehat{\sigma}_W n^{-1/3}$, where $\widehat{\sigma}_W$ is the sample standard deviation of W. $\hat{\beta}_{EPL}$ denotes the proposed estimator, $\hat{\beta}_{CC}$ denotes the complete-case Cox regression analysis, $\hat{\beta}_N$ replaces an unobserved X by W, and $\hat{\beta}_F$ denotes estimator obtained by using true X for all n individuals in the study. The validation fraction is 50 % and the results are based on 500 replications. The term mean $(\widehat{\text{s.e.}})$ denotes the mean of the estimated standard error for each method.

Table 2. Regression Analyses of the Iron Intake in Relationship to Time to Delivery

		Factor		
Analyses		FERRTIN	POVCAT	BMI
Complete-case				
Cox Regression	$\hat{\beta}_{CC}$	0.2485	-0.0541	-0.0099
	s.e.($\hat{\beta}_{CC}$)	0.0748	0.0734	0.0098
	95% C.I.	(0.102, 0.395)	(-0.198, 0.090)	(-0.029, 0.009)
EPL Method				
	$\hat{\beta}_{EPL}$	0.1315	0.0333	-0.0006
	s.e.($\hat{\beta}_{EPL}$)	0.0491	0.03	0.004
	95% C.I.	(0.040, 0.223)	(-0.025, 0.092)	(-0.008, 0.007)

NOTE: FERRITIN is the iron intake reflected by serum ferritin concentration, POVCAT is the poverty category and BMI is pregravid weight status of women.