OPEN ACCESS Freely available online

PLOS ONE

Buckley-James Estimator of AFT Models with Auxiliary Covariates

Kevin Granville[1], **Zhaozhi Fan**[2]*

1 Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, **2** Department of Mathematics and Statistics, Memorial University, St. John's, Newfoundland, Canada

Abstract

In this paper we study the Buckley-James estimator of accelerated failure time models with auxiliary covariates. Instead of postulating distributional assumptions on the auxiliary covariates, we use a local polynomial approximation method to accommodate them into the Buckley-James estimating equations. The regression parameters are obtained iteratively by minimizing a consecutive distance of the estimates. Asymptotic properties of the proposed estimator are investigated. Simulation studies show that the efficiency gain of using auxiliary information is remarkable when compared to just using the validation sample. The method is applied to the PBC data from the Mayo Clinic trial in primary biliary cirrhosis as an illustration.

Citation: Granville K, Fan Z (2014) Buckley-James Estimator of AFT Models with Auxiliary Covariates. PLoS ONE 9(8): e104817. doi:10.1371/journal.pone.0104817

Editor: Xiaofeng Wang, Cleveland Clinic Lerner Research Institute, United States of America

Received April 23, 2014; **Accepted** July 14, 2014; **Published** August 15, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. A hard copy of the data is available from Appendix D of the following book: Fleming, T. R. and Harrington, D. P. (1991) Counting Processes and Survival Analysis. Wiley: New York. PBC data is available online with the R package pbc {survival}. Requests for the authors R code may be sent to the corresponding author.

Funding: The research of Kevin Granville was partially supported by USRA from the National Sciences and Engineering Research Council of Canada. The research of Zhaozhi Fan was partially supported by a Discovery Grant from the National Sciences and Engineering Research Council of Canada. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: zhaozhi@mun.ca

Introduction

It is not uncommon to have one or more missing or mismeasured covariates in large cohort epidemiological studies. There are always cases in medical studies, where it is difficult to obtain an accurate measurement for all patients due to a procedure being too expensive or invasive. Alternatively, some auxiliary measurements which are less precise, but highly related to the target procedure, can be easily collected. In some situations, all of the measurements are error prone, while in other cases, a validation subsample, where the measurements are all accurately taken, is made available.

The former is a pure measurement error problem. The purpose of this paper is to investigate the inference of the latter cases in a failure time setting. In some cases, the validation sample could be large enough on its own, so one could choose to ignore all data from subjects that have missing or mismeasured values for any of the covariates, with just a minor efficiency loss. However, if the validation sample is relatively small, utilizing the auxiliary information will lead to remarkable efficiency gain, as our simulation results will show.

The literature on statistical inference of missing or mismeasured data of failure time regression models is abundant. Ignoring the measurement errors in modeling could lead to severe estimation bias, depending on the magnitude of the measurement error, hence invalidate the whole inference procedure ([1] Prentice, 1982). See also [2] (Rubin, 1976), [3] (Fuller, 1987), [4] (Carrol et al., 1995), and [5] (Wang et al., 1998) among others. The negative influence of mismeasured or missing covariates is largely understood for the Cox proportional hazards model. But the same cannot be said of accelerated failure time models. Details about the Cox model can be seen in the work of [6] (Cox, 1972), [7] (Cox and Oakes, 1984), [8] (Kalbfleisch and Prentice, 2002), [9] (Hu et al., 1998), and [10] (Hu and Lin, 2002), and the references therein. See [11] (Zhou and Pepe, 1995), [12] (Zhou and Wang, 2000), [13] (Liu, Zhou and Cai, 2009), [14] (Fan and Wang, 2009), [15] (Liu, Wu and Zhou, 2010) censored survival models with auxiliary covariates.

However, due to the direct physical interpretation of the AFT models, and the fact that AFT models are robust to model misspecification in the sense that ignoring a covariate will not lead to much bias in estimating the remaining regression coefficients, see [7] (Cox and Oakes, 1984), the biasing effect of covariate measurement error on AFT models deserves further investigation. A recent work on the subject of measurement error in AFT models was done by [16] (He et al., 2007), using a simulation and extrapolation approach. [17] (Yu and Nan 2010) studied the regression calibration approach within the semiparametric framework, assuming a known parametric relationship between the accurately measured covariates and their auxiliaries, up to a few unknown nuisance parameters. [18] (Granville and Fan, 2012) studied the parametric AFT models with auxiliary covariates based on maximum likelihood method.

In this paper, we study the Buckley-James estimator [19] (Buckley and James, 1979) of AFT models with auxiliary covariates. The Buckley-James estimator was shown by [20] (Jin et al., 2006) to be consistent and asymptotically normal when using a consistent estimator as the initial value, due to its asymptotic linearity. Some other insights about the consistency and asymp-

totic theory of this estimator has been investigated by [21] (James and Smith, 1984) and [22] (Lai and Ying,1991), among others. We propose a local polynomial approximation method to handle the missing or mismeasured covariates, through the estimation of the conditional expectation of the unobservable estimating functions. This approach makes neither distributional assumptions about the model error term ε_i, beyond it having mean zero and a finite variance, nor parametric assumptions on the relationship between the correctly measured covariates and their auxiliary variables. The proposed approach will be introduced through a kernel smoothing method, a special case of the local polynomial approximation, see [14] (Fan and Wang, 2009), mainly due to the ease of presentation. See [23] (Nadaraya, 1964), [24] (Watson, 1964), and [25] (Wand and Jones, 1995) for details of kernel smoothing. Intensive simulation studies were conducted to investigate the small sample performance of our proposed method. The results show a remarkable efficiency gain over the method which ignores the auxiliary information.

The remainder of this paper is organized as follows. In the second Section, we introduce Buckley-James estimator for the accelerated failure time model and present our estimation method. Then we investigate the asymptotic properties of our proposed estimator. The Section thereafter contains the results and discussion of our numerical studies, including simulations and the PBC data illustration. In the last Section, we put forth some concluding remarks. The proofs for Theorems were deferred to the appendix.

Inference Methods of Accelerated Failure Time Model

Let T_i and C_i, $i=1,\ldots,n$ be the failure and censoring times for the ith subject in a large study cohort. Due to the censoring, we observe $S_i=\min(T_i,C_i)$ as well as a failure indicator $\delta_i=I(T_i\le C_i)$. Let $\{X_i,Z_i\}$ denote the covariate vector where X_i is the component which is only observed in the validation set and Z_i is the component that is available for the full study cohort. Let W_i be the auxiliary covariate to X_i. Hence the data consists of the validation sample $\{S_i, \delta_i, Z_i, X_i, W_i\}$, and the nonvalidation sample $\{S_i, \delta_i, Z_i, W_i\}$. In this paper we assume that X_i is scalar, mainly because of the simplicity of the presentation, and Z_i may be either a scalar or a vector. In practice, X_i could also be closely correlated with Z_i. A special case is the classical measurement error model $W_i=X_i+U_i$, where U_i is the error encountered when measuring X_i. It is assumed that the U_i's are independent and identically distributed random normal variables, $U_i\sim N(0,\sigma_u^2)$. Of the n observations, the validation sample contains n_V observations, and the non-validation sample contains $n_{\bar{V}}=n-n_V$ observations.

The accelerated failure time model based solely on the validation sample, can be expressed as

$$Y_i=\log(T_i)=\beta_1X_i+\beta_2'Z_i+\varepsilon_i, \tag{1}$$

where $\beta'=(\beta_1,\beta_2')$ is a vector of unknown regression coefficients and the ε_i's are independent and identically distributed with an unspecified distribution F which has mean zero and finite variance. Equation (1) assumes automatically that W_i provides no additional information about the failure time, given $\{X_i, Z_i\}$.

Without making any assumption to the distribution of ε_i, the Buckley-James least squares procedure (Buckley and James, 1979) estimates the regression parameters through the minimization of

$$\sum_{i=1}^{n}(Y_i-\beta_1X_i-\beta_2'Z_i)^2.$$

The least squares estimates of β_1 and β_2' are such that

$$\sum_{i=1}^{n}X_i(Y_i-\beta_1X_i-\beta_2'Z_i)=0, \tag{2}$$

and

$$\sum_{i=1}^{n}Z_i(Y_i-\beta_1X_i-\beta_2'Z_i)=0. \tag{3}$$

In order to deal with censoring, let $Y_i^*=Y_i\delta_i+\mathrm{E}[Y_i|Y_i>\log(C_i)](1-\delta_i)$. Then $\mathrm{E}[Y_i^*]=\beta_1X_i+\beta_2'Z_i$, so $\mathrm{E}[\sum_{i=1}^{n}X_i(Y_i^*-\beta_1X_i-\beta_2'Z_i)]=0$, and $\mathrm{E}[\sum_{i=1}^{n}Z_i(Y_i^*-\beta_1X_i-\beta_2'Z_i)]=0$.

The estimators b_1 and b_2' of β then satisfy

$$\sum_{i=1}^{n}X_i(Y_i^*-b_1X_i-b_2'Z_i)=0, \tag{4}$$

and

$$\sum_{i=1}^{n}Z_i(Y_i^*-b_1X_i-b_2'Z_i)=0. \tag{5}$$

However, the distribution of ε_i is unknown. The distribution of Y_i, and consequently, $\mathrm{E}[Y_i|Y_i>\log(C_i)]$ are both unknown. The censored observations are hence replaced by

$$\tilde{Y}_i(b)=b_1X_i+b_2'Z_i+\frac{\int_{e_i(b)}^{\infty}ud\widehat{F}_b(u)}{1-\widehat{F}_b(e_i(b))},$$

where $e_i(b)=S_i-b_1X_i-b_2'Z_i$, $i=1,\ldots,n$ are the residuals, and

$$\widehat{F_b}(\varepsilon)=1-\prod_{i:e_{(i)}\le\varepsilon}\left(\frac{n-i}{n-i+1}\right)^{\delta_i},$$

is the Kaplan-Meier Product Limit estimator of the distribution function of the residuals, F. $\widehat{F_b}(\varepsilon)$ is a discrete function which will not tend to 1 as ε increases if the largest residual is censored. Therefore, following the convention of Buckley and James, the largest residual is redefined as uncensored for all calculations, if necessary.

Let $\tilde{Y}_i^*(b)=Y_i\delta_i+\tilde{Y}_i(b)(1-\delta_i)$. The estimator for $\beta'=(\beta_1,\beta_2')$ is the solution of the following equation,

$$\beta=\left[\sum_{i=1}^{n}(X_i,Z_i')'(X_i,Z_i')\right]^{-1}\left[\sum_{i=1}^{n}(X_i,Z_i')'\tilde{Y}_i^*(\beta)\right]=\gamma(\beta). \tag{6}$$

It should be noted that $\gamma(\beta)$ depends on X_i, which is available only for the validation sample. For the non-validation sample we

can substitute the estimates of their conditional expectations given the auxiliary and other available covariates. The local polynomial approximation approach can be applied for this purpose, see [14] (Fan and Wang, 2009). For the simplicity of the presentation, we use the kernel smoothing method to estimate the conditional expectation of the unobserved covariates given the auxiliary information.

Note that this simplification does not necessarily lead to efficiency loss. Since the direct estimation of the conditional expectation of the estimating function depends also on the Kaplan-Meier estimation of the survival function of the regression residuals, it could also introduce additional instability into the inference, as compared with imputing the estimated conditional expectation of the mismeasured covariate. Our simulation also revealed this observation (results not included).

The conditional expectation of the mis-measured covariate, denoted by $\widehat{X}_i$, can be estimated as

$$\widehat{X}_i = \frac{\sum_{j\epsilon V} X_j k_h(\Gamma_i - \Gamma_j)}{\sum_{j\epsilon V} k_h(\Gamma_i - \Gamma_j)}, \tag{7}$$

where $\Gamma_i = (W_i, Z_i)'$, $k_h(\cdot) = (h_1 \cdots h_d)^{-1} k(\cdot/h_1, \cdots, \cdot/h_d)$ is the kernel function and $h = (h_1, \cdots, h_d)^T$ is the chosen vector of bandwidth. Using these $\widehat{X}_i$'s in place of the $n_{\overline{V}}$ missing X_i's, we may solve (6) for β using a numerical method, like Broyden's method. Note that, Broyden's method requires two initial values, while the method of [21] (James and Smith, 1984) only requires a single initial value. In this function, $\gamma_n(\beta_{(k)})$ is the value of (6) when calculated using $\beta_{(k)}$. A very natural selection of the initial value of β is the least squares regression estimator calculated from the validation sample.

The standard deviation of these estimators is estimated using bootstrapping. For R replicates, a simple random sample with replacement of the full sample size is taken from the observed data and the above estimation method for β is repeated on each replicate. A sample standard deviation is then calculated to estimate the true standard deviation of the β estimators.

Remark 1 *In order to retain the same quality of information among the replicated estimations, an alternative method of resampling was attempted to keep the proportion of censored observations constant in each replication. We defined* $\delta^* = \sum_{i=1}^n \delta_i$ *to be the total number of observations with uncensored failure times. For R replicates, a simple random sample with replacement of size* δ^* *was taken from these uncensored observations, and a simple random sample with replacement of size* $n - \delta^*$ *was taken from the remaining censored observations. However this alternative method was found to underestimate the true standard deviations and resulted in coverage probabilities that were lower than the nominal level. The reason of this outcome is mostly due to the fact that the independence of the censoring mechanism was broken by the sampling method.*

Defining a Solution

In order to solve the estimating equations for the regression parameters, we use the iterative scheme of $\beta_{(k+1)} = \gamma_n(\beta_{(k)})$. However, as noted by [19] (Buckley and James, 1979) and [21] (James and Smith, 1984), these iterations need not converge. The $\gamma(\beta)$ function is discontinuous and piecewise linear in β so an exact solution may not exist. When this is the case, the iterations can oscillate between two values of β. We define a possible alternate solution which is closest to satisfying $\beta = \gamma(\beta)$, or $\gamma(\beta) - \beta = 0$. If $\beta_{(k)}$ is oscillating between two points due to the lack of an exact solution, we define the alternate solution as $\beta_{(k)}$ that minimizes the modulus of this difference,

$$\min_{\beta_{(k)}} |\beta_{(k+1)} - \beta_{(k)}|. \tag{8}$$

When the iterations do not converge, a cut-off point has to be determined to stop the iterations. It is advised to select a number of iterations that is slightly greater than the amount required for the convergence when a solution typically does exist. For most of our simulations, we have set this point at $k = 11$. In many cases, our simulations converge in three or four steps. So at $k = 5$ or $k = 6$, $\beta_{(k)}$ breaks the loop when checked against $\beta_{(k-1)}$ for convergence, implying the solution being reached at iteration $k - 1$. If the iteration does not converge, the first ten values are checked and whichever value, after, say 5 steps of iteration, satisfies equation (8) is selected as a solution.

When dealing with real data, it is advised to choose an arbitrarily large number for the cut-off point to find the best possible solution.

Asymptotics

In this section, we investigate the asymptotic properties of our proposed estimator. For that sake, we rewrite the estimating function and the Kaplan-Meier estimator of the residual survival function in the counting process frame work. Define a function $U(b,\beta)$ by

$$U(b,\beta) = \sum_{i=1}^n \left[(X_i, Z_i') - (\bar{X}, \bar{Z}_i)\right]' \{\tilde{Y}_i^*(b) - \bar{Y}_i^*(b) - [(X_i, Z_i') - (\bar{X}, \bar{Z}_i)]\beta\},$$

where $\bar{X} = \sum_{i=1}^n X_i/n$, $\bar{Z} = \sum_{i=1}^n Z_i/n$ and $\bar{Y}_i^*(b) = \sum_{i=1}^n \tilde{Y}_i^*(b)/n$. The estimating equation (6) can be rewritten as

$$U(\beta,\beta) \triangleq U(\beta) = 0.$$

The Buckley-James estimator solves the above equation.

When some of the covariates are accurately recorded only for the validation sample, but with relevant auxiliary information available for the whole study cohort, the estimating functions involved those mis-measured covariates belonging to the non-validation sample. We propose to estimate those terms by using the local polynomial approximation approach. Let n denote the size of whole study cohort, η_i, for $i = 1, \ldots, n$ be the validation indicator. Define

$$\widehat{\Gamma}_i = (X_i, Z_i')'\eta_i + (\widehat{X}_i, Z_i')'(1 - \eta_i).$$

The derived estimating equation is then

$$\widehat{U}(\beta) = \sum_{i=1}^n \left[\widehat{\Gamma}_i - \overline{\widehat{\Gamma}}\right] \left\{\tilde{Y}_i^*(\beta) - \bar{Y}_i^*(\beta) - \left[\widehat{\Gamma}_i - \overline{\widehat{\Gamma}}\right]'\beta\right\}.$$

Our proposed estimator of the regression parameter, accommodating the auxiliary information, $\widehat{\beta}$ solves this derived estimating equation.

For a vector a, define $a^{\otimes 0}=1$, $a^{\otimes 1}=a$, $a^{\otimes 2}=aa'$ and $\|a\|=\sup_i|a_i|$. Let $Y_i(t)=I(e_i(\beta)\geq t)$, for $i=1,\ldots,n$ and $Y(t)=\sum_{i=1}^n Y_i(t)$. Let $\Gamma_i^*=(X_i,Z_i')'\eta_i+(X_i^*,Z_i')'(1-\eta_i)$, where $X_i^*=E(X_i|W_i,Z_i')$ for $i\in\overline{V}$.

Let $\widehat{S}_{(k)}(t)=n^{-1}\sum_{i=1}^n Y_i(t)\widehat{\Gamma}_i^{\otimes k}$, $S_{(k)}(t)=n^{-1}\sum_{i=1}^n Y_i(t)\Gamma_i^{\otimes k}$ and

$$s_{(k)}(t)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \Gamma_i^{\otimes k}P(e_i\geq t|\Gamma_i).$$

Denote further $\widehat{S^*}_{(k)}(t)=n^{-1}\sum_{i=1}^n Y_i(t)\widehat{\Gamma^*}_i^{\otimes k}$, $S^*_{(k)}(t)=n^{-1}\sum_{i=1}^n Y_i(t)\Gamma_i^{*\otimes k}$ and

$$s^*_{(k)}(t)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \Gamma_i^{*\otimes k}P(e_i\geq t|\Gamma_i).$$

Without loss of generality, let d be the dimension of Γ_i in the definition of the local polynomial approximation. Suppose further that α is the order of the kernel function K, i.e.

$$\int u^q K(u)du=0,\quad for\quad q=1,2,\cdots,\alpha-1,\quad \int u^\alpha K(u)du\neq 0,$$

and $\int K^2(u)du<+\infty$. The bandwidth conditions are given below.

[**BC**] As $n\to\infty$, $nh^{2\alpha}\to 0$, $nh^{2d}\to\infty$.

The following assumptions, beyond the bandwidth conditions, are necessary for the asymptotic properties of the proposed method.

C.0 The hazard rate function $\lambda(t)$ of $e_i(\beta)$ is such that $\int_{-\infty}^{\infty}\lambda(t)dt<\infty$.

C.1 There exists a constant $B>0$, such that $\|Z_i\|\leq B$, $\|X_i\|\leq B$ and $\|W_i\|\leq B$ for all $i=1,\ldots,n$.

C.2 F has a twice-continuously differentiable density f such that

$$\int_{-\infty}^{\infty} t^2dF(t)<\infty \text{ and } \int_{-\infty}^{\infty}(f'(t)/f(t))^2dF(t)<\infty.$$

C.3 The solution to $U(\beta)=0$ is unique and is an interior point of $\mathbb{B}$, where $\mathbb{B}$ is a compact subset of R^p.

C.4 There exists a function $\pi(t)$ such that, as $n\to\infty$,

$$\sup_{t\in R}\left|\frac{Y(t)}{n}-\pi(t)\right|\xrightarrow{P}0.$$

Remark 2 *The assumption C.3 is proposed just to simplify the proof of the asymptotics of the Buckley-James estimator. This could be violated due to the instability of the Kaplan-Meier estimator of the survival function when getting into the distribution tail. When this violation happens, the tail modification by [22] (Lai and Ying, 1991) should be applied and the method of [26] (Jin, et al., 2006) of selecting a consistent and asymptotically normal estimator as the initial value can be adopted.*

Theorem 1 *Under conditions C.0-C.4 and the bandwidth conditions [BC], $n^{-1/2}\widehat{U}(\beta)$ converges in distribution to a zero-mean normal random vector with covariance matrix $\rho\Sigma(\beta)+(1-\rho)\Sigma_1(\beta)$, where $\rho=\lim_{n\to\infty}n_V/n$,*

$$\Sigma(\beta)=E\left[\int_{-\infty}^{\infty}\left\{(\Gamma_i-\overline{\Gamma})\left(t-\frac{\int_{e_i(\beta)}^{\infty}tdF(t)}{1-F(e_i(\beta))}\right)+\overline{\zeta}_\Gamma(t)\right\}dM_i(t)\right.$$
$$\left.+\frac{1-\rho}{\rho}Q\beta\right]^{\otimes 2},$$

where $Q=\lim_{n\to\infty}\frac{1}{n_V}\sum_{j\in V}Q_j$,

$$Q_j=\int_{-\infty}^{+\infty}\left[E(\Gamma_j|Y_j(t)=1,\overline{V})-\frac{s_{(1)}(t)}{s_{(0)}(t)}\right]$$
$$[\Gamma_j-E(\Gamma_j|Y_j(t)=1,\overline{V})]'\frac{dF(t)}{1-F(t)},$$

and Σ_1 is defined as

$$\Sigma_1(\beta)=\int_{-\infty}^{\infty}\left\{s^*_{(2)}(t)-\frac{s^{*\otimes 2}_{(1)}(t)}{s^*_{(0)}(t)}\right\}\frac{[\int_t^\infty(1-F_\beta(s))ds]^2}{[1-F_\beta(t)]^2}dF_\beta(t).$$

Theorem 2 *Under assumptions C.0-C.4 and the bandwidth conditions [BC], $n^{-1/2}\widehat{U}(\beta)$ is asymptotically linear within the $n^{-1/3}$ neighborhood of β, with probability 1, in the sense that*

$$n^{-1/2}\widehat{U}(b)=n^{-1/2}\widehat{U}(\beta)-[\rho A+(1-\rho)(A^*+Q)]$$
$$\sqrt{n}(b-\beta)+o(\sqrt{n}(b-\beta)+1),$$

if $\|b-\beta\|\leq n^{-1/3}$, where the matrix A is defined as

$$A=\int_{-\infty}^{\infty}\left\{s_{(2)}(t)-\frac{s^{\otimes 2}_{(1)}(t)}{s_{(0)}(t)}\right\}\left[\int_t^\infty(1-F_\beta(s))ds\right]d\lambda(t),$$

and A^* as

$$A^*=\int_{-\infty}^{\infty}\left\{s^*{}_{(2)}(t)-\frac{s^{*\otimes 2}_{(1)}(t)}{s^*_{(0)}(t)}\right\}\left[\int_t^\infty(1-F_\beta(s))ds\right]d\lambda(t).$$

Corollary 1 *Under assumptions C.0-C.4, the bandwidth conditions [BC] and the assumption that A is nonsingular, the solution $\widehat{\beta}$ to $n^{-1/2}\widehat{U}(\beta)=0$ converges in probability to β.*

Theorem 3 *Under assumptions C.0-C.4 and the bandwidth conditions [BC], $\sqrt{n}(\widehat{\beta}-\beta)$ converges in distribution to a zero mean normal random vector with covariance matrix*

$$[\rho A+(1-\rho)(A^*+Q)]^{-1}[\rho\Sigma(\beta)+$$
$$(1-\rho)\Sigma_1(\beta)][\rho A+(1-\rho)(A^*+Q)]^{-1}.$$

Proof of Theorem 1

Let V be the set of the indices of the validation sample, $\bar{V}$ that of the non-validation sample and $\eta_i = I(i \in V)$, $i = 1, \ldots, n$ be the validation indicator.

Let $\bar{\Gamma} = \sum_{i \in V} \Gamma_i / n_V$, $\bar{\Gamma}^{\wedge} = \sum_{i \in \bar{V}} \widehat{\Gamma}_i / n_{\bar{V}}$ and $\bar{\Gamma}^{*} = \sum_{i \in \bar{V}} \Gamma_i^{*} / n_{\bar{V}}$. Let

$$\tilde{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \left(\eta_i \Gamma_i + (1 - \eta_i) \widehat{\Gamma}_i \right).$$

Define

$$e_i(\beta) = y_i - \beta' \Gamma_i^{*},$$

and

$$\widehat{e}_i(\beta) = y_i - \beta' \widehat{\Gamma}_i.$$

Then

$$\widehat{e}_i(\beta) - e_i(\beta) = (1 - \eta_i) \beta' (\widehat{\Gamma}_i - \Gamma_i^{*}).$$

Let $F(\cdot)$ and $F^{\wedge}(\cdot)$ be the distribution functions of e and $\widehat{e}$, Λ and $\Lambda^{\wedge}$ be their cumulative hazard functions. Then

$$M_i(t) = \delta_i I(e_i(\beta) \le t) - \int_{-\infty}^{t} I(e_i(\beta) \ge u) d\Lambda(u),$$

and

$$M_i^{\wedge}(t) = \delta_i I(\widehat{e}_i(\beta) \le t) - \int_{-\infty}^{t} I(\widehat{e}_i(\beta) \ge u) d\Lambda^{\wedge}(u),$$

are martingales with respect to complete σ-field generated by

$$\delta_i, \quad I(e_i(\beta) \le t), \quad I(\widehat{e}_i(\beta) \le t), \quad X_i, \quad W_i, \quad \eta_i, \quad i = 1, \ldots, n.$$

Further,

$$\Lambda^{\wedge}(t) = \Lambda(t + (1 - \eta_i) \beta' (\widehat{\Gamma}_i - \Gamma_i^{*})),$$

and

$$M_i^{\wedge}(t) = M_i(t + (1 - \eta_i) \beta' (\widehat{\Gamma}_i - \Gamma_i^{*})).$$

Let $\widehat{F}_{\beta}(\cdot)$ and $\widehat{F}_{\beta}^{\wedge}(\cdot)$ be the (nominal) Kaplan-Meier estimators of $F_{\beta}(\cdot)$ and $F_{\beta}^{\wedge}(\cdot)$. The estimated Buckley-James estimating function can be rewritten as

$$\widehat{U}(\beta) = \sum_{i=1}^{n} [\eta_i \Gamma_i + (1 - \eta_i) \widehat{\Gamma}_i - \tilde{\Gamma}] \left[\widehat{e}_i(\beta) + (1 - \delta_i) \frac{\int_{\widehat{e}_i(\beta)}^{\infty} u d\widehat{F}_{\beta}^{\wedge}(u)}{1 - \widehat{F}_{\beta}^{\wedge}(\widehat{e}_i(\beta))} \right]$$

$$= \sum_{i=1}^{n} [\eta_i \Gamma_i + (1 - \eta_i) \widehat{\Gamma}_i - \tilde{\Gamma}] \left[e_i(\beta) + (1 - \delta_i) \frac{\int_{e_i(\beta)}^{\infty} u dF(u)}{1 - F(e_i(\beta))} \right]$$

$$+ \sum_{i=1}^{n} (1 - \delta_i) [\eta_i \Gamma_i + (1 - \eta_i) \widehat{\Gamma}_i - \tilde{\Gamma}] \left[\frac{\int_{e_i(\beta)}^{\infty} u d\widehat{F}_{\beta}(u)}{1 - \widehat{F}_{\beta}(e_i(\beta))} - \frac{\int_{e_i(\beta)}^{\infty} u dF(u)}{1 - F(e_i(\beta))} \right]$$

$$+ \sum_{i=1}^{n} (1 - \eta_i) [\eta_i \Gamma_i + (1 - \eta_i) \widehat{\Gamma}_i - \tilde{\Gamma}] \beta' (\widehat{\Gamma}_i - \Gamma_i^{*})$$

$$+ \sum_{i=1}^{n} n (1 - \delta_i) [\eta_i \Gamma_i + (1 - \eta_i) \widehat{\Gamma}_i - \tilde{\Gamma}] \left[\frac{\int_{\widehat{e}_i(\beta)}^{\infty} u d\widehat{F}_{\beta}^{\wedge}(u)}{1 - \widehat{F}_{\beta}^{\wedge}(_i(\beta))} - \frac{\int_{e_i(\beta)}^{\infty} u d\widehat{F}_{\beta}(u)}{1 - \widehat{F}_{\beta}(e_i(\beta))} \right]$$

$$\triangleq I_1 + I_2 + I_3 + I_4.$$

By the martingale representation of Kaplan-Meier estimator of the survival function see [25] (Fleming and Harrington, 1991), also see [20] (Jin, et al., 2006), the bandwidth condition [BC] and the continuity of F (assumption C.2), the first two terms in the above equation can be rewritten as

$$I_1 + I_2 = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \left[\eta_i \Gamma_i + (1 - \eta_i) \widehat{\Gamma}_i - \tilde{\Gamma} \right] \left(t - \frac{\int_{e_i(\beta)}^{\infty} t dF(t)}{1 - F(e_i(\beta))} \right) + \tilde{\zeta}_{\Gamma}(t) \right\} dM_i(t) + o_P(n^{1/2}),$$

where

$$\tilde{\zeta}_{\Gamma}(t) = n^{-1} \sum_{i=1}^{n} (\eta_i \Gamma_i + (1 - \eta_i) \widehat{\Gamma}_i - \tilde{\Gamma})(1 - \delta_i) \xi_i(t),$$

and

$$\xi_i(t) = \int_{e_i(\beta)}^{\infty} S(s) ds \frac{I(t \le e_i(\beta))}{y(t)} + \int_{t}^{\infty} S(s) ds \frac{I(t > e_i(\beta))}{y(t)}.$$

The Kaplan-Meier estimators of the survival functions of $\widehat{e}_i(\beta)$ and $e_i(\beta)$ lead to

$$1 - \widehat{F}_{\beta}^{\wedge}(\widehat{e}_i(\beta)) = 1 - \widehat{F}_{\beta}(e_i(\beta)),$$

and

$$\int_{\widehat{e}_i(\beta)}^{\infty} u d\widehat{F}_{\beta}^{\wedge}(u) - \int_{e_i(\beta)}^{\infty} u d\widehat{F}_{\beta}(u) = -(1 - \eta_i) \beta_1 (\widehat{X}_i - X_i^{*})(1 - \widehat{F}_{\beta}(e_i(\beta))).$$

Hence

$$I_4 = -\sum_{i=1}^{n}(1-\delta_i)(1-\eta_i)[\eta_i\Gamma_i+(1-\eta_i)\widehat{\Gamma}_i-\tilde{\Gamma}]\beta'(\widehat{\Gamma}_i-\Gamma_i^*),$$

and

$$I_3+I_4=\sum_{i=1}^{n}\delta_i(1-\eta_i)[\eta_i\Gamma_i+(1-\eta_i)\widehat{\Gamma}_i-\tilde{\Gamma}]\beta'(\widehat{\Gamma}_i-\Gamma_i^*).$$

This term can be further rewritten as

$$I_3+I_4=\frac{1-\rho}{\rho}\sum_{j\in V}\int_{-\infty}^{+\infty}\left[E(\Gamma_j|\,Y_j(t)=1,\overline{V})-\frac{s_{(1)}(t)}{s_{(0)}(t)}\right]$$
$$(\Gamma_j-E(\Gamma_j|\,Y_j(t)=1,\overline{V}))'\beta\frac{dF(t)}{1-F(t)}+o_p(1)$$

$$=\frac{1-\rho}{\rho}\sum_{j\in V}Q_j\beta+o_p(1)$$

We have

$$\frac{1}{n^{1/2}}\widehat{U}(\beta)=$$

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left\{\left[\eta_i\Gamma_i+(1-\eta_i)\widehat{\Gamma}_i-\tilde{\Gamma}\right]\left(t-\frac{\int_{e_i(\beta)}^{\infty}udF(t)}{1-F(e_i(\beta))}\right)+\tilde{\zeta}_\Gamma(t)\right\}dM_i(t)$$

$$+\frac{1}{n^{1/2}}\frac{1-\rho}{\rho}\sum_{j\in V}Q_j\beta+o_P(1).$$

By assumptions C.0, C.2 and Lenglart's inequality, we have

$$\frac{1}{n^{1/2}}\sum_{i\in\overline{V}}\int_{-\infty}^{\infty}\left[(\widehat{\Gamma}_i-\bar{\Gamma}^{\wedge})-(\Gamma_i^*-\bar{\Gamma}^*)\right]\left(t-\frac{\int_{e_i(\beta)}^{\infty}udF(t)}{1-F(e_i(\beta))}\right)dM_i(t)=o_P(1).$$

Hence the estimating function can be rewritten as

$$\frac{1}{n^{1/2}}\widehat{U}(\beta)=$$
$$\frac{1}{n^{1/2}}\sum_{i\in V}\int_{-\infty}^{\infty}\left\{[\Gamma_i-\bar{\Gamma}]\left(t-\frac{\int_{e_i(\beta)}^{\infty}udF(t)}{1-F(e_i(\beta))}\right)+\bar{\zeta}_\Gamma(t)\right\}dM_i(t)$$

$$+\frac{1}{n^{1/2}}\frac{1-\rho}{\rho}\sum_{j\in V}Q_j\beta$$

$$+\frac{1}{n^{1/2}}\sum_{i\in\overline{V}}\int_{-\infty}^{\infty}\left\{\left[\Gamma_i^*-\bar{\Gamma}^*\right]\left(t-\frac{\int_{e_i(\beta)}^{\infty}udF(t)}{1-F(e_i(\beta))}\right)+\bar{\zeta}_\Gamma^*(t)\right\}dM_i(t)$$

$$=\frac{1}{n^{1/2}}\tilde{U}_V(\beta)+\frac{1}{n^{1/2}}U_{\overline{V}}^*(\beta)+o_P(1).\quad(U)$$

Note that $\tilde{U}_V(\beta)$ is a sum of n_V i.i.d. terms hence central limit theorem applies. By conditions C.0 through C.4 and the martingale central limit theorem, $U_{\overline{V}}^*(\beta)$ converges in distribution to a normal random vector. Further, by independence of $\tilde{U}_V(\beta)$ and $U_{\overline{V}}^*(\beta)$, we have

$$\frac{1}{n^{1/2}}\widehat{U}(\beta)\xrightarrow{D}N(0,\rho\Sigma(\beta)+(1-\rho)\Sigma_1(\beta)).$$

Proof of Theorem 2

From the equation (U) in the proof of Theorem 3.2, and by Theorem 4.1 of Lai and Ying (1991), we have, for $\|b-\beta\|<n^{-1/3}$,

$$\frac{1}{n^{1/2}}U_{\overline{V}}^*(b)=\frac{1}{n^{1/2}}U_{\overline{V}}^*(\beta)+(1-\rho)A^*\sqrt{n}(b-\beta)+o(1+\sqrt{n}(b-\beta))$$

with probability 1. The term $\tilde{U}_V(b)$ consists of two parts, we have

$$\frac{1}{n^{1/2}}\tilde{U}_V(b)=\frac{1}{n^{1/2}}\tilde{U}_V(\beta)+$$
$$(\rho A+(1-\rho)Q)\sqrt{n}(b-\beta)+o(1+\sqrt{n}(b-\beta)),$$

with probability 1. Hence

$$\frac{1}{n^{1/2}}\widehat{U}(b)=\frac{1}{n^{1/2}}\widehat{U}(\beta)+$$
$$(\rho A+(1-\rho)(A^*+Q))\sqrt{n}(b-\beta)+o(1+\sqrt{n}(b-\beta)),$$

with probability 1.

Corollary 1 and Theorem 3 are direct conclusions of Theorems 1 and 2.

Results of Numerical Studies

Simulation Studies

In this section we examine the small sample performance of our proposed estimator. Let $\widehat{\boldsymbol{\beta}}_S$ denote our proposed estimator of the regression coefficients. Its small sample performance is compared with three alternative estimators: the validation estimator ($\widehat{\boldsymbol{\beta}}_V$) which is based solely on the validation sample; the naive estimator ($\widehat{\boldsymbol{\beta}}_N$), which ignores the measurement error by assuming that the unobserved X_i's are equal to the observed W_i's; and the complete case estimator ($\widehat{\boldsymbol{\beta}}_{CV}$), when we assume that X_i are observed for the whole study cohort.

Table 1. Results after 500 simulations for $\beta' = (\log(2), \log(1.5)) = (0.693, 0.405)$ using a standard normal error term.

n	n_V	Censor Rate	σ_u	$\widehat{\beta}$	$\widehat{\beta}_1$	$SD_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_1}$	$CP_{\widehat{\beta}_1}$	$\widehat{\beta}_2$	$SD_{\widehat{\beta}_2}$	$SE_{\widehat{\beta}_2}$	$CP_{\widehat{\beta}_2}$
400	200	0.3	0.5	$\widehat{\beta}_S$	0.693	0.030	0.031	0.948	0.405	0.029	0.029	0.950
				$\widehat{\beta}_V$	0.692	0.041	0.044	0.924	0.407	0.040	0.041	0.938
				$\widehat{\beta}_N$	0.673	0.030	0.030	0.882	0.423	0.029	0.029	0.924
				$\widehat{\beta}_{CV}$	0.693	0.029	0.030	0.954	0.406	0.028	0.028	0.950
400	200	0.5	0.5	$\widehat{\beta}_S$	0.692	0.035	0.036	0.940	0.405	0.032	0.033	0.934
				$\widehat{\beta}_V$	0.692	0.049	0.048	0.942	0.406	0.044	0.045	0.948
				$\widehat{\beta}_N$	0.676	0.035	0.035	0.894	0.424	0.032	0.033	0.902
				$\widehat{\beta}_{CV}$	0.691	0.034	0.035	0.936	0.406	0.031	0.032	0.936
400	200	0.3	0.8	$\widehat{\beta}_S$	0.695	0.031	0.032	0.948	0.404	0.030	0.031	0.928
				$\widehat{\beta}_V$	0.691	0.041	0.040	0.964	0.408	0.040	0.040	0.948
				$\widehat{\beta}_N$	0.645	0.031	0.031	0.650	0.449	0.030	0.030	0.676
				$\widehat{\beta}_{CV}$	0.693	0.029	0.030	0.944	0.405	0.028	0.029	0.938
400	200	0.5	0.8	$\widehat{\beta}_S$	0.695	0.036	0.037	0.956	0.400	0.033	0.035	0.932
				$\widehat{\beta}_V$	0.694	0.049	0.053	0.938	0.403	0.044	0.046	0.942
				$\widehat{\beta}_N$	0.655	0.036	0.036	0.782	0.451	0.033	0.035	0.728
				$\widehat{\beta}_{CV}$	0.694	0.034	0.034	0.958	0.403	0.031	0.033	0.944
250	150	0.3	0.5	$\widehat{\beta}_S$	0.690	0.038	0.039	0.938	0.408	0.036	0.038	0.938
				$\widehat{\beta}_V$	0.688	0.048	0.049	0.938	0.408	0.046	0.045	0.940
				$\widehat{\beta}_N$	0.674	0.038	0.039	0.914	0.422	0.036	0.037	0.910
				$\widehat{\beta}_{CV}$	0.689	0.037	0.038	0.942	0.407	0.035	0.037	0.938
250	150	0.5	0.5	$\widehat{\beta}_S$	0.690	0.044	0.044	0.936	0.402	0.040	0.040	0.944
				$\widehat{\beta}_V$	0.690	0.056	0.058	0.940	0.404	0.050	0.051	0.952
				$\widehat{\beta}_N$	0.677	0.044	0.044	0.920	0.418	0.040	0.040	0.944
				$\widehat{\beta}_{CV}$	0.690	0.044	0.044	0.938	0.403	0.039	0.039	0.950

doi:10.1371/journal.pone.0104817.t001

Table 2. Results after 500 simulations for $\beta' = (\log(2), \log(1.5)) = (0.693, 0.405)$ using an extreme value error term.

n	n_V	Censor Rate	σ_u	$\widehat{\beta}$	$\widehat{\beta}_1$	$SD_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_1}$	$CP_{\widehat{\beta}_1}$	$\widehat{\beta}_2$	$SD_{\widehat{\beta}_2}$	$SE_{\widehat{\beta}_2}$	$CP_{\widehat{\beta}_2}$
400	200	0.3	0.5	$\widehat{\beta}_S$	0.696	0.058	0.056	0.962	0.411	0.054	0.056	0.948
				$\widehat{\beta}_V$	0.694	0.080	0.078	0.946	0.413	0.076	0.080	0.938
				$\widehat{\beta}_N$	0.657	0.055	0.053	0.892	0.411	0.054	0.056	0.946
				$\widehat{\beta}_{CV}$	0.696	0.056	0.054	0.954	0.411	0.054	0.056	0.944
250	150	0.3	0.5	$\widehat{\beta}_S$	0.694	0.073	0.076	0.940	0.401	0.068	0.067	0.960
				$\widehat{\beta}_V$	0.695	0.092	0.095	0.926	0.399	0.086	0.090	0.940
				$\widehat{\beta}_N$	0.663	0.070	0.072	0.920	0.401	0.068	0.067	0.958
				$\widehat{\beta}_{CV}$	0.694	0.071	0.072	0.948	0.402	0.067	0.066	0.960

doi:10.1371/journal.pone.0104817.t002

Table 3. Results after 500 simulations for $\beta' = (\log(2), \log(1.5)) = (0.693, 0.405)$ using a logistic error term.

n	n_V	Censor Rate	σ_u	$\widehat{\beta}$	$\widehat{\beta}_1$	$SD_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_1}$	$CP_{\widehat{\beta}_1}$	$\widehat{\beta}_2$	$SD_{\widehat{\beta}_2}$	$SE_{\widehat{\beta}_2}$	$CP_{\widehat{\beta}_2}$
400	200	0.3	0.5	$\widehat{\beta}_S$	0.689	0.069	0.071	0.934	0.405	0.066	0.067	0.952
				$\widehat{\beta}_V$	0.693	0.095	0.099	0.940	0.403	0.094	0.092	0.956
				$\widehat{\beta}_N$	0.651	0.065	0.068	0.888	0.406	0.066	0.067	0.948
				$\widehat{\beta}_{CV}$	0.688	0.067	0.070	0.940	0.405	0.066	0.067	0.956
400	200	0.3	0.8	$\widehat{\beta}_S$	0.701	0.072	0.074	0.936	0.403	0.067	0.071	0.926
				$\widehat{\beta}_V$	0.701	0.096	0.096	0.950	0.401	0.093	0.096	0.942
				$\widehat{\beta}_N$	0.608	0.064	0.064	0.724	0.403	0.067	0.071	0.932
				$\widehat{\beta}_{CV}$	0.697	0.067	0.070	0.936	0.403	0.066	0.070	0.926

doi:10.1371/journal.pone.0104817.t003

Table 4. AFT model analysis of PBC data, smoothing for $\log(ast)$.

Covariate	$\widehat{\beta}_S$	SD	P-Value	$\widehat{\beta}_V$	SD	P-Value
Intercept	15.5304	2.5729	1.5792e-09	16.1642	2.3047	2.3239e-12
log(ast)	−0.3783	0.1926	4.9482e-02	−0.3364	0.1805	6.2311e-02
age	−0.0278	0.0058	1.8556e-06	−0.0249	0.0061	3.9895e-05
log(albumin)	1.4729	0.5551	7.9733e-03	1.3926	0.5883	1.7931e-02
log(bili)	−0.4800	0.0781	7.7648e-10	−0.4510	0.0781	7.7448e-09
edema05	−0.4387	0.2124	3.8858e-02	−0.3006	0.2221	1.7593e-01
edema1	−0.9190	0.2968	1.9610e-03	−0.9178	0.3063	2.7279e-03
log(protime)	−2.4323	0.8712	5.2415e-03	−2.8227	0.7813	3.0267e-04

doi:10.1371/journal.pone.0104817.t004

The data for the simulations was generated in the following way. The X_i and Z_i are generated from a uniform distribution, $X_i, Z_i \sim \text{uniform}[0,5]$. For each X_i, the auxiliary covariate is defined as $W_i = X_i + U_i$, where U_i is generated from a normal distribution with mean zero and standard deviation σ_u. The value of σ_u determines the magnitude of the measurement error. The failure times were then defined as $T_i = \exp\{Y_i\}$ where $Y_i = \beta_1 X_i + \beta_2' Z_i + \varepsilon_i$. The ε_i's were taken to be independent and identically distributed from either a standard normal, standard extreme value, or logistic distribution, respectively.

Various other parameters are controlled over all simulations. Each run calculates 1000 replicates in the bootstrapping to give consistent estimators of the standard deviations. The parameters were chosen as $\beta' = (\beta_1, \beta_2) = (\log(2), \log(1.5))$. Within a simulation, the censoring times are randomly generated from a uniform distribution with lower limit 0 and an appropriate upper limit to ensure an approximate 30% or 50% censor rate. The n and n_V values are chosen to be either $n = 400$ and $n_V = 200$, having half of the data in the validation set, or $n = 250$ and $n_V = 150$, with the validation set containing 60% of the data. Finally, two values of σ_u are selected, $\sigma_u = 0.5$, and $\sigma_u = 0.8$. For the kernel smoothing used to calculate $\widehat{\beta}_S$ the Gaussian kernel function is selected, which has an order of 2,

$$K(u) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{1}{2}u^2},$$

where $u = (W_i - W_j)/h$. We choose bandwidth $h = 2\sigma_u n^{-1/3}$ as used by [12] (Zhou and Wang, 2000).

The standard error (SE), standard deviation (SD), and coverage probability (CP) are calculated for each set of simulations. The SE values are the sample standard deviations of the β estimates, the SD values are the mean standard deviations generated from the bootstrapping in each simulation, and CP is equal to the percentage of simulations that had the true β value within a 95% confidence interval around its estimate when using the result of the bootstrapping for the standard deviation. The results are presented in Tables 1, 2, and 3.

From Tables 1, 2, and 3, we make the following observations:

(i) Estimators $\widehat{\beta}_S$ and $\widehat{\beta}_V$ are performing very well for each of the three error distributions.

(ii) Naive estimator $\widehat{\beta}_N$ is biased when the measurement error variance σ_u^2 is large.

(iii) The $\widehat{\beta}_S$ estimator is more efficient than $\widehat{\beta}_V$, having standard errors comparable to that of $\widehat{\beta}_{CV}$.

(iv) The proposed method removes the estimation bias in $\widehat{\beta}_N$, both for the regression coefficient of the error-prone covariate and that of the accurately measured covariates.

(v) The bootstrapping procedure results in good estimates of the standard error for all observed cases over the four estimators and three error distributions.

(vi) The coverage probabilities for the 95% confidence intervals are very close to their nominal level, except for $\widehat{\beta}_N$ when σ_u is large, where the estimate is severely biased.

(vii) The model experiences the least variation when the error term follows the standard normal distribution, with

Table 5. AFT model analysis of PBC data, smoothing for $\log(copper)$.

Covariate	$\widehat{\beta}_S$	SD	P-Value	$\widehat{\beta}_V$	SD	P-Value
Intercept	14.6413	2.1482	9.3809e-12	15.1929	1.8216	0.0000e+00
log(copper)	−0.3299	0.0883	1.8663e-04	−0.3105	0.0873	3.7675e-04
age	−0.0250	0.0061	3.9593e-05	−0.0217	0.0061	3.4084e-04
log(albumin)	1.4324	0.5499	9.1876e-03	1.2576	0.5783	2.9666e-02
log(bili)	−0.4218	0.0717	3.9422e-09	−0.4018	0.0739	5.3323e-08
edema05	−0.4285	0.2160	4.7314e-02	−0.3097	0.2226	1.6422e-01
edema1	−0.9021	0.3041	3.0152e-03	−0.9411	0.3113	2.5003e-03
log(protime)	−2.2738	0.8185	5.4687e-03	−2.5324	0.7294	5.1651e-04

doi:10.1371/journal.pone.0104817.t005

standard errors that are approximately half the size as when the error term follows the chosen extreme value or logistic distributions.

(viii). The efficiency gain is ignorable when σ_u^2 is small, such as 0.2 or smaller (simulation results not reported).

Application to PBC Data

To illustrate how to use the smoothing method in practice, we analyze the data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver. PBC is a chronic liver disease that inflames and slowly destroys the bile ducts in the liver, impairing its ability to function properly. It is believed to be a type of autoimmune disorder where the immune system attacks the bile ducts. PBC occurs primarily in women, with approximately 90% of patients being women, most often between the ages of 40 and 60. There is currently no known cure for the disease; the only known way to remove PBC is through a liver transplant, see [27,28].

In the Mayo Clinic trial, 418 patients were eligible. Of these patients, mostly complete data was obtained from the first 312 patients. The remaining 106 patients were not part of the actual clinical trial but had some basic measurements taken and were followed for survival. The variables we used in our regression on the logarithm of time were *age*, patient's age (in years); *albumin*, serum albumin (in mg/dl); *ast*, aspartate aminotransferase (in U/ml), once referred to as SGOT; *bili*, serum bilirunbin (in mg/dl); *copper*, urine copper (ug/day); *edema*, equal to 0 if no edema, 0.5 if untreated or successfully treated, or 1 if there exists edema despite diuretic therapy; *protime*, standardized blood clotting time. Of these, two cases were examined using either ast or copper for our X covariate to be smoothed due to incomplete data, while the others are mostly complete and thus are included in Z.

Edema was split into two categorical variables, *edema05* and *edema1*, defined as

$$edema05 = \begin{cases} 1 & , \quad edema = 0.5, \\ 0 & , \quad \text{otherwise}, \end{cases}$$

and

$$edema1 = \begin{cases} 1 & , \quad edema = 1, \\ 0 & , \quad \text{otherwise}. \end{cases}$$

We also took the log transformation of *albumin*, *ast*, *bili*, *copper*, and *protime*, in the interest of making their marginal distributions closer to normal. For the smoothing of the unobserved $\log(ast)$ and $\log(copper)$ values, $\log(bili)$ was chosen as the auxiliary covariate for both due to its high correlation (>0.5) with both variables. The bandwidth h was calculated using the sample standard deviation of $\log(bili)$, resulting in $\sigma_u \approx 1.020551$ and $h = 0.2734221$ using the same formula as in the numerical simulations, $h = 2\sigma_u n^{-1/3}$. Any observations missing a value for either of the Z covariates were removed, leaving both cases with $n = 416$ while $n_V = 312$ for the model using $\log(ast)$ and $n_V = 310$ for the model using $\log(copper)$.

Examining Tables 4 and 5, we see that both X variables had their estimated standard deviations increase by a small amount due to the error added into the model from smoothing for a missing covariate instead of a mismeasured one. If W was of the form $W_i = X_i + U_i$ like in the simulations, it could have resulted in a higher correlation between the auxiliary variable, W, and the X variable, depending on the magnitude of the measurement error. Despite the small increase in standard deviation the $\log(ast)$ term becomes significant at the 5% confidence level after smoothing, and while $\log(copper)$ was already significant, the p-value did decrease. For the Z variables, we see that they all have a smaller or approximately equal standard deviation after smoothing, which is expected when using the full sample size without needing smoothing for those variables, except for $\log(protime)$ and the intercept term which increased.

Discussion

In this paper we proposed the use of the Buckley-James estimator as a nonparametric method of estimating the regression parameters of an accelerated failure time model with auxiliary covariates. Kernel smoothing was applied using the auxiliary covariates to estimate missing or mismeasured covariates. The Buckley-James method is then applied to the whole study cohort for the inference of the covariates effect. The standard deviations of the estimates of the regression coefficients are estimated through bootstrapping. The proposed estimator is consistent and asymptotically normal.

This method was most effective in the case of mis-measured data due to the naturally high correlation between the corresponding W and X,Z variables which resulted in the estimator involving the smoothing, $\widehat{\beta}_S$, being more efficient than the validation estimator, $\widehat{\beta}_V$, as shown in the numerical simulations. The method should also perform well for the missing variable case given a sufficiently strong correlation. The method was applied to the PBC data as an illustration.

The smoothing model is set up in a general format. In applications, we should only choose those variables which are highly related to the mismeasured one. By doing so we can avoid the situations such as the auxiliary covariates only occupy a narrow region, which could cause instability in the local smoothing, hence the whole model.

Caution should also be taken when the proposed method is applied to a data with extremely small validation sample. A classic measurement error model might be a better option, where one can estimate the measurement error variance using the validation sample.

Acknowledgments

The authors thank the academic editor, Dr. Xiaofeng Wang and two anonymous referees for their careful review of a previous version and insightful comments, which led to a better presentation of this paper.

Author Contributions

Conceived and designed the experiments: KG ZF. Performed the experiments: KG ZF. Analyzed the data: KG. Contributed reagents/materials/analysis tools: ZF. Contributed to the writing of the manuscript: KG ZF.

References

1. Prentice RL (1982) Covariate measurement errors and parameter estimation in failure time regression model, Biometrika 69, pp. 331–342.
2. Rubin DB (1976) Inference and missing data, Biometrika 63. pp. 581–592.
3. Fuller WA (1987) Measurement Error Models, Wiley, NewYork.
4. Carrol RJ, Rupert D, Stefanski LA (1995) Measurement Error in Nonlinear Models, Chapman and Hall, London.

5. Wang N, Lin X, Gutierrez RG, Carrol RJ (1998) Bias analysis and SIMEX approach in generalized linear mixed measurement error models, J. Am. Statist. Assoc. 93, pp. 249–261.
6. Cox DR (1972) Regression models and life-tables (with discussion), J. R. Stat. Soc. Ser. B 34, pp. 187–220.
7. Cox DR, Oakes D (1984) Analysis of Survival Data, Chapman and Hall, London.
8. Kalbfleisch JD, Prentice RL (2002) The Statistical Analysis of Failure Time Data, Second Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA.
9. Hu P, Tsiatis AA, Davidian M (1998) Estimating the parameters in the Cox model when covariate variables are measured with error, Biometrics 54, pp. 1407–1419.
10. Hu C, Lin D (2002) Cox regression with covariate measurement error, Scandinavian Journal of Statistics 29, pp. 637–655.
11. Zhou H, Pepe MS (1995) Auxiliary covariate data in failure time regression analysis, Biometrika, 82, pp. 139–149.
12. Zhou H, Wang CY (2000) Failure Time Regression with Continuous Covariates Measured with Error. Journal of the Royal Statistical Society. Series B (Statistical Methodology) Vol. 62, No. 4, pp. 657–665.
13. Liu Y, Zhou H, Cai J (2009) Estimated pseudopartial-likelihood method for correlated failure time data with auxiliary covariates, Biometrics, 65, pp. 1184–1193.
14. Fan Z, Wang X (2009) Marginal hazards model for multivariate failure time data with auxiliary covariates, Journal of Nonparametric Statistics, 21: 7, pp. 771–786.
15. Liu Y, Wu Y, Zhou H (2010) Multivariate failure time regression with a continuous auxiliary covariate, Journal of Multivariate Analysis, 101, pp. 679–691.
16. He W, Yi GC, Xiong J (2007) Accelerated failure time models with covariates subject to measurement error, Statist. Med., 26, pp. 4817–4832.
17. Yu M, Nan B (2010) Regression Calibration in Semiparametric Accelerated Failure Time Models, Biometrics 66, pp. 405–414.
18. Granville K, Fan Z (2012) Accelarated Failure Time Models with Auxiliary covariates, J Biom Biostat 3: 152. doi:10.4172/2155-6180.1000152.
19. Buckley J, James I (1979) Linear regression with censored data, Biometrika 66 (3), pp. 429–436.
20. Jin Z, Lin DY, Ying Z (2006) On least-squares regression with censored data, Biometrika 93 (1), pp. 147–161.
21. James IR, Smith PJ (1984) Consistency results for linear regression with censored data, The Annals of Statistics Vol. 12, No. 2, pp. 590–600.
22. Lai TL, Ying Z (1991) Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, The Annals of Statistics Vol. 19, No. 3, pp. 1370–1402.
23. Nadaraya TA (1964) On estimating regression, Theory Probab. Applic. 10, pp. 186–190.
24. Watson GS (1964) Smooth Regression Analysis, Sankhyā: The Indian Journal of Statistics, Ser. A 26, pp. 359–372.
25. Wand M, Jones M (1995) Kernel Smoothing, Chapman and Hall, London.
26. Fleming TR, Harrington DP (1991) Counting Processes and Survival Analysis, Jone Wiley & Sons, Inc. New York.
27. National Institutes of Health (2011) "Primary Biliary Cirrhosis." National Digestive Diseases Information Clearinghouse (NDDIC). December 2008. National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health. 29 July 2011. Available: http://digestive.niddk.nih.gov/ddiseases/pubs/primarybiliarycirrhosis/. Accessed 2014 Jul 25.
28. "Primary Biliary Cirrhosis (PBC)." American Liver Foundation. 22 March 2011. 29 July 2011. Available: http://www.liverfoundation.org/abouttheliver/info/pbc/. Accessed 2014 Jul 25.