Regression Calibration in Semiparametric Accelerated Failure Time Models

Menggang Yu1∗ and Bin Nan2

1Department of Medicine, Division of Biostatistics, Indiana University School of Medicine, 410 West 10th Street, Suite 3000, Indianapolis, Indiana 46202, USA
2Department of Biostatistics, University of Michigan, Ann Arbor, 1420 Washington Heights, Ann Arbor, Michigan 48109, USA
∗email: meyu@iupui.edu

SUMMARY. In large cohort studies, it often happens that some covariates are expensive to measure and hence only measured on a validation set. On the other hand, relatively cheap but error-prone measurements of the covariates are available for all subjects. Regression calibration (RC) estimation method (Prentice, 1982, *Biometrika* 69, 331–342) is a popular method for analyzing such data and has been applied to the Cox model by Wang et al. (1997, *Biometrics* 53, 131–145) under normal measurement error and rare disease assumptions. In this article, we consider the RC estimation method for the semiparametric accelerated failure time model with covariates subject to measurement error. Asymptotic properties of the proposed method are investigated under a two-phase sampling scheme for validation data that are selected via stratified random sampling, resulting in neither independent nor identically distributed observations. We show that the estimates converge to some well-defined parameters. In particular, unbiased estimation is feasible under additive normal measurement error models for normal covariates and under Berkson error models. The proposed method performs well in finite-sample simulation studies. We also apply the proposed method to a depression mortality study.

KEY WORDS: Accelerated failure time model; Measurement error; Regression calibration; Two-phase sampling; Validation data.

1. Introduction

In large cohort studies for failure time regression analysis, we often measure many variables. Among them some are easy to obtain while others may be costly to measure. One such example is a depression mortality cohort study conducted in two phases at the Indiana University School of Medicine (Callahan et al., 1998). Phase 1 of the study took place from January 1991 to June 1993 and the whole cohort was screened for depression, dementia, and alcoholism. Phase 2 of the study was conducted following phase 1 on a subsample of the cohort. Patients were selected based on simple random sampling stratified on their depression score measured by the Center for Epidemiologic Studies Depression scale (CES-D). During the second phase, patients were required to complete an extended interview that consisted of the Hamilton Depression Rating Scale (HAM-D) and part of the Sickness Impact Profile (SIP). HAM-D offers higher validity and reliability than CES-D in measuring response to treatment of depression and SIP measures the effect depressive symptoms had on physical and psychosocial functioning of study subjects. Both require more time and effort to collect than CES-D. A question of interest is to determine whether patients’ depression and quality of life level are associated with mortality.

This example can be viewed as a survival analysis problem with covariates subject to measurement error. HAM-D and SIP scores are better measurements to reflect the true morbidity status but are only available for a subset of the cohort (called a validation subset) while CES-D score was measured on all subjects and viewed as a surrogate for HAM-D and SIP. The mortality information was also available on all subjects. The question of interest is to develop an effective way to deal with measurement error or surrogate covariates in the survival analysis framework.

In survival analysis, Cox model is a popular choice. Direct application of the Cox model to the above data detected significant effects of CES-D score, gender, race, and age on survival. However, the proportional hazard assumption is in question. Specifically, the corresponding p-values from the tests of the proportional hazard assumption based on Schoenfeld residuals (Grambsch and Therneau, 1994) are 0.090, 0.0005, 0.187, 0.333, and 0.019 for the four covariates and the overall model, respectively. Because the CES-D scores are considered as surrogates, we also fit the Cox model using HAMD&SIP, gender, race, and age as covariates where HAMD&SIP is sum of the HAM-D and SIP and used as index for a subject’s morbidity level. We assign same weights to HAM-D and SIP scores, we use imputed values via a linear regression (see Section 5.2 for more details). Under such setting, these
p-values from the tests of the proportional hazard assumption now become 0.133, 0.007, 0.213, 0.314, and 0.026. Still the proportional hazard assumption is violated.

This leads us to consider alternative models for analysis. Semiparametric accelerated failure time (AFT) model is a good alternative (Kalbfleisch and Prentice, 2002, Chapter 7). The linear structure of the AFT model and the nice interpretation due to its direct modeling of survival time is especially appealing to medical investigators. Moreover, in our initial model fitting of the same data set, we found that the semiparametric AFT model appears to provide a good fit of the data. The empirical p-values for goodness-of-fit check are 0.692, 0.104, 0.828, and 0.254 for (imputed) HAMD&SIP, gender, race, and age respectively, based on 500 simulated marginal residual score processes (Lin, Robins, and Wei, 1996; Peng and Fine, 2006). Figure 1 provides plots of 50 realizations of the score processes (dashed lines) together with the observed score processes (solid lines). Hence, we consider dealing with measurement error data under the semiparametric AFT model framework.

By viewing measurement error as a missing data problem, the inverse-probability-weighted (IPW) method can be applied for the semiparametric AFT model to deal with covariates subject to measurement error (Yu and Nan, 2006; and Nan, Kalbfleisch, and Yu, 2009). However, the IPW method uses outcome or failure time data from the validation subset only. As we have failure time information for all subjects in the cohort, this can lead to some loss of information. Augmented inverse-probability-weighted (AIPW) estimator has been proposed in the Cox model for censored data (Wang and Chen, 2001; Qi, Wang, and Prentice, 2005) and uncensored data (e.g., Wang and Wang, 2001). Although the class of AIPW estimators includes the semiparametric efficient estimator (Robins, Rotnitzky, and Zhao, 1994), implementation can be extremely challenging and usually involves solving integral equations (Nan, 2004). As a popular alternative, the usual AIPW estimating equation augments IPW equation by conditioning on observed data to reduce computational burden. As observed by Qi et al. (2005, Theorem 2–4) and Wang and Wang (2001, Theorem 1 and 2), such AIPW estimators are asymptotically equivalent to IPW estimators with estimated weights. The improvement in efficiency may be limited.

Corrected score method was proposed for the Cox model by Nanamura (1992) and studied by many others (e.g., Hu and Lin, 2002; Huang and Wang, 2000). These correction methods rely on approximating average terms among risk sets using measurement error data. Such methods may be hard to apply to the semiparametric AFT model setting where risk sets are defined on the residual scale that depends on observed failure time, mismeasured or missing covariates, and the true value of regression coefficients. The residual is defined as the difference between the failure time and its regression mean. In contrast, in the Cox regression, risk sets are usually defined by the difference between the failure time and its regression mean. In linear regression models, mismeasured or missing covariates, and the true value of the error part of linear regression models. As semiparametric AFT models are linear regression models with censored data (see (1) below), this implies that the RC method can also be an intuitively good candidate for them. Carroll et al. (2006, Chapter 4) has provided a comprehensive review of the RC method for analyzing data without censoring. In general, the RC method does not always lead to consistent estimates. So it is an approximate method. On the other hand, nice finite-sample performance has been observed in various applications that has contributed to its popularity especially among epidemiologists (Carroll et al., 2006, Chapter 4).

We also study the stratified random sampling scheme as in our motivating example for selecting the validation sample. This design results in neither independent nor identically distributed observations. Consequently, our development of asymptotic properties of the RC method rely on results for exchangeable random variables. Another sampling method, the i.i.d. Bernoulli sampling, may also be applied as in Wang et al. (1997). Although simpler and analogous results can be obtained from the i.i.d. Bernoulli sampling scheme, we consider the more general stratified random sampling scheme that is known to be more popular in practice.

2. Method
Let \( T^* \) and \( C^* \) be log-transformed failure and censoring times for subject \( i \) in a cohort of size \( N \). We observe \( T_i \equiv T^*_i \wedge C^*_i \), a failure indicator \( \Delta_i \equiv I\{T_i \leq C_i\} \), and a \( d \)-dimensional vector of covariates \( Z_i \). The semiparametric AFT model postulates

\[
T_i = \beta'Z_i + e_i, \quad i = 1, \ldots, N, \tag{1}
\]

where \( e_i \)'s are independent and identically distributed with an unknown distribution \( F_0 \) and density function \( f_0 \). Censoring time \( C_i \) is assumed to be independent of \( e_i \) conditioning on \( Z_i \). This model has been discussed extensively in the literature (e.g., Kalbfleisch and Prentice, 2002, Chapter 7) when \( Z_i \) are measured without error.

Let \( e^i = T_i - \beta'Z_i \) be the observed residual for subject \( i, \hat{N}_i(t, \beta) = I \{e_i^\beta \leq t, \Delta_i = 1\} \) the residual counting...
process, and $Y_i(t, \beta) = I(t^\beta_i \geq t)$ the residual at-risk process. When $(Z_i, T_i, \Delta_i)$ are observed for the entire cohort, Tsiatis (1990) introduced the following rank-based estimating function for $\beta_0$,

$$S_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} \int \rho_N(t, \beta) \left\{ Z_i - \bar{Z}(t, \beta) \right\} dN_i(t, \beta),$$  

(2)

where $\rho_N(t, \beta)$ is a weight function and $\bar{Z}(t, \beta) \equiv D_N^{(1)}(t, \beta)/D_N^{(0)}(t, \beta)$ with $D_N^{(0)}(t, \beta) = N^{-1} \sum_{j=1}^{N} Y_j(t, \beta)$ and $D_N^{(1)}(t, \beta) = N^{-1} \sum_{j=1}^{N} Z_j Y_j(t, \beta)$.

As in the depression mortality study, we consider the situation when part of $Z_i$ is subject to measurement error or missing. We also assume that there is an internal validation subset (Carroll et al., 2006, Chapter 4) on which $Z_i$ are accurately measured. Following the notation of Breslow and Wellner (2007), we focus on the situation when the validation subset is selected via the following two-phase sampling scheme. Let $Z = (Z_1', Z_2')'$ with $Z_1$ subject to measurement error and $Z_2$ always observable. At the first phase, we observe $V = (W, Z_2')' \in \mathcal{V}$ where $W$ is a surrogate for $Z_1$. At the second phase, a subsample is selected and $Z_1$ is fully observed for these subjects. The subset is called a validation set because both $Z_1$ and $W$ are observed. For subjects not in the validation set, $Z_1$ is missing. A popular selection scheme is stratified simple random sampling. Specifically, $\mathcal{V}$ is partitioned into $J$ strata, or $\mathcal{V} = \mathcal{V}_1 \cup \cdots \cup \mathcal{V}_J$. Within each stratum $\mathcal{V}_j$ of size $N_j$, a simple random sample of size $n_j \leq N_j$ are randomly drawn without replacement for $1 \leq j \leq J$. We use a doubly subscripted notation $\eta_{j,i}$ to indicate the $i$th subject in stratum $j$, where $\eta_{j,i} = 1$ if the subject is selected.

Figure 1. Goodness of fit for the depression mortality data.
into the validation set and 0 otherwise. Note that the random variables \((\eta_{j,1}, \ldots, \eta_{j,N})\) are exchangeable and the \(J\) random vectors \((\eta_{j,1}, \ldots, \eta_{j,N})\) are independent. We further assume that censoring variable \(C\) is independent of \(T\) given observed covariate data \(V\). In the depression mortality study, \(Z_1\) contains the HAM-D and SLP scores, \(Z_2\) contains covariates such as gender, race, and age, whereas \(W\) contains CES-D scores. CES-D scores were stratified into two categories at the first phase: one with scores 16 or higher and the other with scores below 16. Simple random samples were then selected from both the categories in the second phase.

Assume that the conditional mean of \(Z_{1i}\) given \(Z_{2i}\) and \(W_i\) can be modeled by \(E(Z_{1i} | Z_{2i}, W_i) = g(Z_{2i}, W_i, \alpha)\) for some nuisance parameter \(\alpha\) of dimension \(q\). The nuisance parameter \(\alpha\) is usually unknown and can be estimated using the validation data. Now denote \(Z_{1i}^{(\alpha)}(\alpha) = (Z_{1i}, \eta_{j,i} + (1 - \eta_{j,i}) g(Z_{2i}, W_{j,i}, \alpha), Z_{2j,i})\). Let \(\epsilon_{j,i} = T_{j,i} - \beta_{i}Z_{2j,i}N_{j,i}(t, \beta) = I(\epsilon_{j,i}^2 \leq t, \Delta_{j,i} = 1), Y_{j,i}(t, \beta) = I(\epsilon_{j,i}^2 \geq t)\) be the residual, its counting process, and at-risk process for subjects with complete data. Let \(\beta_{j,i}^{(\alpha)} = T_{j,i} - \beta_{i}g(Z_{2j,i}, W_{j,i}, \alpha)\) be the residual, its counting process, and at-risk process for at-risk process when \(Z_{1i}\) is replaced by \(g(Z_{2j,i}, W_{j,i}, \alpha)\). Denote

\[
D_N^{(0)}(t, \beta, \alpha) = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \left( \eta_{j,i}Y_{j,i}(t, \beta) + (1 - \eta_{j,i})Y_{j,i}(t, \beta, \alpha) \right).
\]

\[
D_N^{(1)}(t, \beta, \alpha) = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \left( \eta_{j,i}Z_{j,i}Y_{j,i}(t, \beta) + (1 - \eta_{j,i})Z_{j,i}(\alpha)Y_{j,i}(t, \beta, \alpha) \right).
\]

If \(\alpha\) is known, an RC-based estimating function similar to (2) can be formed as

\[
S_N(\beta, \alpha) = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \int \left\{ D_N^{(0)}(t, \beta, \alpha)Z_{j,i} - D_N^{(0)}(t, \beta, \alpha) \right\} dN_{j,i}(t, \beta)
+ \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} (1 - \eta_{j,i}) \int \left\{ D_N^{(0)}(t, \beta, \alpha)Z_{j,i}(\alpha) - D_N^{(0)}(t, \beta, \alpha) \right\} dN_{j,i}(t, \beta, \alpha)
\]

where Gehan’s weight, \(D_N^{(1)}(t, \beta, \alpha)/D_N^{(0)}(t, \beta, \alpha)\), is used for the weighting process \(p_{\nu}\).

We focus on the above Gehan weighted estimating function in this article mainly for notational simplicity. However, our theoretical results should hold for more general weight functions. Our choice is based on two advantages that Gehan weighted estimating function has. First, \(S_N(\beta, \alpha)\) is a monotone function (Fygenson and Ritov, 1994; Yu and Nan, 2006), which is a very nice property to have for root finding. Second, the estimating function does not involve fractions. Consequently, the condition of truncating the stochastic integration (see 3.1 of Tsiatis, 1990) is not needed in our investigation of the asymptotic properties of \(S_N(\beta, \alpha)\). This not only prevents potential information loss due to truncation (Ying, 1993) but also provides less restriction, especially in the presence of measurement error data.

Since \(\alpha\) is unknown, we replace it in (5) by an estimate \(\hat{\alpha}\), which is usually obtainable using existing software based on the validation data. Consequently, we propose using \(S_N(\beta, \hat{\alpha}) = 0\) as our estimating equation for \(\beta\), based on the RC method. To obtain an estimate for \(\beta\), we replace missing \(Z_{1}\) with \(g(Z_{2j}, W, \hat{\alpha})\) and perform a full cohort AFT model analysis to obtain a valid regression estimate \(\hat{\beta}_N\). Despite the nonsmooth nature of \(S_N(\beta, \hat{\alpha})\) in \(\beta\), \(\hat{\beta}_N\) can be readily obtained using existing computation method in full cohort AFT models (e.g., Lin and Geyer, 1992; Jin et al., 2003; and Yu and Nan, 2006).

3. Asymptotic Properties

In this section, we explore asymptotic properties of our proposed estimating function \(S_N(\beta, \hat{\alpha})\) and its corresponding estimate \(\hat{\beta}_N\). We show that \(\hat{\beta}_N\) converges to some well-defined parameter that may or may not be equal to the true parameter. When the magnitude of regression coefficient or the measurement error is small, however, bias is in general very limited (see the comment above Proposition 2). Improvement in efficiency on the other hand can be quite remarkable compared with analysis using only validation data. This will be demonstrated numerically in Section 5.

In Proposition 1, we show uniform convergence of \(D_N^{(0)}(t, \beta, \alpha), D_N^{(1)}(t, \beta, \alpha),\) and \(S_N(\beta, \alpha)\) to their limits:

\[
d^{(0)}(t, \beta, \alpha) = \sum_{j=1}^{J} \nu_j \left[ p_j P_{0|j} \{ \epsilon^2 \geq t \} \right]
+ (1 - p_j) P_{0|j} \{ \epsilon^2 < t \} \right]
\]

\[
d^{(1)}(t, \beta, \alpha) = \sum_{j=1}^{J} \nu_j \left[ p_j P_{0|j} \{ Z(t, \alpha) I(\epsilon^2 \geq t) \} \right]
+ (1 - p_j) P_{0|j} \{ Z(t, \alpha) I(\epsilon^2 < t) \} \right]\]

\[
s(\beta, \alpha) = \sum_{j=1}^{J} \nu_j \left[ \left[ p_j P_{0|j} \{ Z(t, \alpha) I(\epsilon^2 \geq t) \} \right]
+ (1 - p_j) P_{0|j} \{ Z(t, \alpha) I(\epsilon^2 < t) \} \right]\]

Here \(p_j = \eta_{j,i}/N_j, \nu_j\) is the limit of \(N_j/N, P_{0|j}(\cdot) = E(\cdot | V_j \in Y_j)\) is the conditional expectation in stratum \(J, \) under regularity assumptions including the uniqueness of \(\beta^*\) that solves \(s(\beta, \alpha_0) = 0\) at the true \(\alpha_0\). We then have the convergence of \(\hat{\beta}_N\) to \(\beta^*\). In Proposition 2, we show the asymptotic linearity property of \(S_N(\beta, \hat{\alpha})\) and the asymptotic normality of \(\hat{\beta}_N\).

The following regularity conditions are needed for our propositions:
C1. The covariate $Z$ has bounded support.

C2. The number of strata, $J$, is finite and fixed. For any stratum $j$, $v_j > 0$ and $\sqrt{N_j(N_j - v_j)} = O_p(1)$.

C3. For any given set $A \subset \mathbb{R}^d$,

i. The conditional mean of $E(Z_1 | Z_2, W) = g(Z_2, W, \alpha)$ is continuous with respect to $\alpha \in A$ and uniformly bounded.

ii. The class $\{g(Z_2, W, \alpha), \alpha \in A\}$ forms a P-Donsker class.

C4. The true value $\alpha_0$ is an interior point of $A$ such that $\sqrt{N(\alpha - \alpha_0)}$ is asymptotically normal with mean 0.

C5. For each stratum $j$, both $P_{0|j}(e_i^0 \geq t)$ and $P_{0|j}(e_i^\alpha \geq t)$ are continuous in $\beta, t,$ and $\alpha$. The limit functions $d(0)(\epsilon^0, \beta, \alpha), d(0)(\epsilon^\alpha, \alpha, \beta, \alpha), d(1)(\epsilon^0, \beta, \alpha)$, and $d(1)(\epsilon^\alpha, \alpha, \beta, \alpha)$ are continuously differentiable with respect to $\beta$ and $\alpha$.

C6. The solution to $s(\beta, \alpha_0) = 0$ for $\beta$ is unique and is an interior point of $B$, where $B \subset \mathbb{R}^d$ is a compact set of possible parameter values.

Condition C1 commonly holds. Condition C2 is easily satisfied in our case, where the strata are determined by $V$ and $N_j$ follows a binomial distribution with success probability $v_j = P(V \in V_j)$. For Condition C3, suppose $g(Z_2, W, \alpha) = h(\alpha_0 + \alpha_2 Z_2 + \alpha_w W)$, where $h(\cdot)$ is monotone and continuously differentiable, then from Lemma 2.6.15 and 2.6.18 of van der Vaart and Wellner (1996), we know that $h(\alpha_0 + \alpha_2 Z_2 + \alpha_w W): \alpha_0, \alpha_2, \alpha_w \in R \}$ forms a VC-subgraph class. If a VC-subgraph class has an envelope function with finite second moment, it is P-Donsker. This includes generalized linear models for the conditional mean of $E(Z_1 | Z_2, W) = g(Z_2, W, \alpha)$, Condition C4 consists of expected properties for estimates of $\alpha$ using validation data based on stratified simple random sampling, for which various methods have been proposed. Specifically, we have the following asymptotic representation under the stratified simple random sampling (Breslow and Wellner, 2007):

$$\hat{\alpha} - \alpha = N^{-1/2} \sum_{j=1}^{J} \left( \sum_{i=1}^{N_j} \eta_{i,j} e_{i,j} V_{j,i} \right) + o_p(N^{-1/2}),$$

where $H$ is the Hessian matrix, $e_{i,j} = Z_{1,i} - g(Z_{2,i}, W_{j,i}, \alpha)$ is the residual and $V_{j,i} = (Z_{2,i}, W_{j,i})$. Condition C5 is likely to hold in situations when both the failure time and censoring time are continuously distributed, given the continuity condition C3.i for $g$. As a result of C5, $s(\beta, \alpha)$ is continuously differentiable in both its arguments. Condition C6 is essential for the consistency of $\beta_0$.

Proposition 1 provides convergence results of $S_N(\beta, \alpha)$ and Proposition 2 provides the asymptotic linearity of $S_N(\beta, \alpha)$ in an $N^{-1/2}$ neighborhood of $\beta$ and the asymptotic normality of $\beta_0$. We delegate the proofs of the propositions to the Web Appendix due to their length.

**Proposition 1:** Assume that the number of strata $J$ is finite. Let $d(0)(t, \beta, \alpha), d(1)(t, \beta, \alpha),$ and $s(\beta, \alpha)$ be defined as in (6), (7), and (8) correspondingly.

1. Under Conditions C1–C5, we have

$$\sup_{\beta \in B, \alpha \in A, t \in R} \left| D_N(t, \beta, \alpha) - d(0)(t, \beta, \alpha) \right| = o_p(1),$$

$$\sup_{\beta \in B, \alpha \in A, t \in R} \left| D_N(t, \beta, \alpha) - d(1)(t, \beta, \alpha) \right| = o_p(1).$$

Here $p^*$ denotes in outer probability. See Chapter 1 of van der Vaart and Wellner (1996) for the definition of outer probability.

2. Let $\| \cdots \|$ denote the super-norm. Let $\alpha_1, \alpha_2 \in A$ such that $\| \alpha_1 - \alpha_2 \| \to 0$ and $\beta_1, \beta_2 \in B$ such that $\| \beta_1 - \beta_2 \| \to 0$ as $N \to \infty$. Under Conditions C1–C5,$$

\sup_{\beta \in B} \sqrt{N} \left| \{ S_N(\beta, \alpha_1) - S_N(\beta, \alpha_2) \} \right| = o_p(1).

$$

3. If further the condition C6 holds, then the solution $\beta_N$ of $S_N(\beta, \alpha)$ converges to $\beta^*$ with probability $I$ where $\beta^* \in B$ solves $s(\beta, \alpha_0) = 0$.

Note that $\beta^*$ may not be equal to $\beta_0$. In such case, the estimate resulting from the RC method can be at the best viewed as an approximation. In the following two important cases of measurement error models, however, consistent estimation (i.e., $\beta^* = \beta_0$) is obtainable.

The first case is under the general Berkson measurement error model that specifies the conditional distribution of $Z_1$, given $W_1$ and $Z_2$. One such example assumes a linear structure between $Z_1$ and $(W_1, Z_2), Z_1 = \psi_0 + \psi_1 W_1 + \psi_2 Z_2 + U,$ where the error $U$ is independent of $(W_1, Z_2)$. Note that in this case, $T_1$ and imputed covariate vector $(g(Z_2, W_1, \alpha), Z_2)$ still follow a semiparametric AFT model. This can be seen by writing $T_1 = \beta_1 E[Z_1 | W, Z_2] + \beta_2 Z_2 + \beta_3 U + e.$ Hence, if we denote $e^* = \beta_3 U + e$, then $e^*$ is independent of $E[Z_1 | W, Z_2]$ and $Z_2$. This implies $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*.$

The second case is under the additive measurement error model when both $Z$ and its surrogate $W$ are jointly normally distributed. The additive measurement error model specifies a linear structure between $W$ and $(Z_1, Z_2), W = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + U,$ where the error $U$ is independent of $(Z_1, Z_2)$. In this case, we can rewrite $T_1^* = \beta_1 E[Z_1 | W, Z_2] + \beta_2 Z_2 + \beta_1 (Z_1 - E[Z_1 | W, Z_2]) + e.$ Then, due to the joint normality and the fact that $(Z_1 - E[Z_1 | W, Z_2])$ is uncorrelated with both $E[Z_1 | W, Z_2]$ and $Z_2$, we conclude that $(Z_1 - E[Z_1 | W, Z_2])$ is also independent of both $E[Z_1 | W, Z_2]$ and $Z_2$. As a result, we can denote $e^* = \beta_1^* (Z_1 - E[Z_1 | W, Z_2]) + e.$ Hence $T_1^*$ is still linearly related to $E[Z_1 | W, Z_2]$ and $Z_2$, and $e^*$ can be treated as an error term independent of $E[Z_1 | W, Z_2]$ and $Z_2$. Again we have $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*.$

In general cases, we may have asymmetric bias. However such bias is likely to be small when the variation of the term $\beta_1 (Z_1 - E[Z_1 | W, Z_2])$ is relatively small compared to the variation of $e$. This can happen if either $|\beta_1|$ is small or the measurement error is small.
Proposition 2: Assume conditions C1–C6 hold.

1. For any $\alpha \in A$, $S_N(\beta, \alpha)$ is asymptotically linear in an $N^{-1/2}$ neighborhood of $\beta^*$. In other words, if $\sqrt{N}|\beta_N - \beta^*| = O_p(1)$, then
   \[
   \sqrt{N}\{S_N(\beta_N, \alpha) - S_N(\beta^*, \alpha)\} = \sqrt{N}(\beta_N - \beta^*)\hat{s}_N(\beta^*, \alpha) + o_p(1),
   \]
   where $\hat{s}_N(\cdot, \cdot)$ denotes the partial derivative of $s$ with respect to $\beta$.

2. If $S_N(\beta_N, \alpha) - S_N(\beta^*, \alpha)$ is asymptotically equivalent to $\sqrt{N}\{S_N(\beta_N, \alpha_0) - S_N(\beta^*, \alpha_0)\}$ in the sense that
   \[
   \sqrt{N}|\{S_N(\beta_N, \alpha) - S_N(\beta_N, \alpha_0)\} - \{S_N(\beta_N, \alpha_0) - S_N(\beta^*, \alpha_0)\}| = o_p(\sqrt{N}|\beta_N - \beta^*|).
   \]
   As a result, $S_N(\beta, \alpha)$ is asymptotically linear in an $N^{-1/2}$ neighborhood of $\beta^*$. That is, for any $\beta_N \in B$ such that $\sqrt{N}|\beta_N - \beta^*| = O_p(1)$,
   \[
   \sqrt{N}\{S_N(\beta_N, \alpha) - S_N(\beta^*, \alpha)\} = \sqrt{N}(\beta_N - \beta^*)\hat{s}_N(\beta^*, \alpha_0) + o_p(1).
   \]

3. If $\hat{s}_N(\beta^*, \alpha_0)$ is nonsingular, then $\sqrt{N}(\beta_N - \beta^*)$ is asymptotically normal with mean 0.

4. Variance Estimation

From Proposition 2, we see that estimation of the asymptotic variance of $\beta_N$ requires estimation of the asymptotic slope of $S_N(\beta, \alpha)$, which involves estimation of both $f(\cdot)$ and censoring distribution. It is known to be hard to estimate (e.g., Li and Geyer, 1992; Jin et al., 2003). Here, we adopt the method proposed by Huang (2002). Similar algorithms have been used in Nan, Yu, and Kalbfleisch (2006) and Yu and Nan (2006). Numerical stability in variance estimation has been observed from these papers. Specifically, we follow the following three steps.

1. Decompose the estimated asymptotic variance of $N^{-1/2}S_N(\beta, \alpha)$ into $CC^T$, where $C = (c_1, \ldots, c_d)$ and $C^T$ is the transpose of $C$.
2. Solve equations $N^{-1/2}S_N(\beta_j, \alpha) = c_j$ for $\beta_j$, $j = 1, \ldots, d$.
3. Let $D = (\beta_1 - \beta_N, \ldots, \beta_d - \beta_N)$. Then, $NDD^T$ is a consistent variance estimator of $N^{1/2}(\beta_N - \beta^*)$.

For step 2, an effective hybrid Newton method proposed by Yu and Nan (2006) can be applied. Step 1 requires estimation of the asymptotic variance of $N^{-1/2}S_N(\beta, \alpha)$. We delegate the estimation procedure to Web Appendix B. The essence is to express $\sqrt{N}S_N(\beta, \alpha)$ asymptotically as a sum of mutually independent empirical processes. Then, its asymptotic variance can be estimated by using sample estimates corresponding to these empirical processes.

5. Numerical Examples

5.1 Simulations

For simplicity, we consider one-dimensional $Z$ and its surrogate $W$. Sample size $N$ is chosen to be 300, 1000, or 3000. Validation sample size $n$ is chosen to be 90 or 120 when $N = 300, 200, or 400 when $N = 1000$, and 400 or 800 when $N = 3000$. We generate 500 data sets under each of these scenarios.

The AFT model is log $T = \beta Z + \epsilon$ where the true $\beta_1 = 1$. Error distribution is from uniform UNIF$(-3, 3)$. Censoring time $C$ is distributed according to a shifted exponential distribution with its shift and scale chosen so that the censoring rate is approximately 70%. For the covariate $Z$, we consider two distributions: standard normal and Bernoulli. Under the first setting when $Z$ is continuous, we generate the surrogate $W$ according to $W = Z + U$ such that $U \sim N(0, \sigma_W)$. Under the second setting when $Z$ is binary, we generate the surrogate $W$ so that it is a misclassification of $Z$ via concordance probability $p_W = P(W = 1 | Z = 1) = P(W = 0 | Z = 0)$. The conditional mean is estimated via least square regression under the first setting and via sampling frequency under the second setting.

The validation sets are selected via stratified simple random sampling. The stratification is based on whether $W$ is less than 0.5 or not under the first scenario, and whether $W$ is 0 or not under the second scenario. We perform four analyses on each simulated data set to compare its performances under various scenarios. The analyses are: full cohort analysis (Full) where all data are used; validation data only analysis (VA) where only observations in the validation subset are used; naive analysis (Naive) where missing $Z$s are replaced by their surrogates $W$; and the RC analysis (RC).

Results are listed in Tables 1 and 2. We let $\sigma_W$ and $p_W$ vary to investigate how the quality of the surrogate affects the estimation. From the simulation results, we see that full cohort estimates are the most efficient as one would expect. VA estimates are consistent but not very efficient unless the validation size is large. The consistency of the VA method is due to the fact that sampling is based only on covariates that do not change the conditional distribution of the outcome. The Naive method clearly leads to severe bias. On the other hand, RC estimates have very limited bias and very impressive efficiency improvement, particularly when the validation size increases or when measurement error or classification error diminishes. It also has smaller standard error than the VA analysis under all scenarios.

We also carry out simulation studies to compare the IPW and RC methods under the same settings. From Table 3, we observed a significant amount of reduction in empirical mean square errors of the RC estimates in almost all cases. So the RC method can be an attractive alternative to the IPW method.

5.2 The Depression Mortality Study

We analyze the depression mortality study introduced at the beginning of the article. Among 3767 patients screened at the phase 1 of the study, 3155 patients scored below 16 on the CES-D. Among them a simple random sample of 100 was selected. Among 612 patients who scored 16 or higher on the CES-D, 300 were selected. At the phase 2 of the study, the 400 selected patients were required to complete HAM-D and SIP. There are total 3756 subjects in the analysis data set with seven cases deleted due to missing information. The total number of death is 810 and the median follow-up time is 72 months. We study the impact of depression, age, gender, and race on patient survival. In the original study (Callahan et al., 1998), more covariates were considered. Here, we use a subset...
Semiparametric Accelerated Failure Time Models

Table 1
Simulation results based on 500 data sets under the model \( \log T = \beta Z + \epsilon \) (with \( \beta_0 = 1 \)) where \( Z \) is continuous

<table>
<thead>
<tr>
<th>( \sigma_W ) = 0.5</th>
<th>( \sigma_W ) = 1</th>
<th>( \sigma_W ) = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N = 300, n = 90 )</td>
<td>( N = 300, n = 120 )</td>
<td>( N = 1000, n = 400 )</td>
</tr>
<tr>
<td>Avg ( \hat{\beta} )</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td>SE ( \hat{\beta} )</td>
<td>0.210</td>
<td>0.211</td>
</tr>
<tr>
<td>Avg ( \hat{\sigma}(\hat{\beta}) )</td>
<td>0.209</td>
<td>0.211</td>
</tr>
<tr>
<td>90% CP</td>
<td>88.8%</td>
<td>91.8%</td>
</tr>
<tr>
<td>95% CP</td>
<td>95.0%</td>
<td>93.0%</td>
</tr>
</tbody>
</table>

of variables to illustrate our method. There exists a clearly linear relationship between CES-D and HAMD&SIP (sum of HAM-D and SIP) scores among the 400 validation subjects. The estimated correlation coefficient is 0.58. The estimated reliability ratio is 0.653 (note that the reliability ratios in our simulation study under the normal covariates are 0.8, 0.5, and 0.2, corresponding to \( \sigma_W \) of 0.5, 1, and 2).

In our analysis, we first regress HAMD&SIP on CES-D using the 400 validation subjects. The resulted intercept and slope estimates are (9.78, 0.52) with standard errors (0.84, 0.04). Then, we apply the RC method. Both the results of RC and VA methods are listed in Table 4. In the VA analysis, gender is found to be significantly associated with mortality, whereas in the RC analysis, age is found to be significantly associated with the mortality. Note that the coefficients of HAMD&SIP, FEMALE, and AGE have the same direction as those based on the VA analysis, but the sign of the coefficient of BLACK is reversed.

6. Discussion

We have proposed a RC method for estimating parameters in semiparametric AFT models with covariates subject to measurement error. The method performs well in simulations and has clear advantage over ad hoc approaches. As our asymptotic results are derived under very general assumptions between the true covariate vector \( Z_t \) and its surrogate \( W \), this method is also applicable for survival data with missing covariates as long as reliable calibration between missing and observed covariates is feasible. In this case, the RC method is a mean imputation method.
The projection technique studied by Chen (2002) for the Cox model and by Jiang and Zhou (2007) for the additive hazards model should be applicable in our setting. The projection estimator is generally consistent and is guaranteed to have multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

The projection technique is studied by Chen (2002) for the Cox model and by Jiang and Zhou (2007) for the additive hazards model should be applicable in our setting. The projection estimator is generally consistent and is guaranteed to have multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

Due to the limitation of space, we focus in this article on additive measurement error models. Multiplicative errors also arise frequently in practice. While a simple log transformation can convert multiplicative errors to additive errors, sometimes special treatment is needed. We refer readers to Section 4.5 of Carroll et al. (2006) for performing RC methods with multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

Table 2

<table>
<thead>
<tr>
<th>$p_W = 0.9$</th>
<th>$p_W = 0.7$</th>
<th>$p_W = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N = 300, n = 90</strong></td>
<td><strong>N = 300, n = 120</strong></td>
<td><strong>N = 1000, n = 200</strong></td>
</tr>
<tr>
<td><strong>Avg $\hat{\beta}$</strong></td>
<td><strong>Avg $\hat{\beta}$</strong></td>
<td><strong>Avg $\hat{\beta}$</strong></td>
</tr>
<tr>
<td>1.001</td>
<td>1.036</td>
<td>1.005</td>
</tr>
<tr>
<td><strong>SE $\hat{\beta}$</strong></td>
<td><strong>SE $\hat{\beta}$</strong></td>
<td><strong>SE $\hat{\beta}$</strong></td>
</tr>
<tr>
<td>0.448</td>
<td>0.444</td>
<td>0.272</td>
</tr>
<tr>
<td><strong>Avg $\hat{\sigma}(\hat{\beta})$</strong></td>
<td><strong>Avg $\hat{\sigma}(\hat{\beta})$</strong></td>
<td><strong>Avg $\hat{\sigma}(\hat{\beta})$</strong></td>
</tr>
<tr>
<td>0.455</td>
<td>0.456</td>
<td>0.237</td>
</tr>
<tr>
<td><strong>90% CP</strong></td>
<td><strong>90% CP</strong></td>
<td><strong>90% CP</strong></td>
</tr>
<tr>
<td>91.2%</td>
<td>91.4%</td>
<td>92.4%</td>
</tr>
<tr>
<td><strong>95% CP</strong></td>
<td><strong>95% CP</strong></td>
<td><strong>95% CP</strong></td>
</tr>
<tr>
<td>94.6%</td>
<td>96.0%</td>
<td>95.8%</td>
</tr>
</tbody>
</table>

Due to the limitation of space, we focus in this article on additive measurement error models. Multiplicative errors also arise frequently in practice. While a simple log transformation can convert multiplicative errors to additive errors, sometimes special treatment is needed. We refer readers to Section 4.5 of Carroll et al. (2006) for performing RC methods with multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

The projection technique studied by Chen (2002) for the Cox model and by Jiang and Zhou (2007) for the additive hazards model should be applicable in our setting. The projection estimator is generally consistent and is guaranteed to have multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

Due to the limitation of space, we focus in this article on additive measurement error models. Multiplicative errors also arise frequently in practice. While a simple log transformation can convert multiplicative errors to additive errors, sometimes special treatment is needed. We refer readers to Section 4.5 of Carroll et al. (2006) for performing RC methods with multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

The projection technique studied by Chen (2002) for the Cox model and by Jiang and Zhou (2007) for the additive hazards model should be applicable in our setting. The projection estimator is generally consistent and is guaranteed to have multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

Due to the limitation of space, we focus in this article on additive measurement error models. Multiplicative errors also arise frequently in practice. While a simple log transformation can convert multiplicative errors to additive errors, sometimes special treatment is needed. We refer readers to Section 4.5 of Carroll et al. (2006) for performing RC methods with multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

The projection technique studied by Chen (2002) for the Cox model and by Jiang and Zhou (2007) for the additive hazards model should be applicable in our setting. The projection estimator is generally consistent and is guaranteed to have multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.

Due to the limitation of space, we focus in this article on additive measurement error models. Multiplicative errors also arise frequently in practice. While a simple log transformation can convert multiplicative errors to additive errors, sometimes special treatment is needed. We refer readers to Section 4.5 of Carroll et al. (2006) for performing RC methods with multiplicative error models. Extension to semiparametric AFT models may be carried out similarly and warrants further investigation.
Table 3
Simulation results comparing the IPW and RC methods

<table>
<thead>
<tr>
<th>$Z$ continuous</th>
<th>$\sigma_W = 0.5$</th>
<th>$\sigma_W = 1$</th>
<th>$\sigma_W = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = 1000, n = 200$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg $\hat{\beta}$</td>
<td>1.008</td>
<td>0.996</td>
<td>1.007</td>
</tr>
<tr>
<td>SE $\hat{\beta}$</td>
<td>0.263</td>
<td>0.111</td>
<td>0.262</td>
</tr>
<tr>
<td>$N = 1000, n = 400$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg $\hat{\beta}$</td>
<td>0.988</td>
<td>1.005</td>
<td>0.984</td>
</tr>
<tr>
<td>SE $\hat{\beta}$</td>
<td>0.182</td>
<td>0.110</td>
<td>0.175</td>
</tr>
<tr>
<td>$N = 3000, n = 400$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg $\hat{\beta}$</td>
<td>0.998</td>
<td>1.001</td>
<td>0.996</td>
</tr>
<tr>
<td>SE $\hat{\beta}$</td>
<td>0.173</td>
<td>0.105</td>
<td>0.179</td>
</tr>
<tr>
<td>$N = 3000, n = 800$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg $\hat{\beta}$</td>
<td>0.991</td>
<td>0.999</td>
<td>0.995</td>
</tr>
<tr>
<td>SE $\hat{\beta}$</td>
<td>0.127</td>
<td>0.063</td>
<td>0.130</td>
</tr>
<tr>
<td>$Z$ binary</td>
<td>$p_W = 0.5$</td>
<td>$p_W = 0.7$</td>
<td>$p_W = 0.9$</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>$N = 1000, n = 200$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg $\hat{\beta}$</td>
<td>1.010</td>
<td>1.001</td>
<td>1.006</td>
</tr>
<tr>
<td>SE $\hat{\beta}$</td>
<td>0.590</td>
<td>0.531</td>
<td>0.574</td>
</tr>
<tr>
<td>$N = 1000, n = 400$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg $\hat{\beta}$</td>
<td>1.002</td>
<td>1.028</td>
<td>1.007</td>
</tr>
<tr>
<td>SE $\hat{\beta}$</td>
<td>0.390</td>
<td>0.389</td>
<td>0.380</td>
</tr>
<tr>
<td>$N = 3000, n = 400$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg $\hat{\beta}$</td>
<td>1.015</td>
<td>0.945</td>
<td>0.999</td>
</tr>
<tr>
<td>SE $\hat{\beta}$</td>
<td>0.370</td>
<td>0.343</td>
<td>0.362</td>
</tr>
<tr>
<td>$N = 3000, n = 800$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg $\hat{\beta}$</td>
<td>0.983</td>
<td>0.995</td>
<td>0.996</td>
</tr>
<tr>
<td>SE $\hat{\beta}$</td>
<td>0.274</td>
<td>0.261</td>
<td>0.262</td>
</tr>
</tbody>
</table>

Table 4
Results for the depression mortality study

<table>
<thead>
<tr>
<th>Validation only analysis</th>
<th>RC analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
</tr>
<tr>
<td>HAMDS &amp; SIP</td>
<td>-0.007</td>
</tr>
<tr>
<td>FEMALE</td>
<td>0.991</td>
</tr>
<tr>
<td>BLACK</td>
<td>-0.103</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.031</td>
</tr>
</tbody>
</table>

6. Supplementary Materials
Web Appendices referenced in Sections 3 and 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

ACKNOWLEDGEMENTS
The authors would like to thank Dr Christopher M. Callahan for his kind permission to use the depression mortality data and Dr Siu Hui for her insightful discussion during the preparation of this manuscript. The work of Bin Nan was supported in part by NSF grant DMS-07-06700.

REFERENCES


Received September 2008. Revised February 2009. Accepted March 2009.