Nonparametric correction for covariate measurement error in a stratified Cox model

MALKA GORFINE*
Department of Mathematics and Statistics, Bar-Ilan University, Ramat-Gan 52900, Israel

LI HSU, ROSS L. PRENTICE
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA

SUMMARY
Stratified Cox regression models with large number of strata and small stratum size are useful in many settings, including matched case-control family studies. In the presence of measurement error in covariates and a large number of strata, we show that extensions of existing methods fail either to reduce the bias or to correct the bias under nonsymmetric distributions of the true covariate or the error term. We propose a nonparametric correction method for the estimation of regression coefficients, and show that the estimators are asymptotically consistent for the true parameters. Small sample properties are evaluated in a simulation study. The method is illustrated with an analysis of Framingham data.

Keywords: Clayton–Oakes model; Framingham study; Matched case-control family study; Mismeasured covariates; Stratified censored data.

1. INTRODUCTION

A stratified Cox regression model is often useful in biomedical applications. Two important contexts are multicenter clinical trials and matched case-control studies. In a multicenter clinical trial the various centers may have different baseline survival curves. In matched case-control studies, cases and controls are matched on the basis of variables (e.g. neighborhood and age) believed to be associated with the outcome. It is well known that the conditional logistic regression analysis is closely related to the stratified Cox regression analysis when the stratification is on matched set (Prentice and Breslow, 1978; Breslow and Day, 1980, Chapter VII). In these contexts, measurement error may occur in the covariates of main interest. A Cox regression model may also be used to assess the strength of dependence in disease risk between paired relatives in case-control family studies. In this model, the risk factors (e.g. nutrient consumption or physical activity patterns) may be subject to measurement error. However, case/control status is typically known with little or no error.

In follow-up studies it occurs often that some covariate data are not accurately measured. There exists a substantial statistical literature on the effect of the mismeasured covariate on the estimated regression coefficient as well as on various procedures for corrected estimators. Consider the proportional hazards model (Cox, 1972)

\[ \lambda(t|Z) = \lambda_0(t) \exp(\beta^T_0 Z) \]  

(1.1)

*To whom correspondence should be addressed.
where $\lambda(\cdot)$ is a hazard function of a continuously distributed survival time $T$, $Z$ and $\beta_0$ are $p \times 1$ vectors of covariates and corresponding regression coefficient, respectively, and $\lambda_0(\cdot)$ is an unspecified baseline hazard function. Let $C$ be the censoring time, $X = \min(T, C)$, $\delta = I(T \leq C)$ and assume that $C$ and $T$ are independent given $Z$. As usual, assume that the right-censored data consists of iid replicates of $\{X, \delta, Z\}$. Assume that some components of the covariate vector $Z$ are subject to measurement error and consider the following classical measurement error model:

$$W = Z + \epsilon$$  \hspace{1cm} (1.2)

where the errors $\epsilon$ are zero mean variables independent of $Z$, as well as of $X$ and $\delta$. Let the error component be 0 for those covariates that are accurately ascertained.

Prentice (1982) proposed the idea of regression calibration for the proportional hazard model. A regression calibration procedure involves estimating the conditional expectation of the hazard of each individual at each failure time $t$, conditioning on the observed mismeasured covariates and being at risk at $t$, and then applying standard partial likelihood procedures for estimating $\beta_0$. Several corrected estimators have been proposed for the regression coefficient based on the regression calibration method assuming the existence of a validation data set. These estimators arise by assuming that the true covariate $Z$ can be observed in a subset of the sample along with a single measurement of the surrogate $W$ on the entire sample. For example, Zhou and Pepe (1995) presented a consistent estimator for a discrete covariate and Wang et al. (1997) developed a more general kernel-assisted estimator that is approximately consistent. Frequently, however, a validation data set is not practical to obtain but replication data adhering to (1.2) are available; that is, $Z$ is not observed for any study subjects but multiple measurements $W$ are available for some subjects with corresponding measurement errors ($\epsilon$) that are assumed to be independent. For example, the ongoing Women’s Health Initiative (WHI) study (The WHI Study Group, 1998) is designed to address the most common causes of death, disability and impaired quality of life in postmenopausal women. One of the major components of the WHI is an observational study (OS) aimed to identify predictors of disease. In the WHI OS, an array of risk factors was collected on each women. Among these, replicated measurements were taken for dietary and physical activity variables to account for the measurement error. Another example is the Framingham study (Gordon and Kannel, 1968). In Section 5 we use data of this study to examine the effect of the long-term average systolic blood pressure on the risk of dying from coronary heart disease. The systolic blood pressure was measured with error every 2 years after the subject entered the study. The baseline systolic blood pressure and the subsequent measure can be used as two replicates.

Xie et al. (2001) proposed the risk set regression calibration estimator of $\beta_0$ that uses the replication data to recalibrate at each failure time. As typical of a regression calibration estimator, their method introduced some small asymptotic bias. Based on an extensive simulation study, it was shown that their estimator is robust to the true covariate distribution and the true measurement error distribution, though only symmetric distributions were considered.

Another interesting and useful approach for obtaining consistent estimators is through correcting the partial likelihood score function. That approach was used by Nakamura (1992) and Buzas (1998) to avoid a distributional assumption for the covariate. The consistency of the Nakamura (1992) estimator under a normal error distribution was shown by Kong and Gu (1999). Consistent estimators of $\beta_0$ using replicated data were proposed by Huang and Wang (2000) and Hu and Lin (2002). Both estimating procedures are free of distributional assumption on the covariates while the first is also free of distributional assumption on the error terms. In the second approach, a symmetric distribution of the error terms is assumed. Another aspect that distinguishes these two estimation techniques is that Huang and Wang (2000) assume that the existence and number of replicates is independent of an individual’s other data, while Hu and Lin (2002) allow the number of replicates to depend on the individual’s failure time, censoring time and true
covariates. This distinction might be of practical importance in the circumstance where replicate data are available only from a selected subset of the sample.

In the stratified Cox regression model the overall score vector and information matrix become the sum over strata of the stratum-specific score vector and information matrices. Therefore it is not surprising that some of the aforementioned corrected estimators will fail to correct the bias introduced by the mismeasured covariate with small stratum sizes, even when the number of strata is very large as in a matched case-control study. If the sample size in each stratum is moderate to large, as is typical in a multicenter clinical trial, the existing methods for correcting the bias can be expected to perform well, with properties similar to those in a non-stratified setting.

Corrected estimators under the setting of matched logistic model with covariates measurement error were proposed by Armstrong et al. (1989), assuming multivariate normal distribution of covariates. Forbes and Santner (1995) provided a corrected estimator for an accurately measured binary exposure variable in which continuous confounders were measured with error. McShane et al. (2001) proposed a bias-corrected estimator of the log odds ratio from matched case-control data with mismeasured covariates. Asymptotically, the corrected estimator is biased, except under the normality assumption of covariates and measurement error. Their method has numerical convergence problems in the case of a skewed distribution of covariates and normal measurement error. With non-normal measurement error their corrected estimator generally produces a biased estimator.

In this work we consider the stratified Cox proportional hazard model in which the strata divide the subjects into disjoint groups, each of which has a distinct baseline hazard function but common values for the coefficient vector $\beta_0$. The hazard for an individual who belongs to stratum $k$ is therefore

$$\lambda_k(t|Z) = \lambda_{0k}(t) \exp(\beta_0^T Z).$$  \hspace{1cm} (1.3)

We further assume an independent censoring mechanism and the additive measurement error model (1.2). In Section 2 we will show through a simulation study that stratified extensions of the $\beta_0$-estimators of Huang and Wang (2000) and of Hu and Lin (2002) when applied to small strata often failed to converge, even with small values of the error variance. In the samples that converge, these estimators tend to be highly biased. We will also show that the asymptotically biased method of Xie et al. (2001) yields little small sample bias in most cases under symmetrically distributed covariates and error terms, but the bias is non-negligible for skewed distributions. In Section 3, we introduce a nonparametric consistent estimator of $\beta_0$ along with its estimated variance. The term nonparametric refers to the fact that there are no distributional assumptions on the covariates or the error terms. Through a simulation study presented in Section 4, we show that the proposed nonparametric estimator is approximately unbiased in moderate sized samples, even if the true covariates or the error variates are generated from non-symmetric distributions. This estimator is significantly easier to compute than the previously mentioned methods. In Section 5 we provide an illustration based on the Framingham study, and we conclude with some summary comments in Section 6.

2. EXTENSION OF THE EXISTING METHODS TO THE STRATIFIED SETTING


Denote the data in stratum $k$ by $\{X_{ki}, \delta_{ki}, Z_{ki}\}, i = 1, \ldots, n_k$ where $n_k$ follows some positive distribution function, and suppose that these data are iid replicates for $k = 1, \ldots, K$. Also suppose that hazard rates in stratum $k$ adhere to (1.3), $k = 1, \ldots, K$. Consider the situation that replicated measurements are taken for the covariates with measurement error. Let $\{W_{ki1}, \ldots, W_{kim}\}, m \geq 2$ be a finite number of covariate measurements for the subject $ki$ such that $\{\epsilon_{ki1}, \ldots, \epsilon_{kim}\}$ are iid replicates of $\epsilon_{kj}$. The basic idea behind the extensions of the methods by Huang and Wang (2000) and Hu and Lin (2002) is to replace the standard
stratified estimating function

$$
\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\{ Z_{ki} = \frac{\sum_{j=1}^{n_k} Z_{kj} \exp(\beta^T Z_{kj}) I(X_{kj} \geq X_{ki})}{\sum_{j=1}^{n_k} \exp(\beta^T Z_{kj}) I(X_{kj} \geq X_{ki})} \right\} \delta_{ki}
$$

(2.1)

by a corrected one. For each subject $ki$ let $W_{ki}^{(1)}$ and $W_{ki}^{(2)}$ denote distinct selections from $\{W_{k1}, \ldots, W_{kim}\}$. A corrected estimating equation of the type given by Huang and Wang (2000) can be written as

$$
\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\{ A(W_{ki}^{(1)} - \frac{\sum_{j=1}^{n_k} A(W_{kj}^{(1)}) \exp(\beta^T W_{kj}^{(1)}) I(X_{kj} \geq X_{ki})}{\sum_{j=1}^{n_k} A \exp(\beta^T W_{kj}^{(1)}) I(X_{kj} \geq X_{ki})} \right\} \delta_{ki}
$$

where $A$ denotes the operator averaging over all possible sets of $W_{ki}^{(1)}$ or $\{W_{ki}^{(1)}, W_{ki}^{(2)}\}$. By requiring that $W_{ki}^{(1)}$ and $W_{ki}^{(2)}$ be distinct replicates, we get

$$
E[W_{kj}^{(1)} \exp(\beta^T W_{kj}^{(2)}) I(X_{kj} > s)]
$$

$$
= E[I_{kj}^{(1)}] E[\exp(\beta^T I_{kj}^{(2)})] E[\exp(\beta^T Z_{kj}) I(X_{kj} > s)]
$$

$$
+ E[\exp(\beta^T I_{kj}^{(2)})] E[Z_{kj} \exp(\beta^T Z_{kj}) I(X_{kj} > s)]
$$

This is a crucial point to show that the corrected estimating function converges almost surely to (2.1) with large stratum size. For more details see Huang and Wang (2000).

Similarly, an extended estimating equation of the Hu and Lin (2002) type can be defined as

$$
\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\{ A(W_{ki}^{(1)} - \hat{\eta}_1(\beta)/\hat{\eta}_0(\beta)) \frac{\exp(\beta^T W_{ki}^{(1)}) I(X_{ki} \geq X_{ki})}{\sum_{j=1}^{n_k} \exp(\beta^T W_{kj}^{(1)}) I(X_{kj} \geq X_{ki})} \right\} \delta_{ki}
$$

with $\hat{\eta}_0(\beta) = [N^{-1} \sum_{k=1}^{K} \sum_{j=1}^{n_k} A \exp(\beta^T (W_{ki}^{(1)} - W_{ki}^{(2)}))]^{1/2}$, $N = \sum_{k=1}^{K} n_k$ the total number of subjects included in the study, and $\hat{\eta}_1(\beta) = (2N \hat{\eta}_0(\beta))^{-1} \sum_{k=1}^{K} \sum_{j=1}^{n_k} A(W_{ki}^{(1)} - W_{ki}^{(2)} \exp(\beta^T (W_{ki}^{(1)} - W_{ki}^{(2)}))].$

A limitation of these two estimating functions is that in contrast to the nonstratified setting, the summation over ‘risk set’ here, only involves subjects from the same stratum. As a result, these two approaches will usually not correct the bias in small stratum sizes and limited number of replicates $m$. Table 1 shows simulation results for 1000 strata of size 2 with: one standard normal covariate $Z$; exponentially distributed failure time with hazard $\exp(\rho_0 Z)$; 70% censoring rate; three replicates of the mismeasured covariate for each subject and the parameter of interest $\rho_0 = \ln 2$. The results indicate that these two estimators are substantially biased even with small values of the error variance, $\sigma^2$. For moderate and large error variance, the estimating equations for both methods frequently have no roots. Hence, these methods will not be considered further in this paper.

### 2.2 Extension of Xie et al. (2001) estimator

For the moment, suppose that $(Z_{ki}, AW_{ki}^{(1)})^T$ are multivariate normal in the overall risk set that is based on all subjects from all strata, at time $t$. Denote $E(Z_{ki}) = E(AW_{ki}^{(1)}) = \mu(t)$ and the covariance matrix of $(Z_{ki}, AW_{ki}^{(1)})^T$ by

$$
\begin{pmatrix}
\Sigma(t) & \hat{\Sigma}(t) \\
\hat{\Sigma}(t) & \Sigma(t) + \Delta/m
\end{pmatrix}
$$
Table 1. Simulation results for the mean bias of naive (NV) estimator of $\beta_0$, and for extensions of the Huang and Wang-type nonparametric corrected (NPC) and Hu and Lin-type corrected (Hu) estimators of $\beta_0$, based on 1000 strata, two subjects within each stratum, 70% censoring, $\beta_0 = \ln(2)$ and $m = 3$. The number of samples with no root out of 1000, is given in parentheses

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>NV</th>
<th>NPC</th>
<th>Hu</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.701 (0)</td>
<td>0.699 (0)</td>
<td>0.696 (0)</td>
</tr>
<tr>
<td>0.25</td>
<td>0.643 (0)</td>
<td>1.084 (451)</td>
<td>0.849 (929)</td>
</tr>
<tr>
<td>0.75</td>
<td>0.553 (0)</td>
<td>- (1000)</td>
<td>- (1000)</td>
</tr>
<tr>
<td>1.00</td>
<td>0.516 (0)</td>
<td>- (1000)</td>
<td>- (1000)</td>
</tr>
</tbody>
</table>

where $\Sigma(t)$ and $\Delta$ are the covariance matrices of $Z_{ki}$ and $\epsilon$, respectively. The expectation vector and the covariance matrix can be estimated by the replicated samples in a similar way to that presented in Xie et al. (2001). Namely,

$$\hat{\mu}(t) = \left\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbb{I}(X_{ki} \geq t) \right\}^{-1} \left[ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbb{I}(X_{ki} \geq t) (AZ_{ki}^{(1)}) \right]$$

$$\hat{\Sigma}(t) = \left\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbb{I}(X_{ki} \geq t) \right\}^{-1} \left[ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbb{I}(X_{ki} \geq t) (AZ_{ki}^{(1)}) (AZ_{ki}^{(1)})^T \right]$$

Then, $Z_{ki}$ can be estimated by the estimated conditional expectation of $Z_{ki}$ given $AZ_{ki}^{(1)}$,

$$\hat{Z}_{ki}(t) = \hat{\Delta} m^{-1} \left\{ \hat{\Sigma}(t) + \hat{\Delta} m^{-1} \right\}^{-1} \hat{\mu}(t) + \hat{\Sigma}(t) \left\{ \hat{\Sigma}(t) + \hat{\Delta} m^{-1} \right\}^{-1} AZ_{ki}^{(1)}.$$

In this approach, we replace at each failure time $t$, in the estimating function (2.1), the unknown $Z_{ki}$ by its estimator $\hat{Z}_{ki}(t)$ for all subjects of stratum $k$ that belong to the risk set. That is, the extended risk set regression calibration (RRC) estimating function is

$$\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left[ \hat{Z}_{ki}(X_{ki}) - \sum_{j=1}^{n_k} \hat{Z}_{kj}(X_{ki}) \exp \left\{ \beta^T \hat{Z}_{kj}(X_{ki}) \right\} \mathbb{I}(X_{kj} \geq X_{ki}) \right] \delta_{ki}.$$

Note that here we replace the unknown covariate by an estimator that is based on all subjects at risk over all the strata. Therefore the presence of strata, and small number of subjects in each stratum, is expected to have little effect on the asymptotic properties of the estimator of $\beta_0$, comparable to the non-stratified setting. In this procedure we assume a common expectation vector and covariance matrix.
over strata. In the case of non-constant expectation among strata, the method of Xie et al. should be extended so that the estimated covariate \( Z \) will be based on a stratum-specific risk set. But since in this work we are considering the case of small sample stratum size with large number of strata, one would not expect this approach to be efficient. However, based on simulation study, it should be noted that the extended RRC method, that is not based on a stratum-specific risk set, is robust to modest departures from constant expectation over strata, as will be shown in Section 4. In Section 4 we present a simulation study including results of the above extended RRC method. It will be shown that under symmetric distributions of the true covariate and of the error term this estimator performs very well in terms of bias reduction and estimation efficiency. But this is not the case for nonsymmetric distributions of the true covariate or the error term. Therefore, in the following section we present an asymptotically unbiased estimator without making distributional assumptions on the true covariate or the error variate.

3. The Proposed Nonparametric Corrected Estimator and Its Asymptotic Distribution

Adopting the idea of Huang and Wang (2001) for the logistic regression model, we use a weighted estimating equation. At follow-up time \( t \), let \( \omega(\beta, t) \) be some known function of the failure, censoring and true covariates histories of all subjects over \([0, t]\) and consider the weighted estimating function

\[
K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \omega(\beta, X_{ki}) \left( Z_{ki} - \frac{\sum_{j=1}^{n_k} Z_{kj} \exp(\mathbf{Z}^T \mathbf{Z}_{kj}) I(X_{kj} \geq X_{ki})}{\sum_{j=1}^{n_k} \exp(\mathbf{Z}^T \mathbf{Z}_{kj}) I(X_{kj} \geq X_{ki})} \right) \delta_{ki}. \tag{3.1}
\]

It should be noted that any such choice of \( \omega \) will provide an unbiased estimating function if all covariates are accurately ascertainment. Our proposal is to choose

\[
\omega(\beta, t) = Y_{nk}^{-1}(t) \sum_{j=1}^{n_k} \exp(\mathbf{Z}^T \mathbf{Z}_{kj}) I(X_{kj} \geq t)
\]

where \( Y_{nk}(t) = \sum_{j=1}^{n_k} I(X_{kj} \geq t) \) is the number of subjects at risk in stratum \( k \) at time \( t \). Then (3.1) becomes

\[
U_K(\beta) = K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{nk}^{-1}(X_{ki}) \delta_{ki} \sum_{j=1}^{n_k} I(X_{kj} \geq X_{ki}) \exp(\mathbf{Z}^T \mathbf{Z}_{kj})(Z_{ki} - Z_{kj}). \tag{3.2}
\]

Let \( \mathcal{F}_t \) be the \( \sigma \)-algebra generated by all the failure, censoring and true covariate history of all the subjects over \([0, t]\) and assume that study subjects are observed on the time interval \([0, \tau]\), \( 0 < \tau < \infty \). Since for every \( k, i, j, i \neq j \)

\[
E[\mathbf{W}_{ki}^{(1)} \exp(\beta^T \mathbf{W}_{kj}^{(1)}) | \mathcal{F}_\tau] = \mathbf{Z}_{ki} \exp(\beta^T \mathbf{Z}_{kj}) E[\exp(\beta^T \mathbf{w}_{ki}^{(1)})] \]

and for every \( k, i \)

\[
E[\mathbf{W}_{ki}^{(2)} \exp(\beta^T \mathbf{W}_{kj}^{(1)}) | \mathcal{F}_\tau] = \mathbf{Z}_{ki} \exp(\beta^T \mathbf{Z}_{kj}) E[\exp(\beta^T \epsilon_{ki}^{(1)})],
\]

we propose the following nonparametrically corrected estimating function:

\[
CU_K(\beta) = K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{nk}^{-1}(X_{ki}) \delta_{ki} \sum_{j=1}^{n_k} I(X_{kj} \geq X_{ki}) \left\{ A \mathbf{W}_{ki}^{(1)} \exp(\beta^T \mathbf{W}_{kj}^{(1)}) - A \mathbf{W}_{kj}^{(2)} \exp(\beta^T \mathbf{W}_{kj}^{(1)}) \right\}. \tag{3.3}
\]
As a result of the linearity of (3.3) as a function of each stratum risk set, the effect of small stratum size is eliminated. Hence we can expect $\hat{\beta}$, the zero-crossing of (3.3), to be a consistent estimator of the ‘true’ $\beta_0$.

In Appendix A we show that as the number of strata $K$ goes to infinity and $m$, the number of covariate replicates for each study subject, is fixed:

1. $\sup_{\beta \in B} |CU_K(\beta) - U_K(\beta)| \to 0$ almost surely, and for $\beta \in B$, $K^{1/2} \{CU_K(\beta) - U_K(\beta)\}$ is asymptotically normal with mean 0 and a covariance matrix that can be consistently estimated.
2. $K^{1/2} (\hat{\beta} - \beta_0)$ is asymptotically normal with mean 0 and a sandwich type covariance matrix that can be consistently estimated.

4. SIMULATION STUDY

The simulation study examines the performance of the NPCS estimator and the extended RRC estimator in the situation of a common covariance matrix of the true covariate across strata. We considered a single error-prone covariate with $\beta_0 = \ln(2)$; 1000 strata; two subjects within each stratum; $m = 2$ or 3 and two censoring patterns: censoring time is uniformly distributed on either $[0, 0.7]$ or $[0, 6.0]$ yielding approximately 70% and 20% censoring rates, respectively. Two distributions for the true covariate $Z$ and for the error term $\epsilon$ were considered: for $Z$, the standard normal distribution and the standardized truncated (at 5) Chi-square distribution with one degree of freedom; for $\epsilon$, the zero mean normal distribution and the zero-mean truncated (at 5) Chi-square distribution with one degree of freedom.

Each of these two distributions were considered under two different values of the measurement error variance, $\sigma^2 = 0.25, 1.00$. The failure time, $T_{ki}$, was generated according to exponential distribution with $\exp(\beta_0 Z_{ki})$ in the case of homogeneous baseline hazard functions among strata. For the heterogeneous case the exponential distribution considered was with $\exp(\beta_0 Z_{ki}) \exp(\gamma V_k)$, where $V_k$ is a zero mean normally distributed random variable with variance equal to 3, $Z_{ki}$ and $V_k$ are independent, and $\gamma = \ln(2)$. Results are based on 1000 simulation runs for each configuration.

Tables 2 and 3 summarize results of the naive (NV), the extended RRC and the proposed NPCS estimators. The naive estimator of $\beta_0$ was calculated by using the standard stratified estimating equation (2.1) and replacing the unknown covariate $Z_{ki}$ by $AW_{ki}^{(1)}$. For each estimation technique the empirical bias is presented along with their empirical standard deviation, mean of the estimated standard deviation, empirical coverage rate of 95% Wald-type confidence interval and number of samples in which the Newton–Raphson algorithm failed to find a root. Since once measurement error becomes substantial our proposed estimator introduce some outliers in the simulation studies, we summarize the bias of this estimator by both the mean and the median of the empirical bias. For the other estimators under investigation here the discrepancy between the mean and the median is very small and thus only the mean is presented. It should be noted that the same phenomena was noted by Huang and Wang (2001) regarding the presence of outliers in their corrected estimator for the logistic regression model.

In Table 2, the normal distributions for the true covariate and the error term is studied under the aforementioned heterogeneous baseline hazard function configuration. As expected, the naive estimator underestimates the true parameter and suffers from poor empirical coverage rate. The extended RRC and the NPCS estimators perform very well in terms of bias reduction and coverage rate with the former having considerably smaller standard deviation.

Since the existence of different baseline hazard functions among strata has little effect on the performance of the estimators in terms of bias reduction, the following results for nonsymmetric distribution of the true covariate with symmetric error term distribution (Table 3) were conducted under the assumption of equal baseline hazard functions. In contrast to the case of symmetric distributions of the true covariate and the error term, it is evident from Table 3 that the extended RRC estimator failed.
Table 2. Simulation results of the naive (NV), the extended Xie et al. (2001) risk set regression calibration (RRC) and the proposed non-parametric corrected (NPCS) estimators of $\beta_0$. 1000 strata, two subjects within each stratum, $\beta_0 = \ln(2)$, $Z \sim N(0, 1)$ and $\epsilon \sim N(0, \sigma^2_\epsilon)$ with different baseline hazard functions among strata

<table>
<thead>
<tr>
<th>Cens.</th>
<th>$\sigma^2_\epsilon$</th>
<th>$m = 2$</th>
<th>NV</th>
<th>RRC</th>
<th>NPCS</th>
<th>$m = 3$</th>
<th>NV</th>
<th>RRC</th>
<th>NPCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>70%</td>
<td>0.25</td>
<td>B</td>
<td>-0.070</td>
<td>0.015</td>
<td>0.052(0.016)</td>
<td>-0.049</td>
<td>0.019</td>
<td>0.029(0.009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES</td>
<td>0.119</td>
<td>0.134</td>
<td>0.197</td>
<td>0.120</td>
<td>0.126</td>
<td>0.159</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EM</td>
<td>0.132</td>
<td>0.157</td>
<td>0.238</td>
<td>0.136</td>
<td>0.149</td>
<td>0.188</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>84.5</td>
<td>94.7</td>
<td>95.3</td>
<td>90.3</td>
<td>95.6</td>
<td>94.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>B</td>
<td>-0.232</td>
<td>0.015</td>
<td>0.017(-0.008)</td>
<td>-0.170</td>
<td>0.006</td>
<td>0.015(-0.007)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>0.096</td>
<td>0.144</td>
<td>0.223</td>
<td>0.102</td>
<td>0.142</td>
<td>0.197</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.129</td>
<td>0.185</td>
<td>0.214</td>
<td>0.128</td>
<td>0.165</td>
<td>0.189</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>24.6</td>
<td>93.5</td>
<td>94.5</td>
<td>45.3</td>
<td>94.8</td>
<td>94.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0</td>
<td>5</td>
<td>67</td>
<td>0</td>
<td>1</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>0.25</td>
<td>B</td>
<td>-0.087</td>
<td>0.003</td>
<td>0.008(-0.002)</td>
<td>-0.054</td>
<td>0.001</td>
<td>0.010(0.007)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>0.059</td>
<td>0.068</td>
<td>0.083</td>
<td>0.061</td>
<td>0.068</td>
<td>0.081</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.058</td>
<td>0.069</td>
<td>0.086</td>
<td>0.063</td>
<td>0.068</td>
<td>0.084</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>65.1</td>
<td>95.4</td>
<td>95.4</td>
<td>81.8</td>
<td>95.5</td>
<td>94.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>B</td>
<td>-0.255</td>
<td>-0.004</td>
<td>0.020(-0.004)</td>
<td>-0.196</td>
<td>-0.000</td>
<td>0.014(0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>0.049</td>
<td>0.076</td>
<td>0.126</td>
<td>0.053</td>
<td>0.073</td>
<td>0.102</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>0.051</td>
<td>0.080</td>
<td>0.131</td>
<td>0.052</td>
<td>0.078</td>
<td>0.111</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1.0</td>
<td>93.6</td>
<td>95.2</td>
<td>5.3</td>
<td>94.0</td>
<td>95.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B: mean of empirical bias (median of empirical bias); ES: mean of estimated standard errors; EM: empirical standard error. C: empirical coverage of 95% confidence interval; F: number of samples (out of 2000) failed to converge.

To fully correct the bias under these sampling conditions. Specifically, for moderate and high values of $\sigma^2_\epsilon$ it overestimated the true parameter value and the empirical coverage rate is low. The proposed NPCS estimator reduces the bias substantially, has accurate confidence interval coverage but has a large standard error, especially under large values of $\sigma^2_\epsilon$. Similar results were found in the case of nonsymmetric distribution of the true covariate and the error variate (results can be found in the journal website). It should be noted that the samples with no root usually have a parabolic shape for the estimating function with no crossing point at the x-axis. However, under slightly larger stratum sizes (results not shown) the fraction of samples with no root decreases rapidly.

5. Framingham Example

In this section we illustrate our nonparametric correction method by considering an application to the Framingham study (Gordon and Kannel, 1968). The data used here consist of 806 men aged 29–62 years, with mean follow-up of approximately 31 years. We concentrate on examining the effect of the long-term average systolic blood pressure (SBP) on the risk of dying from coronary heart disease (CHD). We constructed a matched case-control study by matching the age at onset of each case with a control whose age is within a year of the case. Hence the data consists of 403 strata of size two each. As
Table 3. Simulation results of the naive (NV), the extended Xie et al. (2001) risk set regression calibration (RRC) and the proposed non-parametric corrected (NPCS) estimators of $\beta_0$. 1000 strata, two subjects within each stratum, $\beta_0 = \ln(2)$, $Z \sim \text{Modified } \chi^2(1)$ and $\epsilon \sim N(0, \sigma^2)$.

<table>
<thead>
<tr>
<th></th>
<th>$\sigma^2$</th>
<th>$m = 2$</th>
<th></th>
<th>$\sigma^2$</th>
<th>$m = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NV    RRC NPCS</td>
<td></td>
<td></td>
<td>NV    RRC NPCS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.064 0.050 0.019(0.010)</td>
<td>-0.043 0.033 0.014(0.002)</td>
<td></td>
<td>-0.071 0.051(0.004)</td>
</tr>
<tr>
<td>70%</td>
<td>0.25</td>
<td>ES 0.078 0.096 0.112</td>
<td>0.083 0.093 0.109</td>
<td>EM 0.081 0.095 0.122</td>
<td>0.082 0.092 0.117</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES 0.067 0.116 0.144</td>
<td>0.072 0.107 0.128</td>
<td>EM 0.066 0.124 0.159</td>
<td>0.072 0.109 0.131</td>
</tr>
<tr>
<td>1.00</td>
<td>0.25</td>
<td>B -0.204 0.145 0.028(0.000)</td>
<td>-0.152 0.102 0.024(0.009)</td>
<td>ES 0.060 0.076 0.088</td>
<td>0.062 0.073 0.085</td>
</tr>
<tr>
<td>20%</td>
<td>0.25</td>
<td>B -0.100 0.050 0.011(0.004)</td>
<td>-0.071 0.037 0.012(0.005)</td>
<td>ES 0.049 0.094 0.113</td>
<td>0.051 0.086 0.106</td>
</tr>
</tbody>
</table>

B: mean of empirical bias (median of empirical bias); ES: mean of estimated standard errors; EM: empirical standard error. C: empirical coverage of 95% confidence interval; F: number of samples (out of 1000) failed to converge.

noted in the introduction, the baseline SBP and the consequent SBP measurement, observed two years after the subject entered the study, will be considered as two replicates, $W_{ki1}$ and $W_{ki2}$. The ratio of the error variance to the covariate variance between subject is observed to be 0.2. Since the performance of the extended RRC method is best with symmetric covariates, we used the transformation $\log\{(SBP - 75)/25\}$ which induces approximate normality (Cornfield, 1962). The estimated SBP coefficient according to the naive method, the extended RRC and the NPCS method are 0.891, 1.130 and 0.982, respectively. The corresponding standard errors are 0.284, 0.366 and 0.335. It is likely that by using the naive estimator the effect of SBP on CHD death is underestimated. Regarding the extended RRC we suspect that the coefficient is overestimated since even after the aforementioned transformation, the covariates are still somewhat skewed. The skewness measures of the first and the second SBP measures after transformation are $-0.198$ and $-0.179$, respectively. Also, both variables are significantly not normal according to the Shapiro–Wilk test. It should be noted that normality assumptions of the true covariate and the error variate cannot be verified by the observed data. With the NPCS approach the attenuation is presumably corrected, and the analysis implies that SBP is positively related with the risk of CHD death.

6. CONCLUDING REMARKS

The proposed NPCS estimator can be readily generalized to time-dependent covariates provided that the value of the covariates at time $t$ are known at that time. The simulation studies presented here under
the general stratified Cox model, aimed also to investigate the performance of the proposed estimator, were done under the extreme case of two subjects within each stratum. It is evident that under a very substantial covariate measurement error the proposed technique may fail in finding a root. But it should also be noted that this problem is eliminated as the number of strata increases. Hence, this nonparametric estimation technique is expected to be valuable mainly in large-scale studies, such as a population based case-control study. The performance of the NPCS method under a small number of strata are satisfying in terms of bias and empirical coverage rate, but less so in terms of root existence. For example, under the setting of a mismeasured covariate with 100 strata and censoring rate of 20%, two replicates for each subject, two subject within each stratum, standard normal true covariate and normally distributed error we get that the biases (median bias, standard error) for \( \sigma^2 = 0.25, 0.75, 1.00 \) are 0.067 (0.013, 0.306), 0.059 (−0.006, 0.349), −0.005 (−0.051, 0.326). The respective empirical coverage rates of 95% Wald-type confidence interval are 94.81, 94.43 and 93.37, and the respective rates of samples with no root are 56/1000, 156/1000 and 216/1000.

It is also evident that under a constant expectation vector over strata the extended RRC estimator yields an estimator with substantially smaller standard error in comparison to that of the proposed nonparametric corrected estimator, and has few numerical convergence problems. But, in the case of skewed distribution the extended RRC estimator may be substantially biased. Since, in practice, the symmetry assumption on the true covariates and the error variates may be impossible to verify, our proposed NPCS method is practically important in providing an estimator that is robust against a wide range of measurement error models. Also note that as heterogeneity in the baseline hazard rates increases, the efficiency of the RRC estimator decreases (results not shown).

Another important application of the stratified Cox proportional hazard model is a matched case-control family study with the constant cross-ratio of Clayton (1978) and Oakes (1989) being the primary parameter of interest and risk factors that are subject to measurement error. The case/control status is assumed to be accurately ascertained. A detailed description of this model along with simulation results of the performance of the naive and NPCS estimators in this setting can be found in the journal website.

ACKNOWLEDGEMENTS

This work is in part supported by the grants (AG14358 and CA53996) from the NIH. This paper uses the Framingham Heart Study (FHS) data supplied by the US NHLBI, NIH, Department of Health and Human Services. The FHS is conducted and supported by the NHLBI in collaboration with FHS investigators. This manuscript was not prepared in collaboration with investigators of the FHS and does not necessarily reflect the opinions or views of the FHS or the NHLBI.

APPENDIX A

Note that \( U_K(\beta) \) in (3.2) is not a derivative of a convex (or concave) function and this estimating function is no longer a score function of a log-likelihood function.

Consider the following conditions:

(A) \( dU_K(\beta)/d\beta \) exists and is continuous in an open neighborhood about \( \beta_0 \).

(B) \(-dU_K(\beta)/d\beta^T|\beta=\beta_0 \) is positive definite with probability going to 1 as \( K \to \infty \).

(C) The convergence of \( dU_K(\beta)/d\beta^T \) to \( \Gamma(\beta) \) is uniform for \( \beta \) in an open neighborhood about \( \beta_0 \).

(D) \( U_K(\beta_0) \to 0 \) as \( K \to \infty \).
It is easily seen that condition (A) holds. In the following it is shown that conditions (D) holds and that condition (C) holds under assumptions (b) and (c) below. For the asymptotic properties of $\hat{\beta}$ the following set of conditions are assumed:

(a) Finite follow-up time $\tau > 0$ satisfies $P(X_{ki} \geq \tau) > 0$ for all $k$ and $i$.
(b) $E(Z_k^T Z_{ki}) < \infty$ and $E(e_k T e_{ki}) < \infty$ for all $k$, $i$ and $l$.
(c) There exists a compact neighborhood $B$ of $\beta_0$ such that $E(\sup_{\beta \in B} Z_k^T Z_{ki} e_k^2 \beta^T Z_k^T) < \infty$ and $E(\sup_{\beta \in B} e_k^2 \beta^T e_{ki}) < \infty$ for all $k$, $i$ and $l$.

The establishment of the uniform convergence of $dU_K(\beta)/d\beta^T$ follows from Theorem 4.1 of Andersen and Gill (1982), which can be verified under conditions (b) and (c). To establish condition (D), rewrite the weighted estimating function $U_K(\beta_0)$ as

$$K^{-1} \sum_{k=1}^K \sum_{i=1}^{n_k} \sum_{j=1}^{n_i} \int_0^\tau Y_{kj}(s) \exp(\beta_0^T Z_{kj})(Z_{kj} - Z_{kj}) dN_{ki}(s), \quad (A.1)$$

where $Y_{kj}(t) = I(X_{ki} > t)$ is the at-risk process, and $N_{ki}(t) = \delta_{ki} I(X_{ki} \leq t)$ is the failure counting process. Also, note that $E(dN_{ki}(t)|F_{t-}) = Y_{kj}(t) \exp(\beta^T Z_{kj}) \lambda_0(t) dt$. Since $U_K(\beta)$ can be written as a sum of $K$ iid random variables, convergence of the above weighted estimating function to zero can be verified by the law of large numbers. Therefore, based on Foutz' Theorem (1977), given conditions (B) and (C), there exists a unique $\hat{\beta}$ such that $U_K(\hat{\beta}) = 0$ with probability going to 1 as $K \to \infty$, and $\hat{\beta} \to \beta_0$ in probability. In the following, $K$ goes to infinity and $m$ is fixed.

**THEOREM 1** Under Conditions (a)–(c), $\sup_{\beta \in B} |CU_K(\beta) - U_K(\beta)| \to 0$ almost surely, and for $\beta \in B$, $K^{1/2} \{CU_K(\beta) - U_K(\beta)\}$ is asymptotically normal with mean 0 and a covariance matrix that can be consistently estimated.

**THEOREM 2** Under conditions (B) and (a)–(c), $K^{1/2} (\hat{\beta} - \beta_0)$ is asymptotically normal with mean 0 and a sandwich-type covariance matrix that can be consistently estimated.

A sketch of the proofs are given in the journal website. A consistent covariance matrix estimator of $K^{1/2} CU_K(\beta)$ is given by $\tilde{\Sigma}(\beta) = K^{-1} \sum_{k=1}^K \xi_k(\hat{\beta}) \xi_k(\hat{\beta})^T$ where

$$\xi_k(\beta) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_i} Y_k^{-1}(X_{ki}) \delta_{ki} I(X_{ki} \geq X_{kj}) \left[ A W_{ki}^{(1)} \exp \left( \beta^T W_{ki}^{(1)} \right) - A W_{kj}^{(2)} \exp \left( \beta^T W_{kj}^{(2)} \right) \right].$$

A consistent estimator of the covariance matrix of $K^{1/2} \hat{\beta}$ is given by

$$I^{-1}(\hat{\beta}) \tilde{\Sigma}(\hat{\beta}) \{I^{-1}(\hat{\beta})\}^T.$$


[Received August 6, 2002; first revision June 12, 2003; second revision July 23, 2003; accepted for publication July 30, 2003]