Journal of Biometrics & Biostatistics

Granville and Fan, J Biom Biostat 2012, 3:6
DOI: 10.4172/2155-6180.1000152

Research Article Open Access

Accelerated Failure Time Models with Auxiliary Covariates

Kevin Granville and Zhaozhi Fan*

Department of Mathematics and Statistics, Memorial University, St. John's, A1C 5S7, Newfoundland, Canada

Abstract

In this paper we study semi-parametric inference procedure for accelerated failure time models with auxiliary information about a main exposure variable. We use a kernel smoothing method to introduce the auxiliary covariate to the likelihood function. The regression parameters are then estimated through maximization of the estimated likelihood function. A consistent estimator of the variance of the estimator of the regression coefficients is proposed. Simulation studies show that the efficiency gain 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.

Keywords: Kernel smoothing; Estimated likelihood function; Accelerated failure time models; Measurement errors; Auxiliary covariates

Introduction

It is quite common when attempting statistical analysis on a set of data that researchers run into the problem of missing or mismeasured observations. This is often the case in medical studies where the tests to get an accurate measurement may be particularly expensive or invasive for the patient. The medical researchers can opt to examine another relevant variable which may be cheaper or easier to measure, even if it does not provide as much information. This can be tested for in place of the original variable or along side when it is possible to do so. Researchers then have the choice to work with either a smaller sample size using just the samples with measurements for the variable of interest or to include the imperfect data in the analysis with the goal of gaining a higher efficiency. For example, in the Primary Biliary Cirrhosis (PBC) study conducted at Mayo Clinic between 1974 and 1984, Aspartate Aminotransferase (AST) was an important predictive variable to the survival time of PBC patients, which was only collected for patients registered to the double blind clinical trial, due to reasons similar to those previously mentioned. But another closely related variable, bilirubin, is recorded for all PBC patients [1]. In order to enhance the efficiency of the statistical analysis regarding the relationship between AST and the patients' survival, it might be worthy to have the available information from all the patients included. Motivated by this example, in this article we propose an inferential method for this kind of survival data, where we replace the missing or mismeasured data using kernel smoothing based on an auxiliary covariate, which is measured for each subject.

When it is possible to have some of the desired data measured accurately, these cases form a validation set. The validation set contains measurements for both the variable of interest and the auxiliary covariate. The rest of the cases are placed into the non-validation set where only the auxiliary covariate is available. In the analysis of this data, if the auxiliary covariate is just the original variable with measurement error, one could be inclined to use it in place of the missing data. Unfortunately this naive method will lead to estimation bias for all regression coefficients in the model which, depending on the magnitude of the error, can be quite large [2]. Hence it is very important for researchers to include as many subjects as possible in their analysis, to aim at a higher efficiency, as well as to correct the estimation bias caused by measurement errors.

Much research has been done in this area in the past. Some research on how to incorporate missing or mismeasured data in models includes the works of Rubin [3], Fuller [4], Carrol et al. [5], Wang et al. [6], Meng and Schenker [7], Cheng and Wang [8] and Yu and Nan [9], to list a few. A common specific statistical model chosen for these situations is the Cox model [10]. For details see Cox and Oakes [11], Kalbfleisch and Prentice [12], Hu et al. [13] and Hu and Lin [14], among others. In this article however, we focus on the parametric accelerated failure time models. When an auxiliary covariate is included in the analysis through an estimated likelihood, the AFT model is more efficient if an appropriate distribution of the failure time is known. Research work based on an estimated partial likelihood function has been conducted by many authors such as Pepe and Flemming [15], Pepe [16], Zhou and Pepe [17], Zhou and Wang [18], Zhou et al. [19], Jiang and Zhou [20], Fan and Wang [21] and Liu et al. [22]. Recently He et al. [23] proposed to use SIMEX method to handle the accelerated failure time models when covariates are subject to measurement error. But investigation about the performance of accelerated failure time models with auxiliary covariates is still limited and deserves to be carried out, due to some reasons such as (1) the AFT models have direct physical interpretation, (2) the AFT models can better predict the survival function of a patient and (3) the AFT models are robust to model misspecification in the sense that ignoring a covariate will not lead to much biased estimates of other regression coefficients [11].

The rest of this article is organized as follows. Section 2 presents the general accelerated failure time model and some special cases which we use in our calculations. Section 3 covers the estimation method. Section 4 discusses the asymptotic properties of our estimator. Section 5 shows the simulation results for finite samples as well as the results from analyzing data from the Mayo Clinic trial in PBC. In Section 6 we put forth our concluding remarks. Finally, we outline the regularity conditions and proof for the theoretical results from Section 4 in the Appendix.

The accelerated failure time model

Let $\{X_i, Z_i\}$ denote the covariate vector where X_i is the component

***Corresponding author:** Zhaozhi Fan, Department of Mathematics and Statistics, Memorial University, St. John's, A1C 5S7, Newfoundland, Canada, Tel: 1-709-864-8076; Fax: 1-709-864-3010; E-mail: zhaozhi@mun.ca

Received August 17, 2012; **Accepted** September 20, 2012; **Published** September 25, 2012

Citation: Granville K, Fan Z (2012) Accelerated Failure Time Models with Auxiliary Covariates. J Biom Biostat 3:152. doi:10.4172/2155-6180.1000152

Copyright: © 2012 Granville K, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

which is only observed in the validation set and Z_i is the component that is always observed. In this case we assume that X_i is scalar and that Z_i is a vector. For every X_i, let W_i be the corresponding auxiliary covariate of the form $W_i = X_i + U_i$ where U_i is the measurement error incurred when attempting to observe X_i. We assume that U_i follows a normal distribution such that $U_i \sim N(0, \sigma_u^2)$. Let T_i, C_i and δ_i repressent the ith failure time, censoring time and censoring indicator, $\delta_i = I_{[T_i \le C_i]}$ We assume that out of the n subjects, the sample size for the validation sample where the X_i's are correctly observed is n_V and the sample size for the non-validation sample where we do not observe X_i's is $n\bar{V} = n - n_V$. The observed data is therefore $\{S_i, \delta_i, Z_i, X_i, W_i\}$ for the validation sample and $\{S_i, \delta_i, Z_i, W_i\}$ for the non-validation sample, where $S_i = \min(T_i, C_i)$.

The accelerated failure time model can be expressed as

$$Y_i = \log(T_i) = \beta_1 X_i + \beta'_2 Z_i + \varepsilon_i, \tag{1}$$

where $\beta' = (\beta_1, \beta'_2)$ is a vector of unknown parameters that we must estimate and ε_i is the random error which has pdf $f_\varepsilon(\varepsilon)$.

Note that the random error term " in model (1) is in its general form. When standardized the scale parameter $1/\sigma$ or b should be included, as in Lawless [24]. Also the equation (1) assumes automatically that if we are given (X_i, Z_i), W_i gives us no additional information about the failure time.

The pdf $f_T(t_i;\ \beta,\ X_i, Z_i)$ of T_i depends on the form of $f_\varepsilon(\varepsilon)$. Once we have $f_T(t_i;\ \beta,\ X_i, Z_i)$, we are able to calculate the survival and hazard functions for failure time T_i as shown below.

$$S(t_i;\beta,X_i,Z_i) = \int_{t_i}^{\infty} f_T(m;\beta,X_i,Z_i)\,dm \tag{2}$$

and $$h(t_i;\beta,X_i,Z_i) = \frac{f_T(t_i;\beta,X_i,Z_i)}{S(t_i;\beta,X_i,Z_i)} \tag{3}$$

The maximum likelihood estimator of the parameters is the maximizer

$$\beta = Argmax_\beta L(\beta)$$

where

$$L(\beta) = \prod_{i=1}^{n} f_T(S_i;\beta,X_i,Z_i)^{\delta_i}\, S(S_i;\beta,X_i,Z_i)^{1-\delta_i}$$

which using (3) can be rewritten as

$$L(\beta) = \prod_{i=1}^{n} h(S_i;\beta,X_i,Z_i)^{\delta_i}\, S(S_i;\beta,X_i,Z_i) \tag{4}$$

Some special cases: There are some special distributions of the survival time which are of specific interests to practitioners in medical research.

The generalized gamma distribution

We begin by demonstrating how to obtain the likelihood function and estimating equations for the generalized gamma distribution model. This is a very useful distribution. It can be reduced into the Weibull, exponential, or log normal models. We may write the general model as

$$Y_i = \log(T_i) = \mu + \beta_1 X_i + \beta'_2 Z_i + \sigma V_i, \tag{5}$$

where σV_i takes the place of ε_i from equation (1) and follows the generalized gamma distribution. The likelihood function is given as

$$L(\beta) = \prod_{i=1}^{n} \begin{matrix} \dfrac{|\theta|}{\theta^{1-\delta_i}\Gamma(1/\theta^2)}\left(\dfrac{1}{\sigma} t_i^{\frac{1}{\sigma\theta}-1}\left[e^{-\frac{\theta}{\sigma}(\mu+\beta_1 X_i+\beta_2' Z_i)}/\theta^2\right]^{1/\theta^2} e^{-t_i^{\frac{\theta}{\sigma}} e^{-\left(\frac{\theta}{\sigma}(\mu+\beta_1 X_i+\beta_2' Z_i)\right)/\theta^2}}\right)^{\delta_i} \\ \times I\left[\dfrac{1}{\theta^2}, \left(t_i^{\frac{\theta}{\sigma}} e^{-\frac{\theta}{\sigma}(\mu+\beta_1 X_i+\beta_2' Z_i)}\right)/\theta^2\right]^{1-\delta_i} \end{matrix} \tag{6}$$

where the function $I[a, x]$ is the incomplete gamma function, defined as

$$I[a,x] = \int_x^\infty u^{a-1} e^{-u}\,du.$$

So

$$I\left[\frac{1}{\theta^2}, \left(t_i^{\frac{\theta}{\sigma}} e^{-\frac{\theta}{\sigma}(\mu+\beta_1 X_i+\beta_2' Z_i)}\right)/\theta^2\right]^{1-\delta_i} = \int_{t_i^{\frac{\theta}{\sigma}} e^{-\frac{\theta}{\sigma}(\mu+\beta_1 X_i+\beta_2' Z_i)}/\theta^2}^{\infty} u^{1/\theta^2-1} e^{-u}\,du.$$

A reduced case, the exponential regression model

When μ= 0, θ= 1, and σ= 1, the likelihood function (6) is reduced to

$$L(\beta) = \prod_{i=1}^{n}\left(e^{-(\beta_1+\beta_2' Z_i)}\right)^{\delta_i} \exp\{-t_i e^{-(\beta_1+\beta_2' Z_i)}\} \tag{7}$$

A proportional odds model

When modeling AFTs, proportional hazards and proportional odds models are frequently used. The above reduced case is a proportional hazards model. Now let us look at the alternative. Letting μ= 0 again, we then let V_i in equation (5) follow the standard logistic distribution.

The likelihood function is

$$L(\beta) = \prod_{i=1}^{n}\left(\frac{1}{\sigma} t_i^{\frac{1}{\sigma}-1} e^{-1/\sigma(\beta_1 X_i+\beta_2' Z_i)}\right)^{\delta_i}\left(1+t_i^{\frac{1}{\sigma}} e^{-1/\sigma(\beta_1 X_i+\beta_2' Z_i)}\right)^{-(\delta_i+1)} \tag{8}$$

Remark 2.1

A suitable model can be chosen by following the routine procedure based on the validation sample. The auxiliary information can be utilized based on the selected parametric model following the estimation method introduced in the following section.

Method of the Estimation

The regression parameters can be estimated through the use of the maximum likelihood method, say $\hat{\beta} = Argmax_\beta l(\beta)$ which solves the estimating equations

$$\frac{\partial l(\beta)}{\partial \beta} = 0,$$

where

$$l(\beta) = \sum_{i=1}^{n}\left(\delta_i \log\left(h(S_i;\beta,X_i,Z_i)\right) + \log\left(S(S_i;\beta,X_i,Z_i)\right)\right) \tag{9}$$

Both the hazard and survival functions depend on X_i, which is available only for the validation sample. The non-validation sample does not contain the X_i measurements. However, there is auxiliary information available. In order to enhance the efficiency of the data analysis, one should take the auxiliary variables into consideration. In this paper we propose to predict the unobserved X_i's from their corresponding auxiliary covariates, the W_i's, by using kernel smoothing and then using these to estimate the hazard and survival functions. For details about kernel smoothing, one can see Nadaraya [25], Watson [26] and Wand and Jones [27]. The equation to estimate the unobserved X_i values is

$$\hat{X}_i = \frac{\sum_{j\in V} X_j k(\frac{W_i - W_j}{h})}{\sum_{j\in V} k(\frac{W_i - W_j}{h})}, \tag{10}$$

where k( ) is the kernel function and h is the chosen bandwidth for smoothing. Note that the selection of the bandwidth should be such that the bandwidth conditions of Theorem 1 be satisfied. Here the optimal bandwidth is chosen as $\mathrm{h} = 2\sigma_{\mathrm{u}}\mathrm{n}^{-1/3}$, as suggested by Zhou and Wang [18].

We can therefore write the estimated likelihood and estimated log-likelihood as

$$EL(\beta) = \prod_{i\in V} h(S_i;\beta,X_i,Z_i)^{\delta_i} S(S_i;\beta,X_i,Z_i) \times \prod_{i\in\hat{V}} \hat{h}(S_i;\beta,W_i,Z_i)^{\delta_i} \hat{S}(S_i;\beta,X_i,Z_i),$$

and

$$\log\left(EL(\beta)\right) = \sum_{i\in V}(\delta_i \log\left(h\left(S_i;\beta,X_i,Z_i\right)\right) + \log(S(S_i;\beta,X_i,Z_i))) +$$
$$\sum_{i\in\bar{V}}(\delta_i \log\left(\hat{h}\left(S_i;\beta,X_i,Z_i\right)\right) + \log(\hat{S}(S_i;\beta,X_i,Z_i))) \tag{11}$$

where $\hat{h}(S_i;\beta,X_i,Z_i) = h\left(S_i;\beta,\hat{X}_i,Z_i\right)$ and $\hat{S}(S_i;\beta,X_i,Z_i) = S\left(S_i;\beta,\hat{X}_i,Z_i\right)$

For our reduced case in Section 2.1.2, equation (11) becomes

$$\log\left(EL(\beta)\right) = \sum_{i\in V}(\delta_i\left(\log\left(\frac{1}{\sigma}t_i^{\frac{1}{\sigma}-1}\right) - \frac{1}{\sigma}\left(\beta_1 X_i + \beta_2' Z_i\right)\right) - t_i^{\frac{1}{\sigma}} e^{-\frac{1}{\sigma}\left(\beta_1 X_i + \beta_2' Z_i\right)})$$
$$\sum_{i\in\bar{V}}(\delta_i\left(\log\left(\frac{1}{\sigma}t_i^{\frac{1}{\sigma}-1}\right) - \frac{1}{\sigma}\left(\beta_1\hat{X}_i + \beta_2' Z_i\right)\right) - t_i^{\frac{1}{\sigma}} e^{-\frac{1}{\sigma}\left(\beta_1\hat{X}_i + \beta_2' Z_i\right)}) \tag{12}$$

and for our proportional odds example in Section 2.1.3, we have

$$\log\left(EL(\beta)\right) = \sum_{i\in V}(\delta_i\left(\log\left(\frac{1}{\sigma}t_i^{\frac{1}{\sigma}-1}\right) - \frac{1}{\sigma}\left(\beta_1 X_i + \beta_2' Z_i\right)\right) - \left(\delta_i + 1\right)\log(1 + t_i^{\frac{1}{\sigma}} e^{-\frac{1}{\sigma}\left(\beta_1 X_i + \beta_2' Z_i\right)}))$$
$$+\sum_{i\in\bar{V}}(\delta_i\left(\log\left(\frac{1}{\sigma}t_i^{\frac{1}{\sigma}-1}\right) - \frac{1}{\sigma}\left(\beta_1\hat{X}_i + \beta_2' Z_i\right)\right) - \left(\delta_i + 1\right)\log(1 + t_i^{\frac{1}{\sigma}} e^{-\frac{1}{\sigma}\left(\beta_1\hat{X}_i + \beta_2' Z_i\right)})) \tag{13}$$

The estimates of the regression parameters are then the maximizers of the estimated log-likelihood function,

$$\hat{\beta}_{EL} = Argmax_{\beta} \log(EL(\beta))$$

Which can be obtained by solving the estimating equations

$$U(\beta) = \frac{\partial \log(EL(\beta))}{\partial\beta} = 0.$$

When the unobservable X_i's, i = 1,…, n_V, are replaced using the proposed kernel smoothing, the unknown parameters can be estimated with existing programs, such as those written in R or SAS. However, the corresponding variance estimates are going to fail due to the estimated unknown X_i's. Hence in this paper we propose to use the Newton-Raphson algorithm to estimate the regression parameters. The variance and covariance matrix of the estimator can be estimated from the calculation process.

Remark 3.1

1. The distribution of the failure time needs be specified in this procedure. An appropriate one can be chosen based on the validation sample by using the routine procedure for parametric model selection. See, for example, Lawless [24].

2. The direct imputation of the unobservable covariate with its kernel smoothing estimation is due to the consideration of model robustness. If there exists slight misspecification of the model, the maximum likelihood estimator of the regression parameters based on the kernel smoothing of unknown expectation in the likelihood function will be inconsistent. This was also observed in our simulation studies but the results will not be reported.

3. The scale parameter, if unknown, can be estimated based on the validation sample routinely, or by adding another equation which is obtained by differentiating the estimated log likelihood with respect to this scale parameter, say,

$$\frac{\partial \log(EL(\beta))}{\partial\sigma} = 0.$$

4. This method can accommodate both missing covariate and mismeasured covariate problems.

5. This method can be extended by using local linear approximation (see Fan and Wang 2009) instead of the equation (10). In nonparametric smoothing, local linear approximation usually performs better than kernel smoothing. The method also accommodates models with instrumental variables.

Asymptotics

Under the regularity conditions (a), (b), (c), and the condition (d) listed in the appendix, our proposed estimates of the regression parameters by maximizing the estimated likelihood function are jointly consistent and asymptotically normally distributed, as described in the following theorem.

Suppose that the order of the kernel function K is, say

$$\int u^q K(u)\,du = 0, forq = 1,2,\ldots,\alpha - 1, \int u^{\alpha} K(u)\,du \neq 0.$$

Theorem 4.1

Under the conditions (a), (b), (c), and (d) in the appendix, and the bandwidth condition that $nh^{2\alpha} \to 0, nh^2 \to \infty$, we have

1. $\hat{\beta}_{EL}$ is a consistent estimator of β.

2. $\sqrt{n}\left(\hat{\beta}_{EL} - \beta\right) \to N(0,\Sigma_{EL})$ in distribution, as $n \to \infty$,

where

$$\Sigma_{EL} = I^{-1}\left(\rho I + \left(1 - \rho\right)\Sigma_{EL}\right)I^{-1},$$

$$I = -E\left\{\frac{\partial^2 y}{\partial\beta'\partial\beta}\left[\delta_i \log\left(h\left(S_i;\beta,X_i,Z_i\right)\right) + \log\left(S\left(S_i;\beta,X_i,Z_i\right)\right)\right]\right\}$$

$$\Sigma = \mathrm{cov}\left\{\frac{\partial}{\partial\beta}\left[\delta_i \log\left(h\left(S_i;\beta,E\left(X_i|W_i\right),Z_i\right)\right) + \log\left(S\left(S_i;\beta,E\left(X_i|W_i\right),Z_i\right)\right)\right]\right\},$$

and $\rho = \lim_{n\to\infty}\frac{n_V}{n}$ is the ratio of the sample size of the validation sample and the total sample size.

The variance and covariance matrix of $\hat{\hat{\beta}}_{EL}$ can be consistently estimated by their sample counterpart from the estimated log-likelihood function,

$$\hat{\Sigma}_{\hat{\beta}_{EL}} = \frac{1}{n}\hat{\Sigma}_{EL} = \frac{1}{n}\hat{I}^{-1}\left(\rho\hat{I} + \left(1 - \rho\right)\hat{\Sigma}\right)\hat{I}^{-1} \tag{14}$$

In equation (14), $\hat{I}$ is the observed Fisher information matrix with elements

$$\hat{I}_{ij} = -\frac{1}{n}\frac{\partial^2 \log(EL(\beta))}{\partial\beta_i\partial\beta_j},$$

when replacing the unknown regression parameters with their estimates.

$\boldsymbol{\Sigma}$ is the sample variance-covariance matrix of the non-validation half of the estimating function $\mathrm{U}(\beta)$ which estimates its corresponding population counterpart. The proof of the theorem is deferred to the appendix.

Results of Numerical Studies

Simulations

In this section we investigate the small sample performance of our proposed estimator. We carry out extensive simulations in order to compare its efficiency and accuracy with other alternative estimation methods. We compare the proposed estimator based on the estimated likelihood method previously discussed $(\hat{\beta}_{EL})$ with three different estimators. The first estimator $(\hat{\beta}_V)$ is based only on the validation sample, ignoring the observations with missing values for X_i. This does not require the estimation of the unobserved data but as a trade-off must deal with a smaller sample size. The second estimator $(\hat{\beta}_N)$ is based on the naive use of the auxiliary covariate as the true covariate in the sample. In this case we assume that for the non-validation sample, the unobserved X_i values are equal to the observed W_i values, ignoring the measurement error. The third estimator $(\hat{\beta}_C)$ is based on a complete knowledge of the data. This is the best case scenario that would exist if we actually observed the X_i values for the non-validation sample and thus are working with a validation sample of the full sample size. We expect the efficiency and accuracy of $(\hat{\beta}_{EL})$ to be better than that of $(\hat{\beta}_V)$ and close to that of $(\hat{\beta}_C)$.

Simulations are done for the cases in Sections 2.1.2 and 2.1.3. For both, the random X_i and Z_i data are generated from a uniform distribution with a lower limit of 0 and an upper limit of 5, X_i, Z_i ~ uniform (0,5). The auxiliary covariate W_i is defined as $W_i = X_i + U_i$ where $Ui \sim \mathrm{N}(0, \sigma_u^2)$ and σ_u^2 determines the size of the measurement error in our sampling. Given X_i and Z_i, the random failure times T_i for the first case are generated from the equations

$$\mathrm{T_i} = \exp\{\mathrm{Y_i}\};$$

and

$$\mathrm{Y_i} = \beta_1 \mathrm{X_i} + \beta_2' \mathrm{Z_i} + \varepsilon_i;$$

where the ε_i's are iid and are following a standard extreme value distribution as discussed in Section 2.1.2. For the proportional odds model, we have

$$\mathrm{Y_i} = \beta_1 \mathrm{X_i} + \beta_2' \mathrm{Z_i} + \sigma \mathrm{V_i};$$

where V_i follows the standard logistic distribution as shown in Section 2.1.3, and we let = 1. The parameters $\beta' = (\beta_1, \beta_2')$ are chosen prior to the simulations. The random censoring times C_i are generated from a uniform distribution, $\mathrm{C_i} \sim \mathrm{uniform}[0;\ \mathrm{c_{lim}}]$, where c_{lim} is chosen such that the results have approximately 30% or 50% of the failure times censored.

For each set of simulations, there are pre-determined n and n_V values and the X_i, W_i, Z_i, T_i, and C_i data is generated as outlined above. We estimate the $n_{\bar{V}} X_i$ values for the non-validation set for use in the estimated likelihood method from the $\mathrm{n_V X_i}$'s in the validation set and the n W_i's by using kernel smoothing as depicted in equation (10). For our calculations, we use the Gaussian kernel function, which has an order of 2,

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

where $u = (W_i - W_j)/h$ and we take bandwidth $h = 2\sigma_u n^{-1/3}$ as used by Zhou and Wang [18]. Then we calculate all of the $\hat{\beta}$'s through the Newton-Raphson Method using the appropriate sets of data for each estimator. By using this method, we are able to solve the equations

$$\frac{\partial l(\beta)}{\partial \beta} = 0,$$

for $\hat{\beta}_V, \hat{\beta}_N, and\ \hat{\beta}_C$ and solve

$$\frac{\partial \log(EL(\beta))}{\partial \beta} = 0$$

for $\hat{\beta}_{EL}$.

For each set of simulations, we calculate the standard error (SE), standard deviation (SD), and the percent of estimators covered when using a 95% confidence interval, the coverage probability (CP). The standard errors are obtained by calculating the sample variance-covariance matrix of the maximum likelihood estimates for the parameters estimated over all simulations. The standard deviations are obtained from the estimated variance using equation (14). The values for CP are obtained by keeping track in each simulation if the true β values are within a 95% confidence interval surrounding the estimates using that simulation's estimated SD value.

The parameter values used in our simulations were $\beta' = (\beta_1, \beta_2)$ = (log(2), log(1.5)). We tested with these values in a few different situations. We used σ_u = 0.2 and σ_u = 0.8, sample sizes n = 200 and n = 500 and censoring rates of 30% and 50%. We chose a constant validation ratio of $\rho = \frac{n_V}{n} = 0.5$ and each simulation is repeated 1000 times. The simulation results are summarized in table 1 for the exponential regression model, and in table 2 for the proportional odds model.

We have also conducted simulations for other parameter settings, such as (1) $\sigma_u = 0.6$; (2) a lower validation rate of 30%; (3) with an unknown but estimated measurement error variance σ_u^2; (4) with an estimated σ in the AFT model. The results were all similar to those reported and are hence skipped.

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

Both $\hat{\beta}_V$ and $\hat{\beta}_{EL}$ are performing very well. The naive estimator $\hat{\beta}_N$ is biased at higher values of measurement error, σ_u.

The $\hat{\beta}_{EL}$ estimator is more efficient than the $\hat{\beta}_V$ estimator in the sense that the latter has bigger standard errors.

If ρ were to increase to 1, the relative efficiencies would go to 1 since aside from having to estimate the unobserved X_i's for the non-validation set versus excluding all of the non-validation data, the methods of estimation are the same.

The proposed variance estimator (14) for $\hat{\beta}_{EL}$ results in a good estimate of the true variance, $\Sigma_{\hat{\beta}_{EL}}$, for both models.

The coverage probabilities of the 95% confidence interval are good for all estimators except $\hat{\beta}_N$ when σ_u is large. In the case where $\sigma_u = 0.8$ they were bad and got worse as we increased the sample size but kept the same ρ since it increased the total data with error in each estimation without lessening its effect with a larger proportion of known X_i values, while the width of the confidence interval is shortened by the increasing sample size.

In comparing the two models, the exponential regression model appears to have smaller SE and SD values for all four estimators, but

	Censor										
n	Rate	σ_u	$\hat{\beta}$	$\hat{\beta}_1$	$SE_{\hat{\beta}_1}$	$SD_{\hat{\beta}_1}$	$CP_{\hat{\beta}_1}$	$\hat{\hat{\beta}}_2$	$SE_{\hat{\beta}_2}$	$SD_{\hat{\beta}_2}$	$CP_{\hat{\beta}_2}$
200	0.3	0.2	V	0.694	0.068	0.067	0.949	0.403	0.063	0.062	0.947
			N	0.690	0.047	0.046	0.939	0.408	0.044	0.043	0.950
			EL	0.693	0.048	0.047	0.938	0.405	0.044	0.043	0.940
			C	0.694	0.047	0.047	0.942	0.404	0.044	0.043	0.950
		0.8	V	0.695	0.065	0.067	0.947	0.403	0.063	0.062	0.952
			N	0.644	0.048	0.043	0.748	0.459	0.048	0.042	0.715
			EL	0.699	0.051	0.050	0.950	0.407	0.048	0.046	0.943
			C	0.694	0.046	0.047	0.946	0.405	0.043	0.043	0.953
	0.5	0.2	V	0.694	0.085	0.085	0.955	0.403	0.076	0.074	0.937
			N	0.691	0.060	0.059	0.954	0.407	0.054	0.052	0.940
			EL	0.694	0.060	0.059	0.949	0.404	0.054	0.052	0.939
			C	0.695	0.060	0.059	0.956	0.404	0.054	0.052	0.940
		0.8	V	0.695	0.086	0.085	0.946	0.406	0.075	0.075	0.951
			N	0.637	0.060	0.054	0.788	0.463	0.055	0.050	0.782
			EL	0.695	0.063	0.061	0.937	0.404	0.056	0.053	0.938
			C	0.695	0.060	0.059	0.944	0.406	0.052	0.052	0.957
500	0.3	0.2	V	0.692	0.043	0.042	0.941	0.406	0.038	0.039	0.957
			N	0.688	0.029	0.029	0.940	0.410	0.027	0.027	0.953
			EL	0.691	0.030	0.029	0.944	0.407	0.027	0.027	0.954
			C	0.691	0.029	0.029	0.944	0.406	0.027	0.027	0.957
		0.8	V	0.696	0.043	0.042	0.942	0.402	0.040	0.039	0.935
			N	0.644	0.032	0.027	0.530	0.458	0.031	0.026	0.473
			EL	0.700	0.033	0.031	0.937	0.405	0.031	0.029	0.935
			C	0.694	0.031	0.029	0.932	0.403	0.028	0.027	0.946
	0.5	0.2	V	0.697	0.052	0.053	0.958	0.404	0.046	0.046	0.949
			N	0.691	0.036	0.037	0.949	0.409	0.032	0.033	0.945
			EL	0.694	0.037	0.037	0.948	0.406	0.033	0.033	0.937
			C	0.695	0.036	0.037	0.953	0.405	0.032	0.033	0.945
		0.8	V	0.696	0.053	0.053	0.946	0.404	0.047	0.046	0.952
			N	0.637	0.037	0.034	0.589	0.462	0.036	0.031	0.566
			EL	0.695	0.039	0.038	0.950	0.404	0.036	0.034	0.937
			C	0.695	0.037	0.037	0.957	0.405	0.033	0.033	0.952

Table 1: Results after 1000 simulations for β' = (log(2),log(1.5)) = (0.693, 0.405) with ρ = 0.5 and $h = 2\sigma_u n^{(-1/3)}$ using the exponential regression model.

the log-logistic regression model does not experience such a dramatic decrease in CP for the $\hat{\beta}_N$ estimator when σ_u was increased. This is likely due to the mentioned larger SD values used in the calculations.

Application to PBC data

We apply the proposed method to analyze data from the Mayo Clinic trial in PBC of the liver. PBC is a chronic liver disease that inflames and slowly destroys the bile ducts in the liver. Bile is a liquid produced in the liver which travels through these bile ducts to assist digestion in the small intestines. When these ducts are damaged, the bile builds up within the liver, causing damage and leading to cirrhosis. Scar tissue will then start to replace healthy liver tissue, impairing its ability to function properly. While the cause of PBC is unknown, it is believed to be a type of autoimmune disorder where the immune system attacks the bile ducts. Approximately 90% of patients who develop PBC are women, most often between the ages of 40 and 60. It is typical for those with PBC to not have any symptoms when diagnosed because it is often diagnosed early from routine blood tests checking the liver. Since it is a slow acting disease, if it is found early the patient may slow the progression of cirrhosis through treatment and still have many years with a healthy lifestyle, and possibly even have a normal life expectancy if their case is not too dire. However, there is currently no known cure for the disease. The only known way to effectively remove PBC is through a liver transplant. If the patient is deemed appropriate for a transplant, steps need to be taken to prevent the immune system from damaging the new liver [28,29].

In the random Mayo Clinic trial, a total of 418 patients were eligible. Of these 418, mostly complete data was obtained from the first 312 patients. The other 106 patients were not part of the actual clinical trial but agreed to have some basic measurements taken and to be followed for survival. The variables that we used for our analysis were time, the number of days between registration and the earlier of death, transplantation, or the study analysis date; status, the indicator of a patient's status at their endpoint in the trial, denoted as 0, 1, or 2, corresponding to censored, transplant, or dead, respectively; Aspartate Aminotransferase (in U/ml), once referred to as SGOT; bili, serum bilirubin (in mg/dl); albumin, serum albumin (in mg/dl); age, patient's age (in years); protime, standardized blood clotting time.

In this clinical trial, one of the variables that were measured only for the first 312 cases was aspartate aminotransferase, due to some

	Censor										
n	Rate	σ_u	$\hat{\beta}$	$\hat{\beta}_1$	X_i's	$SD_{\hat{\beta}_1}$	$CP_{\hat{\beta}_1}$	$\hat{\beta}_2$	$SE_{\hat{\beta}_2}$	$SD_{\hat{\beta}_2}$	$CP_{\hat{\beta}_2}$
200	0.3	0.2	V	0.694	0.098	0.097	0.948	0.408	0.097	0.096	0.955
			N	0.691	0.067	0.068	0.948	0.407	0.066	0.067	0.948
			EL	0.694	0.068	0.069	0.957	0.405	0.067	0.067	0.950
			C	0.694	0.067	0.068	0.947	0.405	0.066	0.067	0.951
		0.8	V	0.692	0.098	0.097	0.953	0.405	0.095	0.096	0.943
			N	0.639	0.069	0.066	0.848	0.447	0.066	0.066	0.909
			EL	0.692	0.073	0.070	0.936	0.405	0.069	0.069	0.952
			C	0.692	0.068	0.068	0.953	0.407	0.065	0.067	0.961
	0.5	0.2	V	0.697	0.112	0.109	0.938	0.407	0.106	0.103	0.943
			N	0.689	0.081	0.076	0.937	0.412	0.076	0.072	0.943
			EL	0.692	0.081	0.076	0.941	0.409	0.076	0.073	0.941
			C	0.693	0.081	0.076	0.937	0.409	0.076	0.072	0.945
		0.8	V	0.697	0.108	0.109	0.954	0.405	0.102	0.104	0.956
			N	0.642	0.076	0.073	0.874	0.447	0.073	0.071	0.912
			EL	0.693	0.080	0.078	0.948	0.403	0.076	0.074	0.946
			C	0.694	0.077	0.076	0.949	0.407	0.074	0.072	0.953
500	0.3	0.2	V	0.695	0.061	0.061	0.948	0.401	0.060	0.060	0.951
			N	0.690	0.043	0.043	0.956	0.406	0.042	0.042	0.948
			EL	0.693	0.043	0.043	0.952	0.403	0.042	0.042	0.948
			C	0.694	0.043	0.043	0.954	0.403	0.042	0.042	0.950
		0.8	V	0.696	0.062	0.061	0.950	0.405	0.061	0.060	0.944
			N	0.643	0.042	0.042	0.776	0.447	0.041	0.042	0.830
			EL	0.696	0.045	0.044	0.945	0.404	0.043	0.043	0.948
			C	0.696	0.043	0.043	0.950	0.406	0.041	0.042	0.958
	0.5	0.2	V	0.694	0.069	0.068	0.952	0.402	0.067	0.065	0.934
			N	0.691	0.048	0.048	0.943	0.407	0.047	0.045	0.939
			EL	0.694	0.049	0.048	0.947	0.404	0.047	0.046	0.940
			C	0.695	0.048	0.048	0.944	0.404	0.047	0.046	0.940
		0.8	V	0.696	0.066	0.068	0.961	0.406	0.066	0.065	0.942
			N	0.640	0.048	0.046	0.753	0.449	0.047	0.045	0.826
			EL	0.691	0.051	0.049	0.942	0.405	0.049	0.046	0.942
			C	0.694	0.049	0.048	0.953	0.408	0.047	0.046	0.956

Table 2: Results after 1000 simulations for β' = (log(2), log(1.5)) = (0.693, 0.405) with ρ = 0.5 and $h = 2\sigma_u n^{(-1/3)}$ using the log-logistic regression model.

difficulties. We are extremely interested in knowing its relationship with the patients' survival. In order to estimate the unobserved AST values for the other 106 patients, which form the non-validation sample in this analysis; we chose serum bilirubin to act as the auxiliary covariate, W. There is data observed for serum bilirubin for every patient and it was therefore available to be used in kernel smoothing. To determine an estimate for σ_u to use in the calculation of the bandwidth, we used the least squares method to the regression equation $X_i = \beta_0 + \beta_1 W_i + \varepsilon_i$ and calculated the MSE so that $\widehat{\sigma_u} = \sqrt{MSE}$ = 0.369. The scale parameter for the AFT models were estimated based only on the validation data and then applied to the analysis using the proposed approach, where we calculated σ = 0.873 for proportional hazards model and σ = 0.676 for proportional odds model.

To test along side of AST, we included the variables serum albumin, age and protime in vector Z. These variables were measured for most of the patients, and thus were good choices for Z. There were two cases in the non-validation set with missing values for protime, so they were omitted. This left us with a validation set of 312 patients and a non-validation set of 104 patients. We decided to not include edema, even though it was measured for all patients, because there was not a single patient in the non-validation set that had edema despite diuretic therapy. For our calculations, we took the logarithms of the data for AST, serum bilirubin, serum albumin, and protime. Also, we treated having a transplant the same as being censored, so a status of 0 or 1 resulted in $\delta = 0$, and thus a status of 2 resulted in $\delta = 1$.

The proportional hazards and the proportional odds models Fit this part of the data equally well, in the sense that we obtained very close AIC values for both. The results of applying these models are hence provided below.

Tables 3 and 4 show the results of the analysis on the PBC data using our estimated likelihood method on all 416 observations and the validation set method on just 312 observations, using both of the previously discussed models. Since we use a separate variable for our auxiliary covariate not just a measurement of X containing error, the naive method is not appropriate for this example. The estimates of the variables' coefficients, their estimated standard deviations, and p-values are listed in the tables.

In Table 3, we see that except for the case of log (albumin), the standard deviations are all smaller in the estimated likelihood method than the validation set method, while every standard deviation is

Method	Variable	$\hat{\beta}$	SD	P-Value
VA	log(AST)	-0.269	0.160	0.093
	log(albumin)	5.128	0.413	<0.001
	age	0.017	0.008	0.035
	log(protime)	1.766|0.474		< 0.001
EL	log(AST)	0.342	0.145	0.018
	log(albumin)	4.737|0.436		< 0.001
	age	0.016	0.007	0.027
	log(protime)	2.112	0.435	< 0.001

Table 3: AFT model analysis of PBC data using validation set and estimated likelihood methods using the exponential regression model.

Method	Variable	$\hat{\beta}$	SD	P-Value
VA	log(AST)	-0.384	0.181	0.034
	log(albumin)	6.455	0.554	<0.001
	age	0.022	0.008	0.008
	log(protime)	1.252|0.527		0.017
EL	log(AST)	0.460	0.160	0.004
	log(albumin)	5.970|0.506		< 0.001
	age	0.021	0.007	0.003
	log(protime)	1.656	0.462	< 0.001

Table 4: AFT model analysis of PBC data using validation set and estimated likelihood methods using the log-logistic regression model.

smaller for the estimated likelihood method in Table 4. In each case, the magnitudes of the estimated coefficients vary between estimation methods, but they show the same relationships between the covariates and time of death. Most importantly however, is that the significance of one of the coefficients differs between estimation methods. For the exponential regression model, we note that the p-value for log (AST) is less than 0.05 only for the estimated likelihood method. Therefore, when using the smaller sample sizes in the validation set method we are unable to conclude that all of the coefficients are significantly different from zero for either model, but all four coefficients become significant when using the estimated likelihood method. This emphasizes the importance of not omitting some of your data since as we have seen, it is possible to accidentally conclude that a significant variable from your analysis is in fact, not significant.

Discussions

In this paper we proposed to use the kernel smoothing method to include the informative auxiliary covariate into the statistical inference of failure time data based on parametric AFT models. An estimator of the regression parameters is obtained through the maximization of an estimated likelihood function. The asymptotics of the proposed estimator is investigated. A consistent estimator of the estimation variance is also proposed. Simulation studies are conducted for the case when the error of the AFT model follows a standard extreme value distribution, as well as a standard logistic distribution. The proposed method is then applied to the PBC data as an illustration.

The motivation of conducting this study is twofold. It is well known that the AFT models are robust to mis-specifications when some of the predictive regressors are ignored. The regression coefficients are invariant, at least for distributions within the Weibull family. Secondly, the partial likelihood method is less efficient in the case of small sized samples, although it is asymptotically efficient when the sample size goes to infinity [11].

The authors are currently investigating semi-parametric AFT models with auxiliary covariates. The outcome is going to be reported in a forthcoming paper.

Acknowledgment

The research of both authors was supported in part by National Sciences and Engineering Research Council of Canada.

References

1. Fleming TR, Harrington DP (1991) Counting processes and survival analysis, John Wiley & Sons, Inc. New York.
2. Prentice RL (1982) Covariate measurement errors and parameter estimation in failure time regression model. Biometrika 69: 331-342.
3. Rubin DB (1976) Inference and missing data. Biometrika 63: 581-592.
4. Fuller WA (1987) Measurement error models. John Wiley & Sons, New York.
5. Carrol RJ, Rupert D, Stefanski LA (1995) Measurement Error in Nonlinear Models. Chapman and Hall, London.
6. Wang NY, Lin XH, Gutierrez RG, Carrol RJ (1998) Bias analysis and SIMEX approach in generalized linear mixed measurement error models. J Am Statist Assoc 93: 249-261.
7. Meng X, Schenker N (1999) Maximum likelihood estimation for linear regression models with right censored outcomes and missing predictors. Comput Stat Data Anal 29: 471-483.
8. Cheng S, Wang N (2001) Linear transformation models for failure time data with covariate measurement error. J Am Statist Assoc 96: 706-716.
9. Yu M, Nan B (2010) Regression calibration in semiparametric accelerated failure time models. Biometrics 66: 405-414.
10. Cox DR (1972) Regression models and life-tables. J R Stat Soc Ser B 34: 187-220.
11. Cox DR, Oakes D (1984) Analysis of survival data. Chapman and Hall, London.
12. Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edn. New York: John Wiley & Sons, Inc.
13. Hu P, Tsiatis AA, Davidian M (1998) Estimating the parameters in the Cox model when covariate variables are measured with error. Biometrics 54: 1407-1419.
14. Hu C, Lin DY (2002) Cox regression with covariate measurement error. Scandinavian Journal of Statistics 29: 637-655.
15. Pepe MS, Flemming TR (1991) A nonparametric method for dealing with mismeasured covariate data. J Am Statist Assoc 86: 108-113.
16. Pepe MS (1992) Inference using surrogate outcome data and a validation sample. Biometrika 79: 355-365.
17. Zhou H, Pepe MS (1995) Auxiliary covariate data in failure time regression. Biometrika 82: 139-149.
18. Zhou H, Wang C-Y (2000) Failure time regression with continuous covariates measured with error. J R Stat Soc Ser B 62: 657-665.
19. Zhou H, Chen J, Cai J (2002) Random effects logistic regression analysis with auxiliary covariates. Biometrics 58: 352-360.
20. Jiang J, Zhou H (2007) Additive hazard regression with auxiliary covariates. Biometrika 94: 359-369.
21. Fan Z, Wang X (2009) Marginal hazards model for multivariate failure time data with auxiliary covariates. J Nonparametr Stat 21: 771-786.
22. Liu Y, Zhou H, Cai J (2009) Estimated pseudo-partial-likelihood method for correlated failure time data with auxiliary covariates. Biometrics 65: 1184-1193.
23. He W, Yi GY, Xiong J (2007) Accelerated failure time models with covariates subject to measurement error. Stat Med 26: 4817-4832.
24. Lawless JF (2003) Statistical models and methods for lifetime data, 2nd edn, Wiley & Sons Inc., Hoboken, NJ.
25. Nadaraya EA (1964) On estimating regression. Theory Probab Appl 9: 141-142.
26. Watson GS (1964) Smooth regression analysis. Sankhya: The Indian Journal of Statistics Ser A 26: 359-372.

27. Wand M, Jones M (1995) Kernel Smoothing. Chapman and Hall, London.

28. (2008) “Primary Biliary Cirrhosis (PBC).” National Digestive Diseases Information Clearinghouse (NDDIC). National Institute of Diabetes and Digestive and Kidney Disease, National Institute of Health.

29. (2011) “Primary Biliary Cirrhosis (PBC).” American Liver Foundation.

30. Cramer H (1951) Mathematical Methods of Statistics. Princeton: Princeton University Press.