Journal of Statistical Software

January 2012, Volume 46, Code Snippet 1. *http://www.jstatsoft.org/*

SIMEX R Package for Accelerated Failure Time Models with Covariate Measurement Error

Wenqing He
University of
Western Ontario

Juan Xiong
University of
Western Ontario

Grace Y. Yi
University of
Waterloo

Abstract

It has been well documented that ignoring measurement error may result in substantially biased estimates in many contexts including linear and nonlinear regressions. For survival data with measurement error in covariates, there has been extensive discussion in the literature with the focus typically centered on proportional hazards models. The impact of measurement error on inference under accelerated failure time models has received relatively little attention, although these models are very useful in survival data analysis. He *et al.* (2007) discussed accelerated failure time models with error-prone covariates and studied the bias induced by the naive approach of ignoring measurement error in covariates. To adjust for the effects of covariate measurement error, they described a simulation and extrapolation method. This method has theoretical advantages such as robustness to distributional assumptions for error prone covariates. Moreover, this method enjoys simplicity and flexibility for practical use. It is quite appealing to analysts who would like to accommodate covariate measurement error in their analysis. To implement this method, in this paper, we develop an R package for general users. Two data sets arising from clinical trials are employed to illustrate the use of the package.

Keywords: accelerated failure time models, measurement error, R package, simulation and extrapolation algorithm, survival data.

1. Introduction

There has been extensive interest in discussing inference methods for survival data with covariates subject to measurement error. It is known that standard inferential procedures may produce biased estimation if measurement error is not properly taken into account (e.g., Carroll *et al.* 2006). With proportional hazards models a number of methods have been proposed to correct bias induced by measurement error (e.g., Prentice 1982; Li and Lin 2003; Yi and Lawless 2007).

Although the impact of covariate measurement error on inferential procedures is well understood for proportional hazards models, there is little discussion about its impact under accelerated failure time (AFT) models, which have proved to be useful in survival analysis (e.g., Lawless 2003).

Unlike the proportional hazards model that focuses modeling on the hazard function, an AFT model directly facilitates the relationship between the failure time (or its transformation) and covariates via a regression model. This formulation allows a direct and transparent interpretation of covariate effects on the change of the failure time. As noted by D. R. Cox (Reid 1994, p. 450), an AFT model is "in many ways more appealing because of its quite direct physical interpretation". In certain applications AFT models could provide better fit to data than proportional hazards models (e.g., Zeng and Lin 2007).

Under Weibull regression models, Giménez *et al.* (1999, 2006) investigated inference methods using the corrected score approach discussed by Nakamura (1992). With general AFT models, He *et al.* (2007) discussed inference procedures to account for effects of covariate measurement error using a simulation-extrapolation (SIMEX) approach. The developed SIMEX method for AFT models is simple to implement and flexible to cover many applications. Moreover, this method is robust in a sense that distributions of covariates, including error-prone covariates, are left unspecified. Because of those features, this method becomes quite appealing to analysts who would like to accommodate covariate measurement error in their analysis of survival data.

Despite that there have been great advances on methodology of addressing covariate measurement error for survival analysis, the methods developed in the literature have not enjoyed widespread use in practice. Reluctance to adopt these methods may be due, in part, to the lack of available software to implement these methods. To address this practical issue, we develop an R package (R Development Core Team 2011), entitled **simexaft**, to implement the SIMEX method discussed in He *et al.* (2007) so that this method can be accessible for general users.

The remainder is organized as follows. Section 2 introduces the notation and model formulation. In Section 3 we describe the SIMEX method and its implementation in R. The developed R package is illustrated in Section 4 with two survival data sets: one arising from a subset of Busselton Health Study (Knuiman *et al.* 1994), and the other from a multi-center clinical trial (Fuchs *et al.* 1994). General discussion is included in the last section. The developed R package is available from the Comprehensive R Archive Network (CRAN) at `http//CRAN.R-project.org/package=simexaft`.

2. Notation and framework

For $i = 1, 2, \ldots, n$, let T_i and C_i be the failure and censoring times for subject i, respectively, and δ_i be the censoring indicator variable taking value 1 if $T_i \leq C_i$ and 0 otherwise. Denote $t_i = \min(T_i, C_i)$. Independent censoring is assumed. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ be the $p \times 1$ covariates subject to possible measurement error, and $\mathbf{z}_i$ the vector of covariates free of error. Response variable $Y_i = \log(T_i)$ is characterized by the AFT model, given by

$$Y_i = \boldsymbol{\beta}_x^\top \mathbf{x}_i + \boldsymbol{\beta}_z^\top \mathbf{z}_i + \epsilon_i \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^\top, \boldsymbol{\beta}_z^\top)^\top$ is the vector of regression parameters of interest, and $\boldsymbol{\beta}_z$ may include

the intercept coefficient. Here ϵ_i assumes a distribution $G(\cdot)$ with parameters $\boldsymbol{\alpha}$. Common choices of the distribution $G(\cdot)$ include Weibull, exponential, Gaussian, logistic, log-normal and log-logistic distributions (e.g., Lawless 2003).

Let $\mathbf{W}_i$ be an observed version of covariate $\mathbf{x}_i$. $\mathbf{W}_i$ and $\mathbf{x}_i$ are assumed to follow a classical additive measurement error model. That is, conditional on $\mathbf{x}_i$ and $\mathbf{z}_i$,

$$\mathbf{W}_i = \mathbf{x}_i + \mathbf{e}_i, \tag{2}$$

where $\mathbf{e}_i$ follows, independent of $\mathbf{x}_i$ and $\mathbf{z}_i$, a normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_e = [\sigma_{jk}]_{p\times p}$. Here we assume nondifferential measurement error, i.e., $f(Y_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{W}_i) = f(Y_i|\mathbf{x}_i, \mathbf{z}_i)$. This mechanism says that the contribution from the observed $\mathbf{W}_i$ is not informative, given the true covariates $\mathbf{x}_i$ and $\mathbf{z}_i$. This mechanism applies commonly to many practical problems, especially when the true and observed covariates occur at a fixed time point and the response is measured at a later time (Carroll *et al.* 2006, p. 36).

The parameters in $\boldsymbol{\Sigma}_e$ can be estimated, for example, when repeated measurements for $\mathbf{x}_i$ are available. In other situations, the parameters in $\boldsymbol{\Sigma}_e$ may be assumed known which is, for instance, based on prior knowledge or other similar studies. When conducting sensitivity analysis to assess the impact of different degrees of measurement error on estimation of the response parameters, the parameters in $\boldsymbol{\Sigma}_e$ are typically specified to be known based on background information about the measurement process.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$ be the parameters for the response model and q be its dimension. Primary interest often centers on estimating parameters $\boldsymbol{\beta}$ in order to study the relationship between the response Y_i and covariates $(\mathbf{x}_i^\top, \mathbf{z}_i^\top)^\top$. Under the true response model (1) let

$$L_i(\boldsymbol{\theta}; y_i, \mathbf{x}_i, \mathbf{z}_i) = [g(y_i - \boldsymbol{\beta}_x^\top \mathbf{x}_i - \boldsymbol{\beta}_z^\top \mathbf{z}_i; \boldsymbol{\alpha})]^{\delta_i} [1 - G(y_i - \boldsymbol{\beta}_x^\top \mathbf{x}_i - \boldsymbol{\beta}_z^\top \mathbf{z}_i; \boldsymbol{\alpha})]^{1-\delta_i}$$

be the likelihood contributed from subject i, where $g(\cdot)$ is the density function corresponding to the distribution function $G(\cdot)$, and $y_i = \log(t_i)$. Denote the log likelihood as

$$\ell(\boldsymbol{\theta}; y, \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}; y_i, \mathbf{x}_i, \mathbf{z}_i)$$

where $\ell_i(\boldsymbol{\theta}; y_i, \mathbf{x}_i, \mathbf{z}_i) = \log L_i(\boldsymbol{\theta}; y_i, \mathbf{x}_i, \mathbf{z}_i)$. If there is no measurement error present in covariates, then the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is obtained by solving

$$\frac{\partial \ell(\boldsymbol{\theta}; y, \mathbf{x}, \mathbf{z})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{3}$$

and this estimator is consistent for $\boldsymbol{\theta}$ and has an asymptotic normal distribution. However, when error is present in covariates, the resulting estimator can be substantially biased (e.g., Li and Lin 2003; Yi and He 2006; He *et al.* 2007).

3. Simulation extrapolation method

3.1. SIMEX algorithm

To conduct valid inference for $\boldsymbol{\theta}$ in the presence of covariate measurement error, He *et al.* (2007) developed a SIMEX method. The basic idea of SIMEX adjustment is to add additional

variability to the observed measurement $\mathbf{W}_i$ in order to establish the trend how measurement error-induced bias may be related to the variance of induced measurement error, and then extrapolate this trend back to the case without measurement error. (Carroll *et al.* 2006, p. 97). This method is robust in a sense that the distribution of $\mathbf{x}_i$ is unspecified. Moreover, it is easy to implement. In this subsection we describe the SIMEX method developed in He *et al.* (2007). Typically, we consider two practical cases for the parameters in $\mathbf{\Sigma}_e$: (i) the parameters in $\mathbf{\Sigma}_e$ are given as fixed values; and (ii) the parameters in $\mathbf{\Sigma}_e$ are not known, but repeated measurements of $\mathbf{x}_i$ are available. The SIMEX method applies to both cases with the same steps except for the first step of data simulation. Details are given as follows.

Given an integer B and a sequence $\mathbf{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ taken from $[0, \lambda_M]$, where $\lambda_1 = 0$, we use the following algorithm to obtain the estimates and associated standard errors of the parameters $\boldsymbol{\theta}$.

Simulation step

In case (i) in which the parameters in $\mathbf{\Sigma}_e$ are known, we generate, for each $i = 1, \ldots, n$, a sequence of variables $\mathbf{u}_{ib} \sim MVN(\mathbf{0}, \mathbf{\Sigma}_e)$ for $b = 1, 2, \ldots, B$. For each $\lambda \in \mathbf{\Lambda}$, set

$$\mathbf{W}_i(b, \lambda) = \mathbf{W}_i + \sqrt{\lambda} \cdot \mathbf{u}_{ib}. \tag{4}$$

The array $\{\mathbf{W}_i(b, \lambda)\}$ of artificially simulated data will be used in the next step for estimation of the parameters.

In case (ii) in which the parameters in $\mathbf{\Sigma}_e$ are unknown but repeated measurements for $\mathbf{x}_i$ are available, we need to modify Equation 4 by making it be computable because $\mathbf{\Sigma}_e$ contains unknown parameters. To be specific, let $\mathbf{V}_{ij}, j = 1, \ldots, m_i$, denote the repeated measurements for true covariate $\mathbf{x}_i$, i.e., $\mathbf{V}_{ij}$ and $\mathbf{x}_i$ are linked by the model

$$\mathbf{V}_{ij} = \mathbf{x}_i + \mathbf{e}_{ij} \quad \mathbf{e}_{ij} \overset{iid}{\sim} MVN(0, \mathbf{\Sigma}_e), \qquad j = 1, \ldots, m_i,$$

where $\mathbf{\Sigma}_e$ is unknown, and $\mathbf{e}_{ij}$'s are independent of $\mathbf{x}_i, \mathbf{z}_i, \mathbf{Y}_i$. Instead of using Equation 4 to generate $\mathbf{W}_i(b, \lambda)$, we set, for given b and λ,

$$\mathbf{W}_i(b, \lambda) = \bar{\mathbf{V}}_i + \sqrt{\lambda / m_i} \sum_{j=1}^{m_i} c_{ij}(b) \mathbf{V}_{ij},$$

where $\bar{\mathbf{V}}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbf{V}_{ij}$, and $\mathbf{c}_i(b) = (c_{i1}(b), \ldots, c_{im_i}(b))'$ is any normalized contrast satisfying $\sum_{j=1}^{m_i} c_{ij}(b) = 0$ and $\sum_{j=1}^{m_i} c_{ij}^2(b) = 1$. A simple way to generate such a contrast $\mathbf{c}_i(b)$ is to use a normal variate generation. That is, for each b and $i = 1, \ldots, n$, independently generate m_i normal random variates $d_{ij}(b), j = 1, \ldots, m_i$, from a standard normal distribution $N(0, 1)$, then calculate $\bar{d}_i(b) = \frac{1}{m_i} \sum_{j=1}^{m_i} d_{ij}(b)$. Setting $c_{ij}(b) = \frac{d_{ij}(b) - \bar{d}_i(b)}{\sqrt{\sum_{l=1}^{m_i} (d_{il}(b) - \bar{d}_i(b))^2}}$ would result in the required contrasts $\mathbf{c}_i(b)$ (Devanarayan and Stefanski 2002).

Estimation step

For given λ and b, we obtain an estimate $\widehat{\boldsymbol{\theta}}(b, \lambda)$ by solving Equation 3 with $\mathbf{x}_i$ replaced by $\mathbf{W}_i(b, \lambda)$. For $r = 1, 2, \ldots, q$, let $\widehat{\theta}_r(b, \lambda)$ denote the rth component of $\widehat{\boldsymbol{\theta}}(b, \lambda)$, and $\widehat{\Omega}_r(b, \lambda)$ be the corresponding variance estimate, given by the rth diagonal element of

$$\left[-\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ell_i(\boldsymbol{\theta}; y_i, \mathbf{W}_i(b, \lambda), \mathbf{z}_i) \big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}(b, \lambda)} \right]^{-1}.$$

The following quantities are then calculated: $\widehat{\theta}_r(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_r(b,\lambda)$, $\widehat{\Omega}_r(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\Omega}_r(b,\lambda)$, $\widehat{S}_r(\lambda) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}_r(b,\lambda) - \widehat{\theta}_r(\lambda))^2$, and $\widehat{\tau}_r(\lambda) = \widehat{\Omega}_r(\lambda) - \widehat{S}_r(\lambda)$.

Extrapolation step

For $r = 1, 2, \ldots, q$, fit a regression model to each of the sequences $\{(\lambda, \widehat{\theta}_r(\lambda)) : \lambda \in \mathbf{\Lambda}\}$ and $\{(\lambda, \widehat{\tau}_r(\lambda)) : \lambda \in \mathbf{\Lambda}\}$, respectively, and extrapolate it to $\lambda = -1$. Let $\widehat{\theta}_r$ and $\widehat{\tau}_r$ denote the corresponding predicted values at $\lambda = -1$. Then $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_q)^\top$ is the SIMEX estimator of $\boldsymbol{\theta}$, and $\sqrt{\widehat{\tau}_r}$ is the associated standard error for the estimator $\widehat{\theta}_r$ $(r = 1, 2, \ldots, q)$.

Although the theory of the SIMEX method is not trivial, an example from simple linear regression can well illustrate the idea of this method. Suppose the response variable Y and the covariate x is modeled as

$$Y = \beta_0 + \beta_x x + \epsilon,$$

where ϵ has mean 0. If replacing x with its observed measurement W, modeled by $W = x + e$ where e has mean 0 and variance σ_e^2, and is independent of ϵ and x, then the resulting least squares estimator $\widehat{\beta}_x^*$ for β_x converges in probability to $\beta_x^* = (\sigma_x^2/(\sigma_x^2 + \sigma_e^2))\beta_x$, instead of β_x. Here σ_x^2 is the variance of x. To see how the bias may be related to the degree of measurement error in x, we perturb W by adding additional error to create $W(b,\lambda) = W + \sqrt{\lambda}u_b$ where u_b is independently generated from a $N(0, \sigma_e^2)$ distribution. Intuitively, if regressing Y over the perturbed version $W(b,\lambda)$, then the resulting estimator $\widehat{\beta}_x(b,\lambda)$ would converge in probability to $\beta_x^*(b,\lambda) = (\sigma_x^2/(\sigma_x^2 + (1+\lambda)\sigma_e^2))\beta_x$. This expression indicates the dependence of the asymptotic bias on the magnitude of measurement error - the less degree of measurement error (equivalently, a smaller λ), the smaller asymptotic bias. In particular, if λ shrinks to 0, $\widehat{\beta}_x(b,0)$ recovers the naive estimator $\widehat{\beta}_x^*$; but if λ takes value -1, then the limit $\beta_x^*(b,-1)$ is identical to the true parameter β_x.

The SIMEX method was initially proposed by Cook and Stefanski (1994) for analyzing complete univariate data with error-prone covariates under parametric models. He *et al.* (2007) generalized this method to handle survival data for which censoring is a typical feature. The SIMEX approach is very appealing because of its simplicity to implement and no requirement of modeling the true covariates $\mathbf{x}_i$ (often not observable). To implement this method, we need to address a few issues. The specification of B or Λ is not unique. Technically speaking, a larger value of B leads to a better SIMEX estimator in the sense that Monte Carlo sampling error in the simulation step can be reduced. For practical use, however, choosing $B = 50, 200$ or 500, and taking Λ to be the equal cut points of interval $[0,1]$ or $[0,2]$ with $M = 5, 10$ or 20, can often lead to fairly reasonable SIMEX estimates (e.g., Carroll *et al.* 2006). Another source of variation in obtaining SIMEX estimators lies in the choice of an extrapolation function. The exact extrapolation function is usually not known. Instead, a user-specified approximation is employed, hence SIMEX estimators are usually approximately consistent. Linear regression or quadratic regression function tends to be the most widely used replacement of the exact extrapolation function. Although SIMEX estimators are often not exactly consistent, they greatly outperform naive estimators for which measurement error is not properly taken into account. The performance of the SIMEX method has been shown superior in some highly nonlinear models (e.g., Carroll *et al.* 1996; Wang *et al.* 1998).

3.2. Implementation in R

The SIMEX procedures described above are implemented in the package **simexaft** (which depends on packages **survival**, Therneau and Lumley 2011, and **mvtnorm**, Genz *et al.* 2011 and Genz and Bretz 2009).

Specifically, the function `simexaft` produces the SIMEX estimates for interesting parameter $\boldsymbol{\beta}$ and other parameters along with their associated SIMEX standard errors and p values. The form of calling function `simexaft` is given by

```
simexaft(formula = formula(data), data = parent.frame(),
  SIMEXvariable = indicator, repeated = FALSE, repind = list(),
  err.mat = err.mat, B = 50, lambda = seq(0, 2, 0.1),
  extrapolation = "quadratic", dist = "weibull")
```

with the arguments being described as follows:

- `formula`: specifies the model to be fitted, with the variables coming with `data`. This argument has the same format as the `formula` argument in the function `survreg` from **survival**, taking the form `Surv(time, censoring indicator) ~ covariates`.
- `SIMEXvariable`: the index of the covariate variables that are subject to measurement error.
- `repeated`: set to `TRUE` or `FALSE` to indicate if there are repeated measurements for the mis-measured variables, i.e., corresponding to cases (i) and (ii) in Section 3.1, respectively.
- `repind`: the index of the repeated measurement variables for each mis-measured variable. It is of an R `list` form. If `repeated = TRUE`, `repind` must be specified.
- `err.mat`: specifies the covariance matrix in error model (2). If `repeated = FALSE`, `err.mat` must be specified.
- `B`: the number of simulated samples for the simulation step. The default is set to be 50.
- `lambda`: the set of $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ with $\lambda_1 = 0$ that is used as the grids for the extrapolation step.
- `extrapolation`: specifies the function form for the extrapolation step. The options are `"linear"` and `quadratic"`. The default is set to be `"quadratic"`.
- `dist`: specifies a parametric distribution that is assumed in AFT model (1). This argument is the same as the `dist` option in the function `survreg`, and it can take distributions such as `"weibull"`, `"exponential"`, `"gaussian"`, `"logistic"`, `"lognormal"`, and `"loglogistic"`.

4. Examples

To illustrate the usage of the developed R package **simexaft**, we apply the package to two real data sets, corresponding to cases with or without repeated measurements for error-contaminated covariates.

The first example is based on a subset of the real data arising from the Busselton Health Study (Knuiman *et al.* 1994). The original data were analyzed in He *et al.* (2007). The data set analyzed here includes survival information for a randomly selected subset of 100 females. The survival time is taken as the age at death, as in He *et al.* (2007). Systolic blood pressure (x_{i1}), cholesterol level (x_{i2}), age at registration (z_{i1}), body mass index (z_{i2}) and smoking status are risk factors related to mortality. Systolic blood pressure is rescaled as $\log(x_{i1} - 50)$, as in Carroll *et al.* (2006, p. 113). Smoking status is classified by two dummy indicators, denoted by z_{i3} and z_{i4}, where $z_{i3} = 1$ indicates an individual is an ex-smoker, and 0 otherwise; $z_{i4} = 1$ represents that an individual is a current smoker, and 0 otherwise. It is known that measurements of risk factors x_{i1} and x_{i2} are subject to substantial error due to the nature of these covariates.

The logarithms of the failure times are postulated by model

$$Y_i = \beta_0 + \beta_{x_1} x_{i1} + \beta_{x_2} x_{i2} + \beta_{z_1} z_{i1} + \beta_{z_2} z_{i2} + \beta_{z_3} z_{i3} + \beta_{z_4} z_{i4} + \epsilon_i/\alpha$$

where error ϵ_i follows a specific distribution. The standard extreme value distribution is assumed for an illustration. We assume that errors in both risk factors x_{i1} and x_{i2} can be represented by model (2).

The developed R package **simexaft** can be downloaded and installed from CRAN. The package can then be loaded in R:

```
R> library("simexaft")
```

Next, load the data that are properly organized with the variable names specified. In the example here, the data set named as `BHS` is included by issuing

```
R> data("BHS")
R> dataset <- BHS
R> dataset$SBP <- log(dataset$SBP - 50)
```

For illustrative purposes, we use settings with $B = 50, \lambda_M = 2$ and $M = 20$. In this example, we assume the parameters in $\mathbf{\Sigma}_e$ are known. This is a typical case when conducting sensitivity analysis. Here set $\sigma_{11}^2 = \sigma_{22}^2 = 0.75^2$ and $\sigma_{12} = \sigma_{21} = 0$ as an example.

The naive AFT approach without considering measurement errors in covariates gives the output:

```
R> formula <- Surv(SURVTIME,DTHCENS) ~ SBP + CHOL + AGE + BMI +
+    SMOKE1 + SMOKE2
R> out1 <- survreg(formula = formula, data = dataset, dist = "weibull")
R> summary(out1)

Call:
survreg(formula = formula, data = dataset, dist = "weibull")
              Value Std. Error      z        p
(Intercept) 12.5302     3.3587  3.731 0.000191
SBP         -1.2524     0.7766 -1.613 0.106807
CHOL        -0.0512     0.1096 -0.467 0.640360
```

```
AGE         -0.0603     0.0223 -2.712 0.006692
BMI          0.0337     0.0400  0.842 0.399920
SMOKE1      -0.7392     0.3993 -1.851 0.064158
SMOKE2      -0.8232     0.4178 -1.970 0.048805
Log(scale)  -0.5142     0.2079 -2.474 0.013375

Scale= 0.598

Weibull distribution
Loglik(model)= -83.5   Loglik(intercept only)= -98.5
        Chisq= 30.02 on 6 degrees of freedom, p= 3.9e-05
Number of Newton-Raphson Iterations: 9
n= 100
```

To adjust for possible effects of measurement error in variables `SBP` and `CHOL`, we call the developed function `simexaft` for the analysis:

```
R> set.seed(120)
R> ind <- c("SBP", "CHOL")
R> err.mat <- diag(rep(0.5625, 2))
R> out2 <- simexaft(formula = formula, data = dataset, SIMEXvariable = ind,
+    repeated = FALSE, repind = list(), err.mat = err.mat, B = 50,
+    lambda = seq(0, 2, 0.1), extrapolation = "quadratic", dist = "weibull")
R> summary(out2)

$coefficients
             Estimate Std. Error      P value
Intercept 16.33008771 3.91664272 3.053897e-05
SBP       -2.40116761 0.93348413 1.010358e-02
CHOL      -0.05630569 0.12982884 6.645124e-01
AGE       -0.04846142 0.02063056 1.882334e-02
BMI        0.05933523 0.04278722 1.655177e-01
SMOKE1    -0.60168913 0.36963556 1.035694e-01
SMOKE2    -0.79819843 0.39230144 4.188551e-02

$scalereg
(Intercept)
  0.5791607

$extrapolation
[1] "quad"

$SIMEXvariable
[1] "SBP"  "CHOL"

attr(,"class")
[1] "summary.simaxaft"
```

Now we demonstrate the use of `simexaft` for the case that the parameters in $\mathbf{\Sigma}_e$ is unknown, but repeated measurements for error-prone covariates are available. This is illustrated by an example from a study of pulmonary exacerbations and rhDNase. Fuchs *et al.* (1994) reported on a double-blind randomized multicenter clinical trial designed to assess the effect of rhDNase, a recombinant deoxyribonuclease I enzyme, versus placebo on the occurrence of respiratory exacerbations among patients with cystic fibrosis. The rhDNase operates by digesting the extracellular DNA released by leukocytes that accumulate in the lung as a result of bacterial infection, and so it was expected that aerosol administration of rhDNase would reduce the incidence of exacerbations (Cook and Lawless 2007, p. 365).

Six hundred and forty five patients were recruited in this trial. Each subject was randomly assigned to treatment or placebo group, and was followed up approximately 169 days for pulmonary exacerbations. Data on the occurrence and resolution of all exacerbations were recorded. The forced expiratory volume (FEV) was considered a risk factor and was measured twice at randomization. The response is defined as the logarithm of the time from randomization to the first pulmonary exacerbation.

To investigate the effect of the FEV on the time to first pulmonary exacerbation, we postulate the model

$$Y_i = \beta_0 + \beta_1 \cdot FEV + \beta_2 \cdot trt + \epsilon_i/\alpha$$

where *trt* is the indicator of treatment, and error ϵ_i follows a specific distribution. The standard extreme value distribution is taken again for illustrations. We assume that measurement errors in risk factors *FEV* can be represented by model (2).

First, load the data, named `rhDNase`, into R by issuing

```
R> data("rhDNase")
```

Two repeated measurements for covariate *FEV*, `fev` and `fev2`, are called in `simexaft` using the option `repeat = TRUE`, along with a list of index of the repeated measurements.

Existing `survreg` can provide the analysis with no measurement error effects properly taken into account, by merely taking the FEV measurement as the average of the two repeated observations:

```
R> rhDNase$fev.ave <- (rhDNase$fev + rhDNase$fev2)/2
R> output1 <- survreg(Surv(time2, status) ~ trt + fev.ave, data = rhDNase,
+    dist = "weibull")
R> summary(output1)

Call:
survreg(formula = Surv(time2, status) ~ trt + fev.ave, data = rhDNase,
    dist = "weibull")
              Value Std. Error     z         p
(Intercept)  4.5183    0.15470 29.21 1.61e-187
trt          0.3570    0.12179  2.93  3.38e-03
fev.ave      0.0193    0.00275  7.00  2.50e-12
Log(scale)  -0.0782    0.05959 -1.31  1.89e-01

Scale= 0.925
```

```
Weibull distribution
Loglik(model)= -1617.5   Loglik(intercept only)= -1652.9
        Chisq= 70.98 on 2 degrees of freedom, p= 3.3e-16
Number of Newton-Raphson Iterations: 5
n= 641
```

Similar analysis results can be obtained if using the `simexaft` function to accommodate covariate error effects. In this example, we note that variation in the two repeated measurements of FEV is too minor to suggest different results obtained from the methods of ignoring or accounting for covariate measurement error. Here we perturb the two repeated observations by adding additional noise, e.g., 15% of sample standard error, and then apply the developed R function to produce the output. This artificial procedure may not be customary when one focuses on a genuine data analysis. However, it is useful for illustration purposes. Moreover, this approach can provide some insights if conducting sensitivity analyses is of prime interest.

```
R> set.seed(120)
R> fev.error <- rhDNase$fev + rnorm(length(rhDNase$fev),
+    mean = 0, sd = 0.15 * sd(rhDNase$fev))
R> fev.error2 <- rhDNase$fev2 + rnorm(length(rhDNase$fev2),
+    mean = 0, sd = 0.15 * sd(rhDNase$fev2))
R> dataset2 <- cbind(rhDNase[, c("time2", "status", "trt")],
+    fev.error, fev.error2)
R> formula <- Surv(time2, status) ~ trt + fev.error
R> ind <- "fev.error"
```

Below is the output obtained from the naive approach that ignores covariate measurement error for perturbed data.

```
R> fev.error.c <- (fev.error + fev.error2)/2
R> output2 <- survreg(Surv(time2, status) ~ trt + fev.error.c,
+    data = rhDNase, dist = "weibull")
R> summary(output2)

Call:
survreg(formula = Surv(time2, status) ~ trt + fev.error.c, data = rhDNase,
    dist = "weibull")
              Value Std. Error     z         p
(Intercept)  4.5303    0.15413 29.39 6.66e-190
trt          0.3555    0.12191  2.92  3.54e-03
fev.error.c  0.0190    0.00273  6.98  3.05e-12
Log(scale)  -0.0772    0.05962 -1.30  1.95e-01

Scale= 0.926

Weibull distribution
Loglik(model)= -1617.9   Loglik(intercept only)= -1652.9
        Chisq= 70.02 on 2 degrees of freedom, p= 6.7e-16
```

```
Number of Newton-Raphson Iterations: 5
n= 641
```

Now we apply the developed function `simexaft` to adjust for the measurement error effects, with the perturbed data analyzed using the repeated measurements option.

```
R> formula <- Surv(time2, status) ~ trt + fev.error
R> output3 <- simexaft(formula = formula, data = dataset2,
+    SIMEXvariable = ind, repeated = TRUE,
+    repind = list(c("fev.error", "fev.error2")),
+    err.mat = NULL, B = 50, lambda = seq(0, 2, 0.1),
+    extrapolation = "quadratic", dist = "weibull")
R> summary(output3)

$coefficients
              Estimate  Std. Error      P value
Intercept   4.51642881 0.155619376 0.000000e+00
rhDNase$trt 0.36127209 0.121934403 3.048152e-03
fev.error   0.01924672 0.002755194 2.836176e-12

$scalereg
(Intercept)
  0.9252959

$extrapolation
[1] "quad"

$SIMEXvariable
[1] "fev.error"

attr(,"class")
[1] "summary.simaxaft"
```

The function `simexaft` can store individual estimated covariate coefficients in the simulation step, and this enables us to show the extrapolation curve through the developed R function `plotsimexaft`. The `plotsimexaft` function plots the extrapolation of the estimate of each covariate effect with the option of `"linear"`, `"quadratic"`, or `"both"` to view the performance of different extrapolants. Figure 1 displays the graph for the variable `SBP` in the first example for which both linear and quadratic extrapolants are applied from the following command

```
R> plotsimexaft(out2, "SBP", "both", ylimit = c(-3, 1))
```

5. Discussion

The impact of measurement error in covariates is well documented for survival data that are typically postulated by proportional hazards models, but there is relatively little discussion on AFT models. Yi and He (2006) explored the measurement error problem for bivariate

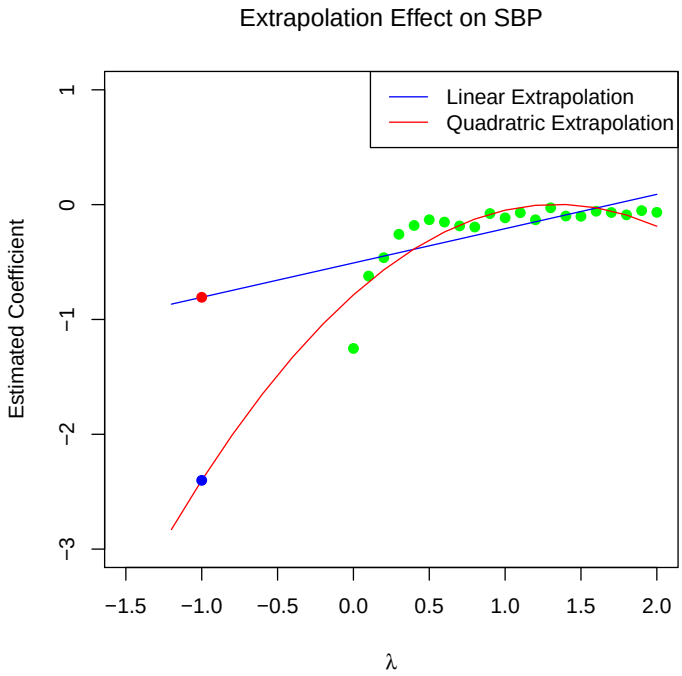

Figure 1: Display of the SIMEX estimate for the first example: Green dots represent the estimates $\hat{\beta}_{x1}(\lambda)$ for different values of λ; the red dot is the SIMEX estimate obtained from the linear extrapolation; and the blue dot is the SIMEX estimate obtained from the quadratic extrapolation.

survival data under AFT models, but their discussion focused on the AFT models with normal error distributions. To accommodate general distributional forms, He *et al.* (2007) describe a simulation based method that is simple to implement. For practical interest, we develop an R package **simexaft** to adjust for biases induced by covariate measurement error under AFT models. Our illustrations show that the developed package is simple to use. It is anticipated that such development is of great interest to data analysts when handling survival data with covariate measurement error. The R package **simexaft** is available from the Comprehensive R Archive Network at `http//CRAN.R-project.org/package=simexaft`.

Acknowledgments

The authors acknowledge the two anonymous reviewers for their comments. The research of He and Yi was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Carroll RJ, Küchenhoff H, Lombard F, Stefanski LA (1996). "Asymptotics for the SIMEX

Estimator in Nonlinear Measurement Error Models." *Journal of the American Statistical Association*, **91**, 242–250.

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006). *Measurement Error in Nonlinear Models*. 2nd edition. Chapman & Hall, London.

Cook J, Stefanski LA (1994). "A Simulation Extrapolation Method for Parametric Measurement Error Models." *Journal of the American Statistical Association*, **89**, 464–467.

Cook RJ, Lawless JF (2007). *The Statistical Analysis of Recurrent Events*. Springer-Verlag.

Devanarayan V, Stefanski LA (2002). "Empirical Simulation Extrapolation for Measurement Error Models with Replicate Measurements." *Statistics and Probability Letters*, **59**, 219–225.

Fuchs HJ, Borowitz DS, Christiansen DH, Morris EM, Nash ML, Ramsey BW, Rosenstein BJ, Smith AL, Wohl ME (1994). "Effect of Aerosolized Recombinant Human DNase on Exacerbations of Respiratory Symptoms and on Pulmonary Function in Patients with Cystic Fibrosis. The Pulmozyme Study Group." *New England Journal of Medidcine*, **331**, 637–642.

Genz A, Bretz F (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2011). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 0.9-9991, URL http://CRAN.R-project.org/package=mvtnorm.

Giménez P, Bolfarine H, Colosimo EA (1999). "Estimation in Weibull Regression Model with Measurement Error." *Communications in Statistics – Theory and Method*, **28**, 495–510.

Giménez P, Bolfarine H, Colosimo EA (2006). "Asymptotic Relative Efficiency of Score Tests in Weibull Models with Measurement Errors." *Statistical Papers*, **47**, 461–470.

He W, Yi GY, Xiong J (2007). "Accelerated Failure Time Models with Covariates Subject to Measurement Error." *Statistics in Medicine*, **26**, 4817–4832.

Knuiman MW, Cullent KJ, Bulsara MK, Welborn TA, Hobbs MST (1994). "Mortality Trends, 1965 to 1989, in Busselton, the Site of Repeated Health Surveys and Interventions." *Australian Journal of Public Health*, **18**, 129–135.

Lawless JF (2003). *Statistical Models and Methods for Lifetime Data*. 2nd edition. John Wiley & Sons, New York.

Li Y, Lin X (2003). "Functional Inference in Frailty Measurement Error Models for Clustered Survival Data Using the SIMEX Approach." *Journal of the American Statistical Association*, **98**, 191–203.

Nakamura T (1992). "Proportional Hazards Model with Covariates Subject to Measurement Error." *Biometrics*, **48**, 829–838.

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

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Reid N (1994). "A Conversation with Sir David Cox." *Statistical Science*, **9**, 439–455.

Therneau T, Lumley T (2011). ***survival**: Survival Analysis, Including Penalised Likelihood*. R package version 2.36-10, URL http://CRAN.R-project.org/package=survival.

Wang N, Lin X, Gutierrez RG, Carroll RJ (1998). "Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models." *Journal of the American Statistical Association*, **93**, 249–261.

Yi GY, He W (2006). "Methods for Bivariate Survival Data with Mis-measured Covariates under an Accelerated Failure Time Model." *Communications in Statistics – Theory and Methods*, **35**, 1539–1554.

Yi GY, Lawless JF (2007). "A Corrected Likelihood Method for the Proportional Hazards Model with Covariates Subject to Measurement Error." *Journal of Statistical Planning and Inference*, **137**, 1816–1828.

Zeng D, Lin D (2007). "Efficient Estimation for the Accelerated Failure Time Model." *Journal of the American Statistical Association*, **102**, 1387–1396.

Affiliation:

Wenqing He, Juan Xiong
Department of Statistical and Actuarial Sciences
University of Western Ontario
London N6A 5B7, Ontario, Canada
E-mail: whe@stats.uwo.ca

Grace Y. Yi
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo N2L 3G1, Ontario, Canada

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 46, Code Snippet 1 *Submitted:* 2008-11-25
January 2012 *Accepted:* 2011-09-01