Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2006
Methodological Issues in Measuring Population Health

Measurement Error in Life History Data

Grace Y. Yi[1] and Wenqing He [2]

Abstract

In practice it often happens that some collected data are subject to measurement error. Sometimes covariates (or risk factors) of interest may be difficult to observe precisely due to physical location or cost. Sometimes it is impossible to measure covariates accurately due to the nature of the covariates. In other situations, a covariate may represent an average of a certain quantity over time, and any practical way of measuring such a quantity necessarily features measurement error. When carrying out statistical inference in such settings, it is important to account for the effects of mismeasured covariates; otherwise, erroneous or even misleading results may be produced. In this paper, we discuss several measurement error examples arising in distinct contexts. Specific attention is focused on survival data with covariates subject to measurement error. We discuss a simulation-extrapolation method for adjusting for measurement error effects. A simulation study is reported.

KEY WORDS: Censoring; Measurement error; Simulation-extrapolation; Survival data.

1. Introduction

Covariate measurement error is often present with various reasons. Sometimes covariates of interest may be difficult to observe precisely due to physical location or cost. For example, the degree of narrowing of coronary arteries may reflect risk of heart failure, but physicians may measure the degree of narrowing in carotid arteries instead due to the less invasive nature of this method of assessment. Sometimes it is impossible to measure covariates accurately due to the nature of the covariates. For example, the level of exposure to potential risk factors for cancer such as radiation can not be measured accurately (Pierce et al. 1992). In other situations, a covariate may represent an average of a certain quantity over time, and any practical way of measuring such a quantity necessarily features measurement error. Covariate measurement error arises commonly in longitudinal studies, case-control studies, survey sampling and survival data analysis. Some specific measurement error examples are described as follows.

Example 1: Longitudinal data

The Framingham Heart Study is an ongoing longitudinal prospective study of risk factors for cardiovascular disease (CVD). The objective of the Framingham Heart Study was to identify common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke (Kannel et al. 1986). A major risk factor, systolic blood pressure, is subject to substantial measurement error. It is impossible to measure long-term systolic blood pressure x . Instead, a specific measurement W during a clinic visit is available. The long-term

[1] Grace Y. Yi, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1, E-mail: yyi@uwaterloo.ca
[2] Wenqing He, Department of Statistics and Actuarial Science, University of Western Ontario, 1151 Richmond Street North, London, Ontario, Canada N6A 5B7, E-mail: whe@stats.uwo.ca

measurement x and the single-visit measurement W are generally different due to daily and seasonal variation and confounding factors.

Example 2: Case-control data

The data discussed in Carroll, Gail, and Lubin (1993) were collected in a case–control study for which the primary objective was to examine the association between invasive cervical cancer and exposure to herpes simplex virus type 2 (HSV-2). Exposure to HSV-2 was assessed both by a refined western blot procedure (x) and by a less accurate western blot procedure (W) for cases ($Y = 1$) and controls ($Y = 0$). But the test result x was only directly observed for less than 6% of the subjects. Measurements based on the less accurate standard western blot test (W) were available for all subjects. In this scenario misclassification in covariate x arises if using all the available measurements W to study the relationship between Y and x .

Example 3: Survey data

Hwang (1986) discussed a survey data set which consists of 5,979 randomly selected households from the United States. Yearly energy consumed by a chosen family was reported. Household conditions, such as the number of windows, enclosed heated area, inches of wall insulation, roof insulation, floor insulation, etc., were collected. Other variables like family income, whether there were persons staying in the house during the day, whether there were certain major appliances, geographic region index, and local weather conditions were also recorded. It is of primary interest to understand the relationship between energy consumption of a household in the United States and its housing characteristics. But to preserve confidentiality, some of the predictors x that might identify household owners have been multiplied by random variables generated by a known distribution. That is, x is not available, instead, a surrogate $W = xU$ was recorded, where U is a random variable generated from a given distribution.

Example 4: Recurrent event data

Recurrent event data are often encountered in biomedical sciences, demographical studies, and industrial research. Examples include seizures of epileptic patients, successive tumors in cancer studies, multiple births in a woman's lifetime, and times to warranty claims for some type of manufactured item. Nutritional Prevention of Cancer (NPC) trial, for instance, was designed to study the long-term safety and efficacy of a daily 200-μg nutritional supplement of selenium (Se) for the prevention of cancer. One outcome of primary interest is squamous cell carcinoma (SCC) of the skin. As individuals may experience multiple recurrences of SCCs over time, it is of interest to study the effect of Se supplementation on SCC (Jiang, Turnbull and Clark 1999). Various factors including demographic, behavioral and medical baseline variables may affect the recurrence of SCC. It is known that some risk factors such as plasma Se status, measured in units of ng/ml, involve biological variability and measurement error.

Example 5: Survival data

The data arising from the Busselton Health Study were collected by a repeated cross-sectional survey in the community of Busselton in Western Australia from 1966 to 1988. One objective of the study was to evaluate the effect of cardiovascular risk factors on the risk of death due to coronary heart disease (Knuiman et al. 1994). The data set analyzed in Yi and Lawless (2006) includes survival information for 2306 spouse pairs. Of these, 2266 pairs have at least one censored response, i.e., at least one member of the couple was still alive at the final observation time. The risk factors for mortality, and especially for mortality due to coronary heart disease, include systolic blood pressure (SBP), cholesterol level (CHOL), age, smoking status and body mass index. It is known that measurements of the risk factors SBP and CHOL are subject to substantial measurement error due to the inherent nature of these covariates.

Measurement error is a common problem in practice. Its effects are complex, and they generally depend on the form of error model and the association of the response with covariates. To conduct valid inference, we need proper adjustments for individual problems. For a comprehensive review of this topic, see Carroll et al. (2006). Below we focus the discussion on survival data with error-prone covariates.

2. Survival Data with Error-Contaminated Covariates

Survival analysis plays an important role in many areas including the biomedical, engineering, and social sciences. For example, in medical studies for potentially fatal diseases we are interested in the survival time of patients with the disease. In those studies we are concerned about the relationship of survival or lifetime of individuals with a number of covariates that have influence on it. Some of those covariates, however, are subject to measurement error. For example, in the study of HIV infection, it is of primary interest to estimate the distribution of the time between HIV infection and AIDS diagnosis. However, an important factor CD4 count contains substantial measurement error, and valid statistical analysis should take this issue into account.

Since the earliest results in survival analysis concerning error-prone covariates discussed in Prentice (1982), there has been increasing research interest in this topic. Under the Cox proportional hazards models for univariate survival data, the so-called regression calibration approach has been proposed by Prentice (1982) to deal with mismeasured covariates, followed by other discussions including Nakamura (1992), Wang et al. (1997), Buzas (1998), Hu, Tsiatis and Davidian (1998), Zhou and Wang (2000), Huang and Wang (2000), Tsiatis and Davidian (2001), Xie, Wang and Prentice (2001), Augustin (2004), and Yi and Lawless (2006). For clustered survival data Li and Lin (2000, 2003) discussed inference methods to account for measurement error in covariates by using frailty models. Hu and Lin (2004) proposed semiparametric regression methods for multivariate failure times. Again the emphasis is on the Cox proportional hazards models that are employed to characterize the marginal distributions of failure times.

The impact of covariate measurement error is well understood for the Cox proportional hazard models. However, there is little discussion on the impact on accelerated failure time (AFT) models, even though these models have a great advantage of transparent interpretation (Lawless 2003). In this paper, we investigate the effect of measurement error on AFT models. We specifically explore a simulation and extrapolation (SIMEX) method to conduct valid inference. Classical additive error model is utilized to facilitate mismeasured covariates. Namely, assume $W = x + e$, where x is the true covariate that is often not precisely observed, W is an observed measurement of x, and e follows a normal distribution with mean 0 (e.g., Nakamura 1992; Li and Lin 2000, 2003).

3. Notation and Model Formulation

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 1 if $T_i < C_i$ and 0 otherwise, $i = 1, 2, ..., n$. Independent censoring is assumed. Let Y_i be the logarithm of the failure time for subject i, x_i be the covariates subject to possible measurement error, and z_i be the vector of covariates free of error, $i = 1, 2, ..., n$. Response variable Y_i is characterized by the AFT model, given by

$$Y_i = \beta_x' x_i + \beta_z' z_i + \varepsilon_i \tag{1}$$

where $\beta = (\beta_x', \beta_z')'$ is the vector of regression parameters of interest, and β_z may include the intercept coefficient. Denote the dimensions of β_x and β as p and q, respectively. Here ε_i's have a distribution $G(.)$ with parameters α, say.

Let W_i be an observed version of covariate x_i. x_i and W_i are assumed to follow a classical additive measurement error model, a model which is perhaps the most widely used in practice. That is, conditional on x_i and z_i,

$$W_i = x_i + e_i \quad (2)$$

where e_i follows, independent of x_i and z_i, a normal distribution with mean 0 and covariance matrix Σ_e=diag(σ_j^2, $j = 1, 2, ..., p$).

Parameters in Σ_e may be estimated from a validation data set or repeated measures of W_i. If neither validation data nor repeated measurements are available, then one can conduct sensitivity analyses based on background information about the measurement process to assess the impact of different degrees of measurement error on estimation of the response parameters. In this paper, error distribution parameters are assumed known, or estimated from an independent sample.

4. Estimation Procedures

In this subsection we describe a simulation-based functional estimation method by adapting the simulation-extrapolation (SIMEX) approach described in Cook and Stefanski (1994).The SIMEX method consists of two steps - a simulation step and a subsequent extrapolation step. The simulation step establishes the naive estimates for the cases when the variance of the error term for each measurement W_{ij}, $j = 1, 2, ..., p$, is $(1+\lambda)\sigma_j^2$. Here W_{ij} is the j th component of W_i, and λ > 0, which takes values in a specified set $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_M\}$, say. For each $j = 1, 2, ..., p$, generate a large number, B, of simulated data, $W_{ij}(b, \lambda)$, $b = 1, 2, ..., B$, by adding to each observed measurement W_{ij} a random variable with mean 0 and variance $\lambda\sigma_j^2$ according to the equation

$$W_{ij}(b, \lambda) = W_{ij} + \sqrt{\lambda}\sigma_j U_{ijb},$$

where the U_{ijb}'s are independent $N(0,1)$ observations. For given λ and b, we implement **R** function **survreg** to fit model (1) to the data set consisting of $\{Y_i, \delta_i, W_i(b, \lambda), z_i\}$. Let $\hat{\beta}(b, \lambda)$ denote the corresponding estimate, and $\hat{\Omega}_r(b, \lambda)$ be the variance estimate for the r th component $\hat{\beta}_r(b, \lambda)$ of $\hat{\beta}(b, \lambda)$, $r = 1, 2, ..., q$. Let

$$\hat{\beta}_r(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_r(b, \lambda), \quad \hat{\Omega}_r(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\hat{\Omega}_r(b, \lambda),$$

$$\hat{S}_r(\lambda) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_r(b, \lambda) - \hat{\beta}(\lambda))^2, \text{ and } \hat{\tau}_r(\lambda) = \hat{\Omega}_r(\lambda) - \hat{S}_r(\lambda).$$

In the extrapolation step, for each component $\hat{\beta}_r(\lambda)$, regress the estimate $\hat{\beta}_r(\lambda)$ on λ and extrapolate the resulting predicted mean to $\lambda = -1$ to obtain the estimate $\hat{\beta}_r$; similarly, for each r, regress the estimate $\hat{\tau}_r(\lambda)$ on λ and extrapolate the resulting predicted mean to $\lambda = -1$ to obtain the variance estimate $\hat{\tau}_r$.

Since the exact form of the extrapolant function is not known, selecting a suitable approximation is a common concern. In principle, any choice between competing extrapolant functions can be based on examining a plot of $\hat{\beta}_r(\lambda)$ (or $\hat{\tau}_r(\lambda)$) versus λ. Practically, common choices of regression functions in the extrapolation step include linear regression, quadratic regression and inverse regression (Cook and Stefanski 1994). We note that in general the SIMEX estimator $\hat{\beta}_r$ is only approximately consistent for β_r, because an approximate (rather than an exact) extrapolant function is used in the extrapolation step.

5. Simulation Study

We conduct a simulation study to investigate the impact of ignoring measurement error on estimation and to assess the performance of the SIMEX approach in contrast to the naive method. In the following simulation study, we set $n = 200$ and generate 1000 simulations for each of the parameter configurations. Take $B = 200$ and $M = 20$ for the SIMEX approach. We generate the failure times from the model

$$Y_i = \beta_0 + \beta_x x_i + \beta_z z_i + \varepsilon_i / \alpha$$

where ε_i follows a standard extreme value distribution with the density function $f(u) = e^u \exp(-e^u)$.

Set α to be 0.5. 30% and 50% censoring are considered, where a fixed censoring time C is generated for each subject. The true covariate x_i is simulated from the normal distribution $N(1,1)$. An observed value W_i is generated from the conditional normal distribution $N(x_i, \sigma^2)$, given x_i. Covariate z_i is generated from $Bin(1, 0.5)$ to represent a balanced design. Take $\beta_0 = 0$, $\beta_x = -\log 2$ and $\beta_z = 0.5$. Different configurations of σ are considered with $\sigma = 0.15, 0.25, 0.75$, featuring minor, moderate, and large measurement error, respectively.

In the table we report on the results of the bias of the estimates, the empirical standard error, the model based standard error, and the coverage probability for 95% confidence intervals. When $\sigma = 0.25$, the impact of measurement error is not striking. However, as the magnitude of error increases, the effect of measurement error is obviously visible. The naive approach fails to provide consistent estimates, which is evident from the estimates for β_x as well as for β_0. The corresponding coverage probabilities for 95% confidence intervals are far away from the nominal level 95%. It is not surprising that the estimates for β_z are not subject to much impact of measurement error, because β_z is the coefficient for precisely observed covariate z_i that is not correlated in the simulation considered here. The SIMEX approach demonstrates a much improved performance, with a lot smaller biases and considerably better coverage probabilities. When measurement error is minor or moderate, the estimates for all the parameters obtained from the SIMEX approach are satisfactory. The finite sample biases are reasonably small, and the coverage probabilities agree with the nominal value 95% reasonably well. If there is sever measurement error, then the performance of the SIMEX method seems less impressive. This could be due to the fact that the exact extrapolation function is not used, but only an approximation is used in the Extrapolation step for the SIMEX method. However, the SIMEX approach does perform a lot better than the naive method, with a certain amount of adjustment for measurement error. It can be seen that the SIMEX method inflates the standard errors, as opposed to the naive approach. The standard errors produced by both the naive and the SIMEX methods increase as measurement error becomes more substantial. It is interesting to see that increasing the proportion of censoring does not necessarily increase bias. However, the precision of the estimates drops.

6. Discussion

The impact of measurement error in covariates is well documented for survival data that are postulated by the Cox proportional hazards models, but there is little discussion on the AFT models, a useful tool for analyzing survival data. In this paper we focus the discussion on the AFT models and investigate the effects of measurement error on estimation of the response parameters. Yi and He (2006) explored the measurement error problem for bivariate survival data under the AFT models, but their discussion focused on the AFT model with normal error distributions. Here our discussion applies to AFT models with general distribution forms, not necessarily restricted to a particular distribution such as normal one. We describe a simulation based method to adjust for the bias induced by the error in covariates. This method is appealing in that it is easy to implement and it does not require the specification of the distribution of the error-prone true covariates that is generally unobservable. Our simulation demonstrates that the SIMEX approach does perform considerably better than the naive method. Its performance is reasonably satisfactory, especially for the case with minor or moderate measurement error.

In the same spirit of Li and Lin (2000) we may explore a likelihood based method to account for measurement error under the framework of AFT models. In such a development we need to assume a distribution form of error-prone covariates. It would be interesting to compare the performance of this approach to that of the SIMEX method.

Acknowledgement

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Augustin, T. (2004), "An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions", *Scandinavian Journal of Statistics,* 31*,* pp.43-50.

Buzas, J. F. (1998), "Unbiased scores in proportional hazards regression with covariate measurement error", *Journal of Statistical Planning and Inference*, 67, pp.247-257.

Carroll, R. L., Gail, M. H., and Lubin, J. H. (1993), "Case-control studies with errors in covariates", *Journal of the American Statistical Association*, 88, pp.185-199.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models*, 2nd ed., Chapman & Hall.

Cook, J. and Stefanski, L. A. (1994), "A simulation extrapolation method for parametric measurement error models", *Journal of the American Statistical Association*, 89, pp. 464-467.

Hu, C. and Lin, D. Y. (2004), "Semiparametric failure time regression with replicates of mismeasured covariates", *Journal of the American Statistical Society*, 99, pp.105-118.

Hu, P., Tsiatis, A. A., and Davidian, M. (1998), "Estimating the parameters in the Cox model when covariate variables are measured with error", *Biometrics*, 54, pp.1407-1419.

Huang, Y. and Wang, C. Y. (2000), "Cox regression with accurate covariates unascertainable: a nonparametric correction approach", *Journal of the American Statistical Association*, 95, pp.1209-1219.

Hwang, J. T. (1986), "Multiplicative errors-in-variables models with application to recent data released by the U.S. department of energy", *Journal of the American Statistical Association,* 81, pp.680-688.

Jiang, W., Turnbull, B. W., and Clark, L. C. (1999), "Semiparametric regression models for repeated events with random effects and measurement error", *Journal of the American Statistical Association,* 94, pp.111-124.

Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler, J., Hulley, S. B., and Kjelsberg, M. O. (1986), "Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for MRFIT", *American Heart Journal*, 112, pp.825-836.

Knuiman, M. W., Cullent, K. J., Bulsara, M. K., Welborn, T. A., and Hobbs, M. S. T. (1994), "Mortality trends, 1965 to 1989, in Busselton, the site of repeated health surveys and interventions", Australian Journal of Public Health, 18, pp.129-135.

Lawless, J. F. (2003), *Statistical Models and Methods for Lifetime Data*, 2nd ed., Johns Wiley & Sons, New York.

Li, Y. and Lin, X. (2000), "Covariate measurement errors in frailty models for clustered survival data", *Biometrika*, 87, pp.849-866.

Li, Y. and 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, pp.191-203.

Nakamura, T. (1992), "Proportional hazards model with covariates subject to measurement error", *Biometric*s, 48, pp.829-838.

Pierce, D. A, Stram, D. O., Vaeth , M., and Schafer, D. (1992), "Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors", *Journal of the American Statistical Association,* 87, pp.351-359.

Prentice, R. L. (1982), "Covariate measurement errors and parameter estimation in a failure time regression model", *Biometrika*, 69, pp.331-342.

Tsiatis, A. A. and Davidian, M. (2001), "A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error", *Biometrika*, 88, pp.447-458.

Wang, C. Y., Hsu, L., Feng, Z. D., and Prentice, R. L. (1997), "Regression calibration in failure time regression", *Biometrics*, 53, pp.131-145.

Xie, S. H., Wang, C. Y., and Prentice, R. L. (2001), "A risk set calibration method for failure time regression by using a covariate reliability sample", *Journal of the Royal Statistical Society B*, 63, pp.855-870.

Yi, G. Y. and He, W. (2006), "Methods for bivariate survival data with mismeasured covariates under an accelerated failure time model", *Communications in Statistics - Theory and Methods*, 35, pp.1539-1554.

Yi, G. Y. and Lawless, J. F. (2006), "A corrected likelihood method for the proportional hazards model with covariates subject to measurement error", To appear in *Journal of Statistical Planning and Inference*.

Zhou, H. and Wang C. Y. (2000), "Failure time regression with continuous covariates measured with error", *Journal of the Royal Statistical Society B*, 62, pp.657-665.

Table: Simulation Results

	σ	Method	β_0				β_x				β_z			
			Bias	SE(M)	SE(E)	CP	Bias	SE(M)	SE(E)	CP	Bias	SE(M)	SE(E)	CP
I	0.25	NAIVE	0.067	0.320	0.309	0.961	-0.071	0.216	0.219	0.916	0.018	0.396	0.391	0.964
	0.50		0.312	0.325	0.319	0.862	-0.276	0.199	0.211	0.647	-0.021	0.405	0.401	0.955
	0.75		0.511	0.326	0.322	0.677	-0.506	0.176	0.181	0.195	0.014	0.408	0.402	0.956
	0.25	SIMEX	-0.018	0.327	0.316	0.967	0.014	0.232	0.236	0.948	0.015	0.398	0.392	0.961
	0.50		0.075	0.341	0.345	0.945	-0.032	0.241	0.265	0.913	-0.032	0.410	0.410	0.955
	0.75		0.190	0.342	0.358	0.912	-0.175	0.231	0.259	0.812	0.001	0.413	0.415	0.944
II	0.25	NAIVE	0.079	0.397	0.411	0.947	-0.074	0.259	0.258	0.924	-0.007	0.473	0.486	0.950
	0.50		0.256	0.403	0.397	0.945	-0.273	0.237	0.235	0.738	0.015	0.479	0.480	0.943
	0.75		0.460	0.410	0.419	0.853	-0.516	0.208	0.213	0.319	0.026	0.481	0.489	0.944
	0.25	SIMEX	0.001	0.401	0.417	0.939	0.010	0.277	0.276	0.952	-0.010	0.475	0.488	0.950
	0.50		0.033	0.412	0.410	0.945	-0.029	0.287	0.294	0.937	0.006	0.483	0.487	0.945
	0.75		0.162	0.416	0.441	0.942	-0.193	0.272	0.298	0.818	0.019	0.486	0.500	0.947

Case I: 30% censoring; Case II: 50% censoring;
SE(M): model based standard error; SE(E): empirical standard error; CP: 95% coverage probability.