Observational Studies 1 (2015) 91-123 Submitted 9/14; Published 7/15

The non-zero mean SIMEX: Improving estimation in the face of measurement error

Nabila Parveen
Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

Erica E. M. Moodie **erica.moodie@mcgill.ca**
Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

Bluma Brenner
Lady Davis Research Institute, Montreal, Quebec, Canada

Abstract

The simulation extrapolation method developed by Cook and Stefanski (1995) is a simulation based technique for estimating and reducing bias due to additive measurement error armed only with knowledge of the variance of the measurement error distribution. However there are many instances in which validation data are not available, and measurement error is known not to have mean zero. For example, in assessing phylogenetic cluster size of HIV viruses, cluster size is systematically underestimated since clustering can only be performed on the viruses of those individuals who have presented for testing. In this setting, it is not possible to obtain validation data; however, using knowledge gleaned from the literature, the distribution of the errors may be estimated. In this work, we extend the simulation extrapolation procedure to accommodate errors with non-zero means, motivated by an interest in determining behavioural correlates of HIV phylogenetic cluster size. We provide theoretical justification for the generalization to the non-zero mean measurement error case, proving its consistency and demonstrating its performance via simulation. We then apply the result to data from a the province of Quebec in Canada to show that findings from a naïve analysis are robust to a substantial range of possible measurement error distributions.

Keywords: SIMEX; non-zero mean measurement error; HIV.

1. Introduction

Since the discovery of the human immunodeficiency virus (HIV) in 1981, HIV has caused nearly 36 million deaths (as of 2012) (amFAR, 2012). While there is no cure or vaccine for HIV, current therapies are highly effective and have dramatically reduced mortality due to HIV. Nevertheless, HIV places an immense burden on individuals and societies, with the annual costs (medical and lost productivity) of new HIV infections in the United States esti-

mated at $16 billion in 2010 (Annual Cost, 2010). There is considerable research activity on HIV in Montreal, Canada. One such study is SPOT (SPOT, 2013), which offers rapid, free and anonymous testing to the community of men who have sex with men (MSM), primarily targeting men who frequent gay social venues. Individuals who are tested at SPOT provide questionnaire data, and for all individuals found to be HIV+, their blood undergoes HIV sequencing. The HIV sequencing information is supplemented with HIV sequencing information from the Quebec genotyping program (Genotype Information, 2010) to determine the size of the sexual network to which the individual belongs, i.e. the number of other HIV+ individuals in the province of Quebec whose HIV sequence fall into the cluster in a phylogenetic analysis. Researchers wish to combine the phylogenetic and epidemiological data to learn about correlates of large phylogenetic clusters (Brenner et al., 2007, 2013; Brenner and Wainberg, 2013). Transmission cluster size (or simply cluster size) is defined as the number of individuals falling into the same HIV phylogenetic grouping. For example, if the HIV sequence of six individuals fall into the same cluster, each will be said to belong to a cluster of size six; if there is an individual whose HIV genome sequence does not cluster with the HIV genome of anyone else in the Quebec genotyping program registry of sequences, this individual is said to belong to a cluster of size one. However, the data available do not include individuals who are HIV+ but are unaware of their status (i.e. have never been tested) nor those who have not had their HIV genotyped (viral load less than 400 copies per ml) (Brenner et al., 2007); there may also be a small number who have been tested outside of the province of Quebec and not yet been seen by a physician in the province. Consequently, measurement error occurs in defining the cluster size. This measurement error is characterized by a systematic *undercounting* of the true cluster size due to the absence of the individuals who have not been tested. Thus, to make correct statistical inference about correlates of sexual network size, this measurement error must be taken into consideration.

There are several approaches to handle measurement error: e.g., method of moments, regression calibration (Carroll and Stefanski, 1990; Gleser, 1990), multiple imputation (Rubin, 1987; Cole et al., 2006), and simulation extrapolation (SIMEX) (Cook and Stefanski, 1995); most require validation data, which is infeasible to collect in the case of phylogenetic or transmission cluster size. Unlike regression calibration and multiple imputation, SIMEX does not require validation data. The approach does, however, require that the measurement error distribution is known or can be well-estimated. In some instances, such as when data arise from a well-understood laboratory assay, the error distribution may be known exactly. In other instances, the distribution may be estimated from validation data if available, or posited based on information available in the literature, or simply assumed (and varied) as in a sensitivity analysis. In the few existing applications of SIMEX in the epidemiological literature, the error distribution has been determined or estimated using a combination of expert judgment and data from the literature (Li and Lin, 2011; Kim and Gleser, 2000; Slate and Bandyopadhyay, 2009; Heid et al., 2009; Costas et al., 2009; Shang, 2012; Allodji et al., 2012; He et al., 2012).

The simulation extrapolation method developed by Cook and Stefanski (Cook and Stefanski, 1995; Stefanski and Cook, 1995; Carroll et al., 1996) is a simulation based technique

for estimating and reducing bias due to additive measurement error. The SIMEX procedure does not require validation data, but does require the distribution of the measurement error to be posited, which may be possible using known properties of a measurement instrument such as a laboratory assay, or from existing literature. SIMEX is a two-step estimation procedure in which additional measurement error is added (in known increments) to the mis-measured data in a resampling-like stage, and a trend between the resulting estimates and the variance of the added measurement errors is established. To date, SIMEX has been limited to mean zero random errors, and will therefore need to be extended to alternative error distributions to be used in the context of under-counted measures. We shall extend the method to accommodate errors with non-zero means, so as to apply it to the SPOT data to determine behavioural correlates of cluster size. In section 2, we develop the theory, then demonstrate its performance in simulations in 3. Next, we apply the method to SPOT. Section 5 discusses the findings.

2. The Simulation Extrapolation (SIMEX) Method

In SIMEX, estimation proceeds in two step: a simulation step and an extrapolation step. Estimates are obtained by *increasing* the measurement error in the mis-measured data in a resampling-like stage, computing estimates from the contaminated data, establishing a trend between these estimates and the variance of the added measurement errors, and extrapolating this trend back to the case of no measurement error. The main idea is to use the information from an incremental addition of measurement error to the mis-measured data using computer-simulated random errors. Adding extra measurement error to the data by simulation allows the researcher to learn about how the estimator's bias is affected by the increase of the measurement error variance. This is the so-called simulation step. In the extrapolation step, the obtained parameter estimates are modelled as a function of the magnitude of the variance of the measurement error and extrapolated to the case of no measurement error. We begin by briefly describing the simulation-extrapolation procedure for zero mean measurement error and then present in detail the extension to non-zero mean measurement error, which we call the non-zero mean SIMEX (NZM-SIMEX), then proceed to derive its large sample properties.

A short description of SIMEX: Suppose U_i, $i = 1, ..., n$, is the unobserved true explanatory variable and an error-prone version X_i is available, where $X_i = U_i + \delta_i$, for $\delta_i \sim N(0, \sigma_\delta^2)$ and it is independent of U_i and Y_i. In the simulation step of SIMEX procedure, artificial measurement error is added to X_i, and B new covariates $X_{i,b}(\lambda_k)$ are generated via $X_{i,b}(\lambda_k) = X_i + \sqrt{\lambda_k}\delta_{ib}$, where $b = 1, ..., B$; $k = 1, ..., K$ and $i = 1, ..., n$ for values of λ_k are chosen by the analyst and $\{\delta_{i,b}\}_{b=1}^{B}$ are independent computer simulated normal random numbers from $N(0, \sigma_\delta^2)$. It can be shown that the variance of $X_{i,b}(\lambda_k)$ is $\sigma_U^2 + (1 + \lambda_k)\sigma_\delta^2$, which increases with λ_k. For each λ_k, let $\hat{\beta}_b(\lambda_k)$ denote the vector of naïve estimates obtained by regressing Y on $X_{i,b}(\lambda_k)$. Using B estimates for each λ_k, an average estimate can be obtained as $B^{-1}\sum_{b=1}^{B}\hat{\beta}_b(\lambda_k)$. By regressing $\hat{\beta}_b(\lambda_k)$ on λ_k, and extrapolating back to $\lambda_k = -1$, we find the estimate $\hat{\beta}(-1)$ corresponding to error $\sigma_U^2 + (1 + \lambda_k)\sigma_\delta^2 = \sigma_U^2$, i.e., to the error free setting. A prototypical example (based on simulated data) on the estimates

$\hat{\beta}(\lambda_k)$ and the extrapolating function that describes the regression of $\hat{\beta}(\lambda_k)$ on λ_k is given in Figure 1 for illustration.

Simulation step

Let us consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i,$$

where the true predictor U_i follows a distribution with finite variance σ_U^2 and $\mathbb{E}[\epsilon_i] = 0$. Suppose X_i is an imperfect measurement of U_i which is defined as

$$X_i = U_i - \delta_i^*,$$

where δ_i^* follows a distribution with $\mathbb{E}[\delta_i^*] = \mu_{\delta^*}$ and $Var[\delta_i^*] = \sigma_{\delta^*}^2$. Also, δ_i^* is independent of Y_i and U_i. For example, in the SPOT data, where U_i is the true value of the count variable 'cluster size', it may be reasonable to assume $\delta_i^* \sim Poisson(\mu)$, so that $\mathbb{E}[\delta_i^*] = Var[\delta_i^*] = \mu$. In other instances we may wish to consider $\delta_i^* = |\delta_i|$, where $\delta_i \sim N(0, \sigma_\delta^2)$, so that δ_i^* follows a folded Normal distribution with $\mathbb{E}[\delta_i^*] = \sigma_\delta\sqrt{\frac{2}{\pi}}$ and $Var[\delta_i^*] = \sigma_\delta^2(1 - \frac{2}{\pi})$. In the simulation step, additional, simulated measurement error is added to the imperfectly measured covariate X_i, and B new covariates $X_{i,b}(\lambda_k)$ are generated using the rule:

$$\begin{aligned} X_{i,b}(\lambda_k) &= X_i - \sqrt{\lambda_k}\delta_{ib}^* + (1 + \sqrt{\lambda_k})\mathbb{E}(\delta_{ib}^*) \\ &= X_i - \sqrt{\lambda_k}\delta_{ib}^* + (1 + \sqrt{\lambda_k})\mu_{\delta^*}, \end{aligned} \tag{1}$$

where $b = 1, \ldots, B$; $k = 1, \ldots, K$ and $i = 1, \ldots, n$. The parameters $\lambda_k \geq 0$ control the variance of the measurement error, and are chosen by the analyst, while $\left\{\delta_{i,b}^*\right\}_{b=1}^{B}$ are artificially introduced random numbers from the distribution of δ_i^*. Note that this is *not* identical to the simulation step in the traditional (mean zero error) SIMEX, but rather an additional term, $(1 + \sqrt{\lambda_k})\mu_{\delta^*}$, has been included in the generation of $X_{i,b}(\lambda_k)$ to account for the non-zero mean of the errors. Carroll et al. (Carroll et al., 1995) recommended taking λ_k as $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_K = 2$. Note that using (1) ensures that

$$\mathbb{E}[X_{i,b}(\lambda_k)] = \mathbb{E}(U_i).$$

The simulation step creates B additional datasets (replication samples to reduce simulation variability) with the same dependent variable Y_i and covariate $X_{i,b}(\lambda_k)$ for each λ_k. The variance of $X_{i,b}(\lambda_k)$ is

$$\begin{aligned} \mathbb{V}\left[X_{i,b}(\lambda_k)\right] &= \mathbb{V}\left[X_i - \sqrt{\lambda_k}\delta_{ib}^* + (1 + \sqrt{\lambda_k})\mu_{\delta^*}\right] \\ &= \mathbb{V}\left[U_i - \delta_i^* - \sqrt{\lambda_k}\delta_{ib}^* + (1 + \sqrt{\lambda_k})\mu_{\delta^*}\right] \\ &= \sigma_U^2 + (1 + \lambda_k)\sigma_{\delta^*}^2 \end{aligned}$$

which increases with the control parameter λ_k. For each λ_k, let $\hat{\beta}_b(\lambda_k)$ denote the vector of naïve estimates obtained by regressing Y on $X_{i,b}(\lambda_k)$. Using B estimates for each λ_k, an average estimate can be obtained as

$$\hat{\beta}^{NZM}(\lambda_k) = \frac{1}{B}\sum \hat{\beta}_b^{NZM}(\lambda_k). \tag{2}$$

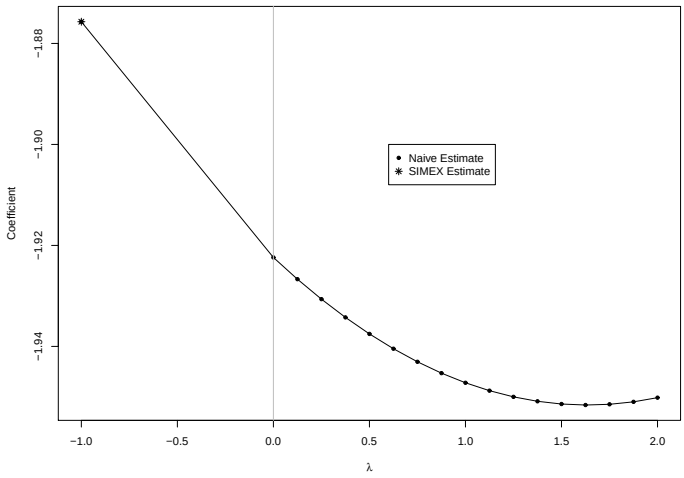


Figure 1: A generic plot of the effect of measurement error of size $(1 + \lambda_k)\sigma_\delta^{*2}$ on parameter estimates when measurement error follows a folded normal distribution. The SIMEX estimate is an extrapolation to $\lambda_k = -1$ whereas the naïve estimate occurs at $\lambda_k = 0$.

Extrapolation step

In the extrapolation step, each component of the vector $\hat{\beta}(\lambda_k)$ is plotted against λ_k for $\lambda_k \geq 0$, and regression techniques are used to fit an extrapolant function. In particular, $\hat{\beta}^{NZM}(\lambda_k)$ is typically regressed on λ_k assuming either a quadratic or a non-linear relationship (e.g., a lowess smoother). The NZM-SIMEX estimator, denoted $\hat{\beta}^{NZM}$, is obtained as the extrapolation of $\hat{\beta}(\lambda_k)$ at $\lambda_k = -1$, which is the ideal case in which there is no measurement error. See Figure 1 for a prototypical figure showing a plot of $\hat{\beta}(\lambda_k)$ against λ_k and the resulting NZM-SIMEX estimate.

Below, we state two key properties of the NZM-SIMEX estimator, $\hat{\beta}^{NZM}$; proofs in the linear regression setting are provided in the Appendix A. As in the zero-mean error distribution setting (Cook and Stefanski, 1995), results hold for more general regression problems, including the fitting of generalized linear models (Li and Lin, 2011), non-linear regression models (Carroll et al., 1996), quantile regression models (Shang, 2012), accelerated failure time models (Wenqing et al., 2012), and even generalized linear mixed models (Wang et al., 2009), but cannot be shown in closed form; results demonstrating the feasibility of the SIMEX in these setting has relied on simulations. As in the previous literature, we provide theorems for the linear regression setting, and demonstrate the performance of the method in the generalized linear regression setting by simulation but not analytically. Both theorems rely on the assumption that the variance of the measurement error is known and finite. The proofs rely extrapolating to the no-error setting; while we can show this explicitly (i.e. in a closed form solution) in a linear regression setting, the extrapolation does not rely on the distribution of Y.

Theorem 1:

The SIMEX estimator for non-zero mean measurement error, $\hat{\beta}^{NZM}$, converges in probability to β.

Theorem 2:

$\hat{\beta}^{NZM}$ is a nonlinear function of λ_k.

3. Simulation study

A simulation study was carried out to empirically evaluate the performance of the NZM-SIMEX procedure under ideal and non-ideal conditions for a variety of outcome and covariate distributions at different sample sizes. In particular, we consider both the case where the error distribution is known exactly, and cases where it is not (e.g. it is known that the error follows a Poisson distribution, but an incorrect mean is assumed). A large range of settings were considered, including but not limited to the Poisson-distributed error setting which will be used in the empirical analysis of Section 4, to showcase the versatility of the methodology across a variety of possible scenarios.

3.1 Design of the simulation study

As the derivation of the NZM-SIMEX is general, we aimed to assess its performance under a variety of conditions specified by the outcome and error distributions. Parameters were chosen to follow those used by Cook and Stefanski (1995). In all instances, we report the bias, standard error (SE) and mean squared error (MSE) of the naïve and NZM-SIMEX estimators based on 1000 simulations. The sample sizes considered were $n = 100, 500$ and 1000. In our motivating data that has been analysed in section 4.2, we have sample size $n = 33$. Therefore, we considered some simulation situations for $n = 33$.

Three outcome distributions were considered: normal, Poisson and Bernoulli distributions. For normally distributed outcomes, data were generated from the model

$$E(Y|U,V) \quad = \quad \beta_0 + \beta_U U + \beta_V V + \beta_{UV} UV.$$

For the Poisson distributed outcomes, data were generated from a log linear regression model

$$log[E(Y|U,V)] \quad = \quad \beta_0 + \beta_U U + \beta_V V + \beta_{UV} UV.$$

For the binary response, data were generated from a logistic regression model

$$logit[P(Y=1|U,V)] \quad = \quad \beta_0 + \beta_U U + \beta_V V + \beta_{UV} UV.$$

Details of the simulation settings for (U,V), $\boldsymbol{\beta} = (\beta_0, \beta_U, \beta_V, \beta_{UV})'$, δ and δ^* are given in Table B1 of Appendix B, with Scenarios 1-10 covering Normally-distributed outcomes; Scenarios 11-12 the Poisson-distributed outcomes, and Scenario 13 the binary outcome.

For NZM-SIMEX procedure, we considered $\lambda_k \in \{0, \frac{1}{8}, \frac{2}{8}, \ldots, \frac{15}{8}, \frac{16}{8}\}$, $b = 200$, and

$$X_b = X - \sqrt{\lambda_k}\delta_b^* + (1 + \sqrt{\lambda_k})E(\delta_b^*),$$

where for the normally distributed outcome only, $\delta_b^* = |\delta_b|$.

3.2 Results of the simulation study

The simulation results are shown in Figure 2, Figure 3 and Tables B2-B5 in Appendix B. It is evident from these results that the NZM-SIMEX procedure leads to a considerable reduction of the bias compared to the naïve estimator.

When the error distribution is correctly specified by the analyst in the NZM-SIMEX method, the bias of the NZM-SIMEX estimator is much less than the naïve estimator. Biases depend on the magnitude of measurement error, whatever the distribution of the measurement error (Tables B2, B4 and B5). However, we also see that the bias reduction in the NZM-SIMEX estimators is less pronounced with increasing degrees of measurement error.

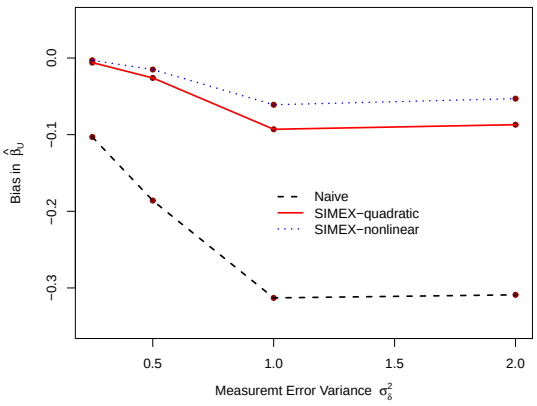
Irrespective of the parametric distribution of the errors (folded normal or Poisson), when parameters of the measurement error distribution are incorrectly specified, it is observed from Table B3 that the NZM-SIMEX estimator performs sub-optimally. However, while the NZM-SIMEX estimator using an incorrect measurement error distribution to generate the simulated errors performs worse than the NZM-SIMEX using the correct measurement error distribution, performance remains superior to that of the naïve estimator. Under-estimation the variability of the measurement error leads to greater bias in the NZM-SIMEX than over-estimation. It is also apparent from the results that, with the very few exceptions, the non-linear fit in NZM-SIMEX procedure yields less biased estimates than quadratic fit.

For discrete Poisson and binary distributed outcomes, it is observed from Table B4 and B5 that for the correctly specified error distribution, the NZM-SIMEX yields a less biased estimator than the naïve approach. In all cases, performance of NZM-SIMEX improves as the sample size increases. Thus, when the distribution of the errors is known, NZM-SIMEX performs well in recovering the true value of the parameter of interest. When the error distribution is mis-specified, the NZM-SIMEX procedure exhibits some bias, but nevertheless significantly outperforms the naïve estimator.

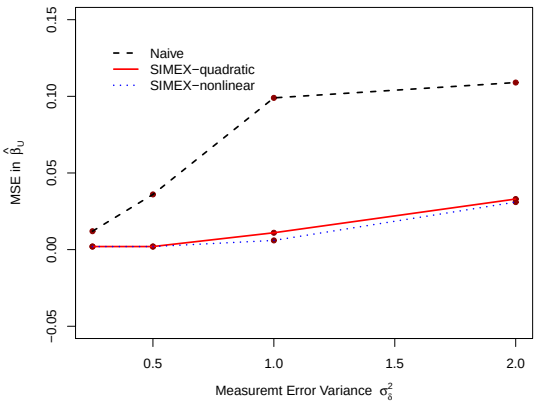
4. Analysis of the SPOT data: Correlating behaviour and cluster size

We now turn back to the motivating question in the analysis of the SPOT data. As noted above, neither SPOT nor the Quebec HIV genotyping program includes HIV+ people who are unaware of their HIV status. We may also fail to capture individuals who underwent testing outside the province of Quebec. This induces measurement error in defining cluster size. In particular, it causes an underestimation of the true cluster size so that, clearly, measurement error in cluster size is not mean zero.

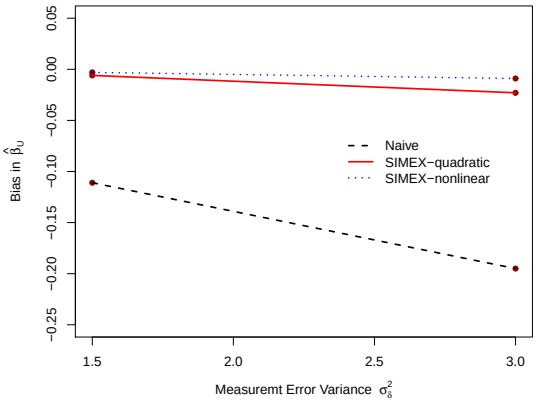
We used data from the SPOT study up until April 2012. At that time, SPOT had tested 1803 MSM, 34 of whom were found to be HIV positive. For all participants, questionnaire data includes several measurements on socio-demographic characteristics, HIV testing behaviour, sexual practices including risk behaviour, history of sexually transmitted infections, and attitudes toward HIV. In this analysis, we focus on the HIV+ individuals and consider whether any of the following variables are correlates of cluster size: age, whether or not a condom was used at last sexual intercourse, number of sex partners, and whether or not an HIV test was taken in the last 24 months. Except for cluster size, one individual's questionnaire data was incomplete; we omit this individual from the analysis, instead analyzing the 33 men with complete data.



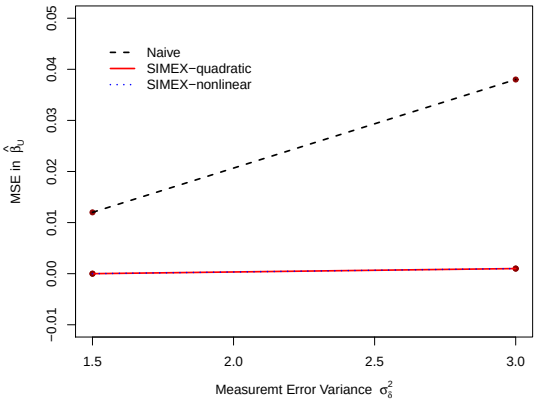
(a) Measurement Error: Folded Normal



(b) Measurement Error: Folded Normal

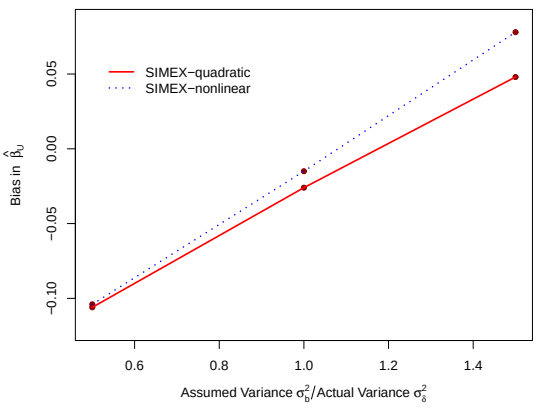


(c) Measurement Error: Poisson

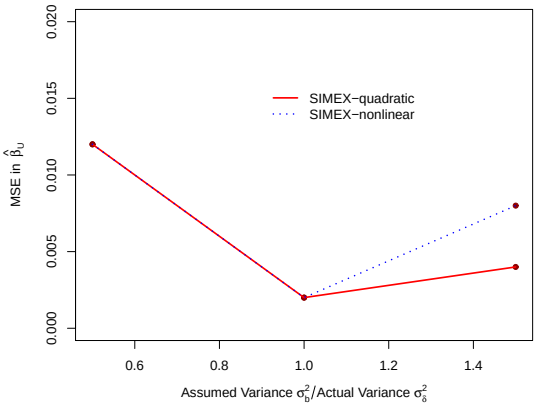


(d) Measurement Error: Poisson

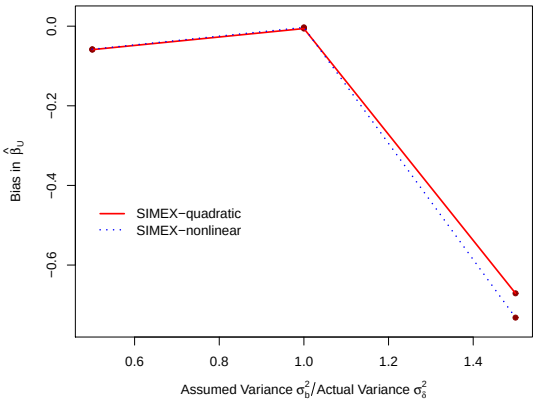
Figure 2: Bias and MSE of the parameter estimator associated with the error prone variable for two different measurement error distributions.



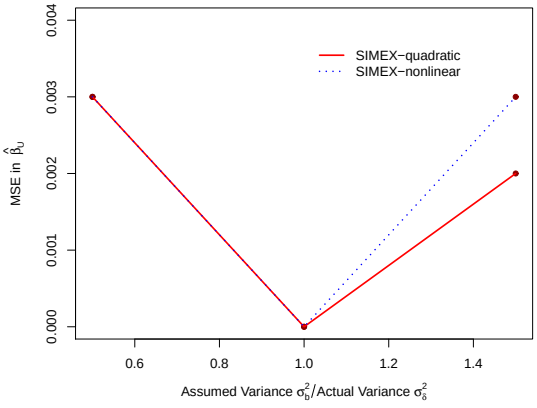
(a) Measurement Error: Folded Normal



(b) Measurement Error: Folded Normal



(c) Measurement Error: Poisson



(d) Measurement Error: Poisson

Figure 3: Bias and MSE of the parameter estimate associated with the error prone variable for two different measurement error distributions.

With the goal of identifying the relationship between cluster size and age, number of sex partners, not using a condom at last sexual intercourse, HIV testing status during last 24 months and number of one night partners we adopted seven distinct regression models. For each variable, both NZM-SIMEX (using quadratic and non-linear extrapolation) and a naïve model were used to obtain estimates. We fit two linear regression models of age on cluster size and number of sex partners on cluster size. We fit two logistic regression models, where in the first model not using a condom at the last sexual intercourse was considered as response variable and in the second model HIV testing status (during last 24 months) was taken as the outcome. Also considering number of sex partners and number of one night partners as count variables, we fit two log-linear models: number of sex partners on cluster size, and number of one night partners on cluster size. Furthermore, considering number of one night partners as a categorical variable (Category 1: < 2 partners, Category 2: $2-4$ partners, and Category 3: ≥ 5 partners), we fit a multinomial regression model considering Category 1 as the reference group. In all models, cluster size was the only covariate.

4.1 Measurement error cluster size

Cluster size is an error-prone covariate; it is cardinal, and hence we assumed the error followed a Poisson distribution. Unfortunately, for data such as SPOT, there is no means of obtaining validation data to inform the distribution of the error short of testing all residents of the province of Quebec, which is both unethical and infeasible. Thus, to specify the mean of this Poisson distribution, we were required to estimate the cluster size distribution for those HIV positive individuals who were not in the Quebec genotyping program because they had not received an HIV test or had been tested outside of Quebec. We now describe the process by which we estimated the distribution of the error in cluster size.

The adult (age > 15) population of Quebec in 2012 was $6,802,700$ (Statistics Canada, 2013) with HIV incidence rate 7 per $100,000$ (HIV statistics Avert). Thus, the total number of *newly* HIV positive individuals in Quebec can be estimated as $6,802,700 \times 0.00007 \approx 476$. In Canada, approximately 25% of people who are living with HIV do not know that they are infected (CATIE, 2011). Therefore, the estimated number of people number of people who are HIV+ but not in the Quebec genotyping cohort in Quebec can be estimated as $(476/0.75) - 476 \approx 159$. These 159 subjects are not included in determining the cluster size. Experts believe that among the MSM community, 15% do not know their status (INSPQ, 2012), so our estimate of 25% may be conservative. Also, because clusters typically persist (Brenner and Moodie, 2012; Brenner et al., 2013) for one or at most two years (i.e. after 12-24 months, few or no new infections are observed with a viral sequence that is genetically very similar), we use the annual HIV incidence rate rather than the prevalence rate to estimate the number of HIV positive individuals in Quebec who are "missing" from our clustering cohort.

We then looked at the cluster size distribution of the 34 HIV positive individuals from the SPOT data (see, Table 1) to estimate the cluster size for 159 unseen HIV+ individuals. In the SPOT study 36% are linked to clusters that are at least of size 2-9, 29% are linked to clusters of size 1 and 35% are linked to clusters of size ≥ 10. Brown et al. (2011) estimated cluster size for MSM from HIV sequences in the United Kingdom. They reported 29% belonged to cluster size 1, 41% are linked to $2-9$ individuals and 29% are part of

cluster size of more than 10 people. Lewis et al. (2008) studied the short term dynamics of the episode among MSM in the United Kingdom. In their analysis they found that 15% belonged to cluster size 1, 60% are linked to $2-9$ and 25% belonged to cluster size ≥ 10. Based on these studies and the SPOT cluster size distribution, we propose a Poisson distribution for the error whose mean, on average, is big enough, to give us a distribution of cluster sizes that is similar to the percentages listed above (i.e. 25-30 % of people in clusters ≥ 10, 40 % in clusters of $2-9$ people). A reasonable Poisson distribution to achieve this would be Poisson(3). Poisson distributions with mean 1, 5, and 10 were also considered to evaluate the sensitivity of the results to the observed measurement error distribution.

4.2 Results

Table 1 shows the summaries of selected characteristics for 33 HIV positive MSM. The mean age of the HIV positive MSM in SPOT is 33. The average number of sex partners is 5.8. About 85% of individuals reported not using a condom on their last sexual intercourse and the majority (88.2%) reported having been tested for HIV in the last two years. Moreover, most (about 62%) belonged to clusters of size 3.

Results from all analyses, whether fit ignoring measurement error or accounting for the error using the NZM-SIMEX, were not significantly different from 0. The lack of significant findings does not appear to be driven by the small sample size leading to highly variable estimators: the estimates themselves were near the null values. For example, log-linear models examining the association between cluster size and number of sex partners (one night or total), point estimates indicate that a one-person increase in the cluster size is associated with a 0.3 - 0.5% increase in the number of sex partners. Considering that the average number of one night partners reported in the SPOT sample is (approximately) 4, one would need to compare groups of men whose cluster size differed by at least 40 people for the expected number of one night partners to increase by one individual to 5.

See Tables C1 to Table C4 in Appendix C for full results. A graphical representation of the SIMEX estimate has also been presented in Figure 4. To obtain standard errors (and p-values for the tests of association) for the NZM-SIMEX estimates, we used a bootstrap procedure with 1000 resamples. That is, both the naïve and NZM-SIMEX (both quadratic and non-linear) approaches yielded the same conclusions (cluster effect is not significant); the estimated parameter were different, but in most cases, not dramatically so. Moreover, different error distributions in all the models produce the similar results ensuring that results are robust to the assumption regarding the mean of measurement error distribution. We therefore conclude that the point estimates appear to be robust to the presence of measurement error. Thus, we conclude We observe that in cluster size is not found to have a statistically significant association with the demographic and behavioural covariates of interest, suggesting that these individual level characteristics are unlikely to be helpful in identifying – and potentially breaking the cycle of HIV transmission within – large clusters.

4.3 Limitations and discussion of the analysis

This ongoing study primarily targets participants who frequent gay social venues and therefore may not be representative of the Montreal MSM population. Therefore, the results from

Table 1: Characteristics of 33 HIV positive MSM. For quantitative variables, mean(SD) are provided; for factor variables, counts (percentage) are reported

Characteristic	Summary Measure
Age	33 (9.5)
No.of Sex Partners	5.8 (4.7)
No condom use	29 (85.3%)
(in last sexual intercourse)	
HIV tested	30 (88.2%)
(during last 24 months)	
Cluster Size	
1	10 (29.4%)
$2-3$	3 (8.8%)
>3	21 (61.7%)
Number of one night partners	4.27 (4.7)
<2	14 (42.4%)
$2-4$	4 (12.1%)
>4	15 (45.5%)

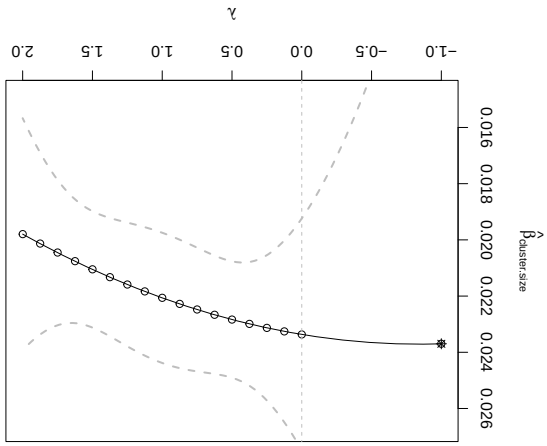


Figure 4: SIMEX estimate a quadratic extrapolation at $\lambda = -1$ from the SPOT analysis relating number of sex partner to cluster size. The naïve estimate occurs at $\lambda = 0$. The 95% pointwise confidence intervals are indicated by dotted (—) lines.

this study may not be generalized to all MSM. More importantly, our conclusions are likely affected by limited power.

It is reasonable to speculate that the data in SPOT may be correlated: it is plausible that the individuals in the study may know one another, and have similar demographic or behavioural characteristics. While the available data provide no means of assessing any correlation beyond the phylogenetic clustering, and approximately half of the individuals in the SPOT study do not share HIV phylogenetic clusters with other SPOT participants, a simple approach did not reveal significant within-cluster pairwise correlation. For example, fitting (naïve) models of the association between each of age and number of one night partners as a function of cluster size via generalized estimating equations positing an exchangeable working covariance reveals a non-significant estimate of the within-cluster correlation of approximately -0.2. The very small size of the SPOT sample creates two challenges in this regard: Lack of significance in the correlation could be driven by lack of power. On the other hand, a larger sample permit the inclusion of more covariates in the mean model, thus affording better assessment of the residual within-cluster correlation. While membership in the same HIV phylogenetic cluster can suggest direct sexual partnership, it is by no means strong evidence of it. Routinely collected sequencing data is not well suited to investigating transmission sources, as an individual whose HIV has not been sequenced may be a common source of infection or missing link in a transmission chain between two individuals in the same cluster with genetically similar viruses, thus creating challenges in identifying the likelihood that two individuals are indeed clustered in some sense beyond that suggested by the phylogeny of the virus which infects them (Eric et al., 2013). Estimators based on analysis that acknowledge the impact of clustering in data tend to be more efficient for factors that vary within cluster, thus it is possible that our analyses missed a significant finding through statistical inefficiency. Given the very small point estimates, however, it seems implausible that any relationship that would be pertinent to public health planning or policy exists in the relationships examined here.

In estimating the error variance of cluster size from the existing literature, it should be noted that measurement error in cluster size was not taken into account in the cited studies (Brown et al., 2011; Lewis et al., 2008). It is possible that our estimates of the error variance are thus too low; for this reason, we considered a range of plausible error distributions, however these did not serve to change the conclusions of our analyses.

All samples in the SPOT study were sequenced on the same platform: ABIPrism 3130xl genetic Analyser; this platform was also used in the Quebec genotyping program for the majority of the cohort's history from 2002 onwards, however the the TrueGene/Bayer HIV platform was used from April 2004 to August 2006. Genome sequence interrelationships were determined using maximum likelihood phylogenies estimated using BioEdit and MEGA2 integrated software and PAUP (version 4, Sinauer Associates). Clusters were then assigned based on high bootstrap values (>98%), short genetic lengths (<1%), and congruent polymorphisms and mutational motifs (Hue et al., 2004). To assess stability of the estimated cluster membership, phylogeny and estimate genetic distance was also estimated using a Bayesian approach via the BEAST (version 1.6.1) software; cluster size and membership estimated this way were similar to the maximum likelihood phylogenies. They were not, however, identical. Thus, both the sequencing platform and the clustering approach are additional sources of error introduced to the variable 'observed cluster size'. The distri-

butional parameters of this error (mean, variance) are unknown, and were not taken into account in our analyses. As noted above, conclusions were unchanged under a range of plausible error distributions, suggesting that taking into account additional sources of error is unlikely to alter the conclusions of the analyses.

Finally, we wish to make two points regarding the interpretation of the our analyses. First, we remind the reader that a cluster is not representative of a sexual or social network. Rather, these are clusterings of the HIV genome taken from an individuals' serum sample at a fixed point in time (fixed for each individual, but varying across individuals). Individuals are then said to cluster if the sequenced HIV genomes are determined to be 'close', in terms of phylogenetic distance. Second, we note that the analyses were undertaken only in an attempt to uncover whether there exists a significant correlation between various individual characteristics and cluster size. Cluster size evolves over time in a highly dynamic fashion, and thus the cluster size used in the analysis may not be reflective of the size of the cluster at the time when an individual was infected with HIV. We do not attempt to attribute any causal interpretation to the associations under investigation. The plausible directionality of the relationship is that individual characteristics could lead to bigger or faster-growing clusters, however as correlation (our estimand of interest) is a symmetric measure, we may 'reverse' the regressor and regressand without compromising its estimation.

5. Discussion

The simulation extrapolation procedure is useful and easily implemented technique to deal with measurement error (Cook and Stefanski 1995), however its development was until now limited to mean zero random errors. In this work, we have extended SIMEX to the case where errors can have non-zero mean errors that can follow any known parametric distribution. This was developed with the goal of analyzing data of HIV infected MSM from the SPOT study, where measurement error occurs in defining the transmission cluster size because of not including people who were unaware of their status or who had been tested outside the province of Quebec.

In this work, we focused on the relatively simple setting of additive error that is independent of both measured covariates and the true, unobserved value of the mismeasured covariate. There are many settings in which this may not be realistic. For example, in the case of the variable cluster size, it is plausible to posit that the error is related to the size of the cluster so that bigger clusters have a greater error variance than smaller clusters. This situation is considerably more challenging, since the error variance is then completely unobservable. We are currently working on extending the NZM-SIMEX to the setting where the error variance depends on observed covariates; extensions to the latent variable setting will follow.

Through a number of simulation studies we evaluated the performance of NZM-SIMEX under ideal and non-ideal conditions for a variety of outcomes and covariate distributions at different sample sizes. Simulation studies showed that NZM-SIMEX performed reasonably well in reducing biases as compared to naïve approach in all cases. The method performs well in recovering the true value of the parameter when the distribution of the measurement errors are known, and offers improvements (reduced bias) over the naïve estimator even when the distribution of errors is known only approximately.

We then applied the method to the SPOT study data, in a first attempt to elucidate correlates of HIV phylogenetic cluster size. However this method is applicable in a number of other settings. For example, in studying the association between mother's age and child mortality using data from Demographic and Health Survey (DHS) of Bangladesh, researchers are faced with the challenge that women in the DHS frequently understate their age. The NZM-SIMEX could be applied to model the relationship between child mortality and mother's age, estimating the distribution of the reporting error through hospital records or other official registries. In other populations, the impact of illicit drug use on a variety of health and quality of life outcomes is of interest. Illicit drug use may be under-reported, and the magnitude of the error could be assessed via hair or urine samples.

The main limitation of the NZM-SIMEX is that it requires knowledge of the measurement error distribution. In case of mis-specified (or, if validation data were available, poorly estimated) error distribution, it may be safer to overestimate variability of measurement error. In such cases, the NZM-SIMEX estimators perform significantly better than the naïve estimators. Thus, to reduce the measurement error bias in a variety of problems, NZM-SIMEX may be considered as a useful and easily implementable approach.

Acknowledgements

This work was supported by Dr. Moodie's Operating Grant from Canadian Institutes of Health Research (CIHR); she is also supported by a Chercheur-Boursier junior 2 career award from the Fonds de recherche du Québec-Santé (FRQ-S).

The authors wish to thank Dr. Michel Roger, and are grateful to the SPOT study group and its participants.

Appendix A. Proofs

A.1 Proof of Theorem 1

Proof Let us again consider the following simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i, \tag{3}$$

where true predictor U_i follows $\mathrm{N}(\mu_U, \sigma_U^2)$ and ϵ_i has mean 0. Suppose X_i is an imperfect measurement of U_i which is defined as

$$X_i = U_i - \delta_i^*, \tag{4}$$

where δ_i^* follows a distribution with mean μ_{δ^*} and variance $\sigma_{\delta^*}^2$, independent of U_i and Y_i. Note that under this measurement error specification, $P(X_i < U_i)$ may be at or near 1, depending on the distribution of U_i and δ_i^*.

As noted above, B new covariates $X_{i,b}(\lambda_k)$ are generated according to equation (1) so that the total measurement error variance is then the variance of $X_{i,b}(\lambda_k)$, i.e. $\sigma_{\delta^*}^2(1+\lambda_k)$. For the b^{th} data set, regressing Y on $X_b(\lambda_k)$ gives the vector of naïve estimates $\hat{\beta}_b^{NZM}(\lambda_k) = (\hat{\beta}_{0,b}(\lambda_k),\ \hat{\beta}_{1,b}(\lambda_k))'$ of $\beta_b(\lambda_k)$ found via ordinary least squares (OLS), with the average estimate at each λ_k computed according to equation (2).

To study the asymptotic mean of the average estimate of slope and intercept, we substitute (4) into (3),which gives

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i + \delta_i^*) + \epsilon_i \\ &= \beta_0 + \beta_1[X_{i,b}(\lambda_k) + \sqrt{\lambda_k}\delta_{ib}^* - (1+\sqrt{\lambda_k})\mu_{\delta^*} + \epsilon_i] \\ &= \beta_0 + \beta_1 X_{i,b}(\lambda_k) + \epsilon_i^*, \end{aligned}$$

where $\epsilon_i^* = \beta_1\{\sqrt{\lambda_k}\delta_{ib}^* - (1+\sqrt{\lambda_k})\mu_{\delta^*} + \epsilon_i\}$. For the b^{th} data set, the naïve estimate of the slope β_1 can be obtained by OLS, which yields

$$\begin{aligned} \hat{\beta}_{1b}^{NZM}(\lambda_k) &= \frac{\sum_{i=1}^{n}(X_{i,b} - \bar{X}_b)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{i,b} - \bar{X}_b)^2} \\ &= \frac{S_{XY} - \sqrt{\lambda_k}S_{Y\delta_b^*}}{S_{XX} + \lambda_k S_{\delta_b^*\delta_b^*} - 2\sqrt{\lambda_k}S_{X\delta_b^*}}. \end{aligned} \tag{5}$$

The naïve estimate of the intercept is

$$\hat{\beta}_{0b}^{NZM}(\lambda_k) = \bar{Y} - \hat{\beta}_{1b}(\lambda_k)\bar{U}. \tag{6}$$

At each λ_k, the expected value of the estimator is

$$\hat{\beta_1}^{NZM}(\lambda_k) = E\left[\hat{\beta}_{1,b}^{NZM}(\lambda_k)|\{Y_i, X_i\}_{i=1}^n\right]$$

and

$$\hat{\beta_0}^{NZM}(\lambda_k) = E\left[\bar{Y} - \hat{\beta}_{1b}^{NZM}(\lambda_k)(\bar{X} + \sqrt{\lambda_k}\bar{\delta^*})|\{Y_i, X_i\}_{i=1}^n\right],$$

where the expectation is in terms of the distribution of $\{\delta_{i,b}\}$ only.

It then follows that

$$E\left[\hat{\beta_1}^{NZM}(\lambda_k)\right] = E\left[\hat{\beta}_{1,b}^{NZM}(\lambda_k)\right]$$

and

$$E\left[\hat{\beta_0}^{NZM}(\lambda_k)\right] = E\left[\hat{\beta}_{0,b}^{NZM}(\lambda_k)\right].$$

Using the fact that

$$\begin{aligned} S_{XY} &\xrightarrow{P} \sigma_{XY}, \\ S_{XX} &\xrightarrow{P} \sigma_{XX}, \\ S_{Y\delta_b} &\xrightarrow{P} \sigma_{Y\delta_b^*}, \\ S_{\delta_b\delta_b} &\xrightarrow{P} \sigma_{\delta_b^*\delta_b^*} \\ \text{and} \quad S_{X\delta_b^*} &\xrightarrow{P} \sigma_{X\delta_b^*}, \end{aligned}$$

we obtain

$$\hat{\beta}_{1,b}^{NZM}(\lambda_k) \xrightarrow{P} \frac{\sigma_{XY} - \sqrt{\lambda_k}\sigma^*_{Y\delta_b}}{\sigma_{XX} + \lambda_k\sigma_{\delta_b^*\delta_b^*} - 2\sqrt{\lambda_k}\sigma_{X\delta_b^*}}$$

and hence

$$\hat{\beta}_{1}^{NZM}(\lambda_k) \xrightarrow{P} \frac{\sigma_{XY} - \sqrt{\lambda}\sigma_{Y\delta_b^*}}{\sigma_{XX} + \lambda_k\sigma_{\delta_b^*\delta_b^*} - 2\sqrt{\lambda_k}\sigma_{X\delta_b^*}}.$$

Here,

$$\begin{aligned}
\sigma_{XY} &= Cov(X,Y) = Cov(U,Y), \\
\sigma_{Y\delta_b^*} &= Cov(Y,\delta_b^*) = 0, \\
\sigma_{XX} &= Var(X) = Var(U+\delta^*) = \sigma_U^2 + \sigma_{\delta^*}^2, \\
\sigma_{\delta_{b^*}\delta_{b^*}} &= Var(\delta_b^*) = \sigma_{\delta^*}^2 \\
\text{and} \quad \sigma_{X\delta_b^*} &= Cov(X,\delta_b^*) = Cov(U+\delta^*,\delta_b^*) = 0
\end{aligned}$$

By substitution into (5), we obtain

$$\begin{aligned}
\hat{\beta}_1^{NZM}(\lambda_k) &\xrightarrow{P} \frac{Cov(U,Y)}{\sigma_U^2 + (1+\lambda_k)\sigma_{\delta^*}^2} \\
&= \frac{Cov(U,Y)}{Var(U)} \frac{Var(U)}{\sigma_U^2 + (1+\lambda_k)\sigma_{\delta^*}^2} \\
&= \beta_1 \left[\frac{\sigma_U^2}{\sigma_U^2 + (1+\lambda_k)\sigma_{\delta^*}^2} \right].
\end{aligned}$$

Hence,

$$\lim_{\lambda_k \to -1} plim\hat{\beta_1^{NZM}}(\lambda_k) = \beta_1.$$

Similarly, considering (6), it can be shown that

$$\lim_{\lambda_k \to -1} plim\hat{\beta}_0^{NZM}(\lambda_k) = \beta_0.$$

■

In the SIMEX extrapolation step, each component of the vector $\hat{\beta}(\lambda_k)$ is modelled as a function of λ_k for $\lambda_k \geq 0$. For example, for the slope parameter, this modelling can be considered as a nonlinear regression problem, with dependent variable $\hat{\beta}_1^{NZM}(\lambda_k)$ and independent variable λ_k having a mean function of the form

$$g(\lambda_k) = \beta_1 \left[\frac{\sigma_U^2}{\sigma_U^2 + (1+\lambda_k)\sigma_{\delta^*}^2} \right].$$

The parameter of interest, β_1, can be obtained from $g(\lambda_k)$ by extrapolation to $\lambda_k = -1$, yielding SIMEX estimate of β. We now demonstrate that the dependence of $\hat{\beta}^{NZM}$ on λ_k is a complex, non-linear form.

A.2 Proof of Theorem 2

Proof For the purposes of the proof, we will consider the slightly more complex and more realistic setting of multiple linear regression:

$$\begin{aligned} Y_i &= \beta_0 + \beta_Z Z_i + \beta_U U_i + \epsilon_i \\ &= \beta_V^t V_i + \beta_U U_i + \epsilon_i, \end{aligned} \tag{7}$$

where now $\beta_V = (\beta_0, \beta_Z)$, $V_i = (1, Z_i)$, and ϵ_i has mean 0. Here Y, V and U denote the response variable, and two covariates measured without error, respectively. As before, instead of the true predictor, U_i, an imperfect measurement X_i is available.

In the multiple linear regression setting, for the b^{th} data set, the regression model (7) can be expressed as

$$\begin{aligned} Y_i &= \beta_V^t V_i + \beta_U X_{bi} + \epsilon_i \\ &= \beta_V^t V_i + \beta_U\{X_i - \sqrt{\lambda_k}\delta_{bi}^* + (1+\sqrt{\lambda_k})\mu_{\delta^*}\} + \epsilon_i \\ &= \beta_V^t V_i + \beta_U\{X_i - \sqrt{\lambda_k}\delta_{bi}^* + a\} + \epsilon_i, \\ &= \left(V_i, \quad X_i - \sqrt{\lambda_k}\delta_{bi}^* + a\right)\begin{pmatrix}\beta_V \\ \beta_U\end{pmatrix} + \epsilon_i \end{aligned}$$

where $a = (1+\sqrt{\lambda_k})\mu_{\delta^*}$. Using OLS to estimate the parameter in (8), we obtain

$$\hat{\beta}_b^{NZM}(\lambda_k) = \left[\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B^{*T}} \\ \boldsymbol{B^*} & \boldsymbol{C^*}\end{pmatrix}\right]^{-1} \binom{k_1}{k_2^*},$$

where

$$\begin{aligned} A &= \sum V_i^{'} V_i, \\ B^* &= \sum V_i^{'} X_i - \sqrt{\lambda_k}\sum V_i^{'}\delta_{bi}^* + a\sum V_i^{'}, \\ C^* &= \lambda_k \sum \delta_{bi}^{*2} + na^2 - 2\sqrt{\lambda_k}\sum X_i^{'}\delta_{bi} + 2a\sum X_i \\ &- 2a\sqrt{\lambda_k}\sum \delta_{bi}, \\ K_1 &= \sum V_i^{'} Y_i, \\ K_2^* &= \sum X_i^{'} Y_i - \sqrt{\lambda_k}\delta_{bi}^* Y_i + a\sum Y_i. \end{aligned}$$

Equation (8) can be expressed as

$$\begin{aligned} \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B^{*T}} \\ \boldsymbol{B^*} & \boldsymbol{C^*}\end{pmatrix}\begin{pmatrix}\hat{\beta}_V(\lambda_k) \\ \hat{\beta}_U(\lambda_k)\end{pmatrix} &= \begin{pmatrix}k_1 \\ k_2^*\end{pmatrix}. \\ \text{or,} \quad \boldsymbol{A}\hat{\beta}_V(\lambda_k) + \boldsymbol{B^{*T}}\hat{\beta}_U(\lambda_k) &= k_1 \qquad (8) \\ \boldsymbol{B^*}\hat{\beta}_V(\lambda_k) + \boldsymbol{C^*}\hat{\beta}_U(\lambda_k) &= k_2^* \qquad (9) \end{aligned}$$

Solving this system of equations, we obtain the following parameters estimates:

$$\begin{aligned}
\hat{\beta}_V^{NZM}(\lambda_k) &= \boldsymbol{A}^{-1}k_1 - \frac{\boldsymbol{A}^{-1}\boldsymbol{B}^{*}k_2^{*} - \boldsymbol{A}^{-1}\boldsymbol{B}^{*}\boldsymbol{B}^{*'}\boldsymbol{A}^{-1}k_1}{\boldsymbol{C}^{*} - \boldsymbol{B}^{*'}\boldsymbol{A}^{-1}\boldsymbol{B}^{*}} \\
\text{and} & \\
\hat{\beta}_U^{NZM}(\lambda_k) &= \frac{k_2^{*} - \boldsymbol{B}^{*'}\boldsymbol{A}^{-1}k_1}{\boldsymbol{C}^{*} - \boldsymbol{B}^{*'}\boldsymbol{A}^{-1}\boldsymbol{B}^{*}} \\
&= \frac{g_1(\sqrt{\lambda_k}) - g_2(\sqrt{\lambda_k})}{g_3(\lambda_k) - g_4(\sqrt{\lambda_k})} \qquad (10)
\end{aligned}$$

Thus, we see that the components of $\hat{\beta}^{NZM}(\lambda_k)$ are non-linear functions of λ_k. ■

The complex dependence of the NZM-SIMEX estimator on λ_k suggests that the estimator may be sensitive to the choice of extrapolating function. We explore this in a comprehensive series of simulations in the section that follows.

Appendix B. Details of Simulation Study

B.1 Design of the Simulation Study

Table B1: Simulation Scenarios

Scenario	Distribution of (U,V)	True δ^*	Y	Assumed δ_b^*
1	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta^* = \lvert\delta\rvert$ and $\delta \sim N(0, 0.25)$	$N(\eta_1^a, 1)$	$\delta_b^* = \lvert\delta_b\rvert$ and $\delta_b \sim N(0, 0.25)$
2	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta^* = \lvert\delta\rvert$ and $\delta \sim N(0, 0.5)$	$N(\eta_1^a, 1)$	$\delta_b^* = \lvert\delta_b\rvert$ and $\delta_b \sim N(0, 0.5)$
3	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta^* = \lvert\delta\rvert$ and $\delta \sim N(0, 1)$	$N(\eta_1^a, 1)$	$\delta_b^* = \lvert\delta_b\rvert$ and $\delta_b \sim N(0, 1)$
4	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta^* = \lvert\delta\rvert$ and $\delta \sim N(0, 2)$	$N(\eta_1^a, 1)$	$\delta_b^* = \lvert\delta_b\rvert$ and $\delta_b \sim N(0, 2)$
5	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(1.5)$	$N(\eta_2^b, 1)$	$\delta_b^* \sim P(1.5)$
6	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(3)$	$N(\eta_2^b, 1)$	$\delta_b^* \sim P(3)$
7	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta \sim N(0, 0.5)$	$N(\eta_1^a, 1)$	$\delta_b \sim N(0, 0.25)$
8	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta \sim N(0, 0.5)$	$N(\eta_1^a, 1)$	$\delta_b \sim N(0, 0.75)$
9	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(1.5)$	$N(\eta_2^b, 1)$	$\delta_b^* \sim P(0.75)$
10	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(1.5)$	$N(\eta_2^b, 1)$	$\delta_b^* \sim P(2.25)$
11	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(1.5)$	$P(\exp(\eta_3^c))$	$\delta_b^* \sim P(1.5)$
12	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(3)$	$P(\exp(\eta_3^c))$	$\delta_b^* \sim P(3)$
13	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(3)$	$Bernoulli(p^d, 1)$	$\delta_b^* \sim P(3)$

$\eta_1^a = -2 + 1 * U + 0.25 * V + 0.25 * UV$

$\eta_2^b = 1 + 1 * U + 1 * V + 0.5 * UV$

$\eta_3^c = 0.25 + 0.5 * U + 0.05 * V + 0.05 * UV$

$p^d = \frac{\exp(\eta_4)}{1+\exp(\eta_4)}$, where $\eta_4 = -2 + 0.25 * U - 1 * V + 0.25 * UV$

B.2 Simulation Results

Table B2: Simulation results for a continuous outcome and a correctly specified error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 1: $Y \sim N(\eta_1, 1)$ and $\delta \sim N(0, 0.25)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$												
$n = 100$												
β_0	-1.631	0.369	0.123	0.151	-1.997	0.003	0.119	0.014	-1.998	0.002	0.119	0.014
β_U	0.900	-0.099	0.121	0.025	0.998	-0.002	0.136	0.018	1.000	0.000	0.136	0.019
β_V	0.389	0.139	0.124	0.035	0.255	0.005	0.121	0.015	0.254	0.004	0.122	0.015
β_{UV}	0.225	-0.025	0.101	0.011	0.243	-0.008	0.110	0.012	0.242	-0.008	0.111	0.012
$n = 500$												
β_0	-1.634	0.366	0.054	0.137	-2.000	-0.000	0.053	0.003	-2.001	-0.001	0.053	0.003
β_U	0.899	-0.101	0.048	0.013	0.996	-0.004	0.054	0.003	0.998	-0.002	0.054	0.003
β_V	0.388	0.138	0.055	0.022	0.252	0.002	0.055	0.003	0.251	0.001	0.055	0.003
β_{UV}	0.232	-0.018	0.041	0.002	0.249	-0.001	0.045	0.002	0.249	-0.001	0.045	0.002
$n = 1000$												
β_0	-1.634	0.366	0.037	0.135	-1.999	0.000	0.037	0.001	-1.999	0.000	0.037	0.001
β_U	0.897	-0.103	0.034	0.012	0.994	-0.006	0.038	0.002	0.997	-0.003	0.039	0.002
β_V	0.389	0.139	0.039	0.021	0.253	0.003	0.038	0.001	0.257	0.002	0.038	0.001
β_{UV}	0.232	-0.018	0.031	0.001	0.249	-0.001	0.033	0.001	0.249	-0.001	0.033	0.001
Scenario 2: $Y \sim N(\eta_1, 1)$ and $\delta \sim N(0, 0.5)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$												
$n = 100$												
β_0	-1.523	0.477	0.132	0.245	-1.997	0.003	0.125	0.016	-1.998	0.002	0.125	0.016
β_U	0.817	-0.183	0.119	0.048	0.979	-0.021	0.147	0.022	0.990	-0.009	0.149	0.023
β_V	0.454	0.204	0.133	0.059	0.264	0.014	0.129	0.017	0.259	0.009	0.131	0.017
β_{UV}	0.209	-0.041	0.101	0.012	0.239	-0.011	0.119	0.014	0.239	-0.010	0.121	0.015
$n = 500$												
β_0	-1.526	0.474	0.058	0.228	-1.999	0.000	0.056	0.003	2.001	-0.001	0.056	0.003
β_U	0.816	-0.184	0.048	0.036	0.976	-0.024	0.058	0.004	0.987	-0.013	0.059	0.004
β_V	0.455	0.205	0.059	0.045	0.261	0.011	0.058	0.004	0.256	0.006	0.058	0.003
β_{UV}	0.217	-0.033	0.042	0.003	0.247	-0.003	0.048	0.002	0.248	-0.002	0.049	0.002
$n = 1000$												
β_0	-1.526	0.474	0.039	0.226	-1.999	0.001	0.039	0.001	-1.999	0.001	0.039	0.002
β_U	0.814	-0.186	0.034	0.036	0.974	-0.026	0.041	0.002	0.985	-0.015	0.042	0.002
β_V	0.456	0.206	0.041	0.044	0.262	0.012	0.039	0.002	0.257	0.007	0.040	0.002
β_{UV}	0.217	-0.033	0.031	0.002	0.247	-0.003	0.036	0.001	0.248	-0.002	0.036	0.001

continued

Table B2: (cont.) Simulation results for a continuous outcome and a correctly specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 3: $Y \sim N(\eta_1, 1)$ and $\delta \sim N(0, 1)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$												
$n = 100$												
β_0	-1.424	0.576	0.146	0.353	-1.994	0.006	0.134	0.018	-1.997	0.003	0.136	0.019
β_U	0.690	-0.309	0.115	0.109	0.914	-0.087	0.158	0.033	0.947	-0.053	0.168	0.031
β_V	0.539	0.289	0.147	0.105	0.293	0.043	0.139	0.021	0.278	0.028	0.144	0.022
β_{UV}	0.183	-0.067	0.101	0.015	0.229	-0.021	0.132	0.018	0.232	-0.018	0.137	0.019
$n = 500$												
β_0	-1.425	0.575	0.063	0.334	-1.996	0.004	0.060	0.0034	-1.999	0.001	0.061	0.004
β_U	0.689	-0.311	0.047	0.099	0.911	-0.089	0.064	0.012	0.943	-0.0567	0.067	0.008
β_V	0.542	0.292	0.064	0.089	0.289	0.039	0.063	0.006	0.275	0.025	0.064	0.005
β_{UV}	0.191	-0.059	0.042	0.005	0.239	-0.011	0.053	0.011	0.243	-0.007	0.055	0.003
$n = 1000$												
β_0	-1.426	0.574	0.043	0.331	-1.995	0.005	0.042	0.002	-1.997	0.003	0.042	0.002
β_U	0.687	-0.313	0.033	0.099	0.907	-0.093	0.044	0.011	0.939	-0.061	0.047	0.006
β_V	0.544	0.294	0.045	0.089	0.292	0.042	0.043	0.004	0.277	0.027	0.044	0.003
β_{UV}	0.192	-0.058	0.030	0.004	0.292	0.042	0.043	0.004	0.243	-0.007	0.039	0.003
Scenario 4: $Y \sim N(\eta_1, 1)$ and $\delta \sim N(0, 2)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$.												
$n = 100$												
β_0	-1.367	0.633	0.163	0.428	-1.986	0.014	0.145	0.021	-1.993	0.008	0.151	0.023
β_U	0.527	-0.473	0.105	0.235	0.769	-0.230	0.161	0.079	0.833	-0.167	0.181	0.061
β_V	0.631	0.381	0.167	0.173	0.357	0.107	0.150	0.034	0.328	0.078	0.159	0.031
β_{UV}	0.147	-0.103	0.097	0.020	0.203	-0.047	0.142	0.022	0.212	-0.038	0.155	0.026
$n = 500$												
β_0	-1.366	0.634	0.0699	0.407	-1.986	0.014	0.065	0.004	-1.992	0.008	0.067	0.005
β_U	0.526	-0.474	0.044	0.227	0.768	-0.232	0.066	0.058	0.829	-0.170	0.074	0.035
β_V	0.637	0.387	0.072	0.155	0.353	0.103	0.067	0.015	0.325	0.075	0.071	0.011
β_{UV}	0.155	-0.095	0.040	0.010	0.215	-0.035	0.057	0.005	0.226	-0.024	0.062	0.004
$n = 1000$												
β_0	-1.427	0.5763	0.1458	0.353	-1.994	0.006	0.134	0.018	-1.997	0.003	0.136	0.019
β_U	0.690	-0.309	0.115	0.109	0.914	-0.087	0.158	0.033	0.947	-0.053	0.168	0.031
β_V	0.539	0.289	0.147	0.105	0.293	0.043	0.139	0.021	0.278	0.028	0.144	0.022
β_{UV}	0.183	-0.067	0.101	0.015	0.229	-0.021	0.132	0.018	0.232	-0.018	0.137	0.019

continued

Table B2: (cont.) Simulation results for a continuous outcome and a correctly specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 5: $Y \sim N(\eta_2, 1)$ and $\delta^* \sim P(1.5)$. True values of the parameters are $\beta_0 = 1$, $\beta_U = 1$, $\beta_V = 1$ and $\beta_{UV} = 0.5$.												
$n = 100$												
β_0	3.672	2.672	0.510	7.402	1.076	0.076	0.696	0.489	1.035	0.035	0.712	0.509
β_U	0.889	-0.111	0.045	0.014	0.994	-0.006	0.055	0.003	0.998	-0.002	0.056	0.003
β_V	2.300	1.300	0.628	2.085	0.995	-0.005	0.856	0.733	0.982	-0.018	0.872	0.761
β_{UV}	0.447	-0.053	0.054	0.006	0.500	0.000	0.066	0.004	0.501	0.001	0.067	0.005
$n = 500$												
β_0	3.667	2.667	0.238	7.169	1.076	0.076	0.308	0.101	1.038	0.038	0.310	0.098
β_U	0.889	-0.111	0.021	0.013	0.994	-0.006	0.024	0.001	0.997	-0.003	0.024	0.001
β_V	2.337	1.337	0.273	1.862	1.049	0.048	0.366	0.136	1.029	0.029	0.373	0.139
β_{UV}	0.444	-0.056	0.024	0.004	0.496	-0.004	0.029	0.002	0.487	-0.002	0.029	0.001
$n = 1000$												
β_0	3.658	2.658	0.166	7.093	1.067	0.067	0.220	0.053	1.028	0.028	0.224	0.051
β_U	0.889	-0.111	0.015	0.012	0.994	-0.006	0.017	0.000	0.998	-0.003	0.017	0.000
β_V	2.333	1.334	0.183	1.812	1.038	0.038	0.244	0.061	1.017	0.017	0.249	0.063
β_{UV}	0.444	-0.056	0.016	0.003	0.497	-0.003	0.019	0.000	0.499	-0.002	0.019	0.000
Scenario 6: $Y \sim N(\eta_2, 1)$ and $\delta^* \sim P(3)$. True values of the parameters are $\beta_0 = 1$, $\beta_U = 1$, $\beta_V = 1$ and $\beta_{UV} = 0.5$.												
$n = 33$												
β_0	5.767	4.767	1.088	23.914	1.324	0.324	1.850	3.531	1.144	0.144	1.971	3.908
β_U	0.801	-0.198	0.108	0.051	0.972	-0.027	0.145	0.021	0.987	-0.012	0.155	0.024
β_V	3.307	2.307	1.246	6.879	1.030	0.030	2.165	4.689	0.923	-0.076	2.341	5.490
β_{UV}	0.406	-0.093	0.122	0.023	0.495	-0.004	0.168	0.028	0.503	0.003	0.183	0.033
$n = 100$												
β_0	5.764	4.764	0.518	22.968	1.287	0.287	0.880	0.858	1.119	0.119	0.934	0.887
β_U	0.804	-0.197	0.052	0.041	0.976	-0.024	0.069	0.005	0.989	-0.010	0.074	0.006
β_V	3.344	2.344	0.643	5.908	1.088	0.088	1.108	1.236	1.015	0.015	1.167	1.362
β_{UV}	0.405	-0.095	0.063	0.013	0.492	-0.008	0.086	0.007	0.498	-0.002	0.091	0.008
$n = 500$												
β_0	5.762	4.762	0.241	22.732	1.286	0.286	0.389	0.233	1.125	0.125	0.402	0.178
β_U	0.804	-0.196	0.024	0.039	0.976	-0.024	0.030	0.002	0.989	-0.010	0.031	0.001
β_V	3.383	2.383	0.282	5.756	1.156	0.156	0.477	0.252	1.076	0.076	0.503	0.259
β_{UV}	0.402	-0.098	0.028	0.010	0.487	-0.013	0.037	0.002	0.494	-0.007	0.039	0.002
$n = 1000$												
β_0	5.749	4.749	0.168	22.579	1.269	0.269	0.276	0.149	1.106	0.106	0.288	0.094
β_U	0.805	-0.195	0.017	0.038	0.977	-0.023	0.022	0.001	0.990	-0.009	0.023	0.001
β_V	3.379	2.379	0.193	5.701	1.143	0.143	0.321	0.123	1.061	0.061	0.334	0.116
β_{UV}	0.402	-0.098	0.019	0.010	0.488	-0.012	0.025	0.001	0.495	-0.006	0.026	0.001

Table B3: Simulation results for a continuous outcome and a mis-specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 7: $Y \sim N(\eta_2, 1)$, $\delta \sim N(0, 0.5)$ and $\delta_b \sim N(0, 0.25)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$.												
$n = 100$												
β_0	-1.523	0.476	0.132	0.245	-1.842	0.157	0.121	0.039	-1.842	0.157	0.121	0.039
β_U	0.817	-0.182	0.119	0.047	0.897	-0.102	0.132	0.028	0.899	-0.100	0.133	0.027
β_V	0.454	0.204	0.132	0.059	0.337	0.087	0.124	0.023	0.336	0.086	0.125	0.023
β_{UV}	0.209	-0.040	0.101	0.011	0.224	-0.025	0.109	0.012	0.223	-0.026	0.109	0.012
$n = 500$												
β_0	-1.525	0.474	0.057	0.228	-1.844	0.155	0.054	0.027	-1.844	0.155	0.054	0.027
β_U	0.816	-0.183	0.047	0.036	0.895	-0.104	0.052	0.013	0.897	-0.102	0.052	0.013
β_V	0.454	0.204	0.058	0.045	0.334	0.084	0.056	0.010	0.334	0.084	0.056	0.010
β_{UV}	0.216	-0.033	0.041	0.002	0.231	-0.018	0.044	0.002	0.231	-0.018	0.044	0.002
$n = 1000$												
β_0	-1.526	0.473	0.039	0.226	-1.844	0.155	0.037	0.025	-1.844	0.155	0.037	0.025
β_U	0.814	-0.185	0.033	0.035	0.893	-0.106	0.037	0.012	0.895	-0.104	0.037	0.012
β_V	0.456	0.206	0.041	0.044	0.336	0.086	0.038	0.008	0.335	0.085	0.038	0.008
β_{UV}	0.216	-0.033	0.030	0.002	0.231	-0.018	0.032	0.001	0.231	-0.018	0.032	0.001
Scenario 8: $Y \sim N(\eta_2, 1)$, $\delta \sim N(0, 0.5)$ and $\delta_b \sim N(0, 0.75)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$.												
$n = 100$												
β_0					-2.136	-0.136	0.131	0.036	-2.143	-0.143	0.133	0.038
β_U					1.054	0.054	0.160	0.028	1.084	0.084	0.168	0.035
β_V					0.198	-0.051	0.135	0.021	0.184	-0.065	0.139	0.023
β_{UV}					0.254	0.004	0.128	0.016	0.256	0.006	0.133	0.017
$n = 500$												
β_0					-2.139	-0.139	0.059	0.023	-2.145	-0.145	0.059	0.024
β_U					1.051	0.051	0.063	0.006	1.081	0.081	0.066	0.010
β_V					0.193	-0.056	0.061	0.006	0.179	-0.070	0.062	0.008
β_{UV}					0.262	0.012	0.051	0.002	0.265	0.015	0.053	0.003
$n = 1000$												
β_0					-2.138	-0.138	0.040	0.020	-2.143	-0.143	0.041	0.022
β_U					1.048	0.048	0.044	0.004	1.078	0.078	0.046	0.008
β_V					0.194	-0.055	0.041	0.004	0.181	-0.068	0.042	0.006
β_{UV}					0.261	0.011	0.038	0.001	0.264	0.014	0.039	0.001
Scenario 9: $Y \sim N(\eta_2, 1)$, $\delta^* \sim P(1.5)$ and $\delta_b^* \sim P(0.75)$. True values of the parameters are $\beta_0 = 1$, $\beta_U = 1$, $\beta_V = 1$ and $\beta_{UV} = 0.5$.												
$n = 100$												
β_0	3.672	2.672	0.510	7.401	2.423	1.423	0.590	2.374	2.417	1.417	0.596	2.363
β_U	0.888	-0.111	0.045	0.014	0.940	-0.059	0.049	0.006	0.941	-0.058	0.049	0.006
β_V	2.300	1.300	0.628	2.085	1.670	0.670	0.729	0.981	1.673	0.673	0.729	0.985
β_{UV}	0.447	-0.052	0.054	0.006	0.473	-0.026	0.059	0.004	0.473	-0.026	0.059	0.004
$n = 500$												
β_0	3.666	2.666	0.237	7.168	2.420	1.420	0.268	2.089	2.415	1.415	0.268	2.076
β_U	0.889	-0.110	0.020	0.012	0.940	-0.059	0.021	0.003	0.941	-0.058	0.021	0.004
β_V	2.336	1.336	0.273	1.861	1.717	0.717	0.313	0.612	1.713	0.713	0.316	0.609
β_{UV}	0.444	-0.055	0.023	0.003	0.469	-0.030	0.025	0.001	0.470	-0.029	0.026	0.001
$n = 1000$												
β_0	3.658	2.658	0.166	7.093	2.411	1.411	0.189	2.028	2.405	1.405	0.190	2.010
β_U	0.889	-0.110	0.014	0.012	0.940	-0.059	0.015	0.003	0.941	-0.058	0.015	0.003
β_V	2.333	1.333	0.183	1.811	1.710	0.710	0.209	0.548	1.705	0.705	0.211	0.542
β_{UV}	0.444	-0.055	0.015	0.003	0.470	-0.029	0.016	0.001	0.470	-0.029	0.017	0.001

continued

Table B3: (cont.)Simulation results: Simulation results for a continuous outcome and a misspecified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 10: $Y \sim N(\eta_2, 1)$, $\delta^* \sim P(1.5)$ and $\delta_b^* \sim P(2.25)$. True values of the parameters are $\beta_0 = 1$, $\beta_U = 1$, $\beta_V = 1$ and $\beta_{UV} = 0.5$.												
$n = 100$												
β_0					-0.343	-1.343	0.817	2.472	-0.464	-1.464	0.859	2.884
β_U					1.047	0.047	0.061	0.005	1.056	0.056	0.064	0.007
β_V					0.283	-0.716	1.001	1.516	0.232	-0.767	1.048	1.687
β_{UV}					0.526	0.026	0.073	0.006	0.530	0.030	0.077	0.006
$n = 500$												
β_0					-0.341	-1.341	0.354	1.925	-0.458	-1.458	0.364	2.260
β_U					1.046	0.046	0.025	0.002	1.055	0.055	0.026	0.003
β_V					0.343	-0.656	0.424	0.611	0.285	-0.714	0.444	0.708
β_{UV}					0.522	0.022	0.031	0.001	0.526	0.026	0.033	0.001
$n = 1000$												
β_0					-0.351	-1.351	0.255	1.892	-0.470	-1.470	0.265	2.232
β_U					1.047	0.047	0.018	0.002	1.056	0.056	0.019	0.003
β_V					0.328	-0.671	0.284	0.531	0.267	-0.732	0.297	0.625
β_{UV}					0.523	0.023	0.020	0.001	0.528	0.021	0.028	0.001

Table B4: Simulation results for a Poisson outcome and a correctly specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 11: $Y \sim P(\exp(\eta_3))$ and $\delta^* \sim P(1.5)$. True values of the parameters are $\beta_0 = 0.25$, $\beta_U = 0.5$, $\beta_V = 0.05$ and $\beta_{UV} = 0.05$.												
$n = 100$												
β_0	1.258	1.008	0.136	1.036	0.302	0.052	0.247	0.064	0.231	-0.018	0.275	0.076
β_U	0.379	-0.120	0.045	0.016	0.483	-0.016	0.056	0.003	0.501	0.001	0.063	0.004
β_V	0.163	0.113	0.136	0.031	0.060	0.010	0.271	0.073	0.054	0.004	0.306	0.094
β_{UV}	0.036	-0.013	0.047	0.002	0.047	-0.002	0.065	0.004	0.047	-0.002	0.074	0.005
$n = 500$												
β_0	1.240	0.990	0.071	0.985	0.268	0.018	0.125	0.016	0.194	-0.055	0.138	0.022
β_U	0.390	-0.109	0.024	0.012	0.491	-0.008	0.028	0.000	0.510	0.010	0.031	0.001
β_V	0.150	0.100	0.072	0.015	0.040	-0.009	0.138	0.019	0.030	-0.019	0.155	0.024
β_{UV}	0.042	-0.007	0.025	0.000	0.052	0.002	0.032	0.001	0.054	0.004	0.036	0.001
$n = 1000$												
β_0	1.241	0.991	0.050	0.985	0.270	0.020	0.084	0.007	0.198	-0.051	0.094	0.011
β_U	0.390	-0.109	0.017	0.012	0.491	-0.008	0.019	0.000	0.510	0.010	0.021	0.000
β_V	0.153	0.103	0.053	0.013	0.051	0.001	0.101	0.010	0.042	-0.007	0.115	0.013
β_{UV}	0.041	-0.008	0.019	0.000	0.049	-0.000	0.023	0.000	0.051	0.001	0.027	0.000
Scenario 12: $Y \sim P(\exp(\eta_3))$ and $\delta^* \sim P(3)$. True values of the parameters are $\beta_0 = 0.25$, $\beta_U = 0.5$, $\beta_V = 0.05$ and $\beta_{UV} = 0.05$.												
$n = 33$												
β_0	2.695	2.445	1.232	7.502	0.415	0.165	2.369	5.642	0.304	0.054	2.970	8.827
β_U	0.411	-0.088	0.100	0.017	0.487	-0.012	0.154	0.024	0.495	-0.004	0.193	0.037
β_V	0.320	0.270	1.238	1.607	0.059	0.009	1.540	1.454	0.075	0.025	1.3011	1.898
β_{UV}	0.039	-0.010	0.104	0.010	0.048	-0.001	0.169	0.028	0.047	-0.002	0.219	0.048
$n = 100$												
β_0	1.888	1.638	0.105	2.694	0.490	0.240	0.320	0.160	0.351	0.101	0.370	0.147
β_U	0.301	-0.198	0.053	0.042	0.430	-0.069	0.071	0.009	0.466	-0.033	0.082	0.007
β_V	0.229	0.179	0.106	0.043	0.090	0.040	0.351	0.124	0.078	0.028	0.425	0.181
β_{UV}	0.027	-0.022	0.054	0.003	0.040	-0.009	0.082	0.006	0.042	-0.007	0.100	0.010
$n = 500$												
β_0	1.889	1.639	0.051	2.691	0.433	0.183	0.173	0.063	0.288	0.038	0.195	0.039
β_U	0.313	-0.186	0.029	0.035	0.445	-0.054	0.038	0.004	0.483	-0.016	0.043	0.002
β_V	0.226	0.176	0.054	0.034	0.053	0.003	0.177	0.031	0.033	-0.016	0.208	0.043
β_{UV}	0.035	-0.014	0.028	0.001	0.049	-0.001	0.040	0.001	0.053	0.003	0.047	0.002
$n = 1000$												
β_0	1.892	1.642	0.035	2.699	0.435	0.185	0.113	0.047	0.292	0.042	0.123	0.017
β_U	0.313	-0.186	0.020	0.035	0.446	-0.053	0.025	0.003	0.482	-0.017	0.027	0.001
β_V	0.226	0.176	0.039	0.032	0.064	0.014	0.128	0.016	0.049	-0.001	0.148	0.022
β_{UV}	0.034	-0.015	0.022	0.001	0.046	-0.003	0.029	0.001	0.049	-0.001	0.034	0.001

Table B5: Simulation results for a Bernoulli outcome and a correctly specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 13: $Y \sim Bernoulli(p)$ and $\delta^* \sim P(3)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 0.25$, $\beta_V = -1$ and $\beta_{UV} = 0.25$.												
$n = 100$												
β_0	-0.836	1.164	0.769	1.946	-2.033	-0.033	1.379	1.905	-2.083	-0.083	1.439	2.079
β_U	0.204	-0.045	0.097	0.011	0.259	0.009	0.131	0.017	0.263	0.013	0.137	0.019
β_V	0.236	1.236	0.929	2.391	-0.953	0.048	1.657	2.748	-0.991	0.009	1.724	2.974
β_{UV}	0.204	-0.046	0.122	0.017	0.260	0.010	0.165	0.027	0.264	0.014	0.171	0.029
$n = 500$												
β_0	-0.767	1.233	0.279	1.599	-1.872	0.128	0.483	0.249	-1.916	0.084	0.499	0.256
β_U	0.189	-0.061	0.034	0.005	0.239	-0.011	0.045	0.002	0.243	-0.007	0.046	0.002
β_V	0.224	1.224	0.349	1.621	-0.864	0.136	0.607	0.386	-0.904	0.096	0.625	0.400
β_{UV}	0.187	-0.063	0.045	0.006	0.238	-0.012	0.059	0.004	0.242	-0.008	0.060	0.004
$n = 1000$												
β_0	-0.777	1.223	0.195	1.533	-1.886	0.114	0.337	0.126	-1.930	0.069	0.345	0.124
β_U	0.189	-0.060	0.024	0.004	0.239	-0.010	0.031	0.001	0.244	-0.006	0.032	0.001
β_V	0.228	1.228	0.253	1.571	-0.858	0.142	0.439	0.213	-0.898	0.102	0.448	0.211
β_{UV}	0.186	-0.064	0.031	0.005	0.238	-0.012	0.042	0.002	0.242	-0.008	0.043	0.002

Appendix C. Additional results from the SPOT analysis

Table C1: Results from simple linear regression. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step. In the first panel, β_1 indicates the expected difference in age between two groups of men whose cluster size differs by one individual, whereas in the second panel, β_1 expected difference in the number of sex partners associated with a one-person difference in cluster size.

		Naïve			SIMEX-Q			SIMEX-NL		
Parameter	μ^*	$\hat{\beta}^{OLS}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value
Model: Relating age to cluster size										
β_0	3	34.283	2.207	0.000	34.607	2.801	0.000	34.733	2.801	0.000
β_1	3	-0.123	0.141	0.388	-0.119	0.149	0.421	-0.131	0.149	0.381
β_0	1				34.441	2.633	0.000	34.374	2.635	0.000
β_1	1				-0.126	0.153	0.409	-0.120	0.154	0.434
β_0	5				34.979	3.142	0.000	34.905	3.143	0.000
β_1	5				-0.129	0.158	0.415	-0.123	0.158	0.438
β_0	10				5.279	1.645	0.001	5.239	1.659	0.002
β_1	10				0.025	0.077	0.746	0.027	0.078	0.729
Model: Relating number of sex partners to cluster size										
β_0	3	5.552	1.119	0.000	5.524	1.064	0.000	5.505	1.064	0.000
β_1	3	0.023	0.071	0.753	0.023	0.065	0.720	0.025	0.065	0.702
β_0	1				5.449	1.164	0.000	5.466	1.167	0.000
β_1	1				0.025	0.067	0.705	0.024	0.067	0.724
β_0	5				5.349	1.321	0.000	5.415	1.320	0.000
β_1	5				0.028	0.069	0.678	0.024	0.069	0.725
β_0	10				1.673	0.299	0.000	1.673	0.299	0.000
β_1	10				0.004	0.013	0.755	0.004	0.013	0.759

* mean of the measurement error distribution, Poisson(μ)

Table C2: Results from the simple logistic regression model. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step. In the top and bottom panels, β_1 represents the difference in the log odds ratio for, respectively, the use of a condom at the last sexual intercourse and having had an HIV last in the last 24 months associated with a one-person difference in cluster size

		Naïve			SIMEX-Q			SIMEX-NL		
Parameter	μ^*	$\hat{\beta}^{OLS}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value
Model: Relating condom use to cluster size										
β_0	3	2.619	0.772	0.001	2.198	1.104	0.012	2.197	1.105	0.012
β_1	3	-0.047	0.036	0.195	-0.013	0.083	0.567	-0.013	0.083	0.559
β_0	1				2.151	0.946	0.005	2.151	0.945	0.005
β_1	1				-0.010	0.083	0.569	-0.010	0.083	0.572
β_0	5				2.229	1.289	0.025	2.229	1.291	0.027
β_1	5				-0.013	0.085	0.564	-0.013	0.085	0.581
β_0	10				2.289	1.739	0.076	2.288	1.746	0.069
β_1	10				-0.014	0.088	0.591	-0.013	0.088	0.565
Model: Relating HIV tests in the last 24 months to cluster size										
β_0	3	1.686	0.732	0.021	1.226	0.654	0.015	1.228	0.655	0.018
β_1	3	0.038	0.067	0.573	0.044	0.054	0.515	0.044	0.054	0.482
β_0	1				1.311	0.564	0.004	1.311	0.564	0.003
β_1	1				0.046	0.057	0.499	0.046	0.057	0.535
β_0	5				1.126	0.781	0.055	1.130	0.781	0.071
β_1	5				0.045	0.058	0.527	0.044	0.058	0.448
β_0	10				0.955	1.082	0.255	0.964	1.095	0.219
β_1	10				0.042	0.058	0.486	0.041	0.059	0.559

* mean of the measurement error distribution, Poisson(μ)

Table C3: Results from log-linear model of number of sex partners on cluster size. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step. β_1 indicates the expected difference the number of sex partners (top panel) and one night sex partners (bottom panel), on the log scale, between two groups of men whose cluster size differs by one individual.

		Naïve			SIMEX-Q			SIMEX-NL		
Parameter	μ^*	$\hat{\beta}^{OLS}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value
Model: Relating number of sex partners on cluster size										
β_0	3	1.716	0.097	0.000	1.709	0.202	0.000	1.700	0.201	0.000
β_1	3	0.004	0.006	0.526	0.003	0.011	0.752	0.004	0.011	0.702
β_0	1				1.714	0.184	0.000	1.709	0.184	0.000
β_1	1				0.004	0.011	0.737	0.004	0.011	0.711
β_0	5				1.692	0.224	0.000	1.705	0.224	0.000
β_1	5				0.004	0.011	0.707	0.003	0.011	0.762
β_0	10				1.659	0.275	0.000	1.659	0.278	0.000
β_1	10				0.005	0.012	0.704	0.005	0.013	0.706
Model: Relating number of one night partners on cluster size										
β_0	3	1.411	0.112	0.000	1.399	0.296	0.000	1.389	0.296	0.000
β_1	3	0.004	0.006	0.568	0.004	0.015	0.797	0.004	0.015	0.761
β_0	1				1.407	0.267	0.000	1.403	0.267	0.000
β_1	1				0.004	0.014	0.787	0.004	0.014	0.767
β_0	5				1.390	0.314	0.000	1.391	0.313	0.000
β_1	5				0.004	0.016	0.801	0.003	0.016	0.803
β_0	10				1.354	0.408	0.001	1.345	0.414	0.001
β_1	10				0.005	0.018	0.789	0.005	0.018	0.778

* mean of the measurement error distribution, Poisson(μ)

Table C4: Results from multinomial model of number of one night partners on cluster size. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step. $\beta_{1(2-4)}$ indicates the expected difference in the log odds of having 2-4 one night partners between two groups of men whose cluster size differs by one individual; $\beta_{1(5+)}$ is the expected difference in the log odds of having at least 5 one night partners between two groups of men whose cluster size differs by one individual.

		Naïve			SIMEX-Q			SIMEX-NL		
Parameter	μ^*	$\hat{\beta}^{OLS}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value
Model: Relating number of one night partners on cluster size										
$\beta_{0(2-4)}$	3	-1.605	0.780	0.039	-1.727	15.143	0.909	-1.740	31.497	0.955
$\beta_{1(2-4)}$	3	0.036	0.051	0.474	0.037	0.573	0.948	0.038	1.179	0.974
$\beta_{0(5+)}$	3	-0.302	0.522	0.562	-0.427	0.760	0.574	-0.441	0.762	0.562
$\beta_{1(5+)}$	3	0.038	0.039	0.331	0.039	0.058	0.499	0.039	0.057	0.488
$\beta_{0(2-4)}$	1				-1.660	12.812	0.897	-1.630	22.317	0.942
$\beta_{1(2-4)}$	1				0.038	0.524	0.942	0.036	0.903	0.968
$\beta_{0(5+)}$	1				-0.347	0.660	0.599	-0.354	0.660	0.591
$\beta_{1(5+)}$	1				0.038	0.055	0.483	0.039	0.055	0.477
$\beta_{0(2-4)}$	5				-1.792	16.775	0.915	-1.817	38.775	0.963
$\beta_{1(2-4)}$	5				0.036	0.591	0.950	0.039	1.351	0.977
$\beta_{0(5+)}$	5				-0.499	0.879	0.569	-0.501	0.881	0.569
$\beta_{1(5+)}$	5				0.038	0.060	0.524	0.038	0.060	0.518
$\beta_{0(2-4)}$	10				-2.012	14.199	0.887	-2.01	34.997	0.954
$\beta_{1(2-4)}$	10				0.039	0.428	0.927	0.038	1.045	0.970
$\beta_{0(5+)}$	10				-0.767	1.182	0.516	-0.746	1.214	0.538
$\beta_{1(5+)}$	10				0.042	0.064	0.511	0.041	0.067	0.544

* mean of the measurement error distribution, Poisson(μ)

References

Allodji, R. S., Thiúbaut, A. C. M., Leuraud, K., Rage, E., Henry, S., Laurier, D. and Bénichou, J. (2012). The performance of functional methods for correcting non-Gaussian measurement error within Poisson regression: corrected excess risk of lung cancer mortality in relation to radon exposure among French uranium miners. *Statistics in Medicine* **31**, 4428-4443.

http://www.amfar.org/about-hiv-and-aids/facts-and-stats/statistics--worldwide/

http://www.cdc.gov/std/stats/sti-estimates-fact-sheet-feb-2013.pdf

http://www.avert.org/canada-hiv-aids-statistics.htm

Brenner, B. G. , Roger, M., Routy, J. P., Moisi, D., Ntemgwa, M., Matte, C., Baril, J. G., Thomas, R., Rouleau, D., Bruneau, J., Leblanc, R., Legault, M., Tremblay, C., Charest, H., Wainberg, M. A., and the Quebec PHI Study Group. (2007). High Rates of Forward Transmission Events Following Acute/Early HIV-1 Infection. *Journal of Infectious Disease* **195(7)**,951-9. PMID: 17330784.

Brenner, B. G., Wainberg, M. A., and Roger, M. (2013). Phylogenetic inferences on HIV -1 transmission: implications for the design of prevention and treatment interventions. *AIDS* **27**, 1045-1057, PMID:23902920.

Brenner, B. G. and Moodie, E. E. M. (2012). HIV Sexual Networks: The Montreal Experience. *Statistical Communications in Infectious Diseases* **4:1**, 1948-4690, DOI: 10.1515/1948-4690.1039.

Brenner, B. G., and Wainberg, M. A. (2013). Future of phylogeny in prevention. *Journal of Acquired Immune Deficiency Syndrome* **2**, S248-54, PMID:23764643.

Brown, A. J. L., Lycett, S. J., Weinert, L., Hughes, G. J., Fearnhill, E. and Dunn, D. T. (2011). Transmission Network Parameters Estimated From HIV Sequences for a Nationwide Epidemic. *The Journal of Infectious Diseases* **204**, 14639.

Carrol, R. J., and Stefanski, L. A. (1990). Approximate qasilikelihood estimation in models with surrogate predictors. *Journal of American Statistical Association* **85**, 652-663.

Carrol, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association* **91**, 242-250.

Carroll, R., Ruppert, D. and Stefanski, L. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

Carroll, R., Ruppert, D., L.A., S. and C.M., C. (2006). *Measurement Error in Nonlinear Models*. Chapman and Hall.

Cole, S .R., Chu, H. and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* **35**, 1074-1081.

Costas, L., Infante-Rivard, C., Zock, J. P., Tongeren, M. V., Boffetta, P., Cusson, A., Robles, C., Casabonne, D., Benavente, Y., Becker, N., Brennan, P., Foretova, L., Maynadie, M., Staines, A., Nieters, A., Cocco, P. and Sanjose, S. D. (2015). Occupational exposure to endocrine disruptors and lymphoma risk in a multu-centric European study. *British journal of Cancer* **112**, 1251-1256.

Cook, J. and Stefanski, L. A. (1995). A simulation extrapolation method for parametric measurement error models. *Journal of American Statistical Association* **89**, 1314-1328.

Erik, M. V. and Simon D. W. F. (2013). Inferring the Source of Transmission with Phylogenetic Data. *Computational Biology*, DOI: 10.1371/journal.pcbi.1003397.

http://publications.gc.ca/collections/collection_2013/aspc-phac/HP37-13-2008-eng.pdf

Gleser, L. J. (1990). Improvements of Naive approach to estimation in nonlinear errors-in-variables regression models. In *Statistical Analysis of measurement Error Models and Application*, P. J. Brown and W. A. Fuller, editors. American mathematics Society, Province.

He, W., Xiong, J. and Yi, G. Y. (2012). SIMEX R package for Accelerated Failure Time Models with Covariate Measurement Error. *Journal of Statistical Software* **46**, Code Snippet 1.

Heid, I. M., Lamina, C., Küchenhoff, H., Fischer, G., Kloop, N., Kolz, M., Grallert, H., Vollmert, C., Wagner, S., Huth, C., Müller, J., Müller, M., Hunt, S. C., Peters, A., Paulweber, B., Wichmann, H. E., Kronenberg, F. and Illig, T. (2008). Estimating the Single Nucleotide Polymorphism Genotype Misclassification From Routine Double Measurements in a Large Epidemiologic Sample. *American Journal of Epidemiology* **168**, 878-889.

Hue, S., Clewley, J. P., Cane, P. A. and Pillay, D. (2004). HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* **18**, 719-728.

urlhttp://www.inspq.qc.ca/publications/notice.asp?E=p&NumPublication=1706

Kim, J. and Gleser, L. J. (2000). SIMEX Approaches to Measurement Error in ROC Studies. *Communications in Statistics - Theory and Methods* **29(11)**, 2473-2491.

Lewis, F., Gareth J. Hughes, G. J., Rambaut, A., Pozniak, A. and Brown, A. J. L. (2008). Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics **5(3)**: e50. doi:10.1371.

Li, Y. and Lin, X. (2011). Functional Inference in Frailty Measurement Error Models for Clustered Survival Data Using the SIMEX Approach. *Journal of the American Statistical Association* **98**, 191-203.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons, Inc.

http://www.catie.ca/en/fact-sheets/epidemiology/epidemiology-hiv-canada?utm_source=google&utm_medium=cpc&utm_content=en&utm_campaign=wad+hiv+stats#hivstatistics

Shang, Y. (2012). Measurement Error Adjustment Using the SIMEX Method: An Application to Student Growth Percentiles. *Journal of Educational Measurement* **49**, 446-465.

Slate, E. H. and Bandyopadhyay, D. (2009). An investigation of the MC-SIMEX method with application to measurement error in periodontal outcomes. *Statistics in Medicine* **28**, 3523-3538.

http://www.spottestmontreal.com/En/default.aspx

http://www.stat.gouv.qc.ca/statistiques/population-demographie/structure/104.htm

Stefanski, L. A. and Cook, J. (1995). Simulation extrapolation: the measurement error jackknife. *Journal of the American Statistical Association* **90**, 1247-1256.

Wang, N., Lin, X., Gutierrez, R. G. and Carroll, R. J. (1998). Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models. *Journal of American Statistical Association* **93**.

Wenqing, H., Juan, X. and Grace, Y. Y. (2012). SIMEX R Package for Accelerated Failure Time Models with Covariate Measurement Error. *Journal of Statistical Software* **46**, 1-14.