Adjusting for Survey Measurement Error with Accuracy Variables

Damião N. da Silva * Chris Skinner †

Abstract

Many sample surveys gather information on the accuracy of the data collection process. For example, for questions related to salary in face–to–face interviews, the interviewer may record if the interviewee consulted the latest payment slip, an early payment slip or did not consulted any other source. Another example is when the interviewer records if a set of questions were answered very accurately, fairly accurately, not very accurately or not at all accurately. Although these types of data can be informative of the presence of measurement error, little is known how to use the accuracy data to produce corrected estimates of the parameters of interest. In this paper, we propose a methodology that incorporates accuracy variables and allows for adjusting for measurement error. Statistical properties of the estimators are examined through a simulation study. The results demonstrate the adjustments outperform the simple estimator that ignores the measurement error.

Key Words: Sample survey, nonsampling bias, cumulative distribution function, pseudo maximum likelihood

1. Introduction

Measurement error is a potential source of nonsampling bias in many surveys. It occurs when the information to be obtained on one or more variables in the study is mismeasured. This happens, for example, as a result of an imprecise or inaccurate data collection instrument, complexities inherent to the variable being measured, and difficulties to the respondent inform the true response properly. This source of error is potentially a concern for the survey users because, if unaccounted for, it could affect the quality of the data collected and, as a possible consequence, distort the inferences for the parameters of interest.

Basic methods to deal with measurement error usually require auxiliary information about the measurement error parameters, replication studies, validation data, or when instrumental variables (Fuller 1987) are available. Other methods make use of latent variable, structural equation or mixed models (Buonaccorsi 2010, p. 172). In this article, we discuss a methodology to adjust for measurement error that is based on variables that collect data on the accuracy of the responses provided by the observed units in a sample. One example of these variables is when the interviewer, after asking questions related to salary or income, records if the interviewee consulted his last pay slip or not. Another example is when the interviewer judges the responses being provided are very accurately, accurately, or not accurately. These variables, which we term as *accuracy variables*, can be a rich source of information to evidence the possibility of measurement error and to adjust for its impact in the survey estimates.

The methodology considered here was introduced by Da Silva and Skinner (2012) to adjust for measurement error an estimator of the cumulative distribution of a gross pay variable measured in the British Household Panel Survey. The accuracy variable adopted corresponds to the indicator that the respondent latest pay slip was not seen by the interviewer. The results obtained in that application illustrate the potential of the adjusted

*Southampton Statistical Sciences Research Institute, University of Southampton, Highfield Campus, Southampton, Hampshire SO17 1BJ (and Universidade Federal do Rio Grande do Norte, Departamento de Estatística, Campus Universitário s/n, Natal, RN, Brazil 59078-970)

†Department of Statistics, The London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, U.K.

estimator to reduce the upward bias of unadjusted based estimates of the cumulative distribution of weekly pay earnings regarding low paid individuals.

The aim of this paper is to investigate empirically statistical properties of the adjusted estimator of the cumulative distribution function considered by Da Silva and Skinner (2012) and to compare those properties to the ones of an estimator that ignores measurement error. The article is organized as follows: in Section 2, we present a measurement error model for the observed data that incorporates an accuracy indicator variable. In Section 3, we demonstrate how the parameters of the model and the cumulative distribution function of the unobserved variable of interest could be estimated in the sample survey setting using pseudo maximum likelihood estimation. Section 4 shows the results of a simulation experiment that demonstrates the finite-sample properties of the proposed methodology in comparison to the unadjusted estimator. Finally, in Section 5, we summarize the main results of the article.

2. Model and sampling setting

Consider a population of N units, denoted by $U = \{1, 2, ..., N\}$, and let A be a sample from U. Let y_i denote the true value of a variable of interest y for i-th unit, y_i^* be its observed value and $\mathbf{x}_i$ a k dimensional vector of auxiliary variables for the i–th unit. We assume the survey contains an accuracy variable indicator a_i, where a_i has the value one, if y_i^* is observed with error, and the value zero, if the y_i^* is observed accurately. This definition of the accuracy variable suggests the basic model for y_i^* given y_i

$$y_i^* = \begin{cases} y_i + \epsilon_i, & a_i = 1, \\ y_i, & a_i = 0, \end{cases} \tag{1}$$

where ϵ_i represents the measurement error associated with the observation y_i^*.

However, because the accuracy variable can itself be observed with error, then a_i becomes unobservable. In this case, if we let a_i^* denote the observed value of a_i, a better model to account for this situation is: whenever $a_i^* = 1$, means that $a_i = 1$ and, therefore, there is measurement error in the observation y_i^*; if $a_i^* = 0$, a_i can assume the value one with a certain probability p and the value zero with probability $1 - p$. Hence, in this second model, $a_i^* = 0$ considers the possibility of cases with measurement error and cases of accurate responses. This model is illustrated as follows

$$a_i^* = \begin{cases} 1 \quad \Rightarrow \quad a_i = 1 \quad \Rightarrow \quad y_i^* = y_i + \epsilon_i \\ 0 \quad \Rightarrow \quad a_i = \begin{cases} 1 \text{ (with probability } p) \quad \Rightarrow \quad y_i^* = y_i + \epsilon_i \\ 0 \text{ (with probability } 1 - p) \quad \Rightarrow \quad y_i^* = y_i \end{cases} \end{cases} \tag{2}$$

In order to describe a probability model for inference in this setting, we consider a framework consisting of a superpopulation measurement error model assumed to generate the finite population data followed by the selection of a probability sample from this population. Since the basic model (1) is a particular case of (2) when $p = 0$, the superpopulation model will be based on the extended model, for a given p. This parameter shall be provided from an external source or, more likely, chosen via a sensitivity analysis. More precisely, we shall assume

(A1) the complete data for the finite population $\{(y_i, y_i^*, a_i, a_i^*, \mathbf{x}_i^\top) : i = 1, 2..., N\}$ corresponds to the first N elements in a sequence of independent random vectors by which

- $y_i \mid \mathbf{x}_i \overset{indep}{\sim} f(y_i \mid \mathbf{x}_i; \boldsymbol{\gamma})$
- $y_i^* \mid \mathbf{x}_i, y_i, a_i = 1 \overset{indep}{\sim} g(y_i^* \mid \mathbf{x}_i, y_i, a_i = 1; \boldsymbol{\eta})$
- y_i is conditionally independent of a_i given $\mathbf{x}_i$
- y_i^* and y_i are conditionally independent of a_i^* given $\mathbf{x}_i$ and a_i
- $P(a_i = 1 \mid a_i^* = 1, y_i, \mathbf{x}_i) = 1$ and $P(a_i = 1 \mid a_i^* = 0, y_i, \mathbf{x}_i) = p$, where p $(0 \leq p < 1)$ is a fixed value

for all $i = 1, 2, ...$ The unknown parameter of this model is $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\eta})^\top$.

(A2) the observed data corresponds to $\{(y_i^*, a_i^*, \mathbf{x}_i^\top) : i \in A\}$ where A is probability sample of size n that is selected from U by a sample design that gives first and second inclusion probabilities π_i and π_{ij}, respectively. Associated with each $i \in A$, there is a sampling weight w_i taken here to be $w_i = \pi_i^{-1}$. The sampling design is such that the distribution of $y_i \mid y_i^*, a_i^*, \mathbf{x}_i, i \in A$ is the same as the distribution of $y_i \mid y_i^*, a_i^*, \mathbf{x}_i$.

Assumption (A1) gives model conditions for the realization of the finite population. The densities f and g specify, respectively, the conditional distribution of y_i given $\mathbf{x}_i$ and the conditional distribution of y_i^* given $\mathbf{x}_i$ and y_i for the cases that are subject to measurement error. The following two independence assumptions are conditions to identify the model parameters. In the last part of (A1), the probabilities involving the true and observed accuracy indicators regards the extended model structure. Assumption (A2) states the observed data relates to a probability sample that is selected by a type of a noninformative sampling design.

Under the present setting, our goal here is to estimate the cumulative distribution function of y relative to the population U, that is

$$F_N(c) = \frac{1}{N} \sum_{i \in U} I(y_i < c),$$

based on the observed data $\{(y_i^*, a_i^*, \mathbf{x}_i^\top) : i \in A\}$. In the following section, we describe a methodology to estimate $F_N(c)$, which requires the estimation of $\boldsymbol{\psi}$. In what follows, we choose the densities $f(y_i \mid \mathbf{x}_i; \boldsymbol{\gamma})$ and $g(y_i^* \mid \mathbf{x}_i, y_i, a_i = 1; \boldsymbol{\eta})$ in (A1) to correspond to the $\mathrm{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$ and $\mathrm{N}(y_i, \tau^2)$ distributions, respectively. Hence, the model parameter is $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\eta}) = (\boldsymbol{\beta}, \sigma^2, \tau^2)^\top$.

3. Estimation

One simple approach to estimate the finite population cdf $F_N(c)$ is to apply the unadjusted estimator

$$\hat{F}_u(c) = \left[\sum_{i \in A} w_i \right]^{-1} \sum_{i \in A} w_i I(y_i^* < c). \tag{3}$$

However, this estimator is usually biased, as it does not account for the measurement error involved in the observed data $\{y_i^* \mid i \in A\}$. An adjusted estimator that takes into account

this measurement error could be constructed by considering an estimator of the model expectation

$$E\left(\hat{F}_d(c) \mid y_U^*, a_U^*, \mathbf{x}_U\right) \doteq \left[\sum_{i\in A} w_i\right]^{-1} \sum_{i\in A} w_i E\left\{I(y_i < c) \mid y_i^*, a_i^*, \mathbf{x}_i\right\},$$

where y_U^* and a_U^* are the population vectors of values of y_i^* and a_i^* and $\mathbf{x}_U$ ={$\mathbf{x}_1$, $\mathbf{x}_2$,..., $\mathbf{x}_N$}. Because

$$E\left\{I(y_i < c) \mid y_i^*, a_i^*, \mathbf{x}_i\right\} = \begin{cases} (1-p_i(\boldsymbol{\psi}))I(y_i^* < c) + p_i(\boldsymbol{\psi})E\left\{I(y_i < c) \mid \mathbf{x}_i, y_i^*, a_i = 1; \boldsymbol{\psi}\right\}, & a_i^* = 0, \\ E\left\{I(y_i < c) \mid \mathbf{x}_i, y_i^*, a_i = 1; \boldsymbol{\psi}\right\} & a_i^* = 1, \end{cases}$$

$p_i(\boldsymbol{\psi}) \equiv \Pr\big[a_i = 1 \mid y_i^*, a_i^* = 0, \mathbf{x}_i\big]$, and, by (A1),

$$E\left\{I(y_i < c) \mid \mathbf{x}_i, y_i^*, a_i = 1; \boldsymbol{\psi}\right\} = \Phi\left(\frac{c - [(1-\rho)\mathbf{x}_i^\top \boldsymbol{\beta} + \rho y_i^*]}{\sigma\sqrt{1-\rho}}\right) \equiv P_{c,i}(\boldsymbol{\psi}), \tag{4}$$

the resulting adjusted estimator of $F_N(c)$ can be written as

$$\hat{F}_a(c) \;=\; \left[\sum_{i\in A} w_i\right]^{-1} \sum_{i\in A} w_i z_i(\hat{\boldsymbol{\psi}}) \tag{5}$$

where $z_i(\hat{\boldsymbol{\psi}}) = (1-a_i^*)\big[(1-p_i(\hat{\boldsymbol{\psi}}))I(y_i^* < c) + p_i(\hat{\boldsymbol{\psi}})P_{c,i}(\hat{\boldsymbol{\psi}})\big] + a_i^* P_{c,i}(\hat{\boldsymbol{\psi}})$ and $\hat{\boldsymbol{\psi}}$ denotes an estimator of $\boldsymbol{\psi}$.

To estimate the model parameter $\boldsymbol{\psi}$, one could apply the pseudo maximum likelihood method. This approach can be described as follows. The pseudo–score function for $\boldsymbol{\psi}$ based on the observed data $\{(y_i^*, a_i^*, \mathbf{x}_i) \,:\, i \in A\}$ can be defined as

$$S_{obs}(\boldsymbol{\psi}) = \sum_{i\in A} w_i \frac{\partial}{\partial \boldsymbol{\psi}} \ln f(y_i^* \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi}) \equiv \sum_{i\in A} w_i S_{obs,i}(\boldsymbol{\psi}), \tag{6}$$

where

$$S_{obs,i}(\boldsymbol{\psi}) \equiv S_{obs}(\boldsymbol{\psi} \mid y_i^*, a_i^*, \mathbf{x}_i) = \frac{\partial}{\partial \boldsymbol{\psi}} \ln f(y_i^* \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi}).$$

By assuming the derivative with respect to $\boldsymbol{\psi}$ and the integral sign could be interchanged, it follows that the i-th (unweighted) component to the score function can be written as

$$\begin{aligned} S_{obs,i}(\boldsymbol{\psi}) &= \frac{1}{f(y_i^* \mid \mathbf{x}_i, a_i^*; \boldsymbol{\psi})} \frac{\partial}{\partial \boldsymbol{\psi}} \int \sum_{a_i=0}^{1} f(y_i^*, y_i, a_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi}) dy_i \\ &= \int \sum_{a_i=0}^{1} \frac{\left[\frac{\partial}{\partial \boldsymbol{\psi}} f(y_i^*, y_i, a_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi})\right]}{f(y_i^*, y_i, a_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi})} \frac{f(y_i^*, y_i, a_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi})}{f(y_i^* \mid \mathbf{x}_i, a_i^*; \boldsymbol{\psi})} dy_i \\ &= \int \sum_{a_i=0}^{1} \left[\frac{\partial}{\partial \boldsymbol{\psi}} \ln f(y_i^*, y_i, a_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi})\right] f(y_i, a_i \mid y_i^*, a_i^*, \mathbf{x}_i; \boldsymbol{\psi}) dy_i \\ &= E\big[S_{com,i}(\boldsymbol{\psi}) \mid y_i^*, a_i^*, \mathbf{x}_i\big], \end{aligned} \tag{7}$$

where

$$S_{com,i}(\boldsymbol{\psi}) \equiv S_{com}(\boldsymbol{\psi} \mid y_i^*, y_i, a_i^*, a_i, \mathbf{x}_i) = \frac{\partial}{\partial \boldsymbol{\psi}} \ln f(y_i^*, y_i, a_i \mid a_i^*, \mathbf{x}_i; \boldsymbol{\psi})$$

is the score function of $\boldsymbol{\psi}$ based on the complete vector of observations, i.e., (y_i^* , y_i, a_i^*, a_i, $\mathbf{x}_i$), for the i–th unit. Thus, the pseudo maximum likelihood estimator of $\boldsymbol{\psi}$ can be obtained as the solution to

$$S_{obs}(\boldsymbol{\psi}) = \sum_{i \in A} w_i S_{obs,i}(\boldsymbol{\psi}) = \sum_{i \in A} w_i E\big[S_{com,i}(\boldsymbol{\psi}) \mid y_i^*, a_i^*, \mathbf{x}_i\big] = \mathbf{0}, \tag{8}$$

where the expectation is taken with respect to the conditional distribution of y_i and a_i given y_i^*, a_i^* and $\mathbf{x}_i$.

An analytical expression for the pseudo–score equations in (8) can derived by first obtaining an expression for $S_{com,i}(\boldsymbol{\psi})$, defined in (7). By (A1), it can be shown that

$$\begin{aligned} S_{com,i}(\boldsymbol{\psi}) \;\; = \;\; & a_i \left[\frac{\partial}{\partial \boldsymbol{\psi}} \ln f(y_i \mid \mathbf{x}_i; \boldsymbol{\gamma}) + \frac{\partial}{\partial \boldsymbol{\psi}} \ln g(y_i^* \mid y_i, a_i = 1, \mathbf{x}_i; \boldsymbol{\eta}) \right] + \\ & (1 - a_i^*)(1 - a_i) \frac{\partial}{\partial \boldsymbol{\psi}} \ln f(y_i^* \mid \mathbf{x}_i; \boldsymbol{\gamma}). \end{aligned}$$

To evaluate the expectation of $S_{com,i}(\boldsymbol{\psi})$, we replace the corresponding normal densities for f and g and use the consequence of (A1) that

$$y_i \mid y_i^*, a_i = 1, \mathbf{x}_i; \psi \;\sim\; \mathrm{N}\big((1-\rho)\mathbf{x}_i^\top \boldsymbol{\beta} + \rho y_i^*,\, \sigma^2(1-\rho)\big),$$

where $\rho = \sigma^2/(\sigma^2 + \tau^2)$. It follows that the solution $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\tau}^2)$ of the pseudo–score equations (8) can be expressed as

$$\begin{aligned} \hat{\boldsymbol{\beta}} \;\; &= \;\; \left\{ \sum_{i \in A} \mathbf{x}_i w_i \hat{w}_{i,1}^* \mathbf{x}_i^\top \right\}^{-1} \sum_{i \in A} \mathbf{x}_i w_i \hat{w}_{i,1}^* y_i^*, \\ \hat{\sigma}^2 \;\; &= \;\; \frac{\sum_{i \in A} w_i (1 - a_i^*)(1 - p_i(\hat{\boldsymbol{\psi}}))(y_i^* - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2}{\sum_{i \in A} w_i (1 - a_i^*)(1 - p_i(\hat{\boldsymbol{\psi}}))} \\ \hat{\tau}^2 \;\; &= \;\; \frac{\sum_{i \in A} w_i \hat{w}_{i,3}^* (y_i^* - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2}{\sum_{i \in A} w_i \hat{w}_{i,3}^*} - \hat{\sigma}^2 = 0, \end{aligned} \tag{9}$$

where $\hat{w}_{i,1}^* = w_{i,1}^*(\hat{\boldsymbol{\psi}}) = a_i^* + (1 - a_i^*)\big[p_i(\hat{\boldsymbol{\psi}}) + (1 - p_i(\hat{\boldsymbol{\psi}}))/\hat{\rho}\big]$, $\hat{w}_{i,3}^* = w_{i,3}^*(\hat{\boldsymbol{\psi}}) = a_i^* + (1 - a_i^*)p_i(\hat{\boldsymbol{\psi}})$,

$$p_i(\hat{\boldsymbol{\psi}}) = \frac{\frac{p}{\sqrt{\hat{\sigma}^2 + \hat{\tau}^2}} \phi\left(\frac{y_i^* - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2 + \hat{\tau}^2}} \right)}{f(y_i^* \mid a_i^* = 0, \mathbf{x}_i; \hat{\boldsymbol{\psi}})},$$

$$f(y_i^* \mid a_i^* = 0, \mathbf{x}_i; \hat{\boldsymbol{\psi}}) = \frac{p}{\sqrt{\hat{\sigma}^2 + \hat{\tau}^2}} \phi\left(\frac{y_i^* - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2 + \hat{\tau}^2}} \right) + \frac{1 - p}{\sigma} \phi\left(\frac{y_i^* - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right)$$

and $\phi(\cdot)$ denotes the standard normal pdf. As $\hat{\boldsymbol{\beta}} = g_1(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\tau}^2)$, $\hat{\sigma}^2 = g_2(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\tau}^2)$ and $\hat{\tau}^2 = g_3(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\tau}^2)$, numerical estimates of these model parameters can be obtained by an iterative procedure given some preliminary estimates. In each iteration, the procedure could first update the estimates $\hat{\sigma}^2$ and $\hat{\tau}^2$ given the current $\hat{\boldsymbol{\beta}}$. Then, these two new estimates could be used to update the estimate $\hat{\boldsymbol{\beta}}$. These update steps should be repeated until a convergence criterion is achieved.

4. Simulation experiment

We now present a simulation experiment to evaluate and compare statistical properties of the adjusted estimator of the cumulative distribution function (5). We also compare the properties of the pseudo maximum likelihood estimator of the model parameters in the extended measurement error model. The experiment is based on a population U of $N = 10{,}000$ units by which the set of observations $\{(x_i, y_i, a_i^*) : i = 1, ..., N\}$ was generated by taking $x_i \overset{i.i.}{\sim} U(0,1)$, $y_i \overset{indep}{\sim} N(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_i, \sigma^2)$ and $a_i^* \overset{indep}{\sim} B(1, \pi(x_i))$, with $\boldsymbol{\beta}_0 = 50$, $\boldsymbol{\beta}_1 = 100$, $\sigma^2 = 9$ and $\pi(x_i) = 1/[1 + \exp(-4.710531 + 9x_i)]$, for all $i = 1, ..., N$. Given the population vectors of the values y_i and a_i^*, a vector of the same size for the observed y_i^* data was generated with $p = 0$ and $p = 0.3$ according to the extended error model. That is, for each fixed p, the realization of the i-th values of the observed variable of interest is obtained by

$$y_i^{*(p)} = (1 - a_i^*)(1 - a_i^{(p)})y_i + (1 - a_i^*)a_i^{(p)}(y_i + \epsilon_i^{(p)}) + a_i^* a_i^{(p)}(y_i + \epsilon_i^{(p)}),$$

where $a_i^{(p)} = a_i^* + (1 - a_i^*)\tilde{a}_i^{(p)}$, $\tilde{a}_i^{(p)} \overset{indep}{\sim} B(1, p)$, $\epsilon_i^{(p)} \overset{indep}{\sim} \mathrm{N}(0, \tau^2)$ and $\tau^2 = 225$, for all $i = 1, ..., N$.

We selected $5{,}000$ simple random samples without replacement (SRSWOR) of size $n = 200$ from the population U. For each sample, the following estimation methods were applied to estimate the model parameter $\boldsymbol{\psi} = (\beta_0, \beta_1, \sigma^2, \tau^2)^\top$:

- Weighted Least Squares (WLS): estimates of β_0, β_1 and σ^2 correspond to the estimated regression coefficients and residual mean square from the weighted linear regression of $y_i^{*(p)}$ on x_i. The weights taken in the fit of the regression model are $w_i = N/n$ for all observations in each sample.

- Method of moments (MM): based on the theoretical results that under the extended model

  $$\mathrm{E}(y_i^* \mid a_i^*, x_i) = \beta_0 + \beta_1 x_i$$

  and

  $$\mathrm{V}(y_i^* \mid a_i^*, x_i) = a_i^*(\sigma^2 + \tau^2) + (1 - a_i^*)(\sigma^2 + p\tau^2),$$

  then a regression of $y_i^{(*p)}$ on x_i among the cases with $a_i^* = 1$ provides estimated coefficients $\mathbf{b}_1$ and residual mean square s_1^2 and, similarly, a regression of $y_i^{(*b)}$ on x_i among the cases with $a_i^* = 0$ provides estimated coefficients $\mathbf{b}_0$ and residual mean square s_0^2. By equating the conditional variances to the residual mean squares and weighting the two regression coefficients by the inverse of variances, the MM estimates can then be given by

  $$\hat{\sigma}^2 = \max\left\{\frac{s_0^2 - p s_1^2}{1 - p}, 0\right\} \text{ and } \hat{\tau}^2 = \max\left\{\frac{s_1^2 - s_0^2}{1 - p}, 0\right\}.$$

  A combined estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)^\top$ can be given by

  $$\hat{\boldsymbol{\beta}} = \frac{s_0^2}{s_1^2 + s_0^2}\mathbf{b}_1 + \frac{s_1^2}{s_1^2 + s_0^2}\mathbf{b}_0.$$

- PMLE: using the MM estimates as initial values, the MLE estimates solve the pseudo–score equations (8) with $\boldsymbol{\psi} = (\beta_0, \beta_1, \sigma^2, \tau^2)^\top$.

Tables 1 and 2 give the the mean, bias, bias ratio, variance, mean square error and root mean square error of the model parameters estimates for $p = 0$ and $p = 0.3$, respectively. The bias ratio was taken as 100 times the absolute bias divided by the standard deviation of the estimator. The results in Tables 1 and 2 show that the WLS estimates of the regres-

Table 1: Monte Carlo Properties of the Estimators for the Model Parameters under the Basic Measurement Error Model ($p = 0$) Based on $5{,}000$ Replicates.

Method	Parameter	Mean	Bias	Bias ratio	Variance	MSE	$\sqrt{\text{MSE}}$
WLS	β_0	50.24	0.24	11.86	3.93	3.99	2.00
	β_1	99.69	-0.31	11.21	7.44	7.54	2.75
	σ^2	131.60	122.60	620.21	390.78	15422.51	124.19
MM	β_0	50.03	0.03	2.45	1.49	1.49	1.22
	β_1	100.00	0.00	0.08	2.76	2.76	1.66
	σ^2	9.04	0.04	2.96	1.72	1.72	1.31
	τ^2	230.04	5.04	14.83	1156.00	1181.42	34.37
PMLE	β_0	50.06	0.06	5.89	1.18	1.19	1.09
	β_1	99.97	-0.03	1.81	2.23	2.23	1.49
	σ^2	8.87	-0.13	10.14	1.66	1.68	1.29
	τ^2	229.64	4.64	13.80	1129.04	1150.55	33.92

sion coefficients are not much affected by ignoring the measurement error. However, the estimates of σ^2 are highly biased upwards. The absolute biases by this method represent about 620% and 749% of the respective standard deviations when $p = 0$ and 0.3, respectively, demonstrating the danger of not adjusting for the measurement error. The MM and

Table 2: Monte Carlo Properties of the Estimators for the Model Parameters under the Extended Measurement Error Model ($p = 0.3$) Based on $5{,}000$ Replicates.

Method	Parameter	Mean	Bias	Bias ratio	Variance	MSE	$\sqrt{\text{MSE}}$
WLS	β_0	50.16	0.16	8.00	4.11	4.14	2.03
	β_1	99.74	-0.26	8.71	9.22	9.29	3.05
	σ^2	162.34	153.34	749.40	418.66	23930.38	154.69
MM	β_0	50.03	0.03	1.09	8.51	8.51	2.92
	β_1	99.91	-0.09	2.23	17.96	17.97	4.24
	σ^2	20.36	11.36	45.59	621.51	750.67	27.40
	τ^2	223.14	-1.86	3.39	3022.07	3025.53	55.00
PMLE	β_0	50.02	0.02	1.30	1.86	1.86	1.36
	β_1	100.02	0.02	1.27	3.61	3.62	1.90
	σ^2	8.64	-0.36	16.62	4.71	4.84	2.20
	τ^2	227.06	2.06	7.25	804.62	808.84	28.44

PMLE methods yield estimates of the regression coefficients with smaller biases than the WLS method, for both values of p. For the case that $p = 0$, both MM and PMLE methods also estimate well the components of variance σ^2 and τ^2. The MM biases are about 3.0% and 14.8% of their standard deviations and these figures for the PMLE method are roughly 10.1% and 13.8%, respectively. However, when $p = 0.3$, the bias ratios in the estimation

of σ^2 and τ^2 are approximately 45.6% and 3.4%, for the MM, and 16.6% and 7.2%, for the PMLE. These analyses for both values of p indicate that the PMLE estimates of σ^2 and τ^2 have negligible biases while the MM becomes biased for $p = 0.3$. This last finding was also observed with simulated results for $p = 0.1$ and $p = 0.2$ (not shown here), suggesting an inferior performance of the MM in relation to the PMLE method to estimate σ^2 for $p > 0$.

The PMLE method is also more efficient than the other two methods in the estimation of the four model parameters for both values of p. The square root of the mean square error of the PMLE estimates are smaller than the figures corresponding to the other two methods. In the estimation of σ^2 and τ^2 with $p = 0.3$, for example, the square root of the mean square errors of the MM estimates are 27.4 and 55.0 while, for the PMLE, they are 2.2 and 28.4.

Table 3: Monte Carlo Properties of the Estimators for the Cumulative Distribution Function under the Basic Measurement Error Model ($p = 0$) Based on $5{,}000$ Replicates.

		$100F_N(c)$					
Method		3	5	10	90	95	97
True data	Mean	3.01	5.00	9.99	90.04	95.04	97.03
	Bias	0.01	0.00	-0.01	0.04	0.04	0.03
	Bias ratio	0.76	0.14	0.46	1.90	2.86	2.39
	Variance	1.49	2.37	4.39	4.34	2.30	1.39
	MSE	1.49	2.37	4.39	4.34	2.30	1.39
	$\sqrt{\text{MSE}}$	1.22	1.54	2.10	2.08	1.52	1.18
Observed data	Mean	7.27	8.74	12.10	89.55	94.72	96.70
	Bias	4.27	3.74	2.10	-0.45	-0.28	-0.30
	Bias ratio	232.31	185.48	91.84	21.33	17.91	24.41
	Variance	3.37	4.07	5.24	4.51	2.45	1.54
	MSE	21.56	18.07	9.66	4.71	2.52	1.64
	$\sqrt{\text{MSE}}$	4.64	4.25	3.11	2.17	1.59	1.28
PMLE	Mean	3.03	5.01	9.92	90.04	95.04	97.01
	Bias	0.03	0.01	-0.08	0.04	0.04	0.01
	Bias ratio	2.35	0.54	3.66	1.98	2.35	0.65
	Variance	1.49	2.50	4.49	4.32	2.31	1.40
	MSE	1.49	2.50	4.49	4.32	2.31	1.40
	$\sqrt{\text{MSE}}$	1.22	1.58	2.12	2.08	1.52	1.18
MM	Mean	3.07	5.04	9.95	90.04	95.04	97.01
	Bias	0.07	0.04	-0.05	0.04	0.04	0.01
	Bias ratio	5.39	2.60	2.18	1.98	2.32	0.63
	Variance	1.63	2.69	4.68	4.32	2.31	1.40
	MSE	1.64	2.69	4.68	4.32	2.31	1.40
	$\sqrt{\text{MSE}}$	1.28	1.64	2.16	2.08	1.52	1.18

Tables 3 and 4 show the results for the estimation of the cumulative distribution function. The properties on these tables are as in Tables 1 and 2. The estimation methods of $F_N(c)$ considered were:

- True Y: uses the unadjusted estimator (3) with the true data y_i in the place of y_i^*.

- Observed Y: uses the unadjusted estimator (3) based on the observed data $y_i^{*(p)}$, for each p.
- PMLE: uses the adjusted estimator (5) by plugging–in the PMLE estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ and $\hat{\tau}^2$, for each p
- MM: uses the adjusted estimator (5) by plugging–in the MM estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ and $\hat{\tau}^2$, for each p

To have a better view of the lower and upper parts of the distribution of variable y, six different values of c in $F_N(c)$ were chosen as population percentiles. First, note the estimator based on the true data is unfeasible in practice according to our framework. Its inclusion in the experiment is for comparison purposes. As expected, it is approximately unbiased and its precision is not affected by the measurement error. The observed data estimator, on

Table 4: Monte Carlo Properties of the Estimators for the Cumulative Distribution Function under the Extended Measurement Error Model ($p = 0.3$) Based on $5{,}000$ Replicates.

		$100F_N(c)$					
Method		3	5	10	90	95	97
True data	Mean	3.00	5.00	9.98	90.03	95.04	97.03
	Bias	0.00	-0.00	-0.02	0.03	0.04	0.03
	Bias ratio	0.34	0.11	0.74	1.62	2.75	2.17
	Variance	1.48	2.37	4.39	4.34	2.30	1.39
	MSE	1.48	2.37	4.39	4.34	2.30	1.39
	$\sqrt{\text{MSE}}$	1.22	1.54	2.09	2.08	1.52	1.18
Observed data	Mean	7.67	9.12	12.39	88.95	93.76	95.39
	Bias	4.67	4.12	2.39	-1.05	-1.24	-1.61
	Bias ratio	249.77	204.50	105.11	48.54	73.43	109.60
	Variance	3.50	4.06	5.19	4.71	2.86	2.16
	MSE	25.33	21.02	10.92	5.82	4.40	4.75
	$\sqrt{\text{MSE}}$	5.03	4.58	2.77	2.41	2.10	2.18
PMLE	Mean	3.06	5.05	9.96	89.89	95.09	96.94
	Bias	0.06	0.05	-0.04	-0.11	0.09	-0.06
	Bias ratio	4.64	2.96	1.90	5.18	5.82	5.27
	Variance	1.87	3.01	5.06	4.30	2.23	1.40
	MSE	1.88	3.01	5.07	4.31	2.24	1.40
	$\sqrt{\text{MSE}}$	1.37	1.74	2.25	2.08	1.50	1.18
MM	Mean	3.61	5.43	10.04	89.94	94.98	96.83
	Bias	0.61	0.43	0.04	-0.06	-0.02	-0.17
	Bias ratio	25.88	16.05	1.45	2.82	1.29	10.73
	Variance	5.49	7.10	9.17	5.19	3.36	2.59
	MSE	5.85	7.28	9.18	5.19	3.36	2.62
	$\sqrt{\text{MSE}}$	2.42	2.70	3.03	2.28	1.83	1.62

the other hand, is highly biased and inefficient. The means of this estimator regarding the values of $100F_N(c)$, i.e., 3, 5, 10, 90, 95, 97, are about 7.3%, 8.7%, 12.1%, 89.6%, 94.7%, 96.7% ($p = 0$) and 7.7%, 9.1%, 12.4%, 89.0%, 93.8% and 95.4% ($p = 0.3$). Hence, with

the observed data estimator, the true cumulative percentages of observations less than or equal to the specified percentiles, c, are overestimated, when c is in the lower tail of the distribution, and underestimates, for c in the upper tail. The bias ratios of these estimates vary from 17.9% to 232.3% ($p = 0$) and from 48.5% to 249.7% ($p = 0.3$).

The adjusted estimator with PMLE method reduces the bias relative to the unadjusted estimator at both parts of the distribution for both cases of p. In the lower part, the adjustment is downwards and, in the upper part, the PMLE estimates tends to be higher on average than those for the unadjusted method. The same pattern is observed with the MM adjusted estimator, although specially when $p = 0.3$, the MM estimates reflect more bias than those based on pseudo maximum likelihood. The variation in the PMLE bias ratios is from 0.5% to 3.7% ($p = 0$) and from 1.9% to 5.8% ($p = 0.3$). The variations for the MM bias ratios are 0.6% to 5.4% ($p = 0$) and 1.3% to 25.9% ($p = 0.3$). In terms of efficiency, the PMLE adjusted estimator is the method with the closest root mean square errors to the ones obtained by the true data method. The MM method has its better efficiency when $p = 0$ than when $p = 0.3$. However, even the former case, the PMLE is more efficient than the MM for estimation in the lower tail.

5. Discussion

In this article, we have studied properties of an adjusted estimator for a cumulative distribution function in the presence of measurement error. The formulation of the estimator, proposed initially by Da Silva and Skinner (2012), depends on the parameters of a model for the variable of interest that allows for the presence of measurement error through the use of one accuracy variable. The estimator of the cumulative distribution function is then defined by replacing the model parameter with a suitable estimate, which was obtained here by a pseudo maximum likelihood estimation method.

A simulation experiment was carried to evaluate the properties of the estimators with respect to the variability of the extended model, described in Section 2, and the sampling design used to select the units to belong to the sample. The parameter p of the extended model took the values 0 and 0.3. The model parameters consisted of two regression coefficients ($\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$), the variance of true variable y around the regression line (σ^2) and the variance of the measurement error (τ^2). The following three methods of estimation for these parameters were explored in the experiment: Weighted Least Square method (WLS), Method of Moments (MM) and Pseudo Maximum Likelihood Estimation (PMLE). The WLS method was considered as a naive approach to estimate the regression coefficients $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and the variance σ^2. The results can be summarized as follows:

- WLS is an unreliable method to be considered in in practice, when there is measurement error. This estimator badly overestimates the variance σ^2 and is also quite inefficient under both values of p.
- PMLE and MM outperform the WLS method by yielding smaller bias ratios and smaller root mean squares errors.
- MM works better under the basic model ($p = 0$) than when it is applied under the extended model with $p > 0$. PMLE works better than MM for both values of p and the gains of the former are greater when $p > 0$.

For the estimation of the cumulative distribution function $F_N(c)$, we compared the properties of an adjusted estimator when the true data and when the observed data of the variable of interest are used. The former was considered for comparison purposes. We also

considered the adjusted–based estimator (5) formed with the replacement of the PMLE and MM estimates. In summary,

- the unadjusted estimates with the observed data have greater biases than when the same estimator is applied for the true (unobserved) values of the variable of interest.
- both adjusted methods (MM and PMLE) decrease, on average, the unadjusted observed data estimates estimates in the lower tail of the distribution. In the upper tail, the adjustment works as an increment to the unadjusted estimates.
- the PMLE method is more successful in achieving greater bias reductions and more efficiency than the MM method, specially when $p = 0.3$.

These findings demonstrate that the pseudo maximum likelihood estimator of the model parameters and the resulting adjusted estimator of cumulative distribution function can be successfully employed in reducing the measurement error bias of the unadjusted estimator. The method of moments, a simple method of estimating variance components, should not be used for situations where there is the possibility of measurement error among the cases belonging to the more accurate group of the accuracy variable.

REFERENCES

Buonaccorsi, J. P. (2010), *Measurement Error: Models, Methods and Applications*, New York: Chapman and Hall/CRC.

Da Silva, D. N., and Skinner, C. (2012), "The Use of Accuracy Indicators to Correct for Survey Measurement Error," *Working paper*.

Fuller, W. A. (1987), *Measurement Error Models*, New York: John Wiley & Sons.