Statistics, Vol. 00, No. 0, Month 2009, 1–9

Taylor & Francis
Taylor & Francis Group

Measurement error models with a general class of error distribution

Alexandre Galvão Patriota and Heleno Bolfarine*

Institute of Mathematics and Statistics, University of São Paulo, Brazil

(*Received 21 October 2007; final version received 30 March 2009*)

In general, the normal distribution is assumed for the surrogate of the true covariates in the classical measurement error model. This paper considers a class of distributions, which includes the normal one, for the variables subject to error. An estimation approach yielding consistent estimators is developed and simulation studies reported.

Keywords: multiple regression; measurement error; corrected score; asymptotic theory; simulation

1. Introduction

The ordinary maximum likelihood (ML) approach in classical regression models, fails when the independent variables are subject to error. The most noticeable and well known problem reported in the literature is the inconsistency of the ML estimators [1]. To solve this problem, a number of alternatives were proposed. The measurement error model (MEM) is the most fashionable of them, but it has some limitations. It is necessary to know some parameters to avoid inconsistencies resulting from unbounded likelihoods (functional version) and non-identifiability (structural version). For more details see, for example [2] and references therein.

Typically [1], it is assumed that the errors are normally distributed, but the normal assumption is not always tenable. Actually, this is a strong assumption that cannot always be satisfied in practice. There are situations where the observed covariate is positive, so that its distribution may not be appropriately approximated by a normal distribution.

The most important contributions of this article are to introduce a multiple regression model in which some covariates subject to error are not necessarily normally distributed and to propose a method for obtaining consistent estimators for all parameters of the model, based on the corrected score approach. The method is computationally simple and can be implemented with any statistical package. It extends the results in [3] that studied an MEM where the surrogate for the unobservable true covariate is the event count per unit time. They regarded the Poisson distribution for the surrogate variable, the rate of which is the unobserved true covariate. The authors justified the proposed model with a medical example.

*Corresponding author. Email: hbolfar@ime.usp.br

ISSN 0233-1888 print/ISSN 1029-4910 online

DOI: 10.1080/02331880903023795
http://www.informaworld.com

The approach considered in this paper assumes certain moment conditions for the surrogate variables which are satisfied by distributions other than the normal, like the gamma, Poisson [3], uniform, Rayleigh, among others. The model structure makes it possible to use the corrected score approach as considered in [4,5] yielding consistent and asymptotically normal estimators which can be used for defining Wald type statistics for hypothesis testing.

The paper is organized as follows. Section 2 defines the model with general distribution for the surrogates and deals with the estimation process for the general setup described above. Examples involving normal, continuous uniform, gamma, Poisson and discrete uniform distributions for the surrogate variable X_i are given in Section 2.1, where consistent estimators are provided in each case. Section 3 studies the asymptotic theory regarding the estimators obtained in Section 2. Section 4 presents simulations studies. Section 5 concludes the paper with final discussions and comments.

2. The general MEM and the corrected score approach

Assume that one observes the triplet $(Y_i, \boldsymbol{X}_i^\top, \boldsymbol{W}_i^\top)$ for each individual $i = 1, \ldots, n$, where Y_i is the response variable, $\boldsymbol{X}_i$ (with dimensions $p \times 1$) and $\boldsymbol{W}_i$ (with dimensions $q \times 1$) are independent vectors with and without measurement error, respectively. That is, $\boldsymbol{X}_i$ is a random vector and it is observed instead of $\boldsymbol{x}_i$ (the true unobservable covariate). The $\boldsymbol{W}_i$ vector may contain indicator variables, e.g., gender, treatment received or continuous variables as age, weight of the ith individual. The following linear relationship among $Y_i, \boldsymbol{W}_i$ and $\boldsymbol{x}_i$ is considered:

$$\begin{aligned} Y_i &= \boldsymbol{\beta}^\top \boldsymbol{W}_i + \boldsymbol{\gamma}^\top \boldsymbol{x}_i + e_i, \\ \boldsymbol{X}_i &\sim \mathcal{G} \in \mathcal{C}(\boldsymbol{x}_i, g_1, g_2), \end{aligned} \tag{1}$$

where $\boldsymbol{W}_i$ and $\boldsymbol{x}_i$ are non-stochastic vectors, $\mathcal{C}(\boldsymbol{x}_i, g_1, g_2)$ is a class of distributions and the functions $g_1(\cdot)$ and $g_2(\cdot)$ satisfy

$$\mathbb{E}[g_1(\boldsymbol{X}_i)] = \boldsymbol{x}_i, \quad \text{and} \quad \mathbb{E}[g_2(\boldsymbol{X}_i)] = \boldsymbol{x}_i \boldsymbol{x}_i^\top \tag{2}$$

where $\boldsymbol{x}_i$ is the true unobservable ($p \times 1$) vector. We also assume that $e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and Y_i is independent of $\boldsymbol{X}_i$. Notice that, the assumptions made for the model (1) are not much restrictive and the functions g_1 and g_2 do not specify any specific distribution, but a family of distributions. Moreover, the functions (2) are sufficient to estimate the model parameters and their covariance matrix using the corrected score approach. To the highest degree of generality we can consider in implementing our approach, we need only to specify the form of the functions $g_1(\cdot)$ and $g_2(\cdot)$ defined in (2) above. Clearly, when a distribution is specified for $\boldsymbol{X}_i$ then the functions $g_k, k = 1, 2$ are naturally specified. For example, in the normal case with $\text{Var}(\boldsymbol{X}_i) = \boldsymbol{\Sigma}$ known and $\mathbb{E}(\boldsymbol{X}_i) = \boldsymbol{x}_i$, we have that $g_1(\boldsymbol{X}_i) = \boldsymbol{X}_i$ and $g_2(\boldsymbol{X}_i) = \boldsymbol{X}_i \boldsymbol{X}_i^\top - \boldsymbol{\Sigma}$. Another interesting example is when the vector $\boldsymbol{X}_i$ has components which follow different distributions, e.g., suppose that $\boldsymbol{X}_i = (X_{1i}, X_{2i})^\top$ such that $X_{1i} \sim \mathcal{N}(x_{1i}, \phi_i)$, $X_{2i} \sim \text{Poisson}(x_{2i})$ and $\text{Cov}(X_{1i}, X_{2i}) = a_i$, where a_i and ϕ_i are known for all $i = 1, \ldots, n$. Then,

$$g_1(\boldsymbol{X}_i) = \boldsymbol{X}_i, \quad g_2(\boldsymbol{X}_i) = \begin{bmatrix} X_{1i}^2 - \phi_i & X_{1i}X_{2i} - a_i \\ X_{1i}X_{2i} - a_i & X_{2i}^2 - X_{2i} \end{bmatrix}.$$

As we can see, the function g_2 may depend on extra parameters which have to be known *a priori* through other studies or extra data (such as replications), because the proposed model does not

allow estimating those extra parameters. However, we must remark that there are many distributions which do not need extra parameters, such as Poisson, exponential, uniform (continuous and discrete), Rayleigh and others.

We assume the functional model, namely no distribution is assumed for the unknown $\boldsymbol{x}_i$ which is considered then an incidental parameter (the structural version, however, can be assumed as well). We define the parameter vector as $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \sigma^2)^\top$. One of the goals is to estimate $\boldsymbol{\theta}$ consistently and make inferences about $(\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$.

One way to deal with the measurement error problem is to replace $\boldsymbol{x}_i$ with $\boldsymbol{X}_i$ in model (1) and then maximize the associated log-likelihood function with respect to $\boldsymbol{\theta}$. This procedure is known as the 'naive' estimation process and $\boldsymbol{X}_i$ being measured with error, the resulting ML estimators are asymptotically biased [1]. In fact, the more inaccurate is the value of $\boldsymbol{X}_i$, the more distorted will be the estimator of $\boldsymbol{\gamma}$ and this can lead to other statistical imprecisions, specially when dealing with confidence intervals and hypothesis testing. To overcome such difficulties, we embrace the corrected score approach to estimate the parameters of the model (1). Nakamura [4] proposed to correct the naive log-likelihood function $\ell(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{X})$ such that its expectation given the response variable is the unobserved log-likelihood function. That is, it requires finding a function $\ell^+(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{X})$ such that

$$\mathbb{E}[\ell^+(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{X})|\boldsymbol{Y}] = \ell(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{x}), \tag{3}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{W} = (\boldsymbol{W}_1^\top, \ldots, \boldsymbol{W}_n^\top)^\top$, $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$ and $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$. If differentiation and integration are interchangeable in (3), then the score function produced by the corrected log-likelihood function, $U^+(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{X}) = \sum_i U^+(\boldsymbol{\theta}, Y_i, \boldsymbol{W}_i, \boldsymbol{X}_i)$ (we write $U^+(\boldsymbol{\theta}, \boldsymbol{X})$ in short) will be unbiased. Thus, under the regularity conditions stated in [6], the estimator $\tilde{\boldsymbol{\theta}}$ such that $U^+(\tilde{\boldsymbol{\theta}}, \boldsymbol{X}) = \boldsymbol{0}$ will be consistent. Following these ideas, the naive log-likelihood function is

$$\ell(\boldsymbol{\theta}) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{\sigma^2}\sum_{i=1}^{n}\{(Y_i - \boldsymbol{\beta}^\top\boldsymbol{W}_i)^2 - 2(Y_i - \boldsymbol{\beta}^\top\boldsymbol{W}_i)\boldsymbol{X}_i^\top\boldsymbol{\gamma} + \boldsymbol{\gamma}^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\gamma}\}.$$

After some algebraic manipulations, it follows that the corrected log-likelihood function is given by

$$\ell^+(\boldsymbol{\theta}) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{\sigma^2}\sum_{i=1}^{n}\{(Y_i - \boldsymbol{\beta}^\top\boldsymbol{W}_i)^2 - 2(Y_i - \boldsymbol{\beta}^\top\boldsymbol{W}_i)g_1(\boldsymbol{X}_i)^\top\boldsymbol{\gamma} + \boldsymbol{\gamma}^\top g_2(\boldsymbol{X}_i)\boldsymbol{\gamma}\}$$

and its derivatives are given by

$$\boldsymbol{U}_1^+(\boldsymbol{\theta}, \boldsymbol{X}) = \frac{\partial\ell^+(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\boldsymbol{U}_{i1}^+(\boldsymbol{\theta}, \boldsymbol{X}) \tag{4}$$

$$\boldsymbol{U}_2^+(\boldsymbol{\theta}, \boldsymbol{X}) = \frac{\partial\ell^+(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}} = \sum_{i=1}^{n}\boldsymbol{U}_{i2}^+(\boldsymbol{\theta}, \boldsymbol{X}) \tag{5}$$

and

$$\boldsymbol{U}_3^+(\boldsymbol{\theta}, \boldsymbol{X}) = \frac{\partial\ell^+(\boldsymbol{\theta})}{\partial\sigma^2} = \sum_{i=1}^{n}\boldsymbol{U}_{i3}^+(\boldsymbol{\theta}, \boldsymbol{X}) \tag{6}$$

where $\boldsymbol{U}_{i1}^+(\boldsymbol{\theta}, \boldsymbol{X}) = 1/\sigma^2\{Y_i - \boldsymbol{\beta}^\top\boldsymbol{W}_i - \boldsymbol{\gamma}^\top g_1(\boldsymbol{X}_i)\}\boldsymbol{W}_i$, $\boldsymbol{U}_{i2}^+(\boldsymbol{\theta}, \boldsymbol{X}) = 1/\sigma^2\{(Y_i - \boldsymbol{\beta}^\top\boldsymbol{W}_i)g_1(\boldsymbol{X}_i) - g_2(\boldsymbol{X}_i)\boldsymbol{\gamma}\}$ and $\boldsymbol{U}_{i3}^+(\boldsymbol{\theta}, \boldsymbol{X}) = -1/(2\sigma^2) + 1/(2\sigma^4)\{(Y_i - \boldsymbol{\beta}^\top\boldsymbol{W}_i)^2 - 2(Y_i - \boldsymbol{\beta}^\top\boldsymbol{W}_i)g_1(\boldsymbol{X}_i)^\top\boldsymbol{\gamma} + \boldsymbol{\gamma}^\top g_2(\boldsymbol{X}_i)\boldsymbol{\gamma}\}$.

It is easy to see that the expectation of these derivatives are equal to zero. Moreover, their expectations given $\boldsymbol{Y}$ are equal to the unobserved score functions. The corrected score estimators are obtained equating (4)–(6) to zero and solving the resulting equations. Thus, they are given by

$$\hat{\boldsymbol{\beta}}_n = \left(\sum_{i=1}^n \boldsymbol{W}_i \boldsymbol{W}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{W}_i [Y_i - \boldsymbol{\gamma}^\top g_1(\boldsymbol{X}_i)], \tag{7}$$

$$\hat{\boldsymbol{\gamma}}_n = \boldsymbol{H}_n^{-1} \left[\sum_{i=1}^n g_1(\boldsymbol{X}_i) Y_i - \sum_{i=1}^n g_1(\boldsymbol{X}_i) \boldsymbol{W}_i^\top \left(\sum_{i=1}^n \boldsymbol{W}_i \boldsymbol{W}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{W}_i Y_i\right] \tag{8}$$

and

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n \{(Y_i - \boldsymbol{\beta}^\top \boldsymbol{W}_i)^2 - 2(Y_i - \boldsymbol{\beta}^\top \boldsymbol{W}_i) g_1(\boldsymbol{X}_i)^\top \boldsymbol{\gamma} + \boldsymbol{\gamma}^\top g_2(\boldsymbol{X}_i) \boldsymbol{\gamma}\}, \tag{9}$$

where $\boldsymbol{H}_n = \sum_i g_2(\boldsymbol{X}_i)^\top - \sum_i g_1(\boldsymbol{X}_i) \boldsymbol{W}_i^\top (\sum_i \boldsymbol{W}_i \boldsymbol{W}_i^\top)^{-1} \sum_i \boldsymbol{W}_i g_1(\boldsymbol{X}_i)^\top$. As we can see, the estimators (7)–(9) have analytical solutions thus not requiring iterative procedures. Section 3 presents general regularity conditions by which the estimators (7)–(9) are consistent. We show some examples in the next subsection to illustrate the usefulness of our approach.

2.1. *Some special cases*

In the examples below, we consider the simple regression model when there are only independent variables subject to error. Assume also that $\boldsymbol{W}_i = 1$, $\boldsymbol{\beta} = \beta$, $\boldsymbol{X}_i = X_i$ and $\boldsymbol{\gamma} = \gamma$ are all scalars. The model (1) reduces to $Y_i = \beta + \gamma x_i + e_i$, where $e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$.

Example 2.1 (Normal distribution) Suppose that $X_i \sim \mathcal{N}(x_i, \phi_i^2)$ with known ϕ_i for all $i = 1, \ldots, n$. Notice that $g_1(X_i) = X_i$ and $g_2(X_i) = X_i^2 - \phi_i^2$, $i = 1, \ldots, n$. Hence, the consistent estimators for β, γ and σ^2 are given by

$$\hat{\beta}_n = \bar{Y} - \hat{\gamma}_n \bar{X}, \quad \hat{\gamma}_n = \frac{\sum_{i=1}^n Y_i X_i - n\bar{Y}\bar{X}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2 - \phi_i^2} \text{ and } \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n \left\{(Y_i - \hat{\beta}_n - \hat{\gamma}_n X_i)^2 - \phi_i^2 \hat{\gamma}_n^2\right\}.$$

Example 2.2 (Uniform distribution) Suppose that $X_i \sim \mathrm{U}(0, x_i)$. Notice that $g_1(X_i) = 2X_i$ and $g_2(X_i) = 3X_i^2$. Then, the consistent estimators for β, γ and σ^2 are given by

$$\hat{\beta}_n = \bar{Y} - 2\hat{\gamma}_n \bar{X}, \quad \hat{\gamma}_n = \frac{\sum_{i=1}^n Y_i X_i - n\bar{Y}\bar{X}}{1.5 \sum_{i=1}^n X_i^2 - 2n\bar{X}^2} \text{ and } \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n \left\{(Y_i - \hat{\beta}_n - \hat{\gamma}_n X_i)^2 - \hat{\gamma}_n^2 X_i^2\right\}.$$

Example 2.3 (Gamma distribution) Suppose that $X_i \sim \mathrm{G}(x_i, \phi)$ with ϕ known, $\mathbb{E}(X_i) = x_i > 0$ and $\mathrm{Var}(X_i) = \phi x_i^2$. Notice that $g_1(X_i) = X_i$ and $g_2(X_i) = X_i^2/(\phi + 1)$ and when $\phi = 1$ it becomes the exponential distribution. Then, the consistent estimators for β, γ and σ^2 are given by

$$\hat{\beta}_n = \bar{Y} - \hat{\gamma}_n \bar{X}, \quad \hat{\gamma}_n = \frac{\sum_{i=1}^n Y_i X_i - n\bar{Y}\bar{X}}{(1+\phi)^{-1} \sum_{i=1}^n X_i^2 - n\bar{X}^2}$$

and

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n \left\{(Y_i - \hat{\beta}_n - \hat{\gamma}_n X_i)^2 - \frac{\phi}{1+\phi} \hat{\gamma}_n^2 X_i^2\right\}.$$

Example 2.4 (Poisson distribution; [3]) Suppose that $X_i \sim \mathrm{P}(x_i)$. Notice that $g_1(X_i) = X_i$ and $g_2(X_i) = X_i^2 - X_i$. Then, the consistent estimators for β, γ and σ^2 are as follows

$$\hat{\beta}_n = \bar{Y} - \hat{\gamma}_n \bar{X}, \quad \hat{\gamma}_n = \frac{\sum_{i=1}^n Y_i X_i - n\bar{Y}\bar{X}}{\sum_{i=1}^n X_i^2 - n\bar{X}(1+\bar{X})} \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \{(Y_i - \hat{\beta}_n - \hat{\gamma}_n X_i)^2 - \hat{\gamma}_n^2 X_i\}.$$

Example 2.5 (Discrete uniform distribution) Suppose that $X_i \sim \mathrm{U}\{0, \ldots, x_i\}$. Notice that $g_1(X_i) = 2X_i$ and $g_2(X_i) = 3X_i^2 - X_i$. Then, the consistent estimators for β, γ and σ^2 are as follows

$$\hat{\beta}_n = \bar{Y} - 2\hat{\gamma}_n \bar{X}, \quad \hat{\gamma}_n = \frac{\sum_{i=1}^n Y_i X_i - n\bar{Y}\bar{X}}{1.5\sum_{i=1}^n X_i^2 - n\bar{X}(0.5+2\bar{X})}$$

and

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \{(Y_i - \hat{\beta}_n - \hat{\gamma}_n X_i)^2 - \hat{\gamma}_n^2 X_i^2\}.$$

Many other distributions may be assigned to X_i. The Rayleigh distribution, for example, has mean $x_i\sqrt{\pi/2}$ and variance $x_i^2(4-\pi)/2$, therefore the functions that are used to estimate consistently the parameters are given by $g_1(X_i) = X_i\sqrt{2/\pi}$ and $g_2(X_i) = (1/2)X_i^2$.

In general, the distribution of X_i may not be specified. It suffices to provide the functions $g_1(\cdot)$ and $g_2(\cdot)$. Notice that if we take $g_1(X_i) = X_i$ and $g_2(X_i) = X_i^2 - \phi^2$, then there are many distributions for X_i that comply with these conditions. For example, if X_i has normal (with mean x_i and variance ϕ^2) or logistic (with mean x_i and variance $\phi^2 = (\pi^2/3)s^2$, where $s > 0$ is the scale parameter known) distributions then the functions $g_1(\cdot)$ and $g_2(\cdot)$ are the same for both cases. Another useful example is the gamma and the normal multiplicative models, namely $X_i \sim \mathcal{N}(x_i, x_i^2\phi)$ and $X_i \sim \mathrm{G}(x_i, \phi)$ generate the same functions $g_1(\cdot)$ and $g_2(\cdot)$. Therefore, this paper considers the distribution of X_i that lies in the family $\mathcal{C}\{x_i, g_1, g_2\}$ which includes all distributions that generate the same functions g_k, $k = 1, 2$.

3. Large sample results

Define $\boldsymbol{U}_i^+(\boldsymbol{\theta}) = (\boldsymbol{U}_{i1}^+(\boldsymbol{\theta}, \boldsymbol{X})^\top, \boldsymbol{U}_{i2}^+(\boldsymbol{\theta}, \boldsymbol{X})^\top, U_{i3}^+(\boldsymbol{\theta}, \boldsymbol{X}))^\top$,

$$\boldsymbol{I}_n^+(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial \boldsymbol{U}_i^+(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \text{and} \quad \bar{\boldsymbol{\Pi}}_n^+(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{U}_i^+(\boldsymbol{\theta})\boldsymbol{U}_i^+(\boldsymbol{\theta})^\top. \tag{10}$$

Further, let

$$\bar{\boldsymbol{\Lambda}}_n^+(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{I}_n^+(\boldsymbol{\theta})] \quad \text{and} \quad \bar{\boldsymbol{\Gamma}}_n^+(\boldsymbol{\theta}) = \mathbb{E}[\bar{\boldsymbol{\Pi}}_n^+(\boldsymbol{\theta})]. \tag{11}$$

We consider the valid regularity conditions stated in [6] regarding the estimating equations (4)–(6). As a result [2], the estimator $\hat{\boldsymbol{\theta}}_n$ is consistent and $n^{1/2}\boldsymbol{L}_n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ converges in distribution to a standard multivariate normal distribution, where $\boldsymbol{L}_n^{1/2}(\boldsymbol{\theta}) = \bar{\boldsymbol{\Gamma}}_n^+(\boldsymbol{\theta})^{-1/2}\bar{\boldsymbol{\Lambda}}_n^+(\boldsymbol{\theta})$ (see also [5] for regularity conditions regarding the asymptotic behaviour of the matrices (10) and (11)).

Therefore, in our proposed model we must compute the matrices $\bar{\boldsymbol{\Lambda}}_n^+(\boldsymbol{\theta})$ and $\bar{\boldsymbol{\Gamma}}_n^+(\boldsymbol{\theta})$ to find a consistent estimator for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_n$. It can be shown that

$$\bar{\boldsymbol{\Lambda}}_n^+(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\boldsymbol{I}_i^+(\boldsymbol{\theta})] = \frac{1}{n\sigma^2}\begin{bmatrix} \sum_i \boldsymbol{W}_i\boldsymbol{W}_i^\top & \sum_i \boldsymbol{W}_i\boldsymbol{x}_i^\top & \boldsymbol{0}_q \\ \sum_i \boldsymbol{x}_i\boldsymbol{W}_i^\top & \sum_i \boldsymbol{x}_i\boldsymbol{x}_i^\top & \boldsymbol{0}_p \\ \boldsymbol{0}_q^\top & \boldsymbol{0}_p^\top & \frac{n}{2\sigma^2} \end{bmatrix}.$$

It also follows that $\bar{\boldsymbol{\Gamma}}_n^+(\boldsymbol{\theta})$ is consistently estimated by $\bar{\boldsymbol{\Pi}}_n^+(\hat{\boldsymbol{\theta}}_n)$ and then we have that a consistent estimator for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_n$ is given by

$$\widehat{\mathrm{Var}(\hat{\boldsymbol{\theta}}_n)} = \frac{1}{n}\bar{\boldsymbol{\Lambda}}_n^+(\hat{\boldsymbol{\theta}}_n)^{-1}\bar{\boldsymbol{\Pi}}_n^+(\hat{\boldsymbol{\theta}}_n)\bar{\boldsymbol{\Lambda}}_n^+(\hat{\boldsymbol{\theta}}_n)^{-\top},$$

which can be implemented using some statistical software, such as R Development Core Team [7].

Thus, we can test the hypothesis H_0: $\boldsymbol{G}\boldsymbol{\theta}_1 = \boldsymbol{d}$, where $\boldsymbol{G}$ is a specified matrix, $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ and $\boldsymbol{d}$ is a vector with appropriate dimensions. The Wald type statistics for testing this hypothesis is given by

$$W_n = (\boldsymbol{G}\hat{\boldsymbol{\theta}}_{1n} - \boldsymbol{d})^\top[\boldsymbol{G}\boldsymbol{P}_n(\hat{\boldsymbol{\theta}}_n)\boldsymbol{G}^\top]^{-1}(\boldsymbol{G}\hat{\boldsymbol{\theta}}_{1n} - \boldsymbol{d}), \tag{12}$$

where $\boldsymbol{P}_n(\hat{\boldsymbol{\theta}}_n) = [\boldsymbol{I}_{c\times c}, \boldsymbol{0}_{c\times 1}]\widehat{\mathrm{Var}(\hat{\boldsymbol{\theta}}_n)}[\boldsymbol{I}_{c\times c}, \boldsymbol{0}_{c\times 1}]^\top$ and $c = p + q$. Here, W_n converges in distribution to a χ^2 distribution with r degrees of freedom, where r is the rank of $\boldsymbol{G}$. As shown in [8], the likelihood ratio statistics based on the corrected likelihood function $l^+(.)$ does not follow central chisquare distribution.

4. Simulation study

To assess the performance of the proposed model and methods, we conducted two simulation studies using the statistical package R Development Core Team [7]. In both simulations, we considered the multiple linear regression model $Y_i = \beta_1 + \beta_2 T_i + \gamma x_i + e_i$, where T_i represents the treatment (taking value one, if the ith experimental unit received the treatment and zero otherwise). The generation was performed considering that the first $n/2$ individuals received the treatment and the remainder $n/2$ the innocuous (placebo) substance. Values for the other parameters were taken as: $\beta_1 = 1$, $\beta_2 = 2$, $\gamma = 1$ and $\mathrm{Var}(e_i) = 5$. We took $x_i \sim \mathrm{Uniform}\{1, \ldots, 10\}$, $i = 1, \ldots, n$. For each of the two scenarios described next, 25,000 samples of sizes n =50, 100 and 200 were generated. They are: (i) $X_i \sim \mathrm{G}(x_i, 0.01)$, that is, X_i has Gamma distribution with $\mathbb{E}(X_i) = x_i$ and $\mathrm{Var}(X_i) = 0.01x_i^2$ and (ii) $X_i \sim \mathrm{P}(x_i)$, that is, X_i has Poisson distribution with $\mathbb{E}(X_i) = x_i$ and $\mathrm{Var}(X_i) = x_i$. For each generated sample, we compute the estimates using the proposed model (assuming known the functions g_1 and g_2 of X_i for each scenario). Table 1 shows the empirical median square error (MSE) and the bias for the estimators in the scenarios (i) and (ii), respectively. That is, let $\delta_{jn} = \hat{\theta}_n^{(j)} - \theta$, where $\hat{\theta}_n^{(j)}$ is the estimate of θ in the jth Monte Carlo simulation when the sample size is n. The MSE is the median of $\{\delta_{1n}^2, \ldots, \delta_{Nn}^2\}$ and the bias is the median of $\{\delta_{1n}, \ldots, \delta_{Nn}\}$. The median was used because it is more robust against discrepant points. Although the asymptotic distribution of $\sqrt{n}\delta_{ni}$ is normal (when n goes to infinity), it can be asymmetric, elliptical or bi-modal for small values of n. We study the behaviour of the estimators (7)–(9) for samples sizes 50, 100 and 200, as reported above.

Table 1. Bias and MSE for parameter estimators under Gamma and Poisson distributions as in scenarios (i) and (ii).

	Sample size	Rates of rejection (%)	Gamma distribution β_1	β_2	γ_1	σ^2
MSE	$n = 50$	0.58	0.6749	0.3428	0.0188	1.6996
bias			−0.2056	0.0450	0.0360	−0.7091
MSE	$n = 100$	0	0.2593	0.1470	0.0103	0.7130
bias			−0.1023	−0.0004	0.0207	−0.3618
MSE	$n = 200$	0	0.1576	0.0842	0.0045	0.4280
bias			−0.0570	0.0009	0.0099	−0.1954
			Poisson distribution			
			β_1	β_2	γ_1	σ^2
MSE	$n = 50$	3.67	0.7792	0.4069	0.0217	2.2102
bias			−0.1877	0.0310	0.0313	−0.7669
MSE	$n = 100$	0.32	0.4037	0.1912	0.0123	1.1485
bias			−0.1351	0.0095	0.0245	−0.4648
MSE	$n = 200$	0.01	0.2221	0.0970	0.0066	0.5713
bias			−0.0716	−0.0075	0.0128	−0.2118

Note: Rejection rates are the percentage of negative estimates $\hat{\sigma}_n^2$.

Table 2. Rejection rates for the hypothesis H_0: $\gamma = \gamma_0$ (at the 5% nominal level) using the Wald type statistics (12) when $n = 50$, 100 and 200 under scenarios (i) and (ii) and under least squares (ordinary model).

		Gamma		Poisson	
		Proposed model	Ordinary model	Proposed model	Ordinary model
$n = 50$					
γ_0	−3	5.70	<0.01	6.00	2.43
	−2	5.08	0.01	6.02	1.56
	−1	4.26	0.01	4.83	0.29
	1	4.50	0.04	4.74	0.24
	2	5.16	<0.01	5.46	1.36
	3	5.31	0.01	5.69	2.24
$n = 100$					
γ_0	−3	5.44	<0.01	5.83	1.32
	−2	5.03	<0.01	5.62	0.63
	−1	4.64	0.02	4.77	0.01
	1	4.86	0.02	4.80	0.03
	2	4.96	<0.01	5.50	0.60
	3	4.98	<0.01	6.06	1.16
$n = 200$					
γ_0	−3	5.16	<0.01	5.88	0.51
	−2	5.03	<0.01	5.30	0.10
	−1	4.71	0.01	4.73	<0.01
	1	4.91	<0.01	4.74	<0.01
	2	4.75	<0.01	5.22	0.14
	3	5.44	<0.01	5.95	0.48

The estimator $\hat{\sigma}_n^2$ obtained using (9) returns negative values for some samples. All such samples were eliminated. Table 1 brings also the percentage of such samples for each sample size. As can be depicted from Table 1, the estimator (9) goes to a positive value closer to σ^2. We also conducted the simulations with other combinations of parameters. They seem to agree with the conclusions

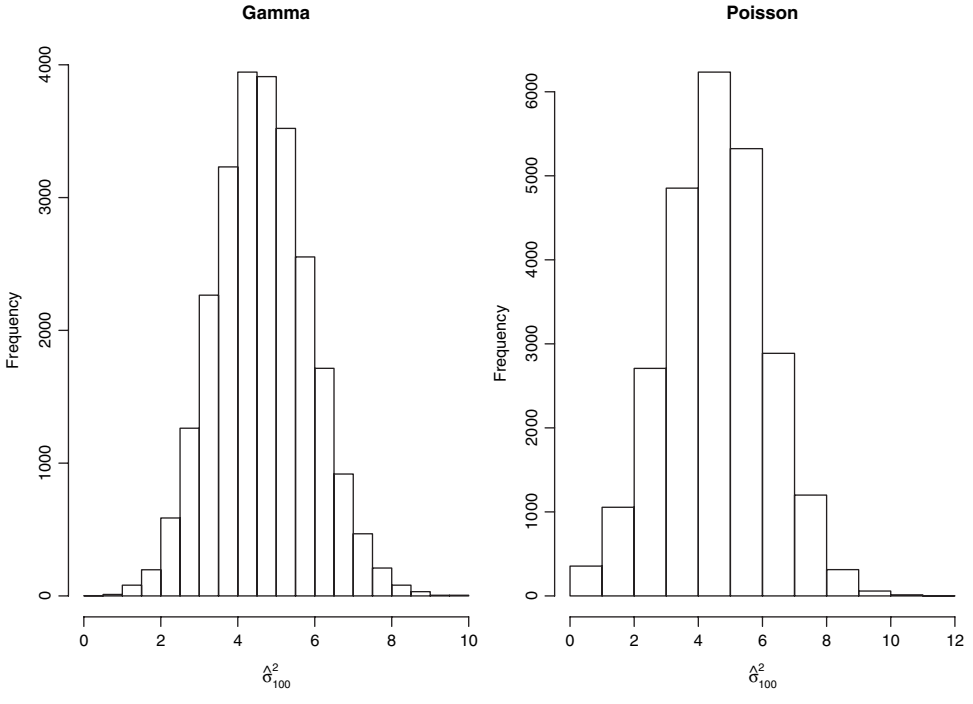

Figure 1. Histograms for $\hat{\sigma}_n^2$ when $n = 100$.

presented above and so they are not reported. We observed in the simulations that the smaller the variance of e_i, σ^2, the greater the percentage of eliminated samples, which is clearly expected. Our approach showed a desirable performance, although for small variances σ^2 and small sample sizes (situations hardly expected in practice) it is possible to have negative variance estimates. Clearly, one way to avoid negative estimates of σ^2 is to use large samples.

Finally, Table 2 depicts the empirical test sizes for testing H_0: $\gamma = \gamma_0$, ($\gamma_0 = -3, -2, \ldots, 2, 3$) for the Wald statistics at the 5% nominal level, when $n = 50$, 100, 200 under scenarios (i) and (ii). The Wald statistics is also computed under the *ordinary model*, that is, using the least squares estimators, without measurement error correction. As can be seen, the empirical test sizes for the approach proposed in this paper is much closer to the real nominal level than the corresponding empirical levels for the Wald statistics using ordinary least squares estimates. Figure 1 depicts the histograms for $\hat{\sigma}_n^2$ when $n = 100$ considering Gamma and Poisson distributions for $X_i|x_i$ (using the same setup described above for the other parameters). As can be seen, the histogram is quite symmetric indicating good agreement with the normal distribution.

5. Final discussion

This paper considered a class of distributions for the surrogate vector of $\boldsymbol{x}_i$ (the unobserved covariate) generalizing previous works in the literature. The simulation studies showed that if the distribution of the surrogate variable is well specified, then the corrected estimators proposed present good behaviour in the sense of decreasing bias and MSE and the Wald statistics based on those estimators present empirical test sizes close to the nominal levels adopted. Such features seem no to be shared by the ordinary least squares estimators which present a much poorer performance. Furthermore, the corrected estimators are easily obtained and iterative procedures

are not required. We emphasize that supplemental data for error correction are not necessary if the surrogate distribution is Poisson, exponential, continuous uniform, discrete uniform or any other distribution having no additional parameters to be estimated. As mentioned by a referee, one can consider σ^2 as a function of the location parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in order to avoid negative estimates for it. We expect to report on such more general situations in incoming papers.

Acknowledgements

The authors are grateful to the referees for their valuable suggestions which led to an improved presentation. The author's research were partially supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

References

[1] W. Fuller, *Measurement Error Models*, Wiley, Chichester, 1987.
[2] C.L. Cheng and J.W. Van Ness, *Statistical Regression with Measurement Error*, Arnold Publishers, London, 1999.
[3] L. Li, M. Palta, and J. Shao, *A measurement error model with a Poisson distributed surrogate*, Stat. Med. 23 (2004), pp. 2527–2536.
[4] T. Nakamura, *Corrected score functions for errors-in-variables models: Methodology and applications to generalized linear models*, Biometrika 77 (1990), pp. 127–137.
[5] P. Gimenez and H. Bolfarine, *Corrected score functions in classical error-in-variables and incidental parameters models*, Aust. J. Stat. 39(3) (1997), pp. 325–344.
[6] P.J. Huber, *The behavior of maximum likelihood estimates under nonstandard conditions*, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. LeCam and J. Neyman, eds., Vol. 1, 1967, pp. 221–233.
[7] R Development Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0. Available at: url: http://www.R-project.org
[8] P. Gimenez and H. Bolfarine, *Hypothesis testing for error-in-variables models*, Ann. Inst. Stat. Math. 52(4) (2000), pp. 698–711.