Article:

Carrasco, JMF, Figueroa-Zuñiga, JI, Leiva, V et al. (2 more authors) (2020) An errors-in-variables model based on the Birnbaum-Saunders and its diagnostics with an application to earthquake data. Stochastic Environmental Research and Risk Assessment, 34 (2). pp. 369-380. ISSN 1436-3240

https://doi.org/10.1007/s00477-020-01767-3

An errors-in-variables model based on the Birnbaum-Saunders and its diagnostics with an application to earthquake data

Jalmar M. F. Carrasco[1], **Jorge I. Figueroa-Zuñiga**[2], **Victor Leiva**[3]*, **Marco Riquelme**[4], **Robert G. Aykroyd**[5]

[1]Department of Statistics, Universidade Federal da Bahia, Brazil
[2]Department of Statistics, Universidad de Concepción, Chile
[3]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Chile
[4]Institute of Statistics, Universidad de Valparaíso, Chile
[5]Department of Statistics, University of Leeds, UK

Abstract

Regression modelling where explanatory variables are measured with error is a common problem in applied sciences. However, if inappropriate analysis methods are applied, then unreliable conclusions can be made. This work deals with estimation and diagnostic analytics in regression modelling based on the Birnbaum-Saunders distribution using additive measurement errors. The maximum pseudo-likelihood and regression calibration methods are used for parameter estimation. We also carry out a residual analysis and apply global and local diagnostic techniques in order to detect anomalous and potentially influential observations. Simulations are conducted to validate the proposed approach and to evaluate performance. A real-world data set, related to earthquakes, is used to illustrate the new approach.

Keywords: Diagnostic techniques; Likelihood methods; Measurement errors; Monte Carlo simulation; Ox and R software; Regression analysis.

1 Introduction

When studying the relationship between a variable of interest (the response) and a set of explanatory variables (the covariates), ignoring possible measurement error in the explanatory variables can cause inconsistent estimators of model parameters; see Stefanski (1985) and Skrondal and Kuha (2012). In this case, the estimators obtained by some usual estimation method, such as least squares or maximum likelihood (ML), when the unobserved covariates are simply replaced by the observed covariates, are called naive estimators. Instead, when variables are subject to measurement error, or are not observed directly, errors-in-variables models should be used, otherwise unreliable inferential results could be obtained; see Stefanski and Carroll (1985).

*Corresponding author: Víctor Leiva. Email: victorleivsanchez@gmail.com; URL: www.victorleiva.cl

There are many reasons why such errors occur, the most common ones being instrument errors. For example, these errors can be present in agriculture and environmental variables, such as rainfall, soil nitrogen content, farm crop acreage; in medical variables, such as blood pressure, pulse rate, temperature, and blood analytics; in management sciences, social sciences and related other fiends, many variables can only be measured with error. In addition, Buonaccorsi (2010, Ch.1, pp. 1-3) mentioned several examples where measurement error occurs. A relevant specific environmental example is described in Fuller (1987, Ch.1, p. 18), where yield of corn is related to the level of nitrogen in the soil, and that this level is measured with error as it is obtained indirectly through laboratory analysis.

In the statistical literature, errors-in-variables regression models are often formulated in terms of a response as a function of covariates, which are measured with error, or are indirectly observed. Thus, in place of true measurements of the covariates, values of another covariate are measured with error. Three forms of modelling are often used when such measurement problems exist: (i) structural modelling, where the unobserved covariate is described by a probability distribution; (ii) functional modelling, where the unknown values of the covariates are treated as parameters and (iii) ultra-structural modelling. Note that the ultra-structural model is a generalization of the structural and functional models; see Gleser (1991). In this paper, we consider a BS errors-in-variables model where the unobserved covariate follows a normal distribution, that is, a structural model, which is a particular case of the ultra-structural model. In addition to theoretical and computational problems, the structural and functional models can suffer from non-identifiability and unbounded likelihood function problems, respectively, as described by (Kendall and Stuart, 2010, Ch. 29, p. 380). Therefore, one of the objectives of the methodology generated from errors-in-variables models is to find consistent estimators of the parameters of interest. Several methods lead to consistent estimators in structural and functional linear models. Some of them involve explicit bias correction of the estimators, while others propose alternative estimators under particular assumptions, as shown by Fuller (1987, Ch.1, p. 18) and Cheng and Van Ness (1999, Ch. 1, pp. 1-48). For the case of non-linear models, some proposed methods are suitable only for estimates under the structural models approach, as they require knowledge of the conditional distribution of the unobserved covariate given the observed covariates; see (Carroll et al., 2006, Ch. 3, p. 65). These estimation methods include maximum pseudo-likelihood techniques and regression calibration; see Guolo (2011).

Errors-in-variables modelling has been addressed using parametric distributions such as the beta and simplex laws; see Carrasco et al. (2014) and Carrasco et al. (2019). A plausible alternative distribution to derive errors-in-covariates models is the Birnbaum-Saunders (BS) distribution, which is skewed to the right and unimodal, having two parameters which modify its shape and scale. The BS distribution has been widely studied and applied in different areas, including engineering and environmental sciences; see Marchant et al. (2013, 2018, 2019), Leiva et al. (2015, 2016), Balakrishnan and Kundu (2019), Martinez et al. (2019), and references therein. In statistical modelling, the BS distribution has received considerable attention. Rieck and Nedelman (1991) developed a BS log-linear model based on the logarithmic version of the BS distribution (in short log-BS), and established a relationship between the BS and log-BS distributions. Subsequently, Villegas et al. (2011) considered an extension of the BS log-linear model, proposed by Rieck and Nedelman (1991), using a BS mixed log-linear model. Leiva et al. (2014) focused modelling on a re-parameterization of the BS distribution. However, although a vast literature on errors-in-variables models exists, formulations of this type based on the BS distribution are still unexplored. We extend the errors-in-variables modelling framework for dealing with covariates measured with errors to include the BS distribution. This adds a new option to the toolbox for applied statistical analysis of error-in-variable problems, which is especially designed for skew measurements.

Diagnostic analytics, a vital step in any modelling, consists of checking model assumptions and identifying departures from these assumptions, as well as identifying the existence of outlying and influential cases. Residuals can be based on their standardized ordinary versions (Leiva et al., 2016), built from deviance components (McCullagh and Nelder, 1983, Ch. 2, p. 35), or using generalized versions (Cox and Snell, 1968). Many studies have used residuals in regression modelling. Pregibon (1981) proposed a deviance component residual in the class of generalized linear models. McCullagh and Nelder (1983, Ch. 6, p. 398) presented a standardization to correct for the effects of skewness and kurtosis. Atkinson (1985) used Monte Carlo methods to construct bands for the residuals called envelopes, which allows appropriate interpretation if the residuals have the expected distribution under the model assumptions. Williams (1987) constructed envelopes in generalized linear models. Fuller (1987, Ch. 1, p. 25), Carroll and Spiegelman (1992) and Buonaccorsi (2010, Ch. 4, p. 94) presented residuals in the presence of measurement errors, suggesting the use of residual plots rather than estimating the predicted values of the unobserved variable. Global and local influence techniques to detect potentially influential cases were proposed by Cook (1977, 1986) and Cook et al. (1988). Some recent papers on the topic are attributed to Santana et al. (2011), Marchant et al. (2016), Garcia-Papani et al. (2017, 2018a,b), Huerta et al. (2018, 2019), Leão et al. (2018), Saulo et al. (2019), and Rodriguez et al. (2020).

The objective of this work is to derive a methodology based on BS errors-in-variables models. The remainder of this paper is organized as follows. In Section 2, we formulate a BS regression model with measurement errors under additivity, whereas its parameter estimation is considered in Section 3. Section 4 presents methods for diagnostic analytics. In Section 5, we describe the numerical results from a simulation study to evaluate the performance of the estimators and a real data illustration to show the potential applications of our methodology. Finally, some conclusions and suggestions for future work are given in Section 6.

2 The model

In this section, we provide background to the BS and log-BS distributions, as well as their modelling. Then, we formulate the new errors-in-variables model based on the log-BS distribution.

2.1 The Birnbaum-Saunders distribution

Consider a random variable T that follows a BS distribution, which is denoted by $T \sim \text{BS}(\alpha, \eta)$, with shape parameter $(\alpha > 0)$ and scale parameter $(\eta > 0)$. The probability density function of T is given by

$$f_T(t; \alpha, \eta) = \frac{t^{-3/2}(t+\eta)}{2\alpha\sqrt{n}}\phi\left(\frac{1}{\alpha}\left(\sqrt{\frac{t}{\eta}} - \sqrt{\frac{\eta}{t}}\right)\right), \quad t > 0,$$

where ϕ represents the probability distribution function of the standard normal distribution, while η is also the median of the distribution. Rieck and Nedelman (1991) developed a sinh-normal (SN) distribution. If the random variable Y follows an SN distribution with shape $(\alpha > 0)$, location

($\mu \in \mathbb{R}$), and scale ($\sigma > 0$) parameters, its probability density function is expressed as

$$f_Y(y; \alpha, \mu, \sigma) = \frac{2}{\alpha\sigma} \cosh\left(\frac{y-\mu}{\sigma}\right) \phi\left(\frac{2}{\alpha} \sinh\left(\frac{y-\mu}{\sigma}\right)\right), \quad y \in \mathbb{R},$$

and then the notation $Y \sim \mathrm{SN}(\alpha, \mu, \sigma)$ is used. If $T \sim \mathrm{BS}(\alpha, \eta)$, then $Y = \log(T) \sim \mathrm{SN}(\alpha, \mu, \sigma = 2)$, where $\mu = \log(\eta)$. For this reason, the SN distribution is also known as the log-BS distribution, where $Y \sim \text{log-BS}(\alpha, \mu)$. Rieck and Nedelman (1991) proposed a fixed-effects log-linear BS regression model with systematic component $\mu_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma}$, for $i = 1, \ldots, n$, where μ_i is the mean of $Y_i \sim \text{log-BS}(\alpha, \mu_i)$, $\boldsymbol{\gamma} \in \mathbb{R}^p$ is the vector of the regression coefficients, and $\boldsymbol{z}_i^\top = (z_{i1}, \ldots, z_{ip})^\top$ is the vector of covariates.

2.2 Birnbaum-Saunders errors-in-variables models

In practice, some covariates may not be directly observed but, instead, are measured with errors. To illustrate this situation in the log-BS regression model, we assume the presence of a single covariate obtained with error. This methodology can then be easily extended to situations in which the data set has more than one covariate measured with error. Specifically, we consider that $\mu_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma} + \beta x_i$, where $\beta \in \mathbb{R}$ is the unknown parameter and x_i is the unobserved true variable. As mentioned above, models with measurement errors can be addressed in three ways. In this work, we study the log-BS regression model with measurement errors under the structural approach. Thus, we leave the analysis under the functional approach to future research.

Suppose $(y_1, w_1), \ldots, (y_n, w_n)$ are pairs of variables observed in a sample of size n — here, we omit the vector of covariates $\boldsymbol{z}_i$ from the notation since they are known and fixed. In addition, recall that $x_1, \ldots, x_n$ are unobserved true variables corresponding to the observed variables $w_1, \ldots, w_n$. Furthermore, let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$ denote the vector of model parameters with $\boldsymbol{\theta}_1$ representing the parameters of interest and $\boldsymbol{\theta}_2$ are irrelevant parameters known as nuisance parameters. The joint probability density function of (Y_i, W_i), for the case i, is obtained by integrating with respect to X_i the joint probability density function of the complete set (Y_i, W_i, X_i), corresponding to

$$f_{Y_i, X_i, W_i}(y_i, x_i, w_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f_{Y_i, X_i | W_i = w_i}(y_i, x_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) f_{W_i}(w_i; \boldsymbol{\theta}_2).$$

Therefore, the associated log-likelihood function is given by

$$\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \log\left(\int f_{Y_i, W_i | X_i = x_i}(y_i, w_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) f_{X_i}(x_i; \boldsymbol{\theta}_2) \mathrm{d}x_i\right). \tag{1}$$

In general, the likelihood function defined in (1) is analytically intractable due to the presence of the integral. An approach used in the literature to approximate the integral is the Gaussian-Hermite quadrature method, which is formulated as

$$\int_{\mathbb{R}} \exp(-x^2) f(x) \mathrm{d}x \approx \sum_{q=1}^{Q} \nu_q f(s_q), \tag{2}$$

where ν_q, s_q are the weights and roots of the Hermite polynomial, respectively, whereas f is the function to be approximated; see (Abramowitz and Stegun, 1972, p. 890). In models with measurement error, practical situations lead us to assume an additive or multiplicative structural link between the observed variable W_i and the unobserved true variable X_i. Here, we assume an additive structure.

Suppose X_i is an unobserved covariate, for $i = 1, \ldots, n$ and the covariate W_i is observed in place of X_i, assuming

$$W_i = \tau_0 + \tau_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $(\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of independent random errors and τ_0, τ_1 are possibly unknown parameters. Carrasco et al. (2014) defined τ_0 and τ_1 as the additive and multiplicative bias of the mechanism of measurement errors, respectively. If $\tau_0 = 0$ and $\tau_1 = 1$, the model reduces to the classical measurement error model. Under the structural approach, we assume that $X_i \sim \mathrm{N}(\mu_X; \sigma_X^2)$ and $\varepsilon_i \sim \mathrm{N}(0; \sigma_\varepsilon^2)$. The log-likelihood function for a sample of size n is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(f_{W_i}(w_i; \boldsymbol{\theta}_2)) + \sum_{i=1}^{n} \log\left(\int f_{Y_i|X_i=x_i}(y_i; \boldsymbol{\theta}) f_{X_i|W_i=w_i}(x_i; \boldsymbol{\theta}_2) \mathrm{d}x_i \right), \tag{3}$$

where $f_{Y_i|X_i=x_i}$ is the log-BS density, $f_{X_i|W_i=w_i}$ is the density of the conditional distribution of X_i given $W_i = w_i$, which is normally distributed with mean and variance defined by

$$\mu_{X|W} = \mu_X + k(w_i - \mu_X) \quad \text{and} \quad \sigma_{X|W}^2 = \sigma_\varepsilon^2 k,$$

for $k = \sigma_X^2/(\sigma_X^2 + \sigma_\varepsilon^2)$, and f_{W_i} is the marginal probability density function of W_i. From (2), and using the standardization transformation $(X - \mu_{X|W})/\sigma_{X|W}$ to reduce the the conditional distribution of X_i given $W_i = w_i$ to a standard normal, the log-likelihood function defined in (3) can be approximated by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(f_{W_i}(w_i; \boldsymbol{\theta}_2)) + \sum_{i=1}^{n} \log\left(\sum_{q=1}^{Q} \frac{\nu_q}{\sqrt{\pi}} f_{Y_i|X_i=\mu_{x|w}+\sqrt{2\sigma_{x|w}^2} s_q}(y_i; \boldsymbol{\theta}) \right).$$

3 Estimation

In this section, we use the maximum pseudo-likelihood and regression calibration estimation techniques. The simulation studies of Carrasco et al. (2014) and Guolo (2011) showed that the maximum pseudo-likelihood estimation method provides the best asymptotic properties for the estimators. However, the regression calibration method, which is widely used because of its computational simplicity, presents slightly biased estimators.

3.1 Maximum pseudo-likelihood

Consider $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$ as defined above. The central idea of the maximum pseudo-likelihood estimation method is to replace the vector of nuisance parameter vector $\boldsymbol{\theta}_2$ with a consistent estimator in the original likelihood function, thereby generating a pseudo-likelihood function. The pseudo-log-likelihood function is maximized in two steps. First, such as in Skrondal and Kuha (2012) and

Carrasco et al. (2014), we estimate $\boldsymbol{\theta}_2$ by maximizing a reduced log-likelihood function defined as

$$\ell_r(\boldsymbol{\theta}_2) = \sum_{i=1}^{n} \log(f_{W_i}(w_i; \boldsymbol{\theta}_2)),$$

which, using the approach defined in Guolo (2011), can be written as

$$\ell_r(\boldsymbol{\theta}_2) = \sum_{i=1}^{n} \log\left(\int f_{W_i|X_i=x_i}(w_i; \boldsymbol{\theta}_2) f_{X_i}(x_i; \boldsymbol{\theta}_2) \mathrm{d}x_i\right). \tag{4}$$

In the model with additive measurement errors, the second step consists of plugging the estimate $\widehat{\boldsymbol{\theta}}_2$ obtained using (4) into the log-likelihood function defined in (3), the result of which is the pseudo log-likelihood function expressed as

$$\ell_p(\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2) = \sum_{i=1}^{n} \log\left(f_{W_i}(w_i; \widehat{\boldsymbol{\theta}}_2)\right) + \sum_{i=1}^{n} \log\left(\int f_{Y_i|X_i=x_i}(y_i; \boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_2) f_{X_i|W_i=w_i}(x_i; \widehat{\boldsymbol{\theta}}_2) \mathrm{d}x_i\right).$$

3.2 Regression calibration

Regression calibration is a simple and widely-used method, which can be applied to any regression model with measurement error to estimate parameters, and it has less computational burden than the ML method; see Thurston et al. (2005), Carroll et al. (2006, Ch. 4, pp. 65-96), Freedman et al. (2008), and Guolo (2011). The central idea of this method is to replace the unobserved variable X_i with an estimate of the conditional expectation of X_i given $W_i = w_i$, $\widehat{\mathrm{E}}(X_i|W_i = w_i)$, in the original log-likelihood function. This allows us to obtain a modified version of the usual log-likelihood function of the BS log-linear regression model expressed as

$$\ell_{rc}(\boldsymbol{\theta}_1) \;=\; -\frac{n}{2}\log(2\pi) + \sum_{i=1}^{n}\log\left(\frac{2}{\alpha}\cosh\left(\frac{y_i - \mu_i}{2}\right)\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{2}{\alpha}\sinh\left(\frac{y_i - \mu_i}{2}\right)\right)^2,$$

where $\mu_i^* = \boldsymbol{z}_i^\top \boldsymbol{\gamma} + x_i^* \beta$, with $x_i^* = \widehat{\mathrm{E}}(X_i|W_i = w_i) = \widehat{\mu}_X + \widehat{k}(w_i - \widehat{\mu}_X)$, $\widehat{k} = \widehat{\sigma}_X^2/(\widehat{\sigma}_X^2 + \widehat{\sigma}_\varepsilon^2)$, and $\widehat{k}$ being known as reliability ratio. In this case,

$$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i, \quad s_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(w_i - \overline{w})^2$$

are the optimal sampling estimators of $\widehat{\mu}_X$ and $\widehat{\sigma}_X^2 + \widehat{\sigma}_\varepsilon^2$, respectively.

4 Diagnostic analysis

In this section, we provide diagnostic methods based on residual analysis and global and local influence techniques for BS errors-in-variables log-linear regression models. Removing cases and re-estimating model parameters is a typical strategy for evaluating the impact of each case on the parameter estimates. The Cook distance (Cook, 1977), originally developed for normal linear models, can be quickly assimilated and extended to different classes of models. However, the elimination of individual cases can lead to a masking effect, as it fails to detect jointly discrepant cases. Another important feature of diagnostic analytics is the detection of influential observations. Cook (1986) proposed assessing the influence of cases by examining the likelihood curvature.

4.1 Residual analysis

This subsection is concerned with finding a measure of the discrepancy between the adjusted model and the data. Thus, one can define a residual as a measure using the difference $y_i - \widehat{\mathrm{E}}(Y_i)$. Then, we define the ordinary residual for the BS regression model with measurement errors as

$$r_i = \frac{y_i - \widehat{\mu}_i^*}{\sqrt{\widehat{\mathrm{Var}}(Y_i)}}, \quad i = 1, \ldots, n,$$

where $\widehat{\mu}_i^* = \boldsymbol{z}_i^\top \widehat{\gamma} + \widehat{X}_i \widehat{\beta}$ and $\widehat{\mathrm{Var}}(Y_i) = \widehat{\alpha}^2\,(1 + 5\widehat{\alpha}^2/4)\exp(\widehat{\mu}_i^*)$, with $\widehat{X}_i = \widehat{\mathrm{E}}(X_i|W_i = w_i)$. Atkinson (1985) suggested that, in order to better interpret the normal probability plot of the proposed residuals, this must be supplemented by envelopes, which are simulated bands obtained by Monte Carlo methods from the adjusted model to assess the existence of serious deviations in the proposed distribution. In a half-normal probability plot, the ith residual value, for $i = 1, \ldots, n$, is compared with the expected values of the order statistics, in absolute value, of the standard normal distribution, given by $\Phi^{-1}((i + n - 1/8)/(2n + 1/2))$, where Φ is the N(0, 1) cumulative distribution function. The graphical plot of the simulated envelope can be used even if the residuals do not have a normal distribution. When this occurs, we do not expect the values to be close to the identity line.

4.2 Global influence

Global influence methods consist of studying the effect of removing the case i of a data set. Consider the log-likelihood function depending on parameter $\boldsymbol{\theta}$ denoted by $\ell(\boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}_{(i)}$ be the estimator of $\boldsymbol{\theta}$ without the case i. Influence of this case can be evaluated as the difference between $\widehat{\boldsymbol{\theta}}_{(i)}$ and $\widehat{\boldsymbol{\theta}}$. If removal of a case causes significant variations in the estimates, more attention should be given to this case. If $\widehat{\boldsymbol{\theta}}_{(i)}$ is far from $\widehat{\boldsymbol{\theta}}$, then the case i is considered to be potentially influential. A first measure of global influence may be defined as a standardized norm and is also known as the generalized Cook distance, defined by

$$\mathrm{CD}_i(\boldsymbol{\theta}) = (\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}})^\top(-\ddot{\boldsymbol{\ell}}(\boldsymbol{\theta}))(\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}), \quad i = 1, \ldots, n,$$

where $\ddot{\boldsymbol{\ell}}(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ is the corresponding Hessian matrix. An alternative measure (Cook et al., 1988) to the Cook distance is the case-deletion likelihood distance (LD_i), which is defined by

$$\mathrm{LD}_i(\boldsymbol{\theta}) = 2(\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_{(i)})), \quad i = 1, \ldots, n,$$

where ℓ is the corresponding log-likelihood function.

4.3 Local influence

The local influence method consists of checking the existence of cases that, under small perturbations, cause significant changes in the results. The method suggested by Cook (1986) is based on the perturbation likelihood distance (LD), which is defined as

$$\mathrm{LD}(\boldsymbol{\delta}) = 2(\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\delta}})),$$

where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\delta}}$ are the ML estimates based on $\ell(\boldsymbol{\theta})$ and on the perturbation log-likelihood function $\ell_{\boldsymbol{\delta}}(\boldsymbol{\theta})$, respectively. Further, let $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^\top$ denote a vector of perturbations and let $\boldsymbol{\delta}_0$ represent the absence of perturbation, so that $\ell(\boldsymbol{\theta}_{\boldsymbol{\delta}_0}) = \ell(\boldsymbol{\theta})$.

Cook (1986) proposed studying the local behaviour of $\mathrm{LD}(\boldsymbol{\delta})$ around $\boldsymbol{\delta}_0$ to evaluate how the geometric surface, called the influence graph, $\breve{\alpha}(\boldsymbol{\delta}) = (\boldsymbol{\delta}, \mathrm{LD}(\boldsymbol{\delta}))^\top$, deviates from the tangent plane at $\boldsymbol{\delta}_0$ as $\boldsymbol{\delta}$ moves slowly away from $\boldsymbol{\delta}_0$ (that is, when small perturbations are introduced into the model). This analysis is performed by examining the curvature of the surface $\breve{\alpha}(\boldsymbol{\delta})$ around $\boldsymbol{\delta}_0$ in direction $\boldsymbol{d}$. Cook (1986) showed that the curvature of the surface, $C_{\boldsymbol{d}}(\boldsymbol{\theta})$, in the direction $\boldsymbol{d}$ is given by $C_{\boldsymbol{d}}(\boldsymbol{\theta}) = 2|\boldsymbol{d}^\top \ddot{\boldsymbol{F}}(\boldsymbol{\theta})\boldsymbol{d}|$, where $\ddot{\boldsymbol{F}}(\boldsymbol{\theta}) = \boldsymbol{\Delta}^\top(-\ddot{\boldsymbol{\ell}}(\boldsymbol{\theta}))^{-1}\boldsymbol{\Delta}$, with $\boldsymbol{\Delta} = \partial^2\ell_{\boldsymbol{\delta}}(\boldsymbol{\theta})/\partial\boldsymbol{\delta}\partial\boldsymbol{\theta}^\top$ being an array of dimension $n(\boldsymbol{\theta}) \times n$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \boldsymbol{\delta} = \boldsymbol{\delta}_0$, and $n(\boldsymbol{\theta})$ representing the dimension of $\boldsymbol{\theta}$. One can express $C_{\boldsymbol{d}}(\boldsymbol{\theta})$ as

$$C_{\boldsymbol{d}}(\boldsymbol{\theta}) = 2\sum_{m=1}^{n} \lambda_m \boldsymbol{v}_m \boldsymbol{v}_m^\top,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n(\boldsymbol{\theta})} \geq \lambda_{n(\boldsymbol{\theta})+1} \geq \cdots \geq \lambda_n$ are the sorted eigenvalues of the array $\ddot{\boldsymbol{F}}(\boldsymbol{\theta})$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are their respective eigenvectors. The interest is in the direction that produces the greatest local influence. This direction, $\boldsymbol{d}_{\max}$, is the normalized eigenvector corresponding to the largest eigenvalue of $\ddot{\boldsymbol{F}}(\boldsymbol{\theta})$. Comparing the graph of the eigenvector components of the $\boldsymbol{d}_{\max}$ with the index of cases is useful in identifying influential observations.

Lesaffre and Verbeke (1998) suggested considering the direction of the case i, the vector $\boldsymbol{d}_i = (0, \ldots, 1, \ldots, 0)^\top$, with the ith element being one. In this sense, a normal curvature, called the total local influence of the case i, is given by

$$C_{\boldsymbol{d},i}(\boldsymbol{\theta}) = 2|\boldsymbol{\Delta}_i^\top(-\ddot{\boldsymbol{\ell}}(\boldsymbol{\theta}))^{-1}\Delta_i|, \quad i = 1, \ldots, n,$$

where Δ_i denotes the ith column of the matrix Δ. In addition, Lesaffre and Verbeke (1998) proposed comparing the graph of $C_{\boldsymbol{d},i}(\boldsymbol{\theta})$ against i to detect influential cases. It is also suggested to use twice the mean value of this measure as the cut-off value on the graph of $C_{\boldsymbol{d},i}(\boldsymbol{\theta})$. Thus, if for the case i the following condition holds

$$C_{\boldsymbol{d},i}(\boldsymbol{\theta}) > \frac{2}{n}\sum_{i=1}^{n} C_{\boldsymbol{d},i}(\boldsymbol{\theta}),$$

then it is classified as potentially influential. In this work, we consider the the diagnostic methods: case-weight, response variable, covariate measured without error, and covariate measured with error. The surfaces for the different schemes of perturbation are calculated numerically using the programming language `Ox`; see Doornik (2006).

5 Numerical results

In this section, we provide the numerical results of our study divided into (i) a Monte Carlo simulation study to evaluate the performance of our proposal, and (ii) an illustration with real data of the BS errors-in-variables model.

5.1 Simulation study

The simulation study presented in this subsection is carried out to understand the asymptotic behaviour of the estimators obtained by using the maximum pseudo-likelihood and regression calibration methods. Our simulation model is given by $Y_i|X_i = x_i \sim$ log-BS(α, μ_i), for $i = 1, \dots, n$, where $\mu_i = \gamma_0 + \gamma_1 z_i + \beta x_i$, $w_i = x_i + \varepsilon_i$, $x_i \sim$ N(μ_X, σ_X^2), $\varepsilon_i \sim$ N$(0, \sigma_\varepsilon^2)$ and $z_i \sim$ U$(0, 6)$. We also assume $\alpha = 0.4$, $\gamma_0 = 12$, $\gamma_1 = -1.5$, $\beta = 2.0$, $\mu_X = 3.0$, $\sigma_X^2 = 2.5$ and $k = 0.50$ (high measurement error), 0.75 (moderate measurement error) and 0.95 (low measurement error). In addition, we consider $Q = 80$ and $n = 25, 50, 100, 200$. Empirical mean, bias, and root of the mean square error (RMSE) of the estimators are calculated using the maximum pseudo-likelihood, calibration regression and naive methods. Tables 1-3 report the results obtained for this scenario when $k = 0.50$, $k = 0.75$ and 0.95, respectively. These tables show the superiority of the maximum pseudo-likelihood method compared to the regression calibration and naive methods when the measurement error is high. In this situation, the estimators of the regression calibration and naive methods seem to be biased, specifically for the parameters α and β, the latter of which is associated with the variable measured with error. These tables also show that as the sample size increases, the maximum pseudo-likelihood estimators become closer to the true values. When the reliability coefficient k is close to one (that is, the variance of the measurement error approaches zero), the estimators based on the maximum pseudo-likelihood and regression calibration methods display good results as the sample size increases, particularly for the parameter β, which is associated with the variable measured with error. However, if we do not assume the presence of measurement errors in the variable, this can lead to misinterpretation, specially when the variability of the measurement error is high. When the variance of the measurement error is small, the regression calibration method is less computationally demanding.

5.2 Empirical illustration

Our illustration analyzes magnitudes of Alaskan earthquakes for the period from 1969 to 1978 taken from Fuller (1987, Ch. 1, p. 56). Three measures of earthquake magnitude have been observed, corresponding to the logarithm of the seismogram amplitude of 20-second surface waves, denoted by Y_i, the logarithm of the seismogram amplitude of longitudinal surface waves, denoted by X_i, and the logarithm of maximum seismogram trace amplitude at short distance, denoted by W_i. The measurement error includes mistakes made in determining the amplitude of ground motion arising from the location of a limited number of observation stations related to the fault plane of the earthquake. Table 4 gives statistical summary including minimum and maximum values, 1st and 3rd quartiles (Q_1, Q_3), median, mean, standard deviation and the coefficients of skewness (CS) and kurtosis (CK). This summary indicates that the variable "surface wave" has moderate skewness indicating that a non-normal distribution is appropriate.

Table 1: Mean, bias and RMSE of the estimator of the indicated parameter and n with $k = 0.50$, where the true parameter values are: $\alpha = 0.4$, $\gamma_0 = 12$, $\gamma_1 = -1.5$, $\beta = 2.0$.

n	Method	Parameter	Mean	Bias	RMSE
25	Naive	α	3.63	-3.23	3.50
		γ_0	15.01	-3.01	3.46
		γ_1	-1.50	0.00	0.41
		β	1.00	1.00	1.05
	Regression calibration	α	3.64	-3.24	3.51
		γ_0	8.92	3.08	21.99
		γ_1	-1.50	0.00	0.41
		β	3.04	-1.04	7.41
	Pseudo likelihood	α	0.49	-0.09	0.71
		γ_0	11.31	0.69	3.12
		γ_1	-1.50	0.00	0.26
		β	2.23	-0.23	1.01
50	Naive	α	4.02	-3.62	3.80
		γ_0	14.98	-2.98	3.27
		γ_1	-1.50	0.00	0.32
		β	1.00	1.00	1.03
	Regression calibration	α	4.02	-3.62	3.80
		γ_0	10.72	1.28	10.71
		γ_1	-1.50	0.00	0.32
		β	2.42	-0.42	3.58
	Pseudo likelihood	α	0.45	-0.05	0.46
		γ_0	11.64	0.35	2.03
		γ_1	-1.50	0.00	0.19
		β	2.12	-0.12	0.65
100	Naive	α	4.31	-3.90	4.00
		γ_0	15.00	-3.00	3.17
		γ_1	-1.50	0.00	0.25
		β	1.00	1.00	1.02
	Regression calibration	α	4.31	-3.91	4.02
		γ_0	11.62	0.38	2.11
		γ_1	-1.50	0.00	0.25
		β	2.13	-0.13	0.64
	Pseudo likelihood	α	0.46	-0.06	0.38
		γ_0	11.89	0.11	0.97
		γ_1	-1.50	0.00	0.14
		β	2.04	-0.04	0.29
200	Naive	α	4.52	-4.12	4.19
		γ_0	15.00	-3.00	3.11
		γ_1	-1.50	0.00	0.20
		β	1.00	1.00	1.01
	Regression calibration	α	4.52	-4.12	4.19
		γ_0	11.84	0.16	1.39
		γ_1	-1.50	0.00	0.20
		β	2.05	-0.05	0.40
	Pseudo likelihood	α	0.49	-0.09	0.32
		γ_0	11.98	-0.01	0.66
		γ_1	-1.50	0.00	0.10
		β	2.00	0.00	0.19

Table 2: Mean, bias and RMSE of the estimator of the indicated parameter and n with $k = 0.75$, where the true parameter values are: $\alpha = 0.4$, $\gamma_0 = 12$, $\gamma_1 = -1.5$, $\beta = 2.0$.

n	Method	Parameter	Mean	Bias	RMSE
25	Naive	α	1.99	-1.59	1.66
		γ_0	13.50	-1.50	1.85
		γ_1	-1.50	0.00	0.24
		β	1.50	0.50	0.55
	Regression calibration	α	1.99	-1.59	1.66
		γ_0	11.59	0.41	1.86
		γ_1	-1.50	0.00	0.24
		β	2.143	-0.14	0.57
	Pseudo likelihood	α	0.42	-0.02	0.42
		γ_0	11.91	0.09	1.13
		γ_1	-1.50	0.00	0.20
		β	2.02	-0.02	0.32
50	Naive	α	2.14	-1.74	1.78
		γ_0	13.47	-1.49	1.70
		γ_1	-1.50	0.00	0.18
		β	1.50	0.50	0.52
	Regression calibration	α	2.14	-1.74	1.78
		γ_0	11.82	0.18	1.13
		γ_1	-1.50	0.00	0.18
		β	2.06	-0.06	0.31
	Pseudo likelihood	α	0.43	-0.03	0.35
		γ_0	11.98	0.02	0.77
		γ_1	-1.50	0.00	0.14
		β	2.00	0.00	0.20
100	Naive	α	2.23	-1.83	1.86
		γ_0	13.50	-1.50	1.62
		γ_1	-1.50	0.00	0.14
		β	1.50	0.50	0.52
	Regression calibration	α	2.23	-1.83	1.85
		γ_0	11.92	0.08	0.79
		γ_1	-1.50	0.00	0.14
		β	2.03	-0.03	0.21
	Pseudo likelihood	α	0.43	-0.03	0.29
		γ_0	12.00	0.00	0.55
		γ_1	-1.50	0.00	0.10
		β	2.00	0.00	0.14
200	Naive	α	2.28	-1.88	1.89
		γ_0	13.50	-1.50	1.57
		γ_1	-1.50	0.00	0.10
		β	1.50	0.50	0.51
	Regression calibration	α	2.28	-1.88	1.89
		γ_0	11.97	0.03	0.58
		γ_1	-1.50	0.00	0.10
		β	2.001	-0.01	0.15
	Pseudo likelihood	α	0.43	-0.03	0.23
		γ_0	12.00	0.00	0.39
		γ_1	-1.50	0.00	0.07
		β	2.00	0.00	0.10

Table 3: Mean, bias and RMSE of the estimator of the indicated parameter and n with $k = 0.95$, where the true parameter values are: $\alpha = 0.4$, $\gamma_0 = 12$, $\gamma_1 = -1.5$, $\beta = 2.0$.

n	Method	Parameter	Mean	Bias	RMSE
25	Naive	α	0.80	-0.40	0.43
		γ_0	12.30	-0.30	0.57
		γ_1	-1.50	0.00	0.10
		β	1.90	0.10	0.15
	Regression calibration	α	0.80	-0.40	0.43
		γ_0	11.96	-0.04	0.52
		γ_1	-1.50	0.00	0.10
		β	2.02	-0.02	0.12
	Pseudo likelihood	α	0.27	0.13	0.25
		γ_0	11.98	0.02	0.48
		γ_1	-1.50	0.00	0.10
		β	2.01	-0.01	0.12
50	Naive	α	0.84	-0.44	0.45
		γ_0	12.30	-0.30	0.45
		γ_1	-1.50	0.00	0.07
		β	1.90	0.10	0.13
	Regression calibration	α	0.84	-0.44	0.45
		γ_0	11.98	0.02	0.35
		γ_1	-1.50	0.00	0.07
		β	2.01	-0.01	0.08
	Pseudo likelihood	α	0.32	0.08	0.20
		γ_0	11.99	0.01	0.34
		γ_1	-1.50	0.00	0.07
		β	2.01	-0.01	0.08
100	Naive	α	0.86	-0.46	0.46
		γ_0	12.30	-0.30	0.38
		γ_1	-1.50	0.00	0.05
		β	1.90	0.10	0.11
	Regression calibration	α	0.86	-0.46	0.47
		γ_0	11.99	-0.01	0.24
		γ_1	-1.50	0.00	0.05
		β	2.00	0.00	0.06
	Pseudo likelihood	α	0.35	0.05	0.15
		γ_0	11.99	0.01	0.24
		γ_1	-1.50	0.00	0.05
		β	2.00	0.00	0.06
200	Naive	α	0.87	-0.47	0.47
		γ_0	12.30	-0.30	0.34
		γ_1	-1.50	0.00	0.03
		β	1.90	0.10	0.11
	Regression calibration	α	0.87	-0.47	0.47
		γ_0	12.00	0.00	0.17
		γ_1	-1.50	0.00	0.03
		β	2.00	0.00	0.04
	Pseudo likelihood	α	0.37	0.03	0.11
		γ_0	12.00	0.00	0.16
		γ_1	-1.50	0.00	0.03
		β	2.00	0.00	0.04

Table 4: Statistical summary of surface wave data.

Min	Q_1	Median	Mean	Q_3	Max	SD	CS	CK
3.60	4.43	5.05	5.08	5.60	7.00	0.79	0.31	-0.52

Here, we consider the maximum pseudo-likelihood method, which was found to give the best results in the simulation. We propose a regression model with BS distributed measurement error, with the structure

$$Y_i|X_i = x_i \sim \text{log-BS}(\alpha, \mu_i), \quad i = 1, \ldots, n,$$

where $\mu_i = \gamma + \beta x_i$, $W_i = \pi_1 + \pi_2 x_i + \varepsilon_i$, $X_i \sim \text{N}(\mu_X, \sigma_X^2)$, and $\varepsilon_i \sim \text{N}(0, \sigma_\varepsilon^2)$, consequently $W_i \sim \text{N}(\pi_1 + \pi_2\mu_X, \pi_2^2\sigma_X^2 + \sigma_e^2)$. To avoid identifiability problems, when considering the structural approach to measurement error models, the vector of parameters $(\sigma_\varepsilon^2, \pi_1, \pi_2)^\top$ can be obtained when we have replications of W_i or using an instrumental variable. Then, this vector can be considered as a nuisance parameter. Thus, the estimate of $(\sigma_\varepsilon^2, \pi_1, \pi_2)^\top$ is obtained when $X_i \sim \text{N}(\mu_X, \sigma_X^2)$ and $\varepsilon_i \sim \text{N}(0, \sigma_\varepsilon^2)$. Therefore, we take $\widehat{\sigma}_\varepsilon^2 = 0.0873$, calculated from the variance of the error (ε) in the model $W_i = \pi_1 + \pi_2 x_i + \varepsilon_i$, with $W_i \sim \text{N}(\pi_1 + \pi_2\mu_X, \pi_2^2\sigma_X^2 + \sigma_e^2)$, $\widehat{\pi}_1 = 2.28835$ and $\widehat{\pi}_2 = 0.55805$. Estimates of the remaining parameters, their corresponding standard errors, z-scores and p-values using naive, maximum pseudo-likelihood, and regression calibration methods are shown in Table 5. From this table, note that the estimates obtained by the naive method are affected by the presence of the measurement error. We can also observe that the parameter γ is not significant when the measurement error is not considered in the model.

Table 5: Estimates, standard errors and p-values of the indicated parameter with earthquake data.

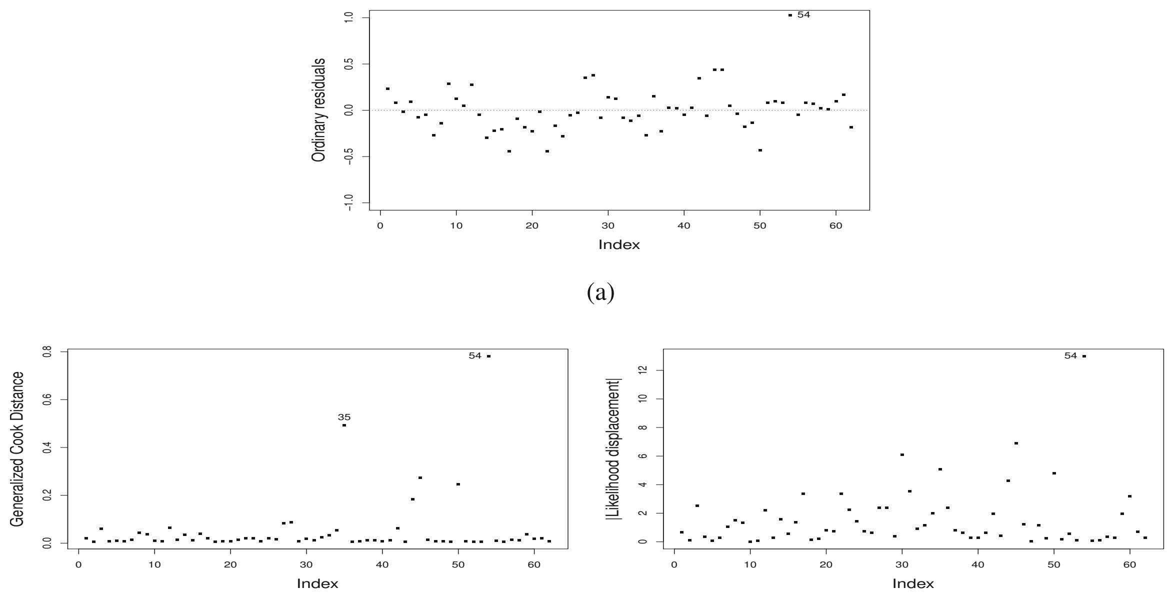
Method	Parameter	Estimate	Standard Error	z-score	p-value
	α	0.5472	0.0491	11.1355	-
Naive	γ	-1.3531	0.7484	-1.8078	0.071
	β	1.2358	0.1433	8.6256	0.000
	α	0.2003	0.2154	0.9297	-
Pseudo likelihood	γ	-6.2210	2.3110	-2.6920	0.007
	β	2.1677	0.4419	4.9049	0.000
	α	0.5472	0.0491	11.1355	-
Regression calibration	γ	-5.9903	1.2848	-4.6624	0.000
	β	2.1251	0.2464	8.6256	0.000

In order to identify outlying and/or influential cases, residual, global and local influence plots are constructed. Figure 1(a) shows the ordinary residuals versus the index of cases. In this graph, we can see that the residuals are randomly distributed around zero without any evidence of lack of fit of the model. Also, note that the case # 54 can be considered as possibly influential.

Graphs of global influence are presented in Figure 1(b)-(c), revealing that cases # 35 and # 54 have an impact on the maximum pseudo-likelihood estimates when they are removed from the data set. In addition, Figures 2 correspond to the measures of local and total local influence for the Alaskan earthquake data on the perturbation schemes of the model, of the response variable and covariate.

From these graphs, we can identify cases #30 and # 45 as being influential.

We complete our diagnostic analytics by finding the percentage relative deviation, PRD $= [(\widehat{\theta} - \widehat{\theta}^*)/\widehat{\theta}] \times 100\%$, where $\widehat{\theta}^*$ represents the estimator of θ obtained after removing one or more outlying and/or influential cases. Table 6 reports estimates, standard errors, z-scores, p-value and PRD when we remove the case # 54 from the data. From this table, the strong changes when deleting the case # 54, specifically in the parameters γ and β, when removing this case, are not significant. Then, we decide to keep these observations in the final predictive BS errors-in-variables model. Once the final model is established, we compare it to the Gaussian (normal) errors-in-variables model (standard model) by means of Akaike information criterion (AIC) and Bayesian information criterion (BIC). Note that the BS model has a better performance (AIC $= 95.50$, BIC $= 101.88$) than the normal model (AIC $= 102.65$, BIC $= 109.04$).



(a)

(b) (c)

Figure 1: Index plot of the (a) ordinary residual, (b) generalized Cook distance and (c) likelihood displacement for the earthquake data.

Table 6: Estimates, standard error, z-value, p-values and PRD (in %) for the indicated parameters when the case # 54 is removed from earthquake data.

Parameter	Estimate	Standard error	z-value	p-value	PRD (%)
α	0.04389	0.06072	0.72284	-	78.0879
γ	-7.59199	5.38881	-1.40884	0.15888	-22.0381
β	2.42193	0.99916	2.42398	0.01535	-11.7281

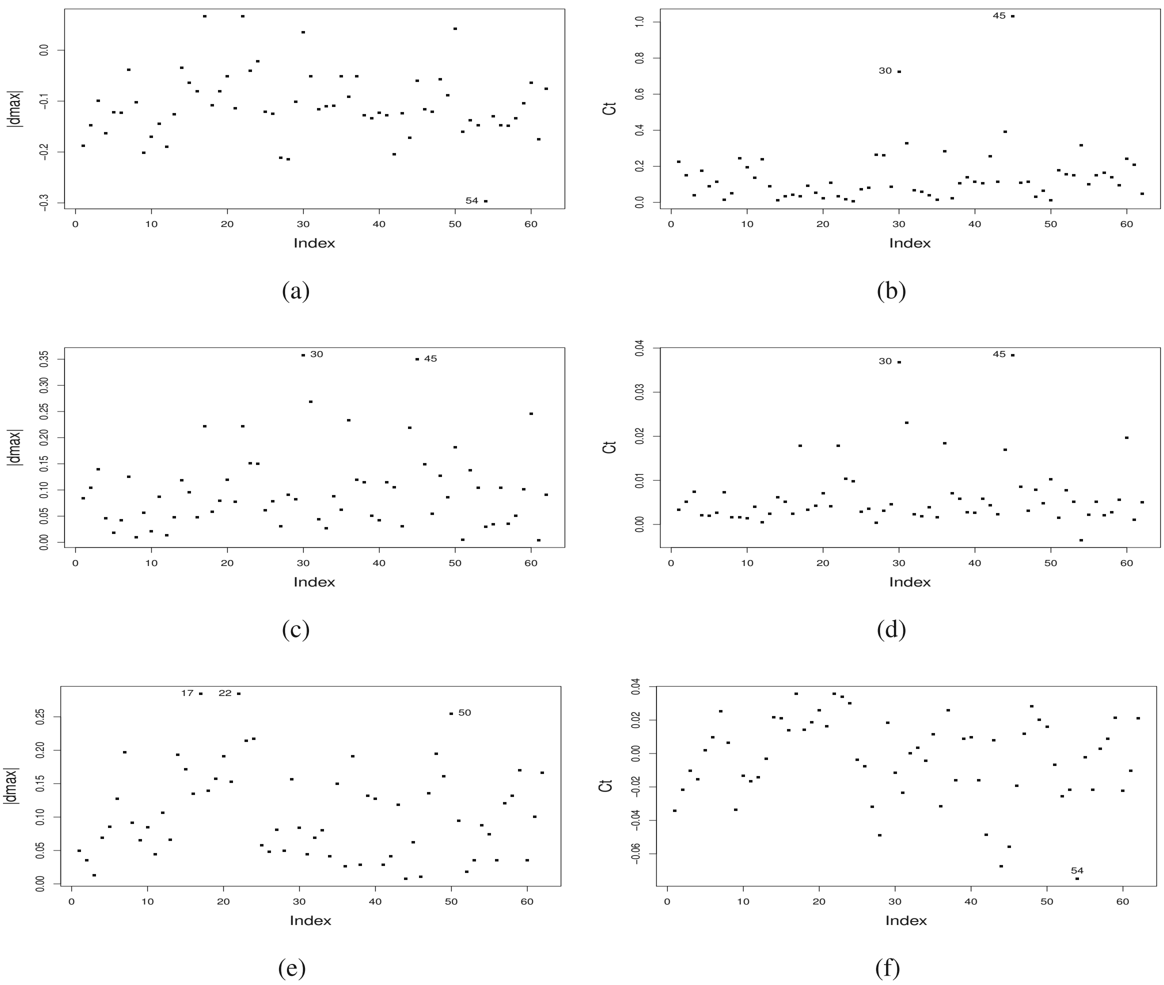


Figure 2: Index plot of $|\boldsymbol{d}_{\max}|$ of (left) local influence and (right) total local influence for perturbation of (a)-(b) case weigh, (c)-(d) response, and (e)-(f) covariate, using earthquakes data.

6 Conclusions

In this work, we studied a model with measurement errors based on the Birnbaum-Saunders distribution. We estimated its parameters using maximum pseudo-likelihood and regression calibration techniques, and also compared them with the method in which measurement errors are not considered (naive likelihood method). The results suggest that not taking into account measurement errors leads to biased estimates, inducing possible inaccurate decisions — this has critical implications for many data-driven scientific studies. We also studied global, local and local total influence under three perturbation schemes, namely perturbation of cases, perturbation of the response variable, and perturbation of the covariate measured with error. Then, we validated the proposed methodology with a real data set and demonstrated that the Birnbaum-Saunders errors-in-variables model has a better performance than the Gaussian errors-in-variables model for these data according to model selection criteria based on loss of information. This suggest that the BS error-in-variables model could also be useful in the analysis of other environmental data sets.

The proposed approach incorporates errors-in-variables modelling which accounts for situations where covariates are measured with error or indirectly. Such modelling leads to better estimation and hence more reliable prediction and inference. The use of the Birnbaum-Saunders distribution allows direct modelling of data sets which are take positive values and follow asymmetric (skewed to the right) distributions. Thus, the present study extended applicability of errors-in-variables modelling beyond the routine symmetric and normal distribution based approaches. Furthermore, the proposed diagnostic analytics complemented the modelling and allowed outlying and influential cases to be identified and hence obtained the final fitted models more robust. Thus, this methodology can have wide ranging applications and has great potential to have significant impact in data analysis. Note that error-in-variables models in general, and in particular our model, can also be used for prediction, considering $x_i^\star$ (an estimate of the conditional expectation of X_i given W_i; see Section 3.2) as the predictor on a future unit.

Further work should include extension of the methodology to functional and ultra-structural modelling approaches to give a full range of techniques. Here, we have only presented the methodology for a single covariate measured with error and hence application to situations in which the data set has more than one covariate measured with error will further highlight the modelling importance. In addition, the approach presented here has a high potential in applied science and there is substantial opportunity for development and validation on other important environmental problems. The method can be added to the toolbox of techniques of data scientists to better model error-in-variables problems and then leading to more reliable and robust decision making.

Acknowledgements

The authors thank the editors and reviewers very much for their constructive comments on an earlier version of this manuscript which resulted in this improved version. This research was supported partially by grant "Fondecyt 1160868" from the National Commission for Scientific and Technological Research of the Chilean government.

References

Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions*. Dover Press, New York, US.

Atkinson, A. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press, Oxford, UK.

Balakrishnan, N. and Kundu, D. (2019). Birnbaum-Saunders distribution: A review of models, analysis and applications. *Applied Stochastic Models in Business and Industry*, 35:4–49.

Buonaccorsi, J.P. (2010). *Measurement Error: Models, Methods and Applications*. Chapman and Hall, London, UK.

Carrasco, J. M.F., Ferrari, S. L.P., and Arellano-Valle, R.B. (2014). Errors-in-variables beta regression models. *Journal of Applied Statistics*, 41:1530–1547.

Carrasco, J. M.F. and Reid, N. (2019). Simplex regression models with measurement error. *Communications in Statistics: Simulation and Computation*, pages in press available at https://doi.org/10.1080/03610918.2019.1626881.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall, New York, US.

Carroll, R.J. and Spiegelman, C.H. (1992). Diagnostics for nonlinearity and heteroscedasticity in errors-in-variables regression. *Technometrics*, 34:186–196.

Cheng, C. and Van Ness, J.W. (1999). *Statistical Regression with Measurement Error*. Oxford University Press, London, UK.

Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19:15–18.

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society B*, 48:133–169.

Cook, R.D., Peña, D., and Weisberg, S. (1988). The likelihood displacement: A unifying principle for influence measures. *Communications in Statistics: Theory and Methods*, 17:623–640.

Cox, D.R. and Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society B*, 2:248–275.

Doornik, J. (2006). *An Object-Oriented Matrix Language*. Timberlake Consultants Press, London, UK.

Freedman, L.S., Midthune, D., Carroll, R.J., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, 27:5195–5216.

Fuller, W.A. (1987). *Measurement Error Models*. Wiley, New York, US.

Garcia-Papani, F., Uribe-Opazo, M.A., Leiva, V., and Aykroyd, R.G. (2017). Birnbaum-Saunders spatial modelling and diagnostics applied to agricultural engineering data. *Stochastic Environmental Research and Risk Assessment*, 31:105–124.

Garcia-Papani, F., Leiva, V., Uribe-Opazo, M.A., and Aykroyd, R.G. (2018a). Birnbaum-Saunders spatial regression models: Diagnostics and application to chemical data. *Chemometrics and Intelligent Laboratory Systems*, 177:114–128.

Garcia-Papani, F., Leiva, V., and Ruggeri, F., and Uribe-Opazo, M.A., (2018b). Kriging with external drift in a Birnbaum-Saunders geostatistical model. *Stochastic Environmental Research and Risk Assessment*, 32:1517–1530.

Gleser, L.J. (1991). Measurement error models. *Chemometrics and Intelligent Laboratory Systems*, 10:45–57.

Guolo, A. (2011). Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica*, 21:1639–1663.

Huerta, M., Leiva, V., Lillo, C., and Rodriguez, M. (2018). A beta partial least squares regression model: diagnostics and application to mining industry data. *Applied Stochastic Models in Business and Industry*, 34:305–321.

Huerta, M., Leiva, V., Liu, S., Rodriguez, M., and Villegas, D (2019). On a partial least squares regression model for asymmetric data with a chemical application in mining. *Chemometrics and Intelligent Laboratory Systems*, 190:55–68.

Kendall, M.G. and Stuart, A. (2010). *The Advanced Theory of Statistics*. Volume II. Wiley, New York.

Leão, J., Leiva, V., Saulo, H., and Tomazella, V. (2018). Incorporation of frailties into a cure rate regression model and its diagnostics and application to melanoma data. *Statistics in Medicine*, 37:4421–4440.

Leiva, V., Ferreira, M., Gomes, M.I., and Lillo, C. (2016). Extreme value Birnbaum-Saunders regression models applied to environmental data. *Stochastic Environmental Research and Risk Assessment*, 30:1045–1058.

Leiva, V., Marchant, C., Ruggeri, F., and Saulo, H. (2015). A criterion for environmental assessment using Birnbaum-Saunders attribute control charts. *Environmetrics*, 26:463–476.

Leiva, V., Santos-Neto, M., Cysneiros, F. J.A., and Barros, M. (2014). Birnbaum-Saunders statistical modelling: A new approach. *Statistical Modelling*, 14:21–48.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, 54:570–582.

Marchant, C., Leiva, V., Cavieres, M.F., and Sanhueza, A. (2013). Air contaminant statistical distributions with application to PM10 in Santiago, Chile. *Reviews of Environmental Contamination and Toxicology*, 223:1-31.

Marchant, C., Leiva, V., Christakos, G., and Cavieres, M.F. (2019). Monitoring urban environmental pollution by bivariate control charts: new methodology and case study in Santiago, Chile. *Environmetrics*, 30:e2551.

Marchant, C., Leiva, V., Cysneiros, F. J.A., and Liu, S. (2018). Robust multivariate control charts based on Birnbaum-Saunders distributions. *Journal of Statistical Computation and Simulation*, 88:182–202.

Marchant, C., Leiva, V., Cysneiros, F. J.A., and Vivanco, J.F. (2016). Diagnostics in multivariate Birnbaum-Saunders regression models. *Journal of Applied Statistics*, 43:2829-2849.

Martinez, S., Giraldo, R., and Leiva, V. (2019). Birnbaum-Saunders functional regression models for spatial data. *Stochastic Environmental Research and Risk Assessment*, 33:1765-1780

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London, UK.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9:705–724.

Rieck, J.R. and Nedelman, J.R. (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics*, 3:51–60.

Rodriguez, M., Leiva, V., Huerta, M., Lillo, M., Ruggeri, F., and Tapia, A. (2020). An asymmetric area model-based approach for small area estimation applied to survey data. *REVSTAT - Statistical Journal*, pages in press available at `https://www.ine.pt/revstat/forthcoming_papers.html`.

Santana, L., Vilca, F., and Leiva, V. (2011). Influence analysis in skew-Birnbaum-Saunders regression models and applications. *Journal of Applied Statistics*, 38:1633–1649.

Saulo, H., Leao, J., Leiva, V., and Aykroyd, R.G. (2019) Birnbaum-Saunders autoregressive conditional duration models applied to high-frequency financial data. *Statistical Papers*, 60:1605–1629.

Skrondal, A. and Kuha, J. (2012). Improved regression calibration. *Psychometrika*, 77:649–669.

Stefanski, L.A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, 78:538–592

Stefanski, L.A. and Carroll, R.J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, 13:1335–1351.

Thurston, S.W., Williams, P.L., Hauser, R., Hu, H., Hernandez-Avila, M., and Spiegelman, D. (2005). A comparison of regression calibration approaches for designs with internal validation data. *Journal of Statistical Planning and Inference*, 131:175–190.

Villegas, C., Paula, G.A., and Leiva, V. (2011). Birnbaum-Saunders mixed models for censored reliability data analysis. *IEEE Transactions on Reliability*, 60:748–758.

Williams, D.A. (1987). Generalized linear models diagnostic using the deviance and single case deletion. *Applied Statistics*, 36:181–191.