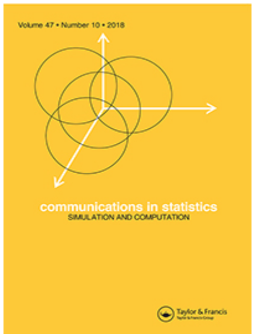

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Simplex regression models with measurement error

Jalmar M. F. Carrasco & Nancy Reid

To cite this article: Jalmar M. F. Carrasco & Nancy Reid (2019): Simplex regression models with measurement error, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1626881

To link to this article: https://doi.org/10.1080/03610918.2019.1626881

Published online: 19 Jun 2019.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1626881

Taylor & Francis
Taylor & Francis Group

Check for updates

Simplex regression models with measurement error

Jalmar M. F. Carrasco[a] and Nancy Reid[b]

[a]Department de Statistics, Federal University of Bahia, Salvador, Brazil; [b]Department of Statistial Sciences, University of Toronto, Toronto, Ontario, Canada

ABSTRACT
This paper considers the simplex regression model when there is measurement error in the covariate. We consider a structural approach where the measurement error follows a normal or gamma distribution. We apply a Monte Carlo EM algorithm to estimate the parameters using a pseudo-likelihood function. A simulation study is used to investigate the impact of ignoring the measurement error. Finally, the results are illustrated with a data set.

ARTICLE HISTORY
Received 6 April 2018
Accepted 28 May 2019

KEYWORDS
Simplex distribution; Gamma distribution; Error-in-variables; Pseudo-likelihood function; Monte Carlo EM algorithm; Proportion data

1. Introduction

Book-length treatments of measurement error and statistical analysis are given in, for instance, Fuller (1987), Carroll et al. (2006) and Buonaccorsi (2010), among others. Causes of measurement error include instrument imprecision, incomplete data and misclassification. The impact of ignoring measurement error varies from problem to problem, and can sometimes be negligible and sometimes drastic. Measurement error models are rapidly gaining importance in many fields, essentially in medical, health and epidemiological studies. Medical variables, such as blood pressure, pulse rate, temperature, and blood chemistries, are measured with non-negligible error; variables in agricultural studies such as precipitation, soil nitrogen content, degree of pest infestation, farm crop acreage allocation, and the like cannot be measured precisely. In management sciences, social sciences, and many others fields some variables can only be measured with error.

If measurement error is ignored, parameter estimates and confidence intervals may suffer serious biases. In addition, measurement error may cause a loss of power for detecting evidence and connections among variables and may mask important features of the data. A number of approaches for analysis of measurement error models have been proposed: for example, correction of moments (Fuller 1987), simulation extrapolation (Cook and Stefanski 1994), regression calibration (Carroll and Stefanski 1990), Bayesian analyses (Gustafson 2004) and inference via maximum pseudo-likelihood (Guolo 2011). Midthune et al. (2016) approached measurement error models using interactions between unobserved and error-free variables. Cheng et al. (2016) studied a

CONTACT Jalmar M. F. Carrasco carrascojalmar@gmail.com Department de Statistics, Federal University of Bahia, Salvador 40110-060, Brazil.

Supplemental data for this article can be accessed here.

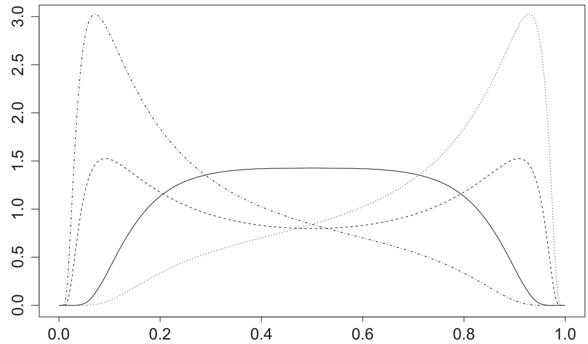

Figure 1. The simplex density function.

method for checking the goodness of fit of the restricted measurement error model and Carrasco et al. (2014) proposed an errors-in-variables beta regression model. The authors assumed a structural additive measurement error model that relates the unobservable covariate with its surrogate, and postulated a normal distribution for the unobservable covariate and the error term.

For proportional data, where the response variable is confined to the interval (0,1), ignoring this may result in misleading conclusions. A review of models for proportional data is given in Kieschnick and McCullough (2003) and they suggest beta or simplex distributions for proportion data. The simplex distribution (Barndorff-Nielsen and Jørgensen 1991), denoted $S^{-}(\mu, \sigma^2)$ has density function

$$p\left(y; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2\{y(1-y)\}^3}} \exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\}, \tag{1}$$

where $\mathrm{E}(Y) = \mu \in (0,1)$ is the mean parameter, $\sigma^2 > 0$ is the dispersion parameter, $d(y, \mu)$ is the deviance,

$$d(y; \mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2},$$

and the variance function is $\nu(\mu) = \mu^3(1-\mu)^3$. The simplex distribution is more flexible than the beta distribution, and can accommodate large left and right skewness. Figure 1 displays some possible shapes of the density function (1) which cannot be captured by the beta density.

Assume we have n independent observations $y_i, i = 1, 2, ..., n$ from a simplex distribution with parameters (μ_i, σ_i^2). The simplex regression model is defined by (1) with $g(\mu_i) = \boldsymbol{z}_i^\top \boldsymbol{\alpha}$ and $h(\sigma_i^2) = \boldsymbol{v}_i^\top \boldsymbol{\delta}$, where $\boldsymbol{z}_i$ and $\boldsymbol{v}_i$, of dimension $1 \times p_1$ and $1 \times p_2$, respectively, are covariates. We assume $\boldsymbol{\alpha} \in \mathbb{R}^{p_1}$ and $\boldsymbol{\delta} \in \mathbb{R}^{p_2}, p_1 + p_2 < n$, and $g(\cdot)$ and $h(\cdot)$ are known monotonic link functions. The simplex regression model is used in Qiu et al. (2008), Song et al. (2004) and in Zhang and Wei (2008). An R (R Core Team 2018) a package simplexreg (Zhang et al. 2016) is available.

We approach the regression model for the simplex distribution under the structural model for errors in variables: we assume a probability distribution for the mismeasurement covariable. We consider both normal and gamma distributions for the error. We discuss the Monte Carlo EM algorithm (Wei and Tanner 1990) to find a maximum pseudo-likelihood estimate, as in Guolo (2011), Carrasco et al. (2014) and Skrondal and Kuha (2012). These last author also showed which the calibration regression estimators are little biased, so we does not used this method to find estimates.

This paper is organized as follows, Sec. 2 introduces the measurement error model based on the simplex distribution and in Sec. 3 we describe inference methodology for the structural errors-in-variables model in which the covariate $\boldsymbol{x}$ can be observed only via a proxy $\boldsymbol{w}$. Results for the normal and gamma model for $\boldsymbol{x}$ are presented in Sec. 4. A simulation study is presented in Sec. 5 and the model is illustrated on a real data set in Sec. 6. Several conclusions are presented in Sec. 7.

2. Simplex error-in-variables regression models

In practice, some explanatory variables may not be directly observed and could be obtained with errors. We extend the simplex regression model of the introduction by

$$g(\mu_i) = \boldsymbol{z}_i^\top \boldsymbol{\alpha} + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad h\left(\sigma_i^2\right) = \boldsymbol{v}_i^\top \boldsymbol{\delta} + \boldsymbol{m}_i^\top \gamma, \tag{2}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{q_1}$ and $\boldsymbol{m}_i \in \mathbb{R}^{q_2}$ are unobserved latent covariates and $\boldsymbol{\beta} \in \mathbb{R}^{q_1}$ and $\gamma \in \mathbb{R}^{q_2}$ are the unknown regression coefficients.

The structural model assumes a probability distribution for the unobserved covariates. Let $\boldsymbol{w}_i = (w_{1i}, ..., w_{q_1 i})^\top$ and $\boldsymbol{x}_i = (x_{1i}, ..., x_{q_1 i})^\top$ be the observed and unobserved variables, respectively. The additive measurement error model is

$$\boldsymbol{w}_i = \boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 \times \boldsymbol{x}_i + \boldsymbol{e}_i, \quad i = 1, ..., n, \tag{3}$$

where $\boldsymbol{e}_i$ is the vector of random errors, $\times$ is the Hadamard product, and $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_1$ are unknown parameters, called additive and multiplicative bias respectively, in Carrasco et al. (2014). If $\boldsymbol{\tau}_0 = \boldsymbol{0}$ and $\boldsymbol{\tau}_1 = \boldsymbol{1}$, (3) is the classical structural model, $\boldsymbol{w}_i = \boldsymbol{x}_i + \boldsymbol{e}_i$.

The multiplicative error structural model is

$$\boldsymbol{w}_i = \boldsymbol{x}_i \times \boldsymbol{e}_i, \quad i = 1, ..., n, \tag{4}$$

which is a classical additive error model after a logarithm transformation. Eckert et al. (1997) consider a general transformed additive error model, i.e. $p(\boldsymbol{w}_i) = p(\boldsymbol{x}_i) + \boldsymbol{e}_i$, where $p(\cdot)$ is a monotone transformation function.

An approach we do not consider is the Berkson error structural models (Berkson 1950), where $\boldsymbol{x}_i = \boldsymbol{w}_i + \boldsymbol{e}_i$, see for example, (Kerber et al. 1993; Rudemo et al. 1989). In all these models $\boldsymbol{e}_i$ is assumed to have independent components and to be independent of the true covariate $\boldsymbol{x}_i$. The vector $\boldsymbol{e}_i$ is also assumed to be independent of any other covariates $\boldsymbol{z}_i$ and of the response variable y_i. This implies a non-differential measurement error model, meaning that y_i and $\boldsymbol{w}_i$ are conditionally independent given $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$.

3. Statistical inference

Suppose a random sample, $(y_1, \boldsymbol{w}_1), ..., (y_n, \boldsymbol{w}_n)$, of size n is observed. We omit the vector of variables $\boldsymbol{z}_i$ from the notation as they are known and fixed. The joint density of $(y_i, \boldsymbol{w}_i)$, observed for the i-th individual, is obtained by integrating the joint density of the complete data $(y_i, \boldsymbol{w}_i, \boldsymbol{x}_i)$, $f(y_i, \boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = f(y_i|\boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_1) f(\boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_2) f(\boldsymbol{x}_i; \boldsymbol{\theta}_3)$, with respect to $\boldsymbol{x}_i$, where the complete parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)^\top$. Following Clayton (1992) the function $f(y_i, \boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ can be view as one function with three parts: (*i*) an outcome function $f(y_i|\boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_1)$, (*ii*) a measurement function $f(\boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_2)$ and (*iii*) an exposure function $f(\boldsymbol{x}_i; \boldsymbol{\theta}_3)$. The logarithm of the likelihood function for the sample of n observations is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int_{\chi} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_1) f(\boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_2) f(\boldsymbol{x}_i; \boldsymbol{\theta}_3) d\boldsymbol{x}_i, \tag{5}$$

where $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_1)$ will be the simplex distribution defined in (1), $f(\boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_2)$ is the conditional distribution of $\boldsymbol{w}_i$ given $\boldsymbol{x}_i$ and $f(\boldsymbol{x}_i; \boldsymbol{\theta}_3)$ is the marginal density of $\boldsymbol{x}_i$.

Usually, the likelihood function in (5) is analytically intractable and it is necessary to use approximations to the integral, for example, Monte Carlo, Gaussian quadrature, stochastic approximation algorithm, etc.

We use the maximum pseudo-likelihood technique, as simulation studies in Carrasco et al. (2014), Skrondal and Kuha (2012) and Guolo (2011) showed that the maximum pseudo-likelihood estimation method provides good asymptotic properties for the estimators. Moreover, the maximum pseudo-likelihood estimate is less computationally intensive. Let $\boldsymbol{\theta}_1$ be the vector of parameters of interest and $\boldsymbol{\theta}_{23} = (\boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ the vector of nuisance parameters, with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_{23})$. The maximum pseudo-likelihood estimation method replaces the vector of nuisance parameters with a consistent estimate in the original likelihood function, thereby generating a pseudo-likelihood function. Then, estimates of the parameters of interest are obtained by using a reliable method such as maximum likelihood (Carrasco et al. 2014; Gong and Samaniego 1981; Guolo 2011; Skrondal and Kuha 2012).

Following Guolo (2011) and Skrondal and Kuha (2012), we estimate the nuisance parameters $\boldsymbol{\theta}_{23}$ by maximizing

$$\ell_r(\boldsymbol{\theta}_{23}) = \sum_{i=1}^{n} \log \int_{\chi} f(\boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_2) f(\boldsymbol{x}_i; \boldsymbol{\theta}_3) d\boldsymbol{x}_i = \sum_{i=1}^{n} \log f(\boldsymbol{w}_i; \boldsymbol{\theta}_{23}), \tag{6}$$

the reduced log-likelihood function. The estimator $\hat{\boldsymbol{\theta}}_{23}$ that maximizes $\ell_r(\boldsymbol{\theta}_{23})$ can be obtained easily using some standard software. The estimate is consistent for $n \to \infty$ under mild regularity conditions (Gourieroux and Monfort 1995a). Moreover, the estimator $\hat{\boldsymbol{\theta}}_{23}$ is asymptotically distributed as a multivariate normal distribution with $\boldsymbol{\theta}_{23}$ mean and covariance matrix $\Sigma^{-1}_{(23,23)} = \mathrm{E}^{-1}[-\partial \ell^2_r(\boldsymbol{\theta}_{23})/\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_3^\top]$. The second step consists in inserting the estimate $\hat{\boldsymbol{\theta}}_{23} = (\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_3)^\top$ obtained in (6) into the log-likelihood function (5)

$$\ell_p(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \log \int_{\chi} f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_1) f\left(\boldsymbol{w}_i|\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_2\right) f\left(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_3\right) d\boldsymbol{x}_i. \tag{7}$$

The estimator of $\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1$, that maximizes $\ell_p(\boldsymbol{\theta}_1)$ is consistent and asymptotically normally distributed (Gong and Samaniego 1981; Gourieroux and Monfort 1995b; Parke 1986). Let $\boldsymbol{U}(\boldsymbol{\theta}) = \partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ be the score function, partitioned as $\boldsymbol{U}(\boldsymbol{\theta}) = (\boldsymbol{U}_{\theta_1}(\boldsymbol{\theta})^\top, \boldsymbol{U}_{\theta_{23}}(\boldsymbol{\theta})^\top)^\top$, and define the mean score $\bar{\boldsymbol{U}}(\boldsymbol{\theta}) = n^{-1}\boldsymbol{U}(\boldsymbol{\theta})$. Let the true parameter value $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^{*\top}, \boldsymbol{\theta}_{23}^{*\top})^\top$. The Fisher information matrix is

$$\mathrm{I}(\boldsymbol{\theta}^*) = \lim_{n\to\infty} \mathrm{E}_{\boldsymbol{\theta}^*}\left[-\frac{\partial \bar{\boldsymbol{U}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right] = \begin{bmatrix} \mathrm{I}_{(1,1)}(\boldsymbol{\theta}^*) & \mathrm{I}_{(1,23)}(\boldsymbol{\theta}^*) \\ \mathrm{I}_{(23,1)}(\boldsymbol{\theta}^*) & \mathrm{I}_{(23,23)}(\boldsymbol{\theta}^*) \end{bmatrix},$$

with partitions corresponding to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_{23}$. It is further assumed that

$$\sqrt{n}\begin{bmatrix} \bar{\boldsymbol{U}}_{\theta_1}(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_{23}^*) \\ \left(\hat{\boldsymbol{\theta}}_{23} - \boldsymbol{\theta}_{23}^*\right) \end{bmatrix} \to N\left(\boldsymbol{0}, \begin{bmatrix} \mathrm{I}_{(1,1)} & \Sigma_{(1,23)} \\ \Sigma_{(23,1)} & \Sigma_{(23,23)} \end{bmatrix}\right).$$

It follow that, the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*)$ is normal with mean zero and variance matrix $\Sigma = \mathrm{I}_{(1,1)}^{-1} + \mathrm{I}_{(1,1)}^{-1}\mathrm{I}_{(23,1)}^\top \Sigma_{(23,23)}^{-1}\mathrm{I}_{(23,1)}\mathrm{I}_{(1,1)}^{-1}$, where $\mathrm{I}_{(1,1)}^{-1}$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_1$ when $\boldsymbol{\theta}_{23}$ is known. The matrix $\mathrm{I}_{(23,1)}$ can be estimated by

$$\mathrm{I}_{(23,1)} = n^{-1}\sum_{i=1}^{n} \boldsymbol{U}_{\theta_{23,i}}(\hat{\boldsymbol{\theta}})\boldsymbol{U}_{\theta_{1,i}}(\hat{\boldsymbol{\theta}})^\top,$$

where $\boldsymbol{U}_{\theta_{23,i}}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{U}_{\theta_{1,i}}(\hat{\boldsymbol{\theta}})$ are the gradients of the reduced and pseudo log-likelihood function for subject i, evaluated at the parameter estimates, respectively. An estimate of $\Sigma_{(23,23)}^{-1}$ can be obtained from the hessian of $\ell_r(\boldsymbol{\theta}_{23})$. For the estimation of $\boldsymbol{\theta}_1$ in (7) we use the Monte Carlo EM algorithm as in Booth and Hobert (1999) and Guolo (2011).

As Guolo (2011), we propose an EM-type algorithm by defining the Monte Carlo estimate, in which

$$\hat{Q}_p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(r)}\right) = M^{-1}\sum_{m=1}^{M}\sum_{i=1}^{n} \kappa_{mi}^{(r)} \log f\left(y_i|\boldsymbol{w}_i, \boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1\right),$$

where $\boldsymbol{x}_{1i}^{*(r)}, \ldots, \boldsymbol{x}_{Mi}^{*(r)}$ are M random samples from $f(\boldsymbol{x}_i|y_i, \boldsymbol{w}_i; \boldsymbol{\theta}_1^{(r)}, \hat{\boldsymbol{\theta}}_{23})$ and $\boldsymbol{\theta}_1^{(r)}$ denotes the value of $\boldsymbol{\theta}$ from the rth iteration. The specification of $f(\boldsymbol{x}_i|y_i, \boldsymbol{w}_i; \boldsymbol{\theta}_1^{(r)}, \hat{\boldsymbol{\theta}}_{23})$ is usually difficult or even impractical in measurement error problems. Guolo (2011) proposed importance sampling, where random samples $\boldsymbol{x}_m^*, m = 1, \ldots, M$ are generated from the importance density $f(\boldsymbol{x}_i; \cdot)$ or $f(\boldsymbol{x}_i|\boldsymbol{w}_i; \cdot)$, assumed known. Then the weight for the ith observation is

$$\begin{aligned}
\kappa_{mi}^{(r)} &= \frac{f\left(\boldsymbol{x}_{mi}^{*(r)}|y_i, \boldsymbol{w}_i; \boldsymbol{\theta}_1^{(r)}\right)}{f\left(\boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1^{(r)}\right)}, \\
&= \frac{f\left(y_i|\boldsymbol{w}_i, \boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1^{(r)}\right) f\left(\boldsymbol{w}_i|\boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1^{(r)}\right)}{\int f\left(y_i|\boldsymbol{w}_i, \boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1^{(r)}\right) f\left(\boldsymbol{w}_i|\boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1^{(r)}\right) f\left(\boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1^{(r)}\right) d\boldsymbol{x}_{mi}^{*(r)}}, \\
&\approx \frac{f\left(y_i|\boldsymbol{w}_i, \boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1^{(r)}\right)}{M^{-1}\sum_{m=1}^{M} f\left(y_i|\boldsymbol{w}_i, \boldsymbol{x}_{mi}^{*(r)}; \boldsymbol{\theta}_1^{(r)}\right)}.
\end{aligned}$$

To simplify the description of the M-step, in the simplex regression model with measurement error, we assume that $\boldsymbol{m}_i = \boldsymbol{x}_i$ in (2). We need to maximize

$$\hat{Q}_p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(r)}\right) = M^{-1}\sum_{m=1}^{M}\sum_{i=1}^{n} \kappa_{mi}^{(r)} \ell_i\left(\mu_i^{(r)}, \sigma_i^{2(r)}\right), \tag{8}$$

where

$$\ell_i\left(\mu_i^{(r)}, \sigma_i^{2(r)}\right) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_i^{2(r)}\right) - \frac{3}{2}\log(y_i(1-y_i)) - \frac{1}{2\sigma_i^{2(r)}} d\left(y_i, \mu_i^{(r)}\right),$$

with $d(y_i, \mu_i^{(r)})$ as defined in (1), $\mu_i^{(r)} = g^{-1}(\boldsymbol{z}_i^\top \boldsymbol{\alpha} + \boldsymbol{x}_{mi}^{\top *(r)} \boldsymbol{\beta})$ and $\sigma_i^{2(r)} = h^{-1}(\boldsymbol{v}_i^\top \boldsymbol{\delta} + \boldsymbol{x}_{mi}^{\top *(r)} \boldsymbol{\gamma})$. We can obtain the score vector for the parameters of interest from $\hat{Q}_p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(r)})$: for $j = 1, ..., p_1$ and $j' = 1, ..., q_1$,

$$U_{\alpha_j}(\boldsymbol{\theta}_1) = \frac{\partial \hat{Q}_p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(r)}\right)}{\partial \alpha_j} = \frac{M^{-1}}{\sigma^2} \sum_{i=1}^{n} \sum_{m=1}^{M} \kappa_{mi}^{(r)} u_i^{(r)} \frac{1}{g'\left(\mu_i^{(r)}\right)} z_{ij},$$

$$U_{\beta_{j'}}(\boldsymbol{\theta}_1) = \frac{\partial \hat{Q}_p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(r)}\right)}{\partial \beta_{j'}} = \frac{M^{-1}}{\sigma^2} \sum_{i=1}^{n} \sum_{m=1}^{M} \kappa_{mi}^{(r)} u_i^{(r)} \frac{1}{g'\left(\mu_i^{(r)}\right)} x_{(m,i)j'}^{*(r)},$$

where $u_i^{(r)} = -d'(y_i, \mu_i^{(r)})/2, d'(y_i, \mu_i^{(r)}) = \partial d(y_i, \mu_i)/\partial \mu_i$ with

$$d'\left(y_i, \mu_i^{(r)}\right) = -\frac{2\left(y_i - \mu_i^{(r)}\right)}{y_i(1-y_i)\mu_i^{2(r)}\left(1-\mu_i^{2(r)}\right)} \left(1 + \frac{\left(y_i - \mu_i^{(r)}\right)\left(1 - 2\mu_i^{2(r)}\right)}{\mu_i^{(r)}\left(1-\mu_i^{2(r)}\right)}\right).$$

For $t = 1, ..., p_2$ and $t' = 1, ..., q_2$,

$$U_{\delta_t}(\boldsymbol{\theta}_1) = \frac{\partial \hat{Q}_p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(r)}\right)}{\partial \delta_j} = \frac{1}{M} \sum_{i=1}^{n} \sum_{m=1}^{M} -\frac{\kappa_{(m,i)}^{(r)}}{2\sigma_i^{2(r)}} \left(1 - \frac{d\left(y_i, \mu_i^{(r)}\right)}{\sigma_i^{2(r)}}\right) \left(\frac{1}{h'\left(\sigma_i^{2(r)}\right)}\right) v_{it},$$

$$U_{\gamma_{t'}}(\boldsymbol{\theta}_1) = \frac{\partial \hat{Q}_p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(r)}\right)}{\partial \gamma_{j'}} = \frac{1}{M} \sum_{i=1}^{n} \sum_{m=1}^{M} -\frac{\kappa_{(m,i)}^{(r)}}{2\sigma_i^{2(r)}} \left(1 - \frac{d\left(y_i, \mu_i^{(r)}\right)}{\sigma_i^{2(r)}}\right) \left(\frac{1}{h'\left(\sigma_i^{2(r)}\right)}\right) x_{(m,i)t'}^{*(r)}.$$

Solving simultaneously the equations $U_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_1) = \boldsymbol{0}, U_{\boldsymbol{\beta}}(\boldsymbol{\theta}_1) = \boldsymbol{0}$ and $U_{\boldsymbol{\sigma}}(\boldsymbol{\theta}_1) = \boldsymbol{0}$ we obtain the pseudo-maximum likelihood estimator for $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\sigma}$. If $\sigma_1^2 = ... = \sigma_n^2 = \sigma^2$, then

$$U_\sigma(\boldsymbol{\theta}_1) = \frac{\partial \hat{Q}_p\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(r)}\right)}{\partial \sigma^2} = \frac{1}{M} \sum_{i=1}^{n} \sum_{m=1}^{M} -\frac{\kappa_{mi}^{(r)}}{2\sigma^2} \left(1 - \frac{d\left(y_i, \mu_i^{(r)}\right)}{\sigma^2}\right).$$

An estimator of the dispersion parameter is

$$\hat{\sigma}^2 = \frac{\sum_{m=1}^{M} \sum_{i=1}^{n} \kappa_{mi}^{(r)} d\left(y_i, \hat{\mu}_i^{(r)}\right)}{M \sum_{m=1}^{M} \kappa_{mi}^{(r)}}.$$

The variance - covariance matrix Σ defined in Sec. 3 is calculated using the approach of Louis (1982). The matrix $I_{(1,1)}$ can be estimated using the expressions

$$\begin{aligned} I_{(1,1)}^1 &= -\frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^\top} Q\left(\boldsymbol{\theta}_1|\hat{\boldsymbol{\theta}}_1\right), \\ &= -\frac{1}{M} \sum_{i=1}^{n} \sum_{m=1}^{M} \kappa_{mi}^{\circ} \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^\top} \ell_p\left(\boldsymbol{\theta}_1; y_i, \boldsymbol{w}_i, \boldsymbol{x}_{mi}^{\circ}, \hat{\boldsymbol{\theta}}_{23}\right), \end{aligned}$$

$$
\begin{aligned}
I^2_{(1,1)} = \sum_{i=1}^{n} & \left\{ \frac{1}{M} \sum_{m=1}^{M} \kappa^{\circ}_{mi} \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell_p\left(\boldsymbol{\theta}_1; y_i, \boldsymbol{w}_i, \boldsymbol{x}^{\circ}_{mi}, \hat{\boldsymbol{\theta}}_{23}\right)\Big|_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1} \right\} \\
& \times \left\{ \frac{1}{M} \sum_{m=1}^{M} \kappa^{\circ}_{mi} \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell_p\left(\boldsymbol{\theta}_1; y_i, \boldsymbol{w}_i, \boldsymbol{x}^{\circ}_{mi}, \hat{\boldsymbol{\theta}}_{23}\right)\Big|_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1} \right\}^{\top}, \\
I^3_{(1,1)} = - \frac{1}{M} & \sum_{i=1}^{n} \sum_{m=1}^{M} \kappa^{\circ}_{mi} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell_p\left(\boldsymbol{\theta}_1; y_i, \boldsymbol{w}_i, \boldsymbol{x}^{\circ}_{mi}, \hat{\boldsymbol{\theta}}_{23}\right) \right\} \\
& \times \left\{ \frac{\partial}{\partial \boldsymbol{\theta}_1} \ell_p\left(\boldsymbol{\theta}_1; y_i, \boldsymbol{w}_i, \boldsymbol{x}^{\circ}_{mi}, \hat{\boldsymbol{\theta}}_{23}\right) \right\}^{\top} \Bigg|_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1},
\end{aligned}
$$

where $\boldsymbol{x}^{\circ}_{mi}$ and κ°_{mi} are the random importance sample and importance weight of the Monte Carlo EM algorithm for the ith subject when the algorithm has converged. The matrix $I_{(23,1)} = I_{(1,23)}$ can be approximated similarly. We used the Package numDeriv[1] in R software to calculate the first and second partial derivatives.

4. Normal and gamma measurement error models

In this section, we consider normal and gamma distributions for a single covariate measured with error. We assume $f(e_i; \cdot)$ known to avoid nonidentifiability problems.

4.1. Normal measurement error models

This additive measurement error model, $w_i = x_i + e_i$ with $e_i \sim N(0, \sigma_e^2)$, follows the hierarchical specification

$$
\begin{aligned}
y_i | \boldsymbol{z}_i, w_i, x_i &\sim S^{-}\left(\mu_i, \sigma_i^2\right), \\
w_i | x_i &\sim N\left(x_i, \sigma_e^2\right), \\
x_i &\sim N\left(\mu_x, \sigma_x^2\right),
\end{aligned} \tag{9}
$$

where $g(\mu_i)$ and $h(\sigma_i^2)$ are defined above. The reduced log-likelihood function (6) is

$$
\begin{aligned}
\ell_r(\boldsymbol{\theta}_{23}) &= \sum_{i=1}^{n} \log f\left(w_i | x_i; \sigma_e^2\right) f\left(x_i; \mu_x, \sigma_x^2\right), \\
&= -\frac{n}{2} \log\left[2\pi\left(\sigma_x^2 + \sigma_e^2\right)\right] - \frac{1}{2\left(\sigma_x^2 + \sigma_e^2\right)} \sum_{i=1}^{n} (w_i - \mu_x)^2,
\end{aligned}
$$

where $\boldsymbol{\theta}_{23} = (\boldsymbol{\theta}_2, \boldsymbol{\theta}_3)^{\top}$, with $\boldsymbol{\theta}_2 = \sigma_e^2$ and $\boldsymbol{\theta}_3 = (\mu_x, \sigma_x^2)^{\top}, \sigma_e^2$ is assumed known, or estimated from supplementary information, such as replicate measurements or partial observation of the error-free covariate (Carroll et al. 2006). The maximum likelihood estimate $\hat{\boldsymbol{\theta}}_3$ of the nuisance parameters solves $\partial \ell_r(\boldsymbol{\theta}_{23}) / \partial \mu_x = \boldsymbol{0}$ and $\partial \ell_r(\boldsymbol{\theta}_{23}) / \partial \sigma_x^2 = \boldsymbol{0}$. Thereby, $\hat{\mu}_x = \bar{w}$ and $\hat{\sigma}_x^2 = n^{-1} \sum_{i=1}^{n} (w_i - \bar{w})^2 - \hat{\sigma}_e^2$ with $\bar{w} = n^{-1} \sum_{i=1}^{n} w_i$. We substitute $\hat{\boldsymbol{\theta}}_3$ into the log-likelihood function (5), giving the pseudo-log-likelihood function (7)

[1]https://cran.r-project.org/web/packages/numDeriv/index.html

that depends on the parameters of interest only. We use the Monte Carlo EM algorithm, described in (3), to estimate $\boldsymbol{\theta}_1$.

4.2. Gamma measurement error models

We consider using a gamma distribution for the true covariate, appropriate for example for skewed data on $\mathbb{R}^+$. Following Sen et al. (2014), we defined a simplex regression model with gamma distribution for a single exposure variable. We suppose the following hierarchical specification

$$\begin{aligned} y_i|z_i, w_i, x_i &\sim S^-(\mu_i, \sigma_i^2), \\ w_i|x_i &\sim Ga(x_i, \phi_e), \\ x_i &\sim Ga(\mu_x, \phi_x), \end{aligned} \tag{10}$$

where $Ga(a, b)$ represents a gamma distribution with mean $a > 0$ and coefficient of variation $b > 0$. If $b \to \infty$, them the random variable y follows a normal distribution. To obtain the model (10), we assumed that $e_i \overset{\text{ind}}{\sim} Ga(1, \phi_e)$ in (4). The joint distribution $f(w_i, x_i; \boldsymbol{\theta}_{23}) = f(w_i|x_i; \boldsymbol{\theta}_2) f(x_i; \boldsymbol{\theta}_3)$ is

$$\begin{aligned} f(w_i, x_i; \boldsymbol{\theta}_{23}) = & \frac{\phi_e^{\phi_e}}{\Gamma(\phi_e)\Gamma(\phi_x)} \left(\frac{\phi_x}{\mu_x}\right)^{\phi_x} x_i^{\phi_x - \phi_e - 1} w_i^{\phi_e - 1} \\ & \times \exp\left(-\phi_e \frac{w_i}{x_i} - \frac{\phi_x}{\mu_x} x_i\right), \end{aligned} \tag{11}$$

where $\Gamma(\cdot)$ is the gamma function, $\boldsymbol{\theta}_{23} = (\boldsymbol{\theta}_2, \boldsymbol{\theta}_3)^\top$, with $\boldsymbol{\theta}_2 = \phi_e$ and $\boldsymbol{\theta}_3 = (\mu_x, \phi_x)^\top$. The marginal distribution is

$$f(w_i; \boldsymbol{\theta}_{23}) = \int_0^\infty f(w_i, x_i; \boldsymbol{\theta}_{23}) dx_i = \int_0^\infty f(w_i|x_i; \boldsymbol{\theta}_2) f(x_i; \boldsymbol{\theta}_3) dx_i.$$

Using the transformation $t_i = \phi_x x_i/\mu_x$ and assuming $\delta = \phi_e\phi_x/\mu_x$,

$$\begin{aligned} f(w_i; \boldsymbol{\theta}_{23}) &= \left(\frac{\phi_x\phi_e}{\mu_x}\right)^{\phi_e} \frac{1}{\Gamma(\phi_x)\Gamma(\phi_e)} w_i^{\phi_e - 1} \int_0^\infty t_i^{\phi_x - \phi_e - 1} \exp\left(-t_i - \frac{\delta w_i}{t_t}\right) dt_i, \\ &= \frac{2}{\Gamma(\phi_e)\Gamma(\phi_x)} \left(\frac{\phi_x\phi_e}{\mu_x}\right)^{\frac{1}{2}(\phi_x + \phi_e)} w_i^{\frac{1}{2}(\phi_x + \phi_e) - 1} \\ &\quad \times K_{\phi_x - \phi_e}\left(2\sqrt{\frac{\phi_x\phi_e}{\mu_x} w_i}\right), \end{aligned} \tag{12}$$

where $w_i > 0$ and $K_b(a)$ is the modified Bessel function of the third kind (Abramowitz and Stegun 1972).

Proposition 1. *Let x and e be independent random variables such that $x \sim Ga(\mu_x, \phi_x)$ and $e \sim Ga(1, \phi_e)$, with $\mu_x > 0, \phi_x > 0$ and $\phi_e > 0$. Let $w = xe$. Then*

(i) $\mathrm{E}(w) = \mu_x$,
(ii) $\mathrm{Var}(w) = (1 - \phi_x + \phi_e)\mu_x^2/\phi_x\phi_e$,
(iii) $\mathrm{Cov}(w) = \mu_x^2/\phi_x$.

From Proposition 1, to avoid nonidentifiability problems, we can easily calculate an estimate of ϕ_e, say $\hat{\phi}_e$, as

$$\hat{\phi}_e = \left[\left(s_w^2\hat{\phi}_x/\bar{w}^2\right)-1\right]/\left(1+\hat{\phi}_x\right), \tag{13}$$

where $\bar{w} = \sum_{i=1}^{n} w_i/n$ and $s_w^2 = \sum_{n}^{i=1} (w_i - \bar{w})^2/(n-1)$.

As in Sec. 4.1, assuming ϕ_e known or estimating by (13), the reduced log-likelihood function

$$\begin{aligned}\ell_r(\boldsymbol{\theta}_{23}) = \sum_{i=1}^{n}\Bigg\{ & \log(2) + \frac{1}{2}\log(\phi_x) + \frac{1}{2}\log(\phi_e) - \frac{1}{2}\log(\mu_x) \\ & - \log\Gamma(\phi_x) - \log\Gamma(\phi_e) + \left(\frac{1}{2}(\phi_x+\phi_e)-1\right)\log(w_i) \\ & + \log\left(K_{\phi_x-\phi_e}\left(2\sqrt{\frac{\phi_x\phi_e}{\mu_x}w_i}\right)\right)\Bigg\}. \end{aligned} \tag{14}$$

We then substitute $\hat{\boldsymbol{\theta}}_2$ and $\hat{\phi}_e$ into the log-likelihood function to obtain the pseudo-likelihood, and use the Monte Carlo EM algorithm to estimate $\boldsymbol{\theta}_1$. The importance density in this case can be $f(x_i;\boldsymbol{\theta}_2)$ or $f(x_i|w_i;\boldsymbol{\theta}_3)$. The conditional density $f(x_i|w_i;\boldsymbol{\theta}_{23})$ can be derived from (11) and (12):

$$\begin{aligned} f(x_i|w_i;\boldsymbol{\theta}_{23}) = & \frac{1}{2}\left(\frac{\phi_x}{\phi_e\mu_x}\right)^{\frac{1}{2}(\phi_x-\phi_e)} x_i^{\phi_x-\phi_e-1} w_i^{-\frac{1}{2}(\phi_x-\phi_e)} \\ & \times \exp\left(-\phi_e\frac{w_i}{x_i} - \frac{\phi_x}{\mu_x}x_i\right) K^{-1}_{\phi_x-\phi_e}\left(2\sqrt{\frac{\phi_x\phi_e}{\mu_x}w_i}\right). \end{aligned}$$

An alternative approach to inference uses (7) directly, approximating the integral numerically, for instance, by Laguerre-Gauss quadrature. The Laguerre-Gauss quadrature approximation is

$$\int_0^\infty e^{-x}f(x)dx \approx \sum_{q=1}^{Q} \nu_q f(\breve{x}_q),$$

where ν_q and $\breve{x}_q$ represent the q-th weight and zero, respectively, of the orthogonal Laguerre polynomial of order Q, where Q is number of quadrature points; see, for example, Abramowitz and Stegun (1972). We can write the pseudo-likelihood in (7) as

$$\begin{aligned} \ell_p(\boldsymbol{\theta}_1) &= \sum_{i=1}^{n} \log \int_0^\infty f\left(y_i, w_i, x_i; \boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{23}\right) dx_i \\ &= \sum_{i=1}^{n} \log \int_0^\infty f(y_i|x_i;\boldsymbol{\theta}_1) f\left(x_i|w_i;\hat{\boldsymbol{\theta}}_{23}\right) f\left(w_i;\hat{\boldsymbol{\theta}}_{23}\right) dx_i \\ &\approx \sum_{i=1}^{n} \log f\left(w_i;\hat{\boldsymbol{\theta}}_{23}\right) + \sum_{i=1}^{n} \log\left\{\sum_{q=1}^{Q} \omega_q \frac{f(y_i|\breve{x}_q;\boldsymbol{\theta}_1) f\left(\breve{x}_q|w_i;\hat{\boldsymbol{\theta}}_{23}\right)}{\exp(\breve{x}_q)}\right\}, \end{aligned}$$

where $\hat{\boldsymbol{\theta}}_{23}$ is find maximized (14). The approximate pseudo-likelihood estimator of $\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1$, is obtained by maximizing the approximate pseudo-log-likelihood function given above. We leave this alternative approach for future studies.

5. Numerical studies

The simulation study presented in this section is carried out to understand the asymptotic behavior of the estimators obtained by using the maximum pseudo-likelihood method. We consider two scenarios with the systematic part of the model given by $g(\mu_i) = \alpha_0 + \alpha_1 z_i + \beta x_i, h(\sigma_i^2) = \delta x_i$, where $g(\cdot)$ and $h(\cdot)$ are the logistic and log-link functions, respectively. In the first scenario, we assume that the variable x_i follows a normal distribution with mean μ_x and variance σ_x^2 and it is structured as shown in §4.1; the true values for the parameters are $\alpha_0 = -0.5, \alpha_1 = 1.0, \beta = 0.5, \delta = 3.0, \mu_x = 0.5, \sigma_x^2 = 0.1$. We also study a case where the dispersion parameter is constant, i.e. $h(\sigma^2) = \delta$. In the second scenario, we consider $x_i \sim Ga(\mu_x, \phi_x)$ as in §4.2, with true values $\alpha_0 = -2.0, \alpha_1 = 0.0, \beta = 0.05, \delta = 5.0, \mu_x = 3.0$, and $\phi_x = 2.0$. The parameters of the measurement error mechanism are known, and we set $\sigma_e^2 = 0.0333$, 0.0052 which corresponds moderate measurement error and low measurement error and $\phi_e = 0.1$, 1.0 in the first and second scenario, respectively. The sample sizes are $n = 25$, 50, 75 and 100, and for the Monte Carlo EM algorithm $M = 120$. All simulation results are based on 1000 (Monte Carlo) replications. We determine the bias, and root mean square error (RMSE) of the estimators. Maximization was performed using the quasi-Newton BFGS method implemented in the function optim the software R. The results are compared with the naïve analysis, ignoring the presence of measurement error. Tables 1 and 2 provide the results obtained for the first scenario. These tables show the superiority of the pseudo-likelihood method compared to the naïve method. In this situation, the estimator of the naïve methods are biased, specifically for parameter β which is associated with the variable measured with error. In addition, these tables show that as the sample size increases, the maximum pseudo-likelihood estimator become closer to the true values. Table 3 give the results obtained for the second scenario. As expected, the naïve estimator is biased, particularly for small sample size for the parameters δ and β, the latter of which is associated with the variable measured with error. However, the root mean square error (RMSE) for parameter β it is a little bigger than naïve estimator, the RMSE of the maximum pseudo-likelihood estimator decreases as the sample size increases. Overall, we conclude that ignoring the measurement error produces misleading inference. Inference based on the pseudo-likelihood methods presents good performance.

6. Data analysis

In this section, we apply the proposed methods to a data set studied by Silva et al. (2018), who investigated 200 individual from a financial institution in Brazil. We will focus on the analysis of proportion of spending (y) used by the customer on his authorized overdraft limit over a fixed period of time. Two measures are observed, w_1 and w_2, which represent the customer's presumed income, obtained from models available on

Table 1. The Bias and RMSE for a simplex regression model with additive measurement error models, in which $x_i \sim N(\mu_x, \sigma_x^2)$. Constant precision model.

σ_e^2	Method	n	Measure	α_0	α_1	β	$\log(\delta)$
0.0333	ℓ_p	25	Bias	0.026	0.010	−0.065	−0.047
			RMSE	0.191	0.188	0.113	0.307
		50	Bias	0.024	0.008	−0.064	−0.039
			RMSE	0.157	0.129	0.099	0.272
		75	Bias	0.018	0.004	−0.063	−0.020
			RMSE	0.127	0.103	0.090	0.190
		100	Bias	0.011	0.002	−0.060	−0.020
			RMSE	0.104	0.089	0.091	0.176
	ℓ_{naive}	25	Bias	−0.311	0.001	0.135	0.285
			RMSE	0.453	0.235	0.171	1.644
		50	Bias	−0.304	0.004	0.137	0.794
			RMSE	0.376	0.144	0.153	1.483
		75	Bias	−0.298	0.003	0.137	0.967
			RMSE	0.349	0.103	0.148	1.420
		100	Bias	−0.300	0.000	0.137	1.051
			RMSE	0.338	0.091	0.145	1.381
0.0052	ℓ_p	25	Bias	0.025	−0.002	−0.029	−0.030
			RMSE	0.207	0.131	0.086	0.141
		50	Bias	0.014	0.002	−0.024	−0.023
			RMSE	0.126	0.098	0.060	0.140
		75	Bias	0.010	0.001	−0.020	−0.015
			RMSE	0.108	0.080	0.049	0.134
		100	Bias	0.010	0.000	−0.020	−0.011
			RMSE	0.092	0.064	0.046	0.122
	ℓ_{naive}	25	Bias	−0.075	0.001	0.017	−0.359
			RMSE	0.434	0.167	0.102	1.469
		50	Bias	−0.063	0.003	0.023	−0.172
			RMSE	0.222	0.113	0.071	1.001
		75	Bias	−0.053	0.006	0.025	0.119
			RMSE	0.182	0.108	0.060	0.880
		100	Bias	−0.049	−0.004	0.023	0.099
			RMSE	0.152	0.078	0.050	0.759

the market. These are treated as replicates with $w = (w_1 + w_2)/2$ being the observed mean customer's presumed income. It is reasonable to assume that the customer's presumed income is measured with error. A binary observed covariate which represent the customer's gender also is consider. Our goal is to model the proportion of spending (y) using the real income of a new costumer (x) as a (latent) covariate measured with error.

The Figure 2a shows the histogram of the proportion of spending with fit using simplex distribution defined in 1, we can observe that the possible shape of the fit cannot be captured by the beta distribution. The Figure 2b and c show the histogram of the customer's presumed income mean and the scatter plot between proportion of spending and customer's presumed income mean classify by gender. We consider the model

$$\begin{aligned} y_i|z_i, w_i, x_i &\sim S^-\left(\mu_i, \sigma_i^2\right), \\ \log\left(\mu_i/(1-\mu_i)\right) &= \alpha_0 + \alpha_1 z_i + \beta x_i, \\ \log\left(\sigma_i^2\right) &= \delta + \gamma x_i, \\ w_i|x_i, \sigma_e^2 &\sim N\left(x_i, \sigma_e^2\right), \end{aligned} \tag{15}$$

$$x_i|\mu_x, \sigma_x^2 \sim N\left(\mu_x, \sigma_x^2\right), \quad \text{for } i = 1, \ldots, 200. \tag{16}$$

We calculated $\hat{\sigma}_e^2 = 0.2263$ following Buonaccorsi and Tosteson (1993, p.231) when have replicate to w. The maximum pseudo-likelihood estimates of the vector

Table 2. The Bias and RMSE for a simplex regression model with additive measurement error models, in which $x_i \sim N(\mu_x, \sigma_x^2)$. Variable precision model.

σ_e^2	Method	n	Mearure	α_0	α_1	β	δ
0.0333	ℓ_p	25	Bias	−0.033	−0.014	0.024	−0.049
			RMSE	0.274	0.346	0.373	0.347
		50	Bias	−0.028	−0.011	0.024	−0.043
			RMSE	0.165	0.214	0.217	0.217
		75	Bias	−0.025	−0.010	0.020	−0.019
			RMSE	0.128	0.157	0.181	0.163
		100	Bias	−0.019	−0.010	0.016	−0.008
			RMSE	0.108	0.126	0.143	0.103
	ℓ_{naive}	25	Bias	0.049	0.046	−0.105	−0.270
			RMSE	0.435	0.682	0.508	0.665
		50	Bias	0.059	0.027	−0.109	−0.172
			RMSE	0.260	0.409	0.326	0.446
		75	Bias	0.065	0.001	−0.109	−0.122
			RMSE	0.210	0.316	0.270	0.339
		100	Bias	0.075	−0.009	−0.103	−0.113
			RMSE	0.180	0.256	0.237	0.307
0.0052	ℓ_p	25	Bias	−0.018	−0.007	0.026	−0.052
			RMSE	0.206	0.316	0.308	0.329
		50	Bias	−0.016	−0.006	0.015	−0.034
			RMSE	0.186	0.190	0.236	0.216
		75	Bias	−0.015	0.004	0.012	−0.027
			RMSE	0.115	0.153	0.168	0.170
		100	Bias	−0.013	0.000	0.011	−0.024
			RMSE	0.087	0.134	0.139	0.145
	ℓ_{naive}	25	Bias	−0.011	0.030	0.009	−0.238
			RMSE	0.386	0.617	0.550	0.604
		50	Bias	0.004	0.012	0.006	−0.107
			RMSE	0.216	0.318	0.315	0.383
		75	Bias	0.001	0.020	−0.008	−0.081
			RMSE	0.184	0.303	0.255	0.310
		100	Bias	0.009	0.000	−0.007	−0.055
			RMSE	0.154	0.253	0.215	0.262

of parameters $(\alpha_0, \alpha_1, \beta, \sigma^2)^\top$ is show in Table 4. The Monte Carlo sample size M was 8000. We also show the naïve estimates that can obtained using the simplexreg packages implemented in R (Zhang et al. 2016). We can see in the Figure 2b, some evidence of nonormality for the variable w. Thereby, Table 5 gives some descriptive measures, indicating that the variable w has an asymmetric distribution, so a non-normal distribution seems appropriate. Thus, we can replace (15) and (16) by

$$w_i|x_i, \phi_e \overset{\text{ind}}{\sim} Ga(x_i, \phi_e) \qquad \text{and}$$
$$x_i|\mu_x, \phi_x \overset{\text{ind}}{\sim} Ga(\mu_x, \phi_x).$$

Under this model the maximum pseudo-likelihood of α_0, α_1, β, δ and γ are showing in Table 6. Here, we calculate using the Eq. (13), $\hat{\phi}_e = 0.1664$. To interpret the estimates in Tables 4 and 6, we use the odds ratios, $\exp(c\hat{\beta})$, where c is the increase in units of the continuous variable. For an increase of $c = 1.0$ meters of w, the odds ratios, when assuming normal measurement error, gamma measurement error and no measurement error naïve method are $\exp(1.0 \times -0.5133) = 0.5985$, $\exp(1.0 \times -0.5247) = 0.5917$, $\exp(1.0 \times -0.5653) = 0.5682$, respectively. In other words, the proportion of

Table 3. The Bias and RMSE for a simplex regression model with multiplicative measurement error models, in which ϕ_e= 0.10 and $x_i \sim Ga(\mu_x, \phi_x)$. Constant precision model.

ϕ_e	Method	n	Measure	α_0	α_1	β	$\log(\delta)$
0.10	ℓ_p	25	Bias	−0.086	−0.010	−0.024	−0.082
			RMSE	0.566	0.554	0.370	0.309
		50	Bias	−0.056	−0.004	0.010	−0.046
			RMSE	0.470	0.420	0.285	0.213
		75	Bias	−0.031	−0.003	0.003	−0.033
			RMSE	0.388	0.306	0.196	0.176
		100	Bias	−0.022	−0.001	0.002	−0.030
			RMSE	0.222	0.107	0.190	0.158
	ℓ_{naive}	25	Bias	−0.039	−0.020	−0.100	−0.168
			RMSE	0.833	1.544	0.245	0.352
		50	Bias	0.070	−0.020	−0.071	−0.080
			RMSE	0.517	0.882	0.100	0.222
		75	Bias	0.118	−0.029	−0.064	−0.046
			RMSE	0.366	0.704	0.075	0.174
		100	Bias	0.120	−0.035	−0.059	−0.037
			RMSE	0.371	0.620	0.067	0.149
1.0	ℓ_p	25	Bias	−0.079	−0.032	0.020	−0.055
			RMSE	0.457	0.437	0.242	0.266
		50	Bias	−0.067	−0.025	0.018	−0.038
			RMSE	0.319	0.316	0.142	0.208
		75	Bias	−0.055	−0.010	0.016	−0.037
			RMSE	0.275	0.270	0.108	0.149
		100	Bias	−0.043	−0.009	0.011	−0.021
			RMSE	0.256	0.247	0.066	0.117
	ℓ_{naive}	25	Bias	0.011	−0.112	−0.064	−0.152
			RMSE	0.811	1.357	0.170	0.332
		50	Bias	0.049	0.020	−0.059	−0.072
			RMSE	0.630	1.004	0.106	0.214
		75	Bias	0.096	0.020	−0.053	−0.045
			RMSE	0.440	0.696	0.084	0.167
		100	Bias	0.118	−0.036	−0.049	−0.023
			RMSE	0.388	0.649	0.074	0.138

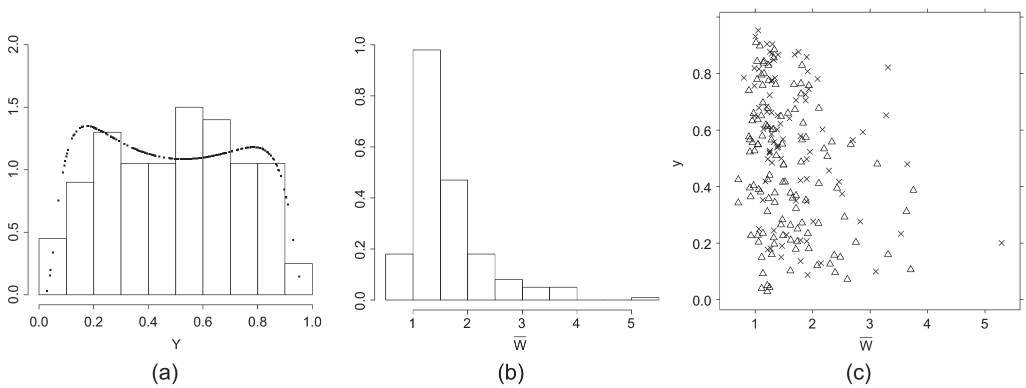

Figure 2. (a) Histogram the proportion of spending with fit simplex distribution, (b) histogram the mean customer's presumed income and (c) Plot of dispersion proportion of spending versus values of mean customer's presumed income by gender (star represent male gender and triangle female gender).

spending (y), decreasing on average by 40.15%, 40.83% and 43.18%. In summary, when we assumes a normal distribution for an unobserved asymmetric variable or ignore measurement error, the interpretation of the odds ratios may change.

Table 4. Estimates and standard errors, for financial data when the normal distribution is assumed for the covariate mismeasurement.

Parameter	Estimate ℓ_p	Stand. error	Estimate ℓ_{naive}	Stand. error
α_0	0.9642	0.1354	0.8783	0.1722
α_1	−0.5318	0.0079	−0.3971	0.0794
β	−0.5133	0.1080	−0.5653	0.1252
δ	2.0423	0.0124	2.5233	0.2609
γ	−0.6012	0.0078	−0.3327	0.1490

Table 5. Summary of the variable mean customer's presumed income.

Minimum	1. Quantile	Median	Mean	3. Quantile
0.7050	1.1919	1.3934	1.6174	1.8820
Maximum	Variance	Stdev	Skewness	Kurtosis
5.2893	0.4526	0.6727	1.8893	4.9610

Table 6. Estimates, standard errors, z stat and p-values for financial data when the gamma distribution is assumed for the covariate mismeasurement.

Parameter	Estimate	Stand. error	z stat	p-value
α_0	1.1974	0.0917	13.0577	0.0000
α_1	−0.4573	0.0275	−16.6290	0.0000
β	−0.5247	0.2383	−2.2018	0.0277
δ	2.3671	0.5321	4.4486	0.0000
γ	−0.1293	0.0671	−2.2644	0.0236

7. Concluding remarks

In this paper we proposed and studied the simplex regression models with measurement error in the covariates. We used a pseudo-likelihood function to estimate the parameter. A Monte Carlo simulation study compared the performance of the estimators in terms of bias and root-mean-square errors and concluding that pseudo-likelihood estimator has good behavior. We considered two distributions for the measurement error model, the normal and the gamma distribution. The gamma distribution is appropriate for a multiplicative error structure. The approach in this paper is easily applied to different distributions for the response variable or different distributions for the covariates with measurement error. For instance, we can consider the skew normal distribution (Azzalini 1985) for the unobserved variable x.

Acknowledgments

The authors are grateful to two anonymous referees and the Editor for very useful comments and suggestions. This work was partially supported by Conselho Nacional de Desenvolvimento Cientco e Tecnolgico (CNPq), Brazil, Programa de Apoio a Pesquisadores Emergentes - UFBA, Brazil and the Department of Statistical Sciences of the University of Toronto, Canada.

References

Abramowitz, M., and I. A. Stegun. 1972. *Handbook of mathematical functions*. New York: Dover.
Azzalini, A. 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12:171–8.

Barndorff-Nielsen, O. E., and B. Jørgensen. 1991. Some parametric models on the simplex. *Journal of Multivariate Analysis* 39 (1):106–16. doi:10.1016/0047-259X(91)90008-P.
Berkson, J. 1950. Are there two regressions? *Journal of the American Statistical Association* 45 (250):164–80. doi:10.1080/01621459.1950.10483349.
Booth, J. G., and J. P. Hobert. 1999. Maximizing generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61 (1):265–85. doi:10.1111/1467-9868.00176.
Buonaccorsi, J. P. 2010. *Measurement error: Models, methods and applications*. London: Chapman and Hall.
Buonaccorsi, J. P., and T. D. Tosteson. 1993. Correcting for nonlinear measurement errors in the dependent variable in the general linear model. *Communications in Statistics - Theory and Methods* 22 (10):2687–702. doi:10.1080/03610929308831179.
Carrasco, J. M. F., S. L. P. Ferrari, and R. B. Arellano-Valle. 2014. Errors-in-variables beta regression models. *Journal of Applied Statistics* 41 (7):1530–47. doi:10.1080/02664763.2014.881784.
Carroll, R. J., and L. A. Stefanski. 1990. Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* 85 (411):652–63. doi:10.1080/01621459.1990.10474925.
Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. *Measurement error in nonlinear models: A modern perspective*. New York: Chapman and Hall.
Cheng, C. L., Shalabh, and G. Garg. 2016. Goodness of fit in restricted measurement error models. *Journal of Multivariate Analysis* 145:101–16.
Clayton, D. G. 1992. *Models for the analysis of cohort and case-control studies with inaccurately measured exposures*. Oxford: University Press.
Cook, J., and L. Stefanski. 1994. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* 89 (428):1314–28. doi:10.1080/01621459.1994.10476871.
Eckert, R. S., R. J. Carroll, and N. Wang. 1997. Transformations to additivity in measurement error models. *Biometrics* 53 (1):262–72.
Fuller, W. A. 1987. *Measurement error models*. New York: John Wiley.
Gong, G., and F. J. Samaniego. 1981. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics* 9 (4):861–9. doi:10.1214/aos/1176345526.
Gourieroux, C., and A. Monfort. 1995a. *Statistics and econometric models. (Vol. 1)*. Cambridge: University Press.
Gourieroux, C., and A. Monfort. 1995b. *Statistics and econometric models. (Vol. 2)*. Cambridge: University Press.
Guolo, A. 2011. Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica* 21:1639–63.
Gustafson, P. 2004. *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Boca Raton, FL: Chapman & Hall.
Kerber, R. A., J. E. Till, S. L. Simon, J. L. Lyon, D. C. Thomas, S. Preston-Martin, M. L. Rallison, R. D. Lloyd, and W. Stevens. 1993. A cohort study of thyroid disease in relation to fallout from nuclear weapons testing. *JAMA: The Journal of the American Medical Association* 270 (17):2076–82. doi:10.1001/jama.1993.03510170066032.
Kieschnick, R., and B. D. McCullough. 2003. Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling: An International Journal* 3 (3): 193–213. doi:10.1191/1471082X03st053oa.
Louis, T. A. 1982. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B* 44:226–133. doi:10.1111/j.2517-6161.1982.tb01203.x.
Midthune, D., R. J. Carroll, L. S. Freedman, and V. Kipnis. 2016. Measurement error models with interactions. *Biostatistics* 17 (2):277–90. doi:10.1093/biostatistics/kxv043.
Parke, W. R. 1986. Pseudo maximum likelihood estimation: the asymptotic distribution. *The Annals of Statistics* 14:335–57.

Qiu, Z., P. X.-K. Song, and M. Tan. 2008. Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics* 35 (4):577–96. doi:10.1111/j.1467-9469.2008.00603.x.

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rudemo, M., D. Ruppert, and J. C. Streibig. 1989. Random-effect models in nonlinear regression with applications to bioassay. *Biometrics* 45 (2):349–62. doi:10.2307/2531482.

Sen, S., R. Lamichhane, and N. Diawara. 2014. A bivariate distribution with conditional gamma and its multivariate form. *Journal of Modern Applied Statistical Methods* 13 (2):169–84. doi:10.22237/jmasm/1414814880.

Silva, E., C. Diniz, J. M. F. Carrasco, and M. de Castro. 2018. A beta regression model with multiplicative log-normal measurement errors. *Communications in Statistics-Simulation and Computation* 47:229–48. doi:10.1080/03610918.2017.1280165.

Skrondal, A., and J. Kuha. 2012. Improved regression calibration. *Psychometrika* 77 (4):649–69. doi:10.1007/s11336-012-9285-1.

Song, P., Z. Qiu, and M. Tan. 2004. Modelling heterogeneous dispersion in marginal model for longitudinal proportional data. *Biometrical Journal* 46 (5):540–53. doi:10.1002/bimj.200110052.

Wei, G. C. G., and M. A. Tanner. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 85 (411):699–704. doi:10.1080/01621459.1990.10474930.

Zhang, P., Z. Qiu, and C. Shi. 2016. simplexreg: An R package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Software* 71:1–21.

Zhang, W., and H. Wei. 2008. Maximum likelihood estimation for simplex distribution nonlinear mixed models via the stochastic approximation algorithm. *Rocky Mountain Journal of Mathematics* 38:1863–75. doi:10.1216/RMJ-2008-38-5-1863.