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
Simplex regression models with measurement error

Jalmar M. F. Carrasco\textsuperscript{a} and Nancy Reid\textsuperscript{b}

\textsuperscript{a}Department de Statistics, Federal University of Bahia, Salvador, Brazil; \textsuperscript{b}Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

\textbf{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.

\textbf{ARTICLE HISTORY}
Received 6 April 2018
Accepted 28 May 2019

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

\section{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
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

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

where $E(Y) = \mu \in (0, 1)$ 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 $(\mu_i, \sigma^2_i)$. The simplex regression model is defined by (1) with $g(\mu_i) = z_i^T \alpha$ and $h(\sigma^2_i) = v_i^T \delta$, where $z_i$ and $v_i$, of dimension $1 \times p_1$ and $1 \times p_2$, respectively, are covariates. We assume $\alpha \in \mathbb{R}^{p_1}$ and $\delta \in \mathbb{R}^{p_2}$, $p_1 + p_2 < n$, 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) 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

\[
\hfill
\text{Figure 1. The simplex density function.}
\]
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 $x$ can be observed only via a proxy $w$. Results for the normal and gamma model for $x$ 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

\[ g(\mu_i) = z_i^T x + x_i^T \beta, \quad h(\sigma_i^2) = v_i^T \delta + m_i^T \gamma, \]

where $x_i \in \mathbb{R}^{q_1}$ and $m_i \in \mathbb{R}^{q_2}$ are unobserved latent covariates and $\beta \in \mathbb{R}^{q_1}$ and $\gamma \in \mathbb{R}^{q_2}$ are the unknown regression coefficients.

The structural model assumes a probability distribution for the unobserved covariates. Let $w_i = (w_{i1}, ..., w_{qi})^T$ and $x_i = (x_{i1}, ..., x_{qi})^T$ be the observed and unobserved variables, respectively. The additive measurement error model is

\[ w_i = \tau_0 + \tau_1 \times x_i + e_i, \quad i = 1, ..., n, \]

where $e_i$ is the vector of random errors, $\times$ is the Hadamard product, and $\tau_0$ and $\tau_1$ are unknown parameters, called additive and multiplicative bias respectively, in Carrasco et al. (2014). If $\tau_0 = 0$ and $\tau_1 = 1$, (3) is the classical structural model, $w_i = x_i + e_i$.

The multiplicative error structural model is

\[ w_i = x_i \times e_i, \quad i = 1, ..., n, \]

which is a classical additive error model after a logarithm transformation. Eckert et al. (1997) consider a general transformed additive error model, i.e. $p(w_i) = p(x_i) + e_i$, where $p(\cdot)$ is a monotone transformation function.

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

3. Statistical inference

Suppose a random sample, $(y_1, w_1), ..., (y_n, w_n)$, of size $n$ is observed. We omit the vector of variables $z_i$ from the notation as they are known and fixed. The joint density of $(y_i, w_i)$, observed for the $i$-th individual, is obtained by integrating the joint density of the complete data $(y_i, w_i, x_i)$, $f(y_i, w_i, x_i; \theta) = f(y_i | w_i, x_i; \theta_1)f(w_i | x_i; \theta_2)f(x_i; \theta_3)$, with
respect to $x_i$, where the complete parameter vector $\theta = (\theta_1, \theta_2, \theta_3)^\top$. Following Clayton (1992) the function $f(y_i, w_i, x_i; \theta)$ can be view as one function with three parts: (i) an outcome function $f(y_i|w_i, x_i; \theta_1)$, (ii) a measurement function $f(w_i|x_i; \theta_2)$ and (iii) an exposure function $f(x_i; \theta_3)$. The logarithm of the likelihood function for the sample of $n$ observations is

$$\ell(\theta) = \sum_{i=1}^{n} \log \int f(y_i|x_i; \theta_1)f(w_i|x_i; \theta_2)f(x_i; \theta_3)dx_i,$$

where $f(y_i|x_i; \theta_1)$ will be the simplex distribution defined in (1), $f(w_i|x_i; \theta_2)$ is the conditional distribution of $w_i$ given $x_i$ and $f(x_i; \theta_3)$ is the marginal density of $x_i$.

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 $\theta_1$ be the vector of parameters of interest and $\theta_{23} = \theta_2, \theta_3$ the vector of nuisance parameters, with $\theta = (\theta_1, \theta_{23})$. 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 $\theta_{23}$ by maximizing

$$\ell_r(\theta_{23}) = \sum_{i=1}^{n} \log \int f(w_i|x_i; \theta_2)f(x_i; \theta_{23})dx_i = \sum_{i=1}^{n} \log f(w_i; \theta_{23}),$$

the reduced log-likelihood function. The estimator $\hat{\theta}_{23}$ that maximizes $\ell_r(\theta_{23})$ 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 $\hat{\theta}_{23}$ is asymptotically distributed as a multivariate normal distribution with $\theta_{23}$ mean and covariance matrix $\Sigma_{(23,23)}^{-1} = E^{-1}[−\partial^2 \ell^2(\theta_{23})/\partial \theta_2 \partial \theta_3^\top]$. The second step consists in inserting the estimate $\hat{\theta}_{23} = (\hat{\theta}_2, \hat{\theta}_3)^\top$ obtained in (6) into the log-likelihood function (5)

$$\ell_p(\theta_1) = \sum_{i=1}^{n} \log \int f(y_i|x_i; \theta_1)f(w_i|x_i; \hat{\theta}_2)f(x_i; \hat{\theta}_3)dx_i.$$

The estimator of $\theta_1, \hat{\theta}_1$, that maximizes $\ell_p(\theta_1)$ is consistent and asymptotically normally distributed (Gong and Samaniego 1981; Gourieroux and Monfort 1995b; Parke 1986). Let $U(\theta) = \partial \ell(\theta)/\partial \theta$ be the score function, partitioned as $U(\theta) = (U_{\theta_1}(\theta)^\top, U_{\theta_{23}}(\theta)^\top)^\top$, and define the mean score $\bar{U}(\theta) = n^{-1}U(\theta)$. Let the true parameter value $\theta^* = (\theta_1^*, \theta_{23}^*)^\top$. The Fisher information matrix is
\[
I(\theta^*) = \lim_{n \to \infty} E_{\theta^*} \left[ -\frac{\partial U(\theta)}{\partial \theta^*} \right] = \begin{bmatrix} I_{(1,1)}(\theta^*) & I_{(1,23)}(\theta^*) \\ I_{(23,1)}(\theta^*) & I_{(23,23)}(\theta^*) \end{bmatrix},
\]

with partitions corresponding to \(\theta_1\) and \(\theta_{23}\). It is further assumed that

\[
\sqrt{n} \begin{bmatrix} \hat{U}_{\theta_1}(\theta_1^*, \theta_{23}^*) \\ (\hat{\theta}_{23} - \theta_{23}^*) \end{bmatrix} \to N \left( \theta_0, \begin{bmatrix} I_{(1,1)} & I_{(1,23)} \\ I_{(23,1)} & I_{(23,23)} \end{bmatrix} \right).
\]

It follows that, the distribution of \(\sqrt{n}(\hat{\theta}_1 - \theta_1^*)\) is normal with mean zero and variance matrix \(\Sigma = I_{(1,1)}^{-1} + I_{(1,1)}^{-1} I_{(23,1)}^{-1} \Sigma_{(23,23)}^{-1} I_{(23,1)}^{-1} I_{(1,1)}^{-1}\) where \(I_{(1,1)}^{-1}\) is the asymptotic covariance matrix of \(\theta_1\) when \(\theta_{23}\) is known. The matrix \(I_{(23,1)}\) can be estimated by

\[
I_{(23,1)} = n^{-1} \sum_{i=1}^{n} \frac{\hat{U}_{\theta_{23,i}}(\hat{\theta}) \hat{U}_{\theta_{1,i}}(\hat{\theta})}{},
\]

where \(\hat{U}_{\theta_{23,i}}(\hat{\theta})\) and \(\hat{U}_{\theta_{1,i}}(\hat{\theta})\) are the gradients of the reduced and pseudo log-likelihood function for subject \(i\), evaluated at the parameter estimates, respectively. An estimate of \(\Sigma_{(23,23)}\) can be obtained from the hessian of \(\ell_i(\theta_{23})\). For the estimation of \(\theta_1\) 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

\[
\hat{Q}_p(\theta_1|\theta_1^{(r)}) = M^{-1} \sum_{m=1}^{M} \sum_{j=1}^{n} \kappa_{mi}^{(r)} \log f(y_i|w_i, x_{mi}^{(r)}; \theta_1),
\]

where \(x_{mi}^{(r)}, \ldots, x_{Mi}^{(r)}\) are \(M\) random samples from \(f(x_i|y_i, w_i; \theta_1^{(r)}, \theta_{23})\) and \(\theta_1^{(r)}\) denotes the value of \(\theta\) from the \(r\)th iteration. The specification of \(f(x_i|y_i, w_i; \theta_1^{(r)}, \theta_{23})\) is usually difficult or even impractical in measurement error problems. Guolo (2011) proposed importance sampling, where random samples \(x_{mi}^{*}, m = 1, \ldots, M\) are generated from the importance density \(f(x_i^*; \cdot)\) or \(f(x_i|w_i; \cdot)\), assumed known. Then the weight for the \(i\)th observation is

\[
\kappa_{mi}^{(r)} = \frac{f(x_{mi}^{*}|y_i, w_i; \theta_1^{(r)})}{f(x_{mi}^{*}|w_i; \theta_1^{(r)})},
\]

\[
= \frac{f(y_i|w_i, x_{mi}^{*}|w_i; \theta_1^{(r)}) f(w_i|x_{mi}^{*}; \theta_1^{(r)})}{\int f(y_i|w_i, x_{mi}^{*}|w_i; \theta_1^{(r)}) f(w_i|x_{mi}^{*}; \theta_1^{(r)}) dx_{mi}^{*}},
\]

\[
\approx \frac{f(y_i|w_i, x_{mi}^{*}|w_i; \theta_1^{(r)})}{M^{-1} \sum_{m=1}^{M} f(y_i|w_i, x_{mi}^{*}|w_i; \theta_1^{(r)})}.
\]

To simplify the description of the M-step, in the simplex regression model with measurement error, we assume that \(m_i = x_i\) in (2). We need to maximize

\[
\hat{Q}_p(\theta_1|\theta_1^{(r)}) = M^{-1} \sum_{m=1}^{M} \sum_{i=1}^{n} \kappa_{mi}^{(r)} l_i(\mu_i^{(r)}, \sigma_i^{2(r)}),
\]

(8)
where
\[ \ell_i^r = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_i^{2(r)}) - \frac{3}{2} \log(y_i(1-y_i)) - \frac{1}{2\sigma_i^{2(r)}} d(y_i, \mu_i^r), \]
with \( d(y_i, \mu_i^r) \) as defined in (1), \( \mu_i^r = g^{-1}(z_i^\top \alpha + x_{mi}^{1(r)} \beta) \) and \( \sigma_i^{2(r)} = h^{-1}(y_i^\top \delta + x_{mi}^{2(r)} \gamma) \). We can obtain the score vector for the parameters of interest from \( \hat{Q}_p(\theta_1|\theta_1^r) \): for \( j = 1, \ldots, p_1 \) and \( j' = 1, \ldots, q_1 \),
\[ U_{\alpha_j}(\theta_1) = \frac{\partial \hat{Q}_p(\theta_1|\theta_1^r)}{\partial \alpha_j} = \frac{M^{-1}}{\sigma^2} \sum_{i=1}^n \sum_{m=1}^M \kappa_{mi}^{(r)} \left( \sigma_i^{2(r)} \right) \frac{1}{y_i(1-y_i) \mu_i^{2(r)}(1-\mu_i^{2(r)})} \right) \left( 1 + \frac{y_i-\mu_i^{(r)}}{\mu_i^{(r)}(1-\mu_i^{(r)})} \right) \right). \]
For \( t = 1, \ldots, p_2 \) and \( t' = 1, \ldots, q_2 \),
\[ U_{\beta_t}(\theta_1) = \frac{\partial \hat{Q}_p(\theta_1|\theta_1^r)}{\partial \beta_t} = \frac{M^{-1}}{\sigma^2} \sum_{i=1}^n \sum_{m=1}^M \kappa_{mi}^{(r)} \left( \sigma_i^{2(r)} \right) \frac{1}{y_i(1-y_i) \mu_i^{2(r)}(1-\mu_i^{2(r)})} \right) \left( 1 + \frac{y_i-\mu_i^{(r)}}{\mu_i^{(r)}(1-\mu_i^{(r)})} \right) \right) \}
Solving simultaneously the equations \( U_{\alpha}(\theta_1) = 0, U_{\beta}(\theta_1) = 0 \) and \( U_{\sigma}(\theta_1) = 0 \) we obtain the pseudo-maximum likelihood estimator for \( \alpha, \beta, \) and \( \sigma \). If \( \sigma_1^2 = \ldots = \sigma_n^2 = \sigma^2 \), then
\[ U_{\sigma}(\theta_1) = \frac{\partial \hat{Q}_p(\theta_1|\theta_1^r)}{\partial \sigma^2} = \frac{1}{M} \sum_{i=1}^n \sum_{m=1}^M \kappa_{mi}^{(r)} \left( \sigma_i^{2(r)} \right) \frac{1}{y_i(1-y_i) \mu_i^{2(r)}(1-\mu_i^{2(r)})} \right) \left( 1 + \frac{y_i-\mu_i^{(r)}}{\mu_i^{(r)}(1-\mu_i^{(r)})} \right) \right). \]
An estimator of the dispersion parameter is
\[ \hat{\sigma}^2 = \frac{\sum_{m=1}^M \sum_{i=1}^n \kappa_{mi}^{(r)} d(y_i, \hat{\mu}_i^{(r)})}{M \sum_{m=1}^M \kappa_{mi}^{(r)}}. \]
The variance - covariance matrix \( \Sigma \) defined in Sec. 3 is calculated using the approach of Louis (1982). The matrix \( I_{1(1)} \) can be estimated using the expressions
\[ I_{1(1)} = -\frac{\partial^2}{\partial \theta_1 \partial \theta_1^\top} Q(\theta_1|\theta_1^r), \]
\[ = -\frac{1}{M} \sum_{i=1}^n \sum_{m=1}^M \kappa_{mi}^{(r)} \frac{\partial^2}{\partial \theta_1 \partial \theta_1^\top} \ell_p(\theta_1; y_i, w_i, x_{mi}^o, \hat{\theta}_23), \]
\[ I_{(1,1)}^2 = \sum_{i=1}^{n} \left\{ \frac{1}{M} \sum_{m=1}^{M} \kappa_{mi} \frac{\partial}{\partial \theta_1} \ell_p \left( \theta_1; y_i, w_i, x_{mi}^0, \hat{\theta}_{23} \right) \right\}_{\theta_1 = \hat{\theta}_1} \times \left\{ \frac{1}{M} \sum_{m=1}^{M} \kappa_{mi} \frac{\partial}{\partial \theta_1} \ell_p \left( \theta_1; y_i, w_i, x_{mi}^0, \hat{\theta}_{23} \right) \right\}_{\theta_1 = \hat{\theta}_1}, \]

where \( x_{mi}^0 \) and \( \kappa_{mi} \) are the random importance sample and importance weight of the Monte Carlo EM algorithm for the \( i \)th 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

\[
\begin{align*}
  y_i | z_i, w_i, x_i & \sim S^{-} \left( \mu_i, \sigma^2_i \right), \\
  w_i | x_i & \sim N(x_i, \sigma^2_e), \\
  x_i & \sim N(\mu_x, \sigma^2_x),
\end{align*}
\]

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

\[
\ell_r(\theta_{23}) = \sum_{i=1}^{n} \log f(w_i | x_i; \sigma^2_e)f(x_i; \mu_x, \sigma^2_x),
\]

\[
= - \frac{n}{2} \log \left[ 2\pi \left( \sigma^2_x + \sigma^2_e \right) \right] - \frac{1}{2 \left( \sigma^2_x + \sigma^2_e \right)} \sum_{i=1}^{n} (w_i - \mu_x)^2,
\]

where \( \theta_{23} = (\theta_2, \theta_3)^\top \), with \( \theta_2 = \sigma^2_e \) and \( \theta_3 = (\mu_x, \sigma^2_x)^\top \), \( \sigma^2_e \) 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 \( \hat{\theta}_3 \) of the nuisance parameters solves \( \partial \ell_r(\theta_{23}) / \partial \mu_x = 0 \) and \( \partial \ell_r(\theta_{23}) / \partial \sigma^2_x = 0 \). Thereby, \( \hat{\mu}_x = \bar{w} \) and \( \hat{\sigma}^2_x = n^{-1} \sum_{i=1}^{n} (w_i - \bar{w})^2 - \hat{\sigma}^2_e \) with \( \bar{w} = n^{-1} \sum_{i=1}^{n} w_i \). We substitute \( \theta_3 \) into the log-likelihood function (5), giving the pseudo-log-likelihood function (7)

\(^1\)https://cran.r-project.org/web/packages/numDeriv/index.html
Proposition 1. Let \( x \) and \( e \) be independent random variables such that \( x \sim \text{Ga}(\mu_x, \phi_x) \) and \( e \sim \text{Ga}(1, \phi_e) \), with \( \mu_x > 0, \phi_x > 0 \) and \( \phi_e > 0 \). Let \( w = xe \). Then

(i) \( \mathbb{E}(w) = \mu_x \),  
(ii) \( \text{Var}(w) = (1 - \phi_x + \phi_e) \mu_x^2 / \phi_x \phi_e \),  
(iii) \( \text{Cov}(w) = \mu_x^2 / \phi_x \).
From Proposition 1, to avoid nonidentifiability problems, we can easily calculate an estimate of $\hat{\phi}_e$, say $\hat{\phi}_e$, as
\[
\hat{\phi}_e = \left[ \left( \frac{s^2_{\phi} \hat{\phi}_x / \tilde{w}^2}{-1} \right) + \left( 1 + \hat{\phi}_x \right) \right],
\]
where $\tilde{w} = \sum_{i=1}^{n} w_i / n$ and $s^2_{\phi} = \sum_{i=1}^{n} (w_i - \tilde{w})^2 / (n-1)$.

As in Sec. 4.1, assuming $\phi_e$ known or estimating by (13), the reduced log-likelihood function
\[
\ell_r(\theta_{23}) = \sum_{i=1}^{n} \left\{ \log (2) + \frac{1}{2} \log (\phi_x) + \frac{1}{2} \log (\phi_e) - \frac{1}{2} \log (\mu_x) - \log \Gamma(\phi_x) - \log \Gamma(\phi_e) + \left( \frac{1}{2} (\phi_x + \phi_e) - 1 \right) \log (w_i) + \log \left( K_{\phi_x-\phi_e} \left( 2 \sqrt{\frac{\phi_x \phi_e}{\mu_x} w_i} \right) \right) \right\}.
\]

We then substitute $\hat{\theta}_2$ and $\hat{\phi}_e$ into the log-likelihood function to obtain the pseudo-likelihood, and use the Monte Carlo EM algorithm to estimate $\theta_1$. The importance density in this case can be $f(x_i; \theta_2)$ or $f(x_i|w_i; \theta_3)$. The conditional density $f(x_i|w_i; \theta_{23})$ can be derived from (11) and (12):
\[
f(x_i|w_i; \theta_{23}) = \frac{1}{2} \left( \frac{\phi_x}{\phi_e \mu_x} \right)^{\frac{1}{2}(\phi_x - \phi_e)} x_i^{\phi_x - \phi_e - 1} w_i^{-\frac{1}{2}(\phi_x - \phi_e)} \times \exp \left( -\phi_e \frac{w_i}{x_i} - \phi_x \frac{x_i}{\mu_x} K_{\phi_x - \phi_e}^{-1} \left( 2 \sqrt{\frac{\phi_x \phi_e}{\mu_x} w_i} \right) \right).
\]

An alternative approach to inference uses (7) directly, approximating the integral numerically, for instance, by Laguerre-Gauss quadrature. The Laguerre-Gauss quadrature approximation is
\[
\int_0^{\infty} e^{-x} f(x) dx \approx \sum_{q=1}^{Q} \nu_q f(\tilde{x}_q),
\]
where $\nu_q$ and $\tilde{x}_q$ 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
\[
\ell_p(\theta_1) = \sum_{i=1}^{n} \log \int_0^{\infty} f(y_i, w_i, x_i; \theta_{1}, \theta_{23}) dx_i
\]
\[
= \sum_{i=1}^{n} \log \int_0^{\infty} f(y_i|x_i; \theta_1) f(x_i|w_i; \theta_{23}) f(w_i; \theta_{23}) dx_i
\]
\[
\approx \sum_{i=1}^{n} \log f(w_i; \theta_{23}) + \sum_{i=1}^{n} \log \left\{ \sum_{q=1}^{Q} \omega_q f(y_i|\tilde{x}_q; \theta_1) f(\tilde{x}_q|w_i; \theta_{23}) \right\},
\]
where $\hat{\theta}_{23}$ is find maximized (14). The approximate pseudo-likelihood estimator of $\theta_1, \hat{\theta}_1$, 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 $g(\mu_i) = \alpha_0 + \alpha_1z_i + \beta x_i, h(\sigma_i^2) = \delta x_i$, 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 $\mu_x$ and variance $\sigma_x^2$ 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 $x_i \sim \text{Ga}(\mu_x, \phi_x)$ 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 $\beta$ 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 $\delta$ and $\beta$, the latter of which is associated with the variable measured with error. However, the root mean square error (RMSE) for parameter $\beta$ 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
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 represents the customer’s gender is also considered. Our goal is to model the proportion of spending \( (y) \) using the real income of a new customer \( (x) \) 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 classified by gender. We consider the model

\[
y_i | z_i, w_i, x_i \sim S^- (\mu_i, \sigma_i^2),
\]

\[
\log (\mu_i/(1-\mu_i)) = \alpha_0 + \alpha_1 z_i + \beta x_i,
\]

\[
\log (\sigma_i^2) = \delta + \gamma x_i,
\]

\[
w_i | x_i, \sigma_i^2 \sim N (x_i, \sigma_i^2),
\]

\[
x_i | \mu_x, \sigma_x^2 \sim N (\mu_x, \sigma_x^2), \text{ for } i = 1, \ldots, 200.
\]

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

<table>
<thead>
<tr>
<th>( \sigma_i^2 )</th>
<th>Method</th>
<th>( n )</th>
<th>Measure</th>
<th>( \alpha_0 )</th>
<th>( \alpha_1 )</th>
<th>( \beta )</th>
<th>( \log (\delta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0333</td>
<td>( \ell_p )</td>
<td>25</td>
<td>Bias</td>
<td>0.026</td>
<td>0.010</td>
<td>-0.065</td>
<td>-0.047</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.191</td>
<td>0.188</td>
<td>0.113</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>Bias</td>
<td>0.024</td>
<td>0.008</td>
<td>-0.064</td>
<td>-0.039</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.157</td>
<td>0.129</td>
<td>0.099</td>
<td>0.272</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>Bias</td>
<td>0.018</td>
<td>0.004</td>
<td>-0.063</td>
<td>-0.020</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.127</td>
<td>0.103</td>
<td>0.090</td>
<td>0.190</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>Bias</td>
<td>0.011</td>
<td>0.002</td>
<td>-0.060</td>
<td>-0.020</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.104</td>
<td>0.089</td>
<td>0.091</td>
<td>0.176</td>
<td></td>
</tr>
<tr>
<td>0.0052</td>
<td>( \ell_p )</td>
<td>25</td>
<td>Bias</td>
<td>0.025</td>
<td>-0.002</td>
<td>-0.029</td>
<td>-0.030</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.207</td>
<td>0.131</td>
<td>0.086</td>
<td>0.141</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>Bias</td>
<td>0.014</td>
<td>0.002</td>
<td>-0.024</td>
<td>-0.023</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.126</td>
<td>0.098</td>
<td>0.060</td>
<td>0.140</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>Bias</td>
<td>0.010</td>
<td>0.001</td>
<td>-0.020</td>
<td>-0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.108</td>
<td>0.080</td>
<td>0.049</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>Bias</td>
<td>0.010</td>
<td>0.000</td>
<td>-0.020</td>
<td>-0.011</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.092</td>
<td>0.064</td>
<td>0.046</td>
<td>0.122</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \ell_{naive} )</td>
<td>25</td>
<td>Bias</td>
<td>-0.075</td>
<td>0.001</td>
<td>0.017</td>
<td>-0.359</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.434</td>
<td>0.167</td>
<td>0.102</td>
<td>1.469</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>Bias</td>
<td>-0.063</td>
<td>0.003</td>
<td>0.023</td>
<td>-0.172</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.222</td>
<td>0.113</td>
<td>0.071</td>
<td>1.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>Bias</td>
<td>-0.053</td>
<td>0.006</td>
<td>0.025</td>
<td>0.119</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.182</td>
<td>0.108</td>
<td>0.060</td>
<td>0.880</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>Bias</td>
<td>-0.049</td>
<td>-0.004</td>
<td>0.023</td>
<td>0.099</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.152</td>
<td>0.078</td>
<td>0.050</td>
<td>0.759</td>
<td></td>
</tr>
</tbody>
</table>

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.
of parameters \((x_0, x_1, \beta, \delta)\)' is shown 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

\[
\begin{align*}
\frac{w_i}{x_i} & \sim \text{Ga}(\lambda_i, \phi_x) \quad \text{and} \\
\frac{x_i}{\mu_x} & \sim \text{Ga}(\lambda_x, \phi_x) .
\end{align*}
\]

Under this model the maximum pseudo-likelihood of \(x_0, x_1, \beta, \delta\) and \(\gamma\) are shown 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, \(\exp (c \beta)\), 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 naive 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
spending ($y$), decreasing on average by 40.15%, 40.83% and 43.18%. In summary, when we assume a normal distribution for an unobserved asymmetric variable or ignore measurement error, the interpretation of the odds ratios may change.

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

<table>
<thead>
<tr>
<th>$\phi_x$</th>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>log ($\phi$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>$\ell_p$</td>
<td>25</td>
<td>Bias</td>
<td>-0.086</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.566</td>
<td>0.554</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>Bias</td>
<td>-0.056</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.470</td>
<td>0.420</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>Bias</td>
<td>-0.031</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.388</td>
<td>0.306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>Bias</td>
<td>-0.022</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.222</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>$\ell_{naive}$</td>
<td>25</td>
<td>Bias</td>
<td>-0.039</td>
<td>-0.020</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.833</td>
<td>1.544</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>Bias</td>
<td>0.070</td>
<td>-0.020</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.517</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>Bias</td>
<td>0.118</td>
<td>-0.029</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.366</td>
<td>0.704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>Bias</td>
<td>0.120</td>
<td>-0.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.371</td>
<td>0.620</td>
</tr>
<tr>
<td>1.0</td>
<td>$\ell_p$</td>
<td>25</td>
<td>Bias</td>
<td>-0.079</td>
<td>-0.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.457</td>
<td>0.437</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>Bias</td>
<td>-0.067</td>
<td>-0.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.219</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>Bias</td>
<td>-0.055</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.275</td>
<td>0.270</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>Bias</td>
<td>-0.043</td>
<td>-0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.256</td>
<td>0.247</td>
</tr>
<tr>
<td></td>
<td>$\ell_{naive}$</td>
<td>25</td>
<td>Bias</td>
<td>0.011</td>
<td>-0.112</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.811</td>
<td>1.357</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>Bias</td>
<td>0.049</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.630</td>
<td>1.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>Bias</td>
<td>0.096</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.440</td>
<td>0.696</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
<td>Bias</td>
<td>0.118</td>
<td>-0.036</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>0.388</td>
<td>0.649</td>
</tr>
</tbody>
</table>

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).
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 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 Científico e Tecnológico (CNPq), Brazil, Programa de Apoio a Pesquisadores Emergentes - UFBA, Brazil and the Department of Statistical Sciences of the University of Toronto, Canada.

References


