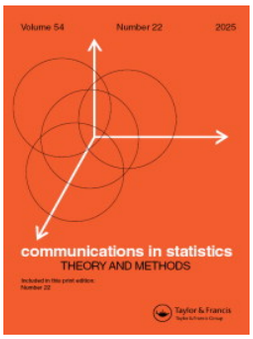

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: www.tandfonline.com/journals/lsta20

Bivariate simplex regression model for rates and proportions data

Lucas S. Vieira, Emerson Amaral, Lizandra C. Fabio, Vanessa Barros & Jalmar M. F. Carrasco

To cite this article: Lucas S. Vieira, Emerson Amaral, Lizandra C. Fabio, Vanessa Barros & Jalmar M. F. Carrasco (09 Sep 2025): Bivariate simplex regression model for rates and proportions data, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2025.2553731

To link to this article: https://doi.org/10.1080/03610926.2025.2553731

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2025.2553731

Bivariate simplex regression model for rates and proportions data

Lucas S. Vieira, Emerson Amaral, Lizandra C. Fabio, Vanessa Barros, and Jalmar M. F. Carrasco

Institute of Mathematics and Statistics, Federal University of Bahia, Salvador–Bahia, Brazil

ABSTRACT

Proportional continuous outcomes, bounded in the unit interval (0,1), are commonly modeled using beta or simplex regression models. While these approaches are effective for univariate data, they may be inadequate for correlated proportional responses. To address this limitation, we propose a bivariate simplex regression model (BSRM) based on the Farlie-Gumbel-Morgenstern (FGM) copula. This model jointly captures the dependence structure between two response variables and simultaneously models both the mean and dispersion parameters. Parameter estimation is conducted via the maximum likelihood method, and the asymptotic properties of the estimators are examined through Monte Carlo simulation studies. Diagnostic tools, including residual analysis and global influence measures such as generalized Cook's distance and the likelihood displacement, are developed to evaluate model adequacy and identify influential observations. The effectiveness of the proposed methodology is illustrated through an application to socioeconomic data from municipalities in Alagoas, Brazil.

ARTICLE HISTORY
14 November 2024
23 August 2025

KEYWORDS
Bivariate simplex distribution; FGM copula; randomized quantile residuals; global influence.

1. Introduction

Proportional continuous outcomes are often derived from experimental or observational scientific studies based on practical situations associated with several knowledge areas (Liu et al. 2020). Linear models are generally unsuitable for estimating the relationship between the explanatory and response variables when the latter are measured within the standard unit interval (0,1). This is because such variables exhibit asymmetry and heteroscedasticity. Ferrari and Cribari-Neto (2004) introduced a class of regression models where the response variable follows a beta distribution with mean μ and dispersion parameters ϕ. Several researchers have further explored inferential and diagnostic procedures for the beta regression model and its extensions. Notable examples include Paolino (2001), Vasconcellos and Cribari-Neto (2005), Simas, Barreto-Souza, and Rocha (2010), Espinheira, Ferrari, and Cribari-Neto (2008), and Rocha and Simas (2011). An alternative for the beta distribution is the simplex distribution. The simplex distribution was first proposed by Barndorff-Nielsen and Jørgensen (1991) and later incorporated into a class of dispersion models by Jorgensen (1997), which extended the framework of generalized linear models (GLMs) (Nelder and Wedderburn 1972). The

CONTACT Jalmar M.F. Carrasco carrascojalmar@gmail.com Institute of Mathematics and Statistics, Federal University of Bahia, Salvador–Bahia, Brazil.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

simplex distribution is particularly convenient and flexible for modeling data confided to the continuous unit interval (0,1), such as proportions, rates, or indices. Let y be a random variable that follows a simplex distribution, with parameters $\mu \in (0,1)$ and $\sigma^2 > 0$. The probability density function (pdf) of the simplex distribution is given by

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

where $0 < y < 1$ and $d(y;\mu) = (y-\mu)^2/y(1-y)\mu^2(1-\mu)^2$ is the regular unitary deviation. The expected value and variance of y are given by $\mathrm{E}(y) = \mu$ and

$$\mathrm{Var}(y) = \mu(1-\mu) - \sqrt{\frac{1}{2\sigma^2}} \exp\left\{\frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\}\Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\},$$

where $\Gamma(a,b) = \int_b^\infty x^{a-1}e^{-x}dx$ is the incomplete gamma function. For regular unitary deviation, the variance function is a function $V : \Omega \to (0,\infty)$, defined by $V(\mu) = 2/[\partial^2 d(y;\mu)/\partial\mu^2]_{y=\mu}$, consequently, the variance function of y is given by $V(\mu) = \mu^3(1-\mu)^3$, where, Ω is the parameter space of μ and σ^2 (Jorgensen 1997). The simplex density function exhibits a variety of shapes depending on the values of its mean (μ) and dispersion (σ) parameters, making it highly flexible for modeling proportion and rate data. When μ is close to 0.5 and σ is small, the density function resembles a symmetric bell-shaped curve, similar to a Gaussian distribution. For values of μ near 1 with moderate to large σ, the distribution becomes left-skewed, concentrating most of the probability mass near 1 with a long left tail. Conversely, when μ is close to 0, the density is right-skewed, with most of the mass near 0 and an extended right tail. As σ increases significantly, the density function approaches a uniform-like shape, spreading probability mass more evenly over the interval $(0,1)$. In contrast, when σ is very small, the function becomes highly concentrated around μ, forming a sharp peak, indicating low dispersion. These variations highlight the adaptability of the simplex distribution in capturing different data patterns, making it a valuable tool in regression modeling. To illustrate these characteristics, Figure 1 presents different density shapes corresponding to the following parameter configurations.

Let assume n independent observations y_i, where $i = 1, 2, \ldots, n$ from a simplex distribution with parameters μ_i and σ_i^2. The simplex regression model is defined by (1), where $g(\mu_i) = \boldsymbol{X}_i^\top\boldsymbol{\beta}$ and $h(\sigma_i^2) = \boldsymbol{Z}_i^\top\boldsymbol{\gamma}$. Here, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $1 \times k_1$ and $1 \times k_2$ covariate vectors, respectively, and $\boldsymbol{\beta} \in \mathbb{R}^{k_1}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{k_2}$, $k_1 + k_2 < n$. The functions $g(\cdot)$ and $h(\cdot)$ are known monotonic link functions. The simplex regression model has been utilized in studies by Qiu, Song, and Tan (2008), Song, Qiu, and Tan (2004), and in Zhang and Wei (2008). Additionally, the package `simplexreg` (Zhang, Qiu, and Shi 2016) is available in `R` (R Core Team 2024). Various authors have proposed extensions to the simplex regression model. For instance, Liu et al. (2020), introduced a simplex regression model inflated with zeros and ones. Espinheira and Silva (2020) proposed local influence analysis techniques for the general class of simplex regression models. Carrasco and Reid (2021) addressed the simplex regression model in the context of measurement error in the covariate, among other contributions.

This article uses data from 102 municipalities in the state of Alagoas, Brazil, including the Human Development Index (HDI), the Social Vulnerability Index (SVI), and the Dependency Ratio (DR), defined as the ratio of the economically dependent population to the economically active population. Data from the 2000 and 2010 censuses are considered independent for convenience. Figure 2 presents histograms for each variable, illustrating that the HDI variable typically exhibits the characteristic shape of the simplex distribution.

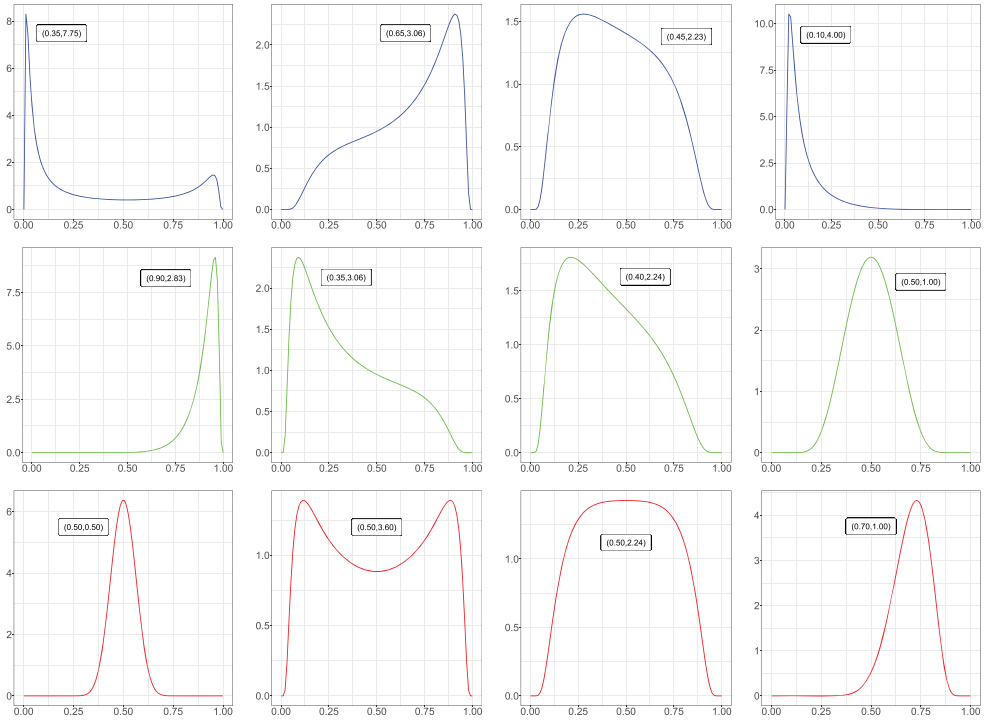

Figure 1. Density graph of the simplex distribution for different parameter values of (μ, σ^2) parameters.

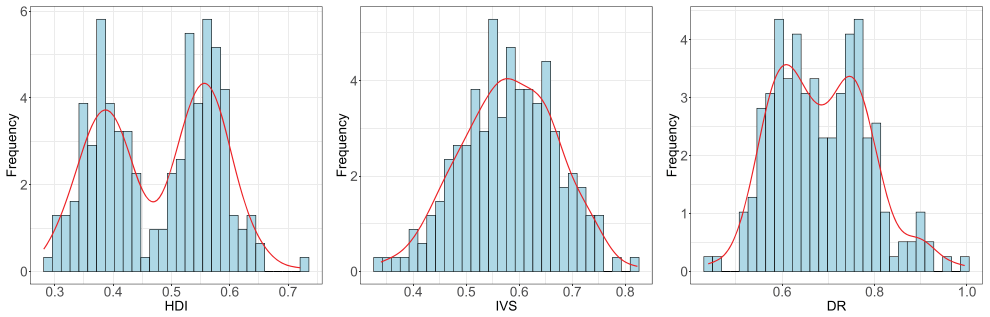

Figure 2. Histogram plots for HDI: The Human Development Index (left panel), IVS: The Social Vulnerability Index (center panel), and DR: The Dependency Ratio variables.

Table 1 presents the estimate, standard error, and p-value for both the simplex and beta distributions. According to the AIC and BIC criteria shown in Table 1, the univariate simplex distribution is identified as the best choice for analyzing the HDI variable.

In practice, we are often interested in examining the relation between variables such as HDI with DR, or SVI with DR. Both HDI and SVI are constrained to the unit interval $(0, 1)$, making simplex and beta distribution commonly used approaches for their analysis. Table 2 displays the estimate, standard error, and p-value for the simplex and beta regression models. Maximum likelihood estimates are obtained by assuming that the outcome variables follow either a simplex or beta distribution, using the `simplexreg` (Zhang, Qiu, and Shi 2016)

Table 1. Estimate, standard error, and *p*-value of the simplex and beta distributions.

Variable	Distribution	Parameter	Estimate	Standard error	*p*-Value
y_1	Simplex	μ	0.477	0.007	<0.001
		σ^2	0.809	0.040	–
	Beta	μ	0.477	0.006	<0.001
		σ^2	0.193	0.009	–
y_2	Simplex	μ	0.581	0.006	<0.001
		σ^2	0.789	0.039	–
	Beta	μ	0.581	0.006	<0.001
		σ^2	0.182	0.008	–

y_1: [Simplex: (AIC = −376.925, BIC = −359.653); Beta: (AIC = −372.475, BIC = −355.203)]; y_2: [Simplex: (AIC = −400.842, BIC = −383.569); Beta: (AIC = −400.873, BIC = −383.569)].

Table 2. Estimates, standard error, and *p*-value of the univariate simplex and beta regression model.

Model	Distribution	Parameters	Estimates	Standard error	*p*-Value
$y_1 \sim x$	Simplex	β_0	2.419	0.084	<0.001
		β_1	−3.675	0.131	<0.001
		γ_0	−2.966	0.395	<0.001
		γ_1	2.873	0.573	<0.001
	Beta	β_0	2.432	0.087	<0.001
		β_1	−3.697	0.136	<0.001
		γ_0	−4.339	0.450	<0.001
		γ_1	2.938	0.656	<0.001
$y_2 \sim x$	Simplex	β_0	−1.794	0.124	<0.001
		β_1	3.110	0.188	<0.001
		γ_0	−2.119	0.345	<0.001
		γ_1	2.110	0.499	<0.001
	Beta	β_0	−1.817	0.128	<0.001
		β_1	3.149	0.195	<0.001
		γ_0	−3.198	0.377	<0.001
		γ_1	1.727	0.545	<0.001

$y_1 \sim x$: [Simplex: (AIC = −694.554, BIC = −681.282, Pseudo-R^2 = 0.793); Beta: (AIC = −690.986, BIC = −677.714, Pseudo-R^2 = 0.794)]; $y_2 \sim x$: [Simplex: (AIC = −575.619, BIC = −562.347, Pseudo-R^2 = 0.584); Beta: (AIC = −578.938, BIC = −565.667, Pseudo-R^2 = 0.590)].

and `betareg` (Cribari-Neto and Zeileis 2010) packages available in `R` (R Core Team 2024) software.

Table 2 shows a similar performance based on the pseudo-R^2 values, with a slight preference for simplex regression model. However, AIC and BIC criteria indicate that the simplex regression model is the best alternative to the beta regression model. Additional details about the dataset are provided in Section 5.

In the literature, there are scientific studies focusing on outcomes within the standard unit interval (0,1), where bivariate or multivariate regression models serve as alternatives to analyzing correlated data. For example, Cepeda-Cuervo, Achcar, and Lopera (2014) proposed a bivariate beta regression model that jointly models the mean and dispersion parameters using the Farlie-Gumbel-Morgenstern (FGM) copula. Souza and Moura (2016) suggested the multivariate beta regression model incorporating various copula functions within Bayesian paradigm. Koochemeshkian, Manouchehri, and Bouguila (2020) introduced the bivariate beta regression model based on a flexible bivariate beta distribution with three shape parameters. Despite these advances, the bivariate simplex regression model remains relatively unexplored, particularly for modeling proportions, such as the proportion of budget allocated to different sectors. Bivariate distributions are frequently constructed using copula functions,

which enable the analysis of dependence structures between two random variables independently of their marginal distributions. This approach provides flexibility in combining various types of marginal distributions. Recently, Amaral et al. (2025) proposed a bivariate simplex distribution using the Farlie-Gumbel-Morgenstern (FGM) copula as an alternative method for analyzing bivariate data constrained to the standard unit interval. This development is a significant advancement in the simplex distribution framework. Consequently, this article aims to introduce a bivariate simplex regression model (BSRM) utilizing the FGM copula as an alternative to existing models for fitting outcomes within the standard unit interval (0,1). Additionally, the joint moments $\mathrm{E}[y_1 y_2]$ of the bivariate simplex distribution can be expressed analytically (Amaral et al. 2025).

This article is structured as follows: In Section 2, the bivariate simplex distribution is introduced. Section 3 defines the bivariate simplex regression model, and diagnostic measures are detailed in Section 3.1. A Monte Carlo simulation study is presented in Section 4. Section 5 includes an analysis of a real dataset. Conclusions are drawn in Section 6. Finally, the references used in this work are listed. R (R Core Team 2024) was used as the computational tool, and the dataset and code required to reproduce the results from Section 5 are available at https://github.com/carrascojalmar/BSRM.

2. Overview

One method for constructing a bivariate distribution is through the use of copula functions. This technique involves combining marginal distribution functions, allowing for the representation of various types of dependencies between variables. A copula is defined as a joint distribution function $C(u_1, u_2, \ldots, u_p) = P(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_p \leq u_p)$, where $0 \leq u_i \leq 1$, and $U_i \sim U(0,1)$, for $i = 1, 2, \ldots, p$. Given a p-dimensional cumulative distribution function $H(\cdot) = C(F_1(y_1), \ldots, F_p(y_p))$ with marginals $F_1, \ldots, F_p$. Nelsen (2006) states that there exists a p-dimensional copula, such as $C(\cdot)$, that is unique if and only $F_1(\cdot), \ldots, F_p(\cdot)$ are all continuous. Based on this method, the joint density function of $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ is given by

$$h(\boldsymbol{y}) = h(y_1, \ldots, y_p) = \frac{\partial^p H(y_1, \ldots, y_p)}{\partial y_1, \ldots, \partial y_p} = c(F_1(y_1), \ldots, F_p(y_p)) \prod_{i=1}^{p} f_i(y_i),$$

where $c(F_1(y_1), \ldots, F_p(y_p)) = \partial^p C(F_1(y_1), \ldots, F_p(y_p))/\partial F_1(y_1) \ldots \partial F_p(y_p)$ and $f_i(y_i) = \partial F_i(y_i)/\partial y_i$, $i = 1, 2, ..., p$. The FGM copula (Farlie-Gumbel-Morgenstern copula) is a relatively simple and classical copula used to model the dependence between two random variables. Although it is not as flexible as other copulas, it provides a basic framework for capturing weak positive or negative dependence. For $p = 2$, with a dependency structure defined by FGM copula, the joint distribution of F_1 and F_2 is given as $H(\boldsymbol{y}) = H(y_1, y_2) = C(F_1(y_1), F_2(y_2)) = F_1(y_1)F_2(y_2) + \lambda F_1(y_1)F_2(y_2)[1 - F_1(y_1)][1 - F_2(y_2)]$, and the joint density function of $\boldsymbol{y} = (y_1, y_2)^\top$ is expressed as $h(\boldsymbol{y}) = h(y_1, y_2) = \partial^2 H(y_1, y_2)/\partial y_1 \partial y_2 = f_1(y_1)f_2(y_2)\{1 + \lambda[1 - 2F_1(y_1)][1 - 2F_2(y_2)]\}$, where $f_1(y_1)$ and $f_2(y_2)$ are marginal density functions of y_1 and y_2, respectively and $\lambda \in [-1, 1]$ is the parameter which measures the association between random variables and is also related to their correlation structure. If $\lambda = 0$, the copula reduces to the independence copula, meaning $C(F_1(y_1), F_2(y_2)) = F_1(y_1) \times F_2(y_2)$, which indicates that the two random variables are independent; for $\lambda > 0$ (Positive dependence), meaning the two variables tend to increase together (though weakly);

for $\lambda < 0$ (negative dependence), meaning that when one variable increases, the other tends to decrease (weak negative dependence).

For the FGM copula, assuming $U_1 \sim U(0,1)$ and $U_2 \sim U(0,1)$, the cumulative distribution yield: $\mathrm{E}(U_1) = \mathrm{E}(U_2) = 1/2$, $\mathrm{Var}(U_1) = \mathrm{Var}(U_2) = 1/12$ and

$$E(U_1U_2) = \int_0^1 \int_0^1 u_1u_2[1 + \lambda(1-2u_1)(1-2u_2)]du_1du_2 = \frac{1}{4} + \lambda\frac{1}{36}.$$

Thus, the correlation between U_1 and U_2 is $\mathrm{Cor}(U_1, U_2) = (1/3)\lambda = [-1/3, 1/3]$, where $\lambda \in [-1, 1]$. While these are useful for the FGM family, they may not be adequate in practical situations with high correlation. In such cases, measures like Kendall's τ, or Spearman's ρ are more appropriate for assessing the strength of association.

Definition 1. (Kendall's τ correlation) Let y_1 and y_2 be continuous random variables with copula $C(\cdot)$. The Kendall's τ correlation measure is given by

$$\tau(y_1, y_2) = \tau(C) = Q(C, C') = 4\int_0^1 \int_0^1 C'(u_1, u_2)dC(u_1, u_2) - 1,$$

where C' denotes the partial derivative of the copula function C. This integral corresponds to the expected value of $C(U_1, U_2)$ when U_1 and U_2 are uniformly distributed random variables. Thus, Kendall's τ can be expressed as $\tau = 4E[C(U_1, U_2)]$.

Definition 2. (Spearman's ρ correlation) Let y_1 and y_2 be continuous random variables with copula $C(\cdot)$ and $\Pi(u_1, u_2) = u_1u_2$. The Spearman's ρ correlation measure for the random vector $\boldsymbol{y} = (y_1, y_2)$ is expressed as:

$$\rho = (y_1, y_2) = 12\int_0^1 \int_0^1 C(u_1, u_2)du_1du_2 - 3.$$

Amaral et al. (2025), assumed that $f_1(y_1; \mu_1, \sigma_1^2)$ and $f_2(y_2; \mu_2, \sigma_2^2)$ are the density functions following the univariate simplex distribution (Jorgensen 1997). The bivariate simplex distribution proposed by Amaral et al. (2025) is given by:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\theta}) = {} & \{2\pi\sigma_1^2[y_1(1-y_1)]^3\}^{-1/2} \exp\left\{-\frac{1}{2\sigma_1^2}d(y_1; \mu_1)\right\} \\ & \times \{2\pi\sigma_2^2[y_2(1-y_2)]^3\}^{-1/2} \exp\left\{-\frac{1}{2\sigma_2^2}d(y_2; \mu_2)\right\} \\ & \times \{1 + \lambda[1 - 2F_1(y_1)][1 - 2F_2(y_2)]\}, \end{aligned} \tag{2}$$

where $F_1(y_1)$ and $F_2(y_2)$ are the cumulative distribution functions of y_1 and y_2, respectively, and $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda)^\top$ is the vector of unknown parameters, with $\lambda \in [-1, 1]$. If $\boldsymbol{y}$ following the bivariate simplex distribution, we denoted it as $\boldsymbol{y} \sim S(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2)^\top$. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$, where $\boldsymbol{y}_i = (y_{1i}, y_{2i})^\top$, for $i = 1, \ldots, n$, follow the bivariate simplex distribution as given in (2). Given a random sample of the n size, the log-likelihood function for the bivariate simplex distribution is $\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}, \boldsymbol{y}_i)$, where

$$\begin{aligned} \ell_i(\boldsymbol{\theta}; \boldsymbol{y}_i) = {} & -\log(2\pi)/2 - \log(\sigma_1^2)/2 - 3\log[y_{1i}(1-y_{1i})]/2 - d(y_{1i}; \mu_1)/2\sigma_1^2 \\ & - \log(2\pi)/2 - \log(\sigma_2^2)/2 - 3\log[y_{2i}(1-y_{2i})]/2 - d(y_{2i}; \mu_2)/2\sigma_2^2 \\ & + \log\{1 + \lambda[1 - 2F_1(y_{1i})][1 - 2F_2(y_{2i})]\}. \end{aligned} \tag{3}$$

Partially differentiating the log-likelihood function, $\ell(\boldsymbol{\theta};\boldsymbol{y})$, with respect to the parameter vector, the elements of the score vector $\boldsymbol{U}(\boldsymbol{\theta}) = (U_{\mu_1}, U_{\mu_2}, U_{\sigma_1^2}, U_{\sigma_2^2}, U_\lambda)^\top$, and the observed information matrix, $J(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta};\boldsymbol{y}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ (Amaral et al. 2025). Under some standard regularity conditions, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is approximately normally distributed with zero mean and variance-covariance matrix $J^{-1}(\boldsymbol{\theta})$.

3. Bivariate simplex regression model (BSRM)

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$, where $\boldsymbol{y}_i = (y_{1i}, y_{2i})^\top$, for $i = 1, \ldots, n$, follow the bivariate simplex distribution given in (2). Then, the bivariate simplex regression model (BSRM) is defined considering the following hierarchical structure:

$$\begin{cases} \boldsymbol{y}_i & \sim & S_2(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2), \\ g(\boldsymbol{\mu}_i) & = & \boldsymbol{X}_i^\top\boldsymbol{\beta}, \\ h(\boldsymbol{\sigma}_i^2) & = & \boldsymbol{Z}_i^\top\boldsymbol{\gamma}, \end{cases}$$

where $\boldsymbol{\mu}_i = (\mu_{1i}, \mu_{2i})^\top$ and $\boldsymbol{\sigma}_i^2 = (\sigma_{1i}^2, \sigma_{2i}^2)^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, $\boldsymbol{\beta}_1 = (\beta_{10}, \ldots, \beta_{1k_1})^\top$, $\boldsymbol{\beta}_2 = (\beta_{20}, \ldots, \beta_{2k_2})^\top$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top)^\top$, $\boldsymbol{\gamma}_1 = (\gamma_{10}, \ldots, \gamma_{1k_1})^\top$, $\boldsymbol{\gamma}_2 = (\gamma_{20}, \ldots, \gamma_{2k_2})^\top$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are vectors containing explanatory variables, $g(\cdot)$ and $h(\cdot)$ are link functions.

For n independent observations of pair of dependent random variables (y_{1i}, y_{2i}), for $i = 1, 2, \ldots, n$, the maximum likelihood function of the bivariate simplex regression model is given by:

$$\begin{aligned} L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{i=1}^{n} & \{2\pi\sigma_{1i}^2[y_{1i}(1-y_{1i})]^3\}^{-1/2} \exp\left\{-\frac{1}{2\sigma_{1i}^2}d(y_{1i};\mu_{1i})\right\} \\ & \times \{2\pi\sigma_{2i}^2[y_{2i}(1-y_{2i})]^3\}^{-1/2} \exp\left\{-\frac{1}{2\sigma_{2i}^2}d(y_{2i};\mu_{2i})\right\} \\ & \times \{1 + \lambda[1-2F_1(y_{1i})][1-2F_2(y_{2i})]\}, \end{aligned} \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \lambda)^\top$, $F_1(\cdot)$ and $F_2(\cdot)$ represent the cumulative distribution functions of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively. The log-likelihood function for the bivariate simplex regression model is given by: $\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta};\boldsymbol{y}_i)$, where $\ell_i(\boldsymbol{\theta};\boldsymbol{y}_i)$ is defined in (3) in which $g(\mu_{ji}) = \sum_{l=0}^{k_1} X_{il}\beta_{jl}$, $h(\sigma_{ji}^2) = \sum_{l=0}^{k2} Z_{il}\gamma_{jl}$ with $i = 1, \ldots, n$ and $j = 1, 2$. The maximum likelihood (ML) estimates $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ are computed by using the quasi-Newton (`BFGS`) method. Alternatively, we can solve the nonlinear equation obtained by setting the components of the score vector equal to zero, that is, $\boldsymbol{U}_\theta = (U_{\boldsymbol{\beta}}^\top, U_{\boldsymbol{\gamma}}^\top, U_\lambda)^\top = \boldsymbol{0}$, in that

$$\begin{aligned} U_{\boldsymbol{\beta}} &= \frac{\ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial\beta_{lj}} = \sum_{i=i}^{n} \frac{\partial\ell_i(\mu_{ji},\sigma_{ji}^2)}{\partial\mu_{ji}} \frac{d\mu_{ji}}{d\eta_{ji}} \frac{\partial\eta_{ji}}{\partial\beta_{jl}}, \\ U_{\boldsymbol{\gamma}} &= \frac{\ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial\gamma_{lj}} = \sum_{i=i}^{n} \frac{\partial\ell_i(\mu_{ji},\sigma_{ji})}{\partial\sigma_{ji}^2} \frac{d\sigma_{ji}^2}{d\xi_{ji}} \frac{\partial\xi_{ji}}{\partial\gamma_{jl}}, \quad \text{and} \\ U_\lambda &= \frac{\ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial\lambda} = \sum_{i=i}^{n} \frac{\partial\ell_i(\mu_{ji},\sigma_{ji})}{\partial\lambda}, \end{aligned}$$

for all $l = 1, \ldots, k_1(k_2)$ and j=1,2. Under certain regularity conditions, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is approximately normally distributed with zero mean and variance-

covariance matrix $J^{-1}(\boldsymbol{\theta})$, where $J(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$. This allows for constructing confidence intervals, hypotheses testing, and making predictions.

Estimates of the parameters associated with the mean in bivariate simplex regression models can be interpreted as follows: Increasing the l-th independent variable by c units while keeping the other variables unchanged, implies that for all $j = 1, 2$ and $i = 1, \ldots, n$, the log odds ratio is given by $\log(\mu^*_{ji}/(1-\mu^*_{ji})) = \beta_{0j} + \ldots + (x_{il} + c)\beta_{lj} + \ldots + x_{ik_1}\beta_{lj}$. Thus, the odds ratio associated with an increase of c units in the l-th covariate is given by:

$$\exp(c\beta_{lj}) = \frac{\mu^*_{ji}/(1-\mu^*_{ji})}{\mu_{ji}/(1-\mu_{ji})}.$$

3.1. Diagnostic analysis

Diagnostic analysis is a crucial phase in statistical modeling. One commonly employed technique is the analysis of residuals, which is essential for assessing model adequacy and identifying atypical observations. In the literature, diagnostic techniques and analysis for the class of regression models limited to the unit interval (0,1) can be found at Espinheira, Ferrari, and Cribari-Neto (2008), Ferrari and Cribari-Neto (2004), Lemonte and Bazán (2016), among others. Alternatively, the randomized quantile residual, introduced by Dunn and Smyth (1996), has been used for its properties. This residual is defined as $r^q_i = \Phi^{-1}\{F(y_i;\boldsymbol{\theta})\}$, $i = 1, \ldots, n$, where $\Phi^{-1}(\cdot)$ represents the inverse of the cumulative distribution function of the standard normal distribution and $F(\cdot)$ the cumulative distribution function of y. This residual can easily be adapted to the proposed model, as it has a known asymptotic distribution (standard normal) under the postulated model. In addition, the marginal distributions are known, and the maximum likelihood estimators have good properties, as evaluated in the simulation study in Section 4. It is, therefore, directly applicable to regression modeling, and its distribution is easy to check using the usual normality test and graphical evaluation procedures.

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top$ be a random vector, where each $\boldsymbol{y}_i = (y_{1i}, y_{2i})^\top$ for $i = 1, \ldots, n$, follows a bivariate simplex distribution. The randomized quantile residuals for the bivariate regression model are given by:

$$\begin{aligned} r^q_i &= \Phi^{-1}\left\{F(\boldsymbol{y}_{ji};\boldsymbol{\theta})\right\} = \Phi^{-1}\left\{\iint_R f(\boldsymbol{y}_{ji};\hat{\boldsymbol{\mu}}_{ji},\hat{\boldsymbol{\sigma}}^2_{ji},\hat{\lambda})\right\}, \\ &= \Phi^{-1}\left\{\iint_R f_1(y_{1i};\hat{\mu}_{1i},\hat{\sigma}^2_{1i}) f_2(y_{2i};\hat{\mu}_{2i},\hat{\sigma}^2_{2i})\{1+\hat{\lambda}[1-2F_1(y_{1i})][1-2F_2(y_{2i})]\}\right\}. \end{aligned}$$

Evaluating the veracity of the hypothesis regarding the probability distribution assumed for the response variable given the covariates is necessary. According to Atkinson (1985), to better interpret the normal probability plot of the proposed residuals, it should be supplemented with envelopes-simulated bands obtained by Monte Carlo methods from the fitted model. These envelopes help assess whether there are significant deviations from 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 $\Phi(\cdot)$ 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.

Alternatively, sensitivity measures under small perturbations in the regression model according to case exclusion (Cook 1977) are commonly used in diagnostic analysis. Excluding observations is a standard method for examining the impact of removing the i-th observation from the data set. For the bivariate simplex regression model presented in (3), let $\ell(\boldsymbol{\theta}_{(i)})$ denote the log-likelihood function with the i-th observation removed, and $\widehat{\boldsymbol{\theta}}^*$ the maximum likelihood estimator of $\boldsymbol{\theta}$ obtained under $\ell(\boldsymbol{\theta}_{(i)})$. The generalized Cook's distance is defined by $GD_i(\boldsymbol{\theta}) = (\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}})^\top[-\ddot{\ell}(\boldsymbol{\theta})](\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}})$, where $-\ddot{\ell}(\boldsymbol{\theta})$ is the observed information matrix. Another commonly used measure is the likelihood ratio, given by $LD_i(\boldsymbol{\theta}) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_{(i)})\}$. In summary, if $\widehat{\boldsymbol{\theta}}_{(i)}$ significantly deviates from $\widehat{\boldsymbol{\theta}}$, the i-th observation may be considered influential (Thomas and Cook 1989).

4. Simulation study

In this section, a Monte Carlo simulation study is conducted to assess the asymptotic behavior of the maximum likelihood estimators. We performed $R = 1,000$ Monte Carlo replicas for sample sizes of 25, 50, 75, and 100. The study evaluates the bias, the root mean square error (RMSE), and 95% confidence interval coverage for the maximum likelihood estimators: $\text{Bias}(\widehat{\theta}) = \sum_{r=1}^{R}(\widehat{\theta}_r - \widehat{\theta})/R$, $\text{RMSE}(\widehat{\theta}) = \sqrt{\sum_{r=1}^{R}(\widehat{\theta}_r - \widehat{\theta})^2/R}$ and $\text{Coverage} = \#(\widehat{\theta} \in \text{IC}[\theta; 1-\alpha])/R$, where $\text{IC}[\theta; 1-\alpha]$ is the confidence interval for θ, α the significance level and $\widehat{\theta}$ the estimation of some element of θ. According to Johnson (1987), random samples from a population $f(\boldsymbol{y}; \boldsymbol{\theta})$ are generated using the following algorithm: (*i*) Generate u_1 and v, which are independent random variables with distribution $U(0,1)$; (*ii*) Compute $A = \lambda(2u_1 - 1) - 1$ and $B = [1 - \lambda(2u_1 - 1)]^2 + 4v\lambda(2u_1 - 1)$; (*iii*) Calculate $u_2 = 2v/(\sqrt{B} - A)$; (*iv*) Obtain the variables $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ using $y_{1i} = F_2^{-1}(u_{1i}|\mu_{1i}, \sigma_{1i}^2)$ and $y_{2i} = F_1^{-1}(u_{2i}|\mu_{2i}^2, \sigma_{2i}^2)$, where $F_j^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the univariate Simplex distribution, for $j = 1, 2$.

The systematic structure is considered as $\log(\mu_{1i}/(1-\mu_{1i})) = \beta_{10} + \beta_{11}x_i$, $\log(\sigma_{1i}^2) = \gamma_{10} + \gamma_{11}x_i$, $\log(\mu_{2i}/(1-\mu_{2i})) = \beta_{20} + \beta_{21}x_i$, and $\log(\sigma_{2i}^2) = \gamma_{20} + \gamma_{21}x_i$, with $x_i \sim U(0,1)$. For the $\boldsymbol{\theta} = (\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}, \lambda)^\top$, we consider 3 vectors of true parameters, namely: $\boldsymbol{\theta}_1 = (-3.5, 1.2, -3.5, 1.2, -0.8, 1.6, -0.8, 1.6, 1.0)^\top$ (the mean of the variables $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are concentrated around zero), $\boldsymbol{\theta}_2 = (-0.5, 1.2. - 0.5, 1.2, -1.5, 1.3, -1.5, 1.3, 1.0)^\top$ (the mean of the variables $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are concentrated around 0.5), $\boldsymbol{\theta}_3 = (2.5, 1.2, 2.5, 1.2, 0.8, 1.6, 0.8, 1.6, 1.0)^\top$ (the mean of the variables $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are concentrated around one). Figures 3–5 display the surface and contour plots for a single random sample size, corresponding to the parameter vectors $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_3$, respectively. The results for these three different parameter vectors are summarized in Tables 3–5.

Table 3 shows that asymptotically, the bias and root mean square error (RMSE) of the maximum likelihood estimators for the parameter vector $\boldsymbol{\beta} = (\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21})^\top$ approach zero as the sample size increases. Although the parameters γ_{10} and γ_{20} exhibit significant bias in small samples, this bias diminishes with large sample sizes. Additionally, the coverage probability for these parameters is slightly below the nominal level of 95% for small samples but converges to this level as the sample size increases. Similar patterns are observed in the results in Tables 4 and 5. These findings indicate that the maximum likelihood estimators possess desirable asymptotic properties, including unbiasedness, efficiency, and consistency.

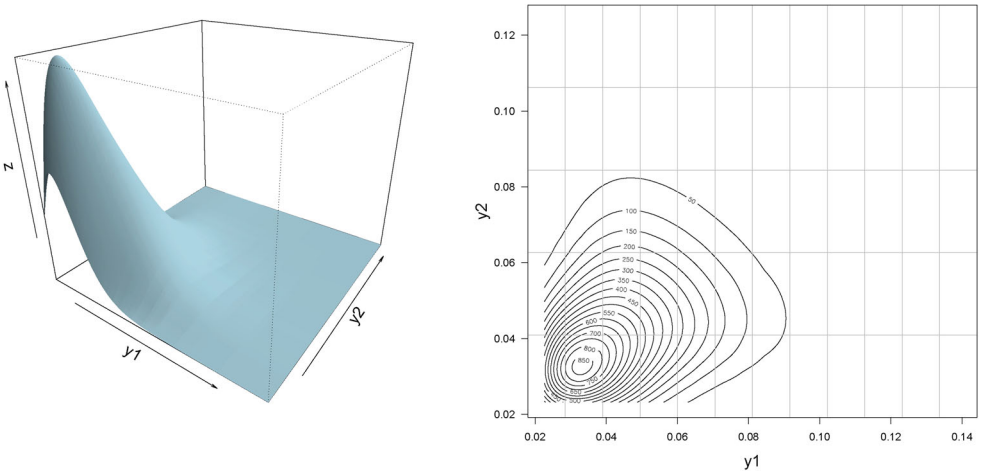

Figure 3. Surface graphs and contour lines for a single sample considering $\boldsymbol{\theta}_1$.

Table 3. Mean, bias, RMSE, and coverage of 95% confidence for $\boldsymbol{\theta}_1$.

n	Measures	β_{10}	β_{11}	β_{20}	β_{21}	γ_{10}	γ_{11}	γ_{20}	γ_{21}	λ
25	Mean	−3.50	1.20	−3.50	1.20	−0.99	1.62	−1.00	1.60	0.46
	Bias	0.00	0.00	0.00	0.00	0.19	−0.02	0.20	0.00	0.04
	RMSE	0.05	0.13	0.05	0.13	0.53	0.99	0.53	0.99	0.43
	Coverage	91.0	92.0	92.0	91.0	91.0	92.0	92.0	91.0	97.0
50	Mean	−3.50	1.20	−3.50	1.20	−0.90	1.64	−0.87	1.58	0.47
	Bias	0.00	0.00	0.00	0.00	0.10	−0.04	0.07	0.02	0.03
	RMSE	0.04	0.11	0.04	0.10	0.38	0.73	0.36	0.69	0.31
	Coverage	92.0	94.0	94.0	94.0	93.0	94.0	94.0	95.0	96.0
75	Mean	−3.50	1.20	−3.50	1.20	−0.86	1.61	−0.87	1.62	0.49
	Bias	0.00	0.00	0.00	0.00	0.06	−0.01	0.07	−0.02	0.01
	RMSE	0.03	0.07	0.03	0.07	0.26	0.47	0.28	0.48	0.26
	Coverage	95.0	94.0	94.0	95.0	96.0	95.0	93.0	95.0	97.0
100	Mean	−3.50	1.20	−3.50	1.20	−0.86	1.64	−0.86	1.64	0.49
	Bias	0.00	0.00	0.00	0.00	0.06	−0.04	0.06	−0.04	0.01
	RMSE	0.02	0.06	0.03	0.06	0.25	0.42	0.24	0.42	0.23
	Coverage	95.0	94.0	94.0	96.0	94.0	94.0	94.0	94.0	96.0

5. Application

In this section, we aim to study the relationship between the variables y_1: the Human Development Index (HDI) by municipality[1], and y_2: the Social Vulnerability Index (SVI) by municipality.[2] We investigate these variables in relation to x, the ratio of the economically dependent population to the economically active population—referred to as the dependency ratio (DR)-which is used as the independent variable. The dependency ratio assumes that younger and older individuals in the population rely economically on others, indicating the burden placed on the productive segment of the population. Table 6 presents some descriptive statistics for the variables under study. It shows that municipalities in Alagoas have an average

[1]The municipal Human Development Index (HDI) is a composite measure consisting of indicators from three dimensions of human development: longevity, education, and income. The index ranges from 0 to 1, where values closer to 1 indicate higher levels of human development, reflecting economic progress and quality of life.

[2]The Social Vulnerability Index (SVI) reflects situations where certain "assets" (employment, housing, human capital, social capital, among others) are absent or insufficient—assets that should ideally be available to all Brazilians. The SVI also ranges from 0 to 1; values closer to 1 indicate higher social vulnerability of the municipality.

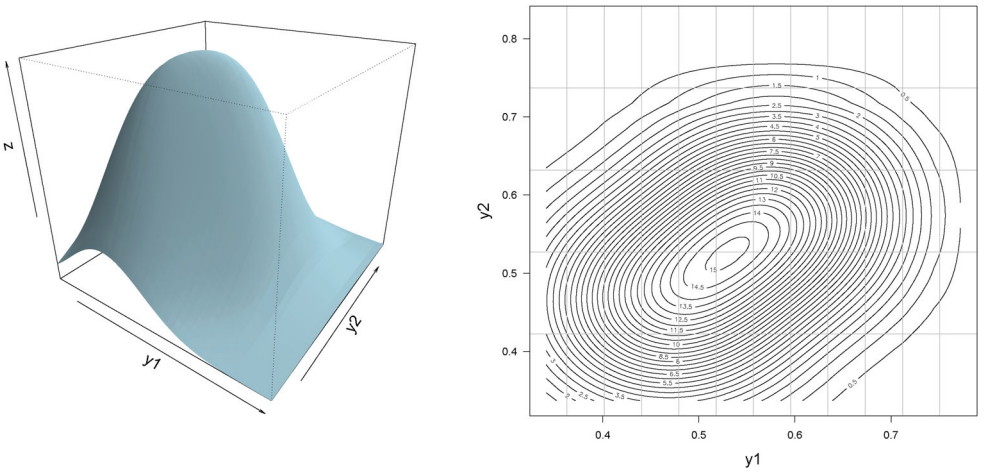

Figure 4. Surface graphs and contour lines for a single sample considering $\boldsymbol{\theta}_2$.

Table 4. Mean, bias, RMSE, and coverage of 95% confidence to $\boldsymbol{\theta}_2$

n	Measures	β_{10}	β_{11}	β_{20}	β_{21}	γ_{10}	γ_{11}	γ_{20}	γ_{21}	λ
25	Mean	−0.50	1.22	−0.50	1.22	−1.62	1.34	−1.62	1.30	0.40
	Bias	0.00	−0.02	0.00	−0.02	0.12	−0.04	0.12	0.00	0.10
	RMSE	0.11	0.23	0.11	0.23	0.63	1.16	0.62	1.17	0.47
	Coverage	92.0	93.0	93.0	93.0	93.0	93.0	94.0	93.0	97.0
50	Mean	−0.51	1.20	−0.51	1.19	−1.56	1.32	−1.58	1.34	0.48
	Bias	0.01	0.00	0.01	0.01	0.06	−0.02	0.08	−0.04	0.02
	RMSE	0.07	0.15	0.07	0.16	0.38	0.76	0.38	0.73	0.35
	Coverage	92.0	93.0	94.0	94.0	94.0	93.0	94.0	94.0	96.0
75	Mean	−0.50	1.20	−0.50	1.20	−1.56	1.31	−1.56	1.32	0.49
	Bias	0.00	0.00	0.00	0.00	0.06	−0.01	0.06	−0.02	0.01
	RMSE	0.05	0.11	0.05	0.11	0.29	0.53	0.29	0.54	0.27
	Coverage	94.0	93.0	94.0	94.0	94.0	95.0	92.0	94.0	97.0
100	Mean	−0.50	1.20	−0.50	1.20	−1.58	1.34	−1.55	1.31	0.50
	Bias	0.00	0.00	0.00	0.00	0.08	−0.04	0.05	−0.01	0.00
	RMSE	0.04	0.08	0.04	0.08	0.25	0.40	0.24	0.40	0.23
	Coverage	94.0	95.0	94.0	94.0	94.0	95.0	95.0	94.0	95.0

municipal HDI and SVI of 48% and 58%, respectively. The average dependency ratio is 68%, suggesting that a sizable portion of the working-age population in Alagoas supports a large number of dependents. Additionally, there is negative skewness (left-tailed) in the HDI and SVI variables and positive skewness (right-tailed) for the DR variable.

Figure 6 displays the relationships among the DR, HDI, and SVI variables. From this point onward, the DR variable is considered in its standardized form. In the left panel, a negative relationship between HDI and DR is observed, indicating that regions with a higher proportion of economically dependent individuals tend to exhibit lower levels of human development. The center panel shows a positive association between SVI and DR, suggesting that greater dependency is linked to increased social vulnerability. Finally, the right panel highlights an inverse relationship between HDI and SVI, reinforcing the idea that lower human development is associated with higher social vulnerability.

Thus, Table 7 presents the estimates, standard errors, and p-value of the parameters for the bivariate simplex regression model using the FGM copula. The table summarizes the estimated coefficients, their corresponding standard errors, and the significance levels of

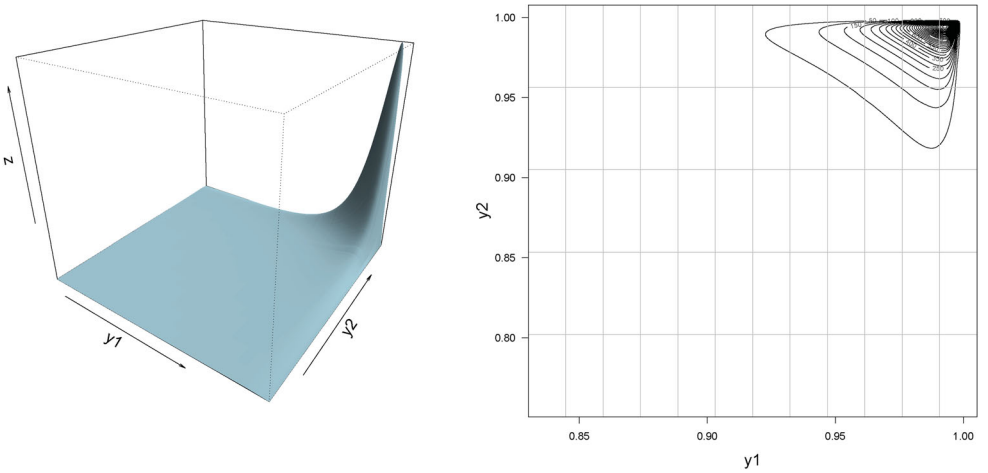

Figure 5. Surface graphs and contour lines for a single sample considering $\boldsymbol{\theta}_3$.

Table 5. Mean, bias, RMSE, and coverage of 95% confidence to $\boldsymbol{\theta}_3$.

n	Measures	β_{10}	β_{11}	β_{20}	β_{21}	γ_{10}	γ_{11}	γ_{20}	γ_{21}	λ
25	Mean	2.51	1.19	2.51	1.20	0.60	1.62	0.62	1.62	0.43
	Bias	−0.01	0.01	−0.01	0.00	0.20	−0.02	0.18	−0.02	0.07
	RMSE	0.11	0.21	0.11	0.21	0.44	0.73	0.43	0.73	0.44
	Coverage	90.0	93.0	92.0	93.0	90.0	93.0	92.0	92.0	96.0
50	Mean	2.51	1.19	2.50	1.19	0.64	1.72	0.68	1.66	0.47
	Bias	−0.01	0.01	0.00	0.01	0.16	−0.12	0.12	−0.06	0.03
	RMSE	0.11	0.19	0.11	0.19	0.42	0.66	0.41	0.68	0.31
	Coverage	92.0	94.0	93.0	94.0	92.0	94.0	93.0	93.0	97.0
75	Mean	2.51	1.19	2.51	1.19	0.75	1.60	0.73	1.62	0.50
	Bias	−0.01	0.01	−0.01	0.01	0.05	0.00	0.07	−0.02	0.00
	RMSE	0.07	0.14	0.07	0.14	0.27	0.50	0.26	0.48	0.26
	Coverage	94.0	95.0	94.0	94.0	94.0	93.0	93.0	94.0	97.0
100	Mean	2.50	1.21	2.51	1.19	0.76	1.60	0.74	1.63	0.50
	Bias	0.00	−0.01	−0.01	0.01	0.04	0.00	0.06	−0.03	0.00
	RMSE	0.07	0.12	0.07	0.12	0.26	0.41	0.25	0.41	0.23
	Coverage	95.0	95.0	94.0	93.0	94.0	95.0	96.0	95.0	96.0

Table 6. Descriptive measures y_1: Human Development Index (HDI), y_2: Social Vulnerability Index (SVI) and x: Dependency Ratio (DR).

Variables	Max.	Min.	1Q.	3Q.	Med.	Md.	DP	Skew	Curt.
y_1	0.72	0.28	0.38	0.56	0.48	0.50	0.09	−0.06	−1.25
y_2	0.82	0.34	0.52	0.64	0.58	0.58	0.09	−0.09	−0.33
x	0.99	0.44	0.60	0.76	0.68	0.67	0.10	0.35	−0.25

each parameter, allowing for a detailed assessment of the model's fit and the relationships between the dependent variables (y_1: HDI and y_2: SVI) and the independent variable x, which represents the standardized Dependency Ratio (DR).

It can be seen from Table 7 that the parameters are significant at the 1% level. As described in Section 3, the estimates associated with the mean parameter can be interpreted in terms of odds ratios. For instance, considering $c = 0.10$ we find that $\exp(0.10 \times \widehat{\beta}_{11}) = \exp(-0.037) \simeq 0.9637$. These results suggest that a 10% increase in the dependency ratio (x) leads, on average, to a 3.63% reduction in the HDI (y_1). In a similar manner, a 10% increase in the dependency

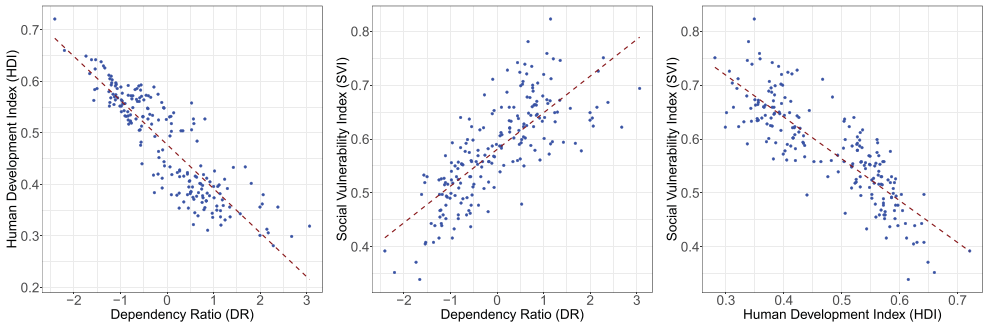

Figure 6. Relationships among the dependency ratio (DR), Human Development Index (HDI), and Social Vulnerability Index (SVI). The left panel displays HDI vs. DR, the center panel shows SVI vs. DR, and the right panel presents HDI vs. SVI.

Table 7. Estimates, standard errors, and *p*-values of the bivariate simplex regression model via the FGM copula.

	Parameters								
	β_{10}	β_{11}	β_{20}	β_{21}	γ_{10}	γ_{11}	γ_{20}	γ_{21}	λ
Estimate	−0.088	−0.370	0.322	0.299	−2.004	0.508	−1.331	0.339	−0.734
Standard Error	0.012	0.014	0.016	0.018	0.103	0.125	0.100	0.103	0.089
p-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000

ratio leads to an estimated 3.04% increase in the social vulnerability index (y_2), given that $\exp(0.10 \times \widehat{\beta}_{12}) = \exp(0.0299) \approx 1.0304$. The parameter λ reflects the dependence between the variables y_1 and y_2, which is related to Kendall's τ measure, given by $\tau = 2\lambda/9$ for the FGM copula. According to Table 7, $\lambda = -0.734$, indicating a negative association. This is consistent with the graphical specification shown in Figure 7 (panel on the right), where we observe an inverse relationship between y_1 and y_2, with Pearson's correlation of approximately $-0.8399 \simeq 83.99\%$ and statistics τ Kendall's equal to approximately $-0.1631 \simeq -16.31\%$.

Regarding the estimates associated with the dispersion parameter, a reduction in dispersion is expected given the negative signs of the estimates for γ_{10} and γ_{20}. Figure 8 displays the graphs of the randomized quantile residuals and the simulated envelope. The randomized quantile residuals are scattered randomly around zero, with no outliers identified. Additionally, the figure indicates a good model fit, as the residuals fall within the bands of the simulated envelope. This suggests that the model adequately captures the data set's underlying structure.

The diagnostic analysis includes the construction of graphs for the global influence measures: generalized Cook's distance and likelihood distance, as illustrated in Figure 9. The analysis reveals several potential influential observations: #38 (municipality Joaquim Gomes; HDI= 0.349, IVS= 0.823 and RD= 0.801), #50 (municipality Minador do Negrão 1; HDI= 0.384, IVS= 0.682 and RD= 0.657), #108 (municipality Belém, HDI= 0.593, IVS= 0.526 and RD= 0.563), #152 (municipality Minador do Negrão 2; HDI= 0.563, IVS= 0.553 and RD= 0.525) and #196 (municipality Maceió, HDI= 0.721, IVS= 0.392 and RD= 0.440). Notably, Joaquim Gomes and Maceió exhibit the highest and lowest SVIs, respectively.

To assess the impact of these observations on the maximum likelihood estimates, we adjusted the bivariate simplex regression model using the FGM copula by sequentially deleting each of the aforementioned observations, as well as all of them together. The impact

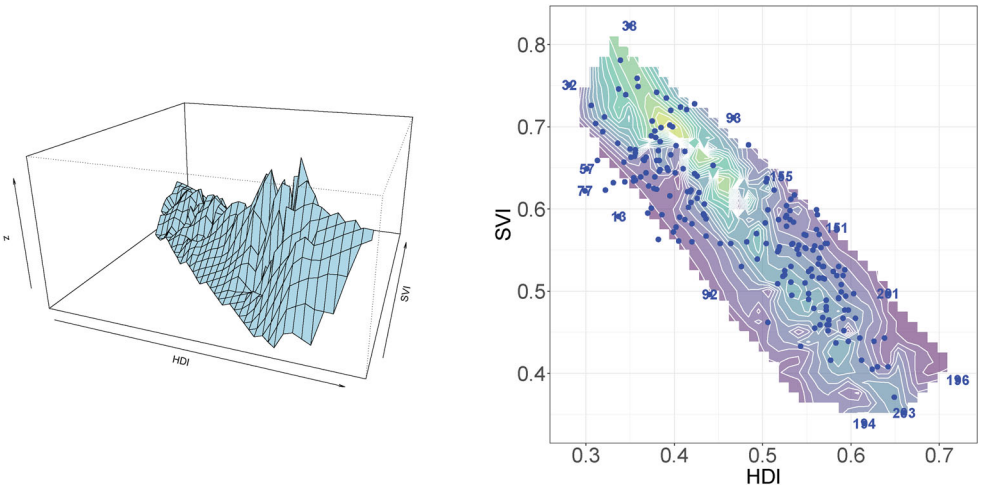

Figure 7. Surface graphs (left panel) and contour lines (right panel) of the variables y_1: human development index and y_2: social vulnerability index.

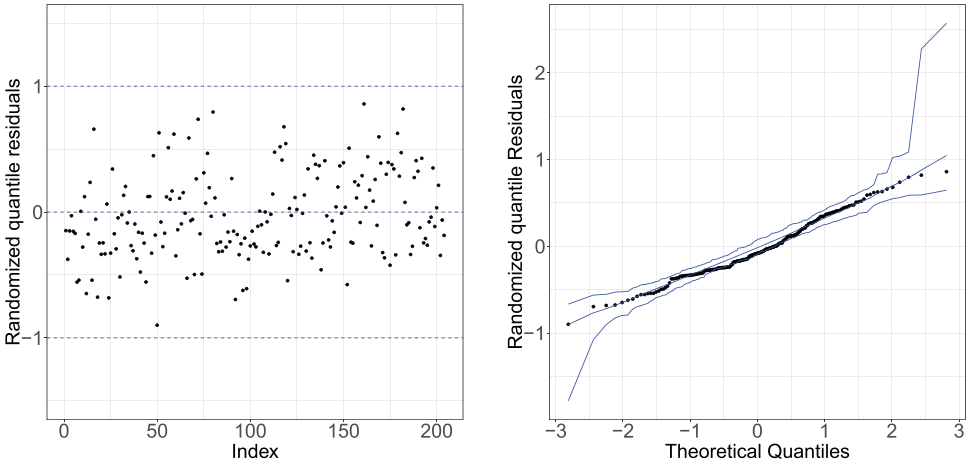

Figure 8. Graph of randomized quantile residuals (left panel) and simulated envelope (right panel).

was measured using the percentage rate of change (PRC), defined as: $\text{PRC} = [(\widehat{\theta}^* - \widehat{\theta})/\widehat{\theta}] \times 100\%$, where $\widehat{\theta}^*$ denotes the parameter estimate excluding the influential observations. With $n = 204$ observations, each observation is expected to influence the parameter estimate by approximately 0.49%.

Table 8 presents the results of the PRC for parameter estimates following the removal of potential influential observations. The analysis shows no significant changes in the estimates of the parameters. However, the most substantial percentage variation was observed in γ_{21} upon deleting observations #38 or #196. Overall, the inferential conclusions remained consistent even after excluding these observations, with all parameters continuing to be statistically significant at the 1% level.

Finally, Figure 7 displays the density surface graph and level curves of the fitted model. These graphs illustrate an inverse relationship between the HDI and SVI variables for the

Table 8. Percentage variation of the parameter estimates of the bivariate simplex regression model via the FGM copula.

	Deleted				
Parameters	#38	#50	#108	#152	#196
β_{10}	-2.71	-1.63	4.86	3.38	6.90
β_{11}	-0.47	0.48	-3.98	1.18	-6.18
β_{20}	-1.57	-0.48	0.55	0.73	1.35
β_{21}	-2.81	0.63	-2.40	1.60	-5.06
γ_{10}	-0.11	1.98	-1.88	1.08	-1.31
γ_{11}	0.53	4.95	-15.96	10.92	-11.04
γ_{20}	2.61	1.64	-2.01	0.86	-3.31
γ_{21}	-17.38	3.71	-8.20	7.47	-17.94
λ	0.02	-1.98	7.30	-2.28	7.58

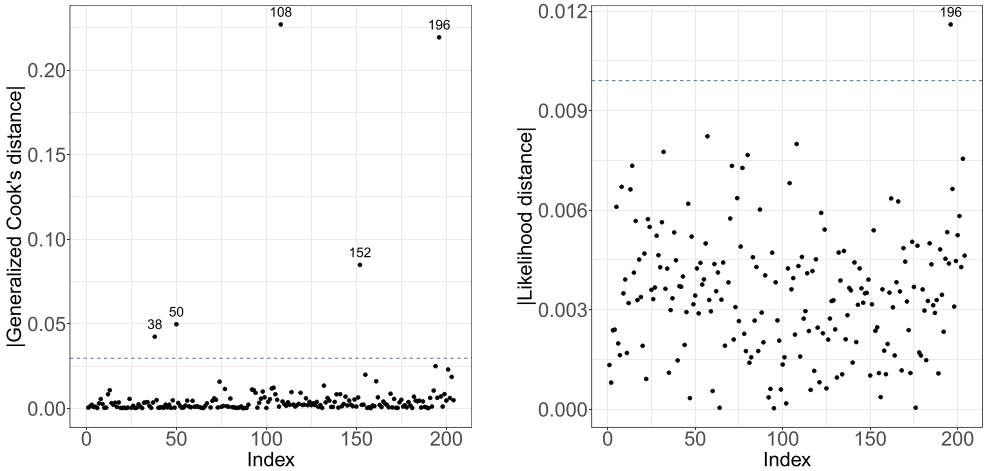

Figure 9. Global influence graphs: Generalized Cook's distance (left panel) and likelihood distance (right panel).

state of Alagoas: as municipal HDI increases, the SVI decreases, and vice versa. In the right panel of Figure 7, the model appears to fit the data well, as the data points are generally enclosed by the contour lines. However, there are nine notable observations that lie outside the contour lines. Observations #13, #32, #38, #57, and #77 correspond to the municipalities of Campo Grande, Inhapi, Joaquim Gomes, Olivença, and Senador Rui Palmeira, respectively. These municipalities are among the least economically developed, characterized by a very low HDI and higher SVI. In contrast, observations #194, #196, #201, and #203 correspond to the municipalities of Barra de São Miguel, Maceió, Rio Largo, and Satuba, respectively. These municipalities are more economically developed, with a higher HDI and lower SVI. Observations that lie outside the contour lines may be considered for removal, as their exclusion does not affect the inferential conclusions regarding the model's fit to the data.

6. Conclusion

In this study, we proposed a bivariate simplex regression model (BSRM) constructed using the Farlie-Gumbel-Morgenstern (FGM) copula to model correlated proportional data. Through theoretical development, maximum likelihood estimation, and a comprehensive simulation

study, we demonstrated the desirable asymptotic properties of the estimators, including consistency and efficiency. The application to socioeconomic data from municipalities in Alagoas, Brazil, provided compelling evidence of the model's practical utility. The analysis revealed a strong inverse relationship between the Human Development Index (HDI) and the Social Vulnerability Index (SVI), with a significant negative dependence parameter ($\lambda = -0.734$), consistent with Kendall's τ and Pearson correlation estimates. The results indicate that a 10% increase in the dependency ratio (DR) was associated with an approximate 3.63% reduction in the HDI and a 3.04% increase in the Social Vulnerability Index. This inverse relationship between HDI and SVI suggests that as a municipality's human development improves, its social vulnerability tends to decrease, and vice versa. The diagnostic measures, including randomized quantile residuals and global influence analysis confirmed that the model fit was adequate, and the robustness of the model in the presence of outlying observations. For future research, several promising directions can be explored. One possibility is to extend the BSRM by incorporating measurement errors in the variables and adapting the model to accommodate different types of copulas. Another important avenue is the application of the local influence method to identify influential observations.

Acknowledgments

The authors sincerely thank the Editor, Associate Editor, and the anonymous reviewer for their valuable feedback and constructive criticism, which have significantly enhanced the presentation and quality of this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The first and second authors gratefully acknowledge the graduate and undergraduate scholarship provided by Fundação de Amparo à Pesquisa do Estado da Bahia – Fapesb and Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, respectively. Jalmar M. F. Carrasco wishes to acknowledge the funding provided by the Institutional Program for Internationalization-CAPES-PrInt (Grant No. 88887.935790/2024 — 00).

References

Amaral, E., L. S. Vieira, V. Barros, L. C. Fabio, and J. M. F. Carrasco. 2025. Bivariate simplex distribution. arXiv preprint arXiv:2504.01210v1.

Atkinson, A. 1985. *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*. Oxford, UK: Clarendon Press.

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.

Carrasco, J. M. F., and N. Reid. 2021. Simplex regression models with measurement error. *Communications in Statistics- Simulation and Computation* 50 (11):3420–35. doi: 10.1080/03610918.2019.1626881.

Cepeda-Cuervo, E., J. A. Achcar, and L. G. Lopera. 2014. Bivariate beta regression models: joint modeling of the mean, dispersion and association parameters. *Journal of Applied Statistics* 41 (3):677–87. doi: 10.1080/02664763.2013.847071.

Cook, R. D. 2000. Detection of influential observation in linear regression. *Technometrics* 42 (1):65–8. doi: 10.1080/00401706.2000.10485981.
Cribari-Neto, F., and A. Zeileis. 2010. Beta regression in r. *Journal of Statistical Software* 32:1–24.
Dunn, P. K., and G. K. Smyth. 1996. Randomized quantile residuals. *Journal of Computational and Graphical Statistics* 5 (3):236–44. doi: 10.1080/10618600.1996.10474708.
Espinheira, P. L., and A. De Oliveira Silva. 2020. Residual and influence analysis to a general class of simplex regression. *TEST* 29 (2):523–52. doi: 10.1007/s11749-019-00665-3.
Espinheira, P. L., S. L. Ferrari, and F. Cribari-Neto. 2008. On beta regression residuals. *Journal of Applied Statistics* 35 (4):407–19. doi: 10.1080/02664760701834931.
Ferrari, S., and F. Cribari-Neto. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31 (7):799–815. doi: 10.1080/0266476042000214501.
Johnson, M. E. 1987. *Multivariate statistical simulation*. New York: John Wiley & Sons.
Jorgensen, B. 1997. *The theory of dispersion models*. New York: Chapman & Hall.
Koochemeshkian, P., N. Manouchehri, and N. Bouguila. 2020. Bivariate beta regression model and its medical applications. In *2020 International Symposium on Networks, Computers and Communications (ISNCC)*. Montreal, QC, Canada.
Lemonte, A. J., and J. L. Bazán. 2016. New class of Johnson SB distributions and its associated regression model for rates and proportions. *Biometrical Journal. Biometrische Zeitschrift* 58 (4):727–46. doi: 10.1002/bimj.201500030.
Liu, P., K. C. Yuen, L.-C. Wu, G.-L. Tian, and T. Li. 2020. Zero-one-inflated simplex regression models for the analysis of continuous proportion data. *Statistics and Its Interface* 13 (2):193–208. doi: 10.4310/SII.2020.v13.n2.a5.
Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* 135 (3):370 doi: 10.2307/2344614.
Nelsen, R. B. 2006. *An introduction to copulas*. New York: Springer.
Paolino, P. 2001. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis* 9 (4):325–46. doi: 10.1093/oxfordjournals.pan.a004873.
Qiu, Z., P. X. 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. 2024. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Rocha, A. V., and A. B. Simas. 2011. Influence diagnostics in a general class of beta regression models. *TEST* 20 (1):95–119. doi: 10.1007/s11749-010-0189-z.
Simas, A. B., W. Barreto-Souza, and A. V. Rocha. 2010. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis* 54 (2):348–66. doi: 10.1016/j.csda.2009.08.017.
Song, P. X., Z. Qiu, and M. Tan. 2004. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal* 46 (5):540–53. doi: 10.1002/bimj.200110052.
Souza, D. F., and F. A. S. Moura. 2016. Multivariate beta regression with application in small area estimation. *Journal of Official Statistics* 32 (3):747–68. doi: 10.1515/jos-2016-0038.
Thomas, W., and R. D. Cook. 1989. Assessing influence on regression coefficients in generalized linear models. *Biometrika* 76 (4):741–9. doi: 10.1093/biomet/76.4.741.
Vasconcellos, K. L., and F. Cribari-Neto. 2005. Improved maximum likelihood estimation in a new class of beta regression models. *Brazilian Journal of Probability and Statistics*:13–31.
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. *Journal of Mathematics* 38:1863–75.