Revista Colombiana de Estadística - Applied Statistics
July 2024, volume 47, issue 2, pp. 283–300
http://doi.org/10.15446/rce.v47n2.111306

Unit Regression Models to Explain Vote Proportions in the Brazilian Presidential Elections in 2018

Modelos de regresión unitaria para explicar las proporciones de votos en las elecciones presidenciales de Brasil en 2018

Renata Rojas Guerra[1,a], Fernando A. Peña-Ramírez[1,b], Tatiane Fontana Ribeiro[3,c], Gauss M. Cordeiro[4,d], Charles Peixoto Mafalda[4,e]

[1]Departamento de Estatística, Centro de Ciências Exatas e da Natureza, Universidade Federal de Santa Maria, Santa Maria, Brasil

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

[3]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil

[4]Departamento de Estatística, Centro de Ciências Exatas e da Natureza, Universidade Federal de Pernambuco, Recife, Brasil

Abstract

In this paper, we aim to identify the covariates associated with the proportion of votes of candidates elected in Brazilian municipalities with a population of more than 300,000 inhabitants. We analyzed the vote proportions from the 2018 presidential runoff election using distributions within the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) class. Unit distributions are quite useful for modeling vote proportions due to their flexibility to accommodate asymmetry and heavy tails. Furthermore, they provide adequate representations of the physiological properties and the empirical distribution of the data. We fit the beta, simplex, unit gamma, and unit Lindley regression models, considering random and fixed effects components to verify spatial correlation among the municipalities. The beta regression with fixed components regarding Brazilian regions is superior. The covariates with significant effects are the proportion of evangelicals, monthly household income per capita, the political spectrum of the

[a]Tenured Assistant Professor. E-mail: renata.r.guerra@ufsm.br
[b]Tenured Assistant Professor. E-mail: fernando.p.ramirez@ufsm.br
[c]Ph.D(c). E-mail: tatianefr@ime.usp.br
[d]Full Professor. E-mail: gauss@de.ufpe.br
[e]Ph.D(c). E-mail: charles1995peixoto@hotmail.com

governors' party elected in 2014 and 2018, and if the municipality is the capital of the state. We note that some Brazilian regions impact the vote proportions' mean and dispersion.

***Key words*:** Blackbeta regression; Brazilian elections; Double-bounded variables; GAMLSS.

Resumen

En este artículo, nuestro objetivo es identificar las covariables asociadas con la proporción de votos de los candidatos electos en municipios brasileños con una población de más de 300,000 habitantes. Analizamos las proporciones de votos de la segunda vuelta de las elecciones presidenciales de 2018 utilizando distribuciones dentro de la clase de Modelos Aditivos Generalizados para localización, Escala y Forma (GAMLSS). Las distribuciones unitarias son muy útiles para modelar proporciones de votos debido a su flexibilidad para acomodar asimetría y colas pesadas. Además, proporcionan representaciones adecuadas de las propiedades fisiológicas y la distribución empírica de los datos. Ajustamos los modelos de regresión beta, simplex, gamma unitario y Lindley, considerando componentes de efectos aleatorios y fijos para verificar la correlación espacial entre los municipios. La regresión beta con componentes fijos respecto a las regiones brasileñas es superior. Las covariables con efectos significativos son la proporción de evangélicos, el ingreso mensual por hogar per cápita, el espectro político del partido de los gobernadores elegidos en 2014 y 2018, y si el municipio es la capital del estado. Notamos que algunas regiones brasileñas impactan en la media y la dispersión de las proporciones de voto.

***Palabras clave*:** Elecciones brasileñas; GAMLSS; Regresión beta; Variables de doble límite.

1. Introduction

Double-bounded variables, such as rates and proportions, usually show asymmetry and heavy tails. The assumption of normality for this type of variables can be inadequate for many appllications. For these situations, one should look for more flexible models that provide adequate representations for the physiological properties and the empirical distribution of the data (Pereira et al., 2014). The beta (Ferrari & Cribari-Neto, 2004) and simplex (Barndorff-Nielsen & Jørgensen, 1991) regressions are common procedures to accommodate these features. These models resemble the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) (Stasinopoulos et al., 2018) by allowing a regression structure in the mean and the dispersion (or precision) parameter of the beta and simplex distributions, respectively. The GAMLSS allows all parameters of the response distribution to vary with explanatory variables and provide a comprehensive framework for easily incorporating nonlinear, random, and spatial effects. One advantage of using this class is its flexibility to accommodate several types of random components. It allows fitting double-bounded variables and can also be considered to describe the behavior of variables with support in the real numbers and asymmetric behaviors, as well as for those that take on positive values and discrete

outcomes. Another advantage is the ease of performing applications through the packages available in the R programming language. In this regard, we can refer to Ribeiro, Seidel, Guerra, Peña-Ramírez & da Silva (2021), Regis et al. (2023), and de Araújo et al. (2022) as some recent contributions in terms of proposing or utilizing models in the GAMLSS class.

Our main goal is to explain the mean variations of the vote proportions received by Jair Bolsonaro in the runoff of the 2018 presidential elections. Some works have used beta regression for previous Brazilian presidential elections. Andrade et al. (2013) assessed the impacts of welfare programs and economic growth on the outcome of the 2006 election. Almeida Junior & Souza (2015) investigated the same in the 2010 election and for the Northeast region. For the 2018 elections, Hunter & Power (2019), and Rennó (2020) carried out some analysis on the political landscape. Yero et al. (2020) analyzed the effects of development indicators in the 2018 elections using machine learning algorithms. They reported that voters residing in less developed regions have left-wing parties as the preferred choice. However, to our best knowledge, a unit regression analysis modeling the 2018 elections has not been carried out. Another novelty of this work is to verify the effect of spatial correlation among the Brazilian municipalities. To this aim, we consider random effects and fixed effects components related to the Region of the municipalities and its latitudes e longitudes.

We consider the beta, simplex, unit gamma (UG) and unit Lindley (UL) regressions since they have mean-based parametrizations. Alternative unit regressions have been proposed in recent years, but most of them focus on quantile parametrizations. We can cite, for example, Bayes et al. (2017) for quantile regression based on the Kumaraswamy distribution and Lemonte & Bazán (2016) for median regression from the Johnson S_B distribution. Other recent advances can be found in Mazucheli et al. (2020), Guerra et al. (2020), Ribeiro, Cordeiro, Pena-Ramírez & Guerra (2021), and Ribeiro et al. (2022). Those models are not considered in the analysis because our interest lies in the effect of socio-demographic and economic indicators in the mean of the vote proportions.

The rest of the paper is divided as follows. Section 2 presents a theoretical background on the beta, simplex, UG, and UL regression models. In Section 3, a descriptive analysis and the data preparation is presented. Section 4 discusses the fitted regressions and the effects of the explanatory variables in the vote proportions. The concluding remarks are outlined in Section 5.

2. Theoretical Background

This section aims to discuss the unit regression models employed in the vote proportion analysis. We present some aspects of parameter estimation, model selection, and diagnostic analysis on unit distributions reparametrized on the mean. By elucidating these fundamental concepts, we aim to establish the theoretical foundation for the subsequent empirical analysis.

2.1. Mean-Based Unit Distributions

The use of mean-based parameterizations enhances the interpretability and applicability of these distributions within the GAMLSS framework. The classical method in this context is the beta regression, as introduced by Ferrari & Cribari-Neto (2004). In this model, the mean response is related to a linear predictor, which involves covariates and unknown regression parameters, through a link function. The model is also indexed by a precision parameter and assumes that the dependent variable has a beta distribution.

The beta regression model has been considered by several authors. For example, Bayer et al. (2018) proposed beta regression control charts to monitor the tire manufacturing process and the relative humidity in Brasília, Brazil. Ghosh (2019) presented a study on robust inference for the model, with application to health studies. Karlsson et al. (2020) introduced a Liu estimator for the beta regression model and performed an application to chemical data. Espinheira et al. (2019) investigated model selection criteria on beta regression for machine learning.

Assuming that the precision parameter is also related to a linear predictor, Simas et al. (2010) formally introduced the varying precision beta regression model. Moreover, an alternative parametrization in terms of a dispersion (not precision) parameter is considered by Cribari-Neto & Souza (2012), Bayer & Cribari-Neto (2017) and Canterle & Bayer (2019). Let Y be a beta random variable indexed by the mean $\mu \in (0,1)$ and the dispersion parameter $\sigma \in (0,1)$, say $Y \sim \text{Beta}\,(\mu,\sigma)$. Its probability density function (pdf) is (for $y \in (0,1)$)

$$f(y;\mu,\sigma) = \frac{\Gamma\left(1/\sigma^2-1\right)}{\Gamma\left(\mu(1/\sigma^2-1)\right)\Gamma\left((1-\mu)(1/\sigma^2-1)\right)} \times y^{\mu(1/\sigma^2-1)-1}(1-y)^{(1-\mu)(1/\sigma^2-1)-1}, \tag{1}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x$ is the gamma function. Under this parameterization, the variance of Y is $\text{Var}(Y) = \sigma^2\mu(1-\mu)$.

The simplex distribution was introduced by Barndorff-Nielsen & Jørgensen (1991) as an alternative to the beta distribution. Let $Y \sim \text{S}(\mu,\sigma^2)$ be a simplex random variable, which pdf is (for $y \in (0,1)$)

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

where $\mu \in (0,1)$ is the mean of Y and σ^2 is a dispersion parameter. Espinheira & Silva (2019) proposed a general class of simplex regression models, in which the mean and the dispersion parameters can be related to a linear predictor. They also perform residual and influence analysis to the simplex regression.

Several studies have been developed considering simplex regression models. Carrasco & Reid (2019) studied measurement errors. López (2013) considered a Bayesian approach to parameter estimation in a simplex regression model and compared with the beta regression. Cordeiro et al. (2020) used the beta and simplex regression models to explain homicides in state Brazilian capitals.

The UG distribution is also a relevant alternative for double-bounded outcomes. Its mean-based parametrization was pionnered by (Mousa et al., 2016), who also introduced the UG regression. Let $Y \sim \mathrm{UG}(\mu, \sigma^2)$ be a UG random variable with pdf (for $y \in (0,1)$)

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

where $\mu \in (0,1)$ is the mean of Y and $\sigma > 0$ can be interpreted as a precision parameter.

The UG regression model has been the focus of several studies, with advancements including investigations by Guedes et al. (2020) into hypotesis testing inferences using a modified likelihood ratio test. de Freitas et al. (2023) explored UG regression models for correlated bounded data. Petterle et al. (2023) proposed a mixed regression models for a response variable following the UG distribution.

Finally, we present the UL distribution as an alternative mean-based regression model for double-bounded random variables. This model regression was recently introduced by Mazucheli et al. (2019) and is an interesting option due to its simplicity, since the UL is an one-parameter distribution that can be defined as its mean. Let $Y \sim \mathrm{UL}(\mu)$ be a UL random variable with pdf (for $y \in (0,1)$)

$$f(y;\mu) = \frac{(1-\mu)^2}{\mu(1-y)^3} \exp\left\{-\frac{y(1-\mu)}{\mu(1-y)}\right\}, \tag{4}$$

where $\mu \in (0,1)$ is the mean of Y. Generalizations of the UL regression include the analysis of correlated data based on estimating equations (Silva et al., 2023) and the UL mixed-effect model (Akdur, 2021).

2.2. The GAMLSS Framework

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ be a set of independent random variables such that $\boldsymbol{Y} \sim \mathbf{D}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, where $\mathbf{D}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ denotes a two-parameter unit distribution and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^\top$ are related to the predictors $\boldsymbol{\eta}_1 = (\eta_{11}, \ldots, \eta_{1n})^\top$ and $\boldsymbol{\eta}_2 = (\eta_{21}, \ldots, \eta_{2n})^\top$, respectively. The GAMLSS structure can be defined in a random effects form as (de Bastiani et al., 2018)

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_1\boldsymbol{\gamma}_1, \tag{5}$$

$$\boldsymbol{\eta}_2 = g_2(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}_2\boldsymbol{\gamma}_2, \tag{6}$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are known design matrices, $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{J_1'1})^\top$ is a parameter vector of length J_1', $\boldsymbol{\beta}_2 = (\beta_{12}, \ldots, \beta_{J_2'2})^\top$ is a parameter vector of length J_2', $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are design matrices of a single factor, or the identity matrix for a random effect at the observational level, $\boldsymbol{\gamma}_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\sigma}_k^2\boldsymbol{I})$ (for $k = 1, 2$) and $\boldsymbol{\sigma}_k^2$ being the variance of the random effect. The functions $g_1(\cdot)$ and $g_2(\cdot)$ are strictly monotonous and twice differentiable link functions.

All the unit regressions are defined are defined as in (5) $g_1 : (0,1) \to \mathbb{R}$, but differing on the assumption of the random component and the mapping required for the $\boldsymbol{\sigma}$ link function. The beta regression has the random component following from (1), in which $g_2 : (0,1) \to \mathbb{R}$ in (6). The simplex and UG regressions have $g_2 : \mathbb{R}^+ \to \mathbb{R}$ in (6) and their random components follow from (2) and (3), respectively. Finally, the UL regression is defined only from (5), with its random component following (4). In this work, we define the logit link function for the parameters supported in the standard unit interval and the log link function for the positive parameters.

It is worth nothing that, while the GAMLSS general framework accommodates any four-parameter distribution, exponential family or non-exponential family, the beta, simplex, UG, and UL distributions entail up to two parameters. Also, considering our purpose is to use random effects to accommodate spatial correlation among the municipalities' regions, we have included only one grouping factor in the random effect components. Hence, Equations (5) and (6) present the framework considered in this work which is simpler than the general case.

The parameter estimation is conducted using the `gamlss` package (Rigby & Stasinopoulos, 2005) in `R` (R Core Team, 2024). We employ the `random` function to estimate the random effects. This function utilizes a penalized quasi-likelihood approach, wherein the model's fitted values are derived from the joint likelihood function. Meanwhile, inference regarding the behavior of the response variable is based on the conditional likelihood given the random effects. Regarding the optimization method, the package provides two basic algorithms for maximizing the penalized log-likelihood: the CG and RS algorithms. The CG algorithm uses the first, second, and cross-derivatives of the penalized log-likelihood concerning the distribution parameters. In contrast, the RS algorithm does not use the cross-derivatives of the log-likelihood. This paper uses the RS algorithm as a default and calls upon the CG algorithm when the first method fails. We refer to Stasinopoulos et al. (2017) for more information on the estimation process. It is worth noting that the beta and simplex distributions are available in the `gamlss` package. This is not the case of the UG and UL distributions. Nevertheless, those models have been made available by Guerra (2024).

2.3. Model Selection and Diagnostics

The selection of the appropriate GAMLSS model comprehend two different stages: fitting and diagnostics (Stasinopoulos et al., 2017). The fitting stage involves the comparison of fitted models from different distributions and variables and can be done using criterion-based methods. At this stage, the generalized Akaike information criterion (GAIC) can be used for comparing non-nested GAMLSS models. This criterion adds to the fitted deviance a penalty for each effective degree of freedom used in the model. Other common measures available for comparing fitted models include the pseudo-R^2.

The diagnostic stage involves the residual analysis, in which it is usually employed the quantile residuals defined as

$$r_i^q = \Phi^{-1}\{F(y_i; \hat{\mu}_i, \hat{\sigma}_i)\}, \tag{7}$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution and $F(\cdot)$ is the cumulative distribution function.

The main advantage of these residuals is that they follows a standard normal distribution when the model is correctly specified. Here, normality tests such as the well-known *Shapiro-Wilk* (SW) test can be used to check this assumption.

For a given distribution, the selection of explanatory variables is one of the most important subjects in statistical modeling and must to be done for all the distribution parameters. There are several functions within the `gamlss` package to assist in this process. In this paper, we use the function `stepGAIC()` to perform a backward procedure in order to select the set of explanatory variables using the AIC. At this step, the likelihood ratio tests (LRT) can also be considered to compare nested models. In addition, citeasnounstasinopoulos2017 suggest performing diagnostic tests after selecting predictors to verify distributional assumptions about model residuals.

3. Data Processing and Descriptive Analysis

The variable of interest is the proportion of valid votes received by Jair Bolsonaro in the 2018 presidential elections runoff (Prop_PSL). We consider the data from Brazilian municipalities with a population greater than 300,000 in the 2010 census, totaling 79 observations. Some socio-demographic indicators and the political spectrum of the governors' party are selected as explanatory variables for the vote proportions. The variables were collected using public data from the Brazilian Tribunal Superior Eleitoral (TSE) (TSE, 2018), Instituto Brasileiro de Geografia e Estatística (IBGE) (SIDRA, 2020), and Instituto de Pesquisa Econômica Aplicada (IPEADATA) (BRASIL, 2020). Table 1 lists all variables and their respective sources.

The variable MHIC is calculated as the ratio between total family income (in nominal terms) and the total number of residents. Income from work and other sources for all residents is considered, including those classified as retired, domestic workers, and relatives of domestic workers (IBGE, 2010). We also analyze the governors' party elected in the Brazilian state where the cities belong. Since the Brazilian political system has many parties, the variables PG_2014 and PG_2018 are defined from the political spectrum of the governors' party elected in 2014 and 2018 elections, respectively. The political spectrum of the party is obtained from Sardinha & Costa (2020), and the governors' party from (BRASIL, 2020). The variable Region is defined since several authors carried out studies on the impact of the Brazilian regions on presidential elections, see Almeida Junior & Souza (2015), Zucco Jr (2013), and Zucco Jr (2015), for instance. The LAT and LONG variables are included since they can be considered as alternatives to the Region to compute the spacial correlation among the municipalities.

Table 1: Description and sources of the variables

Variable	Source	Description
Prop_PSL	TSE	Proportion of valid votes recieved by Jair Bolsonaro in the 2018 presidential elections runoff.
EP	IBGE	Estimated population in the 2010 census, per 100 000.
PE	IBGE	Proportion of evangelicals in the 2010 census. (2010 census).
LR	IBGE	Literacy rate in the 2010 census.
MHIC	IBGE	Monthly household income per capita in reals, R$, in the 2010 census.
DD	IBGE	Demographic density in the 2010 census.
Region	-	Brazilian region to which the municipality belongs (South, Northeast, Southeast, North, and Midwest).
PG_2014	IPEADATA and Sardinha & Costa (2020)	Political spectrum of the governors' party elected in 2014 (left-wing, centre, right-wing).
PG_2018	IPEADATA and Sardinha & Costa (2020)	Political spectrum of the governors' party elected in 2018 (left-wing, centre, right-wing).
Cap_BR	-	Brazilian capital to which the city belongs.
LAT	IBGE	Latitude of the municipality.
LONG	IBGE	Longitude of the municipality.

Table 2 gives some descriptive measures of the response variable and quantitative covariates, except LAT and LONG. The coefficient of variation (CV) indicates that EP has the highest variability and LR has the lowest degree of variability. The mean proportion of votes is 0.63, and the amplitude is 0.53. Since the median is about 0.66, the elected candidate won in most municipalities considered. We have negative skewness and kurtosis coefficients, thus indicating that the larger vote proportion values have a fatter tail than the smallest ones. These features can be confirmed through the plots displayed in Figure 1.

Table 2: Descriptive statistics for the response variable and quantitative covariates

Variable	Mean	Median	Skewness	Kurtosis	Min.	Max.	CV
Prop_PSL	0.628	0.655	−0.527	−0.368	0.314	0.840	19.279
EP	9.020	4.720	5.378	32.976	3.005	112.535	160.160
PE	0.267	0.255	−0.024	−0.291	0.117	0.410	24.688
MHIC	1,100.113	1,042.840	0.723	−0.172	439.720	2,159.170	36.127
DD	2,515.933	1,315.270	1.722	2.490	12.570	13,024.560	119.971
LR	95.328	96.300	−1.710	2.732	85.500	98.300	2.741

From the boxplot of Prop_PSL (Figure 1), we highlight that the cities with the lowest values in relation to the vote proportion are Salvador (Bahia), Caucaia (Ceará), Feira de Santana (Bahia), and Teresina (Piauí), which registered values at 0.314, 0.3667, 0.3715 and 0.3726, respectively. It is noteworthy that all the cities are located in the Northeast of Brazil, and Caucaia is also between the three cities with the lowest MHIC. On the other hand, Besides, Blumenau (Santa Catarina) is the city with the highest values for Prop_PSL and LR.

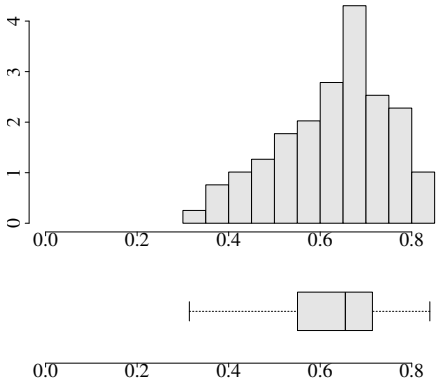


FIGURE 1: Histogram and boxplot of the vote proportion data

Figure 2 displays dispersion plots of the response variable versus the quantitative covariates, except LAT and LONG. Since there is no linear relationship between the variables, we use the Spearman method to compute the correlation matrix (see Table 3). The results indicate that the response variable is positively correlated with most variables. The highest correlation is with LR, followed by the MHIC. This high correlation shows that covariates are essential for the study, as they indicate that LR and MHIC had are associated with the response variable.

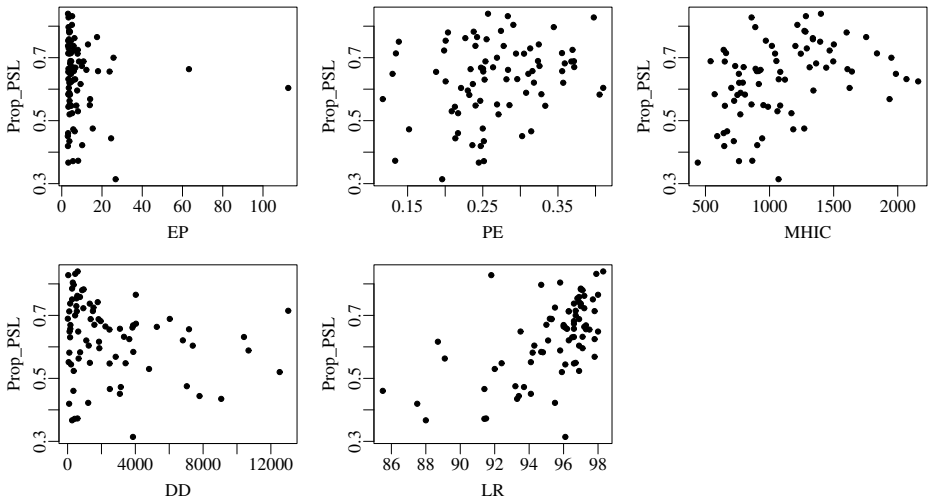


FIGURE 2: Scatter diagram the response variable and quantitative covariates, except LAT and LONG

Table 3: Spearman's correlation matrix with their respective p-values in parenthesis

	Prop_PSL	EP	PE	MHIC	DD
Prop_PSL					
EP	−0.10				
	(0.3650)				
PE	0.23	−0.09			
	(0.0382)	(0.4403)			
MHIC	0.41	0.30	−0.46		
	(0.0002)	(0.0072)	(0.0002)		
DD	−0.24	0.34	0.01	0.02	
	(0.0313)	(0.0022)	(0.9587)	(0.8582)	
LR	0.56	0.12	−0.24	0.76	0.11
	(< 0.0001)	(0.2765)	(0.0316)	(< 0.0001)	(0.3325)

Table 4 presents the frequency distribution for the qualitative variables. About 50% of the municipalities in the study belong to the Southeast region, which is also the most developed. The North and Northeast are the most impoverished and concentrate a relative frequency of 15.19%. In 2014 and 2018, the parties with centre ideology elected more governors. Notice the growing number of right-wing governors from 2014 to 2018. It increased by approximately 26.58%.

Table 4: Frequency distribution for the qualitative variables

Variable	Frequency	Relative frequency (%)
Region		
Northeast	17	21.5189
South	11	13.9241
Southeast	39	49.3672
Northern	6	7.5949
Midwest	6	7.5949
PG_2014		
Left-wing	32	40.5063
Centre	39	49.3671
Right-wing	8	10.1266
PG_2018		
Left-wing	20	25.3164
Centre	30	37.9747
Right-wing	29	36.7089
CA_BR		
Yes	23	29.1139

4. Fitted Regressions

This section discusses the fitted regression models considering the Prop_PSL as the dependent variable. Since Region, PG_2014, PG_2018, and Cap_BR are qualitative variables, we adopt dummies when they are included in the regression analysis with fixed effects. We define the Midwest region as the reference category

for the Region variable. For PG_2014 and PG_2018, the reference is the center spectrum. Finally, the Cap_BR dummy is defined as a variable that equals one if the municipality is capital and zero otherwise.

For each distribution described in Section 2.1, the regression model is specified with three different configurations for the explanatory variables. The first model (Fit 1) does not include the LAT and LONG and considers the Region as a random effect. The second model (Fit 2) differs from the first by considering the Region as a fixed effect through the dummy variables. Finally, we do not consider Region in the third model (Fit 3) and include the LAT and LONG as fixed effects.

Table 5 shows the log-likelihood, degrees of freedom (DF), AIC, and R^2_G measures for all these three configurations and competitive distributions. It also includes the p-values for the Shapiro-Wilk (SW) test to verify the normality assumption for the quantile residuals.

Notably, the UL distribution displays poor performance, which can be justified due to the lack of the parameter σ, which makes it less flexible to other distributions. The classical beta distribution with the Fit 2 configuration presents superior goodness-of-fit measures. This configuration also outperforms Fit 1 and Fit 3 for the models based on the simplex and UG distributions, thus suggesting that considering random effects for the regions is unnecessary. Further, for the Fit 1 settings, the estimates of the random effect parameter σ_1 are all smaller than 0.0001, except for the UG regression model ($\hat{\sigma}_1 = 0.2259$). This indicates that most of the variation is explained by the fixed effects of the model. Moreover, most fitted models do not reject the null hypothesis of normality for the residual distribution at the 1% significance level, as indicated by the SW test results. Exceptions are noted for the beta, simplex, and UG distributions with Fit 1 and Fit 2 configurations. Based on these findings, the beta distribution with the Fit 2 configuration is chosen as the preferred model.

TABLE 5: Goodness-of-fit measures and SW test for the residuals of the fitted models and competitive distributions

Distribution	Model	Log-likelihood	DF	AIC	R^2_G	SW (p-value)
Beta	Fit 1	−140.74	28.00	−225.50	0.88	0.40
	Fit 2	−145.49	30.00	−230.99	0.89	0.04
	Fit 3	−128.08	26.00	−204.16	0.83	0.97
Simplex	Fit 1	−129.65	28.01	−204.95	0.84	< 0.01
	Fit 2	−134.03	30.00	−213.56	0.86	< 0.01
	Fit 3	−115.23	26.00	−190.81	0.80	0.36
UG	Fit 1	−123.28	27.00	−202.12	0.83	0.34
	Fit 2	−133.89	30.00	−209.52	0.86	0.01
	Fit 3	−119.51	26.00	−191.29	0.80	0.35
UL	Fit 1	−45.81	11.00	−76.17	0.22	0.78
	Fit 2	−48.59	15.00	−68.74	0.23	0.89
	Fit 3	−44.87	13.00	−72.30	0.22	0.87

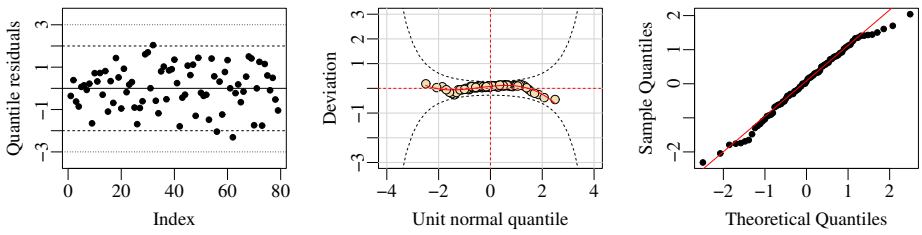
After selecting the beta distribution and Fit 2 configuration as the more appropriate configuration, we employ the `stepGAIC` procedure to identify the optimal

combination of explanatory variables among its nested models. However, upon selection, the model chosen by the algorithm incorporates some non-significant covariates. To address this, we conduct the LRT and find that the model excluding these non-significant variables is favored (p-value = 0.1899). Consequently, we designate this refined model as the final model for our analysis. Table 6 presents the parameter estimates for the final model.

TABLE 6: Estimates, standard errors, and p-values for the selected beta regression model

Effect	Estimate	Std. Error	p-value
Mean submodel			
Intercept	−1.7448	0.2103	< 0.0001
PE	5.0130	0.4523	< 0.0001
MHIC	0.0009	0.0001	< 0.0001
Northeast	0.5853	0.0863	< 0.0001
South	0.3225	0.0330	< 0.0001
PG_2014_Left	−0.1562	0.0290	< 0.0001
PG_2018_Left	−0.5704	0.0463	< 0.0001
PG_2018_Right	0.2087	0.0048	< 0.0001
Cap_BR	−0.3696	0.0729	< 0.0001
Dispersion submodel			
Intercept	−33.3775	4.6129	< 0.0001
EP	−0.1299	0.0078	< 0.0001
PE	−8.5590	1.5984	< 0.0001
LR	0.3800	0.0497	< 0.0001
South	−2.9657	0.4005	< 0.0001
Southeast	−2.2737	0.3056	< 0.0001
PG_2014_Left	0.6144	0.2477	0.0155
Cap_BR	−0.8759	0.2851	0.0030

Figure 3 displays plots of the residuals versus the index, wormplot (Buuren & Fredriks, 2001) and quantile-quantile plot (QQ-plot) of the quantile residuals for the selected beta regression model. These plots confirm that this model is suitable for modeling the vote proportions. Some findings of the vote proportion can follow from Table 6.



(a) Residuals versus index. (b) Wormplot. (c) QQ-plot.

FIGURE 3: Residual analysis for the selected beta regression model

- The PE is a significant covariate, and the average vote proportions tend to be higher in the municipalities with higher proportion of evangelicals. It is also

significant about the dispersion, which tends to decrease as the proportion of evangelicals increases.

- The MHIC has a positive impact on the mean of the vote proportions. This result is consistent with Hunter & Power (2019), which indicates income among the support indicators for Bolsonaro's election. The candidate won among all income groups except for the lowest income levels.

- PG_2014_Left and PG_2018_Left had a negative influence on the mean of the vote proportions. On the other hand, the PG_2018_Right effect on the mean of the vote proportions is positive. In fact, Bolsonaro adopted a position of reducing state intervention in the economy in his campaign, and right-wing voters in Latin America have traditionally been against state intervention in the economy (Rennó, 2020).

- The South and Southeast regions are negatively related to the dispersion of the variable of interest, i.e., the municipality in this region presents a vote proportion less dispersed than those from other regions. Finally, the dispersion tends to decrease for the capitals.

5. Conclusion

In this paper, we conducted a study on unit regression models to quantify the effect of explanatory variables on the proportion of votes of Jair Bolsonaro in the second round of the 2018 presidential elections. We considered data from Brazilian municipalities with a population greater than 300,000 and verified that the elected candidate won in most cities considered. The vote proportion distribution is left-skewed with a few outliers in the left tail. All of them are located in the Northeast region. As explanatory variables, we select some socio-demographic indicators and the political spectrum of the governors' party in the 2014 and 2018 elections. We constructed fitted, simplex, unit gamma, and unit Lindley regression models using the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework to explain the mean variations of the vote proportions. Since the observational units correspond to municipalities, spatial correlation can be expected. Therefore, we considered three different configurations to verify this. The first configuration considers the municipalities's region as a random effect. The second configuration uses dummy variables to compute the regions as fixed effects. Finally, the third configuration includes the latitude and longitude of the municipalities. We concluded that the beta regression under the second configuration is the preferred model. After performing a stepwise procedure, we evaluated significant effects for the proportion of evangelicals, the monthly household income per capita, the political spectrum of the governors' party elected in 2014 and 2018, and Brazilian capital dummy. We also verify that some Brazilian regions impact the vote proportions' mean and dispersion.

Acknowledgements

We thank the two referees and Associate Editor for their valuable comments and suggestions. The author Renata Rojas Guerra acknowledges the support of Serrapilheira Institute/Serra - 2211-41692; FAPERGS/23/2551-0001595-1, FAPERGS/23/2551-0000851-3; and CNPq/306274/2022-1. The author Tatiane Fontana Ribeiro gratefully acknowledge partial financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Data Availability

The data supporting this research is publicly available and can be accessed at the sources referred in Table 2. It is also provided in the following repository `https://github.com/renata-rojasg/vote-prop-PSL2018`, with all the computer codes used in the application.

[Received: September 2023 — Accepted: February 2024]

References

Akdur, H. T. K. (2021), 'Unit-Lindley mixed-effect model for proportion data', *Journal of Applied Statistics* **48**(13-15), 2389–2405.

Almeida Junior, P. M. d. & Souza, T. C. (2015), 'Estimativas de votos da presidente Dilma Rousseff nas eleições presidenciais de 2010 sob o âmbito do bolsa família', *Ciência e Natura* **37**, 12–22.

Andrade, M. V., Noronha, K. V. M. d. S., Menezes, R. d. M., Souza, M. N., Reis, C. d. B., Martins, D. R. & Gomes, L. (2013), 'Desigualdade socioeconômica no acesso aos serviços de saúde no Brasil: um estudo comparativo entre as regiões brasileiras em e 2008', *Economia Aplicada* **17**, 623–645.

Barndorff-Nielsen, O. E. & Jørgensen, B. (1991), 'Some parametric models on the Simplex', *Journal of Multivariate Analysis* **39**, 106–116.

Bayer, F. M. & Cribari-Neto, F. (2017), 'Model selection criteria in beta regression with varying dispersion', *Communications in Statistics-Simulation and Computation* **46**, 729–746.

Bayer, F. M., Tondolo, C. M. & Müller, F. M. (2018), 'Beta regression control chart for monitoring fractions and proportions', *Computers & Industrial Engineering* **119**, 416–426.

Bayes, C. L., Bazán, J. L. & De Castro, M. (2017), 'A quantile parametric mixed regression model for bounded response variables', *Statistics and its Interface* **10**, 483–493.

BRASIL (2020), 'Eleições', Available in: `http://www.ipeadata.gov.br/Default.aspx`.

Buuren, S. v. & Fredriks, M. (2001), 'Worm plot: a simple diagnostic device for modelling growth reference curves', *Statistics in Medicine* **20**, 1259–1277.

Canterle, D. R. & Bayer, F. M. (2019), 'Variable dispersion beta regressions with parametric link functions', *Statistical Papers* **60**, 1541–1567.

Carrasco, J. M. & Reid, N. (2019), 'Simplex regression models with measurement error', *Communications in Statistics-Simulation and Computation* **1**, 1–16.

Cordeiro, G. M., Rocha, E., Figueiredo, D., Fernandes, A., Ortega, E. M. & Prataviera, F. (2020), 'The beta and simplex regression models to explain homicides in state capitals of Brazil', *Model Assisted Statistics and Applications* **15**, 215–224.

Cribari-Neto, F. & Souza, T. C. (2012), 'Testing inference in variable dispersion beta regressions', *Journal of Statistical Computation and Simulation* **82**, 1827–1843.

de Araújo, F. J. M., Guerra, R. R. & Peña-Ramírez, F. A. (2022), 'The Burr XII quantile regression for salary-performance models with applications in the sports economy', *Computational and Applied Mathematic* **41**, 282.

de Bastiani, F., Rigby, R. A., Stasinopoulous, D. M., Cysneiros, A. H. & Uribe-Opazo, M. A. (2018), 'Gaussian markov random field spatial models in gamlss', *Journal of Applied Statistics* **45**(1), 168–186.

de Freitas, J. V. B., Nobre, J. S., Espinheira, P. L. & Rêgo, L. C. (2023), 'Unit gamma regression models for correlated bounded data', *Brazilian Journal of Probability and Statistics* **37**(4), 693–719.

Espinheira, P. L., da Silva, L. C. M., Silva, A. d. O. & Ospina, R. (2019), 'Model selection criteria on beta regression for machine learning', *Machine Learning and Knowledge Extraction* **1**, 427–449.

Espinheira, P. L. & Silva, A. d. O. S. (2019), 'Residual and influence analysis to a general class of simplex regression', *TEST* pp. 1–30.

Ferrari, S. L. P. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **7**, 799–815.

Ghosh, A. (2019), 'Robust inference under the beta regression model with application to health care studies', *Statistical Methods in Medical Research* **28**, 871–888.

Guedes, A. C., Cribari-Neto, F. & Espinheira, P. L. (2020), 'Modified likelihood ratio tests for unit gamma regressions', *Journal of Applied Statistics* **47**, 1562–1586.

Guerra, R. R. (2024), 'Unitdistforgamlss', `https://figshare.com/articles/software/UnitDistForGAMLSS/25328575/1`.

Guerra, R. R., Peña-Ramírez, F. A. & Bourguignon, M. (2020), 'The unit extended Weibull families of distributions and its applications', *Journal of Applied Statistics* **1**, 1–19.

Hunter, W. & Power, T. J. (2019), 'Bolsonaro and Brazil's illiberal backlash', *Journal of Democracy* **30**, 68–82.

IBGE (2010), 'Censo demográfico 2010', *IBGE: Insituto Brasileiro de Geografia e Estatística* .

Karlsson, P., Månsson, K. & Kibria, B. G. (2020), 'A Liu estimator for the beta regression model and its application to chemical data', *Journal of Chemometrics* **34**, e3300.

Lemonte, A. J. & Bazán, J. L. (2016), 'New class of Johnson distributions and its associated regression model for rates and proportions', *Biometrical Journal* **58**, 727–746.

López, F. O. (2013), 'A Bayesian approach to parameter estimation in simplex regression model: A comparison with beta regression', *Revista Colombiana de Estadística* **36**, 1–21.

Mazucheli, J., Menezes, A. F. B. & Chakraborty, S. (2019), 'On the one parameter unit-lindley distribution and its associated regression model for proportion data', *Journal of Applied Statistics* **46**(4), 700–714.

Mazucheli, J., Menezes, A. F. B., Fernandes, L. B., de Oliveira, R. P. & Ghitany, M. E. (2020), 'The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates', *Journal of Applied Statistics* **47**, 954–974.

Mousa, A. M., El-Sheikh, A. A. & Abdel-Fattah, M. A. (2016), 'A gamma regression for bounded continuous variables', *Advances and Applications in Statistics* **49**, 305–326.

Pereira, T. L., Souza, T. C. & Cribari-Neto, F. (2014), 'Uma avaliação da eficiência do gasto público nas regiões do Brasil', *Ciência e Natura* **36**, 23–36.

Petterle, R. R., Taconeli, C. A., da Silva, J. L., da Silva, G. P., Laureano, H. A. & Bonat, W. H. (2023), 'Unit gamma mixed regression models for continuous bounded data', *Journal of Statistical Computation and Simulation* **93**(6), 1011–1029.

R Core Team (2024), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Regis, R. O., Ospina, R., Bernardino, W. & Cribari-Neto, F. (2023), 'Asset pricing in the Brazilian financial market: five-factor GAMLSS modeling', *Empirical Economics* **64**(5), 2373–2409.

Rennó, L. (2020), 'The Bolsonaro voter: Issue positions and vote choice in the 2018 Brazilian presidential elections', *Latin American Politics and Society* **62**, 1.

Ribeiro, T. F., Cordeiro, G. M., Pena-Ramírez, F. A. & Guerra, R. R. (2021), 'A new quantile regression for the COVID-19 mortality rates in the United States', *Computational and Applied Mathematics* **40**, 1–16.

Ribeiro, T. F., Peña-Ramírez, F. A., Guerra, R. R. & Cordeiro, G. M. (2022), 'Another unit Burr XII quantile regression model based on the different reparameterization applied to dropout in Brazilian undergraduate courses', *Plos One* **17**, e0276695.

Ribeiro, T. F., Seidel, E. J., Guerra, R. R., Peña-Ramírez, F. A. & da Silva, A. M. (2021), 'Soybean production value in the Rio Grande do Sulunder the GAMLSS framework', *Communications in Statistics: Case Studies, DataAnalysis and Applications* **7**, 146–165.

Rigby, R. A. & Stasinopoulos, D. M. (2005), 'Generalized additive models for location, scale and shape, (with discussion)', *Applied Statistics* **54**, 507–554.

Sardinha, E. & Costa, S. (2020), 'Direita cresce e engole o centro no congresso mais fragmentado da história', Accessed on: `https://congressoemfoco.uol.com.br/legislativo/direita-cresce-e-engole-o-centro-no-congresso-mais-fragmentado-da-historia/`.

SIDRA (2020), 'Censo demográfico', Accessed on: `https://sidra.ibge.gov.br/tabela/3974`.

Silva, D. V., Akdur, H. T. K. & Paula, G. A. (2023), 'Analysis of correlated unit-Lindley data based on estimating equations', *Statistical Methods & Applications* pp. 1–32.

Simas, A. B., Barreto-Souza, W. & Rocha, A. V. (2010), 'Improved estimators for a general class of beta regression models', *Computational Statistics & Data Analysis* **54**, 348–366.

Stasinopoulos, M. D., Rigby, R. A. & Bastiani, F. D. (2018), 'GAMLSS: a distributional regression approach', *Statistical Modelling* **18**, 248–273.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. & De Bastiani, F. (2017), *Flexible regression and smoothing: using GAMLSS in R*, CRC Press.

TSE (2018), 'Electoral data repository', Results. Presents results of Brazilian general elections. Accessed on: `http://english.tse.jus.br/the-brazilian-electoral-system/statistics`.

Yero, E. J. H., Sacco, N. C. & do Carmo Nicoletti, M. (2020), 'Effect of the municipal Human Development Index on the results of the 2018 Brazilian Presidential elections', *Expert Systems with Applications* p. 114305.

Zucco Jr, C. (2013), 'When payouts pay off: Conditional cash transfers and voting behavior in brazil 2002-10', *American Journal of Political Science* **57**, 810–822.

Zucco Jr, C. (2015), 'The impacts of conditional cash transfers in four presidential elections (2002-2014)', *Brazilian Political Science Review* **9**, 135–149.