UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

TATIANE FONTANA RIBEIRO

ESSAYS ON THE UNIT BURR XII DISTRIBUTION: REGRESSION AND TIME SERIES MODELS

Recife

2020

TATIANE FONTANA RIBEIRO

ESSAYS ON THE UNIT BURR XII DISTRIBUTION: REGRESSION AND TIME SERIES MODELS

Master's thesis submitted to the Programa de Pós-Graduação em Estatística, Centro de Ciências Exatas e da Natureza, Universidade Federal de Pernambuco, as a partial requirement to obtain a Master's degree in Statistics.

Area of concentration: Applied Statistics

Advisor: Prof. Gauss Moutinho Cordeiro

Co-Advisor: Prof. Fernando A. Peña-Ramírez

Recife

2020

Catalogação na fonte
Bibliotecária Mariana de Souza Alves CRB4-2105

R484e Ribeiro, Tatiane Fontana
Essays on the unit Burr XII distribution: regression and time series models / Tatiane Fontana Ribeiro. – 2020.
96f.: il., fig., tab.

Orientador: Gauss Moutinho Cordeiro.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2020.
Inclui referências e apêndices.

1. Estatística Aplicada. 2. Aprendizado estatístico. 3. Distribuições de probabilidade no intervalo unitário. 4. Regressão beta. I. Cordeiro, Gauss Moutinho. (orientador) II. Título.

310 CDD (22. ed.) UFPE-CCEN 2020-200

TATIANE FONTANA RIBEIRO

ESSAYS ON THE UNIT BURR XII DISTRIBUTION: REGRESSION AND TIME SERIES MODELS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 28 de outubro de 2020.

BANCA EXAMINADORA

_____

Prof. Gauss Moutinho Cordeiro
UFPE

_____

Prof. Maria do Carmo Soares de Lima
UFPE

_____

Prof. Edwin Moises Marcos Ortega
ESALQ/USP

To my parents, Teodomiro e Sirlei.

ACKNOWLEDGEMENTS

Initially, I would like to express my gratitude to my advisors Prof. Dr. Gauss Moutinho Cordeiro and Prof. Dr. Fernando A. Peña-Ramírez for their valuable corrections and suggestions about the dissertation, which help improve its quality. I am very grateful for their safe orientation and patience.

In special, I would like to thank my co-advisor, Prof. Dr. Fernando A. Peña-Ramírez, for his huge disposition and unique intelligence to clarify every question that arose in the elaboration of this dissertation. You ever provided me great motivation and immensely contributed to writing this manuscript in English. This is proof you are a great professional, professor, and researcher.

To Prof. Dra Renata Rojas Guerra, who is being my example since 2018. If I got here, it is because you have crossed my path. Also, I am immensely grateful to you by valuable contribution to these chapters of this dissertation.

To my favorite doctorate student, friend, and colleague José Jairo, I am immensely grateful to you for your beautiful friendship, for helping me in my adapting to Recife, and for your immense help and motivation so that I could finish this dissertation in time.

To my parents Teodomiro and Sirlei, for their support, who has always encouraged me to pursue my dreams. Without their support, I wouldn't have come this far. To my dear brothers Taiane, Teodomiro Júnior, and Thiago, for their trust, and so love. I love you guys so much.

To all my friends and colleagues for many and many hours of study shared. I will not cite names, but I would like to thank everyone for so good moments shared and memories that would never be erased from my mind.

I would also like to thank all professors at UFPE, especially professors Francisco Cribari-Neto, Audrey Cysneiros, and Maria do Carmo Soares de Lima.

Finally, I thank to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the financial support.

ABSTRACT

There is an interest in modeling bounded random variables to the standard unit interval in many practical situations, such as rates, proportions, and indexes. We propose two new probability distributions to deal with the uncertainty involved by variables of this type and develop its associated regression models. Both distributions are based on a transformation of the Burr XII random variable. We also introduce a new dynamic model for time series data with support in the interval $(0, 1)$. This dissertation is composed of three main and independent chapters. In the first, we define the unit Burr XII (UBXII) distribution and its quantile regression model. Some of its mathematical and statistical properties are investigated. In the second, the reflexive UBXII distribution is obtained, and the regression model is proposed. The maximum likelihood (ML) method is considered for parameters estimation of both regression models. In the third, we propose the dynamic class of models: UBXII autoregressive moving average (UBXII-ARMA) for time series taking values in the unit interval. The conditional ML method is used to estimate and construct asymptotic confidence intervals of the parameters that index the UBXII-ARMA model. Closed-form expressions for the conditional score vector are derived. Furthermore, Monte Carlo simulation studies, diagnostic analysis tools, model selection criteria, and applications to the real data are presented and discussed for the three proposed models.

Keywords: Beta regression. Quantile regression. Statistical learning. Time series. Unit probability distributions.

RESUMO

Em muitas situações práticas, há interesse em modelar variáveis aleatórias limitidas no intervalo unitário padrão, tais como taxas, proporções e índices. Propomos duas novas distribuições de probabilidade para lidar com a incerteza envolvida por variáveis deste tipo e desenvolvemos seus modelos de regressão associados. Ambas as distribuições são baseadas em uma transformação da variável aleatória Burr XII. Também introduzimos um novo modelo dinâmico para dados de séries temporais com suporte no intervalo $(0, 1)$. Esta dissertação é composta por três capítulos principais e independentes. Na primeira parte, definimos a distribuição Burr XII unitária (UBXII) e o modelo de regressão quantílica associado. Algumas das propriedades estatísticas e matemáticas são investigadas. Na segunda, a distribuição UBXII reflexiva é obtida e o modelo de regressão é proposto. O método de máxima verossimilhança (ML) é considerado para estimação dos parâmetros de ambos os modelos de regressão. Na terceira parte, propomos a classe de modelos dinâmicos: UBXII autorregressivos de médias móveis (UBXII-ARMA) para séries temporais que tomam valores no intervalo unitário. O método de ML condicional é usado para estimar e construir intervalos de confiança dos parâmetros que indexam o model UBXII-ARMA. Expressões em forma fechada para o vetor escore condicional são derivadas. Além disso, estudos de simulação de Monte Carlo, ferramentas de análise de diagnóstico, critérios de seleção de modelos e aplicações a dados reais são apresentadas e discutidas para os três modelos propostos.

Palavras-chave: Aprendizado estatístico. Distribuições de probabilidade no intervalo unitário. Regressão beta. Regressão quantílica.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

1.1 INITIAL CONSIDERATIONS

This dissertation is composed of three main and independent chapters. The dissertation's subject is new classes of models for random variables restricted to the unit interval from transformations of the Burr XII random variable pioneered by (BURR, 1942). Variables with domain in the interval $(0, 1)$ are commonly found in several fields of knowledge. In the context of regression analysis, they are typically studied by the beta regression (FERRARI; CRIBARI-NETO, 2004) and Kumaraswamy regression models. Some alternatives have been introduced in the literature to expand the range of models available as the unit Weibull (UW) quantile regression (MAZUCHELI *et al.*, 2020). We introduce two new classes of quantile regression models. Similarly, dynamic models to analyze bounded-double conditional-response variables have been proposed in the literature, such as beta autoregressive moving average (βARMA) models (ROCHA; CRIBARI-NETO, 2009) and Kumaraswamy autoregressive moving average (KARMA) models (BAYER; BAYER; PUMI, 2017). In this sense, we propose a new time series model for data that assume values in the standard unit interval to the scarce classes of time series models available. In what follows, we present a brief outline of this dissertation.

In Chapter 2, we define the unit Burr XII (UBXII) distribution and its quantile regression model. We provide some of its structural properties. To estimate the parameters that index the model, we consider the maximum likelihood (ML) method and carried out a simulation study to analyze its performance on finite samples. Besides, we derive expressions for the score function and observed information matrix. General techniques of diagnostic analysis and model selection are presented and discussed for the regression model. We empirically show the new model's importance and flexibility through an application to a real dataset, in which the dropout rate of Brazilian undergraduate animal sciences courses is analyzed. Finally, we use a statistical learning method for comparing the proposed model with the Kumaraswamy, unit-Weibull, and beta regressions.

In Chapter 3, a regression model is constructed to identify the variables that affect the mortality rates by COVID-19 in the U.S. states. The mortality rates in these states are computed by considering the total of deaths recorded on 30, 60, 90, and 120 days from the 20th confirmed case. From the reflection of the UBXII variable introduced in Chapter 2, we define the reflexive UBXII distribution and its associated regression model. In the application to

the COVID-19 mortality rates, it is compared to the well-known beta, simplex (JØRGENSEN, 1997b), Kumaraswamy (KUMARASWAMY, 1980), and UW (MAZUCHELI *et al.*, 2020) regressions, which are useful in modeling proportional data. The parameters are estimated by the ML method. We conduct Monte Carlo simulation studies to assess the finite-sample performance of the maximum likelihood estimators. Furthermore, we adapt some regression model adequacy measures and consider an approach of cross-validation to compare the final fitted regression models.

Chapter 4 address a new class of time series models for continuous random variates that assume values in the unit interval. The model arises from the assumption that the random component has conditional UBXII distribution, and it is defined as the UBXII autoregressive moving average. In the introduced model, any quantile can be modeled by a dynamic structure containing autoregressive and moving average terms, time-varying regressors, unknown parameters, and a link function. We consider the conditional ML method for parameter estimation and conduct simulation studies to evaluate the estimates' performance and estimated coverage rates from asymptotic confidence intervals for finite samples. We discuss the goodness-of-fit assessment and forecasting of the new model. We give explicit-form expressions for the conditional score function. To demonstrate our proposal's suitability, an application that uses stocked hydroelectric energy time series data is presented and discussed. Furthermore, we carry out-of-sample forecasting comparisons of the introduced model with the βARMA and KARMA models available for time series taking values in the double bounded interval (a, b).

The notation and terminology used are consistent within each chapter. We consider the `R` programming language (R Core Team, 2020) for Monte Carlo simulations, figures generating, and all remain statistical analysis.

2 THE UNIT BURR XII REGRESSION: PROPERTIES, SIMULATION AND APPLICATION

2.1 INTRODUCTION

Variables like rate, proportions, percentages, and indices which lie on the standard unit interval are commonly found in several fields of knowledge. These variables exhibit extra variation since often present asymmetry and heteroscedasticity; see some examples of them in Kieschnick and McCullough (2003) and Cribari-Neto and Souza (2013). Continuous distributions widely used in the modeling of double-bounded random variables are the Beta and Kumaraswamy (KUMARASWAMY, 1980). However, in some situations, our interest is to know how a set of covariates impact the behavior of a double-bounded response variable in the interval $(0, 1)$. Such influence on the response's mean is usually modeled with the well-known beta regressions (FERRARI; CRIBARI-NETO, 2004). It stands out, mainly due to its flexibility, being able to accommodate asymmetries that are typical to responses of this kind.

Other alternatives to the beta regression to model the response's mean are the simplex (JØRGENSEN, 1997b), Log-Lindley (GÓMEZ-DÉNIZ; SORDO; CALDERÍN-OJEDA, 2014), unit gamma (MOUSA; EL-SHEIKH; ABDEL-FATTAH, 2016), and unit-Lindley regressions (MAZUCHELI; MENEZES; CHAKRABORTY, 2019). Nevertheless, when the dependent variable presents some atypical observations, modeling its median can be more appropriate than its mean, which is more sensitive to outliers (LEMONTE; BAZÁN, 2016). Furthermore, the median-based regression presents desirable properties such as the equivariance-to-monotone transformation and robustness in asymmetrical data (PUMI; RAUBER; BAYER, 2020). A classical alternative in this context is the Kumaraswamy regression under parameterization proposed by Mitnik and Baek (2013). In a more general way, we can also be interested in modeling any quantile of a response in the unit interval to know how covariates impact its different levels. Studies like in Dehbi, Cortina-Borja and Geraci (2016) and Lachos *et al.* (2015), and more recently, the unit-Weibull quantile regression (MAZUCHELI *et al.*, 2020), show that quantile regressions have attracted many researchers' attention. The main advantage of this approach is that it provides classes of regressions quite flexible for modeling data with heterogeneous conditional distributions (BAYES; BAZÁN; CASTRO, 2017), since any quantile of the response, in addition to the median, can be modeled as a function of covariates. However, in many areas and several empirical applications, there is a clear need for new flexible alternatives to the scarce regression classes available.

In this context, we use an approach based on a quantile parameterization to introduce a new probability distribution with the bounded domain on the interval $(0,1)$, and its associated regression. We consider the Burr XII (BXII) distribution, which was pioneered by Burr (1942). Let X be a positive continuous random variable having the BXII distribution. The cumulative distribution function (cdf) and probability density function (pdf) of X are

$$F_X(x;c,d) = 1-(1+x^c)^{-d} \quad x>0, \tag{2.1}$$

and

$$f_X(x;c,d) = c\,d\,x^{c-1}\,(1+x^c)^{-(d+1)}, \tag{2.2}$$

respectively, where $c>0$ and $d>0$ are shape parameters. The pdf in Equation (2.2) is unimodal with mode equal to $[(c-1)/(c\,d+1)]^{1/c}$.

For $c=1$, the BXII distribution is called the Lomax (or Pareto Type II) distribution, mainly in applications to the analysis of business failure data (WATKINS, 2011). By taking $d=1$, it is a special case of the log-logistic distribution. The BXII distribution is a very popular distribution for modeling lifetime data and for modeling phenomenon with monotone failure rates. Our proposal is based on a transformation on X, more specifically, $T(X):(0,\infty)\to(0,1)$.

We provide at least four motivations for this work. First, we propose a new distribution to model bounded random variables in the unit interval, in which one of its parameters is a distribution's quantile. Second, we consider a regression structure for this quantile by assuming that it can be expressed as a function of covariates and, hence, a more general class of regressions is obtained. The third motivation is to use a statistical learning tool for comparing the prediction performance of non-nested models and selecting the most suitable for the data at hand. The fourth motivation is referring to the usefulness of the new regression for modeling the dropout proportion of undergraduate animal science in Brazil.

We are interested in analyzing the university dropout phenomenon because it is a problem with academic, social, and economic implications due to the high cost it inflicts on the students themselves, their families, universities and government (RODRÍGUEZ-MUÑIZ *et al.*, 2019). Organizational variables of the educational institutions provide an explicative approach to the dropout phenomenon (TINTO, 1986). In that idea, several authors studied how aspects of the organizational structure of universities affect student outcomes (see, for instance, Berger (2002) and Sneyers and Witte (2017)). On the other hand, data mining aspects also have been studied. Salloum *et al.* (2020) discusses and reviews the forms of data mining and its importance

in educational research, and Rodríguez-Muñiz *et al.* (2019) uses machine learning techniques to undertake the analysis of dropout in the University of Oviedo (Spain). In the Brazilian scenario, the efficiency and affirmative action policy adoption in higher education institutions (HEIs) are discussed by Zoghbi, Rocha and Mattos (2013) and Vieira and Arends-Kuenning (2019).

In this chapter, we use mining data techniques to obtain the dropout proportion and related institutional variables about Brazilian undergraduate animal science courses. This course has received attention in the literature; see, for instance, Peffer (2011), who sought to identify demographic variables as well as their relation to students' performance and interest areas, and factors associated with enrollment in an introductory animal sciences course.

The rest of chapter is outlined as follows. In Section 2.2, we define a new distribution for modeling bounded random variables on the unit interval as well as a reparametrization in terms of the quantiles. Some of its mathematical and statistical properties are investigated in Section 2.3. We obtain the maximum likelihood of the parameters in Section 2.4. We provide a simulation study in Section 2.5 to evaluate the performance of the estimators. In Section 2.6, we define a regression from the new distribution, discuss the estimation of the parameters and conduct a simulation study, present some diagnostic analysis methods and regression selection criteria. In special, we present a statistical learning tool (cross-validation approach) to compare non-nested regressions. In Section 2.7, we perform an application of the new regression to dropout in Brazilian undergraduate animal sciences courses. We offer some conclusions in Section 2.8. Finally, we provide the observed matrix for the new distribution and Fisher's observed information matrix, and information about data's extraction used in application; see Appendix A.

2.2 THE UNIT BXII DISTRIBUTION

We introduce a new distribution with support on the unit interval based on the BXII distribution. Suppose that X is a random variable having the BXII distribution with cdf and pdf given by (2.1) and (2.2), respectively. We define the *unit Burr XII* (UBXII) distribution through the transformation $Y = \mathrm{e}^{-X}$. Hence, the cdf and pdf of the UBXII distribution are

$$F_Y(y; c, d) = \left(1 + \log^c y^{-1}\right)^{-d}, \qquad 0 < y < 1, \tag{2.3}$$

and

$$f_Y(y; c, d) = c\, d\, y^{-1} \log^{c-1} y^{-1} \left(1 + \log^c y^{-1}\right)^{-(d+1)}, \tag{2.4}$$

respectively, where $c > 0$ and $d > 0$ are shape parameters. Henceforth, if Y is a random variable with pdf (2.4), we write $Y \sim \text{UBXII}\,(c, d)$. The quantile function (qf) of Y follows by inverting Equation (2.3), namely

$$Q_Y(\tau) = \exp\left[-(\tau^{-1/d} - 1)^{1/c}\right], \qquad 0 < \tau < 1. \tag{2.5}$$

The UBXII quantiles can be found from (2.5) by setting values for τ. In particular, the median of Y comes with $\tau = 0.5$. So, we can generate UBXII variates using (2.5). For doing this, we require $Y = Q_Y(\mathcal{T})$, where $\mathcal{T} \sim \mathcal{U}(0, 1)$.

Distributions with direct interpretation parameters are desirable in empirical applications, and for this purpose, several authors have adopted reparameterizations on well-know distributions; see Jørgensen (1997a), Ferrari and Cribari-Neto (2004), Lemonte and Bazán (2016), Mousa, El-Sheikh and Abdel-Fattah (2016) and Mazucheli *et al.* (2020). In general, the mean of the random variable is modeled as proposed by Ferrari and Cribari-Neto (2004); Jørgensen (1997a); and Mousa, El-Sheikh and Abdel-Fattah (2016). However, it is an outlier-sensitive measure, and for the UBXII distribution, we can not obtain a closed-form expression for the mean. Thus, modeling the quantiles is an interesting approach for asymmetric data because they can be outlier-resistant measures (LEMONTE; BAZÁN, 2016), and a smart alternative since the qf of Y has a closed-form. Hence, any quantile can be computed in explicit form. Further, one of the parameters of the UBXII distribution (under a quantile-parameterization) can be interpreted as the τth quantile of Y. Thus, we shall reparameterize Equation (2.3) in terms of the τth quantile $q = Q_Y(\tau)$. By inverting (2.5) and solving for d, we have

$$d = \log \tau^{-1} / \log(1 + \log^c q^{-1}). \tag{2.6}$$

By replacing (2.6) in Equations (2.3) and (2.4), the cdf and pdf of the UBXII distribution (under this parametrization) have the forms

$$F_Y(y; q, c) = \left(1 + \log^c y^{-1}\right)^{\log \tau / \log\left(1 + \log^c q^{-1}\right)}, \qquad 0 < y < 1, \tag{2.7}$$

and

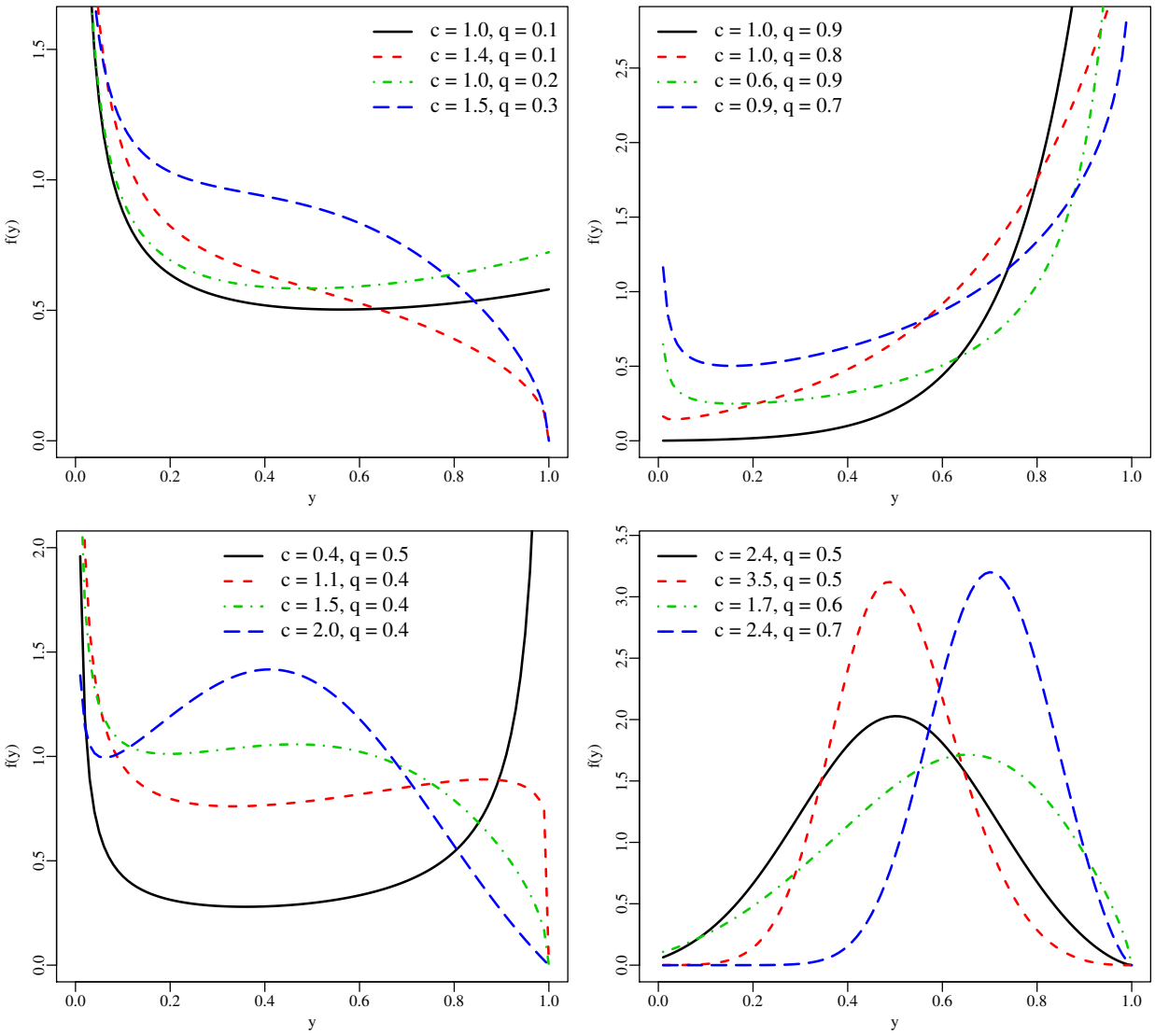
$$f_Y(y; q, c) = \frac{\log \tau^{-c} \log^{c-1} y^{-1}}{y \log\left(1 + \log^c q^{-1}\right)} \left(1 + \log^c y^{-1}\right)^{\log \tau / \log\left(1 + \log^c q^{-1}\right) - 1}, \tag{2.8}$$

respectively. Henceforth, we denote by $Y \sim \text{UBXII}\,(c, q)$ a random variable with density (2.8).

Some UBXII densities (for $\tau = 0.5$) are displayed in Figure 1, which reveal different shapes such as decreasing, increasing, reverse J-shaped, U-shaped, reverse tilde-shaped (decreasing-increasing-decreasing), non-skewed and skewed-left. It is noteworthy that the UBXII density can

accommodate several skew-left shapes and has a reverse tilde-shaped, which is not presented by classical unit distributions.

Figure 1 – Plots of the UBXII density (with $\tau = 0.5$).



Source: Author (2020)

The qf of Y on the new parameterization has the form

$$Q_Y(u) = \exp\left\{-\left[u^{\log(1+\log^c q^{-1})/\log\tau} - 1\right]^{1/c}\right\}, \qquad 0 < \tau < 1. \tag{2.9}$$

So, the UBXII quantiles can be obtained from (2.9) by setting u values. Further, we can generate occurrences for this distribution using (2.9) by the inversion method.

2.3 STRUCTURAL PROPERTIES

In this section, we present some structural properties from the reparameterized UBXII distribution given by (2.8).

2.3.1 Ordinary moments

The ordinary moments are useful to obtain various important characteristics of a continuous distribution. The hth moment of Y is defined as $\mathbb{E}(Y^h) = \int_0^1 y^h f_Y(y;c,q)\,\mathrm{d}y$. We have the following proposition for the UBXII ordinary moments.

Proposition 2.3.1 *The hth ordinary moment of the UBXII distribution has the form*

$$\mathbb{E}(Y^h) = \frac{\log \tau^{-c}}{\log(1+\log^c q^{-1})} \sum_{j=0}^{\infty} \binom{\log\tau/\log(1+\log^c q^{-1})-1}{j} \left(h^{-c(j+1)} \gamma\left(c(j+1),h\right) + h^{c\,[\,j+\log\tau^{-1}/\log(1+\log^c q^{-1})]} \Gamma\left(-c\,[\,j+\log\tau^{-1}/\log(1+\log^c q^{-1})],h\right)\right), \tag{2.10}$$

where $\gamma(a,x) = \int_0^x t^{a-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$ *and* $\Gamma(a,x) = \int_x^{\infty} t^{a-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$ *are the lower and upper incomplete gamma function,respectively.*

Proof. By using the transformation $Y = \mathrm{e}^{-X}$, we can write $\mathbb{E}(Y^h) = \mathbb{E}(\mathrm{e}^{-hX}) = M_X(-h)$, where $M_X(t) = \int_0^{\infty} \mathrm{e}^{tx} f_X(x)\,\mathrm{d}x$ is the moment generating function (mgf) of $X \sim \mathrm{BXII}\ (c,d)$, and d is given by (2.6). Thus, evaluating the mgf of X in $-h$ (see Equation (11) in Guerra *et al.* (2020)), we obtain (2.10). ■

The hth cumulant of Y can be expressed as

$$\kappa_h = \mathbb{E}(Y^h) - \sum_{n=1}^{h-1} \binom{h-1}{n-1} \kappa_n\, \mathbb{E}(Y^{h-n}),$$

where $\kappa_1 = \mathbb{E}(Y)$ and κ_2 are the mean and variance of Y, respectively. The skewness and kurtosis are $\gamma_1 = \kappa_3/\kappa_2^{3/2}$, and $\gamma_2 = \kappa_4/\kappa_2^2$, respectively.

Alternatively, the flexibility of the new distribution can be proved from the Bowley skewness and Moors kurtosis formulas, namely

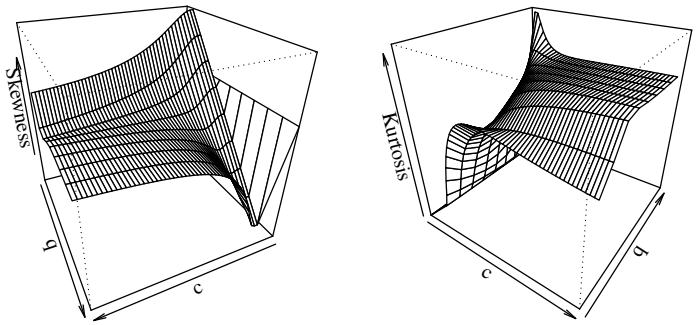
$$B = \frac{Q_Y(3/4) - 2Q_Y(1/2) + Q_Y(1/4)}{Q_Y(3/4) - Q_Y(1/4)}$$

and

$$M = \frac{Q_Y(7/8) - Q_Y(5/8) + Q_Y(3/8) - Q_Y(1/8)}{Q_Y(3/4) - Q_y(1/4)},$$

respectively, where $Q_Y(\cdot)$ is the qf given by (2.9). These measures provide a simple way to figure out the skewness and tail shapes of the distribution. For more details, readers are referred to Kenney and Keeping (1962) and Moors (1988). Figure 2 displays plots for both measures B and M which show that they are sensible to variations of c and q for fixed $\tau = 0.5$.

Figure 2 – The Bowley skewness and Moors kurtosis of the UBXII distribution.



Source: Author (2020)

2.3.2 Incomplete moments

Another important statistical measure is the hth incomplete moment of Y defined as $T_h(z) = \int_0^z y^h f_Y(y;c,q)\,\mathrm{d}y$. We can write from Equation (2.8),

$$T_h(z) = \frac{\log \tau^{-c}}{\log(1+\log^c q^{-1})} \int_0^z y^{h-1} \log^{c-1} y^{-1} \left(1+\log^c y^{-1}\right)^{\log\tau/\log(1+\log^c q^{-1})-1} \mathrm{d}y.$$

Setting $u = \log^c y^{-1}$, we have $\mathrm{d}u/\mathrm{d}y = -c\, y^{-1} \log^{c-1} y^{-1}$ and then

$$T_h(z) = \frac{\log \tau^{-1}}{\log(1+\log^c q^{-1})} \int_{\log^c z^{-1}}^{\infty} \exp\left(-hu^{1/c}\right) (1+u)^{\log\tau/\log(1+\log^c q^{-1})-1} \mathrm{d}u. \tag{2.11}$$

The general case of the binomial theorem is the power series identity

$$(x+a)^\nu = \sum_{j=0}^{\infty} \binom{\nu}{j} x^j a^{\nu-j}, \tag{2.12}$$

where $\binom{\nu}{j} = \Gamma(\nu+1)/[\Gamma(j+1)\Gamma(\nu-j+1)]$ is the generalized binomial coefficient with real arguments, $\nu \in \mathbb{R}$, and $\Gamma(z) = \int_0^\infty x^{z-1}\,\mathrm{e}^{-x}\,\mathrm{d}x$ is the gamma function. This power series converges

since $|x/a| < 1$. We can write from (2.12)

$$(1+u)^{\log\tau/\log(1+\log^c q^{-1})-1} = \sum_{j=0}^{\infty}\binom{\log\tau/\log(1+\log^c q^{-1})-1}{j} \times\left[u^{\log\tau/\log(1+\log^c q^{-1})-1-j}\mathbf{1}_{(\log^c z^{-1},1]}(u)+u^j\mathbf{1}_{(1,\infty]}(u)\right], \quad (2.13)$$

where $\mathbf{1}_\chi(x)$ denotes the indicator function over a given set χ, i.e., $\mathbf{1}_\chi(x) = 1$ if $x \in \chi$ and $\mathbf{1}_\chi(x) = 0$ elsewhere. Combining (2.13) and (2.11), we obtain

$$T_h(z) = \frac{\log\tau^{-1}}{\log(1+\log^c q^{-1})}\sum_{j=0}^{\infty}\binom{\log\tau/\log(1+\log^c q^{-1})-1}{j} \times\left[\int_{\log^c z^{-1}}^{1} u^j\exp(-hu^{1/c})\,\mathrm{d}u+\int_1^{\infty} u^{\log\tau/\log(1+\log^c q^{-1})-1-j}\exp(-hu^{1/c})\,\mathrm{d}u\right].$$

Setting $m = hu^{1/c}$ and, after some algebraic manipulation, we have

$$T_h(z) = \frac{\log\tau^{-c}}{\log(1+\log^c q^{-1})}\sum_{j=0}^{\infty}\binom{\log\tau/\log(1+\log^c q^{-1})-1}{j}\Big(h^{-c(j+1)} \times\left\{\Gamma\left(c(j+1), h\log z^{-1}\right)-\Gamma(c(j+1),h)\right\}+h^{c[\log\tau^{-1}/\log(1+\log^c q^{-1})+j]} \Gamma\left(-c[\log\tau^{-1}/\log(1+\log^c q^{-1})+j], h\right)\Big). \quad (2.14)$$

Equation (2.14) gives the hth incomplete moment of Y in terms of incomplete gamma functions, which is the main result of this section.

For empirical purposes, the shapes of many distributions can be described by the first incomplete moment such as the Lorenz and Bonferroni curves and the mean deviations. The Bonferroni and Lorenz curves are given by $B(\pi) = T_1(p)/(\pi\mu_1')$ and $L(\pi) = T_1(p)/(\mu_1')$, respectively, where $p = Q_Y(\pi)$ comes from (2.5) for a given probability π. These curves have applications in engineering, medicine, economics and several other areas.

Finally, the deviations from the mean $\mu_1' = \mathbb{E}(Y)$ and median $M = Q_Y(0.5)$ of Y can be calculated from well-known formulas

$$\delta_1 = 2\mu_1' F(\mu_1') - 2T_1(\mu_1') \qquad \textit{and} \qquad \delta_2 = \mu_1' - 2T_1(M),$$

where $F_Y(\mu_1')$ and $T_1(p)$ follow from Equations (2.7) and (2.14), respectively.

2.3.3 Generating function

The mgf of a random variable Y is defined by $\mathrm{M}_Y(t) \equiv \mathbb{E}(\mathrm{e}^{tY}) = \int_0^1 \mathrm{e}^{ty} f_Y(y;c,q)\,\mathrm{d}y$ wherever this expectation exists. It is quite used to find the moments of a given random variable.

By considering the pdf (2.8), the UBXII mgf has the form

$$M_Y(t) = \frac{\log \tau^{-c}}{\log\left(1+\log^c q^{-1}\right)} \int_0^1 y^{-1} \mathrm{e}^{ty} \log^{c-1} y^{-1} \left(1+\log^c y^{-1}\right)^{-1+\log\tau/\log\left(1+\log^c q^{-1}\right)} \mathrm{d}y.$$

Setting $u = \log^c y^{-1}$, $\mathrm{d}u/\mathrm{d}y = -\frac{c}{y}\log^{c-1} y^{-1} dy$, and $u = \exp(-u^{1/c})$, it follows that

$$M_Y(t) = \frac{\log \tau^{-1}}{\log(1+\log^c q^{-1})} \int_0^\infty \exp\left[t\exp(-u^{1/c})\right] (1+u)^{\log\tau/\log(1+\log^c q^{-1})-1} \mathrm{d}u.$$

By using the well-known power series: $\mathrm{e}^x = \sum_{k=1}^\infty x^k/k!$, and the generalized binomial theorem in Equation (2.12), we have

$$\begin{aligned} M_Y(t) = &\frac{\log \tau^{-c}}{\log(1+\log^c q^{-1})} \sum_{k=0}^\infty \frac{t^k}{k!} \sum_{j=0}^\infty \binom{\log\tau/\log(1+\log^c q^{-1})-1}{j} \left(k^{-c(j+1)}\gamma\left(c(j+1),k\right)\right. \\ &\left. + k^{c\left[\log\tau^{-1}/\log(1+\log^c q^{-1})+j\right]}\Gamma\left(-c\left[\log\tau^{-1}/\log(1+\log^c q^{-1})+j\right],k\right)\right), \end{aligned}$$

which is the main result of this section.

2.4 ESTIMATION

Various methods can be used to estimate the parameters of a distribution. The maximum likelihood (ML) method is the most commonly used. In what follows, we shall use this method for estimating the parameters of the UBXII distribution.

Let $y_1, \ldots, y_n$ be a random sample of size n from the UBXII distribution, the parameter vector $\boldsymbol{\theta} = (c, q)^\top$, and a known $\tau \in (0,1)$ specified. Based on this sample, the log-likelihood function for $\boldsymbol{\theta}$, $\ell(\boldsymbol{\theta}; \boldsymbol{y}) \equiv \ell(\boldsymbol{\theta})$, has the form

$$\begin{aligned} \ell(\boldsymbol{\theta}) = &n\log(\log\tau^{-c}) - n\log\{\log[t(q)]\} - \sum_{i=1}^n \log y_i + (c-1)\sum_{i=1}^n \log(\log y_i^{-1}) \\ &- \left[1 + \frac{\log\tau^{-1}}{\log[t(q)]}\right] \sum_{i=1}^n \log[t(y_i)], \end{aligned} \tag{2.15}$$

where $t(x) = 1+\log^c x^{-1}$.

Equation (2.15) can be maximized either directly by using well-known plataforms such as the `R` (`optim` function), `SAS` (`PROC NLMIXED`), `Ox` program (`MaxBFGS` sub-routine) or by solving the nonlinear likelihood equations from the differentiation of $\ell(\boldsymbol{\theta})$. By maximizing (2.15), we obtain the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

The components of the score vector are $U(\boldsymbol{\theta}) = [U_c(\boldsymbol{\theta}), U_q(\boldsymbol{\theta})]^\top$, where $U_c(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial c$ and $U_q(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial q$. Setting these components to zero and solving them simultaneously gives

$\hat{\boldsymbol{\theta}}$. The score components are

$$\begin{aligned} U_c(\boldsymbol{\theta}) = & \frac{n}{c} + \sum_{i=1}^{n} \log(\log y_i^{-1}) - \frac{n \log(\log q^{-1})[t(q)-1]}{t(q)\log[t(q)]} - \sum_{i=1}^{n} \frac{[t(y_i)-1]\log(\log y_i^{-1})}{t(y_i)} \\ & - \frac{\log \tau^{-1} \log[t(q)]}{\log^2[t(q)]} \sum_{i=1}^{n} [t(y_i)]^{-1}[t(y_i)-1]\log(\log y_i^{-1}) \\ & + \frac{\log \tau^{-1}[t(q)-1]\log(\log q^{-1})}{t(q)\log^2[t(q)]} \sum_{i=1}^{n} \log[t(y_i)], \end{aligned}$$

and

$$U_q(\boldsymbol{\theta}) = \frac{n\, c \log^{c-1} q^{-1}}{q\, t(q) \log[t(q)]} - \frac{\log \tau^{-c} \log^{c-1} q^{-1}}{q\, t(q) \log^2[t(q)]} \sum_{i=1}^{n} \log[t(y_i)].$$

The MLE of $\boldsymbol{\theta}$ can not be expressed in closed-form by setting $U(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$. However, for fixed c, we note that a MLE semi-closed form of q follows by taking $U_q(\boldsymbol{\theta})|_{q=\hat{q}} = 0$. Hence, it is the solution of

$$\hat{q}(c) = \exp\left(- \left\{ \exp\left[\frac{1}{n} \log \tau^{-1} \sum_{i=1}^{n} \log[t(y_i)] \right] - 1 \right\}^{1/c} \right).$$

By replacing q by $\hat{q}(c)$ in Equation (2.15), we obtain the profile log-likelihood function

$$\begin{aligned} \ell(c) = & -n + n \log(\log \tau^{-c}) - \sum_{i=1}^{n} \log y_i - \sum_{i=1}^{n} \log[t(y_i)] + (c-1) \sum_{i=1}^{n} \log(\log y_i^{-1}) \\ & - n \log \left\{ \frac{1}{n} \log \tau^{-1} \sum_{i=1}^{n} \log[t(y_i)] \right\}. \end{aligned} \tag{2.16}$$

We can compute the score function for c from (2.16)

$$U_c(c) = \frac{n}{c} + \sum_{i=1}^{n} \log(\log y_i^{-1}) - \sum_{i=1}^{n} \frac{[t(y_i)-1]\log(\log y_i^{-1})}{t(y_i)} - \frac{n \sum_{i=1}^{n} \frac{[t(y_i)-1]\log(\log y_i^{-1})}{t(y_i)}}{\sum_{i=1}^{n} \log[t(y_i)]}.$$

However, it is necessary to use a nonlinear optimization method to maximize numerically the profile log-likelihood function (2.16). Typically for the numerical computation of the MLEs, the quasi-Newton algorithm such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is adopted.

Approximate confidence intervals and hypothesis tests for $\boldsymbol{\theta}$ can be constructed by considering its asymptotic distribution of the MLEs. For large samples, $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(0, I^{-1}(\boldsymbol{\theta}))$ approximately assuming that standard regularity conditions (SRCs) hold (see Lehmann and Casella (2011)), where $I(\boldsymbol{\theta})$ is the expected information matrix defined by

$$I(\boldsymbol{\theta}) = \mathbb{E}\left(- \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\top}} \right).$$

The computation of $I(\boldsymbol{\theta})$ may be cumbersome. Nevertheless, when the SRCs are valid, it follows that $I(\boldsymbol{\theta}) = \mathbb{E}[J(\boldsymbol{\theta})]$, where $J(\boldsymbol{\theta}) = -\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ is the observed information matrix. For the UBXII distribution, we can write $J(\boldsymbol{\theta})$ as

$$J(\boldsymbol{\theta}) = -\begin{bmatrix} U_{cc}(\boldsymbol{\theta}) & U_{cq}(\boldsymbol{\theta}) \\ U_{qc}(\boldsymbol{\theta}) & U_{qq}(\boldsymbol{\theta}) \end{bmatrix},$$

where $U_{cc}(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\theta})/\partial c^2$, $U_{qq}(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\theta})/\partial q^2$, and $U_{cq}(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\theta})/(\partial c\,\partial q) = U_{qc}(\boldsymbol{\theta})$. The elements of the matrix $J(\boldsymbol{\theta})$ are given in Appendix A.

Lindsay and Li (1997) proved that the estimated observed information matrix $J(\hat{\boldsymbol{\theta}})$ is a consistent estimator of $I(\boldsymbol{\theta})$ when the sample size is large. It is then possible to obtain the standard errors (SEs) of the MLEs by computing the square roots of the diagonal elements of $J(\hat{\boldsymbol{\theta}})^{-1}$. For instance, we can do large sample inference by building asymptotic confidence intervals with $100\%(1-\alpha)$ nominal coverage for $\boldsymbol{\theta}$ making $\hat{\boldsymbol{\theta}} \pm z_{1-\alpha/2}SE(\hat{\boldsymbol{\theta}})$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ standard normal quantile.

2.5 SIMULATION STUDY

A Monte Carlo simulation study is carried out in the `R` programming language to evaluate the performance of the MLEs of the UBXII parameters that index the distribution. The `Optim` routine (with BFGS quasi-Newton nonlinear optimization algorithm and analytical derivative) is used for maximizing (2.16). The profile log-likelihood function involves a numerical maximization simpler than by using (2.15) since it depends only on the parameter c. We start the root-finding algorithm, using $c = 1$ for the shape parameter.

Different values for the parameter vector $\boldsymbol{\theta}$ are considered according to those presented in Figure 1. Therefore, various combinations of skewness and kurtosis coefficients and density shapes are contemplated. A total of eight scenarios is considered for the sample size $n \in \{25, 75, 150, 300\}$. The inversion method is employed for generating observations, i.e., the qf (2.9) is evaluated in $u \sim \mathcal{U}(0,1)$, being $Q_Y(u) = y$ and, hence, a sample of size n from $Y \sim$ UBXII (c, q) is generated. Each one of the sample sizes is replicated $R = 10,000$ times. We compute quantities as percentage relative bias (RB%) and root mean squared error (RMSE) of the MLEs.

Table 1 reports results from the simulation schemes. As expected, the consistency property of the MLEs holds, i.e., the RMSEs tend to decrease when the sample size increases. Also, it can be noted that the RB%s are smaller for sample size higher, thus indicating that the overall performance of the MLEs is appropriate, as well as they are more accurate and less biased when

n increases. Notice that the biggest RB%s for $\hat{c}$ and $\hat{q}$ are less than 7.38 and 1.62, respectively, even with $n = 25$. In general, the estimate $\hat{q}$ is more accurate when compared with $\hat{c}$. In the scenarios two to six, all the RB%s of $\hat{q}$ are below of 0.84 in absolute value.

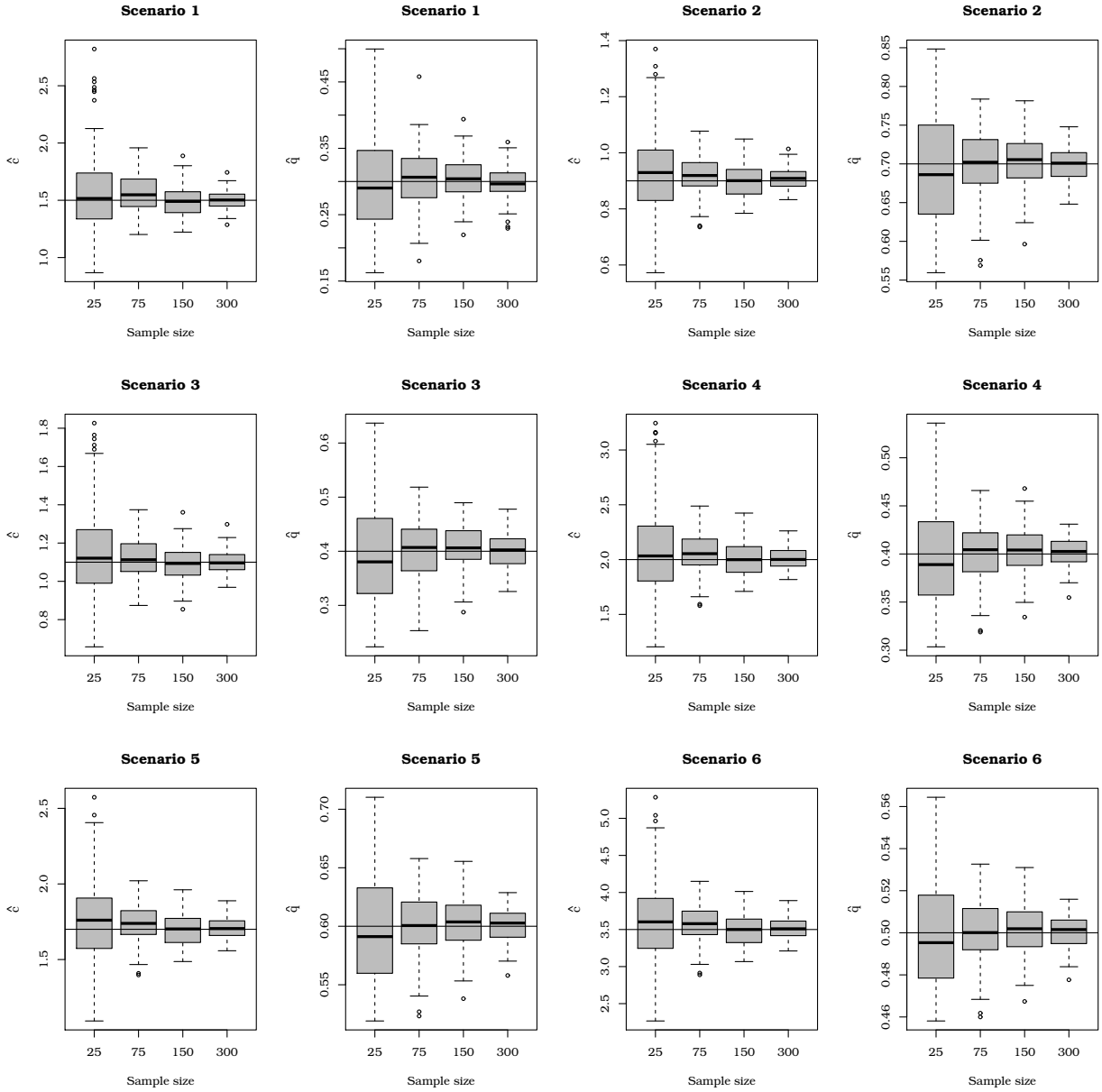
Table 1 – RB%s and RMSEs from the UBXII distribution.

Scenario	c	q	n	RB%		RMSE	
				$\hat{c}$	$\hat{q}(\hat{c})$	$\hat{c}$	$\hat{q}(\hat{c})$
1	1.5	0.3	25	7.3773	1.6170	0.3682	0.0751
			75	2.3671	0.8954	0.1823	0.0440
			150	1.1971	0.5372	0.1251	0.0310
			300	0.6399	0.3746	0.0872	0.0225
2	0.9	0.7	25	5.1296	−0.8436	0.1598	0.0758
			75	1.6937	−0.2415	0.0845	0.0434
			150	0.7708	−0.1126	0.0585	0.0311
			300	0.4013	−0.0741	0.0409	0.0221
3	1.1	0.4	25	6.3920	0.6669	0.2370	0.0967
			75	2.1153	0.6017	0.1216	0.0569
			150	1.1037	0.3580	0.0833	0.0402
			300	0.6429	0.3446	0.0583	0.0290
4	2.0	0.4	25	6.0735	0.2902	0.4203	0.0541
			75	1.9732	0.1424	0.2169	0.0313
			150	0.8997	0.0865	0.1496	0.0224
			300	0.4774	0.0212	0.1049	0.0159
5	1.7	0.6	25	5.0479	−0.1864	0.2940	0.0481
			75	1.6641	−0.0366	0.1555	0.0276
			150	0.7552	−0.0085	0.1075	0.0198
			300	0.3920	−0.0157	0.0755	0.0140
6	3.5	0.5	25	5.0270	0.0122	0.5975	0.0262
			75	1.6559	0.0203	0.3157	0.0150
			150	0.7520	0.0171	0.2182	0.0108
			300	0.3894	0.0012	0.1532	0.0076

Source: Author (2020)

Figure 3 displays boxplots from the first 100 Monte Carlo replications (to favor easy viewing) of the eight current scenarios. We can note that, in most cases, the presence of outliers overestimates the estimates for small sample sizes. However, this fact is attenuated when n increases. Besides, the dispersion of the estimates decreases, and the precision is achieved for larger sample sizes.

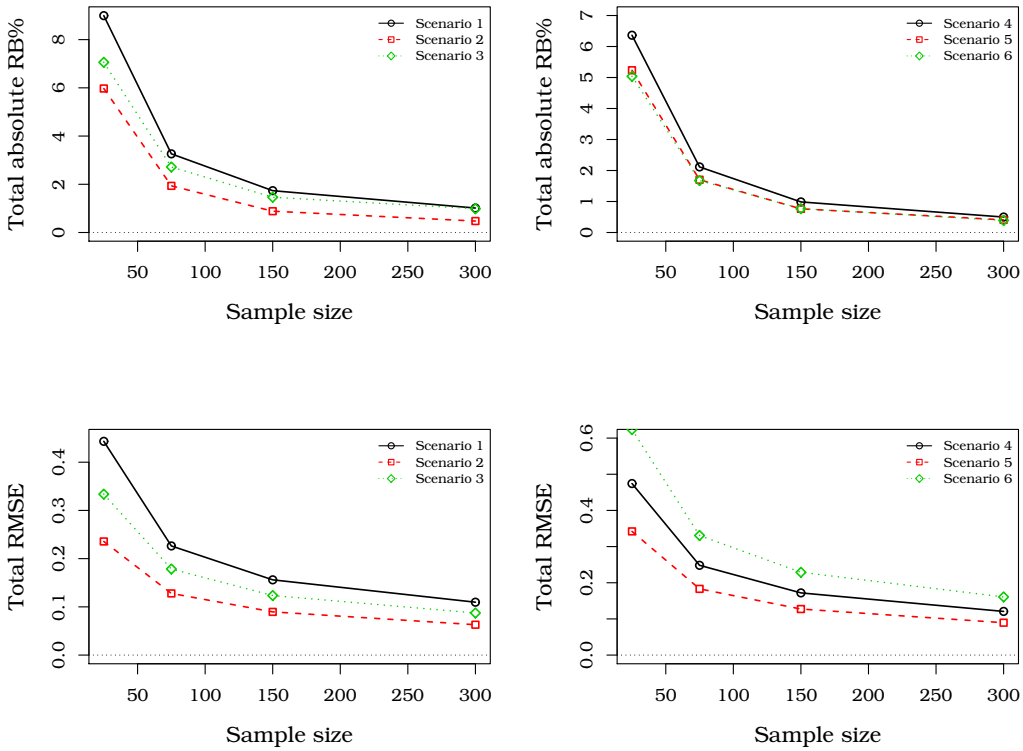
Figure 3 – Boxplots of the first hundred estimates of the Monte Carlo simulation for some sample sizes.



Source: Author (2020)

Figure 4 contains plots of total absolute RB% and total RMSE versus sample sizes for all these scenarios. These quantities are obtained from the sum of the RB% and RMSE of both parameters for each sample size and scenario. Note that those measures decay to zero when n increases in the six scenarios. This shows that the properties of the MLEs (such as asymptotically unbiased and consistent) are held.

Figure 4 – Total absolute RB%s and total RMSE of the MLEs from UBXII distribution with different sample sizes.



Source: Author (2020)

2.6 THE UBXII REGRESSION

Let $Y_1, \ldots, Y_n$ be n independent random variables, where $Y_i \sim \text{UBXII}(q_i, c)$ for $i = 1, \ldots, n$ with shape parameter c and quantile parameter q_i (both unknown) for $0 < \tau < 1$ assumed known. We propose the *UBXII regression* imposing that the quantile q_i of Y_i satisfies the functional relation

$$\boldsymbol{\eta} = g(\boldsymbol{q}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{2.17}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top \in \mathbb{R}^n$ is the n-dimensional vector of linear predictors, $\boldsymbol{q} = (q_1, \ldots, q_n)^\top$ is the vector of quantiles with $q_i \in (0, 1)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k$ is a k-dimensional vector of unknown regression coefficients $(k < n)$, $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$ is the $n \times k$ full column rank matrix, $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ik})$ denotes the ith observation on k covariates which are assumed known, and $x_{i1} = 1, \forall i$. Finally, we shall assume that $g(\cdot)$ is a strictly monotonic and twice differentiable link function which maps $(0, 1)$ into $\mathbb{R}$. By inverting each component of (2.17), we can write

$$q_i = g^{-1}\left(\sum_{j=1}^{k} x_{ij}\beta_j\right) = g^{-1}(\eta_i).$$

There are various possible choices for the link function $g(\cdot)$ such as

- logit: $g(q_i) = \log[q_i/(1-q_i)]$;
- probit: $g(q_i) = \Phi^{-1}(q_i)$, where $\Phi^{-1}(\cdot)$ is the qf of the standard normal random variable;
- complementary log-log: $g(q_i) = \log[-\log(1-q_i)]$;
- log-log: $g(q_i) = -\log[-\log(q_i)]$;
- Cauchy: $g(q_i) = \tan[\pi(q_t - 1/2)]$.

McCullagh McCullagh and Nelder (1989) provides a comparison among some of them.

The choice of the logit link function is the most common by practitioners since the interpretation of the regression parameters becomes quite interesting. Consider increasing the jth regressor at one unit, while the others are kept constant. Let q^* be the quantile of Y under the new value of $\boldsymbol{x}_j$, whereas q denotes the quantile of Y under the original value of this regressor. It can be shown that with the logit link function, we have $\beta_j = \log\{q^*(1-q^*)/[q(1-q)]\}$, i.e., β_j is the log odds ratio (FERRARI; CRIBARI-NETO, 2004). In this context, we will consider the logit link function for $g(\cdot)$ in the UBXII regression. Then, the ith quantile of Y_i is $q_i = \mathrm{e}^{\eta_i}/(1+\mathrm{e}^{\eta_i})$.

2.6.1 Estimation

The parameters estimation in the UBXII regression can also be performed by the ML method. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, c)^\top$ be the vector of $k+1$ unknown parameters to be estimated. The log-likelihood function based on a sample of n independent observations having the UBXII distribution, i.e., $Y_i \sim$ UBXII (q_i, c), can be expressed as

$$\ell(\boldsymbol{\theta}) \equiv \ell(\boldsymbol{\beta}, c) = \sum_{i=1}^{n} \ell_i(q_i, c), \tag{2.18}$$

where $\ell_i(q_i, c)$ is the logarithm of $f_Y(y_i; q_i, c)$ given in Equation (2.8). Hence,

$$\begin{aligned}\ell_i(q_i, c) =& \log(\log \tau^{-c}) - \log y_i + (c-1)\log(\log y_i^{-1}) - \log[t(y_i)] - \log\{\log[t(q_i)]\} \\ & - \frac{\log \tau^{-1} \log[t(y_i)]}{\log[t(q_i)]}.\end{aligned}$$

The score vector, obtained by differentiating the log-likelihood function (2.18) with respect to the unknown parameters β_j, $j = 1, \ldots, k$, and c, is expressed as $\boldsymbol{U} = \left[U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, c)^\top, U_c(\boldsymbol{\beta}, c)\right]^\top$. The components of $\boldsymbol{U}$ can be written in matrix notation. For doing this, we now define some quantities.

Let $q_i^\star = \log^{c-1} q_i^{-1}/\{q_i\, t(q_i) \log[t(q_i)]\}$, $q_i^\dagger = \log \tau^{-1} \log^{c-1} q_i^{-1}/\{q_i\, t(q_i) \log^2[t(q_i)]\}$,

$y_i^\star = \log[t(y_i)]$, and

$$y_i^\sharp = \frac{1}{c} + \log(\log y_i^{-1}) - \frac{\log(\log q_i^{-1})[t(q_i)-1]}{t(q_i)\log[t(q_i)]} - \frac{[t(y_i)-1]\log(\log y_i^{-1})}{t(y_i)} - \frac{\log\tau^{-1}\log[t(q_i)][t(y_i)]^{-1}[t(y_i)-1]\log(\log y_i^{-1})}{\log^2[t(q_i)]} + \frac{\log\tau^{-1}[t(q_i)-1]\log(\log q_i^{-1})\log[t(y_i)]}{t(q_i)\log^2[t(q_i)]}.$$

Then, we have

$$U_{\boldsymbol{\beta}} \equiv U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, c) = c\,\boldsymbol{X}^\top \boldsymbol{D}\,(\boldsymbol{q}^\star - \boldsymbol{q}^\dagger\,\boldsymbol{y}^\star), \tag{2.19}$$

and

$$U_c \equiv U_c(\boldsymbol{\beta}, c) = \mathrm{tr}(\boldsymbol{Y}^\sharp), \tag{2.20}$$

where $\boldsymbol{X}$ is an $n \times k$ matrix whose ith row is $\boldsymbol{x}_i^\top$, $\boldsymbol{D} = \mathrm{diag}\{1/g'(q_1), \ldots, 1/g'(q_n)\}$, $\boldsymbol{q}^\star = (q_1^\star, \ldots, q_n^\star)^\top$, $\boldsymbol{q}^\dagger = (q_1^\dagger, \ldots, q_1^\dagger)^\top$, $\boldsymbol{y}^\star = (y_1^\star, \ldots, y_n^\star)^\top$, and $\boldsymbol{Y}^\sharp = \mathrm{diag}\{y_1^\sharp, \ldots, y_n^\sharp\}$. We provide the calculations of the score components in Appendix A.

Again, the nonlinear Equations $U_{\boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0$ and $U_c|_{c=\hat{c}} = 0$ can not be expressed in closed-form. Hence, a nonlinear optimization method must be used for maximizing the function (2.18) and determine the MLEs $(\hat{\boldsymbol{\beta}}^\top, \hat{c})^\top$. We also provide the observed information matrix for $(\boldsymbol{\beta}^\top, c)^\top$.

To simplify the notation of its components, other quantities are defined as follows

$$m_i = \left\{ \frac{c\log^c q_i^{-1}}{q_i t(q_i)} + \frac{c\log^c q_i^{-1}}{q_i t(q_i)\log[t(q_i)]} - \frac{\log q_i^{-1}}{q_i} - \frac{(c-1)}{q_i} \right\} \frac{c\log^{c-2} q_i^{-1}}{q_i t(q_i)\log[t(q_i)]},$$

$$p_i = \left\{ \frac{(c-1)}{q_i\log[t(q_i)]} + \frac{\log q_i^{-1}}{q_i\log[t(q_i)]} - \frac{2c\log^c q_i^{-1}}{q_i t(q_i)\log^2[t(q_i)]} - \frac{c\log^c q_i^{-1}}{q_i t(q_i)\log[t(q_i)]} \right\} \frac{c\log\tau^{-1}\log^{c-2} q_i^{-1}}{q_i t(q_i)\log[t(q_i)]},$$

$$r_i = \left\{ \log^{c-1} q_i^{-1} + \frac{\log^{c-1} q_i^{-1}}{c\log(\log q_i^{-1})} - \frac{\log^{2c-1} q_i^{-1}}{t(q_i)} - \frac{\log^{2c-1} q_i^{-1}}{t(q_i)\log[t(q_i)]} \right\} \frac{c\log(\log q_i^{-1})}{q_i t(q_i)\log[t(q_i)]},$$

$$u_i = \left\{ \frac{2\log^{2c-1} q_i^{-1}}{t(q_i)\log^2[t(q_i)]} + \frac{\log^{2c-1} q_i^{-1}}{t(q_i)\log[t(q_i)]} - \frac{\log^{c-1} q_i^{-1}}{c\log(\log q_i^{-1})\log[t(q_i)]} - \frac{\log^{c-1} q_i^{-1}}{\log[t(q_i)]} \right\} \times \frac{c\log(\log q_i^{-1})\log\tau^{-1}}{q_i t(q_i)\log[t(q_i)]},$$

$$s_i = \frac{c\log\tau^{-1}\log^{c-1} q_i^{-1}}{q_i t(q_i)\log^2[t(q_i)]} \quad \text{and} \quad y_i^\dagger = \log(\log y_i^{-1})[t(y_i)-1][t(y_i)]^{-1}.$$

Therefore, the observed information matrix can be expressed as (see Appendix A)

$$\boldsymbol{J} = -\begin{pmatrix} \boldsymbol{J}_{\boldsymbol{\beta\beta}} & \boldsymbol{J}_{c\boldsymbol{\beta}} \\ \boldsymbol{J}_{\boldsymbol{\beta}c} & J_{cc} \end{pmatrix}.$$

The quantities $\boldsymbol{J}_{\boldsymbol{\beta\beta}} \equiv \partial^2\ell(\boldsymbol{\beta}, c)/(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top)$ and $\boldsymbol{J}_{\boldsymbol{\beta}c} = \boldsymbol{J}_{c\boldsymbol{\beta}}^\top \equiv \ell(\boldsymbol{\beta}, c)/(\partial c\partial\boldsymbol{\beta})$, and $J_{cc} \equiv \partial^2\ell(\boldsymbol{\beta}, c)/\partial c^2$ are

$$\boldsymbol{J}_{\boldsymbol{\beta\beta}} = \boldsymbol{X}^\top [\,(\boldsymbol{M} + \boldsymbol{P}\boldsymbol{Y}^\star)\,\boldsymbol{D} - c\,(\boldsymbol{Q}^\star - \boldsymbol{Q}^\dagger\boldsymbol{Y}^\star)\,\boldsymbol{T}\,\boldsymbol{D}^\top\boldsymbol{D}\,]\,\boldsymbol{D}\,\boldsymbol{X}, \tag{2.21}$$

$$\boldsymbol{J}_{c\boldsymbol{\beta}}^\top = (\boldsymbol{r} - \boldsymbol{s}\,\boldsymbol{y}^\ddagger + \boldsymbol{u}\,\boldsymbol{y}^\star)^\top\,\boldsymbol{D}\,\boldsymbol{X}, \tag{2.22}$$

and

$$J_{cc} = \mathrm{tr}(\boldsymbol{Y}^\diamond), \tag{2.23}$$

where $\boldsymbol{M} = \mathrm{diag}\{m_1, \ldots, m_n\}$, $\boldsymbol{P} = \mathrm{diag}\{p_1, \ldots, p_n\}$, $\boldsymbol{Q}^\star = \mathrm{diag}\{q_1^\star, \ldots, q_n^\star\}$, $\boldsymbol{Q}^\dagger = \mathrm{diag}\{q_1^\dagger, \ldots, q_n^\dagger\}$, $\boldsymbol{Y}^\star = \mathrm{diag}\{y_1^\star, \ldots, y_n^\star\}$, $\boldsymbol{T} = \mathrm{diag}\{g''(q_1), \ldots, g''(q_n)\}$, $\boldsymbol{r} = (r_1, \ldots, r_n)^\top$, $\boldsymbol{s} = (s_1, \ldots, s_n)^\top$, $\boldsymbol{y}^\ddagger = (y_1^\ddagger, \ldots, y_n^\ddagger)^\top$, and $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$.

We note that the parameters $\boldsymbol{\beta}$ and c are not orthogonal as in the beta (FERRARI; CRIBARI-NETO, 2004), gamma (MOUSA; EL-SHEIKH; ABDEL-FATTAH, 2016), Johnson S_B (LEMONTE; BAZÁN, 2016) regressions for modeling bounded random variables to the standard unit interval and unlike to the generalized linear models discussed by Nelder and Wedderburn Nelder and Wedderburn (1972).

As mentioned in Section 2.4, the matrix $\boldsymbol{J}$ is quite useful for interval estimation and hypothesis testing inference. Assuming that the SRCs hold and the sample size is large,

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{c} \end{pmatrix} \sim \mathcal{N}_{k+1}\left(\begin{pmatrix} \boldsymbol{\beta} \\ c \end{pmatrix}, \boldsymbol{I}^{-1}\right),$$

where $\boldsymbol{I}^{-1}$ is the inverse of $\boldsymbol{I} \equiv \mathbb{E}(\boldsymbol{J})$ is the expected information matrix. It can be estimated of the consistent way by $\hat{\boldsymbol{J}}$, which is computed after replacing the unknown parameters $(\boldsymbol{\beta}^\top, c)^\top$ by the corresponding MLEs.

2.6.2 Simulation study

In this section, a Monte Carlo simulation study is conducted in order to numerically evaluate the finite sample behavior of the MLEs of the UBXII regression's parameters. The

Monte Carlo experiments are performed using the R programming language (R Core Team, 2020). Maximization of the log-likelihood function in (2.18) is carried out using the BFGS quasi-Newton nonlinear optimization algorithm implemented at the `optim` function available in R. We consider the ordinary least squares estimates (OLSEs) as an initial guess for $\boldsymbol{\beta}$ obtained from a linear regression of the transformed responses: $\boldsymbol{z} = [g(q_1), \ldots, g(q_n)]^\top$, i.e., the initial point estimate of $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{z}$. For the shape parameter c, we take the same initial guess in Section 2.5.

The simulations are based on the UBXII regression:

$$\text{logit}(q_i) = \beta_1 + \beta_2 x_{i2}, \qquad i = 1, \ldots, n. \tag{2.24}$$

The covariate $\boldsymbol{x_2}$ is randomly generated from a standard normal. We combine various values of the parameter vector $\boldsymbol{\theta} = (\beta_1, \beta_2, c)^\top$ at six different scenarios. The Monte Carlo replications number adopted and the sample sizes considered are the same from Section 2.5. In each Monte Carlo replication, the inversion method is used to generate n occurrences of a random variable $Y_i \sim \text{UBXII}\ (q_i, c)$. By assuming the regression structure defined in Equation (2.24), it follows that

$$q_i = \frac{\exp(\beta_1 + \beta_2 x_{i2})}{1 + \exp(\beta_1 + \beta_2 x_{i2})},$$

i.e., q_i is equal to the logistic cdf evaluated at $(\beta_1 + \beta_2 x_{i2})$. The statistical quantities computed are also the same of Section 2.5.

Table 2 presents the results of the Monte Carlo simulations. As expected, the RMSEs decrease for all scenarios considered when n increases, thus indicating that the MLEs are consistent. In general, the RB%s are smaller for larger sample sizes. We can note that the most RB% is equal to 10.02 in scenario four for the smallest sample size, and it refers to the estimate of c. For estimates of the parameters β_1 and β_2, all RB%s are below 6.25. In addition, even for $n = 25$, the RMSE values are quite low in any scheme.

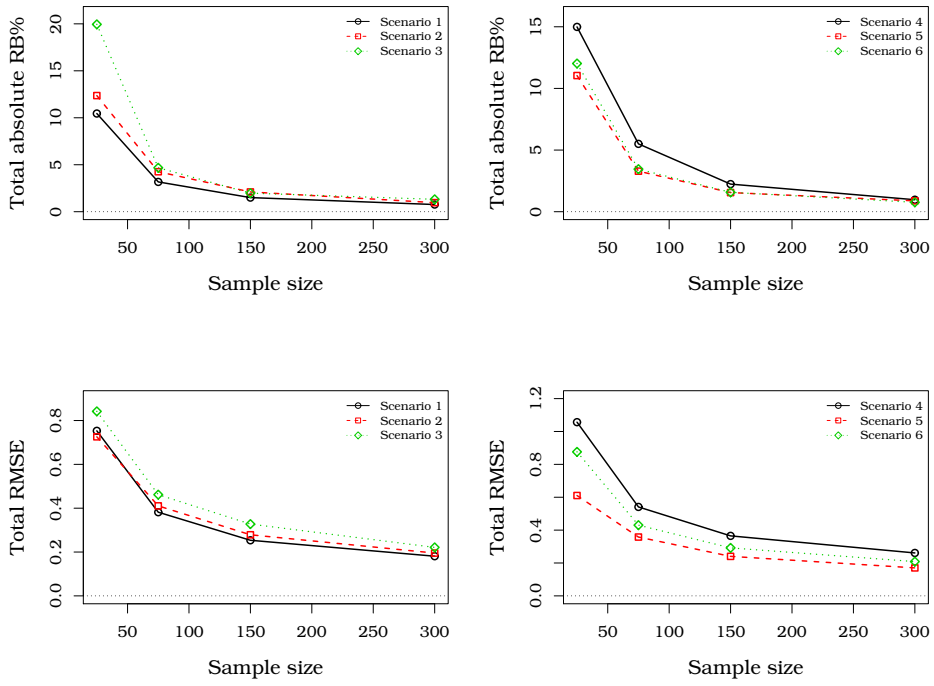
Figure **??** displays plots for the total RB% and total RMSE versus sample sizes. They reveal that the MLEs are consistent, and their biases quickly tend to zero when the sample size grows. Further, the most RB% is about 20, but it decays to less than 5 to the $n = 75$. Thus, as expected, the ML asymptotic properties remain.

Table 2 – RB%s and RMSEs for the UBXII regression.

Scenario	β_1	β_2	c	n	RB%			RMSE		
					$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{c}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{c}$
1	1.3	1.4	2.0	25	−0.3323	0.7091	9.4106	0.1760	0.1499	0.4270
				75	−0.0978	0.3527	2.7208	0.0929	0.0913	0.1975
				150	−0.0392	0.1676	1.2937	0.0644	0.0570	0.1323
				300	−0.0068	0.1248	0.6621	0.0455	0.0451	0.0918
2	0.7	0.4	1.3	25	−1.6493	2.8842	7.8287	0.2680	0.2073	0.2504
				75	−0.1791	1.5523	2.5257	0.1462	0.1397	0.1246
				150	−0.0765	0.8871	1.1396	0.1043	0.0902	0.0843
				300	−0.0671	0.3218	0.5765	0.0732	0.0637	0.0584
3	−0.2	−0.6	1.8	25	6.2548	3.7917	9.8991	0.2599	0.1545	0.4274
				75	1.0153	0.9514	0.1454	0.1454	0.1264	0.1900
				150	0.0641	0.5987	1.3144	0.1001	0.0951	0.1324
				300	0.3399	0.2976	0.6825	0.0724	0.0590	0.0898
4	−0.7	0.4	2.3	25	1.7725	3.1922	10.0246	0.2625	0.2104	0.5838
				75	0.3187	2.0428	3.1417	0.1391	0.1240	0.2780
				150	0.0243	0.7296	1.4848	0.0979	0.0777	0.1894
				300	−0.0512	0.1449	0.7751	0.0691	0.0603	0.1317
5	1.2	−0.5	1.6	25	−0.1732	2.9301	7.9394	0.1860	0.1217	0.3028
				75	−0.0054	0.7715	2.5134	0.1039	0.1069	0.1470
				150	0.0099	0.3839	1.1571	0.0748	0.0663	0.0992
				300	−0.0200	0.2634	0.5976	0.0529	0.0486	0.0687
6	0.4	1.2	2.6	25	−1.4705	0.9087	9.6420	0.1666	0.1364	0.5731
				75	−0.2535	0.4271	2.7730	0.0854	0.0776	0.2673
				150	−0.0936	0.1443	1.3345	0.0597	0.0543	0.1779
				300	0.0397	0.0499	0.6723	0.0431	0.0417	0.1237

Source: Author (2020)

Figure 5 – Total absolute RB%s and total RMSE of the MLEs from UBXII regression with different sample sizes.



Source: Author (2020)

2.6.3 Diagnostic measures and model selection

In order to check the goodness-of-fit and validate the UBXII regression assumptions, we adopt some well-known diagnostic tools that are now discussed. Initially, we use the randomized quantile residuals introduced by Dunn and Smyth (1996). These residuals allow to verify if the model assumptions are satisfied and identifying when the parameter estimations are considerably affected by the presence of atypical observations in the response. If the model is correctly specified, the randomized quantile residuals are standard normally distributed. For the UBXII regression, they are given by

$$r_i = \Phi^{-1}[F_Y(y_i; \hat{q}_i, \hat{c})],$$

where $F_Y(\cdot)$ is the UBXII cdf given in Equation (2.7).

An incorrect functional form specification of the regression and the covariates omission can be identified through the RESET test. This test was initially introduced by Ramsey (1969) as a general misspecification test for the normal linear regression. Afterward, variants of the RESET test for classes of more general regressions were proposed by McCullagh and Nelder (1989) and Pereira and Cribari-Neto (2014). Thus, to determine whether a UBXII regression is misspecified, we propose using a RESET-like misspecification test. Next, we explain how this test can be performed.

The RESET-like test is carried out in two steps. Let $\hat{\boldsymbol{q}}$ be the predicted values vector obtained after fitting a UBXII regression. First, we build testing variables matrix as $\boldsymbol{T} = [\hat{\boldsymbol{q}}^2, \hat{\boldsymbol{q}}^3]$, where the vectors $\hat{\boldsymbol{q}}^2$ and $\hat{\boldsymbol{q}}^3$ are formed by $\hat{\boldsymbol{q}}$ squared and cubed components, respectively. We define the augmented regression

$$g(\boldsymbol{q}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{T}\boldsymbol{\delta}, \tag{2.25}$$

where $\boldsymbol{T}$ is the $n \times 2$ matrix of testing variables, and $\boldsymbol{\delta}$ is a 2×1 vector of parameters. Second, we estimate Equation (2.25) and test the null hypothesis $\mathcal{H}_0 : \boldsymbol{\delta} = \boldsymbol{0}$ against the alternative hypothesis $\mathcal{H}_1 : \boldsymbol{\delta} \neq \boldsymbol{0}$ by using the likelihood ratio (LR) statistic. We compute the LR statistic as $\omega = 2[\ell(\hat{\boldsymbol{\theta}} - \ell(\tilde{\boldsymbol{\theta}})]$, where $\ell(\cdot)$ is the log-likelihood function and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\delta}}^\top, \hat{\boldsymbol{\beta}}^\top, \hat{c})^\top$ is the unrestricted MLE of $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}} = (\boldsymbol{0}^\top, \tilde{\boldsymbol{\beta}}^\top, \tilde{c})^\top$ is the restricted MLE of $\boldsymbol{\theta}$ under the null hypothesis. Under $\mathcal{H}_0$ and the SRCs, ω converge in distribution to chi-square, χ^2_ν, where ν is the number of testing covariates added to the regression ($\nu = 2$ in this case). The non-rejection of the null hypothesis suggests that the regression is correctly specified.

The proportion of the response variable's variability explained by a fitted UBXII regression can be assessed using the generalized (pseudo) R-squared (R^2_G). Nagelkerke *et al.* (1991) defined it as

$$R^2_G = 1 - \exp\left\{-2/n\left[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)\right]\right\},$$

where $\ell(\hat{\boldsymbol{\theta}}_0)$ is the log-likelihood of the null regression, i.e., obtained from the modeling of the response in the covariates absence, and $\ell(\hat{\boldsymbol{\theta}})$ is the log-likelihood of the full regression. A regression with a higher value of R^2_G provides a larger explanation power of the response variable variation.

To select the more suitable model between several nested models, the information criteria such as Akaike information criterion (AIC) (AKAIKE, 1973) and Schwarz information criterion (BIC) (SCHWARZ *et al.*, 1978) can be considered. Both criteria are widely used in practical applications and they are defined by $\text{AIC}(\phi) = 2\left[\, p - \ell(\hat{\boldsymbol{\theta}})\right]$ and $\text{BIC} = p\log n - 2\ell(\hat{\boldsymbol{\theta}})$, where p is the number of estimated parameters.

A way of selecting the best one between different non-nested regressions is to assess its performance in the prediction of the response through statistical learning tools such as the cross-validation approach. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ the vector of n observations of a response variable and $\boldsymbol{X}$ the covariates matrix like in (2.17). In statistical learning methods, a training data set is the observations set in which a model is initially adjusted. An accuracy measure is the *test error*, that result from applying the model fitted to test observations that were not used in training. For example, if we use $(\boldsymbol{y}, \boldsymbol{X})$ as training observations, the test error is $\mathbb{E}[L(Y_0, \hat{y}_0)]$, where $L(\cdot)$ is the loss function and $\hat{y}_0$ is the predicted value using the fitted model from $(\boldsymbol{y}, \boldsymbol{X})$ evaluated in the predictors $\boldsymbol{x}_{\boldsymbol{0}}^\top$ (that does not belong to $\boldsymbol{X}$). To estimate the test error with absolute and quadratic loss, respectively, we consider the mean square error (MSE) defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i$ is the ith predict value by the regression for the ith observation. This statistical measure is small if the predictions of the responses are very close to its true values, and it is large if for some of the observations, the predicted and true responses differ substantially (JAMES *et al.*, 2013).

As cross-validation method we propose the use of the leave-one-out cross-validation (LOOCV). In this approach, we split the ith observation (ith row of a data set in which the response and covariates are disposed by columns) of the other $n-1$ observations that represent the training set whereas the row i is the validation set.

For each removed observation, we use the fitted model with the training set to predict the ith observation of the validation set. After, we estimate the test error by computing the MSEi. Repeating those procedure n times, we obtain $\text{MSE}_1, \ldots, \text{MSE}_n$. The final estimate of the test errors are computed through average of those n statistics as follows (JAMES *et al.*, 2013)

$$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\text{MSE}_i.$$

Hence, we select the regression which provides smaller values for $\text{CV}_{(n)}$.

2.7 APPLICATION

In this section, we assess the UBXII regression performance on real data. The analysis is carried out using the R statistical computing environment (R Core Team, 2020). We fit the UBXII regression and compare it with the Kumaraswamy (Kw), unit-Weibull (UW) (MAZUCHELI *et al.*, 2020), and beta (FERRARI; CRIBARI-NETO, 2004) regressions, which are well-known in the analysis of limited data.

Let Y be a random variable Kw distributed under a median-dispersion parameterization (MITNIK; BAEK, 2013), say $Y \sim \text{Kw}(\omega, d_p)$. The pdf of Y is

$$f(y;\omega,d_p) = \frac{\log 0.5}{d_p \log(1-\omega^{1/d_p})} y^{1/d_p}(1-y^{1/d_p})^{\log 0.5/\log(1-\omega^{1/d_p})-1}, \quad y \in (0,1)$$

where $0 < \omega < 1$ is the median of Y and $d_p > 0$ is a dispersion parameter.

The UW quantile regression was recently introduced by Mazucheli *et al.* (2020). Let $Y \sim \text{UW}(\mu, \gamma)$ be a random variable having the UW distribution under the parameterization given in Mazucheli *et al.* (2020). For $y \in (0,1)$, the random variable Y has density

$$f(y;q,\gamma) = \frac{\gamma}{y}\left(\frac{\log\tau}{\log q}\right)\left(\frac{\log y}{\log q}\right)^{\gamma-1}\tau^{(\log y/\log q)^{\gamma}},$$

where $0 < q < 1$ is the τth quantile, $\gamma > 0$ is a shape parameter, and $\tau \in (0,1)$ is assumed known. Here, it will be considered that $\tau = 0.5$ in order to model the median of Y.

Ferrari and Cribari-Neto (2004) pioneered the beta regression. Different parameterizations can be considered for the beta distribution. We consider a mean-dispersion based parameterization. Let Y be a random variable that follows a beta distribution, say $Y \sim \text{Beta}(\mu, \sigma)$. For $y \in (0,1)$, the Y density is

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

where $0 < \mu < 1$ is the mean of Y, $0 < \sigma < 1$ is a dispersion parameter and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x$ is the complete gamma function. Under this parameterization the variance of Y is $\sigma^2\mu(1-\mu)$.

The regression structure for the Kw, UW, and beta distributions is analogous to (2.24). The main differences are the assumptions under the random components and modeled location parameters. To get the Kw regression, $\boldsymbol{q}$ must be replaced by the median ($\boldsymbol{\omega}$) in Equation (2.24) and supposed that $Y \sim \mathrm{Kw}(\omega_i, d_p)$. The UW regression is obtained by considering the structure (2.24) and assuming that $Y \sim \mathrm{UW}(q_i, \gamma)$. In the beta regression, the location parameter is the mean (μ). Thus, in Equation (2.24), $\boldsymbol{q}$ must be replaced by $\boldsymbol{\mu}$ and supposed that $Y \sim \mathrm{Beta}(\mu_i, \sigma)$.

We get the data from the higher education census conducted yearly by the Brazilian National Institute for Educational Studies and Research "Anísio Teixeira". We are interested in the dropout proportion for animal sciences courses and factors associated with their enrollment and organizational structure. However, the response variable is not directly obtained from the original dataset, and we use mining data techniques to obtain it from other reported variables. After preprocessing and cleaning steps, we select 40 covariates as possible predictors. A detailed description of the data mining tools employed and the final data set are available in appendix A.

The UBXII, Kw, UW, and beta regressions also are used as data mining tools to select a subset of predictors that properly fits the dropout proportion. We test several combinations of predictors using the measures described in Section 6.3. to define the final regressions on each class. In what follows, we describe the response variable and predictive covariates used in our regression analysis.

The response variable is the dropout proportion from 2009 until 2017 of 77 Brazilian undergraduate animal sciences courses. For each course i ($i = 1, \ldots, 77$), we consider three covariates as follows: i) quantity of vacancies offered in the morning shift, denoted by x_{i2}; ii) a dummy variable that equals one if the course guarantees conditions of accessibility for people with disabilities, and zero otherwise, denoted by x_{i3}; and iii) a dummy variable, denoted by x_{i4}, that equals one if the course works on the night shift, and zero otherwise.

Let $\boldsymbol{y} = (y_1, \ldots, y_{77})^\top$ be the vector of the response variable and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4)$ the covariates matrix, where $\boldsymbol{x}_1$ is a vector column with 77 ones and $\boldsymbol{x_j} = (x_{1j}, \ldots, x_{77j})^\top$, with $j = 2, \ldots, 4$. Table 3 provides a descriptive summary of the response variable ($\boldsymbol{y}$) and quantitative covariate ($\boldsymbol{x_2}$), revealing that $\boldsymbol{y}$ has negatively skewed distribution and lighter tails than a normal distribution. Further, its mean is close to the median, the standard deviation (SD) is low, and the values range is sizeable because the minimum and maximum are 0.1077 and 0.9714, respectively. The covariate $\boldsymbol{x}_2$ presents different degrees of variability, skewness, and kurtosis.

Table 3 – Descriptive statistics from the response variable and quantitative covariates.

Var.	Statistics						
	Mean	Median	SD	Skewness	Kurtosis	Min.	Max.
$\boldsymbol{y}$	0.5736	0.5965	0.1818	−0.3449	0.0854	0.1077	0.9714
$\boldsymbol{x}_2$	13.7532	0.0000	29.5449	2.0533	3.2902	0.0000	120.0000

Source: Author (2020)

To study the covariates' effects on the median dropout proportion, we determine $\tau = 0.5$ and specify the UBXII regression as

$$\text{logit}(q_i) = \eta_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4},$$

For comparison purposes, we also fit the Kw, UW, and beta regressions considering the same covariates combination and link function.

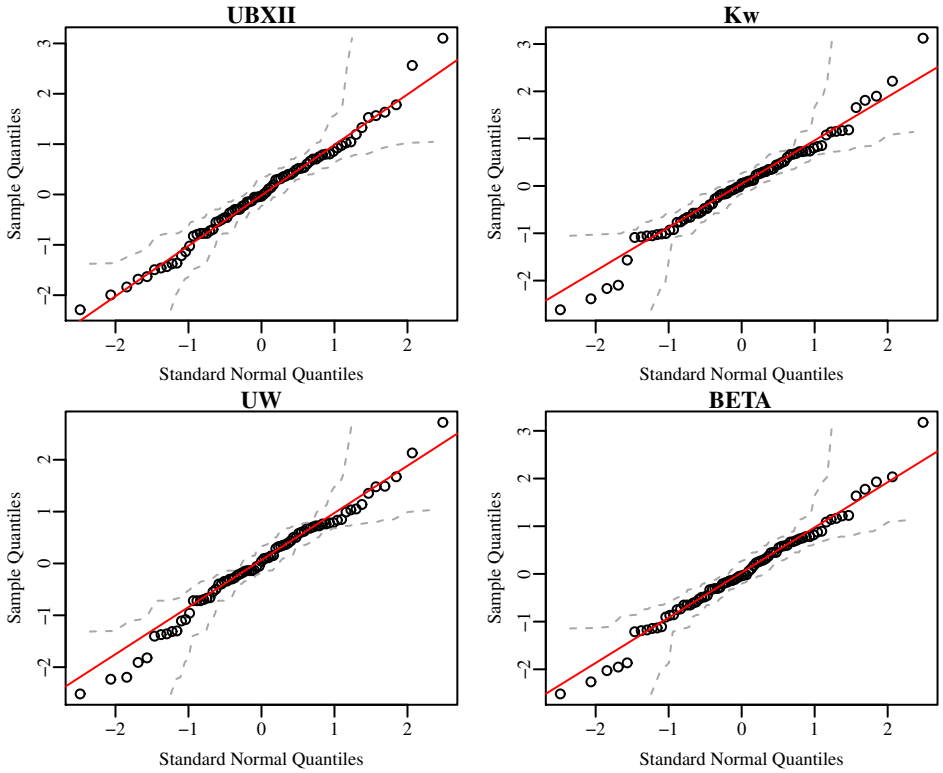
Table 4 brings some goodness-of-fit measures such as AIC, BIC, and R^2_G, the p-values of the Anderson-Darling test (AD) (STEPHENS, 1974) to validate the null hypothesis that errors are normally distributed, the p-values from RESET-like test (RES), and the statistic obtained from the LOOCV approach ($\text{CV}_{(77)}$) that allows assessing the prediction performance of the fitted regressions. We consider $\alpha = 0.05$ as a significance level for all performed hypothesis tests. According to the RESET-like tests, all models are correctly specified. Similarly, the p-values from Anderson-Darling tests indicate is reasonable supposing normality of the errors at each class. It is noteworthy that most of some goodness-of-fit measures suggest that the UBXII regression is more suitable to fit the dropout proportion in the Brazilian zootechnics course between 2009 and 2017 than other considered class of regressions. Moreover, the $CV_{(77)}$ estimate for the fitted UBXII regression is the smallest among all other fitted regressions. This means that the proposed regression leads to better predictions than the classical regressions used in the context of restricted response to the unit interval. Indeed, in Figure 6 it is possible to note that the UBXII regression provides the best fit for this data set since about 97% of the points are under the red line in the QQ-plot of fitted UBXII regression's residuals.

Table 4 – Goodness-of-fit measures and LOOCV statistic for the fitted regressions

Regression	AIC	BIC	R^2_G	AD	RES	$\text{CV}_{(77)}$
UBXII	−55.8423	−44.1233	0.2348	0.8229	0.4334	0.0259
Kw	−48.8064	−37.0873	0.1898	0.2765	0.8354	0.0260
UW	−52.6329	−40.9139	0.2565	0.2795	0.5764	0.0266
BETA	−52.0595	−40.3405	0.2235	0.5433	0.8383	0.0285

Source: Author (2020)

Figure 6 – QQ-plots of the UBXII, Kw, UW, and beta regressions' residuals.



Source: Author (2020)

In Table 5, we provide the estimates of the parameters, standard errors, t statistic value, and p-values for the UBXII regression. The effect of the three considered covariates under the response's median is positive. Further, according to the estimate of β_4, the covariate x_{i4} presents the most impact on the median. That is, the odds ratio increases substantially if the course works on the night shift. Results from other fitted regressions are given in Appendix A; see Table 17.

Table 5 – Fitted UBXII regression for the dropout proportion in the Brazilian zootechnics course.

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
β_1	−0.0509	0.1294	−0.3932	0.6953
β_2	0.0082	0.0024	3.4429	0.0010
β_3	0.5389	0.1560	3.4535	0.0009
β_4	0.8310	0.2665	3.1183	0.0026
c	2.3780	0.2032	–	–

Source: Author (2020)

2.8 CONCLUSIONS

In this chapter, we define the unit Burr XII (UBXII) distribution and its associated regression, which are useful to model continuous random variables in the interval $(0, 1)$. The new distribution is quite flexible, and its density can assume many shapes, including the reverse tilde-shaped. A highlight of this distribution is that one of its parameters, $q(\tau)$, represents the τth quantile of the random variable. The researcher defines the τ value, and assumes a regression structure on $q(\tau)$. Further, we provide structural properties such as ordinary and incomplete moments, and generating function. The maximum likelihood method is used for parameter estimation, and Monte Carlo simulations show that its properties remain. Closed-form expressions for score functions and observed information matrix are also derived. We adapt several diagnostic analysis and model selection techniques that can be employed to check the goodness-of-fit of the estimated model. Finally, the utility of the proposed regression is illustrated with an application to the dropout proportion of Brazilian undergraduate animal sciences courses. The fit of the UBXII regression is superior to the fit of the Kumaraswamy, unit-Weibull, and beta regressions since it provides better prediction performance to the data set at hand. Thus, the UBXII regression is an alternative quite competitive for modeling data that are restricted to the unit interval and can be used when the classical regressions are not suitable.

3 A NEW REGRESSION MODEL FOR THE COVID-19 MORTALITY RATES IN THE UNITED STATES

3.1 INTRODUCTION

Coronavirus disease-2019 (COVID-19), initially so-called 2019-nCoV, belongs to the coronavirus family that are enveloped positive-strand RNA viruses and that have the largest known RNA genomes. This illness infects several species of animals and also humans, causing respiratory tract infections, liver, neurological and gastrointestinal problems, and can range from mild to lethal (GUAN *et al.*, 2003). Its initial source was identified in Wuhan City, Hubei Province of China, in persons exposed to seafood and wet animal wholesale market. The first case was detected in December 2019 (COMMISSION *et al.*, 2019) and has quickly spread all over the world.

In the past two decades, the COVID-19 is the third coronavirus to emerge in the human population, likely characterizing a potentially more novel and severe infectious disease to be revealed. Due to the rapid spread and increase in the number of cases, there are evidence that it is more contagious than the severe acute respiratory syndrome coronavirus (SARS-CoV) and the Middle East respiratory syndrome coronavirus (MERS-CoV) outbreaks, which occurred in 2002 and 2012, respectively (HUANG *et al.*, 2020; MUNSTER *et al.*, 2020). Inclusive, since its similarity with the SARS-CoV, the COVID-19 is also named by SARS-CoV-2.

According to the World Health Organization (WHO) (World Health Organization, 2020b), the COVID-19 spreads through person-to-person transmission (direct contact), contaminated objects or surfaces (indirect contact), or close contact with infected people via mouth and nose secretions. Moreover, some studies show that on surfaces such as plastics and metals, this virus can survive for up to three days, and even in the air, it can survive for more than three hours (CARRATURO *et al.*, 2020; DOREMALEN *et al.*, 2020).

Effective treatment or preventive vaccines have yet not been developed for infections resulting from this virus (HUANG *et al.*, 2020). However, some preventative measures can be taken. The main prophylaxis strategies include social distancing, hand hygiene, and cleaning and disinfection of high-touch surfaces; see Gharpure *et al.* (2020) and Lewnard and Lo (2020), for more details. The WHO (World Health Organization, 2020b) highlights that peoples in close contact with an infected, around one meter, can catch COVID-19 from infectious droplets that may get into their mouth, nose, or eyes.

In April 2020, due to a large number of cases and deaths by the new coronavirus,

New York City has become the new epicenter of the disease in the United States of America (U.S.) (RADMANESH *et al.*, 2020), after Italy. Thenceforward, several other states have experienced a substantial increase in the number of cases and deaths. From January 20 to August 14, 2020, the total of confirmed cases passed five million in the country, being equal to $5,150,407$. In this same period was recorded $164,826$ deaths. Those numbers are equivalent to about 25% of the recorded cases total and 22% of deaths by coronavirus in the world (World Health Organization, 2020a).

Some recent studies present statistical applications to pandemic data in the U.S. Bashir *et al.* (2020) analyzed the correlation between the virus and climate indicators in New York City. They identified that the temperature and air quality are significantly associated with the coronavirus pandemic. Regressive and autoregressive spatial models were examined by Mollalo, Vahedi and Rivera (2020) in order to explain variations of coronavirus in the whole country, considering several environmental, topographic, socioeconomic, behavioral, and demographic factors as predictor variables. Other similar studies can be found in Andersen (2020) and Zhang and Schwartz (2020). However, to our best knowledge, a regression analysis modeling the mortality rate by coronavirus across the U.S. states has no been carried out.

In this context, some regressions are fitted to the coronavirus mortality rates in the 50 American states to determine the demographic, socioeconomic, behavioral, and meteorological explanatory variables (covariates) that affect these rates. Since the response variable has a restricted domain, a new parametric regression is constructed to fit these data. The new regression, based on a transformation on the Burr XII (BXII) random variable, is compared to the beta, simplex, Kumaraswamy (Kw), and unit-Weibull (UW) regressions, which are feasible alternatives to model this kind of data.

The rest of the chapter is structured as follows. In Section 3.2, the incidence of coronavirus disease in the U.S. is reviewed. A new regression to model the mortality rates in the American states is defined in Section 3.3. Further, the estimation of the parameters, a simulation study, and some goodness-of-fit measures to check the adequacy of the proposed regression are discussed. Section 3.4 contains some basic statistics of the data set, carries out the empirical analysis by identifying the best regression to fit the mortality rates, and provides some useful findings. Finally, in Section 3.5, some concluding remarks are addressed.

3.2 COVID-19 IN THE U.S.

The first COVID-19 case in the U.S. was recorded on January 21, 2020. The diagnosed with the coronavirus patient has traveled to Wuhan, China. In Japan, South Korea, and Thailand, the first cases also were reported one day prior (SCHUMAKER, 2020). Nine days after, it was confirmed in the U.S. the first case of person-to-person transmission by the Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention, 2020). On January 31, several events occurred in response to the COVID-19 outbreak, such as: i) the Coronavirus was declared a public health emergency in the U.S. by the Health and Human Services (HHS) and a federal quarantine for 14 days addressed to the 195 American evacuees from Wuhan was decreed; ii) U.S. Airlines suspended all flights between the U.S. and China temporarily; iii) President Donald Trump signed an order whose aim was to deny the entry of foreign nationals to American soil, who had traveled to China during the previous two weeks. On the other hand, on January 31, New York City health officials vehemently denied the rumor about a COVID-19 case in the city. However, on February 02, the mandatory 14 days quarantine was extended to the U.S. citizens, permanent residents, and immediate family who have visited China's Hubei province (Worldometer, 2020).

According to the CDC (Centers for Disease Control and Prevention, 2020) since March 1 until August 15, in the U.S. occurred two peaks in the hospitalization rates among all ages, being the first during the week ending April 18 (10.1 per 100 thousand population) and the second during the week ending July 18 (8.0 per 100 thousand).

From first known deaths recorded in February 2020, the total of confirmed cases had increased in all fifty U.S. states. However, each U.S. state has followed a different SARS-CoV-2 trajectory. During the pandemic's early days, the states of New York and Washington were hit hard. New York quickly became a new epicenter of the pandemic, recording 12,274 new coronavirus cases on April 4, which after one-day, was surpassed by the state of Florida that recorded 15,300 new cases (FREYTAS-TAMURA; ROJAS; FINK, 2020). Nevertheless, from the start of the U.S. pandemic until July 22, New York remained accounting the highest number of confirmed cases, being after was first surpassed by California and later by Florida and Texas (LEWIS, 2020). From this, although in a way more slowly, the total of cases and deaths have risen also in other states.

Until August 25, the five U.S. states with the highest reported death rates are New Jersey, New York, Massachusetts, Connecticut, and Louisiana with 180, 169, 130, 125, and 103 deaths per 100 thousand people (HERNANDEZ *et al.*, 2020). These states count with about 9%, 18%,

5%, 3%, and 3%, respectively, regarding the total of deaths in the country.

After the federal government has noted the threat posed by the coronavirus, on April 11, disaster declarations were approved for all states (GOOD, 2020). Moreover, measures as prohibitions and cancelation of large-scale gatherings, stay-at-home orders, and the closure of schools were taken as state and local responses to the pandemic (DEB; CACCIOLA; STEIN, 2020).

Beginning late April 2020, some U.S. states have begun to reopen their economies after the country into lockdown starting in March. Zhang *et al.* (2020) analyzed the setting of the number of confirmed cases, hospitalizations, and deaths for 11 countries and 40 American states after reopening their commercial activities. They evidenced that 75% of U.S. states increased the number of recorded cases, whereas 17.5% had observed an increase in the total of deaths after reopening. Similarly, according to the American Academy of Pediatrics and the Children's Hospital Association, the total of Coronavirus confirmed cases among children in the U.S. increased at 90% since the reopening of the American schools (CHAVEZ, 2020).

In August 2020, the restrictions on business and social activity have varied by state (Washington Post, 2020). For example, Alaska and West Virginia have not suffered any kind of restriction, contrary to California, which has suffered them more strictly. Thirty-two states present minor restrictions, whereas 15 states are classified at moderate restrictions category. Until effective vaccine and validated treatments against COVID-19 are developed, it is essential to maintain preventive strategies that favor avoiding the exposition to this virus to contain this disease's spread.

3.3 THE PROPOSED REGRESSION

This section aims to introduce a new regression that has much broader applicability in coronavirus mortality rates. This approach's particular feature is that it accommodates double-bounded variables in the unit interval with several types of asymmetry. The proposal is based on the transformation $Z = 1 - \mathrm{e}^{-X}$, where X is a BXII random variable having cumulative distribution function (cdf) and probability density function (pdf)

$$F_X(x; c, d) = 1 - (1 + x^c)^{-d}, \qquad x > 0,$$

and

$$f_X(x; c, d) = c\, d\, x^{c-1} (1 + x^c)^{-(d+1)},$$

respectively, where $c > 0$ and $d > 0$ are shape parameters. It is worth noting that Z can also be seen as a reflexive transformation on W, $Z = 1 - W$, where W is a random variable following a unit Burr XII (UBXII) distribution introduced in Chapter 2. Hence, the cdf and pdf of the *reflexive unit Burr XII (RUBXII) distribution* can be expressed as (for $z \in (0,1)$)

$$F_Z(z;c,d) = 1 - [1 + \log^c(1-z)^{-1}]^{-d}, \tag{3.1}$$

and

$$f_Z(z;c,d) = c\,d\,\frac{(z-1)^{-1}\log^{c-1}(1-z)^{-1}}{[1+\log^c(1-z)^{-1}]^{d+1}}, \tag{3.2}$$

respectively. By inverting (3.1), the quantile function (qf) of Z is

$$Q_Z(u;c,d) = 1 - \exp\left\{-[(1-u)^{-1/d} - 1]^{1/c}\right\}. \tag{3.3}$$

Both the UBXII and RUBXII distributions are special cases of the unit extended Weibull family; see Guerra *et al.* (2020).

In order to introducing a systematic component on a location parameter, the RUBXII distribution is reparameterized in terms of its quantiles. Let $q(\tau) = Q_Z(\tau;c,d)$ be the τth quantile of Z. By evaluating Equation (3.3) in τ and inverting for d,

$$d = \log(1-\tau)^{-1}/\log\left\{1+\log^c[1-q(\tau)]^{-1}\right\}. \tag{3.4}$$

Notwithstanding the quantiles are functions of τ, $q(\tau)$ is just denoted as q to simplify the notation. Then, by replacing (3.4) in Equations (3.1) and (3.2), the cdf and pdf of the RUBXII distribution expressed in terms of a quantile-based parameterization are (for $(y \in (0,1))$)

$$F_Z(z;q,c) = 1 - \left[1+\log^c(1-z)^{-1}\right]^{\frac{\log(1-\tau)}{\log\left[1+\log^c(1-q)^{-1}\right]}}, \tag{3.5}$$

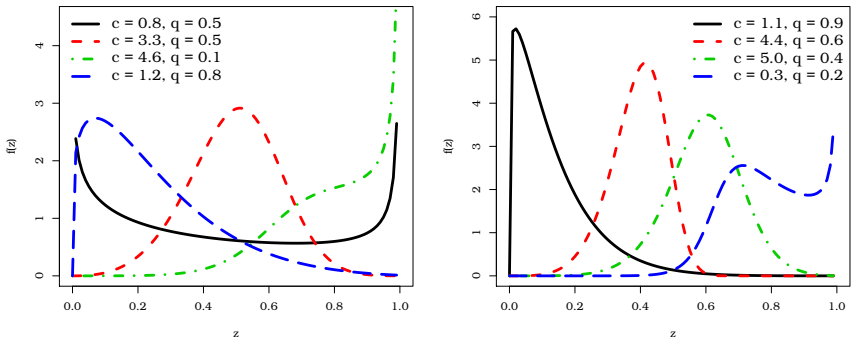
and

$$f_Z(z;q,c) = \frac{\log(1-\tau)^{-c}\log^{c-1}(1-z)^{-1}}{(1-z)\log\left[1+\log^c(1-q)^{-1}\right]}\left[1+\log^c(1-z)^{-1}\right]^{\frac{\log(1-\tau)}{\log\left[1+\log^c(1-q)^{-1}\right]}-1}, \tag{3.6}$$

respectively, where $c > 0$ is a shape parameter and the quantile order $\tau \in (0,1)$ is chosen by the researcher. For example, in the coronavirus mortality rates application of Section 3.4, $\tau = 0.5$, and therefore $q = q(0.5)$ is the median of Z. Henceforth, let $Z \sim \text{RUBXII}(q,c)$ be a random variable having density (3.6).

Figure 7 displays plots of the RUBXII density with $\tau = 0.5$, which have the following forms: U, symmetric, right-skewed, increasing, and increasing-decreasing-increasing. Thus, it is useful for modeling variables with different types of skewness and heavy tails.

Figure 7 – Plots of the RUBXII density ($\tau = 0.5$).



Source: Author (2020)

On the proposed reparameterization, the qf of Z is

$$Q_Z(u) = 1 - \exp\left\{-\left[(1-u)^{\log[1+\log^c(1-q)^{-1}]/\log(1-\tau)} - 1\right]^{1/c}\right\}. \tag{3.7}$$

It is useful to generate observation from the RUBXII distribution by the inversion method since it has a closed-form. So, if U is a random variable having a standard uniform distribution, then $Z = Q_Z(U)$ follows the RUBXII law.

Let $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ be a vector of n independent observations of the variables $Z_i \sim$ RUBXII(q_i, c) (for $i = 1, \ldots, n$). The new regression is proposed assuming that the parameters q_i can be expressed as a function of covariates under the systematic component

$$g(q_i) = \eta_i = \sum_{j=1}^{k} x_{ij}\,\xi_j = \boldsymbol{x}_i^\top \boldsymbol{\xi}, \tag{3.8}$$

where $g : (0,1) \to \mathbb{R}$ is a strictly monotonic and twice differentiable link function, η_i is the linear predictor, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^\top$ is the parameter vector associated with the covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^\top$. The quantities q_i can be obtained by inverting (3.8) as $q_i = g^{-1}(\eta_i)$.

Several link functions can be chosen for $g(\cdot)$ such as the logit, probit, and complementary log-log. In applications, the logit link function is generally considered due to the useful interpretation of the regression coefficients as an odds ratio. It is defined as $g(p) = \log[p/(1-p)]$, and it is used in all fitted regressions in this chapter.

3.3.1 Estimation

The estimation of the parameters that index the RUBXII regression is done by the maximum likelihood (ML) method. Let $\boldsymbol{\theta} = (\boldsymbol{\xi}^\top, c)^\top$ be the $(k+1)$-dimensional parameter vector.

The log-likelihood function based on a sample of n independent observations is

$$\ell(\boldsymbol{\theta}) \equiv \ell(\boldsymbol{\xi}, c) = \sum_{i=1}^{n} \ell_i(q_i, c), \tag{3.9}$$

where q_i satisfies the systematic component (3.8) and $\ell_i(q_i, c)$ is the logarithm of the density $f_Z(z_i; q_i, c)$ given in Equation (3.6), i.e.,

$$\begin{aligned}\ell_i(q_i, c) = &\log(1-z_i)^{-1} - \log\left[\, r(q_i)\,\right] + \log\left[\, \log(1-\tau)^{-c}\,\right] + \log\left[\, \log^{c-1}(1-z_i)^{-1}\right] \\ &+ \left[\log(1-\tau)/r(q_i) - 1\right] r(z_i),\end{aligned}$$

and $r(x) = \log\left[\, 1 + \log^{c}(1-x)^{-1}\right]$.

The components of the score vector $U(\boldsymbol{\theta})$, given in Appendix B, are defined as the partial derivatives of (3.9) with respect to each element of the parameter vector $\boldsymbol{\theta}$. Equalizing its components to zero, $U(\boldsymbol{\theta}) = \mathbf{0}$, and solving the system simultaneously, the maximum likelihood estimators (MLEs) $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\xi}}^{\top}, \hat{c})^{\top}$ of $\boldsymbol{\theta}$ can be found. However, the system of equations is non-linear and cannot be solved analytically. In such a way, the estimators must be obtained through numerical optimization algorithms using well-known programming languages such as the `R` (`optim` function), `SAS` (`PROC NLMIXED`), and `Ox` program (`MaxBFGS` sub-routine).

3.3.2 Simulation study

Some Monte Carlo experiments are carried out to assess the performance of the MLEs on the finite sample. Consider the systematic component for q_i:

$$\log\left(\frac{q_i}{1-q_i}\right) = \eta_i = \xi_1 + \xi_2 x_{i2}, \qquad i = 1, \ldots, n.$$

Four scenarios with different simulation schemes, combining various values for the parameter vector $\boldsymbol{\theta} = (\xi_1, \xi_2, c)^{\top}$, are considered. To evaluate the performance of the MLEs, for each scenario, the samples $\{(z_1, x_{12}), \ldots, (z_n, x_{n2})\}$ are simulated 10,000 times with $n \in \{30, 90, 160, 300\}$. The occurrences of the response $Z_i \sim \text{RUBXII}(q_i, c)$ are obtained by the inversion method through the qf in Equation (3.7). The covariate x_{i2} is generated from a uniform distribution on the interval $(-3, 3)$ (scenarios 1 and 2), and a standard normal distribution (scenarios 3 and 4). The `R` programming language (R Core Team, 2020) is used to perform the simulation study.

The percentage relative bias (RB%) and root mean squared error (RMSE) of the estimates in $\boldsymbol{\theta}$ are determined. Table 6 lists the results for these measures. Low RB values are noted even for small sample sizes. Considering all the scenarios and sample sizes, the RBs of the estimates

of ξ_1 and ξ_2 are less than 4%, and those of c are less than 15%. On the other hand, the RMSE quickly goes to zero when n increases, thus in agreement with the asymptotic properties of the MLEs.

Table 6 – Simulation results from the RUBXII regression.

Scenario	ξ_1	ξ_2	c	n	RB%			RMSE		
					$\hat{\xi}_1$	$\hat{\xi}_2$	$\hat{c}$	$\hat{\xi}_1$	$\hat{\xi}_2$	$\hat{c}$
1	−1.6	1.2	2.3	30	−0.0122	0.4027	7.6418	0.1293	0.0753	0.4343
				90	0.0998	−0.1007	2.4591	0.0757	0.0431	0.2124
				160	0.1782	−0.1204	1.3935	0.0551	0.0336	0.1546
				300	0.1585	−0.1431	0.7695	0.0422	0.0234	0.1098
2	2.5	3.1	3.2	30	−2.9889	−0.9241	13.3777	0.3581	0.1647	0.8805
				90	−2.4068	−0.9072	4.0566	0.1829	0.0874	0.4272
				160	−2.6012	−0.9774	1.6314	0.1445	0.0689	0.2888
				300	−2.7385	−1.0250	0.6371	0.1180	0.0552	0.2042
3	−0.5	−2.8	3.2	30	−3.2219	−0.4907	14.7059	0.2350	0.1103	1.1643
				80	−2.6410	−1.2360	4.6612	0.1528	0.1263	0.6492
				160	−3.9155	−1.9756	1.6002	0.1217	0.0922	0.3960
				300	−4.3493	−2.5438	0.3910	0.1031	0.0878	0.2848
4	1.6	2.3	4.2	30	0.4497	0.1082	8.0971	0.1273	0.1096	0.6221
				90	1.6508	−0.2845	2.7237	0.0731	0.0897	0.3342
				160	1.3309	−0.1371	1.4408	0.0558	0.0512	0.2522
				300	1.8395	−0.2971	0.7331	0.0418	0.0373	0.1709

Source: Author (2020)

3.3.3 Regression model adequacy

In this section, some methods are presented to analyze whether a fitted regression is suitable for a data set. As diagnostic analysis tools of the RUBXII regression, the randomized quantile residuals (qrs) (DUNN; SMYTH, 1996), the generalized pseudo-R^2 (R^2_G), a RESET-type test, and a information criterion are discussed. A cross-validation approach is adopted to assess the prediction performance of the proposed regression and it is compared with other suitable regressions for proportional data.

The randomized qrs for the RUBXII regression are

$$\boldsymbol{r} = \Phi^{-1}[F_Z(\boldsymbol{z}; \hat{\boldsymbol{q}}, \hat{c})],$$

where $F_Z(\cdot)$ is the cdf of the RUBXII distribution given in Equation (3.5) and $\Phi^{-1}(\cdot)$ is the qf of the standard normal distribution. If the fit is adequate, it is expected that the distribution of the qrs be close to the standard normal. To check whether this assumption is satisfied, the well-known Anderson-Darling (AD) test (STEPHENS, 1974) can be performed. The null hypothesis for this test is that the errors are normally distributed. In R, the AD test is implemented in the `nortest` package.

The R^2_G is useful to assess the proportion of the response variable's variation which may be explained by the regression instead of the simple model. It is defined by Nagelkerke *et al.* (1991) as

$$R^2_G = 1 - \exp\left\{-2/n\left[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)\right]\right\},$$

where $\ell(\hat{\boldsymbol{\theta}}_0)$ is the log-likelihood for the null model, i.e., from modeling the response without covariates, and $\ell(\hat{\boldsymbol{\theta}})$ is the log-likelihood of the fitted regression. A regression with a higher value of R^2_G provides a larger explanation power of the response variable's variation.

A RESET-type test introduced by Pereira and Cribari-Neto (2014) can be adopted to detect possible specification errors in the regression. The null hypothesis of this test is that the regression is correctly specified. It may be carried out in the following way: i) fit the regression and obtain the fitted values $\hat{\boldsymbol{q}} = (\hat{q}_1, \ldots, \hat{q}_n)^\top$ of $\boldsymbol{q} = (q_1, \ldots, q_n)^\top$ using (3.8); ii) compute powers of second and third degrees of $\hat{\boldsymbol{q}}$, i.e., get $\hat{\boldsymbol{q}}^2 = (\hat{q}_1^2, \ldots, \hat{q}_n^2)^\top$ and $\hat{\boldsymbol{q}}^3 = (\hat{q}_1^3, \ldots, \hat{q}_n^3)^\top$; and iii) using these powers as additional covariates, fit the augmented regression, and test the significance of them through the likelihood ratio (LR) test.

The LR statistic is $\kappa = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})]$, where $\ell(\hat{\boldsymbol{\theta}})$ and $\ell(\tilde{\boldsymbol{\theta}})$ are the unrestricted and restricted maximized log-likelihood functions, respectively. Under the null hypothesis, κ converges in distribution to a chi-squared with ν degree of freedom, where ν is the number of added test variables ($\nu = 2$ in this case).

The Akaike information criterion (AIC) defined as $\mathrm{AIC} = -2\ell(\boldsymbol{\theta}) + 2\,\phi$ is commonly used for comparing regressions, where ϕ is the number of parameters of the fitted regression (AKAIKE, 1973). The regression that provides the smallest value of AIC is selected. The goodness-of-fit performance of non-nested regressions can be assessed using the leave-one-out cross-validation (LOOCV) approach; see James *et al.* (2013) for details. It involves sequential splitting the data set into two parts. Let $\{(z_1, \boldsymbol{x}_1), \ldots, (z_n, \boldsymbol{x}_n)\}$ be the set of the response Z and its associated vectors of k covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^\top$ (for $i = 1, \ldots, n$). Further, let $\hat{z}_{-i}$ be the estimated value of z_i, excluding the ith observation in the fit, i.e., fitting a regression using

$$\{(z_1, \boldsymbol{x}_1), \ldots, (z_{i-1}, \boldsymbol{x}_{i-1}), \ldots, (z_{i+1}, \boldsymbol{x}_{i+1}), \ldots, (z_n, \boldsymbol{x}_n)\},$$

and then substituting $\boldsymbol{x}_i$ into the fitted regression structure to obtain an estimate of z_i ($\hat{z}_{-i}$).

The mean absolute error (MAE) defined as $\text{MAE} = 1/n \sum_{i=1}^{n} |z_i - \hat{z}_{-i}|$ is taken as a measure of the accuracy of the regression. The lower is the MAE value, the better is the prediction provided by the regression. All techniques discussed in this section can be extended analogously to other regression models as will be addressed in Section 3.4.

3.4 RESULTS AND DISCUSSION

In approximately eight months of the coronavirus advance since its inception, on August 19, 2020 in the U.S., the CDC reported a total of 5,650,176 confirmed cases and 175,789 deaths, putting the disease with 3.1% lethality (Centers for Disease Control and Prevention, 2020). Also, the adoption of systematic non-pharmaceutical interventions seems to have led to a decrease in mortality. Thus, understanding the relationship between demographic, socioeconomic, behavioral, and meteorological variables with the mortality rate became a crucial task. In this sense, this section presents the RUBXII regression's application, concurrently with four other well-known regression models, thus associating the mortality rate with relevant demographic, socioeconomic, behavioral, and meteorological variables.

The amount of information available on the disease is as abundant as it is scattered and unreliable. Therefore, before the analysis, a data mining is built to construct a database described at the beginning of the section, informing the sources from which they are obtained. The regression models chosen in this study consider an essential characteristic of the mortality rate that it belongs to the interval $(0, 1)$.

3.4.1 Descriptive statistical analysis

The response variable is the COVID-19 deaths rate in the U.S. states. This rate is calculated in the 50 states from data available by the Institute for Health Metrics and Evaluation (IHME) (2020). For all states, it is considered the total of deaths per hundred people on 30, 60, 90, and 120 days after the 20th detected case to ensure that the comparisons are made to the same period. In this way, a panel with four observations for each state is structured.

For all states, the population density, Human Development Index, Gini coefficient, hospital

beds, poverty rate, smoking rate, average temperature, and median age, are obtained from the following sources: World Population Review, Global Data Lab, World Atlas, Kaiser Family Foundation, and Iowa Community Indicators Program of the Iowa State University. The response variable and covariates are defined below:

1. MR[1]: Mortality rate (response variable).
2. PD[2]: Population density (p/mi^2).
3. HDI[3]: Human Development Index.
4. GINI[4]: Gini coefficient.
5. BEDS[5]: Hospital beds per 100 thousand inhabitants.
6. PR[6]: Poverty rate (data of 2020).
7. SR[7]: Smoking rate by state (data of 2020).
8. AT[8]: Average temperature (measured in degrees Fahrenheit, ^{o}F).
9. MA[9]: Median age (data of 2020).
10. T_{60}: dummy that is equal to one if the response observation corresponds to mortality rate after 60 days of the 20th confirmed case, and zero otherwise.
11. T_{90}: dummy that is equal to one if the response observation corresponds to mortality rate after 90 days of the 20th confirmed case, and zero otherwise.
12. T_{120}: dummy that is equal to one if the response observation corresponds to mortality rate after 120 days of the 20th confirmed case, and zero otherwise.

Table 7 gives some descriptive measures of these variables. The MR has a high coefficient of variation (CV) for all current time periods, being the most at 60 days with a CV of about 141%. Also, in the four time periods (30, 60, 90, and 120 days), the response presents positive skewness, the mean is not close to the median, and its kurtosis is greater than three indicating that it has a leptokurtic distribution. The HDI, GINI, and MA covariates have the lowest variabilities with CV ranging between about 2% and 6%. On the other hand, the PD covariate has the most CV about at 130% and takes values on a sizeable range since the minimum and maximum are around

[1] It is built from available data at <https://covid19.healthdata.org/united-states-of-america>
[2] <https://worldpopulationreview.com/states>
[3] <https://globaldatalab.org/shdi/shdi/USA/?levels=1%2B4&interpolation=0&extrapolation=0&nearest_real=0&years=2018>
[4] <https://www.worldatlas.com/articles/us-states-by-gini-coefficient.html>
[5] <https://www.kff.org/other/state-indicator/beds-by-ownership/?currentTimeframe=0&selectedDistributions=total&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>
[6] <https://worldpopulationreview.com/state-rankings/poverty-rate-by-state>
[7] <https://worldpopulationreview.com/state-rankings/smoking-rates-by-state>
[8] <https://worldpopulationreview.com/state-rankings/average-temperatures-\by-state>
[9] <https://worldpopulationreview.com/state-rankings/median-age-by-state>

1p/mi^2 (referring to the Alaska state) and 1,215p/mi^2, respectively. The BEDS, PR, SR, and AT covariates have close CVs varying from around 16% to 27%. Moreover, they have a mean close to the median, and kurtosis lower than three. Only the HDI, AT, and MA covariates have negative skewness.

Table 7 – Descriptive statistics

Variable	Statistics						
	Mean	Median	Skewness	Kurtosis	Min.	Max.	CV(%)
MR (30)	0.0044	0.0024	2.3833	4.9741	0.0003	0.0257	131.8649
MR (60)	0.0198	0.0092	2.4975	5.9940	0.0012	0.1299	141.4140
MR (90)	0.0289	0.0145	2.1803	4.3017	0.0012	0.1593	126.7096
MR (120)	0.0340	0.0199	2.0824	3.9428	0.0016	0.1739	114.7390
PD	203.9010	107.7835	2.2166	4.5102	1.2863	1,215.1980	130.1561
HDI	0.9178	0.9220	−0.5294	−0.5138	0.8630	0.9560	2.4331
GINI	0.4522	0.4530	0.1342	−0.4728	0.4190	0.4990	3.9132
BEDS	2.6000	2.4500	0.9717	0.5761	1.6000	4.8000	27.2756
PR	0.1323	0.1322	0.4566	−0.3789	0.0762	0.2007	21.3061
SR	0.1733	0.1715	0.2748	−0.1011	0.0890	0.2600	20.0359
AT	51.9460	51.2000	−0.0132	0.1874	26.6000	70.7000	16.6352
MA	38.3240	38.3000	−0.1342	1.5281	30.7000	44.6000	6.1734

Source: Author (2020)

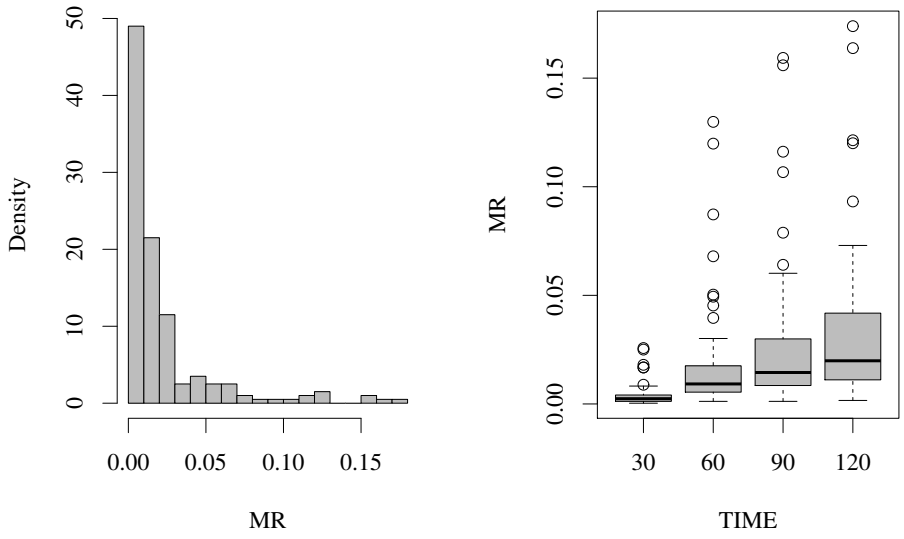
Figure 8 displays the histogram of the MR and box plots from four panel's observations, i.e., MR for 30, 60, 90, and 120 days. The histogram and the four box plots agree to those figures in Table 7. The MR on 30, 60, 90, and 120 days have skewed-right distribution, and it presents quite outliers. Clearly, after 60, 90, and 120 days of the 20th recorded case, the mortality rate has increased substantially according to the box plots.

3.4.1.1 Correlation analysis

The correlation matrix for the current variables is displayed in Figure 9. All correlations are computed by the Spearman method. The response variable is positively correlated to PD which is the most correlation value among the MR and other covariates.

Figure 10 displays dispersion plots of the MR versus each covariate. It can be noted that there is no indication of a linear relationship among them. In this way, to study the significance of the correlations provided in Figure 9, it is carried out a Spearman correlation test and a non-parametric analysis. The null hypothesis ($\mathcal{H}_0$) of this test is that the populational correlation coefficient between two variables is equal to zero, i.e., there is no statistically significant correlation. Under $\mathcal{H}_0$, the computed test statistic converges in distribution to a Student's t-distribution with

Figure 8 – Histogram of the MR and box plots of the MR after 30, 60, 90, and 120 days after the 20th confirmed case.



Source: Author (2020)

$(n-2)$ degrees of freedom, where n is the sample size. The p-values of the test are given in Table 8. In a first analysis, note that there is a statistically significant correlation among the mortality rate and the covariates PD, GINI, and MA.

Table 8 – p-values of the Spearman correlation test between all variables.

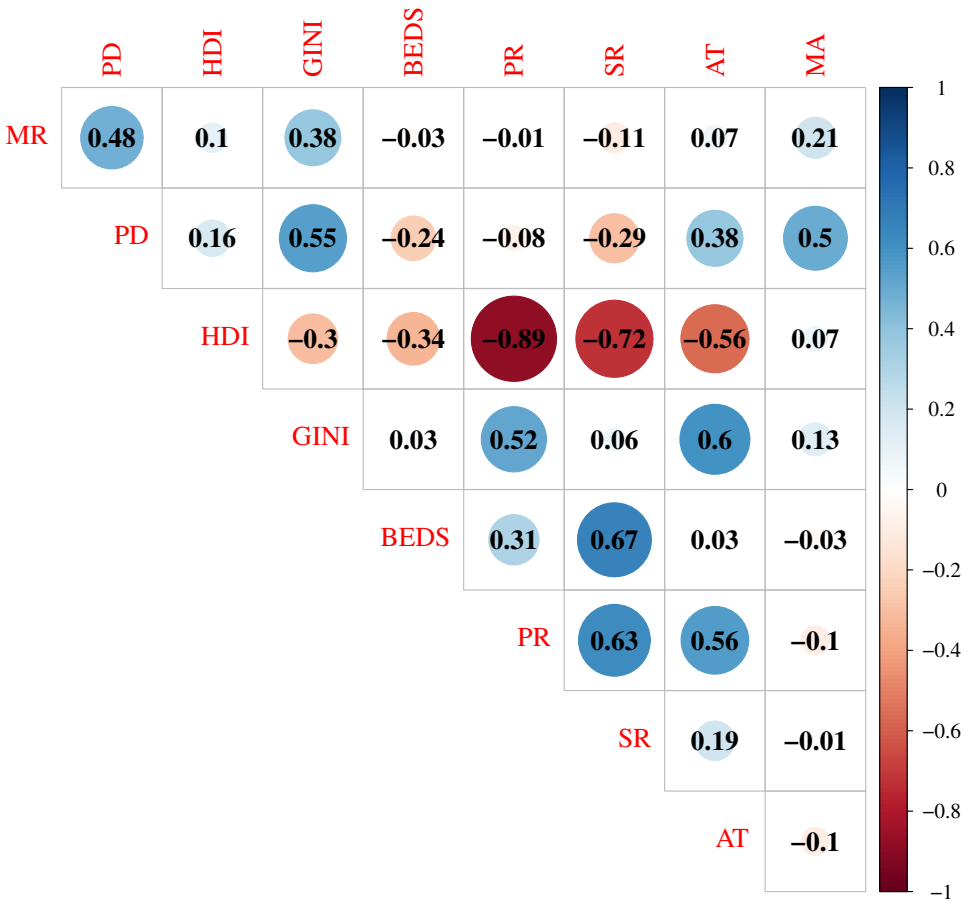
Variables	MR	PD	HDI	GINI	BEDS	PR	SR	AT	MA
MR		< 0.0001	0.1390	< 0.0001	0.6555	0.8885	0.1259	0.3104	0.0036
PD			0.0215	< 0.0001	0.0005	0.2868	< 0.0001	< 0.0001	< 0.0001
HDI				< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
GINI					0.6394	< 0.0001	0.3871	< 0.0001	0.0591
BEDS						< 0.0001	< 0.0001	0.7085	0.6714
PR							< 0.0001	< 0.0001	0.1638
SR								0.0069	0.8619
AT									0.1732
MA									

Source: Author (2020)

3.4.2 Fitted regressions

It is explored more deeply the relationship between covariates and the MR through regression analysis. The prediction performance and goodness-of-fit measures are investigated for the RUBXII regression defined in Section 3.3 with four competitive systematic components to study the effects of some covariates on the mortality rate by coronavirus in the U.S. states. The beta, simplex, and Kw regressions are considered, all of them well-known in the literature, and

Figure 9 – Correlation matrix



Source: Author (2020)

the UW quantile regression recently introduced Mazucheli *et al.* (2020).

The beta regression (FERRARI; CRIBARI-NETO, 2004) is typically the most used to study proportional data. Its density is quite flexible since it has many shapes able to accommodating several types of asymmetry. Some applications of this class of regression can be found in Cribari-Neto and Souza (2013) and Ghosh (2019).

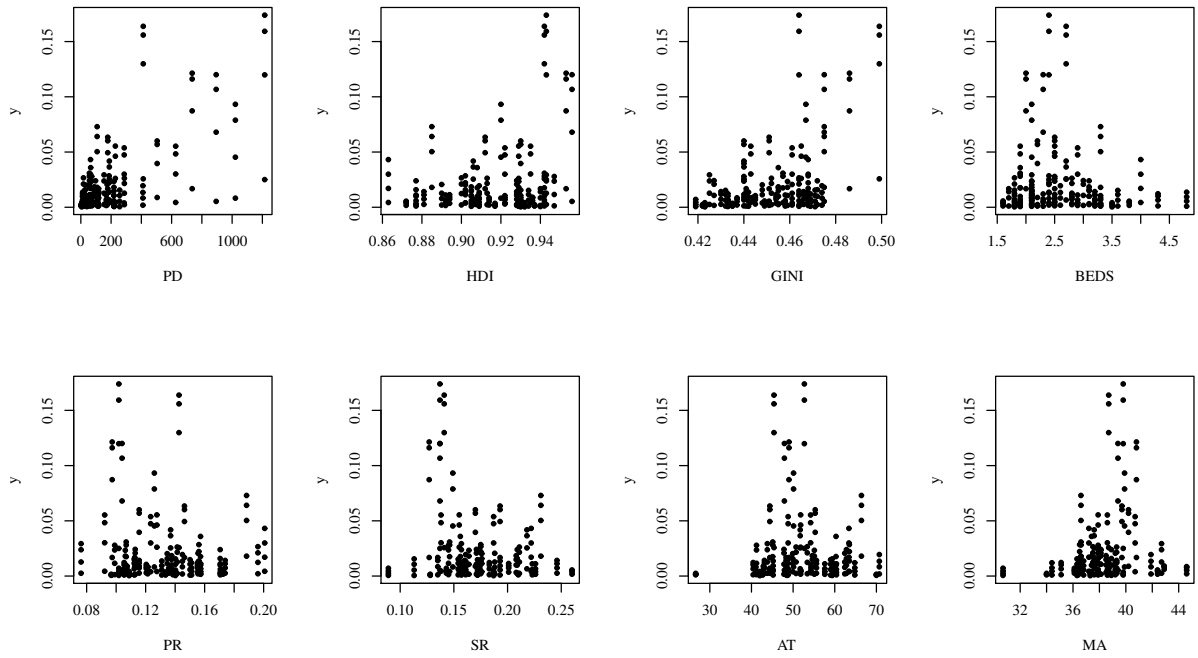
For the beta distribution, it is adopted the parameterization based on the mean parameter μ and dispersion parameter σ. Let $Y \sim \text{Beta}(\mu, \sigma)$ be a random variable having the beta density (for $y \in (0,1)$)

$$f(y;\mu,\sigma) = \frac{\Gamma\left(1/\sigma^2 - 1\right)}{\Gamma\left(\mu(1/\sigma^2-1)\right)\Gamma\left((1-\mu)(1/\sigma^2-1)\right)} y^{\mu(1/\sigma^2-1)-1}(1-y)^{(1-\mu)(1/\sigma^2-1)-1}, \quad (3.10)$$

where $0 < \mu < 1$ is the mean of Y, $0 < \sigma < 1$ is a dispersion parameter and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x$ is the complete gamma function. Under this parameterization the variance of Y is $\sigma^2\mu(1-\mu)$.

Another know regression for response in the unit interval with a systematic component in the mean is the simplex regression defined from Barndorff-Nielsen and Jørgensen (1991). If

Figure 10 – Dispersion plots



Source: Author (2020)

$Y \sim S^{-}(\mu, \sigma^2)$ is a random variable following a simplex distribution, its density is (for $y \in (0,1)$)

$$f(y;\mu,\sigma^2) = \{2\pi\sigma^2[y(1-y)]^3\}^{-1/2}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2\mu^2 y(1-y)(1-\mu)^2}\right\}, \tag{3.11}$$

where $0 < \mu < 1$ is the mean of Y and σ^2 is a dispersion parameter. Further, the unit variance function is $V(\mu) = \mu^3(1-\mu)^3$.

A classical alternative to the beta and simplex regressions is the Kw regression. Its main advantage over the beta and simplex regressions is that it is parameterized in terms of the median. Median-based regressions are more robust against the presence of atypical observations in the response and asymmetries than mean-based (PUMI; RAUBER; BAYER, 2020). Let Y be a random variable that follows a Kw distribution on median-dispersion parameterization proposed by (MITNIK; BAEK, 2013), say $Y \sim \text{Kw}(\omega, d_p)$, with pdf (for $y \in (0,1)$)

$$f(y;\omega,d_p) = \frac{\log 0.5}{d_p \log(1-\omega^{1/d_p})} y^{1/d_p}(1-y^{1/d_p})^{\log 0.5/\log(1-\omega^{1/d_p})-1}, \tag{3.12}$$

where $0 < \omega < 1$ is the median of Y and $d_p > 0$ is a dispersion parameter.

Recently, Mazucheli *et al.* (2020) proposed the UW quantile regression. Let $Y \sim \text{UW}(\mu, \beta)$ be a random variable having the UW density (under the parameterization given in Mazucheli *et al.* (2020)) (for $y \in (0,1)$)

$$f(y;q,\beta) = \frac{\beta}{y}\left(\frac{\log\tau}{\log q}\right)\left(\frac{\log y}{\log q}\right)^{\beta-1}\tau^{(\log y/\log q)^{\beta}}, \tag{3.13}$$

where $0 < q < 1$ is the τth quantile and $\tau \in (0, 1)$ is assumed known. Here, it will be considered that $\tau = 0.5$ in order to model the median of Y.

The four models discussed in this section have a systematic component analogous to that one given in Equation (3.8). They differ only on the assumption of the response distribution and the location parameter. The beta and simplex regressions are parameterized in the mean and defined by replacing q_i by μ_i in Equation (3.8), whereas their random component are given by Equations (3.10) and (3.11), respectively. The beta and simplex regressions are implemented in `R` in the `gamlss` package (RIGBY; STASINOPOULOS, 2005). In a different way, the Kw and UW (when $\tau = 0.5$) regressions have a median-based parametrization, and their associated systematic components are obtained by evaluating Equation (3.8) in ω_i and q_i, respectively. Besides, their random component follow of (3.12) and (3.13), respectively.

The goodness-of-fit measures of the final fitted regressions are reported in Table 9. It is adopted the significance of the estimates as a criterion to choose the variables in the final fits. The RUBXII and Kw regressions are the most competitive to explain the MR, since they have the best adequacy measures. However, the MAE and AIC values for the RUBXII regression are the lowest, showing its superiority in terms of model fit to the current data. The difference of 0.0001 among the MAE values of the RUBXII and Kw regressions is substantial given the scale of the MR as indicated in the histogram in Figure 8. The quotient between the MAE and the mean of the MR ($\overline{\text{MR}}$) is calculated to counteract this effect and obtain a quantity independent of the unit of measurement. The results of the ratio MAE/$\overline{\text{MR}}$ favor to the RUBXII more clearly than those MAE values regards its prediction performance. Further, its R^2_G is the greatest indicating that the fitted RUBXII regression explains 74.05% of the variability of the median response. The p-value of the AD test for the simplex regression's residuals is the only one among the five models lower than 0.05. So, the null hypothesis that the errors' distribution is normal is rejected at a significance level of 5%, and then this regression is not adequate to the current data. On the other hand, according to the p-values of the RESET-type (RES) tests, all fitted regressions are specified correctly at a significance level of 5%.

Table 10 gives the estimates of the parameters, their standard errors, and p-values of the final fitted RUBXII and Kw regressions since they provide the best fits to the coronavirus death rates across the U.S. states according to the adequacy measures in Table 9. For the fitted RUBXII regression, most of the covariates are significant at a significance level of 1%, except for the HDI and BEDS, which are significant at the 5% and 10%, respectively. Otherwise, HDI, BEDS, and PR are not statistically significant in the fitted Kw regression. The remaining covariates are the

Table 9 – Goodness-of-fit measures for the final fitted regressions.

Regression	AIC	R^2_G	p-val (AD)	p-val (RES)	MAE	MAE/$\overline{\text{MR}}$
RUBXII(q_i, c)	$-1,400.6280$	0.7405	0.1055	0.9999	0.0101	0.4652
Kw(ω_i, d_p)	$-1,396.6400$	0.7325	0.0521	0.5484	0.0102	0.4679
Beta(μ_i, σ)	$-1,332.0810$	0.6471	0.3467	0.4778	0.0112	0.5134
UW(q_i, β)	$-1,271.8390$	0.4079	0.1078	0.9999	0.0121	0.5551
Simplex(μ_i, σ^2)	$-1,281.9750$	0.4733	< 0.0001	0.9799	0.0160	0.7340

Source: Author (2020)

same as the RUBXII regression, and all are significant at the level of 1%. The fitted UW quantile and simplex regressions (final models) are addressed in Appendix B; see Table 18.

Table 10 – Fitted regressions for the median of the MR by COVID-19 in the U.S. states.

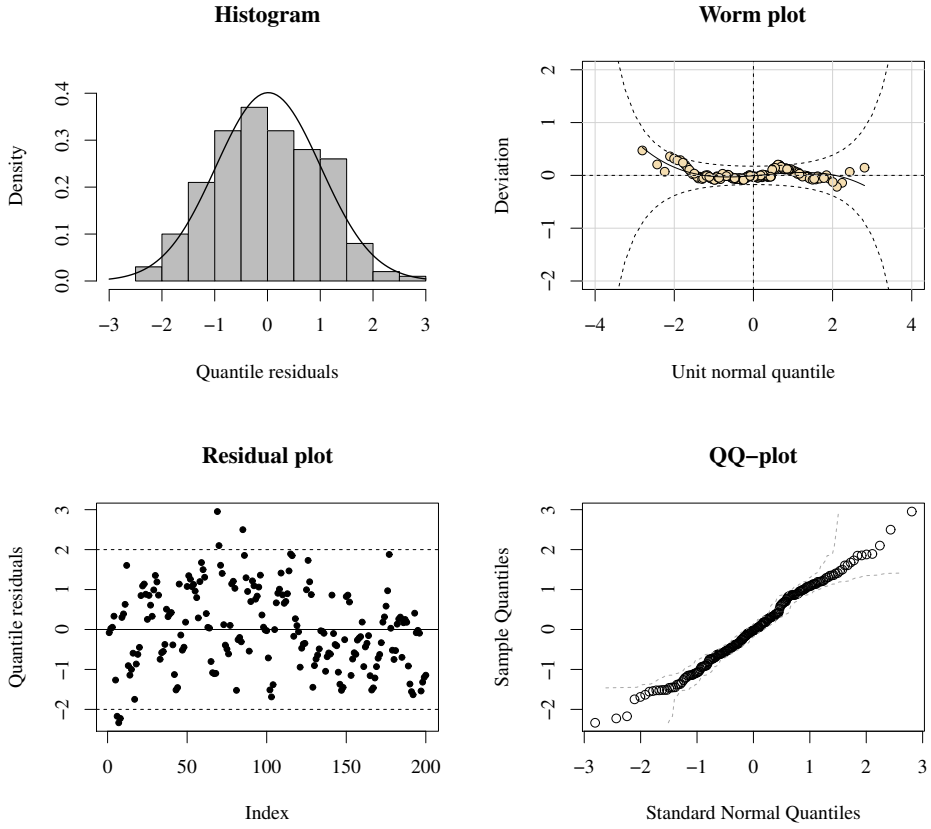
RUBXII(q_i, c)				Kw(ω_i, d_p)			
Coeff.	Estimate	Std. Error	p-value	Coeff.	Estimate	Std. Error	p-value
Int.	-30.1410	6.3543	< 0.0001	Int.	-17.4302	1.2385	< 0.0001
PD	0.0025	0.0002	< 0.0001	PD	0.0025	0.0002	< 0.0001
HDI	17.3367	7.0366	0.0146	GINI	25.6770	3.1104	< 0.0001
GINI	13.4863	4.4055	0.0025	SR	7.6878	1.7282	< 0.0001
BEDS	-0.1673	0.0946	0.0788	AT	-0.0345	0.0073	< 0.0001
PR	14.8157	5.1112	0.0042	T_{60}	1.3236	0.1302	< 0.0001
SR	10.7263	2.6557	0.0001	T_{90}	1.7687	0.1309	< 0.0001
AT	-0.0299	0.0091	0.0013	T_{120}	2.0033	0.1319	< 0.0001
T_{60}	1.3426	0.1254	< 0.0001	d_p	0.6416	0.0357	–
T_{90}	1.7952	0.1264	< 0.0001	–	–	–	–
T_{120}	2.0243	0.1270	< 0.0001	–	–	–	–
c	1.6052	0.0917	–	–	–	–	–

Source: Author (2020)

The residuals from the fitted RUBXII and Kw regressions are now analyzed graphically. Figure 11 provides some residuals plots for the RUBXII regression. The four plots indicate that this fitted regression is suitable. According to the histogram, the residuals have distribution quite close to the standard normal. All points are inside of the confidence bands close to the central line in the worm plot without any trend; see Buuren and Fredriks (2001). In the qrs against the index plot (residual plot), they appear to be randomly scattered around zero. The sample quantiles are within the confidence bands of the quantile-quantile plot (QQ-plot). Analogously, Figure 12 displays residual plots for the fitted Kw regression. The four plots indicate that this fitted regression is suitable, although it is possible to note a lack-of-fit of the residuals to the

standard normal distribution in the histogram. Some substantial departures from the confidence bands and a superior-left-trend can be seen in the worm plot.

Figure 11 – Residuals plots for the fitted RUBXII regression.



Source: Author (2020)

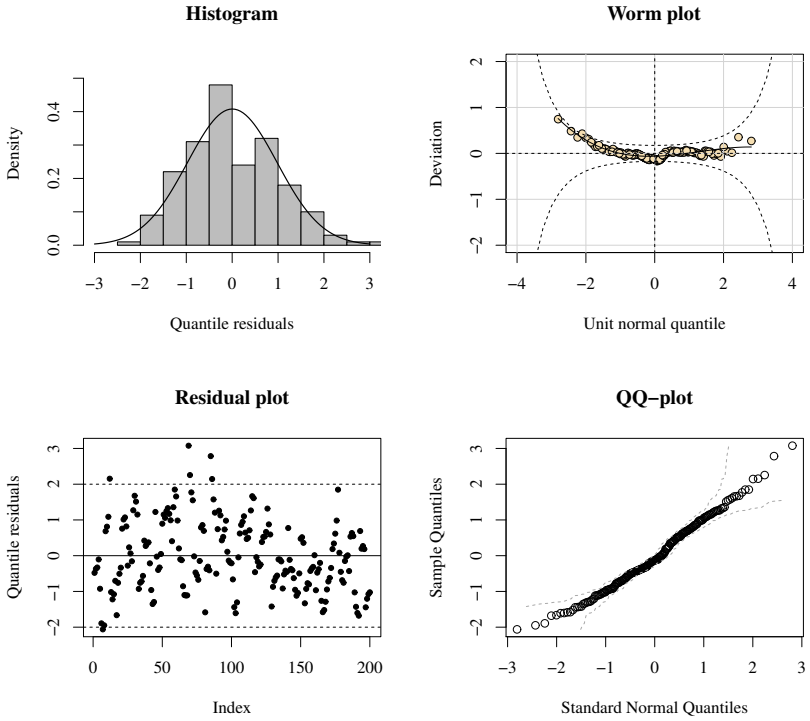
After the above analysis, there are evidence that the RUBXII regression really provides a better quality fit. The smallest value for the considered statistic in the LOOCV approach (see Table 9) indicates that its predictions are more accurate than those of the Kw regression. Therefore, from the estimates of the parameters of the RUBXII regression reported in Table 10, its regression equation can be expressed as

$$\begin{aligned}\log\left[\hat{q}_i/(1-\hat{q}_i)\right] = &-30.1410 + 0.0025\,\mathrm{PD}_i + 17.3367\,\mathrm{HDI}_i + 13.4863\,\mathrm{GINI}_i - 0.1673\,\mathrm{BEDS}_i \\ &+ 14.8157\,\mathrm{PR}_i + 10.7263\,\mathrm{SR}_i - 0.0299\,\mathrm{AT}_i + 1.3426\,\mathrm{T}_{60_i} + 1.7952\,\mathrm{T}_{90_i} \\ &+ 2.0243\,\mathrm{T}_{120_i}.\end{aligned}$$

Based on the fitted RUBXII regression, some findings of the modeling mortality rate's median by COVID-19 in the U.S. states are now presented.

- The PD presents a p-value lower than 0.0001, and its associated estimate is positive, which indicates that the MR is higher in states most densely populated.

Figure 12 – Residuals plots for the fitted Kw regression.



Source: Author (2020)

- The HDI is significant at any significance level since p-value < 0.0001 and unlikely of the expected, the signal of the associated estimate indicates that the MR increases when the HDI increases.
- The GINI coefficient is significant at the 1% level, and its positive estimate means that the MR increases in states with larger Gini coefficient.
- The number of hospital beds is significant at the 10% level. The mortality rate's median decreases when the total hospital beds per 100 thousand inhabitants increases as expected.
- The PR is significant at the 1% level, and the signal of its associated estimate reveals that the MR grows when the PR increases.
- The SR is also significant at any level since p-value$= 0.0001$. The mortality rate's median increases as the SR grows according to the positive signal of its related estimate. This is expected since the immune response of smoking patients potentially decreases (TAGHIZADEH-HESARY; AKBARI, 2020).
- The AT in each state is statistically significant at any usual nominal levels and its signal estimate indicates that the MR decreases when the AT grows. Indeed, Wang *et al.* (2020) showed that high temperatures reduce the COVID-19 viability, i.e., the increase of the temperature suggests a decline in disease spread, and, hence, a decline on the number of

deaths.

- The dummy variables related to the time 60, 90, and 120 days after the 20th confirmed case are significant as expected. As indicated by the box plots in Figure 8, the MR grows steadily during the considered periods.

3.5 CONCLUSION

The COVID-19 characterizes a global pandemic that has been spread across the United States of America (U.S.) since January 2020. In this chapter, it is investigated how demographic, socioeconomic, behavioral, and meteorological variables are related to the mortality rate by COVID-19 in the U.S. states. To reach that aim properly, it is chosen regressions that consider the double bounded characteristic of the mortality rate. It is introduced an alternative model called the reflexive unit Burr XII (RUBXII) regression, which is a useful tool for modeling bounded random variables in the interval $(0, 1)$, such as rates, proportions, and indexes. This proposal is based on a new unit continuous distribution that arises from a transformation on a random variable Burr XII. Further, a more general and useful quantile-parameterization is introduced to define the quantile regression for unit data. The estimation of the parameters, a simulation study to evaluate the performance of the maximum likelihood estimators, and some adequacy measures to check whether the regression's assumptions hold are discussed. After consolidating the data set about the mortality rates and other covariates for the U.S. states, a descriptive statistical analysis and a regression modeling are done. In this way, the new regression is compared with the beta, simplex, Kumaraswamy, and unit-Weibull regressions. The proposed regression is quite competitive compared with other regressions, and it provides the best fit according to some selection criteria since it improves the response's prediction. Thus, from the fitted RUBXII regression, it is possible to identity that the population density, human developed index, Gini coefficient, hospital beds, poverty rate, smoking rate, average temperature (^{o}F) in each state, and during time variables are statistically significant in the modeling of the mortality rate's median by COVID-19 in the U.S. states. The findings in this chapter may improve understanding of coronavirus in the U.S. and assist health-care system readiness for future coronavirus epidemics or pandemics. Since the potentiality of the RUBXII regression to analyze coronavirus data, it is aimed in future research to fit this regression to the mortality rates by coronavirus in other countries of the world.

4 UNIT BURR XII AUTOREGRESSIVE MOVING AVERAGE MODEL FOR TIME SERIES DATA RESTRICTED IN THE UNIT INTERVAL

4.1 INTRODUCTION

Time series data are characterized by sets of observations on the values that a random variable can take over time. Typically, models suitable for this type of data are based on Gaussianity assumptions for the interest variable. Classic examples are the class of autoregressive moving average (ARMA) models and autoregressive integrated moving average (ARIMA) models; see Brockwell, Davis and Fienberg (1991), Box, Jenkins and Reinsel (2011) for more details. However, it is not appropriate to perform Gaussian-based inference for a random variable that has not Gaussian distribution.

According to Cox *et al.* (1981) classification, the issue of extending ARMA time series models to a non-Gaussian framework could be treated under two approaches, such as observation-driven models and parameter-driven (or state-space) models. These approaches mainly differ in the way the dependence structure is incorporated into the model. In the first class, parameters at time t are deterministic functions of lagged dependent variables whereas, in the parameter-driven models, the parameters vary over time as dynamic processes (FRANCO *et al.*, 2019). In the context of state-space models, an example of the most recent approaches that have been used are Markov chain Monte Carlo (MCMC) methods. They aim to obtain posterior distributions for the parameters of these models; see Durbin and Koopman (1997). However, convergence issues and inferential theory for MCMC techniques were not yet fully developed (BENJAMIN; RIGBY; STASINOPOULOS, 2003). On the other hand, from the observation-driven models, the data's likelihood can be expressed explicitly for any fixed set of parameter values. Further, in this approach, it is simplest to carry out the model comparison and diagnostics analysis (BENJAMIN; RIGBY; STASINOPOULOS, 2003). Thus, our focus henceforth is the models of this type.

Several works have been done in the context of observation-driven models. Zeger and Qaqish (1988) considers a class of Markov autoregressive models and discusses a quasi-likelihood (QL) approach to regression analysis with time series data based on the exponential family conditional-response variables, in special distributed as Poisson and gamma. To this class, moving average components were introduced by Li (1994). Afterwards, Benjamin, Rigby and Stasinopoulos (2003) extended the Gaussian ARMA time series models to a non-Gaussian framework by developing dynamic models for random variables in the exponential family, arising the generalized autoregressive moving average (GARMA) models. Moreover, they provided a

formula for the fourth moment of the generalized autoregressive conditional heteroskedasticity (GARCH) model introduced by Bollerslev (1986). Other relevant studies were developed regarding generalized linear autoregressive moving average (GLARMA) models; see Shephard (1995) and Davis, Dunsmuir and Streett (2003).

In order to model serially dependent overtime random variables that assume values on the interval $(0, 1)$, Rocha and Cribari-Neto (2009) pioneered the beta autoregressive moving average (βARMA) models based on the class of beta regression models (FERRARI; CRIBARI-NETO, 2004). They considered a similar approach to those of Benjamin, Rigby and Stasinopoulos (2003) and Shephard (1995). Afterward, Bayer, Bayer and Pumi (2017) proposed the Kumaraswamy autoregressive moving average (KARMA) models for double bounded environmental data. The advantage of this class over βARMA is that they employed a parameterization for the Kumaraswamy distribution in terms of its median by providing more robust models to the presence of atypical observations in the conditional response.

Using suitable models for double-bounded conditional-response variables in the unit interval avoids data transformation before modeling. Moreover, they can naturally accommodate asymmetries and heteroscedasticity, commons to this type's data (ROCHA; CRIBARI-NETO, 2009). Therefore, there is a clear need for new flexible alternatives to the scarcely available classes of time series models.

In this context, we propose a dynamic model for time series data where the conditional-response has support on the standard unit interval. The new model is based on the class of unit Burr XII (UBXII) regression models introduced in Chapter 2. We include additively ARMA terms to the systematic component of the UBXII regression and define the *UBXII-ARMA* model.

The remainder of this chapter is outlined as follows. Section 4.2 introduces the new time series model for conditional variates restricted to the interval $(0, 1)$. Section 4.3 discusses conditional maximum likelihood estimation for the UBXII-ARMA models and provides closed forms for the conditional score vector. Further, we present asymptotic confidence intervals based on the conditional maximum likelihood estimator (CMLE) properties. A Monte Carlo simulation study to assess the finite sample performance of the CMLEs is conducted in Section 4.4. We evaluate the point estimates and the estimated coverage probability from the asymptotic confidence intervals for the UBXII-ARMA model's parameters. Some diagnostic analysis measures and forecasting methods are presented in Section 4.5. An application in stocked hydroelectric energy data is carried out to provide empirical evidence of the proposed model's potentiality. Finally, some conclusions are discussed in Section 4.7.

4.2 THE PROPOSED MODEL

In this section, we introduce a dynamic time series model for the UBXII distribution pioneered in Chapter 2. In this way, it is possible to model serial correlation in its conditional quantiles. We shall consider a similar approach to the employed in the construction of the GARMA (BENJAMIN; RIGBY; STASINOPOULOS, 2003), βARMA (ROCHA; CRIBARI-NETO, 2009), and KARMA (BAYER; BAYER; PUMI, 2017) models.

Let $\{Y_t\}_{t=1}^n$ be a sequence of random variables UBXII-distributed and let $\mathcal{F}_t = \sigma\{Y_t, Y_{t-1}, \ldots\}$ be the σ-algebra generated by the observed information up to time t. Given the previous information set $\mathcal{F}_{t-1}$ (i.e., the smallest σ-algebra such that the variables $Y_1, \ldots, Y_{t-1}$ are measurable), consider that the conditional distribution of each Y_t follows the UBXII distribution, say $Y_t|\mathcal{F}_{t-1} \sim \text{UBXII}(q_t, c)$. Thus, the conditional density of Y_t given $\mathcal{F}_{t-1}$ is

$$f(y_t|\mathcal{F}_{t-1}) = \frac{\log\tau^{-c}\log^{c-1}y_t^{-1}}{y_t\log\left(1+\log^c q_t^{-1}\right)}\left(1+\log^c y_t^{-1}\right)^{\log\tau/\log\left(1+\log^c q_t^{-1}\right)-1}, \quad 0<y_t<1, \tag{4.1}$$

where $0 < q_t < 1$ is a quantile of Y_t and $c > 0$ is a shape parameter. If $\tau = 0.5$, q_t is the median of Y_t. Moreover, the conditional cumulative distribution function (cdf) and conditional quantile function (cqf) are

$$F(y_t|\mathcal{F}_{t-1}) = \left(1+\log^c y_t^{-1}\right)^{\log\tau/\log\left(1+\log^c q_t^{-1}\right)}, \qquad 0<y_t<1, \tag{4.2}$$

and

$$Q(u|\mathcal{F}_{t-1}) = \exp\left\{-[u^{\log(1+\log^c q_t^{-1})/\log\tau}-1]^{1/c}\right\}, \quad 0<\tau<1, \tag{4.3}$$

respectively. Occurrences of the UBXII distribution may be easily generated by the inversion method since the cqf has a simple closed-form expression.

The UBXII distribution is quite versatile since its pdf can assume many different shapes according to the selected parameter values combination. We develop an Open Web application in `Shiny` (R Core Team, 2020) available at <https://unitati.shinyapps.io/UBXII/> that allows viewing dynamic graphics of the UBXII pdf.

To define the dynamic component of the model, we propose the following specification to the conditional quantile q_t

$$\eta_t = g(q_t) = \alpha + \mathbf{x}_t^\top\boldsymbol{\beta} + \sum_{i=1}^{p}\phi_i\left[\,g(y_{t-i}) - \mathbf{x}_{t-i}^\top\boldsymbol{\beta}\,\right] + \sum_{j=1}^{q}\theta_j\, r_{t-j}, \tag{4.4}$$

where $\alpha \in \mathbb{R}$ is a constant, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k$ is a k-dimensional vector of unknown parameters related to the covariates, $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tk})^\top \in \mathbb{R}^k$ is a non-random covariates vector,

being $k < n$, and $\phi_i\,(i = 1,\ldots,p)$ and $\theta_j\,(j = 1,\ldots,q)$ are the autoregressive (AR) and moving average (MA) parameters, respectively. That is, they are the parameters from an ARMA structure with p, $q \in \mathbb{N}$, say ARMA(p,q). The term r_{t-j} corresponds to a random error that can be measured on the predictor scale as $r_t = g(y_t) - g(q_t)$ or on the original scale $r_t = y_t - q_t$. It is only required that r_t be $\mathcal{F}_{t-1}$-measurable. Finally, q_t is related to a linear predictor η_t, through a twice differentiable strictly monotonic link function that maps $(0,1)$ into $\mathbb{R}$, i.e., $g : (0,1) \to \mathbb{R}$ for which the inverse $g^{-1} : \mathbb{R} \to (0,1)$ exists and also is twice continuously differentiable. Some examples of link functions are the logit, probit, and complementary log-log links.

In this way, from (4.1) and (4.4), we define the so-called *UBXII-ARMA*(p,q) dynamic model. In a similar manner of classical ARMA models, we require non-common factors between the AR and MA characteristic polynomials; otherwise the order (p,q) of the model can be reduced. Further, the polynomial AR does not have unit characteristic root. Analogous to the KARMA model, invertibility and causality conditions for the ARMA component are not required; see Brockwell, Davis and Fienberg (1991) for more details.

The dynamic component (see Equation (4.4)) is similar to that one proposed by Rocha and Cribari-Neto (2009) and later by Bayer, Bayer and Pumi (2017). However, there are two sizeable differences. First, the random component is entirely distinct from both proposed since the response variable here has a UBXII distribution. Second, in the class of UBXII-ARMA(p,q) models, it is possible to model any quantile of the response instead of the mean or only the median. Thus, it is a more general time series model and a new alternative that allows analyzing a range of double-bonded conditional responses on the interval $(0,1)$.

4.3 PARAMETER ESTIMATION

The model-fitting procedure described herein is performed out of the conditional maximum likelihood method. Let $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\beta}^\top, \boldsymbol{\phi}^\top, \boldsymbol{\theta}^\top, c)^\top$ be the $(p+q+k+2)$-dimensional parameter vector that index the UBXII-ARMA(p,q) model in a sample $(y_1, \boldsymbol{x_1}), \ldots, (y_n, \boldsymbol{x_n})$, satisfying the specification given in (4.1) and (4.4). Hence, the conditional log-likelihood function can be expressed as

$$\ell \equiv \ell(\boldsymbol{\gamma}; \boldsymbol{y}) = \sum_{t=m+1}^{n} \ell_t(q_t, c), \tag{4.5}$$

where $m = \max\{p, q\} < n$ and $\ell_t(q_t, c)$ is the logarithm of $f(y_t|\mathcal{F}_{t-1})$ given in Equation (4.1). That is,

$$\ell_t(q_t, c) = \log(\log \tau^{-c}) - \log y_t + (c-1)\log(\log y_t^{-1}) - \log[t(y_t)] - \log\{\log[t(q_t)]\} - \frac{\log \tau^{-1} \log[t(y_t)]}{\log[t(q_t)]},$$

where $t(x) = 1 + \log^c(x^{-1})$. It is important to note that $\ell_t(q_t, c) = 0$ for all $t \leq m$.

Upon direct maximizing (4.5), we get the CMLEs $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$. Alternatively, we can obtain the score vector, set its components to zero, and solve the resulting non-linear equation system. In what follows, we compute the score vector by differentiating (4.5) concerning each component of the unknown parameter vector $\boldsymbol{\gamma}$.

4.3.1 Conditional score vector

The conditional score vector, denoted by $U(\boldsymbol{\gamma})$, is composed of the partial derivatives of ℓ with respect to each component of $\boldsymbol{\gamma}$. That is, $U(\boldsymbol{\gamma}) := \partial\ell/\partial\boldsymbol{\gamma} = [U_\alpha(\boldsymbol{\gamma}), U_{\boldsymbol{\beta}}(\boldsymbol{\gamma})^\top, U_{\boldsymbol{\phi}}(\boldsymbol{\gamma})^\top, U_{\boldsymbol{\theta}}(\boldsymbol{\gamma})^\top, U_c(\boldsymbol{\gamma})]^\top$. Let γ_j be the jth component of $\boldsymbol{\gamma}$. Then, the $(k+p+q+1)$ first components of the conditional score vector are obtained using the chain rule as

$$U_{\gamma_j}(\boldsymbol{\gamma}) := \frac{\partial \ell}{\partial \gamma_j} = \sum_{t=m+1}^{n} \frac{\partial \ell_t(q_t, c)}{\partial q_t} \frac{\mathrm{d}q_t}{\mathrm{d}\eta_t} \frac{\partial \eta_t}{\partial \gamma_j}. \tag{4.6}$$

Thus, defining the quantities

$$y_t^\star := \log[t(y_t)], \quad q_t^\star := \frac{c \log^{c-1} q_t^{-1}}{q_t\, t(q_t) \log[t(q_t)]}, \quad \text{and} \quad q_t^\dagger := \frac{\log \tau^{-c} \log^{c-1} q_t^{-1}}{q_t\, t(q_t) \log^2[t(q_t)]},$$

the two first derivatives in (4.6) reduce to

$$\frac{\partial \ell_t(q_t, c)}{\partial q_t} = q_t^\star - q_t^\dagger y_t^\star \quad \text{and} \quad \frac{\mathrm{d}q_t}{\mathrm{d}\eta_t} = \frac{1}{g'(q_t)}.$$

The partial derivatives, $\partial\eta_t/\partial\gamma_j$, are computed recursively as

$$\frac{\partial \eta_t}{\partial \alpha} = 1 - \sum_{j=1}^{q} \theta_j \frac{\partial \eta_{t-j}}{\partial \alpha}, \quad \text{for } r = 1,$$

$$\frac{\partial \eta_t}{\partial \beta_l} = x_{tl} - \sum_{i=1}^{p} \phi_i x_{(t-i)l} - \sum_{j=1}^{q} \theta_j \frac{\partial \eta_{t-j}}{\partial \beta_l}, \quad \text{for } r = 2, \ldots, k+1, \text{ and } l = 1, \ldots, k,$$

$$\frac{\partial \eta_t}{\partial \phi_i} = g(y_{t-i}) - \boldsymbol{x}_{t-i}^\top \boldsymbol{\beta} - \sum_{j=1}^{q} \theta_j \frac{\partial \eta_{t-j}}{\partial \phi_i}, \quad \text{for } r = k+2, \ldots, p+1, \text{ and } i = 1, \ldots, p,$$

and

$$\frac{\partial \eta_t}{\partial \theta_j} = r_{t-j} - \sum_{v=1}^{q} \theta_v \frac{\partial \eta_{t-v}}{\partial \theta_j}, \quad \text{for } r = p+2, \ldots, q+1, \text{ and } j = 1, \ldots, q.$$

The last component of the conditional score vector, $U_c(\gamma)$, follows from direct differentiation of (4.5)

$$\frac{\partial \ell}{\partial c} = \sum_{t=m+1}^{n} \frac{\partial \ell_t(q_t, c)}{\partial c} = \sum_{t=m+1}^{n} y_t^{\sharp},$$

where

$$\begin{aligned} y_t^{\sharp} = \frac{1}{c} + \log(\log y_t^{-1}) - \frac{\log(\log q_t^{-1})[t(q_i) - 1]}{t(q_t)\log[t(q_t)]} - \frac{[t(y_t) - 1]\log(\log y_t^{-1})}{t(y_t)} \\ - \frac{\log \tau^{-1} \log[t(q_t)][t(y_t)]^{-1}[t(y_t) - 1]\log(\log y_t^{-1})}{\log^2[t(q_t)]} \\ + \frac{\log \tau^{-1}[t(q_t) - 1]\log(\log q_t^{-1})\log[t(y_t)]}{t(q_t)\log^2[t(q_t)]}. \end{aligned}$$

Let $\boldsymbol{M}, \boldsymbol{P}, \boldsymbol{R}$ be matrices with dimension $(n-m)\times k$, $(n-m)\times p$ and $(n-m)\times q$, respectively. The (i,j)th element of those matrices are given by

$$\boldsymbol{M}_{i,j} = \frac{\partial \eta_{i+m}}{\partial \beta_j}, \quad P_{i,j} = \frac{\partial \eta_{i+m}}{\partial \phi_j}, \quad \text{and} \quad R_{i,j} = \frac{\partial \eta_{i+m}}{\partial \theta_j},$$

respectively. Then, we can compactly write the score vector's components of $\boldsymbol{\gamma}$ as

$$\begin{aligned} U_{\alpha}(\boldsymbol{\gamma}) &= \boldsymbol{a}^{\top} \boldsymbol{T} (\boldsymbol{q}^{\star} - \boldsymbol{q}^{\dagger} \boldsymbol{y}^{\star}) \\ U_{\boldsymbol{\beta}}(\boldsymbol{\gamma}) &= \boldsymbol{M}^{\top} \boldsymbol{T} (\boldsymbol{q}^{\star} - \boldsymbol{q}^{\dagger} \boldsymbol{y}^{\star}) \\ U_{\boldsymbol{\phi}}(\boldsymbol{\gamma}) &= \boldsymbol{P}^{\top} \boldsymbol{T} (\boldsymbol{q}^{\star} - \boldsymbol{q}^{\dagger} \boldsymbol{y}^{\star}) \\ U_{\boldsymbol{\theta}}(\boldsymbol{\gamma}) &= \boldsymbol{R}^{\top} \boldsymbol{T} (\boldsymbol{q}^{\star} - \boldsymbol{q}^{\dagger} \boldsymbol{y}^{\star}) \\ U_{c}(\boldsymbol{\gamma}) &= \boldsymbol{y}^{\sharp \top} \mathbf{1}, \end{aligned}$$

where $\boldsymbol{a} = (\partial \eta_{m+1}/\partial \alpha, \ldots, \partial \eta_n / \partial \alpha)^{\top}$, $\boldsymbol{T}$ is a diagonal matrix defined as $\boldsymbol{T} = \text{diag}\{1/g'(q_{m+1}), \ldots, 1/g'(q_n)\}$, $\boldsymbol{q}^{\star} = (q^{\star}_{m+1}, \ldots, q^{\star}_n)^{\top}$, $\boldsymbol{q}^{\ddagger} = (q^{\dagger}_{m+1}, \ldots, q^{\dagger}_n)^{\top}$, $\boldsymbol{y}^{\star} = (y^{\star}_{m+1}, \ldots, y^{\star}_n)^{\top}$, $\boldsymbol{y}^{\sharp} = (y^{\sharp}_{m+1}, \ldots, y^{\sharp}_n)^{\top}$, and $\mathbf{1}$ is an $(n-m)$-dimensional vector of ones.

By setting each $U(\boldsymbol{\gamma})$ component equal to zero, i.e., $U_{\alpha}(\boldsymbol{\gamma}) = 0$, $U_{\boldsymbol{\beta}}(\boldsymbol{\gamma}) = \mathbf{0}$, $U_{\boldsymbol{\phi}}(\boldsymbol{\gamma}) = \mathbf{0}$, $U_{\boldsymbol{\theta}}(\boldsymbol{\gamma}) = \mathbf{0}$, $U_c(\boldsymbol{\gamma}) = 0$, and solving these equations simultaneously, the CMLE $\hat{\boldsymbol{\gamma}} = (\hat{\alpha}, \hat{\boldsymbol{\beta}}^{\top}, \hat{\boldsymbol{\phi}}^{\top}, \hat{\boldsymbol{\theta}}^{\top}, \hat{c})^{\top}$ of $\boldsymbol{\gamma}$ is obtained. However, since this system is nonlinear and cannot be solved explicitly, we may maximize the Equation (4.5) through nonlinear optimization methods such as Newton-Raphson or quasi-Newton type algorithms. We consider the quasi-Newton algorithm the so-called

Broyden- Fletcher-Goldfarb-Shanno (BFGS) method; see Press *et al.* (1992). This method is an iterative optimization algorithm, and thus, it requires initialization. We compute the starting values for α, $\boldsymbol{\beta}$, and $\boldsymbol{\phi}$ from an ordinary least squares estimate by considering a linear regression, where the response is $\boldsymbol{Y} = (g(y_{m+1}), \ldots, g(y_n))^\top$, and the covariates matrix is expressed as

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{(m+1)1} & x_{(m+1)2} & \cdots & x_{(m+1)r} & g(y_m) & g(y_{m-1}) & \cdots & g(y_{m-p+1}) \\ 1 & x_{(m+2)1} & x_{(m+2)2} & \cdots & x_{(m+2)r} & g(y_{m+1}) & g(y_m) & \cdots & g(y_{m-p+2}) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nr} & g(y_{n-1}) & g(y_{n-2}) & \cdots & g(y_{n-p}) \end{bmatrix}.$$

For the moving average parameters $\boldsymbol{\theta}$, the starting values are set to zero, and the initial guess for the shape parameter c is one.

From likelihood estimation properties when the usual regularity conditions hold for large sample sizes, we have that

$$\hat{\boldsymbol{\gamma}} \sim \mathcal{N}_{(k+p+q+2)}\left(\boldsymbol{\gamma}, \boldsymbol{K}^{-1}(\boldsymbol{\gamma})\right),$$

where $\boldsymbol{K^{-1}}(\boldsymbol{\gamma})$ is the inverse of the expected information matrix defined as $\boldsymbol{K}(\boldsymbol{\gamma}) = \mathbb{E}[-\partial\ell^2(\boldsymbol{\gamma})/(\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top)]$ and $\mathcal{N}_r$ denotes a r-dimensional normal distribution. That is, the CMLE of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}$, is asymptotically unbiased and normally distributed with covariance matrix equal to the inverse of the Fisher's information matrix. The matrix $\boldsymbol{K}(\boldsymbol{\gamma})$ can be estimated consistently from the observed information matrix evaluated in the CMLE of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}$; see Lindsay and Li (1997).

From the asymptotic normality of $\hat{\boldsymbol{\gamma}}$, it is possible to construct a $100(1-\delta)\%$ approximate confidence interval with $\delta \in (0, 1/2)$ for the elements of $\boldsymbol{\gamma}$, i.e., for γ_i $(i = 1, \ldots, p+q+k+2)$ as follows

$$\left[\hat{\gamma}_i - z_{1-\delta/2}\sqrt{J(\hat{\boldsymbol{\gamma}})^{ii}}; \hat{\gamma}_i + z_{1-\delta/2}\sqrt{J(\hat{\boldsymbol{\gamma}})^{ii}}\right], \tag{4.7}$$

where z_δ is the standard normal upper quantile and $J(\hat{\boldsymbol{\gamma}})^{ii}$ is the (i,i)th element of the $\boldsymbol{J}^{-1}$.

Analogously, based on the asymptotic distribution of the CMLE, we can construct asymptotic test statistics for testing the null hypothesis $\mathcal{H}_0 : \gamma_i = \gamma_i^0$ against $\mathcal{H}_1 : \gamma_i \neq \gamma_i^0$. It can be done via the Wald test (WALD, 1943) defined as (PAWITAN, 2001)

$$Z = \frac{\hat{\gamma}_i - \gamma_i^0}{\sqrt{J(\hat{\boldsymbol{\gamma}})^{ii}}}.$$

Under $\mathcal{H}_0$, the Z statistic has an approximately standard normal distribution. Thus, to the significance level of $\delta\%$, with $(0 < \delta < 1/2)$, we reject the null hypothesis, whether the assumed value by Z, denoted by z, exceeds the quantity $|z_{1-\delta/2}|$.

4.4 SIMULATION STUDY

To assess the finite sample performance of the CMLEs and the asymptotic confidence intervals of the parameters that index the UBXII-ARMA(p,q) model, we conduct a Monte Carlo simulation study. Samples of sizes $n \in \{75, 125, 200, 300\}$ are considered at four distinct scenarios. For each scenario and sample size, we compute 10,000 times the CMLEs and the confidence intervals of the model's parameters.

Several dynamic specifications and different parameter value combinations are selected. For all settings, $\tau = 0.5$ is fixed, and thus q_t, the conditional median of Y_t. We consider the logit link function for $g(\cdot)$ in (4.4). The simulation schemes are

- Scenario 1: UBXII-ARMA$(2,2)$ with parametric values $\alpha = 0.5$, $\phi_1 = 0.6$, $\phi_2 = -0.4$, $\theta_1 = 0.4$, $\theta_2 = 0.1$, and $c = 5.6$.
- Scenario 2: UBXII-ARMA$(1,1)$ with parametric values $\alpha = 0.2$, $\phi_1 = 0.6$, $\theta_1 = 0.1$, and $c = 3.8$.
- Scenario 3: UBXII-ARMA$(2,1)$ with parametric values $\alpha = 0.4$, $\phi_1 = 0.6$, $\phi_2 = -0.4$, $\theta_1 = 0.3$, and $c = 4.5$.
- Scenario 4: UBXII-ARMA$(1,2)$ with parametric values $\alpha = 0.7$, $\phi_1 = -0.7$, $\theta_1 = 0.4$, $\theta_2 = 0.6$, and $c = 3.5$.

For the generation of samples from a UBXII-ARMA(p,q) process, we use the same algorithm employed by Bayer, Bayer and Pumi (2017). For the random component simulation, we adopt the inversion method replacing $u \sim U(0,1)$ in (4.3) and assume the dynamic structure given in Equation (4.4). All Monte Carlo simulations are performed using the R programming language (R Core Team, 2020). Maximization of the conditional log-likelihood function in (4.5) is carried out using the BFGS quasi-Newton nonlinear optimization algorithm implemented at the `optim` function.

To numerically evaluate the behavior of the CMLEs, we compute its percentual relative bias (RB%) and mean squared error (MSE). Monte Carlo results for the different structures are reported in Table 11. We note that the RBs and MSEs are quite small in most scenarios (even those of small samples), thus indicating that the UBXII-ARMA model provides accurate estimates. The largest RB%s (in absolute value) are around 22 and 30 for the estimated means of the parameters of the moving average (θ_2, θ_1) in scenarios 1 and 2, respectively, when the smallest sample size is $n = 75$. All the remaining RBs (in absolute value) are lower than 16.

The RB and MSE values of Scenario 2 are close to the Monte Carlo simulation results for

point estimation in the βARMA$(1,1)$ model carried out by Palm and Bayer (2018) (see Table 3, no corrected estimates). Roughly, it is noteworthy that the UBXII-ARMA$(1,1)$ model yields smaller MSE values by comparing the sample size $n \in \{75, 125\}$ with $n \in \{50, 100\}$.

Overall, we note a result similar to that of Bayer, Bayer and Pumi (2017), in which the estimates of the other parameters tend to present better performance compared to the parameter estimates of the moving average terms. As expected, the MSEs decrease for all scenarios when n increases, thus implying that the performance and accuracy of the CMLEs improve when the sample size increases. Moreover, the MSE values are quite low in any one of the four settings. Therefore, the numerical evaluation indicates that the properties of the CMLEs (asymptotically unbiased and consistent) are remained.

Table 12 brings the estimated coverage probability from the asymptotic confidence intervals for the parameters $\alpha, \phi_1, \phi_2, \theta_1, \theta_2, c$. The limits of the confidence intervals are computed from (4.7) to the usual significance level of $\delta = 0.05$. Overall, the coverage probability of the 95% pointwise confidence intervals of all the parameters is quite close to the considered nominal level. Scenario 1 has the lowest coverage probabilities, mainly for $n = 75$. However, as the sample size increases, coverage probability becomes closer to 95% for all settings and different selected parameter values.

Table 11 – Performance of the CMLEs for the UBXII-ARMA(p,q) model under different ARMA structures and parameter values.

n	Measure	$\hat{\alpha}$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{c}$
				Scenario 1			
75	RB%	2.0185	−5.3152	−2.9294	9.7761	22.0407	−4.9018
	MSE	0.0341	0.1378	0.0293	0.1632	0.0908	0.4605
125	RB%	0.9947	−2.4144	−1.6880	3.9732	7.6033	−3.1744
	MSE	0.0159	0.0722	0.0155	0.0807	0.0427	0.2023
200	RB%	1.8867	−3.8484	−2.0145	6.2767	15.5935	−1.8121
	MSE	0.0102	0.0404	0.0080	0.0454	0.0257	0.1285
300	RB%	1.2304	−2.5762	−1.3755	4.2335	10.9272	−1.2134
	MSE	0.0061	0.0252	0.0051	0.0279	0.0152	0.0735
				Scenario 2			
75	RB%	−10.5114	5.9353	–	−29.8865	–	−3.2580
	MSE	0.0108	0.0178	–	0.0225	–	0.1376
125	RB%	−5.2358	3.0885	–	−15.6325	–	−1.9586
	MSE	0.0054	0.0087	–	0.0119	–	0.0749
200	RB%	−3.4477	2.0306	–	−9.8161	–	−1.1865
	MSE	0.0032	0.0052	–	0.0072	–	0.0431
300	RB%	−2.4488	1.4366	–	−6.7473	–	−0.8043
	MSE	0.0020	0.0033	–	0.0045	–	0.0278
				Scenario 3			
75	RB%	−0.0867	−3.0292	−4.4578	7.5658	–	−4.2000
	MSE	0.0072	0.0344	0.0182	0.0439	–	0.2202
125	RB%	0.1510	−1.4249	−2.1015	3.0766	–	−2.4926
	MSE	0.0038	0.0168	0.0095	0.0195	–	0.1136
200	RB%	0.1408	−0.8888	−1.2725	2.0683	–	−1.4964
	MSE	0.0023	0.0097	0.0057	0.0112	–	0.0649
300	RB%	0.0442	−0.4714	−0.7727	1.2668	–	−0.9823
	MSE	0.0015	0.0065	0.0039	0.0072	–	0.0412
				Scenario 4			
75	RB%	1.1768	1.3360	–	−0.0767	0.1104	−3.2573
	MSE	0.0130	0.0092	–	0.0094	0.0093	0.1349
125	RB%	0.6865	1.1969	–	0.5190	0.6961	−1.8945
	MSE	0.0071	0.0044	–	0.0046	0.0048	0.0730
200	RB%	0.3327	1.1161	–	0.5916	0.8119	−1.2053
	MSE	0.0045	0.0025	–	0.0025	0.0032	0.0443
300	RB%	0.0599	0.8964	–	0.3750	0.5470	−0.8817
	MSE	0.0030	0.0015	–	0.0015	0.0020	0.0285

Source: Author (2020)

Table 12 – Estimated coverage probability from the asymptotic confidence intervals for $\alpha, \phi_1, \phi_2, \theta_1, \theta_2, c$

Parameter	α	ϕ_1	ϕ_2	θ_1	θ_2	c
n	Scenario 1					
75	0.8026	0.7364	0.8642	0.7268	0.7577	0.9165
125	0.8770	0.8310	0.9023	0.8307	0.8441	0.9375
200	0.9020	0.8751	0.9257	0.8766	0.8831	0.9396
300	0.9228	0.9035	0.9352	0.9037	0.9082	0.9445
	Scenario 2					
75	0.9478	0.9484	–	0.9269	–	0.9434
125	0.9520	0.9508	–	0.9377	–	0.9437
200	0.9540	0.9496	–	0.9403	–	0.9482
300	0.9534	0.9533	–	0.9443	–	0.9478
	Scenario 3					
75	0.9243	0.8941	0.8995	0.8841	–	0.9363
125	0.9395	0.9201	0.9205	0.9179	–	0.9426
200	0.9459	0.9363	0.9336	0.9340	–	0.9457
300	0.9486	0.9427	0.9390	0.9407	–	0.9455
	Scenario 4					
75	0.9258	0.9305	–	0.9158	0.9065	0.9229
125	0.9362	0.9408	–	0.9268	0.9284	0.9261
200	0.9388	0.9377	–	0.9353	0.9360	0.9244
300	0.9347	0.9425	–	0.9436	0.9439	0.9190

Source: Author (2020)

4.5 DIAGNOSTIC ANALYSIS AND FORECASTING

After fitting a model, it is important to perform adequacy tests to check whether it fully captures the data dynamics. Since a fitted time series model passes all diagnostic checks, we may use it for out-of-sample forecasting. In what follows, we introduce and discuss some known diagnostic measures and forecasting methods that can be used to identify whether assumptions of a UBXII-ARMA model are satisfied.

We consider the randomized quantile residuals (DUNN; SMYTH, 1996) since they have several advantages over other residuals (PEREIRA, 2019). For the UBXII-ARMA model, the quantile residuals are defined by

$$r_t = \Phi^{-1}\left[F(y_t|\mathcal{F}_{t-1})\right],$$

where $\Phi^{-1}(\cdot)$ is the standard normal quantile function and $F(y_t|\mathcal{F}_{t-1})$ is the cdf given in (4.2), evaluated in the CMLEs, specifically in $\hat{q}_t$ that corresponds to fitted values and in $\hat{c}$. The quantile residuals are roughly normally distributed with mean equal to zero and unit variance when the model is suitable for the data. Furthermore, the index plot of these residuals should not display any noticeable trend.

To picks the most suitable model to fit a data set within other competitive models, the Akaike information criterion (AIC) (AKAIKE, 1973) or, alternatively, the Bayesian information criterion (BIC) (SCHWARZ *et al.*, 1978) can be considered for models selection/comparison. Both criteria are based on maximized conditional log-likelihood function, $\hat{\ell}$, namely

$$\mathrm{AIC}(\nu) = 2\,(\nu - \hat{\ell}) \quad \text{and} \quad \mathrm{SIC}(\nu) = \nu \log n - 2\hat{\ell},$$

where ν is the number of estimated parameters, and n is the sample size. For more details about these criteria, the reader is referred to Choi (2012), who provides their detailed properties.

From elsewhere, suppose that we are interested in forecasting of quantile q_s using origin n ($s > n$). Then, the forecast horizon is $h_0 = s - n$. Initially, we compute the $\{\hat{q}_t\}_{t=m+1}^{n}$ estimates of $\{q_t\}_{t=m+1}^{n}$ sequentially based on the CMLE $\hat{\boldsymbol{\gamma}}$, starting at $t = m+1$, as

$$\hat{q}_t = g^{-1}\left(\hat{\alpha} + \mathbf{x}_t^{\top}\hat{\boldsymbol{\beta}} + \sum_{i=1}^{p}\hat{\phi}_i\,[\,g(y_{t-i}) - \mathbf{x}_{t-i}^{\top}\hat{\boldsymbol{\beta}}\,] + \sum_{j=1}^{q}\hat{\theta}_j\,\hat{r}_{t-j}\right),$$

where

$$\hat{r}_t = \begin{cases} 0 & \text{if } t \leq m. \\ g(y_t) - g(\hat{q}_t) & \text{if } m < t \leq n. \end{cases}$$

Then, for $t = n+1, \ldots, s$, we need to assume that the observations from the covariates $\boldsymbol{x_t}$ are known. Hence, the forecasted values of the conditional quantiles q_s, being $h = 1, \ldots, h_0$, are obtained sequentially from

$$\hat{q}_{n+h} = g^{-1}\left(\hat{\alpha} + \mathbf{x}_{n+h}^{\top}\hat{\boldsymbol{\beta}} + \sum_{i=1}^{p}\hat{\phi}_i\,[\,g(y_{n+h-i}) - \mathbf{x}_{n+h-i}^{\top}\hat{\boldsymbol{\beta}}\,] + \sum_{j=1}^{q}\hat{\theta}_j\,\hat{r}_{n+h-j}\right),$$

where for $t > n$, $\hat{r}_t = 0$ and

$$g(y_t) = \begin{cases} g(\hat{q}_t) & \text{if } t > n, \\ g(y_t) & \text{if } t \leq n. \end{cases}$$

To empirically evaluate the forecasting performance of the UBXII-ARMA model and compare it to the other fitted models, we consider three measures of forecast accuracy, such as the mean square error (MSE), the mean absolute percentage error (MAPE), and the mean absolute scaled error (MASE). These measures allow assessing the difference between the actual value and the predicted value. MSE is largely used due to its theoretical relevance in statistical modeling. However, when the data are supported in positive real, the use of the MAPE is indicated. On the other hand, in the presence of atypical observations in the response, MASE may be preferred

since it is less sensitive to outliers; see Hyndman and Koehler (2006) for a detailed discussion of available measures of univariate time series forecast accuracy. The MSE, MAPE, and MASE are defined by

$$\text{MSE} = \frac{1}{h_0}\sum_{h=1}^{h_0}(y_h - \hat{q}_h)^2,$$

$$\text{MAPE} = \frac{1}{h_0}\sum_{h=1}^{h_0}\frac{|y_h - \hat{q}_h|}{|y_h|}, \quad \text{and}$$

$$\text{MASE} = \frac{1}{h_0}\sum_{h=1}^{h_0}\left(\frac{|y_h - \hat{q}_h|}{\frac{1}{h-1}\sum_{h=2}^{h_0}|y_h - y_{h-1}|}\right),$$

respectively, where the y_h's are the observed values, and $\hat{q}_h$ are the predicted values for the forecast horizon ($h = 1, \ldots, h_0$). Low values for MSE, MAPE and MASE indicate more accurate predictions.

4.6 APPLICATION

This section presents an empirical application study of the UBXII-ARMA model. The data refer to the proportion of stocked hydroelectric energy in Southeast Brazil available at <http://www.ons.org.br/>. The time series is analyzed in the period of May 2000 to April 2019, thus covering 228 months. The last six observations are reserved for assess the forecasting performance of the model. Thus, the series sample size in the fit is $n = 222$ months. Our interest is to model the median; hence, we set $\tau = 0.5$. We use the programming language R (R Core Team, 2020) to carry out the estimations and computations. The conditional log-likelihood functions are maximized using the quasi-Newton algorithm known as BFGS being the considered starting values, as discussed in Section 4.4.

Table 13 brings some descriptive statistic measures of the monthly average proportions of stocked energy in the Southeast of Brazil. These measures corroborate the histogram of the data presented in Figure 13(a). It is noteworthy that the UBXII distribution can accommodating the negative skewness and negative excess kurtosis presented by the data.

Table 13 – Descriptive statistics of the monthly average proportions of stocked energy in the Southeast of Brazil.

Min.	Median	Mean	Max.	Var.	Asymmetry	Exc. Kurtosis
0.1582	0.5547	0.5464	0.8782	0.0411	−0.0515	−1.2480

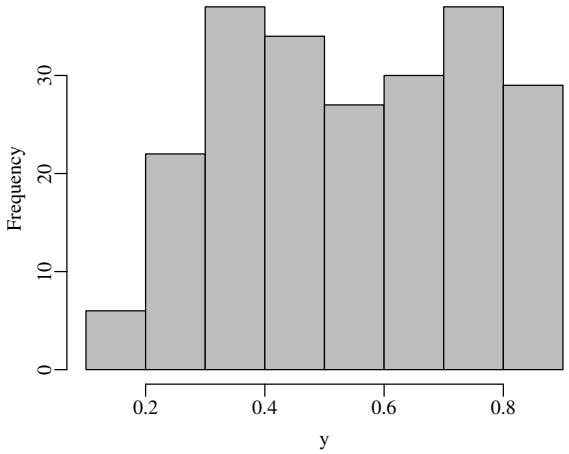
Source: Author (2020)

Figure 13(b) shows evidence of seasonality in the data. The proportions of stocked energy

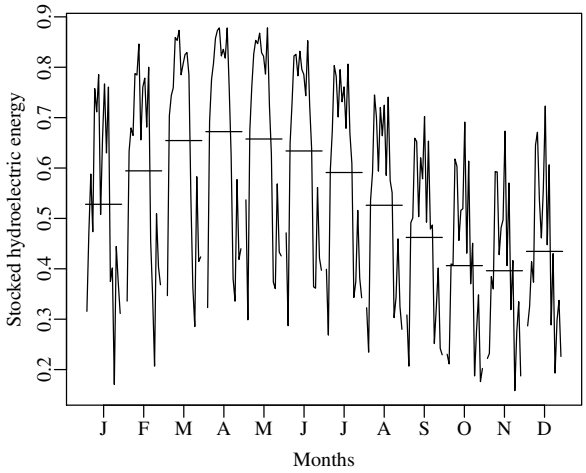
grows until April, after decreases from April to November and again increase from December. A seasonal component can be accommodated of way several. Considering trigonometric functions as covariates is a simple harmonic regression approach; see Bloomfield (2004). We set $\boldsymbol{x}_t = (\cos(2\pi t/12), \sin(2\pi t/12))^\top$ for $t = 1, \ldots, n$, and introduce these covariates in the modeling. Figure 13(c) shows the sample autocorrelation function (ACF) of the time series, whereas Figure 13(d) brings the sample partial autocorrelation function (PACF).

Figure 13 – Observed proportions of stocked hydroelectric energy time series in Southeast of Brazil.

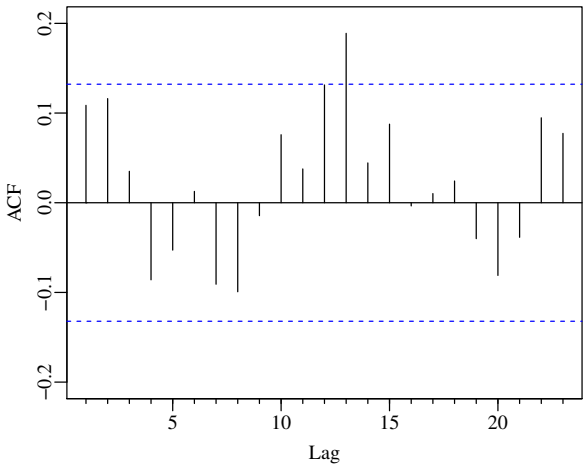
(a) Histogram of the observed data



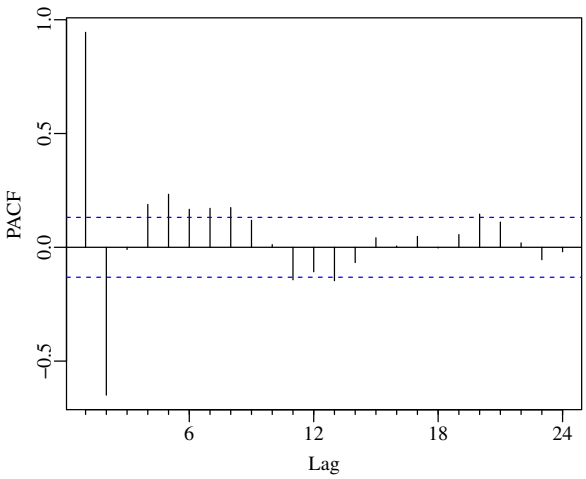
(b) Sazonality



(c) ACF



(d) PACF



Source: Author (2020)

In addition to the UBXII-ARMA model, we also fit the βARMA and KARMA models for comparison purposes. Final models are selected according to the AIC criterion. It was considered all models with autoregressive and moving average dynamics up to the third order and logit link function. The smallest AIC in each class is obtained by the UBXII-AR(2), βAR(2),

and KARMA$(2,2)$ models. Table 14 gives the parameter estimates, standard errors, z statistic value, and p-value of the best fits. Note that, as expected both covariates considered to account for this monthly seasonal component are significant to the usual nominal level of 5%.

Table 14 – Fitted UBXII-AR, βAR, and KARMA models for the proportion of stocked hydroelectric energy in Southeast Brazil.

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
		UBXII-AR(2)		
α	0.0098	0.0135	0.7287	0.4662
β_1	0.4072	0.0469	8.6817	< 0.0001
β_2	0.1015	0.0410	2.4758	0.0133
ϕ_1	1.3390	0.0412	32.5070	< 0.0001
ϕ_2	-0.4119	0.0427	9.6436	< 0.0001
c	11.2294	0.6365	17.6437	< 0.0001
		βAR(2)		
α	0.0074	0.0098	0.7504	0.4530
β_1	0.6201	0.0415	14.9600	< 0.0001
β_2	0.1804	0.0407	4.4287	< 0.0001
ϕ_1	1.3777	0.0513	26.8705	< 0.0001
ϕ_2	-0.4177	0.0517	8.0776	< 0.0001
φ	190.5798	18.1319	10.5107	< 0.0001
		KARMA(2,2)		
α	0.0420	0.0193	2.1728	0.0298
β_1	0.9483	0.0790	12.0013	< 0.0001
β_2	0.2514	0.0984	2.5544	0.0106
ϕ_1	1.3254	0.1783	7.4349	< 0.0001
ϕ_2	-0.4164	0.1636	2.5455	0.0109
θ_1	0.3197	0.1675	1.9084	0.0563
θ_2	0.1803	0.1080	1.6693	0.0951
φ	14.8340	0.7321	20.2628	< 0.0001

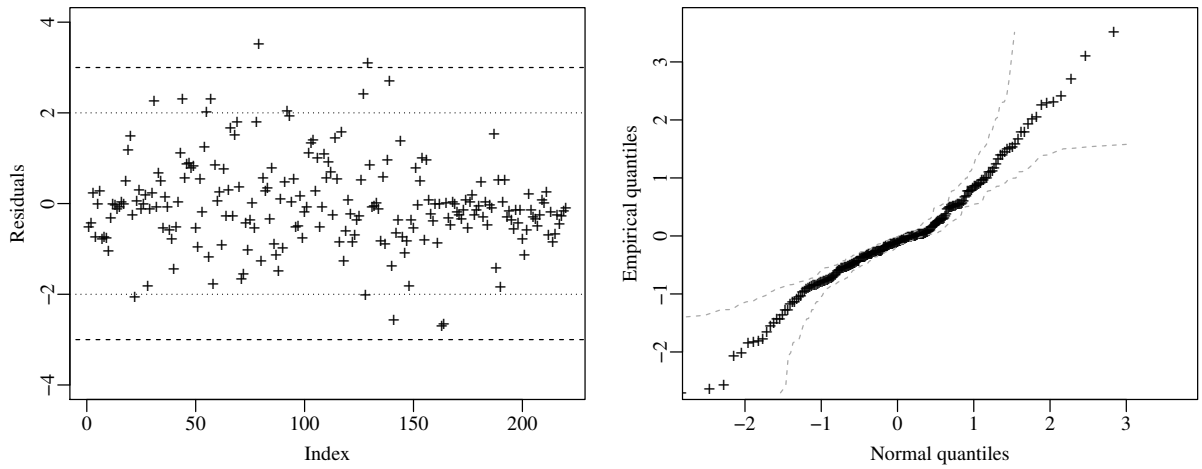
Source: Author (2020)

Two residual diagnostic plots are displayed in Figure 14. In the residuals plot against time (Figure 14(a)), we note that the points distribution has no strong tend, and its behavior is similar to the white noise. The QQ-plot from Figure 14(b) indicates that the quantile residuals have approximately the standard normal distribution. Moreover, from residual ACF and PACF functions, (Figures 14(c) and (d)), it is possible to check the residual white noise hypothesis visually. All these plots and analysis show the fitted UBXII-AR(2) model can be used for out-of-sample forecasting.

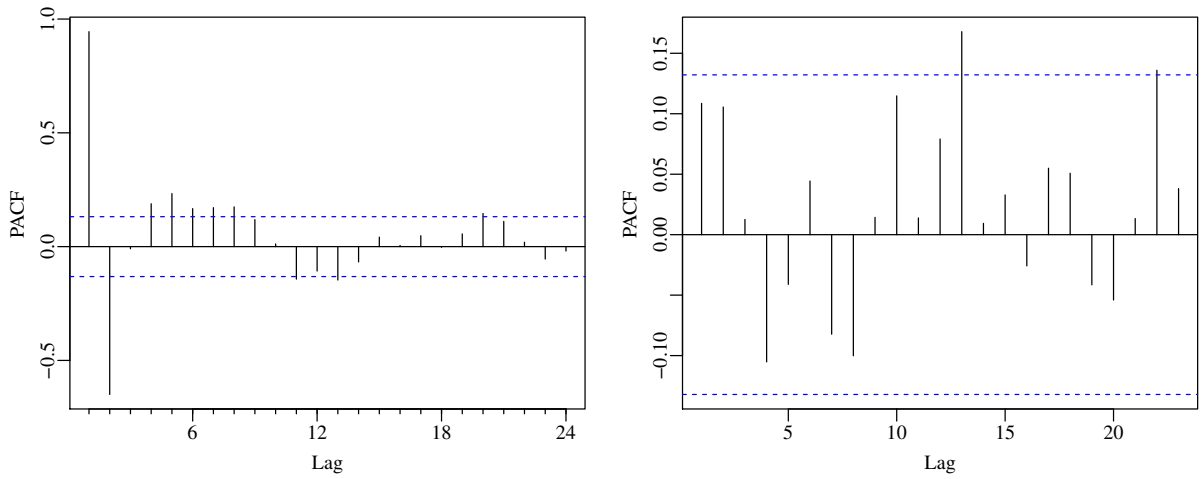
Figure 15(a) gives a plot of the actual values (solid lines) and predicted values (dashed lines) from the fitted UBXII-AR(2) model. Note that the proposed model provides accurate forecasts since the fitted values are quite close to observed data overtime. That is, the UBXII-

Figure 14 – Residual diagnostic plots of the fitted UBXII-AR model for proportion of stocked hydroelectric energy in Southeast Brazil.

(a) Randomized quantile residuals against index (b) QQ-plot



(c) Residual ACF (d) Residual PACF



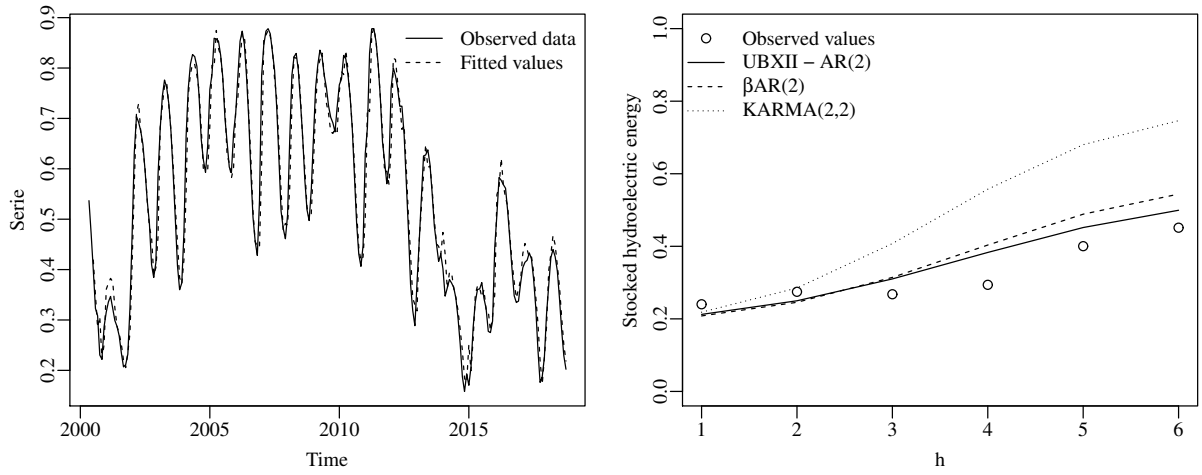
Source: Author (2020)

AR(2) model is suitable to capture the energy stocked proportion data dynamics satisfactorily. Similarly, in the out-of-sample forecasting comparison, the new dynamic model has the best performance; see Figure 15(b).

Table 15 provides values of the accuracy measures (defined in the previous section) for the three final fitted models. According to the results, the UBXII-AR(2) provides the best forecasting since the MSE, MAPE, and MASE presented the smallest values.

Figure 15 – Forecasting performance plots from the UBXII-AR(2) model.

(a) Observed energy stocked proportion and fitted values (b) Out-of-sample forecasting comparison



Source: Author (2020)

Table 15 – Forecasting performance comparison among different the best fitted models in each class

	UBXII-AR(2)	βAR(2)	KARMA(2,2)
MSE	0.0053	0.0127	0.1558
MAPE	0.1505	0.2019	0.4834
MASE	1.0531	1.4777	3.7512

Source: Author (2020)

4.7 CONCLUSION

This chapter proposes a new class of dynamic models: the unit Burr XII autoregressive moving average (UBXII-ARMA) model. This class is appropriate for modeling and forecasting continuous dependent variables over time that assume values in the interval $(0,1)$ such as rates, proportions, and indexes. The new model is very versatile for modeling data of this type since the UBXII density assumes different shapes depending on the values of its parameters. We also perform point estimation by the conditional maximum likelihood method and interval estimation based on asymptotic properties of the conditional maximum likelihood estimators (CMLEs). We derive closed-form expressions for the score function. We assess the finite-sample performance of the CMLE in the UBXII-ARMA framework through Monte Carlo simulation studies. The results show that the CMLEs performs very well, even for small sample sizes.

Moreover, we present some diagnostic analysis and forecasting tools to check whether the proposed model captures the data dynamics. An application study to a dataset regarding the proportion of stocked hydroelectric energy in Southeast of Brazil is presented and discussed.

According to some goodness-of-fit measures, the UBXII-ARMA model has outperformed the KARMA and βARMA models for this dataset. Thus, the introduced model yields the best out-of-sample forecasts for the proportion of stocked hydroelectric energy in Southeast of Brazil in the considered forecast horizon.

5 CONCLUDING REMARKS

This dissertation investigated two transformations of the Burr XII distribution, proposing new probability distributions and its regression models to analyze continuous random variables that assume values in the unit interval. Furthermore, we also introduce a new dynamic model for time series data in the interval $(0, 1)$. In Chapter 2, we define the unit Burr XII (UBXII) distribution and its associated regression model. Some mathematical and statistical properties are investigated. Dropout proportions in Brazilian undergraduate animal sciences courses are modeled by the UBXII regression model that provides more accurate predictions than the beta regression model. In Chapter 3, from the reflection of the random variable UBXII, we propose the reflexive unit Burr XII (RUBXII) distribution and the RUBXII regression. This new model is quite competitive to Kumaraswamy regression and suitable to analyze mortality rates by COVID-19 in the United States. For both regression models, we conduct Monte Carlo simulation studies to assess the maximum likelihood estimators' finite-sample performance. We also provided different tools to perform diagnostic analysis and model selection. Chapter 4 includes autoregressive and average moving components additively to the UBXII regression and defines a time series model useful to analyze dependent variables that assume values in the unit interval called UBXII autoregressive average moving (UBXII-ARMA). From Monte Carlo simulation studies we conducted, it is noteworthy that the conditional maximum likelihood estimators of the parameters that index the model has a good finite-sample performance. The application's to the real data results, computed from the traditional quality measures of prediction, indicate that our proposal provides more accurate forecasts than those provided by the beta autoregressive average moving and Kumaraswamy autoregressive average moving models.

Until now, the contribution of this dissertation is listed below.

- **Chapter 3**: Ribeiro, T. F., Cordeiro, G. M., Peña-Ramírez, F. A., and Guerra, R. R. A new regression model for the COVID-19 mortality rates in the United States. *Statistics in Medicine*. Under Review.

In future work, we shall address the following issues:

- computing the expected information matrices for the UBXII regression, RUBXII regression, and UBXII-ARMA models.
- developing the related asymptotic theory to three proposed models.
- defining the class of inflated UBXII and RUBXII regression models in zeros and ones.
- proposing a RUBXII autoregressive average moving (RUBXII-ARMA) model for time

series data.

- Proposing the class of inflated RUBXII-ARMA and UBXII-ARMA models in zeros and ones.
- introducing the generalized autoregressive score (GAS) with random components being the UBXII and RUBXII distributions.
- applying the novel proposed time serie models to the data of COVID-19.

REFERENCES

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: AKADEMIAI KAIDO. **2nd International Symposium on Information Theory, 1973**. [S.l.], 1973. p. 267–281.

ANDERSEN, M. Early evidence on social distancing in response to COVID-19 in the United States. **Available at SSRN 3569368**, 2020.

BARNDORFF-NIELSEN, O. E.; JØRGENSEN, B. Some parametric models on the simplex. **Journal of multivariate analysis**, Elsevier, v. 39, n. 1, p. 106–116, 1991.

BASHIR, M. F.; MA, B.; KOMAL, B.; BASHIR, M. A.; TAN, D.; BASHIR, M. *et al.* Correlation between climate indicators and COVID-19 pandemic in New York, USA. **Science of The Total Environment**, Elsevier, v. 728, p. 1–4, 2020.

BAYER, F. M.; BAYER, D. M.; PUMI, G. Kumaraswamy autoregressive moving average models for double bounded environmental data. **Journal of Hydrology**, Elsevier, v. 555, p. 385–396, 2017.

BAYES, C. L.; BAZÁN, J. L.; CASTRO, M. D. A quantile parametric mixed regression model for bounded response variables. **Statistics and its interface**, International Press of Boston, v. 10, n. 3, p. 483–493, 2017.

BENJAMIN, M. A.; RIGBY, R. A.; STASINOPOULOS, D. M. Generalized autoregressive moving average models. **Journal of the American Statistical Association**, Taylor & Francis, v. 98, n. 1, p. 214–223, 2003.

BERGER, J. B. The influence of the organizational structures of colleges and universities on college student learning. **Peabody Journal of Education**, Taylor & Francis, v. 77, n. 3, p. 40–59, 2002.

BLOOMFIELD, P. **Fourier analysis of time series: an introduction**. [S.l.]: John Wiley & Sons, 2004.

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. **Journal of Econometrics**, North-Holland, v. 31, n. 3, p. 307–327, 1986.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C. **Time series analysis: forecasting and control**. [S.l.]: John Wiley & Sons, 2011.

BROCKWELL, P. J.; DAVIS, R. A.; FIENBERG, S. E. **Time series: theory and methods**. [S.l.]: Springer Science & Business Media, 1991.

BURR, I. W. Cumulative frequency functions. **The Annals of Mathematical Statistics**, JSTOR, v. 13, n. 2, p. 215–232, 1942.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in Medicine**, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001.

CARRATURO, F.; GIUDICE, C. D.; MORELLI, M.; CERULLO, V.; LIBRALATO, G.; GALDIERO, E.; GUIDA, M. Persistence of SARS-CoV-2 in the environment and COVID-19 transmission risk from environmental matrices and surfaces. **Environmental Pollution**, Elsevier, v. 265, p. 1–6, 2020.

Centers for Disease Control and Prevention. 2020. Online. Accessed 14 August 2020. Available at: <https://www.cdc.gov/>.

CHAVEZ, N. **What's happened in American schools since reopening**. 2020. Online. Accessed 27 August 2020. Available at: <https://www.ctvnews.ca/health/coronavirus/what-s-happened-in-american-schools-since-reopening-1.5065769>.

CHOI, B. **ARMA model identification**. [S.l.]: Springer Science & Business Media, 2012.

COMMISSION, W. Municipal Health *et al.* **Report of clustering pneumonia of unknown etiology in Wuhan City**. 2019.

COX, D. R.; GUDMUNDSSON, G.; LINDGREN, G.; BONDESSON, L.; HARSAAE, E.; LAAKE, P.; JUSELIUS, K.; LAURITZEN, S. L. Statistical analysis of time series: Some recent developments. **Scandinavian Journal of Statistics**, JSTOR, v. 8, n. 2, p. 93–115, 1981.

CRIBARI-NETO, F.; SOUZA, T. C. Religious belief and intelligence: Worldwide evidence. **Intelligence**, Elsevier, v. 41, n. 5, p. 482–489, 2013.

DAVIS, R. A.; DUNSMUIR, W. T.; STREETT, S. B. Observation-driven models for poisson counts. **Biometrika**, Oxford University Press, v. 90, n. 4, p. 777–790, 2003.

de Jonge, E.; WIJFFELS, J.; van der Laan, J. **ffbase: Basic Statistical Functions for Package 'ff'**. [S.l.], 2020. R package version 0.12.8.

DEB, S.; CACCIOLA, S.; STEIN, M. **Sports Leagues Bar Fans and Cancel Games Amid Coronavirus Outbreak. New York Times**. 2020. Online. Accessed 27 August 2020. Available at: <https://www.nytimes.com/2020/03/11/sports/basketball/warriors-coronavirus-fans.html>.

DEHBI, H.-M.; CORTINA-BORJA, M.; GERACI, M. Aranda-ordaz quantile regression for student performance assessment. **Journal of Applied Statistics**, Taylor & Francis, v. 43, n. 1, p. 58–71, 2016.

DOREMALEN, N. V.; BUSHMAKER, T.; MORRIS, D. H.; HOLBROOK, M. G.; GAMBLE, A.; WILLIAMSON, B. N.; TAMIN, A.; HARCOURT, J. L.; THORNBURG, N. J.; GERBER, S. I. *et al.* Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. **New England Journal of Medicine**, Mass Medical Soc, v. 382, n. 16, p. 1564–1567, 2020.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.

DURBIN, J.; KOOPMAN, S. J. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. **Biometrika**, Oxford University Press, v. 84, n. 3, p. 669–684, 1997.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

FRANCO, G. C.; MIGON, H. S.; PRATES, M. O. *et al.* Time series of count data: A review, empirical comparisons and data analysis. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 33, n. 4, p. 756–781, 2019.

FREYTAS-TAMURA, K. d.; ROJAS, R.; FINK, S. **Florida Breaks U.S. Coronavirus Record for Most New Cases in a Day. New York Times**. 2020. Online. Accessed 14 August 2020. Available at: <https://www.nytimes.com/2020/07/12/us/florida-coronavirus-covid-cases.html>.

GHARPURE, R.; HUNTER, C. M.; SCHNALL, A. H.; BARRETT, C. E.; KIRBY, A. E.; KUNZ, J.; BERLING, K.; MERCANTE, J. W.; MURPHY, J. L.; GARCIA-WILLIAMS, A. G. Knowledge and practices regarding safe household cleaning and disinfection for COVID-19 prevention – United States, May 2020. **Morbidity and Mortality Weekly Report**, Centers for Disease Control and Prevention, v. 20, n. 10, p. 705–709, 2020.

GHOSH, A. Robust inference under the beta regression model with application to health care studies. **Statistical Methods in Medical Research**, SAGE Publications Sage UK: London, England, v. 28, n. 3, p. 871–888, 2019.

GÓMEZ-DÉNIZ, E.; SORDO, M. A.; CALDERÍN-OJEDA, E. The Log–Lindley distribution as an alternative to the beta regression model with applications in insurance. **Insurance: Mathematics and Economics**, Elsevier, v. 54, p. 49–57, 2014.

GOOD, D. **U.S. now leads world in deaths, passes 20,000 mark. NEWS**. 2020. Online. Accessed 27 August 2020. Available at: <https://www.nbcnews.com/health/health-news/live-blog/2020-04-11-coronavirus-news-n1181761/ncrd1182006#blogHeader>.

GUAN, Y.; ZHENG, B.; HE, Y.; LIU, X.; ZHUANG, Z.; CHEUNG, C.; LUO, S.; LI, P.; ZHANG, L.; GUAN, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. **Science**, American Association for the Advancement of Science, v. 302, n. 5643, p. 276–278, 2003.

GUERRA, R. R.; PEÑA-RAMIREZ, F. A.; PEÑA-RAMIREZ, M. R.; CORDEIRO, G. M. A note on the density expansion and generating function of the Beta Burr XII. **Mathematical Methods in the Applied Sciences**, Wiley Online Library, v. 43, n. 4, p. 1817–1824, 2020.

HERNANDEZ, S.; O'KEY, S.; WATTS, A.; MANLEY, B.; PETTERSSON, H. **Tracking Covid-19 cases in the US. CNN**. 2020. Online. Accessed 27 August 2020. Available at: <https://edition.cnn.com/interactive/2020/health/coronavirus-us-maps-and-cases/>.

HUANG, C.; WANG, Y.; LI, X.; REN, L.; ZHAO, J.; HU, Y.; ZHANG, L.; FAN, G.; XU, J.; GU, X. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. **The Lancet**, Elsevier, v. 395, n. 10223, p. 497–506, 2020.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. **International Journal of Forecasting**, Elsevier, v. 22, n. 4, p. 679–688, 2006.

Institute for Health Metrics and Evaluation (IHME). **COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections**. Seattle, United States of America, 2020.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [S.l.]: Springer, 2013.

JØRGENSEN, B. Proper dispersion models. **Brazilian Journal of Probability and Statistics**, JSTOR, v. 11, n. 2, p. 89–128, 1997.

JØRGENSEN, B. **The theory of dispersion models**. [S.l.]: CRC Press, 1997.

KENNEY, J.; KEEPING, E. Kurtosis. **Mathematics of Statistics**, Van Nostrand, v. 3, p. 102–103, 1962.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. **Statistical Modelling**, Sage Publications Sage CA: Thousand Oaks, CA, v. 3, n. 3, p. 193–213, 2003.

KUMARASWAMY, P. A generalized probability density function for double-bounded random processes. **Journal of Hydrology**, Elsevier, v. 46, n. 1-2, p. 79–88, 1980.

LACHOS, V. H.; CHEN, M.-H.; ABANTO-VALLE, C. A.; AZEVEDO, C. L. Quantile regression for censored mixed-effects models with applications to hiv studies. **Statistics and its interface**, NIH Public Access, v. 8, n. 2, p. 203, 2015.

LEHMANN, E. L.; CASELLA, G. **Theory of point estimation**. 2nd. ed. [S.l.]: Springer, 2011.

LEMONTE, A. J.; BAZÁN, J. L. New class of Johnson distributions and its associated regression model for rates and proportions. **Biometrical Journal**, Wiley Online Library, v. 58, n. 4, p. 727–746, 2016.

LEWIS, S. **Florida surpasses New York in confirmed COVID-19 cases. CBC NEWS**. 2020. Online. Accessed 27 August 2020. Available at: <https://www.cbsnews.com/news/florida-surpasses-new-york-in-confirmed-covid-19-cases/>.

LEWNARD, J. A.; LO, N. C. Scientific and ethical basis for social-distancing interventions against COVID-19. **The Lancet. Infectious Diseases**, Elsevier, v. 20, n. 6, p. 631–633, 2020.

LI, W. K. Time series models based on generalized linear models: some further results. **Biometrics**, JSTOR, v. 50, n. 2, p. 506–511, 1994.

LINDSAY, B. G.; LI, B. On second-order optimality of the observed Fisher information. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 25, n. 5, p. 2172–2199, 1997.

MAZUCHELI, J.; MENEZES, A. F. B.; CHAKRABORTY, S. On the one parameter unit-Lindley distribution and its associated regression model for proportion data. **Journal of Applied Statistics**, Taylor & Francis, v. 46, p. 700–714, 2019.

MAZUCHELI, J.; MENEZES, A. F. B.; FERNANDES, L. B.; OLIVEIRA, R. P. de; GHITANY, M. E. The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates. **Journal of Applied Statistics**, Taylor & Francis, v. 47, n. 6, p. 954–974, 2020.

MCCULLAGH, P.; NELDER, J. **Generalized linear models**. 2. ed. London: Chapman and Hall, 1989.

MITNIK, P. A.; BAEK, S. The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. **Statistical Papers**, Springer, v. 54, n. 1, p. 177–192, 2013.

MOLLALO, A.; VAHEDI, B.; RIVERA, K. M. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. **Science of the Total Environment**, Elsevier, p. 1–8, 2020.

MOORS, J. A quantile alternative for kurtosis. **Journal of the Royal Statistical Society**, Wiley Online Library, v. 37, p. 25–32, 1988.

MOUSA, A.; EL-SHEIKH, A.; ABDEL-FATTAH, M. A gamma regression for bounded continuous variables. **Advances and Applications in Statistics**, v. 49, n. 4, p. 305–326, 10 2016.

MUNSTER, V. J.; KOOPMANS, M.; DOREMALEN, N. van; RIEL, D. van; WIT, E. de. A novel coronavirus emerging in China – key questions for impact assessment. **New England Journal of Medicine**, Mass Medical Soc, v. 382, n. 8, p. 692–694, 2020.

NAGELKERKE, N. J. *et al.* A note on a general definition of the coefficient of determination. **Biometrika**, Oxford University Press, v. 78, n. 3, p. 691–692, 1991.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

PALM, B. G.; BAYER, F. M. Bootstrap-based inferential improvements in beta autoregressive moving average model. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 47, n. 4, p. 977–996, 2018.

PAWITAN, Y. **In all likelihood: statistical modelling and inference using likelihood**. [S.l.]: Oxford University Press, 2001.

PEFFER, P. A. L. Demographics of an Undergraduate Animal Sciences Course and the Influence of Gender and Major on Course Performance. **NACTA Journal**, North American Colleges and Teachers of Agriculture (NACTA), v. 55, n. 1, p. 26–31, 2011.

PEREIRA, G. H. On quantile residuals in beta regression. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 48, n. 1, p. 302–316, 2019.

PEREIRA, T. L.; CRIBARI-NETO, F. Detecting model misspecification in inflated beta regressions. **Communications in Statistics – Simulation and Computation**, Taylor & Francis, v. 43, n. 3, p. 631–656, 2014.

PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. **Numerical Recipes in C: The Art of Scientific Computing**. 2. ed. USA: Cambridge University Press, 1992.

PUMI, G.; RAUBER, C.; BAYER, F. M. Kumaraswamy regression model with aranda-ordaz link function. **TEST**, Springer, p. 1–21, 2020.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Available at: <https://www.R-project.org/>.

RADMANESH, A.; RAZ, E.; ZAN, E.; DERMAN, A.; KAMINETZKY, M. Brain imaging use and findings in COVID-19: a single academic center experience in the epicenter of disease in the United States. **American Journal of Neuroradiology**, Am Soc Neuroradiology, v. 41, n. 7, p. 1179–1183, 2020.

RAMSEY, J. B. Tests for specification errors in classical linear least-squares regression analysis. **Journal of the Royal Statistical Society: Series B**, Wiley Online Library, v. 31, n. 2, p. 350–371, 1969.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape, (with discussion). **Applied Statistics**, v. 54, n. 3, p. 507–554, 2005.

ROCHA, A. V.; CRIBARI-NETO, F. Beta autoregressive moving average models. **TEST**, Springer, v. 18, n. 3, p. 529–545, 2009.

RODRÍGUEZ-MUÑIZ, L. J.; BERNARDO, A. B.; ESTEBAN, M.; DÍAZ, I. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? **Plos One**, Public Library of Science San Francisco, CA USA, v. 14, n. 6, p. 218–796, 2019.

SALLOUM, S. A.; ALSHURIDEH, M.; ELNAGAR, A.; SHAALAN, K. Mining in Educational Data: Review and Future Directions. In: SPRINGER. **Joint European-US Workshop on Applications of Invariance in Computer Vision**. [S.l.], 2020. p. 92–102.

SCHUMAKER, E. **Timeline: How coronavirus got started The outbreak spanning the globe began in December, in Wuhan, China.** 2020. Online. Accessed 14 August 2020. Available at: <https://abcnews.go.com/Health/timeline-coronavirus-started/story?id=69435165>.

SCHWARZ, G. *et al.* Estimating the dimension of a model. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

SHEPHARD, N. **Generalized linear autoregressions**. [S.l.], 1995. Available at: <https://ideas.repec.org/p/nuf/econwp/0008.html>.

SNEYERS, E.; WITTE, K. D. The interaction between dropout, graduation rates and quality ratings in universities. **Journal of the Operational Research Society**, Taylor & Francis, v. 68, n. 4, p. 416–430, 2017.

STEPHENS, M. A. EDF statistics for goodness of fit and some comparisons. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 69, n. 347, p. 730–737, 1974.

TAGHIZADEH-HESARY, F.; AKBARI, H. The powerful immune system against powerful COVID-19: A hypothesis. **Medical Hypotheses**, Elsevier, v. 140, p. 1–3, 2020.

TINTO, V. Theories of student departure revisited. **Higher education: Handbook of theory and research**, v. 2, n. 359-384, 1986.

VIEIRA, R. S.; ARENDS-KUENNING, M. Affirmative action in Brazilian universities: Effects on the enrollment of targeted groups. **Economics of Education Review**, Elsevier, v. 73, p. 101–931, 2019.

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. **Transactions of the American Mathematical Society**, JSTOR, v. 54, n. 3, p. 426–482, 1943.

WANG, J.; TANG, K.; FENG, K.; LV, W. High temperature and high humidity reduce the transmission of COVID-19. **Available at SSRN 3551767**, 2020.

Washington Post. **Where states reopened and cases spiked after the U.S. shutdown. Washington Post**. 2020. Online. Accessed 27 August 2020. Available at: <https://www.washingtonpost.com/graphics/2020/national/states-reopening-coronavirus-map/>.

WATKINS, A. J. On an integral related to the Burr XII distribution. **Communications in Statistics - Theory and Methods**, Taylor & Francis, v. 40, n. 21, p. 3777–3779, 2011.

WICKHAM, H.; AVERICK, M.; BRYAN, J.; CHANG, W.; MCGOWAN, L. D.; FRANçOIS, R.; GROLEMUND, G.; HAYES, A.; HENRY, L.; HESTER, J.; KUHN, M.; PEDERSEN, T. L.; MILLER, E.; BACHE, S. M.; MüLLER, K.; OOMS, J.; ROBINSON, D.; SEIDEL, D. P.; SPINU, V.; TAKAHASHI, K.; VAUGHAN, D.; WILKE, C.; WOO, K.; YUTANI, H. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

WICKHAM, H.; FRANçOIS, R.; HENRY, L.; MüLLER, K. **dplyr: A Grammar of Data Manipulation**. [S.l.], 2020. R package version 0.8.5.

World Health Organization. 2020. Online. Accessed 14 August 2020. Available at: <https://covid19.who.int/region/amro/country/us>.

World Health Organization. **Q&A: How is COVID-19 transmitted?. World Health Organization**. 2020. Online. Accessed 31 August 2020. Available at: <https://www.who.int/news-room/q-a-detail/q-a-how-is-covid-19-transmitted>.

Worldometer. 2020. Online. Accessed 14 August 2020. Available at: <https://www.worldometers.info/coronavirus/country/us/>.

ZEGER, S. L.; QAQISH, B. Markov regression models for time series: a quasi-likelihood approach. **Biometrics**, JSTOR, v. 44, n. 4, p. 1019–1031, 1988.

ZHANG, C. H.; SCHWARTZ, G. G. Spatial disparities in coronavirus incidence and mortality in the United States: an ecological analysis as of May 2020. **The Journal of Rural Health**, Wiley Online Library, v. 36, n. 3, p. 433–445, 2020.

ZHANG, W.; OLTEAN, A.; NICHOLS, S.; ODEH, F.; ZHONG, F. Epidemiology of reopening in the COVID-19 pandemic in the United States, Europe and Asia. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020.

ZOGHBI, A. C.; ROCHA, F.; MATTOS, E. Education production efficiency: Evidence from Brazilian universities. **Economic Modelling**, Elsevier, v. 31, p. 94–103, 2013.

APPENDIX A – OBSERVED INFORMATION MATRICES AND CHAPTER 2 APPLICATION SUPPLEMENT

Initially, this appendix provides the observed information matrix for the unit Burr XII (UBXII) distribution and a detailed calculation of the observed information matrix for the UBXII regression, defined in Chapter 2. After, it presents supplementary information to the application study carried out in Chapter 2.

OBSERVED INFORMATION MATRICES

In what follows, we present the second-order derivatives of the log-likelihood function to the parameters vector $\boldsymbol{\theta}$, of the UBXII distribution. The elements of the matrix $J(\boldsymbol{\theta})$ presented in Section 2.4 are

$$\begin{aligned}
U_{cc} =& -\frac{n}{c^2} - \frac{n\log^2(\log q^{-1})[t(q)-1]}{t(q)\log[t(q)]} + \frac{n\log^2(\log q^{-1})\log^{2c} q^{-1}}{[t(q)]^2\log[t(q)]} + \frac{n\log^2(\log q^{-1})\log^{2c} q^{-1}}{[t(q)]^2\log^2[t(q)]} \\
&- \sum_{i=1}^{n}\frac{\log^2(\log y_i^{-1})[t(y_i)-1]}{[t(y_i)]^2} - \frac{\log\tau^{-1}}{\log[t(q)]}\left[\sum_{i=1}^{n}\frac{\log^2(\log y_i^{-1})\log^c y_i^{-1}}{t(y_i)}\right. \\
&\left.- \sum_{i=1}^{n}\frac{\log^2(\log y_i^{-1})\log^{2c} y_i^{-1}}{[t(y_i)]^2}\right] + \frac{2\log(\log q^{-1})\log\tau^{-1}[t(q)-1]}{t(q)\log^2[t(q)]} \\
&\times \sum_{i=1}^{n}\frac{\log(\log y_i^{-1})[t(y_i)-1]}{t(y_i)} + \left\{\frac{\log^2(\log q^{-1})\log\tau^{-1}[t(q)-1]}{t(q)\log^2[t(q)]}\right. \\
&\left.- \frac{\log^2(\log q^{-1})\log\tau^{-1}\log^{2c} q^{-1}}{[t(q)]^2\log^2[t(q)]} - \frac{2\log^2(\log q^{-1})\log\tau^{-1}\log^{2c} q^{-1}}{[t(q)]^2\log^3[t(q)]}\right\}\sum_{i=1}^{n}\log[t(y_i)],
\end{aligned}$$

$$\begin{aligned}
U_{qq}(\boldsymbol{\theta}) =& \frac{n\,c^2\log^{2c-2} q^{-1}}{q^2\,[t(q)]^2\log[t(q)]} + \frac{n\,c^2\log^{2c-2} q^{-1}}{q^2\,[t(q)]^2\log^2[t(q)]} - \frac{n\,c\,(c-1)\log^{c-2} q^{-1}}{q^2\,t(q)\log[t(q)]} \\
&- \frac{n\,c\log^{c-1} q^{-1}}{q^2\,t(q)\log[t(q)]} + \left[\frac{c\,(c-1)\log\tau^{-1}\log^{c-2} q^{-1}}{q^2\,t(q)\log^2[t(q)]} + \frac{c\log\tau^{-1}\log^{c-1} q^{-1}}{q^2\,t(q)\log^2[t(q)]}\right. \\
&\left.- \frac{2\,c^2\log\tau^{-1}\log^{2c-2} q^{-1}}{q^2\,[t(q)]^2\log^3[t(q)]} - \frac{c^2\log\tau^{-1}\log^{2c-2} q^{-1}}{\{q\,t(q)\log[t(q)]\}^2}\right]\sum_{i=1}^{n}\log[t(y_i)],
\end{aligned}$$

and

$$\begin{aligned}
U_{cq}(\boldsymbol{\theta}) =& \frac{n\,c\log(\log q^{-1})\log^{c-1} q^{-1}}{q\,t(q)\log[t(q)]} + \frac{n\log^{c-1} q^{-1}}{q\,t(q)\log[t(q)]} - \frac{n\,c\log(\log q^{-1})\log^{2c-1} q^{-1}}{q\,[t(q)]^2\log[t(q)]} \\
&- \frac{n\,c\log(\log q^{-1})\log^{2c-1} q^{-1}}{q\,[t(q)]^2\log^2[t(q)]} - \frac{c\log\tau^{-1}\log^{c-1} q^{-1}}{q\,t(q)\log^2[t(q)]}\sum_{i=1}^{n}\frac{\log(\log y_i^{-1})\log^c y_i^{-1}}{t(y_i)} \\
&+ \left[\frac{2\,c\log(\log q^{-1})\log\tau^{-1}\log^{2c-1} q^{-1}}{q\,[t(q)]^2\log^3[t(q)]} + \frac{c\log(\log q^{-1})\log\tau^{-1}\log^{2c-1} q^{-1}}{q\,[t(q)]^2\log^2[t(q)]}\right.
\end{aligned}$$

$$-\frac{\log\tau^{-1}\log^{c-1}q^{-1}}{q\,t(q)\log^2[t(q)]}-\frac{c\log(\log q^{-1})\log\tau^{-1}\log^{c-1}q^{-1}}{q\,t(q)\log^2[t(q)]}\Bigg]\sum_{i=1}^{n}\log[t(y_i)].$$

Next, we obtain the score function and observed information matrix for the parameter vector $(\boldsymbol{\beta}^\top, c)^\top$ from the regression (2.17). First, we obtain the components of the score vector $\boldsymbol{U}$ in Section 2.6.1. Notice that $U_{\boldsymbol{\beta}} = \left[U_{\beta_1}(\boldsymbol{\beta}, c), \ldots, U_{\beta_k}(\boldsymbol{\beta}, c)\right]^\top$ is the first component of $\boldsymbol{U}$. Invoking the chain rule, we have

$$U_{\beta_j} \equiv \frac{\partial\ell(\boldsymbol{\beta}, c)}{\partial\beta_j} = \sum_{i=1}^{n}\left[\frac{\partial\ell_i(q_i, c)}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i}\frac{\partial\eta_i}{\partial\beta_j}\right], \qquad j = 1, \ldots, k,$$

where

$$\frac{\partial\ell_i(q_i, c)}{\partial q_i} = c\,(q_i^\star - q_i^\dagger\, y_i^\star).$$

We have that $\mathrm{d}q_i/\mathrm{d}\eta_i = 1/g'(q_i)$ and $\partial\eta_i/\partial\beta_j = x_{ij}$. Therefore, the vector $U_{\boldsymbol{\beta}} \equiv \partial\ell(\boldsymbol{\beta}, c)/\partial\boldsymbol{\beta}$ can be written in matrix notation as in Equation (2.19).

Differentiating (2.18) with respect to the parameter c leads to

$$\frac{\partial\ell(\boldsymbol{\beta}, c)}{\partial c} = \sum_{i=1}^{n}\frac{\partial\ell_i(q_i, c)}{\partial c} = \sum_{i=1}^{n} y_i^\sharp,$$

which leads to the second component of $\boldsymbol{U}$ given by (2.20).

We obtain the second-order derivatives $\ell(\boldsymbol{\beta}, c)$ with respect to $\boldsymbol{\beta}^\top$ and c, which compose the observed information matrix $\boldsymbol{J}$ from Section 2.6.1. For $j, p = 1, \ldots, k$, using the chain and product rules, we have

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\beta}, c)}{\partial\beta_p\partial\beta_j} &= \sum_{i=1}^{n}\left\{\frac{\partial}{\partial q_i}\left[\frac{\partial\ell(q_i, c)}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i}\right]\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i}\frac{\partial\eta_i}{\partial\beta_j}\frac{\partial\eta_i}{\partial\beta_p}\right\} \\ &= \sum_{i=1}^{n}\left[\frac{\partial^2\ell_i(q_i, c)}{\partial q_i^2}\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i} + \frac{\partial\ell_i(q_i, c)}{\partial q_i}\frac{\partial}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i}\right]\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i}x_{ij}x_{ip},\end{aligned}$$

where $\partial^2\ell_i(q_i, c)/\partial q_i^2 = m_i + p_i\, y_i^\star$ and

$$\frac{\partial}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i} = -\frac{g''(q_i)}{[g'(q_i)]^2}.$$

It follows that

$$\frac{\partial^2\ell(\boldsymbol{\beta}, c)}{\partial\beta_p\partial\beta_j} = \sum_{i=1}^{n}\left\{\left[(m_i + p_i\, y_i^\star)\frac{1}{g'(q_i)} - c\,(q_i^\star - q_i^\dagger y_i^\star)\frac{g''(q_i)}{g'(q_i)^2}\right]\frac{1}{g'(q_i)}x_{ij}x_{ip}\right\}.$$

In order to simplify the notation, we write $\boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ in matrix form as given by Equation (2.21).

For obtaining the components of $\boldsymbol{J}_{c\boldsymbol{\beta}}^{\top}$ component, we set

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, c)}{\partial c\, \partial \beta_j} = \frac{\partial}{\partial c}\left[\frac{\partial \ell(\boldsymbol{\beta}, c)}{\partial \beta_j}\right] = \frac{\partial}{\partial c}\left\{\sum_{i=1}^{n}\left[\frac{\partial \ell_i(q_i, c)}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i}\frac{\partial \eta_i}{\partial \beta_j}\right]\right\} = \sum_{i=1}^{n}\left[\frac{\mathrm{d}q_i}{\mathrm{d}\eta_i}\frac{\partial \eta_i}{\partial \beta_j}\right]\frac{\partial^2 \ell_i(q_i, c)}{\partial c\, \partial q_i}.$$

The right term can be written as

$$\frac{\partial^2 \ell_i(q_i, c)}{\partial c\, \partial q_i} = r_i - s_i\, y_i^{\ddagger} + u_i\, y_i^{\star}.$$

Hence, analogously to the previous calculates, we have

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, c)}{\partial c\, \partial \beta_j} = \sum_{i=1}^{n}\frac{1}{g'(q_i)}x_{ij}(r_i - s_i\, y_i^{\ddagger} + u_i\, y_i^{\star}).$$

Thus, the quantity $\boldsymbol{J}_{c\boldsymbol{\beta}}^{\top}$ written in matrix notation is just given by (2.22).

For calculating the component J_{cc}, we have

$$\frac{\partial^2 \ell(\boldsymbol{\beta}, c)}{\partial\, c^2} = \frac{\partial}{\partial c}\left[\frac{\partial \ell(\boldsymbol{\beta}, c)}{\partial\, c}\right] = \frac{\partial}{\partial c}\left[\sum_{i=1}^{n}\frac{\partial \ell_i(q_i, c)}{\partial c}\right] = \sum_{i=1}^{n}\frac{\partial^2 \ell_i(q_i, c)}{\partial c^2}.$$

The second-order derivative of $\partial \ell_i(q_i, c)$ with respect to c is expressed as

$$\begin{aligned}
\frac{\partial^2 \ell_i(q_i, c)}{\partial c^2} = &-\frac{1}{c^2} - \frac{\log^2(\log q_i^{-1})[t(q_i) - 1]}{t(q_i)\log[t(q_i)]} + \frac{\log^2(\log q_i^{-1})\log^{2c} q_i^{-1}}{[t(q_i)]^2\log[t(q_i)]} + \frac{\log^2(\log q_i^{-1})\log^{2c} q_i^{-1}}{[t(q_i)]^2\log^2[t(q_i)]} \\
&- \frac{\log^2(\log y_i^{-1})[t(y_i) - 1]}{[t(y_i)]^2} - \frac{\log \tau^{-1}}{\log[t(q_i)]}\left[\frac{\log^2(\log y_i^{-1})\log^{c} y_i^{-1}}{t(y_i)}\right. \\
&\left. - \frac{\log^2(\log y_i^{-1})\log^{2c} y_i^{-1}}{[t(y_i)]^2}\right] + \frac{2\log(\log q_i^{-1})\log \tau^{-1}[t(q_i) - 1]\log(\log y_i^{-1})[t(y_i) - 1]}{t(q_i)t(y_i)\log^2[t(q_i)]} \\
&+ \left\{\frac{\log^2(\log q_i^{-1})\log \tau^{-1}[t(q_i) - 1]}{t(q_i)\log^2[t(q_i)]} - \frac{\log^2(\log q_i^{-1})\log \tau^{-1}\log^{2c} q_i^{-1}}{[t(q_i)]^2\log^2[t(q_i)]}\right. \\
&\left. - \frac{2\log^2(\log q_i^{-1})\log \tau^{-1}\log^{2c} q_i^{-1}}{[t(q_i)]^2\log^3[t(q_i)]}\right\}\log[t(y_i)].
\end{aligned}$$

Let $y_i^{\diamond} = \partial^2 \ell_i(q_i, c)/\partial c^2$. Then, we shall define $\boldsymbol{Y}^{\diamond} = \mathrm{diag}\{y_1^{\diamond}, \ldots, y_n^{\diamond}\}$, and obtain a simpler expression for the component J_{cc}, namely

$$J_{cc} = \sum_{i=1}^{n} y_i^{\diamond} = \mathrm{tr}(\boldsymbol{Y}^{\diamond})$$

as expressed in Equation (2.23).

CHAPTER 2 APPLICATION SUPPLEMENT

Henceforth, we provide a supplementary material that contains information to extract the full database and explains the methodology used in preprocessing and cleaning step. Further, we present a table with the description of the data set's variables used in the application and a table with the results from other fitted Kw, UW, and beta regressions for the considered data set.

Data extraction

The data for this study were otained from the publicly-available higher education census (HEC) microdata. Since 1995, the HEC is conducted yearly by the Brazilian National Institute for Educational Studies and Research "Anísio Teixeira" (INEP) and the data are available at <http://portal.inep.gov.br/web/guest/microdados>. It contains information about the Brazilian higher education system divided into four microdata files, each one presenting students, course, professors and education institution variables. Those files are defined as follows:

1. DM_IES: composed by higher education institutions (HEIs) variables such as the institution's code, administrative category, city, and federation unit, among others;
2. DM_CURSO: contains variables about the undergraduate courses such as the course workload, shift (morning, afternoon, night), number of vacancies, among others;
3. DM_ALUNO: contains variables related to the students such as socio-demographic information from the students, course, admision form, among others;
4. DM_DOCENTE: provides variables related to the professors linked to each HEI, such as socio-demographic and career informations, among others.

We are interested in the dropout proportion for animal sciences courses and factors associated with their enrollment and organizational structure. The DM_ALUNO file provides the information to construct the dropout proportion. The other variables are obtained from the DM_IES and DM_CURSO files. The following section describes the data mining tools employed to obtain the final data set.

Preprocessing and cleaning

Data preprocessing and cleaning involves basic operations to collect and filter the necessary information in order to conduct desired statistical analysis. We perform the data filtering in the R programming language (R Core Team, 2020). We use the `ffbase` (de Jonge; WIJFFELS; van der Laan, 2020), `tidyverse` (WICKHAM *et al.*, 2019), and `dplyr` (WICKHAM *et al.*, 2020) packages, necessary to treat a big database. The population is the cohort of the freshmen animal sciences students in academic year 2009. Each of them has a related unique identification code in the DM_ALUNO file, which allows us to follow them up until 2017, or until the dropout/graduate outcome. The variable CO_ALUNO_SITUACAO identify the student's situation in each census. It is from this variable that we build the dropout proportion.

From the CO_ALUNO_SITUACAO variable, we reclassify the students according to

their last register in the census and construct a new variable under the following categories

1. **dropout**: if the student transferred to another course of the same HEI or who detached from the course, originally encoded as 4 and 5, respectively;
2. **graduate**: if the student completed its undergraduate study, originally encoded as 6;
3. **censoring**: students who have situation likely forming, attending, locked enrollment, or deceased, originally encoded as 1, 2, 3, and 7, respectively.

The response variable for the ith course is given by

$$\text{DROPOUT\_PROPORTION}_i = \frac{\text{number of students with dropout outcome in the } i\text{th course}}{\text{number of students with dropout or graduate outcome in the } i\text{th course}},$$

where $i = 1, \ldots, 78$. The students classified as censoring are not considered since none outcome is observed in this group and one course is eliminated since it had no graduated students until 2017. Thus, we obtain 77 observations corresponding to the dropout proportions of Brazilian animal sciences courses with freshmen students in 2009. We join the organizational variables, from the DM_IES and DM_CURSO files in the HEC of 2009, with the dropout proportion. Finally, we select and clean the those covariates to obtain the final data set. In the cleanning process we i) eliminate some variables with missing observations and identification codes; and ii) join some variables to perform data reduction. The final data set contains the dropout proportion and other 42 covariates.

Table 16 provides the nomenclature (nom.) and a brief description of the response variable, and covariates of the final data set.

Table 16 – Response variable and covariates with its respective description

Nom.	Variable	Description
Y	DROPOUT_PROPORTION (response variable)	Dropout proportion from 2009 until 2017 of Brazilian undergraduate animal sciences courses.
$\boldsymbol{x}_1$	ID[1]	Name of the university to which the course belongs.
$\boldsymbol{x}_2$	QT_VAC_MORNING	Quantity of vacancies offered in the morning shift.
$\boldsymbol{x}_3$	IN_ACCESSIBILITY	Dummy variable that equals one if the course guarantees conditions of accessibility for people with disabilities, and zero otherwise.
$\boldsymbol{x}_4$	IN_NIGHT_COURSE	Dummy variable that equals one if the course works on the night shift, and zero otherwise.
$\boldsymbol{x}_5$	IN_LIBRAS_TRANSLATOR	Dummy variable that equals one if the course provides a translator of Brazilian sign language interpreter (LIBRAS), and zero otherwise.
$\boldsymbol{x}_6$	IN_HIGH_RELIEF	Dummy variable that equals one if the course offers adaptation to high relief of graphics, engravings and figures, and zero otherwise.
$\boldsymbol{x}_7$	IN_AUDIO	Dummy variable that equals one if the course has material in audio, and zero otherwise.
$\boldsymbol{x}_8$	IN_BRAILLE	Dummy variable that equals one if the course has material in Braille, and zero otherwise.
$\boldsymbol{x}_9$	IN_ENL_CHARACTER	Dummy variable that equals one if the course offers material with enlarged characters, and zero otherwise.
$\boldsymbol{x}_{10}$	IN_LIBRAS_DISCIPLINE	Dummy variable that equals one if the course provides translator of LIBRAS, and zero otherwise.
$\boldsymbol{x}_{11}$	IN_GUIDE_INTERPRETER	Dummy variable that equals one if the course makes available guides-interpreter, and zero otherwise.
$\boldsymbol{x}_{12}$	IN_LIBRAS_MATERIAL	Dummy variable that equals one if the course has material in LIBRAS, and zero otherwise.
$\boldsymbol{x}_{13}$	IN_SPEECH_SYNTHESIS	Dummy variable that equals one if the course offers a speech synthesis, and zero otherwise.
$\boldsymbol{x}_{14}$	IN_MORNING_COURSE	Dummy variable that equals one if the course works on the morning shift, and zero otherwise.
$\boldsymbol{x}_{15}$	IN_OTHER_ADM_FORMS	Dummy variable that equals one if the course has alternative forms of admission in addition to the regular ones, and zero otherwise.
$\boldsymbol{x}_{16}$	IN_EVENING_COURSE	Dummy variable that equals one if the course works on the evening shift, and zero otherwise.

(It continues)

Table 17 brings the estimates and p-values of the fitted Kw, UW, and beta regressions for the dropout proportion in the Brazilian zootechnics courses between 2009 and 2017.

[1] Identification variable not considered for modeling.

(Continuation)

$\boldsymbol{x}_{17}$	IN_AGREEMENT	Dummy variable that equals one if the course makes available enter through course of agreement for foreign students, and zero otherwise.
$\boldsymbol{x}_{18}$	IN_DIST_LEARNING	Dummy variable that equals one if the classroom course offers distance learning, and zero otherwise.
$\boldsymbol{x}_{19}$	IN_USE_LAB	Dummy variable that equals one if the course uses the laboratories from the HEI, and zero otherwise.
$\boldsymbol{x}_{20}$	NU_COURSE_LOAD	Course load
$\boldsymbol{x}_{21}$	NU_TIME_COMP	Minimum time to complete the course in number of semesters
$\boldsymbol{x}_{22}$	QT_VAC_INTEGRAL	Quantity of vacancies offered in the integral shift.
$\boldsymbol{x}_{23}$	QT_VAC_NIGHT	Quantity of vacancies offered in the night shift.
$\boldsymbol{x}_{24}$	QT_VAC_EVENING	Quantity of vacancies offered in the evening shift.
$\boldsymbol{x}_{25}$	QT_SEL_PROCESS	Number of students who entered in the course through selection process.
$\boldsymbol{x}_{26}$	QT_SEL_OTHER_FORM	Number of students who entered in the course through other selection forms.
$\boldsymbol{x}_{27}$	IN_CAPITAL_HEI	Dummy variable that equals one if the HEI to which the course belongs is located in the capital, and zero otherwise.
$\boldsymbol{x}_{28}$	IN_ADM_CAT_1	Dummy variable that equals one if the HEI to which the course belongs has a federal administrative category, and zero otherwise.
$\boldsymbol{x}_{29}$	IN_ADM_CAT_2	Dummy variable that equals one if the HEI to which the course belongs has a state administrative category, and zero otherwise.
$\boldsymbol{x}_{30}$	IN_ADM_CAT_3	Dummy variable that equals one if the HEI to which the course belongs has a municipal administrative category, and zero otherwise.
$\boldsymbol{x}_{31}$	IN_ADM_CAT_4	Dummy variable that equals one if the HEI to which the course belongs has a private in the strict sense administrative category, and zero otherwise.
$\boldsymbol{x}_{32}$	IN_ACADEM_ORG_1	Dummy variable that equals one if the HEI's academic organization to which the course belongs is university, and zero otherwise.
$\boldsymbol{x}_{33}$	IN_ACADEM_ORG_2	Dummy variable that equals one if the HEI's academic organization to which the course belongs is university center, and zero otherwise.

(Continuation)

$\boldsymbol{x}_{34}$	IN_ACADEM_ORG_3	Dummy variable that equals one if the HEI's academic organization to which the course belongs is college, and zero otherwise.
$\boldsymbol{x}_{35}$	QT_TEC_INCOMP_ELEM	Number of technical-administrative employees of the HEI (of which the ith course is part) with incomplete elementary education.
$\boldsymbol{x}_{36}$	QT_TEC_HIGH_SCHOOL	Number of technical-administrative employees of the HEI (of which the ith course is part) with high school.
$\boldsymbol{x}_{37}$	QT_TEC_HIGHER_EDUC	Number of technical-administrative employees of the HEI (of which the ith course is part) with higher education.
$\boldsymbol{x}_{38}$	QT_TEC_SPEC	Number of technical-administrative employees of the HEI (of which the ith course is part) with specialization.
$\boldsymbol{x}_{39}$	QT_TEC_MASTER	Number of technical-administrative employees of the HEI (of which the ith course is part) with master's education.
$\boldsymbol{x}_{40}$	QT_TEC_DOC	Number of technical-administrative employees of the HEI (of which the ith course is part) with a doctorate.

Source: Author (2020)

Table 17 – Estimates and p-values of the fitted Kw, UW, and beta regressions for the dropout proportion in the Brazilian zootechnics courses.

Parameter	Kw		UW		Beta	
	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)
β_1	0.0379	0.7251	−0.1778	0.2102	−0.0594	0.5954
β_2	0.0067	0.0378	0.0098	0.0004	0.0082	0.0046
β_3	0.4401	0.0095	0.6801	0.0002	0.5136	0.0023
β_4	0.7169	0.1139	0.9107	0.0055	0.8144	0.0397
d_p, β, σ	0.3116	< 0.0001	1.9202	< 0.0001	−0.6556	< 0.0001

Source: Author (2020)

APPENDIX B – SCORE VECTOR AND OTHER FITTED REGRESSIONS

SCORE VECTOR

Next, it is determined the score vector of the log-likelihood function given by Equation (3.9). It is obtained from the first derivative of the log-likelihood function with respect to the $k+1$ unknown parameters which compose the vector $\boldsymbol{\theta}$. That is, it is defined as $U(\boldsymbol{\theta}) := \left(U_{\boldsymbol{\xi}}(\boldsymbol{\theta})^\top, U_c(\boldsymbol{\theta})\right)^\top$, where

$$U_{\xi_j}(\boldsymbol{\theta}) := \frac{\partial \ell(\boldsymbol{\theta})}{\partial \xi_j} = \sum_{i=1}^{n} \left[\frac{\partial \ell_i(q_i, c)}{\partial q_i} \frac{\mathrm{d}q_i}{\mathrm{d}\eta_i} \frac{\partial \eta_i}{\partial \xi_j} \right]$$

and

$$U_c(\boldsymbol{\theta}) := \frac{\partial \ell(\boldsymbol{\theta})}{\partial c} = \sum_{i=1}^{n} \frac{\partial \ell_i(q_i, c)}{\partial c}$$

with $j = 1, \ldots, k$.

To simplify the notation, the following quantities are considered:

$$a_i := -\frac{c \log^{c-1}(1-q_i)^{-1}}{(1-q_i) r(q_i) \exp[r(q_i)]} + \frac{\log(1-\tau)^{-c} \log^{c-1}(1-q_i)^{-1} r(z_i)}{(1-q_i)[r(q_i)]^2 \exp[r(q_i)]}$$

and

$$b_i := \frac{1}{c} + s(z_i) + \frac{s(z_i) \log^c(1-z_i)^{-1} [\log(1-\tau)/r(q_i) - 1]}{\exp[r(z_i)]} - \frac{\log^c(1-q_i)^{-1} s(q_i)}{r(q_i) \exp[r(q_i)]} - \frac{\log(1-\tau) s(q_i) \log^c(1-q_i)^{-1} r(z_i)}{[r(q_i)]^2 \exp[r(q_i)]},$$

where $s(x) = \log[\log(1-x)^{-1}]$. Observe that

$$\frac{\partial \ell_i(q_i, c)}{\partial q_i} = a_i, \quad \frac{\mathrm{d}q_i}{\mathrm{d}\eta_i} = \frac{1}{g'(q_i)}, \quad \frac{\partial \eta_i}{\partial \xi_j} = x_{ij},$$

$$\text{and} \quad \frac{\partial \ell_i(q_i, c)}{\partial c} = d_i.$$

Hence, the score vector's components can be written compactly in matrix notation as

$$U_{\boldsymbol{\xi}}(\boldsymbol{\theta}) = \boldsymbol{X}^\top \boldsymbol{T} \boldsymbol{a} \quad \text{and} \quad U_c(\boldsymbol{\theta}) = \boldsymbol{b}^\top \mathbf{1},$$

where $\boldsymbol{X}$ is an $n \times k$ covariates matrix, whose ith row is $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^\top$, $\boldsymbol{T} = \mathrm{diag}\{1/g'(q_1), \ldots, 1/g'(q_n)\}$, $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$, $\boldsymbol{b} = (b_1, \ldots, b_n)^\top$, and $\mathbf{1}$ is an n-dimensional vector of 1's.

OTHER FITTED REGRESSIONS

Here, Table 18 reports the estimates, and p-values of the final fitted beta, simplex, and UW regressions.

Table 18 – Some final fitted regressions for the MR by coronavirus in the U.S. states.

Beta(μ_i, σ)			Simplex(μ_i, σ^2)			UW(q_i, β)		
Coeff	Estimate	p-value	Coeff	Estimate	p-value	Coeff	Estimate	p-value
Int	−27.8556	< 0.0001	Int	−23.6492	< 0.0001	Int	−17.1853	< 0.0001
PD	0.0022	< 0.0001	GINI	31.8738	< 0.0001	PD	0.0010	< 0.0001
HDI	21.7485	< 0.0001	BEDS	−0.1952	0.0205	GINI	30.2048	< 0.0001
PR	21.5709	< 0.0001	MA	0.1079	0.0001	AT	−0.0311	0.0001
AT	−0.0122	0.0280	T_{60}	1.3915	< 0.0001	T_{120}	1.2311	< 0.0001
T_{60}	1.0731	< 0.0001	T_{90}	1.8532	< 0.0001	β	5.0111	–
T_{90}	1.4274	< 0.0001	T_{120}	2.0838	< 0.0001	–	–	–
T_{120}	1.6113	< 0.0001	σ^2	2.3355	–	–	–	–
σ	−2.2854	–	–	–	–	–	–	–

Source: Author (2020)