Computational and Applied Mathematics (2021) 40:255
https://doi.org/10.1007/s40314-021-01553-z

# A new quantile regression for the COVID-19 mortality rates in the United States

**Tatiane Fontana Ribeiro[1] · Gauss M. Cordeiro[2] · Fernando A. Peña-Ramírez[3] · Renata Rojas Guerra[3]**

Received: 12 April 2021 / Revised: 12 April 2021 / Accepted: 3 June 2021 /
Published online: 29 September 2021


**Abstract**
An outbreak of coronavirus disease 2019 (COVID-19) has quickly spread worldwide from December 2019, thus characterizing a pandemic. Until August 2020, the United States of America (U.S.) accounted for almost one-fourth of the total deaths by coronavirus. In this paper, a new regression is constructed to identify the variables that affected the first-wave COVID-19 mortality rates in the U.S. states. The mortality rates in these states are computed by considering the total of deaths recorded on 30, 90, and 180 days from the 10th recorded case. The proposed regression is compared to the Kumaraswamy and unit-Weibull regressions, which are useful in modeling proportional data. It provides the best goodness-of-fit measures for the mortality rates and explains 76.57% of its variability. The population density, Gini coefficient, hospital beds, and smoking rate explain the median of the COVID-19 mortality rates in these states. We believe that this article's results reveal important points to face pandemic threats by the State Health Departments in the U.S.

**Keywords** COVID-19 · Pandemic · unit interval · unit Burr XII distribution

**Mathematics Subject Classification** 60E05 · 62J99

## 1 Introduction

Coronavirus disease-2019 (COVID-19), initially so-called 2019-nCoV, belongs to the coronavirus family of enveloped positive-strand RNA viruses. This illness infects several species of animals and humans, causing respiratory tract infections, liver, neurological and gas-

Communicated by Rafael Villanueva.

✉ Tatiane Fontana Ribeiro
tatianefr@ime.usp.br

[1] Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil

[2] Department of Statistics, Federal University of Pernambuco, Recife, PE, Brazil

[3] Department of Statistics, Federal University of Santa Maria, Santa Maria, RS, Brazil

trointestinal problems, ranging 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 (Municipal Health Commission et al. 2019) and has quickly spread worldwide.

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 is 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.

In April 2020, due to many cases and deaths by the new coronavirus, New York City had 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 (World Health Organization 2020). Those numbers are equivalent to about 25% of the documented cases total and 22% of deaths by coronavirus globally (World Health Organization 2020).

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 et al. (2020) to explain variations of coronavirus in the whole country, considering several environmental, topographic, socioeconomic, behavioral, and demographic factors as predictor variables. Duhon et al. (2021) estimated the initial growth rate of COVID-19 for all countries of the world. They used a multiple linear regression model to study the association between the initial growth rate and non-pharmaceutical interventions, demographic, social, and climatic factors. Other similar studies can be found in Andersen (2020) and Zhang and Schwartz (2020).

Although several studies have been done regards to pandemic, to our best knowledge, a regression analysis modeling the first-wave coronavirus mortality rate across the 50 U.S. states has no been conducted. Our goal is to analyze how health care resources, demographic, socioeconomic, and behavioral variables affected the first-wave COVID-19 mortality rate in the U.S. to identify which covariates have a more significant influence on the mortality's initial growth by this disease. This information can be helpful to improve decision-making in the area of public health policy. Moreover, the findings can help understand potential future outbreaks in other countries of the world.

In this context, some regressions are fitted to the first-wave coronavirus mortality rates in the 50 American states to determine the demographic, socioeconomic, health care resources, and behavioral 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 Kumaraswamy (Kw) and unit-Weibull (UW) regressions, which are feasible alternatives to model the median of such data. The main advantage of the proposed regression is that it captures the effect of the associated covariate to health care resources and provides the best regression's adequacy measures. Other similar quantile regressions and unit models recently proposed can be found in Gündüz and Korkmaz (2020), Korkmaz (2020a, b), Korkmaz et al. (2021).

The rest of the paper is structured as follows. A new regression to model the mortality rates in the American states is defined in Sect. 2. Further, the estimation of the parameters, a simulation study, and some goodness-of-fit measures to check the proposed regression's adequacy are discussed. Section 3 contains some basic statistics of the data set, performs an analysis by identifying the best regression to fit the mortality rates, and provides some useful findings. Finally, in Sect. 4, some concluding remarks are addressed.

## 2 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)

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

and

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

respectively, where $c > 0$ and $d > 0$ are shape parameters. It is worth noting that $Z$ can also be seen as a reflected transformation on $W$, $Z = 1 - W$, where $W$ is a random variable following a unit Burr XII (UBXII) distribution pioneered by Korkmaz and Chesneau (2021). Hence, the cdf and pdf of the *reflected unit Burr XII (RUBXII) distribution* can be expressed as (for $z \in (0, 1)$)

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

and

$$f_Z(z; c, d) = c\, d\, \frac{(z-1)^{-1} \log^{c-1}(1-z)^{-1}}{[1 + \log^c(1-z)^{-1}]^{d+1}}, \tag{2}$$

respectively. By inverting (1), the quantile function (qf) of $Z$ is

$$Q_Z(u; c, d) = 1 - \exp\left\{-[(1-u)^{-1/d} - 1]^{1/c}\right\}. \tag{3}$$

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

To introducing a systematic component on a location parameter, the RUBXII distribution is re-parameterized in terms of its quantiles. Let $q(\tau) = Q_Z(\tau; c, d)$ be the $\tau$th quantile of $Z$. By evaluating Equation (3) in $\tau$ and inverting for $d$,

$$d = \log(1-\tau)^{-1} / \log\left\{1 + \log^c\left[1 - q(\tau)\right]^{-1}\right\}. \tag{4}$$

Although the quantiles are functions of $\tau$, $q(\tau)$ is denoted just as $q$ to simplify the notation. Then, by replacing (4) in Equations (1) and (2), the cdf and pdf of the RUBXII distribution expressed in terms of a quantile-based parameterization are (for $z \in (0, 1)$)

$$F_Z(z; q, c) = 1 - \left[1 + \log^c(1-z)^{-1}\right]^{\frac{\log(1-\tau)}{\log\left[1+\log^c(1-q)^{-1}\right]}}, \tag{5}$$

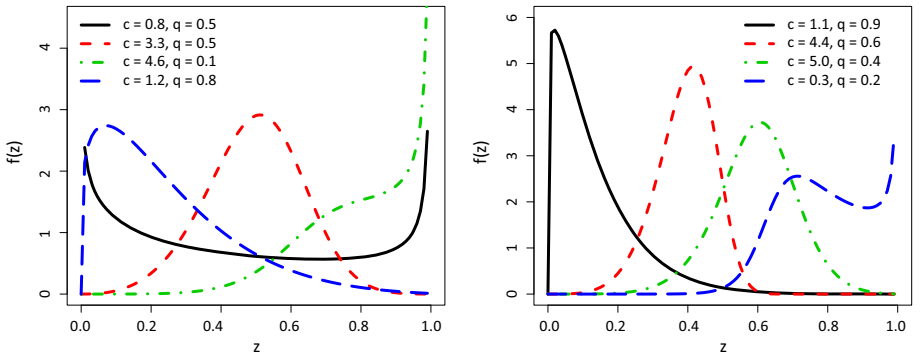


**Fig. 1** Plots of the RUBXII density ($\tau = 0.5$)

and

$$f_Z(z; q, c) = \frac{\log(1-\tau)^{-c}\log^{c-1}(1-z)^{-1}}{(1-z)\log\left[1+\log^c(1-q)^{-1}\right]}\left[1+\log^c(1-z)^{-1}\right]^{\frac{\log(1-\tau)}{\log\left[1+\log^c(1-q)^{-1}\right]}-1}, \tag{6}$$

respectively, where $c > 0$ is a shape parameter and the quantile order $\tau \in (0, 1)$ is chosen by the researcher. Henceforth, let $Z \sim \text{RUBXII}(q, c)$ be a random variable having density (6).

In some cases, median-based regressions are preferable to the mean-based. Median is a more robust measure against the presence of atypical observations and asymmetries at data than the mean. Thus, when data present these features, it is more suitable to consider the median as a measure of location than the mean (Pumi et al. 2020). In the coronavirus mortality rates application of Sect. 3, we consider $\tau = 0.5$, and therefore, $q = q(0.5)$ is the median of $Z$.

Figure 1 displays the RUBXII density plots with $\tau = 0.5$, which have the following forms: U, symmetric, right-skewed, increasing, and increasing-decreasing-increasing (tilde). Thus, it is useful for modeling variables with different types of skewness and heavy tails. Moreover, it can assume shapes (as tilde-shaped) whose densities of classical regressions for modeling unit data do not accommodate.

On the proposed re-parametrization, the qf of $Z$ is

$$Q_Z(u) = 1 - \exp\left\{-\left[(1-u)^{\log[1+\log^c(1-q)^{-1}]/\log(1-\tau)} - 1\right]^{1/c}\right\}. \tag{7}$$

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 $Z_i \sim \text{RUBXII}(q_i, c)$ (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

$$g(q_i) = \eta_i = \sum_{j=1}^{k} x_{ij}\,\xi_j = \boldsymbol{x}_i^\top\,\boldsymbol{\xi}, \tag{8}$$

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

### 2.1 Estimation

The estimation of the parameters of 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

$$\ell(\boldsymbol{\theta}) \equiv \ell(\boldsymbol{\xi}, c) = \sum_{i=1}^{n} \ell_i(q_i, c), \tag{9}$$

where $q_i$ satisfies the systematic component (8) and $\ell_i(q_i, c)$ is the logarithm of the density $f_Z(z_i; q_i, c)$ given in Eq. (6). Thus,

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

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

The components of the score vector $U(\boldsymbol{\theta})$, given in Appendix 1, are defined as the partial derivatives of (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).

### 2.2 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 (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 2021) 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 1 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

**Table 1** Simulation results from the RUBXII regression

| Scenario | $\xi_1$ | $\xi_2$ | $c$ | $n$ | 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 |

estimates of $\xi_1$ and $\xi_2$ 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.

### 2.3 Regression model adequacy

In this section, some methods are presented to analyze whether a fitted regression is suitable for a data set. As goodness-of-fit measures of the RUBXII regression, the maximized log-likelihood value (LL), a normality test for the quantile residuals (Dunn and Smyth 1996), generalized pseudo-$R^2$ ($R^2_G$), and a RESET-type test are considered. The same measures are adopted to compare the proposed regression with other suitable regressions for proportional data.

The quantile residuals for the RUBXII regression are

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

where $F_Z(\cdot)$ is the cdf of the RUBXII distribution given in Eq. (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 quantile residuals is close to the standard normal. To check whether this assumption is satisfied, the well-known Shapiro–Wilk (SW) normality test can be performed.

The $R^2_G$ is useful to assess the proportion of the response variable's variation explained by the regression. It is defined by Nagelkerke (1991) as

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

where $\ell(\hat{\boldsymbol{\theta}}_0)$ is the log-likelihood for the null model, i.e., 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 $R_G^2$ 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 conducted 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 (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 their significance through the likelihood ratio (LR) test.

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

## 3 Results and discussion

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

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

### 3.1 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 CDC (Centers for Disease Control and Prevention 2020). For all states, it is considered the total of deaths per hundred people on 30, 90, and 180 days after the 10th detected case, to ensure that the comparisons are made to the same period. In this way, a panel with three observations for each state is structured.

For all states, the population density, Gini coefficient, hospital beds, smoking rate, poverty rate, and life expectancy, are obtained from the following sources: World Population Review, Global Data Lab, World Atlas, Kaiser Family Foundation, Iowa Community Indicators Program of the Iowa State University, and County Health Rankings and Roadmaps. The response variable and covariates are defined below:

1. MR: Mortality rate (response variable) (Centers for Disease Control and Prevention 2021).
2. PD: Population density (p/mi$^2$) (data of 2020) (World Population Review 2020c).
3. GINI: Gini coefficient (data of 2017) (World Atlas 2017).

**Table 2** Descriptive statistics

| Variable | Statistics | | | | | | |
|---|---|---|---|---|---|---|---|
| | Mean | Median | Skewness | Kurtosis | Min. | Max. | CV(%) |
| MR(30) | 0.0035 | 0.0021 | 2.4538 | 5.2026 | 0.0001 | 0.0191 | 126.2387 |
| MR (90) | 0.0257 | 0.0149 | 2.0870 | 4.0643 | 0.0012 | 0.1375 | 116.6332 |
| MR (180) | 0.0449 | 0.0335 | 1.6478 | 2.9062 | 0.0060 | 0.1800 | 79.7501 |
| PD | 203.9010 | 107.7835 | 2.2110 | 4.4851 | 1.2863 | 1,215.1980 | 130.2652 |
| GINI | 0.4522 | 0.4530 | 0.1339 | −0.4813 | 0.4190 | 0.4990 | 3.9165 |
| BEDS | 2.6000 | 2.4500 | 0.9693 | 0.5641 | 1.6000 | 4.8000 | 27.2984 |
| SR | 0.1733 | 0.1715 | 0.2741 | −0.1108 | 0.0890 | 0.2600 | 20.0528 |
| PR | 0.1323 | 0.1322 | 0.4554 | −0.3877 | 0.0762 | 0.2007 | 21.3240 |
| LE | 78.6960 | 79.1000 | −0.4830 | −0.4259 | 74.8000 | 82.3000 | 2.2677 |

4. BEDS: Hospital beds per 100 thousand inhabitants (data of 2018) (Kaiser Family Foundation 2018).
5. SR: Smoking rate by state (data of 2020) (World Population Review 2020b).
6. PR: Poverty rate (data of 2020) (World Population Review 2020a).
7. LE: Life expectancy (data of 2018) (County Health Rankings & Roadmaps 2018).
8. $T_{90}$: dummy that is equal to one if the response observation corresponds to mortality rate after 90 days of the 10th confirmed case, and zero otherwise.
9. $T_{180}$: dummy that is equal to one if the response observation corresponds to mortality rate after 180 days of the 10th confirmed case, and zero otherwise.

Table 2 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 30 days with a CV of about 126%. Also, in the three time periods (30, 90, and 180 days), the response presents positive skewness, the mean is not close to the median, and at 30 and 90 days its kurtosis is greater than three indicating that it has a leptokurtic distribution. The GINI, and LE covariates have the lowest variabilities with CV ranging between about 2% and 4%. 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 1p/mi$^2$ (referring to the Alaska state) and 1, 215p/mi$^2$, respectively. The BEDS, SR, and PR covariates have close CVs varying from around 21% to 28%. Moreover, they have a mean close to the median, and kurtosis lower than three. Only the LE covariate has negative skewness.

Figure 2 displays the histogram of the MR and box plots from three panel's observations, i.e., MR for 30, 90, and 180 days. The histogram and the three box plots agree to those figures in Table 2. The MR on 30, 90, and 180 days have skewed-right distribution, and it presents some outliers. Clearly, after 90, and 180 days of the 10th recorded case, the mortality rate has increased substantially according to the box plots.

### 3.1.1 Correlation analysis

Initially, we present some dispersion plots of the response variable against each covariate; see Fig. 4. It can be noted that there is no indication of a linear relationship among them. Then Fig. 3 displays the correlation matrix for the current variables by considering the Spearman method. To study the significance of these correlations, it is carried out a Spearman correlation

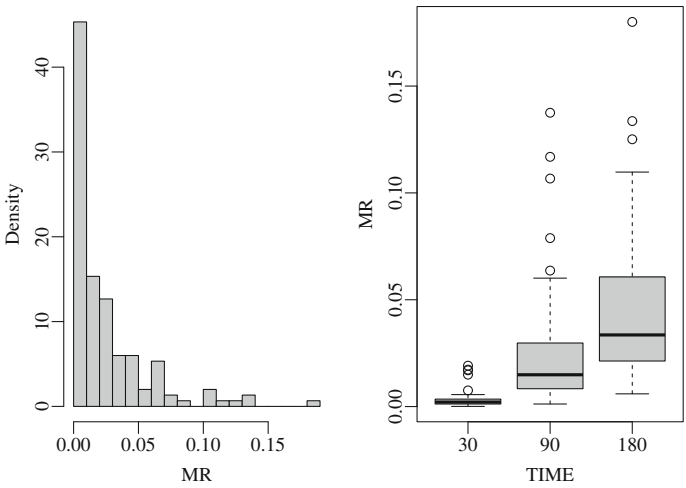


**Fig. 2** Histogram of the MR and box plots of the MR after 30, 90, and 180 days after the 10th confirmed case

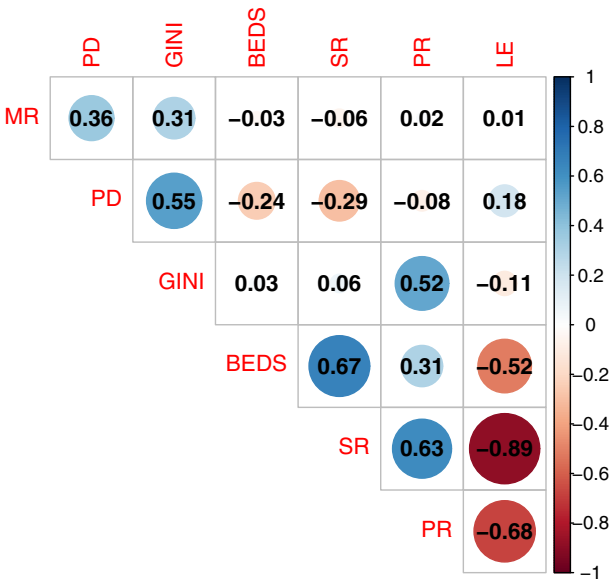


**Fig. 3** Correlation matrix

test and a non-parametric analysis. This test's null hypothesis ($\mathcal{H}_0$) 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 $(n-2)$ degrees of freedom, where $n$ is the sample size. The $p$-values of the test are given in Table 3.

In a first analysis, note that the response variable is positively correlated to PD, presenting the most correlation value with the MR regards to the other covariates (see Figure 3). Moreover, this correlation is significant; see Table 3. Hence, the MR increases as PD. Indeed,

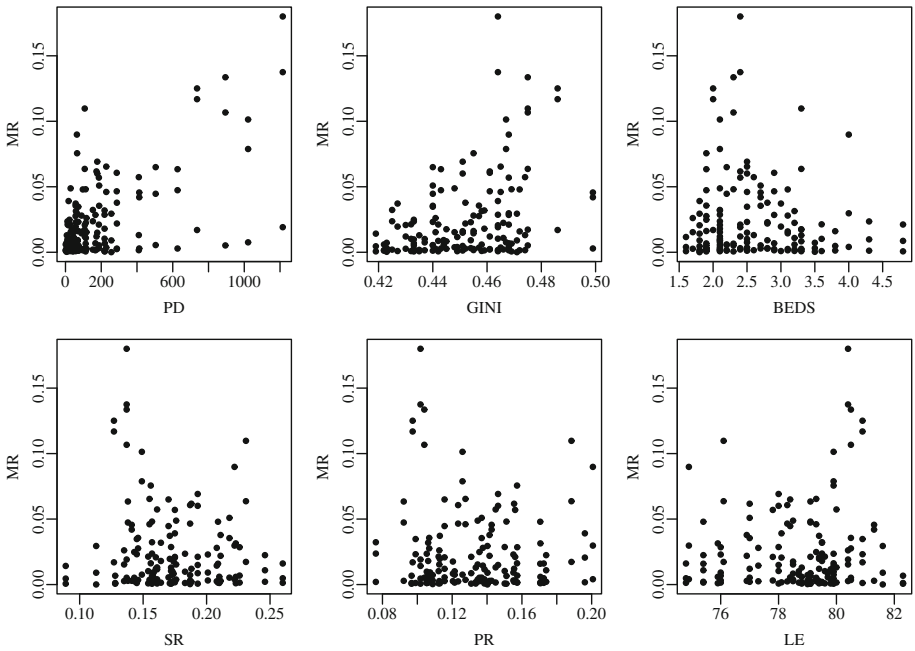


**Fig. 4** Dispersion plots

**Table 3** $p$-values of the Spearman correlation test between all variables

| Variables | MR | PD | GINI | BEDS | SR | PR | LE |
|---|---|---|---|---|---|---|---|
| MR | | $< 0.0001$ | 0.0001 | 0.7088 | 0.4662 | 0.8059 | 0.9015 |
| PD | | | $< 0.0001$ | 0.0028 | 0.0002 | 0.3573 | 0.0255 |
| GINI | | | | 0.6856 | 0.4548 | $< 0.0001$ | 0.1975 |
| BEDS | | | | | $< 0.0001$ | 0.0001 | $< 0.0001$ |
| SR | | | | | | $< 0.0001$ | $< 0.0001$ |
| PR | | | | | | | $< 0.0001$ |
| LE | | | | | | | |

according to Rocklöv and Sjödin (2020), the contact rate by COVID-19 is proportional to population density. Observe also that there is a statistically significant correlation between the MR and the Gini coefficient (Table 3). A similar finding was found in Oronce et al. (2020).

### 3.2 Fitted regressions

In what follows it is explored more deeply the relationship between covariates and the MR through regression analysis. The goodness-of-fit measures are investigated for the RUBXII regression defined in Sect. 2 with two competitive systematic components to study the effects of the covariates given in Sect. 3.1 on the median of the mortality rate by coronavirus in the U.S. states. The well-known Kw regression (Mitnik and Baek 2013) and the UW quantile

**Table 4** Fitted regressions for the median of the MR by COVID-19 in the U.S. states

| Covariate | RUBXII | | Kw | | UW | |
|---|---|---|---|---|---|---|
| | Estimate | $p$-value | Estimate | $p$-value | Estimate | $p$-value |
| Intercept | -13.1507 | $< 0.0001$ | -13.0291 | $< 0.0001$ | -13.4194 | $< 0.0001$ |
| PD | 0.0021 | $< 0.0001$ | 0.0021 | $< 0.0001$ | 0.0020 | $< 0.0001$ |
| GINI | 12.6735 | 0.0007 | 12.7810 | 0.0009 | 13.3897 | 0.0001 |
| BEDS | -0.1905 | 0.0598 | -0.1423 | 0.1774 | 0.0821 | 0.5720 |
| SR | 8.8401 | 0.0003 | 7.3394 | 0.0039 | 3.0877 | 0.1516 |
| $T_{90}$ | 1.8856 | $< 0.0001$ | 1.8695 | $< 0.0001$ | 1.9669 | $< 0.0001$ |
| $T_{180}$ | 2.5751 | $< 0.0001$ | 2.5619 | $< 0.0001$ | 2.8292 | $< 0.0001$ |
| $c$ | 1.5433 | $< 0.0001$ | 0.6408 | $< 0.0001$ | 6.4030 | $< 0.0001$ |

regression (Mazucheli et al. 2020) are considered for comparison purposes. The densities of each competitive regression's random component are given below.

Let $Z$ be a random variable that follows a Kw distribution on median-dispersion parameterization (Mitnik and Baek 2013), say $Z \sim \text{Kw}(q, c)$. Then its pdf is (for $z \in (0, 1)$)

$$f(z; q, c) = \frac{\log 0.5}{c \log(1 - q^{1/c})} z^{1/c} (1 - z^{1/c})^{\log 0.5 / \log(1 - q^{1/c}) - 1}, \tag{10}$$

where $0 < q < 1$ is the median of $Z$ and $c > 0$ is a dispersion parameter.

Recently, Mazucheli et al. (2020) proposed the UW quantile regression. Let $Z \sim \text{UW}(q, c)$ be a random variable having the UW law. Then its pdf is (for $z \in (0, 1)$)

$$f(z; q, c) = \frac{c}{z} \left( \frac{\log \tau}{\log q} \right) \left( \frac{\log z}{\log q} \right)^{c-1} \tau^{(\log z / \log q)^c}, \tag{11}$$

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

Table 4 gives the estimates of the parameters and associated $p$ values of the final fitted RUBXII, Kw, and UW regressions to the coronavirus death rates across the U.S. states. The significance of the estimates is adopted as a criterion to choose the variables in the final fits. The PR and LE covariates were not significant to the usual significance level (1%, 5%, and 10%) at all considered regressions. According to Table 4, when RUBXII regression is fitted, most of the covariates are significant at a significance level of 1%, except for the BEDS, which is significant at 10%. Other fitted regressions do not capture the effect of the covariate BEDS. Besides, the covariate SR is also not statistically significant in the fitted UW regression.

The goodness-of-fit measures of the fitted regressions given in Table 4 are reported in Table 5. The RUBXII regression has the best adequacy measures. It presents the most LL value and $p$-value of SW test upper to the usual nominal level of significance. Further, its $R^2_G$ is the greatest, indicating that the fitted RUBXII regression explains 76.57% of the median response variability. The $p$-value of the SW test for the Kw and UW regressions' residuals are lower than 0.05. Hence, we reject the null hypothesis that the residual distribution is normal at a significance level of 5%. Therefore, these regressions are inadequate to the current data. The $p$-value of the RESET-type (RES) tests indicate that all fitted regressions are specified correctly at usual significance levels. Thus, the results from Table 5 favor the RUBXII more clearly than those Kw and UW regressions by showing its superiority in terms of model fit and significance of the BEDS covariate to the mortality rates by COVID-19 in the U.S. states.

**Table 5** Goodness-of-fit measures for the final fitted regressions

| Regression | LL | $R^2_G$ | $p$-value(SW) | $p$-value(RES) |
|---|---|---|---|---|
| RUBXII | 524.0359 | 0.7657 | 0.1122 | 0.7203 |
| Kw | 523.3649 | 0.7643 | 0.0319 | $> 0.9999$ |
| UW | 521.5016 | 0.7585 | 0.0001 | 0.2498 |

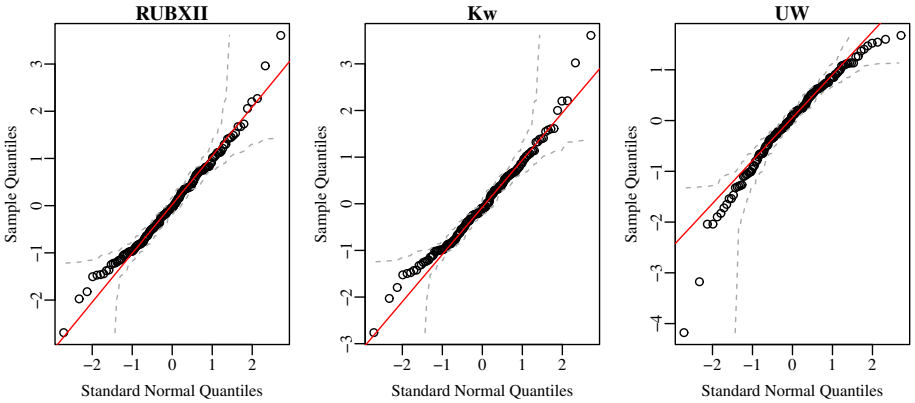


**Fig. 5** Normal Q–Q plot for the quantile residuals of the RUBXII, Kw, and UW fitted regressions

Figure 5 displays a normal Q–Q plot for each fitted regression's quantile residuals to assess if they are normally distributed. The plots corroborate with results from Table 5 by indicating that the RUBXII regression's residuals are more close to a normal distribution since the data points are closely following the straight red line. For the other regressions, mainly the Q–Q plot from the UW regression's residuals, it is possible to note a lack-of-fit of them to the standard normal distribution.

After the above analysis, there is evidence that the RUBXII regression provides a better fit quality. Therefore, from the estimates of the RUBXII regression parameters reported in Table 4, its regression equation can be expressed as

$$\log\left[\hat{q}_i/(1-\hat{q}_i)\right] = -13.1507 + 0.0021\,\mathrm{PD}_i + 12.6735\,\mathrm{GINI}_i - 0.1905\,\mathrm{BEDS}_i + 8.8401\,\mathrm{SR}_i + 1.8856\,\mathrm{T}_{90_i} + 2.5751\,\mathrm{T}_{180_i}.$$

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. Similarly, Wong and Li (2020) showed that population density is an effective predictor of cumulative infection cases in the U.S. at the county level. According to this study, low population density offers a strong protective effect against COVID-19 infection.
- The Gini coefficient is significant at the 1% level, and its positive estimate means that the MR increases in states with a larger Gini coefficient. This finding corroborates with the study of Oronce et al. (2020), who noted that states with higher income inequality had experienced a higher number of deaths by COVID-19.

– 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 increase as expected. According to Janke et al. (2021) U.S. geographic areas with fewer intensive care unit beds, nurses, and general medicine/surgical beds per COVID-19 case were statistically significantly associated with greater deaths in April.
– The SR is mightily significant ($p$-value = 0.0003). The mortality rate's median increases as the SR grows according to the positive signal of its related estimate. This result is expected since the immune response of smoking patients decreases potentially (Taghizadeh-Hesary and Akbari 2020).
– The dummy variables related to the time 90 and 180 days after the 10th confirmed case are significant as expected. As indicated by the box plots in Fig. 2, the MR grows steadily during the considered periods.

## 4 Concluding remarks

The COVID-19 characterizes a pandemic that has been spread across the United States of America (U.S.) since January 2020. This paper investigates how demographic, socioeconomic, health care resources, and behavioral variables are related to the mortality rate by COVID-19 in the U.S. states. To properly reach that aim, it is chosen regressions that consider the double-bounded characteristic of the mortality rate. It is introduced an alternative model called the reflected unit Burr XII (RUBXII) regression, which is a helpful 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 distributed. 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 maximum likelihood estimators' performance 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 regression modeling are done.

In this way, the new regression is compared with the Kumaraswamy and unit-Weibull regressions. The proposed regression is quite competitive compared with those regressions and provides the best fit according to some selection criteria. Thus, from the fitted RUBXII regression, it is possible to identify that the population density, Gini coefficient, hospital beds, and smoking rate are statistically significant in modeling the mortality rate's median by COVID-19 in the U.S. states. This paper's findings may improve understanding of coronavirus in the U.S. and help healthcare system better prepare for the advance of the pandemic or even respond to similar epidemics. Interested readers can access all computational codes at https://github.com/tatianefribeiro/RUBXII_Regression_COVID-19/tree/master. Since the RUBXII regression's potentiality 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

**Acknowledgements** The authors gratefully acknowledge the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) by financial support.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## A: Score vector

In this appendix, it is determined the score vector of the log-likelihood function given by Eq. (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 $i$th row is $\boldsymbol{x}_i = \boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ik})$, $\boldsymbol{T} = \mathrm{diag}\{1/g'(q_1), \ldots, 1/g'(q_n)\}$, $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$, $\boldsymbol{b} = (b_1, \ldots, b_n)^\top$, and $\mathbf{1}$ is an $n$-dimensional vector of 1s.

## References

Andersen M (2020) Early evidence on social distancing in response to COVID-19 in the United States. Available at SSRN 3569368

Bashir MF, Ma B, Komal B, Bashir MA, Tan D, Bashir M et al (2020) Correlation between climate indicators and COVID-19 pandemic in New York, USA. Sci Total Environ 728:138835

Centers for Disease Control and Prevention (2020) https://www.cdc.gov/. Accessed 14 Aug 2020

Centers for Disease Control and Prevention (2021) United States COVID-19 Cases and Deaths by State over Time. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36. Accessed 24 Feb 2021
Municipal Health Commission W, et al. (2019) Report of clustering pneumonia of unknown etiology in Wuhan City
County Health Rankings & Roadmaps (2018) Life expectancy*. https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model/health-outcomes/length-of-life/life-expectancy. Accessed 14 March 2021
Duhon J, Bragazzi N, Kong JD (2021) The impact of non-pharmaceutical interventions, demographic, social, and climatic factors on the initial growth rate of COVID-19: A cross-country study. Sci Total Environ 760:144325
Dunn PK, Smyth GK (1996) Randomized quantile residuals. J Comput Graph Stat 5(3):236–244
Guan Y, Zheng B, He Y, Liu X, Zhuang Z, Cheung C, Luo S, Li P, Zhang L, Guan Y et al (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302(5643):276–278
Guerra RR, Peña-Ramírez FA, Bourguignon M (2020) The unit extended Weibull families of distributions and its applications. J Appl Stat 1–19
Gündüz S, Korkmaz MÇ (2020) A new unit distribution based on the unbounded Johnson distribution rule: the unit Johnson SU distribution. Pak J Stat Oper Res 16(3):471–490
Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 395(10223):497–506
Janke AT, Mei H, Rothenberg C, Becher RD, Lin Z, Venkatesh AK (2021) Analysis of Hospital Resource Availability and COVID-19 Mortality Across the United States. J Hosp Med. https://doi.org/10.12788/jhm.3539
Kaiser Family Foundation (2018) Hospital Beds per 1,000 Population by Ownership Type. 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. Accessed 01 Aug 2020
Korkmaz MÇ (2020a) A new heavy-tailed distribution defined on the bounded interval: the logit slash distribution and its application. J Appl Stat 47(12):2097–2119
Korkmaz MÇ (2020b) The unit generalized half normal distribution: a new bounded distribution with inference and application 82(2):113–140
Korkmaz MÇ, Chesneau C (2021) On the unit Burr-XII distribution with the quantile regression modeling and applications. Comput Appl Math 40(1):1–26
Korkmaz MÇ, Chesneau C, Korkmaz ZS (2021) On the arcsecant hyperbolic normal distribution. Properties, quantile regression modeling and applications. Symmetry 13(1):117
Mazucheli J, Menezes A, Fernandes L, de Oliveira R, Ghitany M (2020) The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates. J Appl Stat 47(6):954–974
Mitnik PA, Baek S (2013) The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. Stat Pap 54(1):177–192
Mollalo A, Vahedi B, Rivera KM (2020) GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Sci Total Environ 728:138884
Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E (2020) A novel coronavirus emerging in China—key questions for impact assessment. N Engl J Med 382(8):692–694
Nagelkerke NJ et al (1991) A note on a general definition of the coefficient of determination. Biometrika 78(3):691–692
Oronce CIA, Scannell CA, Kawachi I, Tsugawa Y (2020) Association between state-level income inequality and COVID-19 cases and mortality in the USA. J Gen Intern Med 35(9):2791–2793
Pereira TL, Cribari-Neto F (2014) Detecting model misspecification in inflated beta regressions. Commun Stat Simul Comput 43(3):631–656
Pumi G, Rauber C, Bayer F (2020) Kumaraswamy regression model with Aranda-Ordaz link function. TEST 1–21
R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
Radmanesh A, Raz E, Zan E, Derman A, Kaminetzky M (2020) Brain imaging use and findings in COVID-19: a single academic center experience in the epicenter of disease in the United States. Am J Neuroradiol 41(7):1179–1183
Rocklöv J, Sjödin H (2020) High population densities catalyse the spread of covid-19. J Travel Med 27(3):taaa038

Taghizadeh-Hesary F, Akbari H (2020) The powerful immune system against powerful COVID-19: a hypothesis. Med Hypotheses 140:109762
Wong DW, Li Y (2020) Spreading of COVID-19: density matters. PLoS ONE 15(12):e0242398
World Atlas (2017) US States by gini coefficient. https://www.worldatlas.com/articles/us-states-by-gini-coefficient.html. Accessed 01 Aug 2020
World Health Organization (2020) https://covid19.who.int/region/amro/country/us. Accessed 14 Aug 2020
World Population Review (2020a) Poverty Rate by State 2021. https://worldpopulationreview.com/state-rankings/poverty-rate-by-state. Accessed 01 Aug 2020
World Population Review (2020b) Smoking Rates by State 2021. https://worldpopulationreview.com/state-rankings/smoking-rates-by-state. Accessed 01 Aug 2020
World Population Review (2020c) US States - Ranked by Population 2020. https://worldpopulationreview.com/states. Accessed 01 Aug 2020
Zhang CH, Schwartz GG (2020) Spatial disparities in coronavirus incidence and mortality in the United States: an ecological analysis as of May 2020. J Rural Health 36(3):433–445