

Article

SimBetaReg Web-Tool: The Easiest Way to Implement the Beta and Simplex Regression Models

Emrah Altun ^{1,*}  and Mahmoud El-Morshedy ^{2,3} ¹ Department of Mathematics, Bartın University, 74100 Bartın, Turkey² Department of Mathematics, College of Science and Humanities in Al-Kharj, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia; m.elmorshedy@psau.edu.sa³ Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt

* Correspondence: emrahaltun@bartin.edu.tr

Abstract: When the response variable is defined on the (0,1) interval, the beta and simplex regression models are commonly used by researchers. However, there is no software support for these models to make their implementation easy for researchers. In this study, we developed a web-tool, named SimBetaReg, to help researchers who are not familiar with programming to implement the beta and simplex regression models. The developed application is free and works independently from the operating systems. Additionally, we model the incidence ratios of COVID-19 with educational and civic engagement indicators of the OECD countries using the SimBetaReg web-tool. Empirical findings show that when the educational attainment, years in education, and voter turnout increase, the incidence ratios of the countries decrease.

Keywords: COVID-19 incidence ratio; beta regression; simplex regression; R Shiny; web-tool



Citation: Altun, E.; El-Morshedy, M. SimBetaReg Web-Tool: The Easiest Way to Implement the Beta and Simplex Regression Models. *Symmetry* **2021**, *13*, 2437. <https://doi.org/10.3390/sym13122437>

Academic Editors: Celia Nunes, Miguel Fonseca and Elsa Moreira

Received: 13 October 2021

Accepted: 8 December 2021

Published: 16 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Bounded data appear in many application areas, such as finance, medical and actuarial sciences (for example, fatal traffic accident ratios, miscarriage pregnancy ratios, and earthquake ratios resulting in loss of life or positive logarithmic returns of investment instruments). To deal with and model these data sets, the beta, Kumaraswamy [1], Topp–Leone [2], and simplex [3] distributions are the first models that come to mind. However, there have been many studies on bounded distributions, especially in the last five years. Two common methods are used to generate new distributions defined on a unit interval, such as log and unit transformations.

These distributions are derived from continuous probability distributions defined on \mathbb{R}^+ . For instance, log-xgamma was derived from the xgamma distribution [4]. Similarly, the log-weighted exponential (log-WE) distribution was derived from the weighted exponential distribution [5]. Some of these distributions can be cited as follows: unit-improved second degree Lindley (unit-ISDL) by Altun and Cordeiro [6], unit-Lindley by Mazucheli et al. [7], log-Bilal by Altun et al. [8], unit-inverse Gaussian by Ghitany et al. [9], transmuted Kumaraswamy by Khan et al. [10], unit-Birnbaum–Saunders by Mazucheli et al. [11], exponentiated Topp–Leone by Pourdarvish [12], and so on. Researchers are still working to generate new distributions for bounded data sets.

When it is desired to explain the change in the dependent variable, defined on the (0,1) interval, by the independent variables, the beta regression model, introduced by Ferrari and Cribari-Neto [13], is the first choice. The second model is the simplex regression, introduced by Kieschnick and McCullough [14], which was extensively studied by Song and Tan [15], Song et al. [16], and Qiu et al. [17]. These models have gained attention from researchers, and several generalizations and alternative models of the beta and simplex distributions have been proposed.

For instance, the unit-Lindley regression model by Mazucheli et al. [7], log-WE, and log-ISDL regression models by Altun [5] and Altun and Cordeiro [6], respectively, unit Burr-XII regression model by Korkmaz and Chesneau [18] and arcsecant hyperbolic normal regression model by Korkmaz et al. [19] are alternative models for the beta and simplex regression models. The open code and data are very important for researchers to reproduce the results given in any scientific study. Even if the codes of the relevant models are made accessible and open, the use of these codes requires partial expertise in R or Python. Therefore, the use of these models has not become widespread.

In the light of the given explanations, the first goal of the presented study is to develop a cloud-based web-tool for the application of the beta and simplex regression models to increase their usage by researchers. We use the R Shiny platform to develop the SimBetaReg web-tool. The name of the application comes from the abbreviations of the beta and simplex regression models. Thanks to the developed application, researchers can easily use the beta and simplex regression models without requiring any software knowledge. In addition, since the developed application is cloud-based, it does not require installation and works independently from the operating systems. As in IBM SPSS, the researchers can upload their data sets and obtain the results of the beta and simplex regression models with residual analysis and goodness of fit statistics.

The second goal of the presented study is to explore the possible relation of the incidence ratio of Coronavirus disease 2019 (COVID-19) with educational and civic engagement indicators for OECD countries.

Many studies have been done by researchers/academicians to analyze the incidence number or incidence ratio of COVID-19. These studies have two different aspects. The first is to forecast the incidence ratio of COVID-19 for a specific region or country by applying the machine learning models. For instance, Mollalo et al. [20] used the artificial neural network to forecast the incidence rate for the United States. However, since only incidence rates were used, the predicted values obtained may deviate from the actual situation.

The second aspect is to seek a correlation or relation between incidence ratio and socio-economic variables. For instance, Karmakar et al. [21] examined the possible relation between incidence and death rates with the social vulnerability index in the United States. Duhon et al. [22] modeled the growth rate of COVID-19 with non-pharmaceutical interventions, social and climatic variables based on the multivariate linear regression.

El-Morshedy et al. [23] modeled the counts of deaths caused by COVID-19 using the count regression models. However, the predictions made by count regression models may lead over-estimated or under-estimated results. Therefore, the modeling of the incidence ratio instead of counts of deaths or cases produces more reliable results. According to the literature review, there is no direct research to model the incidence ratio of COVID-19 with educational and civic engagement indicators. Since the incidence ratio is defined on (0,1) interval, it should be modeled with the beta or simplex regression model by considering the appropriate covariates.

The other parts of the presented study are organized as follows. In Section 2, the mathematical backgrounds of the beta and simplex regression models are summarized. In Section 3, COVID-19 incidence ratio of the OECD countries is modeled with educational and civic engagement indicators by applying the beta and simplex regression models. The results are discussed comprehensively. In Section 4, the implementation of the developed SimBetaReg web-tool is given. The concluding remarks and future works from the presented study are given in Section 5.

2. Regression Models

In this section, we examine the beta and simplex regression models comprehensively.

2.1. Beta Regression

Let Y be a random variable following the beta distribution, defined by

$$f(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1}, \quad (1)$$

where $0 < y < 1$, $\Gamma(\cdot)$ is the gamma function, $\alpha, \beta > 0$. To make the probability distribution function (pdf) in (1) suitable for regression model, it should be re-parametrized based on its mean. Let $\mu = \alpha/(\alpha + \beta)$ and $\phi = \alpha + \beta$ in (1). The resulting pdf is given by

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}. \quad (2)$$

where $0 < \mu < 1$ and $\phi > 0$. Hereafter, the random variable Y is denoted as $Y \sim \text{Beta}(\mu, \phi)$. If the random variable has the pdf in (2), its mean and variance are, respectively

$$\begin{aligned} E(Y) &= \mu, \\ \text{Var}(Y) &= \frac{\mu(1-\mu)}{1+\phi}. \end{aligned}$$

Note that ϕ^{-1} is interpreted as a dispersion parameter since it is an increasing function of the variance. The pdf plots of the beta distribution are plotted in Figure 1. The beta distribution has the following shapes: bathtub, right and left skewed and symmetric.

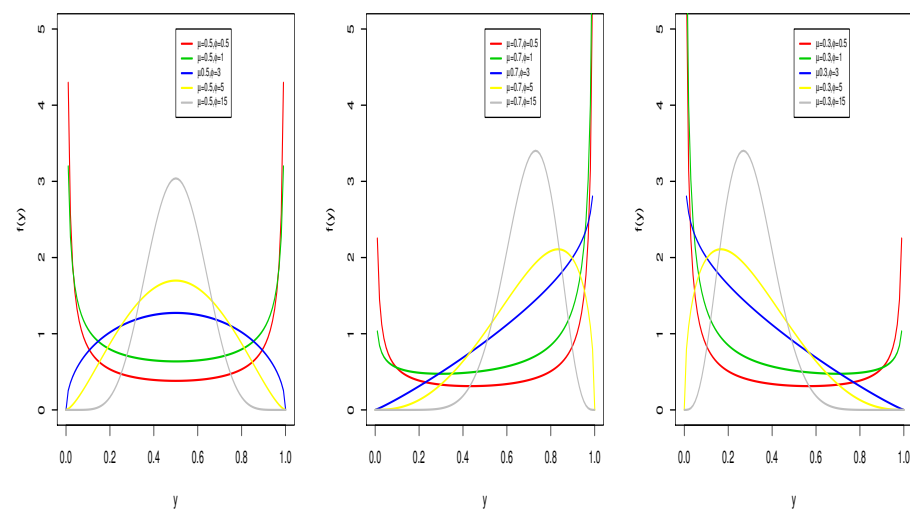


Figure 1. The pdf plots of the beta distribution.

Now, the beta regression model can be defined based on the appropriate link function, which is defined by

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (3)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ is a vector for the regression parameters and $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})$ is a vector for k covariates. The symbol $g(\cdot)$ in (3) represent the link function, such as $g(\cdot) : (0, 1) \rightarrow \mathbb{R}$. Note that the link function is strictly increasing and twice differentiable function. The standard link function in the beta regression model is

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right), \quad (4)$$

which is called the log-link function. Using the inverse of the link function, the variance of the random variable Y is redefined as follows

$$\text{Var}(Y_i) = \frac{g^{-1}(x_i^T \boldsymbol{\beta})(1 - g^{-1}(x_i^T \boldsymbol{\beta}))}{1 + \phi} \quad (5)$$

Based on the density in (2), the log-likelihood function of the beta regression model is given by

$$\begin{aligned} \ell(\boldsymbol{\beta}, \phi) = & \log(\Gamma(\phi)) - \sum_{i=1}^n \log(\Gamma(\mu_i \phi)) - \sum_{i=1}^n \log(\Gamma((1 - \mu_i)\phi)) + \sum_{i=1}^n (\mu_i \phi - 1) \log(y_i) \\ & + \sum_{i=1}^n ((1 - \mu_i)\phi - 1) \log(1 - y_i), \end{aligned} \quad (6)$$

where $\mu_i = \exp(x_i^T \boldsymbol{\beta}) / (1 + \exp(x_i^T \boldsymbol{\beta}))$. The betareg package uses optim function maximizing the log-likelihood in (6) to obtain the maximum likelihood estimate (MLEs) of the unknown parameter vector $(\boldsymbol{\beta}, \phi)$.

2.2. Simplex Regression

Let the random variable Y follows the simplex distribution. The pdf of Y is

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2\{y(1-y)\}^3}} \exp\left\{-\frac{1}{2\sigma^2}d(y; \mu)\right\}, \quad (7)$$

where $0 < y < 1$ and $d(y; \mu)$ is defined as

$$d(y; \mu) = \frac{y(1-y)\mu^2(1-\mu)^2}{(y-\mu)^2}. \quad (8)$$

The random variable having pdf in (7) is denoted by $Y \sim S(\mu, \sigma^2)$. The mean of the simplex distribution is $E(Y) = \mu$ and its variance is given by

$$\text{Var}(Y) = \mu(1-\mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left\{\frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\}, \quad (9)$$

where $\Gamma(\cdot, \cdot)$ is the incomplete gamma function. The pdf plots of the simplex distribution are plotted in Figure 2. The simplex distribution has the following shapes: right and left skewed, symmetric and bimodal shapes.

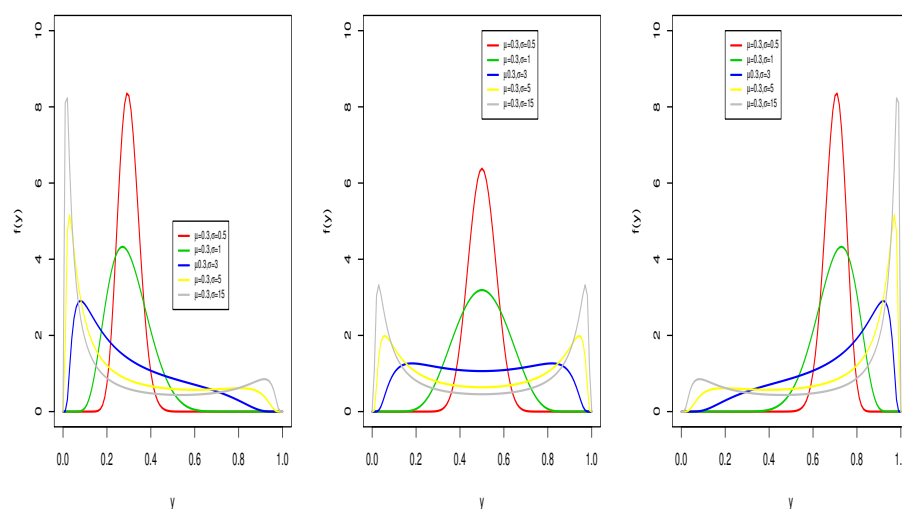


Figure 2. The pdf plots of the simplex distribution.

Using the log-link function given in (3), the log-likelihood function of the simplex regression model is

$$\ell(\beta, \sigma^2) = -\frac{1}{2} \sum_{i=1}^n \log\{2\pi\sigma^2\{y_i(1-y_i)\}^3\} - \sum_{i=1}^n \frac{1}{2\sigma^2} d(y_i; \mu_i). \quad (10)$$

The MLEs of (β, σ^2) are obtained by the VGLM package of the R software, which uses the optim function to maximize the log-likelihood given in (10).

2.3. Randomized Quantile Residuals

The randomized quantile residuals are proposed by Dunn and Smyth [24]. It is defined as

$$r_{q,i} = \Phi^{-1}(u_i), \quad (11)$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the normal distribution and $u_i = F(y_i, \hat{\beta}, \phi)$ for the beta regression and $u_i = F(y_i, \hat{\beta}, \sigma^2)$ for the simplex regression model. The quantile–quantile plot of the residuals as well as goodness-of-fit test can be used to assess the suitability of the fitted model. Kolmogorov–Smirnov (KS) test can be applied to the randomized quantile residuals of the beta and simplex regression models. If the obtained p -value is greater than 0.05%, the randomized quantile residuals follow the standard normal distribution.

2.4. Model Comparison

In general, the R-squared measure is used to know the explanatory power of the fitted model. However, the standard R-squared measure cannot be used for the generalized linear models (GLMs). The pseudo-R-squared measures are available for these types of models. Cox and Snell [25] proposed the following pseudo-R-squared measure for GLMs

$$R_{CS}^2 = 1 - \left(\frac{L_0}{L_M}\right)^{2/n}, \quad (12)$$

where n is the number of observations, L_0 is the likelihood value of the null model with intercept term and L_M is the likelihood value of the fitted model with covariates. The likelihood ratio test is performed to compare the fitted model with null model. The LR test statistic is given by

$$LR = 2(\ell_M - \ell_0), \quad (13)$$

where ℓ_M is the log-likelihood value of the fitted model and ℓ_0 is the log-likelihood value of the null model. The LR statistic is distributed as χ^2 with $df = df_1 - df_2$ where df_1 is the degree of freedom of the null model and df_2 is the degree of freedom of the fitted model with covariates.

3. Empirical Results

In this section, the beta and simplex regression models are used to model the incidence ratio of COVID-19 with some covariates. First, we describe our data set. Then, the estimated parameters of the fitted regression models are given. Later, the model evaluation is done by the residual analysis and goodness-of-fit measures. Finally, we interpret the estimated regression parameters for the best-fitted model.

3.1. Data Set

The data set comes from the indicators of the Better Life Index (BLI), which is calculated for the OECD countries and available in <https://stats.oecd.org/index.aspx?DataSetCode=BLI>. The accessed date for the website is 29 June 2021. Education and civic engagement indicators of the BLI index are selected to model the incidence ratio of

the OECD countries. The research question is “Do the education and civic engagement indicators affect the incidence ratio of COVID-19?”.

We expect that the countries having higher education and civic engagement indicators manage the pandemic process more easily during applying the measures and closures to reduce the incidence of COVID-19. Note that the environmental conditions, government support and amount of resources are also important indicators to help the governments in managing the pandemic process.

The education indicator consists of two variables. These are educational attainment and years in education. Similarly, the civic engagement indicator also has two variables: stakeholder engagement and voter turnout. These four variables are considered covariates. The response variable, the incidence ratio, is calculated as the ratio of the number of positive incidence to the total number of tests. The incidence ratio is calculated based on the available data in (<https://www.worldometers.info/coronavirus/>, date accessed on 29 June 2021).

Figure 3 shows the incidence ratios of different countries. The two countries with the highest incidence ratio are Mexico and Brazil. The three countries with the lowest incidence ratios are New Zealand, Denmark, and Australia. The country with the highest case rate in Europe is Slovenia. The incidence ratios of Germany, France and Italy are very close to each other. The incidence ratio in South Africa are lower than in Poland and Slovenia.

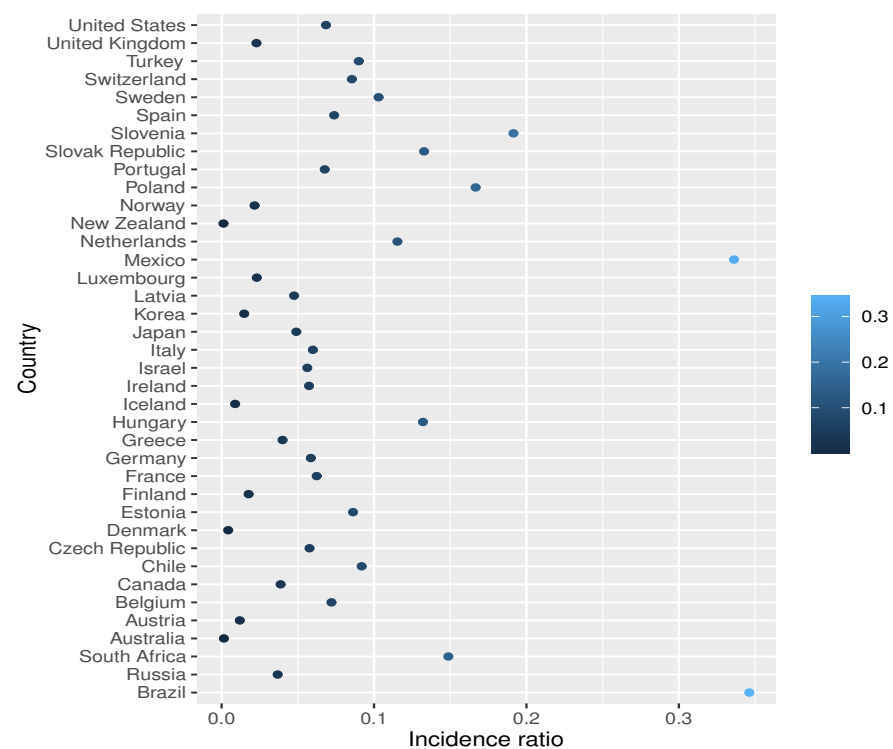


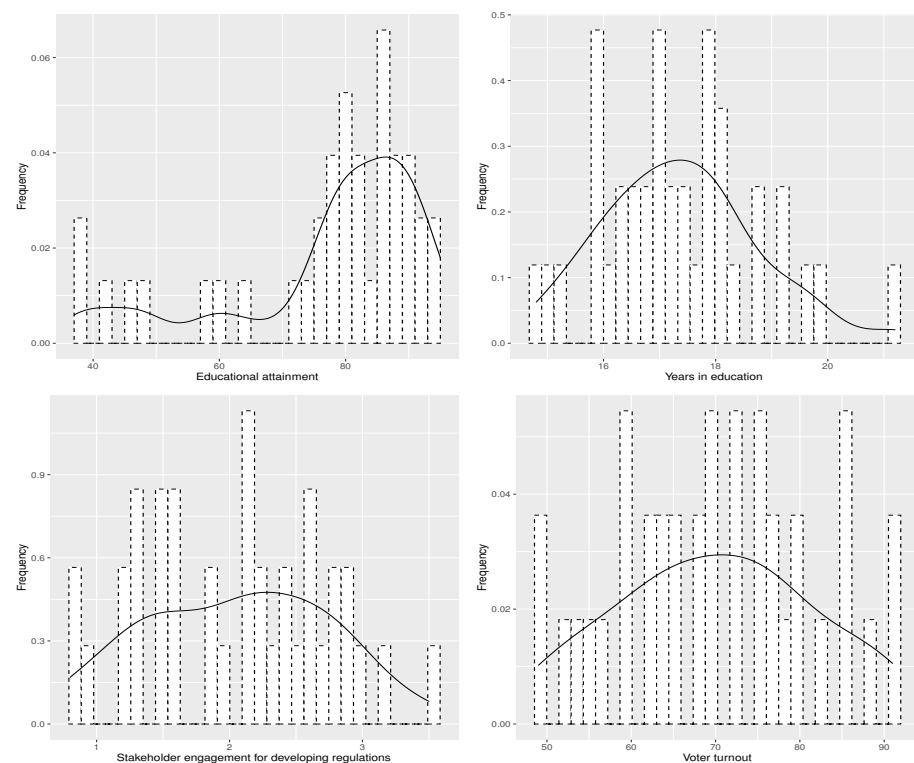
Figure 3. COVID-19 incidence ratio of the OECD countries.

Table 1 shows the descriptive statistics of the variables, such as mean standard deviation (SD), median, minimum and maximum values, range, skewness, and kurtosis. Educational attainment and stakeholder engagement variables are left-skewed and other variables are right-skewed.

Table 1. Summary statistics of the data set.

Variables	n	Mean	SD	Median	Minimum	Maximum	Range	Skewness	Kurtosis
Incidence ratio	38	0.08	0.08	0.06	0	0.35	0.34	1.98	4.10
Educational attainment	38	77.24	16	81.50	37	95	58	−1.23	0.35
Years in education	38	17.38	1.39	17.30	14.80	21.20	6.40	0.44	−0.01
Stakeholder engagement	38	2.05	0.70	2.10	0.80	3.50	2.70	−0.02	−0.96
Voter turnout	38	70.03	11.67	69.50	49	91	42	0.01	−0.93

Figure 4 displays the histogram of the covariates. These figures are also useful to take an information about the characteristics of the data set, such as the skewness and kurtosis.

**Figure 4.** Histograms of the covariates.

3.2. Regression Results

Two regression models are fitted to used data. These are beta and simplex regression models. The statistical backgrounds of the models are presented in Section 2. The dependent variable should be defined on the (0,1) interval. The log-likelihood functions of the regression model given in Section 2 are maximized to obtain the parameter estimates. The incidence ratios of the countries are used as response variable. The covariates are listed below.

- Educational attainment: x_{i1} .
- Years in education: x_{i2} .
- Stakeholder engagement: x_{i3} .
- Voter turnout: x_{i4} .

The fitted regression structure is given by

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}. \quad (14)$$

Since the response variable is defined on the (0, 1) interval, the logit link function is used. Table 2 shows the estimated parameters of the beta and simplex regression models

with corresponding standard errors (SEs) and p -values as well as model selection criteria, such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Based on the calculated AIC and BIC values of the fitted regression model, we conclude that the beta regression model exhibits better modeling ability than the simplex regression model since it has the lowest values of these measures compared with those of the simplex regression model.

Table 2. Results of the fitted regression model for the incidence ratio.

Dependent Variable: Incidence Ratio		Beta			Simplex		
Parameters	Estimates	SEs	p -Value	Estimates	SEs	p -Value	
β_0	4.361	1.572	0.006	12.883	1.780	<0.001	
β_1	−0.021	0.006	0.001	−0.005	0.014	0.714	
β_2	−0.225	0.092	0.015	−0.660	0.131	<0.001	
β_3	0.220	0.148	0.136	−0.678	0.296	0.022	
β_4	−0.027	0.010	0.007	−0.028	0.019	0.129	
ϕ	25.513	6.131	<0.001	—	—	—	
σ	—	—	—	1.978	0.115	<0.001	
$-\ell$		−70.340			−50.723		
AIC		−128.67			−89.4449		
BIC		−118.845			−79.6194		

3.2.1. Comparison

The randomized quantile residuals of the regression models are calculated by using the equation in (11). After that, the quantile–quantile (QQ) plots with simulated envelopes of the randomized quantile residuals for both regression models are plotted in Figure 5. If the model has perfect fit for the data, the plotted points should be on the diagonal line. Thus, it is clear that the plotted points of the beta regression model are more near the diagonal line than those of the simplex regression model.

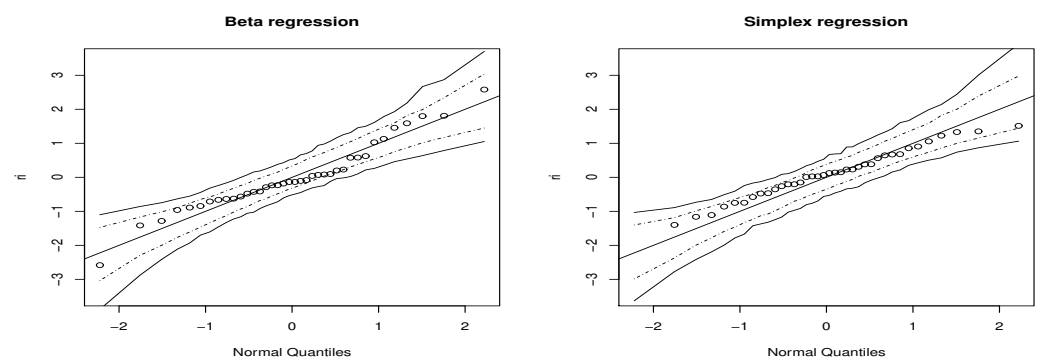


Figure 5. QQ plots of the randomized quantile residuals for the beta and simplex regression models.

To verify the visual results by hypothesis test, the KS test is implemented and results are summarized in Table 3. From these results, it is clear that the normality assumption of the randomized quantile residuals holds for the beta regression model. However, the residuals of the simplex regression model do not hold the normality assumption. It can be evaluated as a piece of evidence for the superiority of the beta regression model over the simplex regression model for the used data set.

Table 3. KS test results of the randomized quantile residuals for the fitted regression models.

Models	KS Test Statistic	<i>p</i> -Value
Beta regression	0.1463	0.355
Simplex regression	0.3525	<0.001

Using the equations in (12) and (13), the Cox–Snell R-squared values and LR test results of the regression models are calculated and reported in Table 4. The pseudo-R-squared value of the beta regression model is higher than the simplex regression model. It is also evident to convince the readers in favor of the beta regression model.

Additionally, the beta regression model provides better results than the null model since the obtained *p*-value of the LR test is $p < 0.001$. However, the simplex regression model does not provide better results than its null model since its *p*-value is higher than 0.05 significance level.

Table 4. Cox–Snell Pseudo R-squared values and LR test results of the fitted regression models.

Models	Cox and Snell Pseudo R^2	LR Test Statistic	<i>p</i> -Value
Beta regression	0.466	23.863	<0.001
Simplex regression	0.202	8.556	0.073

3.2.2. Interpretation

In this subsection, the estimated regression parameters of the beta regression model are interpreted. Educational attainment, years in education, and voter turnout are found statistically significant since their *p*-values are less than 0.05 significance level. In a kind of funny way, when the educational attainment, years in education, and voter turnout increase, the incidence rates of the countries decrease. The stakeholder engagement does not affect the incidence ratios of the countries. These results show that the incidence ratios are lower in countries with higher education levels and democracy. The reason for this situation is that citizens living in these countries comply more with the measures taken during the pandemic process.

Data openness is a crucial issue as all countries have to make information accessible for their institutions. During the pandemic process, this issue has become more popular since some countries have not disclosed COVID-19 cases on time in full. Mak [26] investigated the importance of the data openness for predicting COVID-19 cases based on the five East Asian cities, such as Beijing, Hong Kong, Seoul, Taipei and Tokyo. Mak [26] analyzed the possible relation between pollution and lockdown policies by comparing the pre and post pandemic process and emphasized that the air pollution decreased after COVID-19 waves. Air pollution can be considered as an important explanatory variable to explain the change in COVID-19 cases.

4. SimBetaReg Web-Tool

In this section, we introduce a web-tool of the beta and simplex regression models. The implementations of these models are presented in Section 3. Now, we obtain the same results given in Section 3 with the developed web-tool, SimBetaReg. These models are not available in famous statistical programs, such as IBM SPSS and Minitab. SimBetaReg web-tool is freely accessible at <https://smartstat.shinyapps.io/SimBetaReg/> (date accessed on 1 July 2021). Using the developed application, researchers can do their own analysis by uploading their data sets to the developed web-tool.

Figure 6 shows how to upload one's data set to the SimBetaReg application. Note that the only acceptable data file format is csv (comma-separated values). The separator has to be a semicolon. The first row of the data set should contain the variable names. Using the "Select..." tab, one can easily upload a csv file that contains the data set to be analyzed. The upload data set is displayed in "Data Table" tab. After uploading the data set, users

can select the dependent variable and independent variable(s). Note that the dependent variable should be defined in (0,1). If the dependent variable is not in (0,1) interval, the model does not work.

The screenshot shows the SimBetaReg web application interface. On the left, the 'Load csv file' section displays a file named 'data son.csv' with an 'Upload complete' message. Below this, the 'Bounded regression models' dropdown is set to 'Beta regression'. The 'Go!' button is visible. The 'Select independent variables' dropdown shows 'educat, yearsin, stake, voter' and the 'Select dependent variable' dropdown shows 'cases'. A 'Download Plot' button is at the bottom. On the right, the 'Data Table' tab is active, showing a table with 10 entries for countries like Australia, Austria, Belgium, etc., with columns for country, cases, death, educat, studskil, and year.

Figure 6. Uploading data for SimBetaReg.

Then, selecting the model by “Bounded regression models” and clicking the “GO!” button, users can display the estimated parameters of the model in “Parameter estimates” tab. Figure 7 displays the estimated parameters of the beta regression model obtained by SimBetaReg application. The obtained results by the SimBetaReg web-tool are the same as the results given in Section 3 (see Table 2).

The screenshot shows the SimBetaReg web application interface. On the left, the 'Load csv file' section displays a file named 'data son.csv' with an 'Upload complete' message. Below this, the 'Bounded regression models' dropdown is set to 'Beta regression'. The 'Go!' button is visible. The 'Select independent variables' dropdown shows 'educat, yearsin, stake, voter' and the 'Select dependent variable' dropdown shows 'cases'. A 'Download Plot' button is at the bottom. On the right, the 'Parameter estimates' tab is active, showing the following output:

```
Call:
betareg(formula = reformulate(input$xvar, response = input$yvar), data = data(),
na.action = na.exclude)

Standardized weighted residuals 2:
  Min      IQ  Median      3Q      Max
-3.6281 -0.5007 -0.0509  0.5755  2.4082

Coefficients (mean model with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.360687  1.572425  2.773  0.00555 **
educat      -0.020886  0.006348  -3.290  0.00100 **
yearsin     -0.224761  0.092302  -2.435  0.01489 *
stake       0.220151  0.147629  1.491  0.13590
voter      -0.027337  0.010168  -2.689  0.00717 **

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  25.513    6.131    4.161 3.16e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 70.34 on 6 Df
Pseudo R-squared: 0.4049
Number of iterations: 27 (BFGS) + 6 (Fisher scoring)
```

Figure 7. Displaying the estimated model parameters by SimBetaReg.

The QQ plot of the randomized quantile residuals can be displayed by “QQ plot of the randomized quantile residuals” tab (see Figure 8). This plot can be downloaded in png format by clicking “Download Plot”.

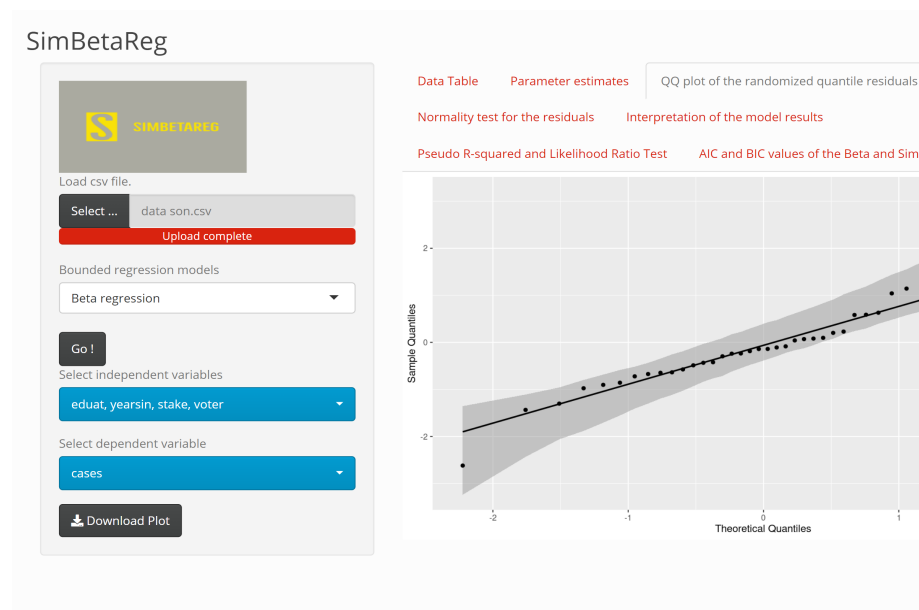


Figure 8. Displaying residual plots by SimBetaReg.

As mentioned before, the randomized quantiles residuals should be distributed as $N(0,1)$ once the model is appropriate for the data set. Thus, it should be checked after fitting the data set. The normality test of the residuals are performed by the KS test, and the results are displayed in “Normality test for the residuals” tab (see Figure 9). The same results are reported in Table 3.

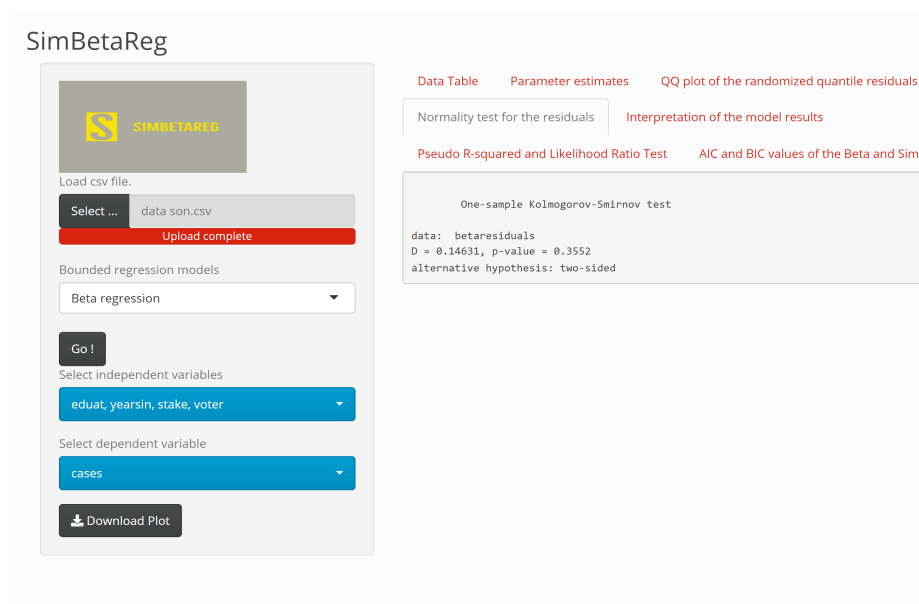


Figure 9. Displaying the KS normality test for residuals by SimBetaReg.

The SimBetaReg interprets the estimated parameters of the models automatically. This property of the SimBetaReg application is very helpful for researchers with less statistical knowledge. This may even be the best specialty of the SimBetaReg application. The interpretations of the estimated parameters are displayed in “Interpretation of the model results” tab (see Figure 10).

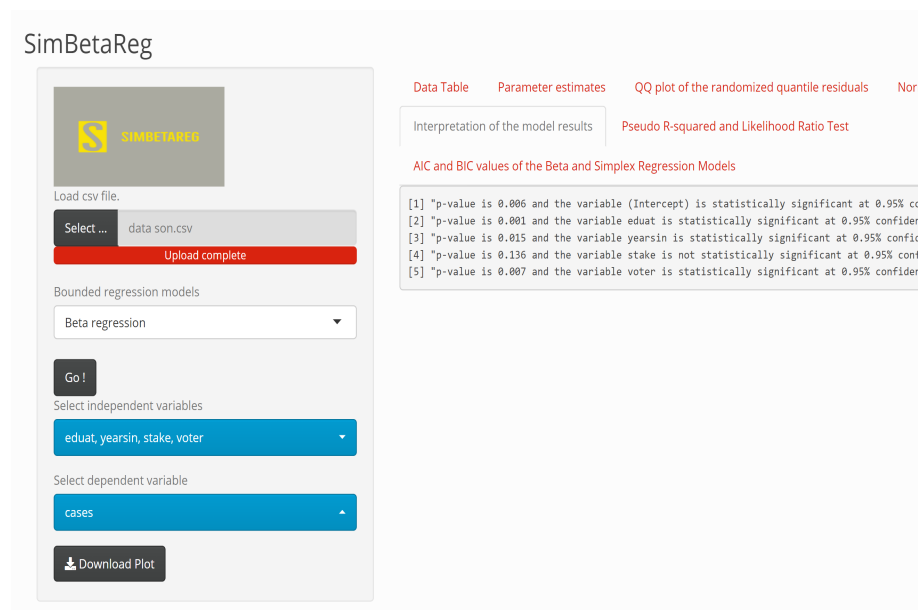


Figure 10. Displaying interpretation of the estimated parameters of the model by SimBetaReg.

The Cox–Snell pseudo R-squared and LR test results are given in “pseudo R-squared and Likelihood Ratio Test” tab (see Figure 11). The pseudo R-squared measure can be used to compare the beta and simplex regression models. The LR test results show the sufficiency of the model against the null model (see Table 4).

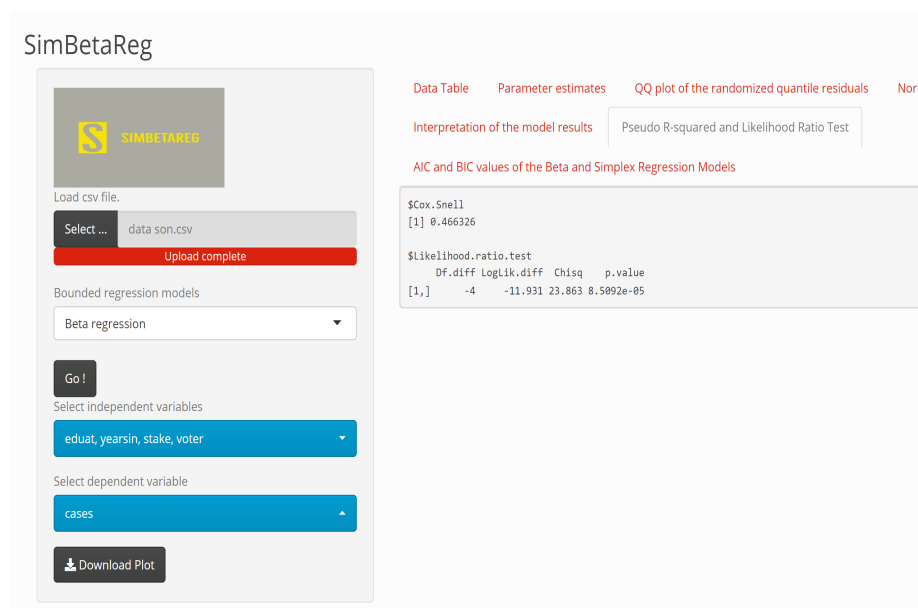


Figure 11. Displaying pseudo R-squared and LR test results by SimBetaReg.

Additionally, the AIC and BIC values of the fitted regression models are given in “AIC and BIC values of the Beta and Simplex Regression Models” tab (see Figure 12). The AIC and BIC statistics are used to select the best model for the data. Thus, the beta regression is selected as the best model in Section 3 (see Table 2).

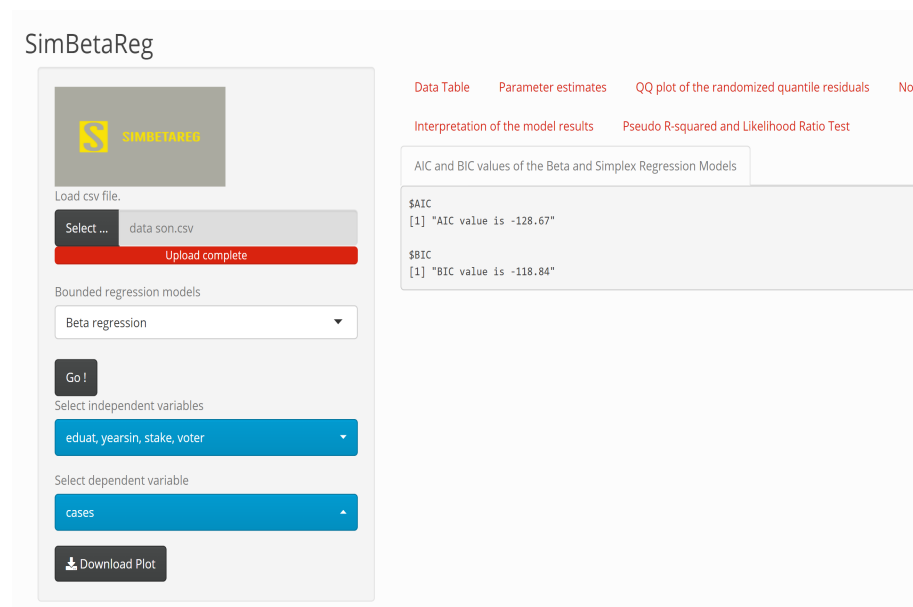


Figure 12. Displaying the AIC and BIC values by SimBetaReg.

Used R Packages

During the development process of the SimBetaReg web-tool, several R packages were used. These were the betareg package by Cribari-Neto and Zeileis [27], VGAM package by Yee [28], ggplot2 package by Wickham [29], gamlss.dist package by Stasinopoulos and Rigby [30] and stats package by R Core Team [31]. The betareg function of the betareg package was used to obtain the parameter estimates of the beta regression model. The vglm function of the VGAM package was used to obtain the parameter estimates of the simplex regression model. The ggplot function of the ggplot2 package was used to draw the QQ plot of the residuals with simulated envelopes. The gamlss.dist package was used for the CDF of the simplex distribution. Finally, the ks.test function of the stats package was used to perform the KS test.

5. Conclusions and Future Work

In this study, we developed a cloud-based web-tool using the R Shiny platform. The developed application, SimBetaReg, is user-friendly and makes the implementation of the beta and simplex regression models easy for researchers and academicians. The parameter estimates, residual analysis, and likelihood ratio tests with pseudo-R-squared values of the beta and simplex regression models are easily obtained by the SimBetaReg web-tool.

Additionally, the incidence ratio of COVID-19 was analyzed with the developed web-tool, and the empirical results are interesting for policy-makers. As future work related to the present study, we plan to develop a new web-tool to model the time-dependent incidence ratio by considering a longitudinal beta regression model. We believe that SimBetaReg will continue to gain attention from researchers, especially those working in the actuarial and medical sciences.

Author Contributions: Conceptualization, E.A.; methodology, E.A.; software, E.A.; validation, E.A.; formal analysis, E.A. and M.E.-M.; investigation, E.A. and M.E.-M.; resources, E.A. and M.E.-M.; data curation, E.A. and M.E.-M.; writing—original draft preparation, E.A. and M.E.-M.; writing—review and editing, E.A. and M.E.-M.; visualization, E.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data set is available upon request from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

COVID-19	Coronavirus disease 2019
pdf	probability distribution function
CDF	cumulative distribution function
MLE	maximum likelihood estimation
unit-ISDL	unit-improved second degree Lindley
log-WE	log-weighted exponential
BLI	Better Life Index
KS	Kolmogorov–Smirnov

References

1. Kumaraswamy, P. A generalized probability density function for double-bounded random processes. *J. Hydrol.* **1980**, *46*, 79–88. [[CrossRef](#)]
2. Topp, C.W.; Leone, F.C. A family of J-shaped frequency functions. *J. Am. Stat. Assoc.* **1955**, *50*, 209–219. [[CrossRef](#)]
3. Barndorff-Nielsen, O.; Jørgensen, B. Some parametric models on the simplex. *J. Multivar. Anal.* **1991**, *39*, 106–116. [[CrossRef](#)]
4. Altun, E.; Hamedani, G.G. The log-xgamma distribution with inference and application. *J. Soc. Française Stat.* **2018**, *159*, 40–55.
5. Altun, E. The log-weighted exponential regression model: Alternative to the beta regression model. *Commun. Stat.-Theory Methods* **2021**, *50*, 2306–2321. [[CrossRef](#)]
6. Altun, E.; Cordeiro, G.M. The unit-improved second-degree Lindley distribution: Inference and regression modeling. *Comput. Stat.* **2019**, *35*, 259–279. [[CrossRef](#)]
7. Mazucheli, J.; Menezes, A.F.B.; Chakraborty, S. On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *J. Appl. Stat.* **2019**, *46*, 700–714. [[CrossRef](#)]
8. Altun, E.; El-Morshedy, M.; Eliwa, M.S. A new regression model for bounded response variable: An alternative to the beta and unit-Lindley regression models. *PLoS ONE* **2021**, *16*, e0245627. [[CrossRef](#)]
9. Ghitany, M.E.; Mazucheli, J.; Menezes, A.F.B.; Alqallaf, F. The unit-inverse Gaussian distribution: A new alternative to two-parameter distributions on the unit interval. *Commun. Stat.-Theory Methods* **2018**, *48*, 3423–3438. [[CrossRef](#)]
10. Khan, M.S.; King, R.; Hudson, I.L. Transmuted kumaraswamy distribution. *Stat. Transit. New Ser.* **2016**, *17*, 183–210. [[CrossRef](#)]
11. Mazucheli, J.; Menezes, A.F.; Dey, S. The unit-Birnbaum–Saunders distribution with applications. *Chil. J. Stat. (ChJS)* **2018**, *9*, 47–57.
12. Pourdarvish, A.; Mirmostafae, S.M.T.K.; Naderi, K. The exponentiated Topp–Leone distribution: Properties and application. *J. Appl. Environ. Biol. Sci.* **2015**, *5*, 251–256.
13. Ferrari, S.; Cribari-Neto, F. Beta Regression for Modelling Rates and Proportions. *J. Appl. Stat.* **2004**, *31*, 799–815. [[CrossRef](#)]
14. Kieschnick, R.; McCullough, B.D. Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Stat. Model. Int. J.* **2003**, *3*, 193–213. [[CrossRef](#)]
15. Song, P.X.-K.; Tan, M. Marginal models for longitudinal continuous proportional data. *Biometrics* **2000**, *56*, 496–502. [[CrossRef](#)] [[PubMed](#)]
16. Song, P.X.K.; Qiu, Z.; Tan, M. Modeling Heterogeneous Dispersion in Marginal Simplex Models for Continuous Longitudinal Proportional Data. *Biom. J.* **2004**, *46*, 540–553. [[CrossRef](#)]
17. Qiu, Z.; Song, P.X.-K.; Tan, M. Simplex Mixed-Effects Models for Longitudinal Proportional Data. *Scand. J. Stat.* **2008**, *35*, 577–596. [[CrossRef](#)]
18. Korkmaz, M.Ç.; Chesneau, C. On the unit Burr-XII distribution with the quantile regression modeling and applications. *Comput. Appl. Math.* **2021**, *40*, 1–26. [[CrossRef](#)]
19. Korkmaz, M.Ç.; Chesneau, C.; Korkmaz, Z.S. On the Arcsecant Hyperbolic Normal Distribution. Properties, Quantile Regression Modeling and Applications. *Symmetry* **2021**, *13*, 117. [[CrossRef](#)]
20. Mollalo, A.; Rivera, K.M.; Vahedi, B. Artificial Neural Network Modeling of Novel Coronavirus (COVID-19) Incidence Rates across the Continental United States. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4204. [[CrossRef](#)]
21. Karmakar, M.; Lantz, P.M.; Tipirneni, R. Association of Social and Demographic Factors with COVID-19 Incidence and Death Rates in the US. *JAMA Netw. Open* **2021**, *4*, e2036462. [[CrossRef](#)] [[PubMed](#)]
22. Duhon, J.; Bragazzi, N.; Kong, J.D. 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.* **2021**, *760*, 144325. [[CrossRef](#)] [[PubMed](#)]
23. El-Morshedy, M.; Altun, E.; Eliwa, M.S. A new statistical approach to model the counts of novel coronavirus cases. *Math. Sci.* **2021**, 1–14. [[CrossRef](#)]
24. Dunn, P.K.; Smyth, G.K. Randomized quantile residuals. *J. Comput. Graph. Stat.* **1996**, *5*, 236–244.

25. Cox, D.R.; Snell, E.J. *Analysis of Binary Data*, 2nd ed.; Chapman Hall: London, UK, 1989.
26. Mak, H.W.L. From COVID-19 Pandemic of Five Selected East Asian Cities to Assessment of Data Openness and Integration for Future City Development. 2021. Available online: https://www.researchgate.net/profile/Hugo-Mak-2/publication/354293725_From_COVID-19_Pandemic_of_Five_Selected_East_Asian_Cities_to_Assessment_of_Data_Openness_and_Integration_for_Future_City_Development/links/612fbc430360302a00734baa/From-COVID-19-Pandemic-of-Five-Selected-East-Asian-Cities-to-Assessment-of-Data-Openness-and-Integration-for-Future-City-Development.pdf (accessed on 1 July 2021).
27. Cribari-Neto, F.; Zeileis, A. Beta Regression in R. *J. Stat. Softw.* **2010**, *34*, 1–24. [CrossRef]
28. Yee, T.W. *Vector Generalized Linear and Additive Models: With An Implementation in R*; Springer: New York, NY, USA, 2015.
29. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
30. Stasinopoulos, M.; Rigby, R. *gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape*. R Package Version 5.3-2. 2021. Available online: <https://CRAN.R-project.org/package=gamlss.dist> (accessed on 1 July 2021).
31. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: <https://www.R-project.org/> (accessed on 1 July 2021).