ISSN: 2754-6705

Journal of Mathematical & Computer Applications

SCIENTIFIC Research and Community

Review Article

Open Access

The Unit Gompertz Regression Model with Applications to COVID-19 Data in Brazil

Lucas David Ribeiro-Reis

Department of Economics, Federal University of Alagoas, Santana do Ipanema-AL, Brazil

ABSTRACT

In this paper, based on the unit Gompertz distribution, a new regression model for the unit interval has been introduced. In this model, the parameterization is done based on the median of the distribution. The estimation of the model parameters is done by maximum likelihood method. Monte Carlo simulations, to show the performance of maximum likelihood estimators are performed. Finally, using Covid-19 data in Brazilian capital cities, the unit Gompertz regression model performance better than the popular beta and Kumaraswamy regression models.

***Corresponding author**

Lucas David Ribeiro-Reis, Department of Economics, Federal University of Alagoas, Santana do Ipanema-AL, Brazil.

Received: June 06, 2024; **Accepted:** June 10, 2024, **Published:** July 30, 2024

Keywords: Unit Gompertz Distribution, Regression Model, Maximum Likelihood, Monte Carlo Simulation, COVID-19, Brazil

Introduction

In applied statistics, there are many phenomena that are in the unit interval, such as rates, proportions, income concentration indices, among others. Seeking to study such phenomena in the context of regression, introduced the beta regression model [1]. This model assumes that the variable response follows beta distribution and its structure is based on the mean. Recently, new papers have appeared, assuming others distributions for the variable response. Some of these regression models are: unit Weibull and unit log-logistic [2-4].

While the beta regression model is parameterized on the mean of the distribution, the unit Weibull and Kumaraswamy regressions models are parameterized in terms of the median of the distribution. Thus, in the presence of very asymmetric data or data presenting outliers, these regressions models are more robust than the beta regression model.

Recently proposed a distribution for the unit interval, called unit Gompertz [5]. So, in this paper based on the unit Gompertz distribution a new regression model for data in $(0,1)$ is proposed. Here, it is assumed that the dependent variable follows the unit Gompertz distribution. Thus, like unit Weibull and Kumaraswamy regression models, this new model has parameterization in the median of the distribution.

Some of the motivations for this new regression model are: (i) this regression is more robust than the regression that has a mean structure, when the data present atypical values; (ii) the model can be used for data with negative or positive skewness; (iii) applications to data on deaths from Covid-19 in Brazil, shows that the model is more appropriate than other models widely used in the literature.

This paper is organized as follows. In Section 2, will present the unit Gompertz distribution, with some properties. In section 3, the regression model is introduced. Parameter estimation by maximum likelihood, the score vector and the observed information matrix are discussed. In Section 4, analysis of diagnostics are presented. In Section 5, Monte Carlo simulations are performed to show the performance of the maximum likelihood estimation for the model parameters. An application to real data on Covid-19 is made in Section 6. Finally, Section 7 presents the final considerations.

Unit Gompertz Distribution

The random variable Y having unit Gompertz distribution has cumulative distribution function (cdf) and probability density function (pdf) given by [5].

$$F(y;\alpha,\lambda)=\exp\left[-\alpha\left(y^{-\lambda}-1\right)\right],\quad 0<y<1$$

and

$$f(y;\alpha,\lambda)=\alpha\lambda y^{-(\lambda+1)}\exp\left[-\alpha\left(y^{-\lambda}-1\right)\right],\quad 0<y<1,\qquad (1)$$

respectively, where $\alpha>0$ and $\lambda>0$. The random variable Y with pdf (1) is denoted by

$$Y\sim UG(\alpha,\lambda)$$

The *k*th moment of is given by [5].

$$\mathbb{E}\left[Y^{k}\right]=\alpha^{k/\lambda}\mathrm{e}^{\alpha}\Gamma\left(1-k/\lambda,\alpha\right),$$

where $\Gamma(p,\alpha)=\int_{\alpha}^{\infty}u^{p-1}\mathrm{e}^{-u}du$ is the upper incomplete gamma function. Therefore, the moments only exist when $k/\lambda<1$ (or $\lambda>k$).

So, the mean and variance *Y* are

$$\mathbb{E}[Y]=\alpha^{1/\lambda}\mathrm{e}^{\alpha}\Gamma(1-1/\lambda,\alpha),\quad \lambda>1$$

and

$$\mathbb{V}[Y] = \alpha^{2/\lambda} e^{\alpha} \{\Gamma(1-2/\lambda, \alpha) - e^{\alpha}\Gamma(1-1/\lambda, \alpha)^2\}, \quad \lambda > 2,$$

respectively. Figure 1 presents some plots of the variance as a function of the parameters, for some values. As can be noted, when the parameter values increase, the variance decreases. Thus, the parameter λ can be interpreted as a precision parameter.

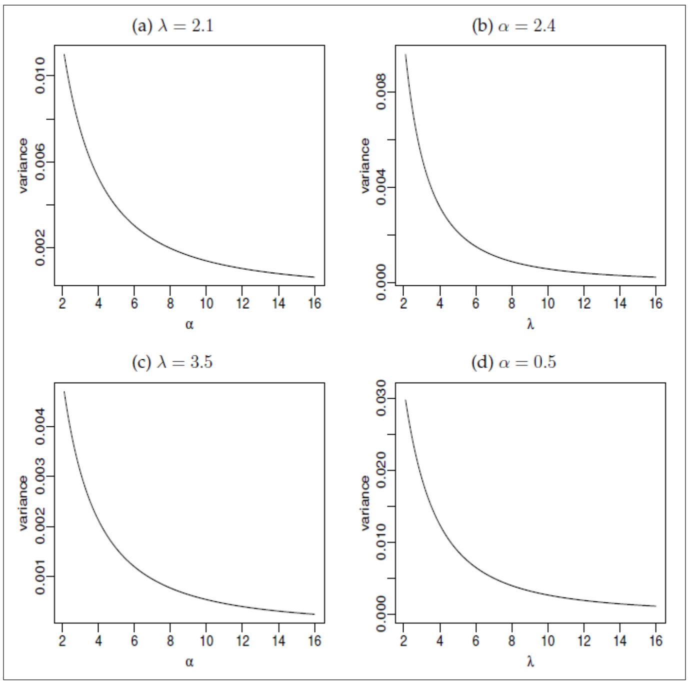

Figure 1: Variance of *Y* for special cases.

Taking $F(y;\alpha,\lambda) = \omega$, the quantile function of the unit Gompertz distribution, say $Q(\omega;\alpha,\lambda) = F^{-1}(\omega;\alpha,\lambda)$, is given as

$$Q(\omega;\alpha,\lambda) = \left[1 - \frac{1}{\alpha}\ln\omega\right]^{-1/\lambda}, \quad 0 < \omega < 1.$$

So, from Equation above, *Y* can be easily simulated. Basically, if *V* is a random variable on unit interval, say $V \sim \mathcal{U}(0,1)$, then the random variable

$$Y = \left[1 - \frac{1}{\alpha}\ln V\right]^{-1/\lambda} \qquad (2)$$

has unit Gompertz distribution.

The Proposed Model

In this Section, based on the unit Gompertz distribution, a new regression model for the unit interval is proposed. This model has parameterization on the median.

Taking $Q(\omega;\alpha,\lambda) = \tau$, and solving for α

$$\alpha = \frac{\ln\omega}{1-\tau^{-\lambda}}.$$

Inserting this expression in (1), the pdf becomes

$$f(y;\omega,\tau,\lambda) = \frac{\lambda\ln\omega}{1-\tau^{-\lambda}} y^{-(\lambda+1)} \exp\left[-\frac{\ln\omega}{1-\tau^{-\lambda}}\left(y^{-\lambda}-1\right)\right], \quad 0 < y < 1,$$

where $0 < \tau < 1$.

Choosing ω = 0.5, then τ represents the median of *Y*, and the pdf becomes

$$f(y;\tau,\lambda) = \frac{\lambda\ln 0.5}{1-\tau^{-\lambda}} y^{-(\lambda+1)} \exp\left[-\frac{\ln 0.5}{1-\tau^{-\lambda}}\left(y^{-\lambda}-1\right)\right], \quad 0 < y < 1. \quad (3)$$

The corresponding cdf is given by

$$F(y;\tau,\lambda) = \exp\left[-\frac{\ln 0.5}{1-\tau^{-\lambda}}\left(y^{-\lambda}-1\right)\right], \quad 0 < y < 1. \quad (4)$$

The random variable *Y* with pdf (3) is denoted by $Y \sim UG(\tau,\lambda)$. Figure 2 presents some plots of the density function of *Y* under parameterization, for some special cases.

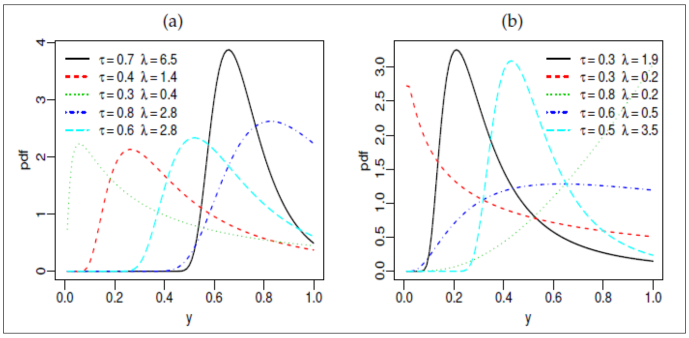

Figure 2: UG density function (3) for special cases

Let the independent random variables $Y_i \sim \mathrm{UG}(\tau_i, \lambda)$, $i = 1, \cdots, n$, with observed values y_i. The proposed model for the median T_i of y_i is given by

$$\mathcal{L}(\beta, \lambda) = \sum_{i=1}^{n} \mathcal{L}_i(\tau_i, \lambda),$$

in which S $\beta = (\beta_1,\ldots \beta_k)^{\mathrm{T}}$ is *k*-vector of unknown parameters $\beta \in \mathbb{R}^k$, $x_{1i}, \cdots, x_{ki}$ are observations on *k* covariates (*k*<*n*), which are assumed fixed and known and η_i is the linear predictor. Here, $g(\bullet)$ is strictly monotonic and twice differentiable link function, such that $g:(0,1) \to \mathbb{R}$. There are several possible choices for the link functions. For example, the logit $g(\tau) = \ln[\tau/(1-\tau)]$, the Gumbel $g(\tau) = -\ln[-\ln\tau]$, the Cauchy $g(\tau) = \tan[\pi(\tau - 0.5)]$

among others.

From Equation (3), the log-likelihood for $(\beta,\lambda)^{\mathrm{T}}$ is

$$\mathcal{L}(\beta, \lambda) = \sum_{i=1}^{n} \mathcal{L}_i(\tau_i, \lambda),$$

where

$$\mathcal{L}_i(\tau_i, \lambda) = \ln\lambda + \ln\left(\frac{\ln 0.5}{1-\tau_i^{-\lambda}}\right) - (\lambda+1)\ln y_i - \frac{\ln 0.5}{1-\tau_i^{-\lambda}}(y_i^{-\lambda} - 1).$$

Differentiating $\mathcal{L}_i(\tau_i,\lambda)$ with respect to τ_i and λ

$$\begin{aligned} \frac{\partial\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i} &= -\frac{\lambda\tau_i^{-\lambda-1}}{1-\tau_i^{-\lambda}} + \frac{\lambda\tau_i^{-\lambda-1}\ln 0.5}{[1-\tau_i^{-\lambda}]^2}(y_i^{-\lambda} - 1) \\ &= -a_i(1-\dot{y}_i), \\ \frac{\partial\mathcal{L}_i(\tau_i,\lambda)}{\partial\lambda} &= \frac{1}{\lambda} - \frac{\tau_i^{-\lambda}\ln\tau_i}{1-\tau_i^{-\lambda}} - \ln y_i + \frac{\tau_i^{-\lambda}\ln\tau_i\ln 0.5}{[1-\tau_i^{-\lambda}]^2}(y_i^{-\lambda} - 1) + \frac{\ln 0.5}{1-\tau_i^{-\lambda}} y_i^{-\lambda}\ln y_i, \\ &= \frac{1}{\lambda} - b_i - \ln y_i + b_i\dot{y}_i + \ddot{y}_i, \end{aligned}$$

where $a_i = \frac{\lambda\tau_i^{-\lambda-1}}{1-\tau_i^{-\lambda}}$, $b_i = \frac{\tau_i^{-\lambda}\ln\tau_i}{1-\tau_i^{-\lambda}}$, $\dot{y}_i = \frac{\ln 0.5}{1-\tau_i^{-\lambda}}(y_i^{-\lambda}-1)$ and $\ddot{y}_i = \frac{\ln 0.5}{1-\tau_i^{-\lambda}} y_i^{-\lambda}\ln y_i$. Once that $\mathbb{E}[\partial\mathcal{L}_i(\tau_i,\lambda)/\partial\tau_i]=0$ and $\mathbb{E}[\partial\mathcal{L}_i(\tau_i,\lambda)/\partial\lambda]=0$, then $\mathbb{E}[\dot{y}_i]=1$ and $\frac{1}{\lambda}-b_i-\mathbb{E}[\ln y_i] + b_i\mathbb{E}[\dot{y}_i] + \mathbb{E}[\ddot{y}_i]=0$, implying $\mathbb{E}[\ddot{y}_i]=\mathbb{E}[\ln y_i]-\frac{1}{\lambda}$.

The differential total is given by

$$\frac{\partial\mathcal{L}(\beta,\lambda)}{\partial\beta_j} = \sum_{i=1}^{n}\frac{\partial\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i}\frac{\partial\tau_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j},$$

$$\frac{\partial\mathcal{L}(\beta,\lambda)}{\partial\lambda} = \sum_{i=1}^{n}\frac{\partial\mathcal{L}_i(\tau_i,\lambda)}{\partial\lambda}.$$

Note that, $\partial\tau_i/\partial\eta_i = 1/g'(\tau_i)$, $\partial\eta_i/\partial\beta_j = x_{ji}$, then

$$U_{\beta_j}(\beta,\lambda) = \frac{\partial\mathcal{L}(\beta,\lambda)}{\partial\beta_j} = -\sum_{i=1}^{n} a_i[1-\dot{y}_i]\frac{1}{g'(\tau_i)}x_{ji},$$

$$U_\lambda(\beta,\lambda) = \frac{\partial\mathcal{L}(\beta,\lambda)}{\partial\lambda} = \sum_{i=1}^{n} s_i,$$

where $s_i = 1/\lambda - b_i - \ln y_i + b_i\dot{y}_i + \ddot{y}_i$.

In matrix form, $U_\beta(\beta,\lambda) = -X^\top Qc$ and $U_\lambda(\beta,\lambda) = s^\top 1_n$, where X is a n x k matrix whose ith row is $x_i^\top$ $Q=\text{diag}\{1/g'(\tau_1),\cdots,1/g'(\tau_n)\}$, $c=(a_1[1-\dot{y}_1],\cdots,a_n[1-\dot{y}_n])^\top$, $s=(s_1,\cdots,s_n)^\top$ and 1_n is a n-dimensional vector of 1's.

The maximum likelihood estimators (MLEs) are obtained by solving the nonlinear system $U_\beta(\beta,\lambda)=U_\lambda(\beta,\lambda)=0$, that has no closed-form. Thus, numerical optimization method is required, such as the BFGS algorithm. This numerical method require starting values for estimates of the parameters. Just like, here the starting values for β are the estimates of least squares method of the regression $g(y)$ on X. So, the guess initial for β is $\beta^{(0)}=(X^\top X)^{-1}X^\top g(y)$ [1]. For the parameter , the initial guess can be taken arbitrarily, e.g. $\lambda^{(0)}=1$

From Equations (7) and (8), the second derivatives of $\mathcal{L}(\beta,\lambda)$ with respect to β_l and λ are given by

$$\begin{aligned}
\frac{\partial^2\mathcal{L}(\beta,\lambda)}{\partial\beta_j\partial\beta_l} &= \sum_{i=1}^{n}\frac{\partial}{\partial\tau_i}\left(\frac{\partial\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i}\frac{1}{g'(\tau_i)}x_{ji}\right)\frac{\partial\tau_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_l}\\
&= \sum_{i=1}^{n}\left[\frac{\partial^2\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i^2}\frac{1}{g'(\tau_i)}x_{ji} + \frac{\partial\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i}\frac{\partial}{\partial\tau_i}\left(\frac{1}{g'(\tau_i)}\right)x_{ji}\right]\frac{1}{g'(\tau_i)}x_{li}\\
&= \sum_{i=1}^{n}\left[\frac{\partial^2\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i^2}\frac{1}{g'(\tau_i)^2}x_{ji}x_{li} - \frac{\partial\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i}\frac{g''(\tau_i)}{g'(\tau_i)^3}x_{ji}x_{li}\right],\\
\frac{\partial^2\mathcal{L}(\beta,\lambda)}{\partial\beta_j\partial\lambda} &= \sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\left(\frac{\partial\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i}\frac{\partial\tau_i}{\partial\eta_i}x_{ji}\right)\\
&= \sum_{i=1}^{n}\left(\frac{\partial^2\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i\partial\lambda}\right)\frac{1}{g'(\tau_i)}x_{ji}\\
\frac{\partial^2\mathcal{L}(\beta,\lambda)}{\partial\lambda^2} &= \sum_{i=1}^{n}\frac{\partial^2\mathcal{L}_i(\tau_i,\lambda)}{\partial\lambda^2}.
\end{aligned}$$

From Equations (5) and (6),

$$\begin{aligned}
\frac{\partial^2\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i^2} &= -(1-\dot{y}_i)\frac{\partial a_i}{\partial\tau_i} + a_i\frac{\partial\dot{y}_i}{\partial\tau_i},\\
&=(1-\dot{y}_i)[(\lambda+1)a_i\tau_i^{-\lambda}+a_i^2]-a_i^2\dot{y}_i,\\
&=h_i,\\
\frac{\partial^2\mathcal{L}_i(\tau_i,\lambda)}{\partial\tau_i\partial\lambda} &= -(1-\dot{y}_i)\frac{\partial a_i}{\partial\lambda} + a_i\frac{\partial\dot{y}_i}{\partial\lambda}\\
&= -\left\{(1-\dot{y}_i)\left[\frac{a_i}{\lambda}-a_ib_i-a_i\ln\tau_i\right]+a_i(b_i\dot{y}_i+\ddot{y}_i)\right\}\\
&= -d_i,\\
\frac{\partial^2\mathcal{L}_i(\tau_i,\lambda)}{\partial\lambda^2} &= -\frac{1}{\lambda^2}-(1-\dot{y}_i)\frac{\partial b_i}{\partial\lambda}+b_i\frac{\partial\dot{y}_i}{\partial\lambda}+\frac{\partial\ddot{y}_i}{\partial\lambda}\\
&= -\left\{\frac{1}{\lambda^2}-(1-\dot{y}_i)(b_i^2+b_i\ln\tau_i)+b_i^2\,\dot{y}_i+2b_i\,\ddot{y}_i+\ddot{y}_i\ln y_i\right\}\\
&= -q_i,
\end{aligned}$$

where $h_i=(1-\dot{y}_i)[(\lambda+1)a_i\tau_i^{-\lambda}+a_i^2]-a_i^2\dot{y}_i$, $d_i=(1-\dot{y}_i)$ $[a_i/\lambda-a_ib_i-a_i\ln\tau_i]+a_i(b_i\dot{y}_i+\ddot{y}_i)$ and $q_i=1/\lambda^2-(b_i^2+b_i\ln\tau_i)[1-\dot{y}_i]+b_i^2\,\dot{y}_i+2b_i\,\ddot{y}_i+\ddot{y}_i\ln y_i$.

Finally,

$$J_{\beta_j\beta_l} = \frac{\partial^2\mathcal{L}(\beta,\lambda)}{\partial\beta_j\partial\beta_l} = \sum_{i=1}^{n}\left(h_i\frac{1}{g'(\tau_i)^2}x_{ji}x_{li} + a_i(1-\dot{y}_i)\frac{g''(\tau_i)}{g'(\tau_i)^3}x_{ji}x_{li}\right),$$

$$J_{\beta_j\lambda} = \frac{\partial^2\mathcal{L}(\beta,\lambda)}{\partial\beta_j\partial\lambda} = -\sum_{i=1}^{n} d_i\frac{1}{g'(\tau_i)}x_{ji}$$

$$J_{\lambda\lambda} = \frac{\partial^2\mathcal{L}(\beta,\lambda)}{\partial\lambda^2} = -\sum_{i=1}^{n} q_i.$$

In matrix form, $J_{\beta\beta}=X^\top PX+X^\top RX$, $J_{\beta\lambda}=J_{\lambda\beta}^\top=-X^\top Qd$ and $J_{\lambda\lambda}=-q^\top 1_n$, where $P=\text{diag}\{h_1/g'(\tau_1)^2,\cdots,h_n/g'(\tau_n)^2\}$, $R=\text{diag}\{a_1(1-\dot{y}_1)g''(\tau_1)/g'(\tau_1)^3,\cdots,a_n(1-\dot{y}_n)g''(\tau_n)/g'(\tau_n)^3\}$ $d=(d_1,\cdots,d_n)^\top$ and $q=(q_1,\cdots,q_n)^\top$

Thus, the observed information matrix is given by

$$J(\beta,\lambda) = -\begin{bmatrix} J_{\beta\lambda} & J_{\beta\lambda}\\ J_{\lambda\beta} & J_{\lambda\lambda}\end{bmatrix} = \begin{bmatrix} -X^\top PX-X^\top RX & X^\top Qd\\ d^\top QX & q^\top 1_n\end{bmatrix}.$$

Under the usual regularity conditions of the maximum likelihood estimators, for n large,

$$\begin{pmatrix}\hat{\beta}\\ \hat{\lambda}\end{pmatrix} \overset{a}{\sim} \mathcal{N}_{k+1}\left(\begin{pmatrix}\beta\\ \lambda\end{pmatrix}, J(\beta,\lambda)^{-1}\right),$$

where $\overset{a}{\sim}$ denotes asymptotic distribution. So, confidence intervals and hypothesis testing can be performed using the normal distribution. Based on asymptotic distribution, the 100(1-γ)% confidence intervals for β_j and λ are given by

$$\hat{\beta}_j \pm z_{(1-\gamma/2)}\sqrt{W_{jj}}, \quad \text{and} \quad \hat{\lambda}\pm z_{(1-\gamma/2)}\sqrt{W_{(j+1)(j+1)}}, \quad j=1,\cdots,k, \tag{9}$$

respectively, where $z_{(1-\gamma/2)}$ is the $(1-\gamma/2)$ quantile of the standard normal distribution and W_{hh} denotes the hth diagonal element of the matrix $J(\beta,\lambda)^{-1}$.

Diagnostic Analysis

An important measure of model adequacy is the analysis of residuals. The residuals show if the estimated model is well-adjusted. Here, two measures of residuals will be discussed, namely: Cox-Snell residuals and Dunn-Smith residuals.

The Cox-Snell residuals are defined as [6].

$$\hat{e}_i = -\ln\left[1 - F\left(y_i;\hat{\tau}_i,\hat{\lambda}\right)\right],$$

where $F\left(y_i;\hat{\tau}_i,\hat{\lambda}\right)$ is the cdf (4) evaluated in $\hat{\tau}_i$ and $\hat{\lambda}$. Note that, $\hat{\tau}_i = g^{-1}\left(\hat{\eta}_i\right)$ If the model is well adjusted, the Cox-Snell residuals follow the standard exponential distribution, that is, exponential distribution with scale parameter 1.

The Dunn-Smith residuals are given by [7].

$$\hat{r}_i = Q_N\left(F\left(y_i;\hat{\tau}_i,\hat{\lambda}\right)\right),$$

where $Q_N(\cdot)$ is the quantile function of the standard normal distribution. If the model is valid, the Dunn-Smith residuals approximately follow the standard normal distribution. Thus, the Dunn-Smith residuals have a behavior around zero with about 95% of the values in the interval (-2,2).

Simulation

A Monte Carlo simulation, with 10,000 replications, is performed to evaluate some asymptotic properties of the MLEs for the parameters of the unit Gompertz regression model. The model proposed, with $x1i$ being a column vector of ’s, is given by

$$g(\tau_i) = \ln\left(\frac{\tau_i}{1-\tau_i}\right) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}, \quad i = 1,\cdots,n,$$

where x_{2i} , x_{3i} and x_{4i} are random samples of the standard uniform distribution, i.e., $x_{mi} \sim \mathcal{U}(0,1)$, $m = 2,3,4$. The true parameters are: $\beta_1 = 3.0$, $\beta_2 = 1.5$, $\beta_3 = -0.8$, $\beta_4 = -2.7$ and $\lambda = 2.4$.
Seven sample size are considered: $n = 30$, 60, 90, 120, 150, 200 and 300.
The simulation mechanism is given as following:

1. Generate $x_{mi} \sim \mathcal{U}(0,1)$, $m = 2,3,4$;

2. Write $\eta_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$ and after define $\alpha_i = \ln 0.5 / \left[1 - \tau_i^{-\lambda}\right]$, where $\tau_i = g^{-1}(\eta_i) = e^{\eta_i} / \left(1 + e^{\eta_i}\right)$

3. Finally, from Equation (2), generate $y_i \sim UG(\alpha_i, \lambda)$

The simulations were performed using the matrix programming language Ox Console with subroutine MaxBFGS and analytical derivatives [8]. The average estimates (AEs), the biases, the mean square errors (MSEs) and the confidence intervals with the coverage rates (CRs) of 90%, 95% and 99% are shown in Table 1. The CRs were calculated from Equation (9).

As can be seen, when the sample grows, the AEs converge to the true parameters and the biases and the MSEs decrease, as expected of the asymptotic theory. Note also that the CRs for the parameters are very close to their nominal levels.

Application

In this Section to show the potentiality of the unit Gompertz regression model, application to real data is considered. The new model is compared with two others well-known models. The competitive models are:

- The beta regression, with pdf given by [1].

$$f_{\text{beta}}(y;\tau,\phi) = \frac{\Gamma(\phi)}{\Gamma(\tau\phi)\Gamma\left((1-\tau)\phi\right)} y^{\tau\phi-1}(1-y)^{(1-\tau)\phi-1}, \quad 0 < y < 1,$$

where $0 < \tau < 1$ denotes the mean, $\phi > 0$ is a precision parameter and $\Gamma(p) = \int_0^\infty u^{p-1} e^{-u} du$, $p > 0$, is the gamma function.

- The Kumaraswamy (Kw) regression model, with pdf given by [2].

$$f_{kw}(y;\tau,\phi) = \frac{\phi \ln 0.5}{\ln\left(1-\tau^\phi\right)} y^{\phi-1}(1-y^\phi)^{\frac{\ln 0.5}{\ln\left(1-\tau^\phi\right)}-1}, \quad 0 < y < 1,$$

where $0 < \tau < 1$ denotes the median and $\phi > 0$ is precision parameter.

- The unit log-logistic (ULL) regression model, in which its density function is given as [4].

$$f_{ull}(y;\tau,\phi) = \frac{\phi}{y(-\ln\tau)}\left(\frac{\ln y}{\ln\tau}\right)^{\phi-1}\left[1 + \left(\frac{\ln y}{\ln\tau}\right)^{\phi}\right]^{-2}, \quad 0 < y < 1,$$

where $0 < \tau < 1$ denotes the median and $\phi > 0$ is shape parameter.

This study will be based on the deaths caused by Covid-19 in the capitals of the Brazil and in the Federal District. It will seek to explain how some socioeconomic variables may be related to these deaths.

The variables are:

- y_i : Covid deaths per 100 thousand inhabitants (collected at December 6, 2021) (https://brasil.io/covid19/);

Table 1: Monte Carlo Simulation Results

n	Par	AE	Bias	MSE	90%	CR	
						95%	99%
30	β1	2.94611	−0.05389	0.56423	82.99	89.03	95.75
	β2	1.51007	0.01007	0.64298	82.37	88.91	95.93
	β3	−0.79344	0.00656	0.37327	82.87	89.23	96.00
	β4	−2.64536	0.05464	0.80334	82.56	88.73	95.47
	λ	4.14087	1.74087	7.25620	76.41	85.24	95.56
60	β1	3.00023	0.00023	0.14618	87.20	92.89	98.03
	β2	1.49478	−0.00522	0.20205	86.69	92.48	97.67
	β3	−0.80569	−0.00569	0.18473	86.52	92.42	97.55
	β4	−2.70340	−0.00340	0.26551	85.69	91.56	97.37
	λ	3.08971	0.68971	1.64351	83.60	90.88	97.70
90	β1	2.99897	−0.00103	0.09108	88.65	93.81	98.39
	β2	1.50440	0.00440	0.12265	88.24	93.42	98.38
	β3	−0.81047	−0.01047	0.11352	87.92	93.14	98.01
	β4	−2.70027	−0.00027	0.14627	87.46	93.33	98.18
	λ	2.80971	0.40971	0.78904	86.23	92.35	98.14
120	β1	3.00236	0.00236	0.06685	88.53	93.78	98.47
	β2	1.50419	0.00419	0.09006	88.10	93.65	98.21
	β3	−0.80623	−0.00623	0.08507	88.17	93.69	98.51
	β4	−2.70897	−0.00897	0.10793	88.52	93.82	98.40
	λ	2.70792	0.30792	0.53813	87.27	93.05	98.42
150	β1	3.00193	0.00193	0.05020	88.62	94.34	98.65
	β2	1.50148	0.00148	0.06737	89.01	94.14	98.56
	β3	−0.80379	−0.00379	0.06528	88.63	94.23	98.55
	β4	−2.70537	−0.00537	0.08450	88.67	93.92	98.34
	λ	2.63776	0.23776	0.39388	88.35	93.54	98.34
200	β1	3.00059	0.00059	0.03905	89.54	94.65	98.79
	β2	1.50514	0.00514	0.05028	88.89	94.31	98.53

	β3	−0.79698	0.00302	0.04945	89.47	94.27	98.59
	β4	−2.70967	−0.00967	0.05708	89.14	94.57	98.81
	λ	2.56115	0.16115	0.24839	88.78	94.13	98.54
300	β1	3.00161	0.00161	0.02679	89.35	94.41	98.80
	β2	1.50423	0.00423	0.03342	89.85	94.70	98.83
	β3	−0.80148	−0.00148	0.03196	89.21	94.48	98.77
	β4	−2.70495	−0.00495	0.03677	89.13	94.32	98.69
	λ	2.50225	0.10225	0.15058	89.27	94.64	98.82

- x_{2i} : Social Vulnerability Index (SVI) is the result of the arithmetic mean of the sub-indices: Urban Infrastructure SVI, Human Capital SVI and Income and Labor SVI, each of which enters the calculation of the final SVI with the same weight. The SVI index varies between 0 and 1. A value of 0 indicates perfect social condition, and 1 indicates a worse situation (http://ivs.ipea.gov.br/).
- x_{3i} : Municipal Human Development Index (MHDI) that is a measure used to classify the degree of economic development and quality of life. This index varies between 0 and 1. The closer to 0, the lower the degree of development. Already for values close to 1, it indicates a high level of development (http://ivs.ipea.gov.br/).
- x_{4i} : Demographic density (DD), that is the ratio between total population and square kilometers (inhabitants/km) (https://www.ibge.gov.br/);
- x_{5i} : GDP per capita, at current prices (R$ 1.00) (https://www.ibge.gov.br/).

Table 2 presents the descriptive statistics of the variables used in this study. As can be seen, for every 100,000 people, the average number by Covid-19 deaths is approximately 348. The number of deaths oscillates between 219 and 573.

Table 2: Descriptive Statistics of the Variables

Variables	mean	SD	min	max
Deaths	347.94149	87.20497	219.13188	573.34774
SVI	0.29526	0.05090	0.17800	0.39300
MHDI	0.77652	0.03520	0.72100	0.84700
DD	2813.41378	2810.59049	15.82103	8601.20441
GDP	34351.86333	14391.11379	20821.46000	80502.47000

SD = standard deviation.

The model is given by

$$g(\tau_i) = \ln\left(\frac{\tau_i}{1-\tau_i}\right) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}, \quad i = 1, \cdots, 27,$$

where τ_i denotes the median for the unit Gompertz and Kumaraswamy regression models, while for the beta regression model, τ_i denotes the mean.

The popular Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan–Quinn Information Criterion (HQIC) measures are used to choose the best model. The model with the lowest values of these statistics is chosen as the best model. These measures are given by

$$\text{AIC} = -2\mathcal{L}(\hat{\boldsymbol{\eta}}) + 2p,\ \text{BIC} = -2\mathcal{L}(\hat{\boldsymbol{\eta}}) + p\ln n \ \text{ and } \ \text{HQIC} = -2\mathcal{L}(\hat{\boldsymbol{\eta}}) + 2p\ln(\ln n),$$

where $\mathcal{L}(\hat{\boldsymbol{\eta}})$ is the log-likelihood evaluated in the MLE $\hat{\boldsymbol{\eta}}$, p is the number of parameters in the model and n is the number of observations.

Another measure that is used in model selection choice is the pseudo R^2, defined by [9]. The pseudo R^2 is given by

$$R^2 = 1 - \exp\left\{-\frac{2}{n}\left[\mathcal{L}(\hat{\boldsymbol{\eta}}) - \mathcal{L}(\hat{\boldsymbol{\eta}}_0)\right]\right\},$$

where $\mathcal{L}(\hat{\boldsymbol{\eta}}_0)$ log-likelihood of the null model (without covariates)

and $\mathcal{L}(\hat{\boldsymbol{\eta}})$ is the log-likelihood of the regression model. Note that and it determines how much the dependent variable is explained by covariates. A high R^2 indicates better predictions. For example, if $\mathcal{L}(\hat{\boldsymbol{\eta}})$ and $\mathcal{L}(\hat{\boldsymbol{\eta}}_0)$ are very close, then the covariates did not add

much to the log-likelihood. So, note that if the difference between $\mathcal{L}(\hat{\boldsymbol{\eta}})$ and $\mathcal{L}(\hat{\boldsymbol{\eta}}_0)$ tends to zero, R^2 also tends to zero. On the other

hand, if $\mathcal{L}(\hat{\boldsymbol{\eta}})$ and $\mathcal{L}(\hat{\boldsymbol{\eta}}_0)$ are very far apart, then the covariates have a significant influence on log-likelihood. Thus, if the distance between $\mathcal{L}(\hat{\boldsymbol{\eta}})$ and $\mathcal{L}(\hat{\boldsymbol{\eta}}_0)$ is very large, then R^2 tends to 1.

Table 3 presents the MLEs estimates for the three models. In the unit Gompertz model only the coefficient β_1 was not statistically significant. For the beta model the coefficients β_1, β_2 and β_4 are not significant, even at the 10% level. In the Kumaraswamy model the coefficients $\beta_1, \beta_2, \beta_3$ and β_4 also are not significant

The coefficients β_3 and β_4, except for β_4 in the ULL model, showed expected signs. Bearing in mind that the greater the development of a municipality, it is expected that there will be a lower death rate. On the other hand, the higher the demographic density, the higher the level of contamination caused by agglomerations. Already the coefficients β_2 and β_5 presented signs contrary to what was expected.

The AIC, BIC, HQIC and R^2 statistics are given in Table 4. All these information criteria point to the unit Gompertz regression model as the best model.

Figure 3(a) presents the plots of the Dunn-Smith residuals. This plot shows that the residuals exhibit a behaviour around zero. Figure 3(b) shows the PP-plot of the Cox-Snell residuals. It can be seen that these residuals are very close to the diagonal line. Thus, this figure shows that the unit Gompertz model is well-adjusted.

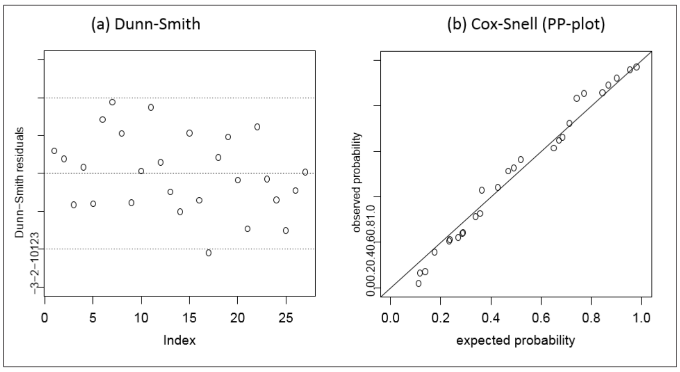

Figure 3: Residuals Plots

Conclusions

In this paper, a new regression model for the unit interval was proposed, based on the unit Gompertz distribution. In this model, the parameterization occurs in terms of the median, which shows an advantage over the models based on the mean, when there is great asymmetry and outliers in the data.

Table 3: Estimates Results

Par	Estimate	Std. Error	z-value	Pr(>|z|)
		unit Gompertz		
β1	−0.225700	2.162300	−0.104380	0.916868
β2	−2.186094	1.248598	−1.750838	0.079974
β3	−6.848625	2.509989	−2.728548	0.006361
β4	0.000023	0.000013	1.851692	0.064070
β5	0.000012	0.000004	3.170348	0.001523
λ	5.651619	0.871264	6.486691	<0.000001
		beta		
β1	−0.981942	2.702201	−0.363386	0.716316
β2	−2.091595	1.561042	−1.339871	0.180287
β3	−5.726490	3.170402	−1.806235	0.070882
β4	0.000008	0.000017	0.465839	0.641331
β5	0.000011	0.000005	2.178832	0.029344
ϕ	5899.959345	1612.342127	3.659248	0.000253
	Kumaraswamy			
β1	−2.793520	2.460990	−1.135121	0.256325
β2	−1.522968	1.437302	−1.059602	0.289326
β3	−3.576979	2.879873	−1.242061	0.214214
β4	−0.000014	0.000016	−0.910465	0.362577
β5	0.000012	0.000005	2.618099	0.008842
ϕ	4.885948	0.733150	6.664326	<0.000001
	unit log-logistic			
β1	−0.211094	2.745867	−0.076877	0.938721
β2	−2.451312	1.512773	−1.620410	0.105144
β3	−6.693367	3.284361	−2.037951	0.041555
β4	0.000012	0.000018	0.662637	0.507563
β5	0.000012	0.000005	2.170893	0.029939
ϕ	44.441313	7.101658	6.257878	<0.000001

Table 4: Information Criteria

Model	AIC	BIC	HQIC	R2
UG	−303.1783	−295.4033	−300.8664	0.3102
Beta	−299.7716	−291.9966	−297.4597	0.1633
Kw	−297.1648	−289.3897	−294.8528	0.1969
ULL	−299.1573	−291.3823	−296.8454	0.1670

The estimates of the parameters are obtained by the maximum likelihood method. The score vector and the observed information matrix were demonstrated in simple forms. Simulation studies are carried out, to show the consistency of the maximum likelihood estimators of the proposed model. The results showed that the model is in agreement with the asymptotic theory.

The usefulness of the model is shown through an application to Covid-19 data in the capital cities of Brazil. The results showed that the unit Gompertz regression model has a better fit than the well-known beta and Kumaraswamy regression models.

For future research, the regression for time series, the regression for data inflated by 0 and 1, correction of MLEs via bootstrap, and the regression structure for the shape parameter can be investigated.

References

1. Ferrari SLP, Cribari Neto F (2004) Beta regression for modelling rates and proportions. Journal of Applied Statistics 31: 799-815.
2. Mitnik PA, Baek S (2013) The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. Statistical Papers 54: 177-192.
3. Mazucheli J, Menezes AFB, Fernandes LB, de Oliveira RP, Ghitany ME (2020) The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates. Journal of Applied Statistics 47: 954-974.
4. Ribeiro Reis LD (2021) Unit log-logistic distribution and unit log-logistic regression model. Journal of the Indian Society for Probability and Statistics 22: 375-388.
5. Mazucheli J, Menezes AFB, Dey S (2019) Unit-Gompertz distribution with applications. Statistica 79: 25-43.
6. Cox DR, Snell EJ (1968) A general definition of residuals. Journal of the Royal Statistical Society: Series B (Methodological) 30: 248-265.
7. Dunn PK, Smyth GK (1996) Randomized quantile residuals. Journal of Computational and Graphical Statistics 5: 236-244.
8. Doornik JA (2018) Ox: an object-oriented matrix programming language. Timberlake Consultants and Oxford, London, United Kingdom https://ora.ox.ac.uk/objects/uuid:242f0a19-0665-4d9a-b863-774a35ce98c7.
9. Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. Biometrika 78: 691-692.