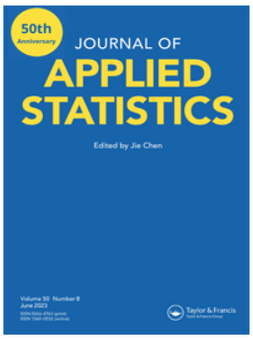

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

The re-parameterized inverse Gaussian regression to model length of stay of COVID-19 patients in the public health care system of Piracicaba, Brazil

E. M. Hashimoto, E. M. M. Ortega, G. M. Cordeiro, V. G. Cancho & I. Silva

To cite this article: E. M. Hashimoto, E. M. M. Ortega, G. M. Cordeiro, V. G. Cancho & I. Silva (2023) The re-parameterized inverse Gaussian regression to model length of stay of COVID-19 patients in the public health care system of Piracicaba, Brazil, Journal of Applied Statistics, 50:8, 1665-1685, DOI: 10.1080/02664763.2022.2036707

To link to this article: https://doi.org/10.1080/02664763.2022.2036707

JOURNAL OF APPLIED STATISTICS
2023, VOL. 50, NO. 8, 1665–1685
https://doi.org/10.1080/02664763.2022.2036707

The re-parameterized inverse Gaussian regression to model length of stay of COVID-19 patients in the public health care system of Piracicaba, Brazil

E. M. Hashimoto[a], E. M. M. Ortega [b], G. M. Cordeiro [c], V. G. Cancho[d] and I. Silva[e]

[a]Department of Mathematics, Federal University of Technology – Paraná, Londrina, PR, Brazil; [b]Department Exact Sciences, University of São Paulo, Piracicaba, SP, Brazil; [c]Department of Statistics, Federal University of Pernambuco, Recife, PE, Brazil; [d]Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, SP, Brazil; [e]Graduate Program in Environmental Engineering, Federal University of Technology – Paraná, Londrina, PR, Brazil

ABSTRACT

Among the models applied to analyze survival data, a standout is the inverse Gaussian distribution, which belongs to the class of models to analyze positive asymmetric data. However, the variance of this distribution depends on two parameters, which prevents establishing a functional relation with a linear predictor when the assumption of constant variance does not hold. In this context, the aim of this paper is to re-parameterize the inverse Gaussian distribution to enable establishing an association between a linear predictor and the variance. We propose deviance residuals to verify the model assumptions. Some simulations indicate that the distribution of these residuals approaches the standard normal distribution and the mean squared errors of the estimators are small for large samples. Further, we fit the new model to hospitalization times of COVID-19 patients in Piracicaba (Brazil) which indicates that men spend more time hospitalized than women, and this pattern is more pronounced for individuals older than 60 years. The re-parameterized inverse Gaussian model proved to be a good alternative to analyze censored data with non-constant variance.

ARTICLE HISTORY
Received 14 June 2021
Accepted 27 January 2022

KEYWORDS
Censored data; inverse Gaussian distribution; regression; SARS-COV-2; SUS

MATHEMATICS SUBJECT CLASSIFICATIONS
62J05; 62J20

1. Introduction

The advantage of associating a distribution to the survival time is that it makes the statistical analysis more precise. However, the literature contains a large range of lifetime continuous distributions [16]. The exponential, Weibull, log-logistic, log-normal, and generalized gamma, are the most often used models in survival analysis [17,18].

On the other hand, the inverse Gaussian (IG) [32] distribution has not been used so often to analyze survival data. This distribution belongs to the class of models to analyze positive asymmetric data [24]. If the parameters of the IG distribution increase, it approximates to the normal distribution, thus making it attractive to analyze asymmetric as well

CONTACT E. M. M. Ortega edwin@usp.br

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2022.2036707

as symmetric random variables. For example, Whitmore [36] presented the adjustment of the IG model to analyze data on the failure of aluminum reduction cells, while Gupta and Kundu [14] adopted the EM algorithm to estimate its parameters. Basak and Balakrishnan [3] showed that the Newton-Raphson method and the EM algorithm provided similar results to estimate the parameters of this distribution. Hanagal and Bhambure [15] analyzed a bivariate dataset utilizing the IG distribution as the frailty model. Suzuki *et al.* [30] proposed a time-promotion model based on the IG distribution to estimate the proportion of cured individuals. Salha and Rasheed [26] showed that the IG kernel estimator presented a smaller mean squared error in comparison with a Gaussian kernel estimator. Finally, Songhua *et al.* [29] proposed an IG model with normal random effects, Punzo [25] defined a re-parametrization based on the mode of this distribution, and Vasconcelos [35] introduced the odd log-logistic generalized inverse Gaussian regression in some applications.

However, the variance of the IG distribution does not allow establishing a direct functional relationship with a linear predictor via a link function. In the literature, Ferrari and Cribari-Neto [11] proposed a re-parametrization of the beta distribution to establish a relation between its mean and a linear predictor, while Nelson [22] utilized a re-parameterized generalized normal distribution so that its expectation is zero and the variance is one. Achim *et al.* [1] showed that the generalized normal distribution, as re-parameterized in Nelson [22], can be obtained from a distribution rewritten in terms of the standard deviation. In view of this, we can add more information to the models, thus allowing that both the variability and the mean be explained by covariates.

In this context, the aim of the present study is to construct a *re-parameterized inverse Gaussian* (RIG) distribution to enable establishment of a functional relationship between a linear predictor and the variance of this distribution, specifically for the analysis of censored data.

The paper is organized as follows. In Section 2, we present the RIG distribution, and some of its structural properties. In Section 3, we define a regression with two systematic components based on the RIG distribution. In Section 4, the diagnostic measures for the new regression are reported to assess its adequacy. In Section 5, a Monte Carlo simulation study is performed to evaluate the maximum likelihood estimates (MLEs) of the parameters, and compare the empirical distribution of the residuals with the standard normal. In Section 6, the new regression is applied to explain the length of hospital stay of COVID-19 patients in the Unified Health System in the City of Piracicaba, Brazil. Finally, we provide the main conclusions of the study about the fitted RIG regression in Section 7.

2. The RIG distribution

A non-negative random variable $T \sim IG(\mu, \lambda)$ follows an IG distribution with parameters $\mu > 0$ and $\lambda > 0$, if its probability density function (pdf) is

$$f(t) = \left(\frac{\lambda}{2\pi t^3}\right)^{1/2} \exp\left\{-\frac{\lambda(t-\mu)^2}{2\mu^2 t}\right\}, \quad t > 0. \tag{1}$$

Equation (1) was pioneered by Schrodinger [27] to describe the time of the first passage in Brownian motion. Subsequently, it was called the IG distribution by Tweedie [32] due to the inverse relationship between the cumulative generating function of the first pass

time distribution and the normal distribution [13]. Tweedie [33,34] presented some properties of this distribution as well as the density curves for some parameter values. Among the basic characteristics, Tweedie [33] showed that the mean and variance of T are given, respectively, by

$$E(T) = \mu \quad \text{and} \quad \text{Var}(T) = \frac{\mu^3}{\lambda}.$$

In addition, the IG distribution belongs to the exponential family and therefore, a functional relationship can be considered for the mean of T and a linear predictor by means of a link function [13,20,24]. However, it is not possible to establish a functional relationship between the variance of T and a linear predictor, since it depends on two parameters. For example, Ferrari and Cribari-Neto [11] proposed a re-parametrization in the beta distribution to establish a relationship between its mean and a linear predictor. On the other hand, Achim *et al.* [1] considered a re-parametrization in the generalized normal distribution in terms of the standard deviation.

Then, motivated by these two works, we propose a re-parametrization of the distribution of T in terms of its variance $\sigma^2 > 0$

$$\text{Var}(T) = \frac{\mu^3}{\lambda} = \sigma^2 \Rightarrow \lambda = \frac{\mu^3}{\sigma^2}. \tag{2}$$

By replacing the parameter λ in Equation (1), the RIG density of T takes the form

$$f(t) = \left(\frac{\mu^3}{2\pi\,\sigma^2\,t^3}\right)^{1/2} \exp\left\{-\frac{\mu(t-\mu)^2}{2\,\sigma^2\,t}\right\}, \quad t > 0, \tag{3}$$

where $E(T) = \mu$ and $\text{Var}(T) = \sigma^2$.

In this way, the RIG distribution allows a linear predictor both in the mean and in the variance through a link function. Henceforth, a random variable $T \sim \text{RIG}(\mu, \sigma^2)$ has the pdf (3), its cumulative distribution function (cdf) has the form

$$F(t) = \Phi\left(\sqrt{\frac{\mu^3}{\sigma^2\,t}}\left(\frac{t}{\mu} - 1\right)\right) + \exp\left(\frac{2\mu^2}{\sigma^2}\right)\Phi\left(-\sqrt{\frac{\mu^3}{\sigma^2\,t}}\left(\frac{t}{\mu} + 1\right)\right), \tag{4}$$

and its survival function is $S(t) = 1 - F(t)$, where $\Phi(\cdot)$ is the standard normal cdf.

Some mathematical properties of T are well-known in the literature. For example, the generating function of T can be expressed as

$$M(s) = E[\mathrm{e}^{s\,T}] = \exp\left\{\frac{\mu^2}{\sigma^2}\left(1 - \sqrt{1 - \frac{2\sigma^2 s}{\mu}}\right)\right\},$$

from which follows the coefficients of skewness and kurtosis.

3. The RIG regression

Let $T_1, \ldots, T_n$ be a random sample from the RIG distribution (3) such that $T_i \sim \text{RIG}(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$. It is assumed that the random variables have different means

and the same variance σ^2. So, assuming that $(t_1, \mathbf{x}_i), \ldots, (t_n, \mathbf{x}_n)$ is the observed sample, a regression can be defined for the mean of T_i as (for $i = 1, \ldots, n$)

$$\mu_i = \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right), \tag{5}$$

where $\mathbf{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$ is the vector of the values of p explanatory variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the vector of unknown parameters.

On the other hand, assuming that the variances are also different, that is, $T_i \sim \text{RIG}(\mu_i, \sigma_i^2)$ (for $i = 1, \ldots, n$), the RIG regression with multiplicative heteroscedastic structure is defined by

$$\sigma_i^2 = \exp\left(\mathbf{z}_i^\top \boldsymbol{\gamma}\right), \tag{6}$$

where $\mathbf{z}_i^\top = (1, z_{i1}, \ldots, z_{iq})$ is the vector of the values of q explanatory variables, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)^\top$ is the vector of unknown parameters. In general, it is very common to choose $\mathbf{z}_i$ as a subset of $\mathbf{x}_i$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ be the parameter vector. The systematic components (5) and (6) define the heteroscedastic RIG regression, and then the density of $T \mid \mathbf{x}_i, \mathbf{z}_i$ can be expressed as

$$f(t_i; \boldsymbol{\theta}) = \left\{\frac{\exp[2(\mathbf{x}_i^\top \boldsymbol{\beta})]}{2\,\pi\, t_i^3 \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}\right\}^{1/2} \exp\left\{-\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})[t_i - \exp(\mathbf{x}_i^\top \boldsymbol{\beta})]^2}{2t_i \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}\right\}. \tag{7}$$

Consequently, the survival function of $T \mid \mathbf{x}_i, \mathbf{z}_i$ follows from (4) as

$$\begin{aligned} S(t_i; \boldsymbol{\theta}) = 1 - \Phi\left[\kappa_i\left(\frac{t_i}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} - 1\right)\right] - \exp\left(\frac{2\exp[2(\mathbf{x}_i^\top \boldsymbol{\beta})]}{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}\right) \\ \times \Phi\left[-\kappa_i\left(\frac{t_i}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} + 1\right)\right], \end{aligned} \tag{8}$$

where

$$\kappa_i = \sqrt{\frac{\exp[3(\mathbf{x}_i^\top \boldsymbol{\beta})]}{t_i \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}}.$$

3.1. Estimation

Let $(t_1, \delta_1, \mathbf{x}_1), \ldots, (t_n, \delta_n, \mathbf{x}_n)$ be observed from $T_i \sim \text{RIG}(\mu_i, \sigma_i^2)$, where δ_i (1=failure, 0=censoring) is the censoring indicator (for $i = 1, \ldots, n$). The log-likelihood function for the parameter vector $\boldsymbol{\theta}$ in the RIG regression defined from Equations (5), (6) and (7) has the form

$$\begin{aligned} l(\boldsymbol{\theta}) = -\frac{r\log(2\pi)}{2} + \frac{3}{2}\sum_{i=1}^{n}\delta_i \mathbf{x}_i^\top \boldsymbol{\beta} - \frac{1}{2}\sum_{i=1}^{n}\delta_i \mathbf{z}_i^\top \boldsymbol{\gamma} - \frac{3}{2}\sum_{i=1}^{n}\delta_i \log(t_i) \\ - \sum_{i=1}^{n}\delta_i \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})[t_i - \exp(\mathbf{x}_i^\top \boldsymbol{\beta})]^2}{2t_i \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} + \sum_{i=1}^{n}(1 - \delta_i) \end{aligned}$$

$$\times \log\left\{1-\Phi\left[\kappa_i\left(\frac{t_i}{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}-1\right)\right]-\exp\left\{\frac{2\exp[2(\mathbf{x}_i^\top\boldsymbol{\beta})]}{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\right\}\right.$$
$$\left.\times\ \Phi\left[-\kappa_i\left(\frac{t_i}{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}+1\right)\right]\right\}, \tag{9}$$

where r is the number of failures and κ_i comes from (8). The MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is determined by maximizing the log-likelihood function (9) or as the solution of the system of nonlinear equations

$$U(\boldsymbol{\theta})=\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\mathbf{0}.$$

However, it is not possible to analytically solve the system of score equations above. So, to maximize the log-likelihood function (9), and find the parameter estimates, we use the `MaxBFGS` subroutine of the matrix programming language `Ox` version `8.00` [10].

The inference procedure for $\boldsymbol{\theta}=(\boldsymbol{\beta}^\top,\boldsymbol{\gamma}^\top)^\top$ can be based on the asymptotic normal approximation

$$\hat{\boldsymbol{\theta}}^\top\sim N_{p+q+2}\left(\boldsymbol{\theta}^\top,[\ddot{L}(\boldsymbol{\theta})]^{-1}\right),$$

where $\ddot{L}(\boldsymbol{\theta})=-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$ is the observed information matrix.

So, the asymptotic confidence interval for θ_j $(j=1,\ldots,p+q+2)$ at a $(1-\alpha)100\%$ confidence level is given by

$$\hat{\theta}_j\pm z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_j)},$$

where $\mathrm{Var}(\hat{\theta}_j)$ is the jth diagonal element of $[\ddot{L}(\boldsymbol{\theta})]^{-1}$ estimated at $\widehat{\boldsymbol{\theta}}$, and $z_{\alpha/2}$ is the value of the standard normal distribution that probably exceeds $\alpha/2$.

4. Diagnostics tools and residual analysis

In this section, we present diagnostic analysis (global and local influence) and residuals for the heteroscedastic RIG regression.

4.1. Influence measures

Global influence

Let $l_{(i)}(\boldsymbol{\theta})$ be the log-likelihood function for $\boldsymbol{\theta}$ defined in (9) by excluding the ith observation and $\hat{\boldsymbol{\theta}}_{(i)}$ be the MLE of $\boldsymbol{\theta}$ obtained by maximizing $l_{(i)}(\boldsymbol{\theta})$. We can use the difference between $\hat{\boldsymbol{\theta}}_{(i)}$ and $\hat{\boldsymbol{\theta}}$ [37] as a measure to assess the influence of the ith case on the estimate $\hat{\boldsymbol{\theta}}$. This measure is a generalization of Cook's distance defined as a standardized form of $\hat{\boldsymbol{\theta}}_{(i)}-\hat{\boldsymbol{\theta}}$, namely

$$GD_i(\boldsymbol{\theta})=(\hat{\boldsymbol{\theta}}_{(i)}-\hat{\boldsymbol{\theta}})^\top\left[\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})\right](\hat{\boldsymbol{\theta}}_{(i)}-\hat{\boldsymbol{\theta}}), \tag{10}$$

where $\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})$ is defined in Section 3.1. Another measure to assess the influence of a case is called the likelihood distance [6]

$$LD_i(\boldsymbol{\theta})=2\left[l(\hat{\boldsymbol{\theta}})-l(\hat{\boldsymbol{\theta}}_{(i)})\right], \tag{11}$$

where $l(\boldsymbol{\theta})$ is the log-likelihood function of the complete sample. Therefore, the ith case is an influential observation if $\hat{\boldsymbol{\theta}}_{(i)}$ is distant from $\hat{\boldsymbol{\theta}}$ in (10) and/or (11).

Local influence

For measures of local influence [7] from small perturbations in the regression, let $\boldsymbol{\omega}$ be the perturbation vector, $l(\boldsymbol{\theta} \mid \boldsymbol{\omega})$ the log-likelihood function (9) of the disturbed model and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ the MLE obtained by maximizing $l(\boldsymbol{\theta} \mid \boldsymbol{\omega})$. Cook [7] showed that the normal curvature of the surface

$$\begin{pmatrix} \boldsymbol{\omega} \\ 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})] \end{pmatrix}$$

in the direction of the vector $\mathbf{d}$ of unit norm is defined as

$$C_{\mathbf{d}}(\boldsymbol{\theta}) = 2|\mathbf{d}^{\top}\boldsymbol{\Delta}^{\top}\ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1}\boldsymbol{\Delta}\mathbf{d}|, \tag{12}$$

where $\boldsymbol{\Delta}$ is the $(p+q+2) \times n$ matrix that depends on the perturbation scheme whose elements are $\Delta_{ji} = \partial^2 l(\boldsymbol{\theta} \mid \boldsymbol{\omega})/\partial\theta_j\partial\omega_i$ (for $i = 1, \ldots, n$ and $j = 1, \ldots, p+q+2$) evaluated at $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$, and $\boldsymbol{\omega}_0$ is the no perturbation vector [7]. Thus, using Equation (12), it is possible to calculate the maximum curvature $\mathbf{C}_{\mathbf{d}_{d\max}}$ in the corresponding direction, say $\mathbf{d}_{\max}$. The quantity $\mathbf{C}_{\mathbf{d}_{\max}}$ corresponds to the largest eigenvalue of the matrix $\mathbf{B} = \boldsymbol{\Delta}^{\top}\ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1}\boldsymbol{\Delta}$, and $\mathbf{d}_{\max}$ is the largest normalized eigenvector. Then, if $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\theta}}$ are close estimates, they indicate that the MLEs are robust with respect to the disturbance in the regression, while relevant differences suggest that the estimates are sensitive to such disturbances.

On the other hand, Lesaffre and Verbeke [19] proposed the statistic $C_{\mathbf{d}_i}(\boldsymbol{\theta})$, where $\mathbf{d}_i$ is a vector of zeros of length n with a value of one in the ith position. In this case, the curvature in the direction $\mathbf{d}_i$ takes the form $C_i = 2|\boldsymbol{\Delta}_i^{\top}\ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1}\boldsymbol{\Delta}_i|$, where $\boldsymbol{\Delta}_i^{\top}$ is the ith row of $\boldsymbol{\Delta}$. Therefore, the observations with values of C_i greater than $2\bar{C}$, where $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$, indicate the possibility of influential points.

Perturbation schemes

Thus, considering the log-likelihood function (9), the following perturbation schemes are adopted:

(a) case-weight perturbation scheme

Let $0 \leq \omega_i \leq 1$ $(i = 1, \ldots, n)$ and $\boldsymbol{\omega}_0 = (1, \ldots, 1)^{\top}$ be the non-perturbation vector of length n. The perturbed log-likelihood function reduces to

$$\begin{aligned} l(\boldsymbol{\theta} \mid \boldsymbol{\omega}) = &-\frac{\log(2\pi)}{2}\sum_{i=1}^{n}\omega_i\delta_i + \frac{3}{2}\sum_{i=1}^{n}\omega_i\delta_i\mathbf{x}_i^{\top}\boldsymbol{\beta} - \frac{1}{2}\sum_{i=1}^{n}\omega_i\delta_i\mathbf{z}_i^{\top}\boldsymbol{\gamma} \\ &- \frac{3}{2}\sum_{i=1}^{n}\omega_i\delta_i\log(t_i) - \sum_{i=1}^{n}\omega_i\delta_i\frac{\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})[t_i - \exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})]^2}{2t_i\exp(\mathbf{z}_i^{\top}\boldsymbol{\gamma})} \\ &+ \sum_{i=1}^{n}\omega_i(1-\delta_i)\log\left\{1 - \Phi\left[\kappa_i\left(\frac{t_i}{\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})} - 1\right)\right]\right. \\ &\left. - \exp\left(\frac{2\exp[2(\mathbf{x}_i^{\top}\boldsymbol{\beta})]}{\exp(\mathbf{z}_i^{\top}\boldsymbol{\gamma})}\right)\Phi\left[-\kappa_i\left(\frac{t_i}{\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})} + 1\right)\right]\right\}. \end{aligned}$$

(b) Response perturbation

Let $\omega_i \in \mathbb{R}$ $(i = 1, \ldots, n)$ and $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$ be the non-perturbation vector of length n. The response variable t_i $(i = 1, \ldots, n)$ is subjected to an additive perturbation scheme such that $t_i^* = t_i + \omega_i M_t$, where M_t it is a scale factor that can be the standard deviation of the response variable [9,28]. In this case, the perturbed log-likelihood function can be expressed as

$$
\begin{aligned}
l(\boldsymbol{\theta} \mid \boldsymbol{\omega}) = & -\frac{r\log(2\pi)}{2} + \frac{3}{2}\sum_{i=1}^{n}\delta_i \mathbf{x}_i^\top \boldsymbol{\beta} - \frac{1}{2}\sum_{i=1}^{n}\delta_i \mathbf{z}_i^\top \boldsymbol{\gamma} - \frac{3}{2}\sum_{i=1}^{n}\delta_i \log(t_i^*) \\
& - \sum_{i=1}^{n}\delta_i \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})[t_i^* - \exp(\mathbf{x}_i^\top \boldsymbol{\beta})]^2}{2t_i^* \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} + \sum_{i=1}^{n}(1-\delta_i) \\
& \times \log\left\{1 - \Phi\left[\kappa_i\left(\frac{t_i^*}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} - 1\right)\right] - \exp\left(\frac{2\exp[2(\mathbf{x}_i^\top \boldsymbol{\beta})]}{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}\right)\right. \\
& \left. \times\ \Phi\left[-\kappa_i\left(\frac{t_i^*}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} + 1\right)\right]\right\}.
\end{aligned}
$$

(c) Explanatory variable perturbation

Suppose that x_j and z_j denote continuous covariates subject to an additive perturbation scheme such that $x_{ij}^* = x_{ij} + \omega_i M_t$ and $z_{ij}^* = z_{ij} + \omega_i M_t$, where M_t is a scale factor that can be the standard deviation of the disturbed covariate [28]. Thus, the perturbed log-likelihood function takes the form

$$
\begin{aligned}
l(\boldsymbol{\theta} \mid \boldsymbol{\omega}) = & -\frac{r\log(2\pi)}{2} + \frac{3}{2}\sum_{i=1}^{n}\delta_i \mathbf{x}_i^{*\top} \boldsymbol{\beta} - \frac{1}{2}\sum_{i=1}^{n}\delta_i \mathbf{z}_i^{*\top} \boldsymbol{\gamma} - \frac{3}{2}\sum_{i=1}^{n}\delta_i \log(t_i) \\
& - \sum_{i=1}^{n}\delta_i \frac{\exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})[t_i - \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})]^2}{2t_i \exp(\mathbf{z}_i^{*\top} \boldsymbol{\gamma})} + \sum_{i=1}^{n}(1-\delta_i) \\
& \times \log\left\{1 - \Phi\left[\kappa_i\left(\frac{t_i}{\exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})} - 1\right)\right] - \exp\left(\frac{2\exp[2(\mathbf{x}_i^{*\top} \boldsymbol{\beta})]}{\exp(\mathbf{z}_i^{*\top} \boldsymbol{\gamma})}\right)\right. \\
& \left. \times\ \Phi\left[-\kappa_i\left(\frac{t_i}{\exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})} + 1\right)\right]\right\},
\end{aligned}
$$

where $\mathbf{x}_i^{*\top}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j[x_{ij} + \omega_i M_t] + \cdots + \beta_p x_{ip}$, $\mathbf{z}_i^{*\top}\boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_{i1} + \cdots + \gamma_j[z_{ij} + \omega_i M_t] + \cdots + \gamma_q z_{iq}$ and $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top$ is the non-perturbation vector of length n.

For the three perturbation schemes, the maximum curvature matrix, namely

$$
\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{vi})_{(p+q+2)\times n} = \left[\frac{\partial^2 l(\boldsymbol{\theta} \mid \boldsymbol{\omega})}{\partial\theta_v \partial\boldsymbol{\omega}_i}\right]_{(p+q+2)\times n}, \quad v = 1, \ldots, p+q+2,\ i = 1, \ldots, n,
$$

is calculated numerically, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ is the perturbation vector of the RIG regression or the observations.

4.2. Residual analysis

The residuals are used to identify the discrepancy between the fitted model and the data set and the presence of discrepant observations. Several types of residuals have been proposed in the literature; for example, Collett [4], Cook and Weisberg [8], and Paula [24]. Specifically, the martingale residuals and deviance residuals have been the most used ones in survival analysis as they take into account the information from censored times [28].

The martingale residuals (r_{M_i}) [12] are defined by

$$r_{M_i} = \delta_i + \log[S(t_i;\hat{\boldsymbol{\theta}})], \quad i = 1,\ldots,n, \tag{13}$$

where δ_i is the censoring indicator, and $S(t_i;\hat{\boldsymbol{\theta}})$ is the estimated survival function. By replacing the survival function (8) in Equation (13), the martingale residuals for the RIG regression take the forms

$$r_{M_i} = \begin{cases} 1 + \log\left\{1 - \Phi\left[\hat{\kappa}_i\left(\dfrac{t_i}{\exp(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})} - 1\right)\right]\right. \\ \quad \left. - \exp\left(\dfrac{2\exp[2(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})]}{\exp(\mathbf{z}_i^\top\widehat{\boldsymbol{\gamma}})}\right)\Phi\left[-\hat{\kappa}_i\left(\dfrac{t_i}{\exp(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})} + 1\right)\right]\right\} & \text{if } \delta_i = 1 \\ \log\left\{1 - \Phi\left[\hat{\kappa}_i\left(\dfrac{t_i}{\exp(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})} - 1\right)\right]\right. \\ \quad \left. - \exp\left(\dfrac{2\exp[2(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})]}{\exp(\mathbf{z}_i^\top\widehat{\boldsymbol{\gamma}})}\right)\Phi\left[-\hat{\kappa}_i\left(\dfrac{t_i}{\exp(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})} + 1\right)\right]\right\} & \text{if } \delta_i = 0, \end{cases} \tag{14}$$

where

$$\hat{\kappa}_i = \sqrt{\frac{\exp[3(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})]}{t_i\exp(\mathbf{z}_i^\top\widehat{\boldsymbol{\gamma}})}}, \quad \text{for } i = 1,\ldots,n.$$

However, the martingale residuals are not symmetrically distributed around zero, i.e. $r_M \in (-\infty, 1]$ [4], which makes difficult to interpret their plots. In order to overcome this problem, Therneau *et al.* [31] introduced the modified martingale residuals (also called the deviance residuals) to make them symmetrically distributed around zero, namely

$$r_{D_i} = \operatorname{sgn}(r_{M_i})\left\{-2\left[r_{M_i} + \delta_i\log(\delta_i - r_{M_i})\right]\right\}, \quad i = 1,\ldots,n, \tag{15}$$

where $\operatorname{sgn}(\cdot)$ is a sign function leading to $+1$ if the argument is positive and -1 if the argument is negative. By replacing the martingale residuals (14) in Equation (15), we obtain the deviance residuals for the RIG regression. The plot of the simulated envelope of the deviance residuals and the residuals versus the adjusted values allow to verify the adequacy of the regression.

5. Simulation study

Monte Carlo simulations are done for different scenarios to evaluate the behavior of the empirical distribution of the deviance residuals for the RIG regression as well as to determine the mean square errors (MSEs) of the MLEs of the parameters. For the simulation

Table 1. Averages of the MLEs and MSEs (in parentheses) of the parameters of the RIG regression.

		Percentages of censoring		
Sample size	Parameter	0%	20%	40%
50	β_0	2.7995 (0.0001)	2.7995 (0.0002)	2.7995 (0.0002)
	β_1	0.4005 (0.0002)	0.4005 (0.0003)	0.4004 (0.0004)
	γ_0	−20.0580 (3.9231)	−21.3330 (5.9023)	−21.5810 (7.9349)
	γ_1	6.1291 (0.2676)	6.1940 (0.3943)	6.2500 (0.5132)
100	β_0	2.7999 (< 0.0001)	2.8004 (0.0001)	2.8002 (0.0001)
	β_1	0.4000 (< 0.0001)	0.3997 (0.0001)	0.4000 (0.0001)
	γ_0	−20.8250 (1.8892)	−20.8470 (2.2419)	−20.9670 (2.9623)
	γ_1	6.0763 (0.13061)	6.0791 (0.1542)	6.1062 (0.1975)
300	β_0	2.8003 (< 0.0001)	2.8000 (< 0.0001)	2.8000 (< 0.0001)
	β_1	0.3998 (< 0.0001)	0.4000 (< 0.0001)	0.3998 (< 0.0001)
	γ_0	−20.6120 (0.5271)	−20.5810 (0.6475)	−20.6160 (0.8340)
	γ_1	6.0254 (0.0367)	6.0178 (0.0452)	6.0265 (0.0565)

study, the following conditions are considered according to the data set reported in the application:

(a) Set sample sizes $n = 50, 100$ and 300 and censoring percentages equal to 0%, 20% and 40% for nine scenarios.
(b) The parameter values are fixed at $\beta_0 = 2.8$, $\beta_1 = 0.4$, $\gamma_0 = -20.5$, $\gamma_1 = 6.0$, so that the two systematic components are

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1}) \quad \text{and} \quad \sigma_i^2 = \exp(\gamma_0 + \gamma_1 z_{i1}),$$

where x_1 and z_1 are generated from a binomial distribution with parameters $n = 1$ and $p = 1/2$ and a uniform distribution in the interval $[3.0, 4.5]$, respectively.
(c) The random values $t_1, \ldots, t_n$ of the RIG distribution (3) are generated using the method by Michael *et al.* [21] as described in Appendix. The censored data are obtained from a uniform distribution in the interval $[11.0, 33.8]$ for 40% censoring and $[11.0, 58.5]$ for 20% censoring.
(d) For each scenario, one thousand replicates are simulated.

Thus, the data are simulated according to the algorithm in Appendix. For each generated sample, the MLEs of the parameters are determined and then, for each fitted regression, we calculate the residuals r_M in (14) and r_D in (15).

The averages of 1000 MLEs and mean squared errors (MSEs) for the fitted regression are reported in Table 1. We can note that the estimates of the parameters are close to the true values, except if the sample size n is small and the censoring percentages are high. The MSE values increase when the percentage of censoring increases and decrease when n increases.

Figure 1 displays the plots of the deviance residuals versus the percentiles of the standard normal distribution for some scenarios. The plots reveal the following findings:

(a) The empirical distribution of the deviance residuals converges to the standard normal distribution when the percentage of censoring decreases.
(b) The empirical distribution also presents a better agreement with the standard normal distribution when the sample size increases.

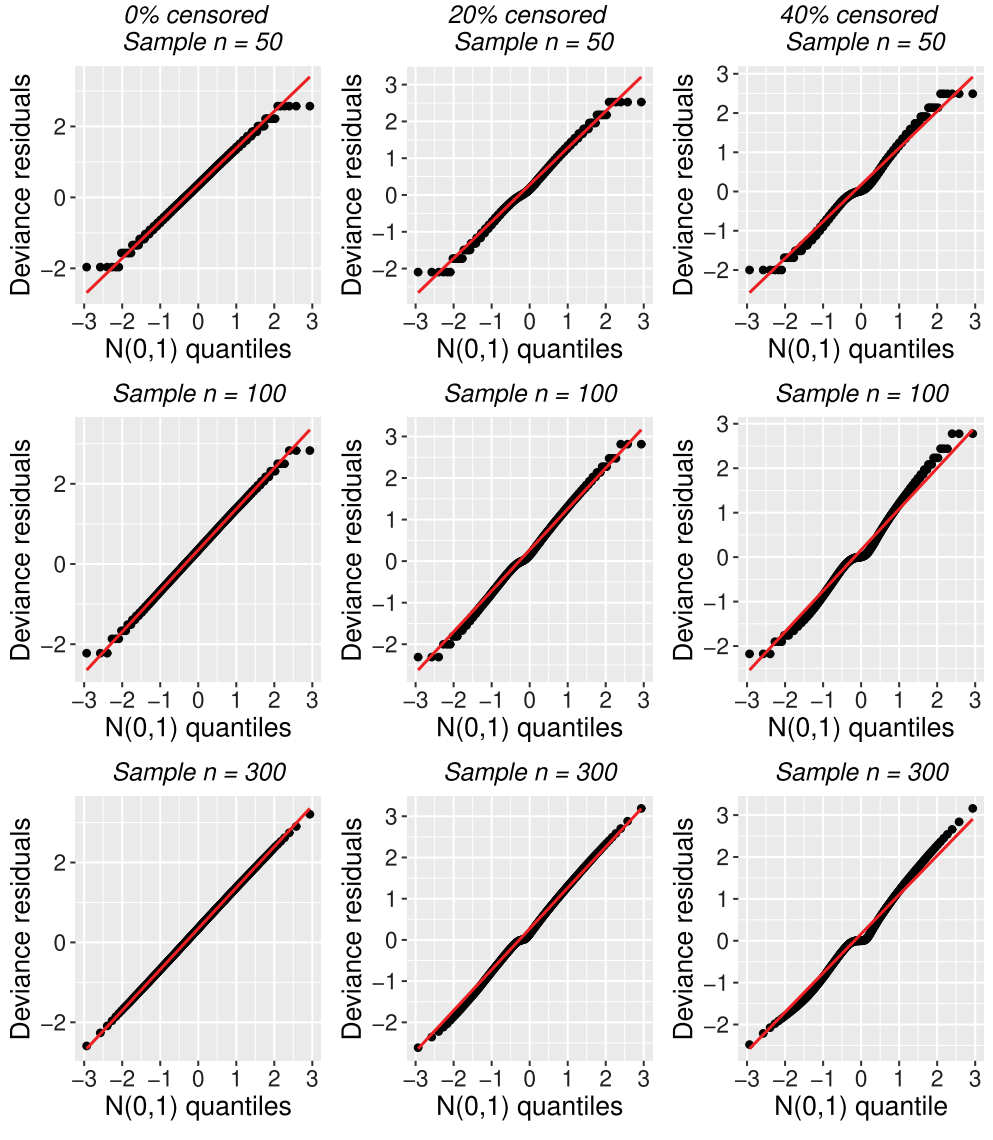

Figure 1. Normal probability plots for the deviance residuals in the RIG regression.

In general, the deviance residuals appear to be consonant to assess the model adequacy. Ortega *et al.* [23] showed that the deviance residuals were used in a data set with approximately 90% censoring, thus indicating that they can be adopted with high censoring percentages.

6. Application: COVID-19 length of stay

The city of Piracicaba is one of the first cities in Brazil to industrialize with the opening of factories in the metal-mechanical sector and to produce equipment for production of sugar, and later alcohol as well. This contributed significantly to the industrial growth of this city

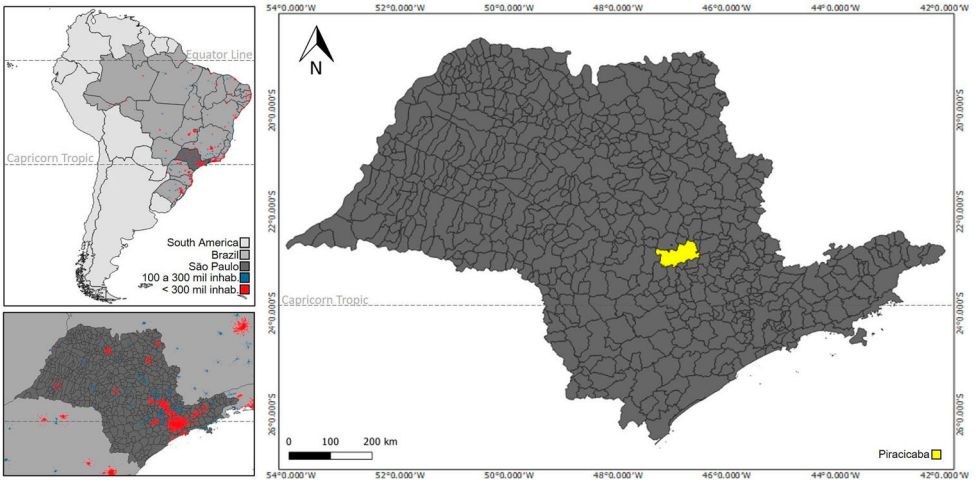

Figure 2. Piracicaba, São Paulo, Brazil.

in the ensuing decades, thus making it attractive to large companies and, consequently, moving the city with services and people. It is located 157 km from the capital of São Paulo (Figure 2), which facilitates the passage of people between the two cities. This heavy circulation of people might have facilitated the entry of the SARS-COV-2 virus in Piracicaba, where the first case of infection was recorded at the end of March 2020. Therefore, we analyze the profile of the patients who were hospitalized due to the new coronavirus.

The COVID-19 data for the city of Piracicaba were obtained from the database of the Unified Health System (SUS), available from the platform of the Department of Informatics of the Ministry of Health,[1] considering the municipal code 353870. Patient data regarding the principal diagnosis as COVID-19 were obtained from the same database according to the International Classification of Diseases (ICD-10), corresponding to code *B*34.2 (infection by coronavirus of unspecified location) and code *B*97.2 (coronavirus as cause of diseases classified in other chapters)[2,3] due to the absence of the category *U*07 in the volumes of ICD-10 in Portuguese (Cidade de São Paulo Saúde, 2020). The sample consists of 55 individuals admitted with COVID-19 to public hospitals (SUS), with censoring of approximately 70%, covering the period from March to May 2020, and the following variables (for $i = 1, \ldots, 55$):

(i) t_i: time (in days) from hospital admission until death.
(ii) δ_i: censoring indicator ($0 =$ censored, $1 =$ failure).
(iii) x_{i1}, z_{i1}: gender ($0 =$ Female, $1 =$ Male).
(iv) x_{i2}, z_{i12}: logarithm of age.

The Kaplan-Meier curves are displayed in Figure 3 to verify the behavior of the length of stay data of these patients. The plots show that the median hospital stay of these patients is 15 days (Figure 3(a)), and there is a more pronounced difference between genders up to 15 days (Figure 3(b)), thus indicating that men have greater chances of survival in the first days of hospitalization.

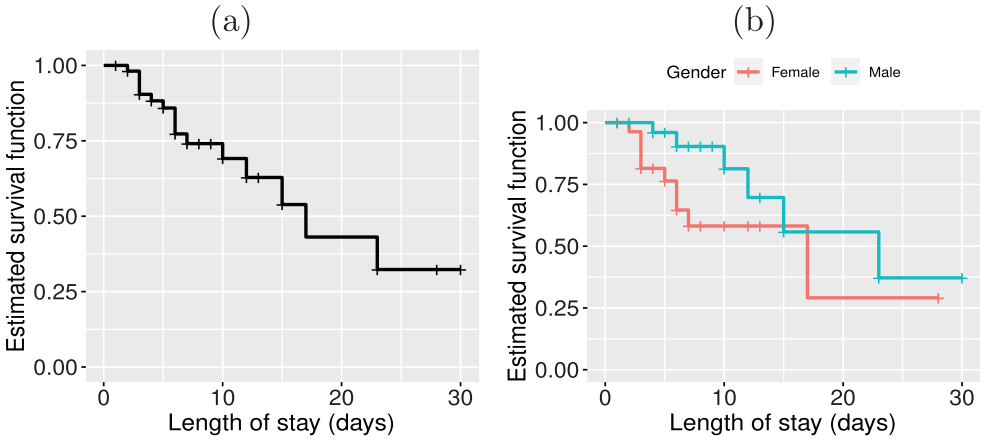

Figure 3. Estimated survival curves by the Kaplan-Meier method for the length of stay (a) and for gender (b).

Table 2. MLEs of the parameters of the RIG distribution for the length of stay of COVID-19 patients.

Parameter	Estimate	Standard error	Interval confidence(95%)
μ	33.7630	6.6585	(20.7126, 46.8133)
σ^2	2762.6396	1674.9181	(0.0000, 6045.4188)

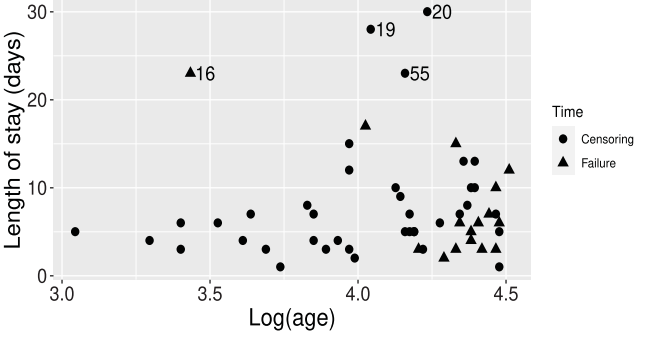

Figure 4. Plot of the hospital length of stay versus the logarithm of age for COVID-19 patients.

First, we fit the RIG density (3) to this data set. The MLEs, their standard errors (SEs) and confidence intervals for its parameters obtained with the initial values $\mu = 0.5$ and $\sigma^2 = 0.5$ are given in Table 2.

It can be noted from Table 2 that the mean length of stay in the hospital is approximately between 21 to 47 days. The variance estimate, indicates that there is a high variability in the data which can also be noted in Figure 4. The observations are more dispersed when the logarithm of the age increases and for this reason, the covariate logarithm of age is candidate to be used in the systematic component of the variance parameter.

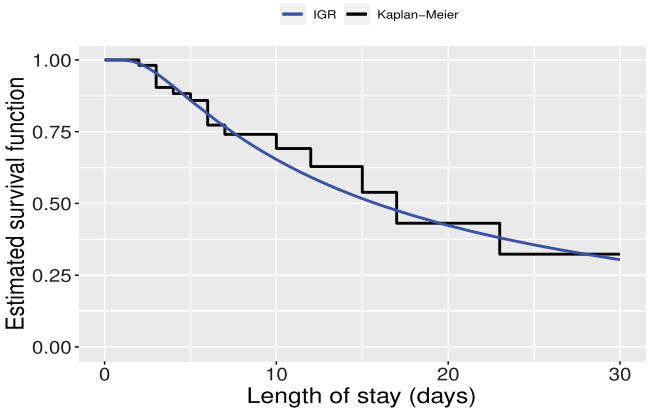

Figure 5. Estimated RIG survival function and the Kaplan-Meier curve for the length of hospital stay of COVID-19 patients.

In addition, the estimated survival function of the RIG distribution follows from (4)

$$\hat{S}(t) = 1 - \Phi\left(\sqrt{\frac{33.7630^3}{2762.6396\,t}}\left(\frac{t}{33.7630} - 1\right)\right) - \exp\left(\frac{2(33.7630^2)}{2762.6396}\right)\Phi\left(-\sqrt{\frac{33.7630^3}{2762.6396\,t}}\left(\frac{t}{33.7630} + 1\right)\right), \tag{16}$$

where t denotes the COVID-19 length of stay (in days).

The estimated RIG survival function (16) and the Kaplan-Meier curve are displayed in Figure 5. The estimated survival function reasonably superpose the empirical curve, which indicates that the density (3) can be used for the analysis of the current data.

The variables of the systematic components (5) and (6) are selected using the method described by Colosimo and Giolo [5]. The steps are as follows:

(a) Step 1: Estimate the RIG regression under a single variable:
- $\mu = \exp(\beta_0 + \beta_1 x_1)$ and $\sigma^2 = \exp(\gamma_0)$,
- $\mu = \exp(\beta_0 + \beta_2 x_2)$ and $\sigma^2 = \exp(\gamma_0)$,
- $\mu = \exp(\beta_0)$ and $\sigma^2 = \exp(\gamma_0 + \gamma_1 z_1)$,
- $\mu = \exp(\beta_0)$ and $\sigma^2 = \exp(\gamma_0 + \gamma_2 z_2)$.

(b) Step 2: The significant covariates in step 1 are estimated together:

$$\mu = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \quad \text{and} \quad \sigma^2 = \exp(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2).$$

We then estimate the reduced RIG regressions, excluding a single covariate:
- $\mu = \exp(\beta_0 + \beta_2 x_2)$ and $\sigma^2 = \exp(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)$,
- $\mu = \exp(\beta_0 + \beta_1 x_1)$ and $\sigma^2 = \exp(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)$,
- $\mu = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ and $\sigma^2 = \exp(\gamma_0 + \gamma_2 z_2)$,
- $\mu = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ and $\sigma^2 = \exp(\gamma_0 + \gamma_1 z_1)$.

Table 3. Selection of covariates.

Step	Model	$-2l(\hat{\boldsymbol{\theta}})$	LR Test	p-value
Step 1	Null	130.6695	–	–
	x_1	124.3520	6.3175	0.0120
	x_2	122.0522	8.6173	0.0033
	z_1	124.4036	6.2659	0.0123
	z_2	119.1318	11.5377	0.0007
Step 2	$x_1 + x_2$ and $z_1 + z_2$	108.6696	–	–
	x_2 and $z_1 + z_2$	108.8340	0.1644	0.6852
	x_1 and $z_1 + z_2$	109.8188	1.1492	0.2837
	$x_1 + x_2$ and z_2	109.0185	0.3489	0.5548
	$x_1 + x_2$ and z_1	116.2177	7.5481	0.0060
Step 3	z_2	119.1318	–	–
	x_1 and z_2	112.8101	6.3217	0.0119
	x_2 and z_2	118.2537	0.8781	0.3487
Final	x_1 and z_2	–	–	–

Table 4. MLEs of the parameters of the RIG regression for the length of stay of COVID-19 patients.

Parameter	Estimate	Standard error	p-value	Confidence interval (95%)
β_0	2.7613	0.1602	0.0000	(2.4473, 3.0753)
β_1	0.3972	0.1480	0.0081	(0.1033, 0.6911)
γ_0	−20.5155	10.1831	0.0439	(−40.4740, −0.5571)
γ_2	6.1270	2.3440	0.0090	(1.5329, 10.7210)

(c) Step 3: The covariates of the systematic component of the mean parameter excluded in step 2 are returned to the model to confirm that they are not statistically significant.

The results of the variable selection procedure are reported in Table 3.

Finally, we consider the following systematic components:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1}) \quad \text{and} \quad \sigma_i^2 = \exp(\gamma_0 + \gamma_2 z_{i2}).$$

The MLEs, their SEs, the confidence intervals and the significance of the parameters for the RIG regression are given in Table 4. The estimates are obtained by choosing $\beta_0 = 5.82275$, $\beta_1 = 0.36873$ and $\gamma_0 = \gamma_2 = 1$ as initial values.

The figures in Table 4 at a significance level of 5% indicate that there is a significant difference between genders in relation to the length of stay and that the mean length of hospital stay of men is higher than women since $\hat{\beta}_1$ is positive.

Continuing the analysis, the results of such influence measure index plots ($GD_i(\boldsymbol{\theta})$ and $LD_i(\boldsymbol{\theta})$) are displayed in Figure 6. These plots reveal that the case ♯16 is a possible influential observation.

The local influence plots, considering the case-weight perturbation, response perturbation and explanatory variable perturbation, reveal that the cases ♯16, ♯19 and ♯24 can be possible influential observations as shown in Figure 7(a–c).

To analyze the impact of these observations on the parameter estimates, the regression is refitted by eliminating each observation individually. Thus, the figures in Table 5 provide the relative changes (in percentages) of each estimated parameter defined by $\mathbf{RC}_{\theta_j} = [(\hat{\theta}_j - \hat{\theta}_{j(i)})/\hat{\theta}_j]100$, where $\hat{\theta}_{j(i)}$ is the jth MLE without the ith observation ($i = 1, \ldots, 55$ and $j = 1, \ldots, 3$).

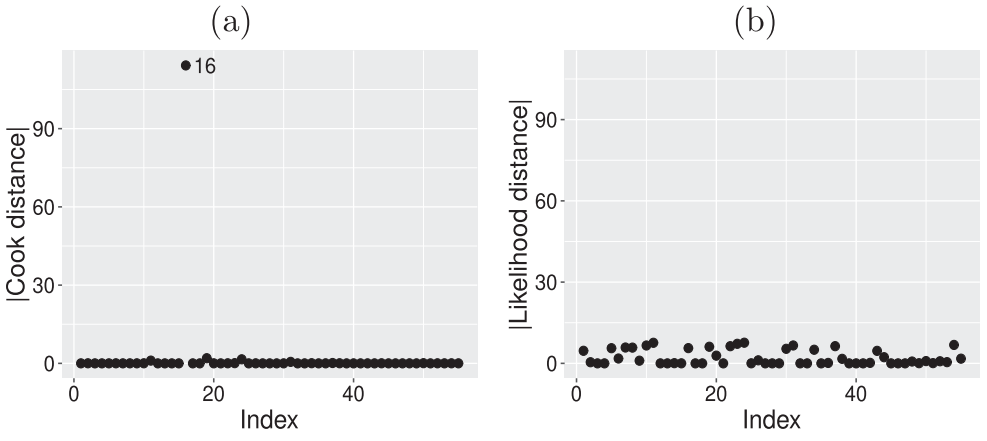

Figure 6. Global influence measures plots for the RIG regression model. (a) $GD_i(\boldsymbol{\theta})$ (b) $LD_i(\boldsymbol{\theta})$.

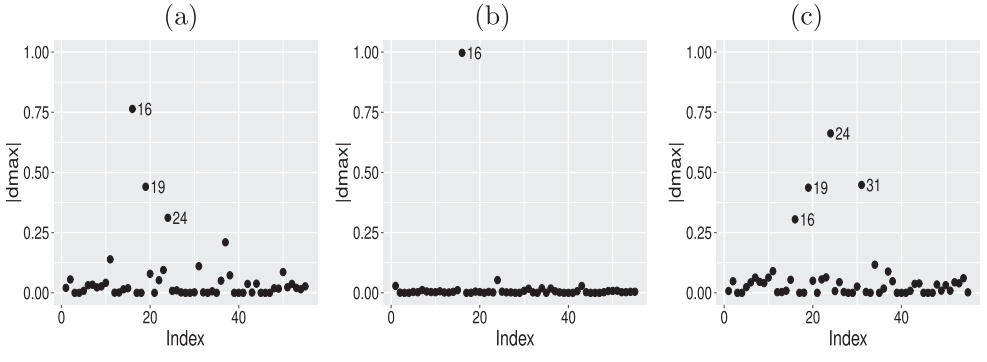

Figure 7. Local influence ($\mathbf{d}_{max}$) plots of the RIG regression for the length of stay of COVID-19 patients. (a) Case-weight perturbation, (b) Response perturbation and (c) Explanatory variable perturbation log(age).

The possible influential observations correspond to patients with the following characteristics:

(i) Patient ♯16: 31-year-old man (youngest in the adult age range in uncensored times) with the longest failure time (23 days) in the uncensored times.
(ii) Patient ♯19: 57-year-old woman (older age in the censored age group) with the second highest censored time (28 days). In relation to the group of women it is the longest.
(iii) Patient ♯24: 67-year-old woman with the failure time equal 3 days among the group of women whose time is not censored.
(iv) Patient ♯31: 73-year-old woman with the first shortest failure time (2 days) among the group of women whose time is not censored.

The results in Table 5 indicate that the MLEs of the parameters of the RIG regression are not very robust in relation to the deletion of influential observations (high percentages of CR). However, the significance of the estimated parameters do not change (at the 5% level

Table 5. Relative changes [-**RC**- in %], MLEs of the parameters and the corresponding p-values (in parentheses) by excluding the observations ♯16, ♯19, ♯24 and ♯31.

Exclusion	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_2$
None	–	–	–	–
	2.7613	0.3972	−20.5155	6.1270
	(0.0000)	(0.0081)	(0.0439)	(0.0090)
♯16	[−19]	[−16]	[20]	[9]
	3.2877	0.4590	−16.3116	5.5745
	(0.0000)	(0.0103)	(0.1393)	(0.0263)
♯19	[5]	[−27]	[−52]	[−39]
	2.6312	0.5060	−31.1382	8.5020
	(0.0000)	(0.0004)	(0.0121)	(0.0030)
♯24	[−3]	[26]	[−38]	[−29]
	2.8481	0.2936	−28.2207	7.9082
	(0.0000)	(0.0542)	(0.0117)	(0.0024)
♯31	[−2]	[17]	[−5]	[−4]
	2.8216	0.3314	−21.6272	6.3734
	(0.0000)	(0.0261)	(0.0300)	(0.0056)

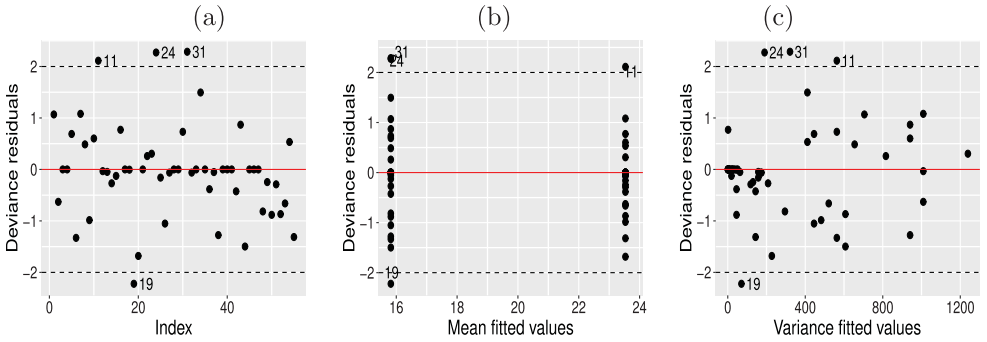

Figure 8. Residual plots of the RIG regression for the length of stay of COVID-19 patients. (a) Index, (b) Values adjusted for the mean and (c) Values adjusted for the variance.

of significance) after removing the cases, except for the parameter β_1 in which the p-value is very close to the significance level limit when ♯24 is removed. In general, it can be considered that there were no major inferential changes after the removal of the observations considered influential in the diagnostic charts. As there were no inferential changes, the observations were maintained, although the estimates are not robust.

The plots of the deviance residuals versus the index of the observations and the deviance residuals versus the adjusted values are displayed in Figure 8. It can be noted in these plots that the residuals are random around zero, which indicates that the regression is reasonably adequate to analyze the current data. In addition, we do not note outliers for the range $(-3, 3)$.

Finally, Figure 9 shows the normal probability plot for the deviance residuals with simulated envelope [2] of the RIG regression. So, there is evidence of a good fit of the regression for the length of hospitalization data, since the points are not outside the simulated envelope. The findings presented by Ortega *et al.* [23] also corroborate the adjustment of the model.

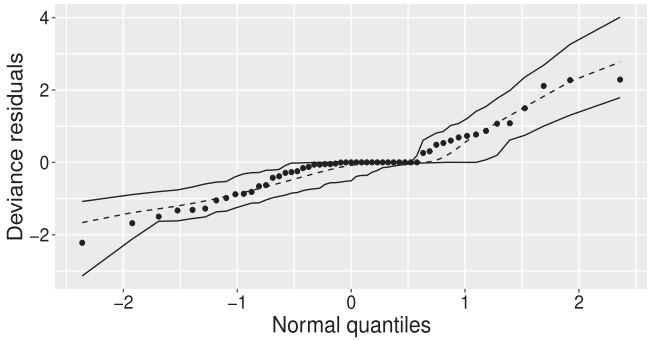

Figure 9. Normal probability plot of the RIG regression for COVID-19 data.

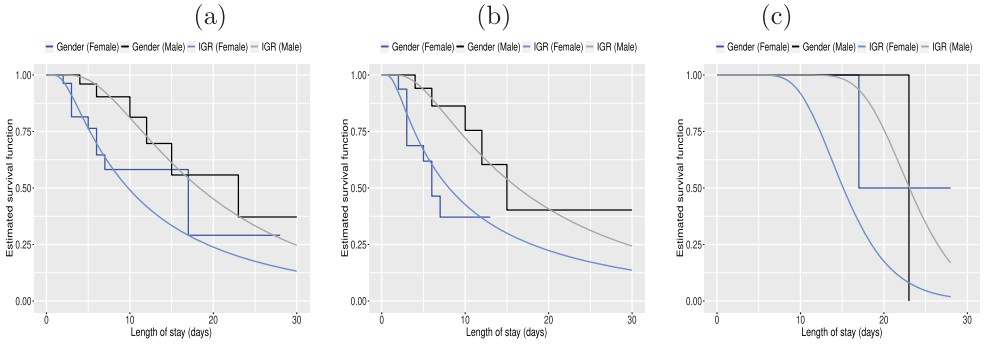

Figure 10. Survival curves estimated by the RIG distribution versus the survival curve estimated by the Kaplan-Meier method for the length of stay of patients with COVID-19. (a) Average age logarithm, (b) Logarithm of mean age greater than or equal to 60 years and (c) Logarithm of mean age less than 60 years.

The estimated survival function (8) for the length of stay data of COVID-19 patients is

$$\hat{S}(t) = \begin{cases} 1 - \Phi\left(\sqrt{\frac{\mu_0^3}{\bar{\sigma}^2 t}}\left(\frac{t}{\mu_0} - 1\right)\right) - \exp\left(\frac{2\mu_0^2}{\bar{\sigma}^2}\right) \\ \quad \times \Phi\left(-\sqrt{\frac{\mu_0^3}{\bar{\sigma}^2 t}}\left(\frac{t}{\mu_0} + 1\right)\right) & \text{if } x_1 = 0, \\ 1 - \Phi\left(\sqrt{\frac{\mu_1^3}{\bar{\sigma}^2 t}}\left(\frac{t}{\mu_1} - 1\right)\right) - \exp\left(\frac{2\mu_{1i}^2}{\bar{\sigma}^2}\right) \\ \quad \times \Phi\left(-\sqrt{\frac{\mu_1^3}{\bar{\sigma}^2 t}}\left(\frac{t}{\mu_1} + 1\right)\right) & \text{if } x_1 = 1, \end{cases} \tag{17}$$

where $\mu_0 = \exp(2.7613)$, $\mu_1 = \exp(3.1585)$, $\bar{\sigma}^2 = \frac{1}{55}\sum_{i=1}^{55} \hat{\sigma}_i^2$ and $\hat{\sigma}_i^2 = \exp(-20.5155 + 6.1270 z_{i2})$.

Figure 10(a) displays the estimated survival curves from (8) and the Kaplan-Meier method stratified by the gender group. Figure 10(b,c) are obtained by separating patients in an age group less than 60 and in an age group greater than or equal to 60 years, respectively. These plots indicate that the RIG regression can be acceptable for estimating the survival function of COVID-19 patients hospitalized since they follow the Kaplan-Meier curve, except for the adult age group. The behavior noted in Figure 10(c) is due to the large amount of censored times in this group.

7. Concluding remarks

We presented a re-parameterized inverse Gaussian (RIG) distribution in terms of the variance parameter with the objective to incorporate a multiplicative heteroscedasticity component in the regression. Based on this distribution, we can include the effects of explanatory variables without using a location-scale model, thus allowing interpretation about the mean without the need to transform the response variable. We defined the RIG regression to analyze the hospitalization time of COVID-19 patients in the city of Piracicaba, São Paulo, Brazil. The results indicated an average hospitalization time of 33 days, and that men tend to remain hospitalized longer than women. Further, some individuals were identified with characteristics outside the pattern of their group, which can help to better understand the effects of the new coronavirus.

Notes

1. DATASUS. Available in: https://datasus.saude.gov.br/. Accessed on: July 20, 2020.
2. Cidade de São Paulo Saúde. Orientações para o preenchimento de declaração de óbito. Available in: https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/saude/arquivos/mortalidade/Orientacoes_Preenchimento_DO_COVID-19-atualizada_2020.pdf. Accessed on: July 20, 2020.
3. Unimed-Rio. Lista de CID10. Available in: https://www.unimed-rio.com.br/static/cid10.xml. Accessed on: July 20, 2020.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the CNPq and CAPES, Brazil.

ORCID

E. M. M. Ortega http://orcid.org/0000-0003-3999-7402
G. M. Cordeiro http://orcid.org/0000-0002-3052-6551

References

[1] A. Achim, A. Loza, D. Bull, and N. Canagarajah, *Statistical modelling for wavelet–domain image fusion*, in *Image Fusion Algorithms and Applications*, Academic Press, Bristol, UK, 2008, pp. 119–138.
[2] A.C. Atkinson, *Plots, Transformations, and Regression*, University Press, Oxford, 1985.

[3] P. Basak and N. Balakrishnan, *Estimation for the three-parameter inverse Gaussian distribution under progressive type-II censoring*, J. Stat. Comput. Simul. 82 (2012), pp. 1055–1072.
[4] D. Collett, *Modelling Survival Data in Medical Research*, Chapman & Hall, London, 2003.
[5] E.A. Colosimo and S.R. Giolo, *Análise de Sobrevivência Aplicada*, Edgard Blacher, São Paulo, 2006.
[6] R.D. Cook, *Detection of influential observations in linear regression*, Technometrics 19 (1977), pp. 15–18.
[7] R.D. Cook, *Assessment of local influence (with discussion)*, J. R. Stat. Soc. 48 (1986), pp. 133–169.
[8] R.D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman & Hall, New York, 1982.
[9] J.N. da Cruz, E.M.M. Ortega, and G.M. Cordeiro, *The log-odd log-logistic Weibull regression model: Modelling, estimation, influence diagnostics and residual analysis*, J. Stat. Comput. Simul. 86 (2016), pp. 1516–1538.
[10] J. Doornik, *Ox: Object-Oriented Matrix Programming Using Ox*, Timberlake Consultants Ltd, London, 2007.
[11] S.L.P. Ferrari and F. Cribari-Neto, *Beta regression for modelling rates and proportions*, J. Appl. Stat. 31 (2004), pp. 799–815.
[12] T.T. Fleming and D.P. Harrington, *Counting Process and Survival Analysis*, Wiley, New York, 1991.
[13] J.L. Folks and R.S. Chhikara, *The inverse Gaussian distribution and its statistical application–a review*, J. R. Stat. Soc. Ser. B (Methodol.) 40 (1978), pp. 263–289.
[14] R.C. Gupta and D. Kundu, *Weighted inverse Gaussian a versatile lifetime model*, J. Appl. Stat. 12 (2011), pp. 2695–2708.
[15] D.D. Hanagal and S.M. Bhambure, *Modeling bivariate survival data using shared inverse Gaussian frailty model*, Comm. Statist. Theory Methods 45 (2016), pp. 4969–4987.
[16] N.L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, John Wiley and Sons, New York, 1994.
[17] J.P. Klein and M.L. Moeschberguer, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York, 2013.
[18] J.F. Lawless, *Statistical Models and Methods for Lifetime Data*, Springer, New York, 2011.
[19] E. Lesaffre and G. Verbeke, *Local influence in linear mixed models*, Biometrics 54 (1998), pp. 570–582.
[20] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, Chapman & Hall, Boca Raton, 1989.
[21] J.R. Michael, W.R. Schuncany, and R.W. Hass, *Generating random variates using transformations with multiple roots*, Am. Stat. 30 (1976), pp. 88–90.
[22] D.B. Nelson, *Conditional heteroskedasticity in asset returns: A new approach*, Econometrica 59 (1991), pp. 347–370.
[23] E.M.M. Ortega, A.J. Lemonte, G.M. Cordeiro, V.G. Cancho, and F.L. Mialhe, *Heteroscedastic log-exponentiated Weibull regression model*, J. Appl. Stat. 45 (2018), pp. 384–408.
[24] G.A. Paula, *Modelos de regressão: Com apoio computacional*, IME/USP, São Paulo, 2013. Available at http://www.ime.usp.br/?giapaula/cursospos.htm. Accessed on: July 30, 2020.
[25] A. Punzo, *A new look at the inverse Gaussian distribution with applications to insurance and economic data*, J. Appl. Stat. 46 (2019), pp. 1260–1287.
[26] R.B. Salha and A.J. Rasheed, *A comparision study between three different Kernel estimators for the Hazard rate function*, Electron. J. Appl. Statist. Anal. 10 (2017), pp. 1–13.
[27] E. Schrodinger, *Zur theorie der fall-und steigversuche und teilchen mit brownscher bewegung*, Phys. Z. 16 (1915), pp. 289–295.
[28] G.O. Silva, E.M.M. Ortega, and V.G. Cancho, *Log-weibull extended regression model: Estimation, sensitivity and residual analysis*, Stat. Methodol. 7 (2010), pp. 614–630.
[29] H. Songhua, Y. Jun, and C. Berenguer, *Degradation analysis based on an extended inverse Gaussian process model with skew-normal random effects and measurement errors*, Reliab. Eng. Syst. Safety 189 (2019), pp. 261–270.

[30] A.K. Suzuki, V.G. Cancho, and F. Louzada, *The Poisson-Inverse-Gaussian regression model with cure rate: A Bayesian approach and its case influence diagnostics*, Statist. Papers 57 (2016), pp. 133–159.
[31] T.M. Therneau, P.M. Grambsch, and T.R. Fleming, *Martingale-based residuals for survival models*, Biometrika 77 (1990), pp. 147–160.
[32] M.C.K. Tweedie, *Inverse statistical variates*, Nature 155 (1945), p. 453.
[33] M.C.K. Tweedie, *Statistical properties of inverse Gaussian distributions. I*, Ann. Math. Statist. 28 (1957), pp. 362–377.
[34] M.C.K. Tweedie, *Statistical properties of inverse Gaussian distributions. II*, Ann. Math. Statist. 28 (1957), pp. 362–377.
[35] J.C.S. Vasconcelos, G.M. Cordeiro, E.M.M Ortega, and E.G. Araujo, *The new odd log-logistic generalized inverse Gaussian regression model*, J. Probab. Stat. 2019 (2019), pp. 1–13.
[36] G.A. Whitmore, *A regression method for censored inverse-Gaussian data*, Canad. J. Statist. 11 (1983), pp. 305–315.
[37] F. Xie and B. Wei, *Diagnostics analysis for log-Birnbaum-Saunders regression models*, Comput. Statist. Data Anal. 51 (2007), pp. 4692–4706.

Appendix. Algorithm

We present below the Algorithm 1 showing the steps to generate random values from the RIG regression.

Algorithm 1: Simulation study with different percentages of censoring

Input: r - number of replications
n - sample size
β_0 - parameter for the mean
β_1 - parameter for the mean
γ_0 - parameter for the variance
γ_1 - parameter for the variance

$j = 1$
while $j \leq r$ **do**
  $g\_mX_1 \sim$ binomial$(1, 0.5)$
  $g\_mX_2 \sim \mathcal{U}(3, 4.5)$
  g_mD $n \times 2$ matrix of ones
  for $i = 1ton$ **do**
    $\mu = \exp(\beta_0 + (\beta_1 * g\_mX_{1i}))$
    $\sigma^2 = \exp(\gamma_0 + (\gamma_1 * g\_mX_{2i}))$
    $\nu_0 \sim \chi_{(1)}$
    $\lambda = \frac{\mu^3}{\sigma^2}$
    $x_1 = \mu + \frac{mu^2\,\nu_0}{2\lambda} - \left[\frac{\mu}{2\lambda}\sqrt{(4\mu\lambda\nu_0) + (\mu^2\nu_0^2)}\right]$
    $x_2 = \frac{\mu^2}{x_1}$
    $p_1 = \frac{\mu}{\mu + x_1}$
    $Y \sim \mathcal{U}(0, 1)$
    if $Y < p_1$ **then**
      $g\_mT = x_1$
    else
      $g\_mT = x_2$
    end
    $g\_mC \sim \mathcal{U}(a, b)$, where a and b are chosen to obtain the censoring percentages
    if $g\_mT < g\_mC$ **then**
      $g\_mD_{i1} = g\_mT$
    else
      $g\_mD_{i1} = g\_mC$
      $g\_mD_{i2} = 0$
    end
  end
  Arrange side by side g_mD, g_mX_1 and g_mX_2 in g_mY
  Obtain the estimate parameters
  if *if the optimization method converges* **then**
    Save parameter estimates to an object
    Determine residues (martingale and deviance) and store them in an object
    update j
  else
    not updated j
  end
end
Determine the average of the r estimates
Determine the mean squared error