ON GAMMA REGRESSION RESIDUALS

María Victoria Cifuentes[a], Martha Corrales[b], Héctor Zarate[c],
Edilberto Cepeda-Cuervo[d]

Departamento de estadistica, Universidad Nacional, Bogotá, Colombia

Abstract

In this paper we propose a new residuals for gamma regression models, assuming that both mean and shape parameters, follow regression structures. The models are summarized and fitted by applying both classic and Bayesian methods as proposed by Cepeda-Cuervo. The residuals are proposed from properties of the biparametric exponential family of distributions and simulated and real data sets are analyzed to determine the performance and behavior of the proposed residuals.

***Key words*:** Gamma regression, Fisher scoring algorithm, Bayesian estimation, residuals.

1. Introduction

The gamma distribution can be used for regression models with more flexibility than other models, such as exponential and poisson, among others. Thus, gamma regression models allow for a monotone, no constant hazard in survival models, and have the reproductive property that the sums of independent gamma distributions are also gamma distributed. Moreover, gamma regression models have the advantage of providing a count-data model with substantially higher flexibility than the Poisson model, which involves very sparse time-series that can be modeled by the gamma regression (Bateson 2009). These models are extended in a wide range of empirical applications, such as in the process of rate setting in the frame-work of heterogeneous insurance portfolios, which is the most important function of insurers (Krishnamoorthy 2006), and in a hospital admissions for rare diseases where time series are very sparse due to infrequency of events (Winklemann 2008).

This paper considers gamma regression models in which both the mean and the dispersion are allowed to depend on unknown parameters and on covariates. Joint modeling of the mean and the shape parameters in gamma regressions were proposed by Cepeda-Cuervo (2001), under both classical and Bayesian approaches. In the former, the parameters are estimated by an alternative iterated maximum likelihood method based on the Fisher scoring algorithm. In the Bayesian approach,

[a] E-mail: mvcifuentesa@unal.edu.co
[b] E-mail: mlcorralesb@unal.edu.co
[c] E-mail: hmzarates@unal.edu.co
[d] E-mail: ecepedac@unal.edu.co

estimations of the regression parameters are obtained by a hybrid Metropolis Hasting algorithm, as in Chib & Greenberg (1995) and Gamerman & Lopes (2006).

Several definitions of residuals are possible for generalized linear models (McCullagh & Nelder 1989). Some uses of generalized residuals include: building goodness of fit measures to check for systematic departure from the model, checking the variance function and the link function, examining them to identify poorly fitting observations, and plotting them to examine effects of new covariates or nonlinear effects of the covariates included in the model. Some of the relevant works related to residuals in generalized linear models are presented in Cox & Snell (1968), Pierce & Schafer (1986) and Dobson (2010).

In this paper we propose and adjust two residuals for gamma regression models. Simulated and real data applications are used to evaluate the benefits and interpretation of the proposed residuals.

After the introduction, this paper includes six sections. In Section 2, a re-parameterization of the gamma distribution is presented. In Section 3, the gamma regression model setting under the classic and Bayesian approaches is summarized. Section 4 presents the residuals obtained under the two-parameter exponential family of properties. Section 5 contains an application based on simulated data. In this Section, we mention two application cases: the first one is based on simulated gamma data and is useful to evaluate residuals' behavior, whereas the second application use data from study presented in McCullagh & Nelder (1989) related with the duration of embryonic stage in fruit fly life, and where we calculated the gamma residuals to measure adjustment of the model proposed by the authors. Finally in Section 6 we present our main conclusions.

2. Re-parameterized gamma distribution

A random variable y has gamma distribution if its density function is given by:

$$f(y;\lambda,\alpha)=\frac{\lambda}{\Gamma(\alpha)}(\lambda y)^{\alpha-1}e^{-\lambda y}I_{(0,\infty)}(y) \tag{1}$$

where $\lambda>0$, $\alpha>0$, $\Gamma(.)$ is the gamma function and $I(.)$ is an indicator function. Under this parameterization, the mean and variance of y are given by $\mu=E(Y)=\alpha/\lambda$ and $\mathrm{Var}(Y)=\alpha/\lambda^2=\mu^2(1/\alpha)$, respectively.

Setting $\lambda=\alpha/\mu$, Cepeda-Cuervo (2001) and Cepeda & Gammerman (2005) write the gamma density function (1), in terms of the mean and shape parameters as follows:

$$f(y)=\frac{1}{y\Gamma(\alpha)}\left(\frac{\alpha y}{\mu}\right)^{\alpha}e^{-\alpha y/\mu}I_{(0,\infty)}(y) \tag{2}$$

Under this re-parameterization, we use $y \sim G(\mu, \alpha)$ to denote that y follows a gamma distribution with $E(y) = \mu$ and α as a shape parameter. The variance of y is now given by $var(y) = \mu^2/\alpha$.

3. Gamma regression models

Let $y_i \sim G(\mu_i, \alpha), i = 1, \ldots, n$, be independent random variables. Then the gamma regression models is defined as

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown regression parameters $(p < n)$, $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})'$ is the vector of p covariates and η_i is a linear prediction. Usually $x_{i1} = 1$, for all i, so that model has a mean intercept. The link function $g(.) : (0, \infty) \to \mathbb{R}$ should be a strictly monotonic twice differentiable function in the classic approach once time differentiable in the Bayesian approach.

Some usual link functions in the gamma regression are: log $g(\mu) = log(\mu)$; identity $g(\mu) = \mu$; and inverse $g(\mu) = 1/\mu$. In the exponential family, the canonical link for the mean is the inverse function (McCullagh & Nelder 1989).

An extension of the gamma regression model proposed by Cepeda-Cuervo (2001) is the variable shape gamma regression model. In this model, the shape parameter is not constant through the observations and it is modeled following a regression structure. That is, $y_i \sim G(\mu_i, \alpha_i)$, $i = 1, \ldots, n$, are independent random variables with gamma distribution, where mean and shape parameters follow a regression structure given by:

$$g(\mu_i) = \eta_{1i} = \boldsymbol{x_i'\beta} \tag{4}$$

$$h(\alpha_i) = \eta_{2i} = \boldsymbol{z_i'\gamma} \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)'$, $p + k < n$, are the vectors of regression parameters related to the mean and dispersion, g is the mean link function, h is the shape link function (usually the log function), η_{1i} and η_{2i} are the linear predictors, and $\boldsymbol{x_i}$ and $\boldsymbol{z_i}$ are the covariates.

3.1. Classic Estimation

Cepeda-Cuervo (2001) proposed a classic approach to fit joint mean and shape gamma regression models using the Fisher scoring algorithm. In that work, he showed that under the reparameterization of the gamma distribution given by (2), the likelihood function of the gamma regression models defined by (4) and (5) is given by:

$$L = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)} \left(\frac{\alpha_i}{\mu_i}\right)^{\alpha_i} y_i^{\alpha_i - 1} \exp\left(-\frac{\alpha_i}{\mu_i} y_i\right) \tag{6}$$

and the log likelihood function by

$$l = \sum_{i=1}^{n} \left\{ -\log[\Gamma(\alpha_i)] + \alpha_i \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - \log(y_i) - \left(\frac{\alpha_i}{\mu_i}\right) y_i \right\} \tag{7}$$

Thus, assuming that $\mu_i = \mathbf{x_i'}\boldsymbol{\beta}$ and $\alpha_i = \exp(\mathbf{z_i'}\boldsymbol{\gamma})$, the components of the score function are:

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} -\frac{\alpha_i}{\mu_i}\left(1 - \frac{y_i}{\mu_i}\right) x_{ij}, j = 1, \dots p \\
\frac{\partial l}{\partial \gamma_k} &= \sum_{i=1}^{n} -\alpha_i \left[\frac{d}{d\alpha_i}\log\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right] z_{ik}, k = 1, \dots, r
\end{aligned}$$

On the other hand, the Hessian matrix is determined by:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_k \beta_j} &= \sum_{i=1}^{n} \frac{\alpha_i}{\mu_i^2}\left(1 - \frac{2y_i}{\mu_i}\right) x_{ij} x_{ik}, j, k = 1, \dots p \\
\frac{\partial^2 l}{\partial \gamma_k \beta_j} &= \sum_{i=1}^{n} -\alpha_i \left[\frac{d}{d\alpha_i}\log\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right] z_{ik}, k = 1, \dots, r \\
\frac{\partial^2 l}{\partial \gamma_k \gamma_j} &= \sum_{i=1}^{n} -\alpha_i \left[\frac{d}{d\alpha_i}\log\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right] z_{ik}, k = 1, \dots, r
\end{aligned}$$

The Fisher information matrix is given by:

$$\begin{aligned}
-E\left(\frac{\partial^2 l}{\partial \beta_k \beta_j}\right) &= \sum_{i=1}^{n} \frac{\alpha_i}{\mu_i^2} x_{ji} x_{ki}, j, k = 1, \cdots, p \\
-E\left(\frac{\partial^2 l}{\partial \gamma_k \beta_j}\right) &= 0, k = 1, \cdots, r; j = 1, \cdots, p \\
-E\left(\frac{\partial^2 l}{\partial \beta_k \beta_j}\right) &= \sum_{i=1}^{n} \alpha_i^2 \left[\frac{d^2}{d\alpha_i^2}\log\Gamma(\alpha_i) - \frac{1}{\alpha_i}\right] z_{ij} z_{ki}, j, k = 1, \cdots, r
\end{aligned}$$

It can be noted that the Fisher information matrix is a block diagonal matrix, where one of the blocks corresponds to the mean regression parameters and the other to the shape regression parameters. Thus the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal, and their maximum likelihood estimators, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, are asymptotically independent. As a consequence of this result,Cepeda-Cuervo (2001) proposed an iterative algorithm to obtain the maximum likelihood estimates of the regression parameters, where given the k-th parameter values $(\boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)})'$, the mean vector of the regression parameters is updapted from:

$$\boldsymbol{\beta^{(k+1)}} = \boldsymbol{(X'W^{(k)}X)^{-1}X'W^{(k)}Y} \tag{8}$$

where $\boldsymbol{W^{(k)}}$ is a matrix with diagonal elements $w_i^{(k)} = (\mu_i^2)^{(k)}/\alpha_i^{(k)}$, and given $(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\gamma}^{(k)})'$, the shape regression parameters $\boldsymbol{\gamma}^{(k+1)}$ updated from the equation:

$$\boldsymbol{\gamma^{(k+1)} = (Z'W^{(k)}Z)^{-1}X'W^{(k)}Y} \tag{9}$$

where $\boldsymbol{W^{(k)}}$ is a matrix with elements $w_i^{(k)} = 1/d_i^{(k)}$, with

$$d_i = \alpha_i^{-2}\left[\frac{d^2}{d\alpha_i^2}\log\Gamma(\alpha_i)\frac{1}{\alpha_i}\right]^{-1} \tag{10}$$

Therefore, given the initial values of the parameters an alternate iterate algorithm can be summarized as follows: Step 1. $\boldsymbol{\beta}^{(k+1)}$ is obtained from equation (8), giving the current values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$; Step 2. $\boldsymbol{\gamma}^{(k+1)}$ is obtained from equation (9), giving the current values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Steps 1 and 2 are repeated until convergence.

3.2. Bayesian estimation

In this section we summarize the Bayesian method proposed Cepeda-Cuervo (2001) to fit gamma regression models, where both mean and shape parameters follows regression structures. In this proposal, without loss of generality, independent normal prior distributions are assumed for the mean and shape regression parameters:

$$\begin{aligned}\boldsymbol{\beta} &\sim N(\mathbf{b}, \mathbf{B})\\ \boldsymbol{\gamma} &\sim N(\mathbf{g}, \mathbf{G})\end{aligned}$$

Let $L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{Y}, \mathbf{X}, \mathbf{Z})$ be the likelihood function and $p(\boldsymbol{\beta}, \boldsymbol{\gamma})$ the joint prior distribution. Given that the posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \sim L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{Y}, \mathbf{X}, \mathbf{Z})p(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and all their conditional distributions $\pi_\beta(\boldsymbol{\beta}|\boldsymbol{\gamma}, \mathbf{Y}, \mathbf{X}, \mathbf{Z})$ and $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}, \mathbf{Z})$ are analytically intractable, an alternate Metropolis Hastings algorithm is proposed to obtain samples of the posterior parameters.

In this algorithm, samples of the conditional posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}, \mathbf{Y}, \mathbf{X}, \mathbf{Z})$ are proposed from the kernel transition function, which is given by:

$$q_1(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = N(\mathbf{b}^*, \mathbf{B}^*) \tag{11}$$

where

$$\begin{aligned}\mathbf{b}^* &= \mathbf{B}^*(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\widetilde{Y})\\ \mathbf{B}^* &= (\mathbf{B}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\end{aligned}$$

For identity and log ($h = log$) mean link functions, the components of the working variables $\widetilde{Y}$ are $\widetilde{y_1}_i = y_i$ and $\widetilde{y_1}_i = \mathbf{x}_i'\boldsymbol{\beta} + y_i/\mu_i - 1$, respectively. $\boldsymbol{\Sigma}$ is a diagonal matrix with $w_i = Var(\widetilde{y_1}_i)$, $i = 1, \ldots, n$, as diagonal elements.

Samples of the posterior conditional distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}, \mathbf{Z})$ are proposed from the kernel transition function

$$q_2(\boldsymbol{\gamma}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = N(\mathbf{g}*, \mathbf{G}*) \tag{12}$$

where

$$\mathbf{g}^* = \mathbf{G}^*(\mathbf{G}^{-1}\mathbf{g} + \mathbf{X}'\Psi^{-1}\widetilde{Y})$$
$$\mathbf{G}^* = (\mathbf{G}^{-1} + \mathbf{X}'\Psi^{-1}\mathbf{X})^{-1}$$

For log link function for the shape, the working variable is $\tilde{y}_{2i} = \mathbf{z}_i'\boldsymbol{\gamma} + y_i/\mu_i - 1$. $\boldsymbol{\Psi}$ is a diagonal matrix with $\varphi_i = Var(\widetilde{y}_{2i})$, $i = 1, \dots, n$.

For more details about this algorithm and its applications, see Cepeda-Cuervo (2001) and Cepeda & Gammerman (2005).

With the kernel transition functions defined by (11) and (12), the hybrid Metropolis Hasting algorithm is defined by the following steps:

1. Begin the chain iteration counter at j=1
2. Set initial chain values $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$ for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively.
3. Propose a new value $\boldsymbol{\phi}$ for $\boldsymbol{\beta}$, generated from 11.
4. Calculate the acceptance probability, $\alpha(\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\phi})$. If the movement is accepted, then $\boldsymbol{\beta}^{(j)} = \boldsymbol{\phi}$. If not accepted, then $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$
5. Propose a new value $\boldsymbol{\phi}$ for $\boldsymbol{\gamma}$, generated from 12.
6. Calculate the acceptance probability, $\alpha(\boldsymbol{\gamma}^{(j-1)}, \boldsymbol{\phi})$. If the movement is accepted, then $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\phi}$. If not accepted, then $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$
7. Change the counter from j to $j+1$ and return to 2 until convergence is reached.

The convergence can be verified empirically in different ways (for details see Gammerman (1997a) and Heidelberger & Welch (1981)).

4. Gamma regression residuals

Residual analysis aims to identify outliers and/or model misspecification. It can be based on ordinary residuals, standardized variants or deviance residuals. Residuals are measures of agreement between the observed responses and the fitted conditional mean. Most residuals are based on the differences between the observed responses and the fitted conditional mean. For the gamma regression, where both mean and shape parameters follows regression structures, we define a first standardized ordinal residual as follows:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var}(y_i)}} \tag{13}$$

where

$$\widehat{var}(y_i) = \frac{\hat{\mu}^2}{\hat{\alpha}_i} \tag{14}$$

A second residual considered in this paper is the deviance residual, which for gamma regression models is given by:

$$r_i^d = -2\sum_{i=1}^{n}\left[\log\left(\frac{y_i}{\widehat{\mu}_i}\right) - \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right] \tag{15}$$

where $\hat{\mu}_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$.

In order to define gamma residuals from the two parameter exponential family we re-parameterized the gamma density function in a natural way, as follows in equation (18), where $\eta_1 = \alpha, T_1 = \log(y), \eta_2 = -\frac{\alpha}{\mu}, T_2 = y, d_0(\eta_1, \eta_2) = \eta_1 \log(\eta_2) - \log\Gamma(\eta_1), S(y) = -\log(y)$.

$$\begin{aligned}
f(y) &= \exp\left[-\log\Gamma(\alpha) + \alpha\log\left(\frac{\alpha y}{\mu}\right) - \frac{\alpha y}{\mu} - \log(y)\right] && (16)\\
&= \exp\left[\alpha\log(y) - \left(\frac{\alpha}{\mu}\right)y + \alpha\log\left(\frac{\alpha}{\mu}\right) - \log\Gamma(\alpha) - log(y)\right] && (17)\\
&= \exp\left[\eta_1 T_1(y) + \eta_2 T_2(y) + \eta 1\log(\eta_2) - \log\Gamma(\eta_1) + S(y)\right] && (18)
\end{aligned}$$

Thus, from the properties of the bi-parametric exponential family of distributions,

$$\begin{aligned}
E(T_1) &= = -\frac{\partial d_0}{\partial \eta_1} = -[\log(\eta_2) - \Psi(\eta_1)] && (19)\\
E(T_2) &= -\frac{\partial d_0}{\partial \eta_2} = -\frac{\eta_1}{\eta_2} = \mu && (20)
\end{aligned}$$

where the digamma function, $\Psi(\eta_1)$, is defined as the derivative of the logarithm of the gamma function

$$\Psi(\eta_1) = \frac{d\log\Gamma(\eta_1)}{d\eta_1} = \frac{\Gamma'(\eta_1)}{\Gamma(\eta_1)} \tag{21}$$

From the same properties of the biparametric exponential family, the variances of T_1 and T_2 are given by:

$$\begin{aligned}
Var(T_1) &= -\frac{\partial^2 d_0}{\partial \eta_1^2} = \Psi'(\eta_1) && (22)\\
Var(T_2) &= -\frac{\partial^2 d_0}{\partial \eta_2^2} = \frac{\eta_1}{\eta_2^2} = \frac{\mu^2}{\alpha} && (23)
\end{aligned}$$

where $\Psi'(\eta_1)$ denotes the derivative of the digamma function estimated on η_1

From this results, two gamma residuals can be proposed. The first one from (19) and (22), is given by:

$$r_i^* = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{v\hat{a}r(y_i^*)}} \tag{24}$$

where $y^* = T_1(y) = log(y)$, $\mu_i^* = E(T_1(y)) = E(y^*)$ and $var(y_i^*) = var(T_1(y)) = \Psi'(\eta_1)$. This residual is computed as the difference between y^* and $\hat{\mu}^*$,the difference between y^* and the estimates of the expected value $\mu^* = E(y^*)$, divided by the squared root of the estimation of the variance $var(y^*)$.

Now from (20) and (23), a second residual can be defined, in this case given by:

$$r_i^+ = \frac{y_i^+ - \hat{\mu}_i^+}{\sqrt{\widehat{var}(y_i^+)}} \tag{25}$$

where $y^+ = T_1(y) = y$, $\mu_i^+ = E(T_1(y)) = E(y^+)$ and $var(y_i^+) = var(T_1(y)) = \mu^2/\alpha)$. This residuals is the same as the ordinary standardized residuals, but is obtained from the properties of the two-parameter exponential family of distributions, as in Lehmann & Casella (1998).

5. Applications

In this section we present two applications: the first from simulated data and the second using data on the duration of the embryonic stage of fruit flies reported by Powsner (1935) and McCullagh & Nelder (1989).

5.1. Simulation data set

In this simulation, 500 values of three covariates were simulated from uniform distributions. Values of the covariates X_2, X_3 and X_4 were generated from uniform distributions $U(0, 30)$, $U(0, 15)$ and $U(10, 20)$, respectively. Values of the covariate X_1 are assumed to be a vector of ones, in order to define mean and shape models with intercept. Values of the response variables, Y, were generated from a gamma distribution with mean and shape parameters given by:

$$\begin{aligned} \hat{\mu}_i &= 15 + 2x_{2i} + 3x_{3i} \qquad (26) \\ \hat{\alpha}_i &= \exp(0.2 + 0.1x_{2i} + 0.3x_{4i}) \qquad (27) \end{aligned}$$

The fitted mean equation and the fitted shape equations, obtained by applying a Bayesian method proposed by Cepeda-Cuervo (2001), are:

$$\begin{aligned} \hat{\mu}_i &= 15.015 + 2.001x_{2i} + 2.998x_{3i} \qquad (28) \\ \hat{\alpha}_i &= \exp(0.360 + 0.104x_{2i} + 0.290x_{4i}) \qquad (29) \end{aligned}$$

We consider residual checks for systematic departure from the model using some informal graphs. From Figure 1, in the second panel, both residuals $r+$ and $r*$ are plotted against the varying mean of the model, $\widehat{\mu}_i$. Typical systematic deviations are absent due to the fact there is neither curvature in the mean nor a systematic change. According to the third panel, where the residuals are plotted against the linear predictor X_3, we conclude there is no appearance of a systematic trend.

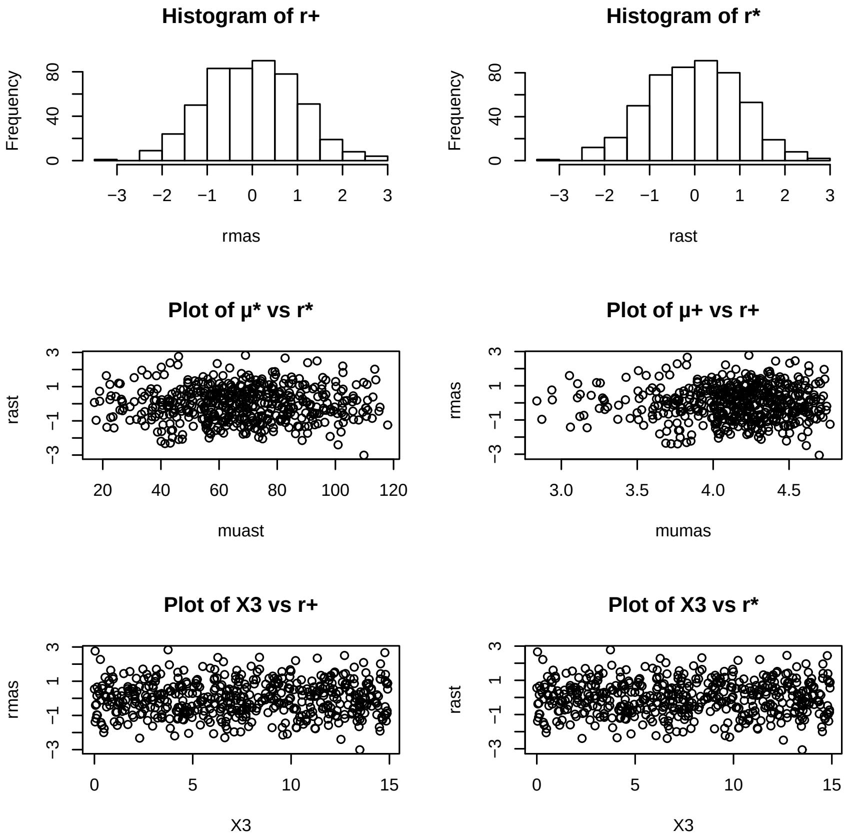

FIGURE 1: Residuals r+ and r*

The normal probability plot in Figure 2 (Q-Q plot for $r+$ and $r*$) suggests a good fit of both residuals $r+$ and $r*$ to the normal distribution. As expected, the analysis of the residual under study did not single out any observation as atypical yield evidence of lack of fit.

Finally, a third plot is the partial residual plot for gamma regression model, which is used to assess the form of a predictor and is thus calculated for each predictor. If the scale is satisfactory, the plot should be approximately linear. If not, its form suggests a suitable alternative. According to Figure 3, the X_2 variable should have curvature and X_3 should be linear.

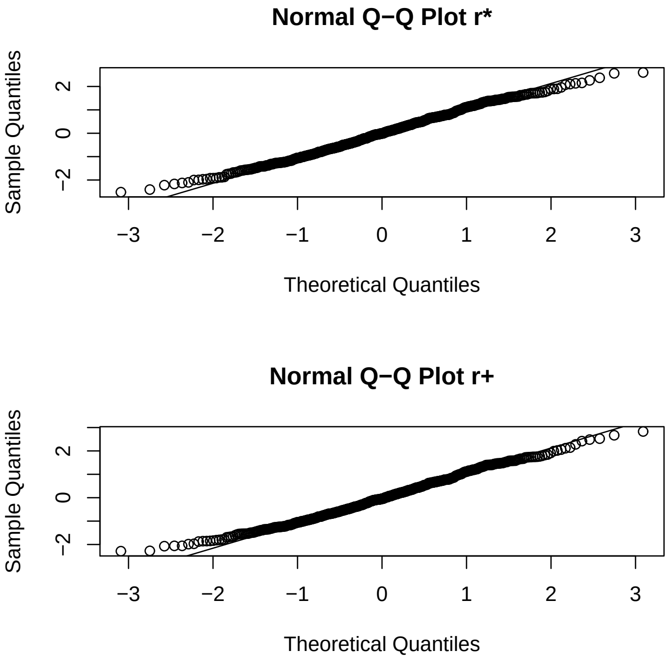

Figure 2: Normal Residuals for r+ and r*

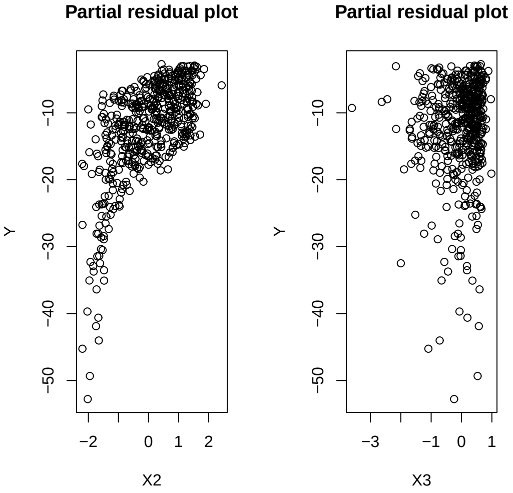

Figure 3: Partial Residuals

5.2. Duration of the embryonic stage of fruit flies

This application is based in an example presented by McCullagh & Nelder (1989). They used a data set collected by (Powsner 1935), to measure the effect of temperature on the duration of the development stages of the fruit fly (Drosophila melanogaster). In his experiment, there are four states: the embryonic, egg-larval, larval and pupal. Only the first is considered here. In this model, observed duration is the response variable, weighted due to batch size.

According to McCullagh & Nelder (1989), the systematic part of the model is considered by rational functions of temperature:

$$\beta_0 + \beta_1 T + \beta_2/(T - \delta) \tag{30}$$

where δ represents an asymptote for the temperature function. The fit of the model takes into account the gamma regression and the identity link was preferred over the inverse and log links respectively. They adjusted this model considering that the coefficient of variation is constant.

The residuals summarized in this article were calculated assuming the model

$$\begin{aligned} \mu_i &= \beta_0 + \beta_1 T_i + \beta_2/T_i & (31)\\ \alpha_i &= \exp(\gamma_0 + \gamma_1 T_i + \gamma_2/T_i), & (32) \end{aligned}$$

for the fruit fly application, and the following parameter estimates (and standard deviations) were observed: $\hat{\beta}_0 = -2.2828(1,4485)$, $\hat{\beta}_1 = 0.04068(0,0298)$, $\hat{\beta}_2 = 36.7313(17,3253)$, $\hat{\gamma}_0 = 3.3718(2,9484)$, $\hat{\gamma}_1 = -0.0529(0,0671)$ and $\hat{\gamma}_2 = -15.8543(31,0588)$

In figure (4), it can the seen that due to the small number of observations, in some panels (such as the fourth one), the residuals (r^*) appear to have linear dependence on μ^*, which means that for this case, r^+ is more dependable than r^*, in order to get a better residual. Regarding the histograms of r^+ and r^*, we can observe that it is not as accurate as the previous application, which was expected since the first data set was generated by a gamma simulation. However in this case the residuals show greater accumulation around zero, but the distribution does not look symmetric like the normal distribution, possibly because of the relatively small number of observations.

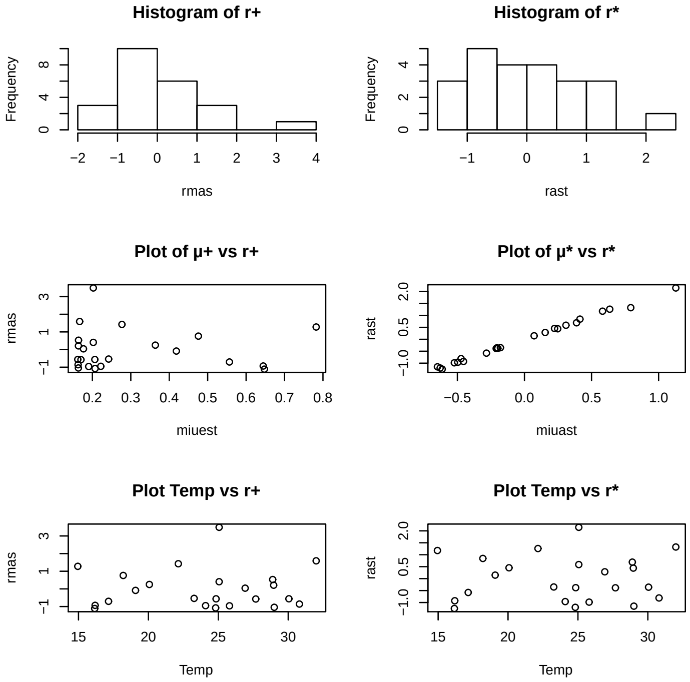

Figure 4: Residuals r+ and r*

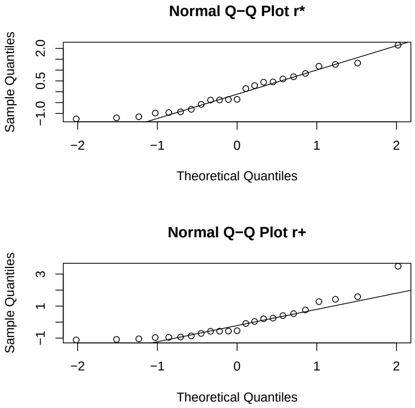

Figure 5: Normality Residuals for r+ and r*

According to the QQ plot, Figure (5), the distribution of both residuals r+ and r* are close to the normal distribution. There is no pattern when we plot the residuals against the covariates.

Finally, plotted Figure (6), where we summarize other residuals calculated from the fruit fly data. There are three. The first and second show the estimated μ against the absolute value of each residual (r^+ and r^*). The second new residual considered was the Pearson residual, which has an irregular and scattered behavior, a desirable property in residuals. The last ones calculated are the deviance residuals for both r^+ and r^*, which are shown in panels 5 and 6 in Figure (6).

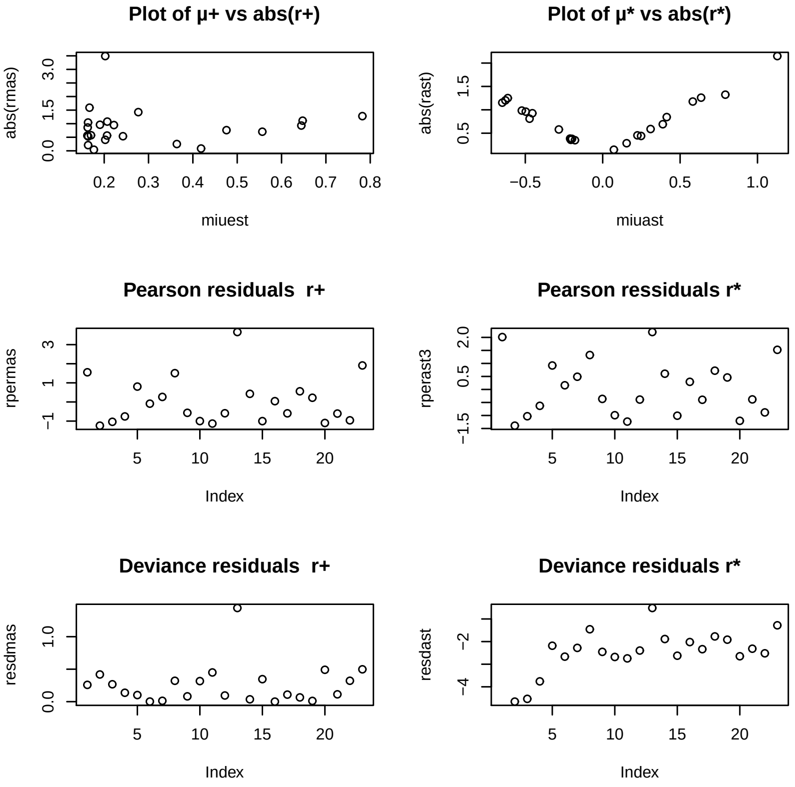

Figure 6: Different Residuals for r+ and r*

6. Conclusions

In this paper, we propose two new residuals for the gamma regression models, for which many link functions can be used. We choose the identity and log link for this evaluation. The new residuals are computed by the difference of the link function responses and their fitted means respectively using Fisher scoring and Bayesian estimation of the parameters. The results suggest that the residuals that we call $r+$ are the same as commonly used ordinary residuals. On other hand, the new residuals $r*$, which come from Fisher scoring iterative algorithm are also approximated by the standard normal distribution and fulfill informal checks for systematic departure from the model. This fact ccan be used to construct more reliable goodness of fit measures and measures of explained variation for gamma regression models.

References

Bateson, T. F. (2009), 'Gamma regression of interevent waiting times versus poisson regression of daily event counts: Inside the epidemiologist's toolboxŨselecting the best modeling tools for the job', *Epidemiology* **20**(2), 202–204.

Cepeda-Cuervo, E. (2001), 'Modelagem de variabilidade em modelos lineares generalizados', *Unpublished Ph.D.thesis, Mathematics Institute, Universidade Federal Rio de Janeiro* .

Cepeda, E. & Gammerman, D. (2005), 'Bayesian methodology for modeling parameters in the two parameters exponential family', *ESTADÍSTICA* **57**(168), 93–105.

Chib, S. & Greenberg, E. (1995), 'Understanding the metropolis-hastings algorithm', *The American Statistician* **49**(4), 327–335.

Cox, P. & Snell, E. (1968), 'A general definition of residuals', *Journal of the Royal Statistical Society, Vol.30* pp. 248–275.

Dobson, A. J. (2010), *An introduction to generalized linear models*, CRC press.

Gamerman, D. & Lopes, H. F. (2006), *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*, CRC Press, address=New York,.

Gammerman, D. (1997a), *Markov Chain Monte Carlo: Sthocastic Simulation for Bayesian Inference*, Chapman and Hall, London.

Heidelberger, P. & Welch, P. D. (1981), 'A spectral method for confidence interval generation and run length control in simulations', *Comm. ACM.* **24**, 233Ũ–245.

Krishnamoorthy, K. (2006), *Handbook of Statistical Distributions with Applications*, Chapman & Hall/CRC, Florida.

Lehmann, E. L. & Casella, G. (1998), *Theory of point estimation*, Springer, New York.

McCullagh, J. & Nelder, J. (1989), *Generalized Linear Models. Second Edition*, Chapman and Hall, London.

Pierce, D. A. & Schafer, D. W. (1986), 'Residuals in generalized linear models', *Journal of the American Statistical Association, Vol.81* **81**(396), 977–986.

Powsner, L. (1935), 'The effects of temperature on the durations of the developmental stages of drosophila melanogaster', *Physiological Zoology* **8**(4), 474–520.

Winklemann, R. (2008), *Econometric analysis of count data*, Springer-Verlag, Berlin, Germany.