Overdispersed Nonlinear Regression Models

María Victoria Cifuentes-Amado ⓘ   **Edilberto Cepeda-Cuervo** ⓘ

Universidad Nacional de Colombia   Universidad Nacional de Colombia

Bogotá   Bogotá

Abstract

In this paper we propose nonlinear regression models in the biparametric family of distributions. In this class of models we propose two new classes of overdispersed nonlinear regression models: the first, defined from the overdispersion family of distributions proposed by Dey, Gelfand, and Peng (1994), and the second from a class of compound distributions. For these models, we develop a Bayesian method in which samples of the posterior distributions are obtained by applying an iterated Metropolis-Hastings algorithm obtained by assuming two groups of parameters, defined by the mean and dispersion regression structures. In the first subclass of models, to improve the performance of the iterated Metropolis-Hastings algorithm, we develop worked variables from the application of Fisher scoring algorithm, to build the kernel transition function. A Bayesian method to fit compound regression models is also proposed. Finally, we present an application to neonatal mortality dataset to illustrate the use of the proposed models and the performance of the Bayesian method to fit the proposed models.

Keywords: nonlinear regression models, overdispersed nonlinear regression, compound regression models, Bayesian methods.

1. Introduction

The binomial and Poisson distributions are usually used to model count datasets. However, count data frequently present overdispersion, which means that the sample variance exceeds the nominal variance of the model and if an appropriate model is not considered, the analysis can present problems in the inference, such as underestimation of the errors and variance, loss of efficiency in the estimations and incorrect inferences about the confidence intervals of the regression parameters (Williams 1982; Cox 1983; Collet 1991). Multiple models have been developed to analyze overdispersed datasets, and have been categorized in two classes (Hinde and Demétrio 1998): models in which a new independent parameter is added to model the variance and/or overdispersion and those in which the parameter of the discrete distribution is assumed to follow an appropriate distribution function. The first class contains models developed by Williams (1982) for binomial count data and by Breslow (1984) for Poisson count data. To this class of models also belong the doubly exponential family of distributions presented by Efron (1986), the exponential dispersion models proposed by Jorgensen (1987) and overdispersed generalized linear models presented by Dey *et al.* (1994). The second class

includes the overdispersion models developed by Crowder (1978), who proposed regression analysis of proportions based on the beta-binomial distribution, and by Margolin, Kaplan, and Zeiger (1981), who formulated the negative binomial distribution for data analysis of the Ames test used for the detection of Salmonella.

The overdispersed models proposed by Dey *et al.* (1994) are linear in the sense that their systematic components are linear functions of the regression parameters. However, frequently a mean linear regression structure is not appropriate to describe the dataset's behavior, when the mean of the variable of interest follows a nonlinear regression structure. Thus, in this paper we propose overdispersed nonlinear regression, model in which the systematic component is given by:

$$\eta_i = S(\boldsymbol{x}_i, \boldsymbol{\beta}), \quad i = 1, ..., n, \tag{1}$$

where $\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})^T$ is the i-th vector of observations of p covariates and $S(\boldsymbol{x}_i, \cdot)$ is a known nonlinear function of $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, a vector of unknown regression parameters.

In the case of count data, Frome (1983) analyzed lung cancer mortality data among British physicians who were regular cigarette smokers, by using a nonlinear Poisson regression model derived from the multistage theory of carcinogenesis, relating the mortality rate with the time of exposure (years of smoking) and with the amount of carcinogen applied per unit time (cigarettes per day). McCullagh and Nelder (1989), in chapter 11, presented an experiment in which the number of grasshoppers that die is observed, from an initial exposed number (a binomial variable of response) to different combinations of doses of an insecticide and a synergist (a substance that potentiates the insecticide) and analyzed some nonlinear functions to model the proportion of dead grasshoppers. Another study, reported in Donoso, Carvajal, Vera, and Poblete (2014), involved the relationship between maternal age and fetal, neonatal, infant and maternal mortality rates in Chile. They found that the lowest mortality rates for both newborns (under 28 days) and infants occurred when the mothers were between 25 and 30 years old, an intermediate age, generating a nonlinear relationship between the rate and maternal age. Other examples are given in Frome, Kutner, and Beauchamp (1973) and Frome and Dufrain (1986), where the effects of radiation on the survival of stem cells or the rate of neutron decay in a Beryllium sample were analyzed using nonlinear regression functions.

In this paper, we propose nonlinear (linear) regression models in the biparametric family of distributions (F^2). In this class of models we propose two new class of overdispersed nonlinear regression models. The first is defined from the family of overdispersed distributions proposed by Dey *et al.* (1994) and the second from a class of compound distributions obtained by assuming that the parameter of a one-parameter distribution follows an appropriate distribution belonging to F^2, in both by assuming that the mean systematic component follows a nonlinear (linear) regression structure.

For these models we develop a Bayesian method in wish samples of the posterior distributions, denoted by $\pi(.)$, are obtained by applying an iterated Metropolis-Hastings (M-H) algorithm obtained by assuming two groups of parameters, defined by the mean and dispersion regression structures. To build the kernel transition functions used to propose samples of $\pi(.)$ in the M-H algorithm, we follow the Bayesian methods proposed in Cepeda-Cuervo (2001) and Cepeda and Gamerman (2005), developing worked variables from the first order Taylor approximation of the nonlinear mean regression models or from the application of the Fisher scoring algorithm. In the first subclass of models, defined from the overdispersed family of distributions of Dey *et al.* (1994), to improve the performance of the iterated M-H algorithm, we develop worked variables from the Fisher scoring algorithm, used to obtain maximum likelihood estimation of the parameters, and use them to build the kernel transition function. Finally, we propose a Bayesian method to fit compound models. In these models, we also propose a M-H iterated algorithm, where the mean kernel transition function is obtained from the combination of the prior distribution and an observational model proposed from a worked variable obtained from the first-order Taylor approximation of the mean nonlinear regression

structure. To define the kernel transition function for the dispersion regression parameters, we propose to use the worked observation variables obtained from the Fisher scoring algorithm of overdispersed models, but assuming a sample of the distribution used to define the compound models, at the current values of the mean and dispersion regression parameters.

The performance of the Bayesian method extended in this paper to fit nonlinear double generalized regression models has been demostrated in the framework of generalized linear models (Cepeda-Cuervo 2001), heteroscedastic Weibull-normal mixture models (Cepeda-Cuervo, Migon, Garrido, and Achcar 2014), heteroscedastic normal-exponential mixture models (Garrido, Cepeda-Cuervo, and Achcar 2011) and heteroscedastic nonlinear regression models (Cepeda-Cuervo and Achcar 2010). Finally, we report some results of an application of the nonlinear negative binomial models in the analysis of the neonatal mortality dataset.

After this introduction, this paper is structured as follows: In Section 2, the nonlinear biparametric regression models are defined, including the classes of overdispersed nonlinear regression models, defined from the Dey *et al.* (1994), and the compound family of nonlinear regression models. In Section 3, a Bayesian method to fit nonlinear regression models in F^2 is defined. The working variables obtained from the application of the Fisher scoring algorithm to fit overdispersed nonlinear regression models are used to define the Bayesian method to fit the proposed models. In Section 4, from the previous maximum likelihood and Bayesian developments, a Bayesian method to fit nonlinear compound regression models is proposed, by introduction of latent parameters. Section 5 presents an application to neonatal mortality dataset. Section 7 is an appendix that include a summary of the main results presented in Cepeda and Gamerman (2005), related to double generalized linear models, and the maximum likelihood and Bayesian methods used to obtain parameter inferences.

2. Nonlinear biparametric regression models

In the framework of the biparametric family of distributions, Gelfand and Dalal (1990) discussed the biparametric exponential family of distributions given by:

$$f(y|\delta, \tau) = b(y)e^{[\delta y + \tau T(y) - \rho(\delta, \tau)]}, \tag{2}$$

where, if Y is a continuous random variable, f is a density function with respect to the Lebesgue measure while if Y is discrete random variable, f is a count measure. The components of $(\delta, \tau) \in \mathcal{S}$ are the parameters of the distribution family (2), where $\mathcal{S} \subseteq \mathcal{R} \times \mathcal{R}$ denotes their parametric space. To obtain orthogonality between the parameters of the distribution, Dey *et al.* (1994) proposed a new parameterization of (2) in terms of the mean of Y, $\mu := E(y|\delta, \tau) = \frac{\partial}{\partial \delta}\rho(\delta, \tau)$, as in (3), where $\Psi^{(i,j)} := \frac{\partial^{i+j}}{\partial \mu^i \partial \tau^j}\Psi$, $\Psi^{(1,0)}(\mu, \tau) = \delta$ and $\Psi(\mu, \tau) = \mu\Psi^{(1,0)} - \rho(\delta, \tau)$.

$$f(y|\mu, \tau) = b(y)e^{[(y-\mu)\Psi^{(1,0)}(\mu,\tau) + \tau T(y) + \Psi(\mu,\tau)]}. \tag{3}$$

Some examples of distributions belonging to the biparametric exponential family of distributions (3), denoted by F_e^2, are the normal $N(\mu, \sigma^2)$, gamma $G(\mu, \alpha)$, lognormal $LN(\mu, \sigma^2)$ and inverse Gaussian $IG(\mu, \lambda)$, among others. The beta distribution is an example of a distribution that does not belong to the exponential family of distribution (2), but the density function of the random variable obtained by the logit transformation $\tilde{Y} = \text{logit}(Y)$ belongs to this family of distributions. The overdispersed exponential family of distributions proposed by Jorgensen (1987) can be seen as part of the double exponential family of distributions proposed by Efron (1986), which is part of the biparametric exponential family proposed by Dey *et al.* (1994). When the dispersion parameter τ is known in (2), a one-parameter exponential family of distributions is obtained, which include the exponential $Exp(\lambda)$, Poisson $\mathcal{P}(\lambda)$, binomial $Bin(m, p)$ and, in general, all the power series distributions introduced by Noack (1950). See McCullagh and Nelder (1989).

Other examples of the one-parameter exponential family of distributions is Lindley's continuous distribution $LD(\theta)$, where $\theta > 0$. This distribution was introduced by Lindley (1958) to describe the lifetime of a process and its density function can be expressed as the equation (2) by calculating $b(y) = y + 1$, $\delta = -\theta$ and $\rho(\delta, 0) = -log(\delta^2/(-\theta + 1))$.

From the one- and the two-parameter exponential family of distributions, we define a compound family of distributions by the combination of a distribution $F_1(\theta)$, belonging to the one-parameter exponential family of distributions, and an appropriate distribution $F_e^2(\mu, \tau)$, belonging to the two-parameter exponential family of distribution (3), by assuming that $\theta \sim F_e^2(\mu, \tau)$. Thus, if a random variable Y follows a distribution $F_1(\theta)$ belonging to the one parameter exponential family of distributions, where θ follows a distribution $F_2(\mu, \tau)$ belonging to exponential family of distributions (3), where $\mu := E(\theta|\mu, \tau) > 0$ and $\tau \in \mathcal{R}$ is a dispersion parameter, the compound distribution of Y, denoted $Y \sim F(\mu, \tau)$, is given by:

$$f_Y(\mu, \tau)(y) = \int_{\mathcal{N}} f_{1(\theta)}(y) f_2(\theta|\mu, \tau) d\theta, \tag{4}$$

where $f_{1(\theta)}(y)$ is the density (probability) function $F_1(\theta)$ and $f_2(\mu, \tau)(y)$ is the density of the distribution function $F_2(\mu_i, \tau_i)$. Henceforth, we use $F(\mu, \tau)$ to denote the compound distribution functions, with density (probability) function given by (4). The compound distributions are also two-parameter distributions.

An important subfamily of the compound model is the compound distribution function, where F_1 is a discrete distribution. Some usual examples of compound discrete distributions are the Poisson normal distribution (Cameron and Trivedi 1998), the (μ, α)-negative binomial distribution and the (μ, ϕ)-beta-binomial distributions (Cepeda-Cuervo and Cifuentes-Amado 2017). The first is obtained from the combination of $Y|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda \sim Lognormal(\mu, \tau)$; the second is a composition between the Poisson and a gamma distributions, belonging to the exponential family (3), and the third is a composition of the binomial and beta distributions. The binomial normal distribution, introduced by Williams (1982), is a compound distribution, where Y follows a binomial distribution, $Y|\pi \sim Bin(m, \pi)$, and $logit(\pi_i)$ follows a normal distribution $logit(\pi) \sim N(\mu, \phi)$. Other examples are: the Poisson Lindley distribution (Munuswamy 1970), where $Y|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda \sim LD(\theta)$); the Poisson Gaussian-Inverse distribution , where $Y|\nu \sim \mathcal{P}(\nu)$ and $\nu \sim IG(\mu, \lambda)$), and the Poisson-Tweedie distribution, where $Y \sim \mathcal{P}(\lambda)$ and $\lambda \sim TW_p(\mu, \phi)$, given that Tweedie distributions belong to the exponential dispersion model family of distributions (Hougaard, Lee, and Whitmore 1997).

Let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ and $\boldsymbol{z}_i = (z_{i1}, ..., z_{ir})^T$, $i = 1, 2, \ldots, n$, be vectors of mean and dispersion explanatory variables, respectively, and let $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_r)^T$, be the vectors of the mean and dispersion regression structures. The biparametric nonlinear (linear) regression models are defined by the following three components:

- The *random component* $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, such that $Y_i \sim F^2(\mu_i, \tau_i)$, $i = 1...n$, are independent random variables, belonging to the F^2 family of distributions.

- The *systematic component* $\boldsymbol{\eta_i} = (\eta_{1i}, \eta_{2i})^T$, such that $\eta_{1i} = S(\boldsymbol{x}_i, \boldsymbol{\beta})$ and $\eta_{2i} = U(\boldsymbol{z}_i, \boldsymbol{\gamma})$, where $S(.,.)$ is a nonlinear function of $\boldsymbol{\beta}$ and $U(.,.)$ can be a linear or nonlinear function of $\boldsymbol{\gamma}$.

- The *link functions* $h(.)$ and $g(.)$, such that $h(\mu_i) := \eta_{1i}$ and $g(\tau_i) = \eta_{2i}$.

The link functions must be strictly monotonous and doubly differentiable to obtain maximum likelihood parameter estimations, and once differentiable in a Bayesian approach. Some usual link functions are the logit, logarithmic and identity functions, depending on the parameter space restrictions of the distribution. The doble generalized linear regression models are defined by the same tree components, assuming that in the *systematic component* $\eta_{1i} = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\eta_{2i} = \boldsymbol{z}_i^T \boldsymbol{\gamma}$ (Cepeda-Cuervo 2001, Cepeda and Gamerman 2005). Some examples of nonlinear biparametric regression models are:

1. *Poisson normal regression model.* Let $Y_i \sim PN(\mu_i, \tau_i)$, $i = 1, ..., n$, be n independent overdispersed random variables that follows Poisson normal distribution. A nonlinear Poisson normal regression model can be defined by assuming that the systematic components are given by: $\eta_{1i} = S(\boldsymbol{x}_i, \boldsymbol{\beta}) + \tau_i/2$ and $\eta_{2i} := \boldsymbol{z}_i^T \boldsymbol{\gamma}$, where $S(\boldsymbol{x}_i, \boldsymbol{\beta})$ is a nonlinear function of $\boldsymbol{\beta}$, for example, as it proposed in Fokianos (2012).

2. *Beta-binomial regression model.* Let $Y_i \sim BB(\mu_i, \tau_i)$, $i = 1, ..., n$, be n independent overdispersed random variables that follow a beta-binomial distribution. A nonlinear beta-binomial regression model can be defined by assuming that the systematic components are given by: $\eta_{1i} := \beta_0 + \beta_1 log(x_{1i} - \beta_2) + \beta_3 x_{2i}/(\beta_4 + x_{2i})$ and $\eta_{2i} := \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}$. The nonlinear regression structure of the systematic component η_{1i} is proposed by McCullagh and Nelder (1989) to analyse the proportion of insects killed by exposure to the mixture of an insecticide and a synergist. Finally, given that $0 < \mu_i < 1$ and $\phi_i > 0$, possible link functions are $logit(\mu_i) = \eta_{1i}$ and $log(\phi_i) = \eta_{2i}$.

3. Bayesian biparametric nonlinear regression models

In this section we extend the Bayesian method proposed by Gamerman (1997) and Cepeda-Cuervo (2001) to obtain samples from the posterior distribution of the linear regression parameters, by applying the Metropolis-Hastings algorithm. This method was proposed by Gamerman (1997), in the framework of generalized linear mixed models, and extended by Cepeda and Gamerman (2000), Cepeda-Cuervo (2001) and Cepeda and Gamerman (2005) to obtain samples from the posterior distribution in the framework of double generalized linear regression models defined in the F_e^2 family of distributions, when functions of the mean and dispersion (variance) parameters follow linear regression structures. Usual examples of the biparametric regression models where both parameters follow linear regression structures are the joint mean and dispersion beta regression models and the joint mean and shape gamma regression models, proposed in Cepeda-Cuervo (2001) and fitting by applying Bayesian methods. As in these works, to define Bayesian nonlinear overdispersed models, we assign normal prior distributions for the regression parameters, and without loss of generality, we assume an independent normal prior distribution $p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta})p(\boldsymbol{\gamma})$ for the regression parameters, given by:

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} \sim N \left[\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{g} \end{pmatrix}, \begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G} \end{pmatrix} \right] \tag{5}$$

Thus, if $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the likelihood function defined from a distribution belonging to exponential family (3) or (4), the posterior distributions is given by $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) \propto L(\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}, \boldsymbol{\gamma})$. However, given that $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is analytically intractable, we propose to obtain samples of the posterior distribution from their posterior conditional distributions $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ and $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta})$, which are also analytically intractable, except for $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ in the case of normal linear regression models (Cepeda and Gamerman, 2001). Thus, we propose to obtain samples from the posterior conditional distribution by applying the M-H algorithm, for which we propose to build kernel transition functions as in Gamerman (1997), Cepeda-Cuervo (2001) or Cepeda and Gamerman (2005), to propose samples of the posterior distributions. These kernels are obtained as a combination of working observation model, defined by assuming that appropriate working variables follow a normal distribution, and the normal prior distributions of the regression parameters. The working variables are obtained by a linearized form of the mean links and nonlinear regression structures. These variables can be obtained from the Fisher scoring algorithm, as in McCullagh and Nelder (1989), for generalized linear models, or in Cepeda and Gamerman (2000) and in Cepeda and Gamerman (2005), for double generalized linear models. The first-order Taylor approximation of the link functions are used to define the kernel transition functions used in the definition of the M-H algorithm applied to fit the joint

mean and dispersion beta regression models proposed in Cepeda-Cuervo (2001). Examples of kernels transition functions in the framework of DGLMs are summarized in Appendix 7.1.

A general approximation to obtain worked variables from which the worked observational models are used to build the kernel transition functions is the Fisher scoring algorithm. In the biparametric nonlinear regression models, if the distribution belongs to the biparametric exponential family (3), the mean worked regression variable is given by (28), as obtained in Section 7.2 by applying the Fisher scoring algorithm. Thus, the worked observational model is obtained by assuming that the worked observational variable

$$\tilde{Y}_i \approx \nabla S(\boldsymbol{x}_i, \boldsymbol{\beta}^{(c)})\boldsymbol{\beta}^{(c)} + h'\left(\mu_i^{(c)}\right)\left(Y_i - \mu_i^{(c)}\right), \tag{6}$$

follows a normal distribution, where $\nabla S(\boldsymbol{x}_i, \boldsymbol{\beta}^{(c)})$ denote the gradient of $S(\boldsymbol{x}_i, .)$, estimate at the current value of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(c)}$, and $h'(.)$ the first order derivative of $h(.)$. The mean and variance of these random variables are given by: $E(\tilde{Y}_{1i}) = \nabla S(\boldsymbol{x}_i, \boldsymbol{\beta}^{(c)})\boldsymbol{\beta}^{(c)}$ and $Var(\tilde{Y}_{1i}) = \left[h'(\mu_i^{(c)})\right]^2 Var(Y_i)$. Thus, the kernel transition function q_1 to obtain samples of $\boldsymbol{\beta}$ is the posterior distribution function obtained by the combination of the conditional prior normal distribution and the worked observational model obtained by assuming that the worked observational variable follows the normal distribution, $\tilde{Y}_i \sim N(\nabla S(\boldsymbol{x}_i, \boldsymbol{\beta}^{(c)})\boldsymbol{\beta}^{(c)}, Var^{(c)}(\tilde{y}_{1i}))$. So, the kernel transition function $q_1(.)$ is given by:

$$q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}) = N(\mathbf{b}^*, \mathbf{B}^*), \tag{7}$$

where $\mathbf{b}^* = \mathbf{B}^*(\mathbf{B}^{-1}\mathbf{b} + \tilde{\mathbf{X}}^T\boldsymbol{\Sigma}_1^{-1}\tilde{\mathbf{Y}})$ and $\mathbf{B}^* = (\mathbf{B}^{-1} + \tilde{\mathbf{X}}^T\boldsymbol{\Sigma}_1^{-1}\tilde{\mathbf{X}})^{-1}$; with $\nabla S(\boldsymbol{x}_i, \boldsymbol{\beta}^{(c)})$ the i-th row of $\tilde{\mathbf{X}}$, $\boldsymbol{\Sigma} = Diag\left(Var^{(c)}\left(\tilde{Y}_i\right)\right)$ and $\tilde{\mathbf{Y}}$ the vector of observed values of (6).

If $U(.,.)$, in the definition of systematic component is a linear function of $\boldsymbol{\gamma}$, $g(\tau_i) = \mathbf{z}_i^T\boldsymbol{\gamma}$, the worked observational variable to define the kernel transition function to propose samples of $\boldsymbol{\gamma}$, can be obtained by applying the Fisher scoring algorithm, as in Section 7, or by first-order Taylor approximation of g(.), where this worked variable is given by:

$$\check{Y}_i \simeq g(\tau_i) + g'(\tau_i)(t_i - \tau_i), \quad i = 1, \dots, n,$$

where $g'(.)$ denote the first-order derivative of $g(.)$ and t_i is a random variable such that $E(t_i) = \tau_i$. For this variable, $E(\check{Y}_i) = \mathbf{z}_i^T\boldsymbol{\gamma}$ and $Var(\check{Y}_i) = [g'(\tau_i)]^2 Var(t_i)$. Thus, if $\boldsymbol{\gamma}^{(c)}$ denotes the current values of $\boldsymbol{\gamma}$, the worked dispersion variables are given by:

$$\check{Y}_i = \mathbf{z}_i^T\boldsymbol{\gamma}^{(c)} + g'\left(g^{-1}\left(\mathbf{z}_i^T\boldsymbol{\gamma}^{(c)}\right)\right)\left[t_i - h^{-1}\left(\mathbf{z}_i^T\boldsymbol{\gamma}^{(c)}\right)\right], \tag{8}$$

and thus the kernel transition function $q_2(.)$ to obtain samples of $\boldsymbol{\gamma}$ is obtained from the combination of the conditional normal prior distribution, $\boldsymbol{\gamma}|\boldsymbol{\beta} \sim N(g, G)$, and the worked observational model obtained from (8) by assuming that it follows the normal distribution, $\check{y}_i \sim N(\mathbf{z}_i^T\boldsymbol{\gamma}^{(c)}, Var^{(c)}(\check{y}_i))$, by applying Bayes' Theorem. Thus, $q_2(.)$ are given by:

$$q_2(\boldsymbol{\gamma}|\boldsymbol{\beta}) = N(\mathbf{g}^*, \mathbf{G}^*), \tag{9}$$

where $\mathbf{g}^* = \boldsymbol{G}^*(\mathbf{G}^{-1}\mathbf{g} + \mathbf{Z}^T\boldsymbol{\Sigma}_2^{-1}\check{\mathbf{Y}})$ and $\mathbf{G}^* = (\mathbf{G}^{-1} + \mathbf{Z}^T\boldsymbol{\Sigma}_2^{-1}\mathbf{Z})^{-1}$, with the variance covariance matrix $\boldsymbol{\Sigma}_2 = \text{diag}(Var^{(c)}(\check{Y}_i))$, $\boldsymbol{Z}$ the matrix values of the dispersion explanatory variables and $\check{\boldsymbol{Y}}$ a vector of observed worked values. The kernel transition function (9) also can be obtained from a combination of a normal prior distribution of γ and the worked observational model given by (29), obtained from the Fisher scoring algorithm.

Thus, with these kernel transition functions, the following Bayesian iterated algorithm is proposed to obtain samples of the posterior parameter distribution of nonlinear regression models.

1. Begin the chain iteration counter at j=1.

2. Set initial chain values $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})^T$ for $(\boldsymbol{\beta}, \boldsymbol{\gamma})^T$.

3. Propose a new value φ for $\boldsymbol{\beta}$, generated from (7).

4. Calculate the acceptance probability, $\alpha(\boldsymbol{\beta}^{(j-1)}, \varphi)$. If the movement is accepted, then $\boldsymbol{\beta}^{(j)} = \varphi$. If it is not accepted, then $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$.

5. Propose a new value φ for $\boldsymbol{\gamma}$, generated from (9).

6. Calculate the acceptance probability, $\alpha(\boldsymbol{\gamma}^{(j-1)}, \varphi)$. If the movement is accepted, then $\boldsymbol{\gamma}^{(j)} = \varphi$. If it is 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.

To fit the orthogonal nonlinear overdispersed regression models defined from the biparametric exponential family of distributions (3), we can apply the Fisher scoring algorithm to obtain the mean and dispersion worked observational variables (28) and (29), used to built the kernel transition functions (7) and (9), respectively. Thus, samples of the posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be obtained by applying the same Bayesian Metropolis-Hastings algorithm algorithm defined above (3).

4. Bayesian method to fit compound regression models

In this section we propose a Bayesian method to fit compound nonlinear (linear) regression models, defined as in the biparametric nonlinear (linear) regression models, but in the compound distribution defined by (4). The idea is the same as presented in Section 3, where samples of the mean and dispersion regression parameters are proposed from the kernel transition functions (7) and (9). However, although the respective worked variables can be obtained from first-order Taylor approximation, here the mean worked variable is given by (28), the same obtained by the first-order Taylor approximation, but the worked variable for the dispersion regression parameters is obtained from simulated values generated from a latent variable. Thus, to obtain a worked variable associated with $\boldsymbol{\gamma}$, simulated values of a random variable θ must be obtained, for which we propose the following procedure: given that $\theta_i|\mu_i, \tau_i \sim F_e^2(\mu_i, \tau_i)$, in each iteration, a sample $\theta_i^{(k)}$ of θ_i is obtained from $F_e^2(\mu_i^{(c)}, \tau_i^{(c)})$, and it is assumed to be a sample of the response variable to obtain values of the worked variable (29), where y_i is now replaced by θ_i.

With the introduction of this new worked variable, the respective kernel transition functions are constructed as in Section 3 and the iterated Bayesian algorithm presented in that section is reformulated as follows:

1. Begin the chain iteration counter at j=1.

2. Set initial chain values $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})^T$ for $(\boldsymbol{\beta}, \boldsymbol{\gamma})^T$.

3. Propose a new value φ for $\boldsymbol{\beta}$, generated from (7).

4. Calculate the acceptance probability, $\alpha(\boldsymbol{\beta}^{(j-1)}, \varphi)$. If the movement is accepted, then $\boldsymbol{\beta}^{(j)} = \varphi$. If it is not accepted, then $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$.

5. Obtain a sample of $\theta_i^{(j)}$ from $F_2(\mu_i^{(j)}, \tau_i^{(j-1)})$, for $i = 1, \ldots, n$.

6. Propose a new value φ for $\boldsymbol{\gamma}$, generated from (9).

7. Calculate the acceptance probability, $\alpha(\boldsymbol{\gamma}^{(j-1)}, \varphi)$. If the movement is accepted, then $\boldsymbol{\gamma}^{(j)} = \varphi$. If it is not accepted, then $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$.

8. Change the counter from j to $j+1$ and return to 2 until convergence is reached.

Example 4.0.1. *(μ, α)-negative binomial nonlinear regression models.* If $y_i \sim NB(\mu_i, \alpha_i), i = 1, \ldots, n$, where $\log(\mu_i) = S(\boldsymbol{x}_i, \boldsymbol{\beta})$ and $\log(\alpha_i) = \boldsymbol{z}_i^T \boldsymbol{\gamma}$, then from (28), the mean random worked variables are given by:

$$\tilde{y}_{1i} = (\nabla S(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}))^T \boldsymbol{\beta}^{(c)} + \frac{y_i}{\mu_i^{(c)}} - 1.$$

where, $S(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}))^T$ denote the gradient of $S(\mathbf{x}_i, .)$ estimated on $\boldsymbol{\beta}^{(c)}$. From the mean and variance of the negative binomial distribution, the mean and variance of $\tilde{y}_{1i}$ are given by $E(\tilde{y}_{1i}) = (\nabla S(\mathbf{x}_i, \boldsymbol{\beta}^{(c)}))^T \boldsymbol{\beta}^{(c)}$ and $Var(\tilde{y}_{1i}) = \frac{\mu_i^{(c)} + \alpha_i^{(c)}}{\mu_i^{(c)} \alpha_i^{(c)}}$, respectively.

For the α regression structures, assuming that θ_i follows gamma distribution $G(\mu_i^{(c)}, \alpha_i^{(c)})$, from (29) the worked variables are given by:

$$
\begin{aligned}
\tilde{y}_{2i} &= \mathbf{z}_i^T \boldsymbol{\gamma}^{(c)} + \frac{(1/\alpha_i)\left[(\theta_i - \mu_i^{(c)})(-1/\mu_i^{(c)}) + \log(\theta_i) + \log(\alpha_i^{(c)}/\mu_i^{(c)}) - \frac{d}{d\alpha_i}\log\Gamma(\alpha_i^{(c)})\right]}{\left[\frac{d^2}{d\alpha_i^2}\log\Gamma(\alpha_i^{(c)}) - \frac{1}{\alpha_i^{(c)}} - \left(\frac{1}{\alpha_i^{(c)}}\right)\left\{E[\log(\theta_i \alpha_i^{(c)}/\mu_i^{(c)})] - \frac{d}{d\alpha_i}\log\Gamma(\alpha_i^{(c)})\right\}\right]} \\
&= \mathbf{z}_i^T \boldsymbol{\gamma}^{(c)} - \frac{1}{\alpha_i^{(c)}}\left[\frac{\theta_i}{\mu_i^{(c)}} - 1 - \log\left(\frac{\theta_i \alpha_i^{(c)}}{\mu_i^{(c)}}\right) + \frac{d}{d\alpha_i}\log\Gamma\left(\alpha_i^{(c)}\right)\right]\left[\frac{d^2}{d\alpha_i^2}\log\Gamma(\alpha_i^{(c)}) - \frac{1}{\alpha_i^{(c)}}\right]^{-1},
\end{aligned}
$$

where, for example, $\frac{d}{d\alpha_i}\log\Gamma\left(\alpha_i^{(c)}\right)$, denote the first-order derivative $\frac{d}{d\alpha_i}\log\Gamma(\alpha_i)$, estimated on $\alpha_i^{(c)}$. The mean and variance of this worked variable are given respectively by:

$$
\begin{aligned}
\mathrm{E}(\tilde{y}_{2i}) &= \mathbf{z}_i^T \boldsymbol{\gamma}^{(c)} \\
\mathrm{Var}(\tilde{y}_{2i}) &= Var\left[\log\left(\frac{\theta_i \alpha_i^{(c)}}{\mu_i^{(c)}}\right) - \frac{\theta_i}{\mu_i^{(c)}}\right]\left\{\alpha_i^{(c)}\left[\frac{d^2}{d\alpha_i^2}\log\Gamma(\alpha_i^{(c)}) - \frac{1}{\alpha_i^{(c)}}\right]\right\}^{-2} \\
&= \left\{\alpha_i^{(c)2}\left[\frac{d^2}{d\alpha_i^2}\log\Gamma\left(\alpha_i^{(c)}\right) - \frac{1}{\alpha_i^{(c)}}\right]\right\}^{-1}
\end{aligned}
$$

The variance of $\tilde{y}_{2i}$ is obtained by applying the properties of the gamma distribution, proven in Cepeda-Cuervo (2001):

$$\mathrm{E}\left[\log\left(\frac{\alpha y}{\mu}\right)\right] = \frac{d}{d\alpha}\log\Gamma(\alpha) \qquad \mathrm{Var}\left[\log\left(\frac{\alpha y}{\mu}\right) - \frac{y}{\mu}\right] = \frac{d^2}{d\alpha^2}\log\Gamma(\alpha) - \frac{1}{\alpha}$$

If a distribution does not belong to the biparametric exponential family distribution (3), such as the beta distribution, the nonlinear regression models also can be defined by assuming a nonlinear regression structure for the mean and a linear (or nonlinear) structure for the dispersion parameter: $h(\mu_i) = S(\mathbf{x}_i, \boldsymbol{\beta})$ and $g(\phi_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$. In this case, the maximum likelihood estimates can be obtained by using the Fisher scoring algorithm and the nonlinear Bayesian regression model can be fitted by application of a iterated Bayesian algorithm, where the kernel transition functions can be developed from worked variables obtained by the first-order Taylor approximation.

5. Applications: Neonatal mortality dataset

The age of a woman when she gives a birth is a factor that influences the risk that the newborn baby will die in the first 28 days of life (neonatal death). To observe the relation between neonatal death and mother's age, data on the annual number of live births (LB) and number

of neonatal deaths (ND) were tabulated between 1999 and 2016 by Donoso *et al.* (2014), from the Vital Statistics Yearbooks of the National Statistics Institute of Chile. In this application, the variable of interest or dependent variable is ND per year, while the explanatory variables are:

- *Maternal age (MA)*, whose values are: (1) if the mother's age is under 15 years old, (2) if the mother's age is between 15 and 19 years, (3) if the mother's age is between 20 and 24 years, (4) if the mother's age is between 25 and 29 years, (5) if the mother's age is between 30 and 34 years, (6) if mother's age is between 35 and 39 years, (7) if mothers age is between 40 and 44 years, and (8) if mother's age is over 45 years.

- Annual number of children born alive, in thousands (LB).

- Year in which observations were made (T). It takes integer values between 1 (year 1999) and 18 (year 2016).

Figure 1 shows the annual neonatal mortality ($1000ND/LB$) for each category of maternal age of 120 observations obtained from 1999 to 2016. From this figure, it is clear that the number of neonatal deaths decreased with the *Maternal Age* from $MA = 1$ to $MA = 4$, after which the neonatal death increased until $MA = 8$, corresponding to maternal age bigger than 45 years. Thus, from the nature of the data and the data behavior, we assume the negative binomial model, where the mean and dispersion regression structure parameters are given by:

$$\mu_i = \beta_1 + \left(\beta_2(MA - \beta_3)^2 + \beta_4\right)LB + \beta_5 T \tag{10}$$

$$\alpha_i = \exp(\gamma_1 + \gamma_2 T), \tag{11}$$

in order to capture the mean change in neonatal death through the categories of the MA, we take into account that the explanatory variable MA is obtained by a partition of the continuous variable "Age of the Mother", assumed to be between 10 and 50 years old, in consecutive equal longitude intervals. The Bayesian nonlinear negative binomial regression model is defined by assuming normal prior distributions for all the parameters, $N(0, 100)$. These nonlinear negative binomial regression models were fitted by applying the Bayesian methods proposed in Section 4, which ensure fast convergence of the chains.

In this model, the mean regression structure is proposed taking into account the behavior of ND as a function of MA, presented in Figure 1, which depends on the LB values, including a scale transformation, plus a linear component in time to detect possible decreasing behavior of mortality. Thus, β_1 is an intercept parameter; β_2 is related with the average rate of change in a parabola with vertex at (β_3, β_4), and β_5 is the parameter related to a linear regression structure in T. Thus, taking into account the behavior of ND as a function of the explanatory variables, we assume a continuous model that can be a good approximation of the mean behavior of ND in the domain of the function, determined by the values of (MA, LB, T).

In this fitting procedure, n=60000 posterior samples were generated, with a burn-in of 10,000 and taking one sample each every 10 to reduce autocorrelation. The posterior parameter estimates, their standard deviations and their 95% confidence intervals are summarized in Table 1, where the Monte Carlo (MC) error is also estimated for each of the parameters. The DIC value for this model was 1013. From these results, we conclude that the neonatal mortality decreased for each category up category 4, corresponding to mother's age between 25 and 29 years, and that increased from the categories 5 to 8, corresponding to mothers between 45 and 49 years old. $\beta_4 = 4.515 > 0$, indicating that ND increased, in each category of MA, when the number of children born alive, LB, increased, being greatest in the MA categories with older ages.

Figure 2 shows the behavior of the standardized Bayesian residuals, indicating good performance of the model fitting the neonatal deaths dataset, showing residuals distributed around zero between -3 and 2, that do not show any trend behavior with the observation numbers.
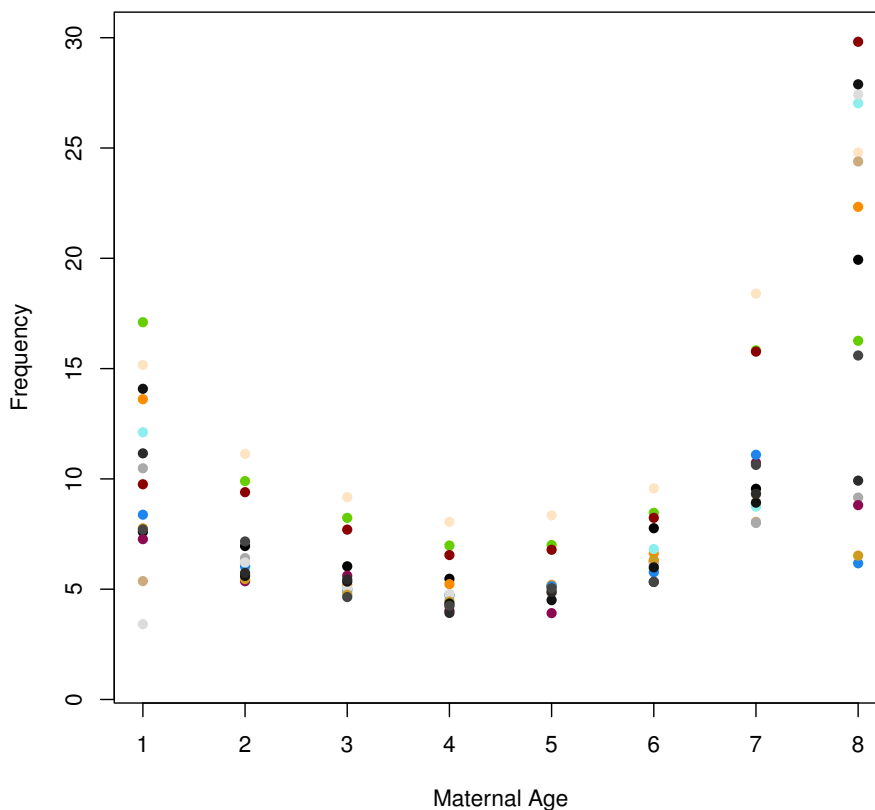
Figure 1: Neonatal mortality versus maternal age in Chile, between 1999 and 2016

The shape of the nonlinear negative binomial model with mean and regression structures given by (10) and (11) confirm the conclusion reported in Donoso *et al.* (2014), where a descriptive analysis of the data indicated parabolic nonlinear relationship of neonatal mortality versus maternal age in Chile. As in Villarroel del Pino (2003), this nonlinear model indicates high risk of maternal and perinatal morbidity of pregnancy in adolescents and women with advanced maternal ages (Donoso and Villarroel 2003).

6. Conclusions

In this paper we propose nonlinear (linear) regression models in F^2. In this class of models, we propose two new classes of overdispersed nonlinear regression models. The first class is defined from the overdispersion family of distributions proposed by Dey *et al.* (1994) and the second from the class of compound distributions. For these models, we develop a Bayesian method, as an extension of those proposed in Cepeda-Cuervo (2001), in which samples of the posterior distributions are obtained by applying an iterated M-H algorithm by assuming two groups of parameters, defined by the mean and dispersion regression structures. In the first class of models, defined in the overdispersed family of distributions of Dey *et al.* (1994), to propose the iterated M-H algorithm, we obtain working variables by applying the Fisher scoring algorithm and use these to build kernel transition functions to propose samples of the posterior parameter distributions. A Bayesian method to fit compound models is also proposed and applied to fit simulated and real datasets.

Particular cases of Bayesian methods proposed here to fit nonlinear biparametric regression models were developed in Cepeda-Cuervo and Achcar (2010) to fit heteroscedastic nonlinear

Table 1: Posterior mean (parameter estimates), standard deviations (SD), and 95% credible intervals of nonlinear negative binomial regression parameter

Parameters	Posterior mean	S.D.	95% Credible Intervals	M.C. Error
β_1	5.141	1.822	(1.838,8.96)	0.1597
β_2	0.5	0.04561	(0.4129,0.5904)	0.00354
β_3	3.976	0.07348	(3.826,4.114)	0.00147
β_4	4.515	0.1229	(4.287,4.51)	0.00539
β_5	-0.2651	0.1361	(-0.5304,-0.0053)	0.01161
γ_1	1.405	0.4019	(0.5435,2.127)	0.03276
γ_2	0.215	0.0455	(0.1296,0.3216)	0.0037
Deviance	1093	3.945	(1088,1103)	0.2251

regression models and in Cepeda-Cuervo (2001) to fit DGLMs. In all of these cases, including the applications developed in this paper, the results show good performance of the Bayesian methods proposed. However, there are many possible extensions that can be developed by practitioners to illustrate the efficiency of the MCMC algorithms developed here. Multiple applications and simulation studies can be developed as extensions of this work to illustrate the use of the proposed models and the performance of the Bayesian method proposed to fit nonlinear regression model.

In the aplication, several chains were generated, starting from different values. All of them exhibited the same qualitative behaviour though iterations after an small initial transient period, providing a rough indication of stationarity. The convergence of the posterior chains can be evaluated by applying Geweke's diagnostics (Geweke 1992).

7. Appendix

7.1. Double generalized linear models (DGLMs)

This section is a summary of the main results presented in Cepeda and Gamerman (2005), where the double generalized linear regression models (DGLMs) are defined in the framework of the biparametric family of distributions. A particular case of these models is defined from the family of distributions defined by Dey *et al.* (1994), were the mean and dispersion parameters are orthogonals, in the sense of Box and Cox, and with systematic components assumed to follow linear regression structures. Some examples of these models, considered in Cepeda-Cuervo (2001) and Cepeda and Gamerman (2005) are the beta regression models, where appropriate functions of the mean and dispersion parameters follow regression structures, or the gamma regression models, where appropriate functions of the mean and variance (or shape) parameters follow regression structures.

Model definition

The double generalized linear regression models are defined by the following three components:

- The *random component:* $\mathbf{Y} = (Y_1, ..., Y_n)^T$, where $Y_i \sim F(\mu_i, \tau_i)$, $i = 1...n$, are independent random variables with distribution belonging to the two-parameter exponential family of distributions (F_e^2).

- The *systematic component:* $\eta_i = (\eta_{1i}, \eta_{2i})^T$, such that $\eta_{1i} = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\eta_{2i} = \boldsymbol{z}_i^T \boldsymbol{\gamma}$.

- The *link functions:* $h(.)$ and $g(.)$, such that $h(\mu_i) = \eta_{1i}$ and $g(\tau_i) = \eta_{2i}$.

The link functions must be strictly monotonous and doubly differentiable to obtain maximum likelihood parameter estimations, and once differentiable in the Bayesian approach.
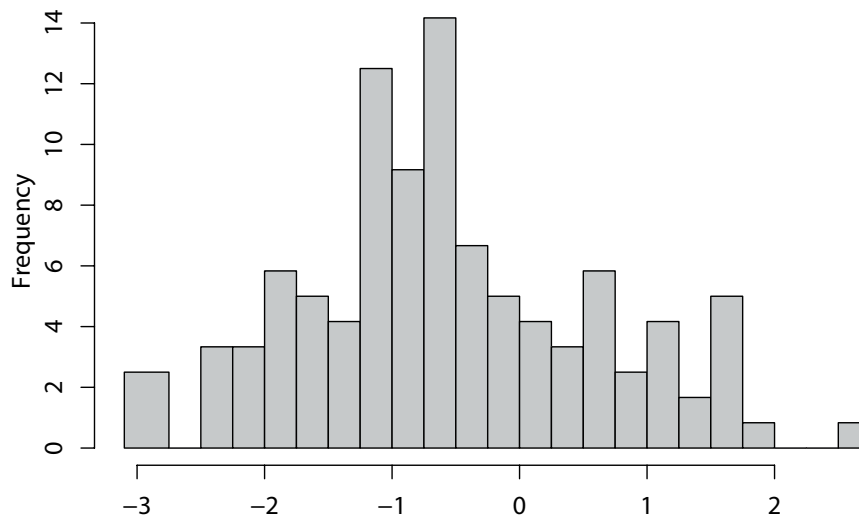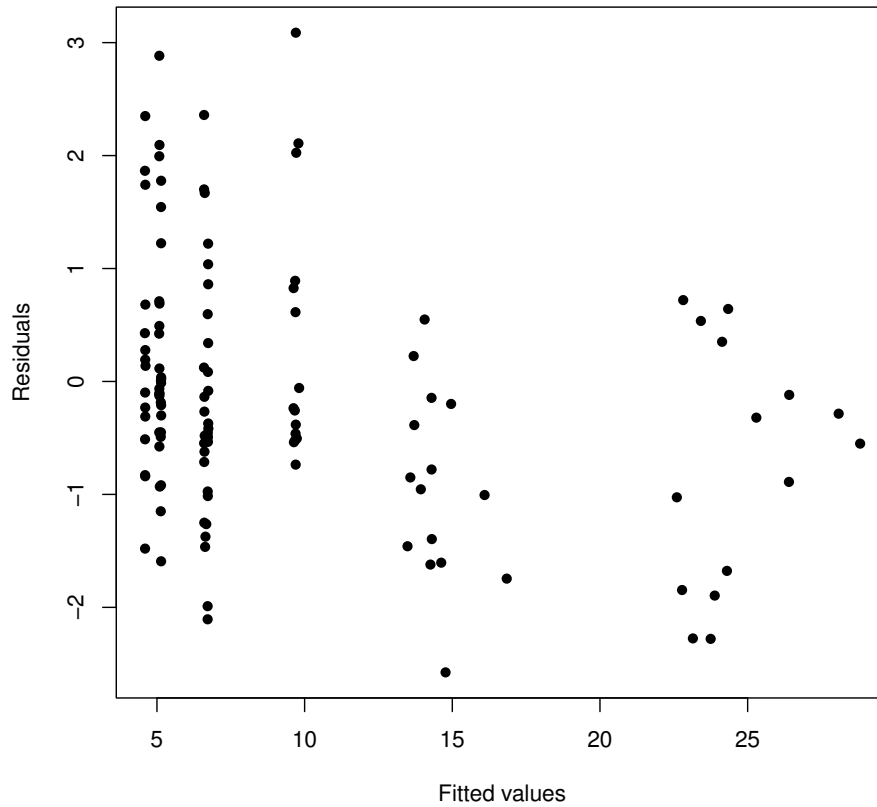
Figure 2: Plot of standardized Bayesian residuals versus fited values and Histogram of residuals

Some examples of the double generalized linear regression models, defined in the orthogonal overdispersed family of distributions of Dey *et al.* (1994) are:

1. *Heteroscedastic normal regression models.* Let $Y_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, ..., n$, be n independent random variables that follow normal distributions. This model is defined by assuming that $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\log(\sigma_i^2) = \boldsymbol{z}_i^T \boldsymbol{\gamma}$.

2. *Mean and shape gamma regression models.* Let $Y_i \sim G(\mu_i, \alpha_i)$, $i = 1, ..., n$, be n independent gamma random variables. These models are defined by assuming that $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\log(\alpha_i) = \boldsymbol{z}_i^T \boldsymbol{\gamma}$.

As examples of non-orthogonal two-parameter regression models, Cepeda-Cuervo (2001) and Cepeda and Gamerman (2005) proposed:

1. *Mean and dispersion beta regression models.* Let $Y_i \sim B(\mu_i, \phi_i)$, $i = 1, ..., n$, be n independent beta random variables, where $\mu_i = a_i/(a_i + b_i)$, $\phi_i = a_i + b_i$. These beta regression models are defined by assuming that:

$$\text{logit}(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{12}$$
$$\log(\phi_i) = \boldsymbol{z}_i^T \boldsymbol{\gamma}. \tag{13}$$

Mean and dispersion beta regression models. Let $Y_i \sim G(\mu_i, \sigma_i^2)$, $i = 1, ..., n$, be n independent gamma random variables, where $\mu_i = E(Y_i)$ and $\sigma_i^2 = Var(Y_i)$. These gamma regression models are defined by assuming that:

$$\text{logit}(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{14}$$
$$\log(\sigma_i^2) = \boldsymbol{z}_i^T \boldsymbol{\gamma}. \tag{15}$$

Bayesian method to fit DGLMs

To obtain posterior parameter estimates using Bayesian methods, the authors assumed normal prior distributions for the regression parameters, denoted by $p(\boldsymbol{\beta}, \boldsymbol{\gamma})$. The posterior parameter distribution, is given by $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) \propto L(\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}, \boldsymbol{\gamma})$, where $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is the likelihood function. Samples of the posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma})$, which are analytically intractable, are easily obtained from the posterior conditional distributions $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ and $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta})$ by applying the M-H algorithm, given that these conditional distributions are analytically intractable, except for $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ in the case of the heteroscedastic normal linear regression models (Cepeda and Gamerman 2000).

To obtain samples from the posterior conditional distributions by applying the M-H algorithm, the authors propose worked variables to approximate $h(\mu)$ and $g(\tau)$ around the current parameter estimates, obtained from the equation of the Fisher scoring algorithm or from the first-order Taylor approximation of $h(.)$ and $g(.)$ around the current values of μ_i and τ_i. Thus, a mean working variables are given by:

$$\tilde{Y}_i \approx h(\mu_i^{(c)}) + h'\left(\mu_i^{(c)}\right)\left(Y_i - \mu_i^{(c)}\right), \tag{16}$$

where $h'(.)$ denotes the first order derivative of $h(.)$, $h(\mu_i^{(c)}) = \boldsymbol{x}_i^T \boldsymbol{\beta}^{(c)}$ and $\mu_i^{(c)} = h^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(c)})$. From (16), a mean working observational model is obtained by assuming that the working variable $\tilde{Y}_i$, for which $E(\tilde{Y}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}^{(c)}$ and $Var(\tilde{Y}_i) = \left[h'(\mu_i^{(c)})\right]^2 Var(Y_i)$, follows a normal distribution. The mean kernel transition function q_1, to obtain samples of $\boldsymbol{\beta}$, is defined as the posterior distribution function obtained by the combination of the conditional prior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ and the working observational models $\tilde{Y}_i \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(c)}, Var^{(c)}(\tilde{Y}_i))$. Thus, the mean kernel transition function $q_1(.)$ is given by:

$$q_1(\boldsymbol{\beta}|\boldsymbol{\gamma}) = N(\mathbf{b}^*, \mathbf{B}^*) \tag{17}$$

where $\mathbf{b}^* = \mathbf{B}^*(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{Y}})$ and $\mathbf{B}^* = (\mathbf{B}^{-1} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}$; with $\boldsymbol{\Sigma} = Diag\left(Var^{(c)}\left(\tilde{Y}_i\right)\right)$ and $\tilde{\mathbf{Y}}$ the vector of observed values of (16).

The full posterior conditional distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is also intractable analytically. Thus, samples from this distribution are obtained by applying the M-H algorithm, for which a kernel transition function is proposed. From a random variable t_i, such that $E(t_i) = \tau_i$, a working dispersion random variable is defined by:

$$\check{Y}_{2i} = g(\tau_i^{(c)}) + g'\left(\tau_i^{(c)}\right)\left(t_i - \tau_i^{(c)}\right), \tag{18}$$

where $E(\check{Y}_{2i}) = \boldsymbol{z}_i^T\boldsymbol{\gamma}^{(c)}$ and $Var(\check{Y}_{2i}) = \left[g'(\tau_i^{(c)})\right]^2 Var(t_i)$, by assuming that it follows a normal distribution. The dispersion (τ) kernel transition function q_2, to obtain samples of $\boldsymbol{\gamma}$, is defined as the posterior distribution function obtained by the combination of the conditional prior distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta})$ and the working observational models $\check{Y}_i \sim N(\boldsymbol{z}_i^T\boldsymbol{\gamma}^{(c)}, Var^{(c)}(\check{y}_i))$. Thus, the kernel transition function $q_2(.)$ is given by:

$$q_2(\boldsymbol{\gamma}|\boldsymbol{\beta}) = N(\mathbf{g}^*, \mathbf{G}^*), \tag{19}$$

where $\mathbf{g}^* = \mathbf{G}^*(\mathbf{G}^{-1}\mathbf{g} + \mathbf{X}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{2}}^{-1}\check{\mathbf{Y}}_{\mathbf{2}})$ and $\mathbf{G}^* = (\mathbf{G}^{-1} + \mathbf{Z}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{2}}^{-1}\mathbf{Z})^{-1}$, with $\boldsymbol{\Sigma}_{\mathbf{2}} = Diag\left(Var^{(c)}\left(\check{Y}_i\right)\right)$ and $\tilde{\mathbf{Y}}$ the vector of observed values of the working variables.

In following examples we present the working variables developed in Cepeda-Cuervo (2001) to build the kernel transition function used in the M-H algorithm to fit beta regression and gamma regression models.

Example 7.1.1. *Mean and shape gamma regression models.* In gamma regression models with $h(\mu_i) = \boldsymbol{x}^T\boldsymbol{\beta}$ and $g(\alpha_i) = \boldsymbol{z}^T\boldsymbol{\gamma}$, if $\boldsymbol{\beta}^{(c)}$ denotes the current values of $\boldsymbol{\beta}$, the mean working variables are given by:

$$\tilde{Y}_i = \boldsymbol{x}_i^T\boldsymbol{\beta}^{(c)} + h'[h^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(c)})][Y_i - h^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(c)})], \tag{20}$$

for which $h(\mu^{(c)}) = \boldsymbol{x}_i^T\boldsymbol{\beta}^{(c)}$ and $\sigma_i^2 = \{h'[h^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(c)})]\}^2 Var(Y_i)$. If $g(.)$ is the identity function, $\tilde{Y}_i = Y_i$

To define a working variable to build a kernel transition function to propose samples of $\boldsymbol{\gamma}$, the random variable $t_i = \lambda Y_i$ is proposed, given that $E(t_i) = \alpha_i$. Thus, if $g(.)$ is the logarithmic function, from the first-order Taylor approximation of $\log(t_i)$ around the current value of $\boldsymbol{\gamma}^{(c)}$, the working shape variable is given by:

$$\check{Y}_i = \boldsymbol{z}_i^T\boldsymbol{\gamma}^{(c)} + \frac{1}{\alpha^{(c)}}(\lambda_i^{(c)}Y_i - \alpha^{(c)}) \tag{21}$$

$$= \boldsymbol{z}_i^T\boldsymbol{\gamma}^{(c)} + \frac{1}{\alpha^{(c)}}(\lambda_i^{(c)}Y_i - \alpha^{(c)}) \tag{22}$$

Example 7.1.2. *Mean and dispersion beta regression models.* In beta regression models with logit mean link function, $h(\mu_i) = \text{logit}(\mu_i)$, and logarithmic dispersion link function, $g(\phi_i) = \log(\phi_i)$, where $\phi = a + b$, the mean working variable, obtained as in example 7.1.1, from the first order Taylor approximation of $\text{logit}(y_i)$ around the current values of $\mu^{(c)}$ is:

$$\tilde{Y}_i = \boldsymbol{x}_i^T\boldsymbol{\beta}^{(c)} + \frac{Y_i - \mu_i^{(c)}}{\mu_i^{(c)}(1 - \mu_i^{(c)})}, \tag{23}$$

for which $g(\mu^{(c)}) = \boldsymbol{z}_i^T\boldsymbol{\gamma}^{(c)}$ and $\sigma_i^2 = \mu_i^{(c)}(1 - \mu_i^{(c)})[1 + exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}_i^{(c)})]$.

To define a working variable to build a kernel transition function to propose samples of $\boldsymbol{\gamma}$, the author starts by defining a random variable $t_i = \frac{(a+b)^2}{a}Y_i$, given that $E(t_i) = \phi_i$. Thus, from the first-order Taylor approximation of $\log(t_i)$ around the current value of $\phi^{(c)}$, the working dispersion variable is given by:

$$\check{Y}_i = \boldsymbol{z}_i^T\boldsymbol{\gamma}^{(c)} + \frac{Y_i}{\mu^{(c)}} - 1 \tag{24}$$

and their associated variance is $\sigma_i^2 = (1 - \mu_i^{(c)})\{(1 - \phi_i^{(c)})\}^{-1}[1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}_i^{(c)})]$.

7.2. Frequentist parameter estimates of nonlinear DGLMs

The Fisher scoring is defined from the first- and second-order derivatives of the logarithm of the likelihood function (25), as follows: let y_i, $i = 1, 2, ..., n$, be n independent observations of n random variables $Y_i \sim F_e^2(\mu_i, \tau_i)$, belonging to the exponential family (3). The logarithm of the likelihood function, $\ell = \log L$, is given by:

$$\ell = \sum_{i=1}^{n} \ell_i = \sum_{i=1}^{n} \{ log(b(y_i)) + (y_i - \mu_i)\Psi^{(1,0)}(\mu_i, \tau_i) + \tau_i T(y_i) + \Psi(\mu_i, \tau_i) \}, \tag{25}$$

where ℓ_i is the logarithm of the i-th likelihood component, $h(\mu_i) = S(\boldsymbol{x}_i^T, \boldsymbol{\beta})$ and $g(\tau_i) = \boldsymbol{z}_i^T \boldsymbol{\gamma}$. Thus, the first-order derivatives of l are given by:

$$\begin{aligned}
\frac{\partial l_i}{\partial \beta_j} &= \frac{\partial l_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_{1i}} \frac{\partial \eta_{1i}}{\partial \beta_j} \quad , j = 1, 2, ..., p \\
&= \left[(y_i - \mu_i)\Psi^{(2,0)}(\mu_i, \tau_i) \right] \frac{\partial \mu_i}{\partial \eta_{1i}} \frac{\partial S(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_j} \\
\frac{\partial l_i}{\partial \gamma_j} &= \frac{\partial l_i}{\partial \tau_i} \frac{\partial \tau_i}{\partial \eta_{2i}} \frac{\partial \eta_{2i}}{\partial \gamma_j}, \quad j = 1, 2, ..., r \\
&= \left[(y_i - \mu_i)\Psi^{(1,1)}(\mu_i, \tau_i) + T(y_i) + \Psi^{(0,1)}(\mu_i, \tau_i) \right] \frac{\partial \tau_i}{\partial \eta_{2i}} z_{ij},
\end{aligned} \tag{26}$$

where $\frac{\partial \tau_i}{\partial \eta_{2i}} = \frac{1}{h'(\tau_i)}$ and $\frac{\partial \mu_i}{\partial \eta_{1i}} = \frac{1}{g'(\mu_i)}$. Thus, the Hessian matrix, obtained by second-order derivatives of ℓ, is given by:

$$\begin{aligned}
\frac{\partial^2 l_i}{\partial \beta_k \beta_j} &= \left[\frac{\partial^2 l_i}{\partial \mu_i^2} \left(\frac{\partial \mu_i}{\partial \eta_{1i}} \right)^2 + \frac{\partial l_i}{\partial \mu_i} \frac{\partial^2 \mu_i}{\partial \eta_{1i}^2} \right] \frac{\partial \eta_{1i}}{\partial \beta_k} \frac{\partial \eta_{1i}}{\partial \beta_j} + \frac{\partial l_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_{1i}} \frac{\partial^2 \eta_{1i}}{\partial \beta_k \partial \beta_j} \\
&= \left[\frac{(y_i - \mu_i)\Psi^{(3,0)}(\mu_i, \tau_i) - \Psi^{(2,0)}(\mu_i, \tau_i)}{h'(\mu_i)^2} - \frac{(y_i - \mu_i)\Psi^{(2,0)}(\mu_i, \tau_i)h''(\mu_i)}{h'(\mu_i)^3} \right] \frac{\partial S(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_k} \frac{\partial S(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_j} \\
&\quad + \left[(y_i - \mu_i)\Psi^{(2,0)}(\mu_i, \tau_i) \right] \frac{\partial \mu_i}{\partial \eta_{1i}} \frac{\partial^2 S(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_k \partial \beta_j} \\
\frac{\partial^2 l_i}{\partial \gamma_k \partial \beta_j} &= \frac{\partial^2 l_i}{\partial \tau_i \partial \mu_i} \frac{\partial \tau_i}{\partial \eta_{2i}} \frac{\partial \mu_i}{\partial \eta_{1i}} \frac{\partial \eta_{2i}}{\partial \gamma_k} \frac{\partial \eta_{1i}}{\partial \beta_j} \\
&= (y_i - \mu_i)\Psi^{(2,1)}(\mu_i, \tau_i) \left(\frac{1}{h'(\mu_i)} \right) \left(\frac{1}{g'(\tau_i)} \right) \frac{\partial S(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_j} z_{ik} \\
\frac{\partial^2 l_i}{\partial \gamma_k \partial \gamma_j} &= \left[\frac{\partial^2 l_i}{\partial \tau_i^2} \left(\frac{\partial \tau_i}{\partial \eta_{2i}} \right)^2 + \frac{\partial l_i}{\partial \tau_i} \frac{\partial^2 \tau_i}{\partial \eta_{2i}^2} \right] \frac{\partial \eta_{2i}}{\partial \gamma_k} \frac{\partial \eta_{2i}}{\partial \gamma_j} \\
&= \left[\frac{(y_i - \mu_i)\Psi^{(1,2)}(\mu_i, \tau_i) + \Psi^{(0,2)}(\mu_i, \tau_i)}{g'(\tau_i)^2} - \frac{\{(y_i - \mu_i)\Psi^{(1,1)}(\mu_i, \tau_i) + T(y_i) + \Psi^{(0,1)}(\mu_i, \tau_i)\}}{g'(\tau_i)^3 / g''(\tau_i)} \right] z_{ik} z_{ij},
\end{aligned} \tag{27}$$

where $k, j = 1, ..., p$ for β and $k, j = 1, ..., r$ for γ. So, the Fisher information matrix, obtained by the expected values of the second-order derivatives, is given by:

$$\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial \beta_k \partial \beta_j} \right] &= \sum_{i=1}^{n} \left[\frac{\Psi^{(2,0)}(\mu_i, \tau_i)}{g'(\mu_i)^2} \right] \frac{\partial S(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_k} \frac{\partial S(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \beta_j} \\
-E\left[\frac{\partial^2 l}{\partial \gamma_l \partial \beta_i} \right] &= 0 \\
-E\left[\frac{\partial^2 l}{\partial \gamma_l \partial \gamma_j} \right] &= \sum_{i=1}^{n} \left[\frac{-\Psi^{(0,2)}(\mu_i, \tau_i)}{h'(\tau_i)^2} + \frac{\{E(T(y_i)) + \Psi^{(0,1)}(\mu_i, \tau_i)\} h''(\tau_i)}{h'(\tau_i)^3} \right] z_{ij} z_{il}
\end{aligned}$$

From these results, the Fisher information matrix is block diagonal, the parameter vectors, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, are globally orthogonal (Cox and Reid 1987) and their maximum likelihood estimates are asymptotically independent.

Thus, given the k-th estimates of the regression parameters, $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$, by the Fisher scoring algorithm, the (k+1)-th estimate of $\boldsymbol{\beta}$ is given by the equation $\mathbf{I}_{\beta}^{(k)}\,\boldsymbol{\beta}^{(k+1)} = \mathbf{I}_{\beta}^{(k)}\,\boldsymbol{\beta}^{(k)} + \mathbf{q}_{\beta}^{(k)}$, where $\mathbf{q}_{\beta}^{(k)} = \left(\frac{\partial l}{\partial \beta_1}, ..., \frac{\partial l}{\partial \beta_p}\right)^T$ and the i-th component of the right hand is given by:

$$
\begin{aligned}
\left(\mathbf{I}_{\beta}^{(k)}\boldsymbol{\beta}^{(k)} + \mathbf{q}_{\beta}^{(k)}\right)_i &= \sum_{m=1}^{p}\left(\sum_{l=1}^{n}\Psi^{(2,0)}(\mu_l^{(k)},\tau_l^{(k)})\left(\frac{1}{g'\left(\mu_l^{(k)}\right)}\right)^2 \frac{\partial S(\boldsymbol{x}_l,\boldsymbol{\beta})}{\partial \beta_i}\frac{\partial S(\boldsymbol{x}_l,\boldsymbol{\beta})}{\partial \beta_m}\right)\beta_m^{(k)} + \frac{\partial l}{\partial \beta_i} \\
&= \sum_{m=1}^{p}\left(\sum_{l=1}^{n}\Psi^{(2,0)}(\mu_l^{(k)},\tau_l^{(k)})\left(\frac{1}{g'\left(\mu_l^{(k)}\right)}\right)^2 \frac{\partial S(\boldsymbol{x}_l,\boldsymbol{\beta})}{\partial \beta_i}\frac{\partial S(\boldsymbol{x}_l,\boldsymbol{\beta})}{\partial \beta_m}\beta_m^{(k)}\right) \\
&\quad + \sum_{l=1}^{n}\Psi^{(2,0)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)\left(y_l - \mu_l^{(k)}\right)\left(\frac{1}{g'\left(\mu_l^{(k)}\right)}\right)\frac{\partial S(\boldsymbol{x}_l,\boldsymbol{\beta}^{(k)})}{\partial \beta_i} \\
&= \sum_{l=1}^{n}\frac{\partial S}{\partial \beta_i}(\boldsymbol{x}_l,\boldsymbol{\beta})\frac{\Psi^{(2,0)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)}{g'\left(\mu_l^{(k)}\right)}\left(\sum_{m=1}^{p}\frac{1}{g'\left(\mu_l^{(k)}\right)}\frac{\partial S(\boldsymbol{x}_l,\boldsymbol{\beta})}{\partial \beta_m}\beta_m^{(k)} + \left(y_l - \mu_l^{(k)}\right)\right)
\end{aligned}
$$

In summary, $\mathbf{I}_{\beta}^{(k)}\,\boldsymbol{\beta}^{(k)} + \mathbf{q}_{\beta}^{(k)} = \tilde{\mathbf{X}}\,\tilde{\mathbf{W}}\,\tilde{\mathbf{Y}}_1$, where the entries of $\tilde{X}$ are given by $\tilde{x}_{ij} = \frac{\partial S}{\partial \beta_i}(x_j,\boldsymbol{\beta}^{(k)})$, $\tilde{\mathbf{W}} = Diag(w_i^{(k)})$, where $\tilde{w}_i = \frac{\Psi^{(2,0)}(\mu_i^{(k)},\tau_i^{(k)})}{g'(\mu_i^{(k)})^2}$, and $\tilde{\mathbf{Y}}_1$ is an (n x 1) worked vector with components $\tilde{y}_{1i}$, $i = 1, 2, \ldots, n$, given by:

$$
\tilde{y}_{1i} = \underbrace{\sum_{m=1}^{p}\frac{\partial S(\boldsymbol{x}_i,\boldsymbol{\beta}^{(k)})}{\partial \beta_m}\beta_m^{(k)}}_{\nabla S\left(\boldsymbol{x}_i^T,\boldsymbol{\beta}^{(k)}\right)\boldsymbol{\beta}^{(k)}} + \left(y_i - \mu_i^{(k)}\right)g'\left(\mu_i^{(k)}\right). \tag{28}
$$

Then, given the k-th estimates of the regression parameters, $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$, by the Fisher scoring algorithm, the (k+1)-th estimate of $\boldsymbol{\gamma}$ is given by the equation $\mathbf{I}_{\gamma}^{(k)}\boldsymbol{\gamma}^{(k+1)} = \mathbf{I}_{\gamma}^{(k)}\boldsymbol{\gamma}^{(k)} + \mathbf{q}_{\gamma}^{(k)}$, where $\mathbf{q}_{\beta}^{(k)} = \left(\frac{\partial l}{\partial \gamma_1}, ..., \frac{\partial l}{\partial \gamma_p}\right)'$, and the i-th component of the right-hand side is given by:

$$
\begin{aligned}
\left(\mathbf{I}_{\gamma}^{(k)}\boldsymbol{\gamma}^{(k)} + \mathbf{q}_{\gamma}^{(k)}\right)_j &= \sum_{m=1}^{r}\left(\sum_{l=1}^{n}\left[\frac{-\Psi^{(0,2)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)}{g'\left(\tau_l^{(k)}\right)^2} + \frac{E\left(T(y_l)\right) + \Psi^{(0,1)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)}{g'\left(\tau_l^{(k)}\right)^3\Big/g''\left(\tau_l^{(k)}\right)}\right]z_{lj}z_{lm}\right) \\
&\quad \times\gamma_m^{(k)} + \frac{\partial l}{\partial \gamma_j} \\
&= \sum_{l=1}^{n}z_{lj}\left\{\left[\frac{-\Psi^{(0,2)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)}{g'\left(\tau_l^{(k)}\right)^2} + \frac{\left\{E\left(T(y_l)\right) + \Psi^{(0,1)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)\right\}g''\left(\tau_l^{(k)}\right)}{h'\left(\tau_l^{(k)}\right)^3}\right]\right. \\
&\quad \left.\times\sum_{m=1}^{r}z_{lm}\gamma_m^{(k)} + \frac{\left(y_l - \mu_l^{(k)}\right)\Psi^{(1,1)}(\mu_l^{(k)},\tau_l^{(k)}) + T(y_l) + \Psi^{(0,1)}(\mu_l^{(k)},\tau_l^{(k)})}{g'\left(\tau_l^{(k)}\right)}\right\} \\
&= \sum_{l=1}^{n}\frac{z_{lj}\left[-\Psi^{(0,2)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)g'\left(\tau_l^{(k)}\right) + \left\{E\left(T(y_l)\right) + \Psi^{(0,1)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)\right\}g''\left(\tau_l^{(k)}\right)\right]}{g'\left(\tau_l^{(k)}\right)^3} \\
&\quad \times\left\{\eta_{2l}^{(k)} + \frac{h'\left(\tau_l^{(k)}\right)^2\left[(y_l - \mu_l^{(k)})\Psi^{(1,1)}(\mu_l^{(k)},\tau_l^{(k)}) + T(y_l) + \Psi^{(0,1)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)\right]}{-\Psi^{(0,2)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)g'\left(\tau_l^{(k)}\right) + \left\{E\left(T(y_l)\right) + \Psi^{(0,1)}\left(\mu_l^{(k)},\tau_l^{(k)}\right)\right\}g''\left(\tau_l^{(k)}\right)}\right\}
\end{aligned}
$$

Thus, the Fisher scoring worked variables used to build the kernel transition function to

obtain samples of $\boldsymbol{\gamma}$ are given by:

$$\tilde{y}_{2i} = \eta_{2i}^{(k)} + \frac{g'\left(\tau_i^{(k)}\right)^2 \left[(y_i - \mu_i)\Psi^{(1,1)}\left(\mu_I^{(k)}, \tau_i^{(k)}\right) + T(y_i) + \Psi^{(0,1)}\left(\mu_i^{(k)}, \tau_i^{(k)}\right)\right]}{-\Psi^{(0,2)}\left(\mu_i^{(k)}, \tau_i^{(k)}\right) g'\left(\tau_i^{(k)}\right) + \left\{E\left(T(y_i)\right) + \Psi^{(0,1)}\left(\mu_i^{(k)}, \tau_i^{(k)}\right)\right\} g''\left(\tau_i^{(k)}\right)}. \quad (29)$$

With these results an iterate maximum likelihood algorithm can be proposed to obtain maximum likelihood estimates of the regression parameters. To fit the Bayesian model obtained by assuming a normal prior distribution for the regression parameters, we can apply the Bayesian method proposed in Section (3), where the mean kernel transition function is given by (19), built with the worked variable (28).

References

Breslow NE (1984). "Extra-Poisson Variation in Log-Linear Models." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **33**(1), 38–44. URL https://doi.org/10.2307/2347661.

Cameron AC, Trivedi PK (1998). *Regression Analysis of Count Data.* Cambridge University Press. ISBN: 9781139013567, URL https://digital.library.tu.ac.th/tu_dc/frontend/Info/item/dc:11207.

Cepeda E, Gamerman D (2000). "Bayesian Modeling of Variance Heterogeneity in Normal Regression Models." *Brazilian Journal of Probability and Statistics*, pp. 207–221. URL https://www.jstor.org/stable/43600978.

Cepeda E, Gamerman D (2005). "Bayesian Methodology for Modeling Parameters in the Two Parameter Exponential Family." *Revista Estadística*, **57**(168, 169), 93–105.

Cepeda-Cuervo E (2001). "Modelagem da Variabilidade em Modelos Lineares Generalizados." Unpublished Ph.D. thesis, Mathematics Institute, Universidade Federal do Río de Janeiro, URL https://repositorio.unal.edu.co/bitstream/handle/unal/11839/modelagem_variabilidade_modelos.pdf?sequence=1&isAllowed=y.

Cepeda-Cuervo E, Achcar J (2010). "Heteroscedastic Nonlinear Regression Models." *Communications in Statistics - Simulation and Computation*, **39**(2), 405–419. doi:10.1080/03610910903480784. URL https://doi.org/10.1080/03610910903480784.

Cepeda-Cuervo E, Cifuentes-Amado MV (2017). "Double Generalized Beta-Binomial and Negative Binomial Regression Models." *Revista Colombiana de Estadística*, **40**(1), 141–163. URL https://doi.org/10.15446/rce.v40n1.61779.

Cepeda-Cuervo E, Migon H, Garrido L, Achcar J (2014). "Generalized Linear Models with Random Effects in the Two-parameter Exponential Family." *Journal of Statistical Computation and Simulation*, **84**(3), 513–525. doi:10.1080/00949655.2012.717292. URL https://doi.org/10.1080/00949655.2012.717292.

Collet D (1991). *Modeling Binary Data.* Chapman Hall, London. ISBN-13: 978-1584883241, URL https://doi.org/10.1201/b16654.

Cox DR (1983). "Some Remarks on Overdispersion." *Biometrika*, **70**(1), 269–274. URL https://doi.org/10.1093/biomet/70.1.269.

Cox DR, Reid N (1987). "Parameter Orthogonality and Approximate Conditional Inference." *Journal of the Royal Statistical Society: Series B (Methodological)*, **49**(1), 1–18.

Crowder MJ (1978). "Beta-Binomial Anova for Proportions." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **27**(1), 34–37. URL http://www.jstor.org/stable/2346223.

Dey D, Gelfand AE, Peng F (1994). "Overdispersed Generalized Linear Models." *Journal of Statistical Planning and Inference*, **64**(1), 93–107. URL https://doi.org/10.1016/S0378-3758(96)00207-8.

Donoso E, Carvajal JA, Vera C, Poblete JA (2014). "La Edad de la Mujer como Factor de Riesgo de Mortalidad Materna, Fetal, Neonatal e Infantil." *Revista médica de Chile*, **142**(2), 168–174. URL http://dx.doi.org/10.4067/S0034-98872014000200004.

Donoso E, Villarroel L (2003). "Reproductive Risk of Women over 40 Years Old." *Revista Médica de Chile*, **131**(1), 55–59. URL https://pubmed.ncbi.nlm.nih.gov/12643219/.

Efron B (1986). "Double Exponential Families and Their Use in Generalized Linear Regression." *Journal of the American Statistical Association*, **81**(395), 709–721. URL http://www.jstor.org/stable/2289002.

Fokianos K (2012). "12 - Count Time Series Models." In T Subba Rao, S Subba Rao, CR Rao (eds.), *Time Series Analysis: Methods and Applications*, volume 30 of *Handbook of Statistics*, pp. 315–347. Elsevier. doi:https://doi.org/10.1016/B978-0-444-53858-1.00012-0. URL https://www.sciencedirect.com/science/article/pii/B9780444538581000120.

Frome EL (1983). "The Analysis of Rates Using Poisson Regression Models." *Biometrics*, pp. 665–674. URL https://doi.org/10.2307/2531094.

Frome EL, Dufrain RJ (1986). "Maximum Likelihood Estimation for Cytogenetic Dose-Response Curves." *Biometrics*, pp. 73–84. URL https://doi.org/10.2307/2531244.

Frome EL, Kutner MH, Beauchamp JJ (1973). "Regression Analysis of Poisson-Distributed Data." *Journal of the American Statistical Association*, **68**(344), 935–940. URL https://doi.org/10.1080/01621459.1973.10481449.

Gamerman D (1997). "Sampling from the Posterior Distribution in Generalized Linear Mixed Models." *Statistics and Computing*, **7**(1), 57–68. URL https://doi.org/10.1023/A:1018509429360.

Garrido L, Cepeda-Cuervo E, Achcar J (2011). "Heteroscedastic Normal–exponential Mixture Models: Bayesian and Classical Approaches." *Applied Mathematics and Computation*, **218**(7), 3635–3648. URL https://doi.org/10.1016/j.amc.2011.09.005.

Gelfand AE, Dalal SR (1990). "A Note on Overdispersed Exponential Families." *Biometrika*, **77**(1), 55–64. URL https://doi.org/10.1093/biomet/77.1.55.

Geweke J (1992). *Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments.* University Pres. ISBN: 9781383023756, URL https://doi.org/10.1093/oso/9780198522669.003.0010.

Hinde J, Demétrio C (1998). "Overdispersion: Models and Estimation." *Computational Statistics & Data Analysis*, **27**(2), 151–170. URL https://doi.org/10.1016/S0167-9473(98)00007-3.

Hougaard P, Lee MT, Whitmore GA (1997). "Analysis of Overdispersed Count Data by Mixtures of Poisson Variables and Poisson Processes." *Biometrics*, pp. 1225–1238. URL https://doi.org/10.2307/2533492.

Jorgensen B (1987). "Exponential Dispersion Models." *Journal of the Royal Statistical Society. Series B (Methodological)*, **49**(2), 127–162. URL http://www.jstor.org/stable/2345415.

Lindley DV (1958). "Fiducial Distributions and Bayes' Theorem." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 102–107. URL https://www.jstor.org/stable/2983909.

Margolin BH, Kaplan N, Zeiger E (1981). "Statistical Analysis of the Ames Salmonella/Microsome Test." *Proceedings of the National Academy of Sciences*, **78**(6), 3779–3783. URL https://doi.org/10.1073/pnas.78.6.3779.

McCullagh P, Nelder JA (1989). *Generalized Linear Models, Second Edition.* Chapman and Hall. ISBN 978-0412317606, URL http://books.google.com/books?id=h9kFH2_FfBkC.

Munuswamy S (1970). "275. Note: The Discrete Poisson-Lindley Distribution." *Biometrics*, **26**(1), 145–149. URL https://doi.org/10.2307/2529053.

Noack A (1950). "A Class of Random Variables with Discrete Distributions." *Annals of Mathematical Statistics*, **21**(1), 127–132. URL https://doi.org/10.1214/aoms/1177729894.

Villarroel del Pino L (2003). "Birth Rates and Reproductive Risk in Adolescents in Chile." *Pan American Journal of Public Health*, **14**(1), 3–8. URL https://doi.org/10.1590/s1020-49892003000600002.

Williams DA (1982). "Extra-Binomial Variation in Logistic Linear Models." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **31**(2), 144–188. URL https://doi.org/10.2307/2347977.

Affiliation:

Maria V Cifuentes-Amado
Facultad de Ciencias
Universidad Nacional de Colombia
Calle 45, Bogota, Colombia
E-mail: mvcifuentesa@unal.edu.co

Edilberto Cepeda-Cuervo
Facultad de Ciencias
Universidad Nacional de Colombia
Calle 45, Bogota, Colombia
E-mail: ecepedac@unal.edu.co