Gamma regression models with the Gammareg R package

Martha Corrales-Bosio
mlcorralesb@unal.edu.co

Edilberto Cepeda-Cuervo
ecepedac@unal.edu.co

Departamento de Estadística
Universidad Nacional de Colombia

Abstract

The class of gamma regression models is based on the assumption that the dependent variable is gamma distributed and that its mean is related to a set of regressors through a linear predictor with unknown coefficients and a link function. This link can be the identity, the inverse or the logarithm function. The model also includes a shape parameter, which may be constant or dependent on a set of regressors through a link function, as the logarithm function. In this paper we describe the Gammareg R-package, which provides the class of gamma regressions in the R system for their statistical computing. The underlying theory is briefly presented and the library implementation illustrated from simulation studies.

Keywords: *Gamma regression, mean regression structures, shape regression structures, Fisher Scoring algorithm, R-package*

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

In this paper we introduce the use of the Gammareg R-package, an R-code developed by us that contains the algorithms and customizable options to fit, under classic methodology, Gamma regression model where both, mean and shape parameters follow regression structures. After the introduction, this paper includes six sections. In Section 2, a mean-shape re-parameterizations of the Gamma distribution is presented. In Section 3 the gamma regression models, where both mean and shape parameters follow regression structures, is presented. In Section 4, a classic method to fit gamma regression models is sumarized. Section 4 presents the Classic Gammareg R-package. Section 6 contains two applications based on simulated data. Finally, in Section 6 we present our main conclusions.

2 Gamma distribution

A random variable Y follows a gamma distribution if its probability density function is given by

$$f(y|\alpha,\lambda) = \frac{\lambda^{\alpha} y^{\alpha-1} e^{-\lambda y}}{\Gamma(\alpha)} \mathbf{I}_{(0,\infty)}(y) \tag{1}$$

where $\alpha, \lambda > 0$ and $\Gamma(.)$ denotes the gamma function; and $\mathbf{I}$ is an indicator function such that $\mathbf{I}_{(0,\infty)}(y) = 1$ if $y \in (0,\infty)$, and zero otherwise. Under this parameterization, the mean and variance of Y are given by $\mathrm{E}(Y) = \alpha/\lambda$ and $\mathrm{Var}(Y) = \alpha/\lambda^2 = \mu^2/\alpha$.

With the re-parameterization of the gamma distribution as a function of the mean, $\mu = \mathrm{E}(Y)$,and the shape parameter,α, as proposed in Cepeda-Cuervo (2001) and Cepeda & Gammerman (2005), setting $\lambda = \alpha/\mu$, the gamma density function can be written as

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

Under this re-parameterization, we use $Y \sim G(\mu,\alpha)$ to denote that Y follows a gamma distribution with $\mathrm{E}(Y) = \mu$ and α as a shape parameter.

From this re-parameterization of the gamma distribution, the joint mean and shape gamma regression were proposed in Cepeda-Cuervo (2001), under classic and Bayesian methodologies, as presented in the next section.

3 Gamma regression models

Let $Y_i \sim G(\mu_i,\alpha)$, $i = 1,\ldots,n$ be a Gamma random sample, of size n. In the gamma regression model with constant shape parameter, the mean regression

structure is defined by

$$g(\mu_i) = \mathbf{x_i}'\beta = \eta_i$$

where g is the link function,$\beta = (\beta_0, \ldots, \beta_p)'$ is the vector of mean regression parameters, $\mathbf{x_i}$ is the i-th vector value of the explanatory variables, and η_i is a linear predictor. Here, $g(.) : (0, \infty) \mapsto \Re$ is a real value function strictly monotonic and twice differentiable (McCullagh & Nelder 1989). 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.

An extension of the gamma regression models is proposed in Cepeda-Cuervo (2001)). In this proposal, the shape parameter is not constant through the observations and it is modeled following a regression structure. That is,the mean and shape parameters follow a regression structures given by:

$$g(\mu_i) = \eta_{1i} = \mathbf{x_i'}\beta \quad (3)$$

$$h(\alpha_i) = \eta_{2i} = \mathbf{z_i'}\gamma \quad (4)$$

where g and h are appropriate real link functions, $\beta = (\beta_0, \ldots, \beta_p)'$ and $\gamma = (\gamma_0, \ldots, \gamma_k)'$, with $p + k < n$, are respectively the mean and shape parameter vectors, $\mathbf{x_i}$ and $\mathbf{z_i}$ are respectively the mean and shape explanatory variables for the i-th observation, and η_{1i}, η_{2i} are the linear predictors. A usual link function for the shape model is the logarithm function.

4 A classic method to fit gamma regression models

(Cepeda-Cuervo 2001) proposed a classical approach to joint modeling the mean and shape parameters using the Fisher scoring algorithm. In that work, he showed that with the gamma reparametrization given by (2), the likelihood function on the gamma regression models defined by (3) and (4), can be written in the form:

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

where $\mu_i = \mathbf{x_i'}\beta$ and $\alpha_i = \exp(\mathbf{z_i'}\gamma)$, and the log likelihood function by:

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

Thus, the score function has components:

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

And the Hessian matrix determined by:

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

Thus, the Fisher information matrix is given by:

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

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 $\boldsymbol{\beta}$ and the other to the shape regression parameter $\boldsymbol{\gamma}$. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are then orthogonal, and the maximum likelihood estimators, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, are asymptotically independent.

Taken in account the structure of the Fisher information matrix, (Cepeda-Cuervo 2001) proposed an iterative algorithm to obtain the maximum likelihood estimates of then regression parameters, where:

1. Given the k-th parameter values $(\boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)})'$, the mean vector $\boldsymbol{\beta}^{(k+1)}$ is updated from :

$$\boldsymbol{\beta^{(k+1)}} = \boldsymbol{(X'W^{(k)}X)^{-1}X'W^{(k)}Y} \tag{5}$$

where $\boldsymbol{W^{(k)}}$ is a matrix with elements $w_i^{(k)} = (\mu_i^2/\alpha_i)$.

2. Given $(\boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)})'$, the shape parameter vector $\boldsymbol{\gamma}^{(k+1)}$ is updated from :

$$\boldsymbol{\gamma^{(k+1)}} = \boldsymbol{(Z'W^{(k)}Z)^{-1}X'W^{(k)}Y} \tag{6}$$

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

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

Therefore, the alternate iterate algorithm can be summarized as follows:

1. Start an iteration count $k = 0$.
2. Give initial values for the parameters $\beta^{(k)}, \gamma^{(k)}$.
3. Obtain $\beta^{(k+1)}$ from equation (5),giving the current values of β and γ.
4. Obtain $\gamma^{(k+1)}$ from equation (6),giving the current values of β and γ.
5. Set the counter iteration $k = k + 1$
6. Go to 3 and 4 are until convergence is achieved.

For other links functions, similar results can be obtained. In particular if $g(.)$ and $h(.)$ are logarithm function, the Fisher information matrix is block diagonal, with two blocks, one for $\boldsymbol{\beta}$ and other for $\boldsymbol{\gamma}$, and thus a similar alternate iterated algorithm can be implemented.

5 Implementation in R: Gammareg package

We can estimate the parameters of a gamma regression as proposed by (McCullagh & Nelder 1989) in R (Team n.d.), using the function glm {stats} for fitting GLMs. An important difference with our proposal is that in **Gammareg** there are potentially two regression structures, one for the mean and other for the shape parameter.

The Gammareg R-package has the computational implementation of the Classical method defined in Section 4. The main model-fitting function in **Gammareg** is Gammareg(), which allows the user to calculate the mean and shape regression parameters in a gamma regression model under classical perspective. The general formula for this function is

Gammareg(formula1,formula2,meanlink)

where

1. ***formula1*** is an object of class formula, that describes the interest variable Y and the regressors X of the mean regression structure.

2. ***formula2*** is an object of class formula that describes the regressors Z for the shape regression structure.

3. ***meanlink*** is the link function for the mean.

The default link is the log (log) link, but the identity (ide) link is also allowed as admisible value.

The returned fitted-model object of the Gammareg class is a list similar to **glm** objects. It provides to the user the regression parameter estimates, $\widehat{\beta}$ and $\widehat{\gamma}$ and their standard deviations. It also provides the estimated covariance matrix for β and γ, the criterion value AIC, the number of iterations to covergence and the value of covergence obtained.

The **Gammareg** R-package has five other functions which allow the user, among other things, to obtain summaries for gamma regression models.

The functions on the package **BayGammareg** are described in the Table 1.

Table 1: Gammareg functions

Function	Description
Gammareg	estimates the media and shape regression parameters
gammahetero1()	performs the classic gamma regression using link log for the mean and link log for the shape.
gammahetero2()	performs the classic gamma regression using identity link for the mean and link log for the shape
summary.Gammareg()	is the standard regression output (coefficient estimates, standard errors, criterions); returns an object of class summary.Gammareg containing the relevant summary statistics (which has a print() method)
print.Gammareg	prints the estimates coefficients and the confidence intervals of a classic gamma regression
print.summary.Gammareg	prints the summary of a classic gamma regression

6 Gammareg in practice

To illustrate the use of **Gammareg** we consider two gamma regression models with simulated data, using two differents links for the mean: log and identity links.

6.1 First simulation

In the first simulation, we consider a gamma regression model with mean and shape structures, given by:

$$\mu_i = \mathbf{x_i}'\beta$$
$$\log(\alpha_i) = \mathbf{z_i'}\gamma$$

First, we generated values of the explanatory variables X_1, X_2, X_3 and X_4. For each of this variables, we simulated $n = 500$ values with $x_{1i} = 1$ for $i = 1, ..., n; x_{2i}$, $i = 1, ..., n$, from a uniform distribution on the interval $(0, 30)$; x_{3i}, $i = 1, ..., n$; from a uniform distribution on the interval $(0, 15)$, and x_{4i}, $i = 1, ..., n$, from a uniform distribution in the interval $(10, 20)$. The values Y_i were generated from a gamma distribution with $\mu_i = 15 + 2x_{2i} + 3x_{3i}$ and $\alpha_i = exp(0.2 + 0.1x_{2i} + 0.3x_4 i)$, as follow:

```
>library(Gammareg)
>X1 <- rep(1,500)
>X2 <- runif(500,0,30)
>X3 <- runif(500,0,15)
>X4 <- runif(500,10,20)
>mui <- 15 + 2*X2 + 3*X3
>alphai <- exp(0.2 + 0.1*X2 + 0.3*X4)
>Y <- rgamma(500,shape=alphai,scale=mui/alphai)
>X <- cbind(X1,X2,X3)
>Z <- cbind(X1,X2,X4)
>formula.mean= Y~X2+X3
>formula.shape= ~X2+X4
>a=Gammareg(formula.mean,formula.shape,meanlink="ide")
>summary(a)
```

The results obtained to apply the Gammareg R-pakage were:

```
#####################################################################
###                  Classic Gamma Regression                     ###
#####################################################################

Call:
Gammareg(formula1 = formula.mean, formula2 = formula.shape, meanlink = "ide")

                  Estimate  L.Intv U.Intv
beta.(Intercept)   15.2231 14.6931 15.753
beta.X2             2.0382  2.0131  2.063
beta.X3             2.8965  2.8483  2.945
gamma.(Intercept)   0.1877  0.1870  0.189
gamma.X2            0.1031  0.1031  0.103
gamma.X4            0.2934  0.2933  0.293

Covariance Matrix for Beta:
             (Intercept)            X2            X3
(Intercept)  0.072747993 -2.615207e-03 -2.819056e-03
X2          -0.002615207  1.637345e-04 -3.582766e-05
X3          -0.002819056 -3.582766e-05  6.024330e-04

Covariance Matrix for Gamma:
```

```
             (Intercept)            X2            X4
(Intercept)  1.485298e-07 -1.243918e-09 -6.456682e-09
X2          -1.243918e-09  4.430374e-11  7.859803e-12
X4          -6.456682e-09  7.859803e-12  3.468175e-10

AIC:
[1] 6606828

Iteration:
[1] 13

Convergence:
[1] 4.660965e-06
```

These results show that all the parameters estimates obtained using the Gammareg R-package are close to the true parameters values of the model. In all of cases with a small standard deviation and the 95% confidence interval contain the true parameter value.

6.2 Second simulation

In the second simulation, it is considered a gamma regression model with mean and shape structures, given by: In the second simulation study we considere the model

$$\begin{aligned} log(\mu_i) &= \mathbf{x_i'}\beta \\ log(\alpha_i) &= \mathbf{z_i'}\gamma \end{aligned}$$

In this case, $n = 500$ values of the explanatory variables X_1, X_2, X_3 and X_4 were generated as in the first simulation, but the values of the interest variable Y were generated from a Gamma distribution with $\mu_i = -5 + 2x_{2i} + 3x_{3i}$ and $\alpha_i = exp(0.2 + 0.1x_{2i} + 0.3x_4 i)$, as follow:

```
>library(Gammareg)
X1 <- rep(1,500)
X2 <- runif(500,0,30)
X3 <- runif(500,0,15)
X4 <- runif(500,10,20)
mui <- exp(-5 + 0.2*X2 -0.03*X3)
alphai <- exp(0.2 + 0.1*X2 + 0.3*X4)
Y <- rgamma(500,shape=alphai,scale=mui/alphai)
X <- cbind(X1,X2,X3)
Z <- cbind(X1,X2,X4)
formula.mean= Y~X2+X3
formula.shape= ~X2+X4
a=Gammareg(formula.mean,formula.shape,meanlink="log")
```

```
summary(a)
```

The results obtained from the application of the Gammareg R-package are:

```
            ######################################################################
            ###                  Classic Gamma Regression                      ###
            ######################################################################

Call:
Gammareg(formula1 = formula.mean, formula2 = formula.shape, meanlink = "log")

                   Estimate   L.Intv U.Intv
beta.(Intercept)  -4.98768 -4.99730 -4.978
beta.X2            0.19953  0.19915  0.200
beta.X3           -0.03021 -0.03083 -0.030
gamma.(Intercept)  0.33678  0.33610  0.337
gamma.X2           0.08687  0.08686  0.087
gamma.X4           0.30849  0.30846  0.309

 Covariance Matrix for Beta:
              (Intercept)            X2            X3
(Intercept)  2.397679e-05 -7.517510e-07 -8.062176e-07
X2          -7.517510e-07  3.746544e-08 -1.793592e-09
X3          -8.062176e-07 -1.793592e-09  1.016449e-07

 Covariance Matrix for Gamma:
              (Intercept)            X2            X4
(Intercept)  1.188862e-07 -8.320284e-10 -5.352878e-09
X2          -8.320284e-10  2.967992e-11  7.473412e-12
X4          -5.352878e-09  7.473412e-12  2.810451e-10

 AIC:
[1] -3469938

 Iteration:
[1] 20

 Convergence:
[1] 1.085736e-08
```

In this case,these results also show that all the parameters estimates obtained using the Gammareg R-package are close to the true parameters values of the model, except by γ_0, but in all cases the 95% confidence interval contain the true parameter value. The standard deviation are small for all parameters, except to γ_0.

7 Conclusions and Extensions

This paper is introduced the Gammareg R-package, that can be used to fit gamma regression models applying the classic method proposed in Cepeda-Cuervo (2001). We use two simulated studies to illustrate the use of the different functions of this package. In all of the cases, there is a speed convergence to the maximum likelihood estimation of the regression parameters, showing the efficient block-iterative alternate algorithm.

There are many possibilities to future works and practical tissues are possible. One is the use of alternative link functions, like the inverse link, that could adjust in a better way some database. Other is the develop a R-package to fit the gamma regression models proposed in Cepeda-Cuervo (2001), where both mean and variance follows regression structures. This are works in development.

References

Bateson, T. F. (2009), 'Gamma regression of interevent waiting times versus poisson regression of daily event counts: Inside the epidemiologist's toolboxselecting 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', *ESTADSTICA* **57**(168), 93–105.

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

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

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

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

Team, R. D. C. (n.d.), *A language and environment for statistical computing. R Foundation for Statistical Computing.*

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