WinBugs CODE For Beta Regression Models

EDILBERTO CEPEDA-CUERVO

Departamento de Estadística [1]

Universidad Nacional de Colombia

Summary

In this paper WinBugs code to fit joint mean and precision (variance) beta regression models is presented. These models are fitted applying Bayesian methodology and assuming normal prior distribution for the regression parameters. Analysis of structural data are included, assuming these models.

Key words: Beta regression, Bayesian methodology, WinBugs CODE

[1]email: ecepedac@unal.edu.co

1 Introduction

In this paper, we present the WinBugs CODE to fit beta regression models. The beta distribution defined in equation (1), has applications in uncertainty or random variation of a probability, fraction or prevalence, among others. Thus, this distribution has many applications in areas such as financial sciences or social sciences as education, where random variables are continuous in a bounded interval which is isomorphic to the interval $[0, 1]$. To mention an example, in studies of the quality of education, a number from 0 to 5 (or any other positive integer bounds) is assigned as a measure of performance for the evaluation of school subjects as math, language, arts, natural sciences or any other scholar area. In these cases, the measure assigned to each student can be expressed as a number from zero to one. Thus, it can be assumed that the level of student performance is a random variable with beta distribution.

The beta p, q distribution function, defined by equation (1) can be re-parametrized as a function of the mean and the so called dispersion parameter as in equation (4), or as function of the mean and variance taking into account equations (5) and (6). This characterization of the beta distribution can be more appropriate. In the first re-parametrization, making $\phi = p + q$ we may see that $p = \mu\phi$, $q = \phi(1 - \mu)$ and $\sigma^2 = \frac{\mu(1-\mu)}{\phi+1}$. In this case, ϕ can be interpreted as a precision parameter in the sense that, for fixed values of μ, larger values of ϕ correspond to smaller values of the variance of Y. This reparametrization presented in Ferrari and Cribari-Neto (2004), was already proposed in the literature, for example in Jorgensen (1997) or in Cepeda (2001, pg 63).

In this case, the mean and dispersion parameters can be modeled as functions of explanatory variables, given that behavior of these parameters can be explained explanatory variables. To cite a few examples, the educational level of mothers could influence students school performance; land concentration can be explained by random variables associated with social and political factors or the proportion of income spent monthly could be explained by the number of persons in the household. At the same time, we can assume that the dispersion parameter changes as a function of the same or other random variables. With these ideas, Bayesian regression, with joint modeling of the mean and dispersion parameters, was initially proposed by Cepeda (2001, pg. 63), under the framework of joint modeling in the biparametric exponential family (see Cepeda and Gamerman 2001, 2005). After that, Ferrari and Cribari-Neto (2004) proposed classical beta regression models, assuming that the dispersion parameter is constant through the rank of the explanatory variables. Further works have been published by Smithson and Verkuilen (2006), Simas et al. (2010) and, Cepeda-Cuervo and Achcar (2010), the latter proposing nonlinear beta regression in the context of Double Generalized Nonlinear Models. The beta regression models were extended in Cepeda et al.(2011), assuming that the observation are spatially correlated.

The rest of the paper is organized as follows: Section 2 includes general concepts on beta distribution. Section 3, the joint mean and precision (variance) beta regression models are defined. Section 4, provides the Winbugs CODE for Beta regression models.

2 Beta Distribution

A random variable Y has beta distribution if its density function is given by

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} I_{(0,1)}(y) \tag{1}$$

where $p > 0$, $q > 0$ and $\Gamma(.)$ denotes the gamma function. The mean and variance of Y, $\mu = E(Y)$ and $\sigma^2 = Var(Y)$, are given by

$$\mu = \frac{p}{p+q} \tag{2}$$

$$\sigma^2 = \frac{p\,q}{(p+q)^2(p+q+1)} \tag{3}$$

Many random variables can be assumed to have beta distribution. For example, income inequality or land distribution when measured using the Gini index proposed by Atkinson(1970), and the performance of students in subjects such as mathematics, natural sciences or literature. In the latter case, if performance X takes values within the interval (a, b), the random variable $Y = (X - a)/(b - a)$ can be assumed to have beta distribution. This performance can be explained by household socioeconomic variables, having fundamental impact on the student cognitive achievement. For example, the level of student achievement is closely related to the educational level of their parents and the number of hours devoted to study a subject. Thus, the beta regression model could be appropriate to explain the behavior of school performance as a function of associated factors. In these applications however, the reparametrization of the beta distribution given in (4) could be more appropriate. In the first, doing $\phi = p + q$ we can see that $p = \mu\phi$, $q = \phi(1 - \mu)$ and $\sigma^2 = \frac{\mu(1-\mu)}{\phi+1}$. Hence, ϕ can be interpreted as a precision

parameter in the sense that, for fixed values of μ, larger values of ϕ correspond to smaller values of the variance of Y. This reparametrizacion that is presented in Ferrari and Cribari-Neto (2004), had already appeared in the literature, for example in Jorgensen (1997) or in Cepeda (2001). With this reparametrization, the density of the beta distribution (1) can be rewritten as

$$f(y|\alpha,\beta) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} I_{(0,1)}(y) \tag{4}$$

In this case, the mean and dispersion parameters can be modeled as function of explanatory variables, for example, as was proposed in Cepeda(2001), given that changes in the precision parameter can be explained by explanatory variables, such as mothers educational level in the case of the student's school performance.

The beta distribution given in (1) can also be reparametrized as a function of the mean and variance, with

$$p = \frac{(1-\mu)\mu^2 - \mu\sigma^2}{\sigma^2} \tag{5}$$

$$q = \frac{(1-\mu)[\mu - \mu^2 - \sigma^2]}{\sigma^2} \tag{6}$$

Although writing (1) as a function of μ and σ^2 can result in a complex expression, joint modeling of the mean and variance can be easily achieved applying the Bayesian methodology proposed in Cepeda(2001), and Cepeda and Gamerman (2005). Sometimes, joint modeling of the mean and variance could be more appropriate than the joint modeling of the mean and the so

called dispersion parameter, given that parameters of the regression models would be more easily interpreted.

3 Joint Mean and Precision (Variance) Beta Regression Models

With the reparametrization of the beta distribution as a function of μ and ϕ, we can define a double generalized beta regression model as proposed in Cepeda (2001). In that research, joint modeling of the mean and dispersion parameters in the beta regression model and a Bayesian methodology to fit the parameters of the proposed model, was defined. Under a general framework, a random sample $Y_i \sim Beta(p_i, q_i)$, $i = 1, 2, \ldots, n$, was assumed, where both, mean and precision parameters, are modeled as a function of explanatory variables. That is,

$$\text{logit}(\mu) = \mathbf{x}_i^t \boldsymbol{\beta} \tag{7}$$

$$\log(\phi) = \mathbf{z}_i^t \boldsymbol{\gamma} \tag{8}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)$ are the vectors of the mean and dispersion regression models and, $\mathbf{x}_i$ and $\mathbf{z}_i$ are the vectors of the mean and dispersion explanatory variables, at the i-th observation, respectively http://www.bdigital.unal.edu.co/5947. Afther Cepeda's work, Ferrari and Cribari-Neto (2004) proposed the same reparametrization of the beta distribution, $\mu = p/(p+q)$ and $\phi = p + q$. In that paper, they assumed that $g(\mu_i) = \mathbf{x}_i^t \boldsymbol{\beta}$, where g is a strictly monotonic and twice differentiable real valued link function defined in the interval $(0, 1)$, assuming that the

dispersion parameter is constant. Although they consider many possible link functions, in the applications they take the logit link function, given that the mean can be interpreted as a function of the odds ratio. The joint mean and dispersion beta regression models proposed by Cepeda(2001), was later studied by Smithson and Verkuilen (2006) and Simas et al. (2010), under a classical perspective. At the same time, a nonlinear beta regression was proposed by Cepeda and Achcar (2010), assuming a nonlinear mean model given by (9) and a dispersion model given by (8), in the context of Double Generalized Nonlinear Models. This model was applied to the schooling rate data analysis in Colombia, for the period ranging from 1991 to 2003 http://www.tandfonline.com/doi/abs/10.1080/03610910903480784.

$$\mu_i = \frac{\beta_0}{1+\beta_1 \exp(\beta_2 x_i)} \tag{9}$$

In this paper, we propose joint mean and variance beta regression models, with the mean modeled as linear or nonlinear function of the parameters, as in (7) or (9), and the variance modeled as a function of the explanatory variables (10), where g is a monotonic and two time differentiable real function, that take into account the positivity of the variance http://www.bdigital.unal.edu.co/6207.

$$g(\sigma_i^2) = \mathbf{z_i}^t \boldsymbol{\gamma} \tag{10}$$

The results of fitting the mean and variance beta regression models are easily interpretable: the mean fitted models have the usual interpretation, but the fitted variance model is easily interpreted directly from data behavior. For example, if the explanatory variable Z_1 is associated to γ_1 and $\gamma_1 > 0$,

increasing behavior of Z_1 is associated with increasing behavior of σ^2. In the same way, the interpretation is applicable when the parameters of the variance models are negative.

In the next sections, structured and real data sets are analyzed applying joint mean and dispersion, and joint mean and dispersion beta regression models to compare the performance of these models, according to the behavior of the data.

4 WinBugs CODE for beta regression

4.1 Joint mean and precision beta regression model

http://www.bdigital.unal.edu.co/5947

```
model
    {
        for( i in 1 : N ) {
            Y[i] ~ dbeta(p[i],q[i])
            p[i]<-mu[i]*tau[i]
            q[i]<-tau[i]-mu[i]*tau[i]
            logit(mu[i])<-b0+  b1*x1[i]+b2*x2[i]
            tau[i] <-exp(c0+ c1*x2[i])

                        }
        b0 ~ dnorm(0.0,1.0E-2)
        b1 ~ dnorm(0.0,1.0E-2)
```

```
        b2 ~ dnorm(0.0,1.0E-2)
        c0 ~ dnorm(0.0,1.0E-2)
        c1 ~ dnorm(0.0,1.0E-2)
            }
```

```
Data
 list(Y=c(0.53, 0.69,0.47,0.87,0.82,0.68,0.53,0.56,
 0.74,0.53,0.06,0.58,0.80,0.64,0.58,0.24),
x1=c(0.441,0.356,0.466,0.177,0.355,0.447,0.591,
0.602,0.326,0.477,0.436,0.345,0.173,0.549,0.156,0.668),
x2=c(0.188,0.428,0.09,0.413,0.312,0.185,0.283,0.362,
0.3918,0.131,0.411,0.336,0.272,0.204,0.169,0.188),
 N= 16)
Inits

list(b0=-1,  b1=-1, b2=0, c0=-3, c1=0)
```

4.2 Joint mean and variance beta regression model

http://www.bdigital.unal.edu.co/6207

```
model
{
for( i in 1 : N ) {
Y[i] ~ dbeta(a[i],b[i])
```

```
a[i]<-((1-mu[i])*mu[i]*mu[i]- mu[i]*sg[i])/sg[i]
b[i]<-(1-mu[i])*(mu[i]-mu[i]*mu[i]-sg[i])/sg[i]
logit(mu[i])  <-b0+ b1*x1[i]
sg[i] <-exp(c0+c1*x1[i])
}
b0 ~ dnorm(0,1)
b1 ~ dnorm(0,0.1)
c0 ~ dnorm(0,0.01)
c1 ~ dnorm(0,0.2)
}
Data
list(Y=c(0.53, 0.69,0.47,0.87,0.82,0.68,0.53,0.56,
0.74,0.53,0.06,0.58,0.80,0.64,0.58,0.24),
x1=c(0.441,0.356,0.466,0.177,0.355,0.447,0.591,
0.602,0.326,0.477,0.436,0.345,0.173,0.549,0.156,0.668),
N= 16)

Inits
list(b0=0,  b1=0,   c0=-3, c1=0)
```

4.3 Nonlinear beta regression modes

http://www.tandfonline.com/doi/abs/10.1080/03610910903480784

```
model
{
for( i in 1 : N ) {
```

```
Y[i] ~ dbeta(a[i],b[i])
a[i]<-mu[i]*tau[i]
b[i]<-tau[i]-mu[i]*tau[i]
mu[i] <-b0/(1+ b1*exp(b2*x1[i]))
tau[i] <-exp(c0+c1*x1[i])
}

b0 ~ dunif(0,1)
b1 ~ dnorm(0,0.1)
b2 ~ dunif(-6, 0)
c0 ~ dnorm(-10,0.01)
c1 ~ dnorm(0,0.2)
}
Data
 list(Y=c(0.2,0.3,0.45,0.55,0.6,0.64,0.7,0.75,0.85,0.9),
x1=c(1,2,3,4,5,6,7,8,9,10), N= 10)

Inits
list(b0=0.88,  b1=3, b2=-2,  c0=-9, c1=0)
```

References

[1] Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*,**2**, 244-263.

[2] Cepeda, E.C. (2001). Variability Modeling in Generalized Linear Models,

Unpublished Ph.D. Thesis. Mathematics Institute, Universidade Federal do Rio de Janeiro.

[3] Cepeda C. E. and Gamerman D. (2001). Bayesian modeling of variance heterogeneity in normal regression models . *Brazilian Journal of Probability and Statistics*, **14**, 207-221.

[4] Cepeda C. E. and Gamerman D. (2005). Bayesian methodology for modeling parameters in the two parameter exponential family. *Estadística*, **57**, 93-105.

[5] Cepeda C. E. and Garrido L. (2012) Bayesian beta regression models: joint mean and precision modeling. *Biblioteca Digital U.N. Repositorio Institucional.* http://www.bdigital.unal.edu.co/5947

[6] Cepeda C. E. (2012) Beta Regression Models: Joint Mean and Variance Modeling. *Biblioteca Digital U.N. Repositorio Institucional.* http://www.bdigital.unal.edu.co/6207

[7] Cepeda-Cuervo, E. and Achcar J. A. (2010). Heteroscedastic Nonlinear Regression Models *Communications in Statistics - Simulation and Computation*, **39**(2):405-419.

[8] Cepeda-cuervo, E., Nuñez-Anton V. (2011). Generalized Econometric Spatial Models. Commun. Stat., Simulation Comput. 39, No. 2, 405-419.

[9] Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions, *Journal of Applied Statistics* **31**, 799-815.

[10] Jorgensen, B. (1997). Proper dispersion models (with discussion). *Brazilian Journal of Probability and Statistics*, **11**, 89-140.

[11] Simas A. B., Barreto-Souza W., Rocha A. V. (2010). Improved Estimators for a General Class of Beta Regression Models, *Computational Statistics & Data Analysis*, **54**(2), 348-366.

[12] Smithson M., Verkuilen J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. *Psychological Methods*, **11**(1),54-71.

[13] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, **64**4, 583-639.