CUSTER-WEIGHTED BETA REGRESSION

M. Alfó[1], L. Nieddu[2] and C. Vitiello[1]

[1] Department of Statistics, Sapienza University of Rome, (e-mail: marco.alfo@uniroma1.it, cecilia.vitiello@uniroma1.it)

[2] UNINT University - Rome, (e-mail: l.nieddu@unint.eu)

ABSTRACT: Beta regression is the standard approach to model a limited dependent variable with range in the unit interval. The model requires two systems of equations, for the mean and the precision parameter, that can be based on the same set of covariates. Nevertheless, the simple linear model for the precision parameter could not be good enough to capture all the heterogeneity in the data. We extend the approach of cluster weighted linear and generalized linear models to beta regression in order to obtain a more flexible model.

KEYWORDS: Cluter weighted regression, beta regression, finite mixtures.

1 Introduction

The analysis of the effect of a set of covariates on the conditional expectation of variables assuming values in the open unit interval $(0,1)$ is a common issue in quantitative research, as limited dependent variables arise in several cases, such as when we look at relative frequencies of an event, (eg proportion of votes gained by a candidate at an election), fractions of a continuous variable (eg the amount of GDP due to a particular economic sector), performance ratings (eg student performances), or limited-valued indeces (eg the Gini index). Various alternative methods for limited value responses have been proposed (see e.g. Papke & Wooldridge, 1996 and Kieschnick & McCullough, 2003), and a naive solution is to transform the dependent variable using an appropriate probit link and then model the transformed response using linear regression (Demsetz & Lehn, 1985). This approach however does not allow for straightforward interpretation of the parameter estimates, and typical OLS regression assumptions are often not met by the transformed data.

A viable alternative is to use a regression model based on the assumption that the response variable follows a conditional Beta distribution $Y \mid \mathbf{x} \sim \mathcal{B}(p,q)$. For modeling purposes, Ferrari & Cribari-Neto, 2004 have proposed an alternative parameterization:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad \mu \in (0,1), \phi > 0$$

where $\mathbb{E}(Y) = \mu = \frac{p}{p+q}$, $Var(Y) = \frac{\mu(1-\mu)}{1+\phi}$ and $\phi = p+q$ represents a *prevision* parameter. The variance of a Beta r.v. is a scaled version of the binomial variance and the precision parameter offers flexibility allowing the Beta distribution to assume different shapes. Model parameter estimates can be obtained using a maximum likelihood (ML) approach (see Ferrari & Cribari-Neto, 2004; Cribari-Neto & Zeileis, 2010). A further parameterization, which can be useful especially in the context we are interested in, has been proposed by Chen, 1999, see also Bagnato & Punzo, 2013.

The aim of this paper is to extend homogeneous beta regression models in the cluster weighted modeling context introduced by Gershenfeld, 1997 and suggest a methodology to capture individual specific unobserved heterogeneity. In this sense, it extends the cluster-weighted GLM framework described by Ingrassia *et al.*, 2015 and Punzo & Ingrassia, 2016.

The structure of the manuscript is as follows. In the next Section the proposed model will be introduced and the EM algorithm for ML estimation of the model parameters will be sketched. In Section 3 the model will be applied to a real dataset. Some conclusions will be drawn in Section 4.

2 The model

Let $(Y, \mathbf{X})$ be a set of random variables (a random response variable Y and a random covariates vector $\mathbf{X}$) with joint density $f(y, \mathbf{x})$, and let us assume that corresponding population can be partitioned into K subpopulations, which in the present context will be referred to as components. Let $\pi_k(\mathbf{x}_i)$ be the prior probability that unit i belongs to component $k = 1, \ldots, K$, and the component indicator $z_{ik} = 1$ if unit i belongs to subset k. The usual assumption is that $\pi_k(\mathbf{x}_i) = \pi_k \, \forall k$ (*assignment independence*).

A more flexible family of mixture models can be obtained releasing the assumption of assignment independence. This approach has been introduced by Gershenfeld, 1997 and is known as cluster-weighted models (CWM), i.e.:

$$f(y_i, \mathbf{x}_i) = \sum_{k=1}^{K} \pi_k g(\mathbf{x}_i \mid z_{ik} = 1) f(y_i \mid \mathbf{x}_i, z_{ik} = 1) \quad . \tag{1}$$

Here, $f(y_i \mid \mathbf{x}_i, z_{ik}=1)$ is the density of the response variable conditional on the set of covariates and the component the unit belongs to, while $g(\mathbf{x}_i \mid z_{ik}=1)$ is the distribution of the covariates in the specific component $k=1,\ldots,K$.

Cluster weighted regression models constitute a flexible family of models to fit the joint density of a set of covariates and a response variable coming from a heterogeneous population, see Ingrassia *et al.*, 2015 and Punzo & Ingrassia, 2016 for cluster-weighted generalized linear models. When the covariates are continuous, it is customary to adopt a multivariate Gaussian density in each component, that is $\mathbf{X} \mid z_{ik}=1 \sim MVN(\upsilon_k, \Sigma_k)$, while for the response a component-specific eta distribution holds $Y \mid \mathbf{X}, z_{ik}=1 \sim \mathcal{B}(\mu_k, \phi_k)$. The location and the dispersion parameter can be linked to the linear predictors as follows:

$$h_1(\mu_i) = \eta_k = \mathbf{x}_i^{'}\beta_k \tag{2}$$

$$h_2(\phi_i) = \xi_k = \mathbf{x}_i^{'}\gamma_k \tag{3}$$

The link functions $h_1(\cdot)$ and $h_2(\cdot)$ for the expected values and the dispersion parameter of the conditional distribution of the response are monotonic and twice differentiable. Suitable candidates for $h_1(\cdot)$ are logit, probit or complementary log-log while for $h_2(\cdot)$ the log or the square root function are usually employed.

The observed data likelihood function for the proposed model is defined by

$$L(\psi) = \prod_{i=1}^{n}\sum_{k=i}^{K} f(y_i \mid \mathbf{x}_i, \beta_k, \gamma_k) g(\mathbf{x}_i \mid \upsilon_k, \Sigma_k)\pi_k$$

where $\theta_k = (\beta_k, \gamma_k)$ is the vector of regression parameters, $\omega_k = (\upsilon_k, \Sigma_k)$, $k = 1,\ldots,K$, represents the set of parameters for the conditional covariates distribution and $\psi = (\theta_1,\ldots,\theta_K,\omega_1,\ldots,\omega_K)$ denotes the global set of parameters. Let $\ell_c(\psi)$ denote the loglikelihood function for the complete data $(\mathbf{y},\mathbf{X},\mathbf{z})$; for fixed K at each iteration, the E step of the EM algorithm is based on the expected value of the loglikelihood function for complete data, based on the observed data and the current parameters estimates, referred to as $Q(\psi \mid \psi^r)$. In the M step of the algorithm, this function is maximized with respect to ψ. The general scheme of the algorithm is given below:

E-step $Q(\psi \mid \psi^{\mathbf{r}}) = \mathbb{E}_{\psi^{(r)},\pi^{(r)}}(\ell_c(\psi \mid \mathbf{X}, y))$
M-step maximize $Q(\psi \mid \psi^{\mathbf{r}})$ w.r.t. (ψ,π) conditional on the weights obtained at the E-step

In the E-step the z_{ik} will be replaced by their conditional expectation w_{ik} $i = 1,\ldots,n, k = 1,\ldots,K$. ML equations for the parameters of the Beta model and the Gaussian density are given by the following expressions

$$\frac{\partial \ell(\cdot)}{\partial \theta_k} = \sum_{i=1}^{n} w_{ik} \frac{\partial log(f(y_i \mid \mathbf{x}_i, \theta_k))}{\partial \theta_k} = 0$$

$$\frac{\partial \ell(\cdot)}{\partial \omega_k} = \sum_{i=1}^{n} w_{ik} \frac{\partial log(g(\mathbf{x}_i \mid \omega_k))}{\delta \omega_k} = 0$$

Conditional to:

$$w_{ik} = \frac{f(y_i \mid \mathbf{x}_i, \theta_k) g(\mathbf{x}_i \mid \omega_k) \pi_k}{\sum_{l=1}^{K} f(y_i \mid \mathbf{x}_i, \theta_l) g(\mathbf{x}_i \mid \omega_l) \pi_l}$$

Both expressions are weighted score equations with weights given by the posterior probabilities w_{ik} that unit i belongs to component k which is common to both the response (conditional) and the covariate distribution.

This leads to standard results for the estimates of parameters υ_k and Σ_k in the Gaussian density:

$$\hat{\upsilon}_k = \frac{\sum_{i=1}^{n} \mathbf{x}_i w_{ik}}{\sum_{i=1}^{n} w_{ik}} \quad \text{and} \quad \hat{\Sigma}_{\mathrm{k}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{ik}} (\mathbf{x}_{\mathrm{i}} - \hat{\upsilon}_{\mathrm{k}})(\mathbf{x}_{\mathrm{i}} - \hat{\upsilon}_{\mathrm{k}})'}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{ik}}} \quad ,$$

while the prior probability estimates can be obtained by solving the following constrained ML problem:

$$\begin{cases} \frac{\partial \ell(\cdot)}{\partial \pi_k} = \sum_{i=1}^{n} \frac{f(y_i|\theta_k) g(\mathbf{X}_i|\omega_k)}{\sum_{k=1}^{K} f(y_i|\theta_k) g(\mathbf{X}_i|\omega_k) \pi_k} + \lambda = 0 \\ \\ \sum_{k=1}^{K} \pi_k = 1 \end{cases} ,$$

leading to $\hat{\pi}_k = \sum_{i=1}^{n} w_{ik}/n$ which is a well known result in finite mixtures.

3 Illustrative example

To give a flavor of the proposed model, acceptance rates in American colleges have been chosen as response variable using instate tuition fees and student/faculty ratio as covariates. The dataset comes from the 1995 U.S. News & World Report's Guide to America's Best Colleges (http://lib.stat.cmu.edu/datasets/colleges/). Out of 1300 colleges, only those records with complete responses and covariates were retained. A binary variable on the legal nature of the institution

Table 1. *MLE for cluster weighted Beta regression model with K=5 components.(location model and precision model)*

	Location Model			Precision Model		
Group	Intercept	X1	X2	Intercept	X1	X2
1	0.4944	0.0001	n.s.	-1.8959	0.0005	n.s.
2	0.8280	n.s.	n.s.	3.5407	-0.0001	-0.0557
3	n.s.	n.s.	0.1212	n.s.	n.s.	n.s.
4	0.9844	n.s.	n.s	n.s.	0.0003	n.s
5	n.s	-0.0001	0.1031	n.s	n.s	n.s

(private/public) and units identification did not contribute to the analysis but gave support in the interpretation. The EM algorithm has been run by $m = 50$ different starting values for each value of $K = 1, \ldots, 10$. For each value of K, the model with the best likelihood has been retained and the BIC Schwarz, 1978 was used to determine the optimal number of components, resulting in the choice of $K = 5$. In table 1 the parameter estimates for the location and dispersion models are reported.

"Instate tuition" (X1) affects the response only in components 1 and 5 while X2 ("student/faculty ratio") has some role in components 3 and 5. Component 1 and 5 mainly include private schools with component 5 associated to very expensive institutions. This last component shows a negative coefficient, probably due to a self selection of the applicants. This is not the case for component 1, where increasing fees correspond to higher acceptance rates. The effect of variable X2 results significant only in components 3 and 5. This variable can considered the structural capacity of the institution to enroll students. A comparison with the standard finite mixture of Beta regressions, see Ferrari & Cribari-Neto, 2004, via the `betareg` package, Grün *et al.*, 2012, will also be discussed.

4 Conclusions

We have proposed a cluster weighted beta regression approach modeling both the location and the precision parameter for a response with a conditional beta distribution and a multivariate Gaussian set of observed covariates. Our proposal embeds the finite-mixture of Beta regressions as a particular case, when the distribution of the observed covariates does not depend on the component membership indicator.

References

BAGNATO, L., & PUNZO, A. 2013. Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. *Computational Statistics*, **28**, 1571–1597.

CHEN, S. X. 1999. Beta kernel estimators for density functions. *Computational Statistics Data Analysis*, **31**, 131–145.

CRIBARI-NETO, F., & ZEILEIS, A. 2010. Beta Regression in R. *Journal of Statistical Software*, **34**(2), 1–24.

DEMSETZ, H., & LEHN, K. 1985. The Structure of Corporate Ownership: Causes and Consequences. *Journal of Political Economy*, **93**(6), 1155–1177.

FERRARI, S., & CRIBARI-NETO, F. 2004. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

GERSHENFELD, N. 1997. Nonlinear inference and Cluster-Wieghted Modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.

GRÜN, B., KOSMIDIS, I., & ZEILEIS, A. 2012. Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software*, **48**, 1–25.

INGRASSIA, S., PUNZO, A., VITTADINI, G., & MINOTTI, S. C. 2015. The generalized linear mixed cluster-weighted model. *Journal of Classification*, **32**, 85–113.

KIESCHNICK, R., & MCCULLOUGH, B.D. 2003. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling*, **3**(3), 193–213.

PAPKE, L., & WOOLDRIDGE, J. 1996. Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates. *Journal of Applied Econometrics*, **11**(6), 619–632.

PUNZO, A., & INGRASSIA, S. 2016. Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, **31**, 989–1013.

SCHWARZ, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.