# CLUSTER WEIGHTED BETA REGRESSION

Luciano Nieddu, Cecilia Vitiello

## 1. Introduction

The analysis of data assuming values in the real open interval (0;1) is a common issue in quantitative research when the effect of selected variables on the conditional expectation of a percentage or rate is considered.

In the literature, various alternative methods to model ratios and percentage data have been proposed (see e.g. Papke and Wooldridge, 1996 and Kieschnick and McCullough, 2003). A possible solution is to transform the dependent variable y, for instance using a logit or a probit transformation, so that it assumes values on the whole real line, and then model the mean of the transformed response as a linear predictor based on a set of covariates applying OLS (Demsez Lehn, 1985) to obtain the parameter estimates. This approach, however, has drawbacks, one of them being the fact that the model parameters cannot be easily interpreted in terms of the average of the original outcome but in terms of the transformed response. Furthermore the assumptions of OLS regression are often not met despite the transformation of the data.

An alternative is to use a regression model that assumes that the response variable follows a beta distribution on the interval (0;1), namely $Y|\mathbf{x} \sim \mathcal{B}(p,q)$:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad y \in (0,1), \; p,q > 0, \tag{1}$$

with $\mathbb{E}(Y) = \frac{p}{p+q}$ and $Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}$.

An alternative parameterization, more convenient for modeling purposes, is the one introduced by Ferrari and Cribari-Neto (2004) were the parameters are expressed in terms of the mean $\mu$ and the precision parameter $\varphi$:

$$\mathbb{E}(Y) = \mu \qquad Var(Y) = \frac{\mu(1-\mu)}{1+\varphi}$$

$$f(y;\mu,\varphi) = \frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma((1-\mu)\varphi)} y^{\mu\varphi-1}(1-y)^{(1-\mu)\varphi-1}, \quad y \in (0,1) \tag{2}$$

The variance of a beta-distributed random variable is a scaled version of the binomial variance and the precision parameter allows for a wide range of shapes for the density.

Estimates of the model parameters can conveniently be obtained using maximum likelihood (ML) estimation (Ferrari and Cribari-Neto, 2004)

Classical beta regression models conveniently account for over dispersion by including a precision parameter $\varphi$ to adjust the conditional variance of the outcome. On the other hand, it is often observed that over dispersion depends on the values of one or more predictor variables (Simas et al., 2010). In this case it is possible to extend the beta regression model by regressing the precision parameter on a subset of the predictor variables.

The aim of this paper is to extend the classical framework of beta regression and suggest a methodology that can help capture unobserved heterogeneity between observations that follow a beta distribution using a cluster weighted modeling approach introduced by Gershenfeld (1997).

In the next Section the proposed model will be introduced and ML estimates for the parameters will be obtained. In Section 3 the model will be applied on a real dataset and in Section 4 some conclusions will be drawn.

## 2. The Model

Finite mixtures of linear regressions are sometime inadequate for some applications (Hennig, 2000), since they assume *assignment independence*, i.e. the prior probability for single unit ($y$, $\mathbf{x}$) generated by one of the components of the mixture is constant over all possible values of the vector of covariates $\mathbf{x}$.

Let ($y$,$\mathbf{x}$) be a set of random variables (a random response variable $y$ and a random vector $\mathbf{x}$) with join density $f(y, \mathbf{x})$. Let's further assume that the support of $\mathbf{x}$ can be partitioned into $K$ subsets.

A more flexible family of mixture models can be obtained assuming that the prior probability for a unit to belong to a cluster depends on the value of the vector of covariates $\mathbf{x}$. This approach was introduced by Gershenfeld (1997) and is known as cluster-weighted models (CWMs), i.e.:

$$f(y, \mathbf{x}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}|k) g(y|\mathbf{x}, k) \tag{3}$$

where $g(y|\mathbf{x}, k)$ is the conditional density of the response variable given the set of covariates and the group the unit belongs to; $f(\mathbf{x}|k)$ is the distribution of the covariates given the group and $\pi_k$ is the prior probability of a unit to belong to group $k$.

Cluster weighted regression models constitutes a flexible family of models to fit the joint density of a set of covariates and a response variable assuming that they are coming from a heterogeneous population.

We will assume that $X|k \sim MVN(\mu_k, \Sigma_k)$ and $Y|X, k \sim \mathcal{B}(\mu'_k, \varphi_k)$.
The location and the dispersion parameter can be linked to the linear predictors as follows:

$g_1(\mu'_k) = \eta_i = \mathbf{x}_i\beta$
$g_2(\varphi_i) = \xi_i = \mathbf{z}_i\gamma$

The functions $g_1(\cdot)$ and $g_2(\cdot)$ are monotonic link functions. Suitable candidates are respectively logit and probit.
The likelihood function for the proposed model is

$$L(\cdot) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \{ f(y_i|\mathbf{x}_i\beta_k\gamma_k) g(\mathbf{x}_i|\mu_k, \Sigma_k) \} \tag{4}$$

let $\theta_k = (\beta_k, \gamma_k)$ and $\omega_k = (\mu_k, \Sigma_k)$, $k=1,\ldots, K$ .

ML equations for the parameters of the Beta model:

$$\begin{cases} \dfrac{\delta \ell(\cdot)}{\delta \theta_k} = \sum_{i=1}^{n} w_{i,k} \dfrac{\delta log(f(y_i|\mathbf{x}_i\theta_k))}{\delta \theta_k} = 0 \\ w_{ik} = \dfrac{\pi_k f(y_i|\theta_k) g(\mathbf{x}_i|\omega_k)}{\sum_{k=1}^{K} \pi_k f(y_i|\theta_k) g(\mathbf{x}_i|\omega_k)} \end{cases}$$

and ML equations for the parameters of the Gaussian process:

$$\begin{cases} \dfrac{\delta \ell(\cdot)}{\delta \omega_k} = \sum_{i=1}^{n} w_{i,k} \dfrac{\delta log(g(x_i|\omega_k))}{\delta \omega_k} = 0 \\ \\ w_{ik} = \dfrac{\pi_k f(y_i|\theta_k) g(\mathbf{x}_i|\omega_k)}{\sum_{k=1}^{K} \pi_k f(y_i|\theta_k) g(\mathbf{x}_i|\omega_k)} \end{cases}$$

are both weighted score equations with weights given by the a posterior probabilities $w_{i,k}$ of unit *i* to belong to component *k*.
This yields to standard results for the estimates of the parameters $\mu_k$ and $\Sigma_k$:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \mathbf{x}_i w_{ik}}{\sum_{i=1}^{n} w_{ik}} \quad \hat{\Sigma}_k = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T w_{ik}}{\sum_{i=1}^{n} w_{ik}}$$

while estimates for the a priori probabilities can be obtained solving the following constrained ML problem:

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

yielding: $\widehat{\pi_k} = \sum_{i=1}^{n} w_{i,k}$.

## 3. Real data example

The U.S. News data contains information on tuition, room and board costs, SAT or ACT scores, application/acceptance rates, graduation rate, student/faculty ratio, spending per student, and a number of other variables for a total of 35 categorical and quantitative variables over a sample of more than 1300 schools.

The dataset is taken from the 1995 U.S. News & World Report's Guide to America's Best Colleges and is freely available from the statlib repository (http://lib.stat.cmu.edu/datasets/colleges/). Most of the data are for the 1993-94 school year. Two third of the schools are private (65.19%).

The rate of accepted applicants has been considered as response variable and "instate tuition" ($X_1$) and "sfratio" (student/faculty ratio" $X_2$) have been used as covariates. Only records with no missing data have been considered.

The proposed model has been fit to the data using BIC to select the optimal number of components. The best value of BIC was obtained in correspondence of K=5.

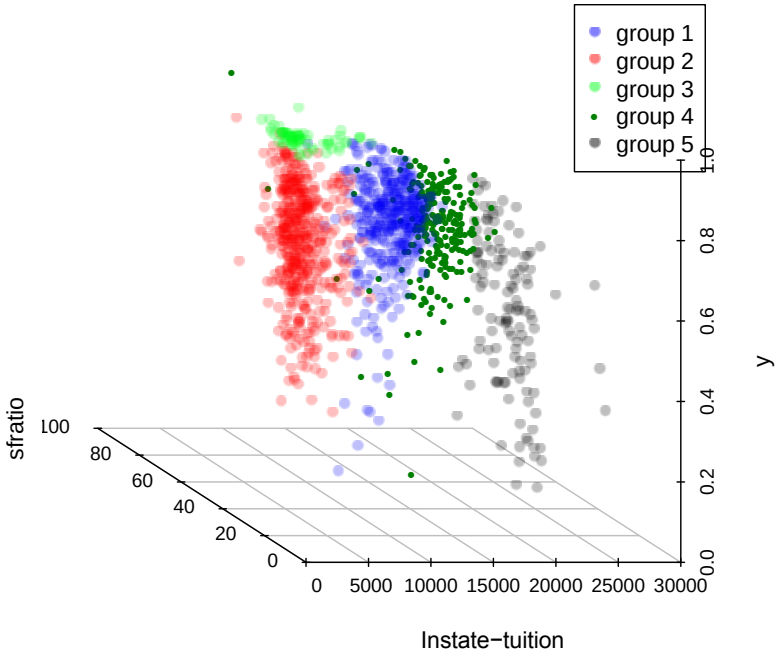
The results of the estimates for location and precision models for the Beta distribution have been reported in Table 1. Only the values that, at a confidence level of $\alpha = 0.05$, were significantly different from zero have been retained.

**Table 1 –** *MLE for cluster weighted Beta regression model with K=5 components.*

| Group | Estimates | Intercept | X1 | X2 |
|---|---|---|---|---|
| 1 | $\beta$ | 0.4944 | 0.0001 | n.s. |
| | $\gamma$ | -1.8959 | 0.0005 | n.s. |
| 2 | $\beta$ | 0.8280 | n.s. | n.s. |
| | $\gamma$ | 3.5407 | -0.0001 | -0.0557 |
| 3 | $\beta$ | n.s. | n.s. | 0.1212 |
| | $\gamma$ | n.s. | n.s. | n.s. |
| 4 | $\beta$ | 0.9844 | n.s. | n.s. |
| | $\gamma$ | n.s. | 0.0003 | n.s. |
| 5 | $\beta$ | n.s. | -0.0001 | 0.1031 |
| | $\gamma$ | n.s. | n.s. | n.s. |

In Figure 1 a 3d-plot of the 5 groups has been displayed to easy the interpretation of the results while in Table 2 the distribution of the schools by group and type (public/private) has been reported.

**Figure 1 –** *Cluster structure with respect to acceptance rate, instate-tuition and sfratio.*




Considering the results for the location parameter, instate-tuition is influential for the rate of acceptance in groups 1 and 5, while student-to-faculty ratio is only influential in Group 3. Group 1 and 5 are mainly private schools (Table 2). Group 5 is made of very expensive and very well known universities, and shows negative coefficient for the variable "instate-tuition": for those famous highly qualified universities, high tuitions means being able to apply a very strict selection of the applicants. In Group 1 we find mainly private colleges with different vocations, for them an increase in tuition increases the proportion of accepted applicants. Therefore instate-tuition are effective for these two groups, representing for one a measure of the selectiveness of the university (Group 5) for the other (Group 1) a measure of the quality of the college (private colleges with very small instate-tuitions could be considered just a way to get a degree).

Variable $X_2$ is influential only on groups 3 and 5. Student-to-faculty ratio can be considered a structural variable indicating the dimension of the school and therefore its capacity to accept students. An increase of $X_2$ in Group 5 increases the

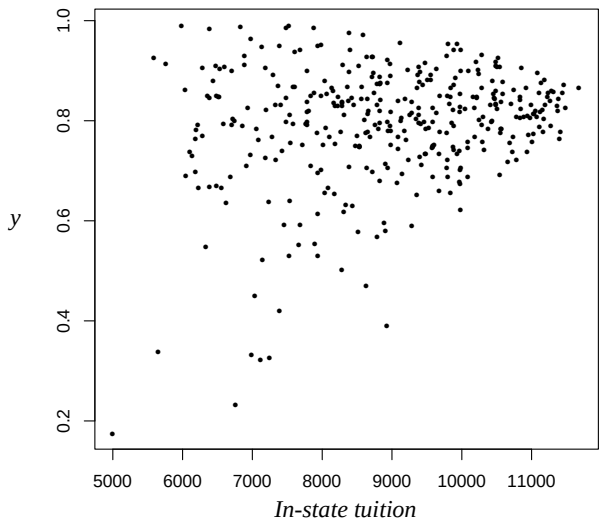
proportion of accepted applicants (the university accepts students up to its structural capacity identified by the predefined student to faculty ratio set by the board of directors). This is valid also for Group 3, where it is the only variable affecting the proportion of applicants. Group 3 is a highly heterogeneous group of schools (Table 2) with very low instate-tuitions and a very high acceptance rate (Figure 1).

**Table 2 –** *Distribution of schools by group and type (public/ private).*

| Group | Public | Private | **Total** |
|---|---|---|---|
| 1 | 1 | 370 | **371** |
| 2 | 389 | 63 | **452** |
| 3 | 48 | 20 | **68** |
| 4 | 1 | 244 | **245** |
| 5 | 0 | 125 | **125** |
| **Total** | **439** | **822** | **1261** |

Let's now consider the effect of those variables on the precision parameter. To better understand the effect of the precision parameter let's consider Group 1 where the precision parameter depends only on instate-tuition and its effect on the precision of the estimate of the response variable may be shown in a two-dimensional plot. In Figure 2 a plot of the outcome variable and the instate-tuition for all colleges of group 1 has been displayed.

**Figure 2 –** *Effect of covariates on precision: instate-tuition vs acceptance rate*



The triangular shape of the cloud implies that there is large variability of the outcome y for small values of the covariate and the variability decreases as the covariate "instate tuition" increases. Therefore the precision of the estimates for *y* is greater for higher values of instate-tuition. Group 2 and Group 4 have two diverging behaviors: they both have an admission rate which does not depends on $X_1$ and $X_2$, with the admission rate of Group 4 being slightly higher than that of Group 2. In Group 4 the precision of the estimates increases with instate-tuitions and does not depend on $X_{2,}$ while in Group 2 it decreases with instate-tuition and with $X_2$.

## 4. Conclusions

We have proposed a Beta regression model based on CWRs that allows for flexibility on modeling both the location and the precision parameter for the beta distribution. Our proposal, which should include the finite-mixture approach as a particular case, not only can be used in presence of over dispersed data but it can also be used as a diagnostic tool to detect a mixture structure in the data. The proposed methodology has been tested on benchmark data yielding very interesting results.

## Riferimenti bibliografici

DEMSETZ H., LEHN K. 1985. The structure of corporate ownership: causes and consequences. *Journal of Political Economy*, 93, 1155–77.

FERRARI S.L.P., CRIBARI-NETO F. 2004. Beta regression for modeling rates and proportions. *Journal of Applied Statistics* 31: 799–815.

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

HENNIG C. 2000. Identifiablity of models for clusterwise linear regression. *Journal of Classification* 17, 273–296.

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

PAPKE L.E., WOOLDRIDGE J.M. 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11: 619–632.

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: 348–366.

## SUMMARY

### Cluster Weighted Beta Regression

Beta regression is the standard method to explore how a response assuming values in (0;1) depends on a set of covariates. With respect to standard regression, in this case, the parametric model requires two systems of equations: one for the mean and the other for the precision parameter that can be based on the same set of covariates.

Therefore for two different sets of covariates and the same value of the linear predictor for the mean we could have different precisions.

Nevertheless a linear model for the precision parameter could not be good enough to capture all the heterogeneity in the data.

We will extend the characteristic approach of cluster weighted linear models to the beta regression problem in order to obtain a flexible model both in analyzing relations between means and covariates and in evaluating prediction precision.



Luciano NIEDDU, UNINT, l.nieddu@unint.eu

Cecilia VITIELLO, Department of Statistics, “Sapienza” University of Rome, cecilia.vitiello@uniroma1.it