A restricted variable dispersion beta regression model: frequentist approach

Luis F. Grajales H.*
Universidad Nacional de Colombia, Bogota, Colombia - lfgrajalesh@unal.edu.co

Oscar O. Melo M.
Universidad Nacional de Colombia, Bogota, Colombia - oomelom@unal.edu.co
Luis A. Lopez P.
Universidad Nacional de Colombia, Bogota, Colombia - lalopezp@unal.edu.co

Abstract

The variable dispersion beta regression model (VDBRM) is useful when the response is measured continuously in the $(0, 1)$ interval (if $y \in (a, b)$, then $\frac{y-a}{b-a} \in (0, 1)$). In this work, in order to take account of constraints on parameters, a restricted variable dispersion beta regression model is proposed, developed, and applied from a frequentist perspective. i) when there are not restrictions, our model coincides with the *variable dispersion BRM*; ii) if there are not restrictions and the dispersion parameter ϕ is assumed constant across observations, our model is the *simple BRM*. First, a penalized likelihood function is proposed, using Lagrange multipliers for restrictions. Second, the restricted maximum likelihood estimators are obtained. Third, the respective inferential analysis is done: hypothesis tests for restrictions and goodness of fit for models. Good results were obtained for simulated and real data. Comparisons with normal and transformed model are done. Also, some Bayesian explorations are presented from an integrated Bayesian/likelihood framework, using flat prior distributions..

Keywords: Restricted beta regression model; variable dispersion beta regression model; fractional factorial experiments.

1. Introduction.

1 Introduction.

Factorial experiments are widely used in industry, engineering, the sciences, and product and process improvement. The 2^k experiments contain k factors at two-levels each, and several factors are evaluated simultaneously. When all possible factors cannot be evaluated, it is usual to perform only a fraction of the complete experiment, which leads to a 2^{k-p} fractional factorial experiment. The most common interest in these experiments is to identify the subset of the factors that has the greatest effect on the response, y. With respect to the response, this work focuses on experiments whose response is measured continuously in the $(0, 1)$ interval (if $y \in (a, b)$, then $\frac{y-a}{b-a} \in (0, 1)$). A lot of empirical data with these conditions are encountered in the statistical literature. Data analysis from factorial experiments by means of three approaches, linear normal model, data transformations, and GLM can be encountered in the literature. Now, considering that y lies in (0,1) and the beta distribution assumes values there, the beta regression model (BRM) might be an option so as to improve analysis or to offer additional comparisons with regard to the three mentioned approaches.
The possibilities of the BRM in factorial and fractional factorial experiments are illustrated afterwards by two examples and one discussion: 1) the simple BRM; 2) the variable dispersion BRM (VDBRM); and 3) the restricted VDBRM.

Example 1. *Justifying the simple BRM in a* 2^4 *experiment. Half-normal plots.*

The *Drill* experiment consists of a 2^4 unreplicated factorial investigating the response variable *advance rate* which assumes values in (0, 100). This experiment is quite well known in the literature. The design matrix

and response data are shown in Table **??** at the end of this chapter. 14 effects are analyzed (4 main effects, 6 two-interaction effects, and 4 three-interaction effects). Initially, Daniel-76 analyzed this data to illustrate the usefulness of normal probability plotting in order to identify large and potentially important factor effects. Also, Torres-93, Montgomery-01, Lewis-01a, Box-05, Montgomery-09, and Myers-2011 reanalyzed this data using linear normal model, logarithm transformation, rank transformation and GLM with *gamma* link. Here, *Half-normal* plots for normal, *log(y)* transformation, and generalized linear models are shown in the three graphics of Figure **??**. This Figure indicates that the B, C, and D main effects appear as *active* in the three models mentioned. Note that BC and CD effects only were active for the normal model. The active effects visually chosen coincide with those reported in the literature, and they are shown in the first part of Table 1. (Notation $A = x_1$,..., $BCD = x_2x_3x_4$).
Therefore, a problem (and opportunity) is latent: different methods can yield different active effects. Could the BRM be another option?. Now, in this work, taking $y/100$ as the response, data is reanalyzed by means

Table 1: Analysis of the *Drill* experiment

Authors (year)	Response	Method	Active factors
Daniel-76	y	Linear reg.	B,C,D, BC, CD
Box-78, Montgomery-01	$log(y)$	Linear reg.	B, C, D
Torres-93	y rank transf.	Plot	B, C, D
Lewis-01a, Myers-2011	y	GLM.gamma	B, C, D

of the simple beta regression model (simple BRM)(Ferrari-04). Initially, upon fitting a simple BRM (fixing dispersion parameter), four link functions were employed for the mean response: *logit*, *probit*, *cloglog* and *cauchit*. The first three link functions present similar results to those considered in Table 1. Instead, the *cauchit* link yielded a very interesting result, the active effects were: B, C, D, BC, BD, CD, and BCD. Therefore, although this is an empirical result, at the same time, it is an encouraging opportunity to pose some questions. With respect to the data analysis from factorial experiments with response in (0,1), two questions arise: if the beta regression model turns out to be empirically appropriate, can the BRM be proposed as a good alternative?. Can the graphical results be supported by the statistical inference of the BRM?. Analyses were done using R software (R-14).
The two previous questions lead to an interesting challenge from theoretical and applied viewpoints. This challenge will be assumed in this work with two additional issues which are commented afterwards (variable dispersion, and restrictions on parameters).
An additional issue for the simple BRM: variable dispersion.
In the simple BRM fitted in Example 1, the dispersion parameter is taken as fixed. The generalization of this model allows for modelling the dispersion to try to explain the variability of the response through several covariates. It is known as the *variable (varying) dispersion beta regression model* (VDBRM). Several authors have discussed the advantages of this model with respect to the simple one. For instance, Bayer-14 present a simulation study to exemplify that efficiency loss takes place when the dispersion parameter is incorrectly taken as a constant. Therefore, in this work the variable dispersion BRM will be employed.

Example 2. *Justifying variable dispersion in the BRM for a* 2^4 *experiment.*

In example 1, the *Drill* data from a 2^4 experiment is described. Starting from Half-normal plots, the best model was the simple BRM with the *cauchit* link. Now, in order to illustrate the advantage of the Variable Dispersion BRM, two BRM's were fitted using the *cauchit* link (g_1) for the mean in both models, as follows:

- Simple BRM:

$$g_1(\mu) = \beta_0 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_{23}x_2x_3 + \beta_{24}x_2x_4 \qquad log(\phi) = \alpha_0 \tag{1}$$

- VDBRM:

$$g_1(\mu) = \beta_0 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_{23}x_2x_3 + \beta_{24}x_2x_4 \qquad log(\phi) = \alpha_0 + \alpha_3x_3 + \alpha_4x_4 \tag{2}$$

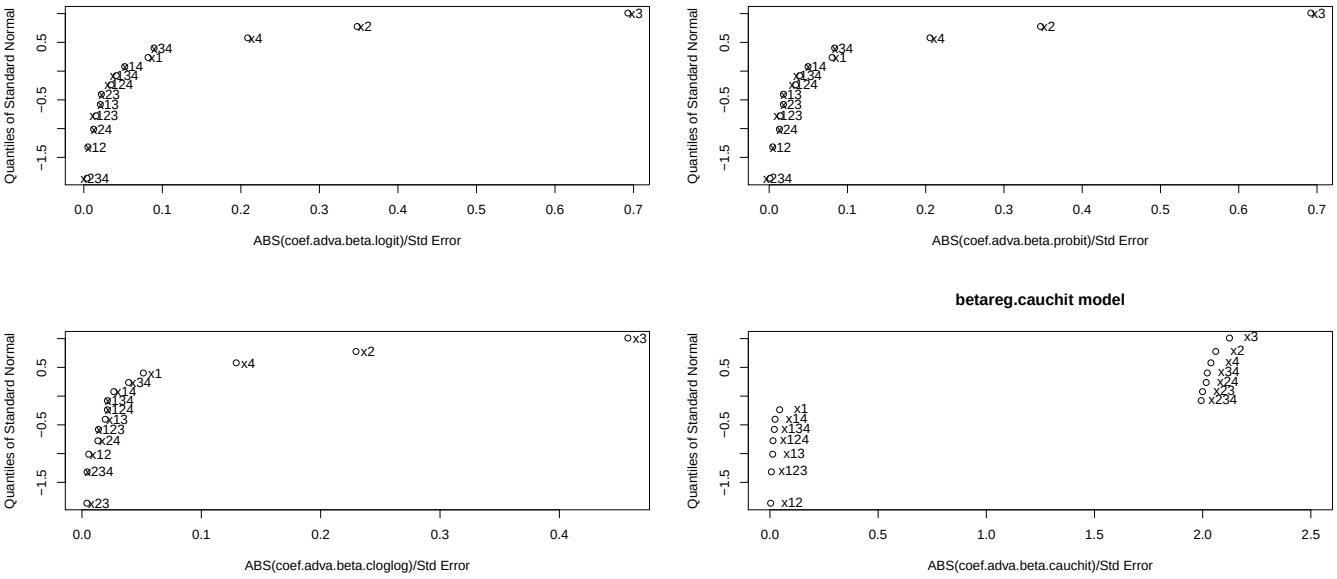

Figure 1: Halfnormal plots Data Advance 4 links Beta Regression

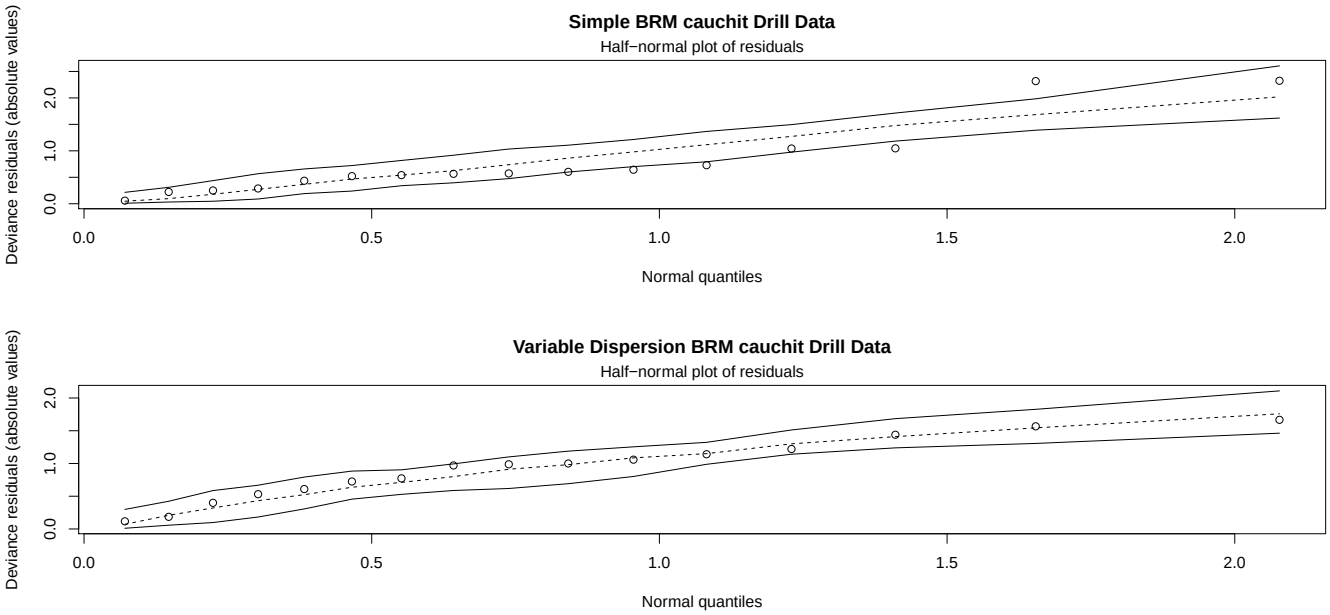

Figure 2: Half-normal residuals for *Advance* data. Simple and variable dispersion BRM

After fitting the models and doing the residual plots, results are shown in figure 2. An inspection of figure 2 indicates that the VDBRM fits better than the simple one. Up to here, the variable dispersion beta regression model (VDBRM) has been motivated to be used in factorial experiments.
The last issue is commented on afterwards.
Restrictions on parameters for the VDBRM.

Under the normal linear model, a common methodology in 2^{k-p} experiments is to assume as *negligible* the higher-order interaction factors (3, 4, 5,... order: this assumption is related to the partial derivatives of Taylor series $\frac{\partial y}{\partial A \partial B \partial C}$, ..., Box-05). When this assumption is done, the estimations of main or low-order interactions effects can be expressed exactly. For instance, if we have the estimation $\widehat{A} + \widehat{BCD}$, and BCD is considered negligible, then we are assuming that BCD=0, and hence $\widehat{A}$ provides the estimation of the main effect A explicitly. In this work, using the VDBRM for analyzing data from 2^{k-p} experiments with response in (a,b), is proposed avoiding this assumption upon considering some restrictions on parameters associated to the higher-order interaction factors, that is to say, a *restricted VDBRM*. Thus, by means of hypothesis tests will be taken decisions as if certain interactions effect can be considered zero or not.
Finally, the objective of this work is to propose, develop, and apply a restricted variable dispersion BRM. After developing the general expressions for the inferential (restricted) results: estimation, hypothesis tests, and goodness of fit, the model will be applied in general problems, and in 2^{k-p} experiments with response in (a,b) also.

2. Beta Distribution

2 Beta distribution

Beta distribution (p, q) is a good alternative when response is in (0,1) because it assumes values in this interval. Many practitioners have used the BRM when the response is in (0,1) and is related to covariates. In order to impose a regression structure for the beta response, several parameterizations for the beta distribution have been proposed.With $\mu = \frac{p}{p+q}$ and $\phi = p + q$, (example Ferrari2004), the probability density of a variable y in terms of its mean μ and its precision ϕ is given by

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\ y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad y \in (0,1), \tag{3}$$

In that case, $E(y) = \mu$ $(0 < \mu < 1)$, $Var(y) = \frac{\mu(1-\mu)}{1+\phi}$, where $\phi > 0$. In the BRM only the mean μ_i is modelled by means of covariates, however, in the variable dispersion beta regression model, both the mean, and the the precision ϕ_i are modelled through covariates. The frequentist estimation in the BRM is similar to the GLM, both based on the maximum likelihood estimation.
In the GLM, the parameters ϕ and β are orthogonal, which does not occur in the BRM. Inference for the BRM is based on asymptotic results under regularity conditions, the solutions do not have closed-form.

3 A Restricted Variable Dispersion Beta Regression Model: a proposal.

Suppose that there are q linearly independent restrictions on the parameter vector $\boldsymbol{\beta}$, yielding the Restricted VDBRM):

$$g_1(\mu_i) = \eta_{1i} = \mathbf{x}_i\boldsymbol{\beta} \qquad g_2(\phi_i) = \eta_{2i} = \mathbf{z}_i\boldsymbol{\alpha} \tag{4}$$

subject to

$$\mathbf{r}_j^T\boldsymbol{\beta} = \delta_j, \quad j = 1, \ldots, q \tag{5}$$

In matrix form,

$$g_1(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}, \quad g_2(\boldsymbol{\phi}) = \boldsymbol{Z}\boldsymbol{\alpha} \qquad \text{s.t.} \quad \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{\delta}$$

Dimension: $\boldsymbol{\beta}_{p_1\times 1}$, $\boldsymbol{\alpha}_{p_2\times 1}$, $\delta_{q\times 1}$ (known fixed numbers), $\boldsymbol{R}_{q\times p_1}$, with q rows linearly independent (known fixed numbers). Condition: $(p_1 + p_2 - q < n)$.
Remark: the columns of $\boldsymbol{Z}$ are usually some columns of $\boldsymbol{X}$.
In that follows: i) a penalized likelihood is proposed in order to estimate the (restricted) parameters, ii) inferential aspects of the model are presented: hypothesis tests, and goodness of fit.
The problem now with the restricted VDBRM is to maximize the log-likelihood (**??**) over $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ under

restrictions (5). Therefore, it is necessary to propose some solutions for the restricted VDBRM (4) - (5). The classical solution is introduced.

4 Frequentist solution for the restricted VDBRM.

One approach to solve restricted optimization problems is the *penalty function* method, [**?**], and it is used for the frequentist solution of (4)-(5). The first step is to consider the quadratic penalty function

$$P(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} l_i(\mu_i, \phi_i) - \frac{1}{2} \sum_{k=1}^{q} \lambda_k (\delta_k - \mathbf{r}_k^T \boldsymbol{\beta})^2 \tag{6}$$

and the second step is to find a solution to the unrestricted problem $max_{\boldsymbol{\beta},\boldsymbol{\alpha}} P(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ for fixed and positive values of λ_k, $k = 1, ..., q$ (Lagrange multipliers). To solve the penalty function (6) from a classical perspective, the *Fisher scoring* method was employed. In this case, for the $(m + 1)$ step

$$K(\mathbf{b}^{(m+1)}, \boldsymbol{\lambda}) \mathbf{b}(\boldsymbol{\lambda})^{(m+1)} = K(\mathbf{b}^{(m)}, \boldsymbol{\lambda}) \mathbf{b}(\boldsymbol{\lambda})^{(m)} + Q(\mathbf{b}^{(m)}, \boldsymbol{\lambda}) \tag{7}$$

Theoretical developments, and applied results for the score vector and the K matrix are shown in Grajales2015.

5 Examples of VDBRM in 2^{k-p} experiment

See Grajales-2012, and Grajales-2015, PhD Thesis.

6 Some conclusions

- A Restricted Variable Dispersion Beta Regression Model (VDBRM) is proposed, developed, and applied.
- Analysis of the simulated VDBRM is done.
- The VDBRM can be used to select models, for example for active effects and interactions in 2^{k-p} experiments with response in (a,b).

References

Box, G.E.P., Hunter, W.G., & Hunter, J.S. (2005). Statistics for Experimenters. John Wiley & Sons.

Cribari-Neto, F. & Zeileis, A. (2010). Beta regression in R. Journal of Statistical Software, 2, 1-24.

Grajales , L.F., Lopez, L.A.,& Melo, O.O. (2015). A Restricted Variable Dispersion Beta Regression Model. Submitted.

Grajales, L.F. (2015). Estimation and Inference for 2^{k-p} Experiments with Beta Response. Statistics PhD Thesis, Universidad Nacional de Colombia, Bogota.

Grajales, L.F., Ospina, R., & Lopez, L.A. (2012). On estimated means for 2^{k-p} Experiments with Beta Response. Journal of Statistics: Advances in Theory and Applications, 8, 105-130.

Lewis, S.L., Montgomery, D.C., & Myers R.H. (2001). Confidence Interval Coverage for Designed Experiments Analyzed with GLM. Journal of Quality Technology, 3, 279-292.

Montgomery, D.C. (2005). Design and Analysis of Experiments. John Wiley & Sons.