Likelihood analysis for a class of Beta mixed models

Wagner Hugo Bonat, Paulo Justiniano Ribeiro Jr[*], Walmes Marques Zeviani

LEG/DEST - Paraná Federal University

Abstract

Beta regression models are a suitable choice for continuous response variables on the unity interval. Random effects add further flexibility to the models and accommodate data structures such as hierarchical, repeated measures and longitudinal, which typically induce extra variability and/or dependence. Closed expressions cannot be obtained for parameter estimation and numerical methods, possibly combined with sampling algorithms, are required. We focus on likelihood inference and related algorithms for the analysis of Beta random effects models motivated by two problems with response variables expressed by indexes taking values in the unit interval. The first is a study on the life quality index of industry workers with data collected following an hierarchical sampling scheme. The second is a study comparing indexes of water quality up and downstream hydroelectric power plants reservoirs with nested effects and a longitudinal data structure. The random effects accounts for the grouped data structures. Model comparisons are used to assess relevant scientific hypothesis. Alternative models and algorithms are compared. Analysis includes data-cloning as alternative to numerical approximations and to assess identifiability. Confidence intervals based on profiled likelihoods are compared with the ones obtained by asymptotic quadratic approximations, showing relevant differences for the parameters related to the random effects.

[*] Corresponding author: paulojus@leg.ufpr.br, Dept. Estatística-UFPR, CP 19.081, Curitiba, PR Brazil, 81.531-990

keywords: hierarchical models, likelihood inference, Laplace approximation, data-cloning, life quality, water quality

1 Introduction

Proportions, rates and indexes are measured in the interval $[0, 1]$ and used as response variables in different subject areas. The usual linear (Gaussian) regression model is inappropriate because observed and predicted values are not confined to the unity domain and is unable to capture asymmetries.

Alternative models are considered in the literature. Kieschnick and McCullough (2003) provides a summary and, based on the results of several case studies, advocates the adoption of Beta regression models under which the distribution of the response variable can assume a diversity of forms.

Regression models for independent and identically distributed Beta variables are proposed by Paolino (2001), Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto (2004). The modelling inherits from the principles of generalised linear models (Nelder and Wedderburn, 1972), with a suitable link function relating covariates to the expectation of the response variable. Simas et al. (2010) extends the models regressing both, the mean and the precision parameters with covariates and also discussing non-linear forms for the predictor. Smithson and Verkuilen (2006) adopts the Beta regression on an analysis of IQ data arguing it provides a prudent and productive alternative to usual choices even if not always providing the best fit. The model properly accounts for data bounded above and below, is able to fit strongly skewed distributions, accommodates heterocedasticy, allows for separately testing hypothesis on location and dispersion whilst being parsimonious with only two parameters as in the Gaussian linear model.

Regression models for independent and identically distributed Beta variables are developed by Paolino (2001), Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto

(2004). The modelling inherits from the principles of generalised linear models (Nelder and Wedderburn, 1972), relating the expected value of the response variable to covariates through a suitable link function. Cepeda (2001), Cepeda and Gamerman (2005) and Simas et al. (2010) extends the models regressing both, the mean and the precision parameters on the covariates. The latter also contemplates non-linear forms for the predictor. Smithson and Verkuilen (2006) explores the Beta regression with an application to IQ data and arguing that, even if not always the best choice, it provides a prudent and productive alternative to choices over usual choices by fitting strongly skewed distributions, accommodating heterocedasticy, allowing for hypothesis on location and dispersion separately and data bounded above and below whilst being parsimonious with two parameters as for the Gaussian linear models.

Methods for likelihood based inference and model assessment are proposed by Espinheira et al. (2008a), Espinheira et al. (2008b) and Rocha and Simas (2010). Bias correction for likelihood estimators are developed by Vasconcellos and Cribari-Neto (2005), Ospina et al. (2006), Ospina et al. (2011) and Simas et al. (2010). Branscum et al. (2007) adopts Bayesian inference analysing virus genetic distances. The Beta regression is implemented by the **betareg** package (Cribari-Neto and Zeileis, 2010) for the R environment for statistical computing (R Development Core Team, 2012). Extended functionality is added for bias correction, recursive partitioning and latent finite mixture (Grün et al., 2011). Mixed and mixture models are further discussed by Verkuilen and Smithson (2011). Time series dependence structure is considered by (McKenzie, 1985), (Grunwald et al., 1993) and (Rocha and Simas, 2010). More recently (da Silva et al., 2011) uses a Bayesian Beta dynamic model for modelling and prediction of time series with an application to the Brazilian unemployment rates.

Dependence structures may arise in other contexts such as groups in the sampling mechanism, hierarchical model structures, longitudinal data and split-plot designs. Correlation can be induced by random effects assigned to observations within the same group and the total variability can be decomposed in within and between groups effects. Beta mixed models therefore allow for dependent and overdispersed data by inclusion of random effects, typically under the assumption they are Gaussian distributed, likewise usual specifications in generalised linear mixed models.

Beta mixed model a suitable choice for the two examples considered here. The first is a study on the life quality index of industry workers with data grouped by the hierarchical structure. The second is a comparison of water quality indexes upstream and downstream hydroelectric power plant reservoirs with data grouped on a longitudinal structure.

The likelihood function involves an integral which cannot be solved analytically. Gaussian Quadrature, Monte Carlo and Laplace approximation were all considered for integrating the random effects an our tests points the latter as the method of choice. We also consider the Markov chain Monte Carlo (MCMC) based algorithm proposed by Lele et al. (2007) for likelihood inference for generalized linear mixed models. Laplace approximation is less demanding on computing time and suitable for model choice whereas the latter can be used for further assessment of best fitted models.

Generalised linear mixed models and Beta regression models are widely discussed in the literature whereas Beta mixed models are recently considered by Figueroa-Zúñiga et al. (2013), under the Bayesian perspective. We focus on likelihood based inference and data cloning (Lele, 2010a) further investigated identifiability of the adopted models. Results obtained by computationally less demanding linear nad non-linear moxed models are included for comparison.

The Beta regression model with random effects is defined in Section 2 and the general setup for likelihood inference is presented in Section 3. The two motivating examples are presented in Section 4, illustrating the flexibility of the model in accounting for relevant features of the data structures which would be neglected under a standard Beta regression assuming independent observations. The examples have different justifications and structures for the random effects. The first specifies two, possibly correlated, random effects and the second has a nested random effects structure as a parsimonious alternative to a fixed effects model. We compare results obtained with different models and algorithms and close with concluding remarks on Section 5.

2 Beta mixed models

The Beta distribution parametrized in terms of mean and precision parameters (Jørgensen, 1997) has density:

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

with $0 < \mu < 1$, $\phi > 0$ and $\Gamma(.)$ is the Gamma function. We denote $Y \sim B(\mu, \phi)$, $E(Y) = \mu$, $V(Y) = \frac{\mu(1-\mu)}{(1+\phi)}$ and ϕ is a precision parameter, the greater its value the lesser the variance of Y.

For random sample from $Y_i \sim B(\mu_i, \phi)$, and assuming ϕ to be constant, the Beta regression model (Ferrari and Cribari-Neto, 2004) is specified by $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \boldsymbol{\eta}_i$, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ a vector of the k unknown regression coefficients, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$ a vector of k known covariates and η_i is a linear prediction. The model specification is completed the choice of a link function $g(\cdot) : (0,1) \rightarrow \Re$. We adopt the *logit* $g(\mu) = \log(\mu/(1-\mu))$ and other usual choices are the *probit*, *complementary log-log* and *cauchit* (Cribari-Neto and Zeileis, 2010).

This model does not contemplates possible dependencies such as induced by multiple measurements on the same observational unit, time or spatial structures. Inclusion of latent random effects on grouped data structure is a parsimonious strategy in comparison to adding parameters to the fixed part of the model, whilst still accounting for nuisance effects.

Denote Y_{ij} an observation $j = 1, \ldots, n_i$ within group $i = 1, \ldots, q$ and $\mathbf{y}_i$ denotes a n_i-dimensional vector of measurements from the i^{th} group. Let $\mathbf{b}_i$ a q-dimensional vector of random effects and assume the responses Y_{ij} are conditionally independent with density

$$f_i(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y_{ij}^{\mu_{ij}\phi-1}(1-y_{ij})^{(1-\mu_{ij})\phi-1}, \tag{2}$$

with a link function $g(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i$, a vectors of known covariates $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ with dimensions p and q, respectively, a p-dimensional vector of unknown regression parameters $\boldsymbol{\beta}$ and the precision parameter ϕ. The model specification is completed by $[\mathbf{b}_i|\Sigma] \sim N(\mathbf{0}, \Sigma)$ assuming Gaussian random effects.

2.1 Parameter estimation

Model parameters can be estimated by maximising the marginal likelihood obtained by integrating the joint distribution $[\mathbf{Y}, \mathbf{b}]$ over the random effects. The contribution to the likelihood from each group is

$$f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i. \tag{3}$$

Assuming independence among the N groups, the full likelihood is given by

$$L(\boldsymbol{\beta}, \Sigma, \phi) = \prod_{i=1}^{N} f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi). \tag{4}$$

Evaluation of (4) requires solving the integral N times. For the simpler model with a single random effect the integrals are unidimensional. More generally, the dimension equals the number of random effects in the model which imposes practical limits to numerical methods and approximations required to evaluate the likelihood. The integrals in our examples have up to five dimensions and solved by Laplace approximation (Tierney and Kadane, 1986) for the reported results. The marginal likelihood is maximised by the algorithm BFGS (Byrd, 1995) as implemented in R (R Development Core Team, 2012).

Alternative methods are available and we report results from other numerical integration methods and also the *data cloning* algorithm (Lele et al., 2007) proposed in the context of maximum likelihood estimation for generalised linear mixed models. Data cloning also provides tools to assess identifiability (Lele, 2010a) which we believe is worth exploring for the Beta mixed model.

The data-cloning algorithm is based on replicating (cloning) $K-times$ the observations $\mathbf{y}_i$ from each group generating $N \times K$ cloned data denoted by $\mathbf{y}_i^K$. The corresponding likelihood $L^K(\boldsymbol{\beta}, \Sigma, \phi)$ has the same maximum as (4) and Fisher information matrix equals K times the original information matrix. The method relies on the Bayesian approach to construct a MCMC algorithm and using the fact the effect of prior vanishes as the number of clones is increased. The model is therefore completed by the specification of priors $\pi(\boldsymbol{\beta})$, $\pi(\Sigma)$ and $\pi(\phi)$, which combined with the cloned likelihood, lead to a posterior of the form

$$\pi_K(\boldsymbol{\beta}, \Sigma, \phi | y_{ij}) = \frac{[\int f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i]^K \pi(\boldsymbol{\beta})\pi(\Sigma)\pi(\phi)}{C(K; y_{ij})} \tag{5}$$

with the normalising constant

$$C(K; y_{ij}) = \int [\int f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i]^K \pi(\boldsymbol{\beta})\pi(\Sigma)\pi(\phi) d\boldsymbol{\beta} d\Sigma d\phi. \tag{6}$$

Monte Carlo Markov chain (MCMC) algorithms (Robert and Casella, 2004) provide a sample from the posterior. By increasing the number K of clones, the posterior mean should converged to the maximum likelihood estimator and K times the posterior variance should correspond to the asymptotic variance of the MLE (Lele, 2010a). Priors are used to run the algorithm without affecting inference as the likelihood can be arbitrarilly weighted by increasing the number of clones to the point that the effect of priors are negligible.

Despite the flexibility of the inferential mechanism, usual concerns on the specification of hierarchical models applies. Realistic and suitable models for the problem and available data can be complex and need to be balanced against identifiability, not often checked nor trivial (Lele, 2010b).

Data cloning provides a straightforward identifiability check which can be used for hierarchical models in general. Lele (2010a) shows that under non-identifiability, the posterior converges to the prior truncated on the non-identifiability space when the number of clones is increased. As a consequence, the largest eigenvector of the parameter's covariance matrix does not converges to zero. More specifically, if identifiable, the posterior variance of a parameter of interest should converge to zero when increasing the number of clones.

2.2 Prediction of random effects

Prediction of random effects are typically required as for the examples considered here. Under the Bayesian paradigm the predictions can be directly obtained from the posterior distribution of the random effects given by

$$f_i(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\beta}, \Sigma, \phi) = \frac{f_i(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\Sigma)}{\int f_i(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i}. \tag{7}$$

which does not have a closed expression for the Beta model. The posterior mode maximizes $f_i(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\Sigma)$ providing a point predictor for $\hat{\mathbf{b}}_i$ and empirical Bayes predictions can be obtained by replacing the unknown parameters by their maximum likelihood estimates.

3 Examples

3.1 Income and life quality of Brazilian industry workers

The Brazilian industry sector *worker's life quality index* (IQVT, acronym in Portuguese) combines 25 indicators from eight thematic areas: housing, health, education, integral health and workplace safety, skill development, work attributed value, corporate social responsibility, participation and performance stimulus. The index is constructed following the same premisses of the united nations human development index[1]. Values are expressed in the unity

[1] http://hdr.undp.org/en/humandev/

interval and the closer to one, the higher the industry's worker life quality.

A pool was conducted by the Industry Social Service[2] in order to assess worker's life quality in the Brazilian industries. The survey included 365 companies on the Federal District and nine out of the 27 Brazilian federative units. IQVT was computed for each company from questionnaires applied to workers according to a sampling design. Companies provided additional information on budget for social benefits and other quality of life related initiatives.

A suitable model is aimed to assess the effects on IQVT of two company related covariates, average *income* and *size*. The first is simply the total of salaries divided by the number of workers expressing the capacity to fulfil individual basic needs such as food, health, housing and education. The second reflects the industry's quality of life management capability. There is a particular interest in learn whether larger companies with 500 or so workers, typically multinational working under regimes of worldwide competition, provide better life standards in comparison with medium (100 to 499 workers) and small (20 to 99 workers) sized industries. The federative unit where the company based is expected to be influential due to varying local legislations, taxing and further economic and political conditions. Plots on Figure 1 suggests IQVT is affect by income, size and federative units. The income is expressed in logarithmic scale centred around their average.

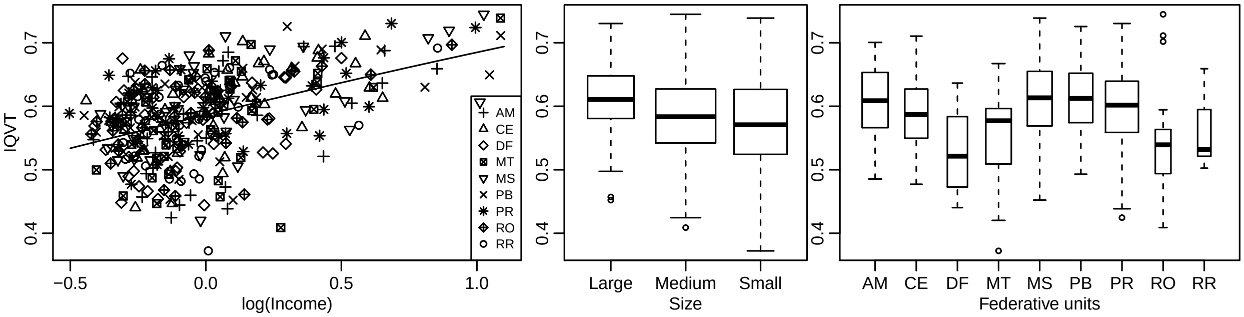

Figure 1: IQVT related to (centred log) average income, company size and federative unit.

[2] Serviço Social da Indústria - SESI

The Beta random effects model for IQVT is

$$
\begin{aligned}
Y_{ij}|\mathbf{b}_i &\sim Beta(\mu_{ij}, \phi) \\
g(\mu_{ij}) &= (\beta_0 + b_{i1}) + \beta_1 Medium_{ij} + \beta_2 Small_{ij} + (\beta_3 + b_{i2}) Income_{ij} \\
\mathbf{b}_i \sim NMV(\mathbf{0}, \Sigma) \text{ with } \Sigma &= \begin{bmatrix} , 1/\tau_1^2 & \rho \\ \rho & 1/\tau_2^2 \end{bmatrix},
\end{aligned}
$$

parametrized such that β_0 is associated with large size companies and β_1 and β_2 are differences with the medium and small sized, respectively. Random intercept b_{i1} and slope b_{i2} associated with *income* account for the effect of the federative units. The link function is the *logit* $g(\mu_{ij}) = \log\{\mu_{ij}/(1-\mu_{ij})\}$. Model parameters to be estimated consists of the regression coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$, the random effects covariance parameters $(\tau_1^2, \tau_2^2, \rho)$ and the precision parameter ϕ.

A sequence of sub-models are defined for testing relevant effects. Model 1 is the null model with simply the intercept. Model 2 includes the covariate *size* and Model 3 the *income*. Model 4 adds random intercepts and Model 5 adds a random slope to *income*. For comparison, we also fit corresponding Gaussian linear and non-linear (mixed) models which are widely used in practice.

A special care was taken to obtain comparable results between the models which does and does not involves numerical integration of the random effects.

Parameter estimates for the Beta models using Laplace approximation for the random effects are given in the top part of Table 1 and maximised log-likelihoods for the five model structures are given in Table 1.

Likelihood computations for models including random effects require solving integrals for which the Laplace approximation is used. The data and R (R Development Core Team, 2012) code will be made available at the paper companion web-page[3].

Results for models 1-3 confirms the effects of the covariates and the increasing values for

[3] http://www.leg.ufpr.br/papercompanion/betamixed

Table 1: Parameter estimates for the Beta models (top) and maximised likelihood for different methods and alternative models (bottom) - IQVT.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
β_0	0.35	0.45	0.43	0.40	0.40
β_1		-0.11	-0.09	-0.07	-0.07
β_2		-0.16	-0.14	-0.13	-0.13
β_3			0.42	0.47	0.47
ϕ	53.97	56.80	72.86	94.19	94.19
τ_1^2				62.36	62.35
τ_2^2					51480.17
ρ					0.85
Method/Model			Maximised likelihood		
Laplace	472.20	481.51	526.94	561.79	561.80
LMM	470.42	479.96	523.85	558.89	558.90
NLMM	470.42	479.96	523.77	558.96	558.96

the estimates of ϕ from 53.97 on model 1 to 72.85 on model 3 confirms further explanation of the data variability. The random intercept clearly improves the model fit expressing the variability of the IQVT among the federative units with an increase of 34.85 in the log-likelihood, however addition of the random slope did not prove relevant. Final choice based on likelihood ratio tests points to Model 4, including the two covariates and just the random intercept. Accuracy of the approximations of the log-likelihood may differ for different combinations of parameter values in particular close to the borders of the parameter space.

The Beta mixed model model is not commonly adopted in the literature and this motivates us to consider the data cloning as distinct approach for likelihood computations and also allows for assessing the model identifiability. The results are reassuring with similar estimates and standard errors obtained by maximization of the approximated marginal likelihood and data cloning as shown in Table 2.

Interval estimates obtained by both, the asymptotic quadratic approximation with standard errors returned by data clone and by profile likelihoods are presented in Table 3. The latter can be asymmetric and with closer to nominal coverage rates. Intervals are similar all the parameters except for τ_1^2 with an artefactual negative lower bound for the quadratic

Table 2: Parameter estimates and standard errors for Model 4 by marginal likelihood and data-cloning - IQVT

Parameter	Marginal likelihood		Data-clone	
	Estimate	Std. error	Estimate	Std. error
β_0	0.40	0.05	0.40	0.05
β_1	-0.07	0.03	-0.07	0.03
β_2	-0.13	0.03	-0.13	0.03
β_3	0.47	0.04	0.47	0.04
ϕ	94.19	7.03	94.17	6.98
τ_1^2	62.36	32.00	62.03	32.08

Table 3: Asymptotic and profile likelihood based confidence intervals, Model 4 - IQVT

Parameter	Asymptotic		Profile	
	2.5%	97.5%	2.5%	97.5%
β_0	0.30	0.50	0.29	0.50
β_1	-0.13	-0.02	-0.13	-0.02
β_2	-0.19	-0.07	-0.19	-0.07
β_3	0.39	0.55	0.39	0.55
ϕ	80.49	107.84	81.09	108.65
τ_1^2	-0.85	124.91	19.74	156.48

approximation.

Identifiability is assessed by the data clone method as described in Section 2. We use the package *dclone* (Sólymos, 2010), with the JAGS (Plummer, 2003) MCMC engine with 1, 5, 10, 20, 30, 40 and 50 clones. For each number of clones we use 3 independent chains of size 6500, and burn-in of 1500. Results are summarised in Figure 2 with chains increasingly concentrated around the maximum likelihood estimate with increasing number of clones.

A flat normal prior (zero mean and precision 0.001) for the regression parameters is not influential. Results for Bayesian inference ($K = 1$) are similar for the original and the 50 fold cloned data. The prior for precision parameters is a Gamma$(0.1, 0.001)$ producing posterior means for ϕ and τ_1^2 for the original data compared with the obtained with cloned data.

Following the data clone idea, under identifiability the posterior variance should converge to zero for increasing number of clones K with variance decreasing at rates $1/k$. Such trend is detected as shown in Figure 3 which uses logarithmic scale to ease the visualisation. Variances decrease satisfactorily at nearly expected rates with a slight but not relevant difference for

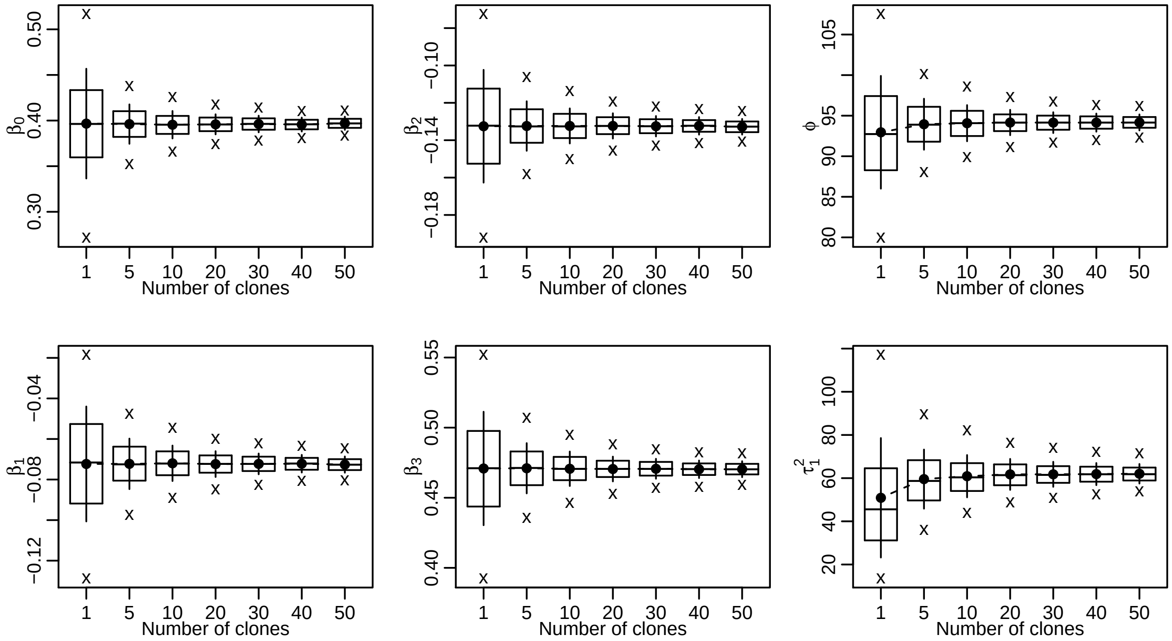

Figure 2: Sampled parameter values for different number of clones, Model 4 - IQVT.

the τ_1^2 parameter supporting the conclusion that the model is identifiable with the current data.

Fitted coefficients support the initial conjectures that the size has a relevant effect on the IQVT with expected decrease of 3.01% and 5.70% changing from large to medium and small sizes, respectively. These are figures obtained setting the other factors to baseline and/or zero values. Increasing income clearly affects positively the IQVT confirming and quantifying an expected behaviour. Finally, allowing for variations between federative units by adopting the random intercept terms increase the log-likelihood on 34.85 units, clearly a significant effect confirming the statements that there is a substantial variation in the quality of life among the federative units. Table 4 summarises the results with the figures of the predicted IQVT for different federative units and sizes and computed of a lower (R$500.00) and higher (R$2,500.00) levels of income.

Table 4 shows positive effects were for Mato Grosso do Sul (MS), Paraná (PR), Amazonas (AM), Ceará (CE) and the best case of Paraìba (PB), with IQVT 9.9% above the global average for small size business with average income of R500.00. Negative effects were

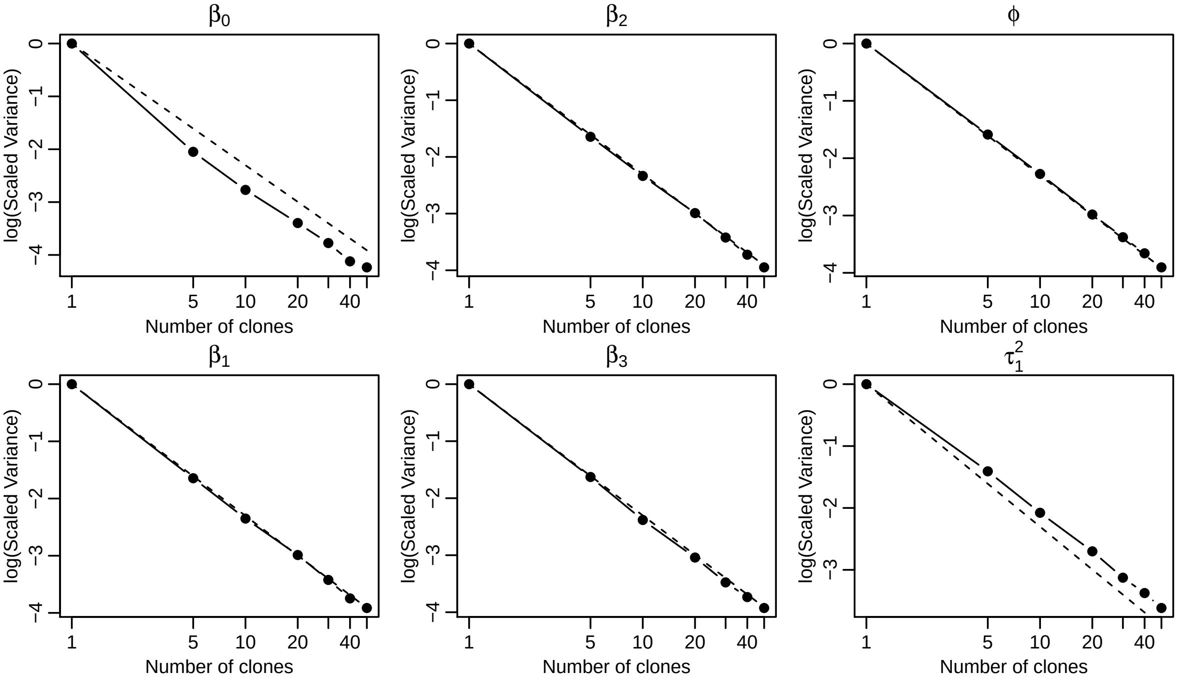

Figure 3: Identificability diagnostics with data-cloning for the Beta model with random intercept - IQVT.

Table 4: Predicted indexes and percentual diferences (within parenthesis) to the global average for Model 4 - IQVT.

Federative		R$ 500,00			R$ 2.500,00	
Unity	Large	Medium	Small	Large	Medium	Small
AM	52.91(1.52)	51.11(1.58)	49.60(1.63)	70.55(0.95)	69.02(1.00)	67.72(1.04)
CE	54.48(4.52)	52.68(4.70)	51.17(4.85)	71.84(2.80)	70.35(2.95)	69.08(3.07)
DF	46.5(-10.77)	44.71(-11.13)	43.23(-11.43)	64.95(-7.06)	63.29(-7.39)	61.88(-7.68)
MT	50.82(-2.49)	49.01(-2.58)	47.51(-2.65)	68.78(-1.58)	67.21(-1.66)	65.87(-1.73)
MS	54.22(4.04)	52.42(4.20)	50.92(4.33)	71.63(2.51)	70.14(2.64)	68.86(2.75)
PB	56.91(9.20)	55.13(9.58)	53.64(9.90)	73.79(5.60)	72.37(5.90)	71.15(6.16)
PR	53.83(3.29)	52.03(3.42)	50.52(3.52)	71.31(2.04)	69.81(2.15)	68.52(2.24)
RO	49.17(-5.66)	47.36(-5.86)	45.86(-6.03)	67.34(-3.64)	65.73(-3.82)	64.36(-3.97)
RR	50.11(-3.85)	48.31(-3.99)	46.80(-4.1)	68.17(-2.45)	66.58(-2.58)	65.22(-2.68)

estimated for Mato Grosso (MT), Roraima (RR), Rondônia (RO) and Distrito Federal (DF), the worse case with IQVT 11.43% below the global average.

There are larger differences to the global average for incomes around R\$500.00 becoming smaller for incomes around R\$2, 500.00, indicating lesser influence of company size and federative unity for increasing incomes. The more pronounced importance of *size* and *State* for low incomes are compatible with Brazilian conditions. There are several governmental supporting policies for low income workers such as social assistance unified system, young agent, social and food security, food support, popular restaurants, community catering, family health, maintenance and development educational fund among other Brazilian governmental social programs[4]. Such programs effectively improves quality of life of low income workers. Additionally companies internal supporting incentives for low income workers such as catering, transportation, basic shopping supply, among others, makes the workplace relevant for the worker quality of life. On the other hand, the greater the income the lesser the dependence on such benefits with the income becoming the main, if not the single, maintainer of the life quality and therefore less influenced by conditions such as size and federative unit. Interpretations based on the fitted model are therefore compatible with the subjective information about the working circumstances in the country. Observed data and fitted values for the random intercept model for each business size is shown on Figure 4.

Figure 4 shows IQVT values concentrated between 0.35 and 0.80 and within this range the relation with the log-income is nearly linear with a satisfactory adherence to the data. Some outliers for small business at Mato Grosso State did not show influence on the overall model fit.

3.2 Water quality on power plant reservoirs

The energy company COPEL operates 16 hydroelectric power plants in Paraná State, Brazil, generating over 4.500 MW. The reservoirs at the power plants are also used for leisure activ-

[4] listed at http://www.portaltransparencia.gov.br

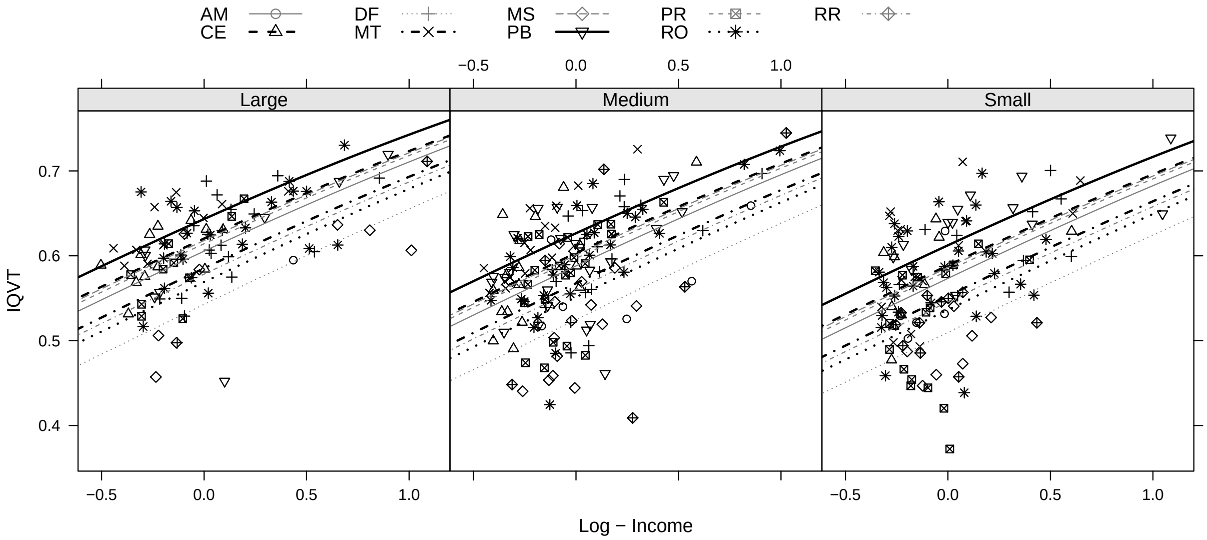

Figure 4: Observed and fitted values for the random intercept model for each business size.

ities, navigation and water supply. Effective functioning of the power plants is related to the water quality which is important on its own and determinant for the growth of organisms and aquatic flora. Assessing the effect of the reservoirs on the water quality is relevant for the water supply and environmental impacts. In compliance with operating licenses, the concessionaire company regularly monitors water quality in the reservoirs, as well as upstream and downstream the dammed rivers.

Monitoring includes comparing nine indicators of water quality agianst reference values given by standards for water suply. The water quality indicators are: dissolved oxygen, temperature, faecal coliform, water pH, biochemical oxygen demand (DBO), total nitrogen, total phosphorus, turbidity and total solids; having the public water supply as reference. The indicators are combined to produce a single water quality index (IQA, acronym in Portuguese) based upon a study conducted in the 70's by the US National Sanitation Foundation and adapted by the Brazilian company CETESB[5].

The main goal of the monitoring and analysis is to identify possible impacts and changes in the water quality possibly attributable to the presence of the dams. The effect is assessed by the comparison of measurements of the water quality between locations considered directly

[5] Companhia de Tecnologia de Saneamento Ambiental

unaffected and affected by the reservoir. Measurements taken upstream the main river are considered unimpacted reference values to be compared with measurements taken at the reservoir and downstream, possibly affected by the water contention and passage through the power plant, respectively.

Water quality indicators are measured quarterly on the 16 operating hydroelectric power plants. We consider here the data collected during the year of 2004. Main interest is in the effect of the covariate *LOCAL*, with levels *upstream*, *reservoir* and *downstream*. Other covariates are the power plant identification (*USINAS*) and the *quarter* of data collection. This amounts to 190 data with 12 measurements (four quarters × tree locals) for each of the 16 power plants with only two missing data.

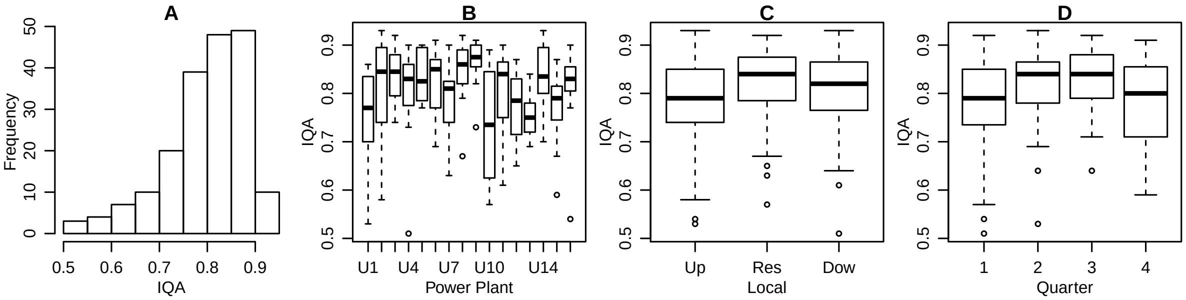

Figure 5: Summaries for the IQA data.

A data summary can be seen in Figure 5. There is a clear left asymmetry on the histogram (5 A), as usual for this kind of data. IQA varies between power plants as seen in Figure 5 B. Figure 5 C suggests an increase from upstream to the reservoir and a decrease from reservoir to downstream. Figure 5 D shows a pattern is expected to be repeated the over years with lower values for first and forth quarters, the warmer periods.

This brief exploratory analysis suggests that in order to investigate the effect of the position relative to the dam represented by the covariate LOCAL, the remaining effects of quarter and plant ID should be accounted for, including the possibility of distinct quarter effects for different plants in the form of an interaction. A further assumption for the analysis is to consider the power plant as a random effect. This is a choice of parsimony since considering main effects and interactions under fixed effects would amount 80 degrees of freedom. These are regarded as a possible sample from a population of environments. Although this can be disputed, the assumption is not only convenient for our intended method of analysis but also has proven sound for this particular data-set.

The model for the IQA data is specified as:

$$\begin{aligned} Y_{ijt}|b_j, b_{j,t} &\sim B(\mu_{ijt}, \phi) \\ g(\mu_{ijt}) &= \beta_0 + \beta_{1,i} + \beta_{2,t} + b_j + b_{j,t} \\ b_j &\sim N(0, \tau_U^2) \ \ ; \ \ b_{jt} \sim N(0, \tau_T^2) \end{aligned}$$

for the i^{th} relative location, j^{th} power plant e t^{th} quarters. Under the adopted parametrization for the fixed effects, $\beta_{1,i}$, $i = 2, 3$ relates to the change from upstream to reservoir and downstream, respectively. Likewise $\beta_{2,t}$, $t = 2, 3, 4$ compares the first quarter with the others. The random intercept b_j captures the deviations of each power plant to the overall mean and $b_{j,t}$ are the effects of each quarter within each power plant. The logit link function is used for $g(\cdot)$.

Hypotheses of interest are tested comparing submodels starting by setting $\beta_{1,i}, \beta_{2,j}, \tau_U^2, \tau_T^2 = 0$ and including each of these parameters, sequentially, up to the full model. Point estimates of the model parameters are presented in Table 5. Numerical estimates are obtained by the BFGS algorithm for maximizing the likelihood and computations use the Laplace approximation for models including random effects.

As expected the likelihood increases with the addition of terms in the model however with no substantial increase from Model 5 to 6. Although formal tests could be considered, in this example it is clear that the model including the random effect b_{jt} is unnecessary. Some criteria can be used to decide the final model, for instance the difference in likelihood is just 1.1091 between models 5 and 6 which under regularity would return a p-value of 0.1363 under a likelihood ratio test.

The likelihood evaluation for the larger model requires the solution of a five dimensional

Table 5: Parameter estimates for the Beta models (top) and maximised likelihood for different methods and alternative models (bottom) - IQA.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
β_0	1.40	1.27	1.14	1.14	1.15	1.15
β_{12}		0.23	0.23	0.24	0.24	0.24
β_{13}		0.15	0.15	0.16	0.15	0.16
β_{22}			0.21	0.22	0.22	0.22
β_{23}			0.29	0.31	0.32	0.32
β_{24}			0.05	0.05	0.06	0.06
ϕ	23.36	24.25	25.78	30.47	42.19	42.20
τ_U^2				28.97		43.54
τ_{UT}^2					11.19	15.04
Method/Model			Maximised likelihood			
Laplace	215.38	218.90	224.62	231.04	237.08	238.19
LMM	198.23	202.12	208.68	213.68	220.39	225.01
NLMM	198.23	202.12	208.72	214.88	223.12	223.91

Table 6: Parameter estimates and standard errors for Model 5 by marginal likelihood and data-cloning - IQA.

Parameter	Marginal likelihodd		Data-clone	
	Estimate	Std. error	Estimate	Std. error
β_0	1.15	0.09	1.15	0.10
β_{12}	0.24	0.05	0.24	0.07
β_{13}	0.15	0.01	0.15	0.07
β_{22}	0.22	0.01	0.22	0.13
β_{23}	0.32	0.03	0.31	0.13
β_{24}	0.06	0.01	0.06	0.13
ϕ	42.19	4.14	42.30	5.32
τ_{UT}^2	11.19	3.31	10.99	3.12

integral for each reservoir. Integral approximations such as Gauss-Hermite, Monte-Carlo integration and Laplace proved time-consuming and accuracy was an issue which could impact the hypotheses tests. Laplace approximation was the method of choice after attempts with these possible options. Alternatively, we have considered the data-cloning algorithm which does not demand integral approximation and numerical maximization. Parameter estimates obtained both ways for Model 5 are presented Table 6.

Point estimates are similar for the mean parameters however with differences for the standard errors. Overall, smaller values are obtained by the likelihood based on the numerical hessian which presented challenges and required adjustments on the finite differences methods in order to obtain numerically valid estimates. Data-cloning is more robust at the expense of a greater computational effort and time.

Quadratic approximation of the likelihood does not hold and symmetric confidence intervals based on the standard deviations are clearly inappropriate for parameters ϕ e τ_T and we turn to intervals based on profile likelihoods. Figure 6 shows the profile likelihoods for the precision parameters reparametrised on log scale for computational convenience. The Figure also includes plots from data-cloning results for diagnostics of identifiability. The profile likelihood for $\log(\tau^2_{UT})$ is asymmetric and this parameter is more sensitive than ϕ to the choice of prior, as indicated by the boxplots. The scaled log-likelihood plots shows a slightly faster decay in variance than expected for the corresponding number of clones, however the larger eigenvalue for the covariance matrix is always smaller than 1.1, an indicator of identifiability.

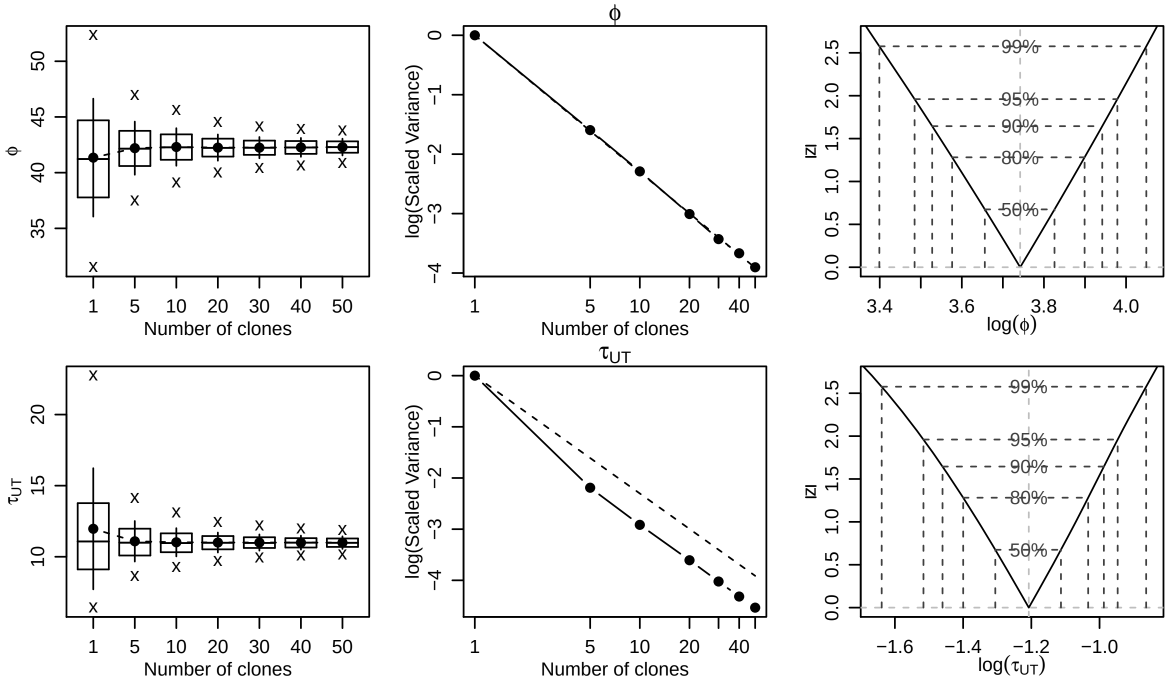

Figure 6: Profile likelihoods for precision parameters and identifiability diagnostics for Model 5.

The model of choice with reassuring results on the identifiability checks provides a confidence interval based on the profile likelihood for the random effects related to the power plants which clearly assures they improve the model fit. Empirical Bayes predictions of the random effects can be obtained and are overlaying the observed values on Figure 7 and separated by the relative position.

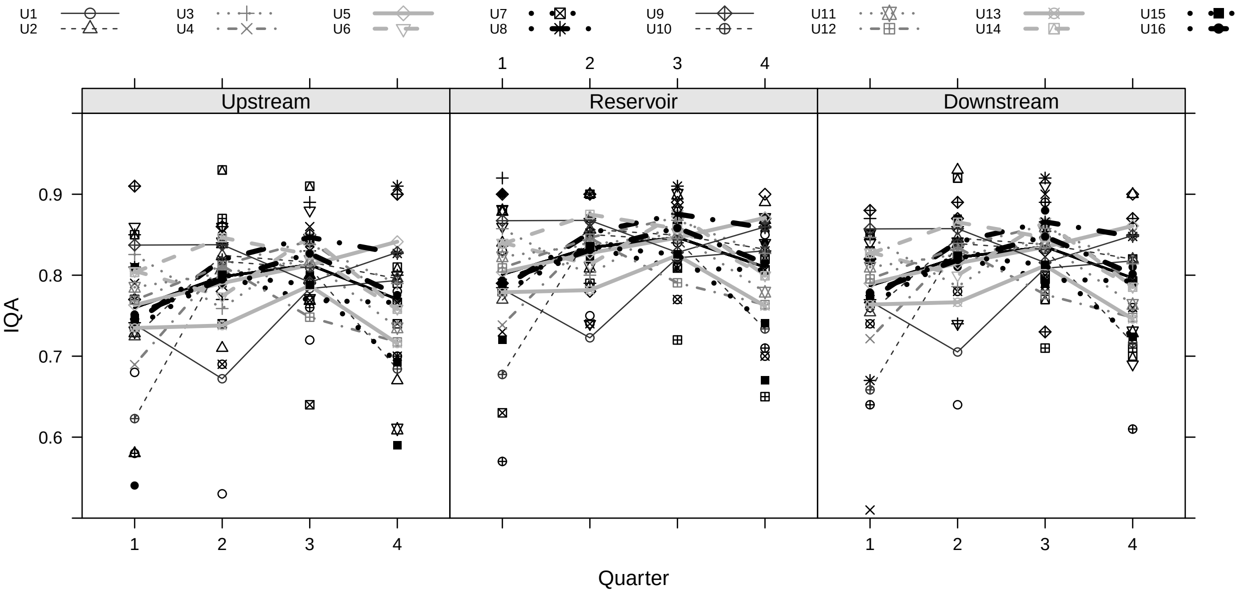

Figure 7: Prediction under the fitted models - IQA.

The fitted model predicts an increase of the IQA of 5.39% and 3.55% from upstream to the reservoir and downstream, respectively. These figures are computed by setting random effects to zero. The analysis confirms lower IQA values for the warmer first and forth quarters compared with the mild second and third quarters. This is likely to be a cyclic behaviour and expected to be repeated over the years. Besides such general behaviour, the random effects imply that the magnitude of the differences in mean values vary between power plants and quarters. The overall pattern is that the IQA substantially improves from upstream to the reservoir however shifting back to the original values downstream, with substantial variation between the power plants.

The adopted model and the algorithms implementing inferential methods proved satisfactory. Some extreme measurements taken upstream in the first and second quarters are smoothed on their fitted values. Differences between quarters suggest a possible temporal structure which could be included if jointly modelling observations from consecutive years, possibly revising the assumption of independent effects for a sequence of years. A greater range of IQA values for the first and forth quarter was detected in exploratory analysis. Accommodating different scale parameters is not worthy for our single year analysis but can be otherwise considered, possibly with interactions with power plant effects. Such addition to the model shall be balanced against the usual difficulties with the increasing of dimensionality of the random effects for the numerical algorithms. Possible workarounds such as quasi-likelihoods, MCMC algorithms, possibly under the Bayesian paradigm, or approximations such as proposed by Rue et al. (2009) need to be tailored for the Beta random effects models.

Our analysis under the Bayesian approach suggest sensitivity to priors seems to be an issue for such model and are likely worsen with the larger numbers of random effects. The data clone proved helpful in eliminating effect of priors and assessing identifiability, at the expense of a greater computational time.

4 Conclusion

A Beta regression mixed model including random effects associated with grouping units on a hierarchical model structure is adopted in the analysis of two datasets with response variables on the unit interval, one on worker's life quality and another on water quality. Different approches were adopted for likelihood computations, numerical (Laplace) approximation and a sampling based strategy by data cloning.

The first analysis shows the Brazilian industry life quality index is influenced by industry size and workers income with relevant random effects associated with the federative units. Findings based on the data analysis are compatible with subjective information validates social science's hypothesis. For the second no negative effects of the damns on the water quality index was detect, which is relevant for licensing power plants operators. The Beta random effects model accommodates environmental effects not fully captured by the measures variables. The random effects allows for a parsimonious model whilst considering extra sources of variation and the grouping structure.

For the data analysis we follow the strategy of fitting and selecting models using likelihood computations via the Laplace approximation followed by a detailed further assessment of the best model by data-cloning.

Likelihood inference methods and algorithms were implemented using numerical approximations to integrate out the random effects on the likelihood computations Results are comparable and we favour Laplace as the method of choice. Confidence intervals based on profile likelihood were obtained with distinct results from the ones obtained by asymptotic quadratic approximations in particular for the parameters associated with random effects. By the time we run our analysis there was no out-of-the box implementation implemented algorithms are made available[6]. In general we obtained stable results on analysis, however computational burden and accuracy of likelihood computations can be prohibitive with increasing number of parameters associated with random effects. Numerically unstable Hessians were found in the analysis of the larger model for water quality index with five random effects parameters.

Numerical marginal likelihood computations were compared with another inference strategy based on a MCMC scheme for cloned data. The data clone algorithm is a relatively new and promising proposal with little programming burden at the cost of increasing computing time, which can be partially alleviated by parallel or multicore computations for the several cloning numbers and chains. A particularly attractive feature is the possibility of investigating identifiability, which holds for both data analysis considered here. Point and interval estimates based on data-clone are comparable with the ones obtained by Laplace approximations. Profiling likelihoods with data cloning requires further developments (Ponciano et al., 2009).

Bayesian analysis is frequently used for analysis of hierarchical models and computationally corresponds to the step of the data cloning algorithm with no replicates of the data.

[6] http://www.leg.ufpr.br/papercompanions/betamix

Sensibility analysis on the prior choice is relevant but attenuated by data-cloning however mixing of MCMC chains and identifiability remains relevant. An alternative is to run data-cloning using integrated nested Laplace approximations (Rue et al., 2009) which can be adjusted to deal with random effects Beta mixed models saving the on computational burden by avoiding the more time demanding MCMC schemes. This also requires checking the Beta model against the usage of improper priors such as the ones commonly used for spatial or temporal models, if not completely avoiding them.

References

Adam J. Branscum, Wesley O. Johnson, and Mark C. Thurmond. Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian and New Zealand Journal of Statistics*, 49(3):287–301, 2007.

Richard H. Byrd. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 35(5):773, 1995.

Edilberto Cepeda. Variability modeling in generalized linear models. Master's thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2001.

Edilberto Cepeda and Dani Gamerman. Bayesian methodology for modeling parameters in the two parameter exponential family. *Estadística*, 57(1):93–105, 2005.

Francisco Cribari-Neto and Achim Zeileis. Beta regression in r. *Journal of Statistical Software*, 34(2), 2010.

Cibele Q. da Silva, Hélio S. Migon, and Leandro. T. Correia. Dynamic Bayesian beta models. *Computational Statistics and Data Analysis*, 55(6):2074–2089, 2011.

Patrícia L. Espinheira, Silvia L.P. Ferrari, and Francisco Cribari-Neto. On beta regression residuals. *Journal of Applied Statistics*, 35(4):407–419, 2008a.

Patrícia L. Espinheira, Silvia L.P. Ferrari, and Francisco Cribari-Neto. Influence diagnostics in beta regression. *Computational Statistics and Data Analysis*, 52(9):4417–4431, May 2008b.

Silvia L.P. Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

Jorge I. Figueroa-Zúñiga, Reinaldo B. Arellano-Valle, and Silvia L.P. Ferrari. Beta mixed regression: a Bayesian perspective. *Computational Statistics and Data Analysis*, 61:137–147, May 2013.

Bettina Grün, Ioannis Kosmidis, and Achim Zeileis. Extended beta regression in R: Shaken, stirred, mixed, and partitioned. Technical Report 22, Universitaet Innsbruck, 2011.

Gary K. Grunwald, Adrian E. Raftery, and Peter Guttorp. Times series of continuous proportions. *Journal of the Royal Statistical Society: Series B*, 9(4):586 – 587, 1993.

Bent Jørgensen. Proper dispersion models (with discussion). *Brazilian Journal of Probability and Statistics*, 11(1):89–140, 1997.

Robert Kieschnick and B. D. McCullough. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling*, 3(3):193–213, 2003.

Subhash R. Lele. Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492):1617–1625, 2010a.

Subhash R. Lele. Model complexity and information in the data: Could it be a house built on sand? *Ecology*, 91(12):3493–3496, 2010b.

Subhash R. Lele, Brian Dennis, and Frithjof Lutscher. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology letters*, 10(7):551–63, 2007.

Ed McKenzie. An autoregressive process for beta random variables. *Management Sciences*, 31(8):988–997, 1985.

John A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384, 1972.

Raydonal Ospina, Francisco Cribari-Neto, and Klaus L.P. Vasconcellos. Improved point and interval estimation for a beta regression model. *Computational Statistics and Data Analysis*, 51(2):960–981, 2006.

Raydonal Ospina, Francisco Cribari-Neto, and Klaus L. P. Vasconcellos. Erratum: Erratum to ”Improved point and interval estimation for a beta regression model” [Comput. statist. data anal. 51 (2006) 960-981]. *Comput. Stat. Data Anal.*, 55(7):2445, 2011.

Philip Paolino. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9(4):325–346, 2001.

Martyn Plummer. JAGS: A program for analysis of Bayesian graphical models using gibbs sampling, 2003.

José M. Ponciano, Mark L. Taper, Brian Dennis, and Subhash R. Lele. Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. *Ecology*, 90(2):356–362, February 2009.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

Andréa V. Rocha and Alexandre B. Simas. Influence diagnostics in a general class of beta regression models. *Test*, 20(1):95–119, 2010.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392, 2009.

Alexandre B. Simas, Wagner Barreto-Souza, and Andréa V. Rocha. Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, 54 (2):348–366, 2010.

Michael J. Smithson and Jay J. Verkuilen. A better lemon squeezer? maximum likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71, 2006.

Peter Sólymos. dclone: Data cloning in R. *The R Journal*, 2(2):29–37, 2010. URL `http://journal.r-project.org/`. R package version: 1.3-0.

Luke Tierney and Joseph Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

Klaus L. P. Vasconcellos and Francisco Cribari-Neto. Improved maximum likelihood estimation in a new class of beta regression models. *Statistics*, pages 13–31, 2005.

Jay J. Verkuilen and Michael Smithson. Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, 2011.