Modelo Beta com efeitos aleatórios

Wagner Hugo Bonat
Paulo Justiniano Ribeiro Jr
Walmes Marques Zeviani

LEG/UFPR - Laboratório de Estatística e Geoinformação

Abstract

O objetivo deste artigo é propor modelos de regressão beta com efeitos aleatórios, para tratar dados no intervalo unitário que apresentem estrutura hierárquica, medidas repetidas, estrutura longitudinal entre outras. São apresentados dois algoritmos para obter estimativas de máxima verossimilhança para os parâmetros do modelo proposto. A primeira usa a maximização da verossimilhança marginal e a segunda utiliza o algoritmo *data clone* que faz uso de técnicas de simulação. Obtemos intervalos baseados em perfil de verossimilhança. Dado a originalidade do modelo, é apresentado um diagnóstico de estimabilidade. O modelo é avaliado com a análise de dois conjuntos de dados reais provenientes das ciências sociais/econômicas e engenharia ambiental. De forma geral, o modelo apresenta alta flexibilidade para tratar dados no intervalo unitário que apresentam algum tipo de correlação decorrente da estrutura do experimento, tal como, observações obtidas ao longo do tempo na mesma unidade experimental. Os procedimentos de inferência mostraram-se adequados e concordantes para os conjuntos de dados analisados. Destaca-se o mal desempenho dos intervalos de confiança obtidos via aproximação quadrática da verossimilhança para os parâmetros de precisão que indexam os efeitos aleatórios, o que foi resolvido pela construção de intervalos baseado em perfil de verossimilhança. Os procedimentos de inferência foram implementados no software estatístico R e essão disponíveis junto ao artigo.

Keywords:

1. Introdução

Muitos pesquisadores de diferentes áreas examinam a influência de covariáveis em uma variável resposta restrita ao intervalo $(0, 1)$, tais como, proporções, taxas ou índices. Nestas situações o modelo de regressão tradicional, assumindo que a resposta tem distribuição Normal, não é adequado, uma vez que esta distribuição tem suporte na reta real e não consegue captar assimetrias próprias de dados restritos ao intervalo unitário.

Para tratar destas limitações diversos modelos têm sido propostos na literatura. Em [1] são apresentadas diversas formas de tratar dados com esta natureza. Após diversos estudos de casos os autores recomendam o uso de modelos de regressão com distribuição beta. A distribuição beta é muito flexível para modelar tais dados, uma vez que, sua densidade pode assumir diversas formas, dependendo da combinação de parâmetros.

Para situações onde a variável resposta é independente e beta distribuída, modelos de regressão foram propostos por [2], [1] e [3]. A forma de construção para tais modelos segue os mesmos princípios dos modelos lineares generalizados [4], onde a esperança da resposta é ligada a um preditor linear por uma função de ligação adequada. Uma versão extendida destes modelos foi proposta por [5], onde não apenas a média é função de covariáveis, mas também o parâmetro de precisão toma a forma de um modelo de regressão. Além disso, os autores contemplam a possibilidade de preditores não-lineares.

Ferramentas para avaliar o ajuste destes modelos dentro de uma perspectiva frequentista (verossimilhança) foram propostas, ver [6], [7] e [8]. Além disso correções para os viéses do estimador de Máxima Verossimilhança nesta classe de modelos foram apresentadas por [9], [10] e [5]. Uma versão Bayesiana foi proposta por [11], para analisar dados de distância genética entre vírus. O pacote **betareg** para o ambiente estatístico [12] é apresentado em [13].

Para dados não independentes algumas propostas de modelagem para proporções contínuas em séries temporais foram encontradas na literatura em [14], [15], [8] e recentemente [16] propôs um modelo beta Bayesiano dinâmico para modelar e prever séries temporais. Os autores fazem uma aplicação a taxa de desemprego mensal Brasileira.

Apesar destes avanços para dados não independentes em séries temporais, nenhum trabalho para modelos com efeitos aleatórios, ou mesmo medidas repetidas em modelos de regressão beta foram encontrados na literatura. Modelos com estes tipos de efeitos são adequados principalmente quando há níveis de amostragem, como em dados longitudinais, experimentos em parcelas subdivididas, são capazes de modelar correlação entre observações decompondo a variação entre os níveis e dentro dos níveis do efeito aleatório, proporcionando modelos mais realísticos. Modelos beta com efeitos aleatórios podem ser naturalmente descritos dentro de um framework de modelos mistos lineares generalizados [17].

O objetivo deste artigo é propor modelos de regressão beta com efeitos aleatórios, ou *beta mixed models* para tratar com dados no intervalo $(0, 1)$ que apresentem superdispersão, medidas repetidas, estrutura longitudinal entre outras. A inferência nesta classe de modelos será feita baseada na função de verossimilhança, que neste caso toma a forma de uma integral que não pode ser resolvida analiticamente. Sendo assim, diferentes estratégias para a resolução numérica desta integral serão consideradas. Recentemente [18] propôs um algoritmo baseado em técnicas MCMC *Markov Chain Monte Carlo* para fazer inferência baseada em verossimilhança em modelos mistos generalizados. Esta abordagem será considerada para comparação com o método tradicional de verossimilhança marginal.

Apesar da distribuição beta ser muito utilizada em análise de dados, a construção de modelos mistos beta não foi encontrada na literatura. Consequentemente, as condições de estimabilidade em modelos deste tipo são desconhecidas, e portanto, também serão consideradas neste artigo. Para isto, será utilizada a abordagem proposta em [19] baseada no algoritmo *data clone*.

O restante do artigo encontra-se dividido da seguinte forma: a Seção 2, introduz o modelo de regressão beta e o expande para a inclusão de efeitos aleatórios. A Seção 3, apresenta o procedimento geral de inferência baseado em verossimilhança marginal. Nesta Seção também é descrito o algoritmo *data clone* e como será realizado o diagnóstico de estimabilidade. Na Seção 4, os modelos serão ajustados a dois conjuntos de dados reais, para demonstrar a flexibilidade do modelo proposto em captar diversos aspectos

em situações práticas, que seriam complicados de obter usando apenas o modelo de regressão beta com efeitos fixos. As estimativas pontuais e intervalares por cada estratégia de inferência serão comparadas. A Seção 5, apresenta as principais conclusões e recomendações para trabalhos futuros. No apêndice é apresentado um exemplo de implementação computacional do modelo proposto.

2. Modelos de regressão beta com efeitos aleatórios

O modelo de regressão beta conforme proposto por [3] é baseado em uma parametrização alternativa da densidade beta em termos de média e precisão. A densidade beta neste caso toma a seguinte forma:

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

onde $0<\mu<1$, $\phi>0$ e $\Gamma(.)$ é a função gama. Denota-se $Y \sim B(\mu,\phi)$, nesta parametrização $E(Y)=\mu$ e $V(Y)=\frac{\mu(1-\mu)}{(1+\phi)}$. O parâmetro ϕ é chamado de parâmetro de precisão uma vez que quanto maior o ϕ menor a variabilidade de Y, assim ϕ^{-1} é um parâmetro de dispersão.

Considere $y_1,\ldots,y_n$ amostra aleatória provenientes de $Y_i \sim B(\mu_i,\phi)$, $i=1,\ldots,n$. O modelo de regressão beta é definido como

$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \boldsymbol{\eta}_i \tag{2}$$

onde $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_k)^T$ é um vetor $k\times 1$ de parâmetros de regressão desconhecidos, $\mathbf{x}_i = (x_{i1},\ldots,x_{ik})^T$ é um vetor de k covariáveis conhecidas e η_i é o preditor linear. O último componente do modelo é a função $g(.):(0,1)\to\Re$ chamada de função de ligação. Neste artigo a menos que dito ao contrário será utilizada a função de ligação *logit* que é definida por $g(\mu)=\log\mu/(1-\mu)$, porém outras funções de ligação como a *probit*, a complemento log-log e Cauchy também podem ser usadas [13].

O modelo descrito acima deixa de ser adequado quando a suposição de independência entre as amostras não for plausível. Isto acontece em diversas situações, por exemplo, quando diversas observações são feitas na mesma unidade experimental, gerando heterogeneidade entre unidades que não pode ser suficientemente descrita por covariáveis. Além disso, as observações podem ser correlacionadas no tempo ou espaço. Nestas situações modelos com a inclusão de efeitos aleatórios latentes têm sido propostos para modelar adequadamente a estrutura de dependência latente nas observações. Outros fatores relevantes na inclusão de efeitos aleatórios são a parsimônia na modelagem, uma vez que a parte fixa não é inflacionada de parâmetros desnecessários, controla para fatores *nuisance* como no caso de blocos, modela efeitos de grupos considerando que são amostras de uma população, mudando assim o nível da inferência.

Para incluir no modelo acima efeitos aleatórios considere Y_{ij} a variável resposta medida no grupo $i=1,\ldots,N$ na repetição $j=1,\ldots,n_i$ e $\mathbf{Y}_i$ um vetor n_i-*dimensional* de todas as medidas disponíveis para o bloco i. Assumindo independência condicional em $\mathbf{b_i}$ um vetor q-dimensional dos efeitos aleatórios tendo distribuição $N(\mathbf{0},\Sigma)$, a resposta Y_{ij} são independentes com densidade da forma

$$f_i(y_{ij}|\mathbf{b_i},\boldsymbol{\mu},\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y_{ij}^{\mu_{ij}\phi-1}(1-y_{ij})^{(1-y_{ij})\phi-1} \tag{3}$$

onde $g(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b_i}$ para uma função de ligação $g(.)$ conhecida, com $\mathbf{x}_{ij}$ e $\mathbf{z}_{ij}$ vetor de covariáveis conhecidas de dimensão p e q respectivamente, $\boldsymbol{\beta}$ um vetor p-dimensional de coeficientes de regressão fixos desconhecidos, e ϕ o parâmetro de precisão. Para completar a especificação do modelo, seja $f(\mathbf{b_i}|\Sigma)$ a densidade da $N(\mathbf{0}, \Sigma)$ distribuição atribuída para os efeitos aleatórios $\mathbf{b_i}$.

3. Inferência em modelos beta com efeitos aleatórios

3.1. Verossimilhança Marginal

A inferência para os parâmetros contidos no modelo beta com efeitos aleatórios pode ser feita maximizando a verossimilhança marginal, obtida após integrar os efeitos aleatórios. A contribuição para a verossimilhança proveniente de cada bloco é dada por

$$f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i, \tag{4}$$

segue que a verossimilhança para $\boldsymbol{\beta}$, Σ e ϕ é dada por

$$L(\boldsymbol{\beta}, \Sigma, \phi) = \prod_{i=1}^{N} f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi). \tag{5}$$

A principal limitação em usar 5 está na necessidade de resolver N integrais que podem ser unidimensionais no caso de modelos com apenas um efeito aleatório com a suposição adicional de independência entre os $\mathbf{b_i}'s$. Caso contrário será necessário a resolução de N integrais q-dimensionais, por exemplo, em um modelo com intercepto e inclinação aleatório não é natural assumir independência entre os efeitos, então será necessário resolver N integrais bidimensionais para contemplar a possível correlação entre os efeitos.

Neste artigo serão considerados modelos onde as integrais contidas na verossimilhança terão no máximo cinco dimensões. Sendo assim, considera-se a aproximação de Laplace, conforme descrita em [20] como método padrão para a resolução das integrais. Para a maximização da verossimilhança será utilizado o algoritmo BFGS [21] implementado na função *optim* do software [12].

Uma forma alternativa de obter estimativas de máxima verossimilhança em modelos mistos lineares generalizados foi proposto por [18] o algoritmo *data cloning*. Um fato muito importante desta metodologia é que dela deriva uma forma para avaliar a estimabilidade de modelos, o que consideramos importante na presente situação, já que, modelos beta com efeitos aleatórios não foram encontrados na literatura, consequentemente a estimabilidade destes modelos é desconhecida. Na próxima Seção, será descrito o algoritmo *data cloning* e como realizar o diagnóstico de estimabilidade.

3.2. Algoritmo data cloning

Para entender a idéia do algoritmo *data cloning* considere que as observações y_{ij} com $i = 1, \ldots, N$ blocos e $j = 1, \ldots, n_i$ repetições em cada bloco, são clonadas $K - vezes$ por blocos, ou seja, N blocos passam a ser $N \times K$ blocos. Denote os dados clonados por y_{ij}^K e a verossimilhança clonada resultante por $L(\boldsymbol{\beta}, \Sigma, \phi)^K$. Dois fatos de fundamental importância nesta nova verossimilhança precisam ser notados. Primeiro, a localização

do ponto de máximo desta função é exatamente o mesmo que o da função $L(\boldsymbol{\beta}, \Sigma, \phi)$. Segundo, a matriz de informação de Fisher baseada na verossimilhança clonada é K vezes a matriz de informação de Fisher baseada na verossimilhança original.

Usando a estrutura hierárquica do modelo misto podemos formular um modelo completo bayesiano designando priori's para todos os elementos do vetor de parâmetros. Denote $\pi(\boldsymbol{\beta})$, $\pi(\Sigma)$ e $\pi(\phi)$ as distribuições a priori de cada componente do vetor de parâmetros do modelo. Combinando com a verossimilhança clonada temos a seguinte distribuição a posteriori de interesse,

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

em que

$$C(K; y_{ij}) = \int [\int f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i]^K \pi(\boldsymbol{\beta})\pi(\Sigma)\pi(\phi) d\boldsymbol{\beta} d\Sigma d\phi \tag{7}$$

é a constante normalizadora. Amostrar desta distribuição a posteriori é possível usando algoritmos padrões de *MCMC* [22]. [19] mostram que quando K aumenta, a média desta distribuição a posteriori converge para o estimador de máxima verossimilhança e K vezes a variância a posteriori converge para a correspondente variância assintótica do MLE. Importante destacar que apesar da necessidade da atribuição de distribuições a priori, a inferência resultante é independente das priori's escolhidas, uma vez que, o algoritmo impõe cada vez mais peso para a verossimilhança clonando os dados o quanto for necessário para que a influência da priori seja desprezível na prática.

Ao propor um modelo hierárquico deve-se ter o cuidado para formular modelos que tenham compatibilidade com a realidade em analise e que possam ter todos os seus parâmetros identificáveis. Em muitas aplicações práticas os modelos são complexos, tornando provas analiticas de identificabilidade extremamente difíceis e raramente são feitas. De forma geral, a análise quase sempre Bayesiana é feita considerando que os parâmetros são de fato identificáveis [23].

Data cloning oferece uma solução simples para o estudo da identificabilidade de modelos. [19] demonstra que se os parâmetros são não identificáveis, quando aumenta-se o número de clones, a distribuição a posteriori converge para a distribuição a priori truncada no espaço de não identificabilidade. Consequentemente, o maior autovalor da matriz de variância-covariância não converge para zero. Este resultado pode ser usado para estudar a falta de identificabilidade de parâmetros em modelos hierárquicos como um todo.

Mais especificamente, se a variância da distribuição a posteriori de um parâmetro de interesse converge para zero, com o aumento do número de clones, este parâmetro é identificável. Desta forma, *data cloning* alerta para o problema de identificabilidade e também ajuda o analista a decidir se um determinado parâmetro é ou não importante para o modelo.

3.3. Predição dos efeitos aleatórios

Um importante componente inferencial quando trabalha-se com modelos mistos é a predição dos efeitos aleatórios. A predição dos efeitos aleatórios é baseada em sua distribuição a posteriori que tem densidade dada por

$$f_i(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\beta}, \Sigma, \phi) = \frac{f_i(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\Sigma)}{\int f_i(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i} \tag{8}$$

infelizmente esta densidade a posteriori não tem forma fechada. Portanto, a moda a posteriori é usada como estimativa pontual para $\mathbf{b}_i$. Mais especificamente, o estimador $\hat{\mathbf{b}}_i$ é o valor de $\mathbf{b}_i$ que maximiza $f_i(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}, \phi)f(\mathbf{b}_i|\Sigma)$, na qual os parâmetros desconhecidos foram substituídos por suas respectivas estimativas provenientes da estimação de máxima verossimilhança. As estimativas obtidas são denominadas de Bayes empirico (EB).

4. Resultados

4.1. Relação entre renda e Qualidade de Vida dos Trabalhadores da indústria brasileira

O índice de qualidade de vida dos trabalhadores (IQVT) da indústria brasileira é composto por 25 indicadores separados em 8 áreas temáticas: Habitação, Saúde, Educação, Saúde Integral e Segurança no Trabalho, Desenvolvimento de competências, Atribuição de valor ao trabalho, Consciência e responsabilidade social, Orientação a participação e ao desempenho. A metodologia para sua construção segue as premissas do Índice de Desenvolvimento Humano - IDH, preconizado pela ONU (Organização das Nações Unidas) em diversos países [1]. O IQVT por construção resulta em valores no intervalo $(0, 1)$, sendo que quanto mais próximo de 1 melhor é a qualidade de vida dos trabalhadores de uma determinada indústria.

Para o caso da indústria brasileira foi realizada uma pesquisa no ano de 2010, pelo Serviço Social da Indústria (SESI) em 365 empresas distribuídas em 8 dos 26 estados brasileiros, além do Distrito Federal. Através de um plano amostral, foram entrevistados trabalhadores das indústrias e o índice foi calculado para cada empresa pertencente a amostra. Além dos trabalhadores a empresa também foi entrevistada sobre diversos aspectos de como trata a questão da qualidade de vida, bem como, seus gastos com benefícios sociais entre outros.

Para a presente análise separamos duas covariáveis de particular interesse em relacionar com a qualidade de vida dos trabalhadores: a renda média da empresa, que indica a capacidade dos trabalhadores em suprir suas necessidades básicas, com alimentação, saúde, habitação, educação entre outras, e o porte da empresa que pode estar indiretamente ligado a sua capacidade em gerenciar e proporcionar qualidade de vida aos seus trabalhadores. Temos interesse em saber se empresas de grande porte (mais de 499 trabalhadores) que em geral são empresas multinacionais, que trabalham em regimes de competição mundiais, contam com trabalhadores com maior qualidade de vida que empresas médias (100 até 499 trabalhadores) e pequenas (de 20 até 99 trabalhadores).

Deseja-se criar um modelo que permita analisar adequadamente a relação destas duas covariáveis de interesse com o IQVT. Duas particularidades desta análise são facilmente notadas. Primeiro, a variável resposta é restrita ao intervalo unitário o que torna o modelo de regressão beta uma ferramenta analítica apropriada. Segundo, tem-se apenas uma amostra dos estados brasileiros 8 mais o Distrito Federal dos 26 estados possíveis. Nesta situação considerar o estado em que a empresa está localizada como um efeito fixo é claramente indesejável, uma vez que, a inferência ficaria restrita a estes estados. Além disso, o estado em que a empresa atua deve ser uma importante fonte de variação na resposta, pois empresas atuantes no mesmo estado estão sujeitas as mesmas políticas tarifárias e públicas como um todo.

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

A Figura 1 apresenta um diagrama de dispersão relacionando a renda e o IQVT e dois conjuntos de Boxplot's comparando os níveis das covariáveis porte e estado. Importante destacar que em todas as análises daqui em diante foi utilizada a covariável renda transformada em escala de logaritmo neperiano e teve sua média subtraída para evitar problemas numéricos.

A Figura 1 mostra uma tendência de aumento do IQVT com o aumento da renda dos trabalhadores. A covariável porte também parece influênciar no IQVT com um decréscimo ao passar de empresas de grande para médio e pequeno porte. Quanto aos estados também parece claro que apresentam diferenças significativas que devem ser levadas em consideração na modelagem estatística.

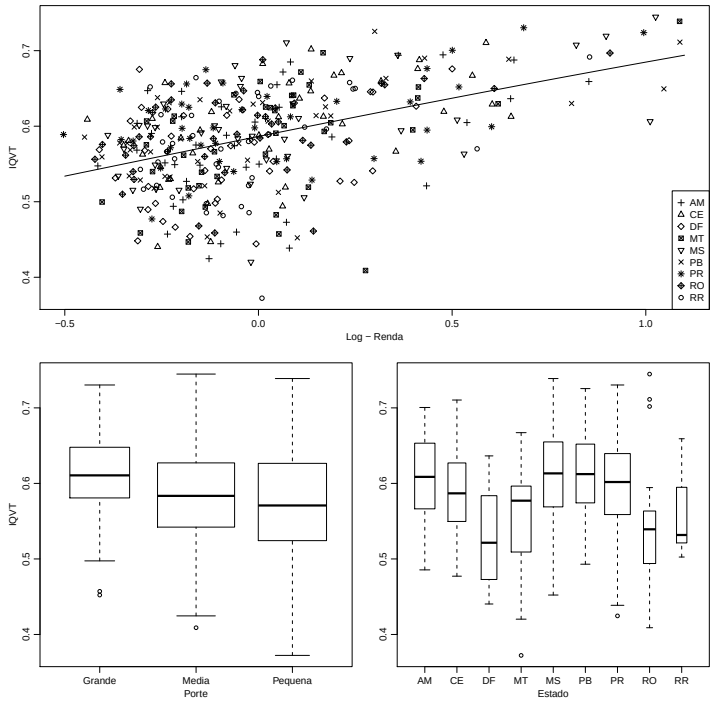

Figura 1: Diagrama de dispersão e boxplot's por covariáveis de interesse. Largura do box é proporcional ao log do número de elementos no grupo.

Dada estas indicações da estatística descritiva e as hipóteses de interesse propomos o seguinte modelo geral para relacionar o IQVT com as covariáveis de interesse:

- $Y_{ij} \sim B(\mu_{ij}, \phi)$;
- $g(\mu_{ij}) = (\beta_0 + b_{i1}) + \beta_1 \text{Média} + \beta_2 \text{Pequena} + (\beta_3 + b_{i2}) \text{Renda}$;
- $b_{ij} \sim NMV(\mathbf{0}, \Sigma)$ com $\Sigma = \begin{bmatrix} 1/\tau_1^2 & \rho \\ \rho & 1/\tau_2^2 \end{bmatrix}$ para $j = 1, 2$.

onde β_0 é o coeficiente associado as grandes indústrias, β_1 e β_2 medem o impacto no IQVT ao sair de uma indústria grande para uma de médio e pequeno porte respectivamente.

Tabela 1: Estimativas pontuais, logaritmo da verossimilhança maximizada e critério de informação de Akaike.

	Modelo.1	Modelo.2	Modelo.3	Modelo.4	Modelo.5
β_0	0.3479	0.4451	0.4338	0.3962	0.3965
β_1		-0.1050	-0.0878	-0.0723	-0.0724
β_2		-0.1608	-0.1443	-0.1326	-0.1329
β_3			0.4184	0.4703	0.4697
ϕ	53.9700	56.7966	72.8577	94.1938	94.1905
τ_1^2				62.3648	62.3464
τ_2^2					51480.4778
ρ					0.8509
ll	463.9274	473.2354	518.6716	553.5231	553.5252
AIC	-923.8548	-938.4708	-1027.3433	-1095.0462	-1091.0504

Efeitos aleatórios são incluídos no intercepto b_{i1} e na inclinação b_{i2} para o efeito da renda. Indicando um desvio da média geral para cada estado (intercepto aleatório) e uma inclinação diferente no impacto da renda, flexibilizando para que a renda tenha diferentes impactos conforme o estado. O vetor de parâmetros do modelo são os coeficientes de regressão $(\beta_0, \beta_1, \beta_2, \beta_3)$, os parâmetros de covariância dos efeitos aleatórios $(\tau_1^2, \tau_2^2, \rho)$ e o parâmetro de dispersão ϕ da Beta.

Alguns casos particulares do modelo geral serão considerados para comparação. No primeiro caso o modelo apresenta apenas β_0, indicando nenhum efeito das covariáveis e sem a presença de efeitos aleatórios é o modelo inicial. Em seguida incluímos o efeito de porte, o efeito de renda, ainda sem a inclusão de efeitos aleatórios. Os modelos sem efeitos aleatórios podem ser pensados como o modelo geral obtido quando se faz $1/\tau_1^2$ e $1/\tau_2^2$ tender a zero. O próximo caso consideramos apenas um intercepto aleatório por estado, é obtido fazendo com que $1/\tau_2^2$ tenda a zero. No último caso estimamos todos os parâmetros envolvidos no modelo geral. Para a covariável porte que é categórica definimos o nível Grande como referência.

Para o ajuste dos modelos foram desenvolvidas rotinas em *R* que estão apresentadas no apêndice. Em todos os modelos a função de ligação é a *logit*, as integrais contidas na verossimilhança foram resolvidas numéricamente pelo método de Laplace, e a maximização numérica da verossimilhança foi feita usando o algoritmo BFGS, implementado na função *optim()*. A Tabela 1 apresenta as estimativas pontuais o logaritmo da verossimilhança maximizada e o Critério de informação de *Akaike* para os cinco modelos considerados.

De acordo com os resultados da Tabela 1 verifica-se um aumento da log verossimilhança com a inclusão das covariáveis de interesse, porte e renda. Verifica-se também um aumento na estimativa pontual do parâmetro ϕ que passou de 53.97 no modelo 1, para 72.85 no modelo 3, indicando que mais variabilidade esta sendo explicada pelo modelo 3 comparado ao modelo 1.

Com relação a inclusão dos efeitos aleatórios, verifica-se que o ajuste melhorou significamente com a inclusão do intercepto aleatório por estado (553.52 − 518.67 = 34.85), mostrando claramente a importância do estado para explicar a variabilidade do IQVT. Quando foi incluído também o slope aleatório não verificou-se melhora significativa no

Tabela 2: Estimativas pontuais e desvio padrão via Verossimilhança Marginal e dClone.

	Pt.Marginal	SD.Marginal	Pt.dclone	SD.dclone
β_0	0.3962	0.0474	0.3970	0.0512
β_1	-0.0723	0.0269	-0.0726	0.0283
β_2	-0.1326	0.0288	-0.1328	0.0296
β_3	0.4703	0.0393	0.4704	0.0402
ϕ	94.1938	7.0256	94.1683	6.9767
τ_1^2	62.3648	31.8706	62.0308	32.0805

ajuste do modelo. Desta forma, fica claro que dentre os modelos ajustados o mais compatível com o conjunto de dados é o modelo 4. O critério de informação de *Akaike* reforça esta conclusão.

É importante notar que os modelos ajustados são aninhados, assim testes de razão de verossimilhança podem ser utilizadas para testar hipóteses relativas aos termos do modelo. Porém, no caso da inclusão dos efeitos aleatórios este aninhando é numa situação limite, já que, o zero não pertence ao espaço paramétrico da variância.

Outro fato relevante nesta análise é que aqui estamos trabalhando com uma aproximação da log-verossimilhança, que apresenta erros de aproximação diferentes de acordo com a combinação de parâmetros que são avaliados. Nesta situação onde temos uma estimativa que está muito próxima da borda do espaço paramétrico, é esperado que as aproximações apresentem um erro maior.

Uma vez que não encontramos resultados na literatura que possam ser comparados com os obtidos neste artigo, optou-se por reajustar o modelo escolhido por uma técnica diferente para nos certificarmos dos resultados obtidos pelo procedimento de inferência. O algoritmo *data clone* foi escolhido, uma vez que parte de algoritmos numéricos bastante distintos dos utilizados por verossimilhança marginal. Além disso, como já foi mencionado ele traz uma forma simples para verificar a estimabilidade do modelo, o que consideramos importante na presente situação. A Tabela 2 apresenta as estimativas pontuais e o desvio padrão associado a cada estimativa componente do modelo pelos dois algoritmos de inferência.

Os resultados apresentados na Tabela 2 mostram grande similaridade, tanto nas estimativas pontuais como nos desvios padrões estimados. Com isso, concluímos que o processo de inferência teve sucesso, já que partindo de duas metodologias distintas chegamos a resultados idênticos. O próximo passo é a construção de intervalos de confiança. Para isso podemos simplesmente usar dos resultados assintóticos do estimador de máxima verossimilhança e obter os intervalos baseados na aproximação quadrática.

Uma outra forma, por vezes mais elegante é construir intervalos baseados em perfil de verossimilhança, que em geral levam a resultados melhores em termos de nível de cobertura por possibilitar a construção de intervalos assimétricos. A Tabela 3 apresenta intervalos de confiança baseados na aproximação quadrática da verossimilhança usando os desvios padrões obtidos pelo algoritmo *data clone* e para completar obtivemos intervalos baseados em perfis de verossimilhança para comparação.

Os intervalos obtidos pela aproximação quadrática da verossimilhança são bastante próximos aos obtidos pela verossimilhança perfilhada para as estimativas dos parâmetros

Tabela 3: Intervalos de confiança assintótico e baseado em perfil de verossimilhança.

	2.5%	97.5%	2.5%	97.5%
β_0	0.2967	0.4973	0.2918	0.4978
β_1	-0.1281	-0.0171	-0.1275	-0.0172
β_2	-0.1909	-0.0747	-0.1910	-0.0741
β_3	0.3916	0.5491	0.3931	0.5480
ϕ	80.4943	107.8424	81.0877	108.6460
τ_1^2	-0.8458	124.9074	19.7383	156.4794

de média e precisão da beta. Para o parâmetro τ_1^2 do efeito aleatório o intervalo aproximado é claramente ruim, uma vez que apresenta valor negativo para um parâmetro estritamente positivo. Este resultado não é incomum quando se usa aproximação quadrática para construir intervalos para parâmetros de precisão/variância. Neste caso, o intervalo obtido pela verossimilhança perfilhada é claramente mais adequado, pois considera o espaço paramétrico.

A última análise a ser feita sobre o ajuste do modelo é a de estimabilidade de seus parâmetros. Para isto, utilizamos o método baseado no algoritmo *data clone* conforme apresentado na Seção 2. Para o ajuste foi utilizado o pacote *dclone* [24] utilizando o JAGS Just another Gibbs sampler [25]. para implementação do algorítimo de MCMC, foram ajustados modelos com 1, 5, 10, 20, 30, 40 e 50 clones. Foram rodadas 3 cadeias simultâneas de tamanho 6500, com uma queima de 1500 para cada conjunto de clones. A figura 2 apresenta BoxPlot's com os valores das cadeias de acordo com o número de clones. O comportamento esperado é que quanto maior o número de clones a cadeia vá ficando cada vez mais concentrada em torno da estimativa de máxima verossimilhança.

Os gráficos seguem o comportamento esperado. É interessante notar o efeito que a priori tem sobre cada parâmetro em análise. Para os parâmetros de média que foi usada uma priori Normal de média zero e precisão 0.001 a estimativa Bayesiana correspondente a não clonar os dados é praticamente a mesma que a com 50 clones, indicando que ela não tem nenhum efeito sobre a inferência destes parâmetros.

Quando olhamos para os parâmetros de precisão onde foram usadas priori Gama$(0.1, 0.001)$ observa-se que a estimativa com um clone teve um leve aumento para o parâmetro ϕ e um aumento mais considerável para o parâmetro τ_1^2 indicando que este parâmetro é mais sensível a escolha da priori. Conforme mencionado na Seção 2, um parâmetro é estimável se com o aumento do número de clones sua variância a posteriori converge para zero, a taxa de queda da variância é $\frac{1}{K}$, esta análise é apresentada na Figura 6, em escala logaritmica para facilitar a visualização.

O comportamento dos gráficos da Figura 6 está conforme o esperado. Com o aumento do número de clones a variância a posteriori dos estimadores está convergindo para zero e a taxa de decaimento é praticamente $\frac{1}{K}$ para todos os parâmetros, excessão ao parâmetro τ_1^2 que apresenta um leve desvio. Com isso, pode-se concluir que todos os parâmetros envolvidos no modelo proposto são estimáveis, e estão compatíveis com o conjunto de dados em análise.

De acordo com o ajuste do modelo, podemos observar que a covariável porte apresenta um forte impacto no IQVT. As estimativas indicam que ao passar de uma indústria de

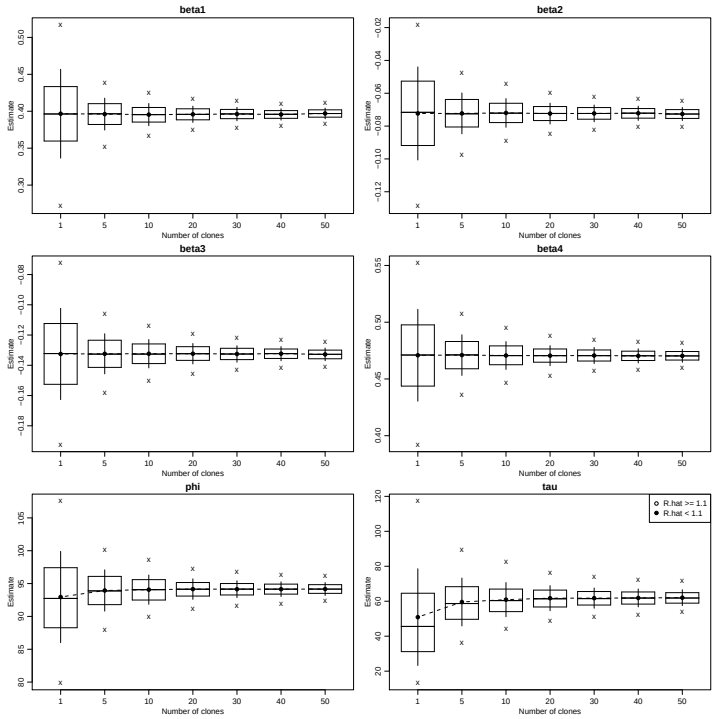

Figura 2: BoxPlot's dos valores simulados para cada parâmetro por número de clones.

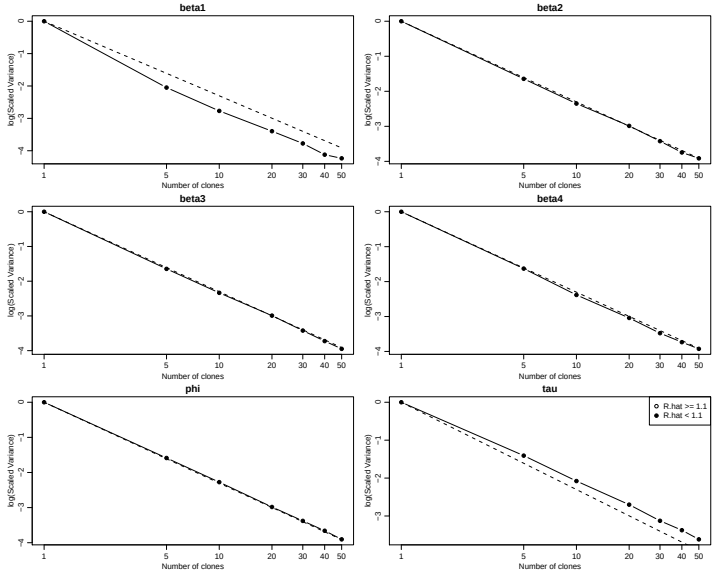

Figura 3: Diagnóstico de estimabilidade, modelo beta com intercepto aleatório.

Tabela 4: Valores preditos por estado, porte e renda. Entre parenteses diferença em percentual em relação a média.

ESTADO	R$ 500,00		
	Grande	Média	Pequena
AM	52.91 (1.52)	51.11 (1.58)	49.6 (1.63)
CE	54.48 (4.52)	52.68 (4.7)	51.17 (4.85)
DF	46.5 (-10.77)	44.71 (-11.13)	43.23 (-11.43)
MT	50.82 (-2.49)	49.01 (-2.58)	47.51 (-2.65)
MS	54.22 (4.04)	52.42 (4.2)	50.92 (4.33)
PB	56.91 (9.2)	55.13 (9.58)	53.64 (9.9)
PR	53.83 (3.29)	52.03 (3.42)	50.52 (3.52)
RO	49.17 (-5.66)	47.36 (-5.86)	45.86 (-6.03)
RR	50.11 (-3.85)	48.31 (-3.99)	46.8 (-4.1)
ESTADO	R$ 2.500,00		
	Grande	Média	Pequena
AM	70.55 (0.95)	69.02 (1)	67.72 (1.04)
CE	71.84 (2.8)	70.35 (2.95)	69.08 (3.07)
DF	64.95 (-7.06)	63.29 (-7.39)	61.88 (-7.68)
MT	68.78 (-1.58)	67.21 (-1.66)	65.87 (-1.73)
MS	71.63 (2.51)	70.14 (2.64)	68.86 (2.75)
PB	73.79 (5.6)	72.37 (5.9)	71.15 (6.16)
PR	71.31 (2.04)	69.81 (2.15)	68.52 (2.24)
RO	67.34 (-3.64)	65.73 (-3.82)	64.36 (-3.97)
RR	68.17 (-2.45)	66.58 (-2.58)	65.22 (-2.68)

Grande porte para uma de Médio porte o IQVT apresenta um decréscimo de 3.01%, já quando vamos para uma indústria de Pequeno porte este decréscimo é de 5.70%. Com relação a covariável renda o modelo indica que ao aumentar a renda média dos trabalhadores de uma determinada indústria é esperado um aumento no IQVT. O último componente do modelo é o estado, pela log verossimilhança, verifica-se que a inclusão deste efeito aumentou a log-verossimilhança em 34.85 unidades, conduzindo um teste de razão de verossimilhança temos um p-valor < 0.0001 mostrando a alta significância deste efeito. Importante destacar que para cada uma destas interpretações as outras covariáveis componentes do modelo foram fixadas em zero.

Para melhorar explorar o efeito do componente estado, bem como, as diferenças no impacto das covariáveis porte e renda de acordo com o estado, a Tabela 4 apresenta o IQVT predito de acordo o estado, porte e duas rendas uma baixa *R*$500.00 e uma alta *R*$2500.00.

Os resultados da Tabela 4 mostram que o estados Paraíba (PB), Mato Grosso do Sul (MS), Paraná (PR), Amazonas (AM) e Ceará (CE) apresentam resultados acima da média geral, destacando o estado da Paraíba que chega a apresentar IQVT até 9.9% maior que a média nacional para empresas de pequeno porte com renda média de *R*$500.00. Por outro lado, os estados Mato Grosso (MT), Roraima (RR), Rondônia (RO) e o Distrito Federal (DF) apresentam resultados abaixo da média geral, destacando o Distrito Federal

que apresenta IQVT até 11.43% menor que a média geral.

A Tabela 4 também mostra que na renda de *R*$500.00 a diferença entre os estados e a média geral são maiores, já quando aumentamos a renda para *R*$2500.00 estas diferenças tendem a diminuir, mostrando que quando a renda aumenta tanto a covariável porte como o estado perdem importância. Porém para renda baixa, o porte e o estado da empresa ganham importância para explicar o IQVT. Este resultado é coerente com a realidade brasileira, uma vez que, para trabalhadores de baixa renda o poder público fornece politicas públicas de apoio, para citar algumas: Sistema Único de Assistência Social (SUAS), Agente Jovem, Segurança Alimentar e Social (SAN), banco de alimentos, restaurante popular, cozinha comunitária, Saúde Familia, fundo de manutenção e desenvolvimento da educação, para mais programas sociais do governo brasileiro consulte [2]. Todos estes programas de alguma forma contribuem para a melhoria da Qualidade de Vida dos Trabalhadores com baixa renda. O mesmo se aplica para a realidade interna da empresa, de forma geral no Brasil quando o trabalhador tem uma renda menor, a empresa fornece alimentação, auxílio transporte, cesta básica entre outros benefícios que tornam o local de trabalho importante para a Qualidade de Vida do Trabalhador.

Por outro lado quando o trabalhador tem uma renda maior em geral ele é menos dependente de tais benefícios sendo a sua renda a principal mantenedora da sua Qualidade de Vida, ganhando assim mais importância perante as outras condições, nesse caso o porte e o estado. Desta forma, o modelo apresenta resultados compatíveis com a realidade social brasileira. Para finalizar a análise a Figura 7 apresenta os dados observados e o ajuste do modelo beta com intercepto aleatório, separado por porte das indústrias.

Pelos gráficos apresentados na Figura 7 é possível ver que o IQVT encontra-se concentrado em valores na faixa de 0.35 até 0.80. Verifica-se também que nesta faixa o relacionamento do IQVT com a log-renda centrada não é muito diferente do linear. De forma geral, o modelo ajustado segue o padrão dos dados. Alguns pontos atípicos aparecem no estado Mato Grosso nas pequenas empresas, porém não tem grande influência no ajuste geral do modelo.

4.2. Qualidade da água em reservatórios operados pela COPEL no estado do Paraná/BR.

A companhia Paranaense de Energia (COPEL) opera no estado do Paraná, dezeseis usinas hidroelétricas, com geração total de mais de 4.500 MW de energia. Os reservatórios constituídos para a geração de energia elétrica têm sido utilizados para inúmeras outras finalizadas, destacando lazer, navegação e captação de água para abastecimento público. A qualidade da água, por si só e como determinante do crescimento de algas, plantas e outros organismos, é fundamental para que tais usinas apresentem máxima funcionalidade.

Pensando em abastecimento de água e no impacto que tais empreendimentos tem sobre o meio-ambiente, é fundamental saber como tais reservatórios atuam sobre a qualidade da água. Neste sentido, a COPEL realiza o monitoramento dos reservatórios, bem como dos rios represados a montante e jusante dos mesmos, em atendimento às condicionantes das licenças de operação destes empreendimentos.

O monitoramento realizado pela concessionária, envolve a avaliação de parâmetros da qualidade da água e o cálculo do Índice de Qualidade da Água - CETESB (IQA), que

[2] www.portaltransparencia.gov.br

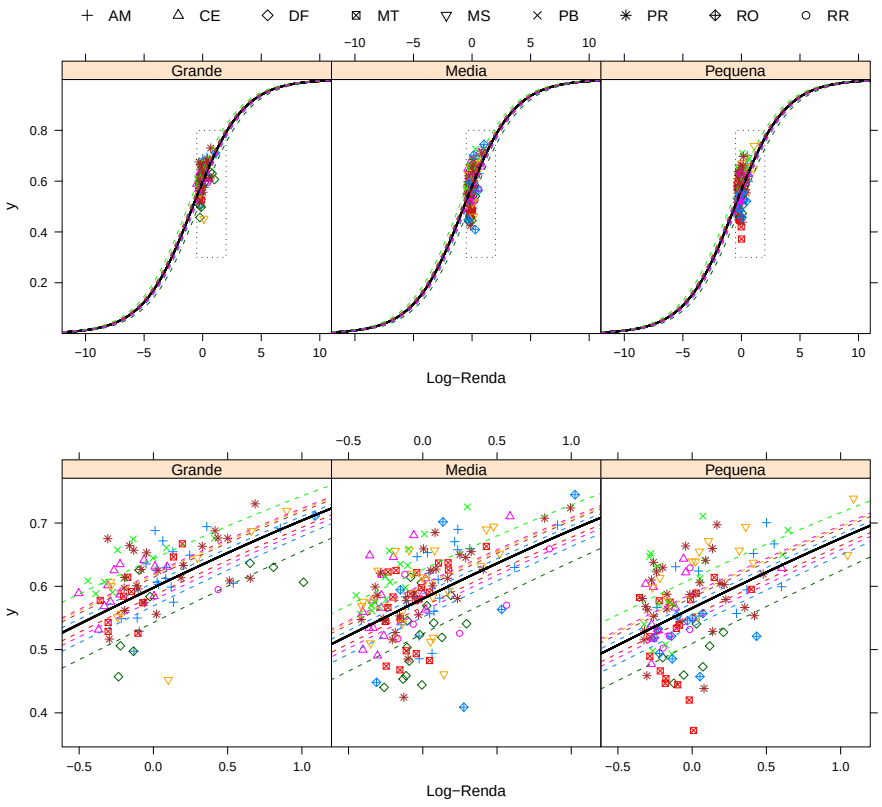

Figura 4: Valores observados e modelo ajustado.

será a variável chave nesta análise. A partir de um estudo realizado em 1970 pela National Sanitation Foundation dos Estados Unidos da América, a CETESB (Companhia de Tecnologia de Saneamento Ambiental) adaptou e desenvolveu o IQA - Índice de Qualidade das Águas, que incorpora nove parâmetros (Oxigênio Dissolvido, Temperatura, Coliformes fecais, pH, DBO, Nitrogênio Total, Fósforo Total, Turbidez, Sólidos totais) considerados relevantes para a avaliação da qualidade das águas, tendo como determinante principal a utilização da mesma para abastecimento público. Por construção o IQA resulta sempre entre valores no intervalo unitário, o que novamente torna o modelo beta uma ferramenta analitica adequada.

O objetivo da análise estatística de dados históricos do monitoramento da qualidade das águas dos reservatórios de usinas hidrolétricas operadas pela COPEL, no estado do Paraná, é identificar possíveis impactos e alterações na qualidade da água decorrente da existência dos reservatórios. A qualidade da água do rio pode ser definida pelas características da estação de montante, onde ainda não há influência do empreendimento no curso da água. Desta forma, utilizou-se dados da estação de montante como situação de referência a qual será comparada com os dados das estações de reservatório e jusante, verificando assim a melhora ou piora da qualidade da água após a passagem pelo reservatório.

O monitoramento realizado pela Concessionária envolve a avaliação de parâmetros da qualidade da água, bem como o cálculo do Índice de Qualidade da Água (IQA), a partir de campanhas realizadas trimestralmente nas dezeseis usinas em funcionamento. Para a análise apresentada foi selecionado o ano de 2004. Temos como covariáveis o LOCAL da coleta (Montante, Reservatório e Jusante) que é a de principal interesse. Além disso, temos 16 usinas e quatro trimestres, que não são de interesse direto porém afetam a variável resposta IQA. Em cada usina foram realizadas 12 observações (4 trimestres em 3 locais), temos duas observações perdidas totalizando 190 amostras para a análise.

A Figura 5 apresenta um resumo gráfico do conjunto de dados.

De acordo com o histograma (5 A) é clara a assimetria a esquerda comum em dados no intervalo unitário. A Figura 5 B mostra claramente que as usinas tem comportamentos diferentes com relação ao IQA. Com relação a covariável LOCAL (Figura 5C), é possível identificar graficamente que o IQA aumenta quando passa-se da Montante para o Reservatório e diminui do Reservatório para a Jusante. Com relação ao trimestre (5 D) o comportamento deve ser ciclico, baixo nos meses de verão (primeiro e quarto semestre) e maior nos meses de inverno (segundo e terceiro semestre).

Temos interesse principal na covariavel LOCAL, porém não podemos desprezar o efeito das condições ambientais representadas pela usina que está sendo avaliada e o período do ano (trimestre) que como mostra os gráficos da Figura 5 deve ter um efeito não desprezível sobre a resposta. Além disso, temos 16 usinas considerar este efeito como sendo fixo irá inflacionar o modelo de parâmetros que poderá influenciar nas estimativas da covariável de principal interesse neste caso LOCAL. Além disso, supor que o efeito de trimestre é o mesmo para todas as usinas não parece ser uma suposição plausível e deverá ser testada, o mais compatível com os dados deve ser um modelo que contemple diferentes efeitos de trimestre dependendo da usina, ou seja, uma interação entre usina e trimestre.

Nesta situação, no caso de considerar todos os efeitos como fixos teríamos que estimar pelo menos 80 parâmetros o que parece muito para estimar com 190 observações. Neste

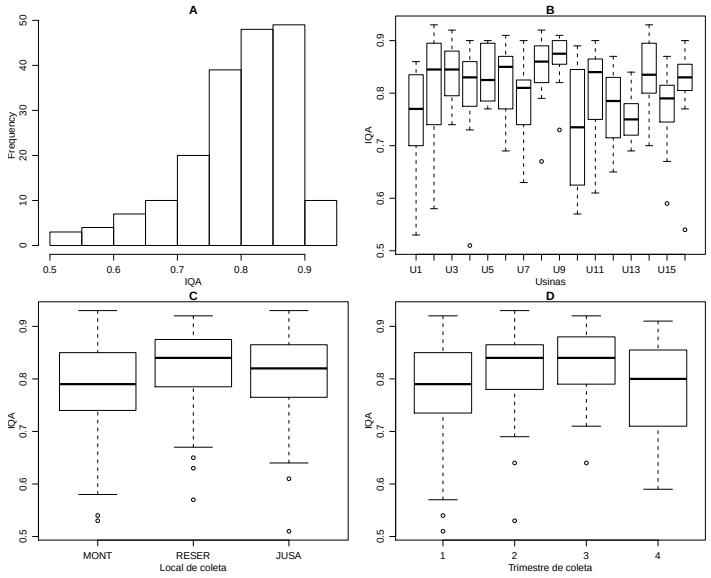

Figura 5: Análise descritiva para o Índice de Qualidade da Água.

exemplo, a vantagem em usar um modelo de efeitos aleatórios é bastante aparente, uma vez que o modelo é mais parcimonioso e permite levar todos os efeitos em consideração, além de testar as hipóteses de interesse.

Tendo estas condições em mente propomos o seguinte modelo geral para descrever o IQA e testar as hipóteses de interesse:

- $Y_{ijt} \sim B(\mu_{ijt}, \phi)$
- $g(\mu_{ijt}) = \beta_0 + \beta_{1,i} + \beta_{2,t} + b_j + b_{j,t}$
- $b_j \sim N(0, \tau_U^2)$
- $b_{jt} \sim N(0, \tau_{UT}^2)$

onde i representa os locais de coleta, j as 16 diferentes usinas e t os quatro trimestres de observação. O termo $\beta_{1,i}$ com $i = 2, 3$ representa o efeito ao mudar da Montante para Reservatório e Jusante respectivamente. O termo $\beta_{2,k}$ com $k = 2, 3, 4$ representa o efeito ao mudar do primeiro para o segundo, terceiro e quarto trimestre. O termo b_j é um intercepto aleatório representando desvios da média geral para cada usina. O último termo $b_{j,t}$ representa um desvio do efeito de trimestre para cada usina em cada trimestre.

Para comparação consideramos alguns casos particulares do modelo geral que representam hipóteses de interese. O primeiro modelo a ser considerado, fixamos $\beta_{1,i}, \beta_{2,j}, \tau_U^2, \tau_{UT}^2 = 0$, ou seja o modelo apenas com intercepto é o ponto inicial. O segundo modelo estimamos $\beta_{1,i}$, e fomos incluindo os termos um de cada vez até chegar ao modelo completo o que resultou em seis modelos para comparação. As estimativas pontuais são apresentadas na Tabela 5, novamente a função de ligação é *logit*, o método para resolução das integrais contidas na verossimilhança foi o de Laplace, o algoritmo de maximização o BFGS.

Tabela 5: Estimativas pontuais, logaritmo da verossimilhança maximizada e critério de informação de Akaike.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6
β_0	1.3959	1.2717	1.1401	1.1356	1.1523	1.1508
β_{12}		0.2271	0.2289	0.2370	0.2400	0.2411
β_{13}		0.1517	0.1477	0.1636	0.1542	0.1582
β_{22}			0.2056	0.2189	0.2226	0.2217
β_{23}			0.2925	0.3093	0.3165	0.3166
β_{24}			0.0475	0.0521	0.0579	0.0581
ϕ	23.3554	24.2535	25.7847	30.4741	42.1862	42.2040
τ^2_U				28.9672		43.5394
τ^2_{UT}					11.1887	15.0378
ll	156.5696	160.0943	165.8117	172.2348	178.2719	179.3810
AIC	-309.1392	-312.1886	-317.6233	-328.4696	-340.5437	-340.7619

Pelos resultados apresentados na Tabela 5 verifica-se o aumento da logLik com a inclusão dos diversos termos no modelo. Como os modelos são aninhados podemos fazer uso do teste da razão de verossimilhança para testar a significância dos diversos termos, partindo do modelo mais simples (Model 1) até o modelo mais complexo (Model 6). A maior logLik é a do Model 6, porém a diferença para o Model 5 é de apenas 1.1091, conduzindo o teste da razão de verossimilhança temos um p-valor de 0.1363 mostrando que a inclusão do intercepto aleatório por usina é desnecessário e consequentemente o parâmetro τ^2_U pode ser removido do modelo.

É importante ressaltar que este modelo apresenta alta complexidade para a estimação por verossimilhança marginal, uma vez que a integral que precisa ser resolvida para cada reservatório tem cinco dimensões. Isso torna algumas técnicas comuns de integração numérica como Gauss-Hermite e Monte Carlo, praticamente não aplicáveis nesta situação. O procedimento via Laplace apresenta dificuldades principalmente com relação ao tempo computacional e acurácia da aproximação do valor da verossimilhança maximizada que é de fundamental importância para comparação dos modelos e testes de hipóteses.

Como método alternativo que não faz uso de integração e maximização numérica foi utilizado o algoritmo *data clone*. Ajustamos o modelo 5, via este algoritmo e comparamos os resultados com o obtido via verossimilhança marginal em termos de estimativas pontuais e desvio padrões, os resultados são apresentados na Tabela 6.

Os resultados apresentados na Tabela 6 mostram que os dois algoritmos de inferência obtiveram resultados bastante similares para os parâmetros de média. A mesma similaridade não ocorre na estimativa dos desvios padrões, de forma geral o método de máxima verossimilhança marginal estimou desvios padrões menores que o algoritmo *data clone*. No caso desta análise o algoritmo numérico apresentou bastante dificuldade em obter o hessiano numérico, sendo que diversos ajustes no algoritmo de diferenças finitas foram necessários para obter estimativas válidas (dentro do espaço paramétrico), mesmo assim as estimativas ficaram bastante diferentes das obtidas pelo algoritmo *data clone*, que apresenta um comportamento robusto ao custo de um tempo computacional muito maior.

Apesar dos desvios padrões terem sido calculados o seu uso para a construção dos

Tabela 6: Point estimates and standard errors for Model 5 obtained by maximasation of the marginal likelihood and data-clonning.

	Pontual Marg	SD Marg	Pontual dClone	SD dClone
β_0	1.1523	0.0880	1.1539	0.1017
β_{12}	0.2400	0.0480	0.2399	0.0678
β_{13}	0.1542	0.0102	0.1542	0.0669
β_{22}	0.2226	0.0072	0.2209	0.1343
β_{23}	0.3165	0.0290	0.3150	0.1330
β_{24}	0.0579	0.0069	0.0571	0.1313
ϕ	42.1862	4.1431	42.2961	5.3180
τ^2_{UT}	11.1887	3.3077	10.9926	3.1216

intervalos de confiança é indesejável principalmente para os parâmetros de precisão, uma vez que a aproximação quadrática em geral apresenta resultados ruins para estimativas deste tipo. Sendo assim, para os parâmetros ϕ e τ_{UT}, obtivemos os intervalos baseado em perfil de verossimilhança. Para concluir o processo de estimação, a Figura 6 apresenta os gráficos de perfil de verossimilhança para os dois parâmetros de precisão, importante destacar que os gráficos são apresentados para os parâmetros reparametrizados (em escala logaritmica), para eficiência computacional. Também é apresentado o diagnóstico de estimabilidade destes parâmetros através das técnicas gráficas, conforme descrito na Seção 2.2.

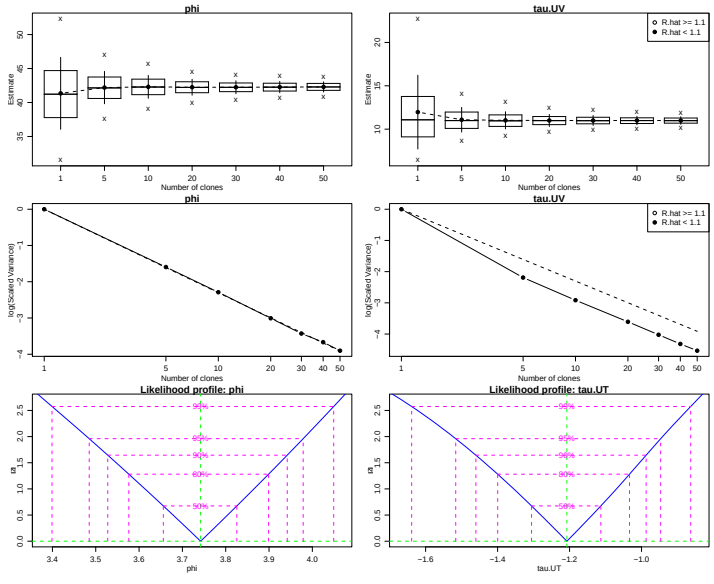

Figura 6: Profile likelihoods for precision parameters and identifiability diagnostics for Model 5.

Os gráficos apresentados na Figura 6 mostram que o parâmetro τ^2_{uT} apresenta forma assimétrica, mesmo na reparametrização utilizada. É possível ver também que este

parâmetro é mais sensível a escolha da priori (ver Boxplot's) que o parâmetro ϕ. Pelos gráficos da log variância escalonada verificamos que a variância para o parâmetro que indexa o efeito aleatório tem um decaimento maior do que o esperado de acordo com o número de clones utilizados na estimação, porém o maior autovalor da matriz de variância-covariância apresentou sempre valores menores que 1.1 indicando que o parâmetro está sendo adequadamente estimado.

Uma vez que o modelo foi escolhido o processo de estimação teve sucesso, todos os parâmetros são estimáveis o intervalo perfilhado para a estimativa do parâmetro de precisão do efeito aleatório indica forte significância deste, podemos fazer a predição dos efeitos aleatórios pela metodologia Bayes Empirica, os resultados da predição são mostrados na Figura 7 sobreposto aos valores observados, de acordo com o local de coleta.

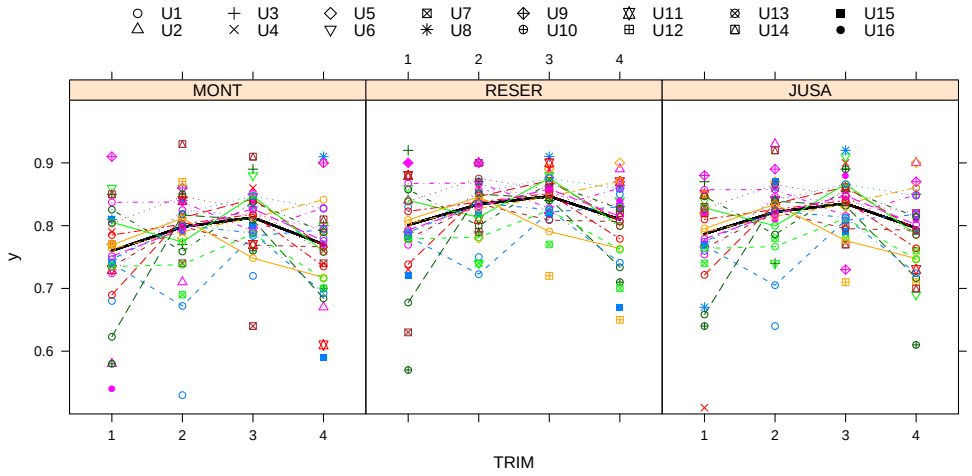

Figura 7: Valores observados e modelo ajustado.

Os resultados do ajuste do modelo indicam que ao passar do ponto de coleta a montante para o ponto dentro do reservatório espera-se um aumento médio no IQA de 5.39%, e da montante para a jusante o aumento é de 3.55%, mantendo as outras condições no estado zero. Com relação aos trimestres verifica-se que os meses de verão (primeiro e quarto) trimestres são os que apresentam os menores valores de IQA. Os resultados indicam que existe um efeito cíclico para o trimestre com queda do IQA nos meses quentes e aumento para os meses de inverno. Os efeitos aleatórios indicam que estes resultados médios são bastante afetados de acordo com o reservatório e trimestre de observação, mostrando uma clara interação entre estes efeitos. De forma geral, podemos concluir que a presença do reservatório proporciona uma melhoria na Qualidade da Água, porém após a passagem pelo reservatório o IQA tende a voltar ao estado original.

O ajuste do modelo mostra-se satisfatório suavizando algumas observações aparentemente atípicas principalmente na estação de montante nos primeiros dois trimestres. Os resultados indicam que os efeitos de trimestres parecem apresentar um comportamento dependente no tempo, porém com apenas um ano de observação não é possível concluir

se isto será persistente no decorrer do período. Neste caso supomos que os efeitos de trimestre são realizações da mesma Normal Multivariada, porém independentes a inclusão de uma estrutura de dependencia, por exemplo, do tipo autoregressiva pode vir a melhorar o ajuste, porém não consideramos esta possibilidade nesta análise.

A suposição de que as quatro realizações são da mesma Normal Multivariada (mesma precisão), pode ser inadequada já que graficamente os trimestres primeiro e quarto mostram maior amplitude no IQA, porém a inclusão de mais três componentes de variância/precisão inflacionária a quantidade de parâmetros no modelo. Consideramos que esta análise poderá ser feita com a replicação do experimento para diferentes anos, contemplando também a possível interação de trimestre com local de coleta. A complexidade de tal modelo necessita de uma melhor avaliação do método de estimação, num modelo com tal estrutura a metodologia aqui apresentada deve ser inadequada pela alta dimensão do vetor de efeitos aleatórios. Métodos mais gerais tais como quasi-likelihood precisam ser adaptados e avaliados para o caso de regressão beta com efeitos aleatórios. Metodologias Bayesianas via simulação (MCMC) e aproximadas como proposto em [26] também precisam ser adaptadas para o caso de regressão beta com efeitos aleatórios.

Pela abordagem Bayesiana é necessário ter uma avaliação do impacto da escolha das distribuições a priori sobre as estimativas, uma vez que mesmo nas situações simples aqui apresentadas a escolha da priori afeta consideravelmente as estimativas da precisão dos efeitos aleatórios, e isso deve ser agravado em dimensões maiores. Neste caso o algoritmo *data clone* é uma ferramenta de fundamental importância, por atentar para tais efeitos e indicar a real estimabilidade dos parâmetros componentes do modelo.

5. Conclusões

Este artigo explora a construção e algoritmos para inferência em modelos de regressão beta com efeitos aleatórios para modelar variáveis respostas que assumem valores no intervalo unitário. Como ilustração da aplicação do modelo proposto foram analisados dois conjuntos de dados reais, referentes a fatores associados ao Índice de Qualidade de Vida do Trabalhador da Indústria Brasileira, sendo uma aplicação voltada a área de ciência sociais e econômicas. O segundo conjunto de dados referente ao Índice de Qualidade da água (IQA) em reservatórios operados pela COPEL no estado do Paraná/BR ilustra uma aplicação na área de engenharia ambiental e áreas relacionadas. Os exemplos foram cuidadosamente escolhidos para demonstrar a aplicação do modelo proposto em áreas bastante distintas. De forma geral, o modelo beta com efeitos aleatórios mostrou-se adequado para analisar ambos os conjuntos de dados. A estimação dos parâmetros envolvidos no modelo foi realizada pelo paradigma de verossimilhança utilizando dois algoritmos distintos, maximização da verossimilhança marginal com aproximação de Laplace para integração dos efeitos aleatórios e clonagem de dados com inferência via algoritmos de MCMC.

Aspectos importantes para inferência na classe de modelos proposto foram explorados e implementados, tais como, diagnósticos de estimabilidade a partir da clonagem de dados e a obtenção de intervalos de confiança baseados em perfis de verossimilhança principalmente para parâmetros de precisão que indexam as distribuições dos efeitos aleatórios, já que, intervalos baseados na aproximação quadrática da verossimilhança foram claramente inadequados produzindo estimativas intervalares fora do espaço paramétrico em alguns casos.

Os resultados da análise para os dados referentes ao Índice de Qualidade de Vida dos Trabalhadores da Indústria Brasileira, mostraram que as covariáveis Porte da empresa e Renda dos trabalhadores são importantes características para explicar o IQVT dos trabalhadores de uma determinada indústria. Estes resultados são compatíveis com a realidade social brasileira, desta forma o modelo estatístico colabora com a construção e validação de teorias sociais relacionadas a Qualidade de Vida dos Trabalhadores industriais.

Com relação ao segundo conjunto de dados referente ao Índice de Qualidade da Água, o interesse principal era quantificar o impacto que os reservatórios têm sobre o IQA. O ajuste do modelo beta permitiu levar em consideração as condições ambientais complexas em que o experimento de campo foi conduzido, bem como responder a questão de interesse, que mostrou o reservatório como tendo um impacto benéfico no índice de qualidade da água. A inclusão dos efeitos aleatórios neste caso também colaborou para a construção de um modelo parcimonioso.

O modelo utilizado não é comum na literatura e não foram encontradas implementações computacionais específicas. Sendo assim, rotinas próprias para os ajustes foram desenvolvidas e são disponibilizadas no apêndice computacional de forma breve e totalmente aberta nos complementos *on line*. O ajuste do modelo proposto apresenta dificuldades computacionais e requer a escolha de métodos de integração numérica, já que a verossimilhança toma a forma de uma integral sob os efeitos aleatórios que não possui solução analítica e um algoritmo de maximização numérica para a obtenção das estimativas de máxima verossimilhança e posterior construção de intervalos de confiança através do perfilhamento da verossimilhança. Para a integração numérica foi utilizado o método de Laplace o algoritmo de maximização escolhido foi o BFGS com implementação utilizando o ambiente R para computação estatística.

É esperado que a metodologia de verossimilhança marginal apresente dificuldades em modelos onde o vetor de efeitos aleatórios seja grande (> 5) o que foi o caso no exemplo do índice de qualidade da água (IQA), no qual um efeito de interação induz um vetor aleatório de cinco dimensões. Os algoritmos produziram estimativas pontuais adequadas, porém não foi capaz de calcular adequadamente o hessiano numérico para alguns dos parâmetros do modelo.

O algoritmo clonagem de dados é relativamente novo e mostrou resultados promissores com um esforço de programação menor que o método de verossimilhança marginal. Entretanto o tempo computacional que depende do número de clones utilizados pode ser elevado para os ajustes apresentados no artigo. Por outro lado a disponibilidade e crescente facilidade de acesso a *clusters* computacionais, algoritmos e recursos de paralelização e máquinas com múltiplos núcleos tendem a reduzir o impacto no tempo computacional, por exemplo rodando múltiplas cadeias MCMC em paralelo. Estimativas obtidas pela maximização da verossimilhança marginal e clonagem de dados são muito próximas assim como erros padrão e estimação intervalar, com as restrições já mencionadas à aproximação quadrática. A construção de intervalos perfilhados e o cálculo da verossimilhança maximizada utilizando a clonagem de dados não foram utilizadas aqui mas são igualmente possíveis como em [27].

Apesar do modelo proposto ser bastante geral, o procedimento de inferência baseado na marginalização da verossimilhança leva a dificuldades numéricas e computacionais com efeitos aleatórios não independentes e de alta dimensionalidade. Por exemplo, modelos para dados espaciais requerem integração numérica de alta dimensão o que torna proibitivo a obtenção da verossimilhança marginal por métodos de integração numérica.

O algoritmo de clonagem de dados pode ser utilizado, porém os algoritmos MCMC em geral não estão livres de problemas de convergência, mistura da cadeia entre outras que merece melhor avaliação em especial nestes contextos.

A abordagem Bayesiana para inferência está incluída nas análises de dados clonados com $K = 1$. Os resultados obtidos aqui apontam o impacto da escolha das distribuições a priori sobre as estimativas, uma vez que mesmo para os modelos utilizados aqui com pouca complexidade na estrutura dos efeitos aleatórios, a escolha de prioris afetou consideravelmente as estimativas da precisão dos efeitos aleatórios, o que deve ser agravado com maiores dimensionalidades. Abordagem Bayesiana via simulação pelo uso de algoritmos MCMC e algoritmo de clonagem da dados compartilham portanto os mesmos potenciais obstáculos de tempo computacional e necessidade de verificação de convergência.

Uma abordagem alternativa é o uso de técnicas de inferência Bayesiana aproximada como proposto por [26] que em diversas situações tem se mostrado extremamente acurada e rápida, porém de implementação altamente especializada, requerendo adaptações para o caso de modelos de regressão beta com efeitos aleatórios. Além disso, o uso de priori's impróprias como é comum no algoritmo INLA - *Integrated Nested Laplace Approximation* para representar efeitos espaciais e temporais, requer verificação da validade da distribuição a posteriori, quando associado a distribuições não convencionais como a beta.

Apêndice A. Programando modelo Beta com efeito aleatório

Neste apêndice apresentamos os principais passos para a construção de um modelo beta com efeitos aleatórios gaussianos. Todas as rotinas foram desenvolvidas na linguagem *R* e encontram-se disponíveis para uso público. Este apêndice foi escrito para ser auto suficiente no sentido de descrever uma análise completa.

Sendo assim, vamos trabalhar com dados simulados, por simplicidade será tratado o modelo Beta com apenas um efeito aleatório, porém em uma estrutura fácil de ser generalizada para efeitos de maior dimensão. Com isso, perde-se em eficiência computacional que no caso de efeito em uma dimensão pode ser feita de forma mais eficiente. O código abaixo apresenta uma função para simular de um modelo Beta com efeito aleatório Gaussiano. O modelo simulado terá apenas três parâmetros $\underline{\theta} = (\beta, \phi, \tau)$, onde β é o intercepto do modelo, ϕ precisão das observações Beta e τ precisão do efeito aleatório de indivíduo. Será simulado 10 blocos com 10 observações cada.

```
> simula.beta <- function(para, n.bloco = 10, n.rep = 5) {
+     bloco <- rep(1:n.bloco, each = n.rep)
+     bloco.efeito <- rep(rnorm(n.bloco, 0, sd = 1/para[3]),
+         each = n.rep)
+     eta <- para[1] + bloco.efeito
+     mu <- exp(eta)/(1 + exp(eta))
+     y <- rbeta(length(mu), mu * para[2], (1 - mu) *
+         para[2])
+     return(data.frame(y = y, ID = bloco))
+ }
```

Simulando o conjunto de dados

```
> set.seed(123)
> dados <- simula.beta(para = c(0.5, 50, 10), n.bloco = 10,
+     n.rep = 10)
```

O próximo passo é fazer uma função que calcule a aproximação de Laplace para uma função genérica, isso é feito no código abaixo.

```
> laplace <- function(funcao, otimizador, n.dim, ...) {
+     integral <- -sqrt(.Machine$double.xmax)
+     inicial <- rep(0, n.dim)
+     temp <- try(optim(inicial, funcao, ..., method = otimizador,
+         hessian = TRUE, control = list(fnscale = -1)))
+     if (class(temp) != "try-error") {
+         integral <- exp(temp$value) * (exp(0.5 *
+             log(2 * pi) - 0.5 * determinant(-temp$hessian)$modulus))
+     }
+     return(integral)
+ }
```

Para facilitar vamos escrever uma função generica onde se passa o modelo escrito de uma forma adequada e ela é capaz de avaliar a verossimilhança marginal, usando a aproximação de Laplace.

```
> verossimilhanca <- function(modelo, formu.X, formu.Z,
+     beta.fixo, prec.pars, otimizador, n.dim, dados) {
+     dados.id <- split(dados, dados$ID)
+     ll <- c()
+     for (i in 1:length(dados.id)) {
+         X <- model.matrix(as.formula(formu.X), data = dados.id[[i]])
+         Z <- model.matrix(as.formula(formu.Z), data = dados.id[[i]])
+         ll[i] <- laplace(modelo, otimizador = otimizador,
+             n.dim = n.dim, X = X, Z = Z, Y = dados.id[[i]]$y,
+             beta.fixo = beta.fixo, prec.pars = prec.pars,
+             log = TRUE)
+     }
+ }
```

Definindo a função de ligação e a sua inversa.

```
> inv.logit <- function(x) {
+     exp(x)/(1 + exp(x))
+ }
```

Escrevendo o modelo Beta para ser passado para a função *verossimilhanca*, note que estamos usando uma reparametrização para facilitar o procedimento numérico. Estamos estimando o $\log(\phi)$ e $\log(\tau)$, o procedimento de volta é feito usando a propriedade de invariância dos estimadores de máxima verossimilhança. Para a construção de intervalos assintóticos usamos o método Delta. Não vamos explicar isso aqui, para o leitor interessado recomendamos a bibliografia referenciada.

```
> modelo <- function(uv, beta.fixo, prec.pars, X, Z,
+     Y, log = TRUE) {
+     phi <- exp(prec.pars[1])
+     tau <- exp(prec.pars[2])
+     if (class(dim(uv)) == "NULL") {
+         uv <- matrix(uv, 1, length(uv))
+     }
+     ll = apply(uv, 1, function(uvi) {
+         preditor <- X %*% beta.fixo + Z %*% as.numeric(uvi)
+         mu <- inv.logit(preditor)
+         sum(dbeta(Y, mu * phi, (1 - mu) * phi, log = TRUE)) +
+             dnorm(uvi[1], 0, sd = 1/tau, log = TRUE)
+     })
+     if (log == FALSE) {
+         ll <- exp(ll)
+     }
+     return(ll)
+ }
```

Com isto, temos o suficiente para escrever a verossimilhança do modelo.

```
> dados$intercepto <- 1
> modelo.ajuste <- function(b1, phi, tau, n.dim, otimizador,
+     dados) {
+     ll = verossimilhanca(modelo = modelo, formu.X = "~ intercepto -1 ",
+         formu.Z = "~ intercepto -1", beta.fixo = b1,
+         prec.pars = c(phi, tau), otimizador = otimizador,
+         n.dim = n.dim, dados = dados)
+     return(-ll)
+ }
```

Obtendo as estimativa de máxima verossimilhança numéricamente, para facilitar vamos usar o pacote *bbmle* que usa internamente a função *optim()*, o algoritmo escolhido foi o BFGS.

```
> ajuste.1 = mle2(modelo.ajuste, start = list(b1 = 0,
+     phi = log(1), tau = log(1)), data = list(otimizador = "BFGS",
+     n.dim = 1, dados = dados))
```

Mostrando as estimativas pontuais e erros padrões e intervalos de confiança assintóticos.

```
> summary(ajuste.1)
> confint(ajuste.1, method = "quad")
```

Podemos facilmente obter os perfis de verossimilhança,

```
> perfil.beta <- profile(ajuste.1)
```

Com isso, mostramos rapidamente os passos para a programação de um modelo Beta com efeito aleatório gaussiano.

Referências

[1] R. Kieschnick, B. D. McCullough, Regression analysis of variates observed on (0, 1): percentages, proportions and fractions, Statistical Modelling 3 (3) (2003) 193–213. doi:10.1191/1471082X03st053oa.
URL http://smj.sagepub.com/cgi/content/abstract/3/3/193

[2] P. Paolino, Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables, Political Analysis 9 (4) (2001) 325–346.
URL http://pan.oxfordjournals.org/cgi/content/abstract/9/4/325

[3] S. Ferrari, F. Cribari-Neto, Beta Regression for Modelling Rates and Proportions, Journal of Applied Statistics 31 (7) (2004) 799–815. doi:10.1080/0266476042000214501.
URL http://www.tandfonline.com/doi/abs/10.1080/0266476042000214501

[4] J. A. Nelder, R. W. M. Wedderburn, Generalized linear models, Journal of the Royal Statistical Society. Series A (General) 135 (3) (1972) pp. 370–384.
URL http://www.jstor.org/stable/2344614

[5] A. B. Simas, W. Barreto-Souza, A. V. Rocha, Improved estimators for a general class of beta regression models, Computational Statistics & Data Analysis 54 (2) (2010) 348–366. doi:10.1016/j.csda.2009.08.017.
URL http://linkinghub.elsevier.com/retrieve/pii/S0167947309003107

[6] P. L. Espinheira, S. L. Ferrari, F. Cribari-Neto, On beta regression residuals, Journal of Applied Statistics 35 (4) (2008) 407–419. doi:10.1080/02664760701834931.
URL http://dx.doi.org/10.1080/02664760701834931

[7] P. Espinheira, S. Ferrari, F. Cribari-Neto, Influence diagnostics in beta regression, Computational Statistics & Data Analysis 52 (9) (2008) 4417–4431. doi:10.1016/j.csda.2008.02.028.
URL http://dl.acm.org/citation.cfm?id=1367150.1367411

[8] A. V. Rocha, A. B. Simas, Influence diagnostics in a general class ofbeta regression models, TEST 20 (1) (2010) 95–119. doi:10.1007/s11749-010-0189-z.
URL http://www.springerlink.com/content/g331633302358003/

[9] K. L. P. Vasconcellos, F. Cribari-Neto, Improved maximum likelihood estimation in a new class of beta regression models, Statistics (2005) 13–31.

[10] R. Ospina, F. Cribari-Neto, K. L. Vasconcellos, Improved point and interval estimation for a beta regression model, Computational Statistics & Data Analysis 51 (2) (2006) 960–981. doi:10.1016/j.csda.2005.10.002.
URL http://dl.acm.org/citation.cfm?id=1647962.1648200

[11] A. J. Branscum, W. O. Johnson, M. C. Thurmond, Baysian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses, Australian & New Zealand Journal of Statistics 49 (3) (2007) 287–301. doi:10.1111/j.1467-842X.2007.00481.x.
URL http://doi.wiley.com/10.1111/j.1467-842X.2007.00481.x

[12] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2009).
URL http://www.R-project.org

[13] F. Cribari-neto, A. Zeileis, Beta regression in R, Journal Of Statistical Software 34 (2).

[14] E. McKenzie, An Autoregressive Process for Beta Random Variables.
URL http://www.jstor.org/pss/2631528

[15] G. K. Grunwald, A. E. Raftery, P. Guttorp, Times Series of Continuous Proportions, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 9 (4) (1993) 586 – 587.
URL http://ideas.repec.org/a/eee/intfor/v9y1993i4p586-587.html

[16] C. Da-Silva, H. Migon, L. Correia, Dynamic Bayesian beta models, Computational Statistics & Data Analysis 55 (6) (2011) 2074–2089. doi:10.1016/j.csda.2010.12.011.
URL http://dl.acm.org/citation.cfm?id=1951000.1951322

[17] J. C. Pinheiro, D. M. Bates, Mixed Effects Models in S and S-Plus, Springer, 2000.
URL http://www.amazon.com/Mixed-Effects-Models-S-S-Plus/dp/0387989579

[18] S. R. Lele, B. Dennis, F. Lutscher, Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods., Ecology letters 10 (7) (2007) 551–63. doi:10.1111/j.1461-0248.2007.01047.x.
URL http://doi.wiley.com/10.1111/j.1461-0248.2007.01047.x

[19] S. R. Lele, Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning, Journal of the American Statistical Association 105 (492) (2010) 1617–1625.

doi:10.1198/jasa.2010.tm09757.
URL http://pubs.amstat.org/doi/abs/10.1198/jasa.2010.tm09757

[20] G. Molenberghs, G. Verbeke, Models for Discrete Longitudinal Data (Springer Series in Statistics), Springer, 2005.
URL http://www.amazon.com/Models-Discrete-Longitudinal-Springer-Statistics/dp/0387251448

[21] R. H. Byrd, A Limited Memory Algorithm for Bound Constrained Optimization, SIAM Journal on Scientific Computing 35 (5) (1995) 773. doi:10.1137/0916069.

[22] C. Robert, G. Casella, Monte Carlo Statistical Methods (Springer Texts in Statistics), Springer, 2004.
URL http://www.amazon.com/Monte-Statistical-Methods-Springer-Statistics/dp/0387212396

[23] S. R. Lele, Model complexity and information in the data: Could it be a house built on sand?, Ecology 91 (12) (2010) 3493–3496. doi:10.1890/10-0099.1.
URL http://www.esajournals.org/doi/abs/10.1890/10-0099.1

[24] P. Sólymos, dclone: Data cloning in R, The R Journal 2 (2) (2010, in press) 29–37, R package version: 1.3-0.
URL http://journal.r-project.org/

[25] M. Plummer, Jags: A program for analysis of bayesian graphical models using gibbs sampling (2003).

[26] H. v. Rue, S. Martino, N. Chopin, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2) (2009) 319–392. doi:10.1111/j.1467-9868.2008.00700.x.
URL http://doi.wiley.com/10.1111/j.1467-9868.2008.00700.x

[27] J. M. Ponciano, M. L. Taper, B. Dennis, S. R. Lele, Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning, Ecology 90 (2) (2009) 356–362. doi:10.1890/08-0967.1.
URL http://www.esajournals.org/doi/abs/10.1890/08-0967.1