Modelo de regressão quantílica bayesiano para estimação em pequenos domínios a nível de unidade

Clarissa de Oliveira Cavalcanti

Orientadora: Kelly Cristina Mota Gonçalves

Universidade Federal do Rio de Janeiro
Instituto de Matemática
Departamento de Métodos Estatísticos

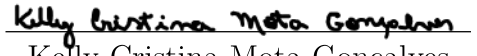
2021

Modelo de regressão quantílica bayesiano para estimação em pequenos domínios a nível de unidade

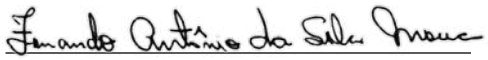
Clarissa de Oliveira Cavalcanti

Dissertação de mestrado submetida ao Departamento de Métodos Estatísticos do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Aprovado por:



Kelly Cristina Mota Gonçalves
IM - UFRJ



Fernando A. S. Moura
IM-UFRJ

Valmária Rocha da Silva Ferraz
CCN - UFPI

Rio de Janeiro, RJ, 14 de dezembro
2021

CIP - Catalogação na Publicação

C376m

Cavalcanti, Clarissa Oliveira
Modelo de Regressão Quantílica bayesiana para estimação em pequenos domínios a nível de unidade / Clarissa Oliveira Cavalcanti. -- Rio de Janeiro, 2021.
87 f.

Orientadora: Kelly Cristina Mota Gonçalves.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós-Graduação em Estatística, 2021.

1. Regressão quantílica bayesiana. 2. Distribuição Laplace assimétrica. 3. Pequenas áreas a nível de unidade. 4. Previsão. 5. Amostrador de gibbs. I. Gonçalves, Kelly Cristina Mota, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

Aos meus pais, Marlene e Cavalcanti, por sempre acreditarem em mim e por terem abdicado de suas vidas em prol das realizações e da felicidade de seus filhos. Ao meu noivo, por todo incentivo, apoio e compreensão nesses dois anos, e as minhas irmãs, por sua preocupação e carinho.

"Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning."(Albert Einstein)

Agradecimentos

A Deus, pela dádiva da vida e por me permitir realizar tantos sonhos. Obrigado por me permitir errar, aprender e crescer, por sua eterna compreensão e tolerância, por seu infinito amor, pela sua voz "invisível" que não me permitiu desistir em tantos desafios no caminho e principalmente por ter me dado uma família tão especial, enfim, obrigado por tudo. A minha orientadora professora Kelly, pela orientação, competência, profissionalismo e dedicação tão importantes. Tantas vezes que nos reunimos e, embora em algumas eu estivesse cansada e desestimulada, bastavam alguns minutos de conversa e umas poucas palavras de incentivo e lá estava eu, com o mesmo ânimo do primeiro dia de aula. Obrigada por acreditar e não desistir de mim. Tenho certeza que não chegaria neste ponto sem o seu apoio. Aos membros da banca examinadora, professor Fernando A. S. Moura e professora Valmária Rocha da Silva Ferraz, que tão gentilmente aceitaram participar e colaborar com esta dissertação e a todos os professores deste departamento por todos os ensinamentos. Aos amigos Ruan, Alana, Andrew, Daiane, e Felipe pelos trabalhos e disciplinas realizados em conjunto e, principalmente, pela preocupação e apoio constantes. A todos os demais amigos e amigas, obrigado pelo convívio, amizade e apoio demonstrado. A Marinha, a assessoria de contratos e em especial aos comandantes, Duarte, Leal, Cesar, Vasconcelos, ao Capitão tenente Valentim, por terem acreditado em mim e pela oportunidade concedida para a realização deste curso e desta dissertação. E aos meus campanhas de bordo, por todo incentivo e esforço em trocas de serviço, atenção, amizade e auxílio para que eu conseguisse assistir as aulas, Gabriella, Alves, Henrique, Flaviana, Bruna e todos os outros que me apoiaram neste processo. À minha mãe e ao meu pai deixo um agradecimento especial, por todas as lições de amor, companheirismo, amizade, caridade, dedicação, abnegação, compreensão e perdão que vocês me dão a cada novo dia. Sinto-me orgulhosa e privilegiada por ter pais tão especiais. E às minhas irmãs queridas, Clênia e Clécia, sempre prontas a me apoiar em tudo nesta vida. A meu noivo Luis Nadal, por todo amor, carinho, compreensão e apoio em tantos momentos difíceis desta caminhada. Obrigado por permanecer ao meu lado,

mesmo sem os carinhos rotineiros, sem a atenção devida e depois de tantos momentos de lazer perdidos. Obrigado pelo presente de cada dia, pelo seu sorriso e por saber me fazer feliz. Por fim, agradeço a todos aqueles que contribuíram, direta ou indiretamente, para a realização desta dissertação, o meu sincero agradecimento.

Resumo

Nesta dissertação apresentamos uma nova classe de modelos para estimação em pequenos domínios a nível de unidade. Pequenos domínios são grupos de unidades para os quais a amostra observada é muito pequena e portanto, não é representativa. O estudo de estimação em pequenos domínios é de extrema importância prática por estar conectado muitas vezes à formulação de políticas, alocação de fundos governamentais, planejamentos, entre outros, necessitando assim de estatísticas cada vez mais precisas. Como o modelo proposto tem a mesma estrutura geral de modelos usuais a nível de unidade, ou seja, uma parcela que incorpora o efeito de covariáveis e outra com efeitos aleatórios específicos de área, diferente da regressão quantílica, além de poder estudar os efeitos das covariáveis para diversos quantis, é possível avaliar os efeitos aleatórios também variando por quantil. O modelo proposto fundamenta-se no paradigma bayesiano de inferência e, portanto, o modelo proposto é construído assumindo a distribuição Laplace assimétrica para a variável de interesse, independente da real distribuição dos dados, cuja representação de mistura normal exponencial permite obter distribuições condicionais completas fechadas e conhecidas. O objetivo deste projeto é inserir a abordagem de regressão quantílica ao contexto de estimação em pequenos domínios, sob o enfoque bayesiano. Desta forma, o modelo proposto permite fazer análises não somente baseadas em medidas de tendência central, como a média, por exemplo, mas para qualquer quantil de interesse. O modelo e as alternativas de previsão propostos foram avaliados através do Vício, EQM e EMA e comparados com modelos já usuais na literatura.

Palavras-Chaves: *Quantil condicional, Distribuição Laplace Assimétrica, Pequenas Áreas, Previsão, Unidade, Amostrador de Gibbs*

Abstract

In this dissertation we present a new class of models for estimating small domains at the unit level. Small domains are groups of unit which an observed sample is too small and therefore, not representative. The study of estimation in small domains is extremely important due to its conection to policies, allocation of administrative funds, planning, among others, requiring increasingly accurate statistics. As the proposed model has the same general structure of usual models at the unit level, that is, a portion that incorporates the effect of covariates and another with area-specific random effects, different from quantile regression, in addition to being able to study the effects of the covariates for different quantiles, it is possible to evaluate the random effects also varying by quantile. The model proposed is based on the Bayesian approach of inference and, therefore, it is built assuming the asymmetric Laplace distribution for the variable of interest, regardless of the real data's distribution, which representation of exponential normal mixture allows to obtain completly closed and marked conditional distributions. The goal of this project is to insert the quantile regression approach to the small domains estimation context, under Bayesian approach. Thus, the proposed model allows measures not only based on measures of central tendency, such as average, for example, but also for any quantile of interest. The proposed model and prediction alternatives were obtained through bias, EQM and EMA compared to models already used in the literature.

keywords*: Conditional quantile, Asymmetric Laplace Distribution, Small Areas, Prediction, Unit, Gibbs Sampling*

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A estimação de pequenas áreas emergiu como área importante para literatura estatística nos últimos anos, devido a demanda por estatísticas confiáveis (Rao e Molina, 2015), uma vez que empresas privadas e órgãos públicos precisam do máximo de informações, na busca de soluções com menor custo orçamentário, estando assim interessados em estatísticas para domínios cada vez menores. O estudo de estimação em pequenos domínios é de extrema importância prática pois os governos estaduais e federal, por exemplo, a fim de identificar sub-regiões menos desenvolvidas, necessitam de informações geográficas mais desagregadas, auxiliando de forma geral em problemas de tomadas de decisão, como a elaboração de planos de desenvolvimento regionais. Tais decisões podem influenciar, por exemplo, em formulação de políticas, alocação de fundos governamentais, planejamentos, entre outros.

Um problema comum neste contexto é que associado ao interesse de estimação em níveis mais desagregados, está o fato de que a amostra observada é muito pequena. Denomina-se pequenos domínios ou pequenas áreas estes grupos de unidades populacionais para os quais a amostra observada é muito pequena e portanto, não é representativa. Dessa forma, estimadores diretos tradicionais de áreas específicas não fornecem precisão adequada.

Pequenos domínios podem se referir tanto a variáveis demográficas como geográficas tais como municípios, distritos ou mesmo bairros, ou até o cruzamento delas. A necessidade de informações e decisões para uma subpopulação para a qual estatísticas de interesse não podem ser produzidas devido a certas limitações dos dados ou porque as estimativas têm erros de amostragem extremamente grandes tem apresentado um crescimento considerável nos últimos anos. Isso torna necessário o emprego de estimadores indiretos que emprestam força de áreas afins, em particular, estimadores indiretos baseados em modelos (Sinha e Rao, 2009).

Há dois tipos de modelos para estimação em pequenas áreas: modelos a nível de área e modelos a nível de unidade. Os modelos a nível de área são baseados em estimativas de levantamento direto da área, enquanto os modelos a nível de unidade são baseados em observações individuais para as unidades dentro das áreas. Em alguns casos, o primeiro é a única alternativa disponível, já que modelos a nível de unidade requerem informações para as unidades individuais na amostra, que podem estar indisponíveis. O modelo de pequenas áreas proposto por Fay e Herriot (1979) é um dos mais populares. Ele empresta força de dados disponíveis de todas as áreas, assumindo uma estrutura hierárquica e usa informações auxiliares a nível de área de outras fontes de dados, como registros administrativos ou censos. Neste contexto, Kott (1989) apresenta o uso de técnicas baseadas em modelo de estimadores consistentes, a partir de modelos de efeitos aleatórios.

A abordagem padrão para estimativa de pequena área usa modelos de regressão para prever as características de interesse e incorpora efeitos aleatórios para cada pequena área (geralmente gaussianos) para medir a variação entre elas, além daquela explicada pelas covariáveis do modelo. Um dos modelos clássicos a nível de unidade é devido ao Battese et al. (1988). Neste trabalho, o objetivo é estimar as áreas cultivadas com milho e soja para doze condados do Centro-Norte de Iowa. Este é um modelo normal de efeitos mistos composto por dois componentes de variância, um correspondendo ao efeito aleatório de área, o outro sendo a variância do erro no nível de unidade. Em contraste, Datta e Ghosh (1991) propõem uma abordagem Bayesiana hierárquica mais geral envolvendo componentes de variância múltiplas sob a suposição de normalidade.

Desta forma, a abordagem padrão usa modelos de regressão e oferece conclusões com base no comportamento da distribuição média. No entanto, incorporar uma abordagem baseada em quantis pode fornecer conclusões mais completas e, portanto, diferentes decisões podem ser feitas dependendo do quantil de interesse. Nesse sentido, a regressão quantílica é uma alternativa robusta (Schoch, 2012). A metodologia é robusta para presença de outliers e acomoda heterocedasticidade dentro de pequenas áreas e permite a investigação de associações entre variáveis em quantis ao invés de médias, o que pode ser mais interessante.

Neste contexto, Chambers e Tzavidis (2006) propõem uma abordagem alternativa para a estimativa de pequenas áreas com base na modelagem de M-quantis de regressão da distribuição condicional. Isso evita as fortes suposições implícitas na abordagem do modelo misto e tem o benefício adicional de não exigir especificação formal dos efeitos aleatórios. Em vez disso, a variabilidade entre áreas é capturada pela variação nos coefi-

cientes dos M-quantis específicos de cada área. Vários desenvolvimentos de sua proposta e aplicações surgiram desde então (Bianchi et al., 2018).

A literatura desenvolvida a respeito destes modelos é ampla, sendo geralmente abordados de forma clássica. Tradicionalmente, os modelos neste contexto assumem que o valor esperado da variável resposta é explicado por variáveis auxiliares e que os erros aleatórios são independentes e identicamente distribuídos, provenientes de alguma distribuição simétrica em torno do zero. Considerando, frequentemente a distribuição gaussiana para o erro aleatório, embora existam propostas que consideram a distribuição Laplace Assimétrica para o erro aleatório, como mostrado em You e Rao (2003) e Rao (2003) a nível de área. Destaca-se o trabalho de Kozumi e Kobayashi (2011), que fazendo uso da representação locação-escala da distribuição de Laplace Assimétrica, como mistura das distribuições normal e exponencial, fornece um esquema de Gibbs com as condicionais completas conhecidas e fáceis de gerar amostras. Assim, a modelagem proposta permite obter distribuições condicionais completas fechadas e conhecidas.

Nesta dissertação apresentamos uma nova classe de modelos flexível de pequenas áreas, através de regressão quantílica bayesiana. A regressão quantílica, vem-se desenvolvendo como uma ferramenta eficiente para a análise das funções quantílicas de uma variável resposta condicionada as covariáveis. A modelagem proposta utiliza a distribuição Laplace Assimétrica cuja representação de mistura normal-exponencial permite obter distribuições condicionais completas fechadas e conhecidas. Nesta nova classe de modelos, foi considerado além de uma avaliação para os diferentes quantis, um modelo de regressão quantílica a nível de unidade, ou seja, uma parcela que incorpopora o efeito de covariáveis e outra com efeitos aleatórios específicos de área. Diferente da regressão na média condicional, a regressão quantílica apresenta vantagens como um conhecimento mais abrangente de como as covariáveis podem influenciar nos quantis da variável resposta, robustez e admitir uma especificação para a distribuição do termo de erro aleatório (Koenker e Ng, 2005; Fabrizi et al., 2020).

Este trabalho organiza-se da forma descrita a seguir. No Capítulo 2, apresentamos os principais modelos de pequenas áreas, segundo Rao e Molina (2015): i) modelo no nível de área, ii) modelo no nível da unidade. Ainda nesse capítulo, apresenta-se uma revisão acerca do modelo de regressão quantílica usando a distribuição Laplace Assimétrica e sua representação de mistura locação-escala. Na Seção 2.2.2, expõe-se o modelo proposto para estimação em pequenos domínios a nível de unidade, mostrando o procedimento de inferência e estudos com base em dados artificiais que ilustram o ajuste adequado do

modelo proposto. Na Seção 2.2.3 apresentamos duas formas de realizar previsão para as unidades que não foram amostradas. A alternativa 1 utiliza a distribuição Laplace Assimétrica não com o propósito de ajustar bem os dados, mas por sua analogia com a função de perda. Além disso, neste caso temos uma estimativa diferente para cada quantil fixado. Em contrapartida, a alternativa 2 permite uma aproximação da média da previsão utilizando a distribuição *a posteriori* dos quantis estimados. No Capítulo 3, três aplicações a dados reais são apresentadas, uma utilizando dados de agricultura, outra com dados de renda e a última relacionada a dados educacionais. Por fim, no Capítulo 4, apresentamos as conclusões e extensões deste trabalho.

Capítulo 2

Regressão quantílica em pequenas áreas

No problema de estimação em pequenos domínios, abordagens baseadas em modelos são amplamente empregadas, já que o mesmo modelo permite emprestar informações dos dados disponíveis em todos os domínios. O tipo de modelagem a ser usada depende dos tipos de variável resposta e auxiliares que estão disponíveis em registros administrativos, censos, etc. Modelos de pequenas áreas são classificados como: (i) modelos a nível de área e (ii) modelos a nível de unidade. Enquanto o primeiro requer informações agregadas a nível de área, o segundo requer informações a nível das unidades dentro das pequenas áreas, o que faz com que modelos a nível de área sejam frequentemente mais usados. No entanto, há casos em que o interesse está em justamente estimar parâmetros populacionais para cada área e que informações a nível de unidade estão disponíveis. O foco deste trabalho está precisamente nestes casos e, portanto modelos a nível de unidade serão considerados.

Tradicionalmente, os modelos neste contexto assumem que o valor esperado da variável resposta é explicado por variáveis auxiliares e que os erros aleatórios são independentes e identicamente distribuídos, provenientes de alguma distribuição simétrica em torno de zero, sendo a principal delas a distribuição normal. No entanto, em cenários com a presença de heteroscedasticidade, valores discrepantes e em que a simetria não é uma suposição razoável, os modelos normais usuais não são razoáveis. Além disso, pode ainda haver um interesse em avaliar o efeito de variáveis auxiliares não no valor esperado, mas em quantis da distribuição, de forma a se ter uma análise mais completa. Para tais cenários propomos neste trabalho um modelo de regressão quantílica a nível de unidade no contexto de estimação em pequenos domínios. Que difere dos trabalhos presentes na literatura, que na maioria das vezes usa regressão quantílica de forma bayesiana a

nível de área (Fabrizi et al., 2020) e de forma frequentista a nível de área (Chambers e Tzavidis, 2006) e unidade (Chambers et al., 2012).

Na Seção 2.1 deste capítulo apresenta-se o problema de estimação em pequenas áreas de forma geral e abordagens tradicionais neste contexto. Na Seção 2.2 o modelo proposto é apresentado, assim como seu procedimento de inferência e duas alternativas de previsão para as unidades que não foram amostras, com objetivo de fazer previsão para a média populacional.

2.1 Estimação em pequenos domínios

O problema de estimação em pequenas áreas está associado a previsão de quantidades descritivas de uma população finita (como médias, totais e índices de concentração) para subpopulações as quais os tamanhos da amostra específicos são, em todos ou na maioria dos casos, insuficiente para permitir inferência precisa usando métodos de estimativa direta.

Um domínio (área) é considerado pequeno se a amostra específica do domínio não for grande o suficiente para produzir "estimativas diretas" de precisão adequada. Para este caso, o uso de modelos torna-se necessário.

Existe uma grande variedade de modelos diferentes, muitas vezes complexos, que podem ser usados, dependendo da natureza da medição da estimativa da área e dos dados auxiliares disponíveis. Uma distinção fundamental na construção do modelo é entre situações onde os dados auxiliares estão disponíveis para as unidades individuais da população e aquelas onde estão disponíveis apenas em nível agregado para cada pequena área. No primeiro caso, os dados podem ser usados em modelos de nível de unidade, enquanto no segundo eles podem ser usados apenas em modelos de nível de área (Rao e Molina, 2015).

Considere uma população U cujas unidades são rotuladas de $1, \ldots, N$. A população U é considerada dividida em m subconjuntos $U_1, \ldots, U_m$ chamados de pequenas áreas. Seja N_i o tamanho de U_i e suponha $U = \bigcup_{i=1}^{m} U_i$, então temos que $N = \sum_{i=1}^{m} N_i$. Associada a população U tem-se uma variável de interesse Y, de forma que para a unidade j dentro da pequena área tem-se a variável Y_{ij} associada, para $j = 1, \ldots, N_i$ e $i = 1, \ldots, m$. Em geral, há interesse na estimação de uma quantidade populacional $\theta = g(Y_{11}, \ldots, Y_{mN_m})$, ou em parâmetros populacionais específicos das áreas, ou seja, $\theta_i = g(Y_{i1}, \ldots, Y_{iN_i})$, para $i = 1, \ldots, m$.

2.1.1 Modelo básico a nível de área

Uma amostra de tamanho n_i é sorteada de cada pequena área, de forma que estimativas diretas de θ_i são obtidas para todo $i = 1, \ldots, m$. O modelo básico a nível de área é dado por:

$$\hat{\theta}_i = z_i^T \beta + b_i v_i + \epsilon_i, \quad i = 1, \ldots, m, \tag{2.1}$$

onde $\hat{\theta}_i$ um estimador direto do parâmetro populacional θ_i da área i, z_i é um vetor $p \times 1$ com $p-1$ covariáveis de nível de área e um intercepto, β é vetor de coeficientes da regressão de dimensão p, v_i são efeitos aleatórios de área, tais que $v_i \sim N(0, \sigma_v^2)$ e ϵ_i são os erros amostrais, tais que $\epsilon_i \sim N(0, \psi_i)$, para $i = 1, \ldots, m$. Os efeitos v_i são independentes dos erros amostrais ϵ_i, b_i é uma constante positiva, que pode ser conhecida, e as variâncias amostrais ψ_i são em geral supostamente conhecidas e dadas pelas respectivas estimativas de variância amostral.

Sob o enfoque bayesiano, o modelo (2.1) é completado com a especificação de uma distribuição a *priori* para o vetor paramétrico $\Theta = (\beta, \sigma_v^2)$. Uma distribuição a *priori* possível neste caso seria $f(\beta) \propto 1$ e $\sigma_v^{-2} \sim Gama(a/2, b/2)$, assumindo independência a *priori* entre os parâmetros do modelo. Mais informações sobre este modelo podem ser encontradas em Rao e Molina (2015).

2.1.2 Modelo básico a nível de unidade

Este tipo de modelo é adequado para o caso em que o interesse está em estimar os parâmetros populacionais θ_i, associado a área i, com base numa amostra de tamanho n_i sorteada dentro de cada pequena área, e quando variáveis auxiliares específicas a nível de unidade $x_{ij} = (1, x_{ij1}, ..., x_{ijp-1})^T$ estão disponíveis para cada elemento j da população em cada pequena área i. Seja y_{ij} a variável de interesse para a unidade j pertencente a área i, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. Assumindo-se que a variável de interesse, y_{ij}, está relacionada com x_{ij}, o modelo a nível de unidade é dado por:

$$y_{ij} = x_{ij}^T \beta + v_i + e_{ij}, \quad j = 1, \ldots, N_i, \; i = 1, \ldots, m.$$

Aqui, os efeitos específicos da área v_i são considerados variáveis aleatórias independentes e indenticamente distribuídas tais que $v_i \sim N(0, \sigma_v^2)$. Os erros aleatórios e_{ij} são também variáveis aleatórias independentes e identicamente distribuídas tais que $e_{ij} \sim N(0, \sigma_e^2)$, independentes dos $v_i's$.

Sob o enfoque bayesiano, é nessário completar o modelo de nível de unidade assumindo uma distribuição *priori* para os parâmetros do modelo $(\beta, \sigma_v^2, \sigma_e^2)$. Uma distribuição *a priori* possível neste caso seria $f(\beta) \propto 1$, $\sigma_v^{-2} \sim Gama(a/2, b/2)$ e $\sigma_e^{-2} \sim Gama(c/2, d/2)$, assumindo independência a *priori* entre os parâmetros do modelo. Este trabalho está particularmente interessado na modelagem a nível de unidade.

2.2 Modelo de regressão quantílica a nível de unidade

Como visto previamente, estimação em pequenas áreas geralmente dependem de modelos de regressão que usam covariáveis e efeitos aleatórios para explicar a variação entre as áreas. No entanto, tais modelos também dependem de suposições fortes como homoscedasticidade, ausência de valores discrepantes, além de permitirem investigar tanto o efeito das covariáveis e efeitos aleatórios específicos das áreas em termos do valor esperado da variável resposta. A regressão quantílica pode ser vista como um alternativa neste caso, pois permitem o ajuste do modelo para diversos quantis, caracterizando assim estes efeitos de uma forma mais completa.

No contexto de estimação em pequenos domínios, Chambers e Tzavidis (2006) introduzem a regressão quantílica como uma classe alternativa de modelos para o mesmo propósito, relaxando algumas das suposições de modelagem convencionais, como a normalidade dos componentes aleatórios e, obtendo estimadores robustos sob a presença de valores discrepantes. Em particular, sua proposta baseia-se na modelagem de coeficientes associados ao quantil da distribuição condicional da variável de interesse dadas as covariáveis, mas sem uso de efeitos aleatórios. A variação entre domínios é caracterizada pela variação em valores específicos da área desses coeficientes associados a quantis. A proposta representa uma alternativa para estimação de efeitos aleatórios de área sem a necessidade de hipóteses paramétricas. Alguns desenvolvimentos dessa abordagem e aplicações surgiram desde então (Bianchi et al., 2018). Mais recentemente, Fabrizi et al. (2020) explorou esta abordagem sob o enfoque bayesiano.

Neste trabalho, temos interesse em propor uma forma alternativa de estimação em pequenos domínios sob a ótica de quantis. Em particular, vamos adotar a abordagem de pseudo-verossimilhança baseda na distribuição Laplace Assimétrica proposta por Yu e Moyeed (2001) e estimar os quantis de interesse separadamente. Abordar uma estimação em quantis nos permite ainda ampliar o escopo para obter estimativas para vários quantis,

fornecendo assim uma imagem mais rica tanto dos efeitos das covariáveis como dos efeitos aleatórios.

2.2.1 Regressão quantílica Bayesiana

A regressão quantílica, vem-se desenvolvendo como uma ferramenta eficiente para a análise das funções quantílicas de uma variável resposta condicionada as covariáveis. Conforme discutido anteriormente, enquanto a regressão usual limita-se em descrever a relação de Y com as covariáveis do estudo em termos de médias condicionais, a regressão quantílica é uma técnica de modelagem estatística que permite analisar essa relação em qualquer quantil de ordem τ de interesse, $\tau \in (0, 1)$.

Ou seja, tomando Y como uma variável aleatória de interesse de natureza contínua e x o vetor p-dimensional de covariáveis. A função de regressão quantílica para o τ-ésimo quantil pode ser expressa por:

$$Q_\tau(y|x) = x^T\beta_\tau, \tag{2.2}$$

que basicamente supõe que o quantil de ordem τ de Y dado X segue um modelo linear com coeficientes β_τ, vetor p-dimensional associado ao vetor de covariáveis.

Sob o ponto de vista da inferência para esta classe de modelos, enquanto no modelo de regressão linear clássico, os coeficientes da regressão são obtidos via minimização da soma dos quadrados dos desvios e isso é equivalente a maximizar uma função de verossimilhança considerando os erros normais, na regressão quantílica os coeficientes são obtidos através da minimização da soma de desvios absolutos ponderados, o que é equivalente a maximizar uma função de verossimilhança considerando erros com distribuição Laplace Assimétrica (Yu e Moyeed, 2001).

Considerando uma amostra aleatória $(y_i; x_i)_{i=1}^N$, o modelo de regressão quantílica pode ser representado por:

$$y_i = x_i^T\beta_\tau + \epsilon_i, \tag{2.3}$$

sendo ϵ_i o termo de erro cuja densidade $f_\tau(.)$ é restrita por acumular probabilidade τ até zero, ou seja,

$$\int_{-\infty}^{0} f_\tau(\epsilon_i)d\epsilon_i = \tau.$$

Na abordagem clássica o vetor de parâmetros β_τ é estimado minimizando a soma:

$$\sum_{i=1}^{N} \rho_\tau(y_i - x_i^T\beta_\tau), \tag{2.4}$$

tal que $\rho_\tau(u) = u[\tau - I(u < 0)]$ a função de perda associada ao quantil τ para todo $u \in R$ e $I(.)$ denota a função indicadora usual. Neste caso, $u = y_i - x_i^T\beta_\tau$.

Yu e Moyeed (2001) introduzem o paradigma bayesiano neste contexto verificando que minimizar a função de perda $\rho_\tau(.)$ é equivalente a maximizar a função de verossimilhança assumindo distribuição Laplace Assimétrica para os erros. A distribuição Laplace Assimétrica com parâmetros $\mu \in \mathbb{R}$ de locação, $\sigma > 0$ de escala e $0 < \tau < 1$ de assimetria, denotada por $Y \sim LA_\tau(\mu, \sigma)$, possui densidade da forma:

$$f_\tau(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\frac{\rho_\tau(y-\mu)}{\sigma}\right\}, \tag{2.5}$$

para $y \in R$. O valor esperado e a variância da variável Y são dados, respectivamente por:

$$E(Y) = \mu + \frac{\sigma(1-2\tau)}{\tau(1-\tau)} \quad \text{e} \quad Var(Y) = \sigma^2\frac{(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2}.$$

No caso em que os parâmetros de locação e escala assumem, respectivamente, os valores 0 e 1, chamamos a distribuição de Laplace Assimétrica Padrão.

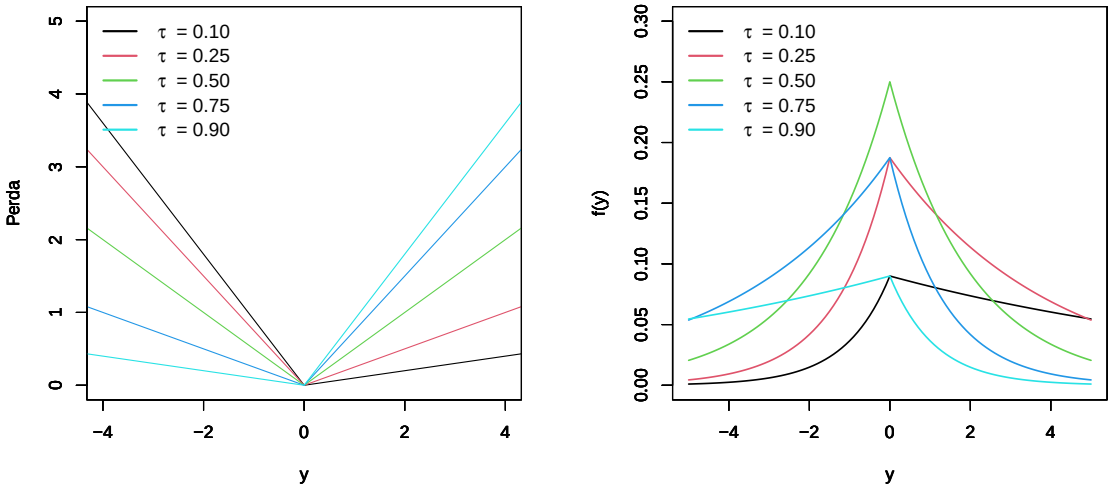
Na Figura 2.2.1 são ilustradas a função de perda ρ_τ em (a) e a função densidade de probabilidade da Laplace Assimétrica padrão em (b), ambos para diferentes valores de τ. Para quantis inferiores e superiores a $\tau = 0.5$ a função de perda penaliza mais a superestimação e subestimação, respectivamente. No quantil $\tau = 0.5$ encontramos a função de perda absoluta. Como a densidade do modelo Laplace Assimétrico possui a função de perda em seu núcleo, a simetria da distribuição se comporta similar a função perda.

Supondo que $\epsilon_i \sim LA_\tau(0, \sigma)$ no modelo descrito em (2.3), pode-se mostrar que a função de verossimilhança é dada por:

$$L(\beta, \sigma) = \frac{\tau^N(1-\tau)^N}{\sigma^N} \exp\left\{-\frac{1}{\sigma}\sum_{i=1}^{N}\rho_\tau(y_i - x_i^T\beta_\tau)\right\}. \tag{2.6}$$

Dessa forma, o estimador de máxima verossimilhança para β é igual ao estimador obtido via minimização da função de perda vista em (2.4). O enfoque bayesiano baseia-se portanto na obtenção da distribuição *a posteriori* do vetor paramétrico (β_τ, σ) obtida com base na função de verossimilhança derivada da suposição de que os erros do modelo (2.3) seguem a distribuição $LA_\tau(0, \sigma)$.

Note que, na regressão quantílica, o vetor paramétrico está indexado pelo quantil que estamos analisando, mas para simplificar notações, vamos omitir o índice τ em β_τ ao longo deste trabalho.



(a) Função de perda

(b) Laplace Assimétrica.

Figura 2.1: Gráfico da função densidade de probabilidade da Laplace Assimétrica padrão e a função de perda para diferentes valores de τ.

Diversas representações da distribuição Laplace Assimétrica em forma de mistura são apresentadas em Kotz et al. (2001a). Uma delas baseia-se na mistura locação-escala entre as distribuições Normal e Exponencial. Ele mostra que uma variável aleatória ϵ com f.d.p. Laplace padrão, adimite a representação:

$$aU + b_\tau \sqrt{U} Z$$

Portanto, se $\epsilon_i* \sim LA_\tau(0,1)$, então $\epsilon_i = \sigma\epsilon_i*$ tem f.d.p. $LA_\tau(0,\sigma)$ Consequentemente, ϵ pode ser representado por:

$$\sigma a U + \sigma b_\tau \sqrt{U} Z \tag{2.7}$$

com $U \sim Exp(1)$ e $Z \sim N(0,1)$. Para $W = \sigma U \sim Exp(\sigma)$, a representação de $\epsilon_i \sim LA_\tau(0,\sigma)$ pode ser rescrita como:

$$aW + b_\tau \sqrt{\sigma W} Z \tag{2.8}$$

Com $Z \sim N(0,1)$ e $W \sim Exp(\sigma)$. Em particular, se uma variável aleatória Y segue uma distribuição Laplace Assimétrica com parâmetros μ, σ e τ, sua distribuição admite a seguinte representação:

$$Y = \mu + a_\tau W + b_\tau \sqrt{\sigma W} Z \tag{2.9}$$

com $a_\tau = \frac{1-2\tau}{\tau(1-\tau)}$ e $b_\tau^2 = \frac{2}{\tau(1-\tau)}$, sendo W uma variável aleatória com distribuição Exponencial de parâmetro σ e Z uma variável aleatória Normal padrão. Isso é equivalente a supor que:

$$Y|\mu, \sigma, \tau, W \sim N(\mu + a_\tau W, b_\tau^2 \sigma W) \tag{2.10}$$

$$W \sim Exp(\sigma). \tag{2.11}$$

A demonstração deste resultado é apresentada no Anexo A e também encontra-se em Kotz et al. (2001a).

O modelo de regressão quantílica dado pela Equação (2.3) com $\epsilon_i \sim LA_\tau(0, \sigma)$ pode, portanto, ser reescrito da forma:

$$y_i = x_i^T \beta + a_\tau W_i + b_\tau \sqrt{\sigma W_i} Z_i, \quad i = 1, \ldots, N. \tag{2.12}$$

com $Z_i \sim N(0, 1)$ e $W_i \sim Exp(\sigma)$. Assim, a função de verossimilhança do modelo para (β, σ) condicional a $w = (w_1, ..., w_N)$ é dada por:

$$L(\beta, \sigma|y, w) = \prod_{i=1}^{N} f(y_i|\beta, \sigma, w_i), \tag{2.13}$$

com $y = (y_1, ..., y_N)$ e $y_i|\beta, \sigma, w_i \sim N(x_i^T \beta + a_\tau w_i, b_\tau^2 \sigma w_i)$ e $W_i|\sigma \sim Exp(\sigma)$. Para completar o modelo bayesiano, utiliza-se a distribuição *priori* $\pi(\beta, \sigma) = \pi(\beta)\pi(\sigma)$ com $\pi(\beta) \propto 1$, e σ com distribuição inversa gama com parâmetros de forma n_0 e de escala s_0, denotada por $\sigma \sim IG(n_0, s_0)$.

A distribuição *a posteriori* conjunta possuir forma analítica complexa e não permite cálculos exatos de médias ou quantis, por exemplo. Neste caso, métodos de Monte Carlo via cadeias de Markov (MCMC) podem ser utilizados para obter amostras da distribuição *a posteriori*. O estimador resultante é tão eficiente quanto o estimador clássico calculado diretamente através da otimização numérica levando-o a receber especial atenção devido a suas propriedades. Além disso, estimativas intervalares podem ser facilmente obtidas a partir de uma amostra da distribuição *a posteriori*.

O uso da representação de mistura locação-escala apresentada na Equação (2.12) resulta em condicionais completas conhecidas, o que facilita a implementação computacional do procedimento de inferência, sendo necessário o uso apenas de amostrador de Gibbs para obtenção de amostras da distribuição *a posteriori* (Kozumi e Kobayashi, 2011).

2.2.2 Modelo proposto em pequenas áreas

Considerando inicialmente a estrutura do modelo básico de pequenas áreas a nível de unidade, confome visto na Seção 2.1 e segundo Rao (2003), temos:

$$y_{ij} = x_{ij}^T \beta_\tau + v_i + \epsilon_{ij}, \tag{2.14}$$

onde y_{ij} é a variável associada ao indivíduo (ou unidade) j, $j = 1, 2, \ldots, N_i$, na área i, $i = 1, \ldots, m$, x_{ij} é o vetor $p \times 1$ de variáveis auxiliares, β é o vetor $p \times 1$ de coeficientes da regressão, v_i é o efeito aleatório da área i e ϵ_{ij} o termo de erro individual aleatório. Os efeitos de área v_i são assumidos independentes com distribuição $N(0, \sigma_v^2)$. Da mesma forma, os erros ϵ_{ij} são supostamente independentes com média zero e variância σ_ϵ. Habitualmente, a distribuição de ϵ_{ij} é Normal, mas distribuições como t-Student ou Normal Assimétrica também são alternativas já propostas na literatura. Além disso, o v_i e o ϵ_{ij} são assumidos como mutuamente independentes.

Para desenvolver regressão quantílica bayesiana neste modelo básico pode-se considerar que o termo de erro aleatório ϵ_{ij} tem distribuição Laplace Assimétrica com parâmetro de assimetria τ, parâmetro de locação zero e parâmetro de escala σ_ϵ, ou seja, $\epsilon_{ij} \sim LA_\tau(0, \sigma_\epsilon)$. Utilizando a representação de mistura da distribuição Laplace Assimétrica, podemos reescrever o modelo como:

$$y_{ij} = x_{ij}^T \beta + a_\tau w_{ij} + b_\tau \sqrt{\sigma_\epsilon w_{ij}} z_i + v_i \tag{2.15}$$

sendo $a_\tau = \frac{1-2\tau}{(1-\tau)}$ e $b_\tau^2 = \frac{2}{\tau(1-\tau)}$, w_{ij} com distribuição exponencial com parâmetro σ_ϵ, denotado por $w_{ij} \sim Exp(\sigma_\epsilon)$, e z_i com distribuição normal padrão, ou seja $z_i \sim N(0, 1)$.

Condicional aos valores fixados de w_{ij} e de v_i, a distribuição de y_{ij} é normal e o modelo em (2.15) pode ser reescrito da forma:

$$y_{ij} | \beta, \sigma_\epsilon, w_{ij}, v_i \sim N(x_{ij}^T \beta + a_\tau w_{ij} + v_i, b_\tau^2 \sigma_\epsilon w_{ij}) \tag{2.16}$$

$$w_{ij} \sim Exp(\sigma_\epsilon) \tag{2.17}$$

$$v_i \sim N(0, \sigma_v^2). \tag{2.18}$$

Note que o modelo proposto é uma extensão do modelo de regressão quantílica pois possui em sua formulação um efeito aleatório a nível de área. Dessa forma, diferente do modelo usual de unidade temos um efeito aleatório v_i, também variando por quantil.

Procedimento de Inferência

O modelo é especificado considerando a distribuição a *priori* $\pi(\beta, \sigma_\epsilon, \sigma_v^2) = \pi(\beta)\pi(\sigma_\epsilon)$ $\pi(\sigma_v^2)$ de modo que a distribuição $\pi(\beta) \propto 1$, a distribuição $\pi(\sigma_v)$ é inversa gama com

parâmetros n_v de forma e s_v de escala, ou seja, $\sigma_v^2 \sim IG(n_v, s_v)$ e $\sigma_\epsilon \sim IG(n_\epsilon, s_\epsilon)$. Denote por $y = (y_{11}, \ldots, y_{1N_1}, \ldots, y_{m1}, \ldots, y_{mN_m})$, $w = (w_{11}, \ldots, w_{1N_1}, \ldots, w_{m1}, \ldots, w_{mN_m})$ e $v = (v_1, \ldots, v_m)$. Considere que uma amostra s de tamanho n é sorteada, sendo esta composta pela união de subamostras de cada pequena área. Dessa forma, $s = \{s_1 \cup s_2 \cup \cdots \cup s_m\}$, onde s_i é a amostra de tamanho n_i sorteada para a pequena área i, $i = 1, \ldots, m$. Sendo assim tem-se que $n = \sum_{i=1}^{m} n_i$. Temos, portanto uma parte de y observada na amostra, a qual chamamos de y_s, e outra não observada, a qual chamamos de $y_{\bar{s}}$. Da mesma forma temos uma parte de w associada às unidades amostradas e outras às unidades não amostradas. Após, observação da amostra a distribuição *a posteriori* de $(\beta, \sigma_\epsilon, \sigma_v^2, w, v)$ é dada por:

$$f(\beta, \sigma_\epsilon, \sigma_w^2, v, w \mid y_s) \propto \left[\prod_{i=1}^{m}\prod_{j=1}^{n_i} f(y_{ij}|\beta, \sigma_\epsilon, v_i, w_{ij})\right]\left[\prod_{i=1}^{m}\prod_{j=1}^{n_i} f(w_{ij}|\sigma_\epsilon)\right] \left[\prod_{i=1}^{m} f(v_i|\sigma_v)\right]\pi(\beta)\pi(\sigma_\epsilon)\pi(\sigma_v^2).$$

Como vemos, a distribuição *a posteriori* não é fácil de manipular algebricamente. Desta forma, para gerar uma amostra da distribuição *a posteriori*, e permitir assim a inferência sobre os parâmetros do modelo recorre-se aos métodos MCMC. Para isso, faz-se necessário obter as distribuições condicionais completas, as quais são dadas por:

- $(\beta|\cdot) \sim NM(m, C)$, sendo $C^{-1} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{x_{ij}x_{ij}^T}{b_\tau^2\sigma_\epsilon w_{ij}}$ e $m = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{x_{ij}(y_{ij}-v_i-a_\tau w_{ij})}{b_\tau^2\sigma_\epsilon w_{ij}} C$;
- $\sigma_\epsilon \sim IG(n_\epsilon, s_\epsilon)$ sendo, $n_e = \frac{3n}{2} + n_\epsilon$ e $\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{(y_{ij}-(x_{ij}^T-v_i-a_\tau w_{ij}))^2}{2b_\tau^2 w_{ij}} + \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij} + s_\epsilon$;
- $(w_{ij}|\cdot) \sim GIG(\frac{1}{2}, \chi, \psi)$ sendo, $\chi = \frac{a_\tau^2}{b_\tau^2\sigma_\epsilon} + \frac{2}{\sigma_\epsilon}$ e $\psi = \frac{(y_{ij}-x_{ij}^T\beta-v_i)^2}{b_\tau^2\sigma_\epsilon}$;
- $(\sigma_v^2|\cdot) \sim IG(n_v, s_v)$, sendo $n_v = \frac{m}{2} + n_v$ e $s_v = \sum_{i=1}^{m} \frac{v_i^2}{2} + s_v$
- $(v_i|\cdot) \sim N(m_v, C_v)$ sendo, $C_v^{-1} = \frac{1}{\sigma_v^2} + \frac{1}{b_\tau^2\sigma_\epsilon w_{ij}}$ e $m_v = \left[\frac{yij-x_{ij}^T-a_\tau w_{ij}}{b_\tau^2\sigma_\epsilon^2 w_{ij}}\right] C_v$.

A notação GIG representa a distribuição gaussiana inversa generalizada cuja função densidade de probabilidade é apresentada no Anexo A. Dado que as condicionais completas são conhecidas e fáceis de gerar seus valores, pode-se utilizar o esquema de amostragem de Gibbs. Para o modelo apresentado não só foram avaliadas as estimativas dos parâmetros como também a previsão para o modelo proposto.

2.2.3 Previsão da média populacional

Um dos principais objetivos de estimação em pequenos domínios é também prever quantidades populacionais, como a média populacional, por exemplo. O parâmetro média populacional da área i é dado por:

$$\theta_i = \frac{\sum_{j \in s_i} y_{ij} + \sum_{j \notin s_i} y_{ij}}{N_i} \tag{2.19}$$

para todo $i = 1, \ldots, m$, onde s_i é o conjunto dos índices das unidades que fazem parte da amostra da pequena área i e N_i é o tamanho da população na área i.

Neste trabalho fizemos a previsão de duas formas: gerando da distribuição preditiva de y_{ij} para todo $j \in s_i$ e usando que $E(y_{ij} \mid y_s)$ pode ser obtida pela integral em τ do quantil de ordem τ.

Alternativa 1: Usando a distribuição preditiva

Para estimar a média populacional de uma determinada área i, podemos usar os valores amostrados do MCMC e prever todos os valores não observados, por meio da distribuição *a posteriori* dos parâmetros e variáveis latentes. Na iteração k, $k = 1, \ldots, K$, do MCMC, o valor amostrado da média populacional na pequena área i, é dado por:

$$\theta_i^{(k)} = \frac{\sum_{j \in s_i} {y_{ij}}^{(k)} + \sum_{j \notin s_i} y_{ij}^{(k)}}{N_i},$$

para todo $i = 1, \ldots, m$. A esperança e a variância *a posteriori* são, repectivamente, dadas por $E(\theta_i|y_s) = \frac{1}{K}\sum_{k=1}^{K} \theta_i^{(k)}$ e $V(\theta_i|y_s) = \frac{1}{K}\sum_{k=1}^{K}[\theta_i^{(k)} - E(\theta_i|y_s)]^2$ para $k = 1, \ldots, K$. Assim, para o indivíduo não observado j da pequena área i, na iteração k, tem-se:

$$y_{ij}^{(k)} \sim LA_\tau(\mu_{ij}, \sigma_\epsilon)$$

onde, segundo modelo proposto $\mu_{ij} = x_{ij}^T \beta_\tau + v_i$. Nessa abordagem utilizamos o modelo Laplace Assimétrico para gerar uma amostra dos y_{ij}, porém, esse modelo não foi proposto com o propósito de ajustar bem os dados e sim por sua analogia com a função de perda. Além disso, neste caso temos uma estimativa para θ_i diferente para cada quantil fixado. Portanto, ainda que estes fatos devam ser levados em consideração na proposta, as aplicações no Capítulo 3, revelam, resultados satisfatórios usando essa abordagem.

Alternativa 2: Usando a distribuição *a posteriori* dos quantis

Outra forma de fazer previsão é utilizando a distribuição *a posteriori* dos quantis. Esta alternativa contorna os dois pontos ressaltados na alternativa 1, ou seja, tanto a questão de ter uma previsão para cada quantil fixado, como o fato de usar um modelo na distribuição preditiva que não foi usado com o próprosito de ajuste.

Considere o quantil de ordem τ de y_{ij} como um parâmetro dado por:

$$q_{\tau,ij} = Q_\tau(y_{ij} \mid x_{ij}) = x_{ij}^T \beta_\tau + v_i.$$

Amostras da distribuição *a posteriori* de q_{ij}^τ pode ser obtida via MCMC, da forma:

$$q_{\tau,ij}^{(k)} = x_{ij}^T \beta_\tau^{(k)} + v_i^{(k)},$$

para a iteração k, $k = 1, \ldots, K$. Com base nesta amostra, podemos obter, por exemplo a média e variância *a posteriori*, dadas, respectivamente, por:

$$E(q_{\tau,ij} \mid y_s) = \frac{1}{K} \sum_{k=1}^{K} {q_{\tau,ij}}^{(k)},$$

$$Var(q_{\tau,ij} \mid y_s) = \frac{1}{K} \sum_{k=1}^{K} \left[{q_{\tau,ij}}^{(k)} - E(q_{\tau,ij} \mid y_s) \right]^2.$$

De forma análoga, intervalos de credibilidade podem ser obtidos com base nos quantis amostrais.

Usando $\hat{q}_{\tau,ij} = E(q_{\tau,ij} \mid y_s)$ como estimativa pontual de $q_{\tau,ij}$ e a propriedade que relaciona valor esperado com função quantílica, obtém-se que:

$$E(y_{ij} \mid y_s) = \int_\tau \hat{q}_{\tau,ij} d\tau. \tag{2.20}$$

Usando integração numérica de Monte Carlo a integral na Equação (2.20) pode ser obtida da seguinte forma:

$$E(y_{ij} \mid y_s) \approx \frac{1}{L} \sum_{l=1}^{L} \hat{q}_{\tau_l,ij},$$

onde L é o número de subdivisões no espaço de $\tau \in (0, 1)$. Sendo assim, o parâmetro populacional pode ser estimado da seguinte forma:

$$E(\theta_i \mid y_s) = \frac{\sum_{j \in s_i} y_{ij} + \sum_{j \notin s_i} E(y_{ij} \mid y_s)}{N_i},$$

para todo $i = 1, \ldots, m$.

2.3 Dados artificiais

Para avaliar empiricamente a metodologia de inferência para o modelo proposto geramos dados artificiais e verificou-se as estimativas obtidas via MCMC para os parâmetros do modelo utilizando o R Development Core Team (2015).

Os dados foram gerados do modelo proposto na Equação (2.15), fixando $p = 4$, $\beta = (10, 5, -2, 1)$, $\sigma_v^2 = 1$ e $\sigma_\epsilon = 1$, para cinco quantis diferentes $\tau \in (0.10, 0.25, 0.50, 0.75$ e $0.90)$. Além disso, as três covariáveis foram geradas de um modelo Normal e assumiu-se $m = 50$ áreas, e n_i variando de 20 a 100 unidades. Uma amostra de 20% foi sorteada dentro de cada pequena área, dessa forma n_i varia entre 4 e 20 unidades. A distribuição a *priori* utilizada para σ_v^2 e σ_ϵ foi $IG(0.01, 0.01)$, caracterizando assim uma distribuição *a priori* vaga.

Foram empregadas vinte e uma mil iterações do algoritimo MCMC em cada um dos casos, considerando um aquecimento de mil iterações e um espaçamento de cinco iterações.

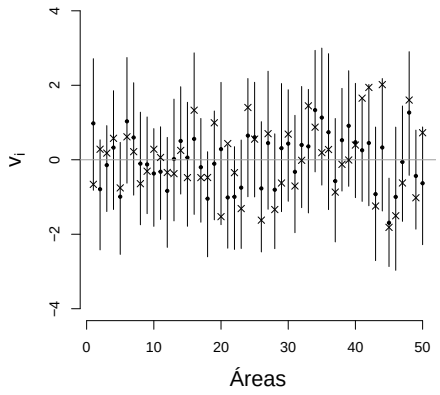
As características das cadeias permitiram verificar visualmente, com valores iniciais aleatoriamente escolhidos, uma convergência quase imediata e autocorrelações que tendem rapidamente para zero. Isto permite obter amostras da distribuição a *posteriori* satisfatórias, com baixas autocorrelações.

A Tabela 2.1, apresenta a média *a posteriori* e intervalos de credibilidade de 95% para todos os parâmetros do modelo, desta forma, avalia-se também a significância. Na grande maioria dos casos apresentados, o valor médio se encontra dentro do intervalo de credibilidade de 95% e a média *a posteriori* próxima do valor verdadeiro. A Figura 2.2 apresenta um sumário da distribuição *a posteriori* para os efeitos aleatórios. Para todas as áreas e quantis os valores verdadeiros encontram-se dentro dos intervalos de credibilidade.

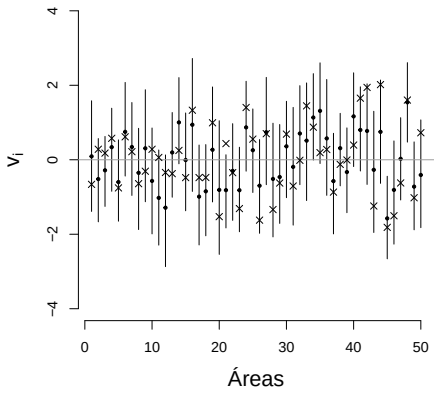
Algumas saídas de monitoramento das cadeias para diferentes quantis e parâmetros estudados nesta simulação, encontram-se nas Figuras B.1 e B.2 do Apêndice B.

Tabela 2.1: Média *a posteriori* e intervalos de credibilidade de 95% para os parâmetros β_i, $i = 0, \ldots, 3$, σ_v^2 e σ_e obtidos a partir do ajuste do modelo proposto a dados artificiais, para alguns quantis.

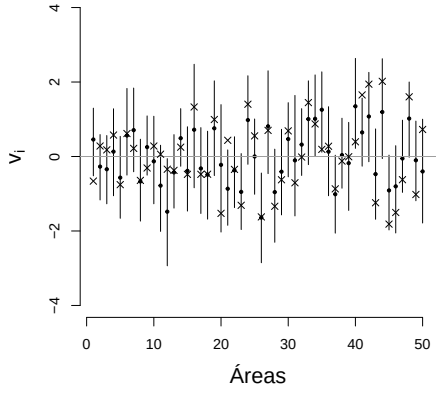
τ	$\beta_0 = 10$	$\beta_1 = 5$	$\beta_2 = -2$	$\beta_3 = 1$	$\sigma_v^2 = 1$	$\sigma_\epsilon = 1$
0.10	10.10	5.51	-1.91	1.03	1.19	1.04
	(9.60,10.60)	(5.14,5.87)	(-2.12,-1.69)	(0.89,1.18)	(0.42,2.36)	(0.97,1.13)
0.25	9.99	5.21	-1.88	1.05	1.03	1.09
	(9.58,10.38)	(4.95 ,5.46)	(-2.041,-1.73)	(0.95 ,1.16)	(0.46,1.88)	(1.00,1.18)
0.50	10.12	5.08	-1.87	1.01	0.93	1.05
	(9.77,10.47)	(4.87,5.27)	(-2.00,-1.74)	(0.93,1.09)	(0.42,1.70)	(0.97,1.13)
0.75	10.11	4.93	-1.93	1.05	1.10	1.01
	(9.71 ,10.49)	(4.71 ,5.17)	(-2.07,-1.78)	(0.95 ,1.15)	(0.54,1.96)	(0.93,1.10)
0.90	9.94	4.69	-1.99	1.12	1.20	1.02
	(9.41,10.16)	(4.31,5.10)	(-2.26,-1.73)	(0.95,1.27)	(0.40,2.41)	(0.94,1.10)



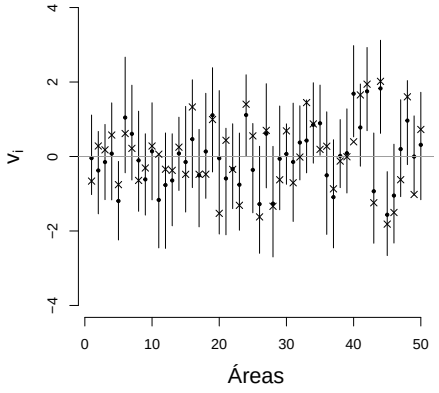
(a) Quantil 0.10



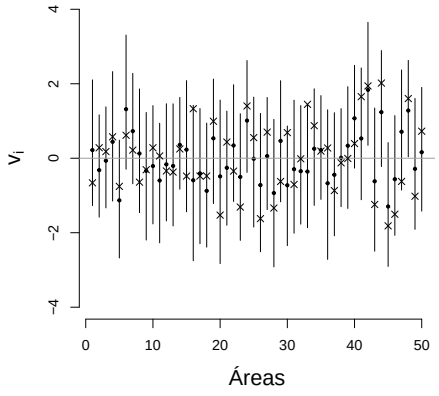
(b) Quantil 0.25



(c) Quantil 0.50



(d) Quantil 0.75



(e) Quantil 0.90

Figura 2.2: Média a *posteriori* e intervalo de credibilidade de 95% para os efeitos aleatórios v_i para cada pequena área para o modelo proposto ajustado para diferentes quantis.

Capítulo 3

Aplicações a dados reais

Neste capítulo o modelo proposto para estimação em pequenas áreas a nível de unidade aplicamos o modelo proposto em três bancos de dados reais. Os algoritmos de amostragem de Gibbs, a manipulação de dados e as análises foram feitas utilizando os programas R Development Core Team (2017) e JAGS (Plummer et al., 2003).

A primeira aplicação é feita a uma base de dados referente a agricultura. O banco é amplamente utilizado na literatura para avaliar modelos de pequenas áreas, com o objetivo de estimar hectares de milho. A segunda aplicação é realizada a uma base de dados relacionada a renda também usada em Moura e Holt (1999) e Silva Ferraz (2011). Nesta tem-se o objetivo de estimar a renda familiar de setores censitários de um município brasileiro. A terceira aplicação usa um banco de dados educacionais e tem como objetivo estimar a média escolar de um teste de português para alunos de uma determinada série do ensino fundamental no município de Rondônia. Ao longo das aplicações o modelo proposto será ajustado para os mesmos quantis dos dados artificiais. Além disso, em todas as aplicações o modelo proposto foi comparado com os modelos Normal, t-Student e Normal-Assimétrico, tanto a nível de estimação dos parâmetros e interpretação, como a nível de previsão. As comparações entre os modelos são feitas usando, em particular, o viés, erro quadrático médio e erro absoluto médio. O modelo Normal-Assimétrico foi proposto por Azzalini (1985) dentro de uma família de distribuições paramétricas que inclui a distribuição Normal padrão como caso particular, porém com um parâmetro extra que permite controlar a assimetria da distribuição. O modelo Normal-Assimétrico a nível de unidade em pequenas áreas foi desenvolvido em Silva Ferraz (2011).

3.1 Aplicação 1: Dados de agricultura dos EUA

Os dados utilizados são referentes a 12 condados de Iowa, obtidos a partir de uma Pesquisa Enumerativa de Junho de 1978 do Departamento de Agricultura dos EUA e de satélites observatórios terrestres (LANDSAT) durante a estação de crescimento de 1978. O número total de segmentos N_i (tamanho da população) dentro de cada condado varia de 402 a 965. O número total de segmentos amostrados (tamanho da amostra) para os 12 condados é 37, e o tamanho da amostra n_i $(i = 1, \ldots, 12)$ dentro cada condado varia de 1 a 6. Nesta aplicação, os condados atuam como domínios e os segmentos de amostra como unidades. Esses dados, utilizados por Battese et al. (1988), estão incluídos em dois conjuntos de dados no R, chamados *cornsoybean* e *cornsoybeanmeans*. O conjunto de dados *cornsoybean* contém códigos dos condados, hectares de milho (*CornHec*) e hectares de grãos de soja (*SoysBeanHec*) em cada segmento da amostra dentro de cada condado, o número de pixels classificado como milho a partir de dados de satélite (*CornPix*) e o número de pixels classificados como grãos de soja (*SoyBeansPix*). O dado *cornsoybeanmeans* foi usado apenas como auxílio para extração de informações sobre a população.

Alguns traballhos como Battese et al. (1988), Prasad e Rao (1990), Arima et al. (2012) e Rao e Molina (2015) utilizaram esse banco de dados para para ilustrar suas análises em pequenas áreas. Desta forma, utilizamos o mesmo banco dessas referências para avaliar o comportamento do modelo proposto nesse trabalho, como uma ilustração preliminar.

Define-se $y_{ij} = CornHec$ para a unidade j no condado i, $x_{1ij} = CornPix$ para a unidade j no condado i e $x_{2ij} = SoyBeansPix$ para a unidade j no condado i.

Sendo assim, nesta aplicação ajustamos o modelo proposto em (2.16) assumindo $p = 3$, logo x_{ij} um vetor 3×1 e $\beta = (\beta_0, \beta_1, \beta_2)^T$. Ajustamos o modelo proposto para 5 diferentes quantis: 0.10; 0.25; 0.50; 0.75 e 0.90. Além disso, ajustou-se os modelos Normal, t-Student e Normal-Assimétrico já conhecidos na literatura, para fins de comparação.

Foram utilizadas distribuições a *priori* vagas e foram rodadas 21 mil iterações do MCMC, com aquecimento de 1 mil e espaçamento de 10, produzindo assim uma amostra final da distribuição *a posteriori* de tamanho 2 mil. A inspeção visual das cadeias levou a concluir que a convergência foi alcançada.

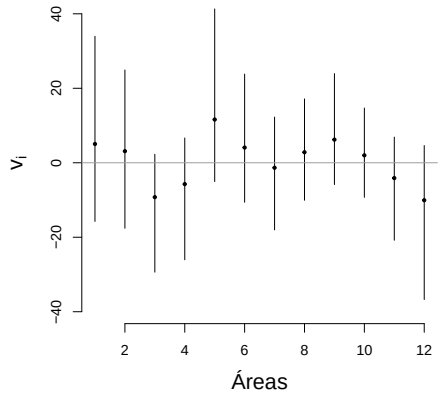
A Tabela 3.1 apresenta a média *a posteriori* e os intervalos de credibilidade de 95% para as componentes do coeficiente da regressão β obtidos para cada modelo. Nesta tabela, nota-se que as estimativas pontuais variam levemente de acordo com o quantil avaliado, ficando mais próximas nos quantis $\tau = 0.25, 0.50, 0.75$. Quando avaliamos o coeficiente β_1, associado ao número de pixels classificado como milho, vemos que este

é positivo e significativo para todos os modelos enquanto o coeficiente β_2, associado ao número de pixels classificado como grão de soja, é negativo e não é significativo para todos os modelos ajustados.

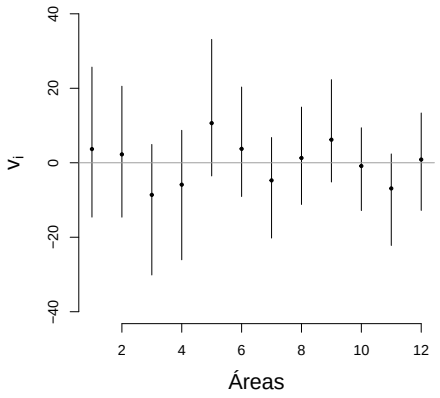
Tabela 3.1: Média *a posteriori* e intervalos de credibilidade de 95% para as componentes do coeficiente da regressão β obtidos a partir do ajuste do modelo proposto para cinco quantis e dos modelos Normal, t-Student e Normal-Assimétrico.

Modelo	β_0	β_1	β_2
$\tau = 0.10$	50.33 (-27.71,131.64)	0.24 (0.07,0.41)	-0.07 (-0.28,0.09)
$\tau = 0.25$	64.65 (-16.95,138.40)	0.21 (0.07,0.39)	-0.09 (-0.28,0.07)
$\tau = 0.50$	62.35 (-15.93,148.14)	0.26 (0.07,0.49)	-0.09 (-0.26,0.06)
$\tau = 0.75$	61.87 (-20.70,148.65)	0.30 (0.10,0.48)	-0.09 (-0.26,0.06)
$\tau = 0.90$	72.42 (-16.27,159.34)	0.29 (0.08,0.49)	-0.11 (-0.28,0.06)
Normal	55.29 (-47.29,98.27)	0.28 (0.15,0.50)	-0.11 (-0.27,0.14)
t-Stud.	49.91 (-61.07,98.13)	0.29 (0.15,0.53)	-0.09 (-0.29,0.15)
N.Assim.	41.27 (-62.33,98.01)	0.29 (0.10,0.54)	-0.11 (-0.34,0.15)

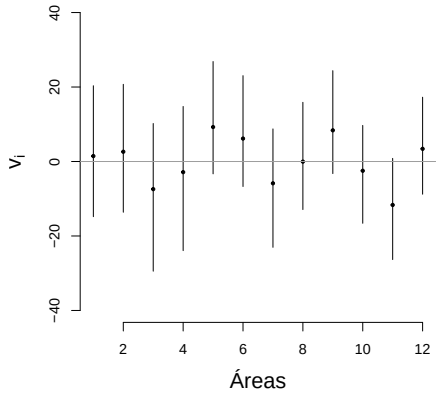
A Figura 3.1 apresenta as médias *a posteriori* e intervalos de credibilidade de 95% para o efeito aleatório v_i para todas as pequenas áreas, com a finalidade de avaliar possíveis mudanças nestes efeitos para diferentes quantis. Em geral, as estimativas destes efeitos são semelhantes para todos os quantis estudados e estes apresentam-se não significativos para todas as áreas. Além disso, estes resultados são semelhantes aos obtidos quando os modelos alternativos Normal, t-Student e Normal-Assimétrico são utilizados.



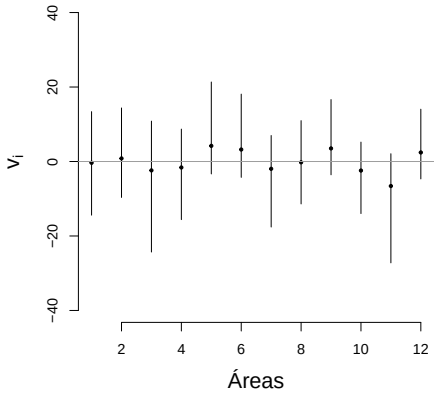
(a) Quantil 0.10



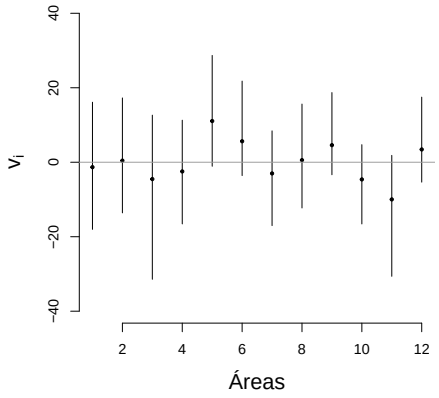
(b) Quantil 0.25



(c) Quantil 0.50



(d) Quantil 0.75



(e) Quantil 0.90

Figura 3.1: Média *a posteriori* e intervalo de credibilidade de 95% para os efeitos aleatórios v_i para cada pequena área nos diferentes quantis.

Além disso, as estimativas para σ_ϵ e σ_v^2 também não variam de forma significativa para os diferentes quantis utilizados.

Os resultados mostrados sugerem que a proposta desta dissertação, quando aplicada neste conjunto de dados, obtém resultados análogos aos obtidos nos demais artigos que utilizaram o mesmo banco, com a vantagem de que a proposta fornece estimação em diferentes quantis, não somente na média como usual, mas também em qualquer quantil da variável resposta.

3.2 Aplicação 2: Dados de renda

Esta seção apresenta uma aplicação em pequenas áreas para dados de renda, os mesmos descritos em Moura e Holt (1999) e Silva Ferraz (2011). Estes dados foram extraídos de um Censo Demográfico Experimental, que consiste de informações sobre 38.740 domicílios em 140 áreas de enumeração (setores censitários) que seriam as pequenas áreas neste caso. O objetivo consiste em estimar a renda familiar de setores censitários de um município brasileiro nas 140 áreas. Da mesma forma que Moura e Holt (1999), assume-se a informação disponível no nível da unidade (domicílio) e duas variáveis foram escolhidas para ser o conjunto de variáveis auxiliares: o número de quartos no domicílio $(1-8)$, x_1, e o nível de escolaridade do chefe de família (escala ordinal de $0-17$), x_2. O número de domicílios por área na população varia de 57 a 588. Desta forma, optou-se por utilizar uma amostra de 5%, conforme sugerido no trabalho de Silva Ferraz (2011) que trabalha com amostras. Portanto, as dimensões das amostras por área (n_i) variam de 3 a 30.

Desta forma, ajustou-se o modelo proposto neste banco de dados, assumindo $p = 3$, logo x_{ij} é um vetor 3×1 e $\beta = (\beta_0, \beta_1, \beta_2)^T$. Uma análise preliminar da variável renda para algumas pequenas áreas pode ser vista na Figura 3.2, a qual revela uma forte assimetria a direita.

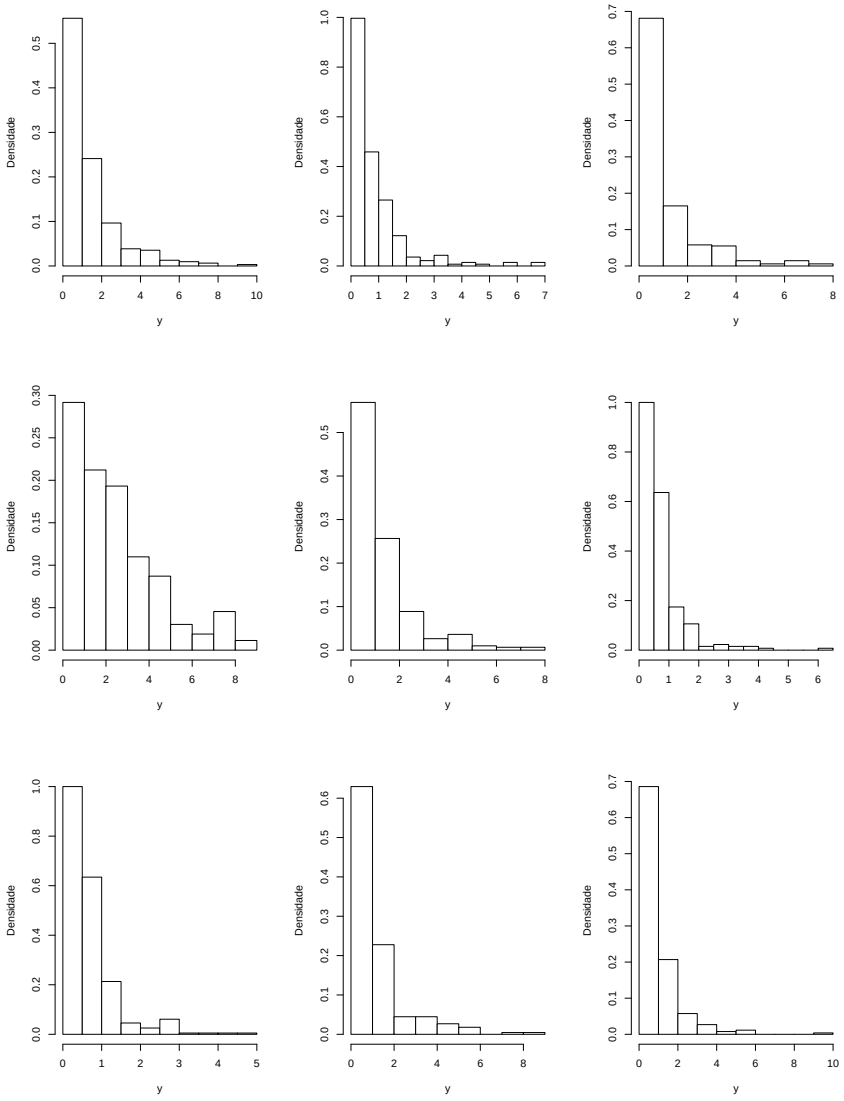


Figura 3.2: Histograma com a distribuição das rendas familiares para algumas pequenas áreas selecionadas aleatoriamente.

O modelo proposto foi ajustado para 5 diferentes quantis: 0.10; 0.25; 0.50; 0.75 e 0.90 assumindo os dados na escala original e na escala logarítmica, devido a visível assimetria. Para fins de comparação foram ajustados também os modelos Normal e t-Student ao logaritmo da renda e o modelo Normal-Assimétrico a renda na escala original. Note que ajustar o modelo Normal na escala logarítmica é equivalente a assumir o modelo logNormal para a variável na escala original.

Foram utilizadas distribuições a *priori* vagas e foram rodadas 11000 iterações do MCMC, com aquecimento de 1000 e espaçamento de 5, produzindo assim uma amostra

final da distribuição *a posteriori* de tamanho 2000. A inspeção visual das cadeias levou a concluir que a convergência foi alcançada. As cadeias das componentes do coeficiente β, σ_ϵ e σ_v^2 podem ser vistas nas Figuras B.4 e B.5, apresentadas no Apêndice B.

Na Tabela 3.2 são apresentados a média *a posteriori* e os intervalos de credibilidade de 95% de todos os coeficientes β_i obtidos para cada modelo para avaliar a significância nos diferentes quantis e modelos avaliados. Para os resultados obtidos usando os dados na escala logarítmica, conclui-se que os efeitos das covariáveis para o modelo proposto são todos significativas e variam pouco entre os quantis. Apenas o quantil 75% revela um ligeiro maior impacto da variável nível de escolaridade do chefe de família no logaritmo da renda. Entretanto, os efeitos de ambas as covariáveis não são significativos quando os modelos Normal e t-Student são ajustados. Um resultado semelhante é obtido quando o modelo proposto e o modelo Normal-Assimétrico são ajustados aos dados na escala original. Os efeitos das covariáveis crescem um pouco à medida que o quantil aumenta e os efeitos obtidos a partir do ajuste do modelo Normal-Assimétrico não são significativos.

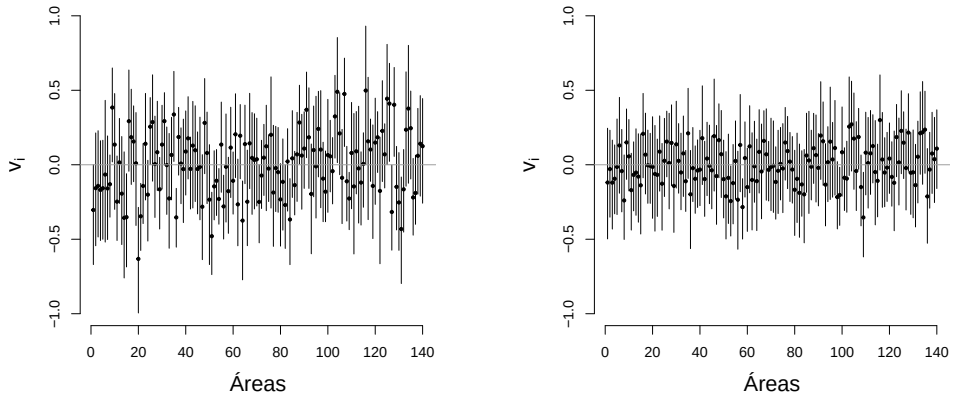
Tabela 3.2: Média *a posteriori* e intervalos de credibilidade de 95% para os parâmetros β_i, $i = 0, \ldots, 2$, obtidos a partir do ajuste do modelo proposto na segunda aplicação.

Modelo	β_0	β_1	β_2
Dados na escala logarítmica			
$\tau = 0.10$	-2.35 (-2.46,-2.25)	0.14 (0.10,0.18)	0.13 (0.13,0.14)
$\tau = 0.25$	-1.93 (-2.03,1.98)	0.16 (0.11,0.20)	0.13 (0.13,0.14)
$\tau = 0.50$	-1.37 (-1.47,1.27)	0.18 (0.14,0.23)	0.12 (0.11,0.12)
$\tau = 0.75$	-1.06 (-1.15,-0.98)	0.27 (0.23,0.28)	0.11 (0.10,0.12)
$\tau = 0.90$	-0.51 (-0.61,-0.41)	0.19 (0.15,0.23)	0.11 (0.10,0.12)
Normal	-0.41 (-0.54,-0.28)	-0.02 (-0.07,0.02)	0.01 (-0.003,0.01)
t-Stud.	-0.40 (-0.53,-0.27)	-0.02 (-0.08,0.02)	0.007 (-0.002,0.016)
Dados na escala original			
N. Assim.	0.07 (0.03,0.10)	-0.003 (-0.01,0.013)	0.0002 (-0.002,0.003)
$\tau = 0.10$	0.007 (-0.03,0.04)	0.03 (0.02,0.05)	0.04 (0.04,0.05)
$\tau = 0.25$	0.01 (-0.03,0.07)	0.06 (0.04,0.09)	0.07 (0.06,0.07)
$\tau = 0.50$	0.01 (-0.04,0.08)	0.13 (0.10,0.16)	0.10 (0.09,0.11)
$\tau = 0.75$	0.12 (0.03,0.22)	0.17 (0.12,0.23)	0.16 (0.15,0.17)
$\tau = 0.90$	0.41 (0.26,0.56)	0.23 (0.17,0.29)	0.21 (0.19,0.22)

O parâmetro de assimetria λ da distribuição Normal assimétrica obteve média *a poste-*

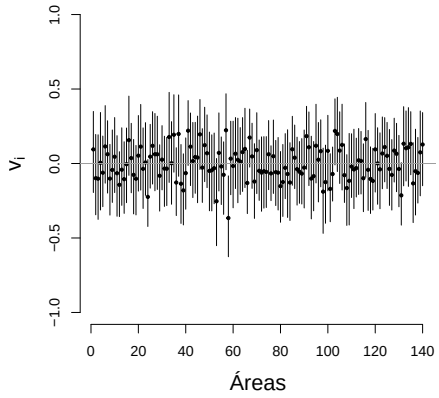
riori 40.49 com intervalo de credibilidade de 95% em $(27.42, 58.91)$, mostrando ser significativo. Este fato comprova a necessidade em ajustar modelos que levem em consideração de forma mais apropriada a assimetria. O grau de liberdade do modelo t-Student, por sua vez, foi estimado em 17.73 com intervalo de credibilidade de 95% de $(9.74, 35.93)$, logo um modelo com cauda pesada pode ser apropriado mas não é estritamente necessário. Sendo assim, principalmente pela assimetria presente, o modelo proposto se mostra como uma alternativa apropriada para este caso, que fornece resultados flexíveis.

As Figuras 3.3 e 3.4 apresentam um sumário da distribuição *a posteriori* dos efeitos aleatórios v_i para cada pequena área obtidos com base no modelo proposto para cada um dos quantis considerados, e para os dados na escala logarítmica e na escala original, respectivamente. Os pontos representam as médias *a posteriori* e os segmentos os respectivos intervalos de credibilidade de 95%. Enquanto a quantidade de efeitos aleatórios significativos aumenta mais nos extremos quando comparado a quantis intermediários quando os dados na escala logarítmica são utilizados, o mesmo só ocorre no quantil 90% para os dados na escala original. Além disso, a incerteza na estimação dos efeitos aleatórios varia por quantil. Na Figura 3.4 a escala para o quantil $\tau = 0.90$ foi modificada para facilitar a visualização.

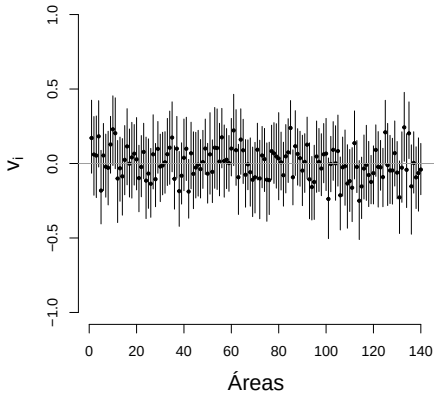


(a) Quantil 0.10

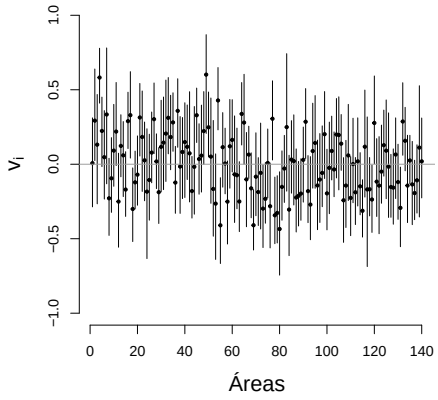
(b) Quantil 0.25



(c) Quantil 0.50

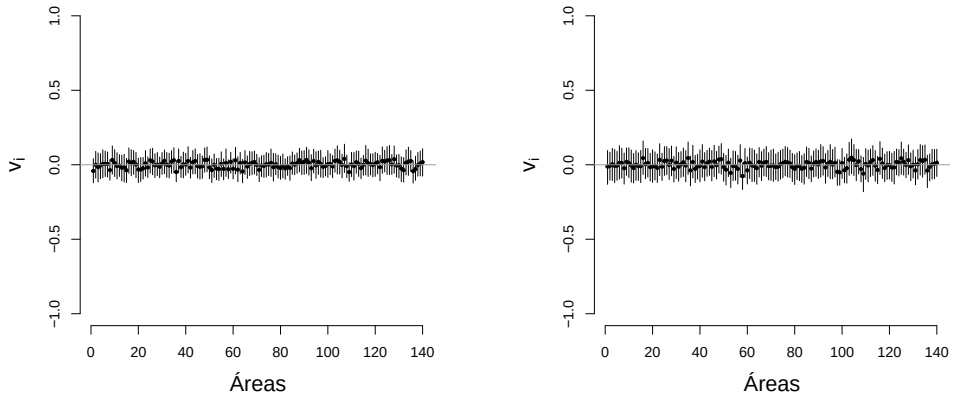


(d) Quantil 0.75



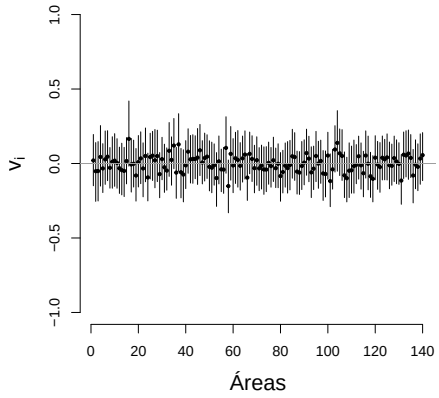
(e) Quantil 0.90

Figura 3.3: Média *a posteriori* e intervalo de credibilidade de 95% para os efeitos aleatórios v_i em cada pequena área para o modelo proposto ajustado para diferentes quantis, com dados na escala logarítmica.

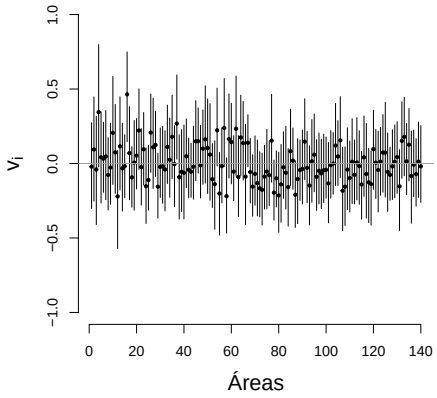


(a) Quantil 0.10

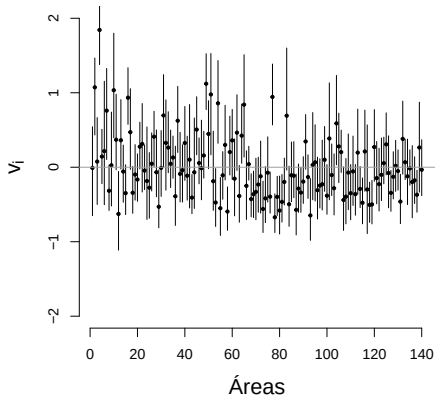
(b) Quantil 0.25



(c) Quantil 0.50



(d) Quantil 0.75



(e) Quantil 0.90

Figura 3.4: Média *a posteriori* e intervalo de credibilidade de 95% para os efeitos aleatórios v_i em cada pequena área para o modelo proposto ajustado para diferentes quantis, com dados na escala original.

A fim de complementar a análise anterior, a Tabela 3.3 mostra a proporção dos efei-

tos aleatórios que são significativos. Como esperado, os resultados variam no modelo proposto para diferentes quantis, tendo uma maior proporção de efeitos aleatórios significativos nos extremos. No ajuste aos dados da escala logarítmica os modelos Normal e t-Student produzem resultados neste caso um pouco distintos do modelo na mediana. Este fato aliado as estimativas dos coeficientes da regressão apresentadas anteriormente sugere que os resultados destes modelos não são facilmente comparáveis aos obtidos com base no ajuste na mediana.

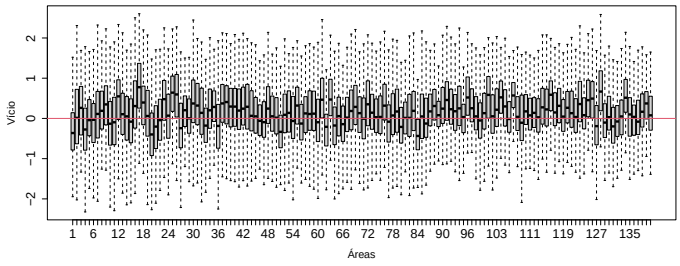
Tabela 3.3: Porcentagem de efeitos aleatórios significativos para cada modelo ajustado.

Dados	Modelo quantílico					Modelos		
	0.10	0.25	0.50	0.75	0.90	Normal	t-Stud.	N.Assim.
$log(y_{ij})$	32.85	5.71	1.43	4.28	40.00	13.57	12.85	-
y_{ij}	0	0	0	4.28	52.86	-	-	0

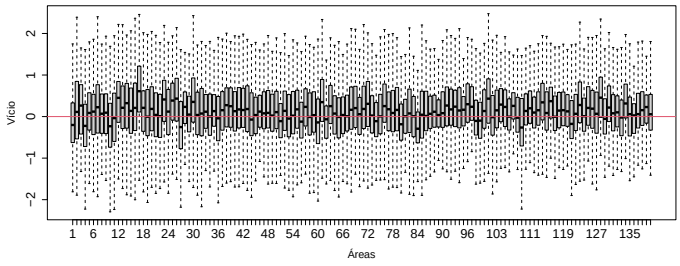
3.2.1 Previsão

Um dos maiores interesses no problema de estimação em pequenas áreas é fazer previsão, neste caso, para as unidades não observadas e média populacional dentro de cada pequena área. Nesta seção realizou-se previsões com base no modelo proposto e comparou-se estas com as previsões obtidas a partir dos modelos t-Student, Normal e Normal assimétrica. Em particular, nesta aplicação utilizou-se apenas a alternativa 1 de previsão apresentada na Seção 2.2.3, considerando cada um dos 5 quantis analisados. Como o valor verdadeiro da média em cada pequena área é conhecido, para fins de comparação os resultados foram resumidos em termos do vício, erro quadrático médio relativo (EQMR) e erro médio absoluto relativo (EMAR) para cada pequena área.

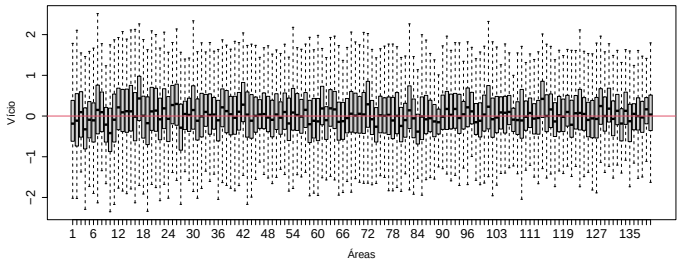
Com o objetivo de uma avaliação inicial mais profunda, nas Figuras 3.5 e 3.6 estão os boxplots dos vícios a nível de unidade, na escala logarítmica e na escala original, dentro de cada área. Quando o ajuste é feito aos dados na escala logarítmica, os boxplots na maioria das vezes encontram-se centrados no zero, indicando que o modelo proposto quando aplicado neste conjunto de dados obtém resultados razoáveis com relação a previsão a nível de unidade para cada quantil ajustado. Em particular, quantis mais próximos da mediana geram ainda menores vícios. Por outro lado, quando o ajuste é feito aos dados na escala original, resultados melhores são obtidos em quantis menores ou iguais a mediana, sendo a mediana ainda o quantil com melhores resultados acerca de previsão a nível de unidade.



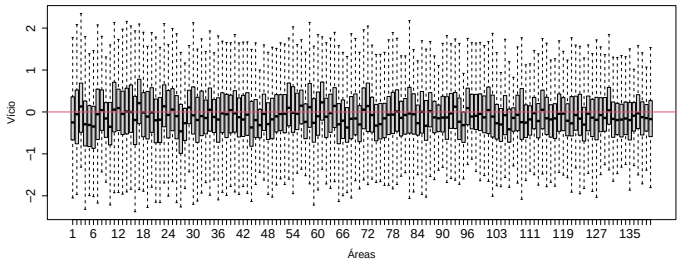
(a) Quantil 0.10



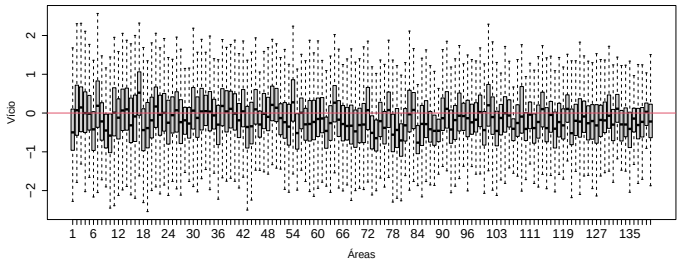
(b) Quantil 0.25



(c) Quantil 0.50

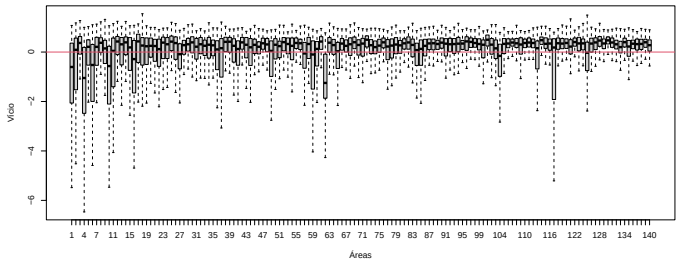


(d) Quantil 0.75

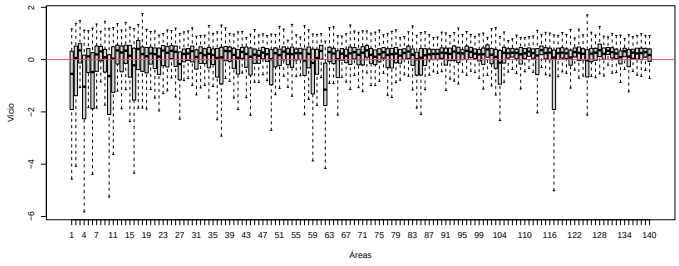


(e) Quantil 0.90

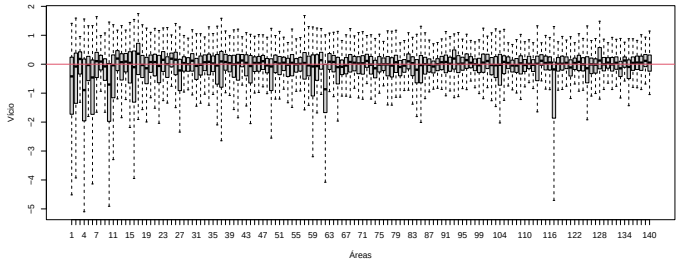
Figura 3.5: Boxplot do vício da previsão para as unidades não observadas em cada pequena área com base no modelo proposto ajustado para os quantis 0.10, 0.25, 0.50, 0.75 e 0.90. Os resultados foram obtidos usando os dados na escala logarítmica.



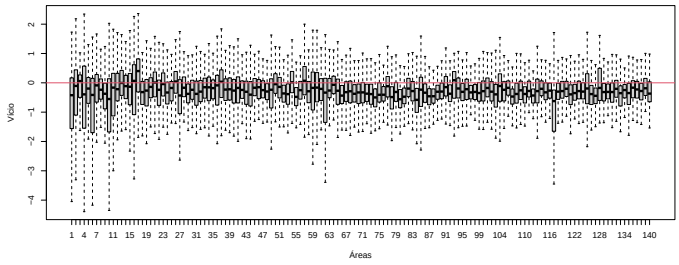
(a) Quantil 0.10



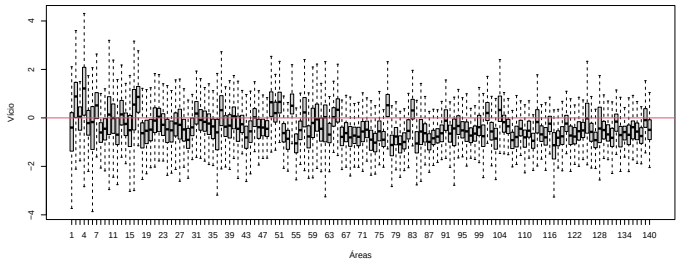
(b) Quantil 0.25



(c) Quantil 0.50



(d) Quantil 0.75



(e) Quantil 0.90

Figura 3.6: Boxplot do vício da previsão para as unidades não observadas em cada pequena área com base no modelo proposto ajustado para os quantis 0.10, 0.25, 0.50, 0.75 e 0.90. Os resultados foram obtidos usando os dados na escala original.

A Figura 3.7 apresenta o boxplot dos vícios relativos para a previsão da média populacional para cada pequena área, para cada um dos modelos ajustados, considerando ainda os dados na escala original e logarítmica. Os boxplots gerados na sua maioria apresentam a mediana próxima do zero, porém é visível a diferença em utilizar a transformação logarítmica na variável resposta para os quantis. Neste caso, os quantis 0.10 e 0.25 apresentam um desempenho inferior aos quantis 0.50 e 0.75, os quais obtiveram melhores resultados. O quantil 0.90 gera maior variabilidade nas previsões para as áreas. Por outro lado, quando o modelo proposto é ajustado aos dados originais, melhores resultados são alcançados nos quantils 0.10, 0.25 e 0.50, o que acredita-se que tenha relação com a assimetria dos dados. Comparando estes dois casos de forma geral, conclui-se que na maioria das vezes há menor variabilidade no ajuste com os dados na escala original.

Os modelos Normal e t-Student, quando ajustados aos dados na escala logarítmica, têm um bom poder preditivo, até superior ao do modelo Normal-Assimétrico. Finalmente, conclui-se que o modelo proposto para os dados na escala original apresenta resultados competitivos nos quantis 0.10, 0.25 e 0.50 com relação aos modelos tradicionais.

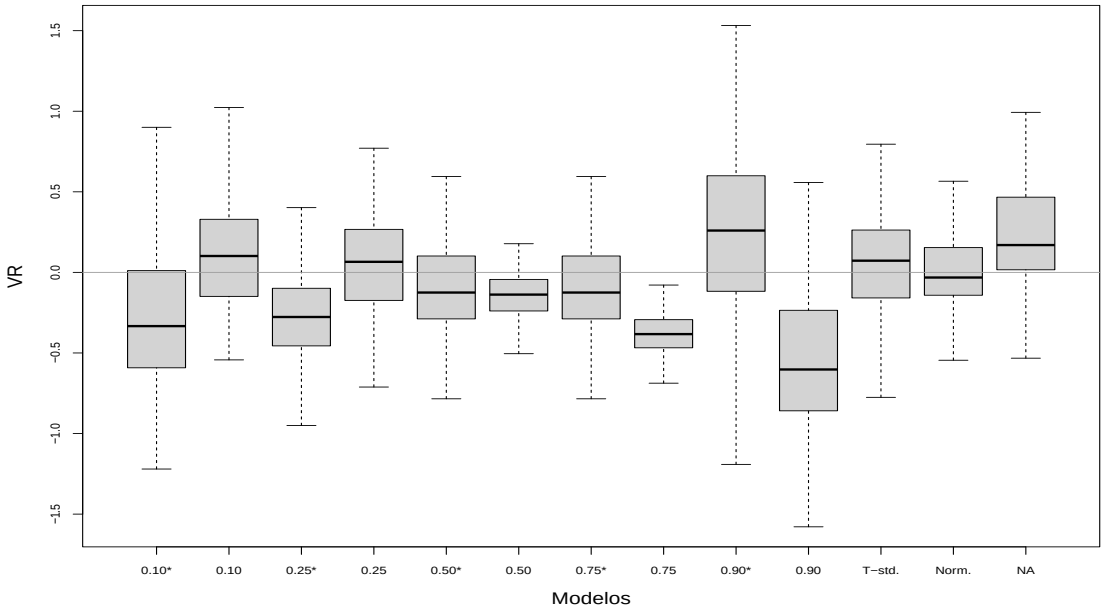


Figura 3.7: Boxplot do vício relativo da previsão da média populacional em cada pequena área com base no modelo proposto ajustado para os quantis 0.10, 0.25, 0.50, 0.75 e 0.90 e nos modelos Normal, t-Student e Normal-Assimétrico. Os quantis com o símbolo * indicam que o ajuste do modelo proposto foi feito aos dados na escala logarítmica, assim como os modelos Normal e t-Student. Os demais resultados foram obtidos usando os dados na escala original.

Complementando a análise anterior, calculou-se a raiz EQMR e o EMAR para as médias em cada área. As Figuras 3.8 e 3.9 mostram os boxplots destas medidas para cada um dos modelos ajustados. Pode-se observar que os ajustes que consideram um modelo de previsão com os dados na escala original usando o modelo proposto apresentam menores valores para a raiz do EQMR e para o EMAR quando comparado ao ajuste usando os dados na escala logarítmica. Além disso, o modelo proposto utilizando os quantis menores ou iguais a mediana, para os dados na escala original, se mostra competitivo quando comparado com os modelos usuais da literatura. Neste caso, ainda melhores resultados são obtidos na mediana. Portanto, há um ganho substancial em termos destas medidas em fazer previsão utilizando regressão quantílica bayesiana para modelos de pequenas áreas a nível de unidade.

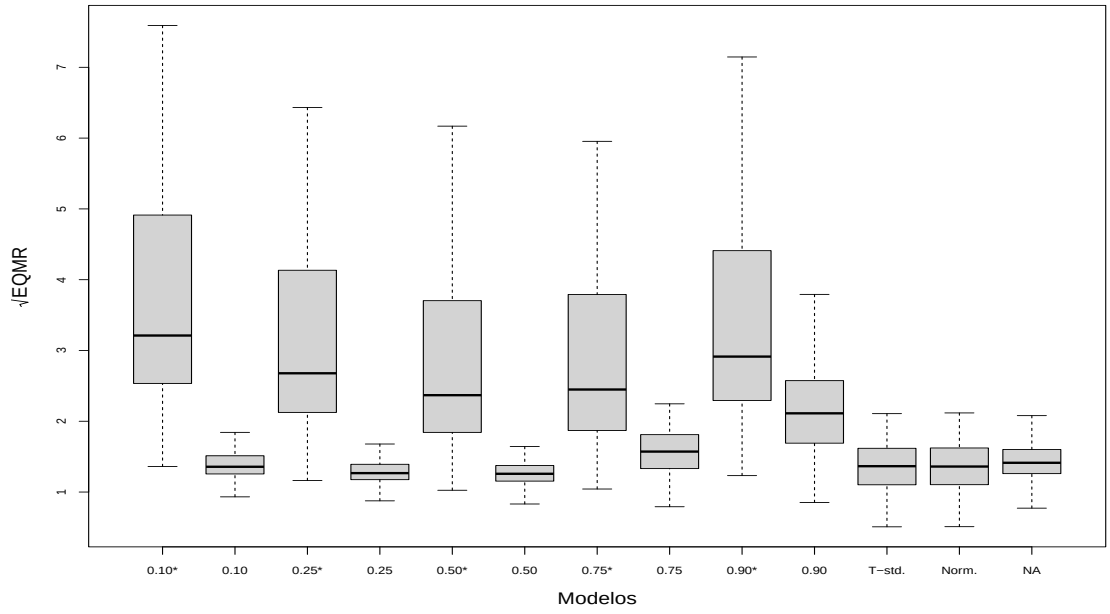


Figura 3.8: Boxplot da raiz do EQMR da previsão da média populacional em cada pequena área com base no modelo proposto ajustado para os quantis 0.10, 0.25, 0.50, 0.75 e 0.90 e nos modelos Normal, t-Student e Normal-Assimétrico. Os quantis com o símbolo * indicam que o ajuste do modelo proposto foi feito aos dados na escala logarítmica, assim como os modelos Normal e t-Student. Os demais resultados foram obtidos usando os dados na escala original.

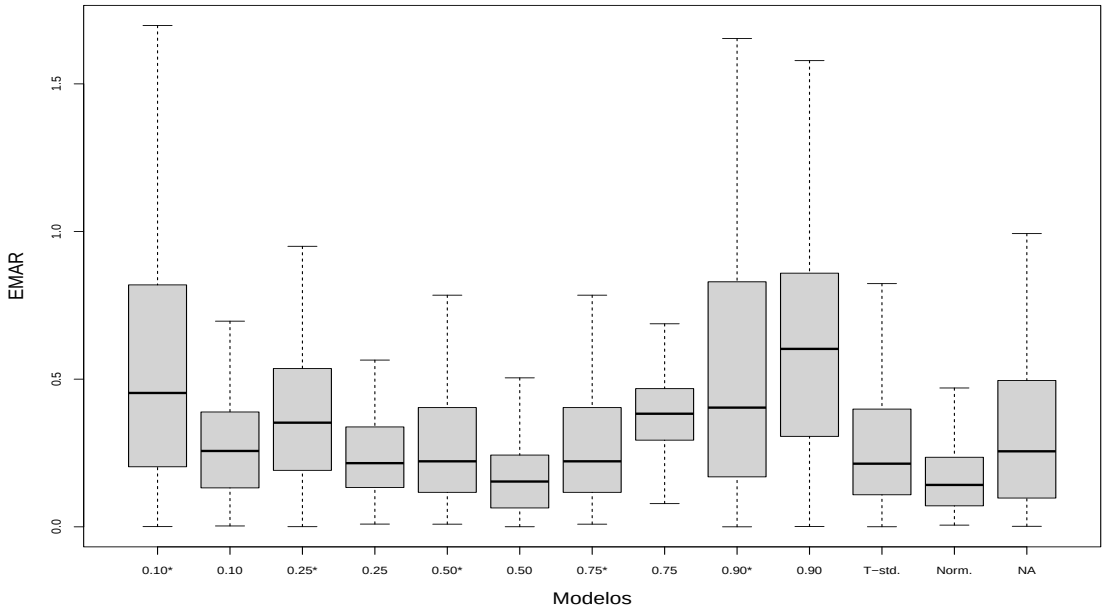


Figura 3.9: Boxplot do EMAR da previsão da média populacional em cada pequena área com base no modelo proposto ajustado para os quantis 0.10, 0.25, 0.50, 0.75 e 0.90 e nos modelos Normal, t-Student e Normal-Assimétrico. Os quantis com o símbolo * indicam que o ajuste do modelo proposto foi feito aos dados na escala logarítmica, assim como os modelos Normal e t-Student. Os demais resultados foram obtidos usando os dados na escala original.

Finalmente, a Tabela 3.4 apresenta as médias do EQMR e do EAMR para as pequenas áreas para os modelos ajustados. Segundo essas medidas o modelo proposto ajustado ao dado na escala original se mostrou competitivo com outros modelos usuais na literatura, quando quantis menores são considerados.

Tabela 3.4: Média da raiz do Erro Quadrático Médio Relativo e Erro Abssoluto Médio Relativo da previsão da média populacional em cada pequena área considerando o modelo proposto para diferentes quantis e os modelos normal, t-student e normal assimétrico.

	Dados	0.10	0.25	0.50	0.75	0.90	Normal	T-stud.	N.Assim
$\sqrt{EQMR}$	$log(y_{ij})$	5.24	4.42	3.85	3.70	4.52	1.37	1.37	-
	y_{ij}	1.38	1.27	1.26	1.57	2.14	-	-	1.40
EMAR	$log(y_{ij})$	1.40	1.13	0.78	0.78	1.45	0.19	0.29	-
	y_{ij}	0.32	0.27	0.16	0.38	0.62	-	-	0.32

3.3 Aplicação 3: Dados educacionais

Os dados utilizados nesta aplicação referem-se à Prova Brasil e estão disponíveis no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).

A Prova Brasil foi criada em 2005 a partir da necessidade de ampliação do Sistema Nacional de Avaliação da Educação Básica (Saeb), passando a avaliar todos os estudantes da rede pública urbana de ensino, de 4ª e 8ª séries do ensino fundamental. Os resultados são disponibilizados para o Brasil, Unidades da Federação, municípios e escolas participantes. Neste trabalho, como ilustração consideramos apenas os estudantes de quinta série do ensino fundamental que fizeram a prova do ano de 2011. Como área foi utilizado os municípios de Rondônia e como unidade as escolas dentro de cada município.

Esses dados consistem de $N = 24900$ escolas em $m = 52$ áreas de enumeração (municípios de Rondônia). De acordo com os objetivos deste tipo de prova, avaliou-se a qualidade da educação e as desigualdades existentes, considerando assim, três variáveis auxiliares: uma dummy indicando se o aluno pertence a escola que temd ependência administrativa estadual ou municípal, x_1, uma dummy indicando se é o aluno estuda em escola da zona urbana ou rural , x_2, e a nota do aluno na proficiência em matemática, x_3. Considerou-se como variável de interesse a proficiência média das escolas em Português, avaliando assim não apenas as desigualdades regionais dentro do estado de Rondônia, como também se há influência na área do aprendizado em matemática em relação ao

da língua portuguesa. O número de escolas por área na população (N_i) varia de 20 a 6440. Desta forma, optou-se por utilizar uma amostra de 10%, ou seja as dimensões das amostras do conjunto de dados (n_i) variam de 2 a 644.

Logo, nesta aplicação, ajustou-se o modelo proposto neste banco de dados, assumindo $p = 4$, logo x_{ij} é um vetor 4×1 e $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$. Uma análise preliminar da distribuição da variável resposta para algumas pequenas áreas selecionadas aleatoriamente pode ser vista na Figura 3.10. Os histogramas revelam uma leve assimetria a direita para algumas áreas.

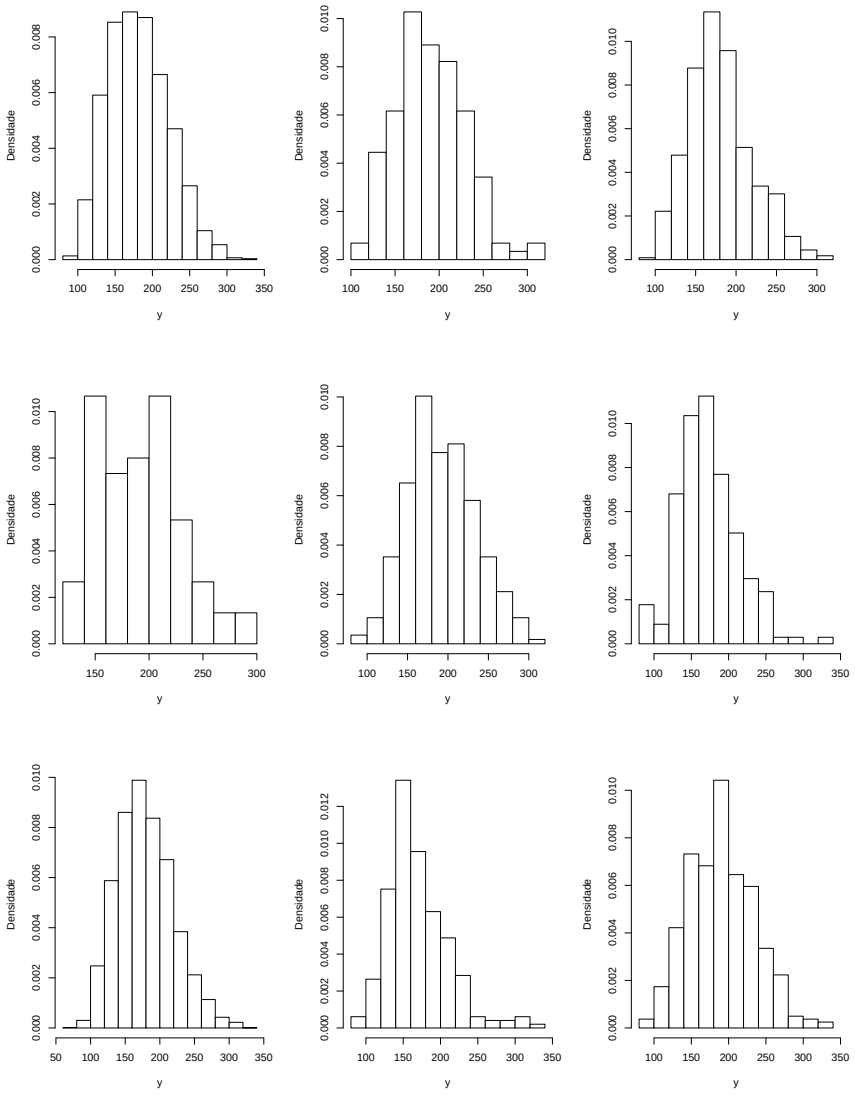


Figura 3.10: Histograma com a distribuição das notas médias das escolas na proficiência em português para algumas pequenas areas selecionadas aleatoriamente.

O modelo proposto foi ajustado para 5 diferentes quantis: 0.10; 0.25; 0.50; 0.75 e 0.90. Além disso, para fins de comparação foram ajustados também os modelos Normal, t-Student e Normal-Assimétrico já conhecidos na literatura. Também nesta aplicação além da previsão usando a distribuição preditiva (alternativa 1), considerou-se a previsão usando a distribuição *a posteriori* dos quantis (alternativa 2), ambas descritas na Seção 2.2.3.

Foram utilizadas distribuições a *priori* vagas e foram rodadas 11000 iterações do MCMC, com aquecimento de 1000 e espaçamento de 10, produzindo assim uma amostra final da distribuição *a posteriori* de tamanho 1000. As Figuras B.6 e B.7 do Apêndice B com os traços das cadeias geradas dos parâmetros β_0, β_1, β_2, β_3 e σ_ϵ e σ_v, respectivamente, sugerem que a convergência foi alcançada.

A Tabela 3.5 apresenta a média *a posteriori* e os intervalos de credibilidade de 95% para os ceoficentes da regressão obtidos para cada um dos modelos ajustados. O coeficiente β_1 relacionado a dependência administrativa da escola, não se mostra significativo para o modelo proposto e se mostra significativo para os demais modelos. O coeficiente β_2 relacionado a localização da escola não é significativo em nenhum quantil para o modelo proposto e para os demais modelos. Já a proficiência em matemática medida pelo coeficiente β_3 está diretamente relacionada com a variável de interesse para todos os modelos, tendo uma variação dependendo do quantil de interesse.

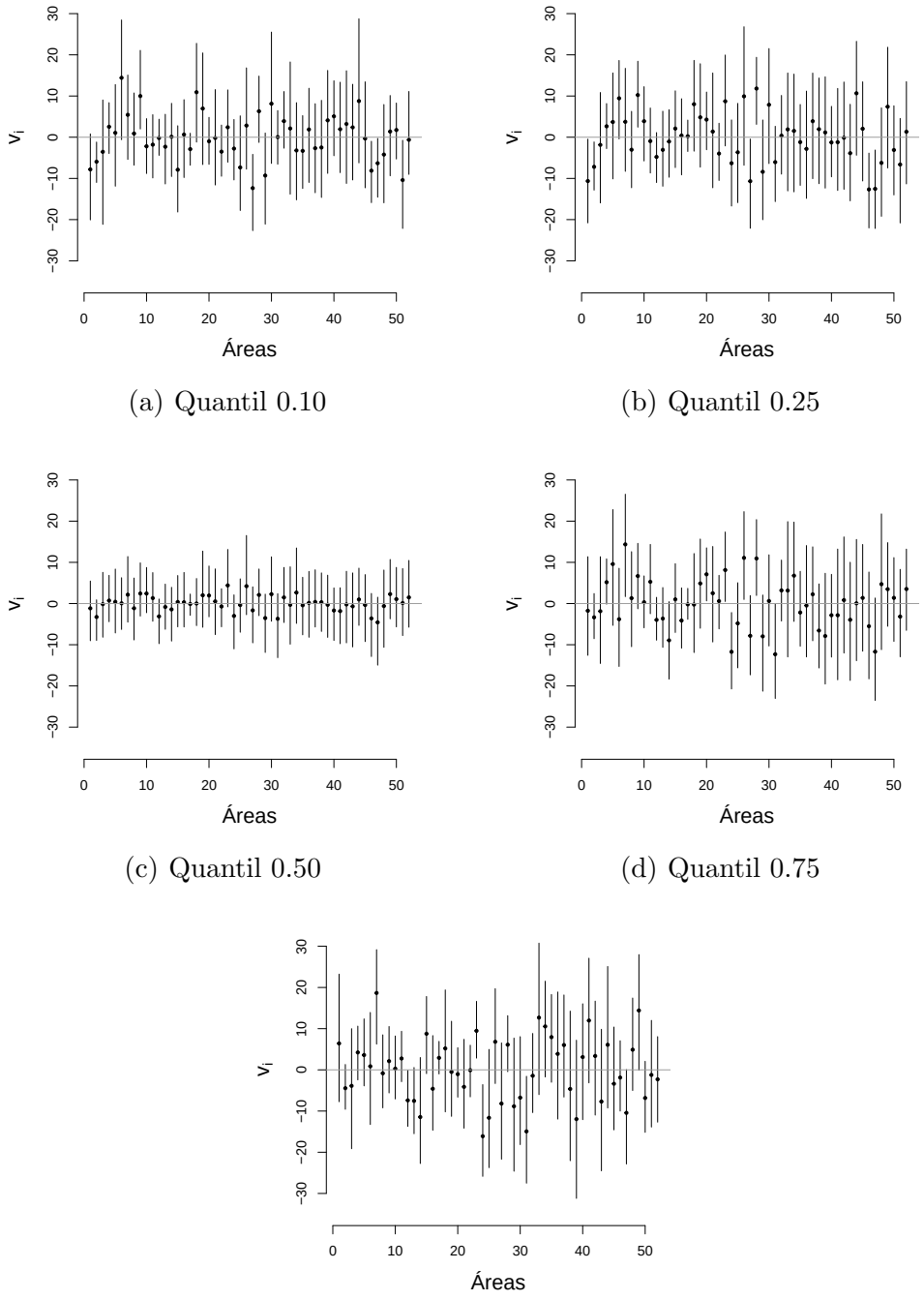
Tabela 3.5: Média *a posteriori* e intervalos de credibilidade de 95% para os parâmetros β_i, $i = 0, \ldots, 3$, obtidos a partir do ajuste do modelo proposto para cinco quantis e dos modelos Normal, t-Student e Normal-Assimétrico.

Modelo	β_0	β_1	β_2	β_3
$\tau = 0.10$	48.28 (37.20,58.22)	0.48 (-3.23,2.45)	-3.38 (-7.14,0.48)	0.49 (0.45,0.52)
$\tau = 0.25$	50.45 (39.98,61.66)	0.72 (-2.35,3.62)	-3.04 (-6.74,0.52)	0.52 (0.52,0.59)
$\tau = 0.50$	50.39 (39.22,61.94)	0.80 (-2.41,3.99)	-3.10 (-7.15,0.67)	0.55 (0.52,0.59)
$\tau = 0.75$	71.93 (61.15,83.08)	0.19 (-2.68,3.36)	-2.10 (-6.11,2.21)	0.66 (0.62,0.69)
$\tau = 0.90$	84.89 (73.00,95.78)	0.03 (-2.66,2.85)	-3.50 (-7.02,-0.07)	0.70 (0.67,0.73)
Norm.	99.16 (96.96,99.97)	15.62 (10.53,20.27)	1.92 (-5.47,10.15)	0.16 (0.12,0.20)
t-Stud.	99.17 (97.12,99.98)	15.58 (10.02 ,21.23)	1.52 (-6.93,10.23)	0.16 (0.12,0.20)
N.Assim.	98.72 (95.34,99.96)	5.81 (1.40,10.24)	-0.27 (-7.21,5.83)	0.10 (0.07,0.13)

O parâmetro de assimetria λ da distribuição Normal assimétrica obteve média *a posteriori* 2.91 com intervalo de credibilidade de 95% em $(2.66, 3.25)$, mostrando ser

significativo e indicando que um modelo que acomode assimetria pode ser interessante. Por outro lado, o parâmetro de graus de liberdade do modelo t-Student foi estimado em 40.26 com intervalo de credibilidade de 95% de $(21.65, 60.58)$, sugerindo que pode haver necessidade de adotar um modelo que acomode caudas pesadas, mas não é estritamente necessário.

A Figura 3.11 apresenta um sumário da distribuição *a posteriori* dos efeitos aleatórios v_i para cada pequena área, obtidos com base no modelo proposto para cada um dos diferentes quantis. Observou-se novamente que a significância do efeito aleatório das pequenas áreas mudam de acordo com o quantil estudado, logo estudar diferentes quantis pode levar a interpretações diferentes e mais completas acerca dos efeitos aleatórios. Por exemplo as áreas 2, 46 e 47 possuem efeitos significativos para os quantis 0.10 e 0.25 e não significativos para os demais. Já as áreas 7 e 23 possuem efeitos de área significativos nos quantis superiores de 0.75 e 0.90 e não significativos nos menores quantis.



(a) Quantil 0.10

(b) Quantil 0.25

(c) Quantil 0.50

(d) Quantil 0.75

(e) Quantil 0.90

Figura 3.11: Média *a posteriori* e intervalo de credibilidade de 95% para os efeitos aleatórios v_i em cada pequena área para o modelo proposto ajustado para diferentes quantis.

A Tabela 3.6 complementa a análise anterior mostrando a proporção dos efeitos

aleatórios que são significativos. Note que estes resultados para o modelo proposto diferem bastante dos modelos usuais, sugerindo também nesta análise que não é possível uma comparação direta de resultados, inclusive na mediana. Os resultados mostram, portanto, que a análise pode ser tornar mais completa ao avaliar diferentes quantis.

Tabela 3.6: Porcentagem de efeitos aleatórios significativos para cada modelo ajustado.

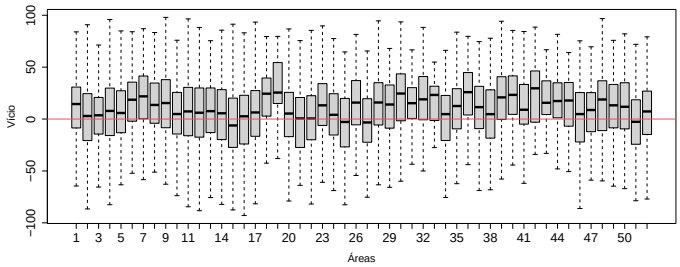
Modelo quantílico					Modelos		
0.10	0.25	0.50	0.75	0.90	t-Stud.	Normal	N.Assim
23.07	26.92	0	36.92	19.23	13.46	11.54	13.46

3.3.1 Previsão

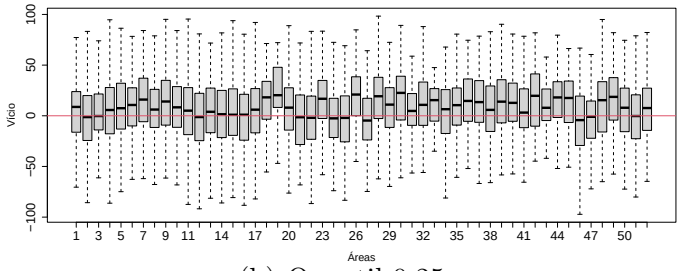
Nesta seção realizou-se previsões com base no modelo proposto e comparou-se estas com as previsões obtidas a partir dos modelos t-Student, Normal e Normal assimétrica. Nesta aplicação foi utilizada a alternativa 1 e alternativa 2 apresentadas na Seção 2.2.3, considerando cada um dos 5 quantis analisados. Como o valor verdadeiro da média em cada pequena área é conhecido, para fins de comparação os resultados foram resumidos em termos do vício, Erro Quadrático Médio (EQM) e Erro Médio Absoluto (EMA) para cada pequena área.

A Figura 3.12 mostra que os boxplots das previsões a nível de unidade utilizando a alternativa 1 de previsão (Seção 2.2.3) se encontram no geral centrados no zero, indicando que o modelo proposto gerou previsões satisfatórias. Nota-se diferenças entre as áreas, tendo por exemplo a área 18 com vício mais distante do zero nos quantis de 0.10 e 0.25. Esse fato se altera principalmente nos quantis de 0.75 e 0.90. O mesmo ocorre para outras áreas, como as áreas 24 e 25 que apesar de terem vícios centrados no zero nos quantis de 0.10, 0.25 e 0.50, se distanciam nos quantis de 0.75 e 0.90. Fazendo uma análise mais profunda notou-se que os resultados estão assossiados a assimetria dos dados nestas áreas. Dessa forma, a regressão quantílica mostra-se promissora neste caso, permitindo ainda a flexibilização de estudar diferentes quantis condicionais a variável resposta.

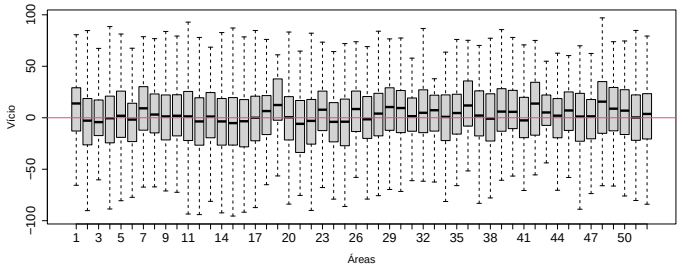
A Figura 3.13 mostra os boxplots do vício a nível de unidade para a previsão utilizando a alternativa 2 apresentada na Seção 2.2.3. Os vícios neste caso também estão centrados no zero com resultados siatisfatórios.



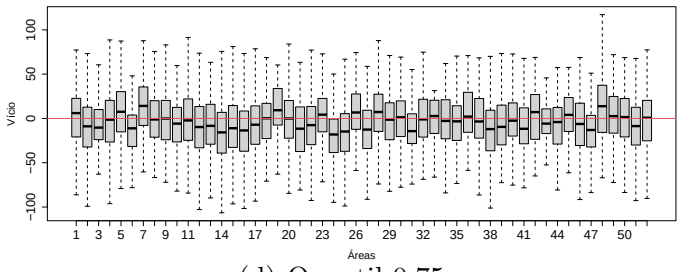
(a) Quantil 0.10



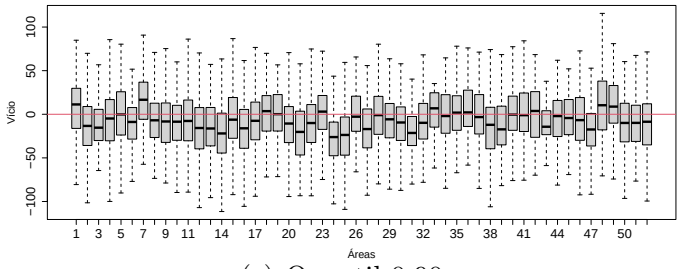
(b) Quantil 0.25



(c) Quantil 0.50



(d) Quantil 0.75



(e) Quantil 0.90

Figura 3.12: Boxplot do vício da previsão utilizando a alternativa 1 para as unidades não observadas em cada pequena área com base no modelo proposto ajustado para os quantis 0.10, 0.25, 0.50, 0.75 e 0.90.

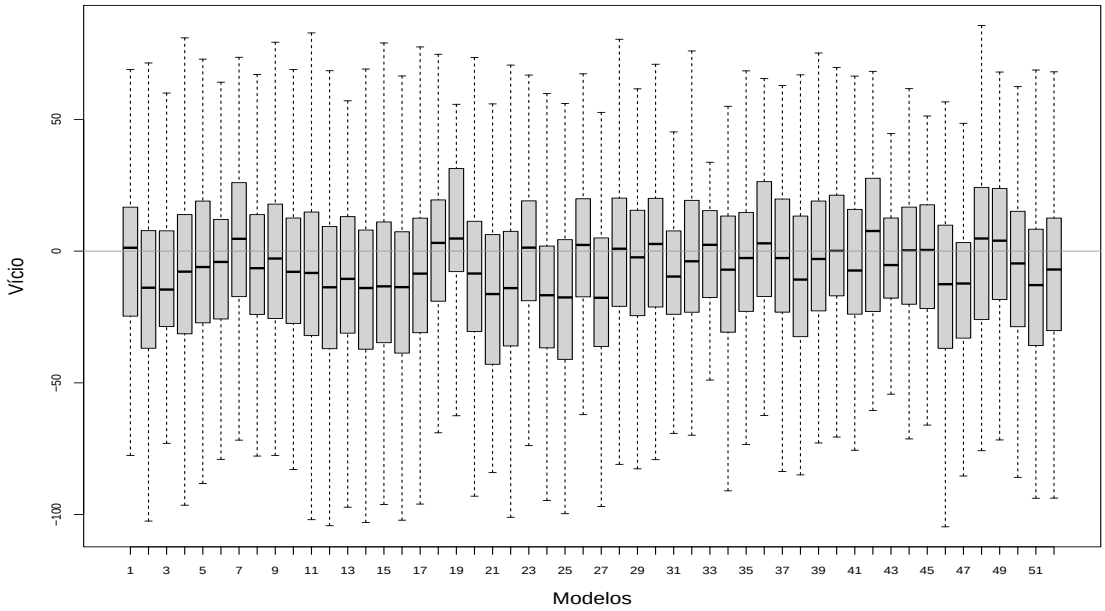


Figura 3.13: Boxplot do vício da previsão utilizando a alternativa 2 da média populacional em cada pequena área a nível de unidade.

Em particular, usando a alternativa 1 de previsão os ajustes que consideram um quantil mais próximo da mediana geram ainda melhores previsões, com respeito ao vício. Sugerimos, portanto, que a previsão seja feito com base em um ajuste do modelo para um quantil mais central neste caso. Desta forma, a Figura 3.14 é uma medida resumo para avaliar qual foi o melhor método de previsão a nível de unidade, dado que ele apresenta boxplots da divisão do melhor quantil avaliado na alternativa 1 ($\tau = 0.50$) sobre a alternativa 2. Essa razão quando menor do que um indica que a previsão feita com a alternativa 2 gerou menor vício do que a alternativa 1. Sobre essa análise temos dentre as 52 áreas, 27 com mediana menor do que 1, e 25 acima de 1. Ou seja, em relação ao quantil estudado, a alternativa 2 se mostrou melhor.

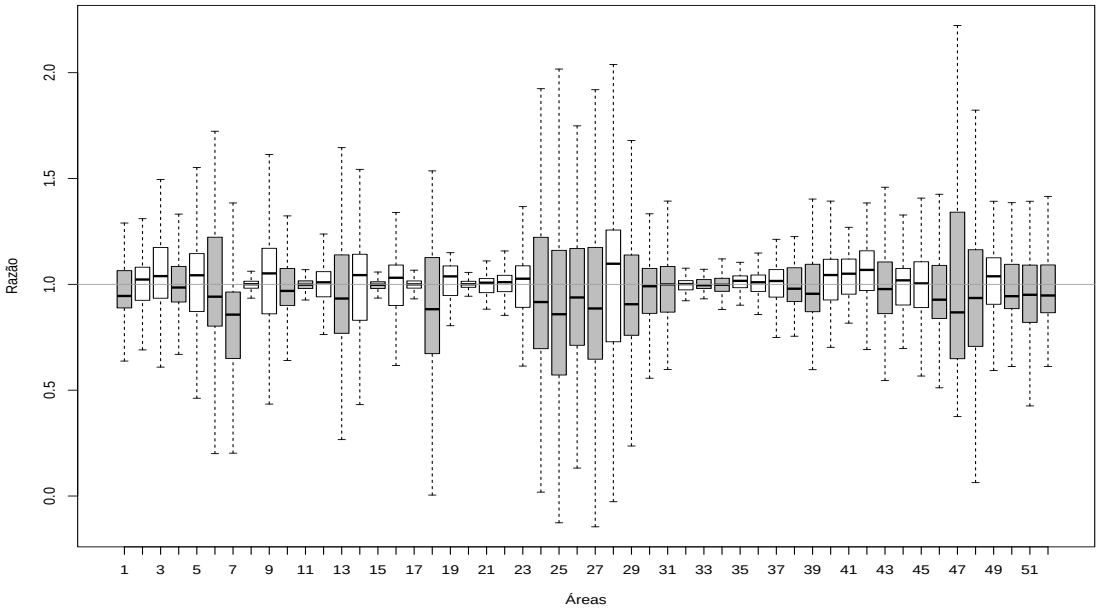


Figura 3.14: Boxplot da razão entre o vício da previsão da média populacional da alternativa 2 em relação a alternativa 1 em cada pequena área. Os boxplots em branco indicam que a divisão das previsões feitas foi maior do que 1, ou seja, que a alternativa 1 gerou menores vícios do que a alternativa 2.

A Figura 3.15 mostra os boxplots dos vícios da previsão das médias populacionais em cada área para o modelo proposto e os modelos t-Student, Normal e Normal assimétrica, com o objetivo de uma avaliação inicial mais profunda. Os boxplots em geral apresentam medianas próximas de zero. O modelo proposto quando aplicado neste conjunto de dados gera menores vícios a nível de previsão tanto na alternativa 1 para os quantis 0.25, 0.50 e 0.75 como na alternativa 2, sendo comparáveis nestes casos com os modelos usuais na literatura. O menor vício é alcançado na previsão usando o modelo ajustado na mediana.

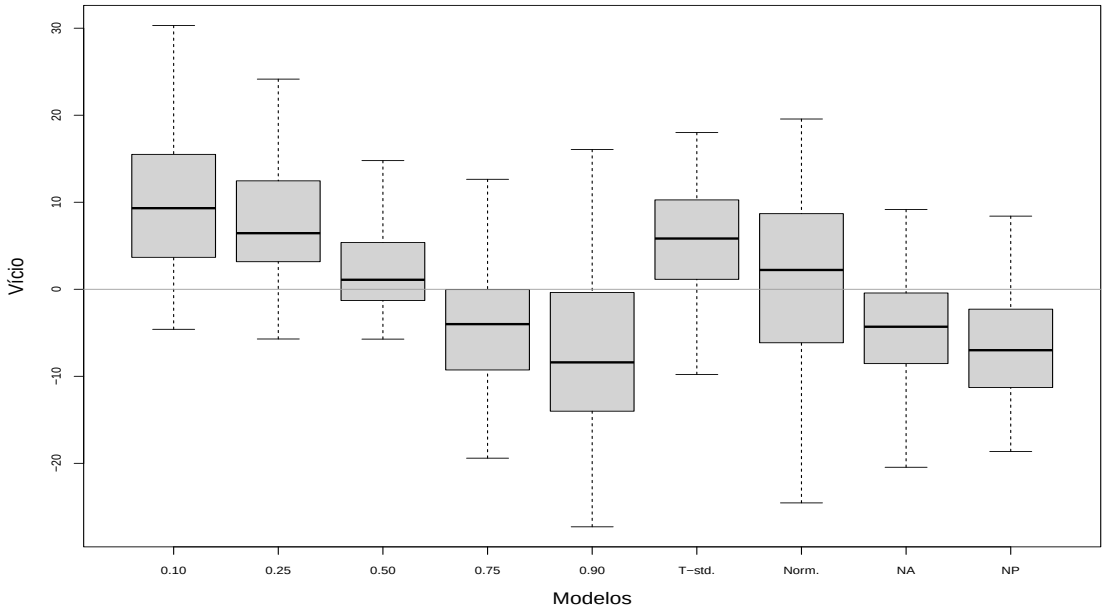


Figura 3.15: Boxplot do vício da previsão da média populacional em cada pequena área com base no modelo proposto considerando a alternativas 1 ajustado para os quantis 0.10, 0.25, 0.50, 0.75 e 0.90 e para o modelo proposto utilizando a alternativa 2 de previsão (NP) e os modelos Normal, t-Student e Normal-Assimétrico.

Complementando a análise anterior, calculou-se o EQM para as médias em cada área. A Figura 3.16 mostra o boxplots destas medidas para cada um dos modelos ajustados. Pode-se observar que as previsões usando a alternativa 2 apresentam menores valores de EQM. Isto ocorre pois essa abordagem faz uma média entre os quantis avaliados, reduzindo a variabilidade. A alternativa 1 para quantis centrais também gera resultantes promissores quando comparado com os modelos usuais da literatura. Pode-se observar que os ajustes que consideram os quantis 0.25, 0.50 e 0.75 são melhores quando comparados com os modelos Normal, t-Student e Normal-Assimétrico.

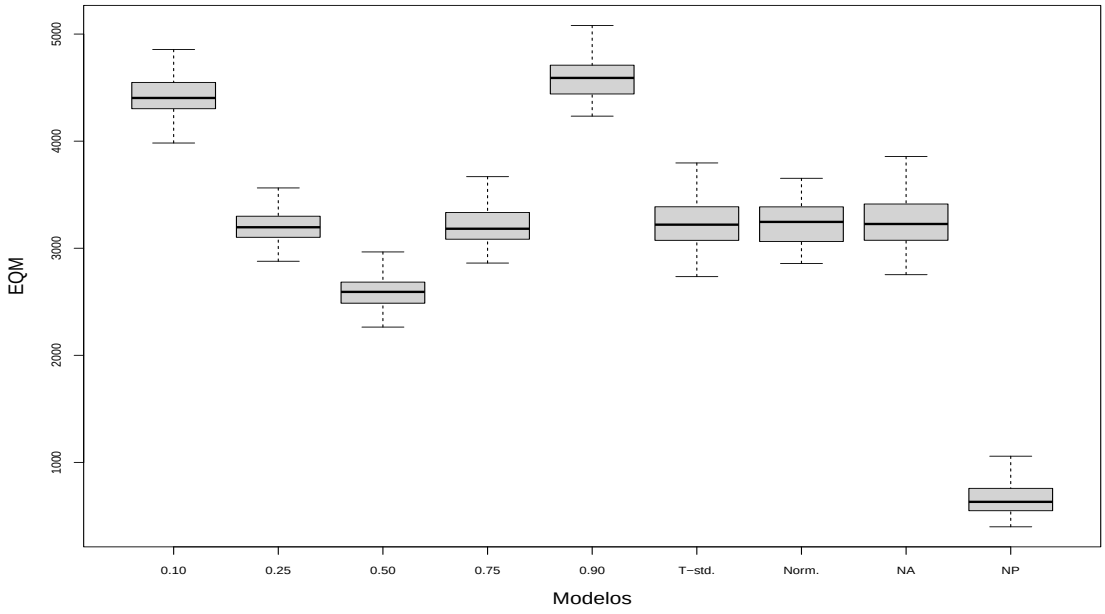


Figura 3.16: Boxplot do EQM da previsão da média populacional em cada pequena área com base no modelo proposto considerando a alternativas 1 ajustado para os quantis 0.10, 0.25, 0.50, 0.75 e 0.90 e para o modelo proposto utilizando a alternativa 2 de previsão (NP) e os modelos Normal, t-Student e Normal-Assimétrico.

Finalmente, a Tabela 3.7 apresenta as médias do EQM para as pequenas áreas para os modelos ajustados. Segundo essa medida a previsão utilizando a alternativa 1 se mostrou competitiva quando comparada com a previsão utilizando outros modelos usuais na literatura, quando ajustado para quantis centrais. Pode-se observar também que o EQM da previsão usando a alternativa 2 apresenta o menor valor quando comparado com os demais modelos. Portanto, há um ganho em fazer previsão no modelo proposto utilizando a alternativa 2 e alternativa 1 para quantis centrais quando comparada aos demais modelos.

Tabela 3.7: Média do Erro Quadrático Médio da previsão da média populacional em cada pequena área considerando o modelo proposto utilizando a alternativa 1 para os diferentes quantis e para o modelo proposto utilizando a alternativa 2 de previsão (NP) e os modelos Normal, t-Student e Normal-Assimétrico.

	0.10	0.25	0.50	0.75	0.90	T-std.	Norm.	N.Assim.	NP
EQM	4428.72	3189.80	2591.91	3207.58	4613.68	3253.67	3273.01	3243.03	654.30

Capítulo 4

Conclusão e trabalhos futuros

Neste trabalho apresentamos uma nova classe de modelos flexível para estimação em pequenos domínios a nível de unidade. A metodologia baseia-se em modelos de regressão quantílica e permite fazer análises não somente baseadas em medidas de tendência central, como a média, por exemplo, mas para qualquer quantil de interesse. A escolha do quantil pode estar associada a algum interesse específico da aplicação prática ou ainda por questões de robustez, na presença de heteroscedasticidade e valores discrepantes.

Como problemas de estimação em pequenas áreas envolvem muitas vezes tomadas de decisão, abordar quantis extremos pode ser mais interessante que a média, por exemplo, na adoção de medidas ou políticas públicas. Por outro lado, também é possível tratar diversos quantis e assim traçar uma cenário mais completo.

Como o modelo proposto tem a mesma estrutura geral de modelos usuais a nível de unidade, ou seja, uma parcela que incorpora o efeito de covariáveis e outra com efeitos aleatórios específicos de área. Porém, na regressão quantílica, além de poder estudar os efeitos das covariáveis para diversos quantis, é possível avaliar os efeitos aleatórios também variando por quantil.

O modelo proposto fundamenta-se no paradigma bayesiano de inferência e, portanto, o modelo proposto é construído assumindo a distribuição Laplace Assimétrica para a variável de interesse, independente da real distribuição dos dados.

Métodos de MCMC são utilizados para obter amostras da distribuição *a posteriori* do vetor paramétrico. Em particular, utilizamos a representação de mistura locação-escala da distribuição Laplace Assimétrica, a fim de usar apenas o amostrador de Gibbs, garantindo assim menor custo computacional.

Como é de grande interesse fazer previsão neste contexto, duas alternativas foram

apresentadas. A primeira baseia-se no modelo Laplace assimétrico e difere dependendo do quantil fixado, a segunda usa a ideia de que o valor esperado pode ser aproximada por uma média de todos os quantis, sendo essa aparentemente mais promissora.

As aplicações apresentadas neste trabalho ilustram que o modelo proposto pode fornecer resultados bem diferentes entre quantis e quando comparado com modelos tradicionais como o normal, t-student e normal assimétrico, no que se refere ao ajuste. Com relação ao poder preditivo, medido com base no vício, EQM e EAM, o modelo proposto apresenta melhores resultados em algumas situações.

Extensões futuras podem ser feitas considerando outras abordagens de regressão quantílica bayesiana que não só usando a distribuição Laplace Assimétrica. Uma alternativa por exemplo seria usar uma extensão mais flexível desse modelo, conhecida como Laplace Assimétrica generalizada (Yan e Kottas, 2017). Por fim, sugere-se também um estudo mais a fundo da alternativa 2 de previsão e propor para este método uma medida de variabilidade, dado que o modelo proposto para este tipo de previsão só estima a média. Neste mesmo método uma questão relevante que desejamos ainda explorar é a previsão com quantis equiespaçados e em maior quantidade, pois acredita-se que poderia fornecer melhores previsões. Contudo, para utilização de um número maior de quantis a rotina computacional possivelmente precisará ser otimizada, dado que o tempo computacional para realizar as aplicações 2 e 3 foram grandes. Principalmente quando foi-se necessário, na aplicação 3, fazer a média *a posteriori* dos quantis na alternativa 2 de previsão.

Referências Bibliográficas

Arima, S., Datta, G. S. e Liseo, B. (2012) Objective bayesian analysis of a measurement error small area model. *Bayesian Analysis*, **7**, 363–384.

Azzalini, A. (1985) A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 171–178.

Battese, G. E., Harter, R. M. e Fuller, W. A. (1988) An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.

Bianchi, A. m., Fabrizi, E., Salvati, N. e Tzavidis, N. (2018) Estimation and testing in m-quantile regression with applications to small area estimation. *International Statistical Review*, **86**.

Chambers, R., Salvati, N. e Tzavidis, N. (2012) M-quantile regression for binary data with application to small area estimation.

Chambers, R. e Tzavidis, N. (2006) M-quantile models for small area estimation. *Biometrika*, **93**, 255–268.

Datta, G. S. e Ghosh, M. (1991) Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics*, 1748–1770.

Fabrizi, E., Salvati, N. e Trivisano, C. (2020) Robust bayesian small area estimation based on quantile regression. *Computational Statistics & Data Analysis*, **145**, 106900.

Fay, R. E. e Herriot, R. A. (1979) Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Koenker, R. e Ng, P. (2005) Inequality constrained quantile regression. *Sankhyā: The Indian Journal of Statistics*, 418–440.

Kott, P. (1989) Robust small domain estimation using random effects modelling. *Survey Methodology*, **15**, 3–12.

Kotz, S., Kozubowski, T. J. e Podgórski, K. (2001a) Asymmetric laplace distributions. 133–178.

— (2001b) Asymmetric multivariate laplace distribution. Em *The Laplace distribution and generalizations*, 239–272. Springer.

Kozumi, H. e Kobayashi, G. (2011) Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, **81**, 1565–1578.

Moura, F. A. d. S. e Holt, D. (1999) Small area estimation using multilevel models. *Survey methodology*, **25**, 73–80.

Plummer, M. et al. (2003) Jags: A program for analysis of bayesian graphical models using gibbs sampling. Em *Proceedings of the 3rd international workshop on distributed statistical computing*, vol. 124, 1–10. Vienna, Austria.

Prasad, N. N. e Rao, J. N. (1990) The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, **85**, 163–171.

R Development Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URLhttp://www.R-project.org. ISBN 3-900051-07-0.

— (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URLhttp://www.R-project.org.

Rao, J. (2003) Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, **2**, 145–169.

Rao, J. N. e Molina, I. (2015) *Small area estimation*. John Wiley & Sons.

Schoch, T. (2012) Robust unit-level small area estimation: A fast algorithm for large datasets. *Austrian Journal of Statistics*, **41**, 243–265.

Silva Ferraz, V. R. (2011) Estimação em pequenas Áreas usando modelos assimétricos. *Tese-Universidade Federal do Rio de Janeiro, IM, DME.*

Sinha, S. K. e Rao, J. (2009) Robust small area estimation. *Canadian Journal of Statistics*, **37**, 381–399.

Yan, Y. e Kottas, A. (2017) A new family of error distributions for bayesian quantile regression. *Relatório técnico*, University of California, Santa Cruz.

You, Y. e Rao, J. (2003) Pseudo hierarchical bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, **111**, 197–208.

Yu, K. e Moyeed, R. A. (2001) Bayesian quantile regression. *Statistics & Probability Letters*, **54**, 437–447.

Anexo A

Resultados úteis

A.1 Distribuição gaussiana inversa generalizada

Como definido em Jorgensen (1982), Y tem distribuição gaussiana inversa generalizada, representada por GIG com parâmetros λ,χ e ψ, denotada por $Y \sim GIG(\lambda, \chi, \psi)$ se a sua f.d.p. é dada por:

$$f(y|\lambda, \chi, \psi) = \frac{(\chi/\psi)^2}{2\kappa_\lambda(\sqrt{\chi\psi})} y^{\lambda-1} exp[\frac{-1}{2}(\chi y^{-1} + \psi y)] \tag{A.1}$$

para $y \epsilon R$, sendo $\lambda \epsilon R$, χ, $\psi \epsilon R$, κ_λ é a função de Bessel modificada do terceiro tipo com índice λ.

A.2 Demostração dos resultados da seção 2.1.2

Demonstração (Kotz et al., 2001b):

Seja $\epsilon \sim LA_\tau(0, 1)$.Então a f.d.p. de ϵ é dada por:

$$f(\epsilon|\tau) = \tau(1-\tau)exp(-\rho_\tau(\epsilon)), sendo$$
$$\rho_\tau(\epsilon) = \epsilon\tau - I(\epsilon \leq 0)$$

A função característica de ϵ pode ser obtida como:

$$\begin{aligned}
\varphi_\epsilon(t) &= E[e^{it\epsilon}] \\
&= \int_{-\infty}^{\infty} e^{it\epsilon}\epsilon f(\epsilon)d\epsilon \\
&= \tau(1-\tau)[\int_{-\infty}^{0} \exp(it\epsilon + (1-\tau)\epsilon)d\epsilon + \int_{0}^{\infty} \exp(it\epsilon - q\epsilon)d\epsilon] \\
&= \tau(1-\tau)[\frac{\exp(it+1-\tau)\epsilon}{it+1-\tau}]_{-\infty}^{0} + [\frac{\exp{(it-\tau)\epsilon}}{it-\tau}]_{0}^{\infty} \\
&= \tau(1-\tau)[\frac{1}{it+1-\tau} - 0] + [0 - \frac{1}{it-\tau}] \\
&= \tau(1-\tau)[\frac{1}{(it+1-\tau)(\tau-it)}] \\
&= [\frac{t^2 - i(1-2\tau)t + \tau(1-\tau)}{^\tau(1-\tau)}]^{-1} \\
&= [\frac{1}{\tau(1-\tau)}t^2 - i(\frac{1-2\tau}{\tau(1-\tau)})t + 1]^{-1}
\end{aligned}$$

Por outro lado, sejam $U \sim e(1)$ e $Z \sim N(0,1)$. A função característica de $\epsilon* = a_\tau U + b_\tau\sqrt{U}Z$ pode ser expressada como:

$$\begin{aligned}
\varphi_{\epsilon*}(t) &= E[e^{it\epsilon*}] \\
&= E[e^{it(a_\tau U + b_\tau\sqrt{U}Z)}] \\
&= E[E[\exp^{it(a_\tau U + b_\tau\sqrt{U}Z)}]|U=u] \\
&= \int_{0}^{\infty} exp(ita_\tau u)E[exp(itb_\tau\sqrt{u}Z)]e^{-u}du.
\end{aligned}$$

Sabe-se que se $Z \sim N(0,1)$, então sua função característica é:

$$\varphi_z(t) = \exp[-\frac{1}{2}t^2]$$

que substituído na Equação vista, resulta em:

$$\begin{aligned}
&= \int_{0}^{\infty} exp(ita_\tau u)exp(-\frac{1}{2}t^2b_\tau^2u)]e^{-u}du. \\
&= \int_{0}^{\infty} \exp{(-u(1+\frac{1}{2}t^2b_\tau^2 - ia_\tau t))}du \\
&= [\frac{exp(-u(1+\frac{1}{2}t^2b_\tau^2 - ia_\tau t)}{-(1+\frac{1}{2}t^2b_\tau^2 - ia_\tau t}]_{0}^{\infty} \\
&= [\frac{b_\tau^2}{2}t^2 - ia_\tau t + 1)]^{-1}
\end{aligned}$$

Finalmente, as duas funções características são equivalentes quando $a_\tau = \frac{1-2\tau}{(1-\tau)}$ e $b_\tau^2 = \frac{2}{\tau(1-\tau)}$,

Apêndice A

Rotina computacional

```
library(rlist)
library(mvtnorm)
library(invgamma)
library(ggplot2)
library(dplyr)
library(tidyr)
library(xtable)
library(gamlss)
require(GeneralizedHyperbolic)
require(ald)

#Modelo quantilico bayesiano a nivel de Unidade #

#set.seed(200)
# Geracao de dados artificiais
m = 50 # número de areas
n_i = sample(20:100,m)
n = sum(n_i)

# coeficientes beta verdadeiros
beta <-c(10, 5, -2, 1)
# gerando matriz de covariaveis
x = list() #
```

```
for(i in 1:m)
{
x[[i]] = matrix(NA,nrow=length(beta),ncol=n_i[i])
for(j in 1:n_i[i])
{
x[[i]][,j] =
c(1,rnorm(1, 0, 1), rnorm(1, 0.5,sqrt(2)),rnorm(1,1,sqrt(5)))
}}

# verdadeiro sigma2_e, sigma2_v
sigma2_e<-1
sigma2_v<-1
vi<-rnorm(m,0,sqrt(sigma2_v))

tau = 0.50 #(quantil)
a_tau = (1-(2*tau))/(tau*(1-tau))
b_tau = 2/(tau*(1-tau))

w=list()
for(i in 1:m)
{
w[[i]] = matrix(NA,nrow=1,ncol=n_i[i])
for(j in 1:n_i[i])
{
w[[i]][,j] =
c((rexp(1,1/sigma2_e)))
}}
# Gerando amostra de y
y = list()
for(i in 1:m)
{
y[[i]] = rep(NA,n_i[i])
for(j in 1:n_i[i])
{
```

```
y[[i]][j] = rnorm(1,x[[i]][,j]%*%beta+vi[i]+a_tau*w[[i]][,j],
sqrt(b_tau*sigma2_e*w[[i]][,j]))
}}
# Ajuste do modelo
iter = 21000
# Hiperparametros
a_v<-0.01
b_v<-0.01
a_e<-0.01
b_e<-0.01

#valores iniciais
beta0 = runif(length(beta),0,10)
sigma2_e0 = 1
sigma2_v0 = 1
vi0 = rnorm(m)

w0=list()
for(i in 1:m)
{
w0[[i]] = matrix(NA,nrow=1,ncol=n_i[i])
for(j in 1:n_i[i])
{
w0[[i]][,j] =
c((rexp(1,2)))
}}

lista1 = x # Covariaveis
lista2 = y # Variavel resposta
lista3 =w0 # valor inicial para wij

# Selecionando a amostra

tamanhos = n_i
```

```
selecionados = list()
ns=NULL
for(i in 1:length(lista1)){
ns[i]=ceiling(0.2*tamanhos[i])
selecionados[[i]] = sort(sample(1:tamanhos[i],ns[i]))
}
xs = x; ys = y ;ws= w0

for(i in 1:length(lista1)){
xs[[i]] = matrix(lista1[[i]][,selecionados[[i]]],
ncol=length(selecionados[[i]]))
ys[[i]] = lista2[[i]][selecionados[[i]]]
ws[[i]] = matrix(lista3[[i]][,selecionados[[i]]],
ncol=length(selecionados[[i]]))
}
xs; ys; ws
ntotals = sum(ns)
betapost = matrix(NA,nrow=length(beta),ncol=iter)
sigmaepost = c() #
v_ipost = matrix(NA,nrow=m,ncol=iter)

wpost=list()
for(i in 1:m)
{
wpost[[i]] = matrix(NA,nrow=iter,ncol=length(ys[[i]]))
for(j in 1:length(ys[[i]]))
{
wpost[[i]][,j] =
c(NA)
}}
sigmavpost = c()

# iteracao 1
vic = vi0
```

```
sigma2_ec = sigma2_e0
betac = beta0
sigma2_vc = sigma2_v0
wijc=ws

betapost[,1]<-beta0
sigmaepost[1]<-sigma2_e0
sigmavpost[1]<-sigma2_v0
v_ipost[,1]<-vi0

for(l in 2:iter ){
xprimex = 0
for(i in 1:m){
for(j in 1:length(ys[[i]])){
xprimex = xprimex + (xs[[i]][,j]%*%t(xs[[i]][,j])/wijc[[i]][,j])
}}
# sampling beta
mediabeta = rep(0,length(betac))
varibeta = solve(xprimex/(b_tau*sigma2_ec))
for(i in 1:m)
{
for(j in 1:length(ys[[i]]))
{
mediabeta = mediabeta + (xs[[i]][,j]/(b_tau*sigma2_ec*wijc[[i]][,j]))*
(ys[[i]][j]-vic[i]-a_tau*wijc[[i]][,j])
}}
mediabeta = varibeta%*%mediabeta

betac <- rmvnorm(n = 1,mediabeta,varibeta)

sigmawij = (a_tau^2/(b_tau*sigma2_ec))+(2/sigma2_ec)

for(i in 1:m)
{
```

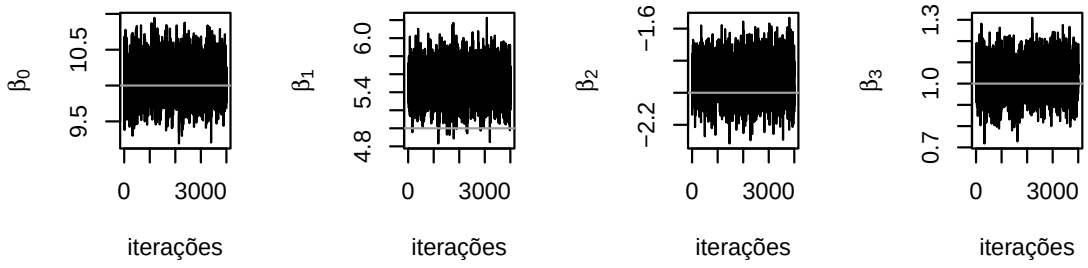
```
for(j in 1:length(ys[[i]]))
{
muwij =((ys[[i]][j]-(xs[[i]][,j])%*%t(betac)-vic[i])^2)/
(b_tu*sigma2_ec)
 wijc[[i]][,j] <- rgig(n = 1,muwij,sigmawij,0.5)
}}
# sampling sigma2_e
alphasige = 3*ntotals/2 + a_e
soma = 0
for(i in 1:m)
{
for(j in 1:length(ys[[i]]))
{
valor=(ys[[i]][j]-xs[[i]][,j]%*%t(betac)-vic[i]-a_tau*
wijc[[i]][,j])^2/(2*b_tau*wijc[[i]][,j])+wijc[[i]][,j]
soma = soma + valor
}}

betasige= soma +b_e
sigma2_ec <- 1/rgamma(n = 1,alphasige,betasige)
# sampling sigma2_v

alphasigv = m/2 + a_v
betasigv = 0.5*sum(vic^2) + b_v
sigma2_vc <- 1/rgamma(n = 1,alphasigv,betasigv)
# sampling v_i
for(i in 1:m)
{
mediav = 0
invariv = (1/(sigma2_vc))
for(j in 1:length(ys[[i]]))
{
invariv = invariv+(1/(b_tau*sigma2_ec*wijc[[i]][,j]))
mediav = mediav+((ys[[i]][j])-xs[[i]][,j]%*%t(betac)-a_tau*
```

```
wijc[[i]][,j])/(b_tau*sigma2_ec*wijc[[i]][,j])
}
variv=1/invariv
mediav = mediav*variv
vic[i] <- rnorm(n = 1,mediav,sqrt(variv))
#   vic[i] <- vi[i]
}
#  guardando as amostras correntes
betapost[,l]<-betac
sigmaepost[l]<-sigma2_ec
sigmavpost[l]<-sigma2_vc
v_ipost[,l]<-vic

for(i in 1:m)
{
for(j in 1:length(ys[[i]]))
{
wpost[[i]][1,j]<-wijc[[i]][,j]
wpost[[i]][l,j]<-wijc[[i]][,j]
}}
print(l)
}
nbur=1000 #
lag = 10
range= seq((nbur+1),iter,lag) # faixa de valores aproveitados

#previsao #
# armazenar o y de previsao#
ypredt=list()
for(i in 1:m)
{
ypredt[[i]] = matrix(NA,nrow=iter,ncol=length(y[[i]]))
for(j in 1:length(y[[i]]))
{
```

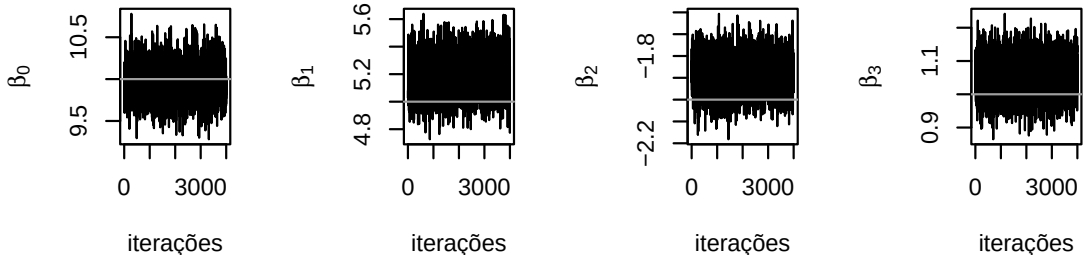
```
ypredt[[i]][,j] =
c(NA)
}}
# Gerando agora da Laplace Assimetrica diretamente#

for(i in 1:m)
{
for(j in (1:n_i[i])[-selecionados[[i]]])
{
for( l in 1:iter){
ypredt[[i]][l,j] = x[[i]][,j]%*%betapost[,l]+v_ipost[i,l]+
rALD(1,0,sqrt(sigmaepost[l]^2),tau)
}}}
```

Apêndice B

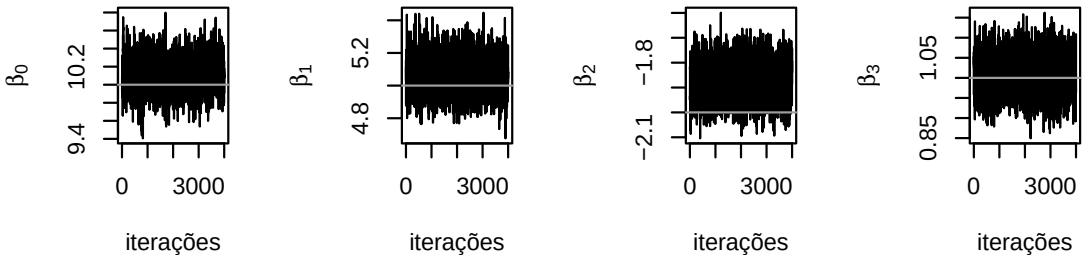
Traços das cadeias dos parâmetros



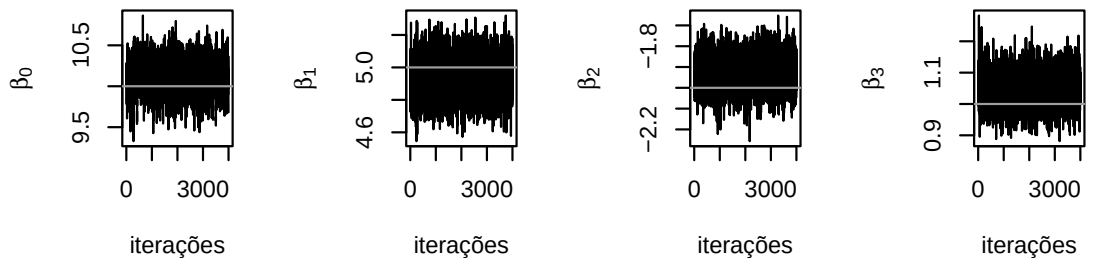
(a) Quantil 0.10



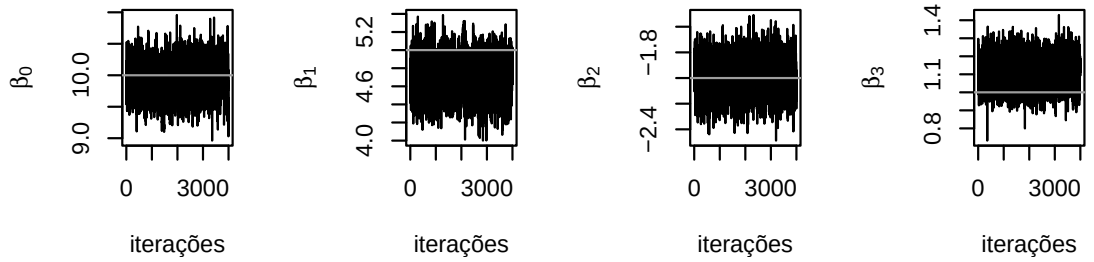
(b) Quantil 0.25



(c) Quantil 0.50

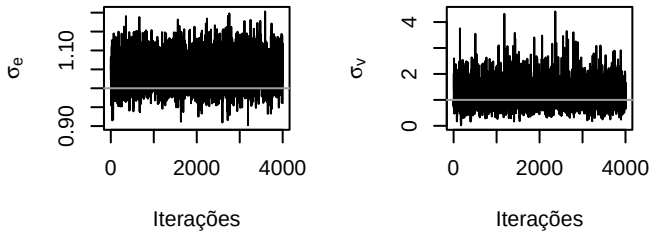


(d) Quantil 0.75

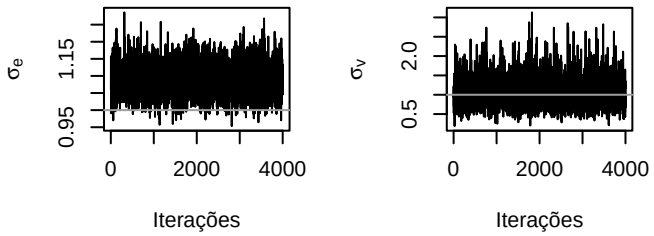


(e) Quantil 0.90

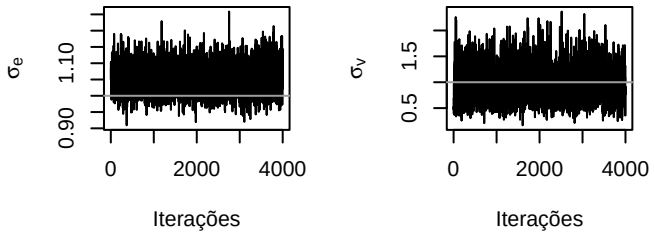
Figura B.1: Monitoramento das cadeias para estimação dos parâmetros β_i do modelo de RQB na aplicação de dados artificiais. Dados gerados dos quantis (0.10, 0.25, 0.50, 0.75 e 0.90).



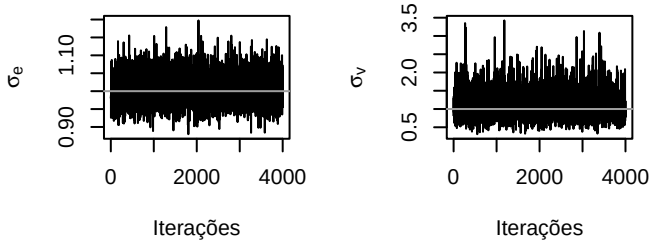
(a) Quantil 0.10



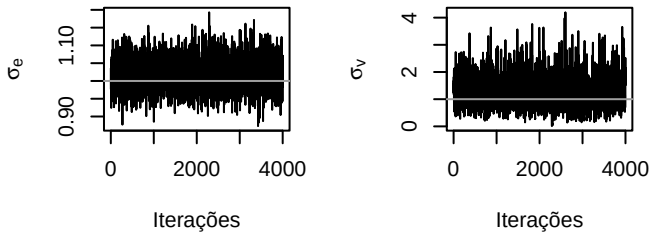
(b) Quantil 0.25



(c) Quantil 0.50

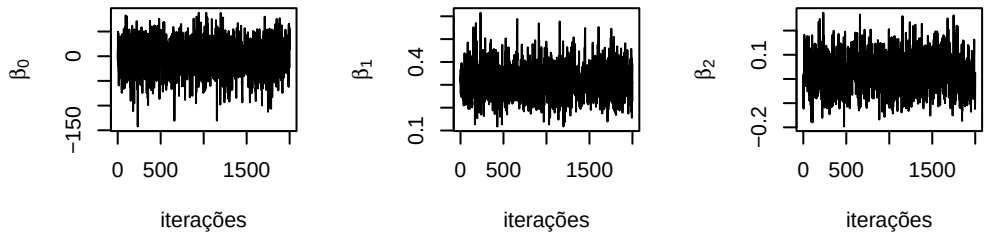


(d) Quantil 0.75

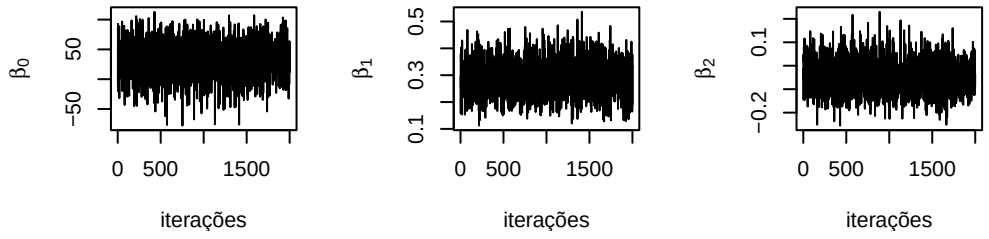


(e) Quantil 0.90

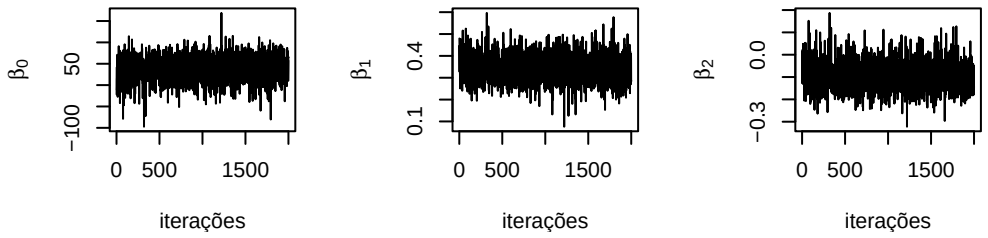
Figura B.2: Monitoramento das cadeias para estimação dos parâmetros β_i do modelo de RQB na aplicação de dados artificiais. Dados gerados dos quantis (0.10, 0.25, 0.50, 0.75 e 0.90).



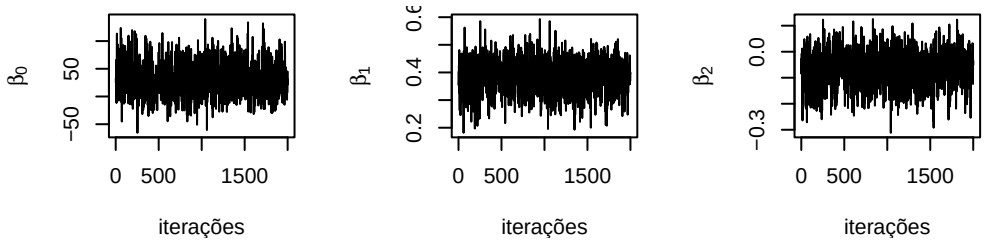
(a) Quantil 0.10



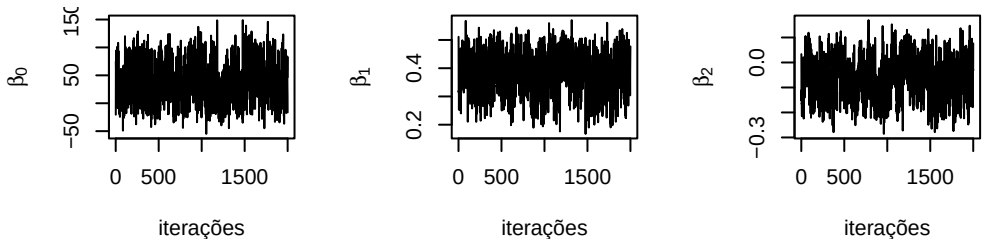
(b) Quantil 0.25



(c) Quantil 0.50

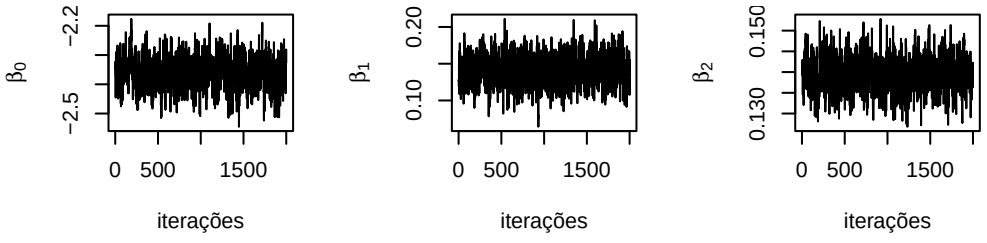


(d) Quantil 0.75

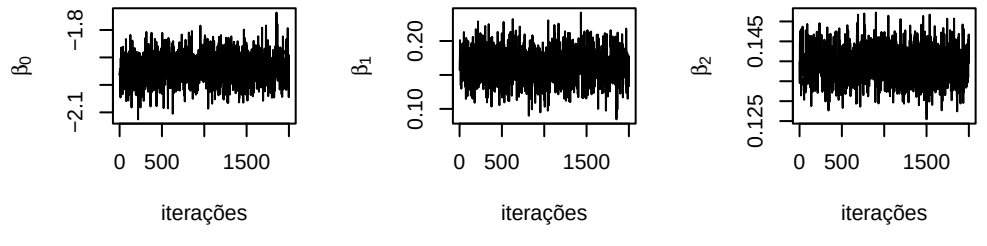


(e) Quantil 0.90

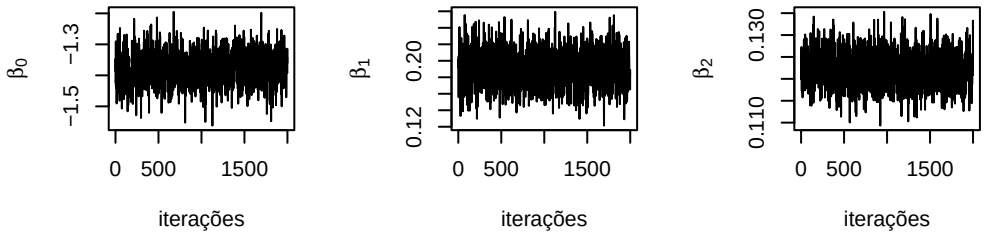
Figura B.3: Monitoramento das cadeias para estimação dos parâmetros β_i no modelo de RQB para a primeira aplicação em dados reais. Dados gerados dos quantis (0.10, 0.25, 0.50, 0.75 e 0.90).



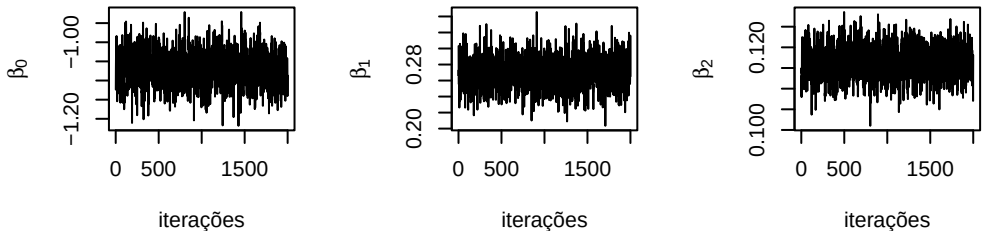
(a) Quantil 0.10



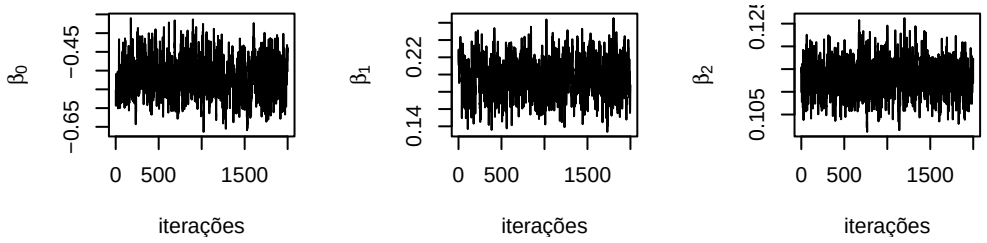
(b) Quantil 0.25



(c) Quantil 0.50

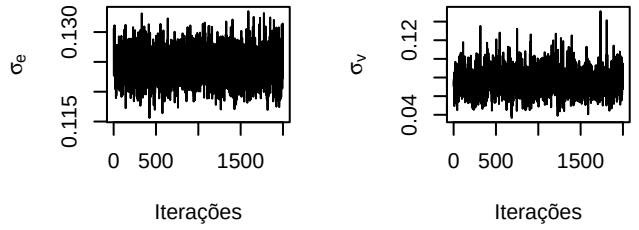


(d) Quantil 0.75

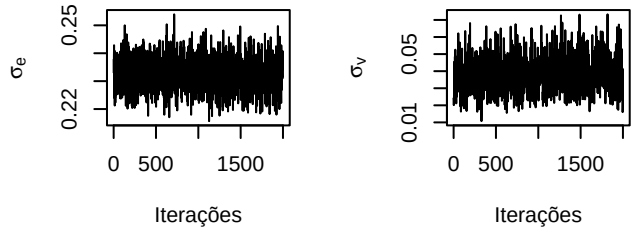


(e) Quantil 0.90

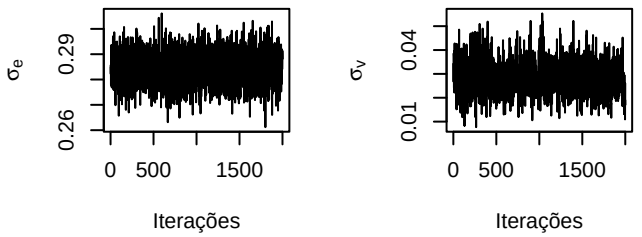
Figura B.4: Monitoramento das cadeias para estimação dos parâmetros β_i do modelo de RQB para a segunda aplicação. Dados gerados dos quantis (0.10, 0.25, 0.50, 0.75 e 0.90).



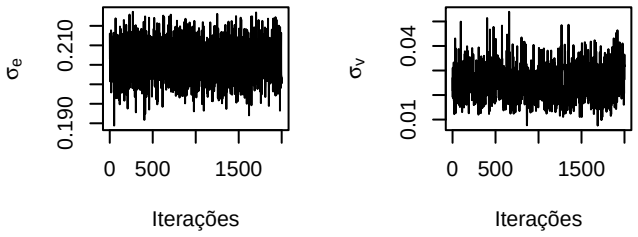
(a) Quantil 0.10



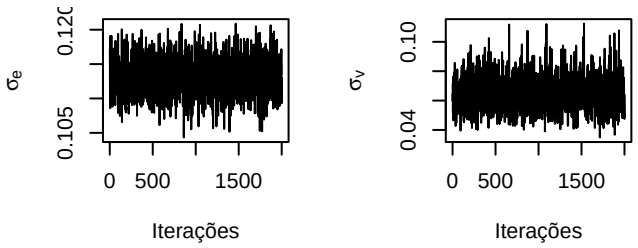
(b) Quantil 0.25



(c) Quantil 0.50

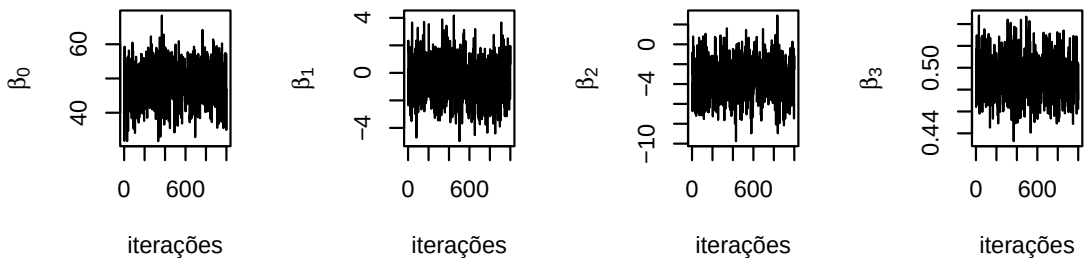


(d) Quantil 0.75

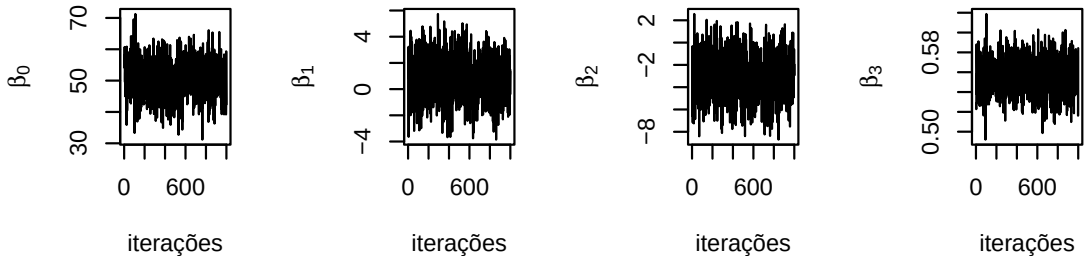


(e) Quantil 0.90

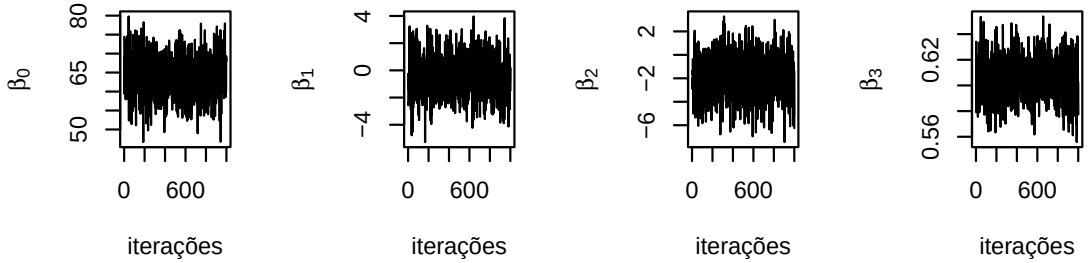
Figura B.5: Monitoramento das cadeias para estimação dos parâmetros σ_ϵ e σ_v do modelo de RQB para segunda aplicação. Dados gerados dos quantis (0.10, 0.25, 0.50, 0.75 e 0.90).



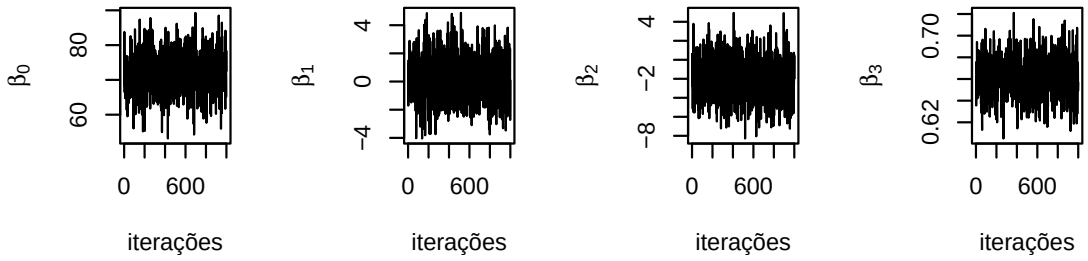
(a) Quantil 0.10



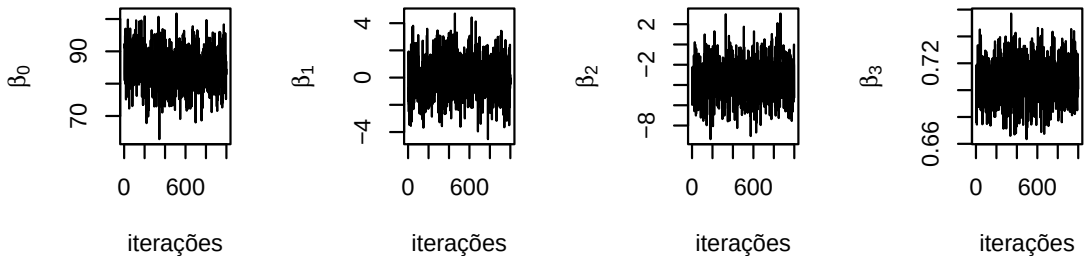
(b) Quantil 0.25



(c) Quantil 0.50

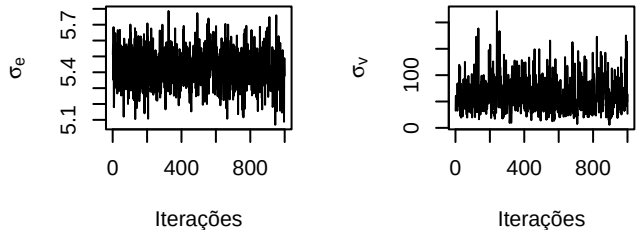


(d) Quantil 0.75

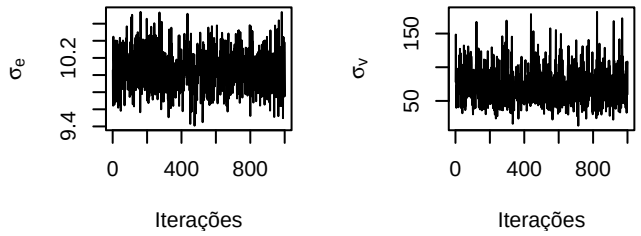


(e) Quantil 0.90

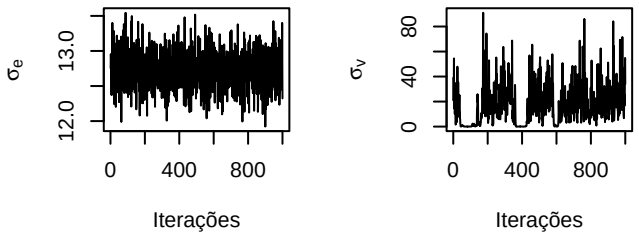
Figura B.6: Monitoramento das cadeias para estimação dos parâmetros β_i do modelo de RQB na terceira aplicação em dados reais. Dados gerados dos quantis (0.10, 0.25, 0.50, 0.75 e 0.90).



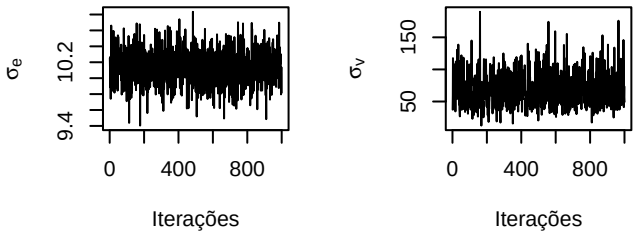
(a) Quantil 0.10



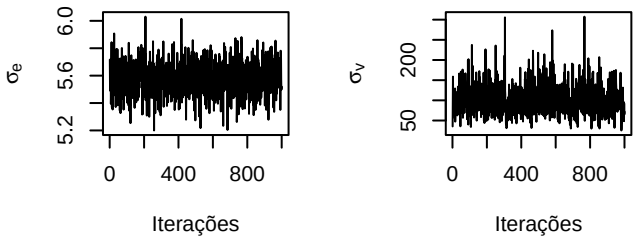
(b) Quantil 0.25



(c) Quantil 0.50



(d) Quantil 0.75



(e) Quantil 0.90

Figura B.7: Monitoramento da autocorrelação das cadeias para estimação dos parâmetros σ_ϵ e σ_v do modelo de RQB para terceira aplicação. Dados gerados dos quantis (0.10, 0.25, 0.50, 0.75 e 0.90).