Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos Não Lineares Truncados Mistos para Locação e Escala

Carolina Costa Mota Paraíba

São Carlos
2015

Carolina Costa Mota Paraíba

Modelos Não Lineares Truncados Mistos para Locação e Escala

Tese apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar como parte dos requisitos para obtenção do título de doutor em estatística.

Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

São Carlos
2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar

P221mL

Paraíba, Carolina Costa Mota.
Modelos não lineares truncados mistos para locação e escala / Carolina Costa Mota Paraíba. -- São Carlos : UFSCar, 2015.
129 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2015.

1. Estatística. 2. Modelos não lineares mistos. 3. Máxima verossimilhança iterativa. 4. Análise de diagnóstico. 5. Análise bayesiana. 6. Diagnóstico bayesiano. I. Título.

CDD: 519.5 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado da candidata Carolina Costa Mota Paraíba, realizada em 14/01/2015:

Prof. Dr. Carlos Alberto Ribeiro Diniz
UFSCar

Prof. Dr. Gustavo Leonel Gilardoni Avalle
UnB

Profa. Dra. Hildete Prisco Pinheiro
UNICAMP

Prof. Dr. José Galvao Leite
UFSCar

Prof. Dr. Victor Hugo Lachos Dávila
UNICAMP

Agradecimentos

Agradeço a minha mãe, Luíza, e ao meu pai, Lourival, pelos muitos anos de amor, apoio e educação.

Ao meu orientador, Professor Carlos Diniz, pelos muitos ensinamentos, críticas e incentivos.

À Aline Maia e ao Lineu Rodrigues por me fornecerem os dados da Bacia do Rio Buriti Vermelho e pela valiosa discussão acerca dos dados experimentais.

Aos professores do Departamento de Estatística da Universidade Federal de São Carlos pelo excelente ambiente de pesquisa e aos funcionários por sempre estarem dispostos a ajudar.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo auxílio financeiro concedido para a realização da minha tese de doutorado.

Resumo

Neste trabalho, apresentamos uma classe de modelos não lineares truncados mistos onde a característica de truncamento dos dados é incorporada ao modelo estatístico assumindo-se que a variável de interesse, isto é, a variável truncada, possui uma função de distribuição truncada que, por sua vez, corresponde a uma função de distribuição condicional obtida ao se restringir o suporte de alguma função de distribuição de probabilidade. A família de modelos não lineares truncados mistos para locação e escala é construída sob a perspectiva de modelos não lineares generalizados mistos e considerando uma classe de distribuições indexadas por parâmetros de locação e escala. Assumimos que o parâmetro de locação da variável resposta é associado a uma função não linear contínua de um conjunto de covariáveis e parâmetros desconhecidos e a efeitos aleatórios não observáveis, e que o parâmetro de escala das respostas pode ser caracterizado por uma função contínua das covariáveis e de parâmetros desconhecidos. Os modelos não lineares truncados mistos para locação e escala, aqui apresentados, são construídos supondo limites de truncamento aleatórios, porém, modelos não lineares truncados mistos com limites fixos e conhecidos são prontamente obtidos como casos particulares desses modelos. Nos modelos construídos sob a suposição de limites de truncamentos aleatórios, a função de verossimilhança é escrita em função dos parâmetros da distribuição da variável resposta truncada e dos parâmetros das distribuições das variáveis de truncamento. Para o caso particular de limites fixos e conhecidos, a função de verossimilhança será apenas uma função dos parâmetros da distribuição truncada assumida para a variável resposta de interesse. As equações de verossimilhança dos modelos, aqui propostos, não possuem soluções analíticas e, sob a perspectiva frequentista de inferência estatística, os parâmetros do modelo são estimados pela maximização direta da função de log-verossimilhança via um procedimento iterativo. Consideramos, também, uma análise de diagnóstico para verificar a adequação do modelo, observações discrepantes e/ou influentes, usando resíduos padronizados e medidas de influência global e influência local. Sob a perspectiva Bayesiana de inferência estatística, as estimativas dos parâmetros dos modelos propostos são definidas como as médias a *posteriori* de amostras obtidas via um algoritmo do tipo cadeia de Markov Monte Carlo das distribuições a *posteriori* dos parâmetros. Para a análise de diagnóstico Bayesiano do modelo, consideramos métricas de avaliação preditiva a *posteriori*, resíduos Bayesianos padronizados e a calibração de casos para diagnóstico

de influência. Como critérios Bayesianos de seleção de modelos, consideramos a soma de log-CPO e um critério de seleção de modelos baseada na abordagem Bayesiana de mistura de modelos. Para ilustrar a metodologia proposta, analisamos dados de retenção de água em solo, que são usados para construir curvas de retenção de água em solo e que estão sujeitos a truncamento pois as medições de umidade de água (a proporção de água presente em amostras de solos) são limitadas pela umidade residual e pela umidade saturada do solo amostrado.

Palavras-chave: distribuições truncadas, modelos não lineares mistos, máxima verossimilhança, algoritmo ECM, diagnóstico, análise Bayesiana, MCMC, diagnóstico Bayesiano.

Abstract

We present a class of nonlinear truncated mixed-effects models where the truncation nature of the data is incorporated into the statistical model by assuming that the variable of interest, namely the truncated variable, follows a truncated distribution which, in turn, corresponds to a conditional distribution obtained by restricting the support of a given probability distribution function. The family of nonlinear truncated mixed-effects models for location and scale is constructed based on the perspective of nonlinear generalized mixed-effects models and by assuming that the distribution of response variable belongs to a truncated class of distributions indexed by a location and a scale parameter. The location parameter of the response variable is assumed to be associated with a continuous nonlinear function of covariates and unknown parameters and with unobserved random effects, and the scale parameter of the responses is assumed to be characterized by a continuous function of the covariates and unknown parameters. The proposed truncated nonlinear mixed-effects models are constructed assuming both random truncation limits; however, truncated nonlinear mixed-effects models with fixed known limits are readily obtained as particular cases of these models. For models constructed under the assumption of random truncation limits, the likelihood function of the observed data shall be a function both of the parameters of the truncated distribution of the truncated variable and of the parameters of the distribution of the truncation variables. For the particular case of fixed known truncation limits, the likelihood function of the observed data is a function only of the parameters of the truncated distribution assumed for the variable of interest. The likelihood equation resulting from the proposed truncated nonlinear regression models do not have analytical solutions and thus, under the frequentist inferential perspective, the model parameters are estimated by direct maximization of the log-likelihood using an iterative procedure. We also consider diagnostic analysis to check for model misspecification, outliers and influential observations using standardized residuals, and global and local influence metrics. Under the Bayesian perspective of statistical inference, parameter estimates are computed based on draws from the posterior distribution of parameters obtained using an Markov Chain Monte Carlo procedure. Posterior predictive checks, Bayesian standardized residuals and a Bayesian influence measures are considered to check for model adequacy, outliers and influential observations. As Bayesian model selection criteria, we consider the sum of log -CPO and a Bayesian model selection procedure

using a Bayesian mixture model framework. To illustrate the proposed methodology, we analyze soil-water retention, which are used to construct soil-water characteristic curves and which are subject to truncation since soil-water content (the proportion of water in soil samples) is limited by the residual soil-water content and the saturated soil-water content.

Key-words: truncated distributions, nonlinear mixed-effects models, maximum likelihood, ECM algorithm, diagnostics, Bayesian analysis, MCMC, Bayesian diagnostics.

Sumário

Capítulo 1

Introdução

Em situações práticas, variáveis de interesse podem estar restritas a um determinado intervalo resultando em um conjunto de dados truncados. Por exemplo, Hausman & Wise (1977) apresentaram um estudo sobre renda onde apenas as famílias cujas rendas estavam abaixo de um certo limiar de pobreza faziam parte da amostra. Já em A'Hearn (2004), o autor analisou dados históricos de altura de soldados do exército norte americano, que são truncados inferiormente uma vez que o exercito impõe uma altura mínima aos recrutas. Há, também, uma grande quantidade de exemplos de dados truncados em estudos ambientais e de fenômenos físicos, tais como em Nadarajah (2008) em que uma distribuição beta invertida truncada foi aplicada a dados diários de medições de níveis de ozônio. Em Flecher *et al.* (2010) os autores destacaram o fato de que variáveis como pH, notas e umidade podem estar sujeitas a limites de truncamento inferior e/ou superior.

Em todos os exemplos acima mencionados, o truncamento surge como resultado de intervenção humana. Porém, o próprio truncamento também pode ser resultante de um processo aleatório, isto é, os limites de truncamento não estão sob controle do investigador, sendo resultantes do processo de amostragem e, assim, a sua ocorrência pode ser representada por um processo estocástico. Por exemplo, em um estudo de sobrevivência de idosos em casas de repouso apresentado por Hyde (1977), a variável de interesse é a idade do paciente no momento da morte, entretanto, não há registro de pacientes que são admitidos na instituição antes de uma certa idade, sendo que a idade de ingresso na instituição varia de paciente para paciente. Em estudos de síndrome de imunodeficiência adquirida (AIDS, em inglês) - ver Lagakos *et al.* (1988), Kalbfleisch & Lawless (1989) e Wang (1989) - são usados dados de transfusão de sangue onde a variável de interesse é o tempo de sobrevivência dos pacientes (tempo transcorrido da infecção pelo vírus até o diagnóstico de AIDS) porém, em geral, a data precisa de infecção do paciente é desconhecida. No entanto, as datas de infecção são conhecidas para pacientes que contraíram o vírus através de transfusão de sangue contaminado, e o momento a partir da infecção por HIV varia para cada paciente.

Tanto em casos de observações truncadas em limites fixos e conhecidos quanto em

casos de observações aleatoriamente truncadas, não consta, no conjunto de dados, um registro de observações fora dos limites de truncamento, sendo esta a característica que melhor distingue dados truncados de dados censurados, nos quais medições parcialmente observadas podem ocorrer e, quando ocorrem, são registradas. Como argumentado em Greene (2003), o truncamento é uma característica da distribuição de probabilidade da qual as observações provêm e, portanto, para explicar a natureza truncada dos dados observados, devemos considerar uma distribuição de probabilidade truncada, que é a parte de uma distribuição que está acima, abaixo, ou entre valores especificados.

Versões truncadas de distribuição de probabilidade têm sido amplamente estudadas. A distribuição normal truncada, que é a distribuição truncada mais utilizada, foi cuidadosamente discutida em Johnson *et al.* (1994) (Seção 10.1). Lee (1979) apresentou a relação entre os dois primeiros momentos de variáveis normais multivariadas sujeitas a truncamento e del Castillo (1994) investigou o método de máxima verossimilhança para a estimação dos parâmetros da distribuição normal truncada. Nadarajah & Kotz (2008) consideraram as distribuições t-Student e F truncadas, derivando expressões explícitas para seus momentos e fornecendo procedimentos de estimativa dos parâmetros pelo método dos momentos e de máxima verossimilhança. Nadarajah (2008) apresentou uma versão truncada da distribuição beta invertida. O autor derivou expressões explícitas para os momentos da distribuição proposta e utilizou o método de máxima verossimilhança para estimativa dos parâmetros. Zhou *et al.* (2010), forneceram um programa desenvolvido no software estatístico `R` (R Development Core Team, 2009) para calcular a média, variância e probabilidade acumulada de variáveis aleatórias normais truncadas. Nadarajah (2009) apresentou as versões truncadas e expressões para os momentos de cinco distribuições com caudas pesadas, sendo elas a t-Student, F, beta invertida, Fréchet e Lévy. Pereira *et al.* (2012) propuseram uma distribuição para estudar proporções que podem assumir valor zero e um, embora não assumindo valores em $(0, c)$, $0 < c < 1$. Os autores apresentaram os momentos, vetores escores e a matriz de informação de Fisher para a distribuição em questão, batizada por eles como distribuição beta truncada inflacionada, uma distribuição resultante da mistura da distribuição trinomial e da distribuição beta definida no intervalo $(c, 1)$. As estimativas dos parâmetros foram obtidas por momentos condicionais e máxima verossimilhança. No entanto, ressaltamos que a versão truncada da distribuição beta apresentada por Pereira *et al.* (2012) difere da construção usual de distribuições truncadas, uma vez que a parte truncada da distribuição beta inflacionada truncada não é obtida como uma parte de uma distribuição que está acima, abaixo, ou entre alguns valores especificados. Em Zaninetti (2013), o autor introduz uma distribuição beta truncada à esquerda derivada da distribuição beta generalizada. A distribuição proposta é construída assumindo-se que a variável aleatória assume valores em um intervalo semiaberto $[a, b)$.

Recentemente, versões truncadas de algumas distribuições assimétricas, como a classe

proposta por Azzalini (1985), foram investigadas por diferentes autores. Flecher *et al.* (2010) derivaram uma forma recursiva para os momentos da distribuição normal assimétrica truncada, propuseram um método de momentos ponderado para a estimação dos parâmetros da distribuição e aplicaram a sua metodologia a dados de umidade relativa. Jamalizadeh *et al.* (2009) discutiram as propriedades de versões truncadas e limitadas das distribuições normal e t-Student assimétricas, utilizando-as para ajustar dados de taxas de câmbio entre a libra britânica e o dólar norte americano. Já Nadarajah & Ali (2004) propuseram uma distribuição t-Student assimétrica truncada construída a partir da versão truncada da distribuição t-Student usual.

No contexto de regressão linear truncada, Heckman (1976) propôs um estimador de mínimos quadrados corrigido, no qual o vício resultante da aplicação dos mínimos quadrados a dados truncados é caracterizado como um erro de especificação ou como um problema de omissão de variável. Este estimador foi construído incluindo a variável omitida como uma variável regressora do modelo. Porém, como os mínimos quadrados usuais são conhecidos por serem viciados quando aplicados a dados sujeitos a truncamento, uma escolha popular para a estimação dos parâmetros do modelo de regressão linear é o método da máxima verossimilhança, ou métodos baseados na verossimilhança. Em Hausman & Wise (1977), os autores propuseram um procedimento de máxima verossimilhança e forneceram um algoritmo do tipo Newton para a obtenção das estimativas dos parâmetros da regressão. Por outro lado, A'Hearn (2004) introduziu um estimador de máxima verossimilhança restrito para modelar dados de altura sujeitos a truncamento. O estimador foi derivado impondo um valor a *priori* para o desvio-padrão da variável resposta e estimando a sua média livremente. O autor usou, ainda, resultados de simulação para mostrar que a sua metodologia de máxima verossimilhança restrita funciona como um método de mínimos quadrados restrito. Nadarajah (2009) ajustou um modelo de regressão linear com erros t-Student truncados a um estudo de psicologia referente a dados de relações extraconjugais. Há, também, uma grande contribuição na área de estimação semiparamétrica e não paramétrica de modelos de regressão truncados tanto com limites fixos e conhecidos como com limites aleatórios (ver Powell (1986), Lagakos *et al.* (1988), Kalbfleisch & Lawless (1989), Wang (1989), Lee (1992), Lee (1993), Newey (2004), Cosslett (2004) e Chen & Zhou (2012)).

Modelos de regressão não lineares truncados têm sido pouco investigados sob a perspectiva de inferência paramétrica. Apenas em Bragato (2004), encontramos um exemplo de estimação paramétrica de um modelo de regressão não linear truncado. No referido trabalho, o autor considera que a variável de interesse segue distribuição normal truncada inferiormente e que seu parâmetro de locação é representado por uma função não linear de um vetor de parâmetros desconhecidos e uma variável explicativa. Exemplos de propostas de metodologias estatísticas para o estudo de modelos de regressão não lineares truncados podem ser apenas encontrados no cenário de inferência não paramétrica e semiparamé-

trica (ver Lee (1992), Ould-Saïd (2006) e de Uña Álvarez *et al.* (2010)). Desta forma, no presente trabalho propomos o desenvolvimento de modelos de regressão não lineares truncados sob uma perspectiva paramétrica, considerando um cenário de limites de truncamento aleatório, que possui como caso particular o caso de limites de truncamento fixos e conhecidos.

Observamos que os modelos de regressão truncados, tanto lineares como não lineares, não correspondem ao modelo de regressão tobit proposto por Tobin (1958). Nos modelos de regressão tobit, a variável de interesse assume valores dentro de limites prefixados, porém, existe a possibilidade de que observações ocorram nos limites da variável. Isto é, o modelo tobit é um modelo de regressão censurado no qual observações que ocorrem fora dos limites são censuradas e assumem valores nos referidos limites. Ressaltamos que modelos de regressão censurados são apropriados para conjuntos de dados onde há um registro de todos os casos, até mesmo daqueles para os quais não foi possível observar a variável resposta (casos censurados). Neste tipo de dados, as covariáveis são sempre observadas para todos os casos, até mesmo para aqueles cuja observação da variável resposta foi censurada. Já nos modelos de regressão truncado, não há informação nem para a variável resposta nem para as covariáveis de casos fora dos limites de truncamento.

Se, além da característica de truncamento, os dados também apresentam padrões de variação que possam ser atribuídos a efeitos não observáveis que provocam uma correlação entre as observações, então as técnicas usuais de regressão truncada não são adequadas já que a correlação entre as observações podem resultar em erros correlacionados. Neste cenário, a classe de modelos mistos (Laird & Ware, 1982; Pinheiro & Bates, 2000) é uma alternativa popular, que elimina a suposição, muitas vezes restritiva, de que as observações sejam independentes, e que também proporciona uma abordagem útil e flexível para a modelagem estatística de dados correlacionados. Dentre a classe de modelos mistos não lineares (Lindstrom & Bates, 1990; Lachos *et al.*, 2013), a média condicional das respostas pode ser parametrizada por uma função não linear dos parâmetros fixos, enquanto que os efeitos aleatórios podem ser introduzidos tanto de forma linear como de forma não linear.

Assim, o objetivo deste trabalho é desenvolver modelos de regressão não lineares truncados mistos para locação e escala, considerando limites de truncamento aleatórios e limites fixos e conhecidos. Os modelos são construídos sob a perspectiva de modelos não lineares generalizados mistos, com o parâmetro de locação da variável resposta sendo associado à função não linear contínua de um conjunto de covariáveis e parâmetros desconhecidos e a efeitos aleatórios não observáveis. Assumimos, ainda, que o parâmetro de escala das respostas pode ser caracterizado por uma função contínua das covariáveis e de parâmetros desconhecidos. Para os modelos não lineares mistos com presença de truncamento aleatório, consideramos as distribuições normal e beta aleatoriamente truncadas, e os modelos não lineares com limites de truncamento fixos e conhecidos são reduzidos às distribuições normal truncada e beta truncada. Ressaltamos que muitas outras dis-

tribuições truncadas poderiam ser utilizadas na construção dos modelos em questão. A escolha de se trabalhar com a distribuição normal se justifica pelo fato da mesma ser tradicionalmente utilizada por pesquisadores ao assumir modelos estatísticos para tratar problemas práticos. Já a distribuição beta é uma alternativa natural para dados que representam proporções, como são os dados de retenção de água em solo utilizados para construir curvas de retenção de água em solo, que é a motivação deste trabalho. Além disso, a distribuição beta pode assumir uma variedade de formas distintas e é amplamente conhecida por sua capacidade de acomodar bem dados com presença de assimetria.

Considerando a abordagem frequentista de inferência estatística, os parâmetros do modelo são estimados pela maximização direta da função de log-verossimilhança, através do algoritmo de otimização não linear de região de confiança (*trust region*) via um procedimento iterativo. Consideramos, também, uma análise de diagnóstico para verificar a adequação do modelo, observações discrepantes (*outliers*) e observações influentes, seguindo a metodologia de diagnóstico da análise de regressão usual. Mais especificamente, consideramos os resíduos padronizados (Cook & Weisberg, 1982) para detecção de *outliers* e para verificar a adequação do modelo. Ademais, consideramos duas medidas de influência global baseadas no princípio de deleção de casos proposto inicialmente por Cook (1977), a distância generalizada de Cook e a distância da verossimilhança, e as medidas de influência local (Cook, 1986) baseadas na perturbação da variável resposta e na perturbação de casos. Como critérios de seleção de modelos, utilizamos o critério de informação de Akaike (Akaike, 1974) e o critério de informação Bayesiano (Schwarz, 1978).

Sob a perspectiva Bayesiana de inferência estatística, as estimativas Bayesianas são computadas a partir de amostras obtidas via um algoritmo do tipo Cadeia de Markov Monte Carlo das distribuições a *posteriori* dos parâmetros e intervalos de credibilidade inter-quantil e HPD são construídos para os parâmetros do modelo. Para a análise de diagnóstico Bayesiano do modelo, consideramos métricas de avaliação preditiva a *posteriori* Gelman *et al.* (2000), dois tipos de resíduos Bayesianos padronizados (Yan & Sedransk, 2010; Albert & Chib, 1995) e a calibração de casos para diagnóstico de influência Peng & Dey (1995); Cho *et al.* (2009). Para a seleção de modelos, consideramos o critério da soma de log-CPO e um critério de seleção baseada na abordagem Bayesiana de mistura de modelos.

Para ilustrar a metodologia proposta, analisamos dados de retenção de água em solo da base de dados do Rio Buriti Vermelho de Rodrigues & Maia (2011). Estes tipos de observações estão sujeitas a truncamento, pois medições de umidade de água são observações referentes à proporção de água presente em amostras de solos que são limitadas pela umidade residual e pela umidade saturada do solo amostrado.

1.1 Motivação: curvas características de retenção de água em solo

As curvas de retenção de água em solo são ferramentas gráficas, que fornecem uma descrição visual da quantidade de água (teor de umidade) remanescente em uma amostra de solo, como uma função da tensão de sucção (potencial matricial). As curvas de retenção são importantes para estudar a relação entre o solo e a água, um fenômeno físico que afeta o uso do solo em muitas e diferentes finalidades. Um dos usos mais comuns deste tipo de curva é na determinação indireta da condutividade hidráulica não saturada, através de modelos estatísticos de distribuição de tamanho de poros (Cornelis *et al.*, 2005).

Curvas de retenção são construídas usando-se dados de retenção de água em solo e a relação entre o conteúdo de água no solo e o potencial matricial é comumente modelada por funções não lineares do tipo $y = \eta(x, \boldsymbol{\beta})$, onde y representa o conteúdo de água no solo observado, quando o mesmo é submetido a um dado potencial matricial x, e $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos. O potencial matricial, ou tensão de sucção, x é uma pressão aplicada às amostras de solo geralmente expressa na unidade de pressão Pascal ou na unidade de pressão atmosfera.

Essas curvas são ajustadas considerando-se pares, (y, x), obtidos através da aplicação de diferentes tensões, x, a uma dada amostra de solo e observando o conteúdo de água no solo, y, remanescente na amostra após a aplicação de cada nível de tensão consideradas. Isto é, as curvas de retenção relacionam y a x. Porém, dada a natureza dos dados de retenção de água em solo, sabe-se que o conteúdo de água no solo observado em um potencial matricial será sempre maior do que o conteúdo de água residual, θ_r, e menor do que o conteúdo de água no solo saturado, θ_s, o que significa que os dados de retenção de água em solo estão sujeitos a truncamento e o conteúdo de água em qualquer amostra de solo é tal que $\theta_r < y < \theta_s$.

Em dados de retenção de água em solo, as amostras de solo são usualmente coletadas ao longo de uma determinada região e, em cada local de coleta, as amostras de solo são amostradas em diferentes profundidades do solo e as medições de teor de umidade são feitas em repetições. Portanto, o conteúdo de água em solo é medido em um nível de profundidade do solo e em um nível experimental. A razão para a coleta de amostras em diferentes profundidades no mesmo local de amostragem é devida ao fato de que a capacidade de retenção de água em solo, em diferentes camadas do mesmo solo, está sujeita a diferentes propriedades hidráulicas e dinâmicas, variando entre as profundidades. Assim, há evidência de uma possível variação dos dados devido a um efeito não observado da profundidade do solo.

A relação entre o conteúdo de água no solo e o potencial matricial não é trivialmente modelado e diversas expressões analíticas não lineares do tipo $y = \eta(x, \boldsymbol{\beta})$, sendo $\boldsymbol{\beta}$ o vetor de parâmetros da curva, podem ser encontradas na literatura. Entre as expres-

sões mais utilizadas estão as propostas por Gardner (1958), Brooks & Corey (1964), van Genuchten (1980) e Fredlund & Xing (1994). Estas expressões são as mais preferidas pois fornecem uma boa aproximação da relação entre a quantidade de água no solo e o potencial matricial. Remetemos o leitor interessado a Leong & Rahardjo (1997) e Sillers *et al.* (2001), para uma revisão detalhada de diferentes expressões propostas para modelar as curvas de retenção. Neste trabalho, consideramos a expressão de Gardner (1958), o modelo proposto por van Genuchten (1980) combinado com a restrição de Burdine (1953) e o modelo de van Genuchten (1980) combinado com a restrição de Mualem (1976).

A expressão de Gardner (Gardner, 1958) é dada por

$$y = \theta_r + \frac{\theta_s - \theta_r}{1 + \beta_1 x^{\beta_2}}, \tag{1.1}$$

em que θ_r e θ_s são os conteúdos de água residual e saturado, respectivamente, β_1 está relacionado ao valor inverso de entrada de ar do solo e β_2 está relacionado à inclinação da curva.

A expressão de van Genuchten (1980) é escrita como

$$y = \theta_r + \frac{\theta_s - \theta_r}{\left[1 + (\beta_1 x)^{\beta_2}\right]^{\beta_3}}, \tag{1.2}$$

sendo que β_1 está relacionado ao valor inverso de entrada de ar no solo, β_2 está relacionado à distribuição de tamanho de poros e β_3 está relacionado à assimetria do modelo.

Burdine (1953) propôs a seguinte relação fixa entre β_2 e β_3,

$$\beta_3 = 1 - \frac{2}{\beta_2}, \tag{1.3}$$

e como $\beta_3 > 0$ segue que $\beta_2 > 2$.

Mualem (1976) propôs uma relação fixa entre β_2 e β_3 que é dada por

$$\beta_3 = 1 - \frac{1}{\beta_2}, \tag{1.4}$$

e como $\beta_3 > 0$, β_2 deve ser maior do que 1.

Em van Genuchten (1980), o autor destacou que θ_s, que é definido como o teor de umidade em $x = 0atm$, é facilmente obtido experimentalmente, estando disponível na maioria das vezes. Por outro lado, θ_r é definido como o teor de umidade em $x = 15atm$ (van Genuchten, 1980), ou ainda como um parâmetro de ajuste igual ao teor de umidade para o qual a primeira derivada de $\eta\left(x, \boldsymbol{\beta}\right)$ com respeito a x é igual a zero (van Genuchten & Nielsen, 1985).

As restrições de Burdine (1953) e Mualem (1976) têm a vantagem de reduzir o número de parâmetros do modelo (1.2), ao mesmo tempo que proporcionam uma aproximação

precisa para as curvas de retenção e sendo aplicáveis a uma variedade de solos. Por outro lado, o uso dessas restrições podem reduzir a flexibilidade da curva ajustada no que diz respeito à sua forma quando comparada com a curva obtida pelo modelo de van Genuchten (1980) (1.2) sem restrição. Além de ser mais flexível do que as suas formas restritas, os parâmetros do modelo de van Genuchten (1980) irrestrito possuem significado físico.

O método mais utilizado para estimar os parâmetros de uma curva de retenção é o método dos mínimos quadrados (ver Yates *et al.* (1992), Dourado-Neto *et al.* (2000), Cornelis *et al.* (2005), Silva *et al.* (2006) e Chao *et al.* (2008)). Porém, por haver uma necessidade de maior precisão na estimação dos parâmetros das curvas de retenção, é importante considerarmos distribuições de probabilidade e assim, considerar o teor de umidade como uma variável resposta de interesse. Além disso, como os dados de retenção de água em solo estão sujeitos a truncamento e, neste cenário, sabe-se que os procedimentos de mínimos quadrados usuais são viciados e ineficientes (ver Maddala (1983)), podendo afetar seriamente a curva estimada e as previsões baseadas na mesma. Desta forma, é muito importante levar em consideração a natureza truncada dos dados e, para isto, podemos considerar distribuições de probabilidade truncadas, que é a parte de uma distribuição que está acima, abaixo ou entre limites especificados. Portanto, neste trabalho é proposta e construída uma classe de modelos não lineares truncados. Esta classe de modelos fornece uma abordagem alternativa para a estimação de curvas de retenção de água em solo, levando em consideração a característica de truncamento presente nos dados.

1.2 Diferenças entre truncamento e censura

O truncamento resulta em amostras cujos valores estão limitados acima, abaixo ou entre limites especificados. O conceito de truncamento é diferente do conceito de censura e os dois não devem ser confundidos. Uma amostra truncada pode ser entendida como uma amostra onde todos os valores fora dos limites de truncamento são omitidos e onde nem mesmo um registro dos casos omitidos é mantido. Por outro lado, em amostras em que o fenômeno de censura ocorre, sempre é feito o registro de todos os casos, mesmo daqueles que são tidos como censurados.

A censura pode, ainda, ser vista como uma característica resultante do processo de coleta dos dados. Já o truncamento é uma característica da população da qual a amostra é coletada. Segundo Greene (2003), a diferença entre estas duas condições é muito sutil, sendo que a segunda condição pode ser gerada da primeira. Para tal, bastaria pensar em exemplos práticos onde o pesquisador descarta as amostras censuradas e mantém apenas os casos não censurados. Nestes cenários, a distribuição dos casos não censurados é truncada com respeito a população de interesse.

Portanto, tanto a censura quanto o truncamento resultam em falta de informação sobre a variável aleatória de interesse. Porém, é necessário destacar que a diferença principal

entre essas duas características dos dados é que no caso da censura há o registro do caso não observado e no truncamento não há registro de casos não observados. Esta diferença estrutural determina se os dados serão abordados como censurados ou truncados.

1.3 Apresentação dos capítulos

Neste primeiro capítulo, introdutório, apresentamos a motivação, contextualização e uma breve descrição da metodologia de construção, estimação e diagnóstico dos modelos não lineares propostos. No Capítulo 2, apresentamos a classe de distribuições de locação e escala truncadas com limites de truncamento fixos e aleatórios. Neste mesmo capítulo, apresentamos as distribuições normal e beta truncadas e as distribuições normal e beta aleatoriamente truncadas. No Capítulo 3, descrevemos a construção da classe de modelos não lineares aleatoriamente truncados mistos para locação e escala, assumindo que a variável resposta segue uma distribuição truncada pertencente à classe considerada, que os limites de truncamento são resultantes do processo de amostragem dos dados e que há presença de fontes de variação não observáveis que provocam correlação entre as observações. Na Seção 3.2, apresentamos brevemente três casos particulares da classe de modelos propostos: o modelo não linear truncado misto para locação e escala, que é obtido ao considerarmos que os limites de truncamento são constantes fixas e conhecidas; o modelo de regressão não linear aleatoriamente truncado para locação e escala, que é obtido se assumirmos que não há efeito de fontes de variação não observáveis; e o modelo de regressão não linear truncado que é obtido ao assumirmos que não há efeito de fontes de variação não observáveis e que os limites de truncamento são fixos e conhecidos. No Capítulo 4, é apresentada a abordagem frequentista para os modelos não lineares aleatoriamente truncados mistos aqui propostos. Os estimadores de máxima verossimilhança dos parâmetros do modelo associados à variável resposta truncada são estimados usando-se um procedimento de máxima verossimilhança iterativa e os estimadores de máxima verossimilhança dos parâmetros dos modelos associados às variáveis de truncamento são obtidos através da maximização direta da função de log-verossimilhança. Intervalos de confiança assintóticos e baseados na razão de log-verossimilhança são calculados para os parâmetros. Na Seção 4.2, uma metodologia de análise de diagnóstico, que compreende uma análise de resíduos, de influência global e de influência local, é considerada para os modelos propostos. No Capítulo 5, é apresentada a metodologia de estimação e métricas de diagnóstico Bayesianas consideradas para o modelo proposto. As estimativas Bayesianas são computadas de amostras obtidas via um algoritmo do tipo cadeia de Markov Monte Carlo das distribuições a *posteriori* dos parâmetros. Intervalos de credibilidade inter-quantil e HPD são construídos para os parâmetros do modelo. Para a análise de diagnóstico Bayesiano do modelo, consideramos métricas de avaliação preditiva a *posteriori*, resíduos Bayesianos padronizados e a métrica de calibração de casos. No Capítulo 6, apresenta-

mos os resultados de simulações para verificar a qualidade dos estimadores de máxima verossimilhança dos modelos propostos e para estudar as suas propriedades frequentistas, bem como para ilustrar e avaliar a eficácia das diferentes métricas de diagnóstico de modelos, consideradas quando os modelos propostos são ajustados a um conjunto de dados perturbados. No Capítulo 7, apresentamos resultados baseados em conjuntos de dados simulados para avaliar as propriedades frequentistas das estimativas Bayesianas e para verificar o desempenho das métricas Bayesianas de diagnóstico sob o modelo proposto na presença de perturbações no conjunto de dados. No Capítulo 8, um conjunto de dados reais é analisado, usando a metodologia proposta sob a perspectiva de inferência frequentista e Bayesiana, respectivamente. Finalmente, no Capítulo 9, são apresentadas algumas conclusões e considerações finais.

Capítulo 2

Distribuições truncadas

O truncamento pode ser definido como uma característica da distribuição de probabilidade da qual uma determinada amostra provém (Greene, 2003). Desta forma, para levar em consideração a natureza truncada dos dados, devemos considerar uma distribuição truncada que corresponde à uma densidade condicional obtida ao se restringir o suporte de uma dada função de densidade de probabilidade.

Na Definição 2.1.1 da Seção 2.1, apresentamos a definição formal de variável aleatória (v.a.) truncada em limites fixos e conhecidos. Nas Seções 2.1.1 e 2.1.2 apresentamos as distribuições normal truncada e beta truncada que podem ser consideradas para construir modelos de regressão não lineares truncados. Na Definição 2.2.1 da Seção 2.2, apresentamos uma definição formal de v.a. truncada com limites de truncamento aleatórios. Nas Seções 2.2.1 e 2.2.2 apresentamos as distribuições normal e beta com truncamento aleatório que serão consideradas na construção dos modelos de regressão não lineares aleatoriamente truncados do Capítulo 3.

2.1 Truncamento com limites fixos e conhecidos

Definição 2.1.1. Seja Y uma v.a. com $f(y;\boldsymbol{\omega})$ sua função de densidade de probabilidade (f.d.p.) e $F(y;\boldsymbol{\omega})$ sua função de distribuição acumulada (f.d.a.), sendo que $\boldsymbol{\omega} \in \Omega$ é um vetor de parâmetros indexadores que compreende, entre outros, um parâmetro de locação (ou média), μ, e um parâmetro de escala (ou dispersão), σ. Considere que Y está restrita a um subintervalo fixo e conhecido (a, b), $a < b$, do domínio de Y. Então a v.a. Y condicionada à $(a < Y < b)$, $Y\,|(a < Y < b)$, é uma v.a. truncada com f.d.p. truncada dada por

$$f(y\,|a < y < b\,;\boldsymbol{\omega}) = \begin{cases} \dfrac{f(y;\boldsymbol{\omega})}{F(b;\boldsymbol{\omega}) - F(a;\boldsymbol{\omega})}, \text{ se } a < y < b \\ 0, \text{ caso contrário,} \end{cases} \tag{2.1}$$

e f.d.a. dada por

$$F\left(y\,|a<y<b\,;\boldsymbol{\omega}\right)=\begin{cases}0,\ \text{se } y<a\\ \dfrac{F\left(y;\boldsymbol{\omega}\right)-F\left(a;\boldsymbol{\omega}\right)}{F\left(b;\boldsymbol{\omega}\right)-F\left(a;\boldsymbol{\omega}\right)},\ \text{se } a\leq y<b\\ 1,\ \text{se } y\geq b.\end{cases} \tag{2.2}$$

A esperança e a variância de uma v.a. truncada $Y\,|(a<Y<b)$ são, respectivamente, definidas por

$$E\left[Y\,|(a<Y<b)\right]=\frac{1}{F\left(b;\boldsymbol{\omega}\right)-F\left(a;\boldsymbol{\omega}\right)}\int_{a}^{b}yf\left(y;\boldsymbol{\omega}\right)dy, \tag{2.3}$$

e

$$Var\left[Y\,|(a<Y<b)\right]=\frac{1}{F\left(b;\boldsymbol{\omega}\right)-F\left(a;\boldsymbol{\omega}\right)}\int_{a}^{b}y^{2}f\left(y;\boldsymbol{\omega}\right)dy-\left[E\left(Y\,|a<Y<b\right)\right]^{2}. \tag{2.4}$$

2.1.1 Distribuição normal truncada

Seja Y uma v.a. normal com parâmetro de locação $\mu\in\mathbb{R}$ e parâmetro de escala $\sigma>0$. Se Y é truncada em (a,b), $-\infty<a<b<+\infty$, então $Y\,|(a<Y<b)$ é uma v.a. normal truncada (Johnson *et al.*, 1994) com função de densidade dada por

$$f\left(y\,|a<y<b\right)=\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\left[\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)\right]^{-1}, \tag{2.5}$$

sendo que ϕ é a função de densidade e Φ é a função de distribuição acumulada normal padrão. A função de densidade normal truncada é denotada por $NT\left(\mu,\sigma,a,b\right)$.

A esperança e variância de uma v.a. normal truncada são dadas por

$$E\left[Y\,|(a<Y<b)\right]=\mu+\sigma\left\{\frac{\phi\left(\frac{a-\mu}{\sigma}\right)-\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}\right\}, \tag{2.6}$$

e

$$\begin{aligned}&Var\left[Y\,|(a<Y<b)\right]\\&=\sigma^{2}\left\{1+\frac{\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right)-\left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}-\left[\frac{\phi\left(\frac{a-\mu}{\sigma}\right)-\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}\right]^{2}\right\},\end{aligned} \tag{2.7}$$

respectivamente.

Na Figura 2.1, são mostrados três exemplos de versões truncadas da distribuição

$N(0;2)$.

Na Seção A.1 do Apêndice A, apresentamos o código implementado no software estatístico R para simular valores de uma v.a. com f.d.p. $NT(\mu, \sigma, a, b)$ usando o método de transformação inversa de geração de v.a.'s.

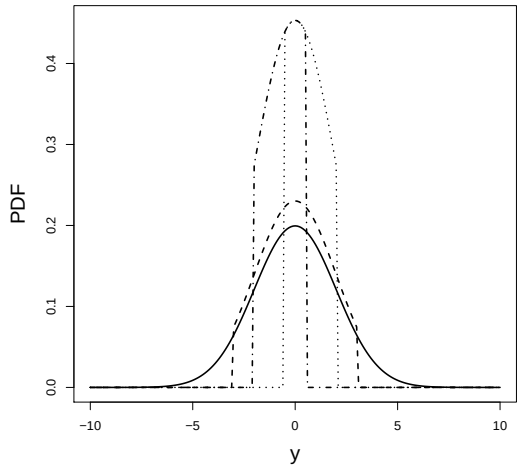


Figura 2.1: Funções de densidade de probabilidade normal truncada: $N(0;2)$ (linha sólida); $NT(0;2;-3;3)$ (linha tracejada); $NT(0;2;-0,5;2)$ (linha pontilhada) e $NT(0;2;-2;0,5)$ (linha tracejada e pontilhada).

2.1.2 Distribuição beta truncada

Consideramos uma versão da distribuição beta truncada construída a partir da expressão para a distribuição beta reparametrizada em termos de um parâmetro de média (locação) e um parâmetro associado à dispersão da v.a.. Assim, suponha que a v.a. Y possui distribuição $Beta(c,d)$, $c,d>0$, com função de densidade dada por

$$f(y) = \frac{1}{B(c,d)} y^{c-1}(1-y)^{d-1},$$

Fazendo $\mu = c/(c+d)$ e $e^{\sigma} = c+d$, obtemos

$$E(Y) = \frac{c}{c+d} = \mu, \tag{2.8}$$

e

$$Var(Y) = \frac{cd}{(c+d)^2(c+d+1)} = \frac{\mu(1-\mu)}{1+e^{\sigma}}. \tag{2.9}$$

Desta forma, μ é o parâmetro de média e σ é o parâmetro associado à dispersão da

v.a. Y. Então, podemos reescrever a função de densidade de Y como

$$f(y) = \frac{y^{\mu e^{\sigma}-1}(1-y)^{(1-\mu)e^{\sigma}-1}}{B(\mu e^{\sigma}, (1-\mu)e^{\sigma})}, \tag{2.10}$$

com $0 < \mu < 1$ e $\sigma \in \mathbb{R}$. Neste trabalho, a distribuição beta como parametrizada em (2.10) é denotada por $Beta(\mu, \sigma)$. Esta reparametrização da distribuição beta tem como inspiração as reparametrizações das distribuições beta-binomial e binomial multiplicativa apresentadas por Lindsey & Altham (1998).

Se Y é truncada em um intervalo (a, b), então a função de densidade de Y dado $(a < Y < b)$, $0 < a < b < 1$, denotada por $BT(\mu, \sigma, a, b)$, é escrita como

$$f(y\,|a < y < b) = \frac{y^{\mu e^{\sigma}-1}(1-y)^{(1-\mu)e^{\sigma}-1}}{B(b; \mu e^{\sigma}, (1-\mu)e^{\sigma}) - B(a; \mu e^{\sigma}, (1-\mu)e^{\sigma})}, \tag{2.11}$$

sendo que $B(\kappa, \tau) = \int_0^1 y^{\kappa-1}(1-y)^{\tau-1}dy$ é a função beta e $B(t; \kappa, \tau) = \int_0^t y^{\kappa-1}(1-y)^{\tau-1}dy$ é a função beta incompleta.

A esperança e variância de uma v.a. beta truncada, sob a forma reparametrizada da distribuição, são dadas por

$$E[Y\,|(a < Y < b)] = \frac{B(a; \mu e^{\sigma}+1, (1-\mu)e^{\sigma}) - B(b; \mu e^{\sigma}+1, (1-\mu)e^{\sigma})}{B(a; \mu e^{\sigma}, (1-\mu)e^{\sigma}) - B(b; \mu e^{\sigma}, (1-\mu)e^{\sigma})}, \tag{2.12}$$

e

$$\begin{aligned} Var[Y\,|(a < Y < b)] &= \frac{B(a; \mu e^{\sigma}+2, (1-\mu)e^{\sigma}) - B(b; \mu e^{\sigma}+2, (1-\mu)e^{\sigma})}{B(a; \mu e^{\sigma}, (1-\mu)e^{\sigma}) - B(b; \mu e^{\sigma}, (1-\mu)e^{\sigma})} \\ &- \left[\frac{B(a; \mu e^{\sigma}+1, (1-\mu)e^{\sigma}) - B(b; \mu e^{\sigma}+1, (1-\mu)e^{\sigma})}{B(a; \mu e^{\sigma}, (1-\mu)e^{\sigma}) - B(b; \mu e^{\sigma}, (1-\mu)e^{\sigma})}\right]^2 \end{aligned} \tag{2.13}$$

respectivamente.

Na Figura 2.2-2.5, são apresentados exemplos de versões truncadas da distribuição beta.

Na Seção A.2 do Apêndice A, apresentamos o código implementado no software estatístico `R` para simular valores de uma v.a. com f.d.p. $BT(\mu, \sigma, a, b)$ usando o método de transformação inversa de geração de v.a.'s.

2.2 Truncamento com limites aleatórios

Em situações práticas, onde a variável de interesse Y está sujeita a limites de truncamento fixos e conhecidos $a < b$, é suficiente especificar apenas a f.d.p. truncada como dada na Definição 2.1.1. Em contrapartida, se a variável de interesse Y está sujeita a limites de truncamento aleatórios, isto é, se os próprios limites de trucamento, A e B, $A < B$,

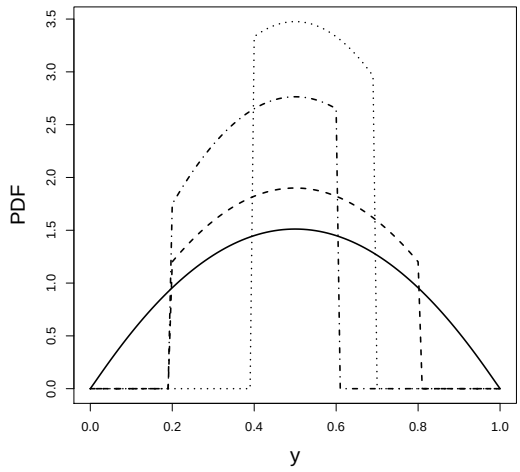


Figura 2.2: Funções de densidade de probabilidade beta truncada: $Beta\,(2;2)$ (linha sólida); $BT\,(2;2;0,2;0,8)$ (linha tracejada); $BT\,(2;2;0,4;0,7)$ (linha pontilhada) e $BT\,(2;2;0,2;0,6)$ (linha tracejada e pontilhada).

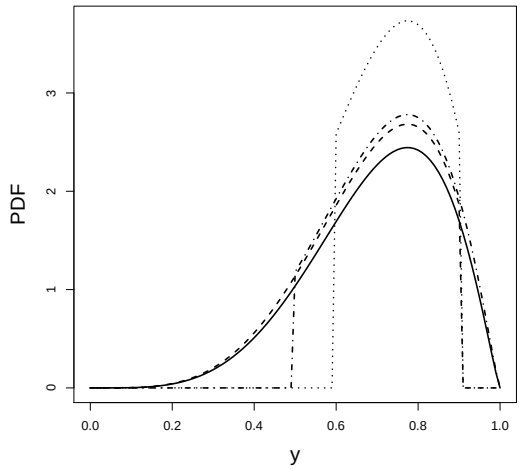


Figura 2.3: Funções de densidade de probabilidade beta truncada: $Beta\,(5,2)$ (linha sólida); $BT\,(5;2;0,1;0,9)$ (linha tracejada); $BT\,(5;2;0,6;0,9)$ (linha pontilhada) e $BT\,(5;2;0,7;1)$ (linha tracejada e pontilhada).

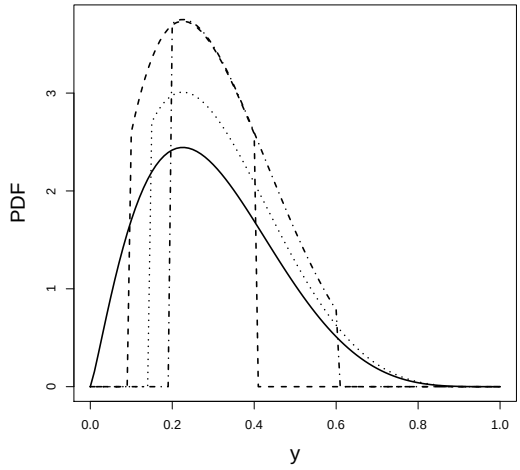


Figura 2.4: Funções de densidade de probabilidade beta truncada: $Beta\,(2;5)$ (linha sólida); $BT\,(2;5;0,1;0,3)$ (linha tracejada); $BT\,(2;5;0,15;0,85)$ (linha pontilhada) e $BT\,(2;5;0,2;0,6)$ (linha tracejada e pontilhada).

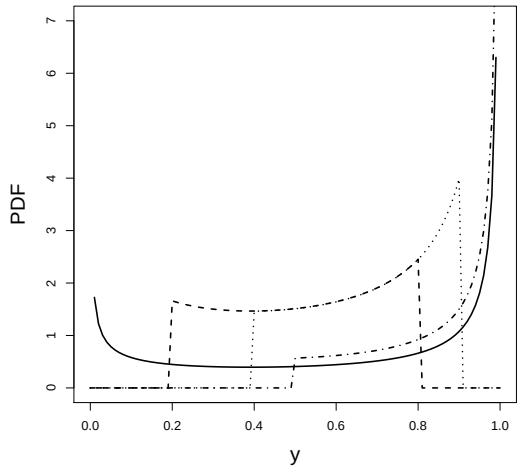


Figura 2.5: Funções de densidade de probabilidade beta truncada: $Beta\,(0,5;0,5)$; $BT\,(0,5;0,5;0,2;0,8)$ (linha tracejada); $BT\,(0,5;0,5;0,4;0,7)$ (linha pontilhada) e $(0,5;0,5;0,1;0,3)$ (linha tracejada e pontilhada).

são variáveis aleatórias, então é necessário especificar a distribuição conjunta de Y, A e B, que corresponde à distribuição do vetor aleatório $(Y, A, B)\,|(A < Y < B)$, que por sua vez, é definida como o produto das distribuições condicionais de $Y\,|(A, B, A < Y < B)$, $A\,|(B, A < Y < B)$ e $B\,|(A < Y < B)$, sendo: $Y\,|(A, B, A < Y < B)$ a variável aleatoriamente truncada condicionada à A, B e $(A < Y < B)$, $A\,|(B, A < Y < B)$ a v.a. de truncamento inferior condicionada à B e $(A < Y < B)$ e $B\,|(A < Y < B)$ a v.a de truncamento superior tal que $(A < Y < B)$.

Definição 2.2.1. Sejam $f(y;\boldsymbol{\omega})$ e $F(y;\boldsymbol{\omega})$ as funções de densidade e distribuição de uma v.a. Y, com $\boldsymbol{\omega} \in \Omega$ o vetor de parâmetros indexadores que compreende, entre outros, um parâmetro de locação e um parâmetro de escala.

Considere $f(a\,|b, a < y < b\,, \boldsymbol{\omega}_A)$ e $F(a\,|b, a < y < b\,, \boldsymbol{\omega}_A)$ a densidade e a distribuição condicional de $A\,|(B, A < Y < B)$, com $\boldsymbol{\omega}_A \in \Omega_A$ o vetor de parâmetros indexadores de f. Considere, ainda, $f(b\,|a < y < b\,, \boldsymbol{\omega}_B)$ e $F(b\,|a < y < b\,, \boldsymbol{\omega}_B)$ a densidade e a distribuição de $B\,|(A < Y < B)$, com $\boldsymbol{\omega}_B \in \Omega_B$ o vetor de parâmetros indexadores de f.

Assuma que a f.d.p. de $Y\,|(A, B, A < Y < B)$ é da forma (2.1). Assim, a função de densidade conjunta de (Y, A, B) dado $(A < Y < B)$, é escrita como

$$\begin{aligned} f(y, a, b\,|a < y < b\,; \boldsymbol{\omega}_T) &= f(y\,|a, b, a < y < b\,; \boldsymbol{\omega})\, f(a\,|b, a < y < b\,; \boldsymbol{\omega}_A)\, f(b\,|a < y < b\,; \boldsymbol{\omega}_B) \\ &= \begin{cases} \dfrac{f(y;\boldsymbol{\omega})\, f(a\,|b, a < y < b\,; \boldsymbol{\omega}_A)\, f(b\,|a < y < b\,; \boldsymbol{\omega}_B)}{F(b;\boldsymbol{\omega}) - F(a;\boldsymbol{\omega})}, \text{ se } a < y < b \\ 0, \text{ caso contrario,} \end{cases} \end{aligned} \tag{2.14}$$

e a f.d.a. conjunta condicional de $(Y, A, B)\,|(A < Y < B)$ é dada por

$$\begin{aligned} F(y, a, b\,|a < y < b\,; \boldsymbol{\omega}_T) &= F(y\,|a, b, a < y < b\,; \boldsymbol{\omega})\, F(a\,|b, a < y < b\,; \boldsymbol{\omega}_A)\, F(b\,|a < y < b\,; \boldsymbol{\omega}_B) \\ &= \begin{cases} 0, \text{ se } y < a \\ \dfrac{F(y;\boldsymbol{\omega}) - F(a;\boldsymbol{\omega})}{F(b;\boldsymbol{\omega}) - F(a;\boldsymbol{\omega})} F(a\,|b, a < y < b\,; \boldsymbol{\omega}_A)\, F(b\,|a < y < b\,; \boldsymbol{\omega}_B), \text{ se } a \leqslant y < b \\ 1, \text{ se } y \geqslant b, \end{cases} \end{aligned} \tag{2.15}$$

com $\boldsymbol{\omega}_T = (\boldsymbol{\omega}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B)'$.

Vale ressaltar que, sob a perspectiva não paramétrica e semiparamétrica, para encontrar os estimadores de $\boldsymbol{\omega}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ é necessário estabelecer condições sob as quais $F(y)$, $F(a)$ e $F(b)$ são identificáveis (Woodroofe, 1985). Para tal, considere $H(z)$ uma função qualquer e sejam $i_H = \inf\{z : H(z) > 0\}$ e $s_H = \inf\{z : H(z) = 1\}$ os limites inferior e superior do suporte de H. Note que a probabilidade de não ocorrer truncamento, definida por $P(A < Y < B)$, será positiva e, portanto, F, F_A e F_B serão identificáveis, quando $i_{F_A} < i_F < i_{F_B}$ e $s_{F_A} < s_F < s_{F_B}$. Algumas propostas de modelagem semi e não paramétrica para dados aleatoriamente truncados podem ser encontradas em Woodroofe (1985), Wang (1989), Moreira & Uña Álvarez (2012) e Shen (2013).

2.2.1 Distribuição normal aleatoriamente truncada

Seja Y uma v.a. normal com parâmetro de locação $\mu \in \mathbb{R}$ e parâmetro de escala $\sigma > 0$. Se a v.a. Y está sujeita a limites de truncamento aleatórios, então é necessário especificar a distribuição conjunta do vetor aleatório $(Y, A, B)\,|(A < Y < B)$ que é definida como o produto das distribuições condicionais de $Y\,|(A, B, A < Y < B)$, $A\,|(B, A < Y < B)$ e $B\,|(A < Y < B)$.

Assuma que $Y\,|(A, B, A < Y < B)$ segue distribuição $NT(\mu, \sigma, a, b)$, e que $A\,|(B, A < Y < B) \sim NT(\mu_A, \sigma_A, -\infty, b)$ e $B\,|(A < Y < B) \sim N(\mu_B, \sigma_B)$, sendo que $N(\mu_B, \sigma_B)$ denota uma distribuição normal com parâmetros de locação μ_B e de escala σ_B. Assim, pela Definição 2.2.1, a distribuição conjunta de $(Y, A, B)\,|(A < Y < B)$ é dada por

$$f(y, a, b\,|a < y < b\,; \boldsymbol{\omega}_T) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right]^{-1}$$
$$\times \frac{1}{\sigma_A}\phi\left(\frac{a-\mu_A}{\sigma_A}\right)\left[\Phi\left(\frac{b-\mu_A}{\sigma_A}\right)\right]^{-1}\frac{1}{\sigma_B}\phi\left(\frac{b-\mu_B}{\sigma_B}\right), \tag{2.16}$$

onde $\boldsymbol{\omega}_T = (\mu, \sigma, \mu_A, \sigma_A, \mu_B, \sigma_B)$.

Um vetor aleatório truncado com distribuição normal aleatoriamente truncada (2.16) é denotado por $(Y, A, B)\,|(A < Y < B) \sim NTA(\mu, \sigma, \mu_A, \sigma_A, \mu_B, \sigma_B)$.

Note que a esperança e variância de $Y\,|(A, B, A < Y < B)$ equivalem a (2.6) e (2.7), respectivamente.

Já a esperança e variância de $A\,|(B, A < Y < B)$ são dadas por

$$E[A\,|(B, A < Y < B)] = \mu_A - \sigma_A\frac{\phi\left(\frac{b-\mu_A}{\sigma_A}\right)}{\Phi\left(\frac{b-\mu_A}{\sigma_A}\right)} \tag{2.17}$$

e

$$Var[A\,|(B, A < Y < B)] = \sigma_A^2\left\{1 - \left(\frac{b-\mu_A}{\sigma_A}\right)\frac{\phi\left(\frac{b-\mu_A}{\sigma_A}\right)}{\Phi\left(\frac{b-\mu_A}{\sigma_A}\right)} - \left[\frac{\phi\left(\frac{b-\mu_A}{\sigma_A}\right)}{\Phi\left(\frac{b-\mu_A}{\sigma_A}\right)}\right]^2\right\}. \tag{2.18}$$

Para $B\,|(A < Y < B)$ temos

$$E[B\,|(A < Y < B)] = \mu_B, \tag{2.19}$$

e

$$Var[B\,|(A < Y < B)] = \sigma_B^2. \tag{2.20}$$

2.2.2 Distribuição beta aleatoriamente truncada

Para construir a distribuição beta aleatoriamente truncada, assumimos que $Y\,|(A,B,A<Y<B)$ segue distribuição $BT\,(\mu,\sigma,a,b)$ e que os limites de truncamento $A\,|(B,A<Y<B)$ e $B\,|(A<Y<B)$ possuem distribuição $BT\,(\mu_A,\sigma_A,0,b)$ e $Beta\,(\mu_B,\sigma_B)$, respectivamente. Assim, pela Definição 2.2.1, a distribuição conjunta de $(Y,A,B)\,|(A<Y<B)$ é dada por

$$f\,(y,a,b\,|a<y<b\,;\boldsymbol{\omega}_T)=\frac{y^{\mu e^{\sigma}}(1-y)^{(1-\mu)e^{\sigma}-1}}{B\,(b;\mu e^{\sigma},(1-\mu)\,e^{\sigma})-B\,(a;\mu e^{\sigma},(1-\mu)\,e^{\sigma})}$$
$$\times\frac{a^{\mu_A e^{\sigma_A}}(1-a)^{(1-\mu_A)e^{\sigma_A}-1}}{B\,(b;\mu_A e^{\sigma_A},(1-\mu_A)\,e^{\sigma_A})}\frac{b^{\mu_B e^{\sigma_B}}(1-b)^{(1-\mu_B)e^{\sigma_B}-1}}{B\,(\mu_B e^{\sigma_B},(1-\mu_B)\,e^{\sigma_B})},\tag{2.21}$$

onde $\boldsymbol{\omega}_T=(\mu,\sigma,\mu_A,\sigma_A,\mu_B,\sigma_B)$.

Um vetor aleatório truncado com distribuição beta aleatoriamente truncada (2.21) é denotado por $(Y,A,B)\,|(A<Y<B)\sim BTA\,(\mu,\sigma,\mu_A,\sigma_A,\mu_B,\sigma_B)$.

A esperança e variância de $Y\,|(A,B,A<Y<B)$ equivalem a (2.12) e (2.13).

Para $A\,|(B,A<Y<B)$ temos

$$E\,[A\,|(B,A<Y<B)]=\frac{B\,(b;\mu_A e^{\sigma_A}+1,(1-\mu_A)\,e^{\sigma_A})}{B\,(b;\mu_A e^{\sigma_A},(1-\mu_A)\,e^{\sigma_A})},\tag{2.22}$$

e

$$Var\,[A\,|(B,A<Y<B)]=\frac{B\,(b;\mu_A e^{\sigma_A}+2,(1-\mu_A)\,e^{\sigma_A})}{B\,(b;\mu_A e^{\sigma_A},(1-\mu_A)\,e^{\sigma_A})}$$
$$-\left[\frac{B\,(b;\mu_A e^{\sigma_A}+1,(1-\mu_A)\,e^{\sigma_A})}{B\,(b;\mu_A e^{\sigma_A},(1-\mu_A)\,e^{\sigma_A})}\right]^2.\tag{2.23}$$

A esperânça e variância de $B\,|(A<Y<B)$ são

$$E\,[B\,|(A<Y<B)]=\mu_B,\tag{2.24}$$

e

$$Var\,[B\,|(A<Y<B)]=\frac{\mu_B\,(1-\mu_B)}{1+e^{\sigma_B}}.\tag{2.25}$$

Capítulo 3

Modelo de regressão não linear aleatoriamente truncado misto para locação e escala

Neste capítulo, desenvolvemos a classe de modelos de regressão não lineares aleatoriamente truncados mistos para locação e escala. Assim como mencionado no Capítulo 1, modelos de regressão não lineares truncados têm sido pouco investigados sob a perspectiva de inferência paramétrica, sendo mais comumente estudados sob as perspectivas semi e não paramétrica. Porém, Bragato (2004) apresenta um caso de modelo de regressão normal inferiormente truncado.

O modelo de regressão não linear aleatoriamente truncado misto para locação e escala, aqui apresentado, é construído assumindo-se limites de truncamento aleatórios e incluindo-se efeitos aleatórios que, por sua vez, possibilitam uma representação estatística apropriada de conjuntos de dados que exibem padrões de variação que podem ser atribuídos a efeitos não observáveis que levam a uma correlação entre as observações.

Vale ressaltar que a classe de modelos mistos (Laird & Ware, 1982; Pinheiro & Bates, 2000) têm como principal objetivo fornecer uma representação estatística da estrutura de regressão para observações feitas em um mesmo indivíduo (unidade experimental), porém, esses modelos também são capazes de fornecer uma representação da estrutura de relação entre múltiplos indivíduos. Esta dupla funcionalidade dos modelos mistos é precisamente alcançada combinando os chamados efeitos fixos das covariáveis - a estrutura que se supõe ser a mesma para todos os indivíduos - e os chamados efeitos aleatórios que são introduzidos no modelo para explicar a variação entre os indivíduos. Como consequência da sua estrutura, os modelos mistos incorporaram naturalmente no coeficiente de regressão a variação existente entre os indivíduos, além de permitirem que a variação na resposta seja particionado, proporcionando, assim, uma forma de compreender as diferentes fontes de variação. Em Matos *et al.* (2013b), são apresentados modelos lineares e não lineares

mistos baseados na distribuição t-Student multivariada. Os modelos são tratados sob a perspectiva de inferência frequentista e são aplicados a conjuntos de dados com presença de censura. Em Lachos *et al.* (2013), os autores propuseram uma abordagem Bayesiana para um modelo não linear misto considerando a distribuição normal/independente.

Como já foi dito no capítulo introdutório, na área de inferência semiparamétrica e não paramétrica é possível encontrar diversos exemplos de propostas de estimação de modelos de regressão truncados com limites aleatórios (ver Powell (1986), Lagakos *et al.* (1988), Kalbfleisch & Lawless (1989), Wang (1989), Lee (1992), Lee (1993), Newey (2004), Cosslett (2004) e Chen & Zhou (2012)).

3.1 Formulação do modelo

Os modelos de regressão para tratar dados sujeitos a truncamento aleatório aqui propostos, baseiam-se na família de distribuições conjuntas condicionais truncadas apresentada na Definição 2.2.1. Consideramos que os dados são medidos em $i = 1, \ldots, M$ grupos, com $j = 1, \ldots, N_i$ replicatas e $k = 1, \ldots, n_{ij}$ observações cada. Assumimos que a fonte de variação não observável está relacionada ao efeito do grupo ao qual a observação pertence. Assim, o modelo de regressão não linear aleatoriamente truncado misto para locação e escala é construído como segue: dado o vetor aleatório (Y, A, B), em que Y é a variável resposta de interesse e A e B são as v.a.'s de truncamento inferior e superior e $\boldsymbol{x} = (x_1, \ldots, x_p)'$ é um vetor de p covariáveis, quatro suposições são assumidas:

1. $(Y, A, B)\,|(A < Y < B)$ possui uma f.d.p. truncada como dada em (2.14);

2. o parâmetro de locação da variável resposta truncada $Y\,|(A, B, A < Y < B)$ é representado por $\eta\,(\boldsymbol{x}_q, \boldsymbol{\beta})$, uma função não linear contínua e duas vezes diferenciável com relação à $\boldsymbol{\beta}$, sendo $\boldsymbol{x}_q$ um subconjunto do vetor de covariáveis $\boldsymbol{x}$;

3. há presença de uma fonte de variação não observável que provoca correlação entre as observações e o efeito de tal variação é introduzido linearmente no modelo através de efeitos aleatórios no parâmetro de locação da variável resposta;

4. o parâmetro de escala da variável resposta truncada $Y\,|(A, B, A < Y < B)$ é representado por $g\,(\boldsymbol{x}_r, \boldsymbol{\alpha})$, uma função contínua e duas vezes diferenciável com relação a $\boldsymbol{\alpha}$, sendo $\boldsymbol{x}_r$ um subconjunto do vetor de covariáveis $\boldsymbol{x}$.

Seja $(Y_{i,jk}, A_{i,jk}, B_{i,jk})\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk})$ o vetor aleatório truncado associado a k-ésima observação da j-ésima replicata do i-ésimo grupo, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, sendo M o número de grupos, N_i o número de replicatas e n_{ij} o número de observações na j-ésima replicata do i-ésimo grupo. Note que $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})$ denota a variável resposta truncada e $A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})$ e

$B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk})$ denotam as v.a.'s de truncamento inferior e superior, respectivamente, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$. O uso da notação i, jk é adotado para ressaltar que o efeito aleatório é introduzido nos $i = 1, \ldots, M$ níveis do grupo.

Supondo uma possível estrutura de dependência entre as observações dentro do mesmo grupo, considere $\boldsymbol{u}_i$ um vetor s-dimensional de efeitos individuais desconhecidos e $\boldsymbol{U}_{i,j}$ uma matriz de delineamento $n_{ij} \times s$ associada aos $\boldsymbol{u}_i$. Seja $\boldsymbol{u} = \left(\boldsymbol{u}_1^{'}, \ldots, \boldsymbol{u}_M^{'}\right)^{'}$, e $\boldsymbol{x} = \left(\boldsymbol{x}_{1,1}^{'}, \ldots, \boldsymbol{x}_{M,N_M}^{'}\right)^{'}$ um conjunto de p covariáveis.

Assuma $\boldsymbol{Y}\,|(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B}, \boldsymbol{u})$ um vetor de $n = \sum_{i=1}^{M}\sum_{j=1}^{N_i} n_{i,j}$ v.a.'s independentes cujos elementos são os vetores $\boldsymbol{Y}_{i,j}\,|(\boldsymbol{A}_{i,j}, \boldsymbol{B}_{i,j}, \boldsymbol{A}_{i,j} < \boldsymbol{Y}_{i,j} < \boldsymbol{B}_{i,j}, \boldsymbol{u}_i)$, compostos por $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i)$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$.

Analogamente, assuma $\boldsymbol{A}\,|(\boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$ um vetor de $n = \sum_{i=1}^{M}\sum_{j=1}^{N_i} n_{i,j}$ v.a.'s independentes cujos elementos são os vetores $\boldsymbol{A}_{i,j}\,|(\boldsymbol{B}_{i,j}, \boldsymbol{A}_{i,j} < \boldsymbol{Y}_{i,j} < \boldsymbol{B}_{i,j})$, compostos por $A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})$, e $\boldsymbol{B}\,|(\boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$ um vetor de $n = \sum_{i=1}^{M}\sum_{j=1}^{N_i} n_{i,j}$ v.a.'s independentes cujos elementos são os vetores $\boldsymbol{B}_{i,j}\,|(\boldsymbol{A}_{i,j} < \boldsymbol{Y}_{i,j} < \boldsymbol{B}_{i,j})$, compostos por $B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk})$, $i = 1, \ldots, M$ e $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$.

Considere que cada vetor aleatório truncado $(Y_{i,jk}, A_{i,jk}, B_{i,jk})\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i)$ possui uma f.d.p. truncada (2.14) pertencente à família descrita na Definição 2.2.1. Suponha que para cada $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i.jk}, \boldsymbol{u}_i)$ o parâmetro de locação $\mu_{i,jk}$ é associado às covariáveis e aos efeitos aleatórios pela relação $\mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i$, onde $\eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right)$ é uma função não linear do vetor de parâmetros desconhecidos q-dimensional $\boldsymbol{\beta}$. Assuma que o parâmetro de escala $\sigma_{i,jk}$ de $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i)$ é associado a um conjunto de covariáveis através da função $g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)$, com $\boldsymbol{\alpha}$ um vetor de parâmetros desconhecidos r-dimensional. $\boldsymbol{x}_{q_{i,jk}}$ e $\boldsymbol{x}_{r_{i,jk}}$ denotam subconjuntos do vetor de p covariáveis $\boldsymbol{x}$.

Assuma que $A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})$ são independentes e identicamente distribuídos com f.d.p. condicional $f\left(a_{i,jk}\,|b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}\,;\boldsymbol{\omega}_A\right)$, $i = 1, \ldots, M$ e $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$. Analogamente, assuma que $B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk})$ são independentes e identicamente distribuídos com f.d.p. $f\left(b_{i,ji}\,|a_{i,jk} < y_{i,jk} < b_{i,jk}\,;\boldsymbol{\omega}_B\right)$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$.

O modelo de regressão não linear aleatoriamente truncado para locação e escala é escrito como

$$
\begin{gathered}
Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i) \sim f\left(y_{i,jk}\,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}, \boldsymbol{u}_i\,;\mu_{i,jk}, \sigma_{i,jk}\right), \\
\mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, \qquad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right), \qquad \boldsymbol{u}_i \sim N_s\left(\boldsymbol{0}, \boldsymbol{\Psi}\right), \\
A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}) \sim f\left(a_{i,jk}\,|b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}\,;\boldsymbol{\omega}_A\right), \\
B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}) \sim f\left(b_{i,jk}\,|a_{i,jk} < y_{i,jk} < b_{i,jk}\,;\boldsymbol{\omega}_B\right), \\
i = 1, \ldots, M, j = 1, \ldots, N_i, k = 1, \ldots, n_{ij},
\end{gathered}
\tag{3.1}
$$

sendo que N_s denota a distribuição normal s-dimensional e $\boldsymbol{\Psi}$ a matriz de covariâncias $s \times s$ dos efeitos aleatórios.

Note que o modelo de regressão não linear aleatoriamente truncado para locação e escala poderia similarmente ser escrito da forma

$$\begin{gathered}(Y_{i,jk}, A_{i,jk}, B_{i,jk}) \,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i) \sim f\left(y_{i,jk}, a_{i,jk}, b_{i,jk} \,|a_{i,jk} < y_{i,jk} < b_{i,jk}, \boldsymbol{u}_i \,; \boldsymbol{\omega}_T\right), \\ \mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, \qquad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right), \qquad \boldsymbol{u}_i \sim N_s\left(\boldsymbol{0}, \boldsymbol{\Psi}\right), \\ i = 1, \ldots, M, j = 1, \ldots, N_i, k = 1, \ldots, n_{ij},\end{gathered} \tag{3.2}$$

com $\boldsymbol{\omega}_T = \left(\mu_{i,jk}, \sigma_{i,jk}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B\right)^{'}$.

Considere $\boldsymbol{y} = \left(\boldsymbol{y}_{1,1}^{'}, \ldots, \boldsymbol{y}_{M,N_M}^{'}\right)^{'}$, $\boldsymbol{a} = \left(\boldsymbol{a}_{1,1}^{'}, \ldots, \boldsymbol{a}_{M,N_M}^{'}\right)^{'}$ e $\boldsymbol{b} = \left(\boldsymbol{b}_{1,1}^{'}, \ldots, \boldsymbol{b}_{M,N_M}^{'}\right)^{'}$ vetores de valores observados de $\boldsymbol{Y}\,|(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B}, \boldsymbol{u})$, $\boldsymbol{A}\,|(\boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$ e $\boldsymbol{B}\,|(\boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$, respectivamente.

Sob o modelo de regressão não linear aleatoriamente truncado misto (3.1), a função de log-verossimilhança dos vetores de efeitos fixos $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}\right)^{'}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$, baseada na distribuição marginal dos dados completos $D^* = (n, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{u})$, é escrita como

$$\begin{gathered}\ell\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B \,|D^*\right) = \\ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left\{f\left(y_{i,jk}, a_{i,jk}, b_{i,jk} \,|a_{i,jk} < y_{i,jk} < b_{i,jk} \,; \boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B\right)\right\},\end{gathered} \tag{3.3}$$

sendo que

$$\begin{aligned}&f\left(y_{i,jk}, a_{i,jk}, b_{i,jk} \,|a_{i,jk} < y_{i,jk} < b_{i,jk} \,; \boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B\right) \\ &= \int f\left(y_{i,jk} \,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}, \boldsymbol{u}_i \,; \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)\right) \\ &\times f\left(a_{i,jk} \,|b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\omega}_A\right) f\left(b_{i,jk} \,|a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\omega}_B\right) \\ &\times f\left(\boldsymbol{u}_i; \boldsymbol{\Psi}\right) d\boldsymbol{u}_i.\end{aligned} \tag{3.4}$$

Assim como ocorre com os modelo não lineares mistos (Lindstrom & Bates, 1990), a integral em (3.4) sob o modelo não linear aleatoriamente truncado misto (3.1) não possui solução analítica e, portanto, devemos considerar algoritmos de otimização iterativos que avaliam a integral numericamente. Outra alternativa poderia ser o uso de procedimentos de otimização baseados em aproximações analíticas para (3.3).

Seguindo o procedimento alternativo proposto por Henderson (1950) e já adotado por vários outros autores (ver Henderson *et al.* (1959), Henderson (1975), Robinson (1991) e Lim *et al.* (2013)) podemos escrever a função de log-verossimilhança do modelo proposto a partir da f.d.p. conjunta do vetor aleatório truncado e dos efeitos aleatórios e assim estimar os parâmetros fixos e predizer os efeitos aleatórios simultaneamente. Isto é, a função de log-verossimilhança do modelo (3.1) pode ser escrita para os vetores de efeitos

fixos $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Psi}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ e os efeitos aleatórios não observados $\boldsymbol{u}$, dado os dados observados $D = (n, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})$, como

$$
\begin{aligned}
&\ell\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B \,|D\right) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log f\left(y_{i,jk}\,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}, \boldsymbol{u}_i\,; \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[C^A_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right)\right] + \sum_{i=1}^{M} \log f\left(\boldsymbol{u}_i; \boldsymbol{\Psi}\right) \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log f\left(a_{i,jk}\,|b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\omega}_A\right) \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log f\left(b_{i,jk}\,|a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\omega}_B\right) \\
&= \ell_1\left(\boldsymbol{\theta}\,|D\right) + \ell_2\left(\boldsymbol{\omega}_A\,|D\right) + \ell_3\left(\boldsymbol{\omega}_B\,|D\right), \qquad (3.5)
\end{aligned}
$$

em que

$$
\begin{aligned}
&C^A_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right) \\
&= F\left(b_{i,jk}; \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)\right) - F\left(a_{i,jk}; \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)\right), \qquad (3.6)
\end{aligned}
$$

e $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u}\right)'$.

Pela função (3.5), é possível observar que os vetores de parâmetros $\boldsymbol{\theta}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ são ortogonais, uma vez que $\ell_1\left(\boldsymbol{\theta}\,|D\right)$ independe de $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$, $\ell_2\left(\boldsymbol{\omega}_A\,|D\right)$ independe de $\boldsymbol{\theta}$ e $\boldsymbol{\omega}_B$ e $\ell_3\left(\boldsymbol{\omega}_B\,|D\right)$ independe de $\boldsymbol{\theta}$ e $\boldsymbol{\omega}_A$. Além disso, vale a pena ressaltar que $\ell_1\left(\boldsymbol{\theta}\,|D\right)$ corresponde à função de log-verossimilhança do modelo estatístico assumido para a variável resposta truncada, $\ell_2\left(\boldsymbol{\omega}_A\,|D\right)$ corresponde à função de log-verossimilhança do modelo estatístico assumido para $\boldsymbol{A}\,|(\boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$ e $\ell_3\left(\boldsymbol{\omega}_B\,|D\right)$ corresponde à função de log-verossimilhança do modelo estatístico assumido para $\boldsymbol{B}\,|(\boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$.

Para a classe de modelos de regressão não lineares aleatoriamente truncados mistos com função de log-verossimilhança (3.5), o vetor escore é dado por

$$
\begin{aligned}
U\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B\right) &= \left(\frac{\partial}{\partial\boldsymbol{\theta}}\ell_1\left(\boldsymbol{\theta}\,|D\right), \frac{\partial}{\partial\boldsymbol{\omega}_A}\ell_2\left(\boldsymbol{\omega}_A\,|D\right), \frac{\partial}{\partial\boldsymbol{\omega}_B}\ell_3\left(\boldsymbol{\omega}_B\,|D\right)\right) \\
&= U\left(U\left(\boldsymbol{\theta}\right), U\left(\boldsymbol{\omega}_A\right), U\left(\boldsymbol{\omega}_B\right)\right). \qquad (3.7)
\end{aligned}
$$

Devido à ortogonalidade dos vetores $\boldsymbol{\theta}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$, a matriz Hessiana do modelo de regressão não linear aleatoriamente truncado misto é dada pela seguinte matriz bloco

diagonal

$$H\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B\right) = diag\left[\frac{\partial^2 \ell_1\left(\boldsymbol{\theta}\,|D\right)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}'}, \frac{\partial^2 \ell_2\left(\boldsymbol{\omega}_A\,|D\right)}{\partial\boldsymbol{\omega}_A\boldsymbol{\omega}'_A}, \frac{\partial^2 \ell_3\left(\boldsymbol{\omega}_B\,|D\right)}{\partial\boldsymbol{\omega}_B\boldsymbol{\omega}'_B}\right]$$
$$= diag\left[H\left(\boldsymbol{\theta}\right), H\left(\boldsymbol{\omega}_A\right), H\left(\boldsymbol{\omega}_B\right)\right]. \tag{3.8}$$

Neste trabalho, o vetor escore (3.7) e a matriz Hessiana (3.8) do modelo não linear aleatoriamente truncado misto são obtidos usando-se o método de derivação numérica de Richardson (Fornberg & Sloan, 1994). O vetor escore é computado usando a função *grad* e a matriz Hessiana é obtida pela função *hessian*, ambas implementadas no pacote *numDeriv* (Gilbert & Varadhan, 2012) do software `R`.

3.2 Casos particulares

Se considerarmos que os limites de truncamento são constantes fixas e conhecidas, obtemos o modelo não linear truncado misto para locação e escala, que é dado por

$$\begin{gathered} Y_{i,jk}\,|(a < Y_{i,jk} < b, \boldsymbol{u}_i) \sim f\left(y_{i,jk}\,|a < y_{i,jk} < b, \boldsymbol{u}_i\,; \mu_{i,jk}, \sigma_{i,jk}\right), \\ \mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, \qquad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right), \qquad \boldsymbol{u}_i \sim N_s\left(\boldsymbol{0}, \boldsymbol{\Psi}\right), \\ i = 1, \ldots, M, j = 1, \ldots, N_i, k = 1, \ldots, n_{ij}. \end{gathered} \tag{3.9}$$

Ao assumir que não há efeito de fontes de variação não observáveis, o modelo de regressão não linear aleatoriamente truncado para locação e escala é obtido como um caso particular do modelo não linear aleatoriamente truncado misto para locação e escala (3.1). Este modelo é escrito como

$$\begin{gathered} Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}) \sim f\left(y_{i,jk}\,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}\,; \mu_{i,jk}, \sigma_{i,jk}\right), \\ \mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right), \qquad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right), \\ A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}) \sim f\left(a_{i,jk}\,|b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}\,; \boldsymbol{\omega}_A\right), \\ B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}) \sim f\left(b_{i,jk}\,|a_{i,jk} < y_{i,jk} < b_{i,jk}\,; \boldsymbol{\omega}_B\right), \\ i = 1, \ldots, M, j = 1, \ldots, N_i, k = 1, \ldots, n_{ij}. \end{gathered} \tag{3.10}$$

Finalmente, um modelo de regressão não linear truncado para locação e escala é obtido como um caso particular do modelo não linear aleatoriamente truncado misto para locação e escala (3.1), ao assumir que não há efeito de fontes de variação não observáveis e que os

limites de truncamento são fixos e conhecidos. Tal modelo é escrito como

$$\begin{gathered} Y_{i,jk}\,|(a < Y_{i,jk} < b) \sim f\left(y_{i,jk}\,|a < y_{i,jk} < b\,;\mu_{i,jk},\sigma_{i,jk}\right), \\ \mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right), \qquad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right), \\ i = 1,\ldots,M, j = 1,\ldots,N_i, k = 1,\ldots,n_{ij}. \end{gathered} \tag{3.11}$$

Os modelos de regressão não lineares aleatoriamente truncados mistos (3.1), com função de log-verossimilhança dada em (3.5), podem ser prontamente singularizados para situações em que a variável resposta é limitada apenas inferior ou superiormente. Nestes casos, devemos assumir uma distribuição para a v.a. de truncamento pertinente (A, se os dados forem truncados inferiormente e B se os dados forem truncados superiormente) e assumir que o outro limite equivale à $+\infty$ ou $-\infty$ (conforme a conveniência). O mesmo vale para os modelo não lineares aleatoriamente truncados (3.10).

Analogamente, o modelo de regressão não linear truncado misto (3.9), pode ser facilmente estendido para os casos nos quais cada observação $y_{i,jk}$ possui limites de truncamento específicos $(a_{i,jk}, b_{i,jk})$. Para isto, basta que para cada observação $y_{i,jk}$, os valores de a e b sejam substituídos por $a_{i,jk}$ e $b_{i,jk}$, respectivamente. O modelo também pode ser singularizado para casos em que ocorrem apenas truncamento inferior ou superior. Para obter o caso particular de truncamento unicamente inferior, basta tomar b como $+\infty$ e, analogamente, para obter o caso de truncamento superior, devemos tomar a como $-\infty$. O mesmo vale para os modelo não lineares truncados (3.11).

O tratamento de modelos de regressão com presença de truncamento aleatório sob a perspectiva de inferência paramétrica é uma área com poucas (ou nenhuma) contribuições disponíveis na literatura atual. Assim, nas Seções 3.3 e 3.4, desenvolvemos os modelos de regressão não linear normal aleatoriamente truncado misto e beta aleatoriamente truncado misto, respectivamente.

3.3 Modelo de regressão não linear normal aleatoriamente truncado misto

Considere $\boldsymbol{u}_i$ um vetor s-dimensional de efeitos individuais desconhecidos e $\boldsymbol{U}_{i,j}$ uma matriz de delineamento $n_{ij} \times s$ associada aos $\boldsymbol{u}_i$. Seja $\boldsymbol{u} = \left(\boldsymbol{u}_1^{'},\ldots,\boldsymbol{u}_M^{'}\right)^{'}$, e $\boldsymbol{x} = \left(\boldsymbol{x}_{1,1}^{'},\ldots,\boldsymbol{x}_{M,N_M}^{'}\right)^{'}$ um conjunto de p covariáveis.

Assuma que os $A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})$ são independentes e identicamente distribuídos com f.d.p. condicional $f\left(a_{i,jk}\,|b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}\,;\boldsymbol{\omega}_A\right)$, $i = 1,\ldots,M$, $j = 1,\ldots,N_i$, $k = 1,\ldots,n_{ij}$. Analogamente, assuma que $B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk})$ são independentes e identicamente distribuídos com f.d.p. $f\left(b_{i,ji}\,|a_{i,jk} < y_{i,jk} < b_{i,jk}\,;\boldsymbol{\omega}_B\right)$, $i = 1,\ldots,M$, $j = 1,\ldots,N_i$, $k = 1,\ldots,n_{ij}$.

Considere que cada vetor aleatório truncado $(Y_{i,jk}, A_{i,jk}, B_{i,jk})\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i)$

possui uma f.d.p. normal aleatoriamente truncada (2.16). Suponha que para cada $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i.jk}, \boldsymbol{u}_i)$ o parâmetro de locação $\mu_{i,jk}$ é associado às covariáveis e aos efeitos aleatórios pela relação $\mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i$, onde $\eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right)$ é uma função não linear do vetor de parâmetros desconhecidos q-dimensional $\boldsymbol{\beta}$. Suponha, ainda, que o parâmetro de escala $\sigma_{i,jk}$ de $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i)$ é associado a um conjunto de covariáveis através da função $g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)$, com $\boldsymbol{\alpha}$ um vetor de parâmetros desconhecidos r-dimensional. $\boldsymbol{x}_{q_{i,jk}}$ e $\boldsymbol{x}_{r_{i,jk}}$ denotam subconjuntos do vetor de p covariáveis $\boldsymbol{x}$.

Sejam $\boldsymbol{y} = \left(\boldsymbol{y}'_{1,1}, \ldots, \boldsymbol{y}'_{M,N_M}\right)'$, $\boldsymbol{a} = \left(\boldsymbol{a}'_{1,1}, \ldots, \boldsymbol{a}'_{M,N_M}\right)'$ e $\boldsymbol{b} = \left(\boldsymbol{b}'_{1,1}, \ldots, \boldsymbol{b}'_{M,N_M}\right)'$ vetores de valores observados de $\boldsymbol{Y}\,|(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B}, \boldsymbol{u})$, $\boldsymbol{A}\,|(\boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$ e $\boldsymbol{B}\,|(\boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$, respectivamente. Então, o modelo não linear normal aleatoriamente truncado misto é dado por

$$\begin{gathered}
(Y_{i,jk}, A_{i,jk}, B_{i,jk})\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i) \sim NTA\left(\mu_{i,jk}, \sigma_{i,jk}, \mu_A, \sigma_A, \mu_B, \sigma_B\right), \\
\mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, \qquad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right), \qquad \boldsymbol{u}_i \sim N_s\left(\boldsymbol{0}, \boldsymbol{\Psi}\right) \qquad (3.12) \\
i = 1, \ldots, M, j = 1, \ldots, N_i, k = 1, \ldots, n_{ij}.
\end{gathered}$$

A função de log-verossimilhança do modelo (3.12) para os vetores de efeitos fixos $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Psi}$, μ_A, σ_A, μ_B e σ_B e para os efeitos aleatórios não observados $\boldsymbol{u}$, dado os dados observados $D = (n, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})$, é escrita como

$$\begin{aligned}
&\ell\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B\,|D\right) \\
&\propto -\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)\right] - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\frac{y_{i,jk} - \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i}{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)}\right]^2 \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[C^A_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right)\right] - M\log\left[|\boldsymbol{\Psi}|\right] - \frac{1}{2}\sum_{i=1}^{M} \boldsymbol{u}'_i\boldsymbol{\Psi}^{-1}\boldsymbol{u}_i \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[\sigma_A\right] - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\frac{a_{i,jk} - \mu_A}{\sigma_A}\right]^2 - \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[\Phi\left(\frac{b_{i,jk} - \mu_A}{\sigma_A}\right)\right] \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[\sigma_B\right] - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\frac{b_{i,jk} - \mu_B}{\sigma_B}\right]^2 \\
&= \ell_1\left(\boldsymbol{\theta}\,|D\right) + \ell_2\left(\boldsymbol{\omega}_A\,|D\right) + \ell_3\left(\boldsymbol{\omega}_B\,|D\right), \qquad (3.13)
\end{aligned}$$

em que

$$C^A_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right) = \Phi\left(\frac{b_{i,jk} - \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i}{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)}\right) - \Phi\left(\frac{a - \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i}{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)}\right),$$

e $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u}\right)'$, $\boldsymbol{\omega}_A = (\mu_A, \sigma_A)$ e $\boldsymbol{\omega}_B = (\mu_B, \sigma_B)$.

O vetor escore (3.7) e a matriz Hessiana (3.8) do modelo não linear normal aleatoriamente truncado misto são obtidos usando-se o método de derivação numérica de Richardson.

Modelo de regressão não linear normal truncado misto

Se os limites de truncamento são constantes fixas e conhecidas, tem-se o modelo não linear normal truncado misto, definido como

$$\begin{gathered} Y_{i,kj}\,|(a<Y_{i,jk}<b,\boldsymbol{u}_i)\sim NT\left(\mu_{i,jk},\sigma_{i,jk},a,b\right), \\ \mu_{i,jk}=\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i, \qquad \sigma_{i,jk}=g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right), \qquad \boldsymbol{u}_i\sim N_s\left(\boldsymbol{0},\boldsymbol{\Psi}\right), \\ i=1,\ldots,M, j=1,\ldots,N_i, k=1,\ldots,n_{ij}. \end{gathered} \tag{3.14}$$

A função de log-verossimilhança do modelo (3.14) para os vetores de efeitos fixos $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ e $\boldsymbol{\Psi}$ e para os efeitos aleatórios não observados $\boldsymbol{u}$, dado os dados observados $D=(n,\boldsymbol{y},\boldsymbol{x},a,b)$, é escrita como

$$\begin{aligned} &\ell\left(\boldsymbol{\theta}\,|D\right) \\ &\propto -\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\log\left[g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)\right]-\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\left[\frac{y_{i,jk}-\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i}{g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)}\right]^2 \\ &-\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\log\left[C_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)\right]-M\log\left[|\boldsymbol{\Psi}|\right]-\frac{1}{2}\sum_{i=1}^{M}\boldsymbol{u}_i^{'}\boldsymbol{\Psi}^{-1}\boldsymbol{u}_i, \end{aligned} \tag{3.15}$$

em que

$$C_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)=\Phi\left(\frac{b-\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i}{g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)}\right)-\Phi\left(\frac{a-\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i}{g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)}\right),$$

e $\boldsymbol{\theta}=\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Psi},\boldsymbol{u}\right)^{'}$.

Modelo de regressão não linear normal aleatoriamente truncado

O modelo de regressão não linear normal aleatoriamente truncado é obtido como um caso particular do modelo (3.12) ao assumirmos $\boldsymbol{u}_i=\boldsymbol{0}$ para todo $i=1,\ldots,M$,

$$\begin{gathered} \left(Y_{i,jk},A_{jk},B_{i,jk}\right)|(A_{i,jk}<Y_{i,jk}<B_{i,jk})\sim RTN\left(\mu_{i,jk},\sigma_{i,jk},\mu_A,\sigma_A,\mu_B,\sigma_B\right), \\ \mu_{i,jk}=\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right), \quad \sigma_{i,jk}=g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right), \\ i=1,\ldots,M, \quad j=1,\ldots,N_i, \quad k=1,\ldots,n_{ij}, \end{gathered} \tag{3.16}$$

e sua função de log-verossimilhança, dado o conjunto de dados $D = (n, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})$, é escrita como

$$
\begin{aligned}
&\ell\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B \,|D\right) \\
&\propto -\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)\right] - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\frac{y_{i,jk} - \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)}\right]^2 \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[C^A_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right] \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[\sigma_A\right] - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\frac{a_{i,jk} - \mu_A}{\sigma_A}\right]^2 - \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[\Phi\left(\frac{b_{i,jk} - \mu_A}{\sigma_A}\right)\right] \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[\sigma_B\right] - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\frac{b_{i,jk} - \mu_B}{\sigma_B}\right]^2 \\
&= \ell_1\left(\boldsymbol{\theta}\,|D\right) + \ell_2\left(\boldsymbol{\omega}_A\,|D\right) + \ell_3\left(\boldsymbol{\omega}_B\,|D\right), \qquad (3.17)
\end{aligned}
$$

em que

$$
C^A_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \Phi\left(\frac{b_{i,jk} - \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)}\right) - \Phi\left(\frac{a_{i,jk} - \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)}\right),
$$

e com $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)^{'}$, $\boldsymbol{\omega}_A = (\mu_A, \sigma_A)$ e $\boldsymbol{\omega}_B = (\mu_B, \sigma_B)$ os vetores de parâmetros do modelo (3.16).

Modelo de regressão não linear normal truncado

Note que se tomarmos $\boldsymbol{u}_i = \boldsymbol{0}$, $a_{i,jk} = a$ e $b_{i,jk} = b$ para todo $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, obtemos como caso particular do modelo de regressão não linear normal truncado misto (3.14), o modelo de regressão não linear normal truncado, que pode ser escrito como

$$
\begin{aligned}
Y_{i,jk}\,|(a < Y_{i,jk} < b) &\sim NT\left(\mu_{i,jk}, \sigma_{i,jk}, a, b\right), \\
\mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right), &\qquad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right), \qquad (3.18) \\
i = 1, \ldots, M, j = 1, \ldots, N_i, k = 1, \ldots, n_{ij}&,
\end{aligned}
$$

e sua função de log-verossimilhança, dado o conjunto de dados $D = (n, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})$, é escrita como

$$\begin{aligned}
&\ell\left(\boldsymbol{\theta}\,|D\right) \\
&\propto -\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[g\left(\boldsymbol{x}_{ri,jk}, \boldsymbol{\alpha}\right)\right] - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\frac{y_{i,jk} - \eta\left(\boldsymbol{x}_{qi,jk}, \boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{ri,jk}, \boldsymbol{\alpha}\right)}\right]^2 \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[\Phi\left(\frac{b - \eta\left(\boldsymbol{x}_{qi,jk}, \boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{ri,jk}, \boldsymbol{\alpha}\right)}\right) - \Phi\left(\frac{a - \eta\left(\boldsymbol{x}_{qi,jk}, \boldsymbol{\beta}\right)}{g\left(\boldsymbol{x}_{ri,jk}, \boldsymbol{\alpha}\right)}\right)\right], \qquad (3.19)
\end{aligned}$$

com $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)^{'}$.

3.4 Modelo de regressão não linear beta aleatoriamente truncado misto

Considere $\boldsymbol{u}_i$ um vetor s-dimensional de efeitos individuais desconhecidos e $\boldsymbol{U}_{i,j}$ uma matriz de delineamento $n_{ij} \times s$ associada aos $\boldsymbol{u}_i$. Seja $\boldsymbol{u} = \left(\boldsymbol{u}_1^{'}, \ldots, \boldsymbol{u}_M^{'}\right)^{'}$, e $\boldsymbol{x} = \left(\boldsymbol{x}_{1,1}^{'}, \ldots, \boldsymbol{x}_{M,N_M}^{'}\right)^{'}$ um conjunto de p covariáveis.

Assuma que os $A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})$ são independentes e identicamente distribuídos com f.d.p. condicional $f\left(a_{i,jk}\,|b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}\,; \boldsymbol{\omega}_A\right)$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$. Analogamente, assuma que $B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk})$ são independentes e identicamente distribuídos com f.d.p. $f\left(b_{i,ji}\,|a_{i,jk} < y_{i,jk} < b_{i,jk}\,; \boldsymbol{\omega}_B\right)$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$.

Considere que cada vetor aleatório truncado $(Y_{i,jk}, A_{i,jk}, B_{i,jk})\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i)$ possui uma f.d.p. beta aleatoriamente truncada (2.21). Suponha que para cada $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i.jk}, \boldsymbol{u}_i)$ o parâmetro de locação $\mu_{i,jk}$ é associado às covariáveis e aos efeitos aleatórios pela relação $\mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i$, onde $\eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right)$ é uma função não linear do vetor de parâmetros desconhecidos q-dimensional $\boldsymbol{\beta}$. Suponha, ainda, que o parâmetro de escala $\sigma_{i,jk}$ de $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i)$ é associado a um conjunto de covariáveis através da função $g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)$, com $\boldsymbol{\alpha}$ um vetor de parâmetros desconhecidos r-dimensional. $\boldsymbol{x}_{q_{i,jk}}$ e $\boldsymbol{x}_{r_{i,jk}}$ denotam subconjuntos do vetor de p covariáveis $\boldsymbol{x}$.

Sejam $\boldsymbol{y} = \left(\boldsymbol{y}_{1,1}^{'}, \ldots, \boldsymbol{y}_{M,N_M}^{'}\right)^{'}$, $\boldsymbol{a} = \left(\boldsymbol{a}_{1,1}^{'}, \ldots, \boldsymbol{a}_{M,N_M}^{'}\right)^{'}$ e $\boldsymbol{b} = \left(\boldsymbol{b}_{1,1}^{'}, \ldots, \boldsymbol{b}_{M,N_M}^{'}\right)^{'}$ vetores de valores observados de $\boldsymbol{Y}\,|(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B}, \boldsymbol{u})$, $\boldsymbol{A}\,|(\boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$ e $\boldsymbol{B}\,|(\boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{B})$, respectivamente. Então, o modelo não linear beta aleatoriamente trun-

cado misto é dado por

$$
\begin{gathered}
\left(Y_{i,jk}, A_{i,jk}, B_{i,jk}\right) | \left(A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u}_i\right) \sim BTA\left(\mu_{i,jk}, \sigma_{i,jk}, \mu_A, \sigma_A, \mu_B, \sigma_B\right), \\
\mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, \qquad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right), \qquad \boldsymbol{u}_i \sim N_s\left(\boldsymbol{0}, \boldsymbol{\Psi}\right), \qquad (3.20) \\
i = 1, \ldots, M, j = 1, \ldots, N_i, k = 1, \ldots, n_{ij}.
\end{gathered}
$$

A função de log-verossimilhança do modelo (3.20) para os vetores de efeitos fixos $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Psi}$, μ_A, σ_A, μ_B e σ_B e para os efeitos aleatórios não observados $\boldsymbol{u}$, dado os dados observados $D = (n, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})$, é dada por

$$
\begin{aligned}
&\ell\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B \,|D\right) \\
&\propto \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\gamma_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right) - 1\right] \log\left(y_{i,jk}\right) + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\lambda_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right) - 1\right] \log\left(1 - y_{i,jk}\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[C^A_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right)\right] - M\log\left[|\boldsymbol{\Psi}|\right] - \frac{1}{2}\sum_{i=1}^{M} \boldsymbol{u}_i^{'}\boldsymbol{\Psi}^{-1}\boldsymbol{u}_i \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left(\mu_A e^{\sigma_A} - 1\right) \log\left(a_{i,jk}\right) + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\left(1 - \mu_A\right) e^{\sigma_A} - 1\right] \log\left(1 - a_{i,jk}\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[B\left(b; \mu_A e^{\sigma_A}, \left(1 - \mu_A\right) e^{\sigma_A}\right)\right] \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left(\mu_B e^{\sigma_B} - 1\right) \log\left(b_{i,jk}\right) + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\left(1 - \mu_B\right) e^{\sigma_B} - 1\right] \log\left(1 - b_{i,jk}\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left\{B\left(\mu_B e^{\sigma_B}, \left(1 - \mu_B\right) e^{\sigma_B}\right)\right. \\
&= \ell_1\left(\boldsymbol{\theta} \,|D\right) + \ell_2\left(\boldsymbol{\omega}_A \,|D\right) + \ell_3\left(\boldsymbol{\omega}_B \,|D\right), \qquad (3.21)
\end{aligned}
$$

com

$$
\gamma_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right) = \left[\eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,j}\boldsymbol{u}_i\right] e^{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)},
$$

$$
\lambda_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right) = \left\{1 - \left[\eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,j}\boldsymbol{u}_i\right]\right\} e^{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)},
$$

$$
C^A_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right) = B\left(b_{i,jk}; \gamma_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right), \lambda_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right)\right) - B\left(a_{i,jk}; \gamma_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right), \lambda_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}\right)\right),
$$

e $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u}\right)^{'}$, $\boldsymbol{\omega}_A = \left(\mu_A, \sigma_A\right)$ e $\boldsymbol{\omega}_B = \left(\mu_B, \sigma_B\right)$.

Assim como descrito na Seção 3, o vetor escore (3.7) e a matriz Hessiana (3.8) do modelo não linear beta aleatoriamente truncado misto são calculados pelo método de aproximação numérica de Richardson.

Modelo de regressão não linear beta truncado misto

Para as situações em que os limites de truncamento são constantes fixas e conhecidas, o modelo não linear beta truncado misto é definido como

$$\begin{gathered} Y_{i,kj}\,|(a<Y_{i,jk}<b,\boldsymbol{u}_i)\sim BT\left(\mu_{i,jk},\sigma_{i,jk},a,b\right), \\ \mu_{i,jk}=\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i, \qquad \sigma_{i,jk}=g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right), \qquad \boldsymbol{u}_i\sim N_s\left(\boldsymbol{0},\boldsymbol{\Psi}\right), \\ i=1,\ldots,M, j=1,\ldots,N_i, k=1,\ldots,n_{ij}. \end{gathered} \tag{3.22}$$

A função de log-verossimilhança do modelo (3.22) para os vetores de efeitos fixos $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ e $\boldsymbol{\Psi}$ e para os efeitos aleatórios não observados $\boldsymbol{u}$, dado os dados observados $D=(n,\boldsymbol{y},\boldsymbol{x},a,b)$, é dada por

$$\begin{aligned} &\ell\left(\boldsymbol{\theta}\,|D\right) \\ &\propto\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\left[\gamma_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)-1\right]\log\left(y_{i,jk}\right)+\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\left[\lambda_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)-1\right]\log\left(1-y_{i,jk}\right) \\ &-\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\log\left[C_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)\right]-M\log\left[|\boldsymbol{\Psi}|\right]-\frac{1}{2}\sum_{i=1}^{M}\boldsymbol{u}_i^{'}\boldsymbol{\Psi}^{-1}\boldsymbol{u}_i, \end{aligned} \tag{3.23}$$

em que

$$\gamma_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)=\left[\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i\right]e^{g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)},$$

$$\lambda_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)=\left\{1-\left[\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i\right]\right\}e^{g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)},$$

$$C_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)=B\left(b;\gamma_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right),\lambda_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)\right)-B\left(a;\gamma_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right),\lambda_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)\right),$$

e $\boldsymbol{\theta}=\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Psi},\boldsymbol{u}\right)^{'}$.

Modelo de regressão não linear beta aleatoriamente truncado

Ao assumirmos $\boldsymbol{u}_i=\boldsymbol{0}$, para todo $i=1,\ldots,M$, em (3.20), obtemos o modelo não linear beta aleatoriamente truncado, definido como

$$\begin{gathered} \left(Y_{i,jk},A_{jk},B_{i,jk}\right)|(A_{i,jk}<Y_{i,jk}<B_{i,jk})\sim BTA\left(\mu_{i,jk},\sigma_{i,jk},\mu_A,\sigma_A,\mu_B,\sigma_B\right), \\ \mu_{i,jk}=\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right), \quad \sigma_{i,jk}=g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right), \\ i=1,\ldots,M \quad j=1,\ldots,N_i, \quad k=1,\ldots,n_{ij}, \end{gathered} \tag{3.24}$$

cuja função de log-verossimilhança, dado o conjunto de dados $D = (n, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})$, é escrita como

$$
\begin{aligned}
&\ell\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B \,|D\right) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\gamma_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right) - 1\right] \log\left(y_{i,jk}\right) + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\lambda_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right) - 1\right] \log\left(1 - y_{i,jk}\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[C_{i,jk}^{A}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right] \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\mu_A e^{\sigma_A} - 1\right] \log\left(a_{i,jk}\right) + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left\{\left[1 - \mu_A\right] e^{\sigma_A} - 1\right\} \log\left(1 - a_{i,jk}\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[B\left(b; \mu_A e^{\sigma_A}, \left(1 - \mu_A\right) e^{\sigma_A}\right)\right] \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left[\mu_B e^{\sigma_B} - 1\right] \log\left(b_{i,jk}\right) + \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \left\{\left[1 - \mu_B\right] e^{\sigma_B} - 1\right\} \log\left(1 - b_{i,jk}\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left\{B\left(\mu_B e^{\sigma_B}, \left(1 - \mu_B\right) e^{\sigma_B}\right)\right. \\
&= \ell_1\left(\boldsymbol{\theta} \,|D\right) + \ell_2\left(\boldsymbol{\omega}_A \,|D\right) + \ell_3\left(\boldsymbol{\omega}_B \,|D\right), \qquad (3.25)
\end{aligned}
$$

com $\gamma_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) e^{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)}$, $\lambda_{i,jk} = \left[1 - \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right)\right] e^{g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)}$,

$$
C_{i,jk}^{A}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right) = B\left(b_{i,jk}; \gamma_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right), \lambda_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right) - B\left(a_{i,jk}; \gamma_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right), \lambda_{i,jk}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right),
$$

e $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)^{'}$, $\boldsymbol{\omega}_A = \left(\mu_A, \sigma_A\right)$ e $\boldsymbol{\omega}_B = \left(\mu_B, \sigma_B\right)$ os vetores de parâmetros do modelo (3.24).

Modelo de regressão não linear beta truncado

Fazendo-se $\boldsymbol{u}_i = \boldsymbol{0}$, $a_{i,jk} = a$ e $b_{i,jk} = b$ para todo $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, obtemos como caso particular do modelo de regressão não linear beta truncado misto (3.22), o modelo de regressão não linear beta truncado, que pode ser escrito como

$$
\begin{aligned}
&Y_{i,kj} \,|\left(a < Y_{i,jk} < b, \boldsymbol{u}_i\right) \sim BT\left(\mu_{i,jk}, \sigma_{i,jk}, a, b\right), \\
&\quad \mu_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right), \quad \sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right), \qquad (3.26) \\
&i = 1, \ldots, M, \quad j = 1, \ldots, N_i, \quad k = 1, \ldots, n_{ij}.
\end{aligned}
$$

Dado o conjunto de dados $D = (n, \boldsymbol{y}, \boldsymbol{x}, a, b)$, a função de log-verossimilhança de $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)'$, o vetor de parâmetros do modelo de regressão não linear beta truncado, é

escrita como

$$\begin{aligned}
&\ell\left(\boldsymbol{\theta}\,|D\right) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\left[\gamma_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right)-1\right]\log\left(y_{i,jk}\right)+\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\left[\lambda_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right)-1\right]\log\left(1-y_{i,jk}\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\log\left[B\left(b;\gamma_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right),\lambda_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right)\right)-B\left(a;\gamma_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right),\lambda_{i,jk}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right)\right)\right], \quad (3.27)
\end{aligned}$$

com $\gamma_{i,jk}=\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)e^{g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)}$ e $\lambda_{i,jk}=\left[1-\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)\right]e^{g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)}$.

Capítulo 4

Metodologia frequentista

Neste capítulo, apresentamos uma metodologia frequentista de estimação e diagnóstico para a classe de modelos não lineares aleatoriamente truncados mistos. Os estimadores de máxima verossimilhança (EMVs) dos parâmetros são obtidos via maximização direta da função de log-verossimilhança. Intervalos de confiança (ICs) de Wald e intervalos de confiança baseados na razão de log-verossimilhança são construídos para os parâmetros do modelo. Na análise de diagnóstico do modelo, consideramos um tipo de resíduo padronizado, duas medidas de influência global e medidas de influência local baseadas na perturbação de casos e na perturbação da variável resposta. Como critérios de seleção de modelos, utilizamos o critério de informação de Akaike (AIC, em inglês) e o critério de informação Bayesiano (BIC, em inglês).

A metodologia, aqui apresentada, é ilustrada no Capítulo 6 usando-se dados simulados. Nas Seções 8.1 e 8.2, um conjunto de dados reais é analisado usando-se esta metodologia. No Apêndice B, são apresentados exemplos dos códigos em `R` das funções usadas para se obter os EMVs dos parâmetros, ICs, resíduos e métricas de diagnósticos descritas a seguir.

4.1 Estimação

Os EMVs $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\omega}}_A$ e $\hat{\boldsymbol{\omega}}_B$ de $\boldsymbol{\theta}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ são calculados usando-se a função de log-verossimilhança $\ell\left(\boldsymbol{\theta}, \boldsymbol{\omega}_A, \boldsymbol{\omega}_B \,|D\right) = \ell_1\left(\boldsymbol{\theta}\,|D\right) + \ell_2\left(\boldsymbol{\omega}_A\,|D\right) + \ell_3\left(\boldsymbol{\omega}_B\,|D\right)$ em (3.5) como segue: $\hat{\boldsymbol{\theta}}$ é obtido de $\ell_1\left(\boldsymbol{\theta}\,|D\right)$ usando-se um procedimento iterativo que, a cada passo, fixa $\boldsymbol{u}$ e $\boldsymbol{\Psi}$ e estima $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ e depois fixa $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ e estima $\boldsymbol{u}$ e $\boldsymbol{\Psi}$, alternando entre esses dois passos até que a convergência seja alcançada, e $\hat{\boldsymbol{\omega}}_A$, e $\hat{\boldsymbol{\omega}}_B$ são obtidos através de um procedimento não linear de maximização direta das funções de log-verossimilhança $\ell_2\left(\boldsymbol{\omega}_A\,|D\right)$ e $\ell_3\left(\boldsymbol{\omega}_B\,|D\right)$, respectivamente.

O procedimento de máxima verossimilhança iterativa aplicado para obter $\hat{\boldsymbol{\theta}}$ é um caso particular do algoritmo ECM apresentado em Meng & Rubin (1993), que, por sua vez, é uma subclasse do algoritmo GEM (Dempster *et al.*, 1977) e que, portanto, possui as mesmas propriedades de convergência, sendo mais adequado para uma série de aplicações

onde a estimativa de máxima verossimilhança é complexa. Matos *et al.* (2013b) desenvolveram o algoritmo ECM para modelos lineares e não lineares mistos com presença de censura e considerando a distribuição t-Student multivariada. Em Matos *et al.* (2013a), os autores propuseram um algoritmo ECM exato para modelos não lineares com presença de censura e realizaram diagnósticos de influência global e local baseando-se na esperança condicional da função de log-verossimilhança do modelo baseada nos dados completos.

Sob condições de regularidade apropriadas (ver Apêndice C), têm-se que a distribuição assintótica de $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ é uma distribuição normal $d(\boldsymbol{\theta})$-dimensional $N_{d(\boldsymbol{\theta})}(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta}))$, em que $d(\boldsymbol{\theta})$ denota a dimensão de $\boldsymbol{\theta}$ e sendo $I(\boldsymbol{\theta})$ a matriz de informação de Fisher definida por $-H(\boldsymbol{\theta})$. Analogamente, $\sqrt{n}\,(\hat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A)$ e $\sqrt{n}\,(\hat{\boldsymbol{\omega}}_B - \boldsymbol{\omega}_B)$ são assintoticamente normais $N_{d(\boldsymbol{\omega}_A)}(\boldsymbol{\omega}_A, I^{-1}(\boldsymbol{\omega}_A))$ e $N_{d(\boldsymbol{\omega}_B)}(\boldsymbol{\omega}_B, I^{-1}(\boldsymbol{\omega}_B))$, com $d(\boldsymbol{\omega}_A)$ a dimensão de $\boldsymbol{\omega}_A$, $d(\boldsymbol{\omega}_B)$ a dimensão de $\boldsymbol{\omega}_B$, e sendo $I(\boldsymbol{\omega}_A)$ e $I(\boldsymbol{\omega}_B)$ matrizes de informação de Fisher definidas por $-H(\boldsymbol{\omega}_A)$ e $-H(\boldsymbol{\omega}_B)$, respectivamente.

Os intervalos de confiança de Wald (IC de Wald) $100(1-\alpha)\%$ para o e-ésimo elemento $\hat{\theta}_e$ do vetor de parâmetros $\hat{\boldsymbol{\theta}}$, para o t-ésimo elemento $\hat{\omega}_{A_t}$ do vetor de parâmetros $\hat{\boldsymbol{\omega}}_A$ e para o o-ésimo elemento $\hat{\omega}_{B_o}$ do vetor de parâmetros $\hat{\boldsymbol{\omega}}_B$ são obtidos de

$$\hat{\theta}_e \pm z_{\alpha/2}\sqrt{-H^{-1}_{(e)}(\hat{\boldsymbol{\theta}})}, \tag{4.1}$$

$$\hat{\omega}_{A_t} \pm z_{\alpha/2}\sqrt{-H^{-1}_{(t)}(\hat{\boldsymbol{\omega}}_A)}, \tag{4.2}$$

e

$$\hat{\omega}_{B_o} \pm z_{\alpha/2}\sqrt{-H^{-1}_{(o)}(\hat{\boldsymbol{\omega}}_B)}, \tag{4.3}$$

respectivamente, com $z_{\alpha/2}$ o quantil $\alpha/2$ da distribuição normal padrão, $H^{-1}_{(e)}(\hat{\boldsymbol{\theta}})$ o e-ésimo elemento da diagonal principal da inversa da matriz Hessiana avaliada em $\hat{\boldsymbol{\theta}}$, $H(\hat{\boldsymbol{\theta}})$, $H^{-1}_{(t)}(\hat{\boldsymbol{\omega}}_A)$ o t-ésimo elemento da diagonal principal da inversa da matriz Hessiana avaliada em $\hat{\boldsymbol{\omega}}_A$, $H(\hat{\boldsymbol{\omega}}_A)$, e $H^{-1}_{(o)}(\hat{\boldsymbol{\omega}}_B)$ o o-ésimo elemento da diagonal principal da inversa da matriz Hessiana avaliada em $\hat{\boldsymbol{\omega}}_B$, $H(\hat{\boldsymbol{\omega}}_B)$, para $e = 1, \ldots, d(\boldsymbol{\theta})$, com $d(\boldsymbol{\theta})$ a dimensão do vetor de parâmetros $\boldsymbol{\theta}$, $t = 1, \ldots, d(\boldsymbol{\omega}_A)$, com $d(\boldsymbol{\omega}_A)$ a dimensão do vetor de parâmetros $\boldsymbol{\omega}_A$ e $o = 1, \ldots, d(\boldsymbol{\omega}_B)$, com $d(\boldsymbol{\omega}_B)$ a dimensão do vetor de parâmetros $\boldsymbol{\omega}_B$.

Calculamos, também, um intervalo de confiança baseados na razão de verossimilhanças (IC-RV) para θ_e, ω_{A_t} e ω_{B_o} que são fundamentados na distribuição qui-quadrado assintótica da estatística teste de razão de log-verossimilhanças (válida sob as condições de regularidade apropriadas enunciadas no Apêndice C).

Para o modelo de regressão não linear aleatoriamente truncado misto para locação e escala (3.1), a função de log-verossimilhança perfilada para o e-ésimo elemento, θ_e, do

vetor de parâmetros $\boldsymbol{\theta}$, é definida como

$$\ell_1^*(\theta_e \,|D) = \max_{\theta_e} \left\{ \ell_1 \left(\hat{\boldsymbol{\theta}}_{(-e)} \,|D \right) \right\}, \tag{4.4}$$

em que $\hat{\boldsymbol{\theta}}_{(-e)}$ é o vetor $\boldsymbol{\theta}$ fixado no EMV $\hat{\boldsymbol{\theta}}$, exceto para o e-ésimo elemento, e $\ell_1(\boldsymbol{\theta}\,|D)$ corresponde à função de log-verossimilhança do modelo estatístico assumido para a variável resposta truncada como dada na equação (3.5).

A função de log-verossimilhança perfilada para o t-ésimo elemento ω_{A_t} de $\boldsymbol{\omega}_A$ é definida como

$$\ell_2^*(\omega_{A_t} \,|D) = \max_{\omega_{A_t}} \left\{ \ell_2 \left(\hat{\boldsymbol{\omega}}_{A_{(-t)}} \,|D \right) \right\}, \tag{4.5}$$

sendo que $\hat{\boldsymbol{\omega}}_{A_{(-t)}}$ é o vetor $\boldsymbol{\omega}_A$ fixado no EMV $\hat{\boldsymbol{\omega}}_A$, exceto para o t-ésimo elemento, e $\ell_2(\boldsymbol{\omega}_A\,|D)$ corresponde à função de log-verossimilhança do modelo estatístico assumido para a v.a. de truncamento inferior como dada na equação (3.5).

A função de log-verossimilhança perfilada para o o-ésimo elemento ω_{B_o} de $\boldsymbol{\omega}_B$ é definida como

$$\ell_3^*(\omega_{B_o} \,|D) = \max_{\omega_{B_o}} \left\{ \ell_3 \left(\hat{\boldsymbol{\omega}}_{B_{(-o)}} \,|D \right) \right\}, \tag{4.6}$$

em que $\hat{\boldsymbol{\omega}}_{B_{(-o)}}$ é o vetor $\boldsymbol{\omega}_B$ fixado no EMV $\hat{\boldsymbol{\omega}}_B$, exceto para o o-ésimo elemento, e $\ell_3(\boldsymbol{\omega}_B\,|D)$ corresponde à função de log-verossimilhança do modelo estatístico assumido para a v.a. de truncamento superior como dada na equação (3.5).

Se θ_e é o verdadeiro valor do parâmetro, então a estatística teste de razão de log-verossimilhanças $2\{\ell_1(\hat{\theta}_e\,|D) - \ell_1^*(\theta_e\,|D)\}$ possui distribuição assintoticamente qui-quadrado com um grau de liberdade (χ^2_1), e um IC-RV de $100(1-\alpha)\%$ para o e-ésimo elemento $\hat{\theta}_e$ de $\boldsymbol{\theta}$ pode ser obtido de

$$\ell_1^*(\theta_e\,|D) \geq \ell_1 \left(\hat{\theta}_e \,|D \right) - 0.5\chi^2_{1,(1-\alpha)}, \tag{4.7}$$

em que $\chi^2_{1,(1-\alpha)}$ é o quantil $(1-\alpha)$ da distribuição χ^2_1.

Analogamente, se ω_{A_t} e ω_{B_t} são os verdadeiros valores dos parâmetros, então $2\{\ell_2(\hat{\omega}_{A_t}\,|D) - \ell_2^*(\omega_{A_t}\,|D)\}$ e $2\{\ell_3(\hat{\omega}_{B_o}\,|D) - \ell_3^*(\omega_{B_t}\,|D)\}$ possuem distribuição assintoticamente qui-quadrado com um grau de liberdade (χ^2_1), e os IC-RVs de $100(1-\alpha)\%$ para o t-ésimo elemento $\hat{\omega}_{A_t}$ de $\boldsymbol{\omega}_A$ e para o o-ésimo elemento $\hat{\omega}_{B_o}$ de $\boldsymbol{\omega}_B$ são computados de

$$\ell_2^*(\omega_{A_t}\,|D) \geq \ell_2(\hat{\omega}_{A_t}\,|D) - 0.5\chi^2_{1,(1-\alpha)}, \tag{4.8}$$

e

$$\ell_3^*\left(\omega_{B_o}\,|D\right) \geq \ell_3\left(\hat{\omega}_{B_o}\,|D\right) - 0.5\chi^2_{1,(1-\alpha)}. \tag{4.9}$$

Um IC-RV $100(1-\alpha)\%$ para $\hat{\theta}_e$ pode ser calculado usando-se o seguinte procedimento:

- considere um conjunto de possíveis valores para θ_e em torno de $\hat{\theta}_e$ e denote este conjunto por Θ_e;
- calcule (4.4) para cada $\theta_e \in \Theta_e$;
- estabeleça como o IC-RV $100(1-\alpha)\%$ de $\hat{\theta}_e$ os valores de $\theta_e \in \Theta_e$ que satisfazem (4.7).

Ressaltamos que os IC-RV's $100(1-\alpha)\%$ para $\hat{\omega}_{A_t}$ e $\hat{\omega}_{B_t}$ são obtidos de forma análoga ao procedimento descrito para $\hat{\boldsymbol{\theta}}$.

O primeiro procedimento considerado para calcular ICs (4.1)-(4.3) é comumente conhecido como IC de Wald. Este tipo de IC pode ter desempenho ruim quando a distribuição do parâmetro de interesse é assimétrica ou quando o erro padrão não é um bom estimador do desvio padrão do parâmetro. Além disso, o IC de Wald também pode ter desempenho ruim para amostras de tamanhos pequenos e moderados. O segundo procedimento (4.7)-(4.9), conhecido como IC perfilado ou IC baseado na razão de verossimilhanças, assume uma distribuição qui-quadrado assintótica para a estatística teste de razão de log-verossimilhanças e não se baseia no pressuposto de normalidade do estimador. A abordagem IC-RV é superior ao método de Wald para amostras de tamanhos pequenos e moderados. Além disso, os ICs de Wald sempre são simétricos com relação aos EMVs, enquanto que os IC-RV's são capazes de captar a assimetria da função de log-verossimilhança com relação às estimativas. Por outro lado, a computação do IC de Wald é fácil e simples, ao passo que o cálculo do IC-RV é um processo iterativo que pode ser dispendioso computacionalmente, dependendo da função de log-verossimilhança do modelo. Quando ambos os intervalos concordam, há evidência de que a aproximação normal funciona bem para o problema em questão. Se os dois ICs são bastante diferentes, então a aproximação normal pode ser entendida como inapropriada e o IC-RV deve ser preferido, já que o mesmo é conhecido por ser mais preciso do que o IC de Wald. O IC-RV também possui a vantagem de ser invariante sob transformações monótonas do parâmetro de interesse.

Note que o procedimento de estimação descrito acima para o modelo (3.1) reduz-se ao caso de inferência sobre os parâmetros dos modelos definidos em (3.9), (3.10) e (3.11). Sob o modelo (3.9), a função de log-verossimilhança corresponde à (3.5) com $\boldsymbol{U}_{i,jk}\boldsymbol{u}_i = \boldsymbol{0}$ para todo $i = 1,\ldots,M$, $j = 1,\ldots,N_i$, $k = 1,\ldots,n_{ij}$ e os EMVs são computados para $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\alpha})'$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$. Para o modelo (3.10), a função de log-verossimilhança dado os dados observados $D = (n,\boldsymbol{y},\boldsymbol{x},a,b)$ corresponde a $\ell_1\left(\boldsymbol{\theta}\,|D\right)$ em (3.5), com $a_{i,jk} = a$

e $b_{i,jk} = b$ para todo $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e estima-se apenas $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u})^{'}$. Já para o modelo (3.11), a função de log-verossimilhança é dada por $\ell_1(\boldsymbol{\theta}\,|D)$ em (3.5), com $a_{i,jk} = a$, $b_{i,jk} = b$ e $\boldsymbol{U}_{i,jk}\boldsymbol{u}_i = \boldsymbol{0}$ para todo $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e estima-se apenas $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})^{'}$.

4.2 Diagnóstico

A seguir, descrevemos as métricas de diagnóstico que serão consideradas para verificar se há má especificação e falta de ajuste sob os modelos propostos. Como o nosso principal objetivo e interesse é o ajuste de modelos de regressão para a variável resposta aleatoriamente truncada, o estudo de diagnóstico é apresentado apenas para o modelo postulado para esta variável. No entanto, caso haja interesse, destacamos que a verificação de modelos também pode ser realizada para os ajustes das variáveis de truncamento inferior e superior usando-se um procedimento análogo ao aqui descrito. Além disso, estas mesmas métricas de diagnóstico também podem ser empregadas no estudo dos modelos (3.9), (3.10) e (3.11) obtidos como casos particulares do modelo de regressão não linear aleatoriamente truncado misto para locação e escala.

4.2.1 Predição

Suponha $\boldsymbol{y}$, $\boldsymbol{a}$ e $\boldsymbol{b}$ vetores n-dimensionais de valores observados de $Y\,|(A, B, A < Y < B)$, $A\,|(B, A < Y < B)$ e $B\,|(A < Y < B)$, respectivamente. Os valores preditos $\hat{y}_{i,jk}$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, são definidos como segue

$$\hat{y}_{i,jk} = \widehat{E}\left[Y_{i,jk}\,\middle|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, \boldsymbol{u})\,; \boldsymbol{x}_{i,jk}, \hat{\boldsymbol{\theta}}\right], \tag{4.10}$$

em que $\hat{\boldsymbol{\theta}}$ é o EMV de $\boldsymbol{\theta}$ e $E\left[Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})\right]$ corresponde à expressão da esperança da distribuição truncada que é assumida para a variável resposta aleatoriamente truncada $Y\,|(A, B, A < Y < B)$.

4.2.2 Resíduos

Consideramos os resíduos padronizados para verificar a adequação do modelo e para identificar *outliers* e possíveis observações influentes. Estes resíduos, para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, são definidos como

$$r^{y}_{i,jk} = \frac{y_{i,jk} - \hat{y}_{i,jk}}{\sqrt{\widehat{Var}\left[Y_{i,jk}\,\middle|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})\,; \boldsymbol{x}_{i,jk}, \hat{\boldsymbol{\theta}}\right]}}, \tag{4.11}$$

sendo que $y_{i,jk}$ é o valor observado da (i, jk)-ésima v.a resposta com $\hat{y}_{i,jk}$ seu valor predito (4.10), $\hat{\boldsymbol{\theta}}$ é o EMV de $\boldsymbol{\theta}$ e $Var\left[Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})\right]$ corresponde à expressão da variância da distribuição truncada que é assumida para a v.a resposta.

4.2.3 Influência global

As métricas de diagnóstico baseadas em deleção de casos fundamentam-se no princípio de que a influência de uma dada observação pode ser verificada comparando-se a diferença entre as estimativas dos parâmetros obtidas ajustando-se o modelo em questão ao conjunto de dados completo, D, e as estimativas dos parâmetros obtidas ajustando-se o modelo aos dados com a (i, jk)-ésima observação deletada $D_{(-i,jk)}$.

Seja $\hat{\boldsymbol{\theta}}_{(-i,jk)}$ o EMV de $\boldsymbol{\theta}$ com o (i, jk)-ésimo caso deletado. Se o (i, jk)-ésimo caso influencia as estimativas obtidas, então $\hat{\boldsymbol{\theta}}_{(-i,jk)}$ é consideravelmente diferente de $\hat{\boldsymbol{\theta}}$ e o (i, jk)-ésimo caso é considerado como influente. Como métricas de diagnóstico de casos influentes, iremos considerar a distância generalizada de Cook e a distância da verossimilhança.

A distância generalizada de Cook para $\boldsymbol{\theta}$ é obtida de

$$GC_{i,jk}\left(\boldsymbol{\theta}\right) = \left(\hat{\boldsymbol{\theta}}_{(-i,jk)} - \hat{\boldsymbol{\theta}}\right)' I\left(\hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_{(-i,jk)} - \hat{\boldsymbol{\theta}}\right), \tag{4.12}$$

e a distância da verossimilhança é definida por

$$LD_{i,jk}\left(\boldsymbol{\theta}\right) = 2\left\{\ell_1\left(\hat{\boldsymbol{\theta}}\,|D\right) - \ell_1\left(\hat{\boldsymbol{\theta}}_{(-i,jk)}\,|D\right)\right\}. \tag{4.13}$$

Em (4.12) e (4.13) $\hat{\boldsymbol{\theta}}_{(-i)}$ é aproximado por

$$\hat{\boldsymbol{\theta}}_{(-i,jk)} = \hat{\boldsymbol{\theta}} + I^{-1}\left(\hat{\boldsymbol{\theta}}\right)V_{(-i,jk)}\left(\hat{\boldsymbol{\theta}}\right), \tag{4.14}$$

em que $V_{(-i,jk)}(\hat{\boldsymbol{\theta}})$ é o elemento do vetor escore (3.7) avaliado em $\hat{\boldsymbol{\theta}}$ com a i-ésima observação deletada e a matriz $I(\hat{\boldsymbol{\theta}})$ é a matriz de informação de Fisher observada e é definida como $-H(\hat{\boldsymbol{\theta}})$, com $H(\boldsymbol{\theta})$ a matriz de (3.8).

O procedimento apresentado a seguir descreve os passos utilizados para calcular as métricas de influência global (4.12) e (4.13) sob o modelo não linear aleatoriamente truncado misto (3.1):

- obtenha o EMV, $\hat{\boldsymbol{\theta}}$, de $\boldsymbol{\theta}$;
- obtenha a aproximação numérica de $V(\hat{\boldsymbol{\theta}})$, a aproximação numérica de $I(\hat{\boldsymbol{\theta}}) = -H(\hat{\boldsymbol{\theta}})$ e a aproximação numérica de $I(\hat{\boldsymbol{\theta}})^{-1} = -H^{-1}(\hat{\boldsymbol{\theta}})$, com $V(\boldsymbol{\theta})$ definido em (3.7) e $H(\boldsymbol{\theta})$ definida em (3.8);
- calcule (4.14), para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$;

- calcule (4.12) para obter os valores da distância generalizada de Cook;
- calcule (4.13) para obter os valores da distância da verossimilhança, sendo que $\ell_1\left(\hat{\boldsymbol{\theta}}\,|D\right)$ corresponde à função de log-verossimilhança do modelo estatístico assumido para a variável resposta truncada avaliada em $\hat{\boldsymbol{\theta}}$, e $\ell_1\left(\hat{\boldsymbol{\theta}}_{(-i,jk)}\,|D\right)$ corresponde à função de log-verossimilhança do modelo estatístico assumido para a variável resposta truncada avaliada em $\hat{\boldsymbol{\theta}}_{(-i,jk)}$, com $\ell_1\left(\cdot\,|D\right)$ definida em (3.5).

Lembrando que, neste trabalho, utilizamos o método de Richardson para obter a aproximação numérica do vetor escore e da matriz Hessiana dos modelos propostos.

4.2.4 Influência local

De acordo com Cook (1986), as métricas de diagnóstico de influência local são úteis para investigar a sensibilidade do modelo a pequenas perturbações nos dados. Ainda segundo Cook (1986), as métricas de diagnóstico de influência global podem ser mascaradas e, portanto, o diagnóstico de influência local deve ser realizado para minimizar inferências incorretas sobre casos influentes.

Seja $\ell_1(\boldsymbol{\theta}\,|D)$ a função de log-verossimilhança do modelo postulado para a variável resposta aleatoriamente truncada, e seja $\boldsymbol{w}$ um vetor de perturbação k-dimensional pertencente ao espaço de perturbação $W \subset \Re^k$. Denote por $\ell_1(\boldsymbol{\theta}\,|D,\boldsymbol{w})$ a função de log-verossimilhança do modelo perturbado e assuma $\boldsymbol{w}_0 \in W$ como sendo o vetor de não perturbação tal que $\ell_1(\boldsymbol{\theta}\,|D,\boldsymbol{w}_0) = \ell_1(\boldsymbol{\theta}\,|D)$. Assim, a influência da perturbação $\boldsymbol{w}$ na estimativa dos parâmetros do modelo pode ser avaliada através do deslocamento da verossimilhança definido como

$$LD_{\boldsymbol{\theta}}\left(\boldsymbol{w}\right) = 2\left\{\ell_1\left(\hat{\boldsymbol{\theta}}\,|D\right) - \ell_1\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{w}}\,|D\right)\right\}, \tag{4.15}$$

em que $\hat{\boldsymbol{\theta}}$ é o EMV de $\boldsymbol{\theta}$ sob $\ell_1(\boldsymbol{\theta}\,|D)$ e $\hat{\boldsymbol{\theta}}_{\boldsymbol{w}}$ é o EMV de $\boldsymbol{\theta}$ sob $\ell_1(\boldsymbol{\theta}\,|D,\boldsymbol{w})$.

A curvatura normal de $LD_{\boldsymbol{\theta}}\left(\boldsymbol{w}_0 + t\boldsymbol{d}\right)$, com $t \in \Re$ e $\boldsymbol{d}$ uma norma de direção unitária, é definida por

$$C_{\boldsymbol{d}}\left(\boldsymbol{\theta}\right) = 2\left|\boldsymbol{d}'\,\boldsymbol{\Delta}'_{\boldsymbol{\theta},\boldsymbol{w}} I(\boldsymbol{\theta})^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta},\boldsymbol{w}}\boldsymbol{d}\right|, \tag{4.16}$$

sendo $\|\boldsymbol{d}\| = 1$, e $I\left(\boldsymbol{\theta}\right)$ a matriz de informação de Fisher e $\boldsymbol{\Delta}_{\boldsymbol{\theta},\boldsymbol{w}} = \partial^2\ell_1\left(\boldsymbol{\theta}\,|D,\boldsymbol{w}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{w}'$, ambas avaliadas em $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{w}_0$.

A direção $\boldsymbol{d}_{\boldsymbol{\theta},max}$, que corresponde ao autovetor de $\boldsymbol{\Delta}'_{\boldsymbol{\theta},\boldsymbol{w}} I(\boldsymbol{\theta})^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta},\boldsymbol{w}}$ avaliada em $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{w}_0$, pode ser usada para avaliar a influência local das observações em $\hat{\boldsymbol{\theta}}$.

Cook (1986) também apresentou uma metodologia para análise de influência local quando apenas uma partição do vetor de parâmetros é de interesse. No caso dos modelos de regressão não lineares aleatoriamente truncados mistos, consideramos a partição $\boldsymbol{\theta} =$

$(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{'}$, com $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{\alpha})^{'}$ e $\boldsymbol{\theta}_2 = (\boldsymbol{\Psi}, \boldsymbol{u})^{'}$ e, para estudar a influência local em $\hat{\boldsymbol{\theta}}_1$, a curvatura normal é dada por

$$C_{\boldsymbol{d},\boldsymbol{\theta}_1}(\boldsymbol{\theta}) = 2\left|\boldsymbol{d}^{'}\boldsymbol{\Delta}_{\boldsymbol{\theta},\boldsymbol{w}}^{'}\left(I(\boldsymbol{\theta})^{-1} - I_{22}\right)\boldsymbol{\Delta}_{\boldsymbol{\theta},\boldsymbol{w}}\boldsymbol{d}\right|,$$

com

$$I_{22} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I(\boldsymbol{\theta}_2)^{-1} \end{bmatrix}, \tag{4.17}$$

e $I(\boldsymbol{\theta}_2) = \partial^2 \ell(\boldsymbol{\theta}\,|D)/\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2^{'}$.

A direção $\boldsymbol{d}_{\boldsymbol{\theta}_1,max}$, que corresponde ao autovetor de $\boldsymbol{\Delta}_{\boldsymbol{\theta},\boldsymbol{w}}^{'}\left(I(\boldsymbol{\theta})^{-1} - I_{22}\right)\boldsymbol{\Delta}_{\boldsymbol{\theta},\boldsymbol{w}}$, pode ser usada para avaliar a influência local das observações em $\hat{\boldsymbol{\theta}}_1$.

Neste trabalho, consideramos o esquema de perturbação de ponderação de casos e o esquema de perturbação da resposta.

Sob o esquema de perturbação por ponderação de casos, o vetor de pesos de não perturbação corresponde a $\boldsymbol{\omega}_0 = \mathbf{1}_n$, sendo que $\mathbf{1}_n$ denota um vetor n-dimensional de elementos iguais a um, $i = 1, \ldots, M$,$j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e $n = \sum_{i=1}^{M}\sum_{j=1}^{N_i} n_{ij}$. Assim, a função de log-verossimilhança perturbada do modelo estatístico assumido para a variável resposta truncada, $\ell_1(\boldsymbol{\theta}\,|D, \boldsymbol{w})$, é escrita como

$$\begin{aligned}
\ell_1(\boldsymbol{\theta}\,|D, \boldsymbol{w}) &= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} w_{i,jk}\ell_{1_{i,jk}}(\boldsymbol{\theta}\,|D) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} w_{i,jk}\log f\left(y_{i,jk}\,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}, \boldsymbol{u}_i\,; \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} w_{i,jk}\log\left[C_{i,jk}^{A}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u})\right] + \sum_{i=1}^{M} w_{i,jk}\log f(\boldsymbol{u}_i; \boldsymbol{\Psi}),
\end{aligned} \tag{4.18}$$

com $C_{i,jk}^{A}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u})$ dada na expressão (3.6) e $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u})^{'}$.

No esquema de perturbação da variável resposta, $\boldsymbol{y}$ é substituído por $\boldsymbol{y}_{\boldsymbol{w}}$, isto é, $y_{i,jk}$ é substituído por y_{i,jk_w}, sendo $y_{i,jk_w} = y_{i,jk} + w_{i,jk}$, $i = 1, \ldots, M$,$j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e o vetor de pesos de não perturbação é $\boldsymbol{w}_0 = \mathbf{0}_n$, sendo que $\mathbf{0}_n$ denota um vetor n-dimensional de elementos iguais a zero, $n = \sum_{i=1}^{M}\sum_{j=1}^{N_i} n_{ij}$. Desta forma, no esquema de perturbação da resposta, a função de log-verossimilhança perturbada associada à variável

resposta truncada é dada por

$$\ell_1(\boldsymbol{\theta}\,|D,\boldsymbol{w}) = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \ell_{1_{i,jk}}(\boldsymbol{\theta}|D_{\boldsymbol{w}})$$
$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log f\left(y_{i,jk_w}\,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk_w} < b_{i,jk}, \boldsymbol{u}_i\,; \eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i, g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right)\right)$$
$$- \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left[C^A_{i,jk}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u})\right] + \sum_{i=1}^{M} \log f(\boldsymbol{u}_i; \boldsymbol{\Psi}), \tag{4.19}$$

com $C^A_{i,jk}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u})$ dada na expressão (3.6), $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u})^{'}$ e $D_{\boldsymbol{w}} = (n, \boldsymbol{x}, \boldsymbol{y}_{\boldsymbol{w}}, \boldsymbol{a}, \boldsymbol{b})$.

Para obter a direção $\boldsymbol{d}_{\boldsymbol{\theta},max}$, correspondente ao autovetor de $\boldsymbol{\Delta}^{'}_{\boldsymbol{\theta},\boldsymbol{w}} I(\boldsymbol{\theta})^{-1}$, e avaliar a influência local de $\boldsymbol{\theta}$ sob o modelo não linear aleatoriamente truncado misto (3.1), utilizamos o procedimento descrito a seguir:

- obtenha o EMV, $\hat{\boldsymbol{\theta}}$, de $\boldsymbol{\theta}$ e calcule a aproximação numérica de $I(\hat{\boldsymbol{\theta}})^{-1} = -H^{-1}(\hat{\boldsymbol{\theta}})$;
- considere o vetor $\boldsymbol{w}_0$ de pesos de não perturbação pertinente para cada caso: $\boldsymbol{\omega}_0 = \mathbf{1}_n$, sob o esquema de perturbação de casos e $\boldsymbol{\omega}_0 = \mathbf{0}_n$, sob o esquema de perturbação da resposta;
- calcule a aproximação numérica da matriz de derivadas de segunda ordem $\boldsymbol{\Delta}_{\hat{\boldsymbol{\theta}},\boldsymbol{w}_0}$;
- obtenha a medida de influência local $\boldsymbol{d}_{\hat{\boldsymbol{\theta}},max}$, o autovetor de $\boldsymbol{\Delta}^{'}_{\hat{\boldsymbol{\theta}},\boldsymbol{w}_0} I(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{\Delta}_{\hat{\boldsymbol{\theta}},\boldsymbol{w}_0}$.

Note que, se o interesse é estudar a influência local na partição $\hat{\boldsymbol{\theta}}_1$, sendo $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{'}$, com $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{\alpha})^{'}$ e $\boldsymbol{\theta}_2 = (\boldsymbol{\Psi}, \boldsymbol{u})^{'}$, basta calcularmos $(I(\hat{\boldsymbol{\theta}})^{-1} - I_{22})$, com I_{22} dada (4.17), e teremos a direção $\boldsymbol{d}_{\hat{\boldsymbol{\theta}}_1,max}$, correspondente ao autovetor de $\boldsymbol{\Delta}^{'}_{\hat{\boldsymbol{\theta}},\boldsymbol{w}_0}\left(I(\hat{\boldsymbol{\theta}})^{-1} - I_{22}\right)\boldsymbol{\Delta}_{\hat{\boldsymbol{\theta}},\boldsymbol{w}_0}$.

Mais uma vez, ressaltamos que neste trabalho utilizamos o método de aproximação numérica de Richardson, para obter as derivadas de primeira e segunda ordem dos modelos propostos.

4.3 Seleção de modelos

Para selecionar um modelo estatístico a partir de um conjunto de modelos candidatos, utilizamos os critérios AIC e BIC que, para os modelos de regressão não lineares aleatoriamente truncados mistos (3.1) do Capítulo 3, são calculados por

$$AIC = -2\ell\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\omega}}_A, \hat{\boldsymbol{\omega}}_B\,|D\right) + 2\left[d\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\omega}}_A, \hat{\boldsymbol{\omega}}_B\right)^{'}\right], \tag{4.20}$$

e

$$BIC = -2\ell\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\omega}}_A, \hat{\boldsymbol{\omega}}_B\,|D\right) + 2\left[d\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\omega}}_A, \hat{\boldsymbol{\omega}}_B\right)^{'}\right]\ln(n)\,. \tag{4.21}$$

Em (4.20) e (4.21), $\ell\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\omega}}_A, \hat{\boldsymbol{\omega}}_B \,|D\right)$ é a função de log-verossimilhança do modelo não linear aleatoriamente truncado misto (3.1) avaliada em $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\omega}}_A$ e $\hat{\boldsymbol{\omega}}_B$, os EMVs de $\boldsymbol{\theta}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$. Ainda em (4.20) e (4.21), $d\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\omega}}_A, \hat{\boldsymbol{\omega}}_B\right)$ denota a dimensão de $\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\omega}}_A, \hat{\boldsymbol{\omega}}_B\right)'$. Em (4.21), $n = \sum_{i=1}^{M} \sum_{j=1}^{N_i} n_{i,j}$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$.

Capítulo 5

Metodologia Bayesiana

Neste capítulo, apresentamos uma metodologia Bayesiana de estimação e diagnóstico para a classe de modelos não lineares aleatoriamente truncados mistos. As estimativas Bayesianas são computadas de amostras obtidas via um algoritmo do tipo Cadeia de Markov Monte Carlo (MCMC) das distribuições a *posteriori* dos parâmetros. Intervalos de credibilidade inter-quantil e HPD (*highest posterior density*) são construídos para os parâmetros do modelo. Na análise de diagnóstico do modelo, consideramos métricas de avaliação preditiva a *posteriori*, dois tipos de resíduos Bayesianos padronizados e a calibração de casos para diagnóstico de influência. Para seleção de modelos, consideramos o critério da soma de log-CPO e um critério de seleção baseada na abordagem Bayesiana de mistura de modelos.

A metodologia aqui apresentada é ilustrada no Capítulo 7, usando-se os dados simulados. Na Seção 8.1, um conjunto de dados reais é analisado usando-se esta metodologia. No Apêndice D, apresentamos exemplos dos códigos em `R` das funções usadas para obter as amostras MCMC das distribuições a *posteriori* e da função implementada para a seleção de modelos baseada em mistura de modelos Bayesianos.

5.1 Estimação

A análise Bayesiana dos modelos de regressão não lineares aleatoriamente truncados mistos para locação e escala, descritos anteriormente, é conduzida assumindo-se independência entre os parâmetros e considerando distribuições a *priori* fracamente informativas $\pi(\boldsymbol{\theta})$, $\pi(\boldsymbol{\omega}_A)$ e $\pi(\boldsymbol{\omega}_B)$ para $\boldsymbol{\theta}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$, respectivamente.

A função de verossimilhança baseada nos dados observados $D = (n, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})$ sob o modelo (3.1), com $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u})^{'}$, $\boldsymbol{\omega}_A = (\mu_A, \sigma_A)$ e $\boldsymbol{\omega}_B = (\mu_B, \sigma_B)$ os vetores de

parâmetros, é escrita como

$$\begin{aligned}
&\mathcal{L}\left(D\,|\boldsymbol{\theta},\boldsymbol{\omega}_A,\boldsymbol{\omega}_B\right)\\
&=\prod_{i=1}^{M}\prod_{j=1}^{N_i}\prod_{k=1}^{n_{ij}}\left\{\frac{f\left(y_{i,jk}\,|a_{i,jk},b_{i,jk},a_{i,jk}<y_{i,jk}<b_{i,jk},\boldsymbol{u}_i\,;\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i,g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)\right)}{C_{i,jk}^{A}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)}\right.\\
&\left.\times f\left(\boldsymbol{u}_i;\boldsymbol{\Psi}\right)f\left(a_{i,jk}\,|b_{i,jk},a_{i,jk}<y_{i,jk}<b_{i,jk}\,;\boldsymbol{\omega}_A\right)f\left(b_{i,jk}\,|a_{i,jk}<y_{i,jk}<b_{i,jk}\,;\boldsymbol{\omega}_B\right)\right\}\\
&=\mathcal{L}_1\left(D\,|\boldsymbol{\theta}\right)\mathcal{L}_2\left(D\,|\boldsymbol{\omega}_A\right)\mathcal{L}_3\left(D\,|\boldsymbol{\omega}_B\right),
\end{aligned}\tag{5.1}$$

onde

$$\begin{aligned}
&C_{i,jk}^{A}\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{u}\right)\\
&=F\left(b_{i,jk};\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i,g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)\right)-F\left(a_{i,jk};\eta\left(\boldsymbol{x}_{q_{i,jk}},\boldsymbol{\beta}\right)+\boldsymbol{U}_{i,jk}\boldsymbol{u}_i,g\left(\boldsymbol{x}_{r_{i,jk}},\boldsymbol{\alpha}\right)\right)
\end{aligned}$$

e $\boldsymbol{\theta}=\left(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Psi},\boldsymbol{u}\right)'$.

Pelo teorema de Bayes, segue que

$$\begin{aligned}
&\pi\left(\boldsymbol{\theta},\boldsymbol{\omega}_A,\boldsymbol{\omega}_B\,|D\right)\propto\pi\left(\boldsymbol{\theta},\boldsymbol{\omega}_A,\boldsymbol{\omega}_B\right)\mathcal{L}\left(D\,|\boldsymbol{\theta},\boldsymbol{\omega}_A,\boldsymbol{\omega}_B\right)\\
&=\pi\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\omega}_A\right)\pi\left(\boldsymbol{\omega}_B\right)\mathcal{L}_1\left(D\,|\boldsymbol{\theta}\right)\mathcal{L}_2\left(D\,|\boldsymbol{\omega}_A\right)\mathcal{L}_3\left(D\,|\boldsymbol{\omega}_B\right)\\
&=\pi_1\left(\boldsymbol{\theta}\,|D\right)\pi_2\left(\boldsymbol{\omega}_A\,|D\right)\pi_3\left(\boldsymbol{\omega}_B\,|D\right),
\end{aligned}\tag{5.2}$$

é a distribuição a *posteriori* conjunta dos parâmetros.

Como $\boldsymbol{\theta}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ são ortogonais, a análise Bayesiana de (5.2) pode ser particionada no estudo de $\pi_1\left(\boldsymbol{\theta}\,|D\right)$, $\pi_2\left(\boldsymbol{\omega}_A\,|D\right)$ e $\pi_3\left(\boldsymbol{\omega}_B\,|D\right)$.

Para os modelos de regressão não lineares aleatoriamente truncados considerados neste trabalho, as distribuições a *posteriori* conjuntas $\pi_1\left(\boldsymbol{\theta}\,|D\right)$, $\pi_2\left(\boldsymbol{\omega}_A\,|D\right)$ e $\pi_3\left(\boldsymbol{\omega}_B\,|D\right)$, bem como as distribuições condicionais completas dos parâmetros, são analiticamente intratáveis, não possuindo formas conhecidas. Desta forma, a análise Bayesiana pode ser realizada considerando-se métodos baseados em MCMC. Neste trabalho, usamos algoritmos do tipo Metropolis-Hastings para gerar amostras de $\pi_1\left(\boldsymbol{\theta}\,|D\right)$, $\pi_2\left(\boldsymbol{\omega}_A\,|D\right)$ e $\pi_3\left(\boldsymbol{\omega}_B\,|D\right)$ e então realizar inferências sobre as distribuições marginais a *posteriori* dos parâmetros.

As amostras MCMC das distribuições a *posteriori* foram obtidas utilizando-se dois tipos de algoritmos: para obter uma amostra MCMC da distribuição a posteriori de $\boldsymbol{\theta}$, utilizamos um algoritmo do tipo Metropolis-Hastings com Gibbs, no qual cada um dos candidatos, para cada um dos parâmetros, é gerado de um passeio aleatório considerando-se uma distribuição normal univariada com desvio-padrão dada pelo elemento da diagonal da matriz de covariâncias definida pela negativa da matriz Hessiana avaliada nos EMVs dos parâmetros. As amostras MCMC de $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ são obtidas usando-se um algoritmo do tipo Metropolis-Hastings, no qual os candidatos são gerados por passeio aleatório considerando-se uma distribuição normal multivariada com matriz de covariâncias dada

pela negativa da matriz Hessiana avaliada nos EMVs dos parâmetros. As convergências das cadeias foram verificadas utilizando-se o critério Geweke (Geweke, 1992). Os algoritmos MCMC Metropolis-Hastings com Gibbs e Metropolis-Hastings foram implementados em `R` e são exemplificados nos Apêndices D.1 e D.2, respectivamente. Os tamanhos das cadeias foram 100 mil com períodos de burn-in de 20 mil e salto de 200.

As estimativas Bayesianas de $\boldsymbol{\theta}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ são denotadas por $\tilde{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\omega}}_A$ e $\tilde{\boldsymbol{\omega}}_B$ e são computadas a partir de amostras MCMC de $\pi_1(\boldsymbol{\theta}\,|D)$, $\pi_2(\boldsymbol{\omega}_A\,|D)$ e $\pi_3(\boldsymbol{\omega}_B\,|D)$, respectivamente. Dois tipos de intervalos de credibilidade são calculados para $\tilde{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\omega}}_A$ e $\tilde{\boldsymbol{\omega}}_B$: um dado pela distância inter-quantil das amostras das distribuições a *posteriori*; e o outro dado pelo intervalo HPD.

Ressaltamos que o procedimento de estimação Bayesiana descrito acima é prontamente adaptável aos casos particulares do modelo (3.1), representados em (3.9), (3.10) e (3.11). Sob o modelo (3.9), a função de verossimilhança corresponde à (5.1) com $\boldsymbol{U}_{i,jk}\boldsymbol{u}_i = \boldsymbol{0}$ para todo $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e as distribuições a *priori* são especificadas para $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})^{'}$, $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$. Note que sob o modelo (3.10) a função de verossimilhança corresponde a $\mathcal{L}_1(D\,|\boldsymbol{\theta})$ em (5.1), com $D = (n, \boldsymbol{y}, \boldsymbol{x}, a, b)$, $a_{i,jk} = a$ e $b_{i,jk} = b$ para todo $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e devemos especificar apenas a distribuição a *priori* de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u})^{'}$. Para o modelo (3.11), a função de verossimilhança é dada por $\mathcal{L}_1(D\,|\boldsymbol{\theta})$ em (5.1), com $a_{i,jk} = a$, $b_{i,jk} = b$ e $\boldsymbol{U}_{i,jk}\boldsymbol{u}_i = \boldsymbol{0}$ para todo $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e a distribuição a priori é definida para $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})^{'}$.

5.2 Diagnóstico: abordagem Bayesiana

A seguir, descrevemos as métricas de diagnóstico Bayesiano de modelos consideradas para verificar a qualidade do ajuste sob os modelos propostos. A avaliação preditiva a *posteriori* (Gelman *et al.*, 1996) é baseada em variáveis de discrepância e é uma ferramenta de diagnóstico útil para verificar a qualidade do ajuste do modelo postulado. Já os resíduos baseados na distribuição preditiva a *posteriori* e os resíduos baseados na distribuição a *posteriori* dos parâmetros são usados para detectar a presença de observações *outliers* e possivelmente discrepantes, mas também servem para avaliar a adequação do modelo em questão. Para detectar observações influentes, consideramos a calibração da divergência de Kullback-Leibler (KL) (Peng & Dey, 1995; Cho *et al.*, 2009), que é baseada na densidade preditiva condicional ordenada.

Salientamos que, assim como o procedimento de estimação Bayesiana descrito acima, as métricas Bayesianas de diagnóstico de modelos, apresentadas nas seções seguintes, podem ser prontamente empregadas no estudo dos modelos (3.9), (3.10) e (3.11) obtidos como casos particulares do modelo de regressão não linear aleatoriamente truncado misto para locação e escala. Além disso, conforme a conveniência, estas mesmas métricas de diagnóstico também podem ser empregadas para verificar o ajuste dos modelos postulados

para as variáveis de truncamento inferior e superior.

5.2.1 Distribuição preditiva a *posteriori*

A distribuição preditiva a *posteriori* é a distribuição de uma observação futura (predição), de um caso não observado, ou até mesmo de uma replicação dos dados, condicionada ao modelo postulado. Para uma observação futura $\tilde{y}$ da variável resposta, e assumindo que $\tilde{y}$ e $\boldsymbol{y}$ são condicionalmente independentes dado $\boldsymbol{\theta}$, a distribuição preditiva a *posteriori* corresponde à verossimilhança de $\tilde{y}$ ponderada sobre a distribuição a *posteriori* e é definida como

$$\pi\left(\tilde{y}\,|D\right) = \int_{\Theta} \pi\left(\tilde{y}\,|\boldsymbol{\theta}\right) \pi_1\left(\boldsymbol{\theta}\,|D\right) d\Theta, \tag{5.3}$$

sendo que

$$\pi\left(\tilde{y}\,|\boldsymbol{\theta}\right) = f\left(\tilde{y}\,\middle|\tilde{a}, \tilde{b}, \tilde{a} < \tilde{y} < \tilde{b}, \boldsymbol{u}\,; \eta\left(\tilde{\boldsymbol{x}}_q, \boldsymbol{\beta}\right) + \boldsymbol{U}\boldsymbol{u}, g\left(\boldsymbol{x}_r, \boldsymbol{\alpha}\right)\right), \tag{5.4}$$

com $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Psi}, \boldsymbol{u}\right)^{'}$, $\tilde{a}$ e $\tilde{b}$ os limites de truncamento e $\tilde{\boldsymbol{x}}_q$ e $\tilde{\boldsymbol{x}}_q$ subconjuntos do vetor de covariáveis de $\tilde{\boldsymbol{x}}$ para a observação futura.

Note que, em (5.4), sob o modelo não linear normal aleatoriamente truncado misto (3.12), $f\left(\tilde{y}\,\middle|\tilde{a}, \tilde{b}, \tilde{a} < \tilde{y} < \tilde{b}, \boldsymbol{u}\,; \cdot\right)$ é a f.d.p. dada em (2.5), e sob o modelo não linear beta aleatoriamente truncado misto (3.20), $f\left(\tilde{y}\,\middle|\tilde{a}, \tilde{b}, \tilde{a} < \tilde{y} < \tilde{b}, \boldsymbol{u}\,; \cdot\right)$ é a f.d.p. dada em (2.11).

5.2.2 Avaliação preditiva a *posteriori*

Verificação de modelos usando a avaliação preditiva a *posteriori* tem como objetivo detectar diferenças entre o modelo ajustado e os dados observados. Assim como descrito em Gelman *et al.* (2000), esta verificação pode ser realizada através da geração de conjuntos de dados replicados, y^{rep}, a partir da distribuição preditiva a *posteriori* (5.3) que, por sua vez, devem ser comparados aos dados observados usando-se uma variável de discrepância, T, uma função dos dados e/ou parâmetros do modelo. A escolha da variável de discrepância é flexível, porém deve ser feita de forma a capturar características importantes dos dados. As variáveis de discrepância podem ser apresentadas graficamente ou resumidas por um p-valor. O leitor interessado pode encontrar uma discussão sobre a escolha da variável de discrepância em Gelman *et al.* (2003). Neste trabalho, consideramos a média e a variância da variável resposta e o *deviance* do modelo como as variáveis de discrepância.

Dada uma amostra MCMC de tamanho L de $\pi\left(\boldsymbol{\theta}\,|D\right)$, a variável de discrepância é computada usando-se o procedimento descrito a seguir:

1. para cada $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e cada $\boldsymbol{\theta}_l$, $l = 1, \ldots, L$, gere y^{rep}_{i,jk_l} de $\pi(\tilde{y}\,|\boldsymbol{\theta}_l)$, com $\pi(\tilde{y}\,|\cdot)$ dada em (5.4), obtendo $\boldsymbol{y}^{rep}_l$, para $l = 1, \ldots, L$;

2. calcule as variáveis de discrepância para cada $\boldsymbol{y}^{rep}_l$, $l = 1, \ldots, L$:

   - discrepância baseada na média amostral da resposta:
     $T^m(\boldsymbol{y}^{rep}_l) = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} y^{rep}_{i,jk_l}$;
   - discrepância baseada na variância amostral da resposta:
     $T^v(\boldsymbol{y}^{rep}_l) = \frac{1}{n-1}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}\left(y^{rep}_{i,jk_l} - \bar{y}^{rep}_l\right)^2$;
   - discrepância baseada no *deviance* do modelo:
     $T^{dev}(\boldsymbol{y}^{rep}_l, \boldsymbol{\theta}_l) = -2\ell_1(D^{rep}_l\,|\boldsymbol{\theta}_l)$, em que $D^{rep}_l = (n, \boldsymbol{y}^{rep}_l, \boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})$ e $\ell_1 = \log\mathcal{L}_1$, com $\mathcal{L}_1$ dada em (5.1);

3. calcule as variáveis de discrepância para os dados observados $T(\boldsymbol{y})$:

   - $T^m(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} y_{i,jk}$;
   - $T^v(\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}}(y_{i,jk} - \bar{y})^2$;
   - $T^{dev}\left(\boldsymbol{y}, \tilde{\boldsymbol{\theta}}\right) = -2\ell_1\left(D\,\middle|\tilde{\boldsymbol{\theta}}\right)$.

Se a variável de discrepância depende dos parâmetros e dos dados, então devemos obter $T(\boldsymbol{y}, \boldsymbol{\theta})$ e $T(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$ para cada amostra MCMC da distribuição a *posteriori*, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$. Gelman *et al.* (2003) sugerem que a comparação entre as variáveis de discrepância calculada para os dados observados e os dados replicados seja apresentada em um histograma das diferenças $T(\boldsymbol{y}, \boldsymbol{\theta}) - T(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$, ou em um gráfico de dispersão de $T(\boldsymbol{y}, \boldsymbol{\theta})$ contra $T(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$. Ao avaliar o histograma, espera-se que o zero esteja contido no mesmo para considerar o modelo como bem ajustado. No caso do gráfico de dispersão, espera-se que o mesmo seja simétrico em torno de uma linha de 45°.

O p-valor preditivo a *posteriori* é definido como $p = P[T(\boldsymbol{y}^{rep}) \geqslant T(\boldsymbol{y})\,|y]$, que, baseado em uma amostra MCMC da distribuição a *posteriori* e nos dados replicados, é estimado por

$$\hat{p} = \frac{\#\{T(\boldsymbol{y}^{rep}) \geqslant T(\boldsymbol{y})\}}{L}. \tag{5.5}$$

Assim como apontado por Lynch & Western (2004), o p-valor Bayesiano deve ser interpretado como a probabilidade de se observar dados ao menos tão extremos quanto os realmente observados, condicionada ao modelo. Logo, p-valores preditivos a *posteriori* próximos a 0 ou 1 sugerem que os dados possuem uma discrepância extrema e que o modelo

ajustado pode ser inapropriado. O p-valor preditivo a *posteriori* não deve ser interpretado como a probabilidade de que o modelo considerado seja verdadeiro condicionado aos dados. Mais conselhos para a interpretação apropriada do p-valor (5.5) podem ser encontradas em Gelman *et al.* (2003).

5.2.3 Predição

O valor predito, $\tilde{y}_{i,jk}$, de uma observação, $y_{i,jk}$, para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, pode ser obtido de uma amostra MCMC de tamanho L, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$, de $\pi_1(\boldsymbol{\theta}\,|D)$ a partir do seguinte procedimento que gera uma amostra da distribuição preditiva a *posteriori* (5.3):

- para cada $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e cada $\boldsymbol{\theta}_l$, $l = 1, \ldots, L$, gere $\tilde{y}_{i,jk_l}$ de $\pi(\tilde{y}\,|\boldsymbol{\theta}_l)$, com $\pi(\tilde{y}\,|\cdot)$ definida em (5.4);
- para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, defina $\tilde{y}_{i,jk}$, o valor predito da (i, jk)-ésima observação como

$$\tilde{y}_{i,jk} = \frac{1}{L}\sum_{l=1}^{L} \tilde{y}_{i,jk_l}. \tag{5.6}$$

5.2.4 Resíduos baseados na distribuição preditiva a *posteriori*

O resíduo preditivo a *posteriori* (Yan & Sedransk, 2010) pode ser usado para verificar a adequação do modelo e a presença de observações *outliers*. Tais resíduos, para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, são definidos como $y_{i,jk} - \tilde{y}_{i,jk}$, sendo que $y_{i,jk}$ é o valor observado da (i, jk)-ésima resposta e $\tilde{y}_{i,jk}$ é seu valor predito obtido de (5.6).

O resíduo preditivo a *posteriori* padronizado, para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, é escrito como

$$r_{i,jk}^{pred} = \frac{y_{i,jk} - \tilde{y}_{i,jk}}{\sqrt{Var\left[Y_{i,jk}\,\middle|\,(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk})\,; D, \tilde{\boldsymbol{\theta}}\right]}}, \tag{5.7}$$

em que $\tilde{\boldsymbol{\theta}}$ é a estimativa Bayesiana de $\boldsymbol{\theta}$, $\tilde{y}_{i,jk}$ é o valor predito de $y_{i,jk}$ dado por (5.6) e $Var\left[Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_i < B_{i,jk})\right]$ é a expressão da variância da distribuição truncada assumida para a variável resposta.

Valores de $r_{i,jk}^{pred}$ centrados em zero indicam que o modelo está bem ajustado aos dados observados.

5.2.5 Resíduos baseados na distribuição a *posteriori*

Os resíduos baseados na distribuição a *posteriori* dos parâmetros foram propostos inicialmente por Albert & Chib (1993), que notaram que sob a perspectiva Bayesiana o resíduo $r^{post} = y - E(Y|D;\boldsymbol{\theta})$ possui uma distribuição a *posteriori* que pode ser usada para detectar observações *outliers*. Em Albert & Chib (1995), os autores destacaram que r^{post} é uma função do vetor de parâmetros $\boldsymbol{\theta}$ e, por consequência, a precisão do conhecimento sobre $\boldsymbol{\theta}$ é refletida na precisão dos resíduos. O resíduo padronizado baseado na distribuição a *posteriori* é dado por

$$r_{i,jk}^{post} = \frac{y_{i,jk} - E\left[Y_{i,jk}\left|\left(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_i < B_{i,jk}\right); D, \boldsymbol{\theta}\right.\right]}{\sqrt{Var\left[Y_i\left|\left(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_i < B_{i,jk}\right); D, \boldsymbol{\theta}\right.\right]}}, \tag{5.8}$$

sendo que $E\left[Y_i\left|\left(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_i < B_{i,jk}\right)\right.\right]$ é a esperança e $Var\left[Y_i\left|\left(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_i < B_{i,jk}\right)\right.\right)$ é a variância do modelo de regressão não linear aleatoriamente truncado misto para locação e escala assumido para a variável resposta.

Dada uma amostra MCMC de tamanho L de $\pi_1(\boldsymbol{\theta}|D)$, uma amostra de (5.8) é obtida usando-se o algoritmo:

- para cada $\boldsymbol{\theta}_l$, $l = 1, \ldots, L$, calcule $E\left[Y_{i,jk}\left|\left(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}\right); D, \boldsymbol{\theta}_l\right.\right]$ e $Var\left[Y_{i,jk}\left|\left(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}\right); D, \boldsymbol{\theta}_l\right.\right]$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$;
- para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, calcule (5.8) para obter uma amostra de tamanho L dos resíduos.

Uma observação é considerada como *outlier* se a média de seus resíduos padronizado baseados na distribuição a *posteriori* estão distantes de zero.

5.2.6 Influência global

A densidade preditiva condicional ordenada (CPO, em inglês) (Gelfand *et al.*, 1992), definida como a densidade preditiva do (i, jk)-ésimo caso condicionado aos dados sem o (i, jk)-ésimo caso, $D_{(-i,jk)}$, para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, é dada por

$$CPO_{i,jk} = \left[\int_{\Theta} \frac{1}{f\left(y_{i,jk}\left|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\theta}\right.\right)} \pi_1(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}\right]^{-1}, \tag{5.9}$$

em que $f\left(y_{i,jk}\left|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\theta}\right.\right)$ é a densidade da distribuição truncada assumida para a variável resposta e $\pi_1(\boldsymbol{\theta}|D)$ é a distribuição a *posteriori* do vetor de parâmetros $\boldsymbol{\theta}$.

Baseado em uma amostra MCMC $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ de $\pi_1(\boldsymbol{\theta}\,|D)$, a aproximação de Monte Carlo de (5.9) é obtida por

$$\widehat{CPO}_{i,jk} = \left[\frac{1}{L}\sum_{l=1}^{L}\frac{1}{f\left(y_{i,jk}\,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\theta}_l\right)}\right]^{-1}. \tag{5.10}$$

A calibração da divergência de KL baseada na CPO usa a perspectiva Bayesiana no diagnóstico de deleção de casos e avalia a influência de uma dada observação nas estimativas dos parâmetros. A divergência de KL, para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, é dada por

$$\begin{aligned} &K\left(\pi_1(\boldsymbol{\theta}\,|D), \pi_1\left(\boldsymbol{\theta}|D_{(-i,jk)}\right)\right) \\ &= -\log\left(CPO_{i,jk}\right) + E_{\boldsymbol{\theta}}\left\{\log f\left(y_{i,jk}\,|a_{i,jk}, b_{i,jk} a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\theta}\right)|D\right\}. \end{aligned} \tag{5.11}$$

Seguindo Peng & Dey (1995) e Cho *et al.* (2009), a calibração, para $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, é escrita como

$$p_{i,jk} = 0.5\left\{1 + \sqrt{1 - \exp\left[-2\hat{K}\left(\pi_1(\boldsymbol{\theta}\,|D), \pi_1\left(\boldsymbol{\theta}|D_{(-i,jk)}\right)\right)\right]}\right\}, \tag{5.12}$$

em que

$$\begin{aligned} &\hat{K}\left(\pi_1(\boldsymbol{\theta}\,|D), \pi_1\left(\boldsymbol{\theta}|D_{(-i,jk)}\right)\right) = \log\left\{\frac{1}{L}\sum_{l=1}^{L}\frac{1}{f\left(y_{i,jk}\,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\theta}_l\right)}\right\} \\ &+ \frac{1}{L}\sum_{l=1}^{L}\log f\left(y_{i,jk}\,|a_{i,jk}, b_{i,jk}, a_{i,jk} < y_{i,jk} < b_{i,jk}; \boldsymbol{\theta}_l\right), \end{aligned} \tag{5.13}$$

é a aproximação de Monte Carlo da divergência de KL entre a distribuição a *posteriori* com os dados completos e a distribuição a *posteriori* com o (i, jk)-ésimo caso deletado, baseada em uma amostra MCMC de tamanho L de $\pi_1(\boldsymbol{\theta}\,|D)$. Valores de (5.12) substancialmente maiores do que 0,5 indicam que a observação é influente.

Para aproximar (5.12) usando uma amostra MCMC de tamanho L de $\pi_1(\boldsymbol{\theta}\,|D)$, consideramos o procedimento a seguir:

- para cada $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e cada $\boldsymbol{\theta}_l$, $l = 1, \ldots, L$, calcule (5.13);
- para cada $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, calcule (5.12) usando os valores de (5.13) obtidos anteriormente.

5.3 Seleção de modelos

A métrica conhecida como soma de log-CPO (Gelfand *et al.*, 1992) e definida como

$$\log\text{-CPO} = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{n_{ij}} \log\left(CPO_{i,jk}\right), \tag{5.14}$$

é um estimador da log-verossimilhança marginal dos dados observados e pode ser usada como um critério de seleção de modelos. Sob este critério, o modelo a ser selecionado é aquele que fornece o maior valor de (5.14) (Carlin & Louis, 2009).

Dada uma amostra MCMC de tamanho L de $\pi_1\left(\boldsymbol{\theta}\,|D\right)$, o seguinte procedimento pode ser usado para calcular (5.14):

- para cada $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$ e cada $\boldsymbol{\theta}_l$, $l = 1, \ldots, L$, calcule (5.10);
- calcule (5.14) usando os valores de (5.10) obtidos anteriormente.

A seguir, descrevemos o procedimento de seleção de modelos baseada no enfoque Bayesiano de modelo de mistura. A motivação para desenvolver esta métrica é justificada como segue: se o nosso interesse é escolher entre Q modelos candidatos, então as estimativas Bayesianas que compõem o vetor de probabilidades de mistura $\tilde{\boldsymbol{\rho}} = (\tilde{\rho}_1, \ldots, \tilde{\rho}_Q)$ podem ser interpretadas como as probabilidades a *posteriori* de que os dados observados D tenham sido gerados por cada um dos q modelos candidatos, $q = 1, \ldots, Q$. Portanto, podemos definir um critério de seleção no qual o modelo com maior valor de $\tilde{\rho}_q$ é tido como o mais adequado para os dados em questão, sendo então selecionado.

Desta forma, suponha que há interesse em escolher entre Q modelos candidatos $\mathcal{M}_1, \ldots, \mathcal{M}_Q$ e denote por $\pi_1\left(\boldsymbol{\theta}_q\,|D\right)$ a distribuição a *posteriori* de $\boldsymbol{\theta}$ sob o q-ésimo modelo candidato, $q = 1, \ldots, Q$. Então, a distribuição a posteriori de $\boldsymbol{\theta}$ pode ser escrita como

$$\pi_1\left(\boldsymbol{\theta}\,|D\right) = \sum_{q=1}^{Q} \rho_q \pi_1\left(\boldsymbol{\theta}_q\,|D\right), \tag{5.15}$$

em que $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_Q)$ são as probabilidades de mistura, com $0 < \rho_q \leq 1$ e $\sum_{q=1}^{Q} \rho_q = 1$, $q = 1, \ldots, Q$.

Para simplificar a interpretação do problema, bem como os cálculos, associamos variáveis latentes $z_{i,jk}$ a cada observação $D_{i,jk} = (y_{i,jk}, \boldsymbol{x}_{i,jk}, a_{i,jk}, b_{i,jk})$ tal que

$$\begin{gathered}\boldsymbol{\theta}\,|D_{i,jk}; Z_{i,jk} = q \sim \pi_1\left(\boldsymbol{\theta}_q\,|D\right), \\ P\left(Z_{i,jk} = q\right) = \rho_q,\end{gathered}$$

isto é, para cada observação no conjunto de dados, associamos uma variável latente que indica o componente de mistura do qual a observação foi gerada. Desta forma, a probabilidade a *posteriori* de que a observação $D_{i,jk}$ tenha sido gerada do q-ésimo modelo é definida por

$$P\left(z_{i,jk} = q \,|D_{i,jk}; \boldsymbol{\theta}, \boldsymbol{\rho}\right) = \frac{\rho_q \pi_1\left(\boldsymbol{\theta}_q | D_{i,jk}\right)}{\sum_{q=1}^{Q} \rho_q \pi_1\left(\boldsymbol{\theta}_q | D_{i,jk}\right)}.$$

Considere que a distribuição a *priori* de $\boldsymbol{\rho}$ é dada por uma distribuição Dirichlet com hiperparâmetros fixos e conhecidos $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_Q)$, $\boldsymbol{\rho}\,|\boldsymbol{\delta} \sim D\left(\delta_1, \ldots, \delta_Q\right)$. Seja $\boldsymbol{z} = \left(\boldsymbol{z}'_{1,1}, \ldots, \boldsymbol{z}'_{M,N_M}\right)'$ um vetor n-dimensional de valores observados de $\boldsymbol{Z}$, composto pelos $Z_{i,jk}$'s, $n = \sum_{i=1}^{M} \sum_{j=1}^{N_i} n_{ij}$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$. Então, a distribuição a *posteriori* de $\boldsymbol{\rho}$ dados D e $\boldsymbol{z}$, denotada por $\pi\left(\boldsymbol{\rho}\,|D, \boldsymbol{z}\right)$, é uma distribuição $D\left(\delta_1^*, \ldots, \delta_Q^*\right)$, com $\delta_q^* = \delta_q + n_q$, $n_q = \#\left\{z_{i,jk} = q\right\}$, $\sum_{q=1}^{Q} n_q = n$, $n = \sum_{i=1}^{M} \sum_{j=1}^{N_i} n_{ij}$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$.

Dadas amostras MCMC de cada modelo candidato $\mathcal{M}_q$ com distribuição a *posteriori* $\pi_1\left(\boldsymbol{\theta}_q\,|D\right)$, o seguinte algoritmo do amostrador de Gibbs pode ser usado para gerar valores da distribuição a *posteriori* $\pi\left(\boldsymbol{\rho}\,|D, \boldsymbol{z}\right)$:

1. calcule as estimativas Bayesianas $\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_q$ sob cada modelo candidato $\mathcal{M}_q$ com distribuição a *posteriori* $\pi_1\left(\boldsymbol{\theta}_q\,|D\right)$, $q = 1, \ldots, Q$;

2. para cada $D_{i,jk}$, calcule as densidades a *posteriori* sob cada modelo candidato $\pi_1\left(\tilde{\boldsymbol{\theta}}_q\,|D_{i,jk}\right)$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$;

3. defina valores iniciais $\boldsymbol{\rho}^{(0)}$;

4. para $t = 1, \ldots$:

   - gere $z_{i,jk}^{(t)}$ de $(q = 1, \ldots, Q)$, com

$$P\left(z_{i,jk}^{(t)} = q \,\middle|D_{i,jk}; \boldsymbol{\theta}, \boldsymbol{\rho}^{(t-1)}\right) = \frac{\rho_q^{(t)} \pi_1\left(\boldsymbol{\theta}_q | D_{i,jk}\right)}{\sum_{q=1}^{Q} \rho_q^{(t)} \pi_1\left(\boldsymbol{\theta}_q | D_{i,jk}\right)};$$

   - gere $\boldsymbol{\rho}^{(t)}$ de $D\left(\delta_1^*, \ldots, \delta_Q^*\right)$, com $\delta_q^* = \delta_q + n_q$, $n_q = \#\left\{z_{i,jk} = q\right\}$, $\sum_{q=1}^{Q} n_q = n$, $n = \sum_{i=1}^{M} \sum_{j=1}^{N_i} n_{ij}$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$.

A estimativa Bayesiana de $\boldsymbol{\rho}$ é denotada por $\tilde{\boldsymbol{\rho}}$ e é obtida da amostra MCMC de $\pi\left(\boldsymbol{\rho}\,|D, \boldsymbol{z}\right)$.

Capítulo 6

Estudo de simulação: metodologia frequentista

Neste capítulo, são apresentados os resultados de estudos de simulação para verificar a qualidade e estudar as propriedades frequentistas dos EMVs obtidos para os modelos não lineares aleatoriamente truncados mistos propostos no Capítulo 3. Apresentamos, também, resultados baseados em dados simulados para ilustrar e avaliar a eficácia das diferentes métricas de diagnóstico de modelos consideradas, e exemplificar a sua capacidade de detectar observações *outliers* e influentes, quando estes mesmos modelos são ajustados a um conjunto de dados perturbados.

Os conjuntos de dados simulados foram gerados baseados em dados de retenção de água em solo, que são geralmente utilizados para construir as curvas de retenção de água em solo, discutidas na Seção 1.1. Lembramos que, neste tipo de dados, o teor de umidade de uma amostra de solo é limitado inferiormente pela umidade residual, θ_r, e é limitada superiormente pela umidade saturada, θ_s. O teor de umidade residual, corresponde à umidade do solo medida na tensão de $15atm$ e o teor de umidade saturada corresponde à umidade do solo medida na tensão de $0atm$. Assim, nos modelos aqui propostos, a variável resposta truncada $Y\,|(A, B, A < Y < B)$ corresponde ao teor de umidade, e os limites de truncamento inferior e superior são $\theta_r = A\,|(B, A < Y < B)$ e $\theta_s = B\,|(A < Y < B)$.

Para cada modelo de regressão a variável explicativa x representa níveis de tensão de sucção variando de $0,01atm$ a $10atm$, $x = (0,01; 0,03; 0,06; 0,1; 0,33; 0,8; 4; 10)$, e portanto, temos $n_{ij} = 8$ observações na j-ésima replicata do i-ésimo grupo, com $k = 1, \ldots, n_{ij}$. Assim como no conjunto de dados reais, os dados são simulados considerando-se que as amostras de solo foram coletadas em três profundidades, $0 - 5cm$, $15 - 20cm$, e $60 - 65cm$, e portanto temos $M = 3$ grupos e $i = 1, 2, 3$.

Os dados de retenção são geralmente medidos em replicatas, e para as simulações apresentadas a seguir, consideramos 12 possibilidades de número de replicatas fazendo $N_i = \{1; 3; 5; 10; 15; 20; 25; 30; 35; 40; 45; 50\}$ replicações e $j = 1, \ldots, N_i$ o que nos fornece tamanhos amostrais $n = \{24; 72; 120; 240; 360; 480; 600; 720; 840; 960; 1080; 1200\}$. Desta

forma, pretendemos verificar a partir de qual tamanho amostral n os EMV dos parâmetros do modelo proposto passam a ter boas propriedades frequentistas, isto é, para qual tamanho amostral as propriedades assintóticas consideradas na Seção 4.1 do Capítulo 4 passam a ser satisfatoriamente válidas.

Os conjuntos de dados foram simulados considerando que o parâmetro de locação $\mu_{i,jk}$ é associado à variável explicativa x e aos efeitos aleatórios pela relação $\mu_{i,jk} = \eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + \boldsymbol{U}_{i,jk}\boldsymbol{u}_i$. A função $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ é dada por uma das expressões da curva de retenção apresentadas na Seção 1.1: na simulação do modelo não linear normal aleatoriamente truncado (3.12), consideramos a expressão de van Genuchten (1980) com a restrição de Mualem (1976), dadas nas equações (1.2) e (1.4); e na simulação do modelo não linear beta aleatoriamente truncado (3.20), consideramos a expressão de Gardner (1958), dada em (1.1). Os efeitos aleatórios representam a variação não observável devida ao nível de profundidade no qual a amostra de solo é obtida e cada $\boldsymbol{u}_i$, $i = 1, 2, 3$ é um vetor unitário, u_i, e portanto temos $s = 1$ e $\boldsymbol{u} = (u_1, u_2, u_3)$. A matriz de delineamento $\boldsymbol{U}_{i,j}$ é uma matriz de zeros e uns, sendo que seu (i, j)-ésimo elemento é igual a um se a amostra pertence ao i-ésimo grupo e zero em caso contrário, para $k = 1, \ldots, n_{ij}$. Logo, podemos escrever $\mu_{i,jk} = \eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i$.

Ressaltamos que os valores do conteúdo de água residual, θ_r, e o conteúdo de água no solo saturado, θ_s nas expressões de retenção de água em solo descritas na Seção 1.1 não são parâmetros de ajuste mas sim os limites inferior e superior de truncamento, e dado que estas duas características do solo variam entre amostras de solo e são diferentes para cada profundidade do solo, elas devem ser tratadas como sendo variáveis aleatórias. Assim, para cada $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, temos que $\theta_{r_{i,jk}} = A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i)$ e $\theta_{s_{i,jk}} = B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk})$ são as variáveis de truncamento inferior e superior, que correspondem às medições de conteúdo de água residual e conteúdo de água no solo saturado, de cada amostra, em cada replicata, em cada nível de profundidade do solo.

No caso do parâmetro de escala, σ, duas estruturas foram consideradas: a estrutura homoscedástica dada por $\sigma_{i,jk} = \sigma$, e a estrutura heteroscedástica dada por $\sigma_{i,jk} = \sigma x_{i,jk}^{\alpha}$. Porém, vale a pena destacar que as variâncias das distribuições normal truncada e beta truncada, apresentadas nas expressões (2.7) e (2.13), dependem do parâmetro de escala (dispersão) e do parâmetro de locação (média) da distribuição. Desta forma, o fato de assumirmos que $\sigma_{i,jk} = \sigma$ para $i = 1, \ldots, M$ e $j = 1, \ldots, N_i$ e $k = 1, \ldots, n_{ij}$, não implica que os dados sejam homoscedásticos, uma vez que a variância de cada observação será uma função de σ e $\mu_{i,jk} = \eta\left(\boldsymbol{x}_{i,jk}, \boldsymbol{\beta}\right) + u_i$.

Os conjuntos de dados simulados foram gerados no software estatístico `R` e os EMVs dos parâmetros dos modelos foram obtidos como descrito na Seção 4.1. Para cada tamanho amostral, foram simulados um total de 200 conjuntos de dados.

6.1 Resultados de simulação para o modelo de regressão não linear normal aleatoriamente truncado misto

Para simular o vetor aleatório $(Y_{i,jk}, A_{i,jk}, B_{i,jk})\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i)$ com distribuição normal aleatoriamente truncada como em (2.16), procedemos como segue: primeiro a v.a. de truncamento superior é obtida de $B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}) \sim N(\mu_B, \sigma_B)$; em seguida, a v.a. de truncamento inferior é gerada como $A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i) \sim NT(\mu_A, \sigma_A, -\infty, b_{i,jk})$; e finalmente a variável resposta aleatoriamente truncada é dada por $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i) \sim NT\left(\eta(x_{i,jk}, \boldsymbol{\beta}) + u_i, \sigma x_{i,jk}^{\alpha}, a_{i,jk}, b_{i,jk}\right)$, com $u_i \sim N(0, \sigma_u)$, $i = 1, 2, 3$, $j = 1, \ldots, N_i$ e $k = 1, \ldots, n_{ij}$. A função $\eta(x_{i,jk}, \boldsymbol{\beta})$ é dada pela expressão de van Genuchten (1980)-Mualem (1976) em (1.2)-(1.4).

Os valores assumidos para os parâmetros foram: $\beta_1 = 55$ e $\beta_2 = 1,45$, $\sigma = 0,01$, $\alpha = -0,15$, $\sigma_u = 0,02$, $\mu_A = 0,25$, $\sigma_A = 0,01$, $\mu_B = 0,5$ e $\sigma_B = 0,05$.

Os resultados obtidos no estudo de simulação são mostrados nas Tabelas 6.1-6.1. A Tabela 6.1 fornece a média dos EMVs obtidos para os conjuntos de dados simulados para cada tamanho amostral. Nesta mesma tabela também são mostrados os verdadeiros valores dos parâmetros fixos, não sendo possível mostrar os valores reais dos efeitos aleatórios u_i, $i = 1, 2, 3$, que mudam para cada conjunto simulado. Na Tabela 6.2 apresentamos o vício dos EMVs e a Tabela 6.3 fornece o erro quadrático médio (EQM) dos EMVs. A Tabela 6.4 fornece a probabilidade de cobertura estimada de ICs de Wald de 95% de confiança, e os erros da cauda à esquerda e à direita dos ICs de Wald. Na Tabela 6.5, apresentamos a probabilidade de cobertura estimada de intervalos IC-RV de 95% e os erros da cauda à esquerda e à direita dos IC-RV's.

Os resultados das Tabelas 6.1-6.5 indicam que os EMVs dos parâmetros fixos da expressão não linear $\eta(x_{i,jk}, \boldsymbol{\beta})$, β_1 e β_2, relacionados aos parâmetros de média $\mu_{i,jk}$ das respostas apresentam boas propriedades frequentistas. Nota-se que, à medida que o tamanho amostral aumenta, tanto o vício (Tabela 6.2) como o EQM (Tabela 6.3) dos EMVs são razoavelmente baixos e que a probabilidade de cobertura estimada aproxima-se da nominal esperada de 95% (Tabela 6.4). Essas mesmas observações valem para os parâmetros σ e α relacionados à dispersão das respostas.

No que diz respeito ao parâmetro σ_u, relacionado aos efeitos aleatórios não observáveis, a média dos EMVs (Tabela 6.1) parece indicar que o mesmo é subestimado e, como sua variância estimada também é pequena, seus intervalos assintóticos de 95% de confiança sistematicamente não contêm o verdadeiro valor do parâmetro, fazendo com que a sua probabilidade de cobertura seja excessivamente baixa (Tabela 6.4). Por outro lado, a falta de precisão ao estimar σ_u não parece afetar as estimativas u_1, u_2 e u_3, cujos vícios e EQMs são pequenos e cujas probabilidades de cobertura estimadas aproximam-se da nominal esperada de 95%.

As probabilidades de coberturas estimadas para os IC-RV's dos parâmetros fixos β_1, β_2, σ, α e dos efeitos aleatórios u_1, u_2 e u_3 estão razoavelmente abaixo da nominal esperada (Tabela 6.5). Porém, por sua característica de capturar a assimetria da função de log-verossimilhança em torno da estimativa do parâmetro, vemos que o IC-RV apresenta uma melhor probabilidade de cobertura estimada para o parâmetro σ_u do que o IC de Wald. A assimetria com relação à σ_u pode também ser notada nos erros das caudas à esquerda e à direita estimados, que para esse parâmetro é observado apenas à direita. Para os demais parâmetros, nota-se que, à medida que o tamanho amostral aumenta, os erros das caudas à esquerda e à direita são relativamente simétricos.

Ainda pelos resultados apresentados nas Tabela 6.1-6.5, nota-se que para os parâmetros dos modelos assumidos para as v.a.'s de truncamento inferior e superior, μ_A, σ_A, μ_B e σ_B, tanto os vícios (Tabela 6.2) quanto os EQMs (Tabela 6.3) observados são pequenos e que os EMVs (Tabela 6.1) estimam os verdadeiros valores dos parâmetros com precisão. Além disso, as probabilidades de cobertura estimadas aproximam-se da nominal esperada de 95% para os ICs de Wald e os IC-RV's (Tabelas 6.4 e 6.5).

De maneira geral, e principalmente para os parâmetros β_1, β_2, σ e α, os resultados do estudo de simulação conduzido para o caso do modelo de regressão não linear normal aleatoriamente truncado misto, apresentado nesta seção, aparenta indicar que a partir do tamanho amostral $n = 720$ os EMVs dos parâmetros do modelo em questão passam a ter boas propriedades frequentistas, isto é, suas estimativas tornam-se satisfatoriamente precisas e tanto o vício quanto o EQM tornam-se razoavelmente pequenos. Assim, na Table 6.6 apresentamos os resultados da simulação para o tamanho amostral $n = 720$.

Tabela 6.1: Resultados da simulação para o modelo normal aleatoriamente truncado misto van Genuchten-Mualem: valor médio dos EMVs.

	Parâmetro											
	β_1	β_2	σ	α	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
n	55,00	1,45	0,01	-0,15	0,05	-	-	-	0,25	0,01	0,50	0,05
24	54,2025	1,4578	0,0097	-0,1120	0,0332	0,0109	0,0041	0,0012	0,2500	0,0095	0,4971	0,0494
72	54,9412	1,4425	0,0098	-0,1453	0,0131	0,0045	0,0058	0,0050	0,2500	0,0098	0,4995	0,0503
120	55,6604	1,4409	0,0099	-0,1454	0,0131	0,0039	0,0072	0,0041	0,2501	0,0099	0,4996	0,0507
240	55,2925	1,4454	0,0099	-0,1465	0,0135	0,0066	0,0066	0,0069	0,2500	0,0100	0,4998	0,0504
360	55,2117	1,4470	0,0099	-0,1494	0,0139	0,0071	0,0071	0,0065	0,2500	0,0100	0,5000	0,0501
480	54,8355	1,4495	0,0099	-0,1505	0,0161	0,0093	0,0070	0,0077	0,2500	0,0100	0,5002	0,0501
600	55,2092	1,4475	0,0099	-0,1493	0,0159	0,0111	0,0084	0,0072	0,2500	0,0100	0,4999	0,0499
720	54,8966	1,4498	0,0100	-0,1503	0,0154	0,0079	0,0092	0,0099	0,2500	0,0100	0,4999	0,0500
840	54,9478	1,4497	0,0099	-0,1488	0,0153	0,0085	0,0085	0,0095	0,2500	0,0100	0,5001	0,0501
960	54,9088	1,4493	0,0100	-0,1504	0,0153	0,0086	0,0094	0,0079	0,2500	0,0100	0,5000	0,0500
1080	55,1471	1,4488	0,0100	-0,1489	0,0144	0,0081	0,0067	0,0078	0,2500	0,0100	0,5000	0,0500
1200	55,0552	1,4496	0,0100	-0,1504	0,0153	0,0079	0,0097	0,0091	0,2500	0,0100	0,4999	0,0500

Tabela 6.2: Resultados da simulação para o modelo normal aleatoriamente truncado misto van Genuchten-Mualem: vício.

	Parâmetro											
n	β_1	β_2	σ	α	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
24	-7,97E-01	7,83E-03	-3,22E-04	3,80E-02	-1,68E-02	-1,78E-03	-1,04E-03	-6,53E-04	2,99E-05	-5,19E-04	-2,89E-03	-6,49E-04
72	-5,88E-02	-7,54E-03	-2,31E-04	4,69E-03	-3,69E-02	-9,66E-04	-1,35E-03	-1,31E-03	-3,67E-05	-1,87E-04	-5,08E-04	3,31E-04
120	6,60E-01	-9,08E-03	-1,27E-04	4,64E-03	-3,69E-02	-1,26E-03	-1,31E-03	-1,37E-03	5,80E-05	-6,13E-05	-4,49E-04	6,55E-04
240	2,93E-01	-4,58E-03	-9,03E-05	3,55E-03	-3,65E-02	-5,23E-04	-7,82E-04	-6,55E-04	2,58E-05	6,84E-06	-2,50E-04	3,81E-04
360	2,12E-01	-2,98E-03	-6,29E-05	6,16E-04	-3,61E-02	-2,80E-04	-4,39E-04	-4,31E-04	-1,38E-05	-1,31E-05	3,24E-05	1,10E-04
480	-1,65E-01	-5,00E-04	-7,02E-05	-5,37E-04	-3,39E-02	-1,59E-04	-2,94E-04	-2,33E-04	-2,67E-05	-1,07E-06	2,32E-04	1,19E-04
600	2,09E-01	-2,54E-03	-5,24E-05	7,44E-04	-3,41E-02	-3,31E-04	-3,03E-04	-2,94E-04	4,65E-05	-3,40E-06	-7,19E-05	-8,96E-05
720	-1,03E-01	-1,86E-04	-3,79E-05	-2,60E-04	-3,46E-02	-2,45E-05	-7,55E-05	-8,13E-05	-1,21E-05	-9,68E-06	-1,32E-04	4,13E-05
840	-5,22E-02	-2,95E-04	-5,73E-05	1,17E-03	-3,47E-02	-1,18E-04	-1,17E-04	-1,12E-04	-1,39E-05	1,05E-05	7,64E-05	1,13E-04
960	-9,12E-02	-7,41E-04	-2,38E-05	-4,01E-04	-3,47E-02	-1,03E-04	-1,42E-04	-2,22E-04	9,39E-06	3,83E-06	-2,33E-05	2,47E-06
1080	1,47E-01	-1,19E-03	-4,73E-06	1,05E-03	-3,56E-02	-1,02E-04	-1,24E-04	-1,30E-04	-3,78E-05	-1,06E-06	-1,91E-05	-1,55E-05
1200	5,51E-02	-3,62E-04	-2,56E-05	-4,46E-04	-3,47E-02	-5,55E-05	-1,65E-05	-8,28E-05	2,82E-05	-1,17E-05	-1,18E-04	3,21E-05

Tabela 6.3: Resultados da simulação para o modelo normal aleatoriamente truncado misto van Genuchten-Mualem: EQM.

	Parâmetro											
n	β_1	β_2	σ	α	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
24	1,29E+02	5,19E-03	4,74E-06	1,55E-02	5,24E-04	1,36E-03	1,22E-03	1,33E-03	3,30E-06	2,13E-06	1,33E-04	4,52E-05
72	2,37E+01	1,01E-03	1,11E-06	2,12E-03	1,45E-03	2,03E-04	2,13E-04	2,92E-04	1,52E-06	6,73E-07	4,22E-05	1,69E-05
120	1,80E+01	7,14E-04	6,74E-07	1,16E-03	1,44E-03	2,07E-04	2,26E-04	2,55E-04	7,89E-07	4,95E-07	1,64E-05	8,42E-06
240	9,38E+00	4,84E-04	3,56E-07	6,43E-04	1,43E-03	2,41E-04	2,47E-04	2,19E-04	3,61E-07	2,24E-07	1,01E-05	4,21E-06
360	5,18E+00	2,71E-04	2,56E-07	3,52E-04	1,42E-03	2,69E-04	2,95E-04	2,12E-04	2,77E-07	1,50E-07	7,66E-06	3,35E-06
480	3,73E+00	1,70E-04	1,71E-07	2,28E-04	1,28E-03	3,15E-04	3,07E-04	3,48E-04	2,33E-07	1,00E-07	5,19E-06	1,72E-06
600	3,35E+00	1,51E-04	1,41E-07	1,93E-04	1,27E-03	2,55E-04	3,07E-04	2,76E-04	1,51E-07	8,29E-08	4,72E-06	2,00E-06
720	2,75E+00	1,10E-04	1,20E-07	1,89E-04	1,34E-03	2,79E-04	3,02E-04	3,35E-04	1,38E-07	7,11E-08	3,72E-06	1,61E-06
840	2,23E+00	9,01E-05	7,89E-08	1,28E-04	1,32E-03	2,95E-04	2,70E-04	2,44E-04	1,27E-07	6,28E-08	3,13E-06	1,46E-06
960	2,26E+00	9,27E-05	8,74E-08	1,23E-04	1,32E-03	2,66E-04	2,74E-04	2,90E-04	1,15E-07	4,82E-08	2,59E-06	1,35E-06
1080	1,92E+00	8,86E-05	7,26E-08	1,08E-04	1,40E-03	2,64E-04	2,42E-04	3,55E-04	9,88E-08	4,44E-08	2,59E-06	9,35E-07
1200	1,82E+00	7,61E-05	6,65E-08	9,76E-05	1,35E-03	3,34E-04	2,73E-04	3,13E-04	7,94E-08	3,90E-08	1,92E-06	1,09E-06

Tabela 6.4: Resultados da simulação para o modelo normal aleatoriamente truncado misto van Genuchten-Mualem: probabilidade de cobertura estimada (PC), erro da cauda à esquerda (EE) e erro da cauda à direita (ED) dos ICs de Wald de 95% de confiança.

	IC de Wald	Parâmetro											
n	95%	β_1	β_2	σ	α	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
	PC	0,8600	0,8650	0,8950	0,8350	0,6800	0,8650	0,8500	0,8300	0,9550	0,8950	0,9250	0,9450
24	EE	0,0200	0,0550	0,0100	0,1400	0,0000	0,0650	0,0900	0,1050	0,0200	0,0000	0,0200	0,0000
	ED	0,1200	0,0800	0,0950	0,0250	0,3200	0,0700	0,0600	0,0650	0,0250	0,1050	0,0550	0,0550
	PC	0,9450	0,9000	0,9450	0,9350	0,1150	0,8900	0,8700	0,8600	0,9300	0,9350	0,9200	0,9400
72	EE	0,0100	0,0200	0,0000	0,0250	0,0000	0,0600	0,0550	0,0500	0,0250	0,0050	0,0250	0,0150
	ED	0,0450	0,0800	0,0550	0,0400	0,8850	0,0500	0,0750	0,0900	0,0450	0,0600	0,0550	0,0450
	PC	0,9200	0,8750	0,9400	0,9400	0,1000	0,8900	0,9100	0,8750	0,9400	0,9200	0,9700	0,9600
120	EE	0,0300	0,0150	0,0050	0,0450	0,0000	0,0200	0,0150	0,0350	0,0250	0,0100	0,0150	0,0300
	ED	0,0500	0,1100	0,0550	0,0150	0,9000	0,0900	0,0750	0,0900	0,0350	0,0700	0,0150	0,0100
	PC	0,9350	0,8650	0,9100	0,9500	0,1300	0,8950	0,8700	0,8600	0,9750	0,9500	0,9550	0,9600
240	EE	0,0300	0,0150	0,0100	0,0300	0,0000	0,0350	0,0300	0,0500	0,0150	0,0050	0,0200	0,0250
	ED	0,0350	0,1200	0,0800	0,0200	0,8700	0,0700	0,1000	0,0900	0,0100	0,0450	0,0250	0,0150
	PC	0,9500	0,9000	0,9250	0,9400	0,1250	0,9100	0,9200	0,9200	0,9550	0,9250	0,9500	0,9600
360	EE	0,0100	0,0200	0,0100	0,0350	0,0000	0,0250	0,0200	0,0350	0,0150	0,0200	0,0250	0,0300
	ED	0,0400	0,0800	0,0650	0,0250	0,8750	0,0650	0,0600	0,0450	0,0300	0,0550	0,0250	0,0100
	PC	0,9550	0,9450	0,9350	0,9650	0,1800	0,9450	0,9150	0,9000	0,9350	0,9650	0,9550	0,9850
480	EE	0,0050	0,0300	0,0100	0,0100	0,0000	0,0350	0,0350	0,0500	0,0250	0,0100	0,0300	0,0150
	ED	0,0400	0,0250	0,0550	0,0250	0,8200	0,0200	0,0500	0,0500	0,0400	0,0250	0,0150	0,0000
	PC	0,9550	0,9400	0,9300	0,9600	0,1650	0,9350	0,9550	0,9250	0,9650	0,9450	0,9350	0,9650
600	EE	0,0250	0,0150	0,0100	0,0250	0,0000	0,0300	0,0250	0,0300	0,0200	0,0150	0,0200	0,0150
	ED	0,0200	0,0450	0,0600	0,0150	0,8350	0,0350	0,0200	0,0450	0,0150	0,0400	0,0450	0,0200
	PC	0,9400	0,9350	0,9050	0,9550	0,1750	0,9200	0,9500	0,9350	0,9700	0,9500	0,9250	0,9550
720	EE	0,0100	0,0300	0,0300	0,0300	0,0000	0,0450	0,0250	0,0200	0,0250	0,0100	0,0150	0,0200
	ED	0,0500	0,0350	0,0650	0,0150	0,8250	0,0350	0,0250	0,0450	0,0050	0,0400	0,0600	0,0250
	PC	0,9700	0,9400	0,9600	0,9750	0,1550	0,9600	0,9500	0,9350	0,9500	0,9500	0,9450	0,9500
840	EE	0,0150	0,0100	0,0000	0,0200	0,0000	0,0200	0,0150	0,0100	0,0200	0,0300	0,0400	0,0400
	ED	0,0150	0,0500	0,0400	0,0050	0,8450	0,0200	0,0350	0,0550	0,0300	0,0200	0,0150	0,0100
	PC	0,9650	0,9450	0,9450	0,9700	0,1700	0,9550	0,9100	0,9200	0,9400	0,9650	0,9550	0,9400
960	EE	0,0250	0,0100	0,0250	0,0150	0,0000	0,0250	0,0350	0,0050	0,0250	0,0100	0,0250	0,0250
	ED	0,0100	0,0450	0,0300	0,0150	0,8300	0,0200	0,0550	0,0750	0,0350	0,0250	0,0200	0,0350
	PC	0,9500	0,9300	0,9500	0,9600	0,1450	0,9300	0,9000	0,9300	0,9250	0,9550	0,9500	0,9750
1080	EE	0,0150	0,0150	0,0150	0,0200	0,0000	0,0350	0,0250	0,0250	0,0400	0,0200	0,0200	0,0150
	ED	0,0350	0,0550	0,0350	0,0200	0,8550	0,0350	0,0750	0,0450	0,0350	0,0250	0,0300	0,0100
	PC	0,9500	0,9200	0,9450	0,9550	0,1400	0,9350	0,9100	0,9450	0,9550	0,9500	0,9550	0,9550
1200	EE	0,0250	0,0200	0,0150	0,0200	0,0000	0,0350	0,0450	0,0150	0,0250	0,0200	0,0150	0,0250
	ED	0,0250	0,0600	0,0400	0,0250	0,8600	0,0300	0,0450	0,0400	0,0200	0,0300	0,0300	0,0200

Tabela 6.5: Resultados da simulação para o modelo normal aleatoriamente truncado misto van Genuchten-Mualem: probabilidade de cobertura estimada (PC), erro da cauda à esquerda (EE) e erro da cauda à direita (ED) dos IC-RV's de 95% de confiança.

	IC-RV	Parâmetro											
n	95%	β_1	β_2	σ	α	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
	PC	0,5500	0,4250	0,8250	0,6021	0,8250	0,5700	0,5650	0,5900	0,9450	0,9300	0,9250	0,9700
24	EE	0,1750	0,3000	0,0950	0,3298	0,0000	0,2100	0,2050	0,2300	0,0400	0,0100	0,0250	0,0050
	ED	0,2750	0,2750	0,0800	0,0681	0,1750	0,2200	0,2300	0,1800	0,0150	0,0600	0,0500	0,0250
	PC	0,6800	0,4000	0,9100	0,8788	0,2150	0,6300	0,6000	0,6200	0,9500	0,9500	0,9250	0,9400
72	EE	0,1450	0,2100	0,0600	0,0707	0,0000	0,1600	0,1600	0,1400	0,0350	0,0200	0,0400	0,0300
	ED	0,1750	0,3900	0,0300	0,0505	0,7850	0,2100	0,2400	0,2400	0,0150	0,0300	0,0350	0,0300
	PC	0,6450	0,4150	0,8700	0,8800	0,2100	0,6400	0,7100	0,6950	0,9200	0,9450	0,9750	0,9450
120	EE	0,2050	0,1950	0,0900	0,0900	0,0000	0,1250	0,0850	0,1000	0,0650	0,0250	0,0150	0,0500
	ED	0,1500	0,3900	0,0400	0,0300	0,7900	0,2350	0,2050	0,2050	0,0150	0,0300	0,0100	0,0050
	PC	0,6500	0,3150	0,8550	0,8550	0,2300	0,6400	0,6200	0,5900	0,9700	0,9550	0,9600	0,9500
240	EE	0,2000	0,3000	0,1000	0,0950	0,0000	0,1850	0,1550	0,1650	0,0300	0,0200	0,0200	0,0350
	ED	0,1500	0,3850	0,0450	0,0500	0,7700	0,1750	0,2250	0,2450	0,0000	0,0250	0,0200	0,0150
	PC	0,6750	0,3850	0,8350	0,9000	0,2400	0,6500	0,6500	0,6100	0,8350	0,9100	0,9350	0,9550
360	EE	0,1800	0,3100	0,1450	0,0600	0,0000	0,1900	0,1950	0,2100	0,1650	0,0550	0,0450	0,0350
	ED	0,1450	0,3050	0,0200	0,0400	0,7600	0,1600	0,1550	0,1800	0,0000	0,0350	0,0200	0,0100
	PC	0,6850	0,4400	0,8550	0,8950	0,3250	0,6450	0,6650	0,6700	0,7350	0,9200	0,9500	0,9750
480	EE	0,1250	0,3250	0,1350	0,0450	0,0000	0,2150	0,2050	0,1950	0,2650	0,0600	0,0400	0,0250
	ED	0,1900	0,2350	0,0100	0,0600	0,6750	0,1400	0,1300	0,1350	0,0000	0,0200	0,0100	0,0000
	PC	0,6250	0,3350	0,8050	0,9100	0,3250	0,6700	0,6800	0,6050	0,6850	0,9000	0,9600	0,9650
600	EE	0,2050	0,3000	0,1900	0,0500	0,0000	0,2100	0,1800	0,2400	0,3100	0,0850	0,0250	0,0300
	ED	0,1700	0,3650	0,0050	0,0400	0,6750	0,1200	0,1400	0,1550	0,0050	0,0150	0,0150	0,0050
	PC	0,6750	0,3750	0,7750	0,8550	0,2800	0,6350	0,6900	0,6250	0,7550	0,8750	0,9050	0,9650
720	EE	0,1500	0,3150	0,2250	0,0700	0,0050	0,2500	0,2400	0,2650	0,2450	0,1050	0,0850	0,0300
	ED	0,1750	0,3100	0,0000	0,0750	0,7150	0,1150	0,0700	0,1100	0,0000	0,0200	0,0100	0,0050
	PC	0,6750	0,4600	0,8103	0,9150	0,3000	0,6650	0,6950	0,6400	0,7850	0,8500	0,8950	0,9400
840	EE	0,1500	0,2900	0,1897	0,0500	0,0000	0,2650	0,2100	0,2650	0,1850	0,1350	0,1000	0,0600
	ED	0,1750	0,2500	0,0000	0,0350	0,7000	0,0700	0,0950	0,0950	0,0300	0,0150	0,0050	0,0000
	PC	0,6100	0,4100	0,7796	0,8750	0,2650	0,6350	0,6150	0,6350	0,7550	0,8900	0,9300	0,9650
960	EE	0,1850	0,2850	0,2204	0,0700	0,0000	0,2650	0,2600	0,2250	0,2300	0,0900	0,0550	0,0300
	ED	0,2050	0,3050	0,0000	0,0550	0,7350	0,1000	0,1250	0,1400	0,0150	0,0200	0,0150	0,0050
	PC	0,6300	0,3950	0,7886	0,9250	0,2350	0,6500	0,5850	0,6250	0,8200	0,8400	0,9350	0,9550
1080	EE	0,2400	0,2850	0,2114	0,0450	0,0050	0,2600	0,2800	0,2500	0,1600	0,1450	0,0500	0,0350
	ED	0,1300	0,3200	0,0000	0,0300	0,7600	0,0900	0,1350	0,1250	0,0200	0,0150	0,0150	0,0100
	PC	0,6400	0,4100	0,8314	0,8900	0,2750	0,6050	0,5950	0,5850	0,8100	0,8300	0,9550	0,9500
1200	EE	0,1900	0,3250	0,1686	0,0550	0,0000	0,2850	0,3050	0,2900	0,1850	0,1450	0,0350	0,0450
	ED	0,1700	0,2650	0,0000	0,0550	0,7250	0,1100	0,1000	0,1250	0,0050	0,0250	0,0100	0,0050

Tabela 6.6: Resultados da simulação para o modelo normal aleatoriamente truncado misto van Genuchten-Mualem.

		Valor	EMV			IC de Wald 95%			IC-RV 95%		
n	Parâmetro	real	(média)	Vício	EQM	PC	EE	ED	PC	EE	ED
	β_1	55,00	54,8966	-1,03E-01	2,75E+00	0,9400	0,0100	0,0500	0,6750	0,1500	0,1750
	β_2	1,45	1,4498	-1,86E-04	1,10E-04	0,9350	0,0300	0,0350	0,3750	0,3150	0,3100
	σ	0,01	0,0100	-3,79E-05	1,20E-07	0,9050	0,0300	0,0650	0,7750	0,2250	0,0000
	α	-0,15	-0,1503	-2,60E-04	1,89E-04	0,9550	0,0300	0,0150	0,8550	0,0700	0,0750
	σ_u	0,05	0,0154	-3,46E-02	1,34E-03	0,1750	0,0000	0,8250	0,2800	0,0050	0,7150
720	u_1	-	0,0079	-2,45E-05	2,79E-04	0,9200	0,0450	0,0350	0,6350	0,2500	0,1150
	u_2	-	0,0092	-7,55E-05	3,02E-04	0,9500	0,0250	0,0250	0,6900	0,2400	0,0700
	u_3	-	0,0099	-8,13E-05	3,35E-04	0,9350	0,0200	0,0450	0,6250	0,2650	0,1100
	μ_A	0,25	0,2500	-1,21E-05	1,38E-07	0,9700	0,0250	0,0050	0,7550	0,2450	0,0000
	σ_A	0,01	0,0100	-9,68E-06	7,11E-08	0,9500	0,0100	0,0400	0,8750	0,1050	0,0200
	μ_B	0,50	0,4999	-1,32E-04	3,72E-06	0,9250	0,0150	0,0600	0,9050	0,0850	0,0100
	σ_B	0,05	0,0500	4,13E-05	1,61E-06	0,9550	0,0200	0,0250	0,9650	0,0300	0,0050

6.2 Resultados de simulação para o modelo de regressão não linear beta aleatoriamente truncado misto

Para o modelo de regressão não linear beta aleatoriamente truncado misto, a v.a. de truncamento superior é obtida de $B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}) \sim Beta\,(\mu_B, \sigma_B)$; em seguida, a v.a. de truncamento inferior é gerada como $A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i) \sim BT\,(\mu_A, \sigma_A, 0, b_{i,jk})$.

Uma vez simulados os limites de truncamento aleatórios, a variável resposta aleatoriamente truncada é dada por $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i) \sim BT\,(\eta\,(x_{i,jk}, \boldsymbol{\beta}) + u_i, \sigma, a_{i,jk}, b_{i,jk})$, com $u_i \sim N\,(0, \sigma_u)$, $i = 1, 2, 3$, $j = 1, \ldots, N_i$ e $k = 1, \ldots, n_{ij}$. A função $\eta\,(x_{i,jk}, \boldsymbol{\beta})$ é dada pela expressão de Gardner (1958) em (1.1).

Os valores assumidos para os parâmetros foram: $\beta_1 = 2,25$ e $\beta_2 = 0,5$, $\sigma = 6$, $\sigma_u = 0,1$, $\mu_A = 0,25$, $\sigma_A = 5,2$, $\mu_B = 0,6$ e $\sigma_B = 4,6$.

Os resultados obtidos no estudo de simulação são mostrados nas Tabelas 6.7-6.7. A Tabela 6.7 fornece a média dos EMVs obtidos para os conjuntos de dados simulados para cada tamanho amostral. Ainda na Tabela 6.7, são mostrados os verdadeiros valores dos parâmetros fixos, não sendo possível mostrar os valores reais dos efeitos aleatórios u_i, $i = 1, 2, 3$, que mudam para cada conjunto simulado. Na Tabela 6.8 temos o vício dos EMVs e na Tabela 6.9 o erro quadrático médio (EQM) dos EMVs. A Tabela 6.10 apresenta a probabilidade de cobertura estimada de ICs de Wald de 95% de confiança, e os erros da cauda à esquerda e à direita dos ICs de Wald. Na Tabela 6.11, são mostrados a probabilidade de cobertura estimada de intervalos IC-RV de 95% e os erros da cauda à esquerda e à direita dos IC-RV's.

Com relação aos parâmetros fixos da expressão não linear $\eta\,(x_{i,jk}, \boldsymbol{\beta})$, β_1 e β_2, relacionados aos parâmetros de média $\mu_{i,jk}$ das respostas, os resultados das Tabelas 6.7-6.11 indicam que os EMVs destes parâmetros apresentam boas propriedades frequentistas. Observa-se, ainda, que, à medida que o tamanho amostral aumenta, tanto o vício (Tabela 6.8) como o EQM (Tabela 6.8) dos EMVs são razoavelmente baixos e que a probabilidade de cobertura estimada aproxima-se da nominal esperada de 95% ((Tabelas 6.10) e 6.11). Essas mesmas observações podem ser feitas para o parâmetro σ relacionado à dispersão das respostas. Por outro lado, há indicativos de problemas de estimação do parâmetro σ_u relacionado aos efeitos aleatórios não observáveis, e a média dos EMVs (Tabela 6.7) obtidos para este parâmetro das amostras simuladas parece indicar que o mesmo é subestimado e, como sua variância estimada também é pequena, seus intervalos assintóticos de 95% de confiança, sistematicamente, não contêm o verdadeiro valor do parâmetro, fazendo com que a sua probabilidade de cobertura seja excessivamente baixa (Tabelas 6.10 e 6.11). Não obstante, as probabilidades de cobertura estimadas dos efeitos aleatórios não observáveis u_1, u_2 e u_3 aproximam-se da nominal esperada de 95%.

Com respeito aos parâmetros μ_A, σ_A, μ_B e σ_B, tanto os vícios (Tabela 6.8) quanto os

EQMs (Tabela 6.9) observados são pequenos e os EMVs (Tabela 6.7) estimam os verdadeiros valores dos parâmetros com precisão. Além disso, as probabilidades de cobertura estimadas aproximam-se da nominal esperada de 95% para os ICs de Wald e os IC-RV's (Tabelas 6.10 e 6.11).

Assim como no caso do modelo de regressão não linear normal aleatoriamente truncado misto, os resultados de simulação obtidos para o caso do modelo de regressão não linear beta aleatoriamente truncado misto parecem indicar que a partir do tamanho amostral $n = 720$ os EMVs dos parâmetros do modelo em questão passam a ter boas propriedades frequentistas, isto é, suas estimativas tornam-se satisfatoriamente precisas e tanto o vício quanto o EQM tornam-se razoavelmente pequenos. Assim, na Table 6.12 apresentamos os resultados da simulação para o tamanho amostral $n = 720$.

Tabela 6.7: Resultados da simulação para o modelo beta aleatoriamente truncado misto Gardner: valor médio dos EMVs.

	Parâmetro										
	β_1	β_2	σ	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
n	2,25	0,50	6,00	0,10	-	-	-	0,25	5,20	0,60	4,60
24	2,3966	0,5097	6,2226	0,0433	0,0039	0,0045	0,0002	0,2500	5,2821	0,6012	4,6900
72	2,2815	0,4990	6,1002	0,0404	0,0000	0,0009	-0,0025	0,2502	5,2037	0,6008	4,6283
120	2,3213	0,5048	6,0568	0,0413	0,0031	0,0005	0,0034	0,2501	5,2026	0,6006	4,6117
240	2,2708	0,5026	6,0339	0,0409	0,0018	-0,0011	0,0102	0,2498	5,1940	0,6000	4,6035
360	2,2580	0,5008	6,0128	0,0399	0,0034	0,0005	0,0028	0,2500	5,2077	0,6000	4,6037
480	2,2542	0,5003	6,0157	0,0393	0,0059	0,0010	0,0039	0,2499	5,2044	0,5999	4,6006
600	2,2478	0,4992	6,0148	0,0400	0,0019	0,0083	0,0010	0,2500	5,2039	0,5998	4,6087
720	2,2506	0,5010	6,0092	0,0401	0,0041	0,0048	0,0042	0,2499	5,2095	0,6000	4,6108
840	2,2457	0,4993	6,0101	0,0376	0,0064	-0,0007	0,0037	0,2499	5,1993	0,6000	4,6072
960	2,2406	0,4988	6,0141	0,0402	0,0068	0,0045	-0,0019	0,2499	5,1989	0,6000	4,6027
1080	2,2502	0,4998	6,0033	0,0394	0,0033	0,0137	-0,0031	0,2500	5,2048	0,6001	4,6063
1200	2,2455	0,4997	6,0020	0,0408	0,0037	0,0044	0,0050	0,2501	5,2061	0,6000	4,6023

Tabela 6.8: Resultados da simulação para o modelo beta aleatoriamente truncado misto Gardner: vício.

	Parâmetro										
n	β_1	β_2	σ	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
24	1,47E-01	9,69E-03	2,23E-01	-5,67E-02	-1,81E-04	4,89E-05	-5,24E-05	3,84E-05	8,21E-02	1,15E-03	9,00E-02
72	3,15E-02	-1,01E-03	1,00E-01	-5,96E-02	-5,92E-04	-2,12E-04	1,18E-04	2,03E-04	3,69E-03	8,47E-04	2,83E-02
120	7,13E-02	4,82E-03	5,68E-02	-5,87E-02	1,06E-03	1,43E-03	1,07E-03	1,37E-04	2,61E-03	5,53E-04	1,17E-02
240	2,08E-02	2,62E-03	3,39E-02	-5,91E-02	3,97E-04	-1,10E-04	7,87E-05	-1,91E-04	-5,96E-03	2,53E-05	3,46E-03
360	7,98E-03	7,90E-04	1,28E-02	-6,01E-02	5,45E-05	1,57E-04	-1,29E-04	-3,40E-06	7,69E-03	-4,27E-05	3,73E-03
480	4,17E-03	3,28E-04	1,57E-02	-6,07E-02	-3,43E-04	-3,18E-04	-1,16E-04	-6,10E-05	4,42E-03	-6,77E-05	6,30E-04
600	-2,21E-03	-7,57E-04	1,48E-02	-6,00E-02	3,15E-05	-4,27E-04	-1,64E-04	-1,55E-05	3,87E-03	-2,46E-04	8,70E-03
720	5,63E-04	9,79E-04	9,21E-03	-5,99E-02	-2,17E-04	-1,19E-04	-2,31E-04	-7,57E-05	9,53E-03	-1,48E-05	1,08E-02
840	-4,28E-03	-7,14E-04	1,01E-02	-6,24E-02	-1,58E-04	8,18E-06	-2,90E-04	-8,44E-05	-6,86E-04	3,50E-05	7,17E-03
960	-9,35E-03	-1,16E-03	1,41E-02	-5,98E-02	-2,04E-04	-1,17E-04	-2,30E-04	-1,49E-04	-1,12E-03	-1,70E-05	2,67E-03
1080	1,70E-04	-1,60E-04	3,29E-03	-6,06E-02	-1,49E-04	-7,32E-05	-1,80E-05	1,79E-05	4,78E-03	7,19E-05	6,31E-03
1200	-4,54E-03	-3,32E-04	1,99E-03	-5,92E-02	-1,90E-04	-3,37E-04	-1,04E-04	5,97E-05	6,08E-03	-7,48E-06	2,32E-03

Tabela 6.9: Resultados da simulação para o modelo beta aleatoriamente truncado misto Gardner: EQM.

	Parâmetro										
n	β_1	β_2	σ	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
24	6,04E-01	2,39E-03	1,60E-01	3,44E-03	2,18E-03	2,19E-03	1,89E-03	4,84E-05	8,51E-02	1,02E-04	9,20E-02
72	1,77E-01	6,17E-04	4,49E-02	3,78E-03	2,20E-03	1,54E-03	1,83E-03	1,33E-05	2,82E-02	3,76E-05	2,61E-02
120	1,22E-01	4,85E-04	2,54E-02	3,65E-03	1,78E-03	1,96E-03	2,01E-03	8,41E-06	1,44E-02	1,91E-05	1,61E-02
240	4,37E-02	2,12E-04	1,07E-02	3,66E-03	1,73E-03	1,67E-03	2,04E-03	3,99E-06	8,19E-03	9,67E-06	8,85E-03
360	2,77E-02	1,04E-04	6,64E-03	3,76E-03	1,69E-03	1,64E-03	1,93E-03	2,70E-06	6,11E-03	5,44E-06	5,56E-03
480	2,79E-02	1,24E-04	5,37E-03	3,87E-03	1,65E-03	1,65E-03	1,89E-03	2,25E-06	3,75E-03	5,16E-06	3,49E-03
600	1,68E-02	7,25E-05	4,25E-03	3,77E-03	1,68E-03	1,78E-03	1,80E-03	1,82E-06	3,98E-03	3,83E-06	4,06E-03
720	1,63E-02	8,38E-05	3,28E-03	3,78E-03	1,71E-03	1,77E-03	1,87E-03	1,61E-06	2,93E-03	3,67E-06	2,53E-03
840	1,30E-02	6,34E-05	3,11E-03	4,07E-03	1,60E-03	1,39E-03	1,76E-03	1,44E-06	2,82E-03	3,18E-06	2,48E-03
960	1,16E-02	4,81E-05	2,69E-03	3,77E-03	1,87E-03	1,96E-03	1,54E-03	1,09E-06	2,21E-03	2,67E-06	1,96E-03
1080	1,12E-02	5,12E-05	2,01E-03	3,86E-03	1,97E-03	1,60E-03	1,48E-03	1,07E-06	2,00E-03	2,00E-06	1,66E-03
1200	9,03E-03	3,86E-05	1,66E-03	3,70E-03	1,93E-03	1,76E-03	1,85E-03	7,35E-07	1,77E-03	2,20E-06	2,00E-03

Tabela 6.10: Resultados da simulação para o modelo beta aleatoriamente truncado misto Gardner: probabilidade de cobertura estimada (PC), erro da cauda à esquerda (EE) e erro da cauda à direita (ED) dos ICs de Wald de 95% de confiança.

	IC de Wald	Parâmetro										
n	95%	β_1	β_2	σ	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
	PC	0,8650	0,9100	0,8550	0,2400	0,8550	0,8450	0,8100	0,9300	0,9250	0,9500	0,9350
24	EE	0,0350	0,0300	0,1450	0,0000	0,0750	0,0650	0,1000	0,0350	0,0650	0,0450	0,0650
	ED	0,1000	0,0600	0,0000	0,7600	0,0700	0,0900	0,0900	0,0350	0,0100	0,0050	0,0000
	PC	0,8850	0,9600	0,8750	0,1500	0,9000	0,9150	0,8950	0,9500	0,9550	0,9300	0,9500
72	EE	0,0300	0,0050	0,1200	0,0000	0,0450	0,0450	0,0650	0,0150	0,0300	0,0550	0,0400
	ED	0,0850	0,0350	0,0050	0,8500	0,0550	0,0400	0,0400	0,0350	0,0150	0,0150	0,0100
	PC	0,9250	0,9250	0,8950	0,1400	0,9050	0,8900	0,9200	0,9650	0,9500	0,9500	0,9450
120	EE	0,0450	0,0400	0,0900	0,0000	0,0700	0,0750	0,0500	0,0200	0,0350	0,0300	0,0300
	ED	0,0300	0,0350	0,0150	0,8600	0,0250	0,0350	0,0300	0,0150	0,0150	0,0200	0,0250
	PC	0,9450	0,9450	0,9400	0,1350	0,9250	0,9300	0,9200	0,9700	0,9400	0,9500	0,9500
240	EE	0,0150	0,0300	0,0450	0,0000	0,0250	0,0250	0,0400	0,0150	0,0400	0,0300	0,0400
	ED	0,0400	0,0250	0,0150	0,8650	0,0500	0,0450	0,0400	0,0150	0,0200	0,0200	0,0100
	PC	0,9600	0,9750	0,9550	0,1200	0,9550	0,9350	0,9400	0,9600	0,9450	0,9600	0,9550
360	EE	0,0100	0,0050	0,0350	0,0000	0,0200	0,0300	0,0200	0,0200	0,0350	0,0150	0,0300
	ED	0,0300	0,0200	0,0100	0,8800	0,0250	0,0350	0,0400	0,0200	0,0200	0,0250	0,0150
	PC	0,9100	0,9350	0,9450	0,1350	0,9250	0,9300	0,9550	0,9350	0,9500	0,9500	0,9500
480	EE	0,0350	0,0200	0,0400	0,0000	0,0300	0,0350	0,0250	0,0200	0,0250	0,0200	0,0250
	ED	0,0550	0,0450	0,0150	0,8650	0,0450	0,0350	0,0200	0,0450	0,0250	0,0300	0,0250
	PC	0,9300	0,9550	0,9300	0,1100	0,9650	0,9400	0,9400	0,9400	0,9400	0,9450	0,9250
600	EE	0,0150	0,0050	0,0550	0,0000	0,0200	0,0200	0,0350	0,0300	0,0400	0,0200	0,0600
	ED	0,0550	0,0400	0,0150	0,8900	0,0150	0,0400	0,0250	0,0300	0,0200	0,0350	0,0150
	PC	0,9450	0,9250	0,9150	0,1400	0,9450	0,9600	0,9450	0,9350	0,9350	0,9250	0,9700
720	EE	0,0200	0,0400	0,0550	0,0000	0,0350	0,0300	0,0350	0,0250	0,0500	0,0300	0,0300
	ED	0,0350	0,0350	0,0300	0,8600	0,0200	0,0100	0,0200	0,0400	0,0150	0,0450	0,0000
	PC	0,9400	0,9450	0,9200	0,0950	0,9450	0,9250	0,9400	0,9300	0,9300	0,9450	0,9400
840	EE	0,0150	0,0200	0,0450	0,0000	0,0200	0,0300	0,0250	0,0250	0,0400	0,0350	0,0350
	ED	0,0450	0,0350	0,0350	0,9050	0,0350	0,0450	0,0350	0,0450	0,0300	0,0200	0,0250
	PC	0,9450	0,9550	0,9050	0,1250	0,9300	0,9400	0,9550	0,9650	0,9400	0,9450	0,9600
960	EE	0,0150	0,0200	0,0800	0,0000	0,0400	0,0300	0,0150	0,0100	0,0400	0,0400	0,0150
	ED	0,0400	0,0250	0,0150	0,8750	0,0300	0,0300	0,0300	0,0250	0,0200	0,0150	0,0250
	PC	0,9200	0,9300	0,9600	0,1250	0,9300	0,9400	0,9500	0,9500	0,9500	0,9600	0,9500
1080	EE	0,0300	0,0200	0,0250	0,0000	0,0200	0,0250	0,0100	0,0350	0,0350	0,0250	0,0400
	ED	0,0500	0,0500	0,0150	0,8750	0,0500	0,0350	0,0400	0,0150	0,0150	0,0150	0,0100
	PC	0,9300	0,9600	0,9650	0,1400	0,9250	0,9350	0,9400	0,9750	0,9250	0,9400	0,9050
1200	EE	0,0350	0,0000	0,0300	0,0000	0,0400	0,0400	0,0250	0,0250	0,0550	0,0350	0,0700
	ED	0,0350	0,0400	0,0050	0,8600	0,0350	0,0250	0,0350	0,0000	0,0200	0,0250	0,0250

Tabela 6.11: Resultados da simulação para o modelo beta aleatoriamente truncado misto Gardner: probabilidade de cobertura estimada (PC), erro da cauda à esquerda (EE) e erro da cauda à direita (ED) dos IC-RV's de 95% de confiança.

	IC-RV	Parâmetro										
n	95%	β_1	β_2	σ	σ_u	u_1	u_2	u_3	μ_A	σ_A	μ_B	σ_B
	PC	0,3150	0,6950	0,8800	0,7400	0,5250	0,5500	0,5000	0,9250	0,9450	0,9400	0,9450
24	EE	0,3500	0,2150	0,1200	0,0000	0,2500	0,2350	0,2450	0,0400	0,0400	0,0550	0,0450
	ED	0,3350	0,0900	0,0000	0,2600	0,2250	0,2150	0,2550	0,0350	0,0150	0,0050	0,0100
	PC	0,3050	0,8150	0,8850	0,6950	0,5550	0,5200	0,5600	0,9350	0,9600	0,9050	0,9650
72	EE	0,3300	0,0800	0,1100	0,0000	0,2250	0,2550	0,2100	0,0400	0,0200	0,0850	0,0250
	ED	0,3650	0,1050	0,0050	0,3050	0,2200	0,2250	0,2300	0,0250	0,0200	0,0100	0,0100
	PC	0,3200	0,7850	0,8950	0,7850	0,5450	0,5250	0,5600	0,9350	0,9500	0,9450	0,9500
120	EE	0,3800	0,1500	0,0900	0,0000	0,2850	0,3150	0,2700	0,0550	0,0350	0,0400	0,0250
	ED	0,3000	0,0650	0,0150	0,2150	0,1700	0,1600	0,1700	0,0100	0,0150	0,0150	0,0250
	PC	0,3800	0,8050	0,9400	0,7250	0,6050	0,5400	0,6050	0,9700	0,9400	0,9350	0,9600
240	EE	0,3450	0,1450	0,0450	0,0000	0,2600	0,2650	0,2300	0,0250	0,0400	0,0650	0,0350
	ED	0,2750	0,0500	0,0150	0,2750	0,1350	0,1950	0,1650	0,0050	0,0200	0,0000	0,0050
	PC	0,3600	0,8400	0,9500	0,7050	0,5850	0,6150	0,6150	0,9550	0,9450	0,9550	0,9550
360	EE	0,3350	0,1150	0,0400	0,0000	0,2450	0,2300	0,2200	0,0450	0,0350	0,0400	0,0300
	ED	0,3050	0,0450	0,0100	0,2950	0,1700	0,1550	0,1650	0,0000	0,0200	0,0050	0,0150
	PC	0,3350	0,7600	0,9450	0,6500	0,6000	0,5350	0,5600	0,9300	0,9450	0,9450	0,9500
480	EE	0,3350	0,1450	0,0400	0,0000	0,2200	0,2250	0,2400	0,0650	0,0250	0,0250	0,0250
	ED	0,3300	0,0950	0,0150	0,3500	0,1800	0,2400	0,2000	0,0050	0,0300	0,0300	0,0250
	PC	0,4300	0,8200	0,9150	0,7200	0,5800	0,6600	0,6650	0,8700	0,9450	0,9400	0,9250
600	EE	0,2950	0,1000	0,0700	0,0000	0,2450	0,1850	0,1800	0,1250	0,0350	0,0300	0,0600
	ED	0,2750	0,0800	0,0150	0,2800	0,1750	0,1550	0,1550	0,0050	0,0200	0,0300	0,0150
	PC	0,3500	0,7450	0,9100	0,7100	0,5950	0,5600	0,6250	0,8400	0,9450	0,9200	0,9700
720	EE	0,3150	0,1800	0,0550	0,0000	0,2050	0,2500	0,2150	0,1550	0,0500	0,0500	0,0300
	ED	0,3350	0,0750	0,0350	0,2900	0,2000	0,1900	0,1600	0,0050	0,0050	0,0300	0,0000
	PC	0,3700	0,7900	0,9200	0,6300	0,6400	0,6050	0,5550	0,8200	0,9350	0,9050	0,9400
840	EE	0,3300	0,1250	0,0600	0,0000	0,2200	0,2400	0,2400	0,1700	0,0450	0,0800	0,0450
	ED	0,3000	0,0850	0,0200	0,3700	0,1400	0,1550	0,2050	0,0100	0,0200	0,0150	0,0150
	PC	0,3500	0,8500	0,9000	0,7000	0,5650	0,5750	0,5850	0,7900	0,9500	0,9200	0,9600
960	EE	0,2900	0,0700	0,0850	0,0000	0,2400	0,2550	0,2350	0,2050	0,0400	0,0800	0,0150
	ED	0,3600	0,0800	0,0150	0,3000	0,1950	0,1700	0,1800	0,0050	0,0100	0,0000	0,0250
	PC	0,3350	0,8150	0,9250	0,6550	0,5900	0,5600	0,5950	0,7200	0,9250	0,8800	0,9450
1080	EE	0,3350	0,1100	0,0600	0,0000	0,2350	0,2500	0,2750	0,2800	0,0600	0,1100	0,0500
	ED	0,3300	0,0750	0,0150	0,3450	0,1750	0,1900	0,1300	0,0000	0,0150	0,0100	0,0050
	PC	0,3900	0,8450	0,9650	0,6900	0,5650	0,6650	0,6400	0,6700	0,9250	0,8650	0,9050
1200	EE	0,2900	0,1000	0,0300	0,0000	0,2250	0,1850	0,2300	0,3300	0,0600	0,1350	0,0700
	ED	0,3200	0,0550	0,0050	0,3100	0,2100	0,1500	0,1300	0,0000	0,0150	0,0000	0,0250

Tabela 6.12: Resultados da simulação para o modelo beta aleatoriamente truncado misto Gardner.

		Valor	EMV			IC de Wald 95%			IC-RV 95%		
n	Parâmetro	real	(média)	Vício	EQM	PC	EE	ED	PC	EE	ED
	β_1	2,25	2,2506	5,63E-04	1,63E-02	0,9450	0,0200	0,0350	0,3500	0,3150	0,3350
	β_2	0,50	0,5010	9,79E-04	8,38E-05	0,9250	0,0400	0,0350	0,7450	0,1800	0,0750
	σ	6,00	6,0092	9,21E-03	3,28E-03	0,9150	0,0550	0,0300	0,9100	0,0550	0,0350
	σ_u	0,10	0,0401	-5,99E-02	3,78E-03	0,1400	0,0000	0,8600	0,7100	0,0000	0,2900
720	u_1	-	0,0041	-2,17E-04	1,71E-03	0,9450	0,0350	0,0200	0,5950	0,2050	0,2000
	u_2	-	0,0048	-1,19E-04	1,77E-03	0,9600	0,0300	0,0100	0,5600	0,2500	0,1900
	u_3	-	0,0042	-2,31E-04	1,87E-03	0,9450	0,0350	0,0200	0,6250	0,2150	0,1600
	μ_A	0,25	0,2499	-7,57E-05	1,61E-06	0,9350	0,0250	0,0400	0,8400	0,1550	0,0050
	σ_A	5,20	5,2095	9,53E-03	2,93E-03	0,9350	0,0500	0,0150	0,9450	0,0500	0,0050
	μ_B	0,60	0,6000	-1,48E-05	3,67E-06	0,9250	0,0300	0,0450	0,9200	0,0500	0,0300
	σ_B	4,60	4,6108	1,08E-02	2,53E-03	0,9700	0,0300	0,0000	0,9700	0,0300	0,0000

6.3 Diagnóstico aplicado a dados simulados

Para ilustrar o estudo de diagnóstico do modelo de regressão não linear truncado misto (3.9), tomamos um conjunto de dados simulado e calculamos os resíduos e as métricas descritas na Seção 4.2. A ideia principal deste procedimento é avaliar a eficácia das diferentes métricas de diagnóstico de modelos consideradas e ilustrar a sua capacidade de detectar observação *outliers* e influentes quando o modelo de regressão não linear truncado misto é ajustado a um conjunto de dados.

O conjunto de dados simulado usado para ilustrar as métricas de diagnóstico é gerado assumindo que a variável resposta truncada segue distribuição beta truncada (2.11) e que os limites de truncamento são fixos e conhecidos e, portanto, consideramos o modelo de regressão não linear beta truncado misto (3.22). Isto é, um conjunto de observações de uma variável resposta com limites de truncamento fixos e conhecidos foi gerado assumindo $Y_{i,jk}\,|(a < Y_{i,jk} < b, u_i) \sim BT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, a, b\right)$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, $M = 3$, $N_i = 8$ e $n_{ij} = 8$, com $a = 0,31$, $b = 0,55$ e $\mu_{i,jk} = \eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i$, sendo $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ a expressão de Gardner (1958) dada em (1.1) e $u_i \sim N\left(0, \sigma_u\right)$, $i = 1, 2, 3$. Os valores dos parâmetros usados na simulação são $\beta_1 = 2,25$ e $\beta_2 = 0,5$, e o parâmetro relacionado à dispersão das respostas foi $\sigma = 6$.

Após gerar os valores da variável resposta, a observação $2, 11$ foi deliberadamente perturbada fazendo-se $y_{2,11} = y_{2,11} + 2,5sd(\boldsymbol{y})$. Além disso, a observação $1, 15$ também foi perturbada pela modificação dos seus limites inferiores e superiores de truncamento. Com isso, pretendemos verificar se o registro equivocado de limites de truncamento podem erroneamente indicar observações não discrepantes da variável resposta como sendo discrepantes. Para tal, os valores dos limites de truncamento fixos e conhecidos da observação $1, 15$ foram alterados de $a_{1,15} = a = 0,31$ e $b_{1,15} = b = 0,55$ para $a_{1,15} = 0,06$ e $b_{1,15} = 0,07$.

O resumo do ajuste do modelo é apresentado na Tabela 6.13. Devido à presença das observações perturbadas, notamos que os parâmetros do modelo não são estimados com precisão. As estimativas dos parâmetros β_1, β_2 e σ estão longe de seus valores reais. Além disso, os ICs de 95% de β_2 e σ não contêm os seus verdadeiros valores. Embora os IC de 95% de β_1 contenham seus valores verdadeiros, seu EMV não é preciso.

Os resíduos padronizados (4.11) são mostrados na Figura 6.1a, onde é possível notar que a observação $2, 11$ é corretamente identificada como *outliers*. O resíduo padronizado da observação $1, 15$ está entre os limites de ± 3 desvios, mas é possível notar que a alteração dos seus limites de truncamento faz com que o seu valor predito seja discrepante (Figura 6.1b). As duas métricas de influência global, a distância generalizada de Cook (4.12) e a distância da verossimilhança (4.13), indicam os casos $1, 15$ e $2, 11$ como influentes. As medidas de influência local consideradas são mostradas na Figura 6.3. Por fim, a métrica de influência local sob o esquema de perturbação da resposta (Figura 6.3a) indica

os casos $1, 15$ e $2, 11$ como tendo uma influência mais pronunciada nas estimativas dos parâmetros do modelo. Já a métrica de influência de local sob o esquema de perturbação de casos (Figura 6.3b) não indica nenhum caso como influente.

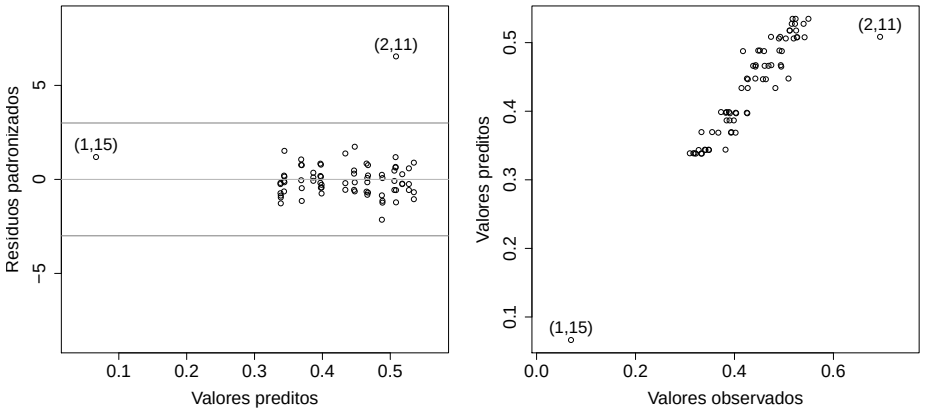
Para verificar a sensibilidade dos EMVs à presença de observações influentes no conjunto de dados, ajustamos o modelo aos dados sem as observações identificadas como influentes. Os EMVs são apresentados na Tabela 6.14, onde pode ser observado que a remoção dos casos influentes do conjunto de dados faz com que os EMVs estejam mais próximos dos valores reais simulados. Além disso, para todos os parâmetros, os valores reais simulados estão contidos no IC de 95%. A Tabela 6.14 também apresenta a diferença relativa entre os EMVs obtidos utilizando os dados perturbados completos e os EMVs obtidos utilizando os dados simulados perturbado com as observações $1, 15$ e $2, 11$ excluídas. Podemos notar que as estimativas dos parâmetros são altamente sensíveis à presença de observações influentes no conjunto de dados, sendo que a maior diferença relativa foi observada para o parâmetro β_0, seguido pelo parâmetro β_2. A menor diferença relativa é observada para o parâmetro β_1. As diferenças relativas (em %) foram calculadas de $|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(-i)}|/\max\{|\hat{\boldsymbol{\theta}}|, |\hat{\boldsymbol{\theta}}_{(-i)}|\}100\%$.

Tabela 6.13: Resumo do modelo beta truncado misto Gardner ajustado aos dados simulados perturbados.

Parâmetro	Valor real	EMV	Desv. pad.	IC de Wal 95%	IC-RV 95%
β_1	2,25	3,4516	1,3231	(0,8584; 6,0449)	(2,8017; 4,2420)
β_2	0,5	0,6813	0,0884	(0,5080; 0,8547)	(0,5988; 0,7764)
σ	6	5,2662	0,2063	(4,8618; 5,6706)	(4,8782; 5,6040)
σ_u	0,1	0,0371	0,0174	(0,0030; 0,0713)	(0,0197; 0,1097)
u_1	0,0517	0,0641	0,0161	(0,0325; 0,0957)	(0,0473; 0,0813)
u_2	-0,0178	-0,0033	0,0154	(-0,0334; 0,0269)	(-0,0190; 0,0124)
u_3	0,0004	-0,0043	0,0154	(-0,0344; 0,0258)	(-0,0200; 0,0114)

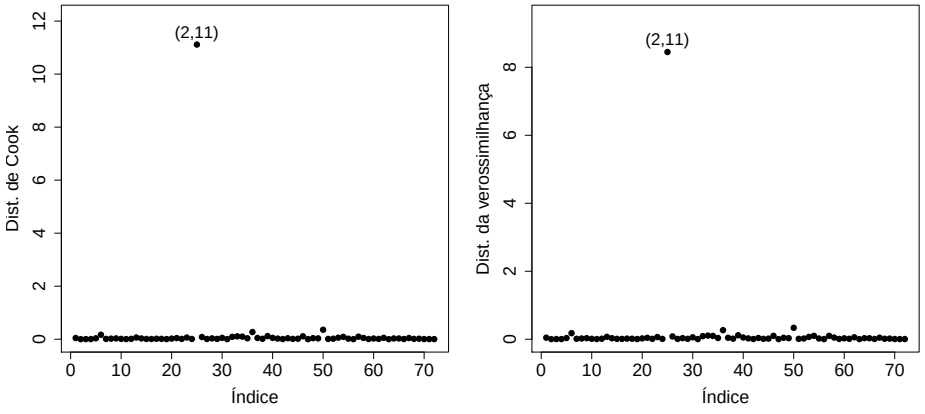
Tabela 6.14: Diferenças relativas (dados simulados perturbados - modelo beta truncado misto Gardner).

Parâmetro	β_1	β_2	σ	σ_u	u_1	u_2	u_3
Valor real	2,25	0,50	6,00	0,10	0,0004	-0,1274	-0,0202
EMV (dados completos)	3,4516	0,6813	5,2662	0,0371	0,0641	-0,0033	-0,0043
EMV (dados com os casos $1, 15$ e $2, 11$ deletados)	2,1098	0,5243	6,0970	0,0299	0,0457	-0,0208	-0,0125
Diferença relativa (%)	39	23	14	20	29	84	66



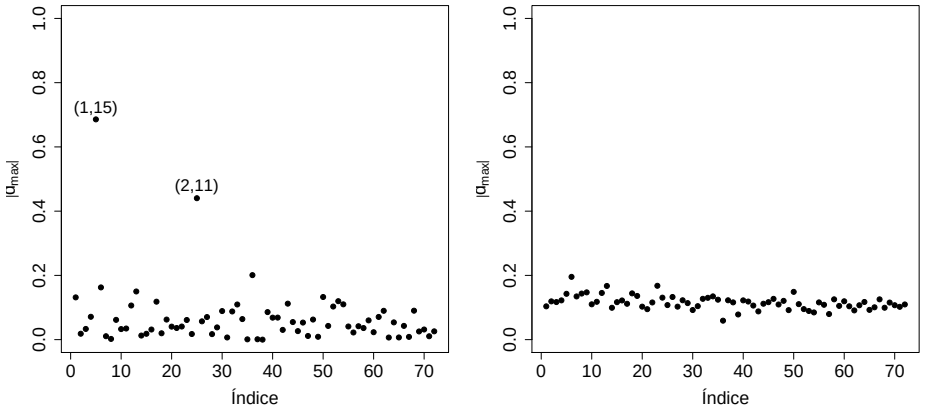
(a) (b)

Figura 6.1: Dados simulados perturbados - modelo beta truncado misto Gardner: (a) resíduos padronizados; (b) valores observados de $\boldsymbol{y}$ contra seus valores preditos.



(a) (b)

Figura 6.2: Dados simulados perturbados - modelo beta truncado misto Gardner: (a) distância generalizada de Cook; (b) distância da verossimilhança.



(a) (b)

Figura 6.3: Dados simulados perturbados - modelo beta truncado misto Gardner: (a) influência local sob o esquema de perturbação da resposta; (b) influência local sob perturbação de casos.

6.4 Condição de normalidade dos resíduos padronizados dos modelos propostos

Com o objetivo de verificar se os resíduos padronizados (4.11), apresentados na Seção 4.2, possuem distribuição aproximadamente normal para os modelos não linear aleatoriamente truncado misto (3.1), tomamos as amostras simuladas nas Seções 6.1 e 6.2 e os respectivos EMVs dos parâmetros e calculamos os resíduos padronizados para cada observação em cada iteração da simulação. Uma vez obtidos os resíduos padronizados para cada uma das observações e para cada um dos conjuntos de dados simulados, o teste de Anderson-Darling (Stephens, 1974) foi computado considerando-se os resíduos padronizados obtidos de todas as simulações para cada observação. Desta forma, e como os valores da covariável x são diferentes para cada observação, é possível verificarmos se há indicativos de que a distribuição dos resíduos se aproximam da distribuição normal para todas as observações. Assim, ao final da simulação temos $n = \sum_{i=1}^{M}\sum_{j=1}^{N_i} n_{i,j}$ valores do teste de Anderson-Darling, um para cada resíduo padronizado $r^{y}_{i,jk}$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$. Para fazer uma analogia com o que ocorreria com os resíduos de um conjunto de dados reais, apresentamos os resultados obtidos para as amostras simuladas com tamanhos iguais a 72 (o mesmo tamanho do conjunto de dados a ser analisado no capítulo de aplicação). Considerando um nível de significância de $0,05$, no caso do modelo não linear normal aleatoriamente truncado misto, a proporção de observações para as quais o teste de Anderson-Darling não rejeitou a hipótese de normalidade foi de $0,6$ e no caso do modelo não linear beta aleatoriamente truncado misto esta mesma proporção foi de $0,54$. Desta forma, podemos concluir que não parece haver evidências de que os resíduos

padronizados (4.11) possuam distribuição normal, tanto sob o modelo não linear normal aleatoriamente truncado misto (3.12) como sob o modelo não linear beta aleatoriamente truncado misto (3.20).

Nas Figuras 6.4 e 6.5 apresentamos o *qqplot* (gráfico quantil-quantil) dos resíduos com envelopes simulados. Esses gráficos são úteis para verificar a adequação do modelo postulado, quando os resíduos não possuem distribuição (aproximadamente) normal. Caso o modelo ajustado aos dados seja adequado, então os resíduos estarão dispostos entre os limites do envelope. Além disso, a presença de pontos externos e distantes do envelope simulado é um indicativo de possíveis casos *outliers* e/ou influentes, e a presença de padrões pode indicar falta de ajuste do modelo em questão. Para obter os envelopes simulados, usamos o seguinte procedimento iterativo baseado no procedimento apresentado em Cook & Weisberg (1994):

Passo 1: ajuste o modelo não linear aleatoriamente truncado misto (3.1) aos dados observados, calcule os resíduos padronizados (4.11) e ordene-os, denotando-os por $d^{y}_{(i,jk)}$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$;

Passo 2: simule um vetor de valores observados da variável resposta truncada, usando os EMVs dos parâmetros obtidos no passo 1, fazendo:

$$\begin{gathered} Y^{*}_{i,jk} \left| \left(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y^{*}_{i,jk} < B_{i,jk}, \hat{\boldsymbol{u}}_i\right) \sim \right. \\ f\left(y^{*}_{i,jk} \left| a_{i,jk}, b_{i,jk}, a_{i,jk} < y^{*}_{i,jk} < b_{i,jk}, \hat{\boldsymbol{u}}_i \, ; \hat{\mu}_{i,jk}, \hat{\sigma}_{i,jk}\right.\right), \\ \hat{\mu}_{i,jk} = \eta\left(\boldsymbol{x}_{q_{i,jk}}, \hat{\boldsymbol{\beta}}\right) + \boldsymbol{U}_{i,jk}\hat{\boldsymbol{u}}_i, \qquad \hat{\sigma}_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \hat{\boldsymbol{\alpha}}\right), \\ i = 1, \ldots, M, j = 1, \ldots, N_i, k = 1, \ldots, n_{ij}, \end{gathered} \tag{6.1}$$

Passo 3: ajuste o modelo não linear aleatoriamente truncado misto (3.1) aos dados simulados do passo 2, calcule os resíduos padronizados $r^{y^*}_{i,jk}$ e ordene-os denotando-os por $d^{y^*}_{(i,jk)}$;

Passo 4: repita os passos 2 e 3 um número suficientemente grande de vezes (neste trabalho, usamos $L = 1000$);

Passo 5: para cada $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, calcule os percentis $2,5\%$ e $97,5\%$ da amostra de resíduos simulados ordenados $d^{y^*}_{(i,jk),1}, \ldots, d^{y^*}_{(i,jk),L}$;

Passo 6: faça um gráfico de dispersão com os resíduos padronizados ordenados $d^{y}_{(i,jk)}$ obtidos no passo 1 contra os quantis teóricos da distribuição normal padrão com os envelopes simulados dados pelos percentis obtidos no passo 5.

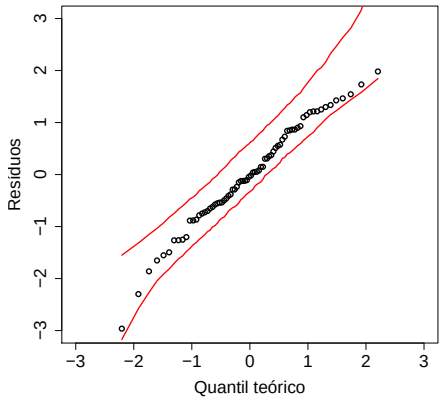


Figura 6.4: Dados simulados - modelo normal aleatoriamente truncado misto van Genuchten heteroscedástico: *qqplot* com envelope simulado.

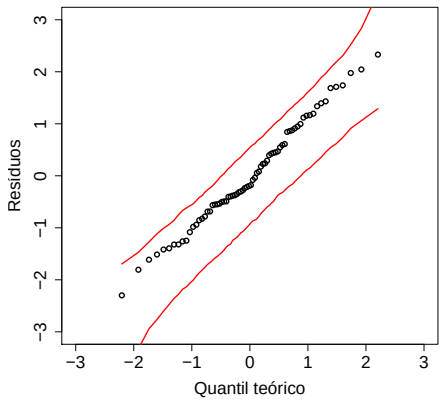


Figura 6.5: Dados simulados - modelo beta aleatoriamente truncado misto Gardner: *qqplot* com envelope simulado.

Capítulo 7

Estudo de simulação: metodologia Bayesiana

Neste capítulo, apresentamos resultados de simulação para ilustrar o processo de estimação e diagnóstico descritos no Capítulo 5, quando aplicados aos modelos propostos no Capítulo 3. Consideramos estudos de simulação com apenas uma amostra para verificar a sensibilidade das estimativas Bayesianas dos parâmetros com relação à escolha das distribuições a *priori*, e para ilustrar e avaliar o desempenho das métricas Bayesianas de diagnóstico de modelo. Apresentamos, ainda, resultados baseados em conjuntos de dados simulados para avaliar as propriedades frequentistas das estimativas Bayesianas obtidas para os modelos não lineares aleatoriamente truncados mistos (3.1), tratados sob o enfoque Bayesiano apresentado na Seção 5.1.

Assim como na Seção 6, os conjuntos de dados simulados foram gerados baseados em dados de retenção de água em solo, nos quais o teor de umidade de amostras de solos é limitado inferiormente pela umidade residual do solo, θ_r, e é limitada superiormente pela umidade do solo saturado, θ_s. Portanto, a variável resposta truncada $Y\,|(A, B, A < Y < B)$ corresponde ao teor de umidade, e as variáveis de truncamento inferior e superior são $\theta_r = A\,|(B, A < Y < B)$ e $\theta_s = B\,|(A < Y < B)$.

Nas Seções 6.1 e 6.2 os resultados dos estudos de simulação apresentados indicam que a partir do tamanho amostral $n = 720$ os EMVs dos parâmetros dos modelos em questão passam a ter boas propriedades frequentistas, isto é, suas estimativas tornam-se satisfatoriamente precisas e tanto o vício quanto o EQM tornam-se razoavelmente pequenos. Assim, nas Seções 7.1.2 e 7.2.2, onde são apresentados os resultados de simulação para avaliar as propriedades frequentistas das estimativas Bayesianas dos parâmetros dos modelos propostos, são simuladas apenas amostras de tamanho $n = 720$, isto é, considerando $M = 3$ grupos com $N_i = 30$ replicatas e $n_{ij} = 8$ observações cada. Para cada modelo, foram simulados um total de 200 conjuntos de dados.

As amostras MCMC das distribuições a *posteriori* foram obtidas utilizando-se dois tipos de algoritmos: para obter uma amostra MCMC da distribuição a posteriori de $\boldsymbol{\theta}$,

utilizamos um algoritmo do tipo Metropolis-Hastings com Gibbs, no qual cada um dos candidatos, para cada um dos parâmetros, é gerado de um passeio aleatório considerando-se uma distribuição normal univariada com desvio-padrão dada pelo elemento da diagonal da matriz de covariâncias definida pela negativa da matriz Hessiana avaliada nos EMVs dos parâmetros. As amostras MCMC de $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ são obtidas usando-se um algoritimo do tipo Metropolis-Hastings, no qual os candidatos são gerados por passeio aleatório considerando-se uma distribuição normal multivariada com matriz de covariâncias dada pela negativa da matriz Hessiana avaliada nos EMVs dos parâmetros. As convergências das cadeias foram verificadas utilizando-se o critério Geweke (Geweke, 1992). Os algoritmos MCMC Gibbs-Metropolis e Metropolis-Hastings foram implementados em `R` e são exemplificados no Apêndice D.2. Os tamanhos das cadeias foram 100 mil com períodos de burn-in de 20 mil e salto de 200.

7.1 Resultados de simulação para o modelo de regressão não linear normal aleatoriamente truncado misto

No estudo de simulação do modelo de regressão não linear beta aleatoriamente truncado misto, as variáveis de truncamento superior e inferior são simuladas de $B_{i,jk}\,|(A_{i,jk} < Y_{i,jk} < B_{i,jk}) \sim N(\mu_B, \sigma_B)$ e $A_{i,jk}\,|(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i) \sim NT(\mu_A, \sigma_A, -\infty, b_{i,jk})$, $i = 1,2,3$, $j = 1,\ldots,N_i$, $k = 1,\ldots,n_{ij}$. Uma vez gerados os limites de truncamento, as observações da variável resposta truncada foram simuladas de $Y_{i,jk}\,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i) \sim NT\left(\eta(x_{i,jk}, \boldsymbol{\beta}) + u_i, \sigma x_{i,jk}^{\alpha}, a_{i,jk}, b_{i,jk}\right)$, com $u_i \sim N(0, \sigma_u)$, $i = 1,2,3$, $j = 1,\ldots,N_i$, $k = 1,\ldots,n_{ij}$. A função $\eta(x_{i,jk}, \boldsymbol{\beta})$ é dada pela expressão de van Genuchten (1980)-Mualem (1976) em (1.2)-(1.4).

7.1.1 Análise de sensibilidade à distribuição a *priori*

A avaliação da sensibilidade do modelo não linear normal aleatoriamente truncado misto, com relação à distribuição a *priori* dos parâmetros, é realizada através de um estudo comparativo entre os resultados obtidos do ajuste do modelo a um conjunto de dados simulado usando-se diferentes distribuições a *priori*.

Para o parâmetro β_1 consideramos uma distribuição a *priori* fracamente informativa inversa-gama com parâmetros τ^{-1} e τ^{-1}, denotada por $IG(\tau^{-1}, \tau^{-1})$. Como na equação de van Genuchten-Mualen (1.2)-(1.4) $\beta_2 > 1$, consideramos uma distribuição a *priori* gama truncada para β_2, $GT(\tau^{-1}, \tau^{-1}, 1, +\infty)$. Para σ consideramos a distribuição a *priori* $IG(\tau^{-1}, \tau^{-1})$.

Para σ_u consideramos a distribuição a *priori* $IG(\tau^{-1}, \tau^{-1})$ e também uma *priori* $Half - t(\nu, \tau)$ proposta por Gelman (2006), que é uma *priori* fracamente informativa mais eficiente do que a inversa-gama, especialmente para casos em que o parâmetro de

escala dos efeitos aleatórios pode ser pequeno. Neste trabalho, consideramos a distribuição half-Cauchy, um caso especial da distribuição $Half-t\left(\nu,\tau\right)$ obtido quando $\nu=1$. Assim, assumimos a reparametrização $\sigma_u^r=\sqrt{|\sigma_u|}$, e a distribuição a *priori* half-Cauchy de σ_u é dada por $\pi\left(\sigma_u\right)\propto\left(\sigma_u^2+\lambda^2\right)^{-1}$, com λ um hiperparâmetro fixo e conhecido.

Para o parâmetro α e os efeitos aleatórios u_i, assumimos distribuições a *priori* fracamente informativas $N\left(0,\tau\right)$. Para μ_A e μ_B assumimos *priori* $N\left(0,\tau\right)$, e para σ_A e σ_B consideramos a distribuição $IG\left(\tau^{-1},\tau^{-1}\right)$.

Consideramos, ainda, o modelo reparametrizado da seguinte maneira: $\beta_1^r=\log\left(\beta_1\right)$ e $\beta_2^r=\log\left(\beta_2-1\right)$, $\sigma^r=\log\left(\sigma\right)$, $\sigma_u^r=\log\left(\sigma\right)$, $\sigma_A^r=\log\left(\sigma_A\right)$ e $\sigma_B^r=\log\left(\sigma_B\right)$. Neste cenário, é assumida uma distribuição a *priori* $N\left(0,\tau\right)$ para cada um dos parâmetros.

Os resumos a *posteriori* obtidos para os parâmetros β_1, β_2, σ, α e σ_u, e para os efeitos aleatórios u_1, u_2 e u_3, sob as três diferentes escolhas de distribuições a *priori*, são mostrados na Tabela 7.1. Nota-se que a média e mediana a *posteriori* e os intervalos de 95% de credibilidade inter-quantil e HPD de todos os parâmetros, com exceção de σ_u, são muito similares, resultando em resumos a *posteriori* razoavelmente iguais sob os três cenários de escolhas de distribuições a *priori*. Por outro lado, observamos que os resultados obtidos para σ_u considerando a *priori* half-Cauchy são relativamente diferentes dos obtidos sob as distribuições a *priori* inversa-gama e normal (modelo reparametrizado), sendo que o parâmetro σ_u parece ser estimado de maneira mais precisa quando consideramos as distribuições a *priori* half-Cauchy. Desta forma, no restante deste trabalho optamos por usar as distribuição a *priori* como definidas no segundo cenário, ou seja, assumindo uma *priori* half-Cauchy fracamente informativa para σ_u.

Os resumos a *posteriori* obtidos para os parâmetros μ_A e σ_A, e μ_B e σ_B são mostrados nas Tabelas 7.2 e 7.3, respectivamente. Pelos resultados obtidos, é possível notar que a média e mediana a *posteriori* e os intervalos de 95% de credibilidade inter-quantil e HPD são muito similares, resultando em resumos a *posteriori* razoavelmente iguais sob as três escolhas de distribuições a *priori*. Desta forma, optamos pelas distribuições a *priori* como definidas nos primeiros cenários apresentados nas Tabelas 7.2 e 7.3 para os parâmetros relacionados com as variáveis de truncamento inferior e superior, respectivamente.

7.1.2 Propriedades frequentistas dos estimadores Bayesianos

Os resultados da avaliação das propriedades frequentistas das estimativas Bayesianas do modelo de regressão não linear normal aleatoriamente truncado misto são baseados em um total de 200 amostras simuladas. Os valores assumidos para os parâmetros foram: $\beta_1=55$ e $\beta_2=1,45$, $\sigma=0,01$, $\alpha=-0,15$, $\sigma_u=0,02$, $\mu_A=0,25$, $\sigma_A=0,01$, $\mu_B=0,5$ e $\sigma_B=0,05$.

Os resultados da simulação são apresentados na Tabela 7.4, onde a quarta coluna mostra a média das médias a *posteriori* obtidas para cada amostra, a quinta coluna mostra

Tabela 7.1: Estudo de sensibilidade com relação a distribuição a *priori* para o modelo Bayesiano normal aleatoriamente truncado misto van Genuchten-Mualem ajustado aos dados simulados: parâmetros β_1, β_2, σ, σ_u, u_1, u_2 e u_3.

Parâmetro/*Priori*	Valor real	Média a *posteriori*	Mediana a *posteriori*	Desv. pad.	Intervalos de credibilidade 95% Inter-quantil	Intervalos de credibilidade 95% HPD
$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$	55	48,8309	48,9723	4,9401	(39,2527; 59,6448)	(38,6145; 58,6356)
$\beta_2 \sim GT(\tau^{-1}, \tau^{-1}, 1, +\infty)$	1,45	1,4733	1,4723	0,0345	(1,4154; 1,5462)	(1,4141; 1,5425)
$\sigma \sim IG(\tau^{-1}, \tau^{-1})$	0,01	0,0117	0,0116	0,0014	(0,0094; 0,0149)	(0,0091; 0,0142)
$\alpha \sim N(0, \tau)$	-0,15	-0,1237	-0,1312	0,0449	(-0,2094; -0,0268)	(-0,2163; -0,0354)
$\sigma_u \sim IG(\tau^{-1}, \tau^{-1})$	0,05	0,0220	0,0171	0,0175	(0,0066; 0,0674)	(0,0054; 0,0539)
$u_1 \sim N(0, \tau)$	0,0136	0,0118	0,0120	0,0046	(0,0025; 0,0203)	(0,0039; 0,0212)
$u_2 \sim N(0, \tau)$	-0,0105	-0,0115	-0,0116	0,0048	(-0,0209; -0,0019)	(-0,0216; -0,0028)
$u_3 \sim N(0, \tau)$	0,0063	0,0089	0,0090	0,0049	(-0,0007; 0,0184)	(0,0009; 0,0196)
$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$	55	47,9933	47,6043	4,6947	(39,4923; 57,9538)	(40,1319; 58,4699)
$\beta_2 \sim GT(\tau^{-1}, \tau^{-1}, 1, +\infty)$	1,45	1,4818	1,4786	0,0351	(1,4222; 1,5570)	(1,4213; 1,5527)
$\sigma \sim Half-t(\nu, \tau)$	0,01	0,0118	0,0117	0,0015	(0,0093; 0,0150)	(0,0090; 0,0146)
$\alpha \sim N(0, \tau)$	-0,15	-0,1252	-0,1267	0,0492	(-0,2202; -0,0186)	(-0,2110; -0,0142)
$\sigma_u \sim Half-t(\nu, \tau)$	0,05	0,0542	0,0272	0,0728	(0,0090; 0,2410)	(0,0061; 0,1763)
$u_1 \sim N(0, \tau)$	0,0136	0,0122	0,0127	0,0047	(0,0030; 0,0204)	(0,0030; 0,0204)
$u_2 \sim N(0, \tau)$	-0,0105	-0,0113	-0,0112	0,0048	(-0,0218; -0,0028)	(-0,0219; -0,0028)
$u_3 \sim N(0, \tau)$	0,0063	0,0100	0,0100	0,0044	(0,0013; 0,0177)	(0,0015; 0,0179)
$\beta_1^r \sim N(0, \tau)$	55	48,5087	48,2197	4,8520	(39,7983; 59,0812)	(40,0661; 59,2933)
$\beta_1^r \sim N(0, \tau)$	1,45	1,4748	1,4720	0,0319	(1,4244; 1,5477)	(1,4231; 1,5411)
$\sigma \sim N(0, \tau)$	0,01	0,0116	0,0114	0,0013	(0,0093; 0,0144)	(0,0088; 0,0138)
$\alpha \sim N(0, \tau)$	-0,15	-0,1255	-0,1295	0,0462	(-0,2200; -0,0287)	(-0,2242; -0,0353)
$\sigma_u^r \sim N(0, \tau)$	0,05	0,0152	0,0118	0,0142	(0,0059; 0,0436)	(0,0050; 0,0349)
$u_1 \sim N(0, \tau)$	0,0136	0,0114	0,0117	0,0041	(0,0038; 0,0192)	(0,0040; 0,0193)
$u_2 \sim N(0, \tau)$	-0,0105	-0,0112	-0,0112	0,0043	(-0,0190; -0,0031)	(-0,0192; -0,0034)
$u_3 \sim N(0, \tau)$	0,0063	0,0092	0,0090	0,0040	(0,0017; 0,0168)	(0,0020; 0,0171)

$\tau = 10^4$, $\lambda = 10$ e $\nu = 1$.

Tabela 7.2: Estudo de sensibilidade com relação a distribuição a *priori* para o modelo Bayesiano normal aleatoriamente truncado misto van Genuchten-Mualem ajustado aos dados simulados: parâmetros μ_A e σ_A.

Parâmetro/*Priori*	Valor real	Média a *posteriori*	Mediana a *posteriori*	Desv. pad.	Intervalos de credibilidade 95% Inter-quantil	Intervalos de credibilidade 95% HPD
$\mu_A \sim N(0, \tau)$	0,25	0,249755	0,249796	0,001161	(0,247505; 0,25205)	(0,247506; 0,25205)
$\sigma_A \sim IG(\tau^{-1}, \tau^{-1})$	0,01	0,009932	0,009819	0,00084	(0,008498; 0,011657)	(0,008568; 0,011702)
$\mu_A \sim N(0, \tau)$	0,25	0,24978	0,249776	0,001165	(0,247465; 0,251901)	(0,247731; 0,252048)
$\sigma_A^r \sim N(0, \tau)$	0,01	0,009877	0,009765	0,000867	(0,008462; 0,01183)	(0,008276; 0,011553)

$\tau = 10^4$, $\lambda = 10$ e $\nu = 1$.

Tabela 7.3: Estudo de sensibilidade com relação a distribuição a *priori* para o modelo Bayesiano normal aleatoriamente truncado misto van Genuchten-Mualem ajustado aos dados simulados: parâmetros μ_B e σ_B.

Parâmetro/*Priori*	Valor real	Média a *posteriori*	Mediana a *posteriori*	Desv. pad.	Intervalos de credibilidade 95% Inter-quantil	Intervalos de credibilidade 95% HPD
$\mu_B \sim N(0, \tau)$	0.5	0.5041	0.5039	0.0055	(0.4923; 0.5148)	(0.4927; 0.5150)
$\sigma_B \sim IG(\tau^{-1}, \tau^{-1})$	0.05	0.0488	0.0484	0.0044	(0.0413; 0.0578)	(0.0413; 0.0579)
$\mu_B \sim N(0, \tau)$	0.5	0.5040	0.5037	0.0057	(0.4935; 0.5157)	(0.4937; 0.5158)
$\sigma_B^r \sim N(0, \tau)$	0.05	0.0487	0.0485	0.0042	(0.0414; 0.0575)	(0.0409; 0.0569)

$\tau = 10^4$, $\lambda = 10$ e $\nu = 1$.

o EQM obtido para as médias a *posteriori*, a sexta coluna mostra o vício obtido para as médias a *posteriori*, a sétima coluna apresenta a mediana das medianas a *posteriori*; a oitava coluna mostra o EQM obtido para as medianas a *posteriori*; a nona coluna apresenta o vício obtido para as medianas a *posteriori*; nas décima e décima primeira colunas temos as probabilidades de cobertura dos intervalos de 95% de credibilidade inter-quantil e HPD, respectivamente.

Pelos resultados apresentados na Tabela 7.4, nota-se que a média a posteriori e a mediana a posteriori dos parâmetros são similares. Além disso, nota-se que o procedimento Bayesiano de estimação fornece resultados precisos para as estimativas dos parâmetros fixos, β_1 e β_2, que compõem a função não linear $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$, relacionada às médias das respostas, e para as estimativas Bayesianas dos parâmetros σ e α, que estão relacionados à dispersão das respostas. Para β_1, β_2, σ e α, tanto o vício como o EQM são pequenos e a probabilidade de cobertura estimada dos intervalos de credibilidade inter-quantil e HPD se aproximam da nominal esperada de 95%. Já a estimativa Bayesiana do parâmetro σ_u é pouco precisa e o mesmo é subestimado. Além disso, as probabilidades de cobertura estimadas para os seus intervalos de credibilidade inter-quantil e HPD estão abaixo de 95%. No entanto, as estimativas Bayesianas dos efeitos aleatórios u_1, u_2 e u_2 são satisfatórias e suas probabilidades de cobertura estimadas estão próximas do esperado.

Tabela 7.4: Resultados da simulação para o modelo Bayesiano normal aleatoriamente truncado misto van Genuchten-Mualem.

		Valor	Média			Mediana			Probabilidade de cobertura	
n	Parâmetro	real	a posteriori	Vício-m	EQM-m	a posteriori	Vício-md	EQM-md	Inter-quantil	HPD
	β_1	55,00	55,0218	2,18E-02	2,84E+00	54,9183	2,18E-02	2,84E+00	0,93	0,92
	β_2	1,45	1,4500	3,94E-05	1,09E-04	1,4494	3,94E-05	1,11E-04	0,97	0,96
	σ	0,01	0,0101	7,47E-05	9,97E-08	0,0101	7,47E-05	9,86E-08	0,95	0,95
	α	-0,15	-0,1480	1,96E-03	1,79E-04	-0,1474	1,96E-03	1,83E-04	0,94	0,94
	σ_u	0,05	0,0789	2,89E-02	7,74E-03	0,0275	2,89E-02	3,75E-03	1,00	0,99
	u_1	-	0,0067	-8,90E-05	3,23E-04	0,0052	-8,90E-05	3,23E-04	0,96	0,94
720	u_2	-	0,0091	-2,11E-05	2,26E-04	0,0080	-2,11E-05	2,26E-04	0,96	0,95
	u_3	-	0,0077	-7,99E-05	3,01E-04	0,0060	-7,99E-05	3,01E-04	0,97	0,98
	μ_A	0,25	0,2500	7,38E-07	1,56E-07	0,2500	7,38E-07	1,56E-07	0,93	0,91
	σ_A	0,01	0,0100	-9,33E-07	8,94E-08	0,0100	-9,33E-07	8,93E-08	0,92	0,91
	μ_B	0,50	0,5000	-4,11E-05	2,54E-06	0,4999	-4,11E-05	2,56E-06	0,98	0,97
	σ_B	0,05	0,0500	5,00E-05	1,69E-06	0,0500	5,00E-05	1,68E-06	0,96	0,95

7.2 Resultados de simulação para o modelo de regressão não linear beta aleatoriamente truncado misto

Nos resultados apresentados para o modelo de regressão não linear beta aleatoriamente truncado misto, a variável de truncamento superior é simulada de $B_{i,jk}\left|\left(A_{i,jk} < Y_{i,jk} < B_{i,jk}\right) \sim B\left(\mu_B, \sigma_B\right)\right.$ e a variável de truncamento inferior é gerada como $A_{i,jk}\left|\left(B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i\right) \sim BT\left(\mu_A, \sigma_A, -\infty, b_{i,jk}\right)\right.$, $i = 1, 2, 3$, $j = 1, \ldots, N_i$,

$k = 1, \ldots, n_{ij}$. Uma vez gerados os limites de truncamento, as observações da variável resposta truncada foram simuladas de $Y_{i,jk} \,|(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i) \sim BT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma x_{i,jk}^{\alpha}, a_{i,jk}, b_{i,jk}\right)$, com $u_i \sim N(0, \sigma_u)$, $i = 1, 2, 3$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$. A função $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ é dada pela expressão de Gardner (1958) em (1.1).

7.2.1 Análise de sensibilidade à distribuição a *priori*

A avaliação da sensibilidade do modelo não linear beta aleatoriamente truncado misto com relação à distribuição a *priori* dos parâmetros é realizada através de um estudo comparativo entre os resultados obtidos do ajuste do modelo a um conjunto de dados simulado usando-se diferentes distribuições a *priori*.

Para os parâmetros β_1 e β_2 consideramos uma distribuição a *priori* $IG(\tau^{-1}, \tau^{-1})$ e para σ consideramos uam distribuição a *priori* $N(0, \tau)$.

Para σ_u, assumimos uma distribuição a *priori* $IG(\tau^{-1}, \tau^{-1})$ e uma *priori* half-Cauchy, $\pi(\sigma_u) \propto (\sigma_u^2 + \lambda^2)^{-1}$, com λ um hiperparâmetro fixo e conhecido, denotada por $Half - t(\nu, \tau)$ com $\nu = 1$, e assumindo a reparametrização $\sigma_u^r = \sqrt{|\sigma_u|}$.

Para os efeitos aleatórios u_i, assumimos a distribuição a *priori* $N(0, \tau)$. Para μ_A e μ_B assumimos distribuição a *priori* $Beta(1/2, 1/2)$, e para σ_A e σ_B consideramos a distribuição $N(0, \tau)$.

Consideramos, ainda, o modelo reparametrizado da seguinte maneira: $\beta_1^r = \log(\beta_1)$ e $\beta_2^r = \log(\beta_2)$, $\sigma_u^r = \log(\sigma)$, $\mu_A^r = \log\{\mu_A/(1-\mu_A)\}$ e $\mu_B^r = \log\{\mu_B/(1-\mu_B)\}$. Neste cenário, para cada um dos parâmetros é assumida uma distribuição a *priori* $N(0, \tau)$.

Os resumos a *posteriori* obtidos para os parâmetros β_1, β_2, σ e σ_u, e para os efeitos aleatórios u_1, u_2 e u_3, sob as diferentes escolhas de distribuições a *priori*, são mostrados na Tabela 7.5. Similarmente, os resumos a *posteriori* obtidos para os parâmetros μ_A e σ_A, e μ_B e σ_B são mostrados nas Tabelas 7.6 e 7.7, respectivamente.

Pelos resultados apresentados na Tabela 7.5, é possível notar que a média e mediana a *posteriori* e os intervalos de 95% de credibilidade inter-quantil e HPD de todos os parâmetros, com exceção de σ_u, são muito similares, resultando em resumos a *posteriori* razoavelmente iguais sob os três cenários de escolhas de distribuições a *priori*. Por outro lado, podemos notar que os resultados obtidos para σ_u considerando a *priori* half-Cauchy são mais precisos do que os obtidos sob as distribuições a *priori* inversa-gama e normal (modelo reparametrizado). Desta forma, no restante deste trabalho optamos por usar as distribuição a *priori* como definidas no cenário dois, ou seja, assumindo uma *priori* half-Cauchy para σ_u.

Os resultados apresentados nas Tabelas 7.6 e 7.7, para os parâmetros (μ_A, σ_A) e (μ_B, σ_B), indicam que os dois cenários de distribuições a *priori* fornecem resumos a *posteriori* bastante similares. Assim, no restante do trabalho usamos as distribuições a *priori* como definidas nos primeiros cenários apresentados nas Tabelas 7.6 e 7.7 para os parâme-

tros relacionados com as variáveis de truncamento inferior e superior, respectivamente.

Tabela 7.5: Estudo de sensibilidade com relação a distribuição a *priori* para o modelo Bayesiano beta aleatoriamente truncado misto Gardner ajustado aos dados simulados: parâmetros β_1, β_2, σ, σ_u, u_1, u_2 e u_3.

Parâmetro/*Priori*	Valor real	Média a *posteriori*	Mediana a *posteriori*	Desv. pad.	Intervalos de credibilidade 95% Inter-quantil	HPD
$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$	2,25	1,8859	1,8634	0,2753	(1,4361, 2,4827	(1,3569, 2,3817)
$\beta_2 \sim IG(\tau^{-1}, \tau^{-1})$	0,5	0,5073	0,5071	0,0242	(0,4591, 0,5574	(0,4577, 0,5525)
$\sigma \sim N(0, \tau)$	6	6,1706	6,1699	0,1691	(5,8357, 6,4989)	(5,8004, 6,4543)
$\sigma_u \sim IG(\tau^{-1}, \tau^{-1})$	0,1	0,0639	0,0498	0,0509	(0,0201, 0,2302)	(0,0156, 0,1681)
$u_1 \sim N(0, \tau)$	0,0493	0,0271	0,0277	0,0101	(0,0078, 0,0457)	(0,0071, 0,0447)
$u_2 \sim N(0, \tau)$	-0,0226	-0,0320	-0,0314	0,0105	(-0,0511, -0,0131)	(-0,0509, -0,0129)
$u_3 \sim N(0, \tau)$	-0,0278	-0,0411	-0,0411	0,0102	(-0,0615, -0,0220)	(-0,0614, -0,0220)
$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$	2,25	1,8899	1,8645	0,2598	(1,4473, 2,4893)	(1,3681, 2,3825)
$\beta_2 \sim IG(\tau^{-1}, \tau^{-1})$	0,5	0,5078	0,5065	0,0248	(0,4584, 0,5630)	(0,4559, 0,5587)
$\sigma \sim N(0, \tau)$	6	6,1597	6,1658	0,1804	(5,7675, 6,5248)	(5,7417, 6,4702)
$\sigma_u \sim Half - t(\nu, \lambda)$	0,1	0,1151	0,0813	0,1040	(0,0265, 0,4523)	(0,0137, 0,3355)
$u_1 \sim N(0, \tau)$	0,0493	0,0276	0,0276	0,0097	(0,0093, 0,0483)	(0,0078, 0,0468)
$u_2 \sim N(0, \tau)$	-0,0226	-0,0322	-0,0313	0,0093	(-0,0527, -0,0155)	(-0,0533, -0,0167)
$u_3 \sim N(0, \tau)$	-0,0278	-0,0415	-0,0414	0,0088	(-0,0583, -0,0247)	(-0,0578, -0,0242)
$\beta_1^r \sim N(0, \tau)$	2,25	1,9181	1,8905	0,2694	(1,4324, 2,5117)	(1,3784, 2,3923)
$\beta_1^r \sim N(0, \tau)$	0,5	0,5065	0,5056	0,0239	(0,4642, 0,5525)	(0,4623, 0,5517)
$\sigma \sim N(0, \tau)$	6	6,1725	6,1823	0,1863	(5,8175, 6,4837)	(5,8589, 6,5097)
$\sigma_u^r \sim N(0, \tau)$	0,1	0,0493	0,0407	0,0379	(0,0195, 0,1200)	(0,0158, 0,1008)
$u_1 \sim N(0, \tau)$	0,0493	0,0288	0,0289	0,0099	(0,0097, 0,0485)	(0,0081, 0,0459)
$u_2 \sim N(0, \tau)$	-0,0226	-0,0303	-0,0301	0,0100	(-0,0503, -0,0117)	(-0,0489, -0,0105)
$u_3 \sim N(0, \tau)$	-0,0278	-0,0397	-0,0390	0,0095	(-0,0593, -0,0235)	(-0,0593, -0,0235)

$\tau = 10^4$, $\lambda = 10$ e $\nu = 1$.

Tabela 7.6: Estudo de sensibilidade com relação a distribuição a *priori* para o modelo Bayesiano beta aleatoriamente truncado misto Gardner ajustado aos dados simulados: parâmetros μ_A e σ_A.

Parâmetro/*Priori*	Valor real	Média a *posteriori*	Mediana a *posteriori*	Desv. pad.	Intervalos de credibilidade 95% Inter-quantil	HPD
$\mu_A \sim Beta(1/2, 1/2)$	0,25	0,2468	0,2469	0,0037	(0,2392; 0,2536)	(0,2392; 0,2536)
$\sigma_A \sim N(0, \tau)$	5,2	5,1584	5,1711	0,1710	(4,7960; 5,4667)	(4,8509; 5,5079)
$\mu_A^r \sim N(0, \tau)$	0,25	0,2473	0,2471	0,0041	(0,2395; 0,2552)	(0,2399; 0,2554)
$\sigma_A \sim N(0, \tau)$	5,2	5,1617	5,1693	0,1775	(4,8296; 5,5043)	(4,8084; 5,4618)

$\tau = 10^4$, $\lambda = 10$ e $\nu = 1$.

7.2.2 Propriedades frequentistas dos estimadores Bayesianos

Os resultados da avaliação das propriedades frequentistas das estimativas Bayesianas do modelo de regressão não linear beta aleatoriamente truncado misto são baseados em um total de 200 amostras simuladas. Os valores assumidos para os parâmetros foram: $\beta_1 = 2,25$ e $\beta_2 = 0,5$, $\sigma = 6$, $\sigma_u = 0,1$, $\mu_A = 0,25$, $\sigma_A = 5,2$, $\mu_B = 0,6$ e $\sigma_B = 4,6$.

Os resultados da simulação são apresentados na Tabela 7.8, onde a quarta coluna mostra a média das médias a *posteriori* obtidas para cada amostra, a quinta coluna mostra

Tabela 7.7: Estudo de sensibilidade com relação a distribuição a *priori* para o modelo Bayesiano beta aleatoriamente truncado misto Gardner ajustado aos dados simulados: parâmetros μ_B e σ_B.

Parâmetro/*Priori*	Valor real	Média a *posteriori*	Mediana a *posteriori*	Desv. pad.	Intervalos de credibilidade 95% Inter-quantil	HPD
$\mu_B \sim Beta\,(1/2, 1/2)$	0,6	0,5980	0,5980	0,0054	(0,5867; 0,6076)	(0,5868; 0,6076)
$\sigma_B \sim N\,(0, \tau)$	4,6	4,7159	4,7151	0,1764	(4,3647; 5,0356)	(4,3657; 5,0391)
$\mu_B^r \sim N\,(0, \tau)$	0,6	0,5985	0,5982	0,0053	(0,5889; 0,6096)	(0,5884; 0,6086)
$\sigma_B \sim N\,(0, \tau)$	4,6	4,7239	4,7361	0,1846	(4,3373; 5,0625)	(4,3444; 5,0681)

$\tau = 10^4$, $\lambda = 10$ e $\nu = 1$.

o EQM obtido para as médias a *posteriori*, a sexta coluna mostra o vício obtido para as médias a *posteriori*, a sétima coluna apresenta a mediana das medianas a *posteriori*; a oitava coluna mostra o EQM obtido para as medianas a *posteriori*; a nona coluna apresenta o vício obtido para as medianas a *posteriori*; nas décima e décima primeira colunas temos as probabilidades de cobertura dos intervalos de 95% de credibilidade inter-quantil e HPD, respectivamente.

Os resultados são apresentados na Tabela 7.8, onde é possível notar que a média a posteriori e a mediana a posteriori dos parâmetros produzem resultados similares. Pelos resultados obtidos, o procedimento Bayesiano de estimação fornece resultados precisos para as estimativas dos parâmetros fixos, β_1 e β_2, que compõem a função não linear $\eta\,(x_{i,jk}, \boldsymbol{\beta})$, relacionada às médias das respostas. As estimativas Bayesianas do parâmetro σ, que está relacionado à dispersão das respostas, também é bastante precisa. Para β_1, β_2 e σ, tanto o vício como o EQM são pequenos e a probabilidade de cobertura estimada dos intervalos de credibilidade inter-quantil e HPD se aproximam da nominal esperada de 95%. Por outro lado, a estimativa Bayesiana do parâmetro σ_u não é tão precisa e este parâmetro parece estar sendo subestimado. Além disso, as probabilidades de cobertura estimadas para seus intervalos de credibilidade inter-quantil e HPD estão um pouco abaixo de 95%. No entanto, as estimativas Bayesianas dos efeitos aleatórios u_1, u_2 e u_2 são satisfatórias e suas probabilidades de cobertura estimadas estão próximas do esperado.

Tabela 7.8: Resultados da simulação para o modelo Bayesiano beta aleatoriamente truncado misto Gardner.

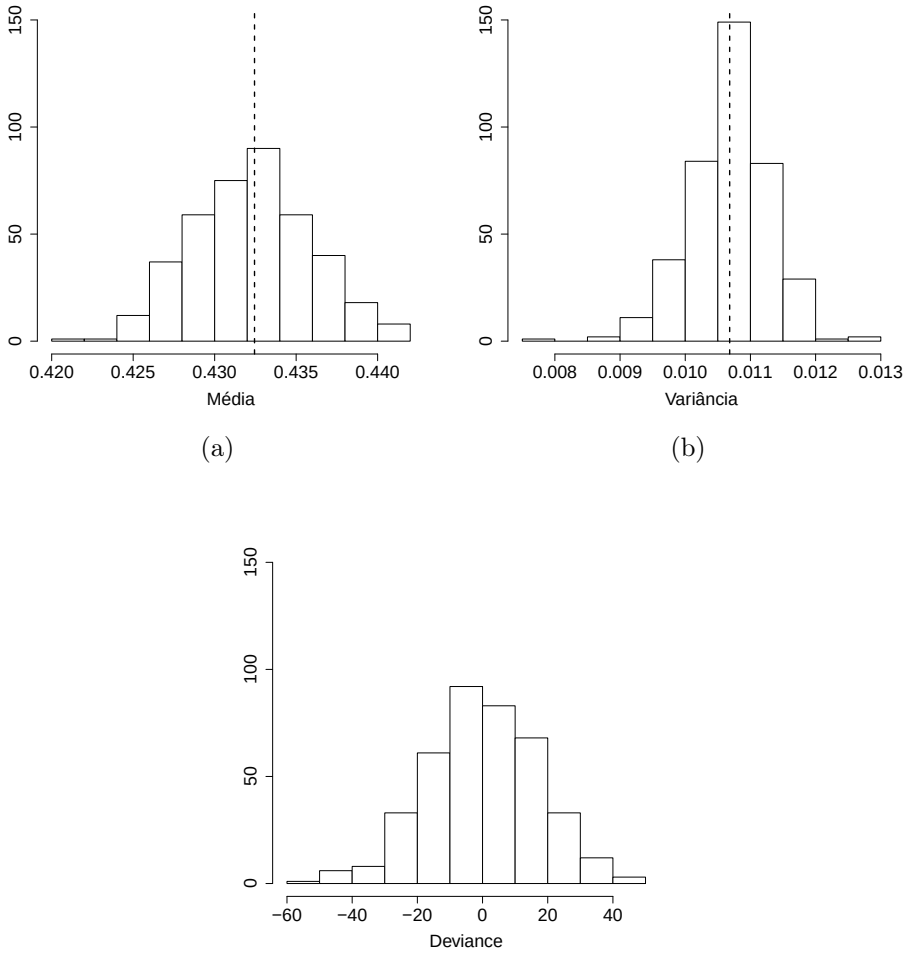
n	Parâmetro	Valor real	Média a posteriori	Vício-m	EQM-m	Mediana a posteriori	Vício-md	EQM-md	Probabilidade de cobertura Inter-quantil	HPD
720	β_1	2,25	2,2658	1,98E-02	1,84E-02	2,2611	1,98E-02	1,81E-02	0,92	0,91
	β_2	0,50	0,5012	1,56E-03	6,75E-05	0,5005	1,56E-03	6,58E-05	0,96	0,96
	σ	6,00	5,9934	-6,40E-03	2,71E-03	5,9913	-6,40E-03	2,74E-03	0,97	0,97
	σ_u	0,10	0,2016	9,94E-02	4,50E-02	0,1086	9,94E-02	3,79E-02	1,00	1,00
	u_1	-	0,0013	-1,21E-03	1,86E-03	-0,0021	-1,21E-03	1,85E-03	0,90	0,89
	u_2	-	0,0164	2,40E-03	2,14E-03	0,0226	2,40E-03	2,14E-03	0,91	0,92
	u_3	-	0,0125	-1,44E-03	1,92E-03	0,0164	-1,44E-03	1,91E-03	0,91	0,90
	μ_A	0,25	0,2501	1,15E-04	1,45E-06	0,2501	1,15E-04	1,44E-06	0,94	0,95
	σ_A	5,20	5,1997	-2,89E-04	2,70E-03	5,2027	-2,89E-04	2,70E-03	0,96	0,96
	μ_B	0,60	0,6001	9,25E-05	3,46E-06	0,6001	9,25E-05	3,49E-06	0,96	0,97
	σ_B	4,60	4,5972	-2,80E-03	2,67E-03	4,5954	-2,80E-03	2,68E-03	0,95	0,94

7.3 Diagnóstico Bayesiano aplicado a dados simulados: avaliação preditiva a *posteriori*

Para verificar se as métricas de avaliação preditiva a *posteriori* são capazes de identificar má especificação da estrutura de regressão assumida para o parâmetro de média das respostas, consideramos um conjunto de dados simulados assumindo $Y_{i,jk}\,|(a < Y_{i,jk} < b, u_i) \sim BT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, a, b\right)$, com $a = 0,31$, $b = 0,55$, $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ a expressão de van Genuchten (1980) dada em (1.2) e $u_i \sim N\left(0, \sigma_u\right)$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, $n_{ij} = 8$, $M = 3$, $N_i = 3$ e $n = 72$. Os valores reais dos parâmetros são $\beta_1 = 70$, $\beta_2 = 2,5$, $\sigma = 4$ e $\sigma_u = 0,1$. Após gerar os dados, o modelo não linear beta truncado misto (3.22) com a função de Gardner (1958) foi ajustado aos dados, isto é, considerando $Y_{i,jk}\,|(a < Y_{i,jk} < b, u_i) \sim BT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, a, b\right)$, com $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ a expressão dada na equação (1.1).

O histograma na Figura 7.1 indica que as discrepâncias baseadas na média, variância e *deviance* do modelo não são capazes de identificar uma diferença significativa entre os dados observados (simulados) e os dados replicados. Isto é, nota-se que a média, variância e *deviance* do modelo obtidos para os dados replicados coincidem com os valores da média, da variância e do *deviance* do modelo dos dados observados. Esta afirmação é corroborada pelos p-valores preditivos a *posteriori* estimados, que foram de 0,49, 0,56 e 0,50, para a média, a variância e *deviance* do modelo.

Desta forma, as métricas de avaliação preditiva a *posteriori* parecem não ser capazes de identificar a falta de ajuste sob o cenário proposto, ou seja, não parece ser possível detectar má especificação da função $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$. Por outro lado, Leong & Rahardjo (1997) mostraram que as expressões de curvas de retenção apresentadas na Seção 1.1 podem ser derivadas de uma expressão genérica única e, portanto, acreditamos que essas expressões sejam suficientemente similares como para não produzir resultados significativamente diferentes para os dados replicados e observados. Ressaltamos que resultados semelhantes foram obtidos ao comparar as outras expressões apresentadas na Seção 1.1 entre si.



(a) (b)

(c)

Figura 7.1: Dados simulados perturbados - modelo Bayesiano beta truncado misto van genuchten: histogramas das avaliações preditivas a *posteriori* baseadas na média (a), variância (b) e *deviance* do modelo (c).

7.4 Diagnóstico Bayesiano aplicado a dados simulados: resíduos e influência

Para ilustrar a detecção de observações *outliers* e/ou influentes do modelo de regressão não linear aleatoriamente truncado misto (3.1) sob o enfoque Bayesiano apresentado na Seção 5.1, tomamos um conjunto de dados simulado e calculamos os dois resíduos Bayesianos (5.7) e (5.8) e a calibração de casos (5.12) descritas na Seção 5.2. O conjunto de dados simulado usado para ilustrar as métricas de diagnóstico é gerado assumindo-se que o vetor aleatório truncado segue distribuição normal aleatoriamente truncada (2.16) e, portanto, consideramos o modelo de regressão não linear normal aleatoriamente truncado misto (3.12). A simulação do conjunto de dados aqui apresentado segue as mesmas configurações descritas nas Seções 7 e 7.1.

A perturbação no conjunto de dados simulados foi causada como segue: primeiro, geramos as observações da v.a. de truncamento superior $\boldsymbol{B}\,|(\boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{A})$ e as observações da v.a. de truncamento inferior $\boldsymbol{A}\,|(\boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{A})$; em seguida, após gerar os valores observados do vetor de respostas truncada $\boldsymbol{Y}\,|(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A} < \boldsymbol{Y} < \boldsymbol{A}, \boldsymbol{u})$, a observação $2, 28$ foi perturbada fazendo-se $y_{2,28} = y_{2,28} + 2, 5sd(\boldsymbol{y})$. Além disso, a observação $1, 15$ também foi perturbada pela modificação dos seus limites inferiores e superiores de truncamento $a_{1,15} =$ e $b_{1,15} =$ para $a_{1,15} = 0, 1$ e $b_{1,15} = 0, 7$.

Pelos resíduos preditivos a *posteriori* padronizados, mostrados na Figura 7.2a, e pelos resíduos baseados na distribuição a *posteriori* dos parâmetros padronizados, apresentados na Figura 7.2b, é possível observar que os casos $1, 15$ e $2, 28$ são corretamente identificados como discrepantes. Pela Figura 7.2c vemos que a calibração da divergência de KL também identifica os casos $1, 15$ ($p_{1,15} = 0, 997$) e $2, 28$ ($p_{2,28} = 0, 999$) como influentes.

O resumo a *posteriori* do ajuste do modelo é apresentado na Tabela 7.9. Nota-se que os parâmetros relacionados ao parâmetro de escala da distribuição da variável resposta truncada não são estimados com precisão. As estimativas Bayesianas dos parâmetros σ e α estão longe de seus valores reais, e no caso do parâmetro α, o mesmo é estimado com sinal inverso. Além disso, os intervalos de credibilidade de 95% inter-quantil e HPD de σ e α não contêm seus verdadeiros valores.

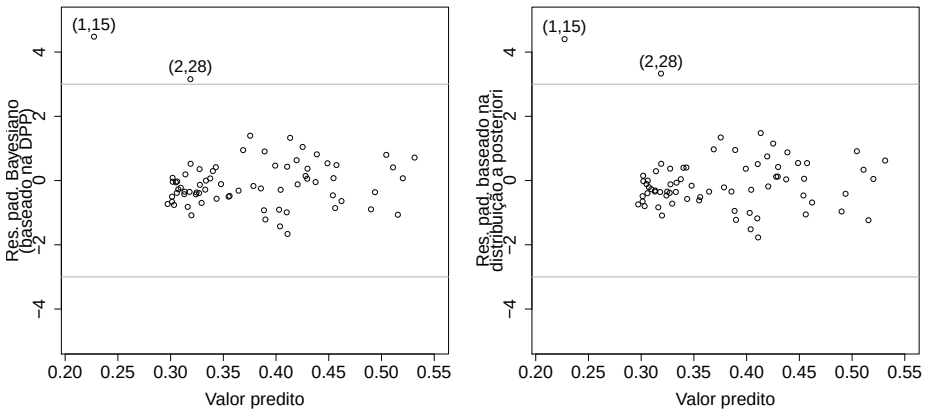
Na Tabela 6.14, apresentamos as estimativas Bayesianas dos parâmetros do modelo ajustado aos dados sem as observações identificadas como influentes. É possível observar que a remoção dos casos influentes do conjunto de dados faz com que as estimativas Bayesianas dos parâmetros tornem-se mais precisas e que os intervalos de credibilidade de 95% inter-quantil e HPD de todos os parâmetros contenham os seus verdadeiros valores. Para os parâmetros fixos do modelo, vemos que a maior discrepância entre as estimativas Bayesianas obtidas utilizando-se os dados perturbados completos, e as estimativas Bayesianas obtidas utilizando-se os dados simulados perturbados com as observações $1, 15$ e $2, 28$ excluídas, ocorrem para os parâmetros α e σ, seguidos por σ_u.

Tabela 7.9: Resumo a *posteriori* do modelo Bayesiano normal aleatoriamente truncado misto van Genuchten ajustado aos dados simulados perturbados.

Parâmetro	Valor real	Média a *posteriori*	Mediana a *posteriori*	Desv. pad.	Intervalos de credibilidade 95% Inter-quantil	Intervalos de credibilidade 95% HPD
β_1	55	54,7000	53,9226	10,8113	(36,3973; 79,5342)	(35,3416; 75,8723)
β_2	1,45	1,4473	1,4442	0,0408	(1,3790; 1,5467)	(1,3712; 1,5274)
σ	0,01	0,0328	0,0325	0,0039	(0,0266; 0,0420)	(0,0258; 0,0407)
α	-0,15	0,2202	0,2180	0,0397	(0,1542; 0,3047)	(0,1463; 0,2952)
σ_u	0,02	0,0787	0,0396	0,1037	(0,0084; 0,4289)	(0,0054; 0,3507)
u_1	-0,0209	-0,0206	-0,0211	0,0077	(-0,0348; -0,0062)	(-0,0349; -0,0062)
u_2	0,0061	-0,0014	-0,0012	0,0078	(-0,0165; 0,0132)	(-0,0167; 0,0131)
u_3	-0,0051	-0,0127	-0,0129	0,0073	(-0,0288; 0,0008)	(-0,0259; 0,0019)
μ_A	0,25	0,2488	0,2486	0,0024	(0,2442; 0,2535)	(0,2441; 0,2535)
σ_A	0,01	0,0203	0,0202	0,0017	(0,0172; 0,0241)	(0,0168; 0,0236)
μ_B	0,5	0,4943	0,4945	0,0065	(0,4808; 0,5061)	(0,4813; 0,5067)
σ_B	0,05	0,0501	0,0494	0,0043	(0,0433; 0,0591)	(0,0429; 0,0586)

Tabela 7.10: Diferenças relativas (dados simulados perturbados - modelo Bayesiano normal aleatoriamente truncado misto van Genuchten).

Parâmetro	β_1	β_2	σ	α	σ_u	u_1	u_2	u_3
Valor real	55	1,45	0,01	-0,15	0,02	-0,0209	0,0061	-0,0051
Média a *posteriori* (dados completos)	54,7000	1,4473	0,0328	0,2202	0,0787	-0,0206	-0,0014	-0,0127
Média a *posteriori* (casos $1, 15$ e $2, 28$ removidos)	53,2290	1,4637	0,0125	-0,0929	0,1591	0,0396	0,0343	0,0115
Diferença relativa (%)	3	1	62	142	51	152	104	190
Mediana a *posteriori* (dados completos)	53,9226	1,4442	0,0325	0,2180	0,0396	-0,0211	-0,0012	-0,0129
Mediana a *posteriori* (casos $1, 15$ e $2, 28$ removidos)	52,6362	1,4594	0,0124	-0,0925	0,0811	0,0398	0,0344	0,0119
Diferença relativa (%)	2	1	62	142	51	153	104	192



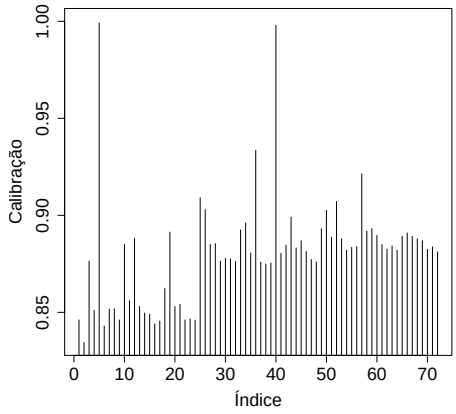


Figura 7.2: Dados simulados perturbados - modelo Bayesiano normal aleatoriamente truncado misto van Genuchten: (a) resíduo preditivo a *posteriori* padronizado; (b) resíduo baseado na distribuição a *posteriori* dos parâmetros padronizado; (c) calibração de casos.

7.5 Seleção de modelos aplicada a dados simulados: modelo de mistura Bayesiano

A seguir, exemplificamos o uso e desempenho das métricas Bayesianas de seleção de modelos descritas na Seção 5.3. No Apêndice D.3, apresentamos os códigos em `R` da função implementada para a seleção de modelos baseada em mistura de modelos Bayesianos.

O conjunto de observações de uma variável resposta com limites de truncamento fixos e conhecidos foi gerada assumindo-se que cada $Y_{i,jk}\,|(a < Y_{i,jk} < b, u_i)$ segue distribuição $NT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma x_{i,jk}^{\alpha}, a, b\right)$, $i = 1, \ldots, M$, $j = 1, \ldots, N_i$, $k = 1, \ldots, n_{ij}$, com $a = 0,31$, $b = 0,55$ e $\mu_{i,jk} = \eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i$, sendo $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ a expressão de van Genuchten (1980)-Mualem (1976) dada em (1.2)-(1.4) e $u_i \sim N\left(0, \sigma_u\right)$, $i = 1, 2, 3$. Os valores dos parâmetros usados na simulação são $\beta_1 = 55$ e $\beta_2 = 1,45$, $\sigma = 0,01$, $\alpha = -0,15$ e $\sigma_u = 0,1$.

Após a geração do conjunto de dados, três outros modelos, além do considerado para simular os dados, foram ajustados aos mesmos; são eles: $Y_{i,jk}\,|(\theta_r < Y_{i,jk} < \theta_s, u_i) \sim BT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, a, b\right)$, sendo $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ a expressão de Gardner (1958) dada em (1.1); $Y_{i,jk}\,|(a < Y_{i,jk} < b, u_i) \sim NT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, a, b\right)$, sendo $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ a expressão de van Genuchten (1980)-Mualem (1976) dada em (1.2)-(1.4); $Y_{i,jk}\,|(a < Y_{i,jk} < b, u_i) \sim BT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, a, b\right)$, sendo $\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$ a expressão de Gardner (1958) dada em (1.1).

O valor da soma de log-CPO de cada modelo ajustado e a estimativa Bayesiana do vetor de probabilidades de mistura $\tilde{\boldsymbol{\rho}}$ são mostrados na Tabela 7.11. Pelos resultados obtidos, o modelo Bayesiano normal truncado misto van Genuchten heteroscedástico é indicado como o melhor ajustado aos dados simulados por ambos os critérios considerados. Isto é, o resultado da simulação indica que tanto o critério da soma de log-CPO quanto o critério de seleção baseado em modelos de mistura Bayesiano são capazes de identificar o modelo correto (do qual os dados foram gerados) como o mais adequado para o conjunto de dados.

Tabela 7.11: Valores dos critérios de seleção de modelos Bayesianos ajustados aos dados simulados.

Modelo	log-CPO	$\tilde{\rho}$
Normal truncado misto Gardner	292,33	0,17
Normal truncado misto van Genuchten	329,36	0,28
Normal truncado misto van Genuchten heteroscedástico*	331,22	0,30
Beta truncado misto Gardner	318,52	0,25

*modelo simulado.

7.6 Algumas comparações entre os resultados de simulação obtidos das metodologias frequentista e Bayesiana

Os estudos de simulação apresentados nas Seções 6.1 e 7.1.2 indicam que as estimativas Bayesianas do modelo de regressão não linear aleatoriamente normal truncado possuem melhores propriedades frequentistas do que os EMVs. Pelos valores de vício e EQM apresentados nas Tabelas 6.6 e 7.4 - compilados na Tabela 7.12 para melhor apreciação - nota-se que o procedimento Bayesiano é capaz de tornar as estimativas dos parâmetros muito mais precisas, ao mesmo tempo em que produz intervalos cuja credibilidade estimada se aproxima da nominal. Esses mesmos comentários são válidos para o modelo não linear beta aleatoriamente truncado, cujos resultados da simulação frequentista e Bayesiana apresentados nas Tabelas 6.12 e 7.8 das Seções 6.2 e 7.2.2 são compilados na Tabela 7.13. Além disso, o procedimento Bayesiano melhora consideravelmente a estimação do parâmetro σ_u relacionado aos efeitos aleatórios não observáveis e os seus intervalos de credibilidade estimados contêm o verdadeiro valor do parâmetro na proporção esperada. Ainda sobre o parâmetro σ_u, ressaltamos que o mesmo é estimado com base na informação de apenas $M = 3$ efeitos aleatórios, e esta poderia ser uma possível causa para os problemas identificados nos resultados de simulação obtidos.

Tabela 7.12: Resultados da simulação (frequentista e Bayesiana) para o modelo normal aleatoriamente truncado misto van Genuchten-Mualem.

		MV	Bayesiano		MV	Bayesiano		Probabilidade de cobertura			
n	Parâmetro	Vício	Vício-m	Vício-md	EQM	EQM-m	EQM-md	IC de Wald	IC-RV	Inter-quantil	HPD
720	β_1	-1,03E-01	2,18E-02	2,18E-02	2,75E+00	2,84E+00	2,84E+00	0,94	0,68	0,93	0,92
	β_2	-1,86E-04	3,94E-05	3,94E-05	1,10E-04	1,09E-04	1,11E-04	0,94	0,38	0,97	0,96
	σ	-3,79E-05	7,47E-05	7,47E-05	1,20E-07	9,97E-08	9,86E-08	0,91	0,78	0,95	0,95
	α	-2,60E-04	1,96E-03	1,96E-03	1,89E-04	1,79E-04	1,83E-04	0,96	0,86	0,94	0,94
	σ_u	-3,46E-02	2,89E-02	2,89E-02	1,34E-03	7,74E-03	3,75E-03	0,18	0,28	1,00	0,99
	u_1	-2,45E-05	-8,90E-05	-8,90E-05	2,79E-04	3,23E-04	3,23E-04	0,92	0,64	0,96	0,94
	u_2	-7,55E-05	-2,11E-05	-2,11E-05	3,02E-04	2,26E-04	2,26E-04	0,95	0,69	0,96	0,95
	u_3	-8,13E-05	-7,99E-05	-7,99E-05	3,35E-04	3,01E-04	3,01E-04	0,94	0,63	0,97	0,98
	μ_A	-1,21E-05	7,38E-07	7,38E-07	1,38E-07	1,56E-07	1,56E-07	0,97	0,76	0,93	0,91
	σ_A	-9,68E-06	-9,33E-07	-9,33E-07	7,11E-08	8,94E-08	8,93E-08	0,95	0,88	0,92	0,91
	μ_B	-1,32E-04	-4,11E-05	-4,11E-05	3,72E-06	2,54E-06	2,56E-06	0,93	0,91	0,98	0,97
	σ_B	4,13E-05	5,00E-05	5,00E-05	1,61E-06	1,69E-06	1,68E-06	0,96	0,97	0,96	0,95

Tabela 7.13: Resultados da simulação (frequentista e Bayesiana) para o modelo normal aleatoriamente truncado misto van Genuchten-Mualem.

		MV	Bayesiano		MV	Bayesiano		Probabilidade de cobertura			
n	Parâmetro	Vício	Vício-m	Vício-md	EQM	EQM-m	EQM-md	IC de Wald	IC-RV	Inter-quantil	HPD
	β_1	2,25E+00	1,98E-02	1,98E-02	5,63E-04	1,84E-02	1,81E-02	1,63E-02	0,35	0,92	0,91
	β_2	5,01E-01	1,56E-03	1,56E-03	9,79E-04	6,75E-05	6,58E-05	8,38E-05	0,75	0,96	0,96
	σ	6,01E+00	-6,40E-03	-6,40E-03	9,21E-03	2,71E-03	2,74E-03	3,28E-03	0,91	0,97	0,97
	σ_u	4,01E-02	9,94E-02	9,94E-02	-5,99E-02	4,50E-02	3,79E-02	3,78E-03	0,71	1,00	1,00
720	u_1	4,09E-03	-1,21E-03	-1,21E-03	-2,17E-04	1,86E-03	1,85E-03	1,71E-03	0,60	0,90	0,89
	u_2	4,84E-03	2,40E-03	2,40E-03	-1,19E-04	2,14E-03	2,14E-03	1,77E-03	0,56	0,91	0,92
	u_3	4,24E-03	-1,44E-03	-1,44E-03	-2,31E-04	1,92E-03	1,91E-03	1,87E-03	0,63	0,91	0,90
	μ_A	2,50E-01	1,15E-04	1,15E-04	-7,57E-05	1,45E-06	1,44E-06	1,61E-06	0,84	0,94	0,95
	σ_A	5,21E+00	-2,89E-04	-2,89E-04	9,53E-03	2,70E-03	2,70E-03	2,93E-03	0,95	0,96	0,96
	μ_B	6,00E-01	9,25E-05	9,25E-05	-1,48E-05	3,46E-06	3,49E-06	3,67E-06	0,92	0,96	0,97
	σ_B	4,61E+00	-2,80E-03	-2,80E-03	1,08E-02	2,67E-03	2,68E-03	2,53E-03	0,97	0,95	0,94

Capítulo 8

Aplicação

Para ilustrar a metodologia proposta, analisamos um perfil de solo da base de dados coletada na Bacia do Rio Buriti Vermelho, localizada na parte oriental do Distrito Federal, Brasil (Rodrigues & Maia, 2011). Esta base de dados contém medições de teor de umidade de água em amostras de solos coletadas em profundidades de $0-5cm$, $15-20cm$, e $60-65cm$ medidas em oito níveis de tensões de sucção, x, variando de $0,01$ a 10 unidades de atmosfera (atm), $x = (0,01; 0,03; 0,06; 0,10; 0,33; 0,80; 4,00; 10,00)$. Cada amostra de solo foi medida em três replicações por nível, o que resulta em um total de 24 medições de teor de umidade para cada nível de profundidade. Isto é, temos $i = 1, 2, 3$, $j = 1, 2, 3$ e $k = 1, \ldots, 8$, com $M = 3$, $N_i = 3$ e $n_{ij} = 8$ e $n = \sum_{i=1}^{M} \sum_{j=1}^{N_i} n_{i,j} = 72$, para cada um dos 17 perfis de solo, que representam os locais onde as amostras de solos foram coletadas ao longo da região considerada. Para cada um dos 17 perfis e em cada profundidade, o teor de umidade do solo saturado, θ_s, foi calculado na tensão de $0atm$ e o teor de umidade residual, θ_r, foi calculado na tensão de $15atm$. Remetemos os leitores interessados a Rodrigues & Maia (2011) para mais detalhes sobre os procedimentos experimentais de coleta das amostras de solo, bem como dos procedimentos de análises laboratoriais usados para medir os teores de umidades das mesmas.

Mais uma vez, ressaltamos que em dados de retenção de água em solo o teor de umidade de uma amostra de solo é limitado inferiormente pela umidade residual, θ_r, e é limitada superiormente pela umidade saturada, θ_s. Isto é, a variável resposta truncada $Y\,|(A, B, A < Y < B)$ corresponde ao teor de umidade, e as variáveis de truncamento inferior e superior são $\theta_r = A\,|(B, A < Y < B)$ e $\theta_s = B\,|(A < Y < B)$.

8.1 Metodologia frequentista: limites de truncamento fixos e conhecidos

Em dados de retenção de água em solo, os limites de truncamento, θ_r e θ_s, são usualmente medidos para cada amostra de solo em cada profundidade. Entretanto, na maioria

dos estudos de estimação de curvas de retenção, é comum que os pesquisadores mantenham um registro dos valores médios observados para o conteúdo de água residual, θ_r, e para o conteúdo de água no solo saturado, θ_s. Quando os dados de retenção de água são armazenados desta maneira, não é possível considerarmos os limites de truncamento como realizações de v.a.'s de truncamento e devemos, então, considerar os limites como sendo fixos conhecidos e constantes.

Para ilustrar o caso do estudo de dados de retenção de água sujeitos a limites fixos e conhecidos, usamos o modelo não linear truncado misto (3.9) e apresentamos os resultados obtidos ajustando tal modelo a um conjunto de dados para o qual os valores observados dos limites de truncamento foram substituídos pelas médias dos valores observados. Em outras palavras, consideramos um conjunto de dados com os limites de truncamento transformados substituindo-se $\theta_{r_{i,j}}$ e $\theta_{s_{i,j}}$ por $\bar{\theta}_r = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\theta_{r_{i,j}}/\sum_{i=1}^{N_i} M$ e $\bar{\theta}_s = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\theta_{s_{i,j}}/\sum_{i=1}^{M} N_i$, $i = 1, \ldots, M$ e $j = 1, \ldots, N_i$. Neste cenário, a variável resposta truncada é denotada por $Y\,|(a < Y < b)$ e os limites de truncamento inferior e superior são $\bar{\theta}_r = a$ e $\bar{\theta}_s = b$.

Nove possíveis modelos foram considerados na análise. Esses modelos são descritos na Tabela 8.1. As expressões das curvas de Gardner (1958), van Genuchten (1980)-Burdine (1953) e van Genuchten (1980)-Mualem (1976) são dadas nas expressões (1.1), (1.2)-(1.4) e (1.2)-(1.3) apresentadas na Seção 1.1.

Os valores de AIC e BIC, mostrados na Tabela 8.2, indicam o modelo normal truncado misto com parâmetro de locação relacionado à expressão de van Genuchten (1980)-Burdine (1953) e aos efeitos aleatórios, e uma estrutura homoscedástica para o parâmetro de dispersão, denotado por NT vGM Hm, como o melhor modelo ajustado aos dados do perfil 219 com os limites transformados.

O resumo do ajuste do modelo NT vGB Hm, apresentado na Table 8.3, mostra todos os parâmetros fixos como estatisticamente significativos com 95% de confiança. O efeito aleatório u_2 relacionado à variação não observável associada com a profundidade $15-20cm$ é o único destacado como estatisticamente significativo com 95% confiança.

A Figura 8.1 mostra as curvas de retenção estimadas para cada profundidade. É possível notar que o conteúdo de água no solo tende a ser maior na profundidade $15-20cm$ e menor na profundidade $60-65cm$. Para as profundidades $15-20cm$ e $60-65cm$ nota-se uma razoável heterogeneidade dos valores de conteúdo de água entre as replicatas. Também podemos observar que a profundidade $0-5cm$ parece ser a com menor heterogeneidade entre as replicatas e que os seus valores de conteúdo de água estão entre as outras duas profundidades. Como esses dados foram obtidos de experimentos laboratoriais utilizando rigorosos processos de medições, as diferenças entre as observações medidas em profundidades diferentes podem estar relacionada com as características físicas do solo, que variam de profundidade para profundidade. Notamos, ainda, que o modelo NT vGB Hm parece fornecer uma boa estimativa para a relação entre o conteúdo de água e o

potencial matricial do solo, usando a expressão de van Genuchten (1980)-Mualem (1976).

Os resíduos padronizados são mostrados na Figura 8.2a. Nota-se que os mesmos estão distribuídos aleatoriamente em torno do zero e que a observação $2,31$ é destacada como *outlier*. Na Figura 8.2b é apresentado o *qqplot* dos resíduos padronizados com envelopes simulados, onde é possível verificar que, apesar da presença do caso discrepante $2,31$, não parece haver falta de ajuste do modelo selecionado. A observação $2,31$ é indicada como influente pela distância generalizada de Cook, na Figura 8.3a, e pela distância da verossimilhança, apresentada na Figura 8.3b.

As métricas de influência local são mostradas na Figura 8.4a e na Figura 8.4b. Sob o esquema de perturbação de casos, a observação $2,31$ é indicada como tendo uma influência mais proeminente nas estimativas dos parâmetros. Por outro lado, sob o esquema de perturbação de casos (Figura 8.4b) nenhuma observação é indicada como influente.

Como estamos lidando com um conjunto de dados reais, a remoção de casos influentes deve ser friamente analisada. Notamos, por exemplo, que a coleta de mais dados talvez pudesse revelar a observação como não sendo influente de fato. Além disso, os dados de retenção de água são obtidos a partir de procedimentos de laboratório altamente precisos e bem calibrados. Por outro lado, nós suspeitamos que a observação $2,31$ está sendo indicada como um *outlier* influente porque seus limites de truncamento inferior e superior foram substituídos pelos valores médios observados no conjunto.

Os dados da observação $2,31$ usados para ajustar os modelos considerados foram $(x_{2,11}, y_{2,11}, a, b) = (0.010, 0.536, 0.271, 0.531)$. Entretanto, os valores originais de $a_{2,11}$ e $b_{2,11}$ eram 0.293 e 0.597, respectivamente, e, como foi exemplificado na Seção 6.3, é possível que uma observação seja convertida erroneamente em discrepante quando seus valores de truncamento são incorretamente registrados. Portanto, o pesquisador precisa estar ciente de que os dados com limites transformados contêm observações que podem ser indicadas como tendo uma influência mais predominante nas estimativas dos parâmetros e que isto pode afetar a curva de retenção ajustada.

Na Tabela 8.4 são apresentadas as diferenças relativas entre os EMVs obtidos com o conjunto de dados completo e os EMVs obtidos removendo-se a observação $2,31$ do conjunto. Podemos notar que os parâmetros são altamente sensíveis à presença da observação influente, sendo que a maior diferença relativa, em %, foi observada para o parâmetro σ, seguido por β_1. A menor diferença ocorre para σ_u. As diferenças relativas foram calculadas de $|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(-i)}|/\max\{|\hat{\boldsymbol{\theta}}|, |\hat{\boldsymbol{\theta}}_{(-i)}|\}100\%$.

Tabela 8.1: Modelos não lineares truncados mistos ajustados aos dados do perfil 219 com limites transformados.

Modelo	Distribuição de $Y_{i,jk}\,\vert(a < Y_{i,jk} < b, u_i)$	$\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$
NT GA Hm		Gardner (1958)
NT vGB Hm	$NT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, a, b\right)$	van Genuchten (1980)-Burdine (1953)
NT vGM Hm		van Genuchten (1980)-Mualem (1976)
NT GA Ht		Gardner (1958)
NT vGB Ht	$NT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma x_{i,jk}^{\alpha}, a, b\right)$	van Genuchten (1980)-Burdine (1953)
NT vGM Ht		van Genuchten (1980)-Mualem (1976)
BT GA Hm		Gardner (1958)
BT vGB Hm	$BT\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, a, b\right)$	van Genuchten (1980)-Burdine (1953)
BT vGM Hm		van Genuchten (1980)-Mualem (1976)

Tabela 8.2: Valores de AIC e BIC dos modelos ajustados aos dados do perfil 219 com limites transformados.

Modelo	AIC	BIC
NT GA Hm	-402,42	-386,48
NT vGB Hm	-418,98	-403,04
NT vGM Hm	-416,90	-400,96
NT GA Ht	-404,44	-386,23
NT vGB Ht	-416,99	-398,78
NT vGM Ht	-414,90	-396,69
BT GA Hm	-403,23	-387,29
BT vGB Hm	-418,36	-402,42
BT vGM Hm	-416,41	-400,47

Tabela 8.3: Resumo do ajuste do modelo NT vGB Hm aos dados do perfil 219 com limites transformados.

Parâmetro	EMV	Desv. pad.	IC de Wald 95%	IC-RV 95%
β_1	175,9011	23,2905	(130,2527; 221,5496)	(158,3132; 195,6504)
β_2	2,2559	0,0201	(2,2165; 2,2953)	(2,2461; 2,2662)
σ	0,0137	0,0012	(0,0114; 0,0161)	(0,0117; 0,0164)
σ_u	0,0125	0,0072	(-0,0016; 0,0266)	(0,0066; 0,0369)
u_1	-0,0097	0,0084	(-0,0261; 0,0068)	(-0,0151; -0,0043)
u_2	0,0066	0,0078	(-0,0087; 0,0219)	(0,0013; 0,0120)
u_3	-0,0182	0,0088	(-0,0354; -0,0010)	(-0,0236; -0,0127)

Tabela 8.4: Dados do perfil 219 com limites transformados - modelo NT vGM Hm: diferença relativa.

Parâmetro	β_1	β_2	σ	σ_u
EMV (dados completos)	145,1977	1,2823	0,0141	0,0103
EMV (dados com a observação $2,31$ deletada)	155,4716	1,2704	0,0130	0,0103
Diferença relativa (%)	6,6082	0,9256	7,3555	0,4401

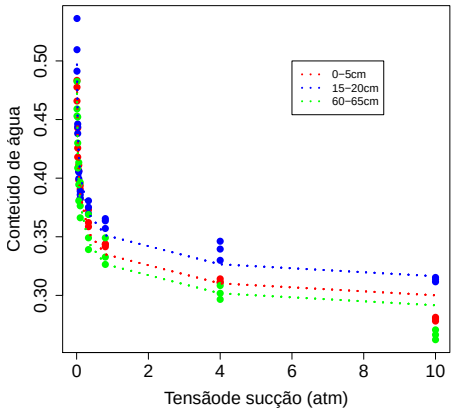
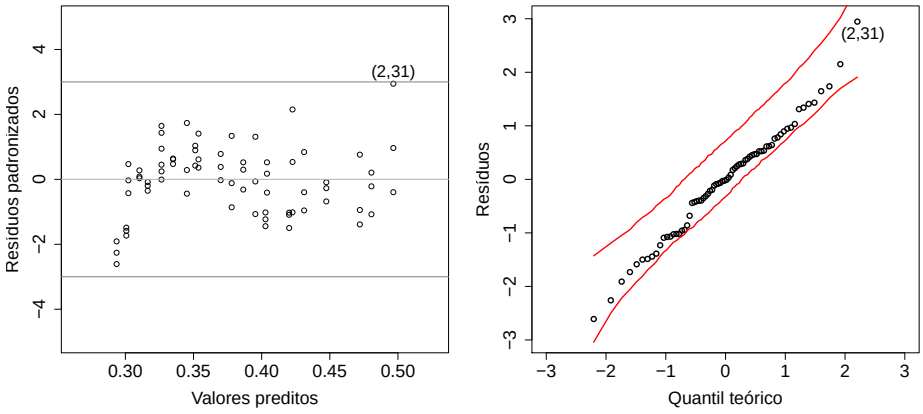
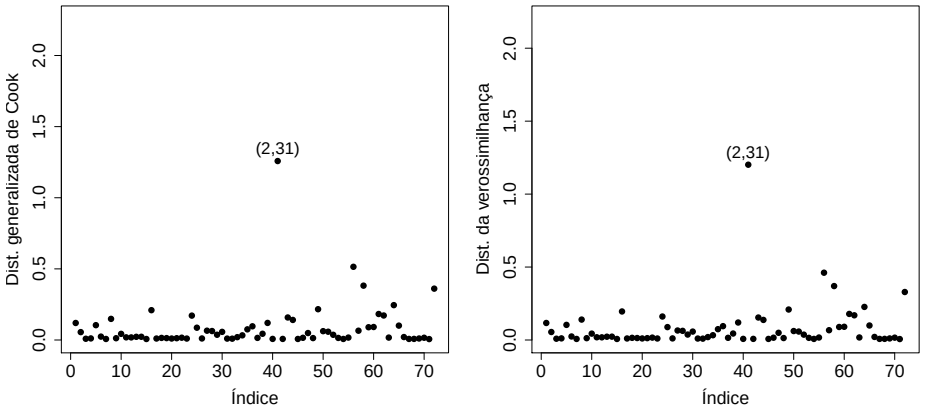


Figura 8.1: Dados do perfil 219 com limites transformados - modelo NT vGM Hm: curva de retenção estimada.



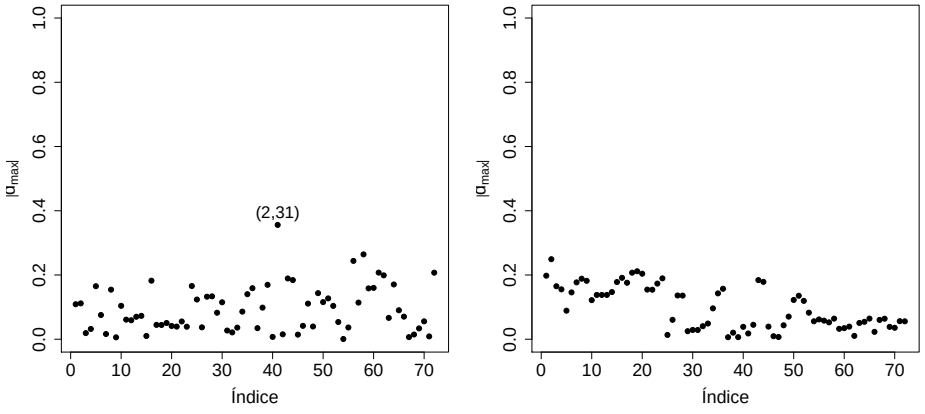
(a) (b)

Figura 8.2: Dados do perfil 219 com limites transformados - modelo NT vGM Hm: (a) resíduos padronizados; (b) *qqplot* com envelope simulado.



(a) (b)

Figura 8.3: Dados do perfil 219 com limites transformados - modelo NT vGM Hm: (a) distância generalizada de Cook; (b) distância da verossimilhança.



(a) (b)

Figura 8.4: Dados do perfil 219 com limites transformados - modelo NT vGM Hm: (a) influência local sob perturbação da resposta; (b) influência local sob perturbação de casos.

8.2 Metodologia frequentista: limites de truncamento aleatórios

O ajuste de curvas de retenção, usando os modelos não lineares aleatoriamente truncados mistos propostos no Capítulo 3, é ilustrado considerando o conjunto de dados do perfil 219 original. Isto é, consideramos os dados do perfil tal qual os mesmos foram observados, sem que os valores de $\theta_{r_{i,j}}$ e $\theta_{s_{i,j}}$ tenham sido substituídos por seus valores médios observados.

Ressaltamos que a análise desse mesmo conjunto de dados com os limites transformados foi apresentada na Seção 8.1, pois os dados de retenção de água usados para construir curvas de retenção são normalmente tratados dessa forma e, portanto, nosso interesse principal era investigar as consequências que dita abordagem poderiam causar na estimação do modelo. Além disso, quando os dados são registrados desta forma, não é possível construir um modelo que leve em consideração as variações nos valores de água residual e água saturada que ocorrem entre as diferentes profundidades de solo. Entretanto, na base de dados da Bacia do Rio Buriti Vermelho, os valores de θ_r e θ_s foram medidos e registrados, o que nos permite considerar um modelo mais apropriado e completo que leva em consideração tanto a natureza truncada da variável resposta quanto o processo de truncamento da mesma e a variação intrínseca dos seus limites que está relacionada, entre outras coisas, com as características físicas e as propriedades hidráulicas do tipo de solo de cada profundidade. Desta forma, a variável resposta truncada $Y\,|(A, B, A < Y < B)$ corresponde ao teor de umidade, e as variáveis de truncamento inferior e superior são $\theta_r = A\,|(B, A < Y < B)$ e $\theta_s = B\,|(A < Y < B)$.

Os nove modelos ajustados aos dados do perfil 219 são descritos na Tabela 8.5 e seus valores de AIC e BIC são apresentados na Tabela 8.6. Nota-se que ambos os critérios indicam o modelo não linear beta aleatoriamente truncado misto, com parâmetro de locação associado à expressão de Gardner (1958) e aos efeitos aleatórios e com uma estrutura homoscedástica para o parâmetro de dispersão (modelo BTA GA Hm), como o de melhor ajuste aos dados. Ainda na Tabela 8.6, as células marcadas com (-) indicam os modelos para os quais não houve convergência na maximização da função de log-verossimilhança.

O resumo do ajuste do modelo BTA GA Hm é apresentado na Tabela 8.7 que mostra todos os parâmetros fixos como estatisticamente significativos com 95% de confiança. O efeito aleatório u_3 relacionado a variação não observável associada com a profundidade $60 - 65$ é o único destacado como estatisticamente significativo com 95% de confiança.

Na Figura 8.5, temos a curva de retenção estimada pelo modelo BTA GA Hm para cada replicata em cada profundidade. Podemos notar que o conteúdo de água no solo tende a ser maior na profundidade $15 - 20cm$ e menor na profundidade $60 - 65cm$, e que a profundidade $0 - 5cm$ apresenta valores de conteúdo de água entre as outras duas profundidades. Para as profundidades $15 - 20cm$ e $60 - 65cm$ nota-se uma razoável

heterogeneidade dos valores de conteúdo de água entre as replicatas e a profundidade $0 - 5cm$ parece ser a com menor heterogeneidade entre as replicatas. Entretanto, o modelo BTA GA Hm parece fornecer uma boa estimativa para a relação entre o conteúdo de água e o potencial matricial do solo usando-se a expressão de Gardner (1958).

Os resíduos padronizados contra os valores preditos são mostrados na Figura 8.6a, onde podemos notar que os mesmos estão aleatoriamente distribuídos em torno do zero e que nenhuma observação se destaca como *outlier*. Na Figura 8.6b, apresentamos *qqplot* dos resíduos padronizados com envelopes simulados, que não indica a presença de casos discrepantes e nem de falta de ajuste do modelo selecionado. As métricas de influência global baseadas na distância generalizada de Cook e na distância da verossimilhança, apresentadas nas Figuras 8.6a e 8.7b, não indicam nenhum caso como influente. Além disso, a métrica de influência local baseada no esquema de perturbação da resposta (Figura 8.8a) e a métrica de influência local baseada na perturbação de casos (Figura 8.8b), não indicam nenhuma observação como tendo um efeito mais pronunciado nas estimativas dos parâmetros. Até mesmo a observação $2,31$, que havia sido indicada como influente na análise dos dados com limites transformados, já não é indicada como influente na Figura 8.8a. Acreditamos que, considerar os limites de truncamento como v.a.'s e levar essa característica em consideração ao construir o modelo nos fornece um ajuste mais adequado, e por sua vez, acomoda os dados observados com maior precisão.

Tabela 8.5: Modelos não lineares aleatoriamente truncados mistos ajustados aos dados do perfil 219.

Modelo	Distribuição de $(Y_{i,jk}, A_{i,jk}, B_{i,jk})\,\vert(A_{i,jk}, B_{i,jk}, A_{i,jk} < Y_{i,jk} < B_{i,jk}, u_i)$	$\eta\left(x_{i,jk}, \boldsymbol{\beta}\right)$
NTA GA Hm		Gardner (1958)
NTA vGB Hm	$NTA\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, \mu_A, \sigma_A, \mu_B, \sigma_B\right)$	van Genuchten (1980)-Burdine (1953)
NTA vGM Hm		van Genuchten (1980)-Mualem (1976)
NTA GA Ht		Gardner (1958)
NTA vGB Ht	$NTA\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma x_{i,jk}^{\alpha}, \mu_A, \sigma_A, \mu_B, \sigma_B\right)$	van Genuchten (1980)-Burdine (1953)
NTA vGM Ht		van Genuchten (1980)-Mualem (1976)
BTA GA Hm		Gardner (1958)
BTA vGB Hm	$BTA\left(\eta\left(x_{i,jk}, \boldsymbol{\beta}\right) + u_i, \sigma, \mu_A, \sigma_A, \mu_B, \sigma_B\right)$	van Genuchten (1980)-Burdine (1953)
BTA vGM Hm		van Genuchten (1980)-Mualem (1976)

Tabela 8.6: Valores de AIC e BIC dos modelos ajustados aos dados do perfil 219.

Model	AIC	BIC
NTA GA Hm	-996,44	-971,40
NTA vGB Hm	-974,19	-949,15
NTA vGM Hm	-979,65	-954,61
NTA GA Ht	-	-
NTA vGB Ht	-951,09	-923,77
NTA vGM Ht	-957,39	-930,07
BTA GA Hm	-1000,35	-975,30
BTA vGB Hm	-979,85	-954,81
BTA vGM Hm	-985,17	-960,13

Tabela 8.7: Resumo do ajuste do modelo BTA GA Hm aos doados do perfil 219.

Parâmetro	EMV	Desv. pad.	IC de Wal de 95%	IC-RV 95%
β_1	4,6716	1,0138	(2,6846; 6,6587)	(4,3384; 5,0340)
β_2	0,5310	0,0362	(0,4601; 0,6018)	(0,5058; 0,5566)
σ	6,8131	0,1766	(6,4671; 7,1592)	(6,4543; 7,1321)
σ_u	0,0190	0,0094	(0,0006; 0,0374)	(0,0101; 0,0561)
u_1	0,0116	0,0082	(-0,0045; 0,0276)	(0,0052; 0,0179)
u_2	-0,0016	0,0083	(-0,0179; 0,0147)	(-0,0082; 0,0049)
u_3	0,0308	0,0075	(0,0160; 0,0455)	(0,0245; 0,0371)
μ_A	0,2713	0,0022	(0,2671; 0,2755)	(0,2671; 0,2756)
σ_A	6,3812	0,1666	(6,0547; 6,7076)	(6,0363; 6,6907)
μ_B	0,5314	0,0049	(0,5218; 0,5410)	(0,5218; 0,5410)
σ_B	4,9669	0,1661	(4,6413; 5,2924)	(4,6229; 5,2758)

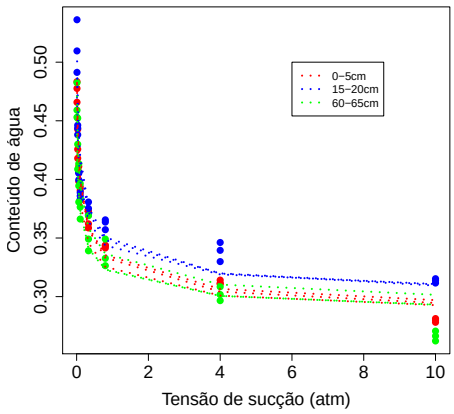
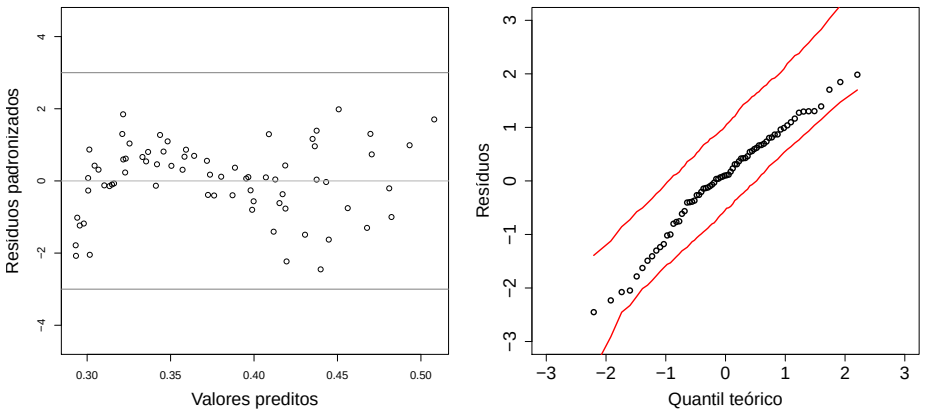
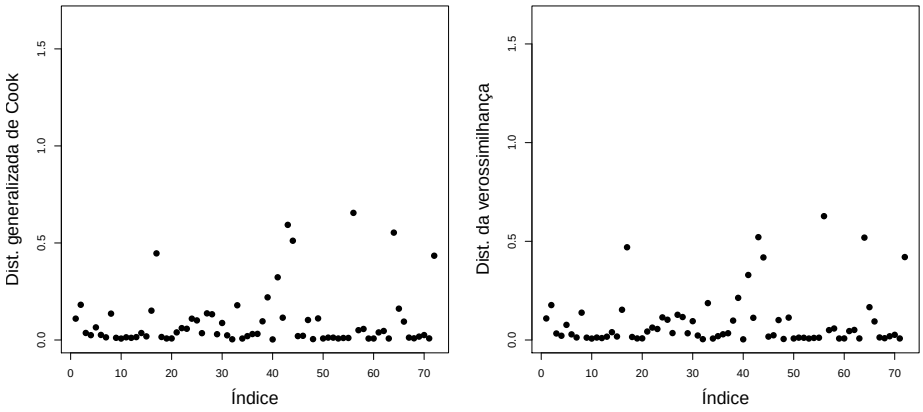


Figura 8.5: Dados do perfil 219 - modelo BTA GA Hm: curva de retenção estimada.



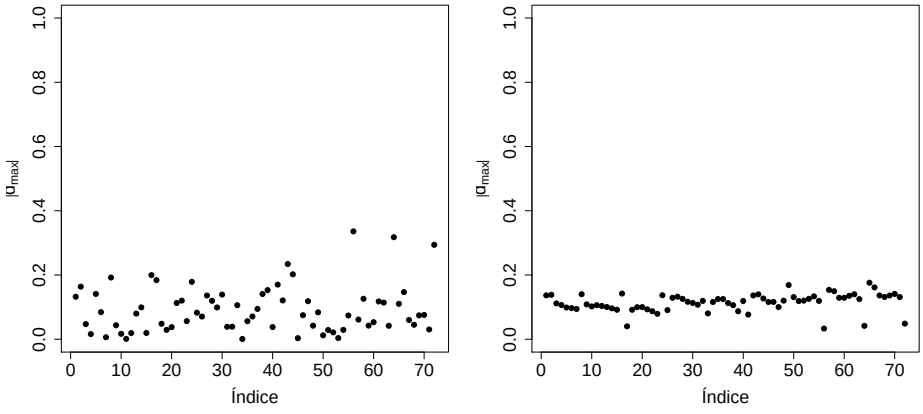
(a) (b)

Figura 8.6: Dados do perfil 219 - modelo BTA GA Hm: (a) resíduos padronizados; (b) *qqplot* com envelope simulado.



(a) (b)

Figura 8.7: Dados do perfil 219 - modelo BTA GA Hm: (a) distância generalizada de Cook; (b) distância da verossimilhança.



(a) (b)

Figura 8.8: Dados do perfil 219 - modelo BTA GA Hm: (a) influência local sob perturbação da resposta; (b) influência local sob perturbação de casos.

8.3 Metodologia Bayesiana

Para ilustrar a metodologia Bayesiana de estimação e diagnóstico dos modelos não lineares aleatoriamente truncados, consideramos o conjunto de dados do perfil 219 e ajustamos os nove modelos na Tabela 8.5 sob o enfoque Bayesiano. As distribuições a *priori* consideradas para cada um dos modelos são mostradas na Tabela 8.8.

Para cada modelo, as amostras MCMC das distribuições a *posteriori* $\pi_1(\boldsymbol{\theta}|D)$, $\pi_2(\boldsymbol{\omega}_A|D)$ e $\pi_3(\boldsymbol{\omega}_B|D)$ foram simuladas como segue: para obter uma amostra MCMC da distribuição a posteriori de $\boldsymbol{\theta}$, utilizamos um algoritmo do tipo Metropolis-Hastings com Gibbs, no qual cada um dos candidatos, para cada um dos parâmetros, é gerado de um passeio aleatório considerando-se uma distribuição normal univariada com desvio-padrão dada pelo elemento da diagonal da matriz de covariâncias definida pela negativa da matriz Hessiana avaliada nos EMVs dos parâmetros. As amostras MCMC de $\boldsymbol{\omega}_A$ e $\boldsymbol{\omega}_B$ são obtidas usando-se um algoritmo do tipo Metropolis-Hastings, no qual os candidatos são gerados por passeio aleatório considerando-se uma distribuição normal multivariada com matriz de covariâncias dada pela negativa da matriz Hessiana avaliada nos EMVs dos parâmetros. As convergências das cadeias foram verificadas utilizando-se o critério Geweke (Geweke, 1992). Os algoritmos MCMC Metropolis-Hastings com Gibbs e Metropolis-Hastings foram implementados em `R` e são exemplificados nos Apêndices D.1 e D.2, respectivamente. Os tamanhos das cadeias foram 100 mil com períodos de burn-in de 20 mil e salto de 200.

Os valores dos critérios de seleção de modelos baseados na soma de log-CPO e a estimativa Bayesiana do vetor de probabilidades de mistura $\tilde{\boldsymbol{\rho}}$ são mostrados na Tabela 8.9. Pelos resultados obtidos, o modelo Bayesiano BTA vGM Hm é indicado como o melhor ajustado aos dados do perfil 219 por ambos os critérios. O resumo a *posteriori* do ajuste do modelo Bayesiano BTA vGM Hm aos dados é apresentado na Tabela 8.10, onde é possível notar que todos os parâmetros de efeitos fixos do modelo são estatisticamente significativos com 95% de credibilidade. O efeito aleatório u_2, relacionado com a variação não observável associada com a profundidade do solo de $15 - 20cm$ é o único indicado como não significativo com 95% de credibilidade.

Os valores obtidos para p-valores preditivos a *posteriori* estimados foram de 0,45, 0,46 e 0,50 para as discrepâncias baseadas na média, variância e *deviance* do modelo Bayesiano BTA vGM Hm. Os histogramas destas discrepâncias são mostrados na Figura 8.10, onde vemos que a média e variância das amostras replicadas coincidem com a média e variância dos dados observados (Figura 8.10a e Figura 8.10b) e que o zero está contido no histograma da diferença entre a *deviance* calculada com os dados observados e a *deviance* calculada com os dados replicados (Figura 8.10c). Assim, a avaliação preditiva a *posteriori* indica um bom ajuste do modelo Bayesiano BTA vGM Hm aos dados do perfil 219.

Na Figura 8.11a e na Figura 8.11b, são mostrados os resíduos preditivos a *posteriori* padronizados e os resíduos baseados na distribuição a *posteriori* dos parâmetros padroni-

zados, respectivamente. Podemos notar que nenhuma observação é indicada como *outlier* por nenhum dos dois tipos de resíduos Bayesianos considerados. Na Figura 8.11c, são apresentados os valores das calibrações e nenhum caso é indicado como influente.

Tabela 8.8: Distribuições a *priori* consideradas para cada modelo Bayesiano ajustado aos dados do perfil 219.*

Modelo	*Priori*
NTA GA Hm	$\beta_1, \beta_2 \sim IG(\tau^{-1}, \tau^{-1})$; $\sigma \sim IG(\tau^{-1}, \tau^{-1})$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$
NTA vGB Hm	$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$; $\beta_2 \sim GT(\tau^{-1}, \tau^{-1}, 2, +\infty)$; $\sigma \sim IG(\tau^{-1}, \tau^{-1})$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$
NTA vGM Hm	$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$; $\beta_2 \sim GT(\tau^{-1}, \tau^{-1}, 1, +\infty)$; $\sigma \sim IG(\tau^{-1}, \tau^{-1})$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$
NTA GA Ht	$\beta_1, \beta_2 \sim IG(\tau^{-1}, \tau^{-1})$; $\sigma \sim IG(\tau^{-1}, \tau^{-1})$; $\alpha \sim N(0, \tau)$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$
NTA vGB Ht	$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$; $\beta_2 \sim GT(\tau^{-1}, \tau^{-1}, 2, +\infty)$; $\sigma \sim IG(\tau^{-1}, \tau^{-1})$; $\alpha \sim N(0, \tau)$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$
NTA vGM Ht	$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$; $\beta_2 \sim GT(\tau^{-1}, \tau^{-1}, 1, +\infty)$; $\sigma \sim IG(\tau^{-1}, \tau^{-1})$; $\alpha \sim N(0, \tau)$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$
BTA GA	$\beta_1, \beta_2 \sim IG(\tau^{-1}, \tau^{-1})$; $\sigma \sim N(0, \tau)$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$
BTA vGB	$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$; $\beta_2 \sim GT(\tau^{-1}, \tau^{-1}, 2, +\infty)$; $\sigma \sim N(0, \tau)$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$
BTA vGM	$\beta_1 \sim IG(\tau^{-1}, \tau^{-1})$; $\beta_2 \sim GT(\tau^{-1}, \tau^{-1}, 1, +\infty)$; $\sigma \sim N(0, \tau)$ $\sigma_u \sim Half-t(\nu, \tau)$; $u_i \sim N(0, \tau)$

$\tau = 10^4$, $\lambda = 10$ e $\nu = 1$.

Tabela 8.9: Valores dos critérios de seleção de modelos Bayesianos ajustados aos dados do perfil 219.

Modelo	log-CPO	$\tilde{\rho}$
NTA GA Hm	597,76	0,10
NTA vGB Hm	620,33	0,13
NTA vGM Hm	618,50	0,13
NTA GA Ht	-	-
NTA vGB Ht	603,20	0,11
NTA vGM Ht	621,80	0,14
BTA GA Hm	593,69	0,10
BTA vGB Hm	626,23	0,14
BTA vGM Hm	628,55	0,15

Tabela 8.10: Resumo a *posteriori* do modelo Bayesiano beta aleatoriamente truncado misto van Genuchten-Mualem ajustado aos dados do perfil 219.

Parâmetro	Média a *posteriori*	Mediana a *posteriori*	Desv pad.	Intervalos de credibilidade 95% Inter-quantil	HPD
β_1	151,7906	149,1771	23,9085	(107,9420; 201,6443)	(109,9681; 203,6075)
β_2	1,2923	1,2918	0,0279	(1,2385; 1,3500)	(1,2412; 1,3520)
σ	6,8243	6,8302	0,1762	(6,4584; 7,1371)	(6,5071; 7,1761)
σ_u	0,0631	0,0377	0,0615	(0,0112; 0,2421)	(0,0078; 0,2012)
u_1	-0,0026	-0,0023	0,0091	(-0,0219; 0,0136)	(-0,0198; 0,0146)
u_2	-0,0157	-0,0150	0,0097	(-0,0369; 0,0009)	(-0,0369; 0,0012)
u_3	0,0191	0,0194	0,0081	(0,0024; 0,0337)	(0,0028; 0,0339)
μ_A	0,2712	0,2713	0,0022	(0,2666; 0,2754)	(0,2669; 0,2754)
σ_A	6,3449	6,3503	0,1566	(6,0170; 6,6291)	(6,0102; 6,6218)
μ_B	0,5317	0,5319	0,0051	(0,5215; 0,5422)	(0,5214; 0,5420)
σ_B	4,9414	4,9427	0,1783	(4,5856; 5,2809)	(4,5781; 5,2595)

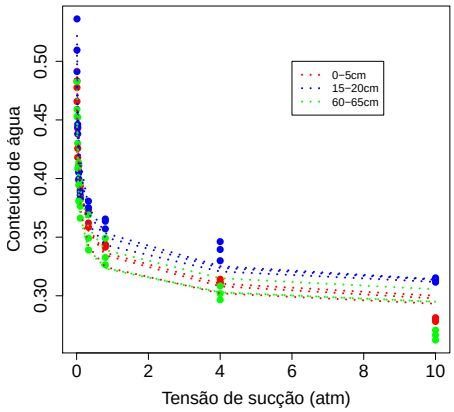
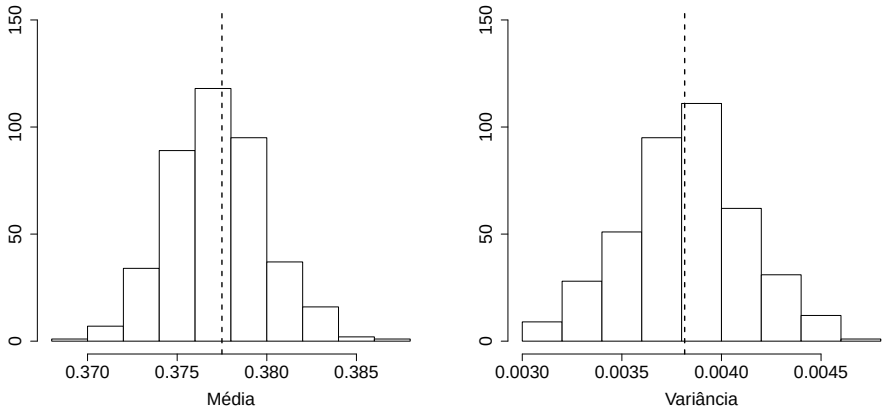
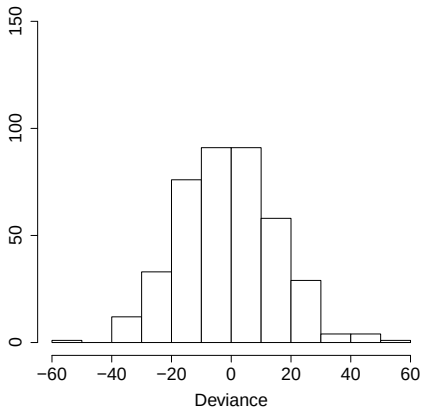


Figura 8.9: Dados do perfil 219 - modelo Bayesiano BTA vGM Hm: curva de retenção estimada.

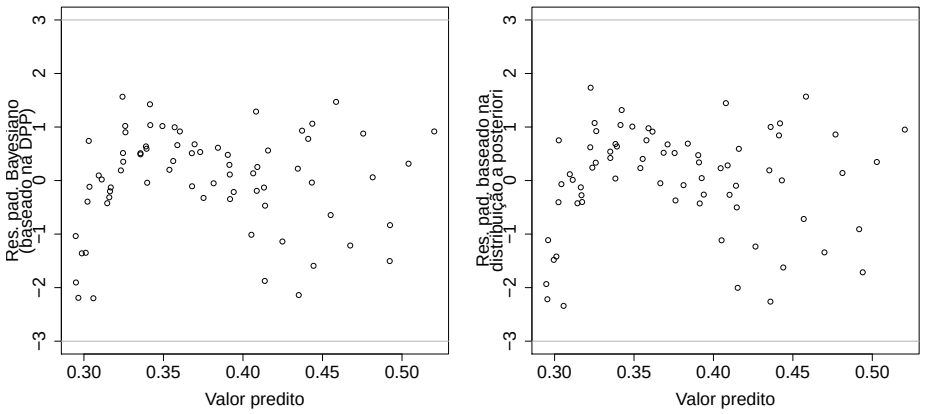


(a) (b)

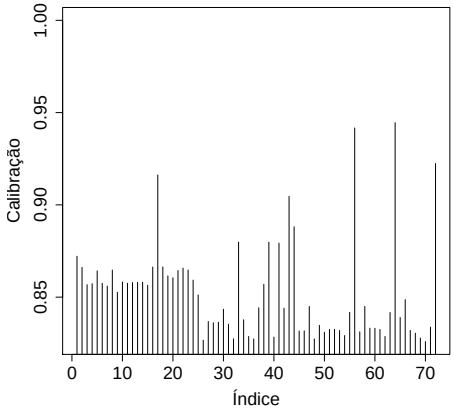


(c)

Figura 8.10: Dados do perfil 219 - modelo Bayesiano beta aleatoriamente truncado misto van Genuchten-Mualem: histogramas das avaliações preditivas a *posteriori* baseadas na média (a), variância (b) e *deviance* do modelo (c).



(a) (b)



(c)

Figura 8.11: Dados do perfil 219 - modelo Bayesiano beta aleatoriamente truncado misto van Genuchten-Mualem: (a) resíduo preditivo a *posteriori* padronizado; (b) resíduo baseado na distribuição a *posteriori* dos parâmetros padronizado; (c) calibração de casos.

8.4 Algumas comparações entre os resultados obtidos para os dados reais usando as metodologias frequentista e Bayesiana

Os resultados de estimação e diagnóstico do modelo de regressão não linear aleatoriamente truncado misto sob a perspectiva frequentista, apresentadosa na Seção 8.2, bem como os resultados de estimação e diagnóstico do modelo de regressão não linear aleatoriamente truncado misto sob a perspectiva Bayesiana, apresentadas na Seção 8.3, indicam que ambos os procedimentos de ajuste e diagnóstico do modelo proposto fornecem resultados similares, o que está de acordo com o esperado pois a análise Bayesiana do modelo foi conduzida considerando-se distribuições a *priori* fracamente informativas. Sob ambas as metodologias, a distribuição beta aleatoriamente truncada é indicada como a que melhor se ajusta aos dados. Por outro lado, de acordo com a metodologia frequentista, o modelo selecionado como o melhor ajustado aos dados do perfil 219 tem o parâmetro de média das respostas modelado pela função de Gardner (1958). Já sob a perspectiva Bayesiana, o modelo selecionado como o de melhor ajuste aos dados tem o parâmetro de média das respostas modelado pela função de van Genuchten (1980)-Mualem (1976). Assim, pelos resultados de simulação apresentados nos Capítulos 6 e 7 e seguindo o que foi discutido na Seção 7.6, os resultados do ajuste Bayesiano do modelo proposto parece ser uma melhor escolha para os dados em questão. Além disso, a metodologia Bayesiana também seria capaz de contornar o problema da falta de convergência do ajuste do modelo NTA GA Ht (Tabelas 8.6 e 8.9). Para isto, bastaria que uma outra estratégia de implementação do MCMC fosse adotada.

Capítulo 9

Algumas conclusões e considerações finais

Neste trabalho, foi proposta e desenvolvida uma classe de modelos não lineares truncados mistos para tratar dados que possuem a característica de truncamento. Esta característica foi traduzida no modelo de regressão por meio da suposição de que a variável resposta possui uma distribuição de probabilidade truncada. O modelo de regressão proposto foi construído considerando-se limites de truncamento aleatórios, isto é, considerando-se os próprios limites de truncamento como v.a.'s, o que faz com que seja necessário assumir distribuições de probabilidade para os mesmos. Este modelo pode ser prontamente reduzido para o caso de limites fixos e conhecidos. Além disso, as informações contidas nas covariáveis foram incorporadas ao modelo assumindo-se que o parâmetro de locação (média) da distribuição da v.a. de interesse era associado à uma função não linear contínua de um vetor de parâmetros desconhecidos e das covariáveis e aos efeitos aleatórios não observáveis. Também foi considerada a possibilidade de que o parâmetro de escala (dispersão) da variável resposta fosse relacionada a um conjunto de covariáveis através de uma função de parâmetros desconhecidos e das covariáveis.

O procedimento de estimação de máxima verossimilhança iterativo dos parâmetros foi aplicado ao modelo proposto com sucesso e os resultados baseados em dados simulados indicaram boas propriedades assintóticas das estimativas de máxima verossimilhança dos parâmetros fixos dos modelos não lineares aleatoriamente truncados mistos. Por outro lado, houve indicativos de problemas na estimação do parâmetro fixo relacionado aos efeitos aleatórios não observáveis, e a média dos EMVs obtidos para este parâmetro das amostras simuladas pareceu indicar que o mesmo era subestimado e a sua probabilidade de cobertura foi excessivamente baixa. Não obstante, não houve problemas na estimação das probabilidades de cobertura dos efeitos aleatórios não observáveis. Ademais, algumas ferramentas de análise de diagnóstico foram utilizadas para verificar os pressupostos do modelo e para detectar observações *outlier* e/ou influentes.

Um procedimento de estimação Bayesiana, para calcular as estimativas dos parâmetros

do modelo, foi aplicado com sucesso aos modelos propostos e os resultados da simulação indicaram boas propriedades frequentistas das estimativas Bayesianas dos parâmetros para o modelo proposto. A metodologia Bayesiana foi capaz de melhorar as propriedades frequentistas do parâmetro de escala relacionado aos efeitos aleatórios não observáveis, que ainda que subestimado, passou a apresentar probabilidades de coberturas estimadas mais próximas da esperada. Ferramentas de diagnóstico Bayesiano foram utilizadas para verificar se há má especificação do modelo e para detecção de *outliers* e observações influentes. Apresentamos, ainda, uma proposta de seleção de modelos baseada na abordagem Bayesiana de mistura de modelos.

Destacamos que apenas duas distribuições truncadas, dentre uma vasta gama de possibilidades, foram consideradas no presente trabalho. Por exemplo, as distribuições t-Student locação-escala truncada, a distribuição de Fréchet truncada, a distribuição de Lévy truncada, entre muitas outras versões truncadas de distribuições da família de locação-escala, poderiam ser consideradas na construção dos modelos não lineares aleatoriamente truncado e com truncamento fixo e conhecido. Além disso, para os limites de truncamento aleatórios, outras distribuições de probabilidade também poderiam ser consideradas e, se o objetivo principal da análise fosse a modelagem estatística desses limites, então poderíamos pensar em estruturas de regressão para os seus parâmetros indexadores em função de covariáveis e parâmetros desconhecidos a serem estimados.

Em resumo, destacamos a pretensão de apresentar e desenvolver uma metodologia de análise de regressão não linear para estudar e modelar variáveis de interesse que são caracterizadas pelo fato de serem truncadas, um tipo de situação que dá origem a diferentes conjuntos de dados nas mais diferentes áreas de aplicação do conhecimento científico. O modelo proposto foi utilizado em um breve estudo de aplicação para estimar curvas de retenção de água em solo, um fenômeno de fundamental interesse nas áreas de planejamento e manejo de recursos hídricos e no agronegócio. Os modelos de regressão não linear truncados são simples, facilmente estimáveis e melhoram consideravelmente a qualidade das inferências fornecidas por levarem em conta a característica de truncamento presente nos dados.

Referências

A'Hearn, A. (2004). A restricted maximum likelihood estimator for truncated height samples. *Economics and Human Biology*, **2**, 5–19.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**(422), 669–679.

Albert, J. H. & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, **82**(4), 747–759.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**(2).

Bragato, P. L. (2004). Regression analysis with truncated samples and its application to ground-motion attenuation studies. *Bulletin of the Seismological Society of America*, **94**(4), 1369–1378.

Brooks, R. H. & Corey, A. T. (1964). Hydraulic properties of porous media. *Hydrology Paper*, **3**(3).

Burdine, N. T. (1953). Relative permeability calculations from pore size distribution data. *Journal of Petroleum Technology*, **5**(3), 71–78.

Carlin, B. P. & Louis, T. A. (2009). *Bayesian methods for data analysis*. Chapman and Hall/CRC, Boca Raton, third edition.

Chao, K., Nelson, J. D., Overton, D. D. & Cumbers, J. M. (2008). Soil water retention curves for remolded expansive soils. In *Unsaturated Soils: Advances in Geo-Engineering, Proceedings of the 1st European Conference, E-UNSAT*.

Chen, S. & Zhou, X. (2012). Semiparametric estimation of a truncated regression model. *Journal of Econometrics*, **167**, 297–304.

Cho, H., Ibrahim, J. G., Sinha, D. & Zhu, H. (2009). Bayesian case in influence diagnostics for survival models. *Biometrics*, **65**(1), 116–124.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**(1), 15–18.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, **48**(2), 133–169.

Cook, R. D. & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman & Hall, New York.

Cook, R. D. & Weisberg, S. (1994). *An introduction to regression graphics*. Wiley, New York.

Cornelis, W. M., Khlosi, M., Hartmann, R., van Meirvenne, M. & de Vos, B. (2005). Comparison of unimodal analytical expressions for the soil-water retention curve. *Soil Science Society of America Journal*, **69**, 1902–1911.

Cosslett, S. (2004). Efficient semiparametric estimation of censored and truncated regressions via smoothed self-consistency equation. *Econometrica*, **72**, 1277–1293.

de Uña Álvarez, J., Liang, H.-Y. & Rodríguez-Casal, A. (2010). Nonlinear wavelet estimator of the regression function under left-truncated dependent data. *Journal of Nonparametric Statistics*, **22**(3), 319–344.

del Castillo, J. (1994). The singly truncated normal distribution: a non-steep exponential family. *Ann. Inst. Statist. Math.*, **46**(1), 57–66.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Dourado-Neto, D., Nielsen, D. R., Hopmans, J. W., Reichardt, K. & Bacchi, O. O. S. (2000). Software to model soil water retention curves (swrc, version 2.00). *Scientia Agricola*, **57**(1), 347–354.

Flecher, C., Allard, D. & Naveau, P. (2010). Truncated skew-normal distributions: moments, estimation by weighted moments and application to climatic data. *International Journal of Statistics*, **LXVIII**(3), 331–345.

Fornberg, B. & Sloan, D, M. (1994). A review of pseudospectral methods for solving partial differential equations. *Acta Numerica*, **3**, 203–267.

Fredlund, D. G. & Xing, A. (1994). Equations for the soil-water characteristic curve. *Canadian Geotechnical Journal*, **31**(3), 521–532.

Gardner, W. R. (1958). Some steady state solutions of unsaturated moisture flow equations with application to evaporation from water table. *Soil Science*, **85**, 228–232.

Gelfand, A. E., D., D. & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). *Bayesian Statistics*, **4**, 147–167.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**(3), 515–533.

Gelman, A., Meng, X. L. & Stern, H. . (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733–807.

Gelman, A., Goegebeur, Y., Tuerlinckx, F. & Van Mechelen, I. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society, Series C*, **49**(2), 247–268.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. B. (2003). *Bayesian data analysis*. Chapman and Hall/CRC Texts in Statistical Science, London, second edition.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, & J. F. M. Smith, editors, *Bayesian Statistics 4*. Oxford University Press, Oxford.

Gilbert, P. & Varadhan, R. (2012). *numDeriv: Accurate Numerical Derivatives*. R package version 2012.9-1.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall, New Jersey, fifth edition.

Hausman, J. A. & Wise, D. A. (1977). Social experimentation, truncated distributions, and efficient estimation. *Econometrica*, **45**(4), 919–938.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5**(4), 475–492.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *The Annals of Mathematical Statistics*, **21**, 309–310.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**(2), 423–447.

Henderson, C. R., Kempthorne, O. & Searle, S. R.and von Krosig, C. M. (1959). The estimation ofeenvironmental and genetic trends from records subject to culling. *Biometrics*, **15**(2), 192–218.

Hyde, J. (1977). Testing survival under right censoring and left truncation. *Biometrika*, **64**, 225–230.

Jamalizadeh, A., Pourmousa, R. & Balakrishnan, N. (2009). Truncated and limited skew-normal and skew-t distributions: properties and an illustration. *Communications in Statistics - Theory and Methods*, **38**, 2653–2668.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994). *Continuous univariate distributions, volume 1*. Wiley, New York, second edition.

Kalbfleisch, J. D. & Lawless, J. F. (1989). Inferences based on retrospective ascertainment: an analysis of the data on transfusion-related aids. *Journal of the American Statistical Association*, **84**, 360–372.

Lachos, V. H., Castro, L. M. & Dey, D. K. (2013). Bayesian inference in nonlinear mixed-effects models using normal independent distributions. *Computational Statistics & Data Analysis*, **64**, 237–252.

Lagakos, S. W., Barraj, L. M. & Gruttolla, V. D. (1988). Nonparametric analysis of truncated survival data with application to aids. *Biometrika*, **75**, 515–524.

Laird, N. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**(4), 963–974.

Lee, L. F. (1979). On the first and second moments of the truncated multi-normal distribution and a simple estimator. *Economics Letters*, **3**, 165–169.

Lee, L. F. (1992). Semiparametric nonlinear least-square estimation of truncated regression models. *Econometric Theory*, **8**, 52–94.

Lee, M. (1993). Quadratic mode regression. *Journal of Econometrics*, **57**, 1–19.

Leong, E. C. & Rahardjo, H. (1997). Review of soil water characteristic curve equations. *Journal of Geotechnical and Geoenvironmental Engineering*, **123**(12), 1106–1117.

Leroy, F., Dauxois, J. Y. & Tubert-Bitter, P. (2013). On the parametric maximum likelihood estimator for independent but non-identically distributed observations with application to truncated data. Technical report.

Lim, H., Song, J. & Jung, B. C. (2013). Score tests for zero-inflation and overdispersion in two-level count data. *Computational Statistics and Data Analysis*, **61**, 15–32.

Lindsey, J. K. & Altham, P. M. E. (1998). Analysis of the human sex ratio by using overdispersion models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **47**(1), 149–157.

Lindstrom, M. J. & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**(3), 673–687.

Lynch, S. M. & Western, B. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods and Reserarch*, **32**(3), 301–335.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge, New York.

Matos, L. A., Lachos, V. H., Balakrishnan, N. & Labra, F. V. (2013a). Influence diagnostics in linear and nonlinear mixed-effects models with censored data. *Computational Statistics & Data Analysis*, **57**, 450–464.

Matos, L. A., Prates, M. O., H., C. M. & Lachos, V. H. (2013b). Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica*, **23**, 1323–1345.

Meng, X. L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: a general framework. *Biometrika*, **80**(2), 267–278.

Moreira, C. & Uña Álvarez, J. (2012). Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics*, **6**, 501–521.

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, **12**, 593–622.

Nadarajah, S. (2008). A truncated inverted beta distribution with application to air pollution data. *Stochastic Environmental Research and Risk Assessment*, **22**(2), 285–289.

Nadarajah, S. (2009). Some truncated distributions. *Acta Applicandae Mathematicae*, **106**(1), 105–123.

Nadarajah, S. & Ali, M. M. (2004). A skewed truncated t distribution. *Mathematical and Computer Modelling*, **40**(9-10), 935–939.

Nadarajah, S. & Kotz, S. (2008). Moments of truncated t and f distributions. *Portuguese Economic Journal*, **7**(1), 63–73.

Newey, W. (2004). Efficient semiparametric estimation via moment restrictions. *Econometrica*, **72**, 1877–1897.

Ould-Saïd, E. Lemdani, M. (2006). Asymptotic properties of a nonparametric regression function estimator with randomly truncated data. *Annals of the Institute of Statistical Mathematics*, **85**(2), 357–378.

Peng, F. & Dey, D. (1995). Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics*, **23**(2), 199–213.

Pereira, G. H. A., Botter, D. A. & Sandoval, M. C. (2012). The truncated inflated beta distribution. *Communications in Statistics - Theory and Methods*, **41**(5), 907–919.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-Plus*. Springer, New York, first edition.

Powell, J. (1986). Symmetrically trimmed least squares estimation for tobit models. *Econometrica*, **54**, 1435–1460.

R Development Core Team (2009). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, **6**(1), 15–32.

Rodrigues, L. N. & Maia, A. H. N. (2011). Funções de pedotransferência para estimar a condutividade hidráulica saturada e as umidades de saturação e residual do solo em uma bacia hidrográfica do cerrado. In *XIX Simpósio brasileiro de recursos hídricos*.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.

Shen, P. S. (2013). Regression analysis of interval censored and doubly truncated data with linear transformation models. *Computational Statistics*, **28**(2), 581–596.

Sillers, W. S., Fredlund, D. G. & Zakerzadeh, N. (2001). Mathematical attributes of soil-water characteristic curves models. *Geotechnical and Geological Engineering*, **19**, 243–283.

Silva, E. M., Lima, E. F. W., Azevedo, J. A. & Rodrigues, L. N. (2006). Valores de tensão na determinação da curva de retenção de água de solos do cerrado. *Pesquisa Agropecuária Brasileira*, **41**(2), 323–330.

Stephens, M. A. (1974). Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**(347), 730–737.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, **26**(1), 24–36.

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, **44**, 892–898.

van Genuchten, M. T. & Nielsen, D. R. (1985). On describing and predicting the hydraulic properties of unsaturated soils. *Annales Geophysicae*, **3**(5), 615–628.

Wang, M. C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association*, **84**(407), 742–748.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics*, **13**(1), 163–177.

Yan, G. & Sedransk, J. (2010). A note on bayesian residuals as a hierarchical model diagnostic technique. *Statistical Papers*, **51**(1), 1–10.

Yates, S. R., van Genuchten, M. T. & Warrick, A. W.; Leij, F. J. (1992). Analysis of measured, predicted, and estimated hydraulic conductivity using the retc computer program. *Soil Science Society of America Journal*, **56**, 347–354.

Zaninetti, L. (2013). The initial mass function modeled by a left truncated beta distribution. *Astrophysical Journal*, **765**(2).

Zhou, M., Yang, D., Wang, Y. & Nadarajah, S. (2010). Programs in r for computing, truncated normal distributions. *Computer Applications in Engineering Education*, **3**(18), 589–592.

Apêndice A

Geração de variáveis aleatórias truncadas

Neste capítulo, apresentamos as funções implementadas em R para simular valores de v.a.'s $Y\,|(a < Y < b) \sim NT\,(\mu, \sigma, a, b)$ descritas na Seção 2.1.1 e para simular valores de v.as's $Y\,|(a < Y < b) \sim BT\,(\mu, \sigma, a, b)$ descritas na Seção 2.1.2. Apresentamos, também, as funções usadas para calcular as esperanças e variâncias destas mesmas variáveis. As funções foram implementadas em R (R Development Core Team, 2009) e podem ser usadas para gerar variáveis que satisfaçam as condições dos casos particulares mencionados na Seção 3.2, considerando-se as restrições pertinentes para cada caso.

A.1 Função em R para a distribuição normal truncada

```
#Função para gerar v.a. normal truncada
rtn = function(n, location, scale, lower, upper){
if (length(location) == 1){location = rep(location, n)}
if (length(location) == n){location = location}
if (length(scale) == 1){scale = rep(scale, n)}; if (length(scale) == n){scale = scale}
if (length(lower) == 1){lower = rep(lower, n)}; if (length(lower) == n){lower = lower}
if (length(upper) == 1){upper = rep(upper, n)}; if (length(upper) == n){upper = upper}
y = NULL
for (i in 1:n){
u = runif(1)
quant = pnorm(lower[i], location[i], scale[i])
+ u*(pnorm(upper[i], location[i], scale[i]) - pnorm(lower[i], location[i], scale[i]))
y = c(y, qnorm(quant, location[i], scale[i]))}
return(y)}

#Função para calcular a esperanção da distribição normal truncada
exp.nt = function(mu, sd, a, b){
fun.exp = function(y){y*dnorm(y, mu, sd)/(pnorm(b, mu, sd) - pnorm(a, mu, sd))}
val = integrate(fun.exp, a, b)$value
```

```
return(val)}

#Função para calcular a variância da distribição normal truncada
var.nt = function(mu, sd, a, b){
exp.y = exp.nt(mu, sd, a, b)
fun.var = function(y){((y - exp.y)^2)*dnorm(y, mu, sd)
/(pnorm(b, mu, sd) - pnorm(a, mu, sd))}
val = integrate(fun.var, a, b)$value
return(val)}
```

A.2 Função em R para a distribuição beta truncada

```
#Função para gerar v.a. beta truncada
rtb = function(n, location, scale, lower, upper){
gamma = location*exp(scale); rho = (1 - location)*exp(scale)
if (length(gamma) == 1){gamma = rep(gamma, n)}; if (length(gamma) == n){gamma = gamma}
if (length(rho) == 1){rho = rep(rho, n)}; if (length(rho) == n){rho = rho}
if (length(lower) == 1){lower = rep(lower, n)}; if (length(lower) == n){lower = lower}
if (length(upper) == 1){upper = rep(upper, n)}; if (length(upper) == n){upper = upper}
y = NULL
for (i in 1:n){
u = runif(1)
quant = pbeta(lower[i], gamma[i], rho[i]) + u*(pbeta(upper[i], gamma[i], rho[i])
- pbeta(lower[i], gamma[i], rho[i]))
y = c(y, qbeta(quant, gamma[i], rho[i]))}
return(y)}

#Função para calcular a esperanção da distribição beta truncada
exp.bt = function(mu, sd, a, b){
gamma = mu*exp(sd)
rho = (1 - mu)*exp(sd)
fun.exp = function(y){y*dbeta(y, gamma, rho)
/(pbeta(b, gamma, rho) - pbeta(a, gamma, rho))}
val = integrate(fun.exp, a, b)$value
return(val)}

#Função para calcular a variância da distribição beta truncada
var.bt = function(mu, sd, a, b){
exp.y = exp.bt(mu, sd, a, b)
gamma = mu*exp(sd)
rho = (1 - mu)*exp(sd)
fun.var = function(y){((y - exp.y)^2)*dbeta(y, gamma, rho)/
(pbeta(b, gamma, rho) - pbeta(a, gamma, rho))}
val = integrate(fun.var, a, b)$value
return(val)}
```

Apêndice B

Metodologia clássica: exemplo de programas em R

A seguir, apresentamos um exemplo dos programas computacionais utilizados na abordagem frequentista dos modelos não lineares aleatoriamente truncados mistos desenvolvidos no Capítulo 3. As funções foram implementadas em R (R Development Core Team, 2009) para obter os EMVs e os ICs de Wald e de RV apresentados na Seção 4.1, e para computar os resíduos os resíduos padronizados, as métricas de detecção de observações influentes e os critérios de seleção de modelo da Seção 4.2 do Capítulo 4.

Ressaltamos que as funções para os procedimentos de estimação, diagnóstico e seleção de modelos sob os casos particulares apresentados na Seção 3.2 são análogos ao aqui exemplificados, fazendo-se as restrições pertinentes para cada caso.

A função implementada em R para obter os EMV $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\omega}}_A$ e $\hat{\boldsymbol{\omega}}_B$ de $\boldsymbol{\theta}$, $\boldsymbol{\omega}_A$ do modelo (3.1) é exemplificada considerando que a variável resposta truncada segue distribuição normal truncada, com parâmetro de locação relacionado à função de Gardner e com parâmetro de escala constante, isto é, consideramos o modelo de regressão não linear normal aleatoriamente truncado misto definido em (3.12) com $\eta\left(\boldsymbol{x}_{q_{i,jk}}, \boldsymbol{\beta}\right)$ dada pela expressão em (1.1) e com $\sigma_{i,jk} = g\left(\boldsymbol{x}_{r_{i,jk}}, \boldsymbol{\alpha}\right) = \sigma$.

```
#Pacotes
library(trust); library(numDeriv)

#Funções de -log-verossimilhanças
lveroY = function(param){
b1 = param[1]; b2 = param[2]; sg = param[3]
sz = param[4]; z = param[5:length(param)]
eta = Tr + (Ts - Tr)/(1 + b1*(x^b2)); sdy = sg; phi = eta + Z.mat%*%z
dens = dnorm(y, phi, sdy); acub = pnorm(Ts, phi, sdy); acua = pnorm(Tr, phi, sdy)
ldensy = log(dens) - log(acub - acua); ldensz = log(dnorm(z, 0, sz))
ll = sum(ldensy) + sum(ldensz)
return(-ll)}
```

```
lvero.y = function(param){
b1 = param[1]; b2 = param[2]; sg = param[3]
eta = Tr + (Ts - Tr)/(1 + b1*(x^b2)); sdy = sg
sz.est = emv.z[1]; z.est = emv.z[-1]
phi = eta + Z.mat%*%z.est
dens = dnorm(y, phi, sdy); acub = pnorm(Ts, phi, sdy); acua = pnorm(Tr, phi, sdy)
ldensy = log(dens) - log(acub - acua); ldensz = log(dnorm(z.est, 0, sz.est))
ll = sum(ldensy) + sum(ldensz)
return(-ll)}

lvero.z = function(param){
sz = param[1]; z = param[2:length(param)]
b1 = emv.y[1]; b2 = emv.y[2]; sg = emv.y[3]
eta = Tr + (Ts - Tr)/(1 + b1*(x^b2)); sdy = sg
phi = eta + Z.mat%*%z
dens = dnorm(y, phi, sdy); acub = pnorm(Ts, phi, sdy); acua = pnorm(Tr, phi, sdy)
ldensy = log(dens) - log(acub - acua); ldensz = log(dnorm(z, 0, sz))
ll = sum(ldensy) + sum(ldensz)
return(-ll)}

lveroA = function(param){
ma = param[1]; sa = param[2]
ll = log(dnorm(Tr, ma, sa)) - log(pnorm(Ts, ma, sa) - pnorm(0, ma, sa))
return(-sum(ll))}

lveroB = function(param){
mb = param[1]; sb = param[2]; ll = log(dnorm(Ts, mb, sb))
return(-sum(ll))}

#Funções usadas para encontrar os EMVs (via trust-region)
lv.trust.y = function(param){
b1 = param[1]; b2 = param[2]; sg = param[3]
if (b1 <= 0 || b2 <= 0 || sg <= 0){return(list(value = Inf))}
if (b1 > 0 && b2 > 0 && sg > 0){
value = lvero.y(param)
gradient = grad(lvero.y, param); hessian = hessian(lvero.y, param)
return(list(value = value, gradient = gradient, hessian = hessian))}}

lv.trust.z = function(param){
sz = param[1]; z = param[2:length(param)]
if (sz <= 0){return(list(value = Inf))}
if (sz > 0){
value = lvero.z(param)
gradient = grad(lvero.z, param); hessian = hessian(lvero.z, param)
return(list(value = value, gradient = gradient, hessian = hessian))}}
```

```
lv.trustA = function(param){
ma = param[1]; sa = param[2]
if (sa <= 0){return(list(value = Inf))}
if (sa > 0){
value = lveroA(param)
gradient = grad(lveroA, param); hessian = hessian(lveroA, param)
return(list(value = value, gradient = gradient, hessian = hessian))}}

lv.trustB = function(param){
mb = param[1]; sb = param[2]
if (sb <= 0){return(list(value = Inf))}
if (sb > 0){
value = lveroB(param)
gradient = grad(lveroB, param); hessian = hessian(lveroB, param)
return(list(value = value, gradient = gradient, hessian = hessian))}}

#Funções de log-verossimilhanças
logveroY = function(param){return(-lveroY(param))}
logveroA = function(param){return(-lveroA(param))}
logveroB = function(param){return(-lveroB(param))}

#Função de -log-verossimilhança para calcular os escores de D(-i,jk)
lv.ddiY = function(param){
b1 = param[1]; b2 = param[2]
sg = param[3]; sz = param[4]; z = param[5:length(param)]
eta = Tri + (Tsi - Tri)/(1 + b1*(xi^b2)); sdy = sg
phi = eta + Z.mati%*%z
dens = dnorm(yi, phi, sdy); acub = pnorm(Tsi, phi, sdy); acua = pnorm(Tri, phi, sdy)
ldensy = log(dens) - log(acub - acua); ldensz = log(dnorm(z, 0, sz))
ll = sum(ldensy) + sum(ldensz)
return(-ll)}

#Funções de -log-verossimilhanças perturbadas
lv.pert.respY = function(param){
b1 = param[1]; b2 = param[2]; sg = param[3]
sz = param[4]; z = param[5:(length(c(nome.parY, nome.parZ))+M)]
eta = Tr + (Ts - Tr)/(1 + b1*(x^b2)); sdy = sg
w = param[(length(c(nome.parY, nome.parZ))+M+1):length(param)]
phi = eta + Z.mat%*%z
dens = dnorm((y + w), phi, sdy); acub = pnorm(Ts, phi, sdy); acua = pnorm(Tr, phi, sdy)
ldensy = log(dens) - log(acub - acua); ldensz = log(dnorm(z, 0, sz))
ll = sum(ldensy) + sum(ldensz)
return(ll)}

lv.pert.casoY = function(param){
b1 = param[1]; b2 = param[2]; sg = param[3]
sz = param[4]; z = param[5:(length(c(nome.parY, nome.parZ))+M)]
```

```
eta = Tr + (Ts - Tr)/(1 + b1*(x^b2)); sdy = sg
w = param[(length(c(nome.parY, nome.parZ))+M+1):length(param)]
phi = eta + Z.mat%*%z
dens = dnorm(y, phi, sdy); acub = pnorm(Ts, phi, sdy); acua = pnorm(Tr, phi, sdy)
ldensy = w*(log(dens) - log(acub - acua)); ldensz = w*log(dnorm(Z.mat%*%z, 0, sz))
ll = sum(ldensy) + sum(ldensz)
return(ll)}

lv.pert.covY = function(param){
w = param[(length(c(nome.parY, nome.parZ))+M+1):length(param)]
b1 = param[1]; b2 = param[2]; sg = param[3]
sz = param[4]; z = param[5:(length(c(nome.parY, nome.parZ))+M)]
eta = Tr + (Ts - Tr)/(1 + b1*((x + w)^b2)); sdy = sg
phi = eta + Z.mat%*%z
dens = dnorm(y, phi, sdy); acub = pnorm(Ts, phi, sdy); acua = pnorm(Tr, phi, sdy)
ldensy = log(dens) - log(acub - acua); ldensz = log(dnorm(z, 0, sz))
ll = sum(ldensy) + sum(ldensz)
return(ll)}

#Função calcular as direções das curvaturas (influência local)
dmaxY = function(param, vc, pert){
if (pert == 1){w0 = rep(0, N); hess = hessian(lv.pert.respY, c(param, w0))}
if (pert == 2){w0 = rep(1, N); hess = hessian(lv.pert.casoY, c(param, w0))}
if (pert == 3){w0 = rep(0, N); hess = hessian(lv.pert.covY, c(param, w0))}
delta = hess[-(1:length(param)),][,-((length(param)+1):(length(param)+N))]
mat = delta%*%vc%*%t(delta); aut = eigen(mat); inl = aut$vectors[,1]
return(inl)}

#Estimação (EMV)
emv.z = c(inicio.sz, inicio.z)
res.mv.y = trust(lv.trust.y, inicio.y, rinit = raio, rmax = 10000, iterlim = 100000,
fterm = 1e-5, mterm = 1e-5, minimize = TRUE, blather = FALSE)
emv.y = res.mv.y$argument
res.mv.z = trust(lv.trust.z, c(inicio.sz, inicio.z), rinit = raio, rmax = 10000,
iterlim = 100000, fterm = 1e-5, mterm = 1e-5, minimize = TRUE, blather = FALSE),
emv.z = res.mv.z$argument

res.mvA = trust(lv.trustA, inicio.a, rinit = 5, rmax = 10000, iterlim = 100000,
fterm = 1e-5, mterm = 1e-5, minimize = TRUE, blather = FALSE)
res.mvB = trust(lv.trustB, inicio.b, rinit = 5, rmax = 10000, iterlim = 100000,
fterm = 1e-5, mterm = 1e-5, minimize = TRUE, blather = FALSE)

estY = try(c(emv.y, emv.z), silent = TRUE)
estA = try(res.mvA$argument, silent = TRUE)
estB = try(res.mvB$argument, silent = TRUE)

dif = 1
```

```
while (dif > tol){

if (fun.esc == 1){raio = 1}
if (fun.esc == 2){raio = 0.1}
if (fun.esc == 2 && fun.mu == 3){raio = 1}

res.mv.y = try(trust(lv.trust.y, inicio.y, rinit = raio, rmax = 10000, iterlim = 100000,
fterm = 1e-5, mterm = 1e-5, minimize = TRUE, blather = FALSE), silent = TRUE)
emv.y = try(res.mv.y$argument, silent = TRUE)
res.mv.z = try(trust(lv.trust.z, c(inicio.sz, inicio.z), rinit = raio, rmax = 10000,
iterlim = 100000, fterm = 1e-5, mterm = 1e-5, minimize = TRUE, blather = FALSE),
silent = TRUE)
emv.z = try(res.mv.z$argument, silent = TRUE)

est.novo = try(c(emv.y, emv.z), silent = TRUE)
mvari = try(min(diag(solve(hessian(lveroY, est.novo)))), silent = TRUE)

if (class(mvari) == "try-error" || is.na(mvari) == TRUE){mvari = -1}

if (class(res.mv.y) == "try-error" || res.mv.y$converged != TRUE ||
class(res.mv.z) == "try-error" || res.mv.z$converged != TRUE || mvari <= 0){
converged = FALSE
dif = 1E-4}

if (class(res.mv.y) != "try-error" && res.mv.y$converged == TRUE &&
class(res.mv.z) != "try-error" && res.mv.z$converged == TRUE && mvari > 0){
converged = TRUE
dif = try(max(abs(estY - est.novo)), silent = TRUE)
estY = est.novo}}

if (converged == TRUE){EMV.NTA = list(emvY = estY, emvA = estA, emvB = estB,
varmvY = diag(solve(hessian(lveroY, estY))), varmvA = diag(solve(hessian(lveroA, estA))),
varmvB = diag(solve(hessian(lveroB, estB))), conv = converged)}
if (converged == FALSE){EMV.NTA = list(conv = converged)}

emvY = EMV.NTA$emvY; emvA = EMV.NTA$emvA; emvB = EMV.NTA$emvB
varmvY = EMV.NTA$varmvY; varmvA = EMV.NTA$varmvA; varmvB = EMV.NTA$varmvB

#IC-RV
nperf = 100000
lv.mvY = lveroY(emvY); sdY = varmvY + 0.5

testeY = matrix(0, nrow = nperf, ncol = length(emvY))
testeY[,1] = seq(-(emvY[1] + sdY[1]), (emvY[1] + sdY[1]), length.out = nperf)
testeY[,2] = seq(-(emvY[2] + sdY[2]), (emvY[2] + sdY[2]), length.out = nperf)
testeY[,3] = seq(-(emvY[3] + sdY[3]), (emvY[3] + sdY[3]), length.out = nperf)
testeY[,4] = seq(-(emvY[4] + sdY[4]), (emvY[4] + sdY[4]), length.out = nperf)
```

```
testeY[,5] = seq(-(emvY[5] + sdY[5]), (emvY[5] + sdY[5]), length.out = nperf)
testeY[,6] = seq(-(emvY[6] + sdY[6]), (emvY[6] + sdY[6]), length.out = nperf)
testeY[,7] = seq(-(emvY[7] + sdY[7]), (emvY[7] + sdY[7]), length.out = nperf)

icrvY = list(length(c(inicio.y, inicio.sz, inicio.z)))
for (k in 1:length(c(inicio.y, inicio.sz, inicio.z))){
icp = NULL; lvp = NULL
for (j in 1:nperf){est = emvY; est[k] = testeY[j,k]; lvp = c(lvp, lveroY(est))}
dif = lv.mvY - (lvp - (1/2)*qchisq(0.95, 1))
dna = which(is.na(dif) == TRUE); dni = which(dif == -Inf); dpi = which(dif == Inf)
if (length(c(dna, dni, dpi)) == 0){dd = seq(1, nperf)}
if (length(c(dna, dni, dpi)) != 0){dd = seq(1, nperf)[-c(dna, dni, dpi)]}
difs = dif[dd]; tt = testeY[dd,k]
icp = rbind (icp, tt[(which(c(sign(difs),0)*c(0,sign(difs)) == -1))])
icrvY[[k]] = icp}
names(icrvY) = colnames(emvY)

lrciY = NULL
for (k in 1:length(c(inicio.y, inicio.sz, inicio.z))){
lrciY = rbind(lrciY, icrvY[[k]])}

#IC de Wald
icwY = NULL
for (k in 1:length(emvY)){
icwY = rbind(icwY, c((emvY[k] - qnorm(1 - (0.05/2))*sqrt(varmvY[k])),
(emvY[k] + qnorm(1 - (0.05/2))*sqrt(varmvY[k]))))}

#Predição
pred.eta = Tr + (Ts - Tr)/(1 + emvY[1]*(x^emvY[2])); pred.sdy = emvY[3]
if (length(pred.sdy) == 1){pred.sdy = rep(pred.sdy, N)}
pred.z = emvY[-(1:length(c(nome.parY, nome.parZ)))]
pred.phi = pred.eta + Z.mat%*%pred.z
y.mat = array(y, dim = c(n, nue*M)); p.mat = array(pred.phi, dim = c(n, nue*M))
glin = 3; gcol = c(rep("red", nue), rep("blue", nue), rep("green", nue))

#Valores preditos
pred.y = NULL
for (i in 1:N){
pred.y = c(pred.y, exp.nt(pred.phi[i], pred.sdy[i], Tr[i], Ts[i]))}

#Variância estimada
vari.y = NULL
for (i in 1:N){
vari.y = c(vari.y, var.nt(pred.phi[i], pred.sdy[i], Tr[i], Ts[i]))}

#Resíduos padronizados
res.y = y - pred.y; resp.y = res.y/sqrt(vari.y)
```

```
#Matriz de informação de Fisher
infoY = hessian(lveroY, emvY)

#Matriz de covarâncias
vcY = solve(hessian(lveroY, emvY))
vcP = matrix(0, nrow = length(c(nome.parY, nome.parZ, nome.prof)),
ncol = length(c(nome.parY, nome.parZ, nome.prof)))
vcP[(length(c(nome.parY, nome.parZ))+1):nrow(vcP),
(length(c(nome.parY, nome.parZ))+1):ncol(vcP)] =
vcY[(length(c(nome.parY, nome.parZ))+1):nrow(vcP),
(length(c(nome.parY, nome.parZ))+1):ncol(vcP)]
vcYP = vcY - vcP

#Influência global: dist. gen. de Cook e dist. da verossimilhança
gc.y = NULL; ld.y = NULL
for (i in 1:N){
xi = x[-i]; yi = y[-i]; Tri = Tr[-i]; Tsi = Ts[-i]; Z.mati = Z.mat[-i,]
grady = grad(lv.ddiY, emvY)
emviY = matrix((emvY + c(vcY%*%as.matrix(grady))), nrow = 1)
gc.y = c(gc.y, (matrix(emviY, nrow = 1) - emvY)
%*%infoY%*%t((matrix(emviY, nrow = 1) - emvY)))
ld.y = c(ld.y, 2*(logveroY(emvY) - logveroY(emviY)))}

#Influência local - perturbacao da resposta
pert = 1; il.resp.y = dmaxY(emvY, vcYP, pert)
#Influeência local - perturbacao de caso
pert = 2; il.caso.y = dmaxY(emvY, vcYP, pert)
```

Apêndice C

Metodologia clássica: condições de regularidade

A seguir, descrevemos as condições de regularidade (Leroy *et al.*, 2013) que devem ser satisfeitas para que os estimadores de máxima verossimilhança sejam consistente e assintoticamente normais quando as observações são provenientes de distribuições independentes não identicamente distribuídas e com presença de truncamento. Nas Suposições 7 e 8 o símbolo $\xrightarrow{\wp}$ denota convergência em probabilidade.

Seja $\boldsymbol{x} = (x_1, \ldots, x_n)$ um conjunto de n observações independentes de $\boldsymbol{X}\,|\boldsymbol{a} < \boldsymbol{X} < \boldsymbol{b} = (X_1\,|a_1 < X_1 < b_1\,, \ldots, X_n\,|a_n < X_n < b_n)$, onde cada $X_i\,|a_i < X_i < b_i$ é uma v.a. truncada com f.d.p. definida como

$$f\left(x_i\,|a_i < x_i < b_i; \boldsymbol{\theta}\right) = \begin{cases} \dfrac{f\left(x_i\,|\boldsymbol{\theta}\right)}{F\left(b_i\,|\boldsymbol{\theta}\right) - F\left(a_i\,|\boldsymbol{\theta}\right)}, a_i < x_i < b_i \\ 0, \text{caso contrário,} \end{cases} \tag{C.1}$$

sendo $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ o vetor de parâmetros indexadores desconhecidos comum a todos os $X_i\,|a_i < X_i < b_i$, $i = 1, \ldots, n$.

Seja $\boldsymbol{\theta} \in \Theta$, Θ um subconjunto aberto de $\mathbb{R}^p$, e $\mathbb{S}_i \in \mathbb{R}$ o suporte de $f\left(x_i\,|a_i < x_i < b_i; \boldsymbol{\theta}\right)$ que é independente de $\boldsymbol{\theta}$, $i = 1, \ldots, n$.

Denote por $\boldsymbol{\theta}^0 = \left(\theta_1^0, \ldots, \theta_p^0\right)$ o verdadeiro valor de $\boldsymbol{\theta}$.

A função de verossimilhança baseada nos dados observados é dada por

$$\mathcal{L}\left(\boldsymbol{\theta}\,|\boldsymbol{x}\right) = \prod_{i=1}^{n} f\left(x_i\,|a_i < x_i < b_i; \boldsymbol{\theta}\right) = \prod_{i=1}^{n} \frac{f\left(x_i\,|\boldsymbol{\theta}\right)}{F\left(b_i\,|\boldsymbol{\theta}\right) - F\left(a_i\,|\boldsymbol{\theta}\right)}, \tag{C.2}$$

e a função de log-verossimilhança é escrita como

$$\ell\left(\boldsymbol{\theta}\,|\boldsymbol{x}\right) = \sum_{i=1}^{n} \left\{\log f\left(x_i\,|\boldsymbol{\theta}\right) - \log\left[F\left(b_i\,|\boldsymbol{\theta}\right) - F\left(a_i\,|\boldsymbol{\theta}\right)\right]\right\}. \tag{C.3}$$

O EMV de $\boldsymbol{\theta}$, denotado por $\hat{\boldsymbol{\theta}}_n$ é tal que

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\left\{\mathcal{L}\left(\boldsymbol{\theta}\,|\boldsymbol{x}\right)\right\} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\left\{\ell\left(\boldsymbol{\theta}\,|\boldsymbol{x}\right)\right\}. \tag{C.4}$$

Suposição 1. *O EMV $\hat{\boldsymbol{\theta}}_n$ é solução das equações normais*

$$\frac{\partial}{\partial\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}\,|\boldsymbol{x}\right) = \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}\,|x_i\right) = \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right) = 0. \tag{C.5}$$

Suposição 2. *As equações normais (C.5) têm solução única.*

Portanto, o EMV $\hat{\boldsymbol{\theta}}_n$ é a solução única da função escore da equação de log-verossimilhança (C.3),

$$S\left(\hat{\boldsymbol{\theta}}_n\right) = \frac{\partial}{\partial\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}\,|\boldsymbol{x}\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ell\left(\hat{\boldsymbol{\theta}}_n\,|x_i\right) = \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,\Big|a_i < x_i < b_i;\hat{\boldsymbol{\theta}}_n\right) = 0 \tag{C.6}$$

e o mesmo corresponde ao máximo local se a matriz Hessiana em $\hat{\boldsymbol{\theta}}_n$

$$H\left(\hat{\boldsymbol{\theta}}_n\right) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\ell\left(\boldsymbol{\theta}\,|\boldsymbol{x}\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = \sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\ell\left(\hat{\boldsymbol{\theta}}_n\,|x_i\right) = \sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,\Big|a_i < x_i < b_i;\hat{\boldsymbol{\theta}}_n\right), \tag{C.7}$$

é negativa definida.

Suposição 3. *Para todo $\boldsymbol{\theta}\in\Theta$, as derivadas parciais $\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)$, $\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)$ e $\frac{\partial^3}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)$ existem para todo x_i, $i = 1,\ldots,n$.*

Suposição 4. *Para todo $\boldsymbol{\theta}\in\Theta$, a derivada parcial $\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)$ é uma função integrável em $\mathbb{S}_i$ e diferenciação e integração são intercambiáveis, isto é,*

$$\int\limits_{S_i}\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)dx_i = \frac{\partial}{\partial\boldsymbol{\theta}}\int\limits_{S_i}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)dx_i,$$

para $i = 1,\ldots,n$.

Suposição 5. *Para todo $\boldsymbol{\theta}\in\Theta$, a derivada parcial $\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)$ é uma função integrável em $\mathbb{S}_i$ e diferenciação e integração são intercambiáveis,*

$$\int\limits_{S_i}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)dx_i = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\int\limits_{S_i}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)dx_i,$$

para $i = 1,\ldots,n$.

Suposição 6. *Para todo* $\boldsymbol{\theta} \in \Theta$, $i = 1, \ldots, n$,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right) \overset{\wp}{\to} \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}E\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)\right\} = 0$$

Suposição 7. *Para todo* $\boldsymbol{\theta} \in \Theta$, $i = 1, \ldots, n$, $E\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,|a < x_i < b;\boldsymbol{\theta}\right)\right\}$ *e* $\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}E\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,|a < x_i < b;\boldsymbol{\theta}\right)\right\}$ *existem, e*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right) \overset{\wp}{\to} \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}E\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)\right\}.$$

Suposição 8. *Para todo* $\boldsymbol{\theta} \in \Theta$, $i = 1, \ldots, n$, $E\left\{\frac{\partial^3}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}\partial\boldsymbol{\theta}}\log f\left(x_i\,|a < x_i < b;\boldsymbol{\theta}\right)\right\}$ *e* $\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}E\left\{\frac{\partial^3}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}\partial\boldsymbol{\theta}}\log f\left(x_i\,|a < x_i < b;\boldsymbol{\theta}\right)\right\}$ *existem, e*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right) \overset{\wp}{\to} \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}E\left\{\frac{\partial^3}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)\right\}. \quad \text{(C.8)}$$

Suposição 9. *Existe* M *tal que para todo* $\boldsymbol{\theta} \in \Theta$, $i = 1, \ldots, n$,

$$\left|\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}E\left\{\frac{\partial^3}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)\right\}\right| < M.$$

Suposição 10. *A matriz*

$$I_i\left(\boldsymbol{\theta}^0\right) = \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}E\left\{-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}\right\}$$

é positiva definida.

Suposição 11. *Seja* $C = \left\{\left[\sum_{j=1}^{p}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,|a < x_i < b;\boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}\right)^2\right]^{1/2} > \varepsilon\sqrt{n}\right\}$. *Para todo* $\varepsilon > 0$,

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}E\left\{\sum_{j=1}^{p}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i\,|a_i < x_i < b_i;\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}\right)^2 I\left(C\right)\right\} = 0,$$

onde $I\left(C\right)$ *é a função indicadora do conjunto* C.

Apêndice D

Metodologia Bayesiana: exemplo de programas em R

A seguir, apresentamos exemplos dos programas computacionais utilizados na abordagem Bayesiana dos modelos não lineares aleatoriamente truncados mistos desenvolvidos no Capítulo 3.

D.1 Algoritmo Metropolis-Hastings com Gibbs

A seguinte função gera e retorna uma cadeia MCMC de um vetor de parâmetros usando um algoritmo do tipo Gibbs com passos de Metropolis e passeio aleatório como descrito no Capítulo 5.

```
MCMC.MHG.SEP.BTA.CRA2 =
function(y, x, Tr, Ts, tau, tau2, nu, R, inicio.y, inicio.sz, inicio.z, vc.mv){

beta1 = numeric(); beta2 = numeric(); sigma = numeric()
sigmaZ = numeric()
Z = matrix(0, nrow = (R + 1), ncol = length(inicio.z))
beta1[1] = inicio.y[1]; beta2[1] = inicio.y[2]; sigma[1] = inicio.y[3]
sigmaZ[1] = inicio.sz; Z[1,] = inicio.z
desv = sqrt(diag(vc.mv))

fdp.ga = function(x, al, be){
fun = x^(al - 1)*exp(-be*x); return(fun)}

fda.ga = function(x, al, be, a, b){
fun = function(x){x^(al - 1)*exp(-be*x)}; val = integrate(fun, a, b)$value
return(val)}

fdp.gt = function(x, al, be, a, b){
fun = function(x){x^(al - 1)*exp(-be*x)}; val = integrate(fun, a, b)$value
```

```
ft = (x^(al - 1)*exp(-be*x))/val
return(ft)}

lvY = function(param){
#Gardner + Homo
if (fun.mu == 1 && fun.esc == 1){
b1 = param[1]; b2 = param[2]; sg = param[3]; sz = param[4]
z = param[5:length(param)]
eta = Tr + (Ts - Tr)/(1 + b1*(x^b2))}
#van Genuchtem sem restricao + Homo
if (fun.mu == 2 && fun.esc == 1){
b1 = param[1]; b2 = param[2]; b3 = param[3]; sg = param[4]; sz = param[5]
z = param[6:length(param)]
eta = Tr+(Ts-Tr)/((1+(b1*x)^b2)^b3)}
#van Genuchten-Burdine + Homo
if (fun.mu == 3 && fun.esc == 1){
b1 = param[1]; b2 = param[2]; sg = param[3]; sz = param[4]
z = param[5:length(param)]
eta = Tr+(Ts-Tr)/((1+(b1*x)^b2)^(1-2/b2))}
#van Genuchten-Mualem + Homo
if (fun.mu == 4 && fun.esc == 1){
b1 = param[1]; b2 = param[2]; sg = param[3]; sz = param[4]
z = param[5:length(param)]
eta = Tr+(Ts-Tr)/((1+(b1*x)^b2)^(1-1/b2))}
#Fredlund-Xing + Homo
if (fun.mu == 5 && fun.esc == 1){
b1 = param[1]; b2 = param[2]; b3 = param[3]; sg = param[4]; sz = param[5]
z = param[6:length(param)]
eta = Ts*(1 - log(1 + x/15)/log(1 + (10^6)/15))*(1/((log(exp(1) + (x/b1)^b2))^b3))}
phi = eta + Z.mat%*%z
gamma = phi*exp(sg); rho = (1 - phi)*exp(sg)
dens = dbeta(y, gamma, rho)
acub = pbeta(Ts, gamma, rho); acua = pbeta(Tr, gamma, rho)
ldensy = log(dens) - log(acub - acua)
ldensz = log(dnorm(z, 0, sqrt(abs(sz))))
ll = sum(ldensy) + sum(ldensz)
return(-ll)}

lpriY = function(param){
if (fun.mu == 1){
lp = sum(-(1/tau + 1)*log(param[1]) - (1/tau)*(1/param[1]))
+ sum(-(1/tau + 1)*log(param[2]) - (1/tau)*(1/param[2]))
+ sum(log(dnorm(param[3], 0, tau)))
+ sum((-1)*log((abs(param[4]))^2 + tau2))
+ sum(log(dnorm(param[5:length(param)], 0, tau)))}

if (fun.mu == 3){
```

```
lp = sum(-(1/tau + 1)*log(param[1]) - (1/tau)*(1/param[1]))
+ sum(log(fdp.ga(param[2], 1/tau, 1/tau)) - log(fda.ga(param[2], 1/tau, 1/tau, 2, Inf)))
+ sum(log(dnorm(param[3], 0, tau)))
+ sum((-1)*log((abs(param[4]))^2 + tau2))
+ sum(log(dnorm(param[5:length(param)], 0, tau)))}

if (fun.mu == 4){
lp = sum(-(1/tau + 1)*log(param[1]) - (1/tau)*(1/param[1]))
+ sum(log(fdp.ga(param[2], 1/tau, 1/tau)) - log(fda.ga(param[2], 1/tau, 1/tau, 1, Inf)))
+ sum(log(dnorm(param[3], 0, tau)))
+ sum((-1)*log((abs(param[4]))^2 + tau2))
+ sum(log(dnorm(param[5:length(param)], 0, tau)))}
return(lp)}

lpostY = function(param){pp = lpriY(param) - lvY(param)
return(pp)}

for (r in 1:R){

b1n = beta1[r] + rnorm(1, 0, desv[1])
b2n = beta2[r] + rnorm(1, 0, desv[2])
sgn = sigma[r] + rnorm(1, 0, desv[3])
szn = sigmaZ[r] + rnorm(1, 0, desv[4])
zn1 = Z[r,1] + rnorm(1, 0, desv[5])
zn2 = Z[r,2] + rnorm(1, 0, desv[6])
zn3 = Z[r,3] + rnorm(1, 0, desv[7])
zn = c(zn1, zn2, zn3)

rb1 = exp(lpostY(c(b1n, beta2[r], sigma[r], sigmaZ[r], Z[r,]))
- lpostY(c(beta1[r], beta2[r], sigma[r], sigmaZ[r], Z[r,])))
if (is.na(rb1) || class(rb1) == "try-error"){beta1[r+1] = beta1[r]}
if (is.na(rb1) != TRUE && class(rb1) != "try-error"){
u = runif(1)
if (min(1, rb1) >= u){beta1[r+1] = b1n}
if (min(1, rb1) < u){beta1[r+1] = beta1[r]}}

rb2 = exp(lpostY(c(beta1[r+1], b2n, sigma[r], sigmaZ[r], Z[r,]))
- lpostY(c(beta1[r+1], beta2[r], sigma[r], sigmaZ[r], Z[r,])))
if (is.na(rb2) || class(rb2) == "try-error"){beta2[r+1] = beta2[r]}
if (is.na(rb2) != TRUE && class(rb2) != "try-error"){
u = runif(1)
if (min(1, rb2) >= u){beta2[r+1] = b2n}
if (min(1, rb2) < u){beta2[r+1] = beta2[r]}}

rsg = exp(lpostY(c(beta1[r+1], beta2[r+1], sgn, sigmaZ[r], Z[r,]))
- lpostY(c(beta1[r+1], beta2[r+1], sigma[r], sigmaZ[r], Z[r,])))
if (is.na(rsg) || class(rsg) == "try-error"){sigma[r+1] = sigma[r]}
```

```
if (is.na(rsg) != TRUE && class(rsg) != "try-error"){
u = runif(1)
if (min(1, rsg) >= u){sigma[r+1] = sgn}
if (min(1, rsg) < u){sigma[r+1] = sigma[r]}}

rsz = exp(lpostY(c(beta1[r+1], beta2[r+1], sigma[r+1], szn, Z[r,]))
- lpostY(c(beta1[r+1], beta2[r+1], sigma[r+1], sigmaZ[r], Z[r,])))
if (is.na(rsz) || class(rsz) == "try-error"){sigmaZ[r+1] = sigmaZ[r]}
if (is.na(rsz) != TRUE && class(rsz) != "try-error"){
u = runif(1)
if (min(1, rsz) >= u){sigmaZ[r+1] = szn}
if (min(1, rsz) < u){sigmaZ[r+1] = sigmaZ[r]}}

rvz = exp(lpostY(c(beta1[r+1], beta2[r+1], sigma[r+1], sigmaZ[r+1], zn))
- lpostY(c(beta1[r+1], beta2[r+1], sigma[r+1], sigmaZ[r+1], Z[r,])))
if (is.na(rvz) || class(rvz) == "try-error"){Z[(r+1),] = Z[r,]}
if (is.na(rvz) != TRUE && class(rvz) != "try-error"){
u = runif(1)
if (min(1, rvz) >= u){Z[(r+1),] = zn}
if (min(1, rvz) < u){Z[(r+1),] = Z[r,]}}}

chain = cbind(beta1, beta2, sigma, sigmaZ, Z)
colnames(chain) = NULL; rownames(chain) = NULL
return(chain)}
```

D.2 Algoritmo Metropolis-Hastings

A seguinte função gera e retorna uma cadeia MCMC de um vetor de parâmetros usando um algoritmo do tipo Metropolis-Hastings com passos de Metropolis e passeio aleatório como descrito no Capítulo 5.

```
MCMC.MHG.MIX = function(y, x, Tr, Ts, tau, R, vc.mv, lpost, fun.rep, fun.mu, fun.esc){

if (fun.rep == 1){
in.mu = c(0.1, 1.1); in.esc = c(0)
inicio.y = c(in.mu, in.esc)ç inicio.sz = c(0.1); inicio.z = rep(0, M)
inicio = c(inicio.y, inicio.sz, inicio.z)}

if (fun.rep == 2){
in.mu = c(0, 0); in.esc = c(0)
inicio.y = c(in.mu, in.esc); inicio.sz = c(0); inicio.z = rep(0, M)
inicio = c(inicio.y, inicio.sz, inicio.z)}

chain = NULL; velho = inicio

for (r in 1:R){
```

```
novo = try(as.numeric(rmultnorm(n = 1, mu = velho, vmat = vc.mv)), silent = TRUE)

ratio = try(exp(lpost(novo) - lpost(velho)), silent = TRUE)

if (is.na(ratio)){chain = rbind(chain, velho); velho = velho}
if (is.na(ratio) != TRUE){
u = runif(1)
if (min(1, ratio) >= u){chain = rbind(chain, novo); velho = novo}
if (min(1, ratio) < u){chain = rbind(chain, velho); velho = velho}}}

colnames(chain) = NULL rownames(chain) = NULL
return(chain)}
```

D.3 Algoritmo de seleção de modelos via mistura Bayesiana

```
eps = 1; n.mod = ncol(dpost)
modelos = seq(1, n.mod); inicio = rep(1/n.mod, n.mod)

SEL.MOD.MD = function(R, eps, n.mod, inicio, dpost){

chain.p = NULL

concat = function(mat){
vet = NULL
for (a in 1:ncol(mat)){vet = c(vet, mat[,a])}
return(vet)}

fun.p = function(p, post.mat){post.vet = concat(matrix(post.mat, nrow = 1))
pp = (p*post.vet)/sum(p*post.vet); return(pp)}

for (r in 1:R){

taui = NULL
for (i in 1:N){taui = rbind(taui, fun.p(inicio, dpost[i,]))}
z = NULL
for (i in 1:N){z = c(z, sample(1:n.mod, size = 1, prob = taui[i,]))}
ni = NULL
for (l in 1:ncol(dpost)){ni = c(ni, length(which(z == modelos[l])))}

p.novo = rdirichlet(1, (eps + ni)); chain.p = rbind(chain.p, p.novo)}

return(chain.p)}
```