UNIVERSIDADE FEDERAL DO AMAZONAS

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE DOUTORADO EM MATEMÁTICA - PDM-UFPA/UFAM

DOUTORADO EM MATEMÁTICA

NELSON LIMA DE SOUZA FILHO

ESTIMAÇÃO BAYESIANA EM MODELOS DE MISTURA DE REGRESSÕES COM CENSURA OU DADOS FALTANTES UTILIZANDO MISTURAS DE ESCALA DE DISTRIBUIÇÕES NORMAIS ASSIMÉTRICAS

MANAUS – AMAZONAS

2023

NELSON LIMA DE SOUZA FILHO

ESTIMAÇÃO BAYESIANA EM MODELOS DE MISTURA DE REGRESSÕES COM CENSURA OU DADOS FALTANTES UTILIZANDO MISTURAS DE ESCALA DE DISTRIBUIÇÕES NORMAIS ASSIMÉTRICAS

Tese apresentada ao Curso de Doutorado em Matemática do Programa de Doutorado em Matemática - PDM-UFPA/UFAM do Instituto de Ciências Exatas da Universidade Federal do Amazonas, como requisito parcial à obtenção do título de doutor em Matemática. Área de Concentração: Estatística

Orientador: Prof. Dr. Celso Rômulo Barbosa Cabral

Co-Orientador: Prof. Dr. Jeremias da Silva Leão

MANAUS – AMAZONAS

2023

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Souza Filho, Nelson Lima de
S729e Estimação bayesiana em modelos de mistura de regressões com censura ou dados faltantes utilizando misturas de escala de distribuições normais assimétricas / Nelson Lima de Souza Filho . 2023
115 f.: il.; 31 cm.

Orientador: Celso Rômulo Barbosa Cabral
Coorientador: Jeremias da Silva Leão
Tese (Doutorado em Matemática) - Universidade Federal do Amazonas.

1. Estimação bayesiana. 2. Mistura de modelos de regressão. 3. Mistura de escala da normal assimétrica. 4. Dados censurados. 5. Dados faltantes. I. Cabral, Celso Rômulo Barbosa. II. Universidade Federal do Amazonas III. Título

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Matemática

FOLHA DE APROVAÇÃO

"*ESTIMAÇÃO BAYESIANA EM MODELOS DE MISTURA DE REGRESSÃO COM CENSURA OU DADOS FALTANTES UTILIZANDO MISTURA DE ESCALA DE DISTRIBUIÇÕES NORMAIS ASSIMÉTRICAS*"

NELSON LIMA DE SOUZA FILHO

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Dr. Celso Rômulo Barbosa Cabral - UFAM - (Presidente)

Profª. Drª. Francyelle de Lima Medina - UFPE - (Membro Externo)

Profª. Drª. Larissa Avila Matos - UNICAMP - (Membro Externo)

Profª. Drª. Cibele Maria Russo Novelli - USP - (Membro Externo)

Prof. Dr. Aldo William Medina Garay - UFPE - (Membro Externo)

Manaus, 29 de Março de 2023

Documento assinado eletronicamente por **Cibele Maria Russo Novelli**, **Usuário Externo**, em 30/03/2023, às 12:34, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Larissa Avila Matos**, **Usuário Externo**, em 30/03/2023, às 12:47, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **FRANCYELLE DE LIMA MEDINA**, **Usuário Externo**, em 30/03/2023, às 16:37, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **ALDO WILLIAM MEDINA GARAY**, **Usuário Externo**, em 30/03/2023, às 16:40, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Aos meus pais Nelson e Sorteme com extrema gratidão.

Ao meu irmão Djalma Araújo pelo apoio e companheirismo.

À minha esposa Mariana Balbino pelo amor incondicional.

AGRADECIMENTOS

Agradeço,

À Deus que me trouxe do pouco e me fez um professor doutor.

Ao Professor Celso Rômulo pela ajuda, confiança, incentivo, disponibilidade e pela excelente orientação.

Aos meus amigos Rodrigo, David e Maurício pela amizade, ao meu irmão Djalma pelo exemplo de bondade e caráter e à minha amiga Camila pela grande ajuda com esse texto.

Ao meu grupo de amigos "do RPG", que sempre torceram por mim diretamente e sempre me ajudaram a distrair a mente nos momentos difíceis, seja matando dragões seja acertando um amigo em uma falha crítica ou até gastando todas suas magias para que o personagem do amigo não morra durante a sessão. Espero que sejamos eternos companheiros de batalha.

À Kezia, minha irmã de coração e a minha sogra Márcia Balbino por todo o apoio nas horas de tristeza e pouca fé.

Ao meu pai e minha mãe pela educação que me foi dada e pelo exemplo de pessoas humildes e honradas.

A todos os amigos que já se foram cujas presenças sempre serão lembradas e as contribuições para a formação do meu caráter, jamais esquecidas.

A todos os meus alunos e ex alunos que sempre desejaram o sucesso deste trabalho.

Aos professores do Departamento de Estatística da UFAM pelos ensinamentos.

"Senhor, fazei de mim um instrumento da vossa paz. Onde há ódio, que eu leve o amor. Onde há ofensa, que eu leve o perdão. Onde há discórdia, que eu leve a união. Onde há dúvida, que eu leve a fé. Onde há erro, que eu leve a verdade. Onde há desespero, que eu leve a esperança. Onde há tristeza, que eu leve a alegria. Onde há trevas, que eu leve a luz. Ó Mestre, Fazei que eu procure mais consolar que ser consolado; compreender que ser compreendido; amar que ser amado. Pois é dando que se recebe, é perdoando que se é perdoado, é morrendo que se vive para a vida eterna."

(São Francisco de Assis)

RESUMO

A utilização de modelos de misturas de regressões vem da necessidade de estudar dados com comportamento heterogêneo, em que temos a existência de populações distintas (grupos), cujas relações lineares entre a variável resposta e as varáveis preditoras diferenciam-se, entre os grupos, pelos coeficientes do modelo de regressão. Nesse contexto, é muito comum a utilização de modelos de misturas em que as componentes têm distribuição normal, porém a utilização de distribuições oriundas de uma família de misturas de escala de distribuições normais assimétricas - SMSN (*Scale Mixture of Skew-Normal*), no lugar da distribuição normal, é uma prática comum quando os dados possuem características como assimetria e caudas pesadas, aspectos esses que o modelo normal não comporta.

A ausência ou perda de algumas observações em um conjunto de dados é um padrão muito importante e muito abordado na literatura. Se tais dados não forem tratados de forma correta, por exemplo, quando são ignorados, podem gerar grandes prejuízos para as estimações dos parâmetros. Por isso, em parte desse texto, propomos a utilização de um modelo de misturas de regressões cujos erros têm distribuição oriunda da família SMSN, como forma de ajustar esse tipo de dados, especificamente quando as ausências se encontram tanto na variável resposta como nas covariáveis.

Outro problema bastante recorrente diz respeito à existência de uma estrutura de censura nas variáveis respostas dentro de cada grupo. Propomos lidar com esses problemas utilizando uma mistura de modelos tobit com erros aleatórios distribuídos na família SMSN. A modelagem utilizando essa família também acomoda possíveis comportamentos multimodais gerados pela estrutura dos grupos.

Desenvolvemos um algoritmo MCMC para realizar a estimação Bayesiana. Os modelos propostos são comparados com seus equivalentes simétricos, como os contidos na família SMN (*scale mixtures of Normal*), por meio de alguns critérios de seleção de modelos. Mostramos a eficiência do método proposto através da análise de dados simulados e reais.

Palavras-chave: Estimação Bayesiana, Mistura de Modelos de Regressão, MCMC, Distribuição Normal Assimétrica, Mistura de Escala da Normal Assimétrica, Dados Censurados, Dados Faltantes.

ABSTRACT

The use of mixed regression models comes from the need to study data with heterogeneous behavior, where there are distinct populations (groups), whose linear relationships between the response variable and the predictor variables differ, between the groups, by the coefficients of the regression model. In this context, it is very common to use models of mixtures in which the components have a normal distribution, but the use of distributions from a family of Scale Mixture of Skew-Normal - SMSN, instead of the normal distribution, it is a common practice when the data have characteristics such as asymmetry and heavy tails, aspects that the normal model does not support.

The absence or loss of some observations in a data set is a very important pattern and is much discussed in the literature. If such data are not treated correctly, for example, when they are ignored, they can cause great damage to the parameter estimates. For this reason, in part of this text, we propose the use of mixtures of regression models whose errors are distributed from the SMSN family, as a way of adjusting this type of data, specifically when absences are found in the response variable and in the covariates.

Another enough recurrent problem refers to the existence of a censored structure in the response variables within each group. We propose to deal with these problems using a mixture of tobit models with random errors distributed in the SMSN family. The modeling using this family also accommodates possible multimodal behaviors generated by the structure of the groups. We developed an MCMC algorithm to perform Bayesian estimation. The proposed models are compared with their symmetric equivalents, such as those contained in the SMN (scale mixtures of normal) family, using some model selection criteria. We show the efficiency of the proposed method through the analysis of simulated and real data.

Keywords: Bayesian Estimation, Mixture Regression Models, MCMC, Skew Normal Distribution, Scale Mixture of Skew Normal, Censored Data, Missing Data.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A suposição de normalidade, geralmente, não consegue comportar dados que contenham observações atípicas (*outliers*), as quais podem ter uma forte influência sobre as estimativas dos parâmetros do modelo. Observações atípicas não são exatamente observações influentes, mas podem ser, e se assim o forem podem prejudicar o ajuste dos modelos, principalmente os que não comportam esse tipo de de observações (o estudo de influência não será o foco desse texto mas foi tratado em alguns trabalhos como em Massuia *et al.* (2017)). Há, portanto, uma necessidade de explorarmos modelos mais flexíveis que consigam acomodar esses valores sem prejudicar as inferências.

Nessa linha de estudo de modelos mais robustos - entende-se, aqui, robustez como a tolerância do modelo a valores discrepantes dos dados e não a sensibilidade de suas estimativas com relação as observações atípicas-. Em nossa linha de estudo, temos os modelos baseado em misturas de escala da distribuição normal (SMN - Scale Mixture of Normal), desenvolvida por Andrews e Mallows (1974), tal estudo propôs um conjunto de distribuições que contêm não só a normal, mas também a Slash e a T de Student, que diferente no modelo normal são capazes de incorporar características de caudas pesadas, conseguindo, assim, abranger um número maior de possibilidades para a modelagem dos dados. Ainda na evolução dessa ideia, temos a família de distribuições SMSN, proposta por Branco e Dey (2001), que é uma mistura de escala da distribuição normal assimétrica, esse tipo de família de distribuições contém membros que têm caudas mais pesadas do que a normal, como a t de Student assimétrica, e também contém membros que possuem parâmetros de forma que conseguem regular a assimetria, no caso da normal assimétrica por exemplo. Um caso especial da família SMSN é exatamente a família de distribuições SMN.

No contexto de modelos de regressão, uma característica muito comum é a heterogeneidade da distribuição da variável resposta, modelos baseados em misturas de distribuições finitas conseguem capturar, com qualidade, essa característica dos dados reais. Embora os modelos de mistura de distribuições normais tenham a capacidade de aproximar qualquer tipo de distribuição, o número de componentes na mistura necessários para dar conta de distribuições de caudas pesadas pode ser muito grande. Mistura de modelos de regressão baseados em distribuições mais flexíveis, como a t de Student, podem ser considerados uma boa abordagem para modelar dados com caudas pesadas, como podemos ver com mais detalhes em Yao, Wei e Yu (2014).

Modelos de mistura de regressões em que os erros aleatórios têm distribuição contida na família SMSN surgem da necessidade de modelarmos a distribuição de dados não só com caudas pesadas e heterogeneidade mas também com assimetria.

Existem muitas motivações para se estudar modelos de misturas de regressões, dentre elas a capacidade desse modelo em ajustar dados com diferentes coeficientes do modelo de regressão para os grupos. Esses modelos são amplamente utilizados para investigar a relação entre variáveis provenientes de vários grupos homogêneos desconhecidos. Quandt (1972) introduziu pela primeira vez esses modelos sob o nome de "regressão de comutação" e, mais tarde, Späth (1979) denominou-os de "regressão linear de cluster".

Uma boa leitura sobre esse tema pode ser feita em McLachlan e Peel (2000), e para um ponto de vista Bayesiano, temos Frühwirth-Schnatter (2006). Em se tratando de um contexto frequentista de estimação via máxima verossimilhança, temos um grande número de trabalhos em diferentes áreas do conhecimento. As aplicações incluem marketing (Quandt e Ramsey,1978 e DeSarbo e Cron,1988), economia (Cosslett e Lee,1985), agricultura (Turner,2000) e psicometria (Liu e Lin,2014).

Neste trabalho, pretendemos explorar dois modelos particulares: Um modelo de mistura de regressões com censura na resposta e um modelo de mistura de regressões em que a variável resposta e as variáveis regressoras contêm dados ausentes, em ambos os modelos pretendemos ajustar os erros aleatórios com distribuições pertencentes a família SMSN. Pouco se falará desses modelos aqui, eles serão abordados nos capítulos 2 e 3. Uma informação importante, diz respeito aos modelos pertencentes a família SMSN aos quais iremos focar nosso trabalho. São eles: Normal Assimétrico, Slash Assimétrico, t de Student Assimétrico, Normal Contaminado Assimétrico e os seus correspondentes simétricos pertencentes à família SNM.

O problema de identificabilidade em modelos estatísticos se refere à capacidade de determinar valores únicos para os parâmetros do modelo a partir dos dados observados. Este problema é recorrente em modelos com misturas. Quando um modelo é identificável, é possível distinguir um conjunto específico de parâmetros de outros possíveis valores que o modelo poderia assumir. No entanto, em casos em que o modelo não é identificável, dois ou mais conjuntos diferentes de valores dos parâmetros do modelo podem gerar resultados semelhantes ou idênticos, o que pode levar a estimativas imprecisas ou errôneas dos parâmetros e consequentemente à interpretação inadequada dos resultados. Neste trabalho, iremos utilizar uma técnica inovadora chamada Clonagem de Dados (*Data Cloning*), proposta por Lele, Nadeem e Schmuland (2010), que tem a vantagem de permitir a identificação de modelos não identificáveis. Além disso, a

Clonagem de Dados pode ser aplicada a uma ampla variedade de modelos estatísticos, incluindo modelos hierárquicos, modelos lineares generalizados e modelos de mistura.

Para estimação Bayesiana de parâmetros em modelos mais complexos como os nossos, aproximações numéricas ou via amostragem de Monte Carlo tradicional não são a melhor alternativa. Uma boa forma de tratar esse tipo de problema é através de métodos de amostragem computacionalmente mais intensivos, como Markov Monte Carlo Chains (MCMC), que nos fornece uma classe de algoritmos para amostragem aleatória sistemática de distribuições de probabilidade de alta dimensão, como em Cabral, Souza e Leão (2022), que usaram esses métodos computacionais para o estudo de um modelo de erro nas variáveis, cuja covariável latente $\mathbf{x}_i$ era ajustada por uma mistura finita de distribuições SMSN. Para as nossas pretensões, o amostrador de Gibbs, que é um método MCMC, veja Gelfand (2000), se mostra como uma boa alternativa, pois nos permite a geração de amostras da distribuição a posteriori através das distribuições condicionais completas. O algoritmo do tipo Gibbs é uma das abordagens mais comuns para a amostragem via MCMC, principalmente quando o foco é o ajuste de modelos Bayesianos mais complexos, como em Massuia *et al.* (2017) e Nascimento e Abanto-Valle (2022), que estudaram o modelo de mistura de regressões na família SMSN.

Desenvolvemos um algoritmo do tipo Gibbs para inferência Bayesiana a posteriori e discutimos algumas medidas de qualidade de ajuste, que podem ser facilmente calculadas utilizando as amostras geradas pelo algoritmo de Gibbs, tais medidas são úteis para verificarmos qual modelo, dentre os muitos membros da família SMSN, é o melhor para ajustar um conjunto de dados em particular, mostrando principalmente as vantagens de se trabalhar um modelo de misturas quando a variável resposta nos dá indícios de uma estrutura com grupos distintos.

2 MISTURA DE ESCALA DA DISTRIBUIÇÃO NORMAL ASSIMÉTRICA (SMSN - SCALE MIXTURE SKEW-NORMAL)

Essa classe de distribuições foi proposta por Branco e Dey (2001) e contém versões estendidas de distribuições simétricas clássicas, como a t de Student assimétrica e slash assimétrica, além de toda a família de distribuições SMN, para mais detalhes veja Andrews e Mallows (1974) e Lange e Sinsheimer (1993). Antes de definirmos a classe SMSN, apresentamos o conceito fundamental de distribuição normal assimétrica (SN-*skew-normal*), que é uma extensão da distribuição normal, com a introdução de um parâmetro regulador de forma, que controla a assimetria da distribuição. Para mais detalhes, podemos ver Valle (2004) e Azzalini (2013).

Definição 1. *Uma variável aleatória Y tem uma distribuição normal assimétrica com parâmetro de localização μ, parâmetro de escala $\sigma^2 > 0$ e parâmetro de forma λ, denotado por $Y \sim \text{SN}(\mu, \sigma^2, \lambda)$, se sua função densidade de probabilidade é dada por*

$$\text{SN}(y|\mu, \sigma^2, \lambda) = 2\phi(y|\mu, \sigma^2)\Phi(\lambda p(y)), \quad y \in \mathbb{R} \tag{2.1}$$

em que $p(y) = (y-\mu)/\sigma$, $\Phi(\cdot)$ é a função de distribuição de uma variável aleatória com distribuição normal padrão, e $\phi(y; \mu, \sigma^2)$ é a função de densidade de uma distribuição normal com média μ e variância σ^2.

Antes de introduzirmos a família SMSN iremos definir o conceito de mistura de densidades.

Definição 2. *Sejam Y e S variáveis aleatórias, tal que S tem função de densidade $h(\cdot)$ e $P(S \in \mathscr{S}) = 1$, ambas definidas em um mesmo espaço de probabilidade. Para $s \in \mathscr{S}$, seja $g(\cdot|s)$ a função de densidade de $Y|S = s$. Então a função de densidade*

$$f(y) = \int_{\mathscr{S}} g(y|s)h(s)ds, \tag{2.2}$$

é chamada uma mistura de escala de densidades da família $\{g(\cdot|s); s \in \mathscr{S}\}$. S é chamado de fator de escala e $h(\cdot)$ é chamada de função de densidade da mistura.

As Definições 1 e 2 são de suma importância para a definição 3 que veremos a seguir, em que introduziremos a família SMSN.

Definição 3. *Uma variável aleatória Y pertence à família SMSN ou segue uma distribuição SMSN se*

$$Y = \mu + U^{-1/2}Z, \tag{2.3}$$

em que μ é o parâmetro de localização, $Z \sim \text{SN}(0,\sigma^2,\lambda)$ e U é uma variável aleatória com função de densidade $h(u|\nu)$, sendo que Z e U são independentes e $P(U>0)=1$.

A Definição 3 é um caso particular da Definição 2, em que a distribuição de Y é uma mistura de escala de distribuições normais assimétricas em que U é o fator de escala e $h(u|\nu)$ é a função de densidade da mistura, em que diferentes $h(u|\nu)$ levam Y a seguir diferentes distribuições dentro da família. Se U é uma variável degenerada em 1, então, a variável Y terá distribuição normal assimétrica com parâmetros μ,σ^2 e λ. A notação adotada para v.a. Y que segue uma distribuição pertencente a família SMSN é $Y \sim \text{SMSN}(\mu,\sigma^2,\lambda;h)$ e para sua função de densidade $\text{SMSN}(\cdot|\mu,\sigma^2,\lambda;h)$.

Quando condicionamos Y em U teremos (2.3) da forma

$$Y|U=u \sim \text{SN}(\mu,\sigma^2u^{-1},\lambda).$$

Assim, podemos escrever a distribuição marginal de Y como

$$\text{SMSN}(y|\mu,\sigma^2,\lambda;h) = \int_0^\infty \phi(y|\mu,\sigma^2u^{-1})\Phi\left(\frac{\lambda(y-\mu)}{\sigma u^{-1/2}}\right)h(u|\nu)du. \tag{2.4}$$

Se $\lambda=0$ temos que Y segue uma distribuição pertencente a família SMN.

O Teorema a seguir nos dá uma outra forma de definir a distribuição normal assimétrica, e que consequentemente auxiliará a definir as distribuições que pertencem à família SMSN, porém antes vamos apresentar a representação $\text{NT}(\mu,\sigma^2;(a,b))$ que denota a distribuição normal truncada no intervalo (a,b), ou seja, a distribuição de $Z|Z\in(a,b)$ em que $Z\sim\text{N}(\mu,\sigma^2)$. Henze (1986) criou uma outra forma de definir uma variável Y com distribuição normal assimétrica. Essa forma é dada abaixo pelo seguinte teorema, cuja prova pode ser encontrada no próprio artigo.

Teorema 1. *Se $Y\sim\text{SN}(\mu,\sigma^2,\lambda)$ então existem T e V independentes tais que*

$$Y=\mu+\Delta T+\tau^{1/2}V, \tag{2.5}$$

em que $T\sim\text{NT}(0,1;(0,\infty))$, $V\sim\text{N}(0,1)$, $\delta=\frac{\lambda}{\sqrt{1+\lambda^2}}$, $\Delta=\sigma\delta$ e $\tau=(1-\delta^2)\sigma^2$.

A partir do Teorema 1, temos a seguinte representação estocástica para a família SMSN

$$\begin{aligned} Y|T=t,U=u &\sim \text{N}(\mu+\Delta t,u^{-1}\tau),\\ T|U=u &\sim \text{NT}(0,u^{-1};(0,\infty)),\\ U &\sim h(\cdot|\nu). \end{aligned}$$

Essa representação é muito importante para a construção de algoritmos MCMC, como podemos ver em Cabral, Bolfarine e Pereira (2008) ou do algoritmo EM em Cabral, Lachos e Prates (2012) . Também chamada de estrutura de dados aumentados ou representação de dados aumentados, essa representação será o ponto de partida para a construção do nosso algoritmo do tipo Gibbs.

A recuperação dos parâmetros originais pode ser feito de forma direta com o sistema abaixo

$$\sigma^2 = \tau + \Delta^2 \quad \text{e} \quad \lambda = \Delta\tau^{-1/2}. \tag{2.6}$$

Note que diferentes distribuições para o fator de escala U geram diferentes membros da família SMSN.

Uma quantidade importante que iremos utilizar com frequência após a definição do modelo de regressão é $k_m = \mathrm{E}[U^{-1/m}]$. Os cálculos de k_m para diferentes membros da família SMSN podem ser encontrados de forma mais detalhada em Basso *et al.* (2010) e Lachos, Ghosh e Arellano-Valle (2010). A seguir iremos apresentar alguns diferentes membros da família SMSN.

Normal assimétrica (SN-Skew-Normal): Quando $U = 1$ com probabilidade 1, gerando $k_m = 1$ e sua função de densidade dada por (2.1).

Slash assimétrica - (SSL-Skew-Slash): Quando $U \sim \mathrm{B}(\nu, 1)$, que é a distribuição Beta com média $\frac{\nu}{\nu+1}$, assim teremos

$$k_m = \frac{\nu}{\nu - \frac{m}{2}}, \quad \nu > \frac{m}{2}. \tag{2.7}$$

Se uma variável Y segue uma distribuição slash assimétrica, então, denotamos por $Y \sim \mathrm{SL}(\mu, \sigma^2, \lambda, \nu)$ com função densidade de probabilidade

$$\mathrm{SSL}(y|\mu, \sigma^2, \lambda, \nu) = 2\nu \int_0^1 u^{\nu-1} \phi(y|\mu, \sigma^2 u^{-1}) \Phi\left(u^{\frac{1}{2}} \lambda p(y)\right) du, \quad y \in \mathbb{R}, \tag{2.8}$$

em que ν representa os graus de liberdade da distribuição.

t de Student assimétrica - (ST-Skew-t): Quando $U \sim \mathrm{Gamma}(\nu/2, \nu/2)$ (distribuição Gamma com média 1), logo temos

$$k_m = \frac{(\nu/2)^{(m/2)} \Gamma((\nu - m)/2)}{\Gamma(\nu/2)}. \tag{2.9}$$

Se Y segue uma distribuição t de Student assimétrica, denotamos que $Y \sim \text{ST}(\mu, \sigma^2, \lambda, \nu)$. Outra forma de definir essa distribuição é por meio de sua função de densidade

$$\text{ST}(y|\mu,\sigma^2,\lambda,\nu) = \frac{2\Gamma(\frac{\nu+1}{2})\left(1+\frac{p(y)^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}}{\Gamma(\nu/2)(\pi\nu\sigma^2)^{1/2}} T\left(\frac{\lambda p(y)(\nu+1)^{\frac{1}{2}}}{(\nu+p(y)^2)^{\frac{1}{2}}}|\nu+1\right), \quad (2.10)$$

em que $T(\cdot|\nu)$ é a função de distribuição de uma variável aleatória que segue distribuição t de Student com média zero, parâmetro de escala 1 e grau de liberdade ν. Além disso, quando $\nu \to \infty$, a distribuição $\text{ST}(\mu, \sigma^2, \lambda, \nu)$ aproxima-se da distribuição $\text{SN}(\mu, \sigma^2, \lambda)$.

Normal contaminada assimétrica - (SCN- Skew Contaminated Normal): Quando U é uma variável discreta tal que $P(U = \eta) = \rho$ e $P(U = 1) = 1 - \rho$. Assim temos que,

$$k_m = \frac{\rho}{\eta^{m/2}} + 1 - \rho.$$

A função de densidade de uma variável aleatória com distribuição Normal contaminada assimétrica é dada por

$$\begin{aligned}\text{SCN}(y|\mu,\sigma^2,\lambda,\rho,\eta) = 2\Bigg\{&\rho\phi(y|\mu,\eta^{-1}\sigma^2)\Phi\left(\frac{\lambda(y-\mu)}{\eta^{-1/2}\sigma}\right) \\ &+(1-\rho)\phi(y|\mu,\sigma^2)\Phi\left(\frac{\lambda(y-\mu)}{\sigma}\right)\Bigg\},\end{aligned} \quad (2.11)$$

em que $0 < \rho < 1$ e $0 < \eta < 1$. A distribuição normal contaminada assimétrica reduz-se à distribuição normal assimétrica quando $\eta = 1$.

2.1 MISTURA FINITA DE MODELOS DE REGRESSÃO

Nesta seção, iremos apresentar os conceitos de mistura finita e também como incorporar esse conceito ao modelo de regressão com erros na família SMSN. Para isso vamos definir o conceito de mistura finita de densidades.

A distribuição de mistura pode assumir algumas formas como na Definição 3, porém neste ponto, estamos interessados no caso particular em que a variável aleatória S, da Definição 2, segue uma distribuição discreta. A definição 4 descreve a variável Z como o fator de escala.

Definição 4. *Sejam Y e Z variáveis aleatórias, tal que $P(Z = j) = p_j$ com $j = 1, \ldots, G$, ambas definidas em um mesmo espaço de probabilidade. Seja $g(\cdot|Z = j) = g_j(\cdot)$ a função de densidade condicional de $Y|Z = j$. Então, a função de densidade*

$$f(y) = \sum_{j=1}^{G} p_j g_j(y), \quad (2.12)$$

é chamada uma mistura finita de densidades.

A função de densidade $g_j(y)$ é denominada j-ésima componente da mistura e p_j é conhecido como j-ésimo peso da mistura. Z é uma variável classificadora, não observada, que indica de qual das G componentes y pertence. Se $Z = j$ então y pertence à j-ésima componente da mistura, $g_j(\cdot)$, e terá peso p_j tal que $\sum_{j=1}^{G} p_j = 1$.

Uma extensão natural para modelos de mistura finita é construída de tal forma que os parâmetros específicos dos componentes sejam relacionados, de alguma forma, com algumas covariáveis, como é o caso dos modelos de mistura de regressões. Para esse tipo especial de modelo, o coeficiente de regressão associado as covariáveis é diferente para cada componente da mistura.

Na definição 5 iremos ver o conceito de modelo de misturas finitas de regressões com componentes na família SMSN.

Definição 5. *Uma mistura finita de modelos de regressão cujas componentes são SMSN é dada por (2.12) em que,*

$$g_j(y_i) = \text{SMSN}(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}_j + \Delta_j b, \sigma_j^2, \lambda_j; h), \tag{2.13}$$

tal que $\mathbf{x}_i^\top$ *é a i-ésima linha da matriz de planejamento para* $i = 1, \cdots, n$, $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj})^\top$ *é um vetor de parâmetros da regressão p-dimensional,* $b = -k_1\sqrt{\frac{2}{\pi}}$ *e* $\Delta_j = \delta_j\sigma_j$ *em que* $\delta_j = \frac{\lambda_j}{\sqrt{1+\lambda_j^2}}$.

A quantidade b é definida de tal forma que $\text{E}[Y_i|Z_i = j] = \mathbf{x}_i^\top \boldsymbol{\beta}_j$, chamamos isso de centralização do modelo, veja Basso *et al.* (2010) e Massuia *et al.* (2017). Ao modelo de mistura de regressões em (2.13) daremos o nome de FMR-SMSN, isso nos auxiliará quanto à referência do modelo no decorrer do trabalho.

Sendo G o número de componentes da mistura e pela Definição 5, temos que a função de densidade de Y_i é dada por

$$f(y_i|\mathbf{x}_i, \boldsymbol{\psi}) = \sum_{j=1}^{G} p_j \text{SMSN}(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}_j + b\Delta_j, \sigma_j^2, \lambda_j; h), \tag{2.14}$$

em que $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}, \mathbf{p}, \nu)^\top$ ou $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}, \mathbf{p}, \rho, \eta)^\top$ quando estivermos tratando do modelo SCN, tal que $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_G^\top)^\top$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_G^2)^\top$, $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_G)^\top$ e $\mathbf{p} = (p_1, \cdots, p_G)^\top$.

A função de log-verossimilhança do modelo FMR-SMSN é dada por

$$\ell(\mathbf{y}|\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \sum_{j=1}^{G} p_j \text{SMSN}(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}_j + b\Delta_j, \sigma_j^2, \lambda_j; h).$$

Iremos redefinir $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{iG})^\top$ de tal forma que

$$Z_{ij} = \begin{cases} 1, \text{ se } Y_i \text{ pertence a componente } j \\ 0, \text{ c.c,} \end{cases} \tag{2.15}$$

em que $\sum_{j=1}^{G} Z_{ij} = 1$. Cada vetor aleatório $\mathbf{Z}_i$ segue uma distribuição multinomial, com probabilidades $p_1, \cdots, p_G$, ou seja,

$$P(\mathbf{Z}_i = \mathbf{z}_i) = \prod_{j=1}^{G} p_j^{z_{ij}},$$

e seja também $\mathbf{Z} = (\mathbf{Z}_1^\top, \cdots, \mathbf{Z}_n^\top)^\top$ em que os n vetores que compõem $\mathbf{Z}$ são independentes entre si.

Por (2.14) e (2.15) podemos afirmar que

$$Y_i | Z_{ij} = 1 \sim \text{SMSN}(\mathbf{x}_i^\top \boldsymbol{\beta}_j + b\Delta_j, \sigma_j^2, \lambda_j; h).$$

Considerando o Teorema 1 temos que a representação estocástica do modelo FMR-SMSN é

$$\begin{aligned} Y_i | T_i = t_i, U_i = u_i, Z_{ij} = 1 &\sim \text{N}(\mathbf{x}_i^\top \boldsymbol{\beta}_j + \Delta_j(t_i + b), u_i^{-1}\tau_j), \\ T_i | U_i = u_i &\sim \text{NT}(0, u_i^{-1}; (0, \infty)), \\ U_i &\sim h(\cdot | \nu), \\ \mathbf{Z}_i &\sim \text{Multinomial}(1, p_1, \cdots, p_G), \end{aligned} \tag{2.16}$$

(2.17)

para $i = 1, \cdots, n$ e $j = 1, \cdots, G$.

Note que os parâmetros originais, σ_j^2 e λ_j, podem ser recuperados com (2.6).

3 MODELO DE MISTURA FINITAS DE REGRESSÕES COM CENSURA

3.1 INTRODUÇÃO

O interesse em estudar modelos de regressão quando a variável dependente pode ser censurada é um dos objetivos desse capítulo. Limitações nos equipamentos de medição ou no desenho experimental são algumas das várias situações nas quais pode ocorrer censura.

Um exemplo clássico de censura pode ser visto em Breen *et al.* (1996) em que uma classe é submetida a um exame escolar cuja nota para um aluno ser aprovado é de 40 pontos. Cada aluno recebe um atestado no qual consta se ele foi aprovado ou não. Porém a nota no exame aparece somente no atestado dos alunos aprovados, assim a nota de um determinado aluno só é completamente observada se for maior ou igual a 40, caso contrário observamos o valor 39. Esse tipo de experimento pode ser denotado como censurado à esquerda. Suponha que queiramos estudar a relação entre a pontuação dos alunos com variáveis como: idade, escolaridade e região em que nasceu, logo fica evidente a necessidade de um modelo de regressão que incorpore essa forma da variável resposta. Neste caso, um modelo de regressão com respostas censuradas pode ser uma alternativa adequada.

O tema de variáveis censuradas é muito comum em econometria. Veja, por exemplo, Al-Malkawi (2007) e Komrattanapanya e Suntraruk (2013). Estudos envolvendo variáveis como remuneração podem, podem ser desenvolvidos por meio de um modelo de censura à esquerda, em que o zero pode ser considerado como censura para pessoas que não são remuneradas, ou seja, taxa salarial observada zero para pessoas que não trabalham, e taxa salarial registrada com o valor real positivo para os indivíduos que trabalham. Um estudo muito popular envolvendo remuneração de mulheres pode ser visto em Mroz *et al.* (1987).

A abordagem de modelos de regressão normal censurados, também conhecidos como modelos Tobit, tornou-se bastante comum na literatura, veja por exemplo Nelson (1977), Vaida e Liu (2009), Guo, Sayed e Essa (2020), Hou, Huo e Leng (2020) e Lemus *et al.* (2021). No entanto como o modelo Tobit supõe que os erros seguem distribuição normal, existe uma necessidade de estudar uma estrutura mais flexível para a distribuição destes erros, incluindo características como caudas mais pesadas e assimetria. Em Arellano-Valle *et al.* (2012) e Massuia *et al.* (2015) podemos ver modelos de regressão censurados com resposta seguindo distribuição t de Student univariada e em Matos *et al.* (2018) o caso multivariado. Ainda nesse contexto, podemos ver os trabalhos Garay *et al.* (2015) e Garay *et al.* (2017), que estenderam

esse conceito para variável resposta pertencendo à família SMN tanto no contexto Bayesiano quando frequentista respectivamente e da mesma forma Massuia *et al.* (2017) e Matos *et al.* (2018) que estudaram o modelo de regressão com respostas censuradas e erros com distribuição na família SMSN, no contexto Bayesiano e frequentista respectivamente. Ainda na área de estudo envolvendo censuras, podemos citar Alencar *et al.* (2022) que trabalharam um modelo de mistura de efeitos mistos para dados longitudinais com censura e Alencar, Matos e Lachos (2022) que trabalharam um modelo de mistura finita de dados censurados e ausentes usando a distribuição normal assimétrica multivariada.

Em Karlsson e Laitila (2014) foi sugerido, utilizando abordagem frequentista, a utilização de uma mistura finita de modelos Tobit, para estimar modelos de regressão com variáveis com resposta censurada. Zeller *et al.* (2019) estenderam esses conceitos, e estudaram o modelo de mistura de regressões com erros na família SMN e respostas censuradas em um contexto frequentista utilizando um algoritmo do tipo EM. Ainda nesse sentido e também em um contexto frequentista, Mirfarah, Naderi e Chen (2021) propuseram um modelo de mistura de regressões censurado com erros tendo distribuição na família SMN, em que os pesos são modelados por uma regressão logística multinomial utilizando as covariáveis.

Nesse capítulo temos a intenção de estender o trabalho de Massuia *et al.* (2017), porém para o caso de heterogeneidade na distribuição da resposta.

3.2 MISTURA FINITA DE MODELOS DE REGRESSÃO COM CENSURA NA RESPOSTA

A definição de censura que vamos apresentar agora é um dos conceitos de censura existentes, intitulado de censura à esquerda.

Seja Y_i uma variável aleatória parcialmente observável, de tal forma que observamos V_i que é descrita como,

$$V_i = \begin{cases} c_i, & \text{se } Y_i \leq c_i \\ Y_i, & \text{se } Y_i > c_i, \end{cases} \tag{3.1}$$

em que c_i é um valor conhecido e $i = 1, \ldots, n$.

Optamos por definir apenas a censura à esquerda, pois a extensão desse conceito para a censura à direita é natural se definirmos o nível de censura c_i como $-c_i$ e a resposta Y_i como $-Y_i$. Já para o conceito de censura intervalar podemos ver Pu e Li (1999) e Li e Pu (2003).

Chamaremos de FMR-SMSN-CR o modelo dado por (2.13) e (3.1), que é o nosso

modelo proposto para esse capítulo, tem sua função de log-verossimilhança dada por,

$$\ell(\boldsymbol{\upsilon},\boldsymbol{\rho}|\boldsymbol{\psi}) = \sum_{i=1}^{n} \log\left\{ \sum_{j=1}^{G} p_j \left[F_{\text{SMSN}}\left(\frac{c_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta_j b}{\sigma_j} \right) \right]^{\xi_i} \right. \\ \left. \times \left[\text{SMSN}(\upsilon_i | \mathbf{x}_i^\top \boldsymbol{\beta}_j + \Delta_j b, \sigma_j^2, \lambda_j, \nu) \right]^{1-\xi_i} \right\}, \tag{3.2}$$

na qual $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_n)^\top$ denota o vetor de observações de $\mathbf{V} = (V_1, \ldots, V_n)^\top$ e $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)^\top$ é um vetor de indicadores em que cada elemento assume valor 1 se a observação é censurada e 0 caso contrário, com $i = 1, \ldots, n$. A notação F_{SMSN} corresponde à função de distribuição de uma SMSN$(0, 1, \lambda_j, \nu)$. Para o nosso modelo optamos por utilizar os graus de liberdade iguais para as G populações, essa escolha tem principalmente duas motivações, primeiramente teremos um modelo com menos parâmetros, o que é bom levando em consideração a parcimônia e também teremos uma redução significativa no tempo computacional.

3.3 MODELAGEM BAYESIANA

Para a estimação do modelo FMR-SMSN-CR, utilizaremos um algoritmo do tipo Gibbs. A construção desse algoritmo utilizará principalmente a representação estocástica (2.16) para o desenvolvimento das condicionais completas, assim como apresentado em Garay *et al.* (2015) e Massuia *et al.* (2017).

3.3.1 Distribuições a Priori

Para a construção de um modelo sob o enfoque Bayesiano é necessário especificarmos distribuições a priori para o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\tau}, \mathbf{p}, \nu)^\top$. Note que $\boldsymbol{\theta}$ é o vetor utilizando a reparametrização dada no Teorema 1, em que $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_G)^\top$, $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_G)^\top$. Assim como em Nascimento e Abanto-Valle (2022) faremos a suposição de que para cada componente da mistura $\boldsymbol{\phi}_j = (\boldsymbol{\beta}_j, \Delta_j) \sim \text{N}_{p+1}(\mathbf{b}_0, \mathbf{B}_0)$, $\tau_j | f_\tau \sim \text{InvGamma}(e_\tau, f_\tau)$, que denota a distribuição gama inversa com média $e_\tau / f_\tau - 1$ e $f_\tau \sim \text{Gamma}(g, h)$, que denota a distribuição gama com média g/h. Essa hierarquia em $\tau_j | f_\tau$ é uma sugestão dada por Richardson e Green (1997) para um modelo de misturas de normais univariadas. Os hiperparâmetros $\mathbf{b}_0$, $\mathbf{B}_0$, e_τ, g e h são conhecidos.

As escolhas de valores de hiperparâmetros das distribuições a priori contidas nesse trabalho, se baseiam na ideia de expressar pouco conhecimento prévio para o parâmetro em questão, ou seja, parâmetros como a matriz de covariância $\mathbf{B}_0$ serão assumidos como sendo

diagonal, com valores iguais a 100. O hiperparâmetro de localização $\mathbf{b}_0$ será considerado nulo. Para o hiperparâmetro e_τ geralmente utilizamos 2 e consideramos g e h pequenos e positivos em geral iguais a 0,01. Para o vetor de pesos, utilizaremos uma suposição comum $\mathbf{p} \sim \text{Dir}(\kappa_1, \cdots, \kappa_G)$, uma distribuição Dirichlet com hiperparâmetros $\kappa_1 = \kappa_2 = \cdots = \kappa_G = 1$.

O parâmetro de graus de liberdade ν oriundo da distribuição $h(\cdot|\nu)$, que está ligada as diferentes distribuições da família SMSN, tem prioris que dependem do modelo em questão. Assim como em Cabral, Lachos e Madruga (2012), para o modelo FMR-ST-CR, definiremos $\nu|\gamma \sim \text{Exp}(\gamma)$, que denota a distribuição exponencial com média $1/\gamma\,(\gamma > 0)$ e $\gamma \sim \text{Unif}(a_\gamma, b_\gamma)$, ou seja, uma distribuição uniforme no intervalo (a_γ, b_γ). Para o modelo FMR-SSL-CR utilizaremos $\nu \sim \text{Gamma}(\alpha_\nu, \gamma_\nu)$. Assim como em Cabral, Souza e Leão (2022), utilizaremos $a_\gamma = 0,04$, $b_\gamma = 0,5$ e $\alpha_\nu = \gamma_\nu = 0,01$.

Não especificaremos distribuições a priori para os parâmetros ρ e η dos modelos FMR-CN-RC e FMR-SCN-RC pois serão considerados conhecidos, o motivo dessa decisão se dá por um problema que enfrentamos a cerca da convergência desses dois parâmetros, nos estudos de recuperação de parâmetros os mesmos convergiram para zero. Um problema similar pode ser visto em Rosa, Padovani e Gianola (2003) e Liu (1996), que discutiu o problema e observou que isso seria resultado de uma "confusão" entre o vetor de fatores de escala $\mathbf{U} = (U_1, \cdots, U_n)^\top$ e η. Liu (1996) propôs integrar os fatores de escala obtendo a distribuição conjunta de $\mathbf{U}$ e η e usou um método de aceitação-rejeição para extrair amostras da distribuição condicional resultante de η, enquanto Rosa, Padovani e Gianola (2003) utilizaram um algortimo de Metropolis-Hastings para amostrar da distribuição conjunta de $\mathbf{U}$ e η. A nossa forma de resolver esse problema, será bem diferente e se baseará na escolha dos valores destes parâmetros por meio de um estudo, no qual avaliaremos várias combinações de ρ e η, comparando os ajustes do modelo por meio de critérios de seleção.

Assumiremos independência entre os vetores de parâmetros de $\boldsymbol{\theta}$, de tal maneira que a distribuição a priori conjunta será dada por,

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\Delta})\pi(\mathbf{p})\pi(\nu|\gamma)\pi(\gamma)\prod_{j=1}^{G}\pi(\tau_j|f_\tau)\pi(f_\tau).$$

Segundo Massuia *et al.* (2017), essa suposição de independência levará a boas propriedades matemáticas como distribuições conjugadas, gerando assim uma facilidade maior na implementação de um amostrador de Gibbs, e se a suposição de independência entre os parâmetros não for válida, a distribuição a posteriori corrigirá o problema sem prejudicar o processo inferencial. Usando softwares Bayesianos existentes como JAGS (Plummer *et al.*

(2003)) ou Stan (Carpenter *et al.* (2017)), um algoritmo MCMC pode ser implementado através da representação estocástica (2.16), outra escolha pode ser o desenvolvimento de condicionais completas e sua implementação utilizando o software R (Team *et al.* (2013)), a qual iremos explanar nessa parte do trabalho.

3.3.2 Algoritmo MCMC

Em modelos complexos, como é o nosso caso, existe uma natural impossibilidade de obtermos estimadores por meio de medidas estatísticas da distribuição a posteriori, como valor esperado ou mediana, isso se dá pelo fato dessas medidas serem integrais complexas. Uma boa saída para isso é a utilização de um algoritmo do tipo MCMC para aproximar essas medidas. Essa técnica já bem difundida, utiliza a geração de amostras das distribuições a posteriori dos parâmetros. Com essas amostras podemos proceder com uma estimação aproximada utilizando alguma medida de tendência central como média ou mediana amostral.

O algoritmo do tipo Gibbs que utilizaremos consiste em um gerador de amostras que atualiza cada parâmetro de forma isolada, um por cada vez. São geradas amostras da distribuição a posteriori dos parâmetros a partir das distribuições condicionais completas, extraída da representação de dados aumentados fornecida em (2.16).

Para mais detalhes sobre do amostrador de Gibbs e resultados teóricos, consulte Gelfand (2000). Para obter amostras da distribuição a posteriori do modelo FMR-SMSN-CR, basta seguir os passos do algoritmo abaixo.

Antes de darmos prosseguimento a construção do nosso algoritmo, seja o conjunto de índices $A_j = \{i \in \{1,...,n\}; Z_{ij} = 1\}$ e $\mathbf{T} = (T_1, \ldots, T_n)^\top$, $U = (\mathbf{U}_1, \ldots, \mathbf{U}_n)^\top$ e n_j como o cardinal de A_j, com $j = 1, \cdots, G$.

Passo 1. Gere $\mathbf{p}$ a partir de $\pi(\mathbf{p}|\mathbf{z}_1, \cdots, \mathbf{z}_n)$, que representa uma Dirichilet da forma

$$\text{Dir}(n_1 + \kappa_1, \cdots, n_G + \kappa_G).$$

Passo 2. Para cada $i = 1, \cdots, n$; se $\upsilon_i = c_i$ gere de forma independente y_i a partir da distribuição $\pi(y_i|\upsilon_i, t_i, u_i, z_{ij} = 1, \boldsymbol{\beta}_j, \Delta_j, \tau_j, \nu)$, que representa uma distribuição normal truncada

$$\text{NT}(\mathbf{x}_i^\top \boldsymbol{\beta}_j + \Delta_j t_i, u_i^{-1}\tau_j; (-\infty, c_i]),$$

caso contrário $\upsilon_i = y_i$ com $j = 1, \cdots, G$.

Passo 3. Pra cada $i = 1, \cdots, n$, gere Z_{ij} independentemente a partir da distribuição discreta a seguir

$$p(Z_{ij} = 1 | y_i, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\tau}, \nu, \mathbf{p}) = \frac{p_j \text{SMSN}(y_i | \mathbf{x}_i^\top \boldsymbol{\beta}_j + \Delta_j b, \Delta_j, \tau_j; h)}{\sum_{k=1}^{G} p_k \text{SMSN}(y_i | \mathbf{x}_i^\top \boldsymbol{\beta}_k + \Delta_k b, \Delta_k, \tau_k; h)}, \quad j = 1, \cdots, G.$$

Passo 4. Gere $\boldsymbol{\phi}_j = (\boldsymbol{\beta}_j, \Delta_j)^\top$ a partir da distribuição $\pi(\boldsymbol{\phi}_j | \mathbf{y}, \mathbf{z}, \mathbf{t}, \mathbf{u}, \tau_j, \Delta_j, \nu)$, que representa uma normal multivariada $\text{N}_{p+1}(\boldsymbol{\mu}_{\phi_j}, \boldsymbol{\Sigma}_{\phi_j})$ em que

$$\begin{aligned} \boldsymbol{\Sigma}_{\phi_j} &= \left[\mathbf{H}_j^\top \boldsymbol{\Sigma}_{U_j} \mathbf{H}_j \tau_j^{-1} + \mathbf{B}_0^{-1} \right]^{-1} \\ \boldsymbol{\mu}_{\phi_j} &= \boldsymbol{\Sigma}_{\phi_j} \left[\mathbf{H}_j^\top \boldsymbol{\Sigma}_{U_j} (\mathbf{Y}_j - \Delta_j b) \tau_j^{-1} + \mathbf{B}_0^{-1} \mathbf{b}_0 \right] \end{aligned}$$

sendo $\mathbf{H}_j = [\mathbf{X}_j \;\; \mathbf{T}_j]$, $\boldsymbol{\Sigma}_{U_j} = \text{diag}(\mathbf{U}_j)$, $\mathbf{X}_j$ uma matriz cujas linhas são $\mathbf{x}_i^\top$, com $i \in A_j$, $\mathbf{T}_j$, $\mathbf{U}_j$ e $\mathbf{Y}_j$ são vetores com elementos T_i, U_i e Y_i, $i \in A_j$, respectivamente.

Passo 5. Para cada $i = 1, \cdots, n$ gere de forma independente t_i a partir da distribuição $\pi(t_i | y_i, u_i, \boldsymbol{\beta}_j, \Delta_j, \tau_j, z_{ij} = 1)$, que representa uma distribuição normal truncada

$$\text{TN}\left(\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - b\Delta_j)\Delta_j}{\Delta_j^2 + \tau_j}, \frac{\tau_j}{u_i(\Delta_j^2 + \tau_j)}; (0, \infty) \right).$$

Passo 6. Gere τ_j a partir da distribuição $\pi(\tau_j | \mathbf{y}, \mathbf{t}, \mathbf{u}, \mathbf{z}, \boldsymbol{\beta}_j, \Delta_j, \nu, f_\tau)$, que representa uma gama inversa

$$\text{InvGamma}\left(\frac{n_j}{2} + e_\tau, f_\tau + \frac{1}{2} \sum_{i \in A_j} u_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta_j (t_i + b))^2 \right),$$

em que o primeiro parâmetro é de forma e o segundo de taxa.

Passo 7. Gere f_τ a partir da distribuição $\pi(f_\tau | \gamma)$, que representa uma Gamma$(e_\tau G + g; \sum_{j=1}^{G} \tau_j^{-1} + h)$.

Passo 8. Para cada $i = 1, \cdots, n$ gere de forma independente u_i a partir da distribuição $\pi(u_i | y_i, t_i, \boldsymbol{\beta}_j, \Delta_j, \tau_j, \nu, z_{ij} = 1)$, que representa uma distribuição,

1. Para o modelo FMR-ST-CR

$$\text{Gamma}\left(\frac{\nu}{2} + 1, \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta_j (t_i + b))^2 \tau_j^{-1} + t_i^2 + \nu}{2} \right).$$

2. Para o modelo FMR-SSL-CR

$$\mathrm{TG}\left(1+\nu, \frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}_j-\Delta_j(t_i+b))^2\tau_j^{-1}+t_i^2}{2}; (0,1)\right),$$

em que TG representa a a distribuição gamma truncada.

3. Para o modelo FMR-SCN-CR uma distribuição discreta dicotômica assumindo η com probabilidade $p_2^*/p_1^*+p_2^*$ e 1 com probabilidade $1-p_2^*(1+p_1^*)/p_1^*$, tal que

$$\begin{aligned} p_1^* &= \rho\eta\exp\left\{-\frac{\eta}{2}\left[\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}_j-\Delta_j(t_i+b))^2}{\tau_j}+t_i^2\right]\right\} \\ p_2^* &= (1-\rho)\exp\left\{-\frac{(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}_j-\Delta_j(t_i+b))^2+t_i^2\tau_j}{2\tau_j}\right\}. \end{aligned}$$

Passo 9. Para a geração de ν iremos analisar os modelos FMR-ST-CR e FMR-SSL-CR individualmente,

1. Para o modelo FMR-SSL-CR gere ν de $\pi(\nu|\mathbf{u})$, que representa a distribuição

$$\mathrm{Gamma}(\alpha_\nu+n, \gamma_\nu-\sum_{i=1}^n\log u_i).$$

2. No modelo FMR-ST-CR é impossivel conseguir uma distribuição condicional completa conjugada para ν logo utilizaremos um algoritmo de Metropolis-Hastings (MH).

   (i) Gere γ a partir da distribuição $\pi(\gamma|\nu)$ que é $\mathrm{TG}(2,\nu;(a_\gamma,b_\gamma))$, gama truncada no intervalo (a_γ,b_γ) com parâmetro de forma 2 e parâmetro de taxa ν.

   (i) Gere ν, utilizando um passo de MH, com densidade marginal condicional

$$\pi(\nu|\gamma,\mathbf{y},\mathbf{z},\beta_j,\Delta_j,\tau_j,\mathbf{p}) \propto e^{-\gamma\nu}\sum_{i=1}^n\log\left\{\sum_{j=1}^G p_j\mathrm{ST}(y_i|\mathbf{x}_i^\top\boldsymbol{\beta}_j,\sigma_j^2,\lambda_j,\nu)\right\}. \quad (3.3)$$

   Dada a observação $\nu^{(l-1)}$ obtida na iteração $l-1$, gere um candidato a nova observação ν^* utilizando como distribuição proposta $\mathrm{LN}(\log\nu^{(l-1)},\delta_\nu^2)$, que representa uma logNormal com densidade,

$$\pi(\nu^*|\nu^{(l-1)}) = \frac{1}{\nu^*\delta_\nu(2\pi)^{1/2}}\exp\left(\frac{(\log\nu^*-\nu^{(l-1)})^2}{2\delta_\nu^2}\right).$$

   A nova observação ν^* é aceita com probabilidade

$$\min\left\{\frac{\pi(\nu^*|\cdots)\nu^*}{\pi(\nu^{(l-1)}|\cdots)\nu^{(l-1)}},1\right\},$$

em que $\pi(\nu^*|\cdots)$ é a densidade (3.3) utilizando os valores atualizados dos parâmetros envolvidos. O hiperparâmetro conhecido δ_ν^2 será regulado tal que a taxa de aceitação dos graus de liberdade seja algo no intervalo (0,15; 0,3). Veja Liu (1994), Garay *et al.* (2015) e Massuia *et al.* (2017) para uma discussão detalhada.

3.4 COMPARAÇÃO DE MODELOS

Considerando a necessidade de comparar modelos dentre todos vistos nesse capítulo, iremos definir critérios que selecionem um melhor ajuste de um modelo Bayesiano, considerando suas diversas formas e particularidades. Esses critérios são úteis para comparar modelos que ajustam um mesmo conjunto de dados.

Um dos métodos mais utilizados para comparar modelos é o da estatística ordenada preditiva condicional (CPO), que se baseia no critério de validação cruzada.

Seja $\mathbf{Z} = (z_1, \cdots, z_n)^\top$ uma amostra aleatória de $\pi(\cdot|\boldsymbol{\theta})$, em que a verossimilhança é denotada por $\pi(\mathbf{z}|\boldsymbol{\theta})$, o CPO_i é definido como,

$$\text{CPO}_i = \int \pi(z_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{z}_{(-i)})d\boldsymbol{\theta} = \left(\int \frac{\pi(\boldsymbol{\theta}|\mathbf{z})}{\pi(z_i|\boldsymbol{\theta})}d\boldsymbol{\theta}\right)^{-1} = \left[\text{E}_{\boldsymbol{\theta}|\mathbf{z}}\left(\frac{1}{\pi(z_i|\boldsymbol{\theta})}\right)\right]^{-1}, \qquad (3.4)$$

em que $\mathbf{Z}_{(-i)}$ é a amostra sem a i-ésima observação. Para o modelo proposto, o CPO_i consiste em uma complexa integral. Uma aproximação de Monte Carlo pode ser obtida utilizando uma amostra MCMC, $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_Q\}$, da distribuição a posteriori,

$$\widehat{\text{CPO}}_i = \left(\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{\pi(z_i|\boldsymbol{\theta}_q)}\right)^{-1},$$

em que $\boldsymbol{\theta}_q$ é a q-ésima geração a posteriori do vetor $\boldsymbol{\theta}$ em uma amostra MCMC, tal que $q = 1, \cdots, Q$.

Uma estatística resumida dos CPO_i's é o chamado logaritmo da verossimilhança pseudo marginal (LPML), definida por $\text{LPML} = \sum_{i=1}^{n}\log\widehat{\text{CPO}_i}$. Quando utilizamos o LPML como critério para comparar modelos, valores maiores de LPML indicam melhores ajustes.

Agora iremos definir uma quantidade chamada de desvio, que normalmente compõe os critérios de seleção,

$$\text{D}(\boldsymbol{\theta}) = -2\log\pi(\mathbf{z}|\boldsymbol{\theta}) = -2\log\left(\prod_{i=1}^{n}\pi(z_i|\boldsymbol{\theta})\right).$$

Os critérios em sua maioria, medem ao mesmo tempo a qualidade do ajuste e a complexidade do modelo, penalizando os mais complexos com base na parcimônia. O critério de informação de desvio (DIC) proposto por Spiegelhalter *et al.* (2002) é extremamente utilizado para comparação de modelos Bayesianos.

Antes de definirmos o DIC definiremos a medida de complexidade ρ_{DIC}, que é uma mensuração do *número efetivo de parâmetros no modelo* dada por

$$\rho_{\mathrm{DIC}} = \overline{\mathrm{D}}(\boldsymbol{\theta}) - \mathrm{D}(\widetilde{\boldsymbol{\theta}}),$$

em que $\overline{\mathrm{D}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{n} \mathrm{E}[\log \pi(z_i|\boldsymbol{\theta})|\mathbf{z}]$, e $\mathrm{D}(\widetilde{\boldsymbol{\theta}})$ é o desvio avaliado em algum estimador $\widetilde{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$, geralmente aproximado pela média ou mediana da amostra MCMC. Então

$$\mathrm{DIC} = 2\rho_{\mathrm{DIC}} + \mathrm{D}(\widetilde{\boldsymbol{\theta}}) = 2\overline{\mathrm{D}}(\boldsymbol{\theta}) - \mathrm{D}(\widetilde{\boldsymbol{\theta}}). \tag{3.5}$$

Podemos aproximar $\overline{\mathrm{D}}(\boldsymbol{\theta})$ pela média a posteriori da amostra dos desvios

$$\widehat{\overline{\mathrm{D}}}(\boldsymbol{\theta}) = -2\frac{1}{Q}\sum_{q=1}^{Q}\left[\log\prod_{i=1}^{n}\pi(z_i|\boldsymbol{\theta}_q)\right],$$

então temos

$$\widehat{\mathrm{DIC}} = 2\widehat{\overline{\mathrm{D}}}(\boldsymbol{\theta}) - \mathrm{D}(\widetilde{\boldsymbol{\theta}}).$$

Outro critério muito utilizado e de extrema importância para o desenvolvimento da nossa proposta de modelo é o Watanabe-Akaike (WAIC), introduzido por Watanabe e Opper (2010). Primeiramente vamos definir a log densidade preditiva pontual *(log pointwise predictive density-lppd)*, dada por

$$\mathrm{lppd} = \log\left(\prod_{i=1}^{n}\pi(z_i|\mathbf{z})\right) = \sum_{i=1}^{n}\log\int\pi(z_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta} = \sum_{i=1}^{n}\log\mathrm{E}\left[\pi(z_i|\boldsymbol{\theta})|\mathbf{z}\right].$$

O WAIC é definido como o lppd mais uma correção para o número efetivo de parâmetros a serem ajustados. Essa correção penaliza modelos com muitos parâmetros. Existem duas abordagens diferentes para calcular essa correção, para tal definamos

$$\rho_{\mathrm{WAIC}_1} = 2\mathrm{lppd} + \overline{\mathrm{D}}(\boldsymbol{\theta}) \quad \text{e} \quad \rho_{\mathrm{WAIC}_2} = \sum_{i=1}^{n} Var(\log\pi(z_i|\boldsymbol{\theta})|\mathbf{z}),$$

então o WAIC finalmente é dado por

$$\mathrm{WAIC}_k = 2\rho_{\mathrm{WAIC}_k} - 2\mathrm{lppd}, \quad k = 1;2.$$

Pode-se aproximar o valor do WAIC utilizando uma amostra MCMC da seguinte forma

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log\left(\frac{1}{Q}\sum_{q=1}^{Q}\pi(z_i|\boldsymbol{\theta}_q)\right), \quad \text{então} \quad \widehat{\rho_{\text{WAIC}_1}} = 2\widehat{\text{lppd}} + \widehat{\overline{\text{D}}}(\boldsymbol{\theta}).$$

Para calcularmos uma aproximação para a variância de $\log\pi(z_i|\boldsymbol{\theta})$ a posteriori, primeiramente vamos definir

$$V_{q=1}^{Q}(x) = \frac{1}{Q-1}\sum_{q=1}^{Q}(x_q - \bar{x})^2, \quad \text{em que} \quad \bar{x} = \frac{1}{Q}\sum_{q=1}^{Q} x_q,$$

então

$$\widehat{\rho_{\text{WAIC}_2}} = \sum_{i=1}^{n} V_{q=1}^{Q}(\log\pi(z_i|\boldsymbol{\theta}_q)).$$

Via MCMC podemos calcular os WAICs como

$$\widehat{\text{WAIC}_k} = 2\widehat{\rho_{\text{WAIC}_k}} - 2\widehat{\text{lppd}}.$$

Em Celeux *et al.* (2006), o critério WAIC_1 foi denominado como DIC_{obs} (DIC observado) e visto como uma versão modificada do critério DIC original (3.5), uma vez que a definição original de DIC não se aplica a modelos de misturas. Isso porque uma condição essencial para o uso adequado do DIC é que a média a posteriori seja uma boa estimativa. Necessitamos dessa premissa para que a quantidade

$$\text{D}(\widetilde{\boldsymbol{\theta}}) = -2\log\left(\prod_{i=1}^{n}\pi(z_i|\widetilde{\boldsymbol{\theta}})\right)$$

seja adequada para o uso no DIC, o que não necessariamente acontece no caso de misturas devido a um fenômeno chamado *label switching*. Para mais detalhes veja Frühwirth-Schnatter (2006) e Stephens (2000). O DIC_{obs} ou WAIC_1 servem como uma alternativa para contornarmos esse problema, sendo assim a melhor opção para a nosso modelo.

Como alternativa aos critérios de seleção, vamos apresentar um outro método para avaliar a adequação de um modelo. Para isso utilizamos uma medida baseada na distribuição preditiva a posteriori. A forma da distribuição preditiva a posteriori que nos utilizaremos é dada por $\pi(\mathbf{z}_{pr}|\mathbf{z}) = E[\pi(\mathbf{y}_{pr}|\boldsymbol{\theta})|\mathbf{z}]$, em que $\mathbf{z}_{pr}$ são os dados replicados que iriamos obter caso o experimento que produziu $\mathbf{Z}$ fosse repetido com o mesmo modelo e o mesmo valor de $\boldsymbol{\theta}$.

A ideia é que, se o modelo se ajustar bem, os dados replicados gerados por meio do modelo, devem ser semelhantes aos dados observados. Em outras palavras, os dados observados devem parecer plausíveis sob a distribuição preditiva a posteriori.

Se as observações forem muito diferentes em relação à distribuição preditiva a posteriori, existem alguns problemas no ajuste do modelo aos dados. A medida da diferença entre o modelo e os dados é então calculada por meio de uma estatística de resumo. Seguiremos a ideia de Gelman *et al.* (2013), que utiliza a função desvio, denotado por $T(\mathbf{z},\boldsymbol{\theta}) = \mathrm{D}(\boldsymbol{\theta})$, uma função da amostra e do vetor de parâmetros.

O p-valor Bayesiano preditivo a posteriori, proposto por Rubin (1984) é definido como

$$P(T(\mathbf{z}_{pr},\boldsymbol{\theta}) \geq T(\mathbf{z},\boldsymbol{\theta})|\mathbf{z}),$$

De acordo com Gelman *et al.* (2013), a diferença média entre $T(\mathbf{z}_{pr},\boldsymbol{\theta})$ e $T(\mathbf{z},\boldsymbol{\theta})$ em L sorteios simulados, será de importância prática se seu p-valor estiver próximo de 0 ou 1. Nesse caso o modelo será considerado inapropriado para ajustar os dados. Os p-valores extremos significam que não se pode esperar que o modelo capture as características dos dados. Um p-valor muito pequeno ou muito grande (por exemplo, $< 0,05$ ou $> 0,95$) indica uma especificação incorreta do modelo, ou seja, o padrão observado pode não ser visto quando os dados forem replicados.

3.5 ESTUDOS DE SIMULAÇÃO

3.5.1 Estudo 1 - Comportamento da Estimação

Nesta seção, utilizamos a amostra MCMC gerada pelo nosso algoritmo, dado na Seção 3.3.2, para avaliar o desempenho do modelo FMR-SMSN-CR quanto a estimação dos parâmetros. Este estudo de simulação foi desenvolvido com o intuito de analisar o comportamento das estimativas quando utilizamos diferentes tamanhos de amostra (n = 300 e 500) e níveis de censura à esquerda (5%, 10%, 20% e 30%). Fixado um tamanho amostral e um nível de censura geramos os dados artificialmente a partir de cada modelo FMR-SMSN-CR com a seguinte configuração: $G = 2$ e $\mathbf{x}_i^\top = (1, x_{i1}, x_{i2})$, tal que $x_{i1} \sim \mathrm{Unif}(1,5)$ e $x_{i2} \sim \mathrm{TN}(0,1;(2,6))$, os valores para os parâmetros utilizados foram $\boldsymbol{\beta}_1^\top = (\beta_{01};\beta_{11};\beta_{21}) = (25;0,8;0,5)$, $\boldsymbol{\beta}_2^\top = (\beta_{02};\beta_{12};\beta_{22}) = (1,5;-1;-1)$, $\boldsymbol{\sigma}^2 = (1;1)^\top$, $p_1 = 0,7$ e para os modelos assimétricos, $\boldsymbol{\lambda}^\top = (2;-2)$, $\nu = 5$ e ainda $\rho = 0,7$ e $\eta = 0,3$ para os modelos FMR-CN-CR e FMR-SCN-RC. Logo estimamos os parâmetros utilizando 66 mil amostras MCMC, retirando as primeiras 10 mil amostras e armazenando amostras utilizando um espaçamento de 20 unidades. Esse procedimento foi repetido 100 vezes, para cada combinação de tamanho amostral e nível de censura, sob os diversos

modelos FMR-SMSN-CR (FMR-T-CR, FMR-SL-CR, FMR-N-CR, FMR-NC-CR, FMR-ST-CR, FMR-SSL-CR, FMR-SNC-CR e FMR-SN-CR). Os valores iniciais foram escolhidos de forma arbitrária e são os seguintes: para $\boldsymbol{\beta}_1^\top$ e $\boldsymbol{\beta}_2^\top$ utilizamos um vetor com todos os valores iguais entre si e iguais a 10, $\boldsymbol{\sigma}^2$ e λ iguais a 5, ν igual a 10 e os pesos iguais a 0,5.

Para o estudo de simulação, o nível de censura geralmente é obtido censurando uma amostra por meio de seus percentis, utilizando os mesmos como ponto de corte. Por exemplo, para um nível de censura à esquerda de 5% encontramos o quinto percentil amostral e fazemos todas as observações abaixo desse percentil assumirem o valor do mesmo, no entanto se o problema fosse de censura à direita e com uma taxa de 5% , o percentil que deveria ser encontrado seria o nonagésimo quinto, em que todas as observações acima deste tomariam o valor do mesmo. É claro que se utilizarmos exatamente essa ideia para todas as 100 amostras, teremos diferentes níveis de censura para uma taxa de censura fixada, contrapondo um princípio natural de controle do experimento aleatório. Para resolver este problema encontramos, por meio de testes, um nível de censura que resultou uma taxa de censura média desejada. Os testes se basearam na geração das 100 amostras e a imposição de vários níveis de censura fixos para todas as amostras, em que para cada nível de censura observamos a taxa de censura média das amostras e escolhemos aquele nível que nos retornou a taxa de censura média desejada. Para os modelos FMR-ST-RC e FMR-SSL-RC esses níveis de censura são: -17,3 para 5%, -14,5 para 10%, -11,2 para 20% e -5 para 50%, que geraram uma taxa média de censura de 5,109%, 10,130%, 20,196% e 50,210% no modelo FMR-ST-RC e 4.842%, 9.855%, 20,051% e 50,238% para o modelo FMR-SSL-RC. Um estudo semelhante pode ser visto em Zeller *et al.* (2019), onde temos a inclusão do tamanho amostral $n = 50$ e uma quantidade de 500 repetições, a quantidade de taxas é a mesma porém os ajustes se baseiam em um algoritmo EM.

As Figuras 1, 2 e 3 mostram gráficos de boxplots das estimativas de alguns parâmetros para o modelo FMR-ST-CR, FMR-SSL-CR e FMR-SCN-CR (os parâmetros ρ e η dos modelos FMR-CN-CR e FMR-SCN-CR, neste estudo, serão considerados conhecidos). Os gráficos para os demais modelos podem ser encontrados no Apêndice A (Figuras de 28 a 32).

Em geral para um nível de censura fixo, observamos nesse estudo que, quando o tamanho da amostra passa de 100 para 500 a qualidade das estimativas melhora - entende-se, aqui, qualidade das estimativas como a proximidade delas para o verdadeiro valor do parâmetro -, enquanto sua variabilidade diminui. Além disso para um tamanho de amostra fixo, é possível notar que o crescimento do nível de censura gera um crescimento na variabilidade e uma piora da qualidade das estimativas dos parâmetros, principalmente com 50% de censura. Isso ocorre

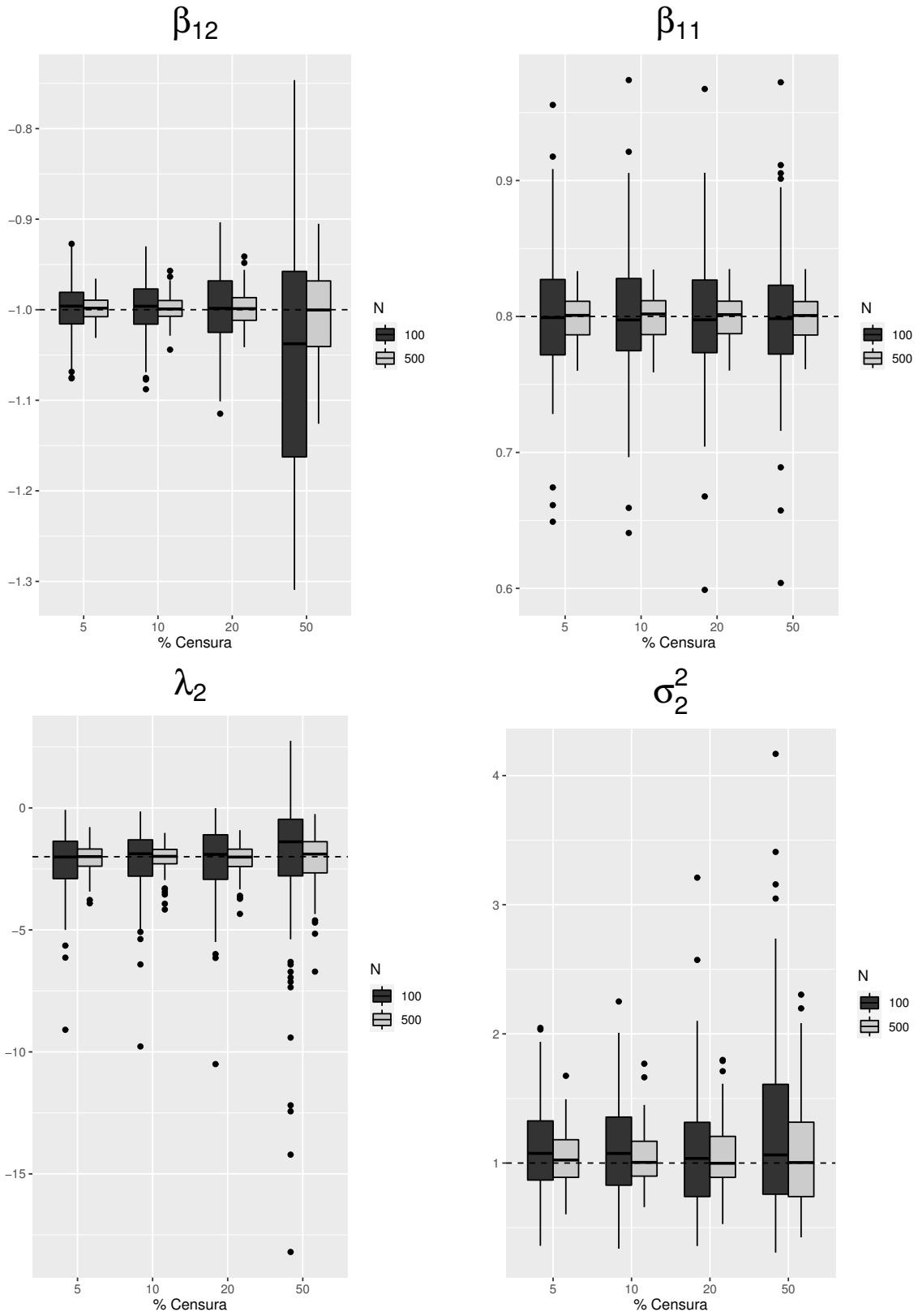

Figura 1 – Boxplots das estimativas de β_{11}, β_{12}, λ_2 e σ_2^2 (a linha horizontal indica o verdadeiro valor do parâmetro) para o modelo FMR-ST-CR com diferentes taxas de censura (reta das abscissas) e tamanhos amostrais (N).

exclusivamente na estimação dos parâmetros pertencentes a segunda população, isso se dá pelo fato dessa população conter toda a quantidade de dados censurados da amostra, pois estamos utilizando uma censura à esquerda. Os gráficos de traço da primeira amostra da posteriori, gerada no estudo de simulação, para todos os modelos FMR-SMSN-CR, utilizando $n = 100$, podem ser encontrados no Apêndice A.

Em um computador com processador Intel Core i5-3210M CPU 2.50GHz, 8 GB de memória RAM e sistema operacional de 64 bits, utilizando o software JAGS, o tempo computacional de cada réplica desse experimento (em um total de 100) foi, em média, de 1 hora para os modelos assimétricos e 42 minutos para os modelos simétricos, lembrando que cada replica conta com dois diferentes tamanhos amostrais e quatro diferentes taxas de censura. Particularmente para $n = 500$, com uma taxa de censura fixa em 20%, o tempo computacional foi de 13 minutos para os modelos assimétricos e 9 minutos para os modelos simétricos.

3.5.2 Estudo 2 - Flexibilidade do Modelo

Nesta seção, mostraremos a flexibilidade do modelo FMR-SMSN-CR, para acomodar dados oriundos de modelos de natureza diferente da família SMSN. Criaremos um modelo de mistura de regressões no qual os erros aleatórios seguem uma distribuição, cujas propriedades sejam completamente diferentes da família SMSN. A distribuição que utilizaremos para esse propósito é chamada de distribuição normal inversa gaussiana (NIG). Utilizando a Definição 2, podemos mostrar que a distribuição NIG é uma mistura de escala da distribuição normal e uma inversa gaussiana. Um estudo semelhante pode ser visto em Cabral, Souza e Leão (2022), onde o mesmo utilizou um modelo com erro nas variáveis.

Então vamos supor um modelo FMR-NIG-CR, em que as componentes da mistura seguem uma distribuição normal inversa gaussiana.

Definição 6. *Dizemos que uma variável aleatória U segue uma distribuição inversa gaussiana se sua função de densidade é dada por*

$$g(u) = \frac{\delta}{\sqrt{2\pi}} u^{-3/2} \exp\left\{ -\frac{1}{2}\left(\frac{\delta^2}{u} + \gamma^2 u - 2\delta\gamma \right) \right\}, \ u > 0,$$

em que $\gamma > 0$ e $\delta > 0$.

Neste caso, utilizamos a notação $U \sim \text{IG}(\gamma, \delta)$. Mais detalhes podem ser vistos em Barndorff-Nielsen (1997).

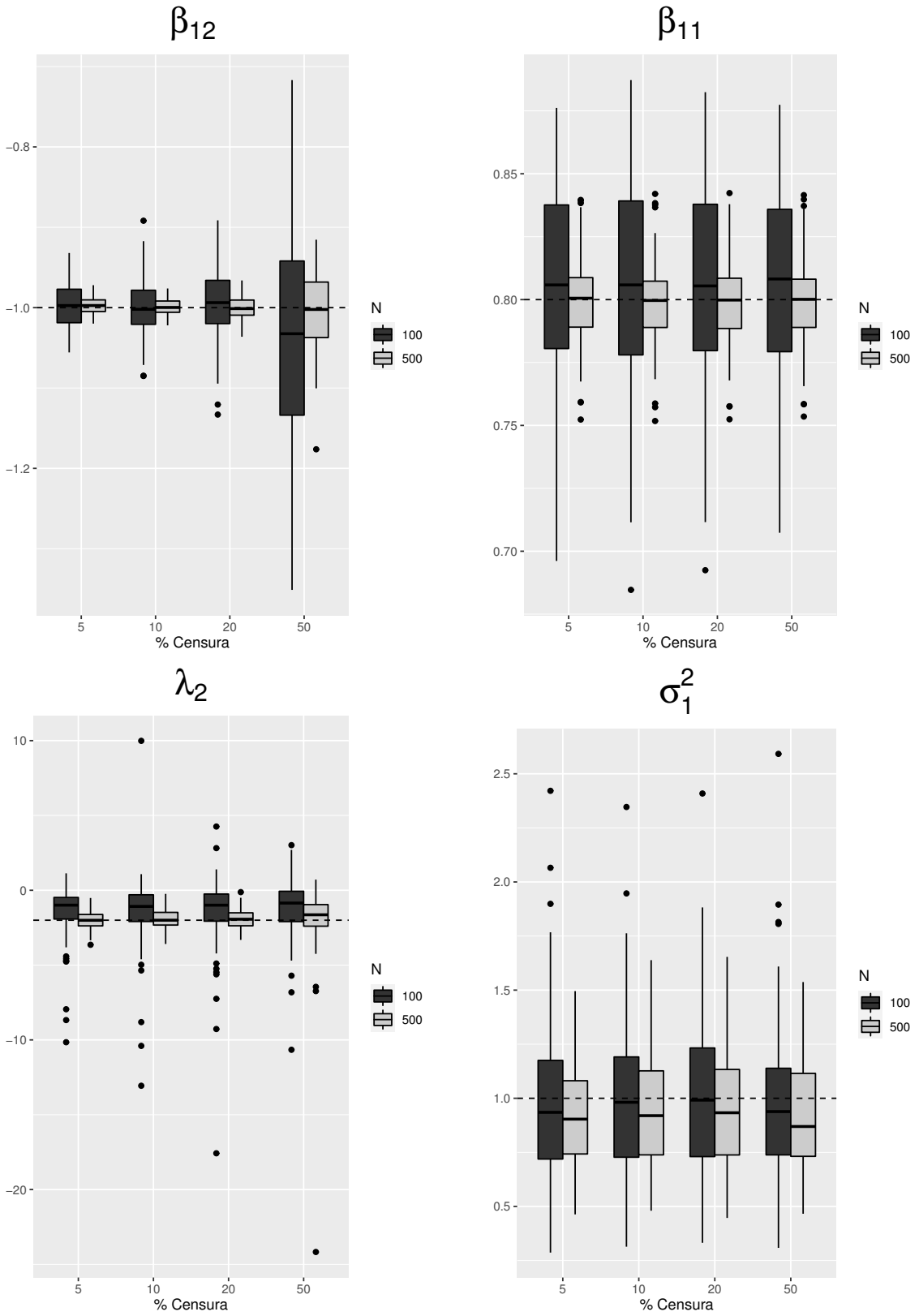

Figura 2 – Boxplots das estimativas de β_{12}, β_{11}, σ_1^2 e λ_2. (a linha horizontal indica o verdadeiro valor do parâmetro) para o modelo FMR-SSL-CR com diferentes taxas de censura (reta das abscissas) e tamanhos amostrais (N).

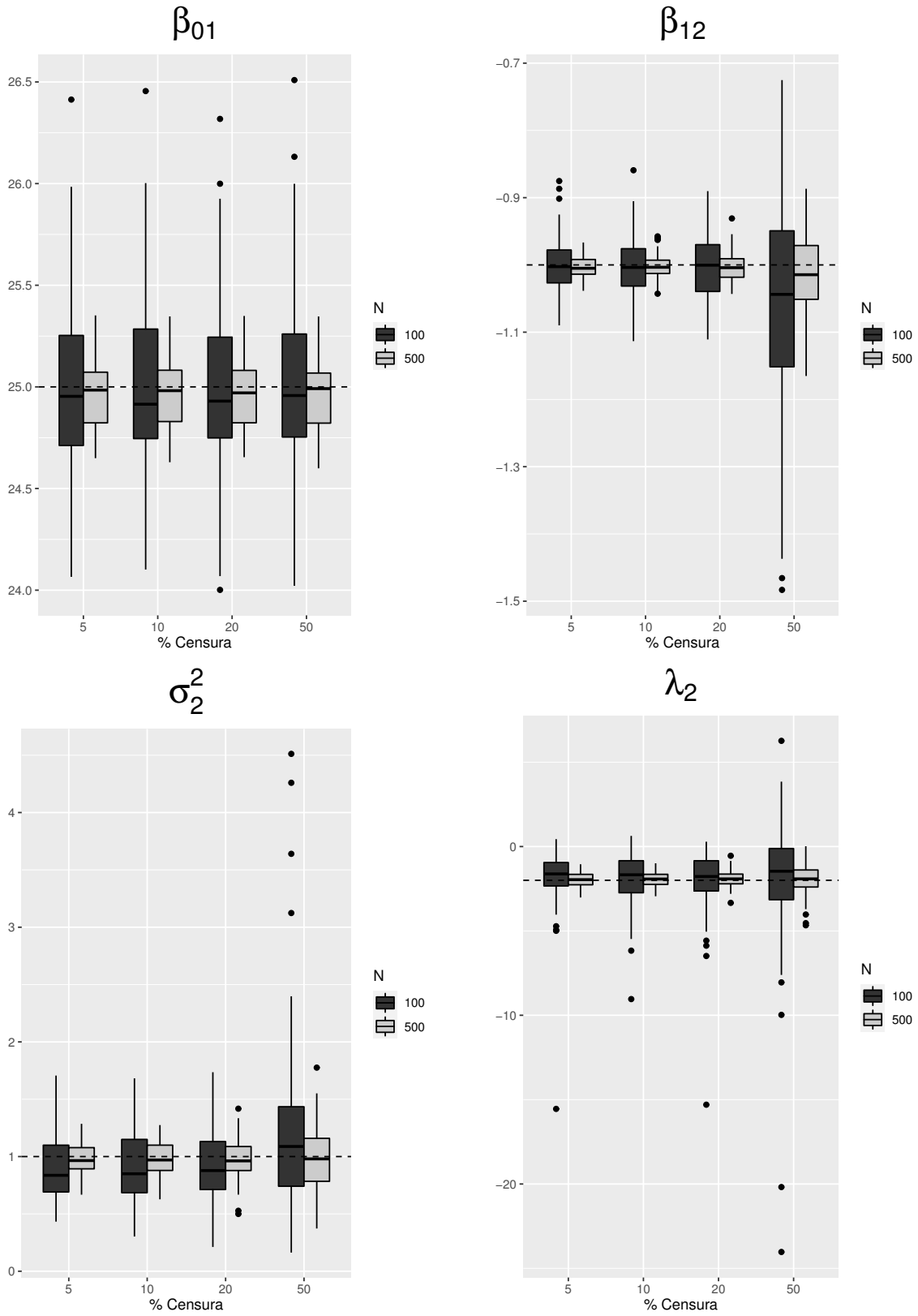

Figura 3 – Boxplots das estimativas de β_{01}, β_{12}, σ_2^2 e λ_2. (a linha horizontal indica o verdadeiro valor do parâmetro) para o modelo FMR-SCN-CR com diferentes taxas de censura (reta das abscissas) e tamanhos amostrais (N).

Definição 7. *Dizemos que uma variável aleatória X segue uma distribuição NIG se admite a seguinte representação estocástica*

$$X|U=u \sim \mathrm{N}(\mu+u\Delta\lambda, u\Delta),\ U \sim \mathrm{IG}(\gamma,\delta),$$

em que μ *e* λ *são parâmetros pertencentes aos reais e* $\Delta > 0$.

A notação para uma variável com distribuição NIG será $X \sim \mathrm{NIG}(\mu,\Delta,\lambda,\gamma,\delta)$ e a função de densidade $\mathrm{NIG}(\cdot|\mu,\Delta,\lambda,\gamma,\delta)$. Para assegurar a identificabilidade utilizaremos a restrição $\Delta = 1$, sob essas condições a média e variância são dadas por,

$$\mathrm{E}(X) = \mu + (\delta/\gamma)\lambda \quad \text{e} \quad \mathrm{Var}(X) = (\delta/\gamma^3)(\gamma^2+\lambda^2). \tag{3.6}$$

A distribuição NIG apresenta caudas pesadas e assimetria. Por esse motivo utilizaremos a mesma como um teste para verificar se os modelos FMR-SMSN-CR, que teoricamente ajustam com qualidade essa características, conseguem absorver com propriedade a assimetria e caudas pesadas oriundas de um modelo com uma sistemática totalmente diferente da família SMSN. A distribuição normal é um caso especial da distribuição NIG quando $\lambda = 0$, $\gamma \to \infty$ e $\delta \to \infty$.

Para a definição do modelo FMR-NIG-CR iremos assumir que a distribuição de Y_i é uma mistura finita de densidades, em que a j-ésima densidade condicional de $Y|Z=j$ dada na Definição 4 é da forma

$$g_j(y_i) = \mathrm{NIG}(y_i|x_i^\top \boldsymbol{\beta}_j + b_j^*, 1, \lambda_j, \gamma_j, \delta_j), \tag{3.7}$$

em que $b_j^* = -(\delta_j/\gamma_j)\lambda_j$. Considerando que a variável resposta é como em (3.1), temos a seguinte representação hierárquica para a distribuição de Y_i, denotada como FMR-NIG-CR, dada por

$$\begin{aligned} Y_i|U_i=u_i, Z_{ij}=1 &\sim \mathrm{N}(x_i^\top \boldsymbol{\beta}_j + b_j^* + u_i\lambda_j, u_i), \\ U_i|Z_{ij}=1 &\sim \mathrm{IG}(\gamma_j,\delta_j), \\ \mathbf{Z}_i &\sim \mathrm{Multinomial}(1,p_1,\cdots,p_G),\ i=1,\cdots,n \ \text{ e } \ j=1,\cdots,G. \end{aligned} \tag{3.8}$$

Fizemos um estudo computacional em que geramos uma amostra de tamanho $n = 100$ e outra com $n = 500$ de um modelo FMR-NIG-CR com $G = 3$, $p_1 = p_2 = 0,3$, $\boldsymbol{\beta}_1^\top =$

$(\beta_{01};\beta_{11}) = (-5;1)$, $\boldsymbol{\beta}_2^\top = (\beta_{02};\beta_{12}) = (0;2)$, $\boldsymbol{\beta}_3^\top = (\beta_{03};\beta_{13}) = (10;4)$, $\boldsymbol{\lambda}^\top = (-2;1;-2)$, $\boldsymbol{\delta}^\top = (\delta_1;\delta_2;\delta_3) = (0,5;0,5;0,5)$, $\boldsymbol{\gamma}^\top = (\gamma_1;\gamma_2;\gamma_3) = (1;1;1)$ e $x_{i1} \sim \text{Unif}(1,6)$ para $i = 1,\cdots,n$, com uma censura à esquerda, cujo ponto de censura corresponde ao décimo percentil da amostra, ou seja, uma taxa de censura de 10%. O procedimento de ajuste dos modelos foi feito pelo algoritmo dado em 3.3.2 e as especificações quanto a geração das amostras MCMC foram as mesmas utilizadas na seção 3.5.1. A Figura 4 mostra o histograma da variável resposta, na qual claramente podemos observar a censura à esquerda.

A esse conjunto de dados ajustamos todos os modelos da família SMSN estudados nesse capítulo, com três componentes. Após o ajuste dos modelos calculamos os critérios de seleção WAIC_1 e WAIC_2.

O método mais comum para monitorar a convergência de uma amostra MCMC em relação à sua distribuição estacionária é o gráfico de traço, que exibe os valores simulados da amostra a posteriori nas diferentes iterações do algoritmo MCMC. A análise desse gráfico se baseia na observação do mesmo em janelas de tempo distintas, no qual podemos identificar possíveis mudanças na cadeia ao longo do suporte da distribuição a posteriori. Podemos observar à direita da Figura 4 os gráficos de traço de algumas amostras a posteriori dos parâmetros, onde vemos que todos que foram apresentados percorrem de maneira homogênea os seus suportes.

Tabela 1 – WAIC_1 e WAIC_2 para os modelos FMR-SMSN-CR cujo conjunto de dados foi gerado pelo modelo FMR-NIG-CR.

	Tamanho amostral			
	n=100		n=500	
Modelo	WAIC_1	WAIC_2	WAIC_1	WAIC_2
FMR-N-CR	550,2036	549,3956	2723,901	2730,207
FMR-T-CR	484,7163	485,4316	2325,813	2325,911
FMR-SL-CR	491,4093	492,1413	2363,503	2363,681
FMR-SN-CR	478,3624	482,3757	2360,061	2366,735
FMR-ST-CR	456,2796	458,5176	**2221,949**	**2222,407**
FMR-SSL-CR	**449,9010**	**453,7181**	2289,661	2286,687

Na Tabela 1 é notório que modelos delineados para ajustar assimetria e caudas pesadas ao mesmo tempo, como FMR-ST-CR e FMR-SSL-CR, são considerados melhores ajustes, por meio dos critérios de seleção, em comparação com os demais modelos que não comportam essas características.

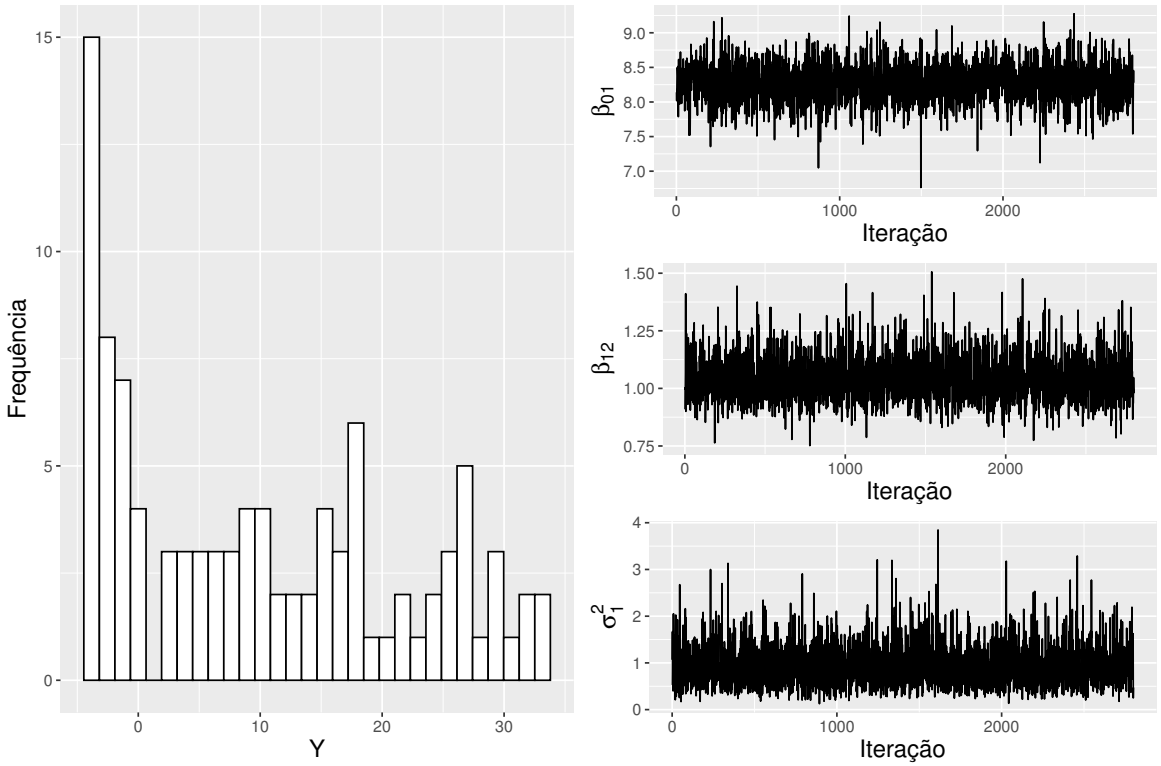

Figura 4 – Histograma da variável resposta com $n = 100$, gráficos de traço e gráficos de autocorrelação de β_{01}, β_{12} e σ_1^2 para dados gerados por meio do modelo FMR-NIG-CR e ajustados pelo modelo FMR-ST-CR.

3.5.3 Estudo 3 - Clonagem de Dados e Identificabilidade

Quando estudamos misturas de modelos de regressão geralmente enfrentamos um dilema quando o assunto é provar a identificabilidade do modelo. Isso ocorre pela dificuldade em provar esse comportamento analiticamente e pela escassez de trabalhos abordando tal tema. Quando falamos de modelos baseados em misturas de escala da distribuição normal assimétrica a dificuldade é ainda maior, não há muitos trabalhos relacionados a este tema. Veja Zeller *et al.* (2019) para uma breve discussão.

A falta de identificabilidade pode levar a problemas na interpretação dos resultados e na estimação dos parâmetros pois podem haver múltiplas combinações de valores de parâmetros que produzem a mesma distribuição de probabilidade, o que dificulta a interpretação das estimativas dos parâmetros e torna difícil inferir sobre o significado dos resultados obtidos. Outro problema é referente à comparação de modelos. Pode ser difícil compará-los de forma justa, uma vez que ambos podem produzir resultados semelhantes, mas com valores de parâmetros diferentes, tornando a modelagem probabilística menos eficaz e mais suscetível a erros. Essas incertezas quanto a estimação e comparação dos modelos também se estendem ao uso das propriedades assintóticas dos estimadores de máxima verossimilhança.

O algoritmo de Clonagem de Dados é uma forma alternativa de obtermos evidências de que não há problemas de identificabilidade com os parâmetros do modelo. Esse algoritmo foi proposto por Lele, Nadeem e Schmuland (2010). A contribuição dessa técnica para a metodologia proposta é de extrema importância pois garantirá a confiança nas estimativas dos parâmetros do modelo.

Sejam $\mathbf{y} = (y_1, \cdots, y_n)^\top$ os dados reais, $\pi(\mathbf{y}|\boldsymbol{\theta})$ a verossimilhança e $\boldsymbol{\theta}$ o vetor $r-$dimensional de parâmetros do modelo. O método de Clonagem de Dados se baseia na ideia de replicar os dados reais, dando origem ao que chamamos de clones.

Segundo Lele, Nadeem e Schmuland (2010), para entendermos melhor o método, imagine que um indivíduo realiza um experimento estatístico exatamente igual ao que gerou $\mathbf{y}$, porém não apenas uma vez, mas sim k vezes, de forma simultânea e independentemente. Suponha também que além disso, cada um dos k experimentos produz réplicas, por acaso, exatamente iguais aos dados reais $\mathbf{y}$. Definimos por $\mathbf{y}^{(K)} = (\mathbf{y}, \cdots, \mathbf{y})^\top$ o conjunto formado por todas as réplicas idênticas de $\mathbf{y}$, o índice K representa a quantidade de vezes que os dados foram replicados, note que $\mathbf{y}^{(K)}$ tem dimensão $(K \times N)$.

A função de verossimilhança baseada na combinação dos dados desses K experimentos independentes é dada por $[\pi(\mathbf{y}|\boldsymbol{\theta})]^K$ e o máximo de $\pi(\mathbf{y}^{(K)}|\boldsymbol{\theta})$ é exatamente igual ao máximo da função $\pi(\mathbf{y}|\boldsymbol{\theta})$. Com esse algoritmo conseguimos aproximar as estimativas de máxima verossimilhança e o inverso da matriz de informação de Fisher de $\boldsymbol{\theta}$ utilizando amostras MCMC de uma distribuição a posteriori modificada.

Sob condições adequadas (ver apêndice de Lele, Nadeem e Schmuland (2010)) é possível mostrar que para K grande, a distribuição a posteriori de $\boldsymbol{\theta}|\mathbf{y}^{(K)}$ é aproximada por uma distribuição $\mathrm{N}_r(\widehat{\boldsymbol{\theta}}, (1/K)I^{-1}(\widehat{\boldsymbol{\theta}}))$ em que $\widehat{\boldsymbol{\theta}}$ é a estimativa de máxima verossimilhança do vetor $\boldsymbol{\theta}$ e $I(\widehat{\boldsymbol{\theta}})$ é a matriz de informação de Fisher. Logo uma aproximação direta de $\widehat{\boldsymbol{\theta}}$ e $I(\widehat{\boldsymbol{\theta}})$ pode ser obtida utilizando a média das amostras MCMC extraídas da distribuição a posteriori de $\boldsymbol{\theta}|\mathbf{y}^{(K)}$ e K vezes a matriz de covariância dessas amostras a posteriori respectivamente.

Lele, Nadeem e Schmuland (2010) mostraram que, se a matriz de covariância da distribuição a posteriori de $g(\boldsymbol{\theta})|\mathbf{y}^{(K)}$ tem seu maior autovalor λ_K convergindo para zero, quando K aumenta, então $g(\boldsymbol{\theta})$ é estimável. Esta convergência para zero tem a mesma taxa de $1/K$.

Seja $\lambda_K^S = \lambda_K/\lambda_1$ o maior autovalor padronizado. A convergência de λ_K pode ser observada pela análise de um gráfico de λ_K^S em função de K comparando-o com o valor de $1/K$.

A Figura 5 mostra esses gráficos para os modelos assimétricos quando $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$, em que $\mathbf{y}$ é uma amostra artificial gerada do modelo FMR-SMSN-CR, com tamanho $n = 100$,

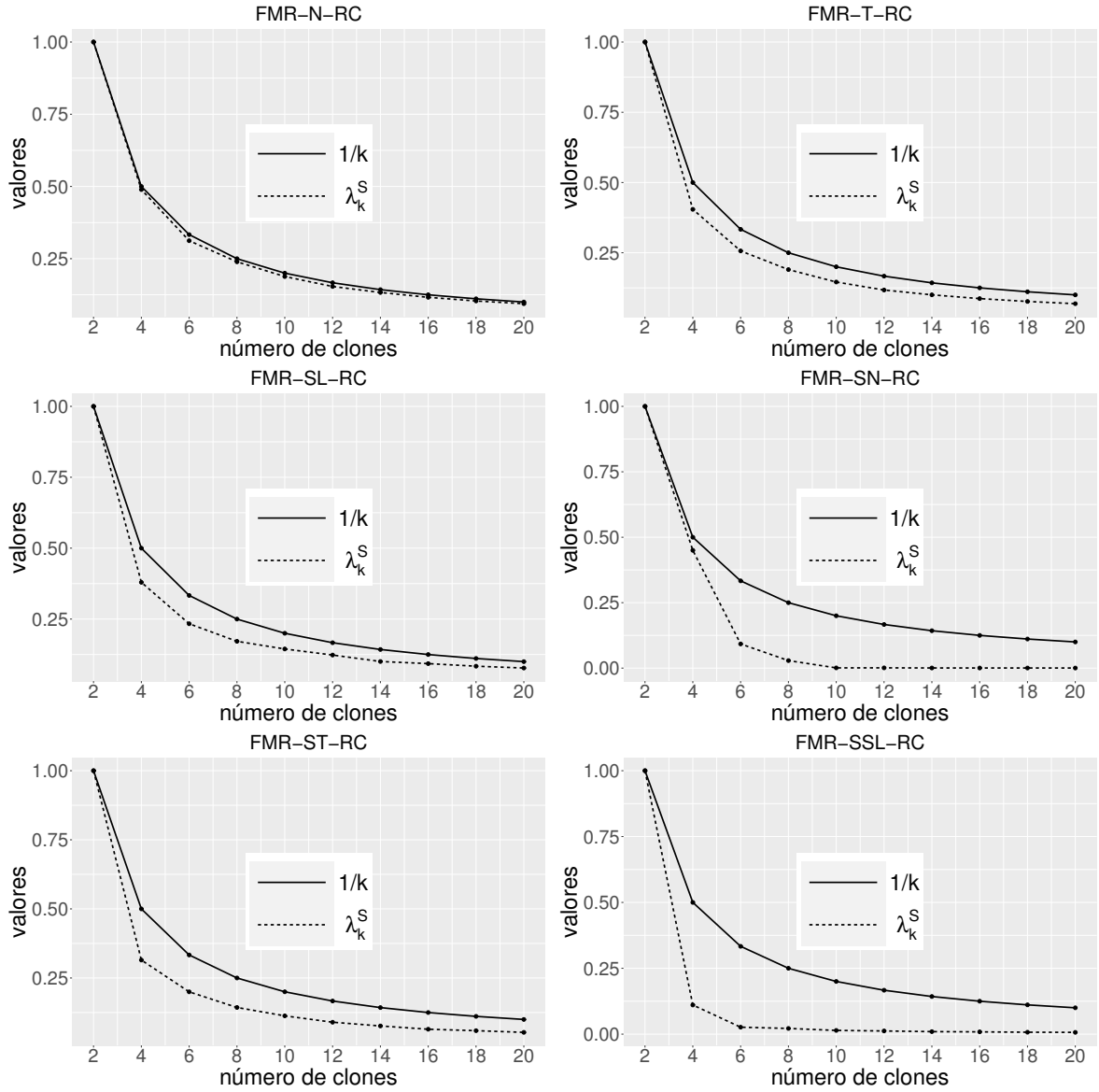

Figura 5 – Estudo de simulação 3. Checando a identificabilidade utilizando clonagem de dados.

$G = 2$, $p_1 = 0{,}5$, $\boldsymbol{\beta}_1^\top = (\beta_{01};\beta_{11};\beta_{21}) = (25;0{,}8;0{,}5)$, $\boldsymbol{\beta}_2^\top = (\beta_{02};\beta_{12};\beta_{22}) = (1{,}5;-1;-1)$, $\boldsymbol{\lambda}^\top = (2;-2)$, $\boldsymbol{\sigma}^2 = (2;2)^\top$ e $\nu = 2$, para $i = 1,\cdots,n$. As especificações quanto a geração das amostras MCMC, foram as mesmas utilizadas na seção 3.5.1.O procedimento de ajuste dos modelos foi feito utilizando o algoritmo dado na seção 3.3.2. O estudo de Clonagem de Dados foi realizado utilizando um pacote do R chamado **dclone** (Sólymos,2010) o qual utiliza nosso algoritmo para gerar varias cadeias sob diferentes quantidades de tamanhos de clones, calculando λ_K^S e construindo os gráficos. Em todos os gráficos podemos notar uma convergência de λ_K^S para zero, quando K aumenta. Essa convergência tem taxas mais próximas de $1/K$ para o modelo FMR-ST-CR e para os modelos simétricos. Os modelos FMR-SN-CR e FMR-SSL-CR tiveram uma convergência mais rápida assim como o FMR-SSL-CR, que já atingiu um valor muito

próximo de zero com apenas 6 clones. Assim, todos os gráficos sugerem fortes evidências de identificabilidade em todos os modelos.

3.5.4 Estudo 4 - Influência dos Dados com Observações Atípicas

Quando falamos de modelar dados com outliers, é esperado que modelos com caudas pesadas sejam menos suscetíveis a mudanças nas estimativas quando esses outliers estão presentes em um conjunto de dados. Em nosso conjunto de modelos temos o FMR-T-CR, FMR-SL-CR, FMR-ST-CR e FMR-SSL-CR como exemplo desse tipo de modelo. O objetivo deste estudo é analisar a influência de algumas observações discrepantes nas estimativas dos parâmetros. Podemos ver em Cabral, Souza e Leão (2022) um estudo semelhante porém sem um contexto de mistura de modelos de regressão.

Para esse estudo verificaremos o comportamento dos modelos FMR-ST-CR e FMR-SSL-CR, que teoricamente acomodam melhor pontos localizados nas caudas da distribuição, em comparação ao modelo FMR-SN-CR que não tem essa característica. Para a construção desse estudo, escolheremos de forma aleatória 4 pontos dentro de um conjunto de dados simulado. Esses pontos serão perturbados gradativamente da seguinte forma: aumentando os valores das covariáveis e da resposta em $\Lambda\%$ de seu valor original, em que $\Lambda = 10, 20, 30, ..., 150$, Λ é definida como taxa de perturbação. Isso significa que a variável original x será perturbada e assumirá o valor x^* da seguinte forma

$$x^* = \left(1 + \frac{\Lambda}{100}\right) x.$$

Os dados foram gerados pelo modelo FMR-SN-CR com censura à esquerda e taxa de 10%, $G = 2$, $p_1 = 0,7$, $\boldsymbol{\beta}_1^\top = (\beta_{01}; \beta_{11}) = (2; 6)$, $\boldsymbol{\beta}_2 = (\beta_{02}; \beta_{12})^\top = (0,5; 1)$, $\boldsymbol{\lambda}^\top = (2; -2)$, $\boldsymbol{\sigma}^2 = (0,1; 0,5)^\top$ e $x_{i1} \sim \text{Unif}(1, 6)$ para $i = 1, \cdots, n$. Chamaremos de I o conjunto de pontos, não censurados, escolhidos para serem perturbados. Estes pontos foram: #79 ($y_{79} = 14,100$), #29 ($y_{29} = 22,471$), #11 ($y_{11} = 23,314$) e #57 ($y_{57} = 34,065$). As especificações quanto a geração das amostras MCMC foram as mesmas utilizadas na seção 3.5.1

Para avaliar as mudanças que as estimativas dos parâmetros apresentaram com as perturbações feitas nos dados, utilizamos a mudança relativa definida por

$$RC = \left| \frac{\widehat{\alpha} - \widehat{\alpha}_I}{\widehat{\alpha}} \right| \times 100,$$

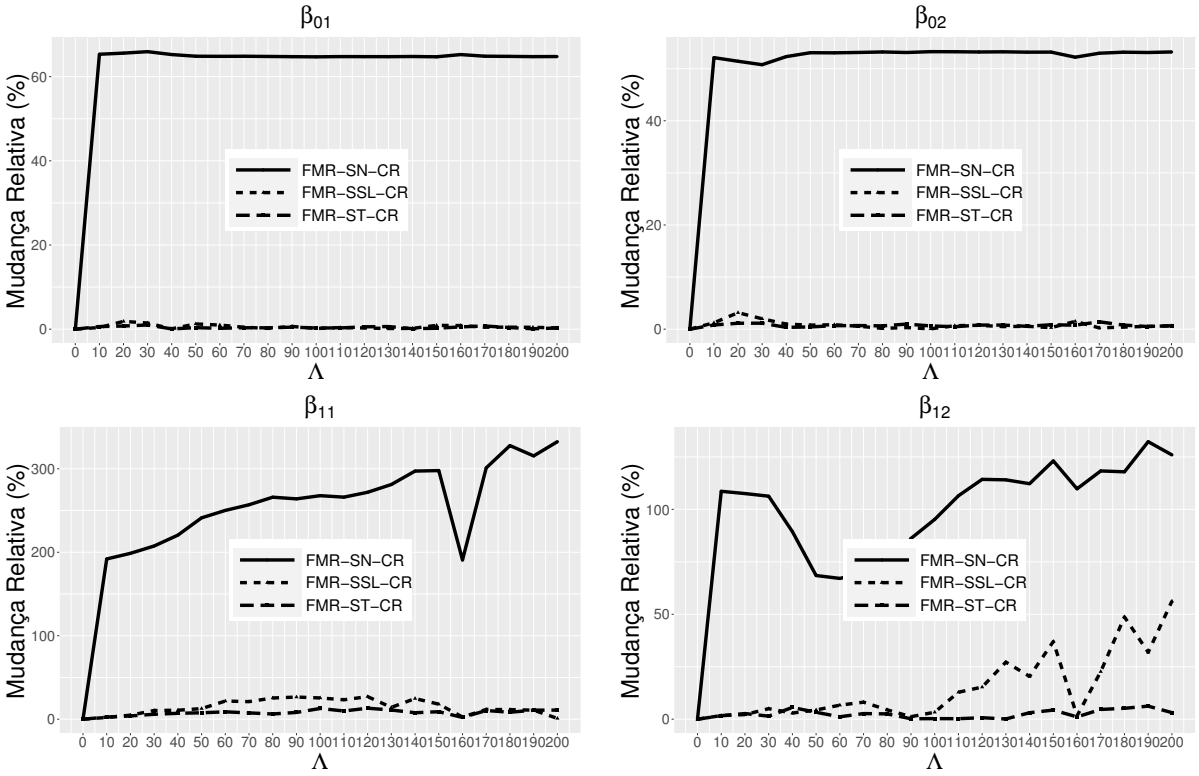

Figura 6 – RC (em %) para $\beta_{01}, \beta_{11}, \beta_{02}$ e β_{12} nos modelos FMR-SN-CR, FMR-ST-CR e FMR-SSL-CR com diferentes níveis de perturbação Λ.

em que $\widehat{\alpha}$ e $\widehat{\alpha}_I$ denotam as medianas a posteriori de cada parâmetro antes e depois que um conjunto I de observações foi perturbado, respectivamente.

Na Figura 6 temos os efeitos da perturbação dos dados nos coeficientes do modelo de regressão. Em cada gráfico temos as mudanças relativas para os três modelos FMR-SN-CR (linha continua), FMR-ST-CR e FMR-SSL-CR (linhas seccionadas) e o comportamento dessa mudança relativa com o avanço da taxa de perturbação (Λ). No gráfico referente ao coeficiente β_{01} podemos observar que, com uma taxa de perturbação de 10% (eixo das abcissas), o modelo FMR-SN-CR já atingiu uma mudança relativa maior que 60% (eixo das ordenadas) e se manteve assim mesmo com o aumento das taxas de perturbação. Isso ocorreu de forma similar para o parâmetro β_{02}, enquanto que, para os parâmetros β_{11} e β_{12}, as mudanças relativas do modelo FMR-SN-CR tiveram um crescimento gradativo com o aumento das taxas de perturbação, atingindo até 300% de mudança relativa em β_{11}, quando $\Lambda = 200\%$. Os modelos FMR-ST-CR e FMR-SSL-CR se mantiveram sempre com pouca mudança relativa, mesmo com o aumento significativo das taxas de pertubações, tal que apenas o modelo FMR-SSL-CR, no parâmetro β_{12}, se mostrou um pouco suscetível para taxas de perturbações acima de 130%. Logo, diante desse experimento, concluímos que os modelos assimétricos FMR-ST-CR e FMR-SSL-CR são robustos a outliers, robustes esta que está ligada ao fato de não sofrerem mudanças severas em

suas estimativas quando alguns pontos discrepantes são introduzidos ao banco de dados, o que naturalmente não acontece com o modelo FMR-SN-RC. Os demais parâmetros dos modelos não foram mostrados aqui por uma questão de espaço porém obtivemos comportamentos parecidos para todos.

3.6 DADOS REAIS

Para observamos e analisarmos os modelos propostos, iremos estudar sua aplicação em um conjunto de dados reais chamado *wage rate* que está descrito em Mroz *et al.* (1987). Os dados correspondem a um estudo em que foi avaliado o salário de 753 mulheres brancas casadas, com idades entre 30 e 60 anos em 1975. Das 753 mulheres consideradas neste estudo, 428 trabalharam em algum momento durante o ano de 1975, enquanto as demais mulheres têm medidas nulas para as variáveis referentes aos seus respectivos empregos, incluindo o salário, pois não trabalharam durante esse ano. Esse banco de dados pode ser encontrado no R com os nomes `PSID1976`, `Mroz87` ou `MROZ.RAW` nos pacotes `AER`, `sampleSelection` e `ssmrob` respectivamente.

Para o nosso ajuste, utilizaremos "Horas anuais de trabalho" como variável dependente e consideraremos a nulidade dessas horas trabalhadas como uma censura em zero. Um estudo parecido que utiliza a nulidade da variável horas trabalhadas como ponto de censura pode ser encontrado na seção 19.3.3 de William (2008). Neste estudo os pesquisadores estavam interessados em saber se as esposas cujos casamentos eram estatisticamente mais propensos a se dissolver, diminuíam essa possibilidade passando mais tempo trabalhando.

Na Figura 7 podemos notar uma possível bimodalidade para a variável resposta, nos dando assim um direcionamento inicial para especificarmos uma quantidade de componentes no modelo de mistura igual a dois. Abaixo faremos uma descrição das variáveis que iremos utilizar no estudo.

- y_i: Horas anuais de trabalho da esposa fora de casa divididas por 1000. Se a mulher não trabalhou no ano, então logicamente suas horas anuais de trabalho serão iguais a zero. Assim consideraremos essa observação como censurada em zero. Essas observações são chamadas de censuras à esquerda e acumulam uma quantidade de 43,16% de amostras censuradas.

Assim como em Massuia *et al.* (2017), utilizaremos como variáveis explicativas:

- x_{i1}: Idade da esposa;

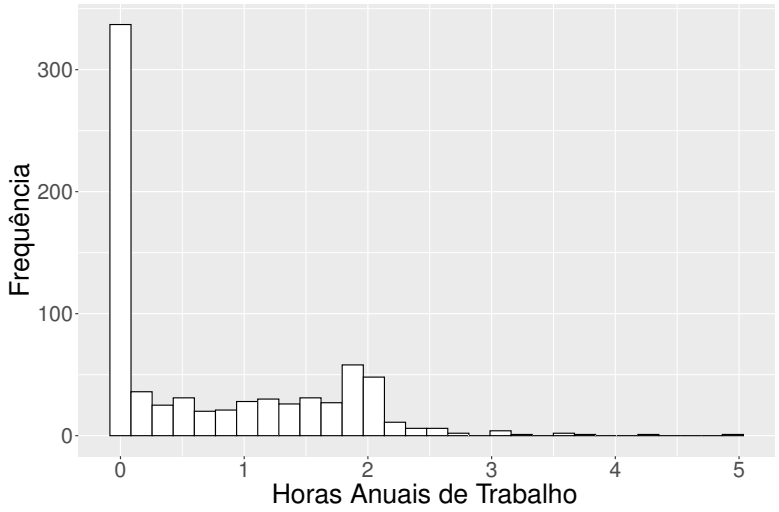

Figura 7 – Histograma da variável resposta Horas anuais de Trabalho.

- x_{i2}: Anos de escolaridade;
- x_{i3}: Número de crianças menores de seis anos que moram na casa;
- x_{i4}: Número de crianças entre seis e dezenove anos que moram na casa.

Nesse estudo consideramos que os dados são provenientes de uma mistura de duas componentes ($G = 2$) e geramos 66 mil amostras MCMC, retirando as primeiras 10 mil amostras, armazenando-as utilizando um espaçamento de 20 unidades e utilizando valores iniciais iguais aos utilizados na seção 3.5.1.

Para essa análise ajustaremos todos os modelos FMR-SMSN-CR, em especial para os modelos FMR-SCN-RC e FMR-CN-RC construiremos um grid de possíveis valores para ρ e η. Sendo o conjunto $\mathbf{B} = \{0,11; 0,22; 0,33; 0,44; 0,55; 0,66; 0,77; 0,88; 0,99\}$, consideramos todos os pares (ρ, η) que estão no produto cartesiano de $\mathbf{B} \times \mathbf{B}$, gerando um total de 81 pares de valores, após isso ajustamos os modelos FMR-SCN-RC e FMR-CN-RC e calculamos o WAIC_2 para os 81 ajustes. Para o modelo FMR-CN-RC o melhor ajuste ficou sob a responsabilidade do par de valores dos parâmetros $\rho = 0,22$ e $\eta = 0,55$ gerando um $\text{WAIC}_2 = 1809,29$ e para o modelos FMR-SCN-RC os valores $\rho = 0,66$ e $\eta = 0,44$ foram os que geraram o menor $\text{WAIC}_2 = 1806,031$.

Na Tabela 2 temos as medianas a posteriori (MD), desvios padrão (Dp) e intervalos de credibilidade (HPD) dos parâmetros (Par) do modelo após o ajuste dos diferentes modelos FMR-SMSN-CR. A escolha da mediana amostral como forma de estimar os parâmetros é proveniente do fato de que vários deles possuem distribuições a posteriori com comportamento assimétrico e de caudas pesadas como podemos ver na Figura 8. Podemos notar nesta Tabela 2 que as estimativas dos coeficientes de regressão são semelhantes em todos os modelos ajustados.

Tabela 2 – Dados Wage rate. Estimativas dos parâmetros (Par) para os modelos FMR-SMSN-CR. Md representa mediana e Dp representa o Desvio padrão das amostras MCMC.

	Modelos											
	FMR-N-CR			FMR-T-CR			FMR-SL-CR			FMR-CN-CR		
Par	Md	Dp	HPD (95%)	Md	Dp	HPD (95%)	Md	Dp	HPD (95%)	Md	Dp	HPD (95%)
β_{01}	1,818	0,208	(1,482; 2,306)	1,931	0,265	(1,497; 2,478)	2,046	0,277	(1,572;2,634)	1,845	0,246	(1,470; 2,421)
β_{11}	0,003	0,002	(-0,003; 0,007)	0,003	0,005	(-0,006; 0,009)	0,002	0,004	(-0,005; 0,009)	0,002	0,002	(-0,003; 0,007)
β_{21}	-0,002	0,013	(-0,031; 0,021)	-0,013	0,017	(-0,049; 0,015)	-0,020	0,017	(-0,055; 0,010)	-0,003	0,015	(-0,042; 0,018)
β_{31}	-2,630	1,380	(-6,274; -1,802)	-0,736	0,211	(-0,983; -0,110)	-0,709	0,254	(-0,922; 0,029)	-2,583	1,300	(-5,799;-1,739)
β_{41}	0,025	0,019	(-0,015; 0,060)	0,028	0,042	(-0,043; 0,074)	0,023	0,029	(-0,041; 0,069)	0,022	0,022	(-0,026; 0,062)
β_{02}	0,773	0,584	(-0,432; 1,788)	0,730	0,819	(-0,531; 1,998)	0,783	0,587	(-0,260; 2,004)	0,818	0,573	(-0,279; 1,887)
β_{12}	-0,045	0,009	(-0,063; -0,028)	-0,047	0,120	(-0,068; -0,025)	-0,048	0,009	(-0,066;-0,030)	-0,046	0,009	(-0,065;-0,029)
β_{22}	0,127	0,027	(0,071; 0,179)	0,133	0,259	(0,075; 0,195)	0,130	0,026	(0,077; 0,179)	0,125	0,026	(0,074; 0,175)
β_{32}	-0,746	0,158	(-1,076; -0,454)	-1,091	0,613	(-1,481; -0,712)	-1,104	0,162	(-1,428;-0,799)	-0,772	0,159	(-1,087;-0,456)
β_{42}	-0,093	0,048	(-0,187; 0,002)	-0,094	0,488	(-0,203; 0,020)	-0,100	0,050	(-0,200;-0,009)	-0,092	0,047	(-0,192;-0,006)
σ_1^2	0,014	0,007	(0,005; 0,030)	0,020	0,124	(0,005; 0,051)	0,019	0,010	(0,006; 0,041)	0,014	0,008	(0,005; 0,033)
σ_2^2	1,619	0,146	(1,337; 1,907)	1,322	25,75	(0,905; 1,748)	0,961	0,199	(0,603; 1,366)	1,369	0,128	(1,126; 1,621)
p_1	0,165	0,024	(0,120; 0,212)	0,158	0,092	(0,108; 0,218)	0,164	0,026	(0,118; 0,215)	0,170	0,025	(0,124; 0,225)
ν	-	-	-	8,861	8,124	(3,196; 23,951)	2,408	1,598	(1,400; 4,923)	-	-	-
	FMR-SN-CR			FMR-ST-CR			FMR-SSL-CR			FMR-SCN-CR		
Par	Md	Dp	HPD (95%)	Md	Dp	HPD (95%)	Md	Dp	HPD (95%)	Md	Dp	HPD (95%)
β_{01}	1,823	0,240	(1,438; 2,407)	1,680	0,231	(1,185; 2,103)	1,572	0,225	(1,047; 1,895)	1,776	0,217	(1,400; 2,245)
β_{11}	0,004	0,003	(-0,002; 0,010)	0,004	0,003	(-0,001; 0,010)	0,003	0,003	(-0,002; 0,009)	0,003	0,003	(-0,004; 0,009)
β_{21}	-0,012	0,018	(-0,045; 0,020)	-0,001	0,013	(-0,029; 0,020)	0,007	0,010	(-0,015; 0,024)	-0,007	0,014	(-0,035; 0,015)
β_{31}	-0,679	0,364	(-0,956; 0,285)	-2,026	1,096	(-5,354;-1,731)	-2,047	0,612	(-3,702;-1,741)	-0,678	0,214	(-0,996;-0,167)
β_{41}	0,022	0,025	(-0,027; 0,066)	0,014	0,020	(-0,026; 0,049)	0,005	0,022	(-0,033; 0,048)	0,019	0,024	(-0,029; 0,063)
β_{02}	0,685	0,556	(-0,406; 1,805)	0,752	0,589	(-0,557; 1,797)	0,543	0,462	(-0,035; 1,532)	0,760	0,609	(-0,428; 1,961)
β_{12}	-0,046	0,009	(-0,064;-0,028)	-0,048	0,009	(-0,067;-0,031)	-0,045	0,009	(-0,061;-0,031)	-0,047	0,009	(-0,065;-0,029)
β_{22}	0,137	0,027	(0,085; 0,188)	0,128	0,029	(0,074; 0,187)	0,126	0,026	(0,087; 0,184)	0,132	0,030	(0,076;-0,192)
β_{32}	-1,039	0,163	(-1,363;-0,725)	-0,705	0,179	(-1,069;-0,360)	-0,529	0,156	(-0,855;-0,241)	-1,059	0,165	(-1,390; 0,733)
β_{42}	-0,090	0,051	(-0,194; 0,005)	-0,093	0,051	(-0,190; 0,005)	-0,063	0,046	(-0,145; 0,031)	-0,098	0,054	(-0,204; 0,008)
σ_1^2	0,064	0,043	(0,011; 0,145)	0,089	0,071	(0,013; 0,251)	0,132	0,215	(0,009; 0,836)	0,054	0,060	(0,006; 0,211)
σ_2^2	2,569	0,534	(1,587; 3,505)	1,631	0,552	(0,984; 2,978)	2,311	0,512	(1,295; 3,388)	1,685	0,719	(0,512; 3,312)
p_1	0,157	0,023	(0,113; 0,205)	0,789	0,037	(0,709; 0,852)	0,246	0,049	(0,171; 0,372)	0,164	0,040	(0,110; 0,243)
λ_1	-3,691	16,939	(-30,578;3,150)	-5,275	18,286	(-20,24; 0,572)	-7,528	4,319	(-14,968;-0,871)	-4,417	0,011	(-30,72; 1,946)
λ_2	1,395	2,886	(-1,313; 4,760)	-0,095	0,978	(-1,889; 1,737)	14,608	23,930	(-2,525;48,461)	0,373	1,539	(-1,798; 3,846)
ν	-	-	-	8,363	6,174	(2,477;21,961)	4,718	3,401	(1,400;12,140)	-	-	-

É importante notarmos também que os parâmetros $\beta_{11}, \beta_{21}, \beta_{41}$ e β_{02} contiveram o zero em seus intervalos de credibilidade, isso naturalmente nos chama atenção para a possibilidade das variáveis “Idade da esposa”, “Anos de escolaridade” e “Número de crianças entre seis e dezenove anos que moram na casa” não terem relação com a variável resposta, porém essa afirmação precisa ser avaliada com bastante atenção e para melhor compreendermos esse fato é importante observarmos a Figura 8, nessa figura temos os gráficos de traço (tal que podemos observar que todos os parâmetros cobriram de forma satisfatória os seus respectivos suportes) e as estimativas de densidade kernel do modelo FMR-ST-CR (os gráficos para os demais modelos assimétricos estão em anexo), para tanto podemos observar que apesar de conterem o zero em seus intervalos de credibilidade, apenas β_{21} tem o zero como moda da sua distribuição a posteriori estimada. Talvez um melhor modelo para ajustar esses dados seria desconsiderando a variável “Anos

de escolaridade" porém, apenas na primeira componente. O parâmetros λ_2 para o modelo FMR-SSL-CR e λ_1 e λ_2 para os modelos FMR-SN-CR e FMR-ST-CR também possuem o zero em seus intervalos de credibilidade, porém com uma variabilidade muito grande, levantando evidências contra a hipótese de que esses parâmetros sejam nulos e consequentemente os modelos simétricos sejam mais adequados para essa análise. Podemos ver isso mais claramente na Figura 8, principalmente para λ_1.

Tabela 3 – Dados Wage rate. Critérios de seleção e p-valor Bayesiano (p_B) dos modelos FMR-SMSN-CR. A nomenclatura (2) refere-se aos modelos com duas componentes na mistura e (1) aos modelos com apenas uma componente na mistura.

	Critérios			
Modelos	LPML	$WAIC_1$	$WAIC_2$	p_B
FMR-N-CR(2)	-918,2694	1836,5080	1836,2470	0,6110
FMR-T-CR(2)	-905,8365	1812,0030	1810,1850	0,6578
FMR-SL-CR(2)	-903,7427	1808,8540	1806,5790	0,6525
FMR-CN-CR(2)	-902,6855	1809,7830	1809,2900	0,6110
FMR-SN-CR(2)	-905,0430	1816,3300	1806,0700	0,6575
FMR-ST-CR(2)	**-891,5914**	**1782,6500**	**1782,9320**	**0,5307**
FMR-SSL-CR(2)	-892,3793	1783,6900	1784,9930	0,6071
FMR-SCN-CR(2)	-901,7500	1803,8700	1806,0310	0,6572
FMR-N-CR(1)	-953,6760	1907,1890	1907,3560	0,6971
FMR-T-CR(1)	-954,7615	1909,3580	1909,5250	0,6224
FMR-SL-CR(1)	-953,8774	1907,5660	1907,7610	0,6322
FMR-SN-CR(1)	-953,8943	1906,3650	1907,7470	0,5882
FMR-ST-CR(1)	-935,2677	1870,2060	1870,5530	0,5382
' FMR-SSL-CR(1)	-941,3766	1882,3350	1882,7370	0,5432

Na Tabela 3 podemos ver os critérios de seleção de modelos LPML, $WAIC_1$, $WAIC_2$ e o p-valor Bayesiano calculados para todos os modelos FMR-SMSN-CR descritos nesse capítulo. Vemos que o modelo que obteve o maior LPML e menores $WAIC_1$ e $WAIC_2$ foi o modelo FMR-ST-CR sugerindo a possibilidade de que este é o modelo que melhor ajusta esse conjunto de dados. O modelo normal foi o pior ajuste levando em consideração os critérios de seleção.

Uma ideia prospectiva que pode emergir durante a análise modelos de misturas finitas, é se realmente vale a pena o uso de uma estrutura com mais de uma componente quando estamos analisando um conjunto de dados, considerando essa abordagem ajustamos todos os modelos FMR-SMSN-CR com apenas uma componente na mistura, ou seja, ajustamos o modelo proposto por Massuia *et al.* (2017). A Tabela 3 apresenta os critérios de seleção de modelos

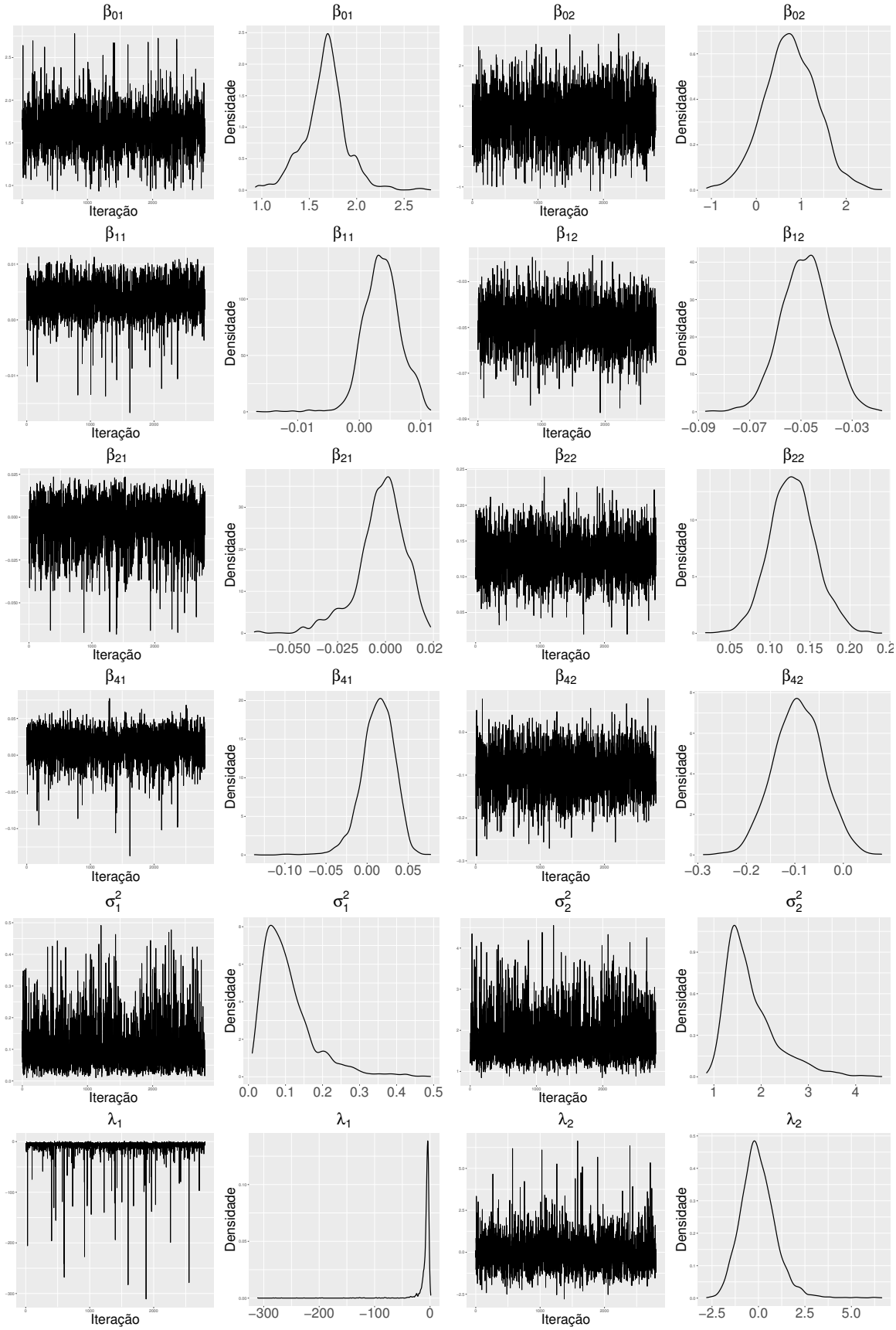

Figura 8 – Gráficos de traço e de Kernel de alguns parâmetros do modelo FMR-ST-CR ajustado nos dados *wage rate*.

e o p-valor Bayesiano para os modelos com e sem mistura. Segundo os critérios de seleção, é interessante notarmos que nenhum dos modelos sem mistura obteve melhor comportamento que os modelos considerando duas densidades na mistura. Apesar disso, todos modelos mostraram uma boa adequação quando observamos particularmente os p-valores envolvidos, inclusive os sem misturas.

Neste capítulo, propusemos a classe de distribuições SMSN, como uma substituta para a escolha convencional da distribuição normal, em misturas de modelos TOBIT. Generalizando os trabalhos de Massuia *et al.* (2017) e Zeller *et al.* (2019), utilizando uma abordagem Bayesiana. Para explorar as propriedades estatísticas dos modelos propostos, criamos um algoritmo eficiente do tipo Gibbs. Quatro estudos de simulação foram realizados. O primeiro estudo revelou uma boa eficiência e precisão das estimativas dos parâmetros em todos os níveis de censura e tamanhos amostrais, analisando os modelos assimétricos e seus correspondentes simétricos. No segundo estudo de simulação, vimos uma performance muito satisfatória dos modelos assimétricos ao ajustar dados oriundos do modelo FMR-NIG-CR. No terceiro estudo de simulação, utilizamos o algoritmo de Clonagem de Dados a fim de indentificarmos indícios de que o nosso modelo é identificável, o que aconteceu exatamente dessa forma. No quarto estudo de simulação, mostramos que o desempenho das estimativas dos parâmetros para os modelos FMR-ST-CR e FMR-SSL-CR é melhor do que no modelo FMR-SN-CR quando perturbamos, de forma gradativa, observações de indivíduos escolhidos ao acaso. Essa perturbação se dá no aumento em porcentagem das observações de cada unidade amostral escolhida. Finalmente, analisamos um conjunto de dados reais, no qual vimos que os modelos que consideram duas componentes na mistura tiveram melhores ajustes em comparação aos modelos com apenas uma componente, além de observarmos melhores ajustes para os modelos assimétricos em comparação aos demais modelos.

4 MODELO DE MISTURA DE REGRESSÕES COM DADOS AUSENTES

4.1 INTRODUÇÃO

Dados ausentes são recorrentes na literatura e se forem tratados de forma inadequada podem levar a inferências tendenciosas e ineficientes. Vários tipos de experimentos podem conter dados ausentes, em particular citamos os dados longitudinais. A ausência de observações é algo muito recorrente em estudos longitudinais, principalmente pelas desistências ou abandonos do estudo antes da sua conclusão pretendida, embora a maioria dos estudos longitudinais sejam projetados para coletar cada indivíduo, ou unidade amostral, em todos os momentos do acompanhamento.

Há uma extensa literatura que aborda e propõe diversos métodos e modelos que consideram dados ausentes. Dentre os livros didáticos que abordam esse tema, estão: Daniels e Hogan (2008), Molenberghs *et al.* (2014) e Little e Rubin (2019). Técnicas simples como análise completa de caso (também chamada de exclusão de lista), métodos de ponderação e métodos de imputação múltipla, podem ser utilizados para considerar na análise a ausência de dados, mas podem levar a erros referentes a estimação dos parâmetros. Nessa área de estudo os trabalhos avançam principalmente na construção de modelos flexíveis considerando dados ausentes, como Wang *et al.* (2004) que utilizaram o modelo t de Student multivariado ou Lim, Narisetty e Cheon (2017), que estenderam esse estudo para mistura de modelos de regressões t de Student multivariado, utilizando uma abordagem frequentista.

O estudo de dados ausentes pode ser separado em métodos para dados ausentes ignoráveis e não ignoráveis (esses conceitos serão abordados na próxima seção, porém para um aprofundamento no tema veja Daniels e Hogan (2008)). Aqui podemos citar alguns estudos realizados considerando a existência de dados ausentes não ignoráveis em um contexto Bayesiano, como Zhao e Duan (2022), que construíram um algoritmo do tipo Gibbs para a estimação de um modelo de regressão normal, com dados ausentes na variável resposta, e Cai, Song e Hser (2010) que por sua vez fizeram um estudo Bayesiano utilizando um modelo de misturas de equações estruturais multivariadas em um contexto de dados faltantes nas variáveis respostas e nas covariáveis. Outro trabalho muito interessante nesse âmbito é o de Ma e Chen (2018), que revisa os recentes desenvolvimentos e aplicações de métodos Bayesianos para lidar com dados ausentes ignoráveis e não ignoráveis.

Neste trabalho, o objetivo é introduzir uma extensão ainda maior para o modelo de

regressão com dados ausentes. A ideia é ajustar um modelo de mistura de regressões que leve em consideração uma estrutura de dados ausentes e que os erros aleatórios tenham distribuição SMSN. Nesse contexto, há vários trabalhos que propõem misturas de modelos regressão com dados ausentes, em que os mesmos se encontram apenas na variável resposta, como em Lim, Narisetty e Cheon (2017). Nesse capítulo propomos um modelo que leva em consideração dados ausentes tanto na resposta como nas covariáveis, assim como em Cai, Song e Hser (2010).

De uma perspectiva Bayesiana, os dados ausentes são tratados como quantidades desconhecidas adicionais para as quais uma distribuição a posteriori pode ser estimada. Portanto, a abordagem Bayesiana não faz distinção fundamental entre dados ausentes e parâmetros, pois ambos são quantidades aleatórias desconhecidas. Nós apenas precisamos especificar um modelo conjunto apropriado para os dados observados e ausentes e os parâmetros do modelo.

Segundo Molenberghs *et al.* (2014), quando estamos falando de dados faltantes, a inferência sobre parâmetros geralmente requer suposições não testáveis na distribuição dos dados faltantes. Logo, as conclusões da análise com dados incompletos necessitam de um pouco mais de cuidado e as razões que levaram as observações serem faltantes devem ser cuidadosamente consideradas. Essas afirmações se mostram bem coerentes quando precisamos definir distribuições para as covariáveis que possuem observações ausentes e a escolha dos diferentes mecanismos de geração de dados ausentes, os quais iremos ver na sequencia desse trabalho.

4.2 MECANISMOS DE GERAÇÃO DE DADOS AUSENTES

Quando estudamos modelos para ajustar dados que contêm observações faltantes, é importante entender e considerar o mecanismo gerador desses dados ausentes. Geralmente as ausências são geradas por mecanismos que não são controlados no experimento e sim analisados posteriormente, pois as ausências podem ocorrer de maneira espontânea e não previsível. Por exemplo, um participante pode se recusar a responder uma determinada questão, o que levaria a uma ausência de dados que não pode ser controlada no momento da coleta. Definiremos três mecanismos que descrevem a probabilidade de uma unidade amostral ser considerada faltante: (i) Faltantes completamente ao acaso - MCAR (*Missing Completely at Random*), (ii) Faltantes aleatoriamente - MAR (*Missing at Random*) e (iii) Faltantes não aleatoriamente - MNAR (*Missing Not at Random*). Rubin (1976) desenvolveu essa nomenclatura com o intuito de auxiliar o estudo de dados ausentes.

Para um melhor entendimento, suponha o caso em que temos dados ausentes tanto na resposta como nas covariáveis, no contexto de modelos de regressão. Considere os vetores aleatórios $\mathbf{Y} = (Y_1, \cdots, Y_n)^\top$ e $\mathbf{x}_i = (x_{i1}, ..., x_{iq})^\top$, note que os mesmos podem ser escritos como $\mathbf{Y} = \{\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}\}$ e $\mathbf{x}_i = \{\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}_i^{\mathrm{m}}\}$ em que $\mathbf{Y}^{\mathrm{o}}$ e $\mathbf{x}_i^{\mathrm{o}}$ denotam a parte observável de $\mathbf{Y}$ e $\mathbf{x}_i$ e $\mathbf{Y}^{\mathrm{m}}$ e $\mathbf{x}_i^{\mathrm{m}}$ a parte não observável de $\mathbf{Y}$ e $\mathbf{x}_i$, respectivamente.

Para introduzirmos os conceitos de dados faltantes, vamos supor que $Y_1, ..., Y_n$ são observações independentes, tais que $Y_i \sim \sum_{j=1}^{G} p_j \mathrm{SMSN}(\mathbf{x}_i^\top \boldsymbol{\beta}_j + b\Delta_j, \sigma_j^2, \lambda_j; h)$ para $i = 1, ..., n$, também definimos $\mathbf{r}^y = (r_1^y, \cdots, r_n^y)^\top$ e $\mathbf{r}_i^x = (r_{i1}^x, \cdots, r_{iq}^x)^\top$, que são os indicadores de observações ausentes para $\mathbf{Y}$ e $\mathbf{x}_i$, tais que

$$r_i^y = \begin{cases} 0, \text{ se } Y_i \text{ for observado} \\ 1, \text{ c.c.} \end{cases} \quad \text{e} \quad r_{ik}^x = \begin{cases} 0, \text{ se } x_{ik} \text{ for observado, } k = 1, \cdots, q \\ 1, \text{ c.c.} \end{cases} \tag{4.1}$$

e finalmente definamos $\boldsymbol{\varphi}^y = (\varphi_0^y, \varphi_1^y)^\top$ e $\boldsymbol{\varphi}^x = (\varphi_0^x, \varphi_1^x, \cdots, \varphi_q^x)^\top$ parâmetros relacionados a distribuições de r_i^y e $\mathbf{r}_i^x$, respectivamente.

Missing Completely at Random (MCAR): É um tipo de mecanismo de dados faltantes que implica que a probabilidade de uma variável ser ausente é independente de todas as outras variáveis disponíveis, incluindo as variáveis dependentes e independentes em termos que um modelo de regressão, ou seja,

$$\pi(r_i^y | Y_i^{\mathrm{m}}, Y_i^{\mathrm{o}}) = \pi(r_i^y) \qquad \text{e} \qquad \pi(\mathbf{r}_i^x | \mathbf{x}_i^{\mathrm{m}}, \mathbf{x}_i^{\mathrm{o}}) = \pi(\mathbf{r}_i^x).$$

Em outras palavras, a ausência de dados não está relacionada a nenhuma característica do objeto de estudo ou das variáveis em questão. Este tipo de mecanismo é considerado o mais benéfico em termos de análise, pois não introduz viés ou distorções nas estimativas. No entanto, é importante destacar que o MCAR é difícil de ser verificado e é considerado raro em situações reais.

Missing at Random (MAR): Em contraste com o MCAR, diz-se que os dados estão ausentes aleatoriamente (MAR) se a probabilidade da ausência de uma observação em um conjunto de dados for dependente apenas dos valores observados e não está relacionada aos valores ausentes. Essa abordagem permite estudar os valores ausentes com base nas informações disponíveis ou observadas. Em termos de probabilidade dizemos que os dados ausentes obedecem ao mecanismo MAR se,

$$\pi(r_i^y|Y_i^{\mathrm{m}},Y_i^{\mathrm{o}}) = \pi(r_i^y|Y_i^{\mathrm{o}}) \qquad \text{e} \qquad \pi(\mathbf{r}_i^x|\mathbf{x}_i^{\mathrm{m}},\mathbf{x}_i^{\mathrm{o}}) = \pi(\mathbf{r}_i^x|\mathbf{x}_i^{\mathrm{o}}).$$

Um exemplo de mecanismo MAR pode surgir quando um experimento exige que um sujeito seja removido do estudo assim que o valor de uma variável cair fora de um determinado intervalo de valores, por exemplo, admita que duas variáveis, pressão sistólica (Y_i) e idade (X_i) são observadas em alguns indivíduos, porém ao final da coleta de dados serão descartados as pressões arteriais de indivíduos com uma idade acima de 60 anos. Nesse caso, as ausências em Y_i dependem dos valores observados de X_i.

O uso do mecanismo MAR pode ser considerado uma suposição mais plausível sobre dados ausentes em muitas aplicações, isso se dá porque o mecanismo MAR é muito menos restritivo para modelos de regressão com dados ausentes do que o mecanismo MCAR. Molenberghs *et al.* (2014) afirma que é indiscutível que a suposição MAR deve ser a suposição padrão para a análise de dados parcialmente ausentes, a menos que haja uma razão forte e convincente para apoiar a suposição MCAR.

Missing Not at Random (MNAR): É um mecanismo de dados faltantes que assume que a ausência de dados não é aleatória, ou seja, depende dos valores específicos que deveriam ter sido obtidos, além dos realmente obtidos. Em outras palavras, a probabilidade de um dado faltante é influenciada pelos valores dos outros dados tanto observados como faltantes.

Por exemplo, ainda na mesma ideia do experimento hipotético citado para ilustrar o mecanismo MAR. Suponha o conjunto de dados antes da exclusão de alguns Y_i e agora suponha que algumas idades (X_i) serão descartadas, se as pressões sistólicas associadas a elas estiverem abaixo de um valor especificado, ao final teremos tanto Y_i com X_i como observações ausentes. É claro que algumas exclusões de X_i serão condicionadas a valores de Y_i que foram considerados ausentes, isso nos dá a ideia de que dessa forma o método de exclusão está sob um mecanismo MNAR pois o fato de uma observação ser ausente depende tanto das partes observadas como das partes ausentes dos dados.

Nesta situação, a probabilidade de um dado faltante pode ser modelada utilizando uma distribuição condicional,

$$\pi(r_i^y|Y_i) = \pi(r_i^y|Y_i^{\mathrm{m}},Y_i^{\mathrm{o}}) \qquad \text{e} \qquad \pi(\mathbf{r}_i^x|\mathbf{x}_i) = \pi(\mathbf{r}_i^x|\mathbf{x}_i^{\mathrm{m}},\mathbf{x}_i^{\mathrm{o}}).$$

Qualquer método inferencial válido sob o mecanismo NMAR requer a especificação

de um modelo para o mecanismo de dados ausentes. Um mecanismo NMAR é muitas vezes chamado de falta "não ignorável" porque o mecanismo de dados ausentes não pode ser ignorado quando o objetivo é fazer inferências sobre a distribuição dos dados completos (ver seção 1.4 de Molenberghs e Kenward (2007)).

4.3 MISTURA FINITA DE MODELOS DE REGRESSÃO COM DADOS AUSENTES

Existe uma grande variedade de possibilidades quando estamos falando de estruturas de dados ausentes em modelos de regressão. Na literatura, podemos encontrar trabalhos em que os dados faltantes são tratados unicamente na variável resposta como em Zhao e Duan (2022), outra possibilidade são estudos em que os dados faltantes estão apenas nas covariáveis como em Garcia, Ibrahim e Zhu (2010). Esse capítulo tratará exatamente do problemas de dados ausentes na variável resposta e em todas as covariáveis, em que a variável resposta será univariada.

Para estudarmos os dados sob um contexto não ignorável tanto na resposta como nas covariáveis, precisamos definir uma estrutura apropriada para modelar as distribuições de $r_i^y|Y_i$ e $\mathbf{r}_i^x|\mathbf{x}_i$. Cai, Song e Hser (2010) atenta-nos para o cuidado que devemos ter com a utilização de modelos complexos ou com muitas variáveis, pois facilmente podem tornar o modelo não identificável.

Assim como em Zhao e Duan (2022) e Cai, Song e Hser (2010), consideraremos que $r_{ik}^x|\mathbf{x}_i$ são independentes para todo $k = 1, \cdots, q$, com funções de probabilidade dos indicadores dadas por

$$\pi(r_i^y|y_i, \boldsymbol{\varphi}^y) = \pi(r_i^y = 1|y_i, \boldsymbol{\varphi}^y)^{r_i^y}[1 - \pi(r_i^y = 1|y_i, \boldsymbol{\varphi}^y)]^{1-r_i^y}, \tag{4.2}$$

$$\pi(\mathbf{r}_i^x|\mathbf{x}_i, \boldsymbol{\varphi}^x) = \prod_{k=1}^{q} \pi(r_{ik}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}^y)^{r_{ik}^x}[1 - \pi(r_{ik}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}^y)]^{1-r_{ik}^x}. \tag{4.3}$$

Para tanto, iremos definir $\pi(r_i^y|y_i, \boldsymbol{\varphi}^y)$ e $\pi(\mathbf{r}_i^x|\mathbf{x}_i, \boldsymbol{\varphi}^x)$ como modelos de regressão logística da seguinte forma,

$$\text{logit}\left[p(r_i^y = 1|y_i, \boldsymbol{\varphi}^y)\right] = \varphi_0^y + \varphi_1^y y_i = (\boldsymbol{\varphi}^y)^\top \mathbf{y}_i^*, \tag{4.4}$$

$$\text{logit}\left[p(r_{ik}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}^x)\right] = \varphi_0^x + \sum_{l=1}^{q} \varphi_k^x x_{il} = (\boldsymbol{\varphi}^x)^\top \mathbf{x}_i^*, \tag{4.5}$$

em que $\mathbf{y}_i^* = (1, y_i)^\top$ e $\mathbf{x}_i^* = (1, \mathbf{x}_i)^\top$. Logo, as funções de probabilidade dos indicadores podem ser escritas como

$$\pi(r_i^y|y_i,\boldsymbol{\varphi}^y) = \exp\left\{(\boldsymbol{\varphi}^y)^\top \mathbf{y}_i^* r_i^y - \log\left(1+\exp\left\{(\boldsymbol{\varphi}^y)^\top \mathbf{y}_i^*\right\}\right)\right\}, \tag{4.6}$$

$$\pi(\mathbf{r}_i^x|\mathbf{x}_i,\boldsymbol{\varphi}^x) = \exp\left\{(\boldsymbol{\varphi}^x)^\top \mathbf{x}_i^* \sum_{k=1}^{p} r_{ik}^x - p\log\left(1+\exp\left\{(\boldsymbol{\varphi}^x)^\top \mathbf{x}_i^*\right\}\right)\right\}. \tag{4.7}$$

Note que se $\varphi_1^y = \varphi_1^x = \cdots = \varphi_q^x = 0$ em (4.4) e (4.5) então teremos exatamente o mecanismo MCAR com $p(r_i^y = 1|y_i,\varphi_0^y) = \exp\{\varphi_0^y\}$ e $p(r_{ik}^x = 1|\mathbf{x}_i,\boldsymbol{\varphi}^x) = \exp\{\varphi_0^x\}$.

Quando as covariáveis e a resposta estão sob o mecanismo MNAR, a verossimilhança é dada por,

$$\pi(y_i,\mathbf{x}_i,r_i^y,\mathbf{r}_i^x|\boldsymbol{\psi},\boldsymbol{\varphi}^x,\boldsymbol{\varphi}^y) = \pi(y_i|\mathbf{x}_i,\boldsymbol{\psi})\pi(\mathbf{x}_i|\boldsymbol{\alpha})\pi(r_i^y|y_i,\boldsymbol{\varphi}^y)\pi(\mathbf{r}_i^x|\mathbf{x}_i,\boldsymbol{\varphi}^x). \tag{4.8}$$

note que $\mathbf{x}_i$ é aleatório com densidade ou função de probabilidade $\pi(\mathbf{x}_i|\boldsymbol{\alpha})$ em que $\boldsymbol{\alpha}$ é um vetor de parâmetros conhecido.

Sendo $\mathbf{X} = (\mathbf{x}_1,\cdots,\mathbf{x}_n)^\top$, $\mathbf{r}^x = (\mathbf{r}_1^x,\cdots,\mathbf{r}_n^x)^\top$ e $\boldsymbol{\Psi} = (\boldsymbol{\psi},\boldsymbol{\varphi}^x,\boldsymbol{\varphi}^y)^\top$ então temos a verossimilhança dada por,

$$\pi(\mathbf{y},\mathbf{x},\mathbf{r}^x,\mathbf{r}^y|\boldsymbol{\Psi}) = \prod_{i=1}^{n}\pi(y_i|\mathbf{x}_i,\boldsymbol{\psi})\pi(\mathbf{x}_i|\boldsymbol{\alpha})\pi(r_i^y|y_i,\boldsymbol{\varphi}^y)\pi(\mathbf{r}_i^x|\mathbf{x}_i,\boldsymbol{\varphi}^x). \tag{4.9}$$

Segundo Rubin (1976), se as covariáveis e/ou a variável resposta são MAR ou MCAR, ou seja, modelos ignoráveis, então as quantidades $\pi(r_i^y|y_i,\boldsymbol{\varphi}^y)$ e/ou $\pi(\mathbf{r}_i^x|\mathbf{x}_i,\boldsymbol{\varphi}^x)$, que representam os mecanismos de dados faltantes, podem ser ignoradas de (4.9). Ver seção 1.3.2 de Molenberghs e Kenward (2007) ou Corolário 6.1A de Little e Rubin (2019).

Segundo Ma e Chen (2018) quando há mais de uma covariável com observações ausentes no conjunto de dados, uma possibilidade é modelar todas as q covariáveis ausentes utilizando distribuições multivariadas, por exemplo, utilizando uma distribuição normal multivariada $\mathrm{N}_q(\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)$ se todas as covariáveis ausentes forem contínuas (a forma da matriz $\boldsymbol{\Sigma}_x$ dependerá das intenções da modelagem, podendo ter uma estrutura de dependência entre as covariáveis ou, de uma forma mais simples, sendo diagonal supondo independência entre as mesmas, como fizemos na sessão 4.6.1) , ou alguma distribuição multivariada discreta se todas as covariáveis forem discretas ou até uma regressão probit multivariada para covariáveis binárias correlacionadas. Para facilitar o método abordaremos apenas o caso em que as covariáveis são absolutamente contínuas. Para uma leitura mais aprofundada sob outras diferentes formas utilizadas para modelar as covariáveis sugerimos Garcia, Ibrahim e Zhu (2010) e Ma e Chen (2018) . Na seção 4.6.1, faremos um estudo de comparação entre duas distribuições diferentes para as covariaveis, normal e uniforme, em um modelo de regressão com duas regressoras.

Ao modelo cuja verossimilhança é dada por (4.9) daremos o nome de FMR-SMSN-MD no qual $\pi(y_i|\mathbf{x}_i,\boldsymbol{\psi})$ é dado em (2.14) e pode ser escrito como em (2.16).

4.3.1 Inferência Bayesiana e Distribuição a Posteriori

Vamos definir inicialmente o vetor $\boldsymbol{\Theta}=(\boldsymbol{\theta},\boldsymbol{\varphi}^x,\boldsymbol{\varphi}^y)^\top$ de parâmetros reparametrizados. Essa reparametrização ajudará bastante na construção do algoritmo do tipo Gibbs, podendo ter os parâmetros originais recuperados por (2.6).

Para trabalharmos os conceitos inferenciais no modelo FMR-SMSN-MD, é necessário gerarmos observações aleatórias a partir da distribuição a posteriori $\pi(\boldsymbol{\Theta}|\mathbf{D}^{\mathrm{o}})\propto\pi(\mathbf{D}^{\mathrm{o}}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})$, no qual $\mathbf{D}^{\mathrm{o}}=\{\mathbf{Y}^{\mathrm{o}},\mathbf{X}^{\mathrm{o}},\mathbf{r}^{\mathrm{y}},\mathbf{r}^{\mathrm{x}}\}$ denota os dados observados tal que $\mathbf{X}^{\mathrm{o}}=(\mathbf{x}_1^{\mathrm{o}},\cdots,\mathbf{x}_n^{\mathrm{o}})^\top$, $\pi(\mathbf{D}^{\mathrm{o}}|\boldsymbol{\Theta})$ é a verossimilhança para os dados observados e $\pi(\boldsymbol{\Theta})$ representa a distribuição a priori de $\boldsymbol{\Theta}$. A distribuição a posteriori é dada por

$$\begin{aligned}\pi(\boldsymbol{\Theta}|\mathbf{D}^{\mathrm{o}}) &\propto \pi(\mathbf{D}^{\mathrm{o}}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta}),\\ &\propto \left[\int\int \pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\alpha})\pi(\mathbf{r}^x|\mathbf{x},\boldsymbol{\varphi}^x)\pi(\mathbf{r}^y|\mathbf{y},\boldsymbol{\varphi}^y)d\mathbf{y}^{\mathrm{m}}d\mathbf{x}^{\mathrm{m}}\right]\pi(\boldsymbol{\Theta}),\\ &\propto \left[\int\int \prod_{i=1}^{n}\pi(y_i|\mathbf{x}_i,\boldsymbol{\theta})\pi(\mathbf{x}_i|\boldsymbol{\alpha})\pi(\mathbf{r}_i^x|\mathbf{x}_i,\boldsymbol{\varphi}^x)\pi(r_i^y|y_i,\boldsymbol{\varphi}^y)d\mathbf{y}^{\mathrm{m}}d\mathbf{x}^{\mathrm{m}}\right]\pi(\boldsymbol{\Theta}),\end{aligned} \tag{4.10}$$

em que $\pi(y_i|\mathbf{x}_i,\boldsymbol{\theta})=\sum_{j=1}^{G}p_j\mathrm{SMSN}(y_i|\mathbf{x}_i^\top\boldsymbol{\beta}_j+b\Delta_j,\Delta_j,\tau_j;h)$.

É notável que a complexidade do modelo proposto, que é agravada com a existência de um mecanismo de dados faltantes não ignorável, gera a distribuição a posteriori dada em (4.10), que contém integrais de alta dimensão e sem expressão analítica.

Com o intuito de desenvolver uma maneira eficaz para ajustar o modelo FMR-SMSN-MD, definimos um conjunto de dados aumentamos, que inclui os dados observados, $\mathbf{D}^{\mathrm{o}}$, os vetores de valores ausentes $\mathbf{Y}^{\mathrm{m}}$ e $\mathbf{X}^{\mathrm{m}}=\{\mathbf{x}_1^{\mathrm{m}},\cdots,\mathbf{x}_n^{\mathrm{m}}\}$ e os vetores de variáveis latentes $\mathbf{T}$ e $\mathbf{U}$, dados pela representação estocástica (2.16). A ideia é criar o conjunto de dados completos $\mathbf{D}=\{\mathbf{Y},\mathbf{X},\mathbf{T},\mathbf{U},\mathbf{Z},\mathbf{r}^x,\mathbf{r}^y\}$, assim, as estimativas Bayesianas de $\boldsymbol{\Theta}$ podem ser obtidas gerando amostras da distribuição a posteriori conjunta $\pi(\boldsymbol{\Theta},\mathbf{D}^{\mathrm{m}}|\mathbf{D}^{\mathrm{o}})$ tal que $\mathbf{D}^{\mathrm{m}}=\{\mathbf{Y}^{\mathrm{m}},\mathbf{X}^{\mathrm{m}},\mathbf{T},\mathbf{U},\mathbf{Z}\}$.

A especificação das distribuições a priori para os parâmetros será feita para os parâmetros reparametrizados. Definiremos para cada componente da mistura, no mecanismo MNAR, $\boldsymbol{\varphi}^x\sim\mathrm{N}_p(\boldsymbol{\mu}^0_{\varphi^x},\boldsymbol{\Sigma}^0_{\varphi^x})$ e $\boldsymbol{\varphi}^y\sim\mathrm{N}_2(\boldsymbol{\mu}^0_{\varphi^y},\boldsymbol{\Sigma}^0_{\varphi^y})$, e para todos os mecanismos teremos, $\boldsymbol{\phi}_j=(\boldsymbol{\beta}_j,\Delta_j)\sim\mathrm{N}_{p+1}(\mathbf{b}_0,\mathbf{B}_0)$, $\tau_j|f_\tau\sim\mathrm{InvGamma}(e_\tau,f_\tau)$ e $f_\tau\sim\mathrm{Gamma}(g,h)$. Os hiperparâmetros $\mathbf{b}_0$, $\mathbf{B}_0$, $\boldsymbol{\mu}^0_{\varphi^x}$, $\boldsymbol{\Sigma}^0_{\varphi^x}$, $\boldsymbol{\mu}^0_{\varphi^y}$, $\boldsymbol{\Sigma}^0_{\varphi^y}$, e_τ, g e h são conhecidos.

Assim como no modelo com censuras as escolhas dos valores dos hiperparâmetros das distribuições a priori, serão escolhido com o propósito de expressar pouco conhecimento prévio para o parâmetro em questão. Por esse motivo as matrizes de covariância $\mathbf{B}_0$, $\boldsymbol{\Sigma}^0_{\varphi^x}$ e $\boldsymbol{\Sigma}^0_{\varphi^y}$ serão assumidas como sendo diagonais com valores iguais a 100. Os hiperparâmetros $\mathbf{b}_0$, $\boldsymbol{\mu}^0_{\varphi^y}$ e $\boldsymbol{\mu}^0_{\varphi^x}$ serão considerados nulos. Os hiperparâmetro e_τ, g e h assim como as priores de $\mathbf{p}$ e ν, e seus respectivos hiperparâmetros, serão as mesmas utilizadas no modelo FMR-SMSN-RC (seção 3.3.1) assim como as especificações dos parâmetros ρ e η dos modelos FMR-CN-MD e FMR-SCN-MD.

Para a distribuição a priori de $\boldsymbol{\Theta}$ assumiremos independência entre os parâmetros, ou seja, a distribuição a priori conjunta é dada por,

$$\pi(\boldsymbol{\Theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\Delta})\pi(p)\pi(\nu|\gamma)\pi(\gamma)\pi(\boldsymbol{\varphi}^x)\pi(\boldsymbol{\varphi}^y)\prod_{j=1}^{G}\pi(\tau_j|f_\tau)\pi(f_\tau).$$

4.4 ALGORITMO DO TIPO GIBBS

A ideia do algoritmo do tipo Gibbs, para o modelo FMR-SMSN-MD, é gerar amostras aleatórias de $\{\boldsymbol{\Theta}, \mathbf{D}^{\mathrm{m}}\} = \{\boldsymbol{\theta}, \boldsymbol{\varphi}^x, \boldsymbol{\varphi}^y, \mathbf{Y}^{\mathrm{m}}, \mathbf{X}^{\mathrm{m}}, \mathbf{T}, \mathbf{U}, \mathbf{Z}\}$ a partir da distribuição a posteriori dos parâmetros simulando de forma interativa e utilizando as distribuições condicionais completas. As gerações das amostras da posteriori dos parâmetros podem ser feitas de forma independente para cada componente.

As condicionais completas do algoritmo do tipo Gibbs são dadas abaixo por meio de passos que formam um algoritmo que nós auxiliará na obtenção de amostras da distribuição a posteriori do modelo FMR-SMSN-MD.

Passo 1. Gere $\mathbf{p} = (p_1, \cdots, p_G)^\top$ a partir de $\pi(\mathbf{p}|\mathbf{z}_1, \cdots, \mathbf{z}_n)$, que é uma Dirichlet da forma

$$\mathrm{Dir}(n_1 + \kappa_1, \cdots, n_G + \kappa_G).$$

Passo 2. Para cada $i \in \{i \in \{1 \cdots, n\}; r_i^y = 1\}$ gere Y_i^{m} a partir da distribuição $\pi(Y_i^{\mathrm{m}}|r_i^y, t_i, U_i, \mathbf{x}_i, z_{ij} = 1, \boldsymbol{\theta}_j)$ que é proporcional a

$$\exp\left\{-\frac{u_i}{2\tau_j}\left(y_i^{\mathrm{m}} - \mathbf{x}_i^\top\boldsymbol{\beta}_j - \Delta_j(t_i+b)\right)^2 + (\boldsymbol{\varphi}^y)^\top\mathbf{y}_i^* r_i^y - \log\left(1+\exp\left\{(\boldsymbol{\varphi}^y)^\top\mathbf{y}_i^*\right\}\right)\right\}.$$

Passo 3. Para cada $i = 1, \cdots, n$ gere $\mathbf{x}_i^{\mathrm{m}}$ de $\pi(\mathbf{x}_i^{\mathrm{m}}|\mathbf{r}_i^x, \mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\alpha})$ que é proporcional a

$$\exp\left\{(\boldsymbol{\varphi}^x)^\top\mathbf{x}_i^* \sum_{k=1}^{q} r_{ik}^x - p\log\left(1+\exp\left\{(\boldsymbol{\varphi}^x)^\top\mathbf{x}_i^*\right\}\right)\right\}\pi(\mathbf{x}_i^{\mathrm{m}}|\boldsymbol{\alpha}).$$

Passo 4. Pra cada $i = 1, \cdots, n$, gere Z_{ij} independentemente a partir da distribuição discreta a seguir,

$$p(Z_{ij} = 1|\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}_j, \boldsymbol{\varphi}^x, \boldsymbol{\varphi}^y) = \frac{p_j \pi_j(y_i, \mathbf{x}_i, r_i^y, \mathbf{r}_i^x|\boldsymbol{\theta}_j, \boldsymbol{\varphi}^x, \boldsymbol{\varphi}^y)}{\sum_{k=1}^{G} p_k \pi_k(y_i, \mathbf{x}_i, r_i^y, \mathbf{r}_i^x|\boldsymbol{\theta}_j, \boldsymbol{\varphi}^x, \boldsymbol{\varphi}^y)}, \quad j = 1, \cdots, G,$$

em que

$$\begin{aligned}\pi_j(y_i, \mathbf{x}_i, r_i^y, \mathbf{r}_i^x, z_{ij} = 1|\boldsymbol{\theta}_j, \boldsymbol{\varphi}^x, \boldsymbol{\varphi}^y) &= \exp\left\{(\boldsymbol{\varphi}^y)^\top \mathbf{y}_i^* r_i^y - \log\left(1 + \exp\left\{(\boldsymbol{\varphi}^y)^\top \mathbf{y}_i^*\right\}\right)\right\} \\ &\times \text{SMSN}(y_i|\mathbf{x}_i^\top \boldsymbol{\beta}_j + b\Delta_j, \Delta_j, \tau_j; h) \\ &\times \exp\left\{(\boldsymbol{\varphi}^x)^\top \mathbf{x}_i^* \sum_{k=1}^{q} r_{ik}^x - p \log\left(1 + \exp\left\{(\boldsymbol{\varphi}^x)^\top \mathbf{x}_i^*\right\}\right)\right\} \pi(\mathbf{x}_i^{\text{m}}|\boldsymbol{\alpha}).\end{aligned}$$

Passo 5. Gere $\boldsymbol{\phi}_j = (\boldsymbol{\beta}_j, \Delta_j)^\top$ a partir da distribuição $\pi(\boldsymbol{\phi}_j|\mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{t}, \mathbf{u}, \tau_j, \Delta_j, \nu)$, que representa uma normal multivariada $\text{N}_{p+1}(\boldsymbol{\mu}_{\boldsymbol{\phi}_j}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}_j})$ tal que

$$\begin{aligned}\boldsymbol{\Sigma}_{\phi_j} &= \left[\mathbf{H}_j^\top \boldsymbol{\Sigma}_{U_j} \mathbf{H}_j \tau_j^{-1} + \mathbf{B}_0^{-1}\right]^{-1} \\ \boldsymbol{\mu}_{\phi_j} &= \boldsymbol{\Sigma}_{\phi_j}\left[\mathbf{H}_j^\top \boldsymbol{\Sigma}_{U_j}(\mathbf{Y}_j - \Delta_j b)\tau_j^{-1} + \mathbf{B}_0^{-1}\mathbf{b}_0\right]\end{aligned}$$

sendo $\mathbf{H}_j = [\mathbf{X}_j \ \ \mathbf{T}_j]$, $\boldsymbol{\Sigma}_{U_j} = \text{diag}(\mathbf{U}_j)$, $\mathbf{X}_j$ uma matriz cujas linhas são $\mathbf{x}_i^\top$, com $i \in A_j$, $\mathbf{T}_j$, $\mathbf{U}_j$ e $\mathbf{Y}_j$ são vetores com elementos T_i, U_i e Y_i, $i \in A_j$, respectivamente.

Passo 6. Para cada $i = 1, \cdots, n$ gere de forma independente t_i a partir da distribuição $\pi(t_i|y_i, \mathbf{x}_i, u_i, \boldsymbol{\theta}_j, z_{ij} = 1)$, que representa uma distribuição normal truncada

$$\text{TN}\left(\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - b\Delta_j)\Delta_j}{\Delta_j^2 + \tau_j}, \frac{\tau_j}{u_i(\Delta_j^2 + \tau_j)}; (0, \infty)\right).$$

Passo 7. Gere τ_j a partir da distribuição $\pi(\tau_j|\mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{t}, \mathbf{u}, \boldsymbol{\beta}_j, \Delta_j, \nu, f_\tau)$, que representa uma gama inversa

$$\text{InvGamma}\left(\frac{n_j}{2} + e_\tau, f_\tau + \frac{1}{2}\sum_{i \in A_j} u_i(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta_j(t_i + b))^2\right),$$

em que o primeiro parâmetro é de forma e o segundo de taxa.

Passo 8. Gere f_τ a partir da distribuição $\pi(f_\tau|\boldsymbol{\tau})$, que representa uma $\text{Gamma}(e_\tau G + g; \sum_{j=1}^{G} \tau_j^{-1} + h)$.

Passo 9. Para cada $i = 1, \cdots, n$ gere de forma independente u_i a partir da distribuição $\pi(u_i|y_i, t_i, \boldsymbol{\theta}_j, z_{ij} = 1)$, que representa uma distribuição,

1. Para o modelo FMR-ST-CR

$$\text{Gamma}\left(\frac{\nu}{2}+1, \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta(t_i+b))^2 \tau_j^{-1} + t_i^2 + \nu}{2}\right).$$

2. Para o modelo FMR-SSL-CR

$$\text{TG}\left(1+\nu, \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta(t_i+b))^2 \tau_j^{-1} + t_i^2}{2}; (0,1)\right),$$

em que TG representa a a distribuição gamma truncada no intervalo (0,1).

3. Para o modelo FMR-SCN-CR uma distribuição discreta dicotômica assumindo η com probabilidade $p_2^*/p_1^* + p_2^*$ e 1 com probabilidade $1 - (p_2^*/p_1^* + p_2^*)$, em que

$$\begin{aligned} p_1^* &= \rho\eta \exp\left\{-\frac{\eta}{2}\left[\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta_j(t_i+b))^2}{\tau_j} + t_i^2\right]\right\} \\ p_2^* &= (1-\rho)\exp\left\{-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta_j(t_i+b))^2 + t_i^2\tau_j}{2\tau_j}\right\} \end{aligned}$$

Passo 10. Para a geração de ν iremos analisar os modelos FMR-ST-CR e FMR-SSL-CR individualmente,

1. Para o modelo FMR-SSL-CR gere ν de $\pi(\nu|\mathbf{u})$, que representa a distribuição

$$\text{Gamma}(\alpha_\nu + n, \gamma_\nu - \sum_{i=1}^{n} \log u_i).$$

2. No modelo FMR-ST-CR é impossivel conseguir uma distribuição condicional completa conjugada para *nu* logo utilizaremos um algoritmo de Metropolis-Hastings (MH).

   (i) Gere γ a partir da distribuição $\pi(\gamma|\nu)$ que representa $\text{TG}(2, \nu; (a_\gamma, b_\gamma))$, gama truncada no intervalo (a_γ, b_γ) com parâmetro de forma 2 e de taxa ν.

   (i) Gere ν, utilizando um passo de MH, com densidade

$$\pi(\nu|\gamma, \mathbf{y}, \mathbf{x}, \mathbf{z}, \beta_j, \Delta_j, \tau_j, \mathbf{p}) \propto e^{\gamma\nu}\pi(\mathbf{y}, \mathbf{x}, \mathbf{r}^x, \mathbf{r}^y|\boldsymbol{\Theta}), \tag{4.11}$$

em que $\pi(\mathbf{y}, \mathbf{x}, \mathbf{r}^x, \mathbf{r}^y|\boldsymbol{\Theta})$ é dado em (4.9) porém de forma reparametrizada. A ideia é a seguinte: dada a observação $\nu^{(l-1)}$ obtida na iteração $l-1$, gere

um candidato a nova observação ν^* utilizando como distribuição proposta $\text{LN}(\log \nu^{(l-1)}, \delta_\nu^2)$, que é uma logNormal com densidade,

$$\pi(\nu^*|\nu^{(l-1)}) = \frac{1}{\nu^* \delta_\nu (2\pi)^{1/2}} \exp\left(-\frac{(\log \nu^* - \log \nu^{(l-1)})^2}{2\delta_\nu^2}\right).$$

A nova observação ν^* é aceita com probabilidade

$$\min\left\{\frac{\pi(\nu^*|\cdots)\nu^*}{\pi(\nu^{(l-1)}|\cdots)\nu^{(l-1)}}, 1\right\},$$

em que $\pi(\nu^*|\cdots)$ é a densidade (4.11) trabalhada utilizando os valores atualizados dos parâmetros envolvidos. O hiperparâmetro conhecido δ_ν^2 será regulado tal que a taxa de aceitação dos graus de liberdade seja algo no intervalo (0,15; 0,3).

Passo 11. Gere $\boldsymbol{\varphi}^x$ a partir da distribuição $\pi(\boldsymbol{\varphi}^x|\mathbf{y}, \mathbf{x}, \mathbf{r}^x)$, que é proporcional a

$$\exp\left\{\sum_{i=1}^{n}\left[(\boldsymbol{\varphi}^x)^\top \mathbf{x}_i^* \sum_{k=1}^{q} r_{ik}^x - p\log\left(1+\exp\left\{(\boldsymbol{\varphi}^x)^\top \mathbf{x}_i^*\right\}\right)\right] - \frac{1}{2}\left(\boldsymbol{\varphi}^x - \boldsymbol{\mu}_{\varphi^x}^0\right)^\top (\boldsymbol{\Sigma}_{\varphi^x}^0)^{-1}\left(\boldsymbol{\varphi}^x - \boldsymbol{\mu}_{\varphi^x}^0\right)\right\}.$$

Passo 12. Gere $\boldsymbol{\varphi}^y$ a partir da distribuição $\pi(\boldsymbol{\varphi}^y|\mathbf{y}, \mathbf{x}, \mathbf{r}^y)$, que é proporcional a

$$\exp\left\{\sum_{i=1}^{n}\left[(\boldsymbol{\varphi}^y)^\top \mathbf{y}_i^* r_i^y - \log\left(1+\exp\left\{(\boldsymbol{\varphi}^y)^\top \mathbf{y}_i^*\right\}\right)\right] - \frac{1}{2}\left(\boldsymbol{\varphi}^y - \boldsymbol{\mu}_{\varphi^y}^0\right)^\top (\boldsymbol{\Sigma}_{\varphi^y}^0)^{-1}\left(\boldsymbol{\varphi}^y - \boldsymbol{\mu}_{\varphi^y}^0\right)\right\}.$$

O algoritmo acima é desenvolvido especialmente para o mecanismo MNAR, para adaptar o mesmo ao mecanismo MAR devemos relembrar que para esse mecanismo, podemos ignorar, na função de verossimilhança dada em (4.9), as distribuições dos indicadores de dados ausentes. De forma prática, essa adaptação consiste em excluir a parte referente aos valores ausentes, do passo 2 e 3 e também as condicionais completas de $\boldsymbol{\varphi}^y$ e $\boldsymbol{\varphi}^x$, passo 11 e 12.

As distribuições condicionais completas dadas nos passos 2, 3, 11 e 12 não são distribuições conjugadas, utilizaremos o algoritmo MH para simular observações a partir delas. Os detalhes na implementação do algoritmo MH para esses passos não são apresentados nesse capítulo por uma questão de economia de espaço. As distribuições de $\mathbf{x}_i$ podem ser facilmente escolhidas dependendo do problema em questão.

4.5 COMPARAÇÃO DE MODELOS

Quando estamos trabalhando o DIC para o modelo FMR-SMSN-MD em particular, existem alguns ajustes que necessariamente devem ser feitos para adaptá-lo as características

de misturas e de dados ausentes, principalmente porque a quantidade $\widehat{\boldsymbol{\Theta}}$ (média ou mediana a posteriori) não é um bom estimador no contexto de misturas.

Em Celeux *et al.* (2006) podemos observar uma grande quantidade de opções de DIC's para diferentes contextos estruturais. Em particular o DIC_4 se mostrou interessante no caso de modelos de misturas com dados faltantes pois consegue incorporar essas características em sua construção.

Inicialmente vamos definir um DIC para dados completos, definindo $\text{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Theta}|\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}\right]$ como o estimador para dados completos, que é um estimador robusto a problemas de identificabilidade uma vez que os componentes são identificados por $\mathbf{D}^{\text{m}}$, e então o DIC para o modelo completo é,

$$\text{DIC}(\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}) = -4\text{E}_{\boldsymbol{\Theta}}\left[\log\pi(\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}|\boldsymbol{\Theta})|\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}\right]+2\log\pi(\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}|\text{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Theta}|\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}\right]),$$

então,

$$\begin{aligned}
\text{DIC}_4 &= \text{E}_{\mathbf{D}^{\text{m}}}\left[\text{DIC}(\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}})|\mathbf{D}^{\text{o}}\right] \\
\text{DIC}_4 &= -4\text{E}_{\boldsymbol{\Theta},\mathbf{D}^{\text{m}}}\left[\log\pi(\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}|\boldsymbol{\Theta})|\mathbf{D}^{\text{o}}\right] \\
&\quad +2\text{E}_{\mathbf{D}^{\text{m}}}\left[\log\pi\left(\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}|\text{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Theta}|\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}}\right]\right)|\mathbf{D}^{\text{o}}\right] \\
&= -4\text{E}_{\boldsymbol{\Theta},\mathbf{D}^{\text{m}}}\left[\log\pi(\mathbf{D}|\boldsymbol{\Theta})|\mathbf{D}^{\text{o}}\right]+2\text{E}_{\mathbf{D}^{\text{m}}}\left[\log\pi\left(\mathbf{D}|\text{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Theta}|\mathbf{D}\right]\right)|\mathbf{D}^{\text{o}}\right],
\end{aligned} \tag{4.12}$$

em que $\log\pi(\mathbf{D}|\boldsymbol{\Theta})$ é a log-verossimilhança completa que pode ser escrita da seguinte forma

$$\begin{aligned}
&\log\pi(\mathbf{D}|\boldsymbol{\Theta}) = \\
&\log\prod_{j=1}^{G}\left\{\prod_{i=1}^{n}\left[\pi(y_i|\boldsymbol{\theta}_j,t_i,u_i,\mathbf{x}_i)\pi(t_i|u_i)f(u_i|\nu)\pi(\mathbf{x}_i|\boldsymbol{\alpha})\pi(r_i^y|y_i,\boldsymbol{\varphi}^y)\pi(\mathbf{r}_i^x|\mathbf{x}_i,\varphi^x)\right]^{z_{ij}}\right\}\pi(\mathbf{z}|\mathbf{p}) \\
&= c-\frac{1}{2}\sum_{j=1}^{G}n_j\log\tau_j+\sum_{i=1}^{n}\log h(u_i|\nu)-\frac{1}{2}\sum_{j=1}^{G}\sum_{i=1}^{n}\frac{z_{ij}u_i}{\tau_j}(y_i-\mathbf{x}_i^{\top}\boldsymbol{\beta}_j-\Delta_j(t_i-b))^2 \\
&\quad+\sum_{i=1}^{n}\pi(\mathbf{x}_i|\boldsymbol{\alpha})-\frac{1}{2}\sum_{i=1}^{n}u_it_i^2+\sum_{i=1}^{n}(\boldsymbol{\varphi}^y)^{\top}\mathbf{y}_i^*r_i^y-\sum_{i=1}^{n}\log\left(1+\exp\left\{(\boldsymbol{\varphi}^y)^{\top}\mathbf{y}_i^*\right\}\right) \\
&\quad+\sum_{i=1}^{n}(\boldsymbol{\varphi}^x)^{\top}\mathbf{x}_i^*\sum_{k=1}^{q}r_{ik}^x-p\sum_{i=1}^{n}\log\left(1+\exp\left\{(\boldsymbol{\varphi}^x)^{\top}\mathbf{x}_i^*\right\}\right)+\sum_{j=1}^{G}n_j\log p_j,
\end{aligned}$$

em que c é uma constante.

As esperanças do DIC_4 podem ser aproximadas utilizando a amostra MCMC, $\{(\mathbf{D}^{\text{m}(q*)},\boldsymbol{\Theta}^{(q*)});q*=1,\cdots,Q\}$. A primeira esperança de (4.12) pode ser aproximada por

$$\frac{1}{Q}\sum_{q=1}^{Q}\log\pi\left(\mathbf{D}^{\text{o}},\mathbf{D}^{\text{m}(q*)}|\boldsymbol{\Theta}^{(q)}\right).$$

Seja $\boldsymbol{\Theta}^{(q*,l)}$, $l = 1, \cdots, L$ observações geradas de $\pi\left(\boldsymbol{\Theta}|\mathbf{D}^{\mathrm{o}}, \mathbf{D}^{\mathrm{m}(q*)}\right)$ por meio do MCMC, então, temos a seguinte aproximação,

$$\mathrm{E}_{\boldsymbol{\Theta}}\left[\boldsymbol{\Theta}|\mathbf{D}^{\mathrm{o}}, D^{\mathrm{m(q*)}}\right] \approx \boldsymbol{\Theta}^{-(q*)} = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\Theta}^{(q*,l)},$$

ao final temos o DIC_4 aproximado dado por

$$\widehat{\mathrm{DIC}_4} = -\frac{4}{Q}\sum_{q*=1}^{Q}\log\pi\left(\mathbf{D}^{\mathrm{o}}, \mathbf{D}^{\mathrm{m(q)}}|\boldsymbol{\Theta}^{(q*)}\right) + \frac{2}{Q}\log\sum_{q*=1}^{Q}\pi\left(\mathbf{D}^{\mathrm{o}}, \mathbf{D}^{\mathrm{m}(q)}|\boldsymbol{\Theta}^{-(q*)}\right).$$

4.6 ESTUDOS DE SIMULAÇÃO

4.6.1 Estudo 1 - Recuperação de Parâmetros

Nessa seção estudaremos o desempenho do modelos FMR-SMSN-MD baseado nas estimações dos parâmetros. Com esse objetivo, utilizaremos as simulações de Monte Carlo para avaliar o desempenho das estimativas. Este estudo de simulação tem como objetivo de observar as mudanças nas estimativas quando alteramos o tamanho amostral, taxa de observações faltantes e diferentes formas de distribuição para a covariável em questão.

Todos os modelos FMR-SMSN-MD foram gerados com duas regressoras e com a seguinte configuração: $G = 2$ e $\mathbf{x}_i = (1, x_{i1}, x_{i2})^\top$ em que x_{i1} e x_{i2} seguem uma distribuição Unif$(2, 12)$, com $i = 1, \cdots, n$, os valores para os parâmetros utilizados foram $\boldsymbol{\beta}_1 = (\beta_{01}; \beta_{11}; \beta_{21})^\top = (25; 0,8; 0,5)^\top$, $\boldsymbol{\beta}_2 = (\beta_{02}; \beta_{12}; \beta_{22})^\top = (1,5; -1; -1)^\top$, $\boldsymbol{\sigma}^2 = (\sigma_1^2; \sigma_2^2) = (2; 2)^\top$ e $p_1 = 0,7$, para os modelos assimétricos $\boldsymbol{\lambda} = (\lambda_1; \lambda_2)^\top = (2; -2)^\top$ e graus de liberdade $\nu = 4$, além de $\rho = 0,7$ e $\eta = 0,3$ para os modelos FMR-CN-MD e FMR-SCN-MD.

Nesse estudo em particular, utilizaremos o mecanismo de dados faltantes não ignorável MNAR dado em (4.2) e (4.3) de tal forma a obtermos três taxas diferentes de dados faltantes, para tanto, a qual faremos $\boldsymbol{\varphi}^y = (a_y; b_y) = (-6,5; 0,1)^\top$, $\boldsymbol{\varphi}^x = (a_{x_1}; b_{x_1}; b_{x_2}) = (-6,5; 0,4; 0,4)^\top$ que gerará algo em torno de 10% de taxa de dados faltantes, $\boldsymbol{\varphi}^y = (-5,2; 0,1)^\top$, $\boldsymbol{\varphi}^x = (-5,2; 0,4;$ $0,4)^\top$ para algo próximo de 30% de dados faltantes e $\boldsymbol{\varphi}^y = (-4,4; 0,1)^\top$ e $\boldsymbol{\varphi}^x = (-4,4; 0,4; 0,4)^\top$ para algo em torno de 50% de faltantes. Para todos os estudos consideraremos que uma observação i é ausente se Y_i é ausente e/ou x_{ik} é ausente para algum $k = 1, \cdots, p$. Utilizaremos também diferentes tamanhos de amostra ($n = 100, 300$ e 500) e diferentes distribuições $\pi(\mathbf{x}_i|\boldsymbol{\alpha})$. Faremos de duas formas: (1) Utilizando a distribuição normal $\mathrm{N}_2(\overline{\mathbf{x}}^{\mathrm{o}}, \mathbf{S}_x^{2\mathrm{o}})$ em que $\overline{\mathbf{x}}^{\mathrm{o}}$ é um

vetor de médias em que cada coordenada corresponde a média amostral utilizando a parte observável de cada covariável e $\mathbf{S}_x^{2o}$ matriz diagonal na qual os elementos da diagonal são as variâncias amostrais extraídas dos valores observados das covariáveis. (2) Utilizando a distribuição $\text{Unif}(\min(X_1^o), \max(X_1^o))$ para x_{i1} e $\text{Unif}(\min(X_2^o), \max(X_2^o))$ para x_{i2}, tal que $\min(X_l^o)$ e $\max(X_l^o)$ representam o mínimo e o máximo da covariável x_{il} com $l = 1, 2$ respectivamente.

Para cada combinação de diferentes tamanhos amostrais, níveis de dados ausentes e distribuições da covariável faremos 100 repetições. Para a geração das amostras MCMC utilizamos o software JAGS, após a geração das amostras da posteriori aproximamos a estimação dos parâmetros utilizando a mediana amostral em que todas as cadeias terão tamanho 66 mil, retirando as primeiras 10 mil gerações, selecionando as gerações utilizando um espaçamento de 20 unidades e com valores iniciais iguais aos utilizados na seção 4.6.1.

Na Figura 9 temos os boxplots das 100 estimativas dos parâmetros do modelo FMR-ST-MD (os outros modelos não foram apresentados aqui por uma questão de espaço, porém os gráficos referentes a esse estudo para os demais modelos assimétricos se encontram no Apêndice A nas Figuras 36, 39 e 42), podemos observar que cada gráfico representa um parâmetro que é dividido em dois blocos, à esquerda os boxplots referentes ao tamanho amostral $n = 100$, no meio $n = 300$ e à direita $n = 500$. É fácil observar que tamanhos amostrais maiores geram variabilidades menores nas estimativas e algumas vezes as tornam mais próximas do valor real (linha pontilhada) , principalmente quando temos uma taxa de dados ausentes menor. Dentro de cada bloco, no eixo das ordenadas, temos a diferenciação por taxa de dados ausentes, é interessante percebermos que taxas menores geram estimativas mais próximas dos valores reais, porém isso não é tão evidente para tamanhos amostrais menores.

Outro comportamento que podemos notar na Figura 9 é referente a comparação das distribuições atribuídas à covariável, em que aparentemente a distribuição uniforme parece ter uma variabilidade um pouco menor quando temos $n = 100$, porém se levarmos em consideração apenas a mediana das estimativas não se pode optar por uma ou outra, utilizando como referência o verdadeiro valor do parâmetro. Já para $n = 300$ e 500, aparentemente, apenas em ν notamos a existência de uma pequena tendência de variabilidade menor e melhores estimativas para a distribuição normal, principalmente quando temos taxas dados ausentes menores. Embora encontremos alguns comportamentos, já citados anteriormente, concluímos que existe pouca diferença entre as distribuições escolhidas para a covariável em estudo, levando em consideração as estimativas desses parâmetros, mostrando que o nosso modelo nos fornece estimativas próximas dos valores verdadeiros e robustas a escolha dessas distribuições.

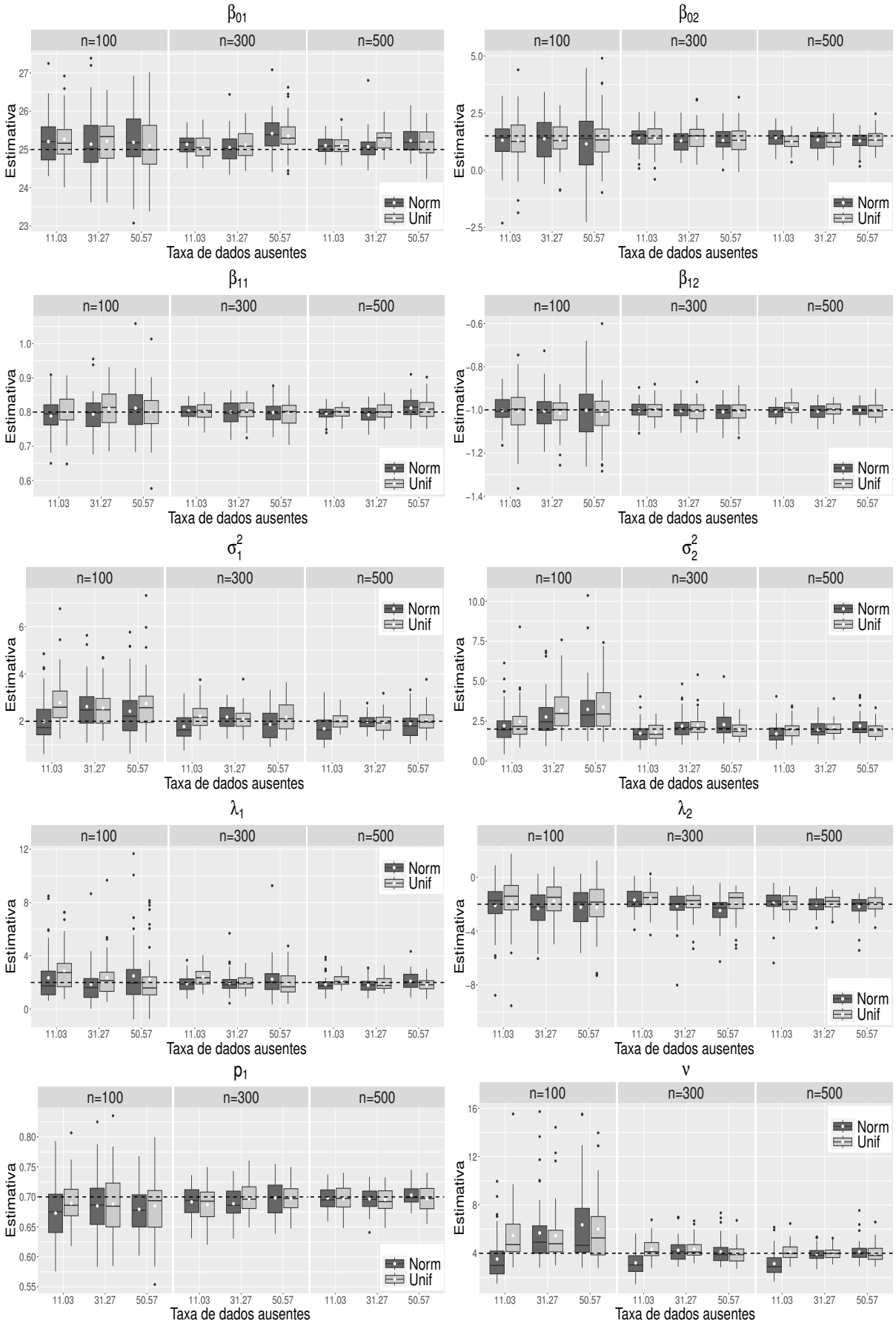

Figura 9 – Boxplots das estimativas dos parâmetros para o modelo FMR-ST-MD, comparando diferentes distribuições para a covariável (normal ou uniforme), diferentes tamanhos amostrais e diferentes taxas de dados ausentes. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

Nas Figuras 10 e 11 temos as 100 estimativas dos parâmetros referentes ao mecanismo de dados ausentes não ignorável para o modelo FMR-ST-MD, com diferentes distribuições da covariável e diferentes tamanhos amostrais. Nos referenciando pelo valor real dos parâmetros (linha pontilhada) é possível observar que a variabilidade das estimativas diminui e a proximidade para o valor verdadeiro aumenta, quando aumentamos o tamanho amostral. Com 31,27% de taxa média, Figura 10, verificamos um comportamento parecido para as diferentes distribuições da covariável. Já na Figura 11, onde temos uma taxa média de 50,57%, além de termos a melhoria das estimativas oriundas do aumento do tamanho amostral é interessante citarmos a melhora da proximidade das estimativas dos parâmetros para os valores reais, nos modelos comparando com a taxa de 31,27% . Quanto as distribuições das covariáveis os modelos que utilizaram a normal em sua maioria mostraram uma menor variabilidade e para a_{x_1}, b_{x_1} e b_{x_2} uma maior proximidade dos valores verdadeiros.

Em um computador com processador Intel Core i5-3210M, CPU 2.50GHz, 8 GB de memória RAM e sistema operacional de 64 bits, utilizando o software JAGS, o tempo computacional de cada réplica (de um total de 100) foi, em média, de 38 minutos para os modelos simétricos e 42 minutos para os modelos assimétricos, lembrando que cada replica conta com três diferentes tamanhos amostrais e dois diferentes taxas de dados faltantes. Para n=500, com uma taxa de dados faltantes fixa em 11,03%, o tempo computacional para os modelos assimétricos foi de 19,2 minutos e dos simétricos 18 minutos.

Além do estudo de recuperação dos parâmetros vistos anteriormente, o desempenho das estimativas dos parâmetros foi avaliada pelo viés relativo e MSE relativo comparando as estimativas dos coeficientes do modelo de regressão com os coeficientes verdadeiros, nas 100 repetições, da seguinte forma

$$\text{Vício Relativo}(\%) = \frac{1}{Q}\sum_{q=1}^{Q}\frac{\widehat{\beta}^{(q)}-\beta}{|\beta|}\times 100, \tag{4.13}$$

$$\text{MSE Relativo}(\%) = \frac{1}{Q}\sum_{q=1}^{Q}\frac{\left(\widehat{\beta}^{(q)}-\beta\right)^2}{\beta^2}\times 100, \tag{4.14}$$

tal que β e $\widehat{\beta}^{(q)}$ representam o valor verdadeiro de algum coeficiente de $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \cdots, \beta_{p1})^\top$ e a sua a q-ésima estimativa posteriori respectivamente.

As Tabelas 4 e 5 mostram o viés relativo e MSE relativo para os modelos FMR-SMSN-MD com 11,03% e 50,57% de taxa de observações faltantes no mecanismo MNAR, respectivamente. A partir dos resultados mostrados nessas tabelas, fazemos as seguintes pontua-

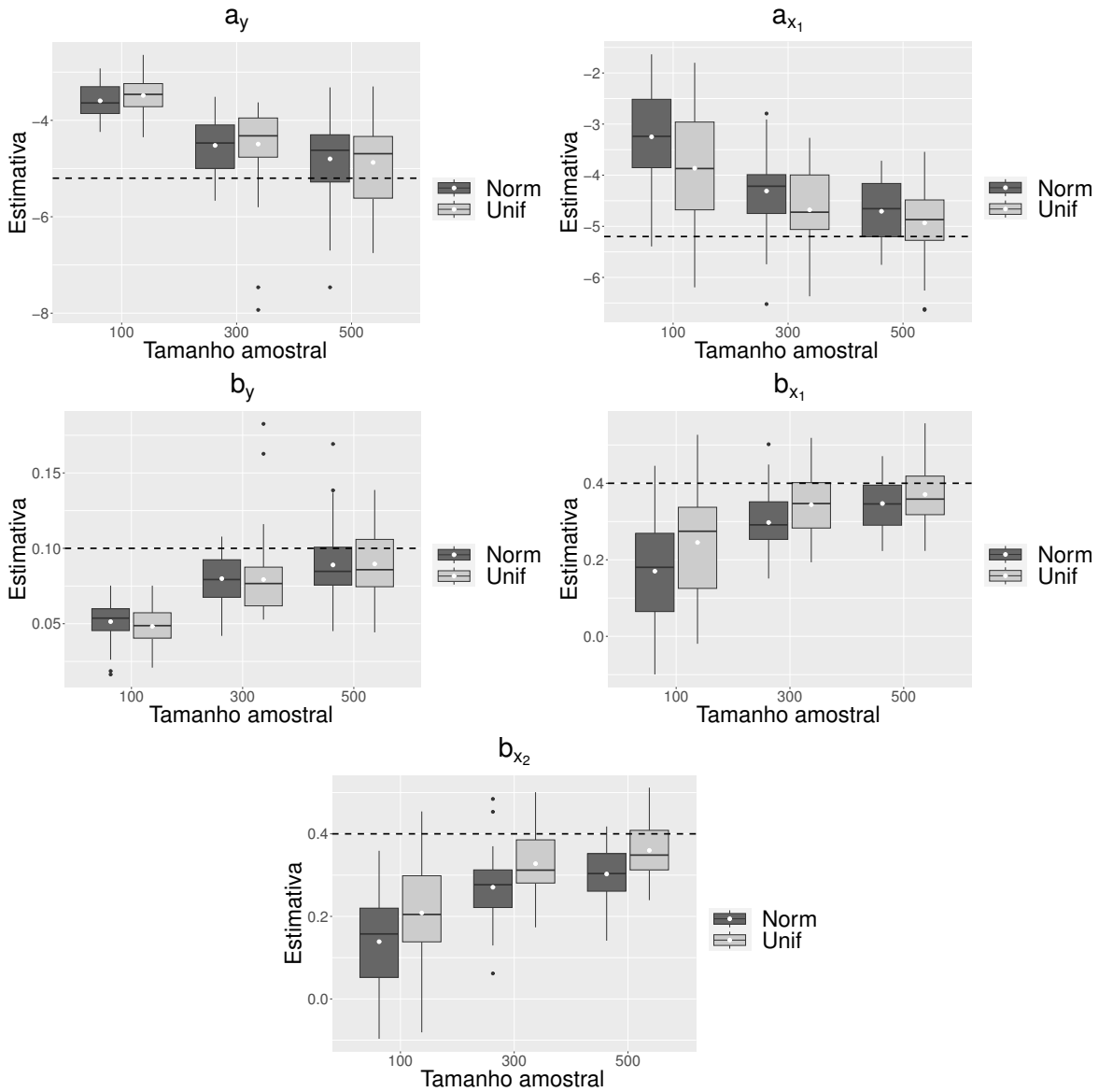

Figura 10 – Boxplots das estimativas dos parâmetros da regressão logística no modelo FMR-ST-MD, referente ao mecanismo MNAR, com taxa de dados ausentes de 31,27%, com diferentes distribuições para a covariável (normal ou uniforme) e tamanhos amostrais. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

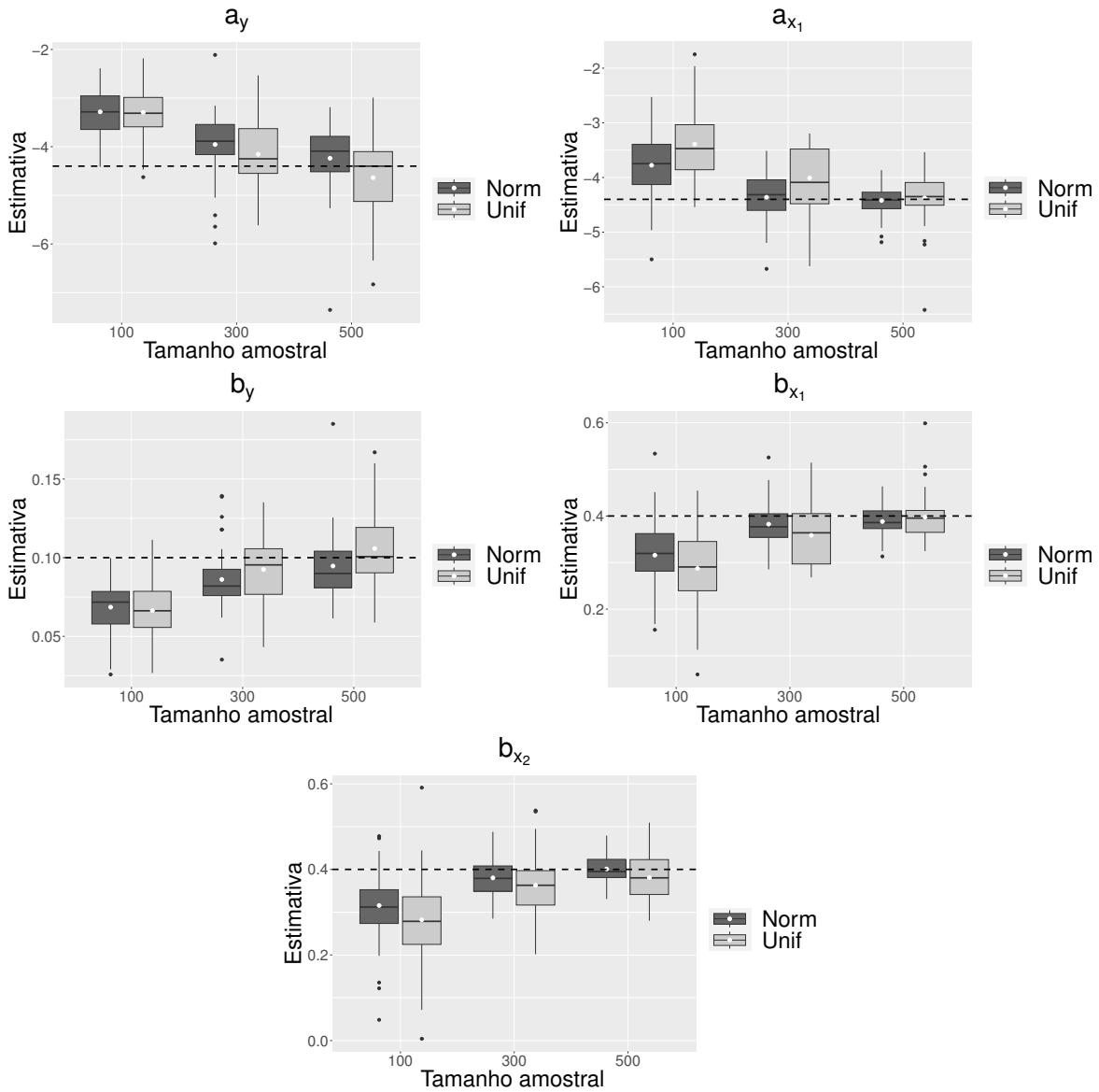

Figura 11 – Boxplots das estimativas dos parâmetros da regressão logística no modelo FMR-ST-MD, referente ao mecanismo MNAR, com taxa de dados ausentes de 50,07%, com diferentes distribuições para a covariável (normal ou uniforme) e tamanhos amostrais. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

ções:(i) Os erros tendem a diminuir com o aumento do tamanho amostral e a diminuição das taxas de dados ausentes em ambas as tabelas. (ii) Em sua maioria, os parâmetros dos modelos assimétricos parecem ter erros menores para 11,03% e 50,57% de taxa. (iii) O parâmetro β_{02} tem erros maiores que os demais parâmetros, principalmente para os modelos simétricos, inclusive para $n = 500$, em que tivemos um MSE, para covariável com distribuição normal, de 33,86% para uma taxa de 11,03% (modelo FMR-CN-MD) e 15,9% para uma taxa de 50,57% (modelo FMR-T-MD), enquanto nos modelos assimétricos, para n=500, esses erros são bem menores, apesar disso podemos observar que β_{02} diminui naturalmente os seus erros com o aumento do tamanho amostral assim como os demais parâmetros.

4.6.2 Estudo 2 - Flexibilidade do Modelo

Nessa seção, demonstraremos, assim como em na seção 3.5.2, a flexibilidade do modelo FMR-SMSN-MD, para acomodar dados oriundos de modelos de natureza diferente da família SMSN.

Usaremos o modelo 3.7 tal que a verossimilhança de $\mathbf{y}$ é dada por 4.9, a esse modelo daremos o nome de FMR-NIG-MD. Para o mecanismo de dados faltantes MNAR utilizaremos os conceitos vistos na seção 4.3.

Fizemos um estudo computacional em que trabalhamos um modelo de misturas de regressões com tamanho amostral $n = 100$ e $n = 500$ de um modelo FMR-NIG-MD com $G = 3$, $p_1 = p_2 = 0,3$, $\boldsymbol{\beta}_1 = (\beta_{01};\beta_{11};\beta_{21})^\top = (-5;1;1)$, $\boldsymbol{\beta}_2 = (\beta_{02};\beta_{12};\beta_{22})^\top = (2;2;1)$, $\boldsymbol{\beta}_3 = (\beta_{03};\beta_{13};\beta_{23})^\top = (18;4;1)$, $\boldsymbol{\lambda} = (-2;1;-2)$, $\boldsymbol{\delta} = (0,5;0,5;0.5)^\top$, $\boldsymbol{\gamma} = (1;1;1)^\top$, $x_{i1} \sim \mathrm{N}(2,6)$ e $x_{i2} \sim \mathrm{N}(2,10)$ para $i = 1,\cdots,n$. Denotaremos X_1 e X_2 as duas variáveis regressoras envolvidas no modelo.

Aqui o estudo se divide em três partes, trabalharemos as formas MCAR, MAR e MNAR para a seleção das observações faltantes, em particular os dados faltantes serão introduzidos apenas na variável resposta e na regressora X_1 enquanto a variável X_2 permanecerá totalmente observável. No contexto MCAR utilizaremos uma taxa de 30% de dados faltantes. Para os parâmetros referentes aos dados ausentes, no contexto MNAR, faremos $\boldsymbol{\varphi}^y = (-5;0,1)^\top$ e $\boldsymbol{\varphi}^x = (-5;0,5)^\top$, tal que $\boldsymbol{\varphi}^x$ é definido para o mecanismo de retirada na variável X_1, tais coeficientes geraram uma taxa de dados faltantes próxima de 30%. Já no mecanismo MAR, para uma taxa próxima de 30%, as observações de Y e X_1 serão consideradas ausentes dependendo dos valores observados em X_2, da seguinte forma: Seja q_d o d-ésimo quartil da distribuição empírica

		n=100		n=300		n=500	
Modelos	Par	Normal	Unif	Normal	Unif	Normal	Unif
FMR-N-MD	β_{01}	0,01 (0,06)	-0,13 (0,06)	-0,33 (0,02)	0,01 (0,02)	-0,16 (0,01)	0,02 (0,01)
	β_{02}	-3,66(45,14)	4,53(46,27)	-5,02(15,63)	-5,70 (6,76)	-5,62 (5,69)	0,82 (8,48)
	β_{11}	1,75 (0,71)	2,34 (0,52)	0,49 (0,21)	-0,21 (0,15)	0,66 (0,13)	-0,36 (0,09)
	β_{12}	-0,14 (0,71)	-3,01 (0,95)	1,33 (0,38)	0,92 (0,16)	0,51 (0,15)	-0,02 (0,24)
	β_{21}	-0,24 (1,24)	-0,56 (1,72)	1,54 (0,49)	0,55 (0,53)	0,76 (0,26)	0,78 (0,24)
	β_{22}	1,01 (0,93)	1,09 (0,76)	-0,48 (0,34)	0,13 (0,20)	0,08 (0,14)	-0,29 (0,19)
FMR-SN-MD	β_{01}	-0,52 (0,03)	-0,43 (0,03)	0,05 (0,01)	0,07 (0,01)	-0,11 (0,01)	0,06 (0,01)
	β_{02}	-7,04(21,89)	1,02(21,15)	2,61 (7,35)	0,44 (5,98)	0,64 (5,03)	2,54 (4,03)
	β_{11}	1,52 (0,28)	1,42 (0,29)	-0,05 (0,11)	-0,46 (0,12)	0,31 (0,04)	-0,50 (0,05)
	β_{12}	0,51 (0,58)	0,01 (0,42)	-0,14 (0,12)	-0,39 (0,16)	-0,34 (0,08)	-0,72 (0,10)
	β_{21}	1,77 (0,81)	1,49 (0,61)	-0,13 (0,23)	0,34 (0,23)	0,50 (0,18)	0,38 (0,13)
	β_{22}	-0,03 (0,54)	-0,37 (0,42)	-0,29 (0,17)	-0,01 (0,11)	-0,06 (0,10)	0,21 (0,07)
FMR-T-MD	β_{01}	0,81 (0,10)	0,40 (0,08)	0,07 (0,03)	0,19 (0,03)	-0,08 (0,02)	-0,13 (0,02)
	β_{02}	-8,39(73,25)	-10,7(78,37)	-5,66(16,36)	-9,04(17,87)	-3,74 (12,1)	5,72 (8,60)
	β_{11}	-0,14 (1,02)	1,68 (0,75)	0,08 (0,25)	-0,75 (0,15)	0,41 (0,13)	0,28 (0,14)
	β_{12}	0,80 (1,07)	2,16 (1,72)	0,16 (0,29)	0,08 (0,41)	-0,31 (0,21)	-0,74 (0,22)
	β_{21}	-2,60 (1,85)	-3,29 (3,02)	0,09 (0,76)	-0,82 (0,92)	-0,11 (0,57)	1,45 (0,32)
	β_{22}	0,59 (1,44)	-1,74 (1,15)	0,74 (0,50)	2,17 (0,52)	0,93 (0,27)	-0,84 (0,28)
FMR-ST-MD	β_{01}	1,07 (0,07)	0,83 (0,07)	0,21(0,01)	0,53 (0,02)	0,37 (0,01)	0,41 (0,01)
	β_{02}	-13,31(55,58)	-12,49(42,29)	-7,18(14,56)	-4,35 (11,91)	-16,24(8,52)	-4,71 (6,54)
	β_{11}	0,02 (0,40)	-1,41 (0,50)	0,31(0,10)	0,37 (0,07)	0,05 (0,05)	-0,44 (0,07)
	β_{12}	-0,50 (1,16)	-0,24 (0,50)	0,17(0,19)	-0,12 (0,17)	1,21 (0,13)	-0,79 (0,10)
	β_{21}	-0,75 (0,92)	0,64 (1,15)	1,47(0,37)	-1,48 (0,26)	0,21 (0,22)	0,25 (0,13)
	β_{22}	0,88 (0,58)	-0,29 (0,77)	-0,44(0,22)	-0,68 (0,24)	0,03 (0,11)	0,49 (0,08)
FMR-SL-MD	β_{01}	0,61 (0,09)	0,67 (0,10)	-0,53 (0,03)	-0,09 (0,02)	-0,12 (0,01)	0,06 (0,02)
	β_{02}	-5,99(43,48)	7,06 (85,49)	-0,47 (16,5)	5,07(13,54)	-4,12 (8,58)	-1,38 (7,42)
	β_{11}	-0,57 (0,81)	1,87 (0,81)	-0,40 (0,23)	-0,70 (0,25)	-0,07 (0,09)	-0,01 (0,15)
	β_{12}	0,30 (0,99)	-0,08 (1,66)	1,39 (0,40)	-0,96 (0,34)	0,70 (0,19)	-0,40 (0,19)
	β_{21}	-2,52 (1,91)	-4,98 (2,86)	4,38 (0,81)	2,74 (0,73)	1,01 (0,38)	-0,04 (0,43)
	β_{22}	-0,75 (1,04)	-3,06 (1,56)	-1,29 (0,40)	-0,60 (0,32)	-0,06 (0,17)	0,02 (0,24)
FMR-SSL-MD	β_{01}	0,03 (0,06)	0,45 (0,05)	-0,31 (0,02)	-0,64 (0,01)	-0,38 (0,01)	-0,30 (0,01)
	β_{02}	-1,36 (31,07)	15,44 (41,5)	4,98 (9,61)	-1,34 (7,28)	8,37 (7,11)	1,48 (3,88)
	β_{11}	0,10 (0,31)	-1,93 (0,52)	0,02 (0,14)	0,42 (0,15)	-0,05 (0,08)	-0,07 (0,11)
	β_{12}	1,71 (0,69)	-0,71 (0,84)	0,10 (0,22)	0,79 (0,17)	-0,31 (0,11)	0,23 (0,09)
	β_{21}	-1,75 (1,84)	-0,66 (1,10)	-0,45 (0,31)	1,34 (0,25)	0,39 (0,19)	-0,72 (0,17)
	β_{22}	-0,04 (0,51)	-2,15 (0,87)	-0,05 (0,18)	1,17 (0,20)	0,00 (0,10)	0,97 (0,10)
FMR-CN-MD	β_{01}	0,78 (0,22)	3,30 (0,39)	0,20 (0,07)	0,59 (0,06)	0,59 (0,03)	0,07(0,03)
	β_{02}	1,38(115,7)	-4,30(53,29)	7,21(32,67)	5,50(32,24)	-0,01(33,86)	-11,37(16,7)
	β_{11}	-0,54 (1,61)	-4,88 (2,25)	1,56 (0,57)	-0,22 (0,50)	-0,25 (0,28)	-0,31(0,25)
	β_{12}	0,74 (2,16)	-0,31 (1,95)	0,99 (0,67)	-2,38 (0,77)	-0,13 (0,47)	-0,16(0,49)
	β_{21}	0,23 (4,23)	-9,71 (6,48)	-1,91 (1,81)	-1,66 (1,55)	-2,93 (0,74)	1,2(0,83)
	β_{22}	-5,54 (3,72)	-0,48 (3,48)	-2,14 (0,55)	-0,20 (0,69)	-0,27 (0,54)	1,85(0,37)
FMR-SCN-MD	β_{01}	0,18 (0,08)	0,12 (0,11)	-0,11 (0,02)	0,33 (0,03)	-0,02 (0,02)	0,15 (0,02)
	β_{02}	5,84(69,72)	-12,47(77,02)	4,76(19,16)	-4,89(16,32)	-4,72 (9,91)	0,99 (8,46)
	β_{11}	-0,40 (0,55)	1,18 (1,16)	0,76 (0,26)	-0,33 (0,30)	0,42 (0,20)	-1,00 (0,17)
	β_{12}	0,45 (1,41)	-0,93 (1,30)	-1,23 (0,33)	1,09 (0,43)	0,87 (0,28)	-0,78 (0,14)
	β_{21}	-0,27 (2,33)	-2,15 (2,38)	0,10 (0,47)	-1,57 (0,47)	-0,08 (0,40)	-0,01 (0,32)
	β_{22}	-2,26 (1,56)	0,71 (1,82)	-0,10 (0,36)	-0,24 (0,29)	-0,63 (0,20)	0,35 (0,20)

Tabela 4 – Vício Relativo (MSE relativo) para todos os modelos FMR-SMSN-MD sob o mecanismo MNAR com taxa média de dados ausentes de 11,03%, para os parâmetros(Par) referentes aos coeficientes do modelo de regressão e os pesos da mistura, com diferentes distribuições para a covariável (normal ou uniforme) e diferentes tamanhos amostrais.

		n=100		n=300		n=500	
Modelos	Par	Normal	Unif	Normal	Unif	Normal	Unif
FMR-N-MD	β_{01}	-1,11 (0,12)	-0,16 (0,12)	-0,33 (0,03)	-0,29 (0,04)	-0,47 (0,02)	-0,30 (0,01)
	β_{02}	9,44(42,28)	-20,05(63,4)	-5,08(17,42)	2,87(22,58)	-7,97 (12,0)	1,28(10,98)
	β_{11}	5,01 (1,21)	2,42 (0,79)	0,28 (0,34)	3,43 (0,48)	1,20 (0,22)	2,51 (0,25)
	β_{12}	-1,67 (1,19)	0,70 (1,37)	-0,29 (0,32)	-1,71 (0,50)	1,60 (0,24)	-1,41 (0,26)
	β_{21}	2,88 (2,90)	3,67 (3,61)	3,64 (0,83)	2,13 (0,82)	2,03 (0,62)	3,89 (0,47)
	β_{22}	-2,74 (1,01)	-1,16 (1,29)	1,35 (0,46)	-2,70 (0,63)	-0,47 (0,18)	-2,10 (0,32)
FMR-SN-MD	β_{01}	-0,04 (0,04)	-0,09 (0,06)	-0,33 (0,02)	-0,26 (0,02)	-0,09 (0,01)	-0,14 (0,01)
	β_{02}	0,38 (34,4)	-9,99 (27,8)	1,28 (7,51)	8,30 (7,39)	-3,29 (7,60)	-0,81 (6,27)
	β_{11}	0,28 (0,42)	1,08 (0,40)	0,85 (0,12)	1,79 (0,25)	0,06 (0,08)	1,08 (0,14)
	β_{12}	-0,69 (0,90)	-0,78 (0,52)	-0,73 (0,17)	-1,43 (0,22)	0,14 (0,11)	-0,38 (0,13)
	β_{21}	1,94 (1,13)	2,57 (1,98)	1,56 (0,49)	2,36 (0,53)	0,61 (0,23)	2,87 (0,35)
	β_{22}	0,54 (0,82)	1,31 (0,73)	-0,04 (0,16)	-2,62 (0,31)	0,51 (0,13)	-1,39 (0,18)
FMR-T-MD	β_{01}	0,22 (0,16)	0,17 (0,16)	0,07 (0,04)	-0,24 (0,05)	-0,08 (0,02)	-0,27 (0,02)
	β_{02}	-29,18(85,49)	15,99(73,49)	-5,80 (16,7)	4,33(32,17)	-7,79 (15,9)	6,64(15,03)
	β_{11}	-1,25 (1,51)	2,51 (1,24)	0,25 (0,41)	3,17 (0,50)	0,89 (0,21)	2,58 (0,28)
	β_{12}	0,16 (1,78)	-4,20 (2,07)	0,25 (0,39)	-2,17 (0,66)	1,19 (0,34)	-2,18 (0,35)
	β_{21}	5,65 (4,86)	4,45 (4,73)	0,75 (1,18)	3,96 (1,24)	0,64 (0,63)	4,34 (0,69)
	β_{22}	3,59 (1,98)	-4,70 (1,99)	-0,12 (0,40)	-2,34 (0,72)	0,19 (0,33)	-2,73 (0,49)
FMR-ST-MD	β_{01}	0,39 (0,10)	0,74 (0,10)	1,40 (0,05)	1,67 (0,06)	0,89 (0,03)	0,92 (0,04)
	β_{02}	-0,32 (51,52)	-23 (103,68)	-13,48(17,25)	-13,11(15,06)	-10,55(8,83)	-14,28(8,53)
	β_{11}	0,03 (0,72)	1,44 (0,79)	-0,25 (0,21)	-0,18 (0,22)	1,15 (0,17)	1,46 (0,19)
	β_{12}	-1,97 (1,28)	-0,20 (1,88)	-0,64 (0,29)	-0,95 (0,27)	-0,40 (0,14)	-0,08 (0,12)
	β_{21}	3,17 (1,95)	2,90 (2,06)	-0,10 (0,61)	1,12 (0,48)	-0,48 (0,42)	2,96 (0,60)
	β_{22}	-1,07 (0,76)	-1,76 (1,73)	0,07 (0,33)	-2,41 (0,31)	0,13 (0,14)	-1,46 (0,14)
FMR-SL-MD	β_{01}	-0,49 (0,14)	0,27 (0,15)	0,03 (0,03)	-0,07 (0,05)	-0,18 (0,02)	-0,14 (0,04)
	β_{02}	6,65(107,72)	-5,71(63,96)	2,46(16,64)	5,09(29,71)	1,66(12,95)	4,38(13,51)
	β_{11}	4,29 (2,11)	2,74 (1,27)	1,76 (0,37)	4,05 (0,74)	-0,35 (0,27)	2,83 (0,45)
	β_{12}	-1,55 (1,95)	-4,29 (2,00)	0,99 (0,49)	-2,50 (0,55)	0,00 (0,31)	-2,44 (0,41)
	β_{21}	1,66 (2,96)	2,99 (4,75)	-1,37 (1,03)	1,94 (0,76)	2,08 (0,63)	2,91 (0,95)
	β_{22}	-2,97 (1,94)	-0,79 (1,61)	-1,37 (0,37)	-2,29 (0,68)	-0,71 (0,35)	-1,60 (0,37)
FMR-SSL-MD	β_{01}	-0,69 (0,07)	0,05 (0,07)	-0,43 (0,02)	0,09 (0,03)	-0,42 (0,01)	0,07 (0,01)
	β_{02}	4,20 (38,4)	-1,73(53,45)	5,24 (13,16)	1,82 (10,43)	12,78 (7,08)	-5,13 (6,52)
	β_{11}	0,44 (0,74)	3,87 (0,78)	0,27 (0,24)	2,37 (0,20)	-0,29 (0,12)	2,07 (0,18)
	β_{12}	0,42 (0,81)	-1,78 (1,20)	-0,29 (0,28)	-2,05 (0,36)	-0,96 (0,18)	-0,87 (0,18)
	β_{21}	3,68 (1,58)	1,83 (2,30)	0,11 (0,56)	0,79 (0,61)	0,95 (0,30)	1,18 (0,36)
	β_{22}	-2,75 (0,96)	-3,15 (1,30)	-0,21 (0,31)	-1,58 (0,41)	-0,48 (0,13)	-0,36 (0,16)
	β_{01}	0,65 (0,17)	0,68 (0,23)	-0,14 (0,07)	0,21 (0,09)	-0,29 (0,06)	-0,04 (0,06)
	β_{02}	-17,27(209,5)	-23,23(169,7)	-6,70(34,13)	-9,46(33,34)	1,04(24,84)	-5,35(27,62)
	β_{11}	1,68 (2,05)	2,60 (2,76)	0,58 (0,58)	2,80 (0,89)	0,75 (0,57)	3,50 (0,68)
	β_{12}	2,65 (4,57)	-1,82 (3,38)	0,42 (0,92)	-0,57 (1,00)	-0,61 (0,58)	-2,16 (0,71)
	β_{21}	-3,84 (6,47)	2,00 (5,17)	2,48 (1,99)	5,02 (2,97)	3,38 (1,83)	4,28 (1,82)
	β_{22}	-1,55 (4,71)	-0,30 (2,85)	-0,36 (0,70)	-3,24 (0,92)	-0,02 (0,50)	-1,82 (0,56)
FMR-SCN-MD	β_{01}	-0,71 (0,11)	1,42 (0,15)	-0,37 (0,05)	0,65 (0,05)	-0,08 (0,02)	0,32 (0,02)
	β_{02}	-4,44(86,14)	-16,00(84,0)	6,18(15,91)	5,43(19,25)	5,93(14,34)	-9,22(16,54)
	β_{11}	3,31 (1,48)	1,36 (1,60)	1,16 (0,36)	-0,68 (0,34)	0,54 (0,17)	1,98 (0,29)
	β_{12}	-1,43 (1,78)	-2,68 (1,51)	-2,10 (0,59)	-1,72 (0,54)	-1,24 (0,29)	-0,50 (0,32)
	β_{21}	2,36 (2,84)	-4,93 (3,28)	2,49 (1,20)	2,86 (1,38)	1,13 (0,56)	1,17 (0,66)
	β_{22}	-1,80 (1,76)	0,37 (1,70)	-0,09 (0,60)	-3,42 (0,62)	-1,05 (0,24)	-0,78 (0,34)

Tabela 5 – Vício Relativo (MSE relativo) para todos os modelos FMR-SMSN-MD sob o mecanismo MNAR com taxa média de dados ausentes de 50,07%, para os parâmetros(Par) referentes aos coeficientes do modelo de regressão e os pesos da mistura, com diferentes distribuições para a covariável (normal ou uniforme) e diferentes tamanhos amostrais.

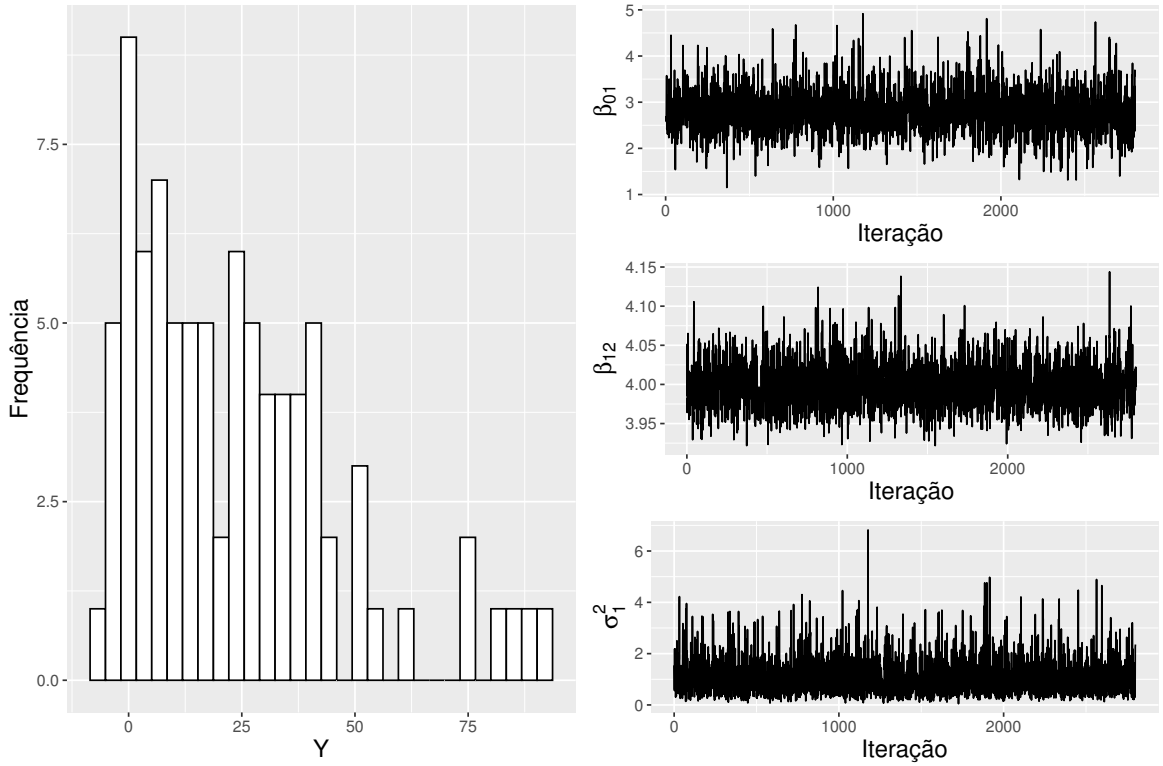

Figura 12 – Histograma da variável responta com n=100 e traceplots para β_{01}, β_{12} e σ_1^2 do modelo FMR-NIG-MD com o método MCAR.

de X_2 e definindo $\mathbf{V} = (V_1, V_2, V_3, V_4)^\top = (10\%, 7\%, 5\%, 0\%)^\top$, as probabilidades de se ter uma observação ausente para Y e X_1 são de V_1 se $X_2 < q_1$, V_2 se $X_2 \in [q_1, q_2)$, V_3 se $X_2 \in [q_2, q_3]$ e V_4 se $X_2 > q_3$. Para ambos os mecanismos utilizaremos a distribuição uniforme, vista na seção 4.6.1, como distribuição da covariável. A ideia particular dessa estrutura para o mecanismo MAR foi retirada de Lim, Narisetty e Cheon (2017).

As especificações quanto a geração da amostra MCMC e o algoritmo utilizado foram os mesmos utilizados na seção 4.6.1. Para esses conjuntos de dados foram ajustados todos os modelos da família SMSN, com três componentes. Após esses ajuste calculamos os critérios de seleção DIC.

A Figura 12 mostra o histograma de Y em que podemos notar uma clara ideia de multimodalidade que é um apelo natural para o uso de um modelo de misturas, podemos observar também, a direita da figura, os gráficos de traço para as amostras da distribuição a posteriori de alguns parâmetros, em que vemos que todos que foram apresentados percorrem de maneira homogênea os seus suportes.

Na Tabela 6, é notório que modelos que levam em conta assimetria, caudas pesadas ao mesmo tempo, como ST e SSL, superam os demais modelos que não comportam essas características. E, esse cenário é inerente ao tamanho amostral.

Tabela 6 – DIC's dos ajustes dos modelos FMR-SMSN-MD para os dados gerados sob o modelo FMR-NIG-MD.

	Sample size					
	n=100			n=500		
Modelos	MNAR	MCAR	MAR	MNAR	MCAR	MAR
FMR-N-MD	605,8257	597,0522	670,0832	3217,2822	2999,2833	2988,7589
FMR-T-MD	580,4208	579,6867	586,8371	2802,2825	2700,1756	2726,1225
FMR-SL-MD	573,0127	538,3986	624,8345	2725,2839	2911,7253	2975,7588
FMR-SN-MD	553,5727	592,4875	589,0938	2775,1472	2779,1458	2686,1422
FMR-ST-MD	549,8192	570,6949	571,2838	2657,2536	2668,6428	2543,7588
FMR-SSL-MD	**536,1833**	**516,8745**	**527,1758**	**2178,8596**	**2477,1425**	**2475,1425**

4.6.3 Estudo 3 - Clonagem de Dados e Identificabilidade

Nessa seção faremos um estudo similar ao visto na seção 3.5.3, nesse utilizaremos o modelo FMR-SMSN-MD.

A Figura 13 mostra os gráficos de λ_k^S em função de K para os modelos assimétricos quando $g(\boldsymbol{\Theta}) = \boldsymbol{\Theta}$, em que $\mathbf{y}$ é uma amostra artificial gerada do modelo FMR-SMSN-MD através dos mecanismos MCAR e MNAR com tamanho amostral $n = 100$, $G = 2$, $p_1 = 0,7$, $\boldsymbol{\beta}_1 = (\beta_{01};\beta_{11})^\top = (25;0,8)$, $\boldsymbol{\beta}_2 = (\beta_{02};\beta_{12})^\top = (2,5;-1)$, $\boldsymbol{\sigma}^2 = (2;2)^\top$, $\boldsymbol{\lambda} = (2;-2)$, $\nu = 2$ e $x_{i1} \sim \mathrm{N}(0,1)$, para $i = 1,\cdots,n$. Para o mecanismo MNAR utilizaremos $\boldsymbol{\varphi}^y = (-5;0,17)^\top$ e $\boldsymbol{\varphi}^x = (-2;-0,01)^\top$, gerando uma quantidade de dados ausentes perto de 25%, para o mecanismo MCAR também utilizaremos uma taxa de dados ausentes de 25%. A distribuição utilizada para modelar a covariável, em todos os mecanismos, as especificações da cadeia e o algoritmo de geração da amostra MCMC foram as mesmas utilizadas na seção 4.6.2. Assim como no capítulo anterior, o procedimento de Clonagem de Dados foi realizado utilizando o pacote do R chamado dclone (Sólymos (2010)).

Na Figura 13 os gráficos (a), (b) e (c) se referem ao método MNAR e (d), (e) e (f) ao MCAR, todos eles, independentes do tipo de mecanismo, sugerem fortes evidências de identificabilidade em todos os casos considerados.

4.6.4 Estudo 4 - Influência dos Dados com Observações Atípicas

Repetiremos aqui, o estudo de influência feito na seção 3.5.4, utilizaremos a mesma ideia e conceitos porém no contexto do modelo FMR-SMSN-MD. Chamaremos de I o conjunto

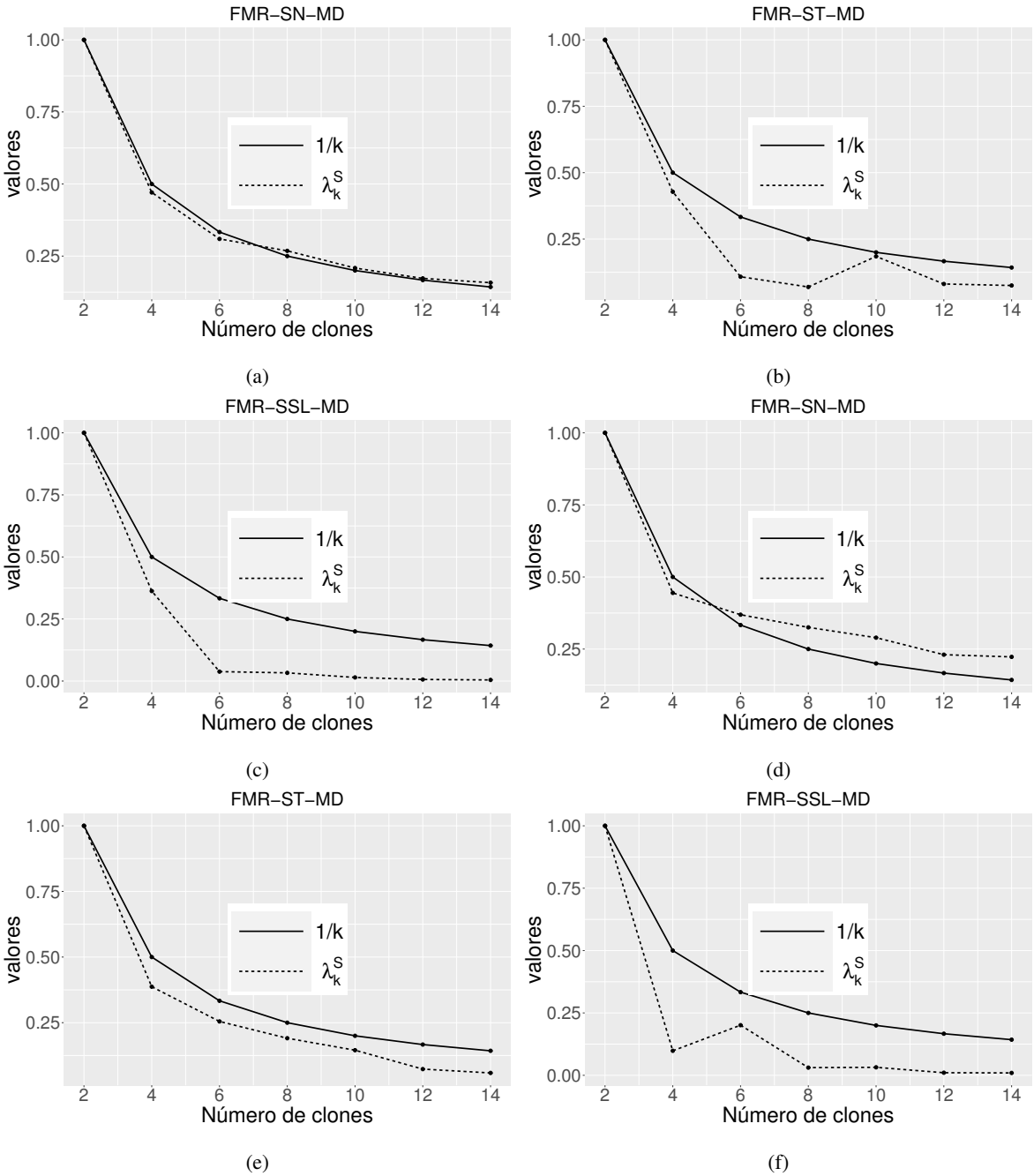

Figura 13 – Estudo de Simulação 3. Checando a identificabilidade utilizando Clonagem de dados. (a), (b) e (c) método MNAR e (d), (e) e (f) método MCAR

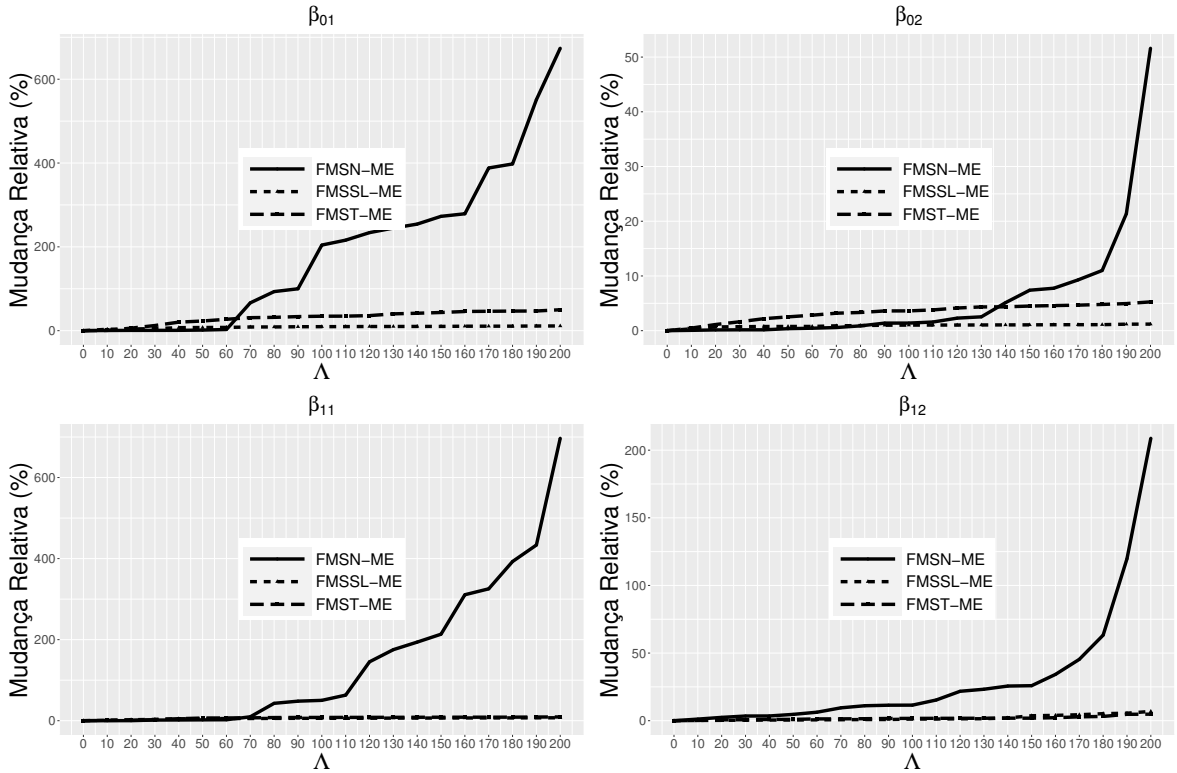

Figura 14 – RC (em %) para $\beta_{01}, \beta_{11}, \beta_{02}$ e β_{12} para os modelos FMR-SN-MD, FMR-ST-MD e FMR-SSL-MD ajustados sob o mecanismo MNAR com diferentes níveis de pertubação Λ.

de pontos escolhidos para serem contaminados, são eles: #78 ($y_{78} = 26,848$); #94 ($y_{94} = 28,082$); #35 ($y_{35} = 27,768$); #11 ($y_{11} = 29,166$) e #29 ($y_{29} = 28,266$) dados esses gerados a partir do modelo FMR-SN-CR com mecanismo MNAR para uma taxa de dados faltantes próxima de 20% com $n = 100$, $G = 2$, $p_1 = 0,7$, $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11})^\top = (2,6)$, $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12})^\top = (0,5;1)$, $\boldsymbol{\lambda} = (2,-2)$, $\boldsymbol{\sigma}^2 = (0,1;0,5)^\top$, $\boldsymbol{\varphi}^y = \boldsymbol{\varphi}^x = (-4;0,1)^\top$ e $x_{i1} \sim \text{Unif}(1,6)$ para $i = 1, \cdots, n$. O procedimento de ajuste dos modelos foi feito utilizando o software JAGS. A distribuição da covariável e as especificações quanto a geração da cadeia MCMC foram as mesmas utilizadas na seção 4.6.2.

Na Figura 14 temos os efeitos da perturbação dos dados para os coeficientes do modelo de regressão. Podemos observar que os parâmetros β_{01} e β_{11}, para o modelo FMR-SN-MD, com uma perturbação de 12% já atingiram uma variação relativa maior que 100% se mantendo sempre crescente até o final, enquanto os parâmetros β_{02} e β_{12} tiveram um crescimento mais lento e gradativo, para o modelo FMR-SN-MD, atingindo ate 200% de variação relativa em β_{12}. Os modelos FMR-ST-MD e FMR-SSL-MD se mantiveram sempre com pouca variação mesmo com o aumento significativo das pertubações.

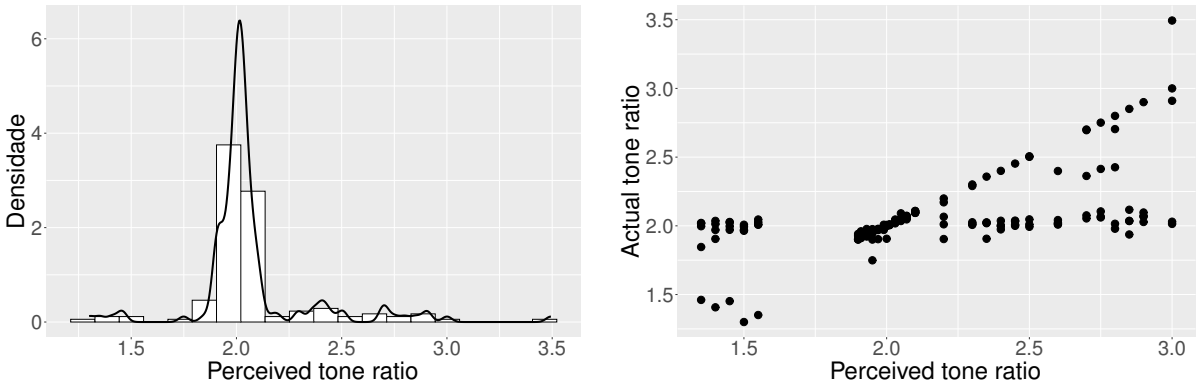

Figura 15 – Histograma da variável *Perceived tone ratio* e gráfico de dispersão das variáveis *Perceived tone ratio* e *Acttual tone ratio* para o banco de dados Tons

4.7 DADOS REAIS

Ilustramos nossos métodos propostos por meio de um conjunto de dados obtido de Cohen (1984), os quais representam a percepção de tons musicais, por músicos. Chamaremos esses banco de dados de Tons. Neste experimento de percepção de tom, um tom fundamental puro, com sobretons gerados eletronicamente adicionados, foi tocado para um músico treinado. Foi pedido ao músico que afinasse um tom ajustável para uma oitava acima do tom fundamental, foram registradas 150 tentativas do mesmo músico. Os sobretons foram determinados por uma razão de alongamento, que é a relação entre o tom ajustado e o tom fundamental, ao final tivemos o registro da relação de tom percebida(Y) e a relação de tom real(X). Duas tendências distintas emergem claramente (veja a Figura 15) que se relaciona com duas hipóteses exploradas em Cohen (1984), chamadas de hipótese de memória intervalar e hipótese de correspondência parcial. Muitos artigos analisaram esse conjunto de dados utilizando uma mistura de modelos de regressões lineares, Hunter e Young (2012), Viele e Tong (2002), Veaux (1989). Yao, Wei e Yu (2014) analisaram este conjunto de dados propondo um modelo de regressão de mistura robusta utilizando a distribuição t de Student e Zeller, Cabral e Lachos (2016) estendeu esse estudo para a família SMSN. A Figura 15 apresenta um histograma dos dados e mostra que há um padrão aparentemente não simetrico e com caudas pesadas para a variável *Perceived tone ratio*.

O banco de dados Tones não contém dados ausentes os mesmos serão criados artificialmente. Geraremos os índices que indicarão quais dados serão considerados como faltantes, por meio dos mecanismos MCAR e MNAR, com $\boldsymbol{\varphi}^y = \boldsymbol{\varphi}^x = (-5; 1,5)^\top$ para uma taxa próxima de 20%.

O ajuste foi feito pelos modelos FMR-SMSN-MD. Utilizando o software JAGS para a geração das amostras MCMC, com tamanho 66 mil, retirando as primeiras 10 mil gerações,

selecionando as gerações utilizando um espaçamento de 20 unidades e valores iniciais iguais aos utilizados na seção 4.6.1. Quanto a distribuição da covariável, utilizaremos a distribuição uniforme vista na seção 4.6.1.

Em especial para os modelos FMR-SCN-RC e FMR-CN-RC construiremos um grid, ajustaremos os modelos FMR-SCN-RC e FMR-CN-RC e calcularemos os DIC's para as varias combinações de ρ e η, assim como feito na seção 3.6 porém utilizando os dois mecanismos MCAR e MNAR tal que: (1) Sob o mecanismo MCAR e modelo FMR-CN-RC o melhor ajuste ocorreu com a utilização do par de parâmetros $\rho = 0,1109$ e $\eta = 0,001$ por meio dos quais obtivemos um DIC $= -338,873$ e para o modelo FMR-SCN-RC os parâmetros $\rho = 0,1109$ e $\eta = 0,2208$ foram os que retornaram o menor DIC $= -348,900$. (2) Sob o mecanismo MNAR para o modelo FMR-CN-RC o melhor ajuste ocorreu com a utilização do par de parâmetros $\rho = 0,0010$ e $\eta = 0,1109$ com DIC $= -358,371$ e para o modelo FMR-SCN-RC os parâmetros $\rho = 0,1109$ e $\eta = 0,2208$ foram os que retornaram o menor DIC $= -360,744$.

Na Tabela 7, temos as estimativas dos parâmetros de todos os modelos FRM-SMSN-MD para os mecanismos MNAR e MCAR. As estimativas dos coeficientes de regressão se mostraram bem parecidas, principalmente para β_{01} e β_{12}. Nos modelos simétricos, na maioria das vezes, tivemos pouca variabilidade nas estimativas, com exceção do modelo FMR-SL-MD sob o mecanismo MCAR. Para os modelos assimétricos comumente tivemos baixa variabilidade, exceto para os parâmetros de forma. É interessante notarmos que o parâmetro β_{02} é sempre muito próximo de zero, porém podemos ver nas figuras 16 e 16 que apenas nos modelos SN e ST a posteriori estimada de β_{02} têm sua moda no zero, nessa mesma figura podemos observar os gráficos de traços para a amostra a posteriori dos coeficientes do modelo de regressão, de um modo geral os gráficos de traços percorrem de forma homogênea seus respectivos suportes.

Os pesos e os graus de liberdade foram sempre muito parecidos para todos os modelos trabalhados.

Para compararmos os ajustes dos modelos utilizaremos o critério DIC descrito na seção 4.5, os modelos como os menores DICs serão considerados melhores. Na Tabela 8, podemos observar os critérios de seleção, tal que os menores DICs foram destinado para o modelo FMR-ST-MD, sob o mecanismo MNAR, e FMR-SSL-MD para o mecanismo MCAR. É importante salientar a superioridade dos modelos assimétricos sob os modelos simétricos no que diz respeito a comparação dos ajustes via critérios de seleção, em que essa superioridade foi robusta a escolha do mecanismo de dados faltantes.

Para essa ultima análise faremos algo como um processo de recuperação de parâme-

	Modelos							
	FMR-N-MD		FMR-T-MD		FMR-SL-MD		FMR-CN-MD	
	MNAR	MCAR	MNAR	MCAR	MNAR	MCAR	MNAR	MCAR
Par	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)
β_{01}	1,919(0,027)	1,920(0,027)	1,964(0,022)	1,953(0,028)	1,958(0,022)	1,923(0,032)	1,903(0,027)	1,924(0,259)
β_{02}	-0,037(0,152)	-0,059(0,115)	0,002(0,006)	0,002(0,005)	0,002(0,006)	-0,382(1,089)	-0,099(0,061)	0,003(0,006)
β_{11}	0,041(0,012)	0,042(0,012)	0,024(0,010)	0,028(0,012)	0,027(0,010)	0,042(0,014)	0,047(0,012)	0,038(0,113)
β_{12}	1,002(0,066)	1,005(0,051)	1,000(0,003)	0,999(0,002)	1,000(0,003)	-1,764(9,302)	1,035(0,028)	0,998(0,002)
σ_1^2	0,002(0,001)	0,002(0,001)	0,001(0,001)	0,001(0,001)	0,001(0,001)	0,001(0,001)	0,001(0,000)	0,001(0,000)
σ_2^2	0,027(0,010)	0,012(0,004)	0,001(0,001)	0,001(0,001)	0,001(0,001)	5,196(2,376)	0,003(0,001)	0,002(0,000)
p_1	0,706(0,048)	0,692(0,053)	0,606(0,050)	0,659(0,051)	0,609(0,050)	0,767(0,016)	0,688(0,051)	0,627(0,048)
ν	-	-	1,097(0,192)	1,499(0,351)	0,507(0,082)	0,463(0,071)	-	-
	FMR-SN-MD		FMR-ST-MD		FMR-SSL-MD		FMR-SCN-MD	
	MNAR	MCAR	MNAR	MCAR	MNAR	MCAR	MNAR	MCAR
Par	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)	Média(Dp)
β_{01}	1,921(0,025)	1,926(0,026)	1,928(0,044)	1,931(0,036)	1,919(0,000)	1,918(0,030)	1,914(0,025)	1,767(0,047)
β_{02}	-0,028(0,167)	-0,024(0,056)	-0,008(0,015)	0,001(0,007)	0,001(0,007)	0,004(0,010)	-0,034(0,027)	-0,387(0,081)
β_{11}	0,039(0,011)	0,040(0,011)	0,026(0,010)	0,030(0,012)	0,041(0,014)	0,040(0,013)	0,042(0,011)	0,038(0,010)
β_{12}	0,999(0,071)	0,981(0,024)	0,998(0,003)	0,998(0,002)	0,999(0,003)	0,998(0,006)	0,999(0,013)	0,988(0,020)
σ_1^2	0,005(0,001)	0,003(0,001)	0,001(0,001)	0,001(0,001)	0,001(0,001)	0,002(0,001)	0,003(0,001)	0,003(0,001)
σ_2^2	0,043(0,026)	0,017(0,007)	0,001(0,001)	0,001(0,001)	0,002(0,001)	0,001(0,001)	0,004(0,001)	0,016(0,005)
λ_1	-1,906(1,436)	-0,898(1,399)	-0,385(0,463)	-0,569(0,616)	0,001(0,603)	-0,729(0,767)	-2,047(1,782)	-2,410(0,511)
λ_2	0,623(2,701)	-13,551(12,881)	-3,542(2,510)	-1,473(3,264)	-7,300(8,554)	-1,243(7,137)	-2,690(8,274)	-2,740(0,699)
p_1	0,699(0,048)	0,652(0,054)	0,612(0,048)	0,652(0,051)	0,636(0,046)	0,670(0,049)	0,639(0,049)	0,663(0,005)
ν	-	-	1,280(0,155)	1,620(0,351)	1,451(0,050)	1,523(0,301)	-	-

Tabela 7 – Estimativas dos parâmetros (Mediana(DP)) para diferentes mecanismos de retirada de observações nos dados Tons.

	Mecanismos	
Modelos	MNAR	MCAR
FMR-N-MD	-238,2456	-276,8215
FMR-T-MD	-357,2960	-364,0360
FMR-SL-MD	-353,6661	-365,5135
FMR-CN-MD	-358,3710	-338,8730
FMR-SN-MD	-240,0954	-320,0241
FMR-ST-MD	**-363,7630**	-373,0636
FMR-SSL-MD	-363,7010	**-417,9336**
FMR-SCN-MD	-360,7440	-348,9898

Tabela 8 – DIC para modelos FMR-SMSN-MD sob os mecanismos MCAR e MNAR, dados Tons.

tros, em que consideraremos os parâmetros estimados nos modelos sem dados ausentes (FMR-SMSN), como uma espécie de valor verdadeiro, que serão utilizados como referência para a comparação com os modelos FMR-SMSN-MD, para essa comparação utilizaremos o mecanismo MNAR com $\boldsymbol{\varphi}^y = \boldsymbol{\varphi}^x = (-5;1,5)^\top$ para uma taxa próxima de 10% e $\boldsymbol{\varphi}^y = \boldsymbol{\varphi}^x = (-6;1,5)^\top$ para algo perto de 30%, e utilizaremos o mecanismo MCAR também com taxas de 30% e 10% para dados faltantes.

As especificações da amostra MCMC serão as mesmas utilizadas na primeira parte

dessa seção. Repetiremos a cadeia 100 vezes para cada conjunto de taxa de dados ausentes, tamanho amostral e mecanismo de dados ausentes. A comparação entre os modelos com e sem dados ausentes ficará por conta da medida de mudança relativa da forma

$$RC = \left| \frac{\widehat{\alpha}^{\mathrm{o}} - \widehat{\alpha}^{\mathrm{m}}}{\widehat{\alpha}^{\mathrm{o}}} \right| \times 100, \tag{4.15}$$

em que $\widehat{\alpha}^{\mathrm{o}}$ e $\widehat{\alpha}^{\mathrm{m}}$ correspondem a algum parâmetro de FMR-SMSN (modelo sem dados ausentes) e FMR-SMSN-MD, respectivamente.

Na Figura 18, temos os boxplots para as 100 mudanças relativas de β_{02}, β_{11} e β_{12} para os mecanismos MNAR e MCAR, diante dessa figura podemos observar que os modelos assimétricos tiveram as menores mudanças relativas quando comparados com os modelos simétricos, também é importante notar que taxas maiores de dados ausentes geraram mudanças relativas maiores e mais dispersas. Em sua maioria, os modelos simétricos tiveram mudanças maiores e com maior variabilidade do que os modelos assimétricos. Os demais parâmetros foram abstraídos por questão de espaço, porém tiveram comportamentos bem parecidos.

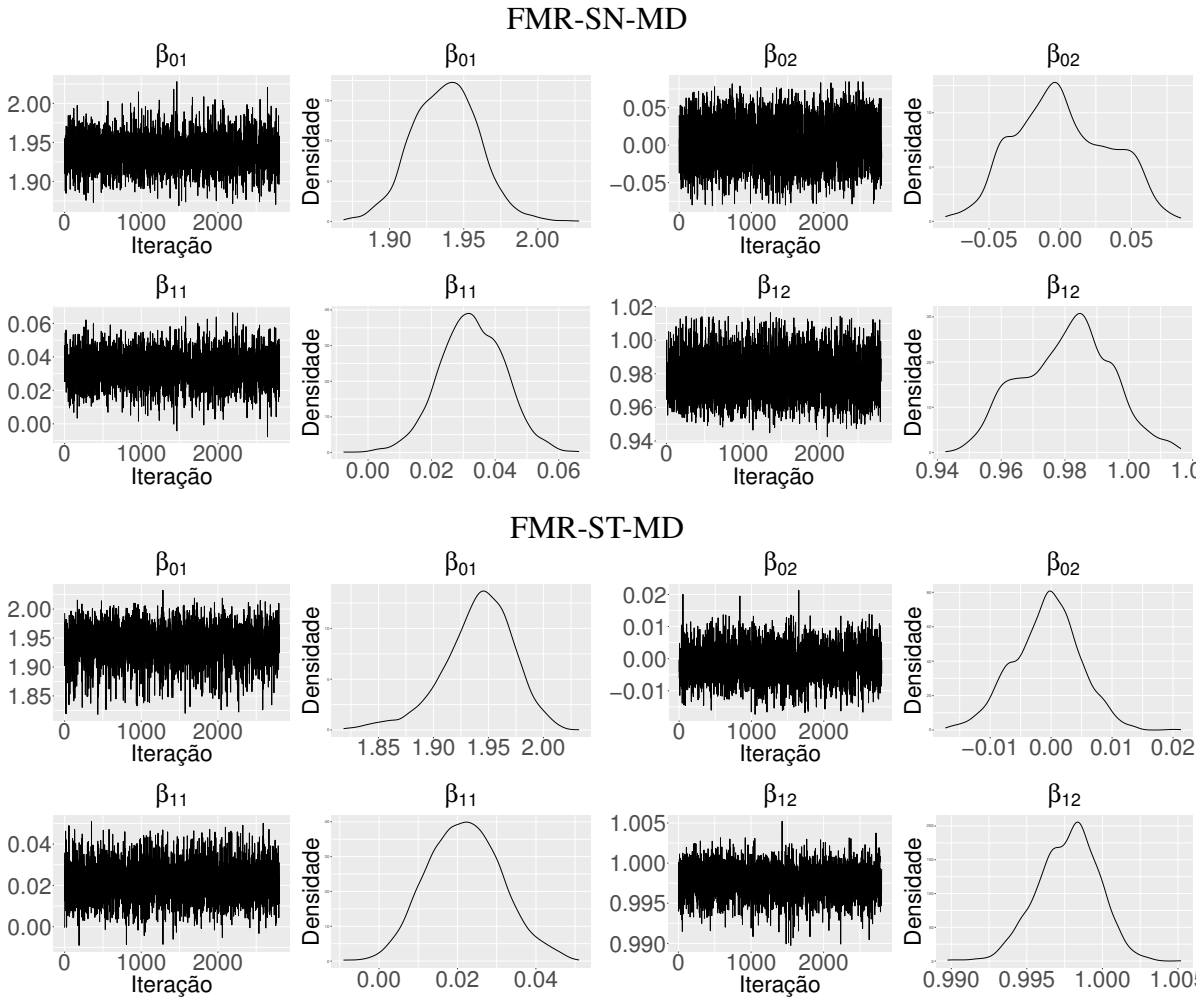

Figura 16 – Gráficos de traço e de Kernel de alguns parâmetros dos modelo FMR-SN-MD e FMR-ST-MD sob o mecanismo MCAR, ajustado nos dados *Tons*.

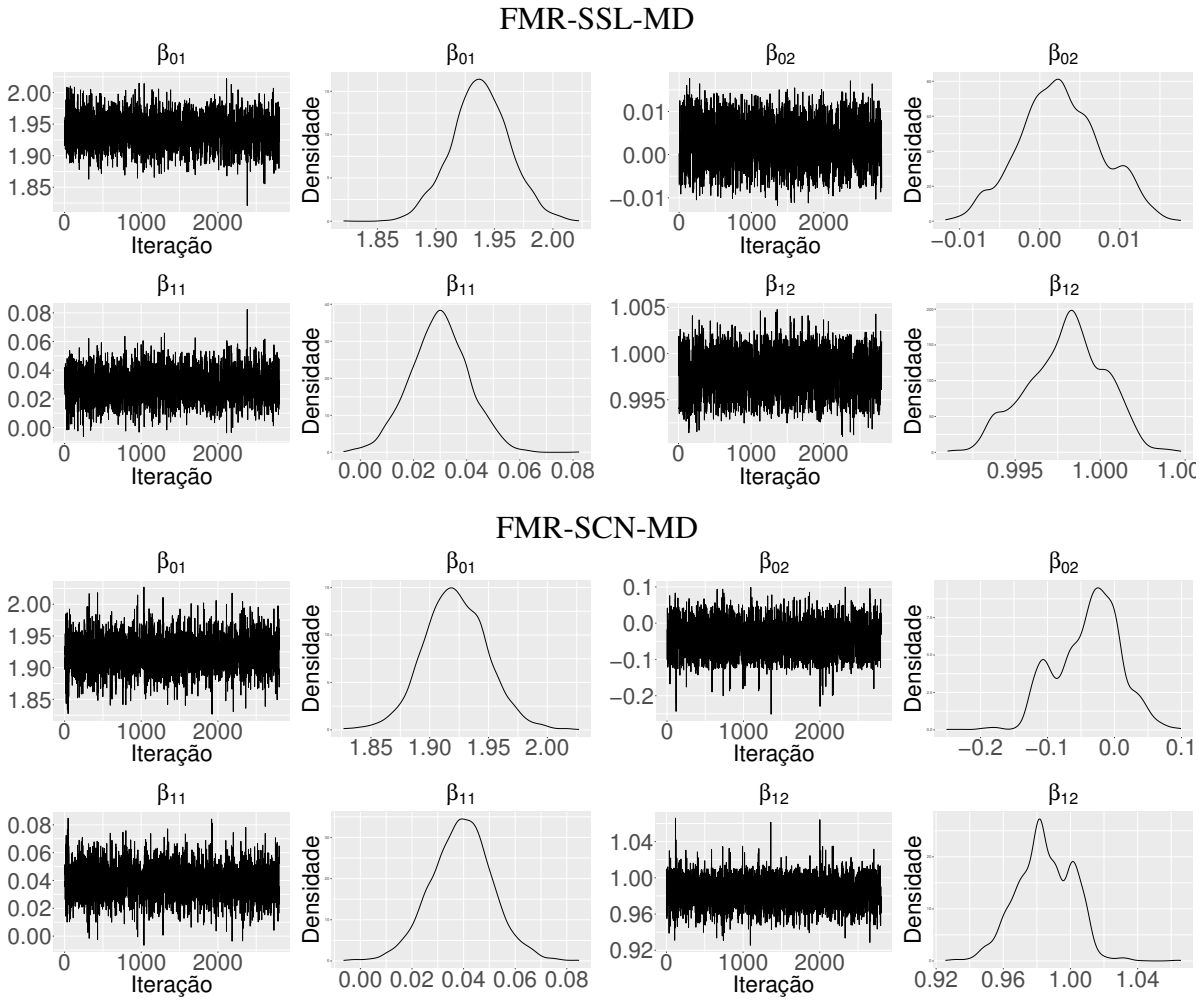

Figura 17 – Gráficos de traço e de Kernel de alguns parâmetros dos modelo FMR-SSL-MD e FMR-SCN-MD sob o mecanismo MCAR, ajustado nos dados *Tons*.

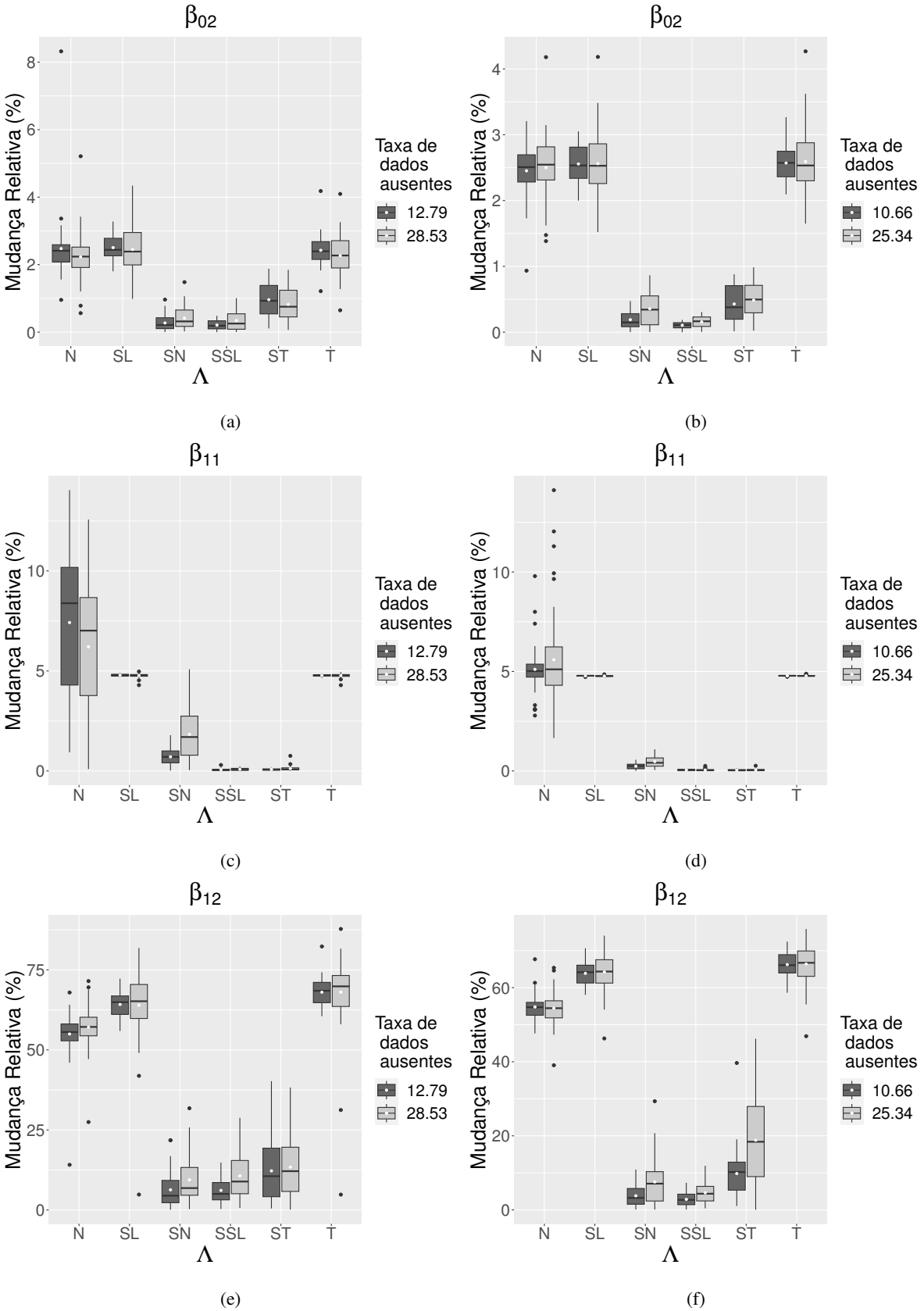

Figura 18 – Mudança relativa para alguns parâmetros sob os mecanismos MNAR (a, c, e) e MCAR (b, d, e), dados Tons.

5 CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho, propomos dois modelos, onde ambos são modelos de misturas de regressões cujos erros aleatórios seguem uma distribuição pertencente a uma classe flexível de distribuições. O primeiro, o qual denotamos por FMR-SMSN-CR, foi uma extensão do modelo tobit clássico com erros normais, que se mostrou flexível o suficiente para ajustar simultaneamente assimetria e caudas pesadas, além de acomodar o comportamento de grupos para os quais temos distintos coeficientes de regressão. O modelo proposto foi uma mistura de modelos de regressão, cujos erros aleatórios seguem distribuição oriunda de uma família de misturas de escala da normal assimétrica, e cuja variável resposta possui censuras, generalizando os trabalhos de Massuia *et al.* (2017), Zeller *et al.* (2019) e Nascimento e Abanto-Valle (2022). No segundo modelo, o qual denotamos por FMR-SMSN-MD, trabalhamos o problema de observações ausentes, onde criamos uma abordagem muito geral em modelos de regressão, propondo um modelo de misturas de regressões que incorporou a possibilidade dessas observações ausentes ocorrerem na resposta e nas covariáveis de maneira simultânea, por meio de mecanismos distintos que explicam essas ausências, também supondo que os erros aleatórios associados a variável resposta são modelados por uma mistura finita de distribuições normais assimétricas.

Para explorar as propriedades estatísticas dos modelos propostos, sugerimos um algoritmo do tipo Gibbs eficiente para cada modelo, FMR-SMSN-CR e FMR-SMSN-MD, que pode ser implementado com softwares estatísticos existentes, como R, outras formas de se obter as amostras MCMC podem ocorrer com a utilização de softwares como STAN e JAGS, que foi o escolhido para estimação do modelo FMR-SMSN-MD.

A metodologia aplicada foi a análise de dados artificiais através de estudos de simulação que mostraram a eficiência dos modelos em: (i) recuperar os verdadeiros valores dos parâmetros sob vários panoramas diferentes, como tamanho amostral, taxa de dados censurados (modelo FMR-SMSN-CR), taxa de dados ausentes (modelo FMR-SMSN-MD) e covariáveis seguindo diferentes distribuições de probabilidade; (ii) ajustar dados oriundos de modelos de natureza diferente da família SMSN, e; (iii) ajustar dados com outliers, tal que um pequeno estudo de influência local foi conduzido mostrando a flexibilidade dos modelos de caudas pesadas ao ajustar pontos discrepantes. Ao final, trabalhamos o algoritmo de Clonagem de Dados onde vimos que o mesmo apontou evidências de que os nossos modelos são identificáveis.

Para os dados reais, os modelos se mostraram flexíveis o suficiente para ajustar com qualidade dados com características de caudas pesadas, assimetria e multimodalidade onde os

modelos assimétricos sempre se mostraram mais vantajosos com relação aos simétricos, por meio do critério de seleção DIC para os modelos FMR-SMSN-MD e $WAIC_1$, $WAIC_2$, LPML e p-valores Bayesianos como no caso do modelo FMR-SMSN-CR.

Os programas de computador e dados utilizados nas aplicações e nas simulações podem ser encontrados no seguinte repositório https://github.com/nelfilho?tab=repositories.

Aqui faremos um apanhado de trabalhos futuros em nossa linha de estudo: (1) Para os modelos FMR-SMSN-CR e FMR-SMSN-MD supor que o número de componentes na mistura é desconhecido e desenvolver uma metodologia para estimá-lo. (2) Testar a robustez dos modelos para casos de má-especificação do mecanismos geradores de dados faltantes. (3) Proceder uma análise de diagnóstico dos modelos FMR-SMSN-CR e FMR-SMSN-MD. (4) Criar modelos FMR-SMSN-CR e FMR-SMSN-MD com erros sob uma estrutura auto-regressiva. (5) Estudar um modelo de mistura de modelos de regressão quantílica com erros na família SMSN e censura ou dados ausentes.

REFERÊNCIAS

AL-MALKAWI, H.-A. N. Determinants of corporate dividend policy in jordan: an application of the tobit model. **Journal of Economic and Administrative Sciences**, Emerald Group Publishing Limited, v. 23, p. 44–70, 2007.

ALENCAR, F. H. de; GALARZA, C. E.; MATOS, L. A.; LACHOS, V. H. Finite mixture modeling of censored and missing data using the multivariate skew-normal distribution. **Advances in Data Analysis and Classification**, Springer, v. 16, p. 521–557, 2022.

ALENCAR, F. H. de; MATOS, L. A.; LACHOS, V. H. Finite mixture of censored linear mixed models for irregularly observed longitudinal data. **Journal of Classification**, Springer, p. 1–24, 2022.

ANDREWS, D. F.; MALLOWS, C. L. Scale mixtures of normal distributions. **Journal of the Royal Statistical Society, Series B**, v. 36, p. 99–102, 1974.

ARELLANO-VALLE, R. B.; CASTRO, L. M.; GONZÁLEZ-FARÍAS, G.; MUÑOZ-GAJARDO, K. A. Student-t censored regression model: properties and inference. **Statistical Methods & Applications**, Springer, v. 21, p. 453–473, 2012.

AZZALINI, A. **The skew-normal and related families**. [S.l.]: Cambridge University Press, 2013.

BARNDORFF-NIELSEN, O. E. Normal inverse Gaussian distributions and stochastic volatility modelling. **Scandinavian Journal of Statistics**, v. 24, p. 1–13, 1997.

BASSO, R. M.; LACHOS, V. H.; CABRAL, C. R. B.; GHOSH, P. Robust mixture modeling based on scale mixtures of skew-normal distributions. **Computational Statistics & Data Analysis**, Elsevier, v. 54, p. 2926–2941, 2010.

BRANCO, M. D.; DEY, D. K. A general class of multivariate skew-elliptical distributions. **Journal of Multivariate Analysis**, v. 79, p. 99–113, 2001.

BREEN, R. *et al.* **Regression models: Censored, sample selected, or truncated data**. [S.l.]: Sage, 1996.

CABRAL, C. R. B.; BOLFARINE, H.; PEREIRA, J. R. G. Bayesian density estimation using skew student-t-normal mixtures. **Computational Statistics & Data Analysis**, Elsevier, v. 52, p. 5075–5090, 2008.

CABRAL, C. R. B.; LACHOS, V. H.; MADRUGA, M. R. Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population. **Journal of Statistical Planning and Inference**, Elsevier, v. 142, p. 181–200, 2012.

CABRAL, C. R. B.; LACHOS, V. H.; PRATES, M. O. Multivariate mixture modeling using skew-normal independent distributions. **Computational Statistics & Data Analysis**, Elsevier, v. 56, p. 126–142, 2012.

CABRAL, C. R. B.; SOUZA, N. L. de; LEÃO, J. Bayesian measurement error models using finite mixtures of scale mixtures of skew-normal distributions. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 92, p. 623–644, 2022.

CAI, J.-H.; SONG, X.-Y.; HSER, Y.-I. A bayesian analysis of mixture structural equation models with non-ignorable missing responses and covariates. **Statistics in Medicine**, Wiley Online Library, v. 29, p. 1861–1874, 2010.

CARPENTER, B.; GELMAN, A.; HOFFMAN, M. D.; LEE, D.; GOODRICH, B.; BETANCOURT, M.; BRUBAKER, M.; GUO, J.; LI, P.; RIDDELL, A. Stan: A probabilistic programming language. **Journal of Statistical Software**, Columbia Univ., New York, NY (United States); Harvard Univ., Cambridge, MA, v. 76, p. 1–10, 2017.

CELEUX, G.; FORBES, F.; ROBERT, C. P.; TITTERINGTON, D. M. Deviance information criteria for missing data models. **Bayesian Analysis**, v. 1, p. 651–674, 2006.

COHEN, E. A. Some effects of inharmonic partials on interval perception. **Music Perception**, University of California Press, v. 1, p. 323–349, 1984.

COSSLETT, S. R.; LEE, L.-F. Serial correlation in latent discrete variable models. **Journal of Econometrics**, Elsevier, v. 27, p. 79–97, 1985.

DANIELS, M. J.; HOGAN, J. W. **Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis**. [S.l.]: chapman and hall/CRC, 2008.

DESARBO, W. S.; CRON, W. L. A maximum likelihood methodology for clusterwise linear regression. **Journal of classification**, Springer, v. 5, p. 249–282, 1988.

FRÜHWIRTH-SCHNATTER, S. **Finite Mixture and Markov Switching Models**. [S.l.]: Springer Verlag, 2006.

GARAY, A. M.; BOLFARINE, H.; LACHOS, V. H.; CABRAL, C. R. Bayesian analysis of censored linear regression models with scale mixtures of normal distributions. **Journal of Applied Statistics**, Taylor & Francis, v. 42, p. 2694–2714, 2015.

GARAY, A. M.; LACHOS, V. H.; BOLFARINE, H.; CABRAL, C. R. Linear censored regression models with scale mixtures of normal distributions. **Statistical Papers**, Springer, v. 58, p. 247–278, 2017.

GARCIA, R. I.; IBRAHIM, J. G.; ZHU, H. Variable selection for regression models with missing data. **Statistica Sinica**, NIH Public Access, v. 20, p. 149–165, 2010.

GELFAND, A. E. Gibbs sampling. **Journal of the American statistical Association**, Taylor & Francis, v. 95, p. 1300–1304, 2000.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian data analysis**. [S.l.]: Taylor & Francis, 2013.

GUO, Y.; SAYED, T.; ESSA, M. Real-time conflict-based bayesian tobit models for safety evaluation of signalized intersections. **Accident Analysis & Prevention**, Elsevier, v. 144, p. 105660, 2020.

HENZE, N. A probabilistic representation of the'skew-normal'distribution. **Scandinavian Journal of Statistics**, JSTOR, 1986.

HOU, Q.; HUO, X.; LENG, J. A correlated random parameters tobit model to analyze the safety effects and temporal instability of factors affecting crash rates. **Accident Analysis & Prevention**, Elsevier, v. 134, p. 105326, 2020.

HUNTER, D. R.; YOUNG, D. S. Semiparametric mixtures of regressions. **Journal of Nonparametric Statistics**, Taylor & Francis, v. 24, p. 19–38, 2012.

KARLSSON, M.; LAITILA, T. Finite mixture modeling of censored regression models. **Statistical Papers**, Springer, v. 55, p. 627–642, 2014.

KOMRATTANAPANYA, P.; SUNTRARUK, P. Factors influencing dividend payout in Thailand: A tobit regression analysis. **International Journal of Accounting and Financial Reporting**, Macrothink Institute Inc., v. 3, p. 255, 2013.

LACHOS, V. H.; GHOSH, P.; ARELLANO-VALLE, R. B. Likelihood based inference for skew normal independent linear mixed models. **Statistica Sinica**, v. 20, p. 303–322, 2010.

LANGE, K.; SINSHEIMER, J. S. Normal/independent distributions and their applications in robust regression. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 2, p. 175–198, 1993.

LELE, S. R.; NADEEM, K.; SCHMULAND, B. Estimability and likelihood inference for generalized linear mixed models using data cloning. **Journal of the American Statistical Association**, Taylor & Francis, v. 105, p. 1617–1625, 2010.

LEMUS, M. N.; LACHOS, V. H.; GALARZA, C. E.; MATOS, L. A. Estimation and diagnostics for partially linear censored regression models based on heavy-tailed distributions. **Statistics and Its Interface**, International Press of Boston, v. 14, p. 165–182, 2021.

LI, L.; PU, Z. Rank estimation of log-linear regression with interval-censored data. **Lifetime Data Analysis**, Springer, v. 9, p. 57–70, 2003.

LIM, H. K.; NARISETTY, N. N.; CHEON, S. Robust multivariate mixture regression models with incomplete data. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, p. 328–347, 2017.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. [S.l.]: John Wiley & Sons, 2019.

LIU, C. Bayesian robust multivariate linear regression with incomplete data. **Journal of the American Statistical Association**, Taylor & Francis, v. 91, p. 1219–1227, 1996.

LIU, J. S. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. **Journal of the American Statistical Association**, Taylor & Francis, v. 89, p. 958–966, 1994.

LIU, M.; LIN, T.-I. A skew-normal mixture regression model. **Educational and Psychological Measurement**, SAGE Publications Sage CA: Los Angeles, CA, v. 74, p. 139–162, 2014.

MA, Z.; CHEN, G. Bayesian methods for dealing with missing data problems. **Journal of the Korean Statistical Society**, Springer, v. 47, p. 297–313, 2018.

MASSUIA, M. B.; CABRAL, C. R. B.; MATOS, L. A.; LACHOS, V. H. Influence diagnostics for student-t censored linear regression models. **Statistics**, Taylor & Francis, v. 49, p. 1074–1094, 2015.

MASSUIA, M. B.; GARAY, A. M.; CABRAL, C. R.; LACHOS, V. Bayesian analysis of censored linear regression models with scale mixtures of skew-normal distributions. **Statistics and its Interface**, International Press of Boston, v. 10, p. 425–439, 2017.

MATOS, L. A.; CASTRO, L. M.; CABRAL, C. R.; LACHOS, V. H. Multivariate measurement error models based on student-t distribution under censored responses. **Statistics**, Taylor & Francis, v. 52, p. 1395–1416, 2018.

MCLACHLAN, G.; PEEL, D. **Finite Mixture Models**. [S.l.]: John Wiley & Sons, 2000.

MIRFARAH, E.; NADERI, M.; CHEN, D.-G. Mixture of linear experts model for censored data: A novel approach with scale-mixture of normal distributions. **Computational Statistics & Data Analysis**, Elsevier, v. 158, p. 107–182, 2021.

MOLENBERGHS, G.; FITZMAURICE, G.; KENWARD, M. G.; TSIATIS, A.; VERBEKE, G. **Handbook of missing data methodology**. [S.l.]: CRC Press, 2014.

MOLENBERGHS, G.; KENWARD, M. **Missing data in clinical studies**. [S.l.]: John Wiley & Sons, 2007.

MROZ, T. A. *et al.* The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. **Econometrica**, Econometric Society, v. 55, p. 765–799, 1987.

NASCIMENTO, M. G. L.; ABANTO-VALLE, C. A. Flexible robust mixture regression modeling. **REVSTAT-Statistical Journal**, v. 20, p. 101–115, 2022.

NELSON, F. D. Censored regression models with unobserved, stochastic censoring thresholds. **Journal of econometrics**, Elsevier, v. 6, p. 309–327, 1977.

PLUMMER, M. *et al.* JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: VIENNA, AUSTRIA. **Proceedings of the 3rd international workshop on distributed statistical computing**. [S.l.], 2003. v. 124, p. 1–10.

PU, Z.; LI, L. Regression models with arbitrarily interval-censored observations. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 28, p. 1547–1563, 1999.

QUANDT, R. E. A new approach to estimating switching regressions. **Journal of the American statistical association**, Taylor & Francis, v. 67, p. 306–310, 1972.

QUANDT, R. E.; RAMSEY, J. B. Estimating mixtures of normal distributions and switching regressions. **Journal of the American statistical Association**, Taylor & Francis, v. 73, p. 730–738, 1978.

RICHARDSON, S.; GREEN, P. J. On Bayesian analysis of mixtures with an unknown number of components. **Journal of the Royal Statistical Society, Series B**, v. 59, p. 731–792, 1997.

ROSA, G.; PADOVANI, C. R.; GIANOLA, D. Robust linear mixed models with normal/independent distributions and bayesian mcmc implementation. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, Wiley Online Library, v. 45, p. 573–590, 2003.

RUBIN, D. B. Inference and missing data. **Biometrika**, Oxford University Press, v. 63, p. 581–592, 1976.

RUBIN, D. B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. **The Annals of Statistics**, JSTOR, v. 12, p. 1151–1172, 1984.

SÓLYMOS, P. dclone: Data cloning in r. **R Journal**, v. 2, p. 29–37, 2010.

SPÄTH, H. Algorithm 39 clusterwise linear regression. **Computing**, Springer, v. 22, p. 367–373, 1979.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. van der. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society, Series B**, v. 64, p. 583–639, 2002.

STEPHENS, M. Dealing with label switching in mixture models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 62, n. 4, 2000.

TEAM, R. C. *et al.* R: A language and environment for statistical computing. Vienna, Austria, 2013.

TURNER, T. R. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 49, p. 371–384, 2000.

VAIDA, F.; LIU, L. Fast implementation for normal mixed effects models with censored response. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 18, p. 797–817, 2009.

VALLE, A. D. **The skew-normal distribution**. [S.l.]: Chapman and Hall/CRC, 2004.

VEAUX, R. D. D. Mixtures of linear regressions. **Computational Statistics & Data Analysis**, Elsevier, v. 8, p. 227–245, 1989.

VIELE, K.; TONG, B. Modeling with mixtures of linear regressions. **Statistics and Computing**, Springer, v. 12, p. 315–330, 2002.

WANG, H. X.; ZHANG, Q. bing; LUO, B.; WEI, S. Robust mixture modelling using multivariate t-distribution with missing information. **Pattern Recognition Letters**, Elsevier, v. 25, p. 701–710, 2004.

WATANABE, S.; OPPER, M. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of Machine Learning Research**, v. 11, p. 3571–3594, 2010.

WILLIAM, H. G. **Econometric Analysis**. [S.l.]: Pearson Education Limited, 2008.

YAO, W.; WEI, Y.; YU, C. Robust mixture regression using the t-distribution. **Computational Statistics & Data Analysis**, Elsevier, v. 71, p. 116–127, 2014.

ZELLER, C. B.; CABRAL, C. R.; LACHOS, V. H. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. **Test**, Springer, v. 25, p. 375–396, 2016.

ZELLER, C. B.; CABRAL, C. R. B.; LACHOS, V. H.; BENITES, L. Finite mixture of regression models for censored data based on scale mixtures of normal distributions. **Advances in Data Analysis and Classification**, Springer, v. 13, p. 89–116, 2019.

ZHAO, Y.; DUAN, X. Bayesian adaptive lasso for regression models with nonignorable missing responses. **Journal of Mathematics**, Hindawi, v. 2022, p. 1–12, 2022.

APÊNDICE A –

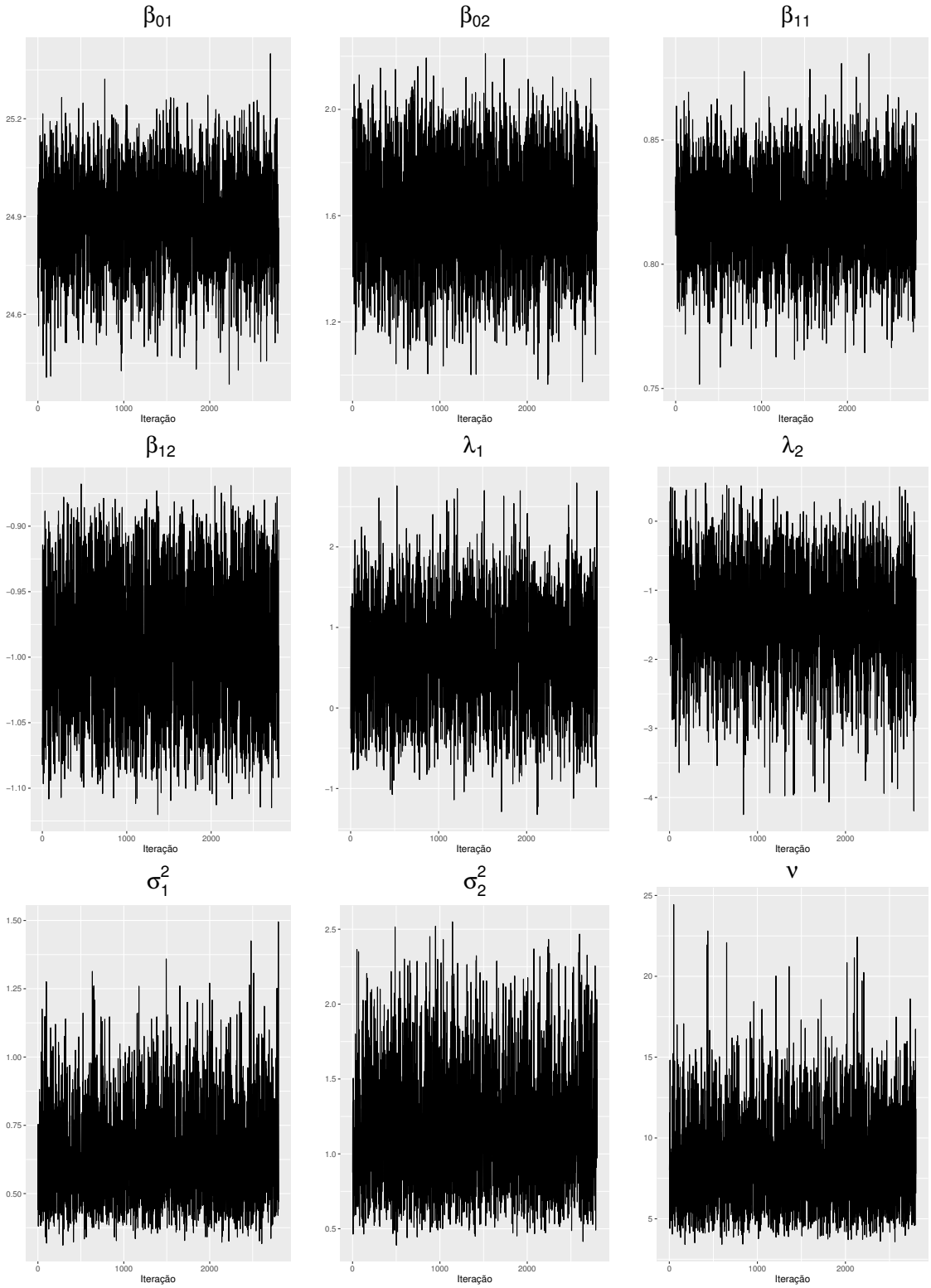

Figura 19 – Gráficos de traço feitos para os parâmetros do modelo FMR-ST-CR, considerando $n = 100$ e taxa de censura de 20%.

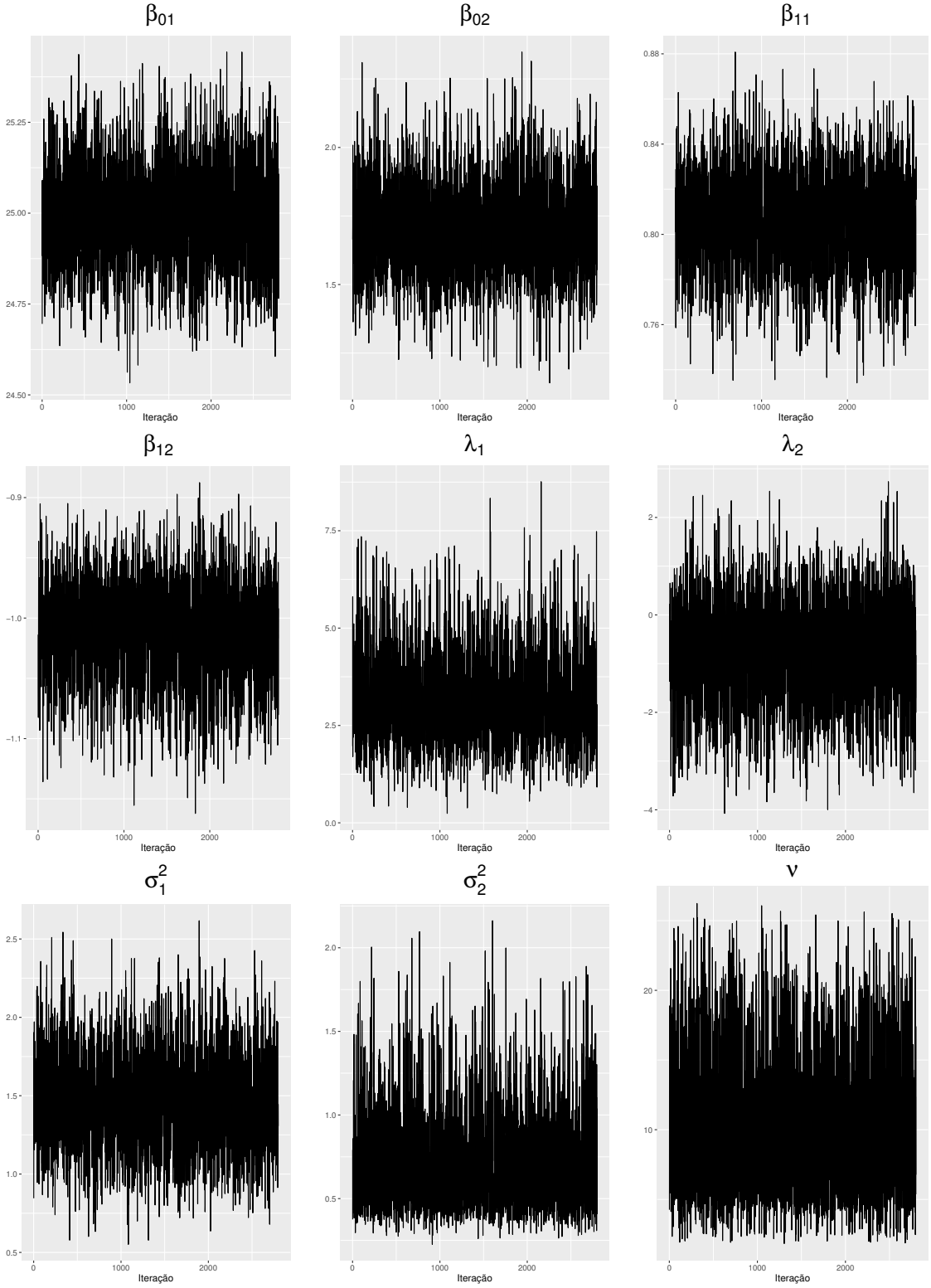

Figura 20 – Gráficos de traço feitos para os parâmetros do modelo FMR-SSL-CR, considerando $n = 100$ e taxa de censura de 20%.

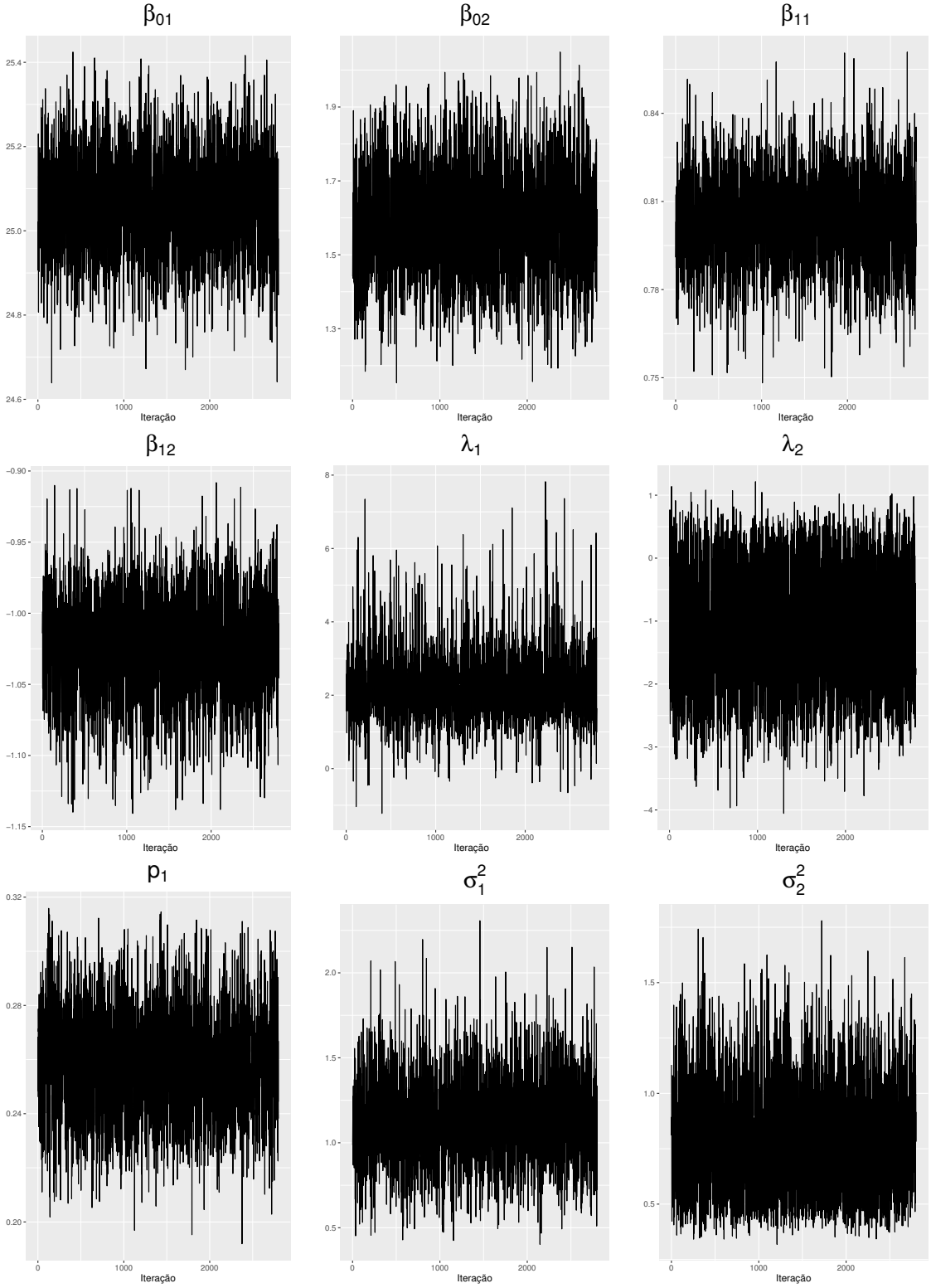

Figura 21 – Gráficos de traço feitos para os parâmetros do modelo FMR-SN-CR, considerando $n = 100$ e taxa de censura de 20%.

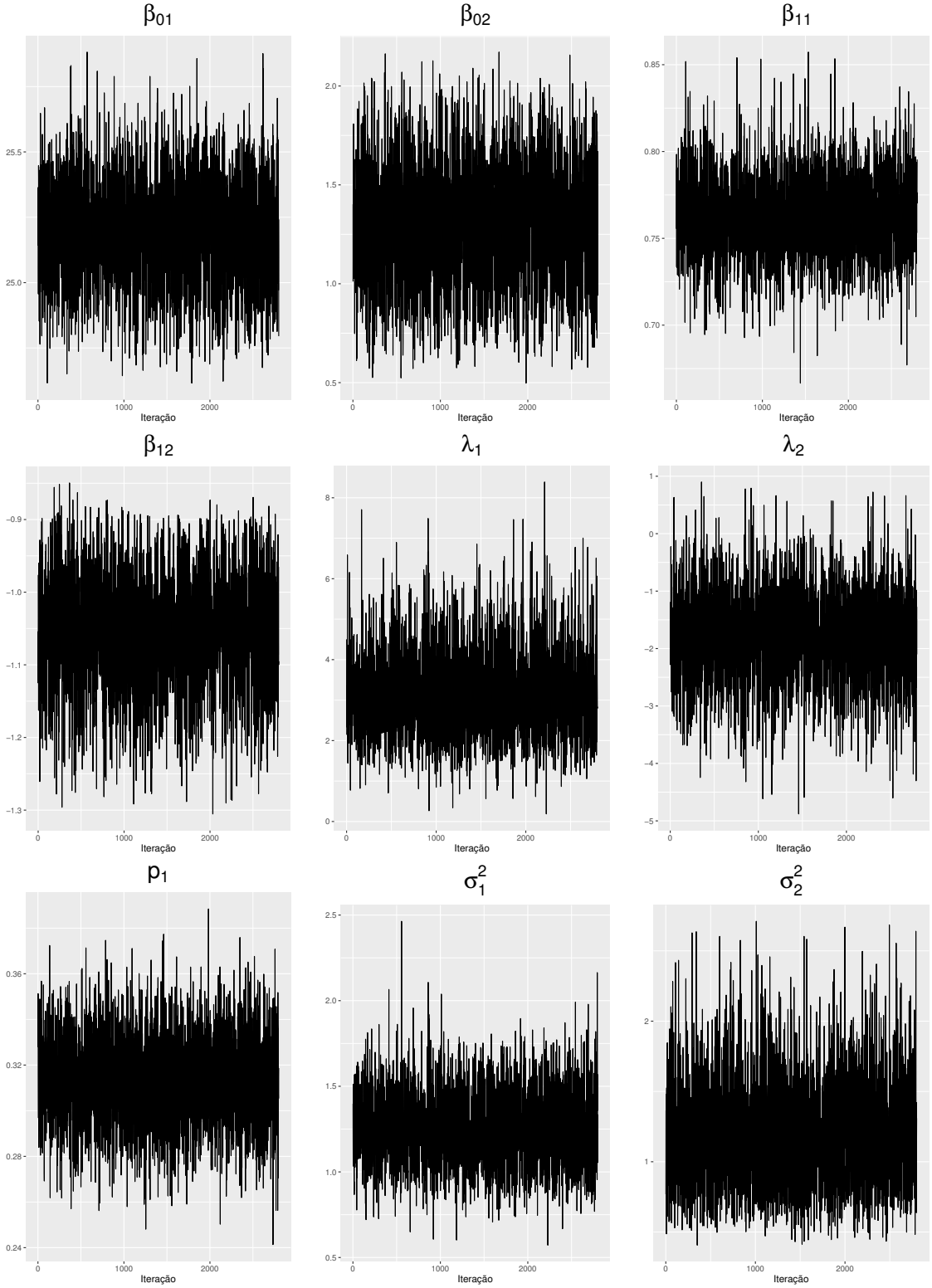

Figura 22 – Gráficos de traço feitos para os parâmetros do modelo FMR-SCN-CR, considerando $n = 100$ e taxa de censura de 20%.

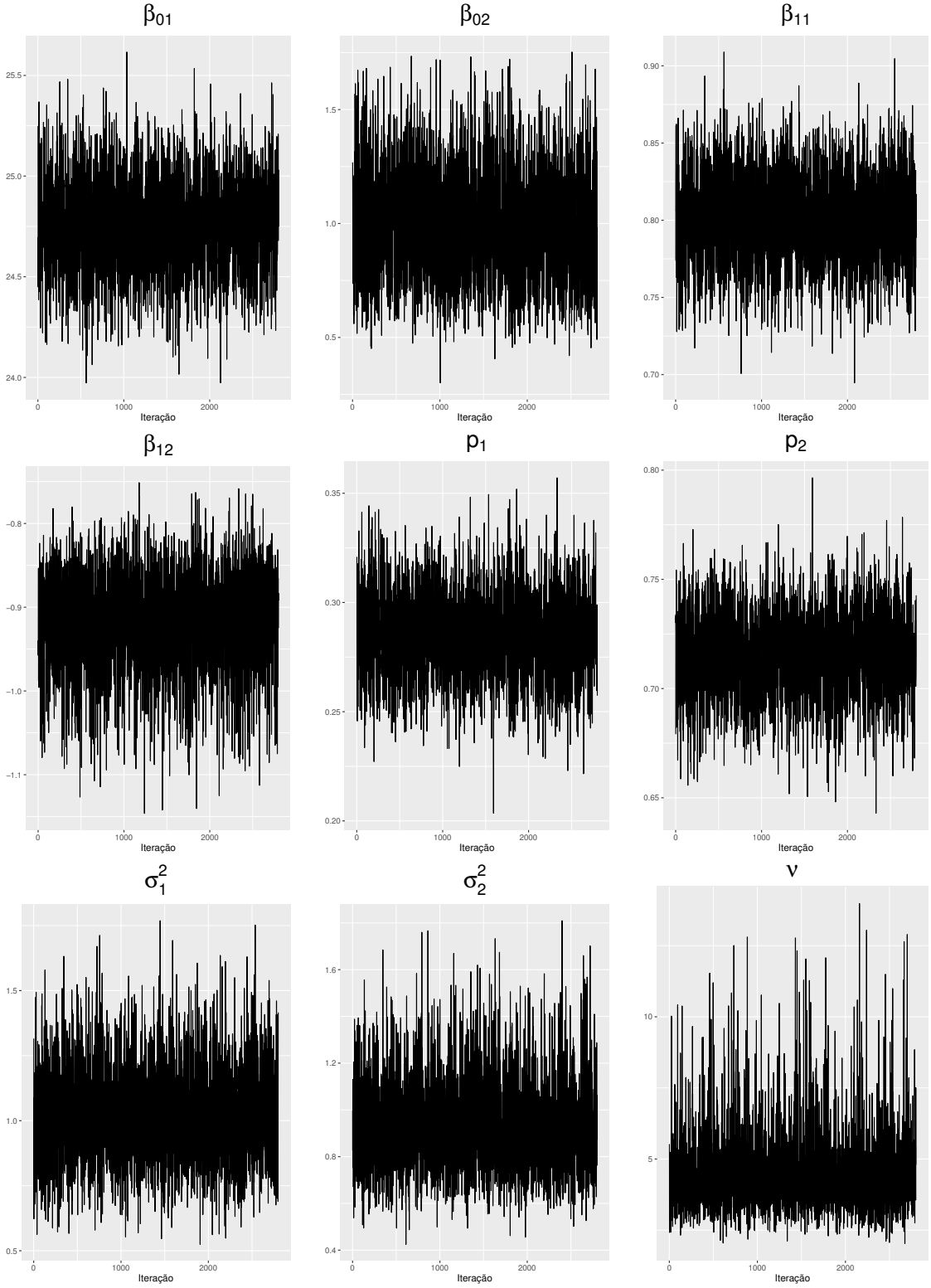

Figura 23 – Gráficos de traço feitos para os parâmetros do modelo FMR-T-CR, considerando $n = 100$ e taxa de censura de 20%.

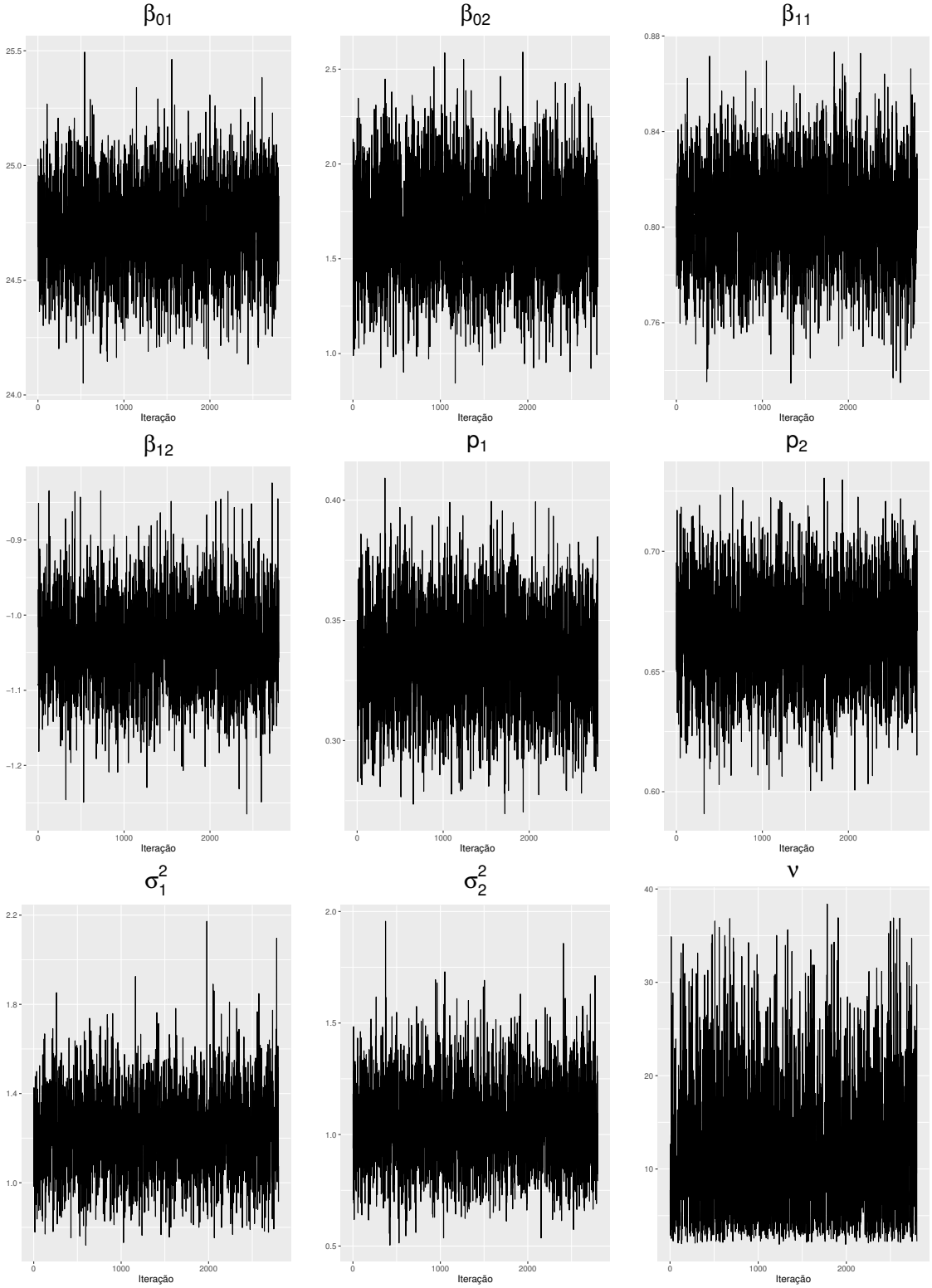

Figura 24 – Gráficos de traço feitos para os parâmetros do modelo FMR-SL-CR, considerando $n = 100$ e taxa de censura de 20%.

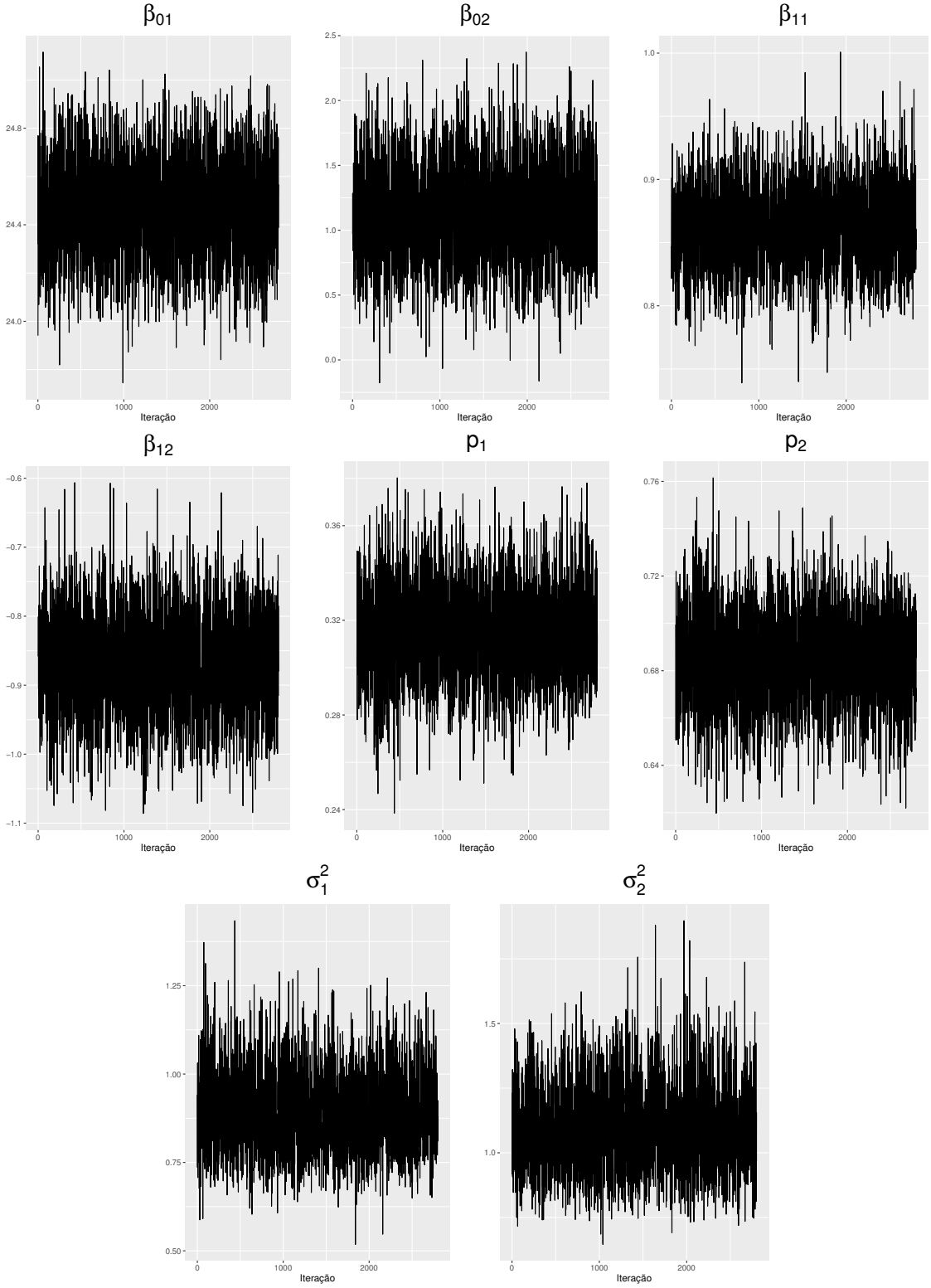

Figura 25 – Gráficos de traço feitos para os parâmetros do modelo FMR-CN-CR, considerando $n = 100$ e taxa de censura de 20%.

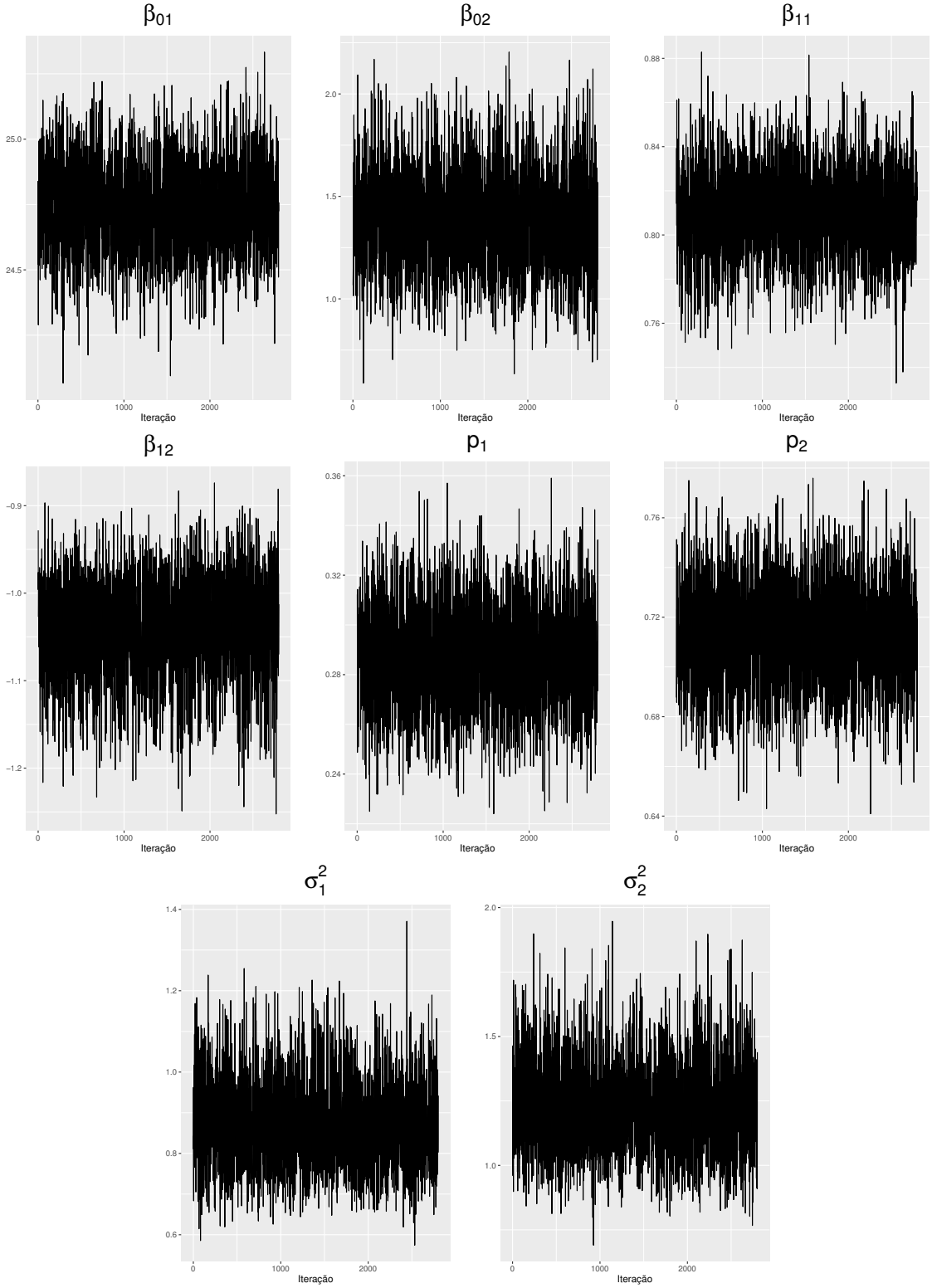

Figura 26 – Gráficos de traço feitos para os parâmetros do modelo FMR-N-CR, considerando $n = 100$ e taxa de censura de 20%.

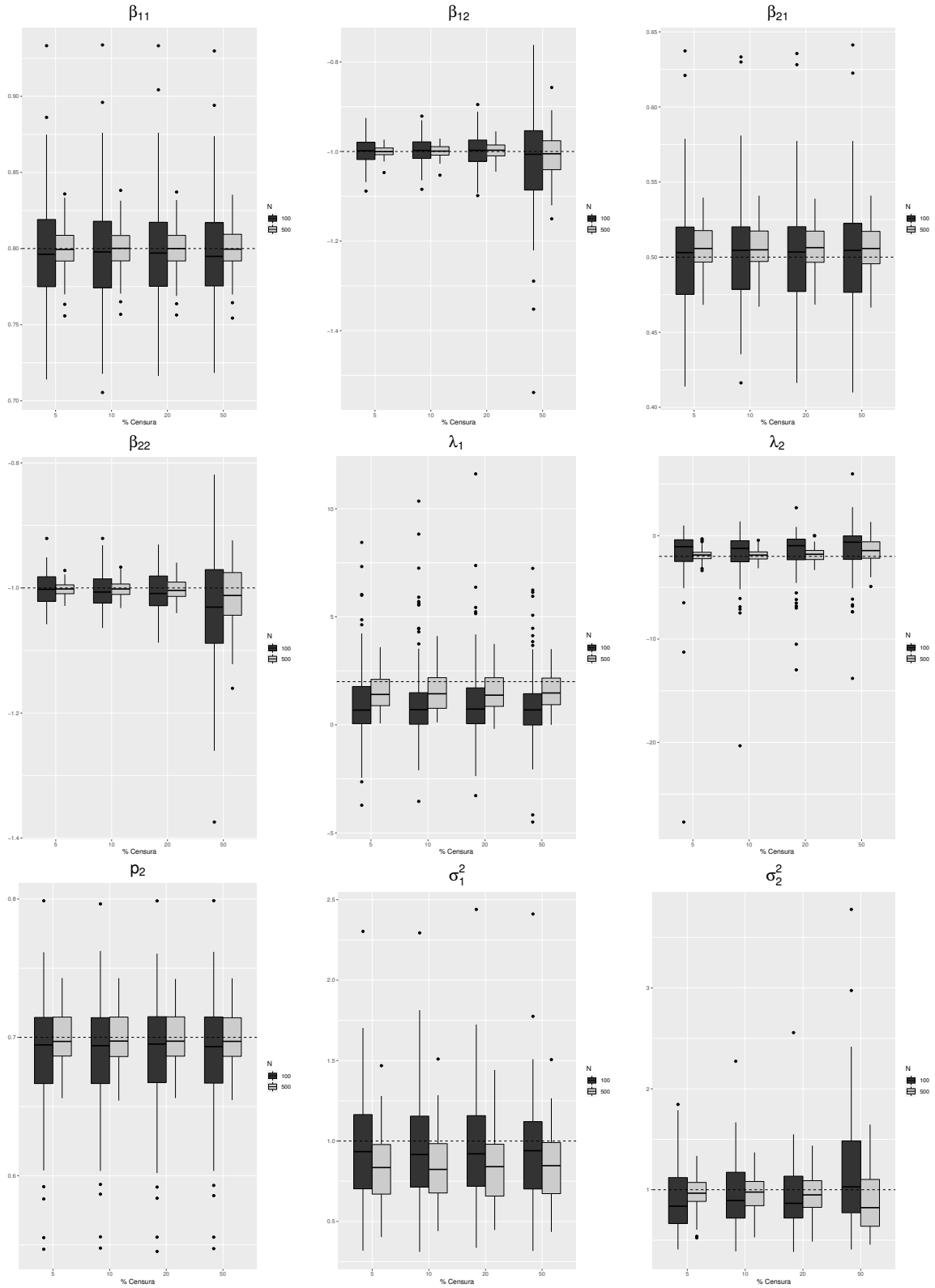

Figura 27 – Boxplots das estimativas de alguns parâmetros (a linha horizontal indica o verdadeiro valor do parâmetro) para o modelo FMR-SN-CR, considerando diferentes taxas de censura (reta das abscissas) e tamanhos amostrais (N).

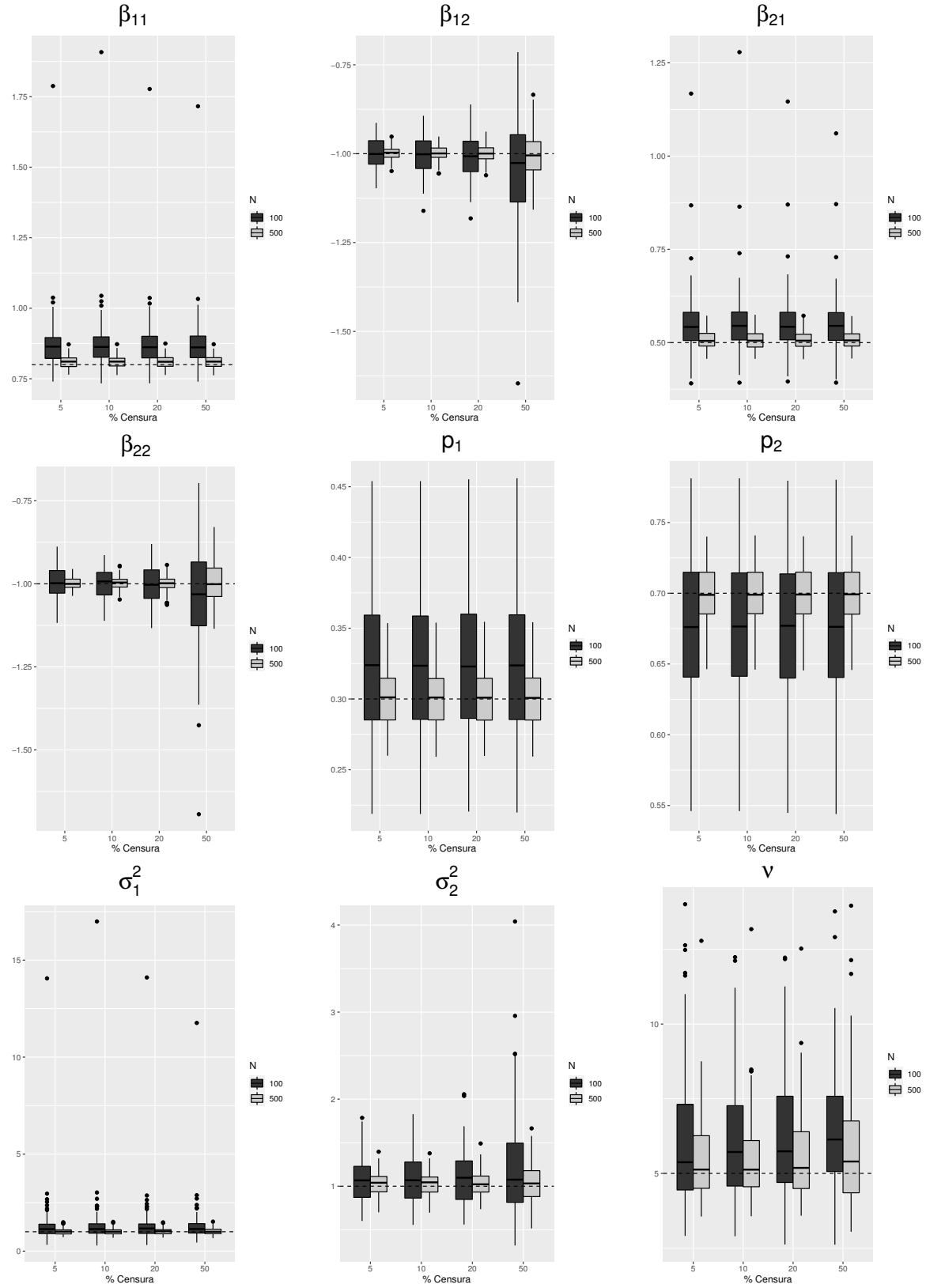

Figura 28 – Boxplots das estimativas de alguns parâmetros (a linha horizontal indica o verdadeiro valor do parâmetro) para o modelo FMR-T-CR, considerando diferentes taxas de censura (reta das abscissas) e tamanhos amostrais (N).

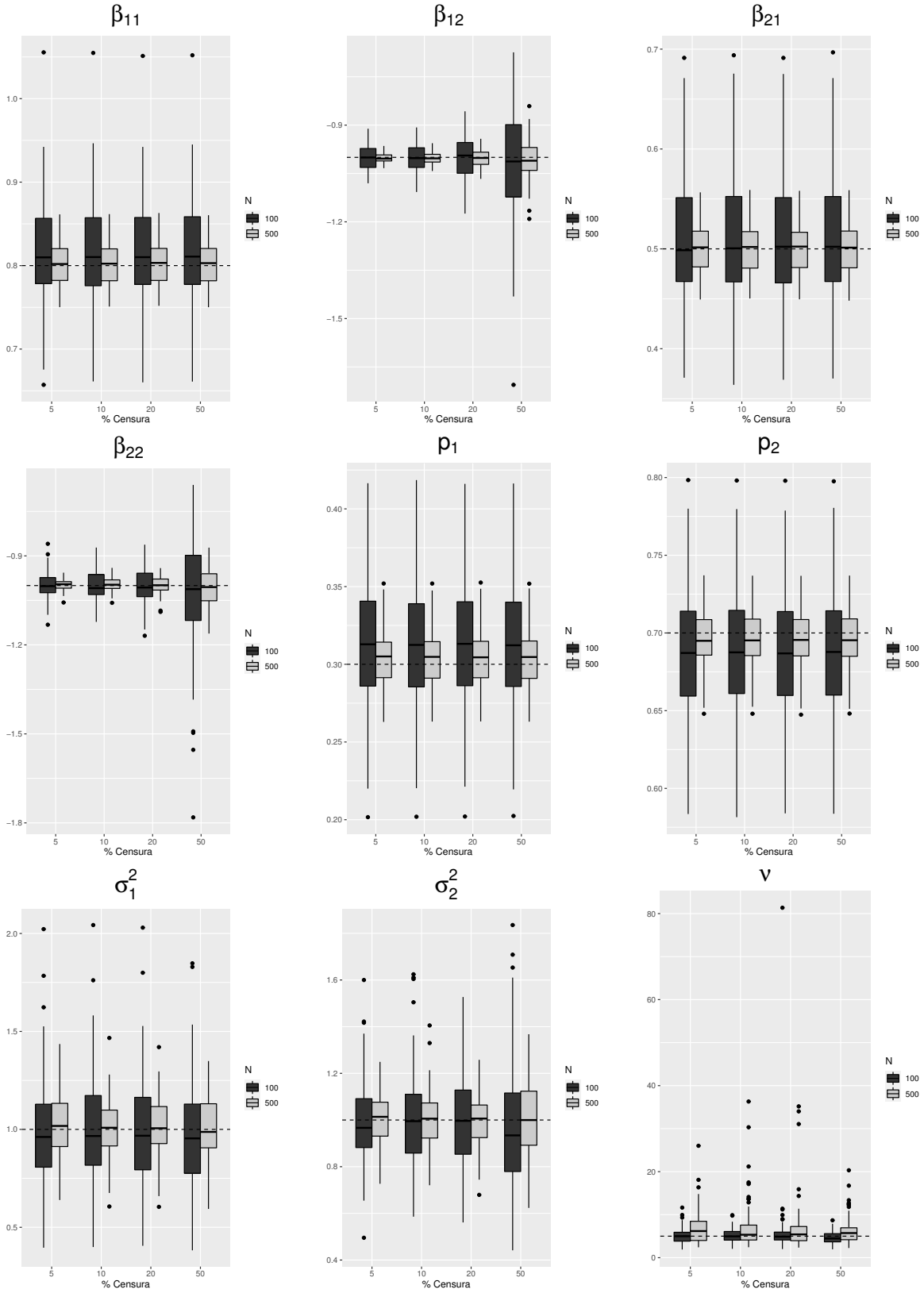

Figura 29 – Boxplots das estimativas de alguns parâmetros (a linha horizontal indica o verdadeiro valor do parâmetro) para o modelo FMR-SL-CR, considerando diferentes taxas de censura (reta das abscissas) e tamanhos amostrais (N).

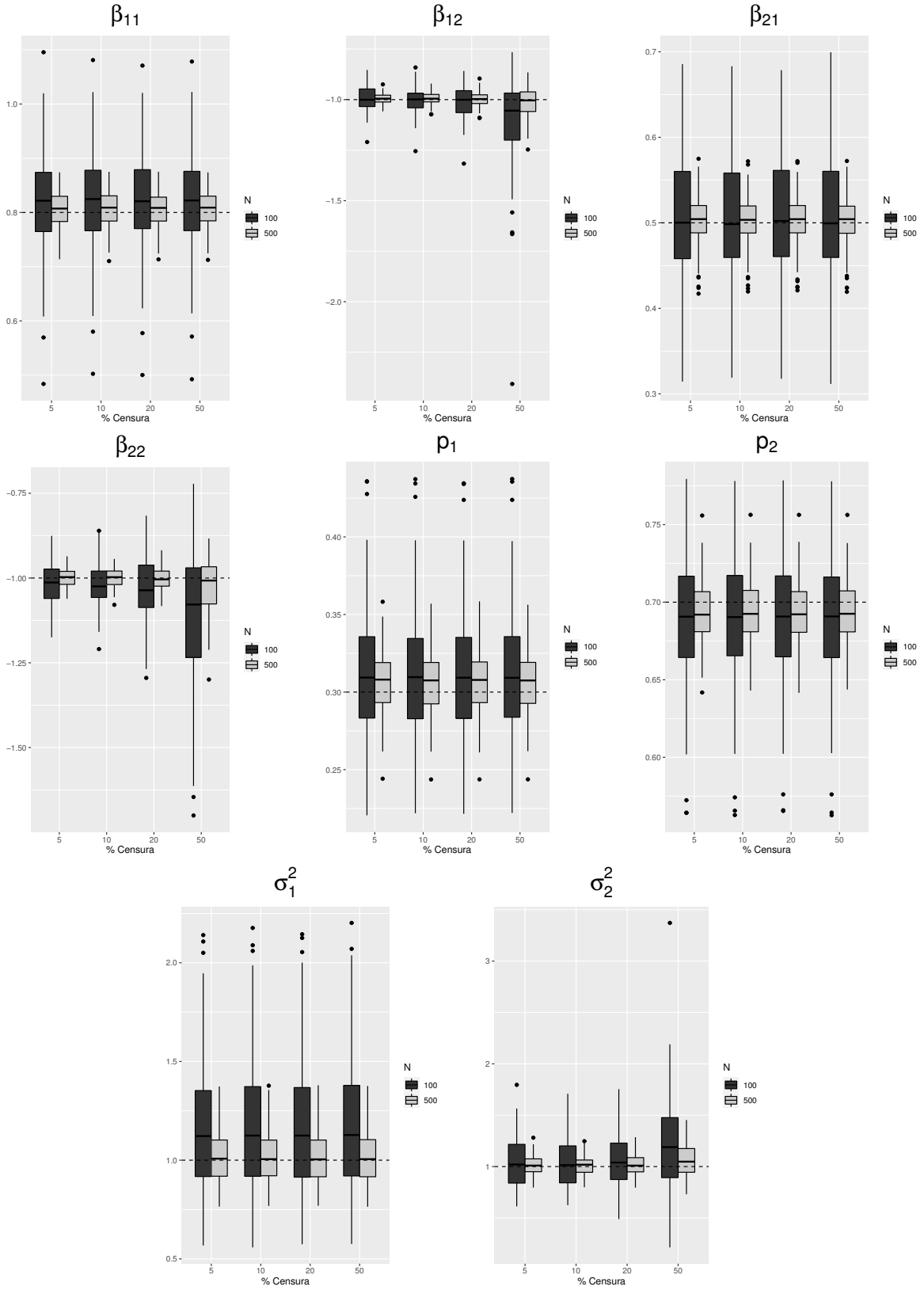

Figura 30 – Boxplots das estimativas de alguns parâmetros (a linha horizontal indica o verdadeiro valor do parâmetro) para o modelo FMR-CN-CR, considerando diferentes taxas de censura (reta das abscissas) e tamanhos amostrais (N).

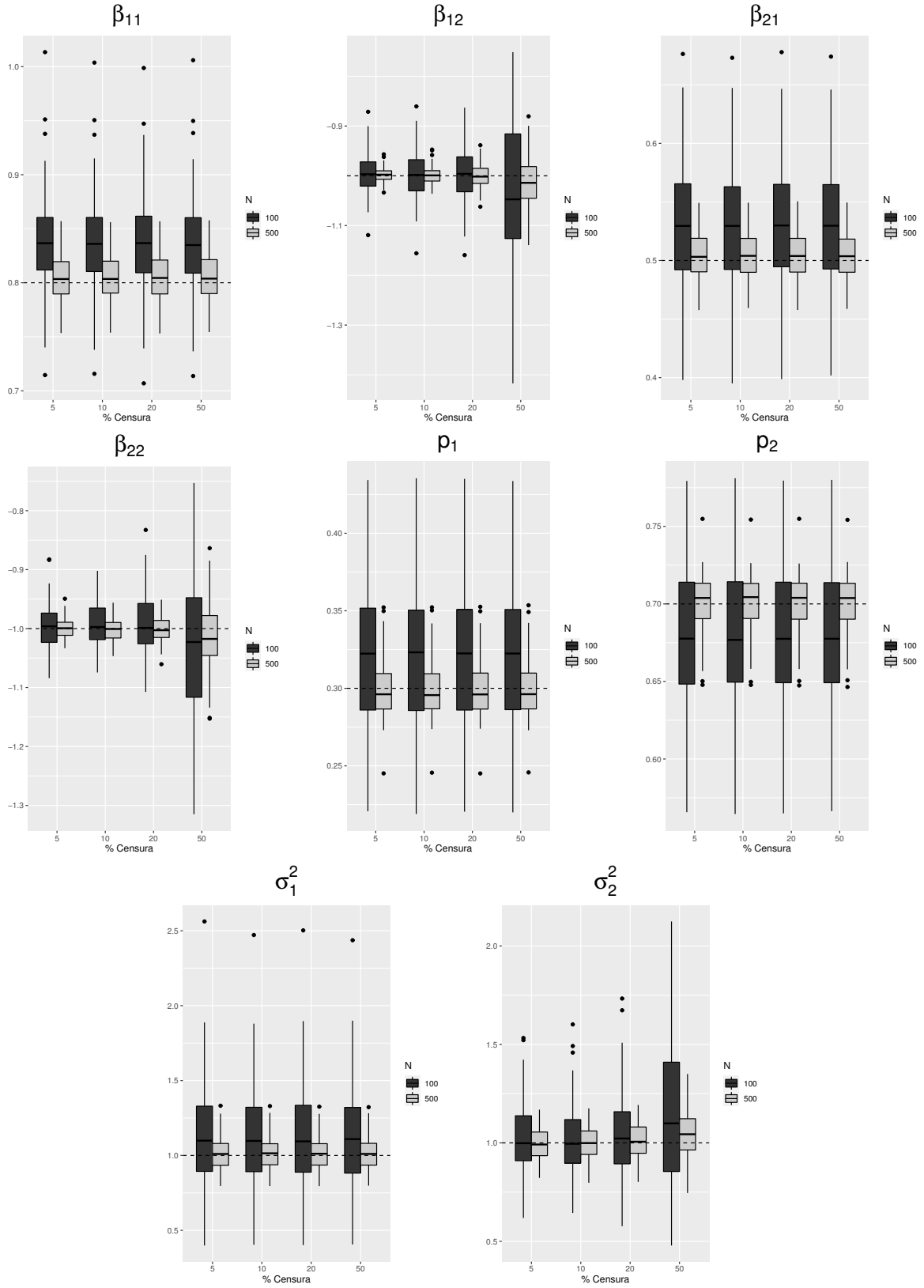

Figura 31 – Boxplots das estimativas de alguns parâmetros (a linha horizontal indica o verdadeiro valor do parâmetro) para o modelo FMR-N-CR, considerando diferentes taxas de censura (reta das abscissas) e tamanhos amostrais (N).

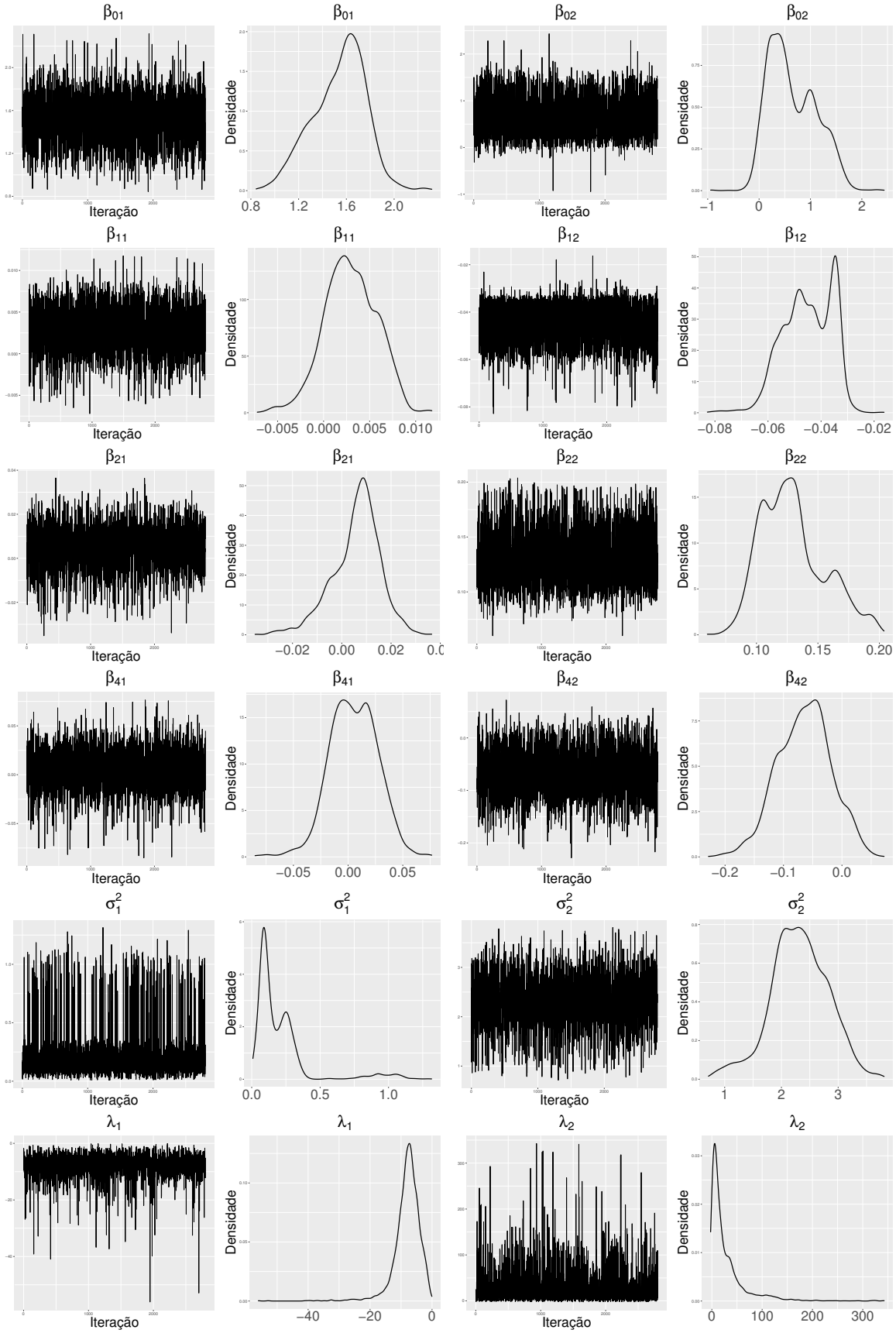

Figura 32 – Gráficos de traço e de Kernel de alguns parâmetros do modelo FMR-SSL-CR ajustado nos dados *wage rate*.

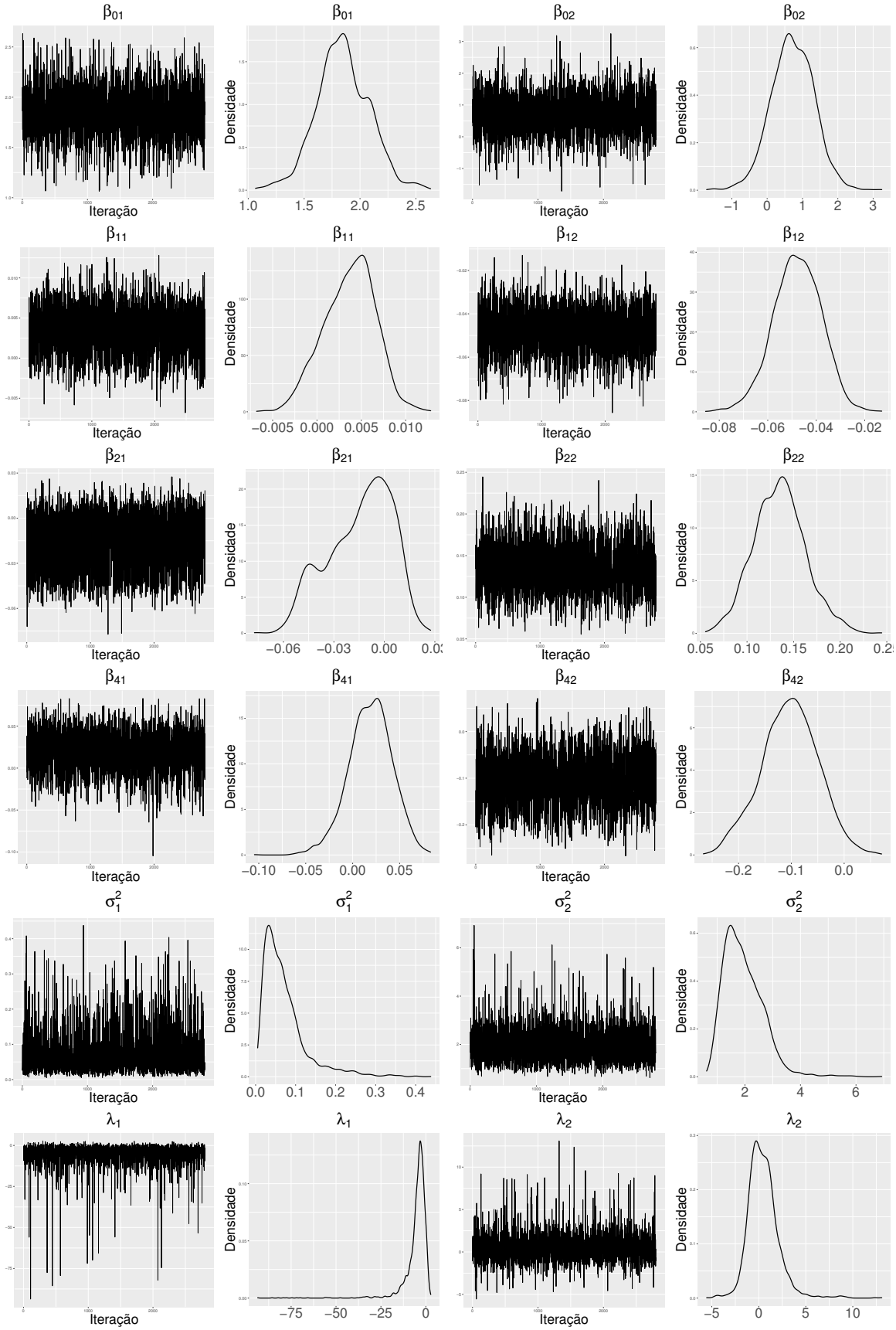

Figura 33 – Gráficos de traço e de Kernel de alguns parâmetros do modelo FMR-SCN-CR ajustado nos dados *wage rate*.

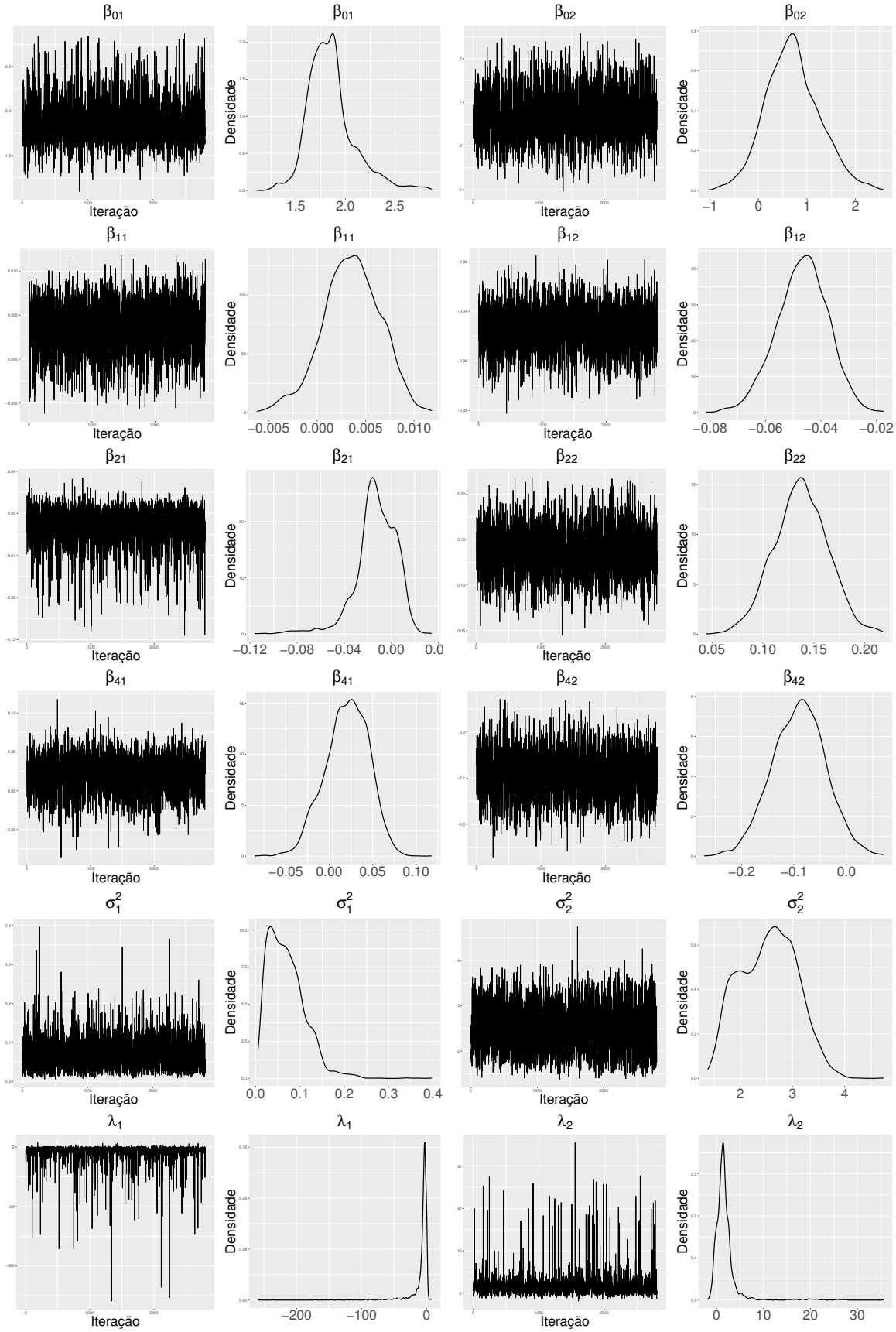

Figura 34 – Gráficos de traço e de Kernel de alguns parâmetros do modelo FMR-SN-CR ajustado nos dados *wage rate*.

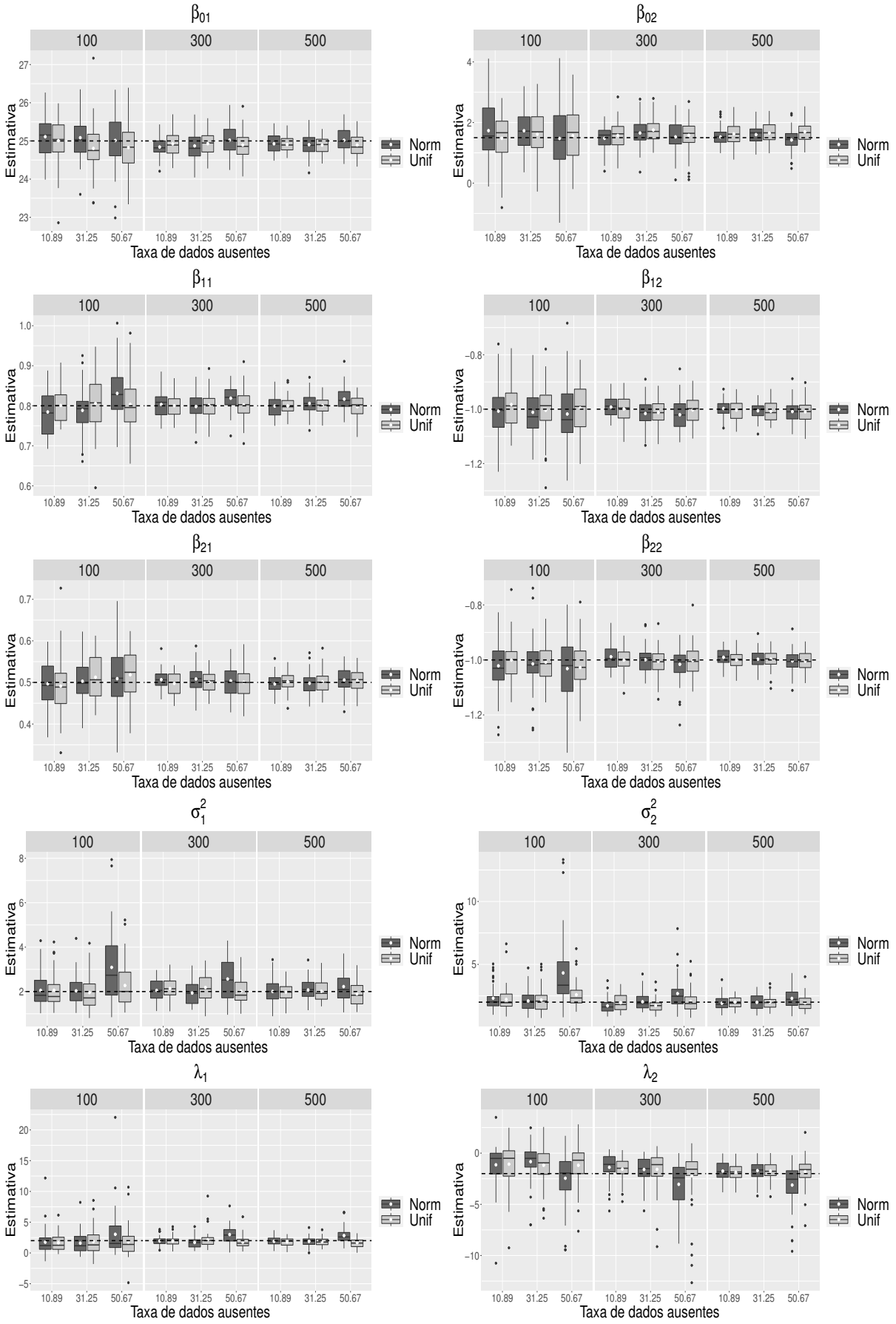

Figura 35 – Boxplots das estimativas dos parâmetros para o modelo FMR-SSL-MD, comparando diferentes distribuições para a covariável, diferentes tamanhos amostrais e diferentes taxas de dados ausentes. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

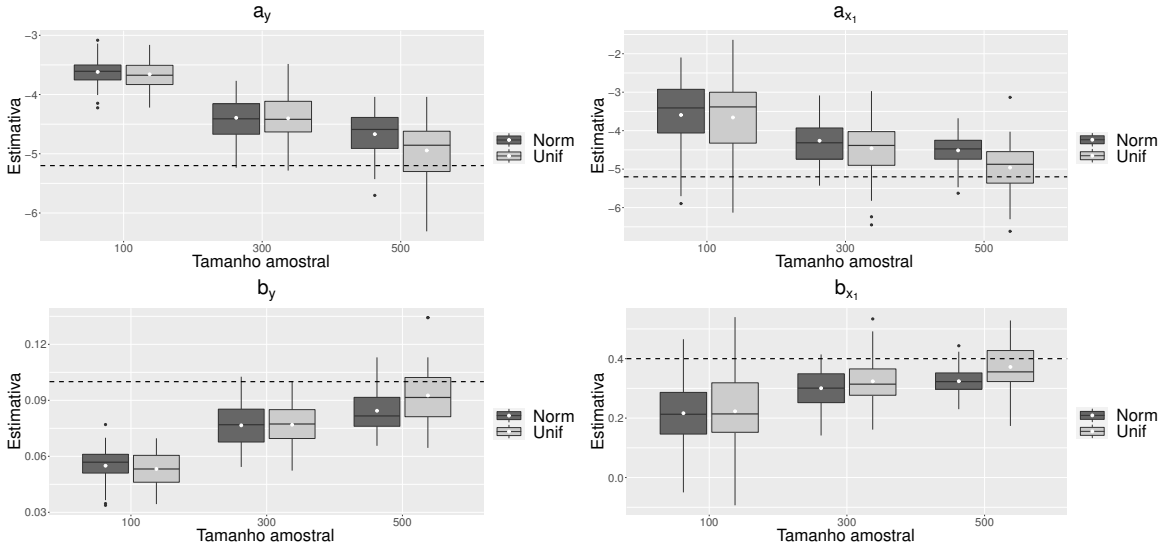

Figura 36 – Boxplots das estimativas dos parâmetros da regressão logística no modelo FMR-SSL-MD, referente ao mecanismo MNAR, com taxa de dados ausentes de 30%, com diferentes distribuições para a covariável e tamanhos amostrais. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

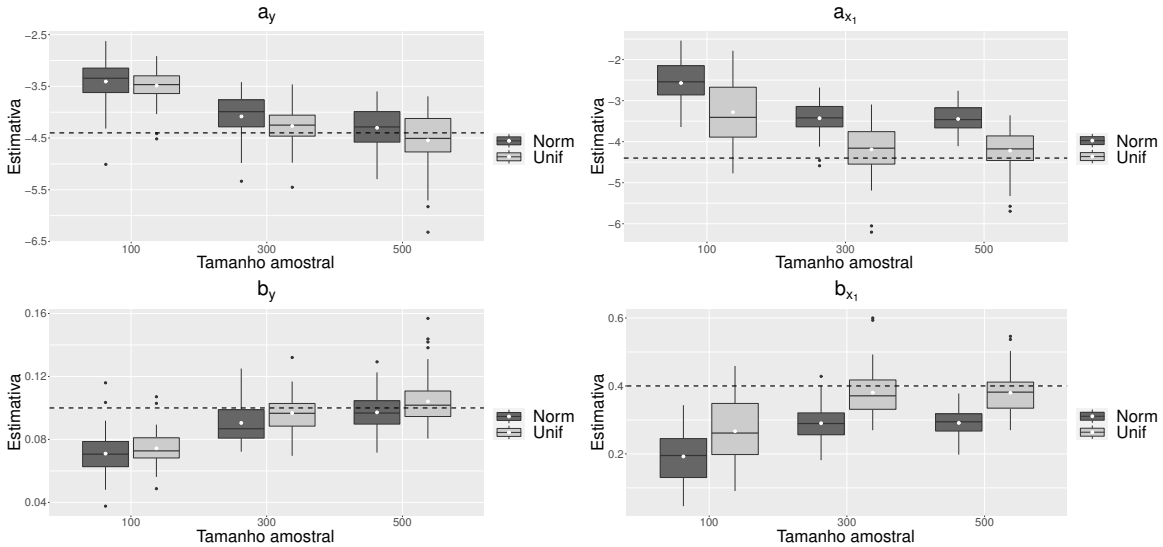

Figura 37 – Boxplots das estimativas dos parâmetros da regressão logística no modelo FMR-SSL-MD, referente ao mecanismo MNAR, com taxa de dados ausentes de 50%, com diferentes distribuições para a covariável e tamanhos amostrais. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

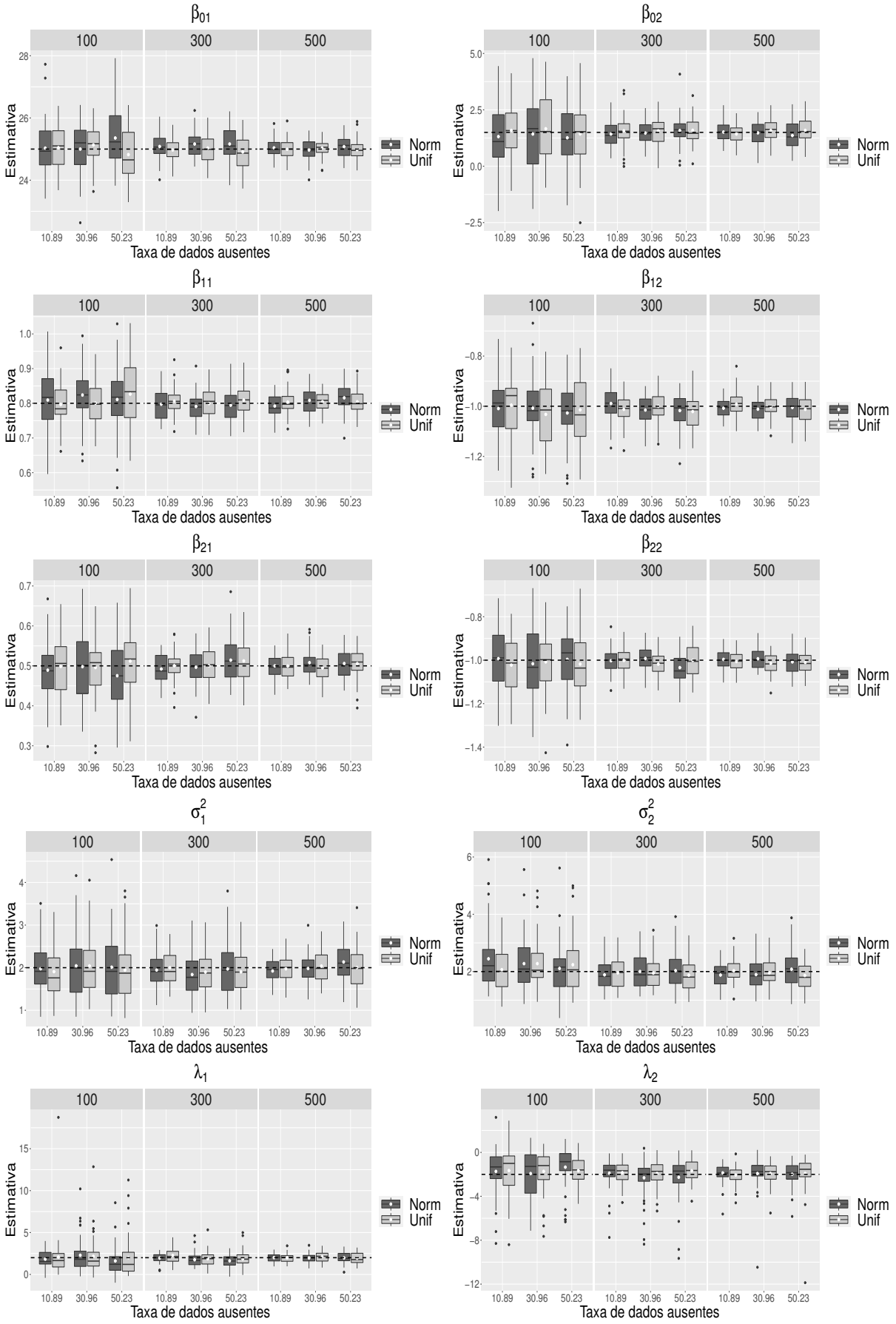

Figura 38 – Boxplots das estimativas dos parâmetros para o modelo FMR-SCN-MD, comparando diferentes distribuições para a covariável, diferentes tamanhos amostrais e diferentes taxas de dados ausentes. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

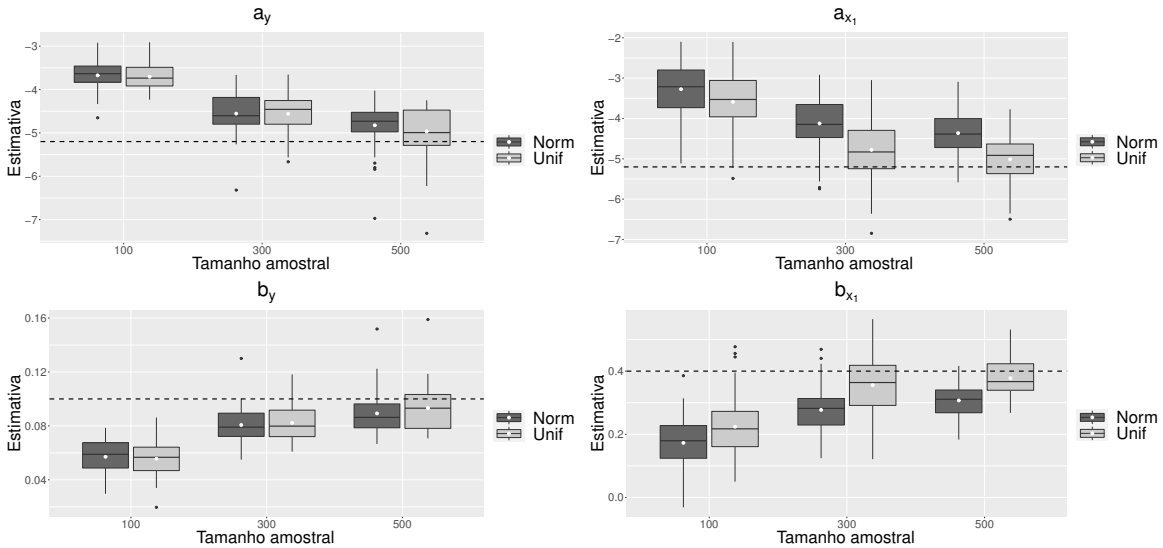

Figura 39 – Boxplots das estimativas dos parâmetros da regressão logística no modelo FMR-SCN-MD, referente ao mecanismo MNAR, com taxa de dados ausentes de 30%, com diferentes distribuições para a covariável e tamanhos amostrais. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

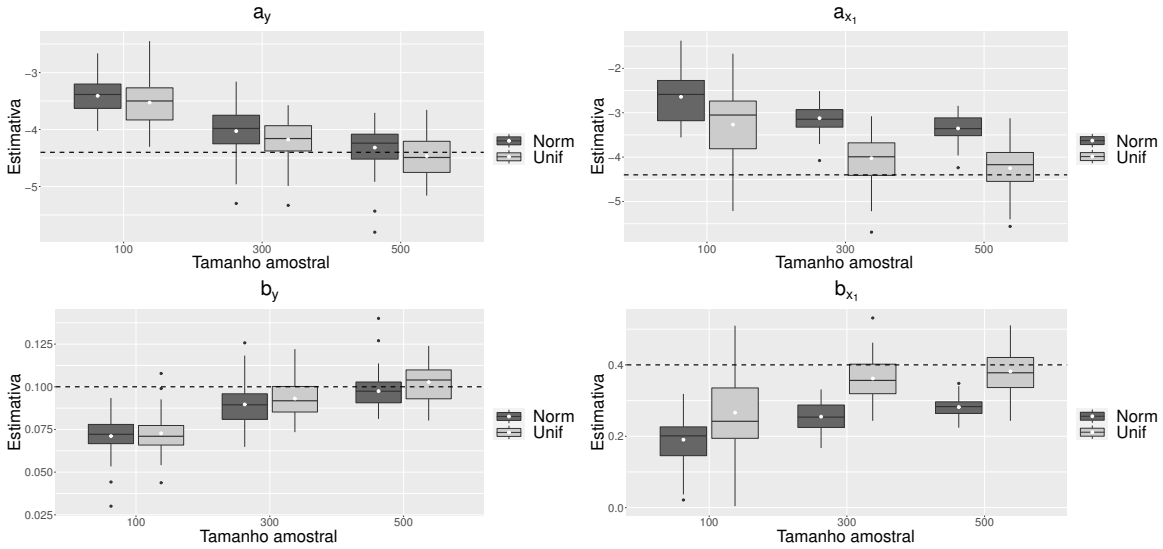

Figura 40 – Boxplots das estimativas dos parâmetros da regressão logística no modelo FMR-SCN-MD, referente ao mecanismo MNAR, com taxa de dados ausentes de 50%, com diferentes distribuições para a covariável e tamanhos amostrais. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

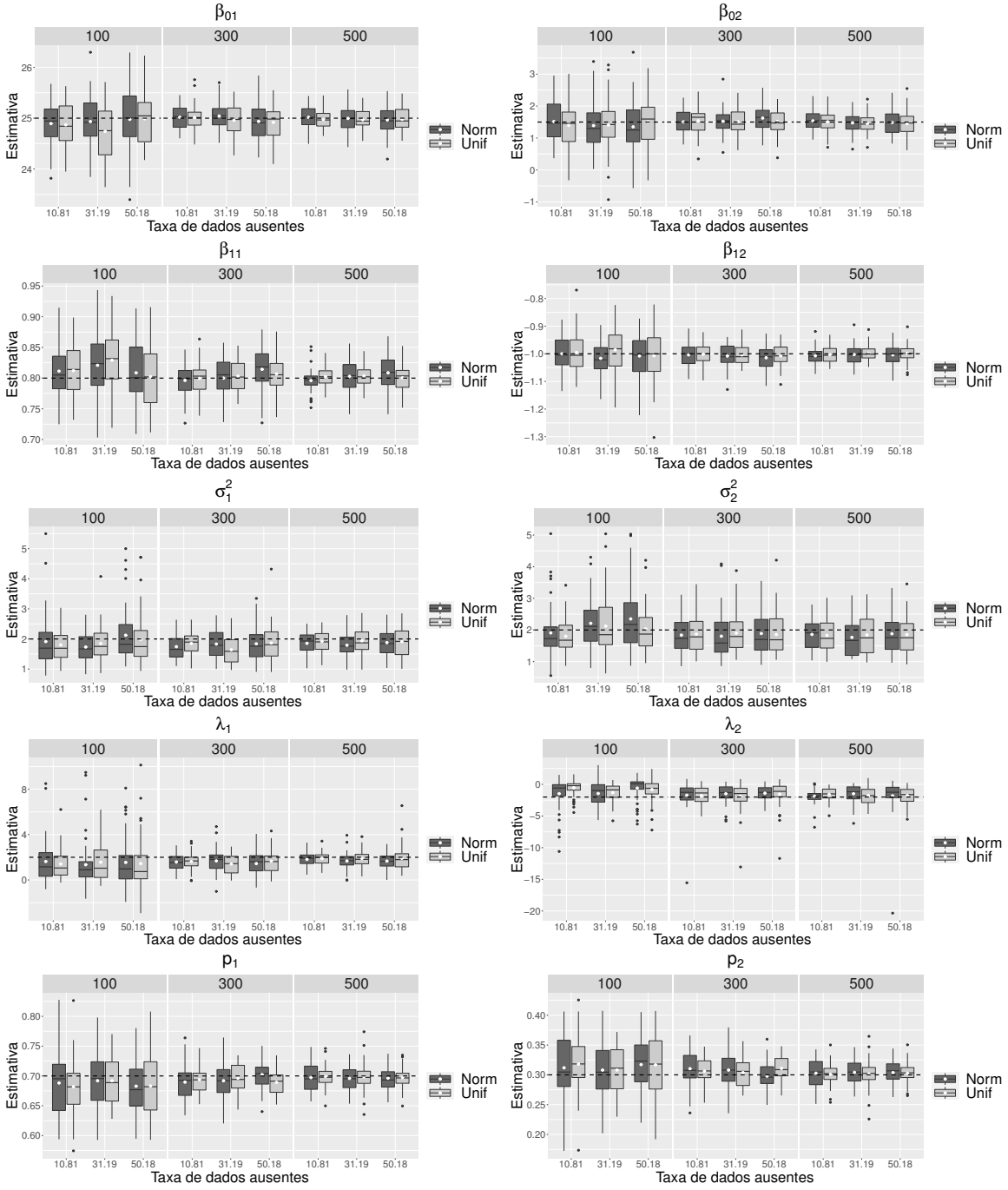

Figura 41 – Boxplots das estimativas dos parâmetros para o modelo FMR-SN-MD, comparando diferentes distribuições para a covariável, diferentes tamanhos amostrais e diferentes taxas de dados ausentes. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

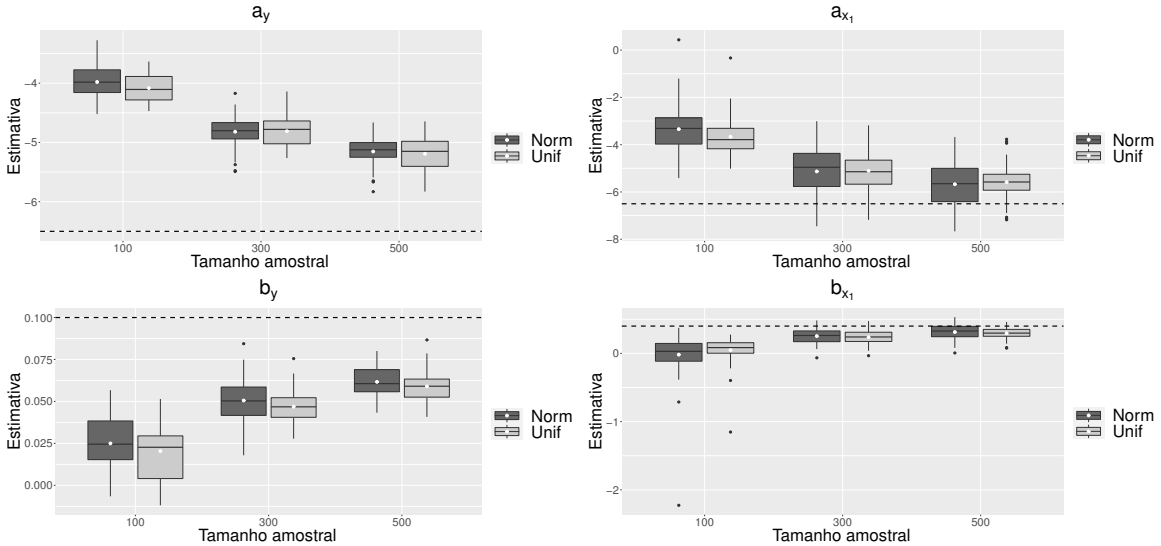

Figura 42 – Boxplots das estimativas dos parâmetros da regressão logística no modelo FMR-SN-MD, referente ao mecanismo MNAR, com taxa de dados ausentes de 9.699%, com diferentes distribuições para a covariável e tamanhos amostrais. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.

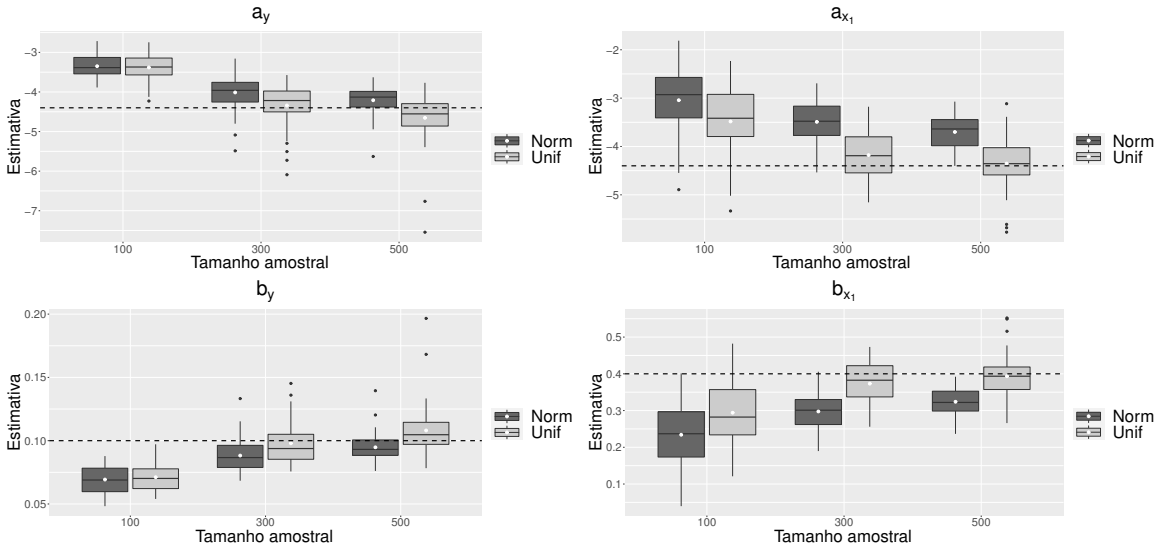

Figura 43 – Boxplots das estimativas dos parâmetros da regressão logística no modelo FMR-SN-MD, referente ao mecanismo MNAR, com taxa de dados ausentes de 50%, com diferentes distribuições para a covariável e tamanhos amostrais. A linha pontilhada corresponde ao verdadeiro valor do parâmetro. A média das estimativas é dada por um ponto branco dentro dos boxplots.