UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

MARIA YESSENIA ALVAREZ GIL

MODELOS DE REGRESSÃO LINEAR PARA DADOS INCOMPLETOS UTILIZANDO DISTRIBUIÇÕES ASSIMÉTRICAS

Recife

2024

MARIA YESSENIA ALVAREZ GIL

MODELOS DE REGRESSÃO LINEAR PARA DADOS INCOMPLETOS UTILIZANDO DISTRIBUIÇÕES ASSIMÉTRICAS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador: Prof. Dr. Aldo William Medina Garay

Coorientador: Prof. Dr. Víctor Hugo Lachos Dávila

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Gil, Maria Yessenia Álvarez.
Modelos de regressão linear para dados incompletos utilizando distribuições assimétricas / Maria Yessenia Álvarez Gil. - Recife, 2024.
70f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Programa de Pós-graduação em Estatística, 2024.
Orientação: Aldo William Medina Garay.
Coorientação: Víctor Hugo Lachos Dávila.

1. Modelos de regressão censurados; 2. Distribuições de caudas pesadas; 3. Algoritmo ECME; 4. Distribuições de misturas de escala da distribuição normal assimétrica. I. Garay, Aldo William Medina. II. Dávila, Víctor Hugo Lachos. III. Título.

UFPE-Biblioteca Central

MARIA YESSENIA ALVAREZ GIL

Modelos de Regressão Linear para Dados Incompletos utilizando Distribuições Assimétricas

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestra em Estatística.

Aprovada em 23 de julho de 2024.

BANCA EXAMINADORA

Prof. Dr. Aldo William Medina Garay
Presidente/Orientador, UFPE

Prof. Dr. Francisco José de Azevedo Cysneiros
Examinador Interno

Prof. Dr. Celso Rômulo Barbosa Cabral
Examinador Externo à Instituição, UFAM

Com muito amor, à minha mãe Andrea, ao meu pai Julián
e aos meus irmãos Yeferson e Julián Andrés.

AGRADECIMENTOS

Meu mais profundo agradecimento a todas as pessoas que contribuíram de forma significativa neste processo. Sobretudo, agradeço a Deus pela sua orientação e fortaleza.

De maneira especial, quero agradecer:

Aos meus queridos pais, Andrea e Julián, pelo amor e apoio incondicional. Vocês são grande parte da motivação para alcançar minhas metas.

Aos meus irmãos, por compartilharem comigo momentos de alegria e estarem sempre presentes. Obrigada por sempre acreditarem em mim.

Aos meus orientadores, Dr. Aldo William Medina Garay e Dr. Víctor Hugo Lachos, pela oportunidade de trabalhar sob sua orientação. Sua experiência, compreensão e paciência foram fundamentais para o desenvolvimento deste trabalho. Agradeço profundamente sua confiança, dedicação e valiosos conselhos, que me permitiram crescer academicamente e profissionalmente.

À Dra. Francyelle de Lima Medina, pelo seu tempo, dedicação e ajuda na revisão e correção deste trabalho, especialmente no que se refere ao idioma.

Aos professores do CCEN, com quem tive a oportunidade de aprender e crescer academicamente.

Às minhas avós Isabel e Victoria, e aos meus tios, especialmente ao Carlos, pelo amor, conselhos, apoio emocional e palavras de incentivo.

Aos meus amigos colombianos, tanto os que estiveram presentes quanto aqueles que me acompanharam à distância, pela companhia, amizade e apoio em cada etapa deste longo e desafiador caminho. Compartilhar com vocês momentos de alegria e estresse foi fundamental para me manter motivada e focada em alcançar minha meta.

Aos meus amigos e colegas do Brasil, pelo caloroso acolhimento neste país. Seu tempo, boa disposição e ajuda para me adaptar à nova cultura e estilo de vida fizeram da minha estadia uma etapa enriquecedora e cheia de experiências inesquecíveis. Sua amabilidade e companheirismo transformaram o Brasil em um segundo lar para mim.

À Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) pelo apoio financeiro

RESUMO

As distribuições de misturas de escala da distribuição normal assimétrica (SMSN) são uma classe de distribuições assimétricas com caudas pesadas, que inclue distribuições como a normal assimétrica, t de Student assimétrica e normal contaminada assimétrica. Este trabalho propõe um modelo de regressão linear, com censura intervalar, supondo que os erros seguem distribuições da classe SMSN, o que oferece uma alternativa mais flexível aos modelos de regressão censurados tradicionais que assumem distribuição normal para os erros. Implementamos um algoritmo para a estimação dos parâmetros via maximização condicional da função de verossimilhança (ECME), que apresenta expressões analíticas para o passo E. Essas expressões se baseiam em fórmulas para a média e variância de distribuições de misturas de escala da distribuição normal assimétrica truncadas, que podem ser calculadas numericamente utilizando o pacote `MomTrunc` disponível no software `R`. Ilustramos a aplicação e adequação da metodologia proposta por meio de estudos de simulação e análise de três conjuntos de dados reais.

Palavras-chaves: Modelos de regressão censurados. Distribuições de caudas pesadas. Algoritmo ECME. Distribuições de misturas de escala da distribuição normal assimétrica.

ABSTRACT

The scale mixture of skew-normal (SMSN) distributions is a class of asymmetric distributions with heavy tails, which includes distributions such as skew-normal, skew-t, and skew-contaminated normal. This work proposes a linear regression model with interval censoring, assuming that errors follow distributions from the SMSN class, resulting in more robust and flexible models than censored regression models that assume normal distribution for errors. We implemented an algorithm for parameter estimation via conditional maximization of the likelihood function (ECME), which provides analytical expressions for the E step. These expressions are based on formulas for the mean and variance of truncated scale mixtures of skew-normal distributions, which can be computed using the `MomTrunc` package available in the `R` software. We illustrate the application and adequacy of the proposed methodology through simulation studies and analysis of three real data sets.

Keywords: Censored regression models. Heavy-tailed distributions. ECME algorithm. Scale mixtures of skew-normal distributions.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Em diferentes áreas de pesquisa, surge o desafio de estimar os parâmetros de um modelo de regressão, quando a variável de interesse não está completamente observada. Este cenário pode ocorrer em diversas áreas de estudo, como por exemplo em econometria, na qual por motivos de confidencialidade, os dados de renda são informados até um certo limite; engenharia, em que os testes de fadiga de materiais frequentemente resultam em dados censurados, pois a detecção de uma falha é restrita a um intervalo entre períodos de inspeção; em medicina, em alguns estudos clínicos, os dados podem ser limitados pelas restrições dos equipamentos para medir abaixo ou acima de certos limites; entre outras situações. Esses estudos apresentam grandes desafios e destacam a necessidade de desenvolver modelos estatísticos robustos e eficientes para analisar esse tipo de dados.

Na literatura estatística, o modelo de regressão censurada (CR) tem se destacado devido à sua diversidade de aplicações. Uma suposição comum nos modelos CR para dados contínuos é que os erros aleatórios seguem uma distribuição Normal (N-CR). No entanto, é amplamente discutido que essa distribuição é sensível a valores atípicos. Consequentemente, nos últimos anos, foram propostos vários modelos paramétricos, que proporcionam alternativas na modelagem de dados censurados. Por exemplo, Arellano-Valle et al. (2012) propuseram o uso da distribuição t de Student em modelos de regressão truncada, Massuia et al. (2015), desenvolveram medidas diagnósticas para modelos de regressão censurados, utilizando a mesma distribuição, incluindo a implementação de um algoritmo do tipo EM para a estimação de máxima verossimilhança. Posteriormente, Garay et al. (2015) e Garay et al. (2017) propuseram um modelo CR no qual os erros seguem uma distribuição de mistura de escala normal (SMN-CR), sob os enfoques bayesianos e frequentistas, respectivamente. Esta classe de distribuições simétricas, proposta por Andrews e Mallows (1974), inclui casos especiais como a distribuição normal, a distribuição t de Student, a distribuição normal contaminada, entre outras.

Por outro lado, a classe de distribuições de misturas de escala da distribuição normal assimétrica (SMSN), desenvolvida por Branco e Dey (2001), tem atraído crescente atenção nos últimos anos devido à sua capacidade de capturar simultaneamente assimetria e alta curtose. Esta classe inclui como casos especiais a classe SMN e as distribuições normal assimétrica (skew-normal, SN), t de Student assimétrica (skew-t, ST) e normal contaminada assimétrica

(skew-normal contaminada, SCN), entre outras, tornando-se assim uma opção mais razoável para a inferência. Sob enfoque bayesiano, Massuia et al. (2017) propuseram um algoritmo de Monte Carlo via cadeias de Markov (MCMC), baseado no amostrador de Gibbs para estimar os parâmetros dos modelos CR sob a classe de distribuições SMSN (SMSN-CR), o qual está implementado no pacote **BayesCR** (GARAY; MASSUIA; LACHOS., 2017) no software estatístico R (R Core Team, 2024). Além disso, sob o enfoque da inferência baseada na verossimilhança, Mattos, Garay e Lachos (2018) propuseram um algoritmo eficiente de Monte Carlo EM (MCEM) para calcular estimadores de máxima verossimilhança dos parâmetros do modelo SMSN-CR, utilizando a aproximação estocástica do algoritmo EM (SAEM). No entanto, os algoritmos SAEM e o amostrador de Gibbs são custosos em termos computacionais, devido à combinação de simulações de Monte Carlo (MC) e outros processos iterativos, o que torna difícil a avaliação da convergência e requer o uso de uma quantidade considerável de tempo. Recentemente, Lachos et al. (2022) estudaram os modelos CR baseados na distribuição t de Student assimétrica (ST-CR) e demonstraram a robustez do modelo ST-CR frente à assimetria e caudas pesadas (alta curtose), utilizando o algoritmo EM para obter estimadores de máxima verossimilhança, os quais se baseiam no cálculo dos primeiros dois momentos da distribuição t de Student assimétrica truncada.

Neste trabalho, propomos um algoritmo do tipo EM alternativo e analiticamente simples, para calcular os estimadores de máxima verossimilhança dos parâmetros do modelo SMSN-CR. Seguindo a abordagem de Lachos et al. (2022), mostramos que o passo E se reduz a calcular os primeiros dois momentos das distribuições SMSN truncadas, com parâmetros específicos. As fórmulas gerais para estes momentos foram desenvolvidas recentemente por Lachos, Garay e Cabral (2020), e utilizaremos o pacote **MomTrunc** (GALARZA; KAN; LACHOS., 2022) disponível no software R, para sua implementação. Por outro lado, para calcular os erros padrão aproximados dos estimadores de máxima verossimilhança dos parâmetros, seguimos a estratégia utilizada por autores como Meilijson (1989), Lin (2010), Garay et al. (2017), entre outros, que se baseia na matriz de informação empírica.

Nossa proposta apresenta vantagens sobre os métodos de estimação existentes, especificamente em comparação com os algoritmos do amostrador de Gibbs, MCEM e SAEM, considerando que nosso algoritmo do tipo EM é mais eficiente em termos computacionais, pois obtém expressões analíticas nos passos E e M, reduzindo a necessidade de simulações extensivas.

Além disso, diferentemente do amostrador de Gibbs, nosso método elimina a necessidade de múltiplas iterações de simulações para alcançar a convergência e, em comparação com

o MCEM e o SAEM, evita a dependência de simulações de Monte Carlo e aproximações estocásticas no passo E, o que pode resultar em um menor número de iterações para atingir a convergência.

1.2 ORGANIZAÇÃO DO TRABALHO

Os resultados obtidos neste trabalho estão organizados em 4 capítulos da seguinte forma:

No segundo capítulo, apresentamos os conceitos fundamentais e algumas definições necessárias para compreender o modelo proposto.

No terceiro capítulo, propomos o modelo de regressão linear com censura intervalar sob a classe de distribuições SMSN, denotado por SMSN-ICR, abordando sua formulação matemática e os métodos de estimação dos parâmetros. Além disso, conduzimos dois estudos de simulação para avaliar a adequação do modelo proposto em diferentes cenários, analisando o desempenho dos estimadores dos parâmetros em amostras de tamanho finito e a precisão dos parâmetros estimados na presença de observações atípicas. Também apresentamos três exemplos práticos da análise para conjuntos de dados reais sobre rendimentos semanais, fluidez na nomeação de letras e taxas salariais de mulheres casadas. Essas aplicações reforçam a utilidade do modelo SMSN-ICR em diversas áreas.

No quarto capítulo, são apresentadas as conclusões finais por meio de uma síntese dos resultados obtidos. Adicionalmente, discutimos possíveis direções para pesquisas futuras.

2 CONCEITOS FUNDAMENTAIS

Neste capítulo, serão apresentados alguns conceitos e definições fundamentais para a compreensão do modelo proposto. Exploramos uma série de resultados que servirão de base para a formulação do modelo SMSN-ICR. Esses resultados abrangem tópicos como censura, truncamento, distribuições de misturas de escala da distribuição normal assimétrica (SMSN) e o algoritmo EM.

2.1 CENSURA E TRUNCAMENTO

De acordo com Cameron e Trivedi (2005), a censura e o truncamento são dois conceitos cruciais na análise de regressão, quando se trabalha com dados incompletos. A censura ocorre quando a informação sobre a variável dependente está parcialmente observada para alguns indivíduos, devido a restrições ou limitações específicas, como na medição ou na forma como os dados foram coletados, enquanto a informação das variáveis regressoras está completamente disponível. Um exemplo típico de censura é quando se coletam dados sobre rendimentos, mas para algumas pessoas essa informação é reportada apenas como valores superiores a um certo limite por razões de confidencialidade. Em comparação, o truncamento refere-se à perda de observações tanto da variável dependente quanto das regressoras para certos valores. Por exemplo, se em um estudo só se incluem pessoas de baixos rendimentos, as observações de indivíduos com maiores recursos econômicos são completamente perdidas na análise. Cameron e Trivedi (2005) destacam que o truncamento implica uma maior perda de informação comparado à censura, o que pode afetar significativamente as estimativas dos parâmetros dos modelos de regressão. Detalhes adicionais podem ser encontrados em Garay et al. (2015) e Garay et al. (2017).

Quando uma observação é censurada, apenas um intervalo pode ser observado no qual o verdadeiro valor da variável está contido. Este intervalo pode ser do tipo: $\lfloor a, \infty)$, indicando censura à direita; $(-\infty, a\rfloor$, representando censura à esquerda; e $\lfloor a, b\rfloor$, referente à censura intervalar, em que a e b são constantes conhecidas pertencentes aos números reais, com $a < b$ e $\lfloor a, b\rfloor$ indicando que cada extremo do intervalo pode ser aberto ou fechado. Por exemplo, em estudos de sobrevivência, se for observado que um paciente sobrevive além de um certo período de tempo, mas o tempo exato de sobrevivência não é registrado, tem-se censura à direita. Em

testes de detecção de temperatura, se o equipamento não pode medir valores abaixo de um determinado limite, o valor mínimo registrado será o limite inferior, representando uma censura à esquerda. Por fim, em alguns estudos sobre a extensão dos rios em uma determinada região, podem surgir restrições geográficas ou técnicas que impossibilitam a medição exata, indicando apenas que a extensão do rio está entre dois valores, resultando em censura intervalar.

Além da censura, é fundamental também compreender o conceito de dados faltantes, ou *missing data*, denotados aqui como NA, nos quais os valores observados não estão disponíveis para algumas variáveis dependentes, podendo ocorrer por diferentes razões, como erros de medição ou respostas omitidas. Neste trabalho, consideramos que os dados faltantes podem ser tratados como dados sujeitos à censura do tipo intervalar, em que seu verdadeiro valor está contido em um intervalo que varia de menos infinito a mais infinito, $(-\infty, \infty)$. Esta consideração permite que tanto a censura intervalar quanto os dados faltantes sejam tratados dentro da mesma estrutura, tornando o modelo proposto aplicável em situações em que qualquer um desses fenômenos esteja presente, ou ambos simultaneamente. Além disso, supomos que esses dados estão ausentes de forma aleatória, seguindo a teoria descrita por Rubin (1976), sobre inferência estatística na presença de dados faltantes.

Neste trabalho, daremos ênfase a modelos para dados com censura intervalar, já que este inclui censura à esquerda, tomando $\lfloor a, b \rfloor = (-\infty, b\rfloor$ e à direita com $\lfloor a, b \rfloor = \lfloor a, \infty)$, além de dados faltantes em que $\lfloor a, b \rfloor = (-\infty, \infty)$.

A seguir, apresentamos a definição de distribuição truncada, que será útil ao longo do estudo.

Definição 1. *Seja $X \sim D$, uma variável aleatória com distribuição D e suporte χ. Denotamos por $f(\cdot)$ e $F(\cdot)$ sua função de densidade de probabilidade (pdf) e sua função de distribuição acumulada (cdf), respectivamente. Se $\mathbb{P}(a < X < b) > 0$ com $a < b$, uma variável aleatória V segue uma distribuição D truncada no intervalo $\lfloor a, b \rfloor \subset \chi$, denotada por $V \sim TD$, se tiver a mesma distribuição que $X \,|\, X \in \lfloor a, b \rfloor$.*

Assim, a pdf da variável aleatória V é dada por:

$$f_V(v \,|\, \lfloor a, b \rfloor) = \frac{f(v)}{F(b) - F(a)} \mathbb{1}_{\lfloor a, b \rfloor}(v),$$

em que $\mathbb{1}_{\mathbb{A}}(\cdot)$ denota a função indicadora do conjunto $\mathbb{A}$, isto é, $\mathbb{1}_{\mathbb{A}}(v) = 1$ se $v \in \mathbb{A}$ e $\mathbb{1}_{\mathbb{A}}(v) = 0$ caso contrário.

2.2 A CLASSE DE DISTRIBUIÇÕES DE MISTURA DE ESCALA DA DISTRIBUIÇÃO NORMAL ASSIMÉTRICA (CLASSE SMSN)

No decorrer deste estudo, $X \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denota que a variável aleatória X segue uma distribuição normal p-variada, em que $\boldsymbol{\mu}$ é o vetor das médias e $\boldsymbol{\Sigma}$ é a matriz de variância–covariância, com pdf e cdf representadas por $\phi_p\left(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ e $\Phi_p(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, respetivamente. Além disso, $X \sim \mathrm{T}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ refere-se a uma variável aleatória X que segue uma distribuição t de Student p-variada com vetor de médias $\boldsymbol{\mu}$, matriz de escala $\boldsymbol{\Sigma}$ e ν graus de liberdade, neste caso sua pdf e cdf são representadas por $t_p\left(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right)$ e $\mathrm{T}_p\left(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right)$, respectivamente. No caso univariado, $p = 1$, omitimos o índice p, de forma que se $\mu = 0$ e $\sigma^2 = 1$, para o caso da normal, utilizamos $\phi(\cdot)$ e $\Phi(\cdot)$ para a pdf e cdf, respectivamente, e para o caso da t de Student, $t\left(\cdot \,|\, \nu\right)$ e $\mathrm{T}\left(\cdot \,|\, \nu\right)$.

Para compreender o modelo que será analisado posteriormente, é essencial fornecer uma breve descrição e destacar certas propriedades das distribuições normal assimétrica (SN), t de Student assimétrica (ST) e normal contaminada assimétrica (SCN). Estas pertencem à classe de distribuições de misturas de escala da distribuição normal assimétrica (SMSN), desenvolvida por Branco e Dey (2001), baseando-se na teoria apresentada por Azzalini (1985) para a distribuição normal assimétrica. É importante destacar que, dentro desta categoria, também encontramos a família de distribuições de mistura de escala normal (SMN), proposta por Andrews e Mallows (1974). Em relação à notação, usaremos $\psi_{SMSN}(\cdot)$ para nos referirmos a pdf e $\Psi_{SMSN}(\cdot)$ para a cdf das distribuições da classe SMSN.

Definição 2. *Uma variável aleatória Z segue uma distribuição normal assimétrica, se sua função de densidade de probabilidade é dada por:*

$$\psi_{SN}\left(z \,|\, \mu, \sigma^2, \lambda\right) = 2\phi(z \,|\, \mu, \sigma^2)\Phi\left(\frac{\lambda(z-\mu)}{\sigma}\right). \tag{2.1}$$

em que $\mu \in \mathbb{R}$ é o parâmetro de localização, $\sigma^2 > 0$ é o parâmetro de escala e $\lambda \in \mathbb{R}$ é o parâmetro de forma, relacionado à assimetria da distribuição. Denotamos essa distribuição por $Z \sim SN(\mu, \sigma^2, \lambda)$. Valores positivos de λ indicam assimetria à direita, enquanto valores negativos indicam assimetria à esquerda. Se $\mu = 0$ e $\sigma^2 = 1$, dizemos que Z segue uma distribuição normal assimétrica padrão com parâmetro de forma λ.

Definição 3. *Uma variável aleatória Y segue uma distribuição pertencente à família SMSN*

se ela pode ser escrita da seguinte forma:

$$Y = \mu + \kappa(U)^{1/2} Z, \quad U \perp Z, \tag{2.2}$$

em que μ é o parâmetro de locação, $Z \sim SN(0, \sigma^2, \lambda)$, $\kappa(\cdot)$ é uma função positiva, U é uma variável aleatória com função de distribuição $H(\cdot\,|\,\boldsymbol{\nu})$ e densidade $h(\cdot\,|\,\boldsymbol{\nu})$ e $\boldsymbol{\nu}$ é um escalar ou vetor de parâmetros que indexa a distribuição de U. $U \perp Z$ denota que as variáveis aleatórias U e Z são independentes. Neste caso, $Y \sim$ SMSN$(\mu, \sigma^2, \lambda; H)$.

A variável aleatória U é conhecida como o *fator de escala* e sua cdf $H(\cdot\,|\,\boldsymbol{\nu})$ é chamada de *função de distribuição de mistura*. A relação entre a classe SMSN e a distribuição normal assimétrica torna-se evidente por meio da Equação (2.2). Especificamente, a distribuição condicional de Y dado $U = u$, segue uma distribuição normal assimétrica,

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

Assim, a função de densidade marginal de Y, é dada por:

$$\psi_{\mathsf{SMSN}}(y\,|\,\mu, \sigma^2, \lambda; H) = 2 \int_0^\infty \phi(y\,|\,\mu, \kappa(u)\sigma^2) \Phi\left(\frac{\lambda(y-\mu)}{\sigma\kappa(u)^{1/2}}\right) dH(u\,|\,\boldsymbol{\nu}). \tag{2.3}$$

Quando $\lambda = 0$, a família de distribuições SMSN reduz-se à classe simétrica SMN. Nesse caso particular, a pdf é definida como:

$$\psi_{\mathsf{SMN}}(y\,|\,\mu, \sigma^2; H) = \int_0^\infty \phi(y\,|\,\mu, \kappa(u)\sigma^2)\, dH(u\,|\,\boldsymbol{\nu}).$$

Como apresentado por Basso et al. (2010) no contexto de misturas finitas de distribuições SMSN para dados completos, direcionamos nossa análise para o caso em que $\kappa(u) = 1/u$, já que várias distribuições da classe SMSN, especialmente as utilizadas neste trabalho, são obtidas mediante essa escolha. Além disso permite desenvolver algumas propriedades, como a função geradora de momentos, a esperança e a variância.

Por outro lado, seguindo Basso et al. (2010), existe outra representação estocástica para $Y \sim$ SMSN$(\mu, \sigma^2, \lambda; H)$, descrita como:

$$Y = \mu + \Delta T + u^{-1/2} \Gamma^{1/2} T_1, \tag{2.4}$$

em que $\Delta = \sigma\delta$, $\Gamma = \sigma^2(1 - \delta^2)$, $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$ e $T = u^{-1/2}\,|T_0|$. T_0 e T_1 são variáveis aleatórias normais padrão independentes e $|\cdot|$ denota valor absoluto.

Esta representação estocástica, demonstrada em Basso et al. (2010), é particularmente útil por várias razões. Em primeiro lugar, facilita a geração de números pseudoaleatórios, o

que é essencial para a realização de simulações numéricas e estudos de Monte Carlo. Em segundo lugar, permite a obtenção de propriedades analíticas, proporcionando uma compreensão mais profunda do comportamento dessas distribuições. Além disso, é crucial para a inferência estatística, especialmente na implementação do algoritmo EM.

É importante notar que, a partir da Equação (2.4), podemos obter a seguinte representação hierárquica:

$$\begin{aligned} Y \,|\, T = t, U = u &\sim \mathrm{N}\left(\mu + \Delta t\,,\, u^{-1}\Gamma\right), \\ T \,|\, U = u &\sim \mathrm{TN}\left(0, u^{-1}; [0,\infty)\right), \\ U &\sim \mathrm{H}\left(\cdot\,; \boldsymbol{\nu}\right), \end{aligned}$$

em que $X \sim \mathrm{TN}\left(0, u^{-1}; [0,\infty)\right)$ denota que a variável aleatória X segue uma distribuição normal truncada no intervalo $[0,\infty)$, na qual 0 é a média e u_i^{-1} é a variância da normal antes do truncamento.

A seguir, o Lema obtido em Massuia et al. (2017), apresenta a expressão da cdf da família de distribuições SMSN.

Lema 2.2.1. *Se $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$, então a cdf de Y pode ser escrita como:*

$$\Psi_{SMSN}\left(y \,|\, \mu, \sigma^2, \lambda; H\right) = \int_0^\infty 2\Phi_2\left(\boldsymbol{y}(u)^* \,|\, \boldsymbol{\omega}, \boldsymbol{\Sigma}\right) dH(u \,|\, \boldsymbol{\nu}),$$

$$em\ que\ \ \boldsymbol{y}(u)^* = \left(u^{1/2}y, 0\right)^\top, \quad \boldsymbol{\omega} = (\mu, 0)^\top \ \ e\ \ \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & -\delta\sigma \\ -\delta\sigma & 1 \end{pmatrix}. \tag{2.5}$$

Prova. *Veja o Apêndice A em Massuia et al. (2017).*

Conforme mencionado anteriormente, neste trabalho consideramos três membros da classe SMSN. A seguir, apresentamos cada uma dessas distribuições juntamente com sua pdf e cdf, obtidas a partir da Equação (2.3) e do Lema 2.2.1. É importante destacar que cada distribuição é determinada pela escolha do fator de escala U, que pode ser discreto ou contínuo. Adicionalmente, é apresentado o valor $k_m = E[U^{-m/2}]$, em que $m \in \mathbb{Z}$.

(i) A distribuição normal assimétrica

Denotada por $Y \sim \mathrm{SN}\left(\mu, \sigma^2, \lambda\right)$, a variável aleatória Y é obtida quando U segue uma distribuição degenerada em 1, com $\mathbb{P}(U = 1) = 1$. A pdf é dada pela Equação (2.1) e

a cdf é:

$$\Psi_{\mathrm{SN}}\left(y\,|\,\mu,\sigma^2,\lambda\right) = 2\Phi_2\left(\boldsymbol{y}^*\,|\,\boldsymbol{\omega},\boldsymbol{\Sigma}\right) \quad \text{e}$$
$$k_m = 1$$

(ii) A distribuição t de Student assimétrica

Representada por $Y \sim \mathrm{ST}\left(\mu,\sigma^2,\lambda,\nu\right)$, a variável aleatória Y é obtida quando $U \sim \mathrm{Gamma}(\nu/2,\nu/2),\ \nu>0$. A pdf e a cdf são expressas, respectivamente, por:

$$\begin{aligned}\psi_{\mathrm{ST}}\left(y\,|\,\mu,\sigma^2,\lambda,\nu\right) &= \frac{2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{d(y)^2}{\nu}\right)^{-\frac{\nu+1}{2}} \\ &\quad\times \mathrm{T}\left(\lambda\, d(y)\sqrt{\frac{\nu+1}{\nu+d(y)^2}}\,\middle|\,\nu+1\right), \\ \Psi_{\mathrm{ST}}\left(y\,|\,\mu,\sigma^2,\lambda,\nu\right) &= 2\mathrm{T}_2\left(\boldsymbol{y}^*\,|\,\boldsymbol{\omega},\boldsymbol{\Sigma},\nu\right) \quad \text{e} \\ k_m &= \left(\frac{\nu}{2}\right)^{m/2}\frac{\Gamma\left((\nu-m)/2\right)}{\Gamma\left(\nu/2\right)},\end{aligned}$$

em que $d(y) = (y-\mu)/\sigma$.

(iii) A distribuição normal contaminada assimétrica

Denotada por $Y \sim \mathrm{SCN}\left(\mu,\sigma^2,\lambda,\nu_1,\nu_2\right)$, neste caso a variável aleatória Y é obtida quando U é uma variável aleatória discreta que assume dois estados: ν_2 com probabilidade ν_1 e 1 com probabilidade $1-\nu_1$. Consequentemente, sua pdf e cdf são dadas, respectivamente, por:

$$\begin{aligned}\psi_{\mathrm{SCN}}\left(y\,|\,\mu,\sigma^2,\lambda,\boldsymbol{\nu}\right) &= 2\left\{\nu_1\phi(y\,|\,\mu,\nu_2^{-1}\sigma^2)\Phi(\nu_2^{1/2}A) + (1-\nu_1)\phi(y\,|\,\mu,\sigma^2)\Phi(A)\right\}, \\ \Psi_{\mathrm{SCN}}\left(y\,|\,\mu,\sigma^2,\lambda,\boldsymbol{\nu}\right) &= 2\left\{\nu_1\Phi_2\left(\nu_2^{1/2}\boldsymbol{y}^*\,|\,\boldsymbol{\omega},\boldsymbol{\Sigma}\right) + (1-\nu_1)\Phi_2\left(\boldsymbol{y}^*\,|\,\boldsymbol{\omega},\boldsymbol{\Sigma}\right)\right\} \quad \text{e} \\ k_m &= \nu_1\nu_2^{-m/2} + 1 - \nu_1,\end{aligned}$$

sendo $A = \lambda(y-\mu)/\sigma$ e $\boldsymbol{\nu} = (\nu_1,\nu_2)^\top$, com $0<\nu_1<1$ e $0<\nu_2<1$,

em que $\boldsymbol{y}^* = (y,0)^\top$, $\boldsymbol{\omega}$ e $\boldsymbol{\Sigma}$ são definidos na Equação (2.5).

A seguir, apresentamos um Lema que fornece a esperança e a variância de uma variável aleatória que segue uma distribuição pertencente à classe SMSN (BASSO et al., 2010).

Lema 2.2.2. *Seja* $Y \sim SMSN\left(\mu,\sigma^2,\lambda,H\right)$.

a) Se $k_1 = E[U^{-1/2}] < \infty$, *então* $E[Y] = \mu + \sqrt{\frac{2}{\pi}}k_1\Delta$;

b) Se $k_2 = E[U^{-1}] < \infty$, *então* $\mathrm{Var}[Y] = \sigma^2\left(k_2 - \frac{2}{\pi}k_1^2\delta^2\right)$

em que δ *e* Δ *são definidos como foi feito para a Equação* (2.4).

Prova. *Para a demonstração, utilizaremos a função geradora de momentos da variável aleatória* $X \sim SN(0,1,\lambda)$. *Dessa forma, primeiro vejamos que* $M_X(t) = 2e^{\frac{t^2}{2}}\Phi(\delta t)$, $t \in \mathbb{R}$.

$$\begin{aligned}
M_X(t) = E[e^{tX}] &= \int_{-\infty}^{\infty} 2e^{tx}\phi(x)\Phi(\lambda x)dx \\
&= 2\int_{-\infty}^{\infty} e^{\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-t)^2}\Phi(\lambda x)dx \\
&= 2e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\phi(w)\Phi(\lambda w + \lambda t)dw, \quad w = x - t, dw = dx \\
&= 2e^{\frac{t^2}{2}}E[\Phi(\lambda W + \lambda t)] \\
&= 2e^{\frac{t^2}{2}}\Phi(\lambda t\,|\,0, 1+\lambda^2) \qquad (2.6) \\
&= 2e^{\frac{t^2}{2}}\Phi(\delta t), \quad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}.
\end{aligned}$$

A Equação (2.6) é obtida a partir do Lema A.1 apresentado em Basso et al. (2010), considerando $\mathbf{a} = \lambda t$, $\mathbf{B} = \lambda$, $\boldsymbol{\mu} = \boldsymbol{\eta} = 0$ *e* $\boldsymbol{\Sigma} = \boldsymbol{\Omega} = 1$.

Desse resultado, tem-se que $E[X] = \sqrt{\frac{2}{\pi}}\delta$ *e* $\mathrm{Var}[X] = 1 - \frac{2}{\pi}\delta^2$ *e dado que* $Z = \mu + \sigma X \sim SN(\mu, \sigma^2, \lambda)$, *então* $E[Z] = \mu + \sigma\delta\sqrt{\frac{2}{\pi}}$ *e* $\mathrm{Var}[Z] = \sigma^2\left(1 - \frac{2}{\pi}\delta^2\right)$. *Finalmente, da representação dada na Equação* (2.2) *e como* $U \perp Z$, *obtêm-se* $E[Y]$ *e* $\mathrm{Var}[Y]$.

2.3 O ALGORITMO EM

O algoritmo EM "Expectation-Maximization" é um metodo iterativo introduzido por Dempster, Laird e Rubin (1977) utilizado para obter estimadores de máxima verossimilhança de parâmetros em modelos estatísticos, em particular para situações que envolvem dados faltantes ou incompletos. Cada iteração consiste em dois passos: um passo de esperança (Passo-E) e um passo de maximização (Passo-M).

O algoritmo EM tem sido amplamente utilizado em diversas aplicações; no entanto, sua implementação prática pode variar. Específicamente, quando o Passo-M se torna analiticamente complexo, é comum recorrer a extensões como ECM (MENG; RUBIN, 1993), ECME (LIU; RUBIN, 1994) ou SAEM (DELYON; LAVIELLE; MOULINES, 1999). Neste trabalho, utilizamos o algoritmo ECME, devido à sua eficiência computacional, rápida convergência, flexibilidade na estrutura do modelo e facilidade de implementação. O processo pode ser resumido da seguinte forma:

Inicialmente, consideremos o vetor de parâmetros de interesse $\boldsymbol{\theta}$, que pertence ao espaço paramétrico Θ. Seja $\mathbf{y}_{\mathsf{c}} = (\mathbf{y}_{\mathsf{obs}}, \mathbf{y}_{\mathsf{mis}})$ o vetor de dados completos, em que $\mathbf{y}_{\mathsf{mis}}$ representa os dados não observados ou faltantes e $\mathbf{y}_{\mathsf{obs}}$ os dados observados. A função de log-verossimilhança dos dados completos é denotada por $\ell_{\mathsf{c}}(\boldsymbol{\theta} \mid \mathbf{y}_{\mathsf{c}})$. Assim, considerando $\widehat{\boldsymbol{\theta}}^{(t)}$ como o estimador de $\boldsymbol{\theta}$ na iteração t do algoritmo, com $t = 0, 1, 2, \ldots$, e $\widehat{\boldsymbol{\theta}}^{(0)}$ sendo um valor inicial, na iteração $(t+1)$ temos os seguintes dois passos:

Passo-E Calcular a esperança condicional da função de log-verossimilhança dos dados completos com respeito à distribuição condicional de $\mathbf{y}_{\mathsf{mis}}$ dado $\mathbf{y}_{\mathsf{obs}}$ e o parâmetro estimado $\widehat{\boldsymbol{\theta}}^{(t)}$, ou seja,

$$Q\left(\boldsymbol{\theta} \,|\, \widehat{\boldsymbol{\theta}}^{(t)}\right) = \mathrm{E}\left[\ell_{\mathsf{c}}\left(\boldsymbol{\theta} \,|\, \mathbf{y}_{\mathsf{c}}\right) \,|\, \mathbf{y}_{\mathsf{obs}}, \widehat{\boldsymbol{\theta}}^{(t)}\right], \tag{2.7}$$

Passo-M Obter $\widehat{\boldsymbol{\theta}}^{(t+1)}$ que maximiza $Q\left(\boldsymbol{\theta} \,|\, \widehat{\boldsymbol{\theta}}^{(t)}\right)$ em relação a $\boldsymbol{\theta}$, ou seja,

$$Q\left(\widehat{\boldsymbol{\theta}}^{(t+1)} \,|\, \widehat{\boldsymbol{\theta}}^{(t)}\right) \geq Q\left(\boldsymbol{\theta} \,|\, \widehat{\boldsymbol{\theta}}^{(t)}\right) \quad \text{para todo} \quad \boldsymbol{\theta} \in \Theta.$$

No algoritmo ECME, o passo-M é substituído por vários passos de maximização condicional (CM), que são computacionalmente mais simples. Esses passos maximizam a expressão apresentada na Equação (2.7) (passo-CMQ) ou a função de verossimilhança $\ell_{\mathsf{c}}(\boldsymbol{\theta} \,|\, \mathbf{y}_{\mathsf{c}})$ (passo-CML).

Estes dois passos são repetidos alternadamente até que alguma medida de convergência seja satisfeita. Por exemplo, $\left|\ell(\boldsymbol{\theta}^{(k+1)} \,|\, \mathbf{y}_{\mathsf{obs}}) - \ell(\boldsymbol{\theta}^{(k)} \,|\, \mathbf{y}_{\mathsf{obs}})\right|$ ou $\left|\ell(\boldsymbol{\theta}^{(k+1)} \,|\, \mathbf{y}_{\mathsf{obs}}) / \ell(\boldsymbol{\theta}^{(k)} \,|\, \mathbf{y}_{\mathsf{obs}}) - 1\right|$ atinja um valor abaixo de um determinado limiar predefinido.

Por outro lado, é importante destacar que o algoritmo ECME se destaca pela sua eficiência computacional em comparação com o algoritmo SAEM implementado em Mattos, Garay e Lachos (2018). Isso ocorre porque o SAEM é uma abordagem estocástica que combina simulação de Monte Carlo com outros procedimentos iterativos.

Para mais detalhes sobre essas e outras extensões e variações do algoritmo EM, consulte Meng e Dyk (1997), os quais apresentaram um arcabouço teórico mais abrangente.

2.4 ANÁLISE DE RESÍDUOS

Na análise de modelos de regressão, é essencial avaliar a qualidade do ajuste do modelo e identificar possíveis valores atípicos que possam influenciar os resultados. Uma maneira eficaz de realizar essa avaliação é por meio da análise de resíduos.

No contexto específico de dados censurados, de acordo com Ortega, Bolfarine e Paula (2003a) (ver mais detalhes em Therneau, Grambsch e Fleming (1990)) , tem-se que os *resíduos martingale*, r_{M_i}, são definidos da seguinte forma:

$$r_{M_i} = \rho_i + \log(S(y_i; \widehat{\boldsymbol{\theta}})),$$

em que ρ_i é uma variável indicadora que assume o valor 0 se a observação é censurada e 1 caso contrário. $S(y_i, \widehat{\boldsymbol{\theta}}) = P(Y_i > y_i)$ representa a função de sobrevivência avaliada em y_i, e $\widehat{\boldsymbol{\theta}}$ é o estimador do vetor de parâmetros de interesse $\boldsymbol{\theta}$. No entanto, é importante observar que esses resíduos possuem uma distribuição assimétrica, assumindo valores no intervalo $(-\infty, +1]$, o que pode dificultar a interpretação e análise. Assim, para lidar com essa assimetria, Therneau, Grambsch e Fleming (1990) propuseram a seguinte transformação:

$$r_{MT_i} = \text{sign}(r_{M_i})\sqrt{-2[r_{M_i} + \rho_i \log(\rho_i - r_{M_i})]},$$

para $i = 1, \ldots, n$, em que $\text{sign}(r_{M_i})$ denota o sinal do resíduo martingale, r_{M_i}.

Essa transformação, motivada pelos resíduos de desvio utilizados em modelos lineares generalizados (MCCULLAGH; NELDER, 1989), tem como objetivo obter resíduos distribuídos simetricamente ao redor de zero, tornando mais fácil a identificação de valores atípicos e a validação dos pressupostos do modelo. Em particular, como em Garay et al. (2017), Lachos et al. (2022), Mattos, Garay e Lachos (2018) e outros, utilizamos essa abordagem para o modelo de regressão linear com censura utilizando a classe de distribuições SMSN.

Finalmente, é importante ressaltar que, devido à falta de independência e normalidade da transformação dos resíduos martingale r_{MT_i}, uma prática comum é adicionar envelopes nos gráficos de probabilidade normal correspondentes, conforme sugerido por Atkinson (1981).

2.5 CRITÉRIOS PARA COMPARAR MODELOS

Ao realizar uma análise comparativa de diversos modelos, é crucial empregar critérios que nos permitam avaliar e selecionar o modelo mais adequado para descrever o comportamento dos dados observados. Nesse contexto, recorremos a uma variedade de critérios de informação que são amplamente reconhecidos na literatura científica.

Suponha que temos n observações de um modelo específico, no qual $\widehat{\boldsymbol{\theta}}$ representa o estimador do vetor de parâmetros de interesse e $\ell(\boldsymbol{\theta})$ denota a função de log-verossimilhança desse modelo. Para a comparação entre os diferentes modelos considerados neste estudo, utilizamos o Critério de Informação de Akaike (AIC) (AKAIKE, 1974), o Critério de Informação

Bayesiano (BIC) (SCHWARZ et al., 1978), o AIC consistente (CAIC) (BOZDOGAN, 1987) e o Critério de Informação Hannan-Quinn (HQIC) (BURNHAM; ANDERSON, 2002). Estes critérios são definidos por:

$$\begin{aligned}
\mathrm{AIC} &= -2\ell(\widehat{\boldsymbol{\theta}}) + 2\rho, \\
\mathrm{BIC} &= -2\ell(\widehat{\boldsymbol{\theta}}) + \rho\log(n), \\
\mathrm{CAIC} &= -2\ell(\widehat{\boldsymbol{\theta}}) + \rho\left(\log(n)+1\right) \quad \text{e} \\
\mathrm{HQIC} &= -2\ell(\widehat{\boldsymbol{\theta}}) + 2\rho\log\left(\log(n)\right),
\end{aligned}$$

em que ρ representa o número de parâmetros livres no modelo. Valores mais baixos desses critérios indicam um melhor ajuste do modelo aos dados observados, demonstrando uma maior capacidade do modelo de representar a estrutura subjacente dos dados.

Adicionalmente, é importante mencionar que o AIC e o BIC são amplamente utilizados devido à sua simplicidade e eficácia na comparação de modelos. No entanto, existem situações em que outros critérios, como o CAIC e o HQIC, podem ser também adequados ou fornecer informações adicionais. Por esses motivos, e seguindo Lachos et al. (2022), incluímos o CAIC e o HQIC para realizar uma comparação mais abrangente entre os modelos considerados e contrastar alguns resultados. Isso garante que diferentes aspectos e complexidades dos modelos sejam adequadamente avaliados, proporcionando uma análise mais completa.

A seguir, apresentamos o modelo de regressão linear com censura intervalar sob a classe de distribuições SMSN, detalhando suas características principais e algumas aplicações.

3 MODELO SMSN-ICR

Neste capítulo, apresentamos o modelo de regressão linear com censura intervalar sob a classe de distribuições SMSN, denominado SMSN-ICR. Abordaremos a estimação dos parâmetros do modelo utilizando o algoritmo do tipo EM, o cálculo dos erros padrão dos estimadores de máxima verossimilhança, além de dois estudos de simulação para avaliar a adequação do modelo proposto em diferentes cenários. Também serão apresentados três exemplos práticos de análise utilizando conjuntos de dados reais.

3.1 O MODELO

Considere o modelo de regressão linear com distribuições SMSN da seguinte forma:

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} \text{SMSN}\left(-\sqrt{\frac{2}{\pi}}k_1\Delta, \sigma^2, \lambda; H\right),\ i = 1, \ldots, n, \tag{3.1}$$

em que Y_i é a variável resposta, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ é o vetor de parâmetros de regressão e $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ é o vetor de variáveis explicativas. Como definido por Basso et al. (2010) e Garay, Lachos e Abanto-Valle (2011), consideramos que o parâmetro de localização do erro aleatório seja igual a $-\sqrt{\frac{2}{\pi}}k_1\Delta$, de modo que $E[\varepsilon_i] = 0$. Assim, temos que:

$$Y_i \sim \text{SMSN}\left(\mathbf{x}_i^\top \boldsymbol{\beta} + b_1\Delta, \sigma^2, \lambda; H\right), \tag{3.2}$$

em que $b_1 = -\sqrt{\frac{2}{\pi}}k_1$ e as expressões de k_1 para cada uma das distribuições SMSN são apresentadas na Seção 2.2.

Por outro lado, estamos interessados na situação em que ocorre censura intervalar na variável resposta Y_i, com $i \in \{1, \ldots, n\}$. Neste caso, os dados observados para o i-ésimo sujeito são representados por (C_i, v_i), em que C_i é o indicador de censura e v_i pode ser um intervalo ou um valor observado. Especificamente, quando a i-ésima observação é censurada $(C_i = 1)$, temos que $v_i = \lfloor v_{1i}, v_{2i} \rfloor$, o que significa que $v_{1i} \leq Y_i \leq v_{2i}$. Por outro lado, se a i-ésima observação não é censurada $(C_i = 0)$, então $v_i = Y_i$.

É importante notar que as observações com censura intervalar abrangem outros tipos, como censura à direita representada por $\lfloor v_{1i}, v_{2i} \rfloor = \lfloor v_{1i}, \infty)$, censura à esquerda quando $\lfloor v_{1i}, v_{2i} \rfloor = (-\infty, v_{2i} \rfloor$, e também inclue a presença de dados faltantes indicada por $\lfloor v_{1i}, v_{2i} \rfloor = (-\infty, \infty)$. O modelo (3.1)–(3.2), que incorpora censura intervalar nas observações, é definido como o modelo SMSN-ICR.

3.2 ESTIMAÇÃO DOS PARÂMETROS DESCONHECIDOS

Para desenvolver o algoritmo do tipo EM e obter o estimador de máxima verossimilhança do vetor de parâmetros $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \sigma^2, \lambda, \nu\right)^\top$, utilizamos a estratégia de dados aumentados, a qual inclui os dados observados juntamente com algumas variáveis latentes. Para isso, precisamos de uma representação do modelo em termos de estrutura de dados incompletos, ou seja:

Das Equações. (2.4) e (3.2), temos a seguinte estrutura hierárquica:

$$\begin{aligned} Y_i &\sim \mathrm{SMSN}\left(\mathbf{x}_i^\top\boldsymbol{\beta} + b_1\Delta, \sigma^2, \lambda; H\right), \\ Y_i \,|\, T_i = t_i, U_i = u_i &\sim \mathrm{N}\left(\mathbf{x}_i^\top\boldsymbol{\beta} + \Delta t_i\,,\, u_i^{-1}\Gamma\right), \\ T_i \,|\, U_i = u_i &\sim \mathrm{TN}\left(b_1, u_i^{-1}; [b_1, \infty)\right), \\ U_i &\sim \mathrm{H}\left(\cdot\,; \boldsymbol{\nu}\right), \end{aligned}$$

em que $\Gamma = \sigma^2\left(1-\delta^2\right)$, $\Delta = \sigma\delta$, $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$ e $\mathrm{TN}\left(b_1, u_i^{-1}; [b_1, \infty)\right)$ representa uma distribuição normal truncada no intervalo $[b_1, \infty)$, na qual b_1 é a média e u_i^{-1} é a variância da normal antes do truncamento.

Definindo $\mu_i^* = \mathbf{x}_i^\top\boldsymbol{\beta} + b_1\Delta$ e utilizando o resultado obtido em Basso et al. (2010), conforme implementado em Garay, Lachos e Abanto-Valle (2011), Mattos, Garay e Lachos (2018), tem-se a distribuição condicional de T_i, dado Y_i e U_i, representada por $T_i \,|\, Y_i = y_i, U_i = u_i \sim \mathrm{TN}\left(\mu_{T_i} + b_1, u_i^{-1}M_T^2; \lfloor b_1, \infty)\right)$, em que μ_{T_i} e M_T^2 são definidos por:

$$\mu_{T_i} = \frac{\Delta}{\Delta^2 + \Gamma}(y_i - \mu_i^*) \quad \text{e} \quad M_T^2 = \frac{\Gamma}{\Delta^2 + \Gamma}.$$

Suponha que a amostra observada $\mathbf{y}_{\text{obs}}$, de tamanho $n_0 = n - m$, em que m representa a quantidade de dados faltantes, seja dividida em dois subconjuntos. Um desses subconjuntos contém p dados censurados, enquanto o outro compreende $n_0 - p$ dados não censurados. Dessa forma, $\mathbf{y}_{\text{obs}} = \{\mathbf{y}_{\text{cens}}, y_{p+1}, \ldots, y_{n_0}\}$, com $\mathbf{y}_{\text{cens}} = \{v_1, \ldots, v_p\}$ e $v_i = (v_{1i}, v_{2i})$, para $i = 1, \ldots, p$. A função de log-verossimilhança de $\boldsymbol{\theta}$, dada a amostra observada e denotada por $\ell(\boldsymbol{\theta} \,|\, \mathbf{y}_{\text{obs}})$, é expressa como:

$$\begin{aligned} \ell(\boldsymbol{\theta} \,|\, \mathbf{y}_{\text{obs}}) = \log\Bigg\{\prod_{i=1}^{n_0}\Bigg[\Psi_{SMSN}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda; H\right) - \Psi_{SMSN}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda; H\right)\Bigg]^{C_i} \\ \times \Big[\psi_{SMSN}(y_i \,|\, \mu_i^*, \sigma^2, \lambda; H)\Big]^{1-C_i}\Bigg\}, \end{aligned} \tag{3.3}$$

em que C_i é um indicador de censura, com $C_i = 1$ se a i-ésima observação for censurada, e $C_i = 0$ caso contrário.

Considerando que temos $q = p+m$ dados incompletos (censurados e faltantes), o aspecto fundamental para o desenvolvimento do nosso algoritmo do tipo EM é trabalhar com os dados completos $\mathbf{y}_c = \{\mathbf{y}_L, y_{p+1}, \ldots, y_{n_0}, u_1, \ldots, u_n, t_1, \ldots, t_n\}$, ou seja, tratamos o problema como se as variáveis latentes $\mathbf{Y}_L = \{Y_1, \ldots, Y_q\}$, $\mathbf{U} = \{U_1, \ldots, U_n\}$ e $\mathbf{T} = \{T_1, \ldots, T_n\}$ fossem observadas. Assim, a função de log-verossimilhança completa $\ell_c(\boldsymbol{\theta}\,|\,\mathbf{y}_c)$ pode ser expressa como:

$$\begin{aligned}\ell_c(\boldsymbol{\theta}\,|\,\mathbf{y}_c) = &-n\log\pi - \frac{n}{2}\log\Gamma + \sum_{i=1}^{n}\log u_i - \frac{1}{2\Gamma}\sum_{i=1}^{n} u_i\left(y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \Delta t_i\right)^2 \\ &- \frac{1}{2}\sum_{i=1}^{n} u_i(t_i - b_1)^2 + \sum_{i=1}^{n}\log h(u_i\,|\,\boldsymbol{\nu}),\end{aligned}$$

em que $h(\cdot\,|\,\boldsymbol{\nu})$ é a pdf da variável aleatória U.

A seguir, $\boldsymbol{\theta}^{(k)}$ representa o vetor dos parâmetros estimados na k-ésima iteração. O algoritmo do tipo EM para estimar os parâmetros pode ser apresentado por meio dos seguintes passos:

(i) **Passo-E**

Aqui, obtemos a função $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})$, dada por:

$$\begin{aligned}Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)}) &= E\left[\ell_c(\boldsymbol{\theta}\,|\,\mathbf{y}_c)\,|\,\mathbf{y}_{\mathsf{obs}}, \boldsymbol{\theta}^{(k)}\right] \\ &= -n\log\pi - \frac{n}{2}\log\Gamma - \frac{1}{2\Gamma}\sum_{i=1}^{n}\Big[\mathcal{E}_{02i}(\boldsymbol{\theta}^{(k)}) - 2\mathcal{E}_{01i}(\boldsymbol{\theta}^{(k)})\mathbf{x}_i^\top\boldsymbol{\beta} \\ &+ \mathcal{E}_{00i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_i^\top\boldsymbol{\beta})^2 - 2\Delta\mathcal{E}_{11i}(\boldsymbol{\theta}^{(k)}) + 2\Delta\mathcal{E}_{10i}(\boldsymbol{\theta}^{(k)})\mathbf{x}_i^\top\boldsymbol{\beta} + \Delta^2\mathcal{E}_{20i}(\boldsymbol{\theta}^{(k)})\Big] \\ &+ \sum_{i=1}^{n}\mathrm{E}\left[\log U_i\,|\,y_{\mathsf{obs}_i}, \boldsymbol{\theta}^{(k)}\right] + \sum_{i=1}^{n}\mathrm{E}\left[U_i(T_i - b_1)^2\,|\,y_{\mathsf{obs}_i}, \boldsymbol{\theta}^{(k)}\right] \\ &+ \sum_{i=1}^{n}\mathrm{E}\left[\log\{h(U_i\,|\,\boldsymbol{\nu})\}\,|\,y_{\mathsf{obs}_i}, \boldsymbol{\theta}^{(k)}\right].\end{aligned}$$

A expressão da função $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})$ é completamente determinada pelo conhecimento das seguintes esperanças:

$$\mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)}) = \mathrm{E}\left[U_i T_i^r Y_i^s\,|\,y_{\mathsf{obs}_i}, \boldsymbol{\theta}^{(k)}\right] \quad \text{e} \quad \mathrm{E}\left[\log\{h(U_i\,|\,\boldsymbol{\nu})\}\,|\,y_{\mathsf{obs}_i}\right],$$

em que $r, s \in \{0, 1, 2\}$. Para todas as distribuições da classe SMSN consideradas neste trabalho, essas expressões são obtidas de forma analítica, conforme detalhado a seguir:

- ***Para uma observação não censurada "i":***

  Neste caso, temos que $C_i = 0$, ou seja, $v_i = y_i$, assim para $r, s \in \{0, 1, 2\}$,

$$\mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)}) = y_i^s\mathcal{E}_{r0i}(\boldsymbol{\theta}^{(k)}),$$

em que, de acordo com Basso et al. (2010) e Garay et al. (2017), utilizando as propriedades da esperança condicional, temos que:

$$\mathcal{E}_{10i}(\boldsymbol{\theta}^{(k)}) = \mathcal{E}_{00i}(\boldsymbol{\theta}^{(k)})(\mu_{T_i}^{(k)} + b_1) + M_T^{(k)}\tau_i^{(k)},$$
$$\mathcal{E}_{20i}(\boldsymbol{\theta}^{(k)}) = \mathcal{E}_{00i}(\boldsymbol{\theta}^{(k)})(\mu_{T_i}^{(k)} + b_1)^2 + M_T^{2(k)} + M_T^{(k)}(\mu_{T_i}^{(k)} + 2b_1)\tau_i^{(k)},$$

em que

$$\tau_i^{(k)} = \mathrm{E}\left[U_i^{1/2} W_\Phi\left(\frac{U_i^{1/2}\mu_{T_i}^{(k)}}{M_T^{(k)}}\right) \mid y_i, \boldsymbol{\theta}^{(k)}\right] \quad \text{e} \quad W_\Phi(x) = \frac{\phi(x)}{\Phi(x)}, \text{ para } x \in \mathbb{R}.$$

As expressões $\mathcal{E}_{00i}(\boldsymbol{\theta}^{(k)})$ e $\tau_i^{(k)}$, sob as distribuições SMSN, foram obtidas de Basso et al. (2010) e são apresentadas no Apêndice A.

- ***Para uma observação incompleta "i":***

  Neste caso, temos que $C_i = 1$, ou seja $v_{1i} \leq Y_i \leq v_{2i}$, assim,

$$\mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)}) = \mathrm{E}[U_i T_i^r Y_i^s \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}^{(k)}], \quad \text{para} \;\; r, s \in \{0, 1, 2\},$$

  que podem ser obtidas para as diferentes distribuições, utilizando os resultados da seguinte proposição desenvolvida para este caso:

  Proposição 1. *Seja* $Y_i \sim SMSN\left(\mathbf{x}_i^\top\boldsymbol{\beta} + b_1\Delta, \sigma^2, \lambda; H\right)$. *Logo, para* $v_{1i} < v_{2i}$, $\mathcal{E}_{rsi}(\boldsymbol{\theta}) = E[U_i T_i^r Y_i^s \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}]$, *para* $r, s \in \{0, 1, 2\}$, *é dado por:*

$$\mathcal{E}_{10i}(\boldsymbol{\theta}) = \frac{\Delta}{\Delta^2 + \Gamma}\left(\mathcal{E}_{01i}(\boldsymbol{\theta}) - \mathcal{E}_{00i}(\boldsymbol{\theta})\mu_i^*\right) + b_1\mathcal{E}_{00i}(\boldsymbol{\theta}) + \sqrt{\frac{\Gamma}{\Delta^2 + \Gamma}}\mathbf{W}_\Psi^0(\boldsymbol{\theta}),$$

$$\begin{aligned}\mathcal{E}_{20i}(\boldsymbol{\theta}) = {} & \left(\frac{\Delta}{\Delta^2 + \Gamma}\right)^2\left(\mathcal{E}_{02i}(\boldsymbol{\theta}) - 2\mathcal{E}_{01i}(\boldsymbol{\theta})\mu_i^* + \mathcal{E}_{00i}(\boldsymbol{\theta})(\mu_i^*)^2\right) \\ & + 2b_1\left(\frac{\Delta}{\Delta^2 + \Gamma}\right)\left(\mathcal{E}_{01i}(\boldsymbol{\theta}) - \mathcal{E}_{00i}(\boldsymbol{\theta})\mu_i^*\right) + b_1^2\mathcal{E}_{00i}(\boldsymbol{\theta}) + \frac{\Gamma}{\Delta^2 + \Gamma} \\ & + \sqrt{\frac{\Gamma}{\Delta^2 + \Gamma}}\left\{\left(\frac{\Delta}{\Delta^2 + \Gamma}\right)\mathbf{W}_\Psi^1(\boldsymbol{\theta}) - \left(\left(\frac{\Delta}{\Delta^2 + \Gamma}\right)\mu^* - 2b_1\right)\mathbf{W}_\Psi^0(\boldsymbol{\theta})\right\},\end{aligned}$$

$$\mathcal{E}_{11i}(\boldsymbol{\theta}) = \left(\frac{\Delta}{\Delta^2 + \Gamma}\right)\left(\mathcal{E}_{02i}(\boldsymbol{\theta}) - \mathcal{E}_{01i}(\boldsymbol{\theta})\mu_i^*\right) + b_1\mathcal{E}_{01i}(\boldsymbol{\theta}) + \sqrt{\frac{\Gamma}{\Delta^2 + \Gamma}}\mathbf{W}_\Psi^1(\boldsymbol{\theta}),$$

  em que $b_1 = -\sqrt{\frac{2}{\pi}}k_1$, $\Gamma = \sigma^2(1 - \delta^2)$, $\Delta = \sigma\delta$, $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$ *e* $\mu_i^* = \mathbf{x}_i^\top\boldsymbol{\beta} + b_1\Delta$.

  Prova. *A prova destes resultados pode ser obtida utilizando a propriedade da esperança condicional, expressa por:*

$$E[U_i T_i^r Y_i^s \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}] = E[Y_i^s E[U_i E[T_i^r \,|\, U_i, Y_i] \,|\, Y_i] \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}],$$

  com $r, s \in \{0, 1, 2\}$. *É importante enfatizar que os momentos das distribuições TSMSN resultantes podem ser encontrados em Lachos, Garay e Cabral (2020).*

Os resultados para $\mathcal{E}_{0ri}(\boldsymbol{\theta})$ e $\mathbf{W}_{\Psi}^{k}(\boldsymbol{\theta})$, com $k \in \{0,1\}$, para cada uma das distribuições SMSN são apresentados no Apêndice A.

Observação: No caso em que a observação "i" é faltante, temos que $-\infty \leq Y_i \leq \infty$, assim $\mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)}) = \mathrm{E}[U_i T_i^r Y_i^s \,|\, \boldsymbol{\theta}^{(k)}]$ para $r, s \in \{0,1,2\}$.

O passo CM é implementado da seguinte forma:

(ii) **Passo-CMQ 1:**

Fixe $\Delta = \Delta^{(k)}$, obtenha $\boldsymbol{\theta}^{(k+1)}$ maximizando $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})$ em relação a $\boldsymbol{\beta}$, o que resulta em:

$$\boldsymbol{\beta}^{(k+1)} = \left(\sum_{i=1}^{n} \mathcal{E}_{00i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_i\mathbf{x}_i^{\top})\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i \left(\mathcal{E}_{01i}(\boldsymbol{\theta}^{(k)}) - \Delta^{(k)}\mathcal{E}_{10i}(\boldsymbol{\theta}^{(k)})\right);$$

(iii) **Passo-CMQ 2:**

Fixe $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$ e atualize $\Delta^{(k)}$ maximizando $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})$ em relação a Δ, dada por:

$$\Delta^{(k+1)} = \frac{\sum_{i=1}^{n} \mathcal{E}_{11i}(\boldsymbol{\theta}^{(k)}) - \sum_{i=1}^{n} \mathcal{E}_{10i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_i^{\top}\boldsymbol{\beta}^{(k+1)})}{\sum_{i=1}^{n} \mathcal{E}_{20i}(\boldsymbol{\theta}^{(k)})};$$

(iv) **Passo-CMQ 3:**

Obtenha $\Gamma^{(k+1)}$ maximizando $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})$ em relação a Γ, considerando $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$ e $\Delta = \Delta^{(k+1)}$. Dessa forma, obtém-se:

$$\begin{aligned}\Gamma^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \Big[&\mathcal{E}_{02i}(\boldsymbol{\theta}^{(k)}) - 2\mathcal{E}_{01i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_i^{\top}\boldsymbol{\beta}^{(k+1)}) \\ &+ \mathcal{E}_{00i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_i^{\top}\boldsymbol{\beta}^{(k+1)})^2 - 2\Delta^{(k+1)}\mathcal{E}_{11i}(\boldsymbol{\theta}^{(k)}) \\ &+ 2\Delta^{(k+1)}\mathcal{E}_{10i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_i^{\top}\boldsymbol{\beta}^{(k+1)}) + (\Delta^{(k+1)})^2\mathcal{E}_{20i}(\boldsymbol{\theta}^{(k)})\Big].\end{aligned}$$

(v) **Passo-CML:**

Estimamos $\boldsymbol{\nu}$ maximizando a função de log-verossimilhança dada na Equação (3.3), especificamente:

$$\begin{aligned}\boldsymbol{\nu}^{(k+1)} = \underset{\boldsymbol{\nu}}{\operatorname{argmax}} \Bigg\{ &\sum_{i=1}^{n_0} \log\Bigg[\Psi_{\mathrm{SMSN}}\left(v_{2i}\,|\,\mathbf{x}_i^{\top}\boldsymbol{\beta}^{(k+1)}, \sigma^{2(k+1)}, \lambda^{(k+1)}; H\right) \\ &- \Psi_{\mathrm{SMSN}}\left(v_{1i}\,|\,\mathbf{x}_i^{\top}\boldsymbol{\beta}^{(k+1)}, \sigma^{2(k+1)}, \lambda^{(k+1)}; H\right)\Bigg]^{C_i} \\ &+ \sum_{i=1}^{n_0} \log\Big[\psi_{\mathrm{SMSN}}(y_i\,|\,\mathbf{x}_i^{\top}\boldsymbol{\beta}^{(k+1)}, \sigma^{2(k+1)}, \lambda^{(k+1)}; H)\Big]^{1-C_i}\Bigg\}.\end{aligned}$$

Conforme mencionado por Garay et al. (2017) e Mattos, Garay e Lachos (2018), é possível realizar o Passo-CML utilizando a rotina **optim** disponível no software R.

Por outro lado, observe que $\sigma^{2(k+1)}$ e $\lambda^{(k+1)}$ podem ser obtidos através das seguintes reparametrizações:

$$\sigma^{2(k)} = \Gamma^{(k)} + \Delta^{2(k)} \quad \text{e} \quad \lambda^{(k)} = \frac{\Delta^{(k)}}{\sqrt{\Gamma^{(k)}}}.$$

Este processo é iterado até que a diferença entre duas avaliações sucessivas da função de log-verossimilhança, como mencionado na Seção 2.3, se torne suficientemente pequena, $|\ell(\boldsymbol{\theta}^{(k+1)} \,|\, \mathbf{y}_{\text{obs}})/\ell(\boldsymbol{\theta}^{(k)} \,|\, \mathbf{y}_{\text{obs}}) - 1| < \epsilon$. Neste trabalho consideramos $\epsilon = 10^{-6}$.

Observações de implementação

Em nosso algoritmo, adotaremos a seguinte estratégia para obter os valores iniciais: consideramos apenas as observações não censuradas, ou seja, excluímos os dados faltantes e censurados. Essa abordagem simplifica o processo inicial de estimação, ao focar nos dados observados disponíveis, o que proporciona um ponto de partida razoável para as iterações subsequentes.

Desta forma, para obter os valores iniciais dos parâmetros do modelo, representados pelo vetor $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\beta}^{(0)}, \sigma^{2(0)}, \lambda^{(0)}, \boldsymbol{\nu}^{(0)}\right)$, procedemos assim:

- Utilizamos os estimadores de mínimos quadrados ordinários (OLS) para obter estimativas iniciais dos coeficientes $\boldsymbol{\beta}^{(0)}$ aplicando a função `lm()` disponível no software R.
- Aplicamos o método dos momentos para obter $\sigma^{2(0)}$.
- Para inicializar o parâmetro de assimetria $\lambda^{(0)}$, consideramos três vezes o sinal do coeficiente de assimetria dos resíduos.
- Quanto às distribuições ST e SCN, $\nu^{(0)} = 3$ e $\boldsymbol{\nu}^{(0)} = (0.5, 0.5)$ são estabelecidos, respectivamente.

3.3 ERROS PADRÃO APROXIMADOS

Para calcular os erros padrão dos estimadores de máxima verossimilhança, utilizamos a matriz de informação empírica, conforme discutido por Garay et al. (2017), Mattos, Garay e Lachos (2018) e Lachos et al. (2022). Assim, seja $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \lambda)^\top$, $\mathbf{Y}_{\text{obs}}$, as variáveis latentes

$\mathbf{Y}_L = \{Y_1, \ldots, Y_q\}$, $\mathbf{U} = \{U_1, \ldots, U_n\}$ e $\mathbf{T} = \{T_1, \ldots, T_n\}$ e $\mathbf{Y}_c = (\mathbf{Y}_{obs}, \mathbf{Y}_L, \mathbf{U}, \mathbf{T})$, então a matriz de informação empírica é definida por:

$$\mathbf{I}_e\left(\boldsymbol{\theta}\,|\,\mathbf{y}_{\text{obs}}\right) = \sum_{i=1}^{n} \mathbf{s}\left(y_{\text{obs}_i}\,|\,\boldsymbol{\theta}\right)\mathbf{s}^{\top}\left(y_{\text{obs}_i}\,|\,\boldsymbol{\theta}\right) - \frac{1}{n}\mathbf{S}\left(\mathbf{y}_{\text{obs}}\,|\,\boldsymbol{\theta}\right)\mathbf{S}^{\top}\left(\mathbf{y}_{\text{obs}}\,|\,\boldsymbol{\theta}\right),$$

em que $\mathbf{S}^{\top}\left(\mathbf{y}_{\text{obs}}\,|\,\boldsymbol{\theta}\right) = \sum_{i=1}^{n}\mathbf{s}\left(y_{\text{obs}_i}\,|\,\boldsymbol{\theta}\right)$.

Como mostrado por Louis (1982), o escore individual $\mathbf{s}(y_{\text{obs}_i}\,|\,\boldsymbol{\theta})$ pode ser determinado por:

$$\mathbf{s}(y_{\text{obs}_i}\,|\,\boldsymbol{\theta}) = \frac{\partial Q_i(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})}{\partial\boldsymbol{\theta}}, \quad i = 1, \ldots, n, \tag{3.4}$$

com $Q_i(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)}) = E\left[\ell_c(\boldsymbol{\theta}\,|\,\mathbf{y}_{c_i})\,|\,y_{\text{obs}_i}, \boldsymbol{\theta}^{(k)}\right]$.

Substituindo os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ na Equação (3.4), a matriz de informação empírica se reduz a: $\mathbf{I}_e\left(\widehat{\boldsymbol{\theta}}\,|\,\mathbf{Y}_{\text{obs}}\right) = \sum_{i=1}^{n}\widehat{\mathbf{s}}_i\widehat{\mathbf{s}}_i^{\top}$, em que $\widehat{\mathbf{s}}_i = \left(\widehat{\mathbf{s}}_{\beta_i}, \widehat{\mathbf{s}}_{\sigma_i^2}, \widehat{\mathbf{s}}_{\lambda_i}\right)$ é o vetor escore individual com

$$\begin{aligned}
\widehat{\mathbf{s}}_{\beta_i} &= \frac{1+\widehat{\lambda}^2}{\widehat{\sigma}^2}\left(\mathbf{x}_i\mathcal{E}_{01i}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{00i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i\mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}} - \widehat{\sigma}\frac{\widehat{\lambda}}{\sqrt{1+\widehat{\lambda}^2}}\mathbf{x}_i\mathcal{E}_{10i}(\widehat{\boldsymbol{\theta}})\right),\\
\widehat{\mathbf{s}}_{\sigma_i^2} &= -\frac{1}{2\widehat{\sigma}^2} + \frac{1+\widehat{\lambda}^2}{2\widehat{\sigma}^4}\left(\mathcal{E}_{02i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{01i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}} + \mathcal{E}_{00i}(\widehat{\boldsymbol{\theta}})(\mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}})^2\right)\\
&\quad - \frac{\widehat{\lambda}\sqrt{(1+\widehat{\lambda}^2)}}{2\widehat{\sigma}^3}\left(\mathcal{E}_{11i}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{10i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}}\right),\\
\widehat{\mathbf{s}}_{\lambda_i} &= \frac{\widehat{\lambda}}{1+\widehat{\lambda}^2} - \frac{\widehat{\lambda}}{\widehat{\sigma}^2}\left(\mathcal{E}_{02i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{01i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}} + \mathcal{E}_{00i}(\widehat{\boldsymbol{\theta}})(\mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}})^2\right)\\
&\quad + \frac{1+2\widehat{\lambda}^2}{\widehat{\sigma}\sqrt{1+\widehat{\lambda}^2}}\left(\mathcal{E}_{11i}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{10i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}}\right) - \widehat{\lambda}\mathcal{E}_{20i}(\widehat{\boldsymbol{\theta}}).
\end{aligned}$$

As esperanças condicionais $\mathcal{E}_{rsi}(\boldsymbol{\theta}^{(k)}) = \mathrm{E}[U_iT_i^rY_i^s\,|\,\mathbf{y}_{\text{obs}_i}, \boldsymbol{\theta}^{(k)}]$, para $r,\ s \in \{0,1,2\}$, podem ser obtidas diretamente a partir do nosso algoritmo ECME proposto.

3.4 ESTUDOS DE SIMULAÇÃO

Nesta seção, ajustamos modelos SMSN-ICR a conjuntos de dados artificiais para verificar sua capacidade de modelar estruturas complexas envolvendo assimetria, dados incompletos (dados censurados e/ou faltantes) e observações atípicas.

Consideramos o modelo dado pela Equação (3.2), com diferentes níveis de censura $p \in \{0\%, 8\%, 15\%, 20\%, 35\%\}$ e porcentagem de valores faltantes $m \in \{0\%, 5\%\}$ na variável resposta. $\mathbf{x}_i^{\top} = (1, x_i)^{\top}$ em que, como sugerido por Garay et al. (2017), Labra et al. (2012), Massuia et al. (2017), Mattos, Garay e Lachos (2018), Lachos et al. (2022), os valores de x_i foram gerados de forma independente a partir de uma distribuição uniforme, no intervalo $[1, 30]$,

permanecendo constantes ao longo dos experimentos. Para cada uma dessas combinações de níveis de censura e valores faltantes, geramos $R = 500$ amostras de Monte Carlo.

Por outro lado, é importante mencionar o método implementado neste trabalho para induzir os dados incompletos, o qual é realizado da seguinte forma:
Para censura intervalar:

1. A partir da amostra gerada de tamanho n, da variável resposta y, selecionamos aleatoriamente uma quantidade de observações a serem censuradas, determinada por p.

2. Calculamos o desvio padrão, denotado por Sd, para o conjunto de observações selecionadas no passo 1 e substituímos os valores por os intervalos $[y - \mathrm{Sd}, y + \mathrm{Sd}]$.

Para dados faltantes:

3. Das observações restantes, não censurados nos passos anteriores, selecionamos uma quantidade correspondente ao porcentagem m e os substituímos por NA.

Todos os procedimentos computacionais foram implementados usando o software R (R Core Team, 2024).

3.4.1 Estudo I

Este estudo foi conduzido para avaliar o desempenho dos estimadores propostos para os parâmetros em amostras de tamanho finito. Os tamanhos das amostras foram fixados em $n \in \{80, 160, 300,\ 500, 700, 1000\}$ e foram gerados a partir dos modelos SMSN-ICR (modelos SN-ICR, ST-ICR e SCN-ICR) considerando os seguintes parâmetros: $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (1.5, 2)^\top$, $\sigma^2 = 1.5$, e $\lambda = 2.4$. Para o modelo ST-ICR, consideramos $\nu = 3$, e para o modelo SCN-ICR, utilizamos $\boldsymbol{\nu} = (0.1, 0.1)$.

Para avaliar as estimativas obtidas pelo nosso algoritmo ECME proposto (com $\boldsymbol{\nu}$ fixo, devido a limitações de tempo computacional), comparamos o viés (Bias), o erro quadrático médio (MSE) e o viés relativo (RBias) para o parâmetro $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^\top = (\beta_1, \beta_2, \sigma^2, \lambda)^\top$ ao longo de $R = 500$ réplicas. Essas medidas são definidas por:

$$\mathrm{Bias}(\theta_i) = \frac{1}{R}\sum_{j=1}^{R}(\widehat{\theta}_i^{(j)} - \theta_i), \quad \mathrm{MSE}(\theta_i) = \frac{1}{R}\sum_{j=1}^{R}(\widehat{\theta}_i^{(j)} - \theta_i)^2 \quad \text{e}$$

$$\mathrm{RBias}(\theta_i) = \frac{1}{R}\sum_{j=1}^{R}\left(\frac{\widehat{\theta}_i^{(j)} - \theta_i}{\theta_i}\right), \quad \text{para} \quad i = 1, 2, 3, 4,$$

em que $\widehat{\theta}_i^{(j)}$ é a estimativa de θ_i a partir da j-ésima amostra.

Conforme mostrado nas Figuras 1 - 2, em todos os parâmetros do modelo SMSN-ICR, considerando um nível de censura de 15% e dados faltantes de 5%, o Bias e o MSE diminuem conforme o tamanho da amostra aumenta (geralmente, a partir de $n \geq 300$), indicando que as estimativas baseadas em nosso algoritmo do tipo EM proposto possuem boas propriedades assintóticas. Além disso, a partir da Tabela 1, observamos que, sob o modelo SN-ICR, em geral o viés relativo (RBias) tende a zero conforme n aumenta. É importante ressaltar que realizamos simulações com outros quatro níveis de censura $p \in \{0\%,\ 8\%,\ 20\%,\ 35\%\}$ e os padrões de convergência se comportaram de forma similar, como pode ser visto nas Figuras 11 - 18 e as Tabelas 10 e 11 fornecidas no Apêndice B.

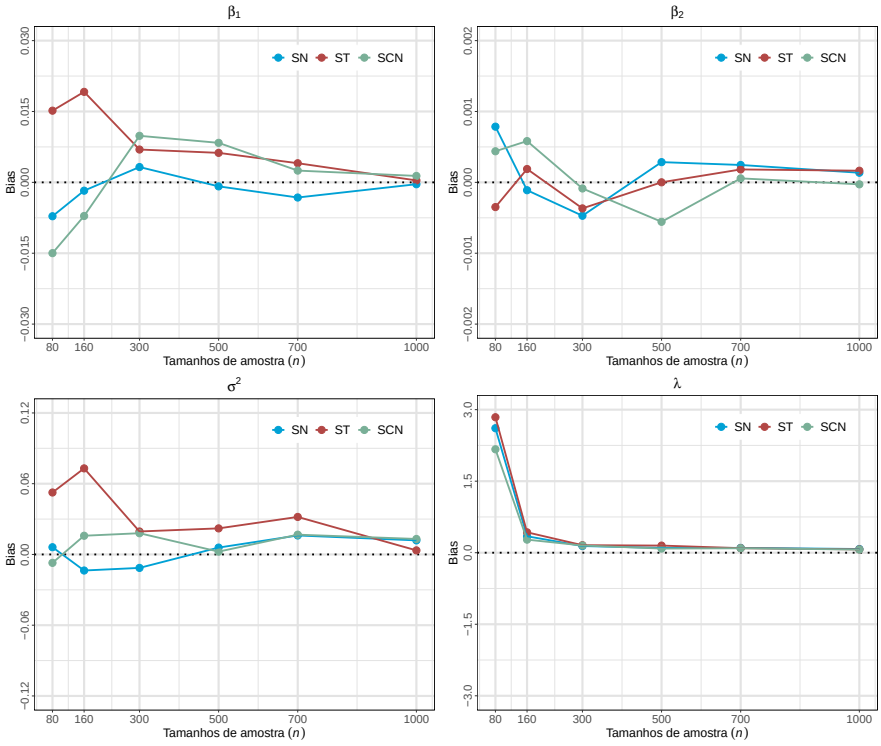

Figura 1 – Estudo de simulação 1. Bias das estimativas dos parâmetros dos modelos SMSN-ICR, com níveis de censura e dados faltantes de 15% e 5%, respectivamente.

3.4.2 Estudo II

O objetivo deste estudo de simulação é avaliar o desempenho da metodologia proposta para estimar os parâmetros dos modelos SMSN-ICR, na presença de observações atípicas na variável resposta. O tamanho da amostra foi fixado em $n = 300$, e os erros ε_i, $i = 1, 2, \ldots, n$,

Tabela 1 – Simulação 1. Resultados do Bias, MSE e RBias das estimativas dos parâmetros do modelo SN-ICR com diferentes tamanhos de amostra (n), níveis de censura intervalar (p) e de dados faltantes (m).

SN-ICR								
Nível Cens. (Falt.)		**Medida**	**Tamanhos de amostra (n)**					
			80	**160**	**300**	**500**	**700**	**1000**
0% (0%)	β_1	Bias	-0.0114	0.0058	0.0059	0.0056	-0.0011	-0.0007
		MSE	0.0369	0.0165	0.0085	0.0057	0.0037	0.0026
		RBias	-0.0076	0.0039	0.0040	0.0037	-0.0008	-0.0005
	β_2	Bias	0.0004	-0.0001	-0.0001	-0.0003	0.0001	0.0002
		MSE	0.0001	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	0.0002	0.0001^*	0.0001^*	-0.0001	0.0001^*	0.0001
	σ^2	Bias	-0.0093	0.0096	0.0001^*	0.0002	0.0017	0.0055
		MSE	0.1633	0.0773	0.0420	0.0244	0.0184	0.0120
		RBias	-0.0062	0.0064	0.0001^*	0.0001	0.0011	0.0037
	λ	Bias	0.9909	0.2591	0.0972	0.0776	0.0528	0.0586
		MSE	33.421	0.8369	0.2900	0.1855	0.1175	0.0895
		RBias	0.4129	0.1079	0.0405	0.0323	0.0220	0.0244
8% (5%)	β_1	Bias	0.0091	-0.0036	-0.0016	0.0041	-0.0021	-0.0038
		MSE	0.0405	0.0153	0.0107	0.0061	0.0038	0.0030
		RBias	0.0061	-0.0024	-0.0011	0.0027	-0.0014	-0.0025
	β_2	Bias	-0.0002	0.0005	0.0004	-0.0002	0.0002	0.0002
		MSE	0.0001	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	-0.0001	0.0002	0.0002	-0.0001	0.0001	0.0001
	σ^2	Bias	-0.0189	-0.0310	0.0092	-0.0029	0.0045	0.0061
		MSE	0.1863	0.0950	0.0498	0.0295	0.0223	0.0128
		RBias	-0.0126	-0.0207	0.0062	-0.0020	0.0030	0.0041
	λ	Bias	1.8162	0.1947	0.1455	0.0726	0.0614	0.0671
		MSE	102.99	1.0432	0.4308	0.2124	0.1576	0.0987
		RBias	0.7567	0.0811	0.0606	0.0303	0.0256	0.0280
15% (5%)	β_1	Bias	-0.0072	-0.0018	0.0032	-0.0008	-0.0032	-0.0004
		MSE	0.0475	0.0205	0.0116	0.0072	0.0047	0.0035
		RBias	-0.0048	-0.0012	0.0022	-0.0006	-0.0021	-0.0003
	β_2	Bias	0.0008	-0.0001	-0.0005	0.0003	0.0002	0.0001
		MSE	0.0002	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	0.0004	-0.0001	-0.0002	0.0001	0.0001	0.0001
	σ^2	Bias	0.0061	-0.0136	-0.0114	0.0058	0.0161	0.0119
		MSE	0.2312	0.1002	0.0539	0.0338	0.0199	0.0146
		RBias	0.0041	-0.0091	-0.0076	0.0039	0.0107	0.0080
	λ	Bias	2.6113	0.3449	0.1403	0.1013	0.1013	0.0783
		MSE	157.34	2.1401	0.5443	0.2550	0.1558	0.1064
		RBias	1.0880	0.1437	0.0584	0.0422	0.0422	0.0326
20% (5%)	β_1	Bias	-0.0054	0.0042	0.0035	0.0028	0.0022	0.0016
		MSE	0.0425	0.0230	0.0115	0.0073	0.0049	0.0040
		RBias	-0.0036	0.0028	0.0023	0.0019	0.0015	0.0011
	β_2	Bias	0.0004	0.0001^*	0.0001^*	-0.0002	-0.0002	-0.0001
		MSE	0.0002	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	0.0002	0.0001^*	0.0001^*	-0.0001	-0.0001	0.0001^*
	σ^2	Bias	-0.0411	-0.0248	-0.0149	-0.0043	0.0024	0.0146
		MSE	0.2249	0.1113	0.0533	0.0322	0.0231	0.0154
		RBias	-0.0274	-0.0165	-0.0100	-0.0029	0.0016	0.0098
	λ	Bias	2.2130	0.2654	0.1358	0.0714	0.0647	0.0892
		MSE	184.04	1.8489	0.5088	0.2432	0.1679	0.1089
		RBias	0.9221	0.1106	0.0566	0.0297	0.0270	0.0372
35% (5%)	β_1	Bias	-0.0029	0.0084	0.0001	-0.0021	0.0001^*	0.0011
		MSE	0.0633	0.0276	0.0156	0.0086	0.0062	0.0042
		RBias	-0.0019	0.0056	0.0001^*	-0.0014	0.0001^*	0.0007
	β_2	Bias	-0.0003	-0.0001	0.0003	0.0001^*	0.0001^*	0.0001^*
		MSE	0.0002	0.0001	0.0001	0.0001^*	0.0001^*	0.0001^*
		RBias	-0.0002	-0.0001	0.0001	0.0001^*	0.0001^*	0.0001^*
	σ^2	Bias	-0.0410	-0.0307	-0.0194	-0.0017	-0.0100	0.0064
		MSE	0.2646	0.1635	0.0702	0.0406	0.0331	0.0191
		RBias	-0.0273	-0.0204	-0.0129	-0.0011	-0.0066	0.0043
	λ	Bias	2.0922	0.4258	0.0964	0.0849	0.0385	0.0661
		MSE	74.096	4.5405	0.6078	0.3087	0.2255	0.1234
		RBias	0.8718	0.1774	0.0402	0.0354	0.0160	0.0276

0.0001^* indica que o número é menor que 0.0001 (< 0.0001).

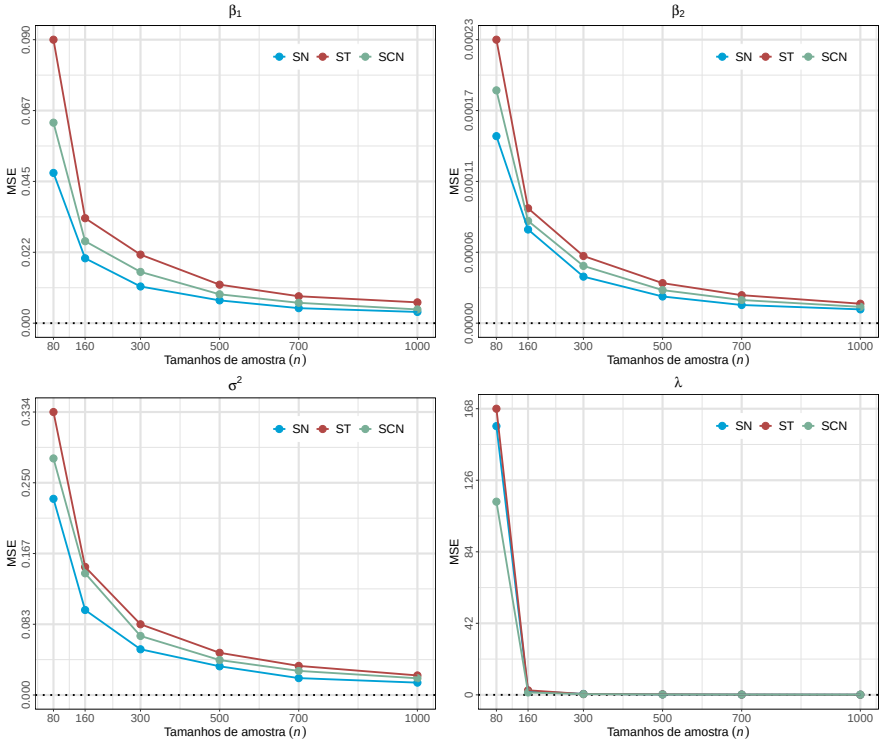

Figura 2 – Estudo de simulação 1. MSE das estimativas dos parâmetros dos modelos SMSN-ICR, com níveis de censura de 15% e dados faltantes de 5%, respectivamente.

foram gerados independentemente a partir da distribuição $\mathrm{SN}\left(\mathbf{x}_i^\top \boldsymbol{\beta} - \sqrt{\frac{2}{\pi}}\Delta, \sigma^2, \lambda\right)$ em que $\Delta = \sigma\delta,\ \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$. Adicionalmente, foram considerados dados incompletos com os níveis de censura $p = 15\%$ e $m = 5\%$.

Para ter as perturbações na variável resposta, em cada réplica, selecionamos três observações: **(i)** $y_{\text{min}} = \min\{y_1, y_2, \ldots, y_{n_1}\}$, **(ii)** $y_{\text{max}} = \max\{y_1, y_2, \ldots, y_{n_1}\}$ e **(iii)** uma observação aleatória y_{aleat}, das n_1 observações sem censura. Posteriormente, substituímos essas observações da seguinte forma: $y_{\text{min}}(\Lambda) = y_{\text{min}} - \Lambda$, $y_{\text{max}}(\Lambda) = y_{\text{max}} + \Lambda$ e $y_{\text{aleat}}(\Lambda) = y_{\text{aleat}} + \Lambda$, em que $\Lambda \in \{1, 2, \ldots, 9, 10\}$.

Para cada amostra gerada, estimamos os parâmetros dos modelos SN-ICR, ST-ICR $(\nu = 3)$ e SCN-ICR $(\boldsymbol{\nu} = (0.1, 0.1))$. Conforme sugerido por Mattos, Garay e Lachos (2018), estamos interessados em avaliar a mudança relativa nas estimativas, definida por:

$$RC\left(\widehat{\theta}_i(\Lambda)\right) = \left|\frac{\widehat{\theta}_i(\Lambda) - \widehat{\theta}_i}{\widehat{\theta}_i}\right|,$$

em que $\widehat{\theta}_i(\Lambda)$ e $\widehat{\theta}_i$ representam as estimativas dos parâmetros, com e sem perturbações, respectivamente. A Figura 3 apresenta os valores médios das mudanças relativas na estimativa para as R réplicas. Observe que, conforme Λ aumenta, as estimativas do modelo SN-ICR são mais sensíveis a essa perturbação, apresentando maiores mudanças relativas em comparação

com os modelos ST-ICR e SCN-ICR, caracterizados por terem caudas pesadas. Isso indica que os modelos ST-ICR e SCN-ICR são mais robustos do que o modelo SN-ICR na presença de observações atípicas. Além disso, foram realizadas simulações com outros níveis de censura (8%, 20% e 35%). Os resultados dessas simulações, apresentados nas Figuras 19 - 21 do Apêndice B, mostram padrões semelhantes. Esses resultados concordam com os encontrados por Garay, Lachos e Abanto-Valle (2011) e Mattos, Garay e Lachos (2018) os quais também abordam considerações semelhantes dentro do mesmo contexto assimétrico.

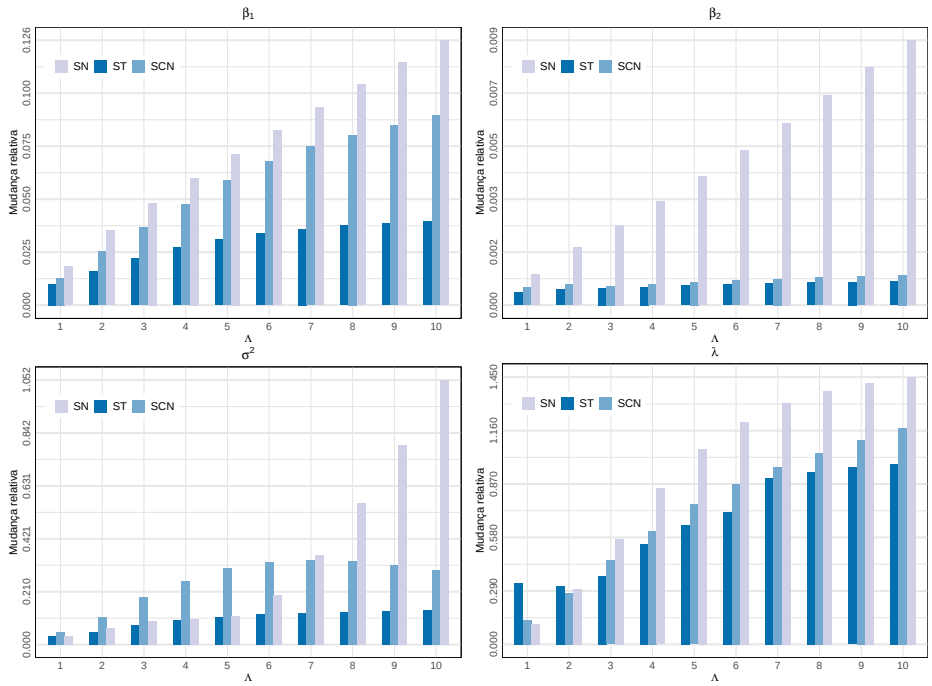

Figura 3 – Estudo de Simulação 2. Mudanças relativas médias das estimativas, para diferentes perturbações Λ e nível de censura de 15% e dados faltantes de 5%.

3.4.3 Estudo III

O terceiro estudo de simulação tem como objetivo avaliar a consistência do método de aproximação proposto na Seção 3.3 para calcular os erros-padrão (SE) dos estimadores de máxima verossimilhança $\widehat{\boldsymbol{\theta}} = \left(\widehat{\beta_1}, \widehat{\beta_2}, \widehat{\sigma^2}, \widehat{\lambda}\right)^\top$ dos modelos SMSN-ICR. Para isso, foram considerados quatro níveis de censura $p \in \{0\%,\ 8\%,\ 20\%,\ 35\%\}$ e porcentagens de dados faltantes $m \in \{0\%,\ 5\%\}$.

Os parâmetros utilizados na simulação foram definidos como $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (1.5, 2)^\top$, $\sigma^2 = 1.5$, e $\lambda = -2.4$. O tamanho da amostra foi fixado em $n = 400$. Os erros foram gerados

a partir dos modelos SMSN-ICR. Para o modelo ST-ICR, fixou-se $\nu = 3$, e para o modelo SCN-ICR, $\boldsymbol{\nu} = (0.1, 0.1)$.

Em cada amostra, foram obtidas as estimativas de $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^\top = (\beta_1, \beta_2, \sigma^2, \lambda)^\top$, os respectivos erros padrão conforme descrito na Seção 3.3 e os intervalos de confiança assintóticos de 95% para cada parâmetro, ou seja, $[\widehat{\theta} - 1.96 \times \text{SE}, \widehat{\theta} + 1.96 \times \text{SE}]$.

A partir de todas as estimativas obtidas nas 500 amostras, foram calculados:

- O desvio padrão de Monte Carlo de $\widehat{\theta}_i$, conhecido como MC-Sd:

$$\text{MC-Sd} = \sqrt{\frac{1}{499}\left[\sum_{j=1}^{500}\left(\widehat{\theta}_i^{(j)}\right)^2 - 500\left(\overline{\theta}_i\right)^2\right]} \quad \text{em que} \quad \overline{\theta}_i = \frac{1}{500}\sum_{j=1}^{500}\widehat{\theta}_i^{(j)}.$$

- A média dos erros padrão aproximados das estimativas obtidas pelo método ECME usando a matriz de informação empírica, denominada AV-SE.

- A proporção de vezes em que os intervalos de confiança cobriram o valor verdadeiro do parâmetro (COV MC).

Na Tabela 2, são apresentados os resultados obtidos. De maneira geral, a proporção de cobertura (COV MC) dos intervalos de confiança assintóticos para os parâmetros se mantém próxima a 95% na maioria dos cenários, o que indica que os intervalos construídos a partir dos SE aproximados conseguem capturar adequadamente os valores verdadeiros dos parâmetros. No entanto, em alguns cenários, observa-se que a cobertura tende a se desviar um pouco de 95%, especialmente para o parâmetro λ. Ainda assim, considera-se que os níveis de cobertura permanecem satisfatórios, o que confirma a eficácia do método proposto para lidar com a censura e dados faltantes na construção de intervalos de confiança.

Tabela 2 – Simulação 3. Resultados do MC-Sd, AV-SE e COV MC para as estimativas dos parâmetros do modelo SMSN-ICR com diferentes níveis de censura intervalar (p) e de dados faltantes (m).

		SN-ICR			**ST-ICR**		
Nível Cens. (Falt.)		**MC Sd**	**AV SE**	**COV MC**	**MC Sd**	**AV SE**	**COV MC**
0% (0%)	β_1	0.0849	0.0823	95.2%	0.1089	0.1111	95.0%
	β_2	0.0049	0.0048	93.2%	0.0055	0.0055	95.4%
	σ^2	0.1664	0.1780	96.2%	0.2125	0.2174	94.6%
	λ	0.4635	0.4881	96.6%	0.4928	0.4989	98.0%
8% (5%)	β_1	0.0854	0.0883	95.0%	0.1201	0.1189	94.0%
	β_2	0.0050	0.0051	93.8%	0.0061	0.0059	93.4%
	σ^2	0.1859	0.1920	95.6%	0.2293	0.2318	94.4%
	λ	0.5203	0.5286	97.2%	0.5009	0.5275	96.6%
15% (5%)	β_1	0.0956	0.0920	94.2%	0.1245	0.1245	94.6%
	β_2	0.0055	0.0053	94.4%	0.0064	0.0062	94.4%
	σ^2	0.1799	0.2005	97.0%	0.2429	0.2434	96.0%
	λ	0.5323	0.5609	97.8%	0.5496	0.5635	97.2%
20% (5%)	β_1	0.0950	0.0952	96.0%	0.1272	0.1284	96.0%
	β_2	0.0054	0.0055	95.8%	0.0067	0.0064	93.0%
	σ^2	0.2001	0.2073	94.6%	0.2348	0.2530	96.4%
	λ	0.5721	0.5850	96.4%	0.5638	0.5863	97.6%
35% (5%)	β_1	0.1099	0.1071	95.0%	0.1474	0.1436	94.4%
	β_2	0.0063	0.0062	96.2%	0.0072	0.0072	93.2%
	σ^2	0.2137	0.2352	97.2%	0.2867	0.2818	94.0%
	λ	0.6446	0.6739	98.8%	0.7271	0.6734	94.8%
		SCN-ICR					
0% (0%)	β_1	0.0913	0.0961	95.4%	-	-	-
	β_2	0.0052	0.0052	95.8%	-	-	-
	σ^2	0.1985	0.2065	96.0%	-	-	-
	λ	0.5029	0.5103	97.0%	-	-	-
8% (5%)	β_1	0.1042	0.1030	95.0%	-	-	-
	β_2	0.0055	0.0056	95.6%	-	-	-
	σ^2	0.2210	0.2199	95.0%	-	-	-
	λ	0.5587	0.5495	96.4%	-	-	-
15% (5%)	β_1	0.1044	0.1075	95.2%	-	-	-
	β_2	0.0056	0.0058	95.4%	-	-	-
	σ^2	0.2260	0.2309	95.0%	-	-	-
	λ	0.5388	0.5747	96.2%	-	-	-
20% (5%)	β_1	0.1167	0.1110	94.0%	-	-	-
	β_2	0.0063	0.0060	94.0%	-	-	-
	σ^2	0.2447	0.2370	93.2%	-	-	-
	λ	0.6040	0.6009	95.6%	-	-	-
35% (5%)	β_1	0.1276	0.1242	95.4%	-	-	-
	β_2	0.0072	0.0067	93.4%	-	-	-
	σ^2	0.2605	0.2671	94.8%	-	-	-
	λ	0.6606	0.6818	96.8%	-	-	-

3.5 APLICAÇÕES

Nesta seção, aplicamos o modelo proposto a três conjuntos de dados reais para demonstrar sua eficácia e versatilidade no tratamento de dados que envolvem censura à direita, à esquerda e intervalar, apresentando desta forma, exemplos de sua utilidade prática.

3.5.1 Pesquisa domiciliar (OHS99)

Para esta primeira aplicação, utilizamos o conjunto de dados da Pesquisa Domiciliar de Outubro de 1999 (OHS99) (Central Statistical Service, 2000), conduzida pela Statistics South Africa (StatsSA), que apresenta observações com censura intervalar. Essas pesquisas consideram informações demográficas, socioeconômicas e laborais de um país e são cruciais para estudar características de uma população. Em particular, o conjunto de dados referenciado daqui em diante como, dados OHS99, oferece uma visão detalhada da situação econômica e social da África do Sul em 1999, servindo como uma ferramenta importante para governos, pesquisadores e ONGs no, desenvolvimento de políticas públicas apropriadas.

Como mencionado por Fintel (2007), a obtenção de informações precisas e verdadeiras em pesquisas sociais e econômicas enfrenta vários desafios, tais como a desconfiança dos entrevistados, a incerteza quanto às posições dos membros da família, bem como a falta de informações, devido à recusa de alguns entrevistados ou por outras razões.

Uma maneira de superar esse problema é incluir opções de intervalo de renda nos questionários. Essa prática permite obter informações de forma abrangente, tanto daqueles que preferem permanecer anônimos quanto daqueles que só podem fornecer valores aproximados de seus rendimentos. Isso é importante porque, se a análise de dados considerasse apenas dados pontuais, excluiria uma quantidade de respostas sobre renda, o que pode afetar a distribuição dos dados e a confiabilidade das análises. Assim, sugere-se o uso de um modelo de regressão censurado, que seja capaz de lidar com observações em intervalos e dados faltantes.

O conjunto de dados consiste em $n = 2376$ trabalhadores em tempo integral, que trabalharam em 1999 (não incluindo autônomos), com idades entre 15 e 65 anos. Assim, as variáveis utilizadas no estudo foram:

- Y_i : Renda semanal, em rand sul-africano (ZAR)[1], considerando o valor original dividido

[1] O rand sul-africano é a moeda oficial da África do Sul

por 100.

- x_{i1} : Gênero (1 = Masculino, 0 = Feminino).
- x_{i2} : Idade.
- x_{i3} : Zona de residência (1 = Rural, 0 = Urbana).
- x_{i4} : Raça (1 = Africano/Negro, 0 = Pardo, Indiano/Asiático, Branco).
- x_{i5} : Filiação sindical (1 = Sim, 0 = Não), para $i = 1, \ldots, n$.

Observamos que, em relação à variável de renda semanal, 1517 indivíduos (63.85%) forneceram um valor, enquanto 736 (Aprox. 31%) relataram um intervalo e 123 indivíduos (5.2%) não relataram nenhuma informação (NA). É importante mencionar que os valores fornecidos não são necessariamente exatos, podendo haver erros adicionais além do erro aleatório. No entanto, nosso modelo não contempla essa possibilidade.

Na Tabela 3 são apresentadas algumas estatísticas descritivas das variáveis explicativas utilizadas no estudo. Para a variável idade, são fornecidas medidas como a média, a mediana e o desvio padrão. Para as variáveis binárias (dummies), são apresentadas as proporções correspondentes das categorias. Essas estatísticas estão divididas em três grupos distintos: indivíduos que reportaram sua renda com um valor específico (**Não censurados**), indivíduos que reportaram um intervalo como renda (**Censurados**) e indivíduos que não reportaram nenhuma informação sobre a renda (**Faltantes**).

Tabela 3 – Conjunto de dados OHS99. Resumo das variáveis explicativas por grupos: não censurados, censurados e ausentes.

Variável	Estatísticas	Não censurados	Censurados	Faltantes	Total
Idade	# de casos	1517	736	123	2376
	média	36	36	38	36
	mediana	35	35	38	35
	desvio padrão	10.54	9.98	11.78	10.44
Gênero	Masculino	64.2%	62.3%	57.7%	63.3%
Zona	Rural	44.6%	24.3%	13.8%	36.7%
Raça	Africano/Negro	57.0%	53.0%	50.4%	55.4%
Sindicato	Sim	31.7%	48.0%	56.1%	38.1%

Na Figura 4 são apresentados os histogramas das frequências relativas da variável resposta para dois grupos de dados: os censurados e os não censurados. Nestes, pode-se observar uma

assimetria positiva na distribuição da renda semanal. Por outro lado, no eixo x são representados os intervalos utilizados na pesquisa, em que cada ponto marca o início de um intervalo e o ponto seguinte indica o final. Esta representação segmentada, facilita a visualização da distribuição dos dados em cada intervalo específico, permitindo uma comparação entre ambos os grupos. Além disso, o gráfico destaca a importância de considerar os dados com censura intervalar no estudo, já que os dados não censurados, por si só, podem não refletir adequadamente a informação real da população.

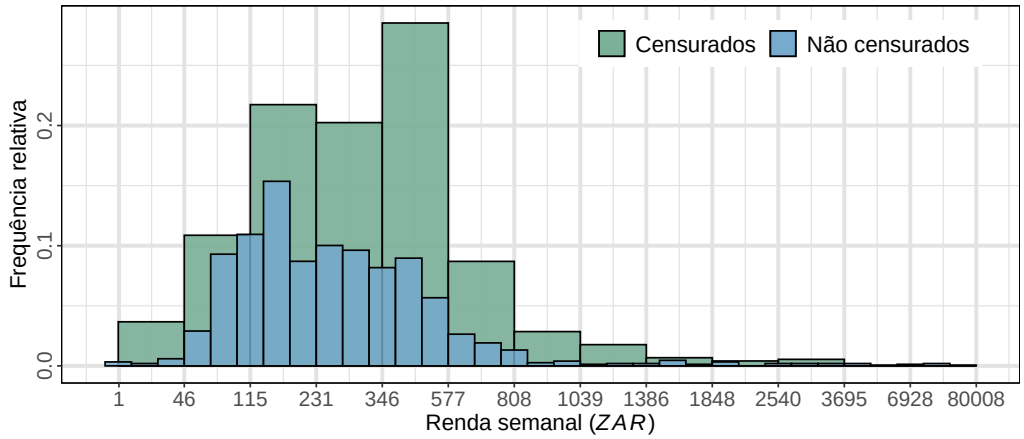

Figura 4 – Conjunto de dados OHS99. Histograma das frequências relativas da renda semanal para os dados censurados e para os não censurados (observados).

Como a variável resposta, renda semanal, apresenta censura intervalar e dados faltantes, utilizamos nosso modelo SMSN-ICR neste conjunto de dados, conforme apresentado ao longo deste capítulo, com o objetivo de fornecer inferências sobre a relação entre renda semanal e os fatores socioeconômicos e demográficos dos indivíduos.

A Tabela 4 contém as estimativas de máxima verossimilhança para os parâmetros dos modelos SMN-ICR e SMSN-ICR, bem como os erros padrão (SE) estimados, obtidos por meio da matriz de informação empírica - veja a Seção 3.3 - e os intervalos de confiança assintóticos de 95% $\lfloor \mathrm{LL},\ \mathrm{UL} \rfloor$ dos parâmetros.

A partir dos intervalos de confiança, observamos que todos os parâmetros dos modelos ajustados: T-ICR, CN-ICR, ST-ICR e SCN-ICR, são significativamente diferentes de zero, dado que nenhum dos intervalos inclui o valor zero. Essa observação sugere que os parâmetros têm efeitos significativos nos modelos. Em comparação, nos modelos ajustados N-ICR e SN-ICR, todos os intervalos de confiança incluem o valor zero, indicando que os parâmetros nesses modelos não são significativamente diferentes de zero. Além disso, observamos que, de forma

Tabela 4 – Conjunto de dados OHS99. Estimativas de máxima verossimilhança (Est), erros padrão aproximados (SE) e intervalos de confiança assintótica de 95% para os parâmetros ($\lfloor \mathrm{LL} ; \mathrm{UL} \rfloor$), dos modelos SMN-ICR e SMSN-ICR.

	N-ICR				**T-ICR**				**CN-ICR**			
	Est	SE	$\lfloor$LL	UL$\rfloor$	Est	SE	$\lfloor$LL	UL$\rfloor$	Est	SE	$\lfloor$LL	UL$\rfloor$
β_0	1.423	8.230	-14.71	17.55	2.340	0.127	2.091	2.590	2.409	0.406	1.612	3.205
β_1	0.418	2.254	-3.999	4.835	0.522	0.061	0.403	0.642	0.724	0.225	0.282	1.166
β_2	0.071	0.187	-0.296	0.438	0.011	0.003	0.005	0.016	0.018	0.010	-0.001	0.036
β_3	-0.494	3.669	-7.685	6.697	-1.512	0.066	-1.641	-1.383	-1.662	0.248	-2.148	-1.176
β_4	-1.210	3.478	-8.028	5.608	-0.209	0.060	-0.327	-0.092	-0.392	0.201	-0.787	0.003
β_5	2.410	3.363	-4.182	9.002	0.850	0.060	0.732	0.968	0.957	0.199	0.568	1.347
σ^2	476.6	6.129	464.6	488.6	1.076	0.059	0.961	1.192	7.783	0.052	7.681	7.885
ν	-	-	-	-	2.01	-	-	-	-	-	-	-
ν_1	-	-	-	-	-	-	-	-	0.024	-	-	-
ν_2	-	-	-	-	-	-	-	-	0.01	-	-	-
	SN-ICR				**ST-ICR**				**SCN-ICR**			
	Est	SE	$\lfloor$LL	UL$\rfloor$	Est	SE	$\lfloor$LL	UL$\rfloor$	Est	SE	$\lfloor$LL	UL$\rfloor$
β_0	9.998	6.079	-1.916	21.912	3.457	0.127	3.209	3.705	4.277	0.197	3.891	4.663
β_1	0.526	1.929	-3.254	4.306	0.415	0.059	0.300	0.530	0.458	0.099	0.264	0.652
β_2	0.037	0.133	-0.224	0.298	0.008	0.003	0.003	0.013	0.009	0.004	0.001	0.018
β_3	-0.831	2.546	-5.820	4.159	-0.902	0.071	-1.041	-0.763	-0.963	0.144	-1.246	-0.680
β_4	-0.625	2.503	-5.530	4.280	-0.247	0.058	-0.361	-0.133	-0.277	0.092	-0.458	-0.096
β_5	1.525	2.335	-3.052	6.102	0.694	0.057	0.583	0.806	0.686	0.097	0.496	0.875
σ^2	494.3	4.680	485.1	503.5	5.871	0.298	5.288	6.455	18.64	0.083	18.47	18.80
λ	6.243	3.766	-1.138	13.62	4.715	0.413	3.906	5.525	6.871	0.659	5.579	8.163
ν	-	-	-	-	2.924	-	-	-	-	-	-	-
ν_1	-	-	-	-	-	-	-	-	0.02	-	-	-
ν_2	-	-	-	-	-	-	-	-	0.01	-	-	-

geral, as estimativas dos parâmetros do modelo de regressão são semelhantes entre os modelos T-ICR, CN-ICR, ST-ICR e SCN-ICR ajustados.

Por outro lado, os valores estimados para ν e ν_1 nos modelos ST-ICR e SCN-ICR, respectivamente, sugerem que a suposição dos modelos SN-ICR (N-ICR) pode não ser apropriada para este conjunto de dados. Isso porque, conforme mencionado em (BASSO et al., 2010), a distribuição ST (t de Student) converge para a distribuição SN (N) à medida que ν tende a ∞, e a distribuição SCN (CN) converge para a mesma quando $\nu_1 = \nu_2 = 1$.

É importante ressaltar que, conforme mencionado por Mattos, Garay e Lachos (2018), as estimativas para os parâmetros σ^2 e λ não são comparáveis, pois estão em uma escala diferente.

Os critérios de seleção de modelo, apresentados na Tabela 5, indicaram que os modelos com caudas mais pesadas do que os modelos N-ICR e SN-ICR produzem estimativas mais

Tabela 5 – Conjunto de dados OHS99. Critérios de seleção de modelos para os modelos SMSN-ICR e SMN-ICR.

Critérios	SN-ICR	ST-ICR	SCN-ICR	N-ICR	T-ICR	CN-ICR
Log-verossimilhança	-8571.81	**-4289.33**	-4840.91	-9869.32	-4314.18	-5554.93
AIC	17159.62	**8596.65**	9701.81	19752.64	8646.37	11129.86
BIC	17205.81	**8648.61**	9759.54	19793.06	8698.32	11187.59
CAIC	17213.81	**8657.61**	9769.54	19800.06	8707.32	11197.59
HQIC	17176.44	**8615.57**	9722.82	19767.35	8665.28	11150.87

precisas. Em particular, a distribuição ST-ICR se ajusta melhor aos dados do que as outras distribuições assimétricas (SN-ICR e SCN-ICR) e distribuições simétricas SMN (N-ICR, T-ICR e CN-ICR), devido a apresentar a maior log-verossimilhança e os menores valores nos critérios de seleção.

Como sugerido por Ortega, Bolfarine e Paula (2003b), Barros et al. (2010), Garay et al. (2017), Mattos, Garay e Lachos (2018), Lachos et al. (2022), para identificar observações atípicas e/ou especificação inadequada do modelo, analisamos os gráficos da transformação dos resíduo do tipo martingale - veja a Seção 2.4 - com envelopes simulados (ATKINSON, 1981), conforme apresentado na Figura 5. Esta figura indica claramente que o modelo ST-ICR é mais adequado para modelar o conjunto de dados do que os modelos assimétricos (SN-ICR e SCN-ICR) e simétricos (N-ICR, T-ICR e CN-ICR), dado que somente quatro resíduos se encontram fora dos envelopes. Adicionalmente a isso, é importante mencionar que realizamos uma análise mais aprofundada dos dados associados a esses valores, e não encontramos nenhuma característica específica que distinguisse esses indivíduos com salários tão altos das outras observações.

3.5.2 Fluência de nomes de letras (LNF)

Nesta subseção analisamos os dados de fluidez no nomeamento de letras (Letter-Name Fluency - LNF), que apresentam censura à direita, obtidos de estudantes peruanos por meio da Avaliação da Leitura nos Primeiros Anos (Early Grade Reading Assessment - EGRA), parte do instrumento RTI-FDA (2008). O LNF é uma avaliação padronizada, administrada individualmente, que avalia a rapidez com que as pessoas podem nomear uma série de letras em um período específico, neste caso, um minuto, e contribui na medição do conhecimento dos nomes

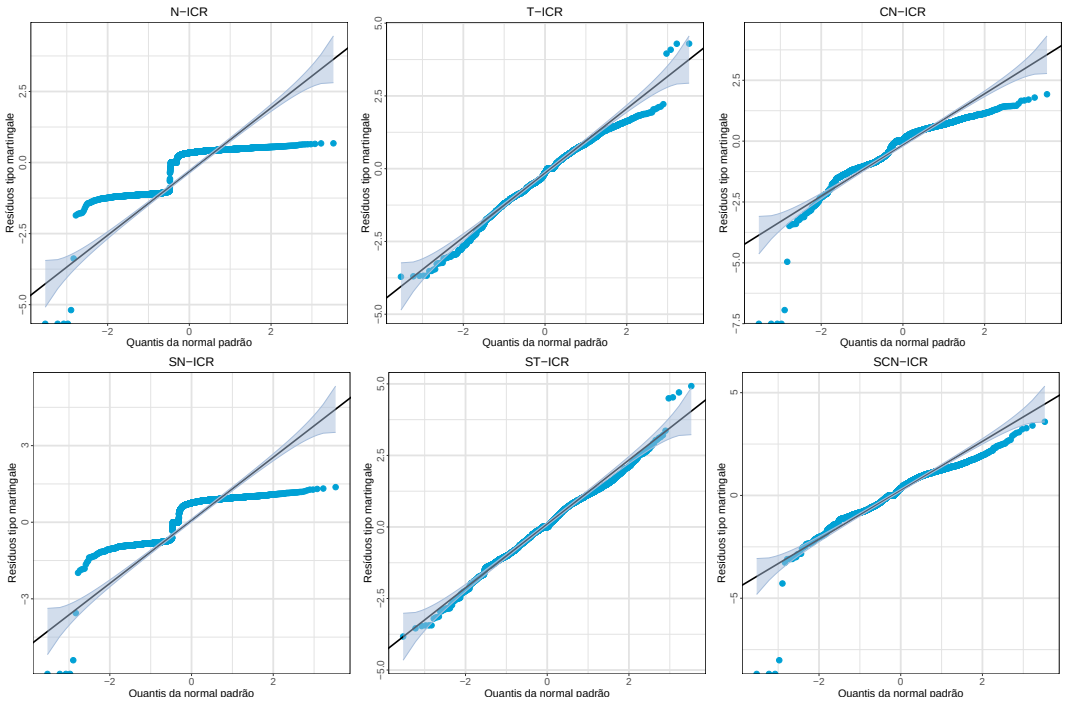

Figura 5 – Conjunto de dados OHS99. Envelopes dos resíduos do tipo martingale dos modelos SMSN-ICR e SMN-ICR.

das letras (Letter-Name Knowledg - LNK). Estes dados são importantes em contextos educacionais e psicológicos, pois permitem aos professores medir o desempenho dos alunos, prever habilidades de leitura, ortografia, consciência fonológica e inteligência geral, e assim identificar aqueles em risco de enfrentar desafios acadêmicos, apoiando efetivamente seu desenvolvimento linguístico e acadêmico.

A censura à direita nesses dados se deve à limitação de tempo imposta durante a avaliação. Isso significa que os alunos que concluem a tarefa em menos de um minuto podem ter lido mais letras/palavras/frases/parágrafos, o que afeta a média relatada de letras corretamente identificadas. Portanto, há a possibilidade de subestimar a resposta média de LNF, para um grupo devido à presença de observações censuradas. Para abordar esse problema, sugere-se um modelo de regressão censurada, que pode levar em conta observações abaixo ou acima de um limite para estimar a verdadeira resposta média de LNF para diferentes grupos de interesse.

O conjunto de dados, referido daqui em diante como, dados LNF, é composto por $n = 511$ estudantes e as variáveis definidas no estudo foram:

- Y_i : Número de letras lidas corretamente pelo aluno em um minuto.
- x_{i1} : Zona de residência (1 = Rural, 0 = Urbana).

- x_{i2} : Nível escolar (1 = 3º ano, 0 = 2º ano).
- x_{i3} : Gênero (1 = Feminino, 0 = Masculino).

Temos que, em relação ao tempo, 479 estudantes utilizaram o minuto completo na prova, enquanto 32 (6.26%) a terminaram antes do minuto, o que indica observações com censura à direita. Na Figura 6, apresenta-se o boxplot do número de letras lidas corretamente em relação à zona de residência e ao ano escolar. Nota-se que a mediana da variável de resposta é ligeiramente maior para as escolas urbanas e o terceiro ano. Além disso, é possível distinguir a partir desta figura que a mediana é superior para as escolas urbanas em comparação com as rurais, assim como para o terceiro ano em comparação com o segundo ano.

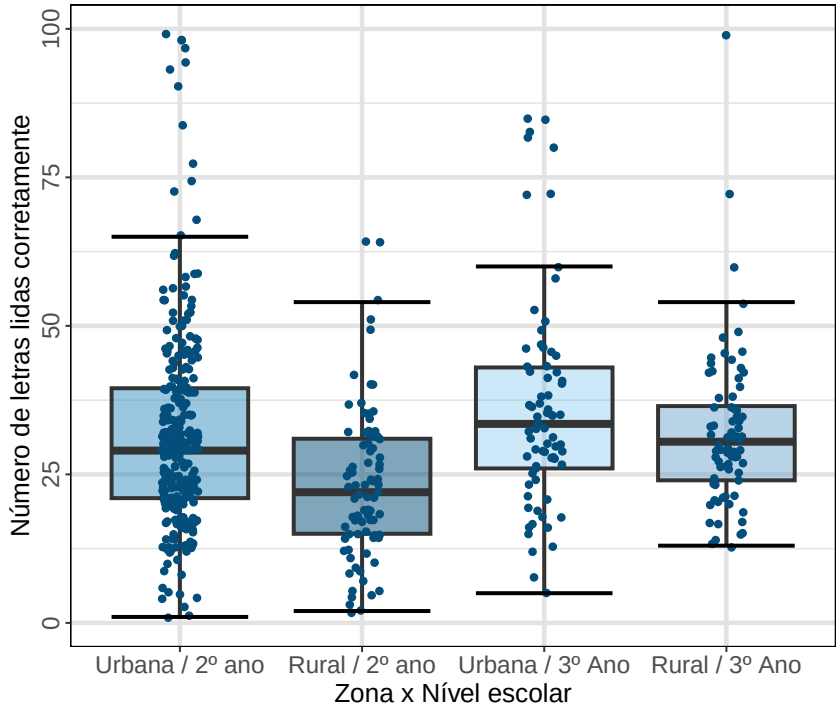

Figura 6 – Conjunto de dados LNF. Boxplot da variável de resposta em relação às covariáveis Zona de residência e Nível escolar.

Esses dados foram analisados por Lachos et al. (2022), no qual compararam o ajuste dos modelos N-CR, SN-CR, T-CR e ST-CR. Além disso, Oliveira, Oliveira e Lachos (2023) propuseram procedimentos de diagnóstico para o modelo de regressão linear ST com resposta censurada. Neste estudo, revisamos o mesmo conjunto de dados para avaliar o desempenho do nosso modelo SMSN-ICR, incluindo também os modelos CN-ICR e SCN-ICR. Segundo Lachos et al. (2022), a variável de resposta apresenta assimetria positiva e curtose, o que indica um desvio da distribuição normal, como pode ser observado na Figura 7. Além disso, eles apontam

que as observações censuradas têm uma média e um desvio padrão mais altos em comparação com as não censuradas, sugerindo uma subestimação da média da variável resposta Y_i.

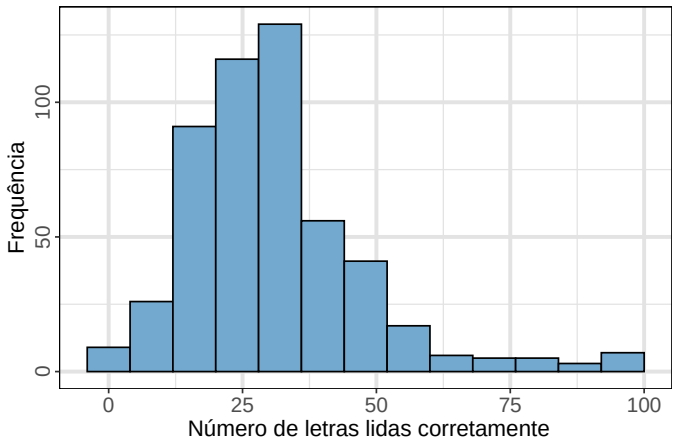

Figura 7 – Conjunto de dados LNF. Histograma da quantidade de letras lidas corretamente por estudantes em um minuto.

A Tabela 6 apresenta as estimativas de máxima verossimilhança para os parâmetros dos modelos SMN-ICR e SMSN-ICR, assim como os erros padrão (SE) estimados, obtidos por meio da matriz de informação empírica - veja a Seção 3.3 - e os intervalos de confiança assintóticos de 95% $\lfloor \mathrm{LL},\ \mathrm{UL} \rfloor$. Observamos que, em todos os modelos ajustados, o intervalo de confiança do parâmetro β_4 inclui o zero, indicando que não há evidência suficiente para afirmar que a covariável Gênero, tem um efeito significativo sobre a variável dependente no modelo de regressão proposto. Por outro lado, os demais parâmetros são significativamente diferentes de zero, conforme os resultados obtidos por Lachos et al. (2022). Além disso, considerando que na distribuição ST (t de Student), quando ν tende a ∞, e na distribuição SCN (CN) quando $\nu_1 = \nu_2 = 1$, ambas convergem para a distribuição SN (N) (BASSO et al., 2010), temos que os valores estimados para ν e ν_1 dos modelos ST-ICR e SCN-ICR, sugerem que a suposição do modelo SN-ICR (N-ICR) pode não ser apropriada para este conjunto de dados.

Os critérios de seleção de modelo, apresentados na Tabela 7, indicaram que os modelos com caudas mais pesadas em comparação com SN-ICR e as distribuições simétricas SMN (N-ICR, T-ICR e CN-ICR) proporcionam estimativas mais precisas e se ajustam melhor aos dados. Essas conclusões podem ser reforçadas pela Figura 8, na qual são apresentados gráficos da transformação do resíduos do tipo martingale - veja a Seção 2.4 - com envelopes simulados (ATKINSON, 1981), nos quais podemos observar também que os modelos ST-ICR e SCN-ICR têm, em sua maioria, os resíduos dentro dos envelopes.

Tabela 6 – Conjunto de dados LNF. Estimativas de máxima verossimilhança (Est), erros padrão aproximados (SE) e intervalos de confiança assintótica de 95% para os parâmetros ($\lfloor \mathrm{LL} ; \mathrm{UL} \rfloor$), dos modelos SMN-ICR e SMSN-ICR.

	N-CR				**T-CR**				**CN-CR**			
	Est	SE	$\lfloor$LL	UL$\rfloor$	Est	SE	$\lfloor$LL	UL$\rfloor$	Est	SE	$\lfloor$LL	UL$\rfloor$
β_1	32.52	1.370	29.84	35.21	29.17	1.210	26.80	31.54	29.29	1.134	27.07	31.52
β_2	-7.288	1.927	-11.064	-3.511	-5.804	1.434	-8.615	-2.993	-5.547	1.411	-8.313	-2.782
β_3	6.349	1.747	2.926	9.773	6.315	1.453	3.468	9.162	5.728	1.442	2.901	8.555
β_4	-0.136	1.529	-3.133	2.861	0.426	1.246	-2.015	2.867	0.486	1.218	-1.901	2.872
σ^2	283.1	14.61	254.4	311.7	138.4	15.36	108.3	168.5	141.9	13.04	116.4	167.6
ν	-	-	-	-	3.485	-	-	-	-	-	-	-
ν_1	-	-	-	-	-	-	-	-	0.126	-	-	-
ν_2	-	-	-	-	-	-	-	-	0.096	-	-	-
	SN-ICR				**ST-ICR**				**SCN-ICR**			
	Est	SE	$\lfloor$LL	UL$\rfloor$	Est	SE	$\lfloor$LL	UL$\rfloor$	Est	SE	$\lfloor$LL	UL$\rfloor$
β_1	32.31	1.051	30.25	34.37	31.79	1.065	29.71	33.88	31.75	1.060	29.67	33.83
β_2	-4.647	1.517	-7.619	-1.674	-4.719	1.242	-7.152	-2.285	-4.678	1.245	-7.117	-2.238
β_3	5.952	1.324	3.358	8.546	6.256	1.218	3.868	8.644	6.161	1.227	3.757	8.565
β_4	-0.167	1.207	-2.533	2.199	0.042	1.105	-2.123	2.208	0.085	1.103	-2.077	2.246
σ^2	658.6	44.42	571.6	745.7	323.5	41.08	242.9	404.0	304.5	38.95	228.1	380.8
λ	3.866	0.615	2.660	5.072	2.486	0.479	1.548	3.424	2.645	0.506	1.653	3.637
ν	-	-	-	-	4.368	-	-	-	-	-	-	-
ν_1	-	-	-	-	-	-	-	-	0.219	-	-	-
ν_2	-	-	-	-	-	-	-	-	0.192	-	-	-

Tabela 7 – Conjunto de dados LNF. Critérios de seleção de modelos para os modelos SMSN-ICR e SMN-ICR.

Critérios	**SN-ICR**	**ST-ICR**	**SCN-ICR**	**N-ICR**	**T-ICR**	**CN-ICR**
Log-verossimilhança	-2007.683	-1995.22	**-1994.016**	-2058.962	-2023.112	-2018.188
AIC	4027.367	4004.44	**4004.032**	4127.924	4060.223	4052.376
BIC	4052.785	**4034.095**	4037.923	4149.105	4089.878	4086.267
CAIC	4058.785	**4041.095**	4045.923	4154.105	4096.878	4094.267
HQIC	4033.694	**4016.066**	4017.318	4136.228	4071.849	4065.662

3.5.3 Taxa salarial (Mroz)

A terceira aplicação trata de dados econômicos censurados pela esquerda, especificamente no conjunto de dados que aborda as taxas salariais de mulheres casadas, conforme descrito por Mroz (1987). Esse tipo de dados é importane na análise econômica e sociológica, pois fornece uma compreensão mais profunda do efeito causal de certas variáveis no mercado de trabalho

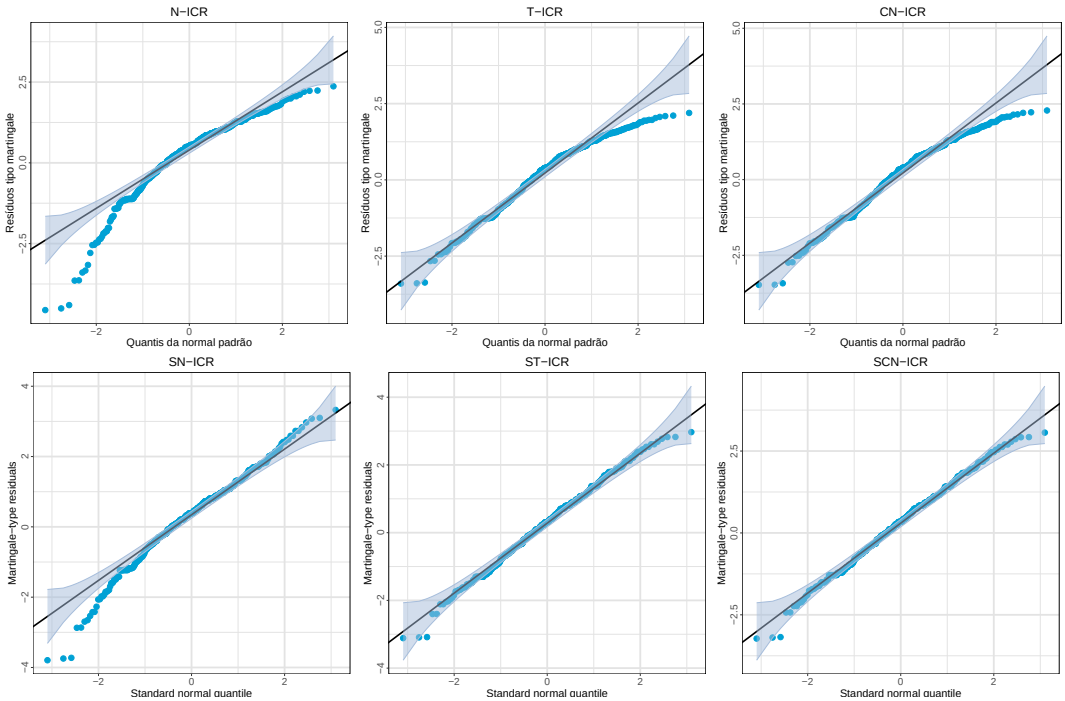

Figura 8 – Conjunto de dados LNF. Envelopes dos resíduos do tipo martingale dos modelos SMSN-ICR e SMN-ICR.

para mulheres casadas. Além disso, pode ter implicações significativas para a formulação de políticas públicas destinadas a enfrentar desafios no mercado de trabalho.

Os dados, referenciados daqui em diante como, dados Mroz, podem ser obtidos por meio do pacote SMNCensReg (GARAY; MASSUIA; LACHOS, 2022), disponível no software R. Este conjunto de dados inclui $n = 753$ observações de mulheres brancas casadas entre 30 e 60 anos em 1975. As variáveis analisadas no estudo são as seguintes:

- Y_i : Ganho potencial [2] das mulheres brancas casadas (em dólares americanos do ano de 1975).
- x_{i1} : Idade em anos.
- x_{i2} : Nível de escolaridade em anos.
- x_{i3} : Número de crianças menores de 6 anos no domicílio.
- x_{i4} : Anos de experiência anterior no mercado de trabalho.

[2] O ganho potencial se refere ao lucro ou rendimento máximo que uma pessoa ou empresa poderia ter obtido em determinado período, se todas as oportunidades tivessem sido aproveitadas sem restrições ou limitações.

Sabe-se que 428 delas trabalharam durante o ano de 1975, assim seu ganho potencial é sua própria renda (salário médio por hora), enquanto 325 (43.16%) não trabalharam e relataram um salário médio por hora de zero, resultando assim na censura à esquerda em zero para essas observações. Isso ocorre porque, para as mulheres que não trabalharam, não podemos observar diretamente o seu ganho potencial e sabemos apenas que este valor pertence ao intervalo $(-\infty, 0)$, representando a renda que deixaram de obter. Na Figura 9, podemos ver o histograma dos salários e notar o comportamento assimétrico dessa variável, o que reflete a presença de valores iguais a zero e a censura à esquerda.

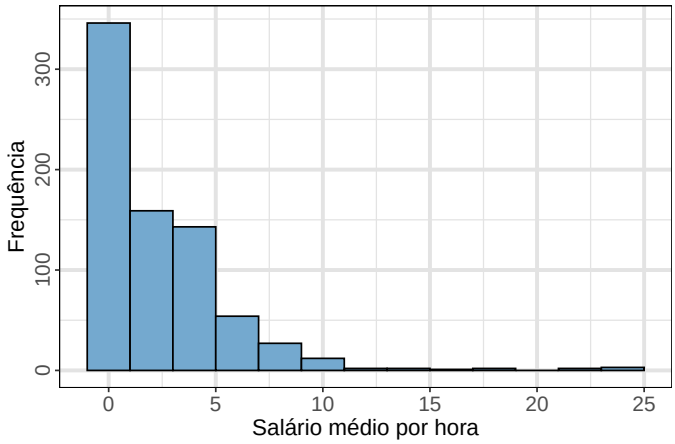

Figura 9 – Conjunto de dados Mroz. Histograma do salário médio por hora das esposas.

Esses dados têm sido amplamente utilizados em diversos estudos. Entre estes estão implementações do modelo CR Student-t (ARELLANO-VALLE et al., 2012), análise de modelos SMN-CR (GARAY et al., 2017), avaliação de desempenho de modelos SMSN-CR sob uma perspectiva bayesiana (MASSUIA et al., 2017) e avaliação do desempenho do algoritmo SAEM (MATTOS; GARAY; LACHOS, 2018). Mais recentemente, Oliveira, Oliveira e Lachos (2023) aplicaram técnicas diagnósticas, ilustrando a robustez do modelo ST-CR em comparação com seus concorrentes (N-CR, SN-CR, T-CR). Neste estudo, utilizamos o conjunto de dados para avaliar o desempenho do algoritmo do tipo EM proposto.

As estimativas de máxima verossimilhança dos parâmetros dos modelos SMSN-ICR, juntamente com seus erros padrão aproximados (SE) calculados a partir da Matriz de Informação Empírica - conforme descrito na Seção 2.3 - e os intervalos de confiança assintóticos de 95% $\lfloor \mathrm{LL},\ \mathrm{UL} \rfloor$, são apresentados na Tabela 8. Destaca-se que, com relação ao parâmetro estimado β_1, correspondente ao intercepto, ele é significativamente diferente de zero apenas no modelo ST-ICR, pois o intervalo de confiança não inclui o valor zero. Por outro lado, os demais parâ-

Tabela 8 – Conjunto de dados Mroz. Estimativas de máxima verossimilhança (Est), erros padrão aproximados (SE) e intervalos de confiança assintótica de 95% para os parâmetros ($\lfloor \mathrm{LL} \ ; \ \mathrm{UL} \rfloor$), dos modelos SMSN-ICR.

	SN-ICR				**ST-ICR**				**SCN-ICR**			
	Est	SE	$\lfloor \mathrm{LL}$	$\mathrm{UL} \rfloor$	Est	SE	$\lfloor \mathrm{LL}$	$\mathrm{UL} \rfloor$	Est	SE	$\lfloor \mathrm{LL}$	$\mathrm{UL} \rfloor$
β_1	0.058	1.414	-2.713	2.829	-2.783	1.133	-5.003	-0.563	-0.968	1.142	-3.206	1.270
β_2	-0.188	0.025	-0.238	-0.139	-0.139	0.019	-0.177	-0.101	-0.162	0.021	-0.203	-0.120
β_3	0.601	0.071	0.461	0.740	0.591	0.052	0.489	0.694	0.591	0.053	0.488	0.695
β_4	-2.951	0.398	-3.732	-2.170	-2.298	0.301	-2.889	-1.707	-2.644	0.323	-3.277	-2.011
β_5	0.225	0.023	0.179	0.270	0.195	0.019	0.159	0.232	0.211	0.019	0.173	0.249
σ^2	32.17	1.522	29.19	35.16	10.69	1.864	7.034	14.34	9.734	1.449	6.894	12.57
λ	2.572	0.616	1.365	3.780	-1.034	0.235	-1.494	-0.573	0.312	0.398	-0.468	1.092
ν	-	-	-	-	2.147	-	-	-	-	-	-	-
ν_1	-	-	-	-	-	-	-	-	0.053	-	-	-
ν_2	-	-	-	-	-	-	-	-	0.056	-	-	-

metros são significativamente diferentes de zero em todos os modelos ajustados. Além disso, os valores estimados para ν e ν_1 nos modelos ST-ICR e SCN-ICR, respectivamente, sugerem que a suposição do modelo SN-ICR (N-ICR) pode não ser apropriada para este conjunto de dados. Isso se deve ao fato de que, na distribuição ST (t de Student), quando ν tende a ∞, e na distribuição SCN (CN) quando $\nu_1 = \nu_2 = 1$, ambas convergem para a distribuição SN (N) (BASSO et al., 2010).

Tabela 9 – Conjunto de dados Mroz. Critérios de seleção de modelos para os modelos SMSN-ICR.

Critérios	**SN-ICR**	**ST-ICR**	**SCN-ICR**
Log-verossimilhança	-1417.061	-1367.248	**-1366.06**
AIC	2848.123	2750.496	**2750.12**
BIC	2880.491	**2787.488**	2791.736
CAIC	2887.491	**2795.488**	2800.736
HQIC	2860.593	**2764.747**	2766.152

A Tabela 9 apresenta os critérios de seleção de modelos obtidos para os modelos SN-ICR, ST-ICR e SCN-ICR. Observamos que os modelos com caudas mais pesadas apresentam valores menores em comparação com o modelo SN-ICR, indicando que oferecem estimativas mais precisas e um melhor ajuste aos dados. Esta afirmação também é confirmada pela Figura 10, que apresenta os gráficos da transformação dos resíduos tipo martingale com os envelopes

simulados, conforme detalhado na Seção 2.4. Ao analisar especificamente os gráficos dos modelos ST-ICR e SCN-ICR na figura, observa-se que, na maioria dos casos, os resíduos encontram-se dentro dos envelopes.

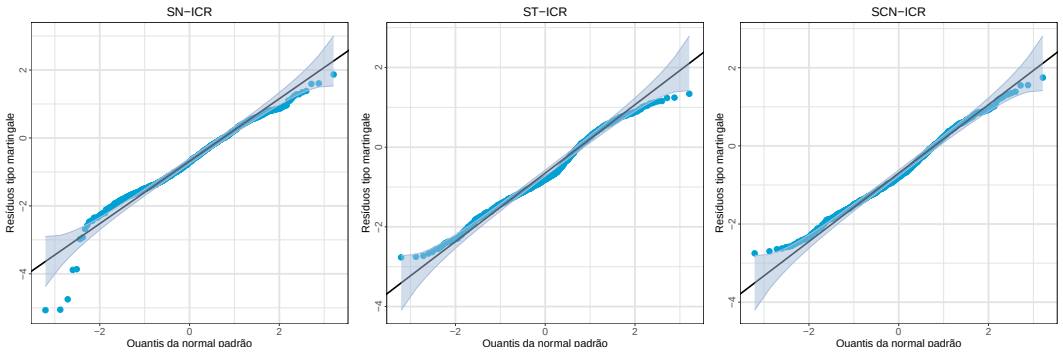

Figura 10 – Conjunto de dados Mroz. Envelopes dos resíduos do tipo martingale dos modelos SMSN-ICR

A seguir, apresentamos as conclusões obtidas a partir das análises realizadas, destacando os principais resultados. Além disso, discutimos as perspectivas futuras deste trabalho, sugerindo possíveis extensões do modelo SMSN-ICR.

4 CONCLUSÕES E PERSPECTIVAS FUTURAS

4.1 CONCLUSÕES

Neste trabalho, propusemos um modelo de regressão com censura intervalar sob a classe de distribuições SMSN, denominado SMSN-ICR, como uma alternativa à suposição de que os erros aleatórios seguem uma distribuição normal em modelos lineares censurados. Para essa estrutura complexa, desenvolvemos um inovador algoritmo ECME para obter estimativas dos parâmetros de máxima verossimilhança, que foram implementadas utilizando o software estatístico R. O modelo e os métodos aqui propostos são uma extensão dos trabalhos de Barros et al. (2010), Garay, Lachos e Abanto-Valle (2011), Garay et al. (2017), Garay et al. (2015), Massuia et al. (2015), Mattos, Garay e Lachos (2018) e Lachos et al. (2022), sob uma perspectiva frequentista.

Realizamos dois estudos de simulação utilizando nosso algoritmo do tipo EM proposto, considerando diferentes níveis de censura intervalar e dados faltantes. O primeiro estudo avaliou o desempenho das estimativas dos parâmetros em amostras de tamanho finito, indicando boas propriedades assintóticas. O segundo estudo analisou a precisão das estimativas dos parâmetros sob os modelos SMSN-ICR na presença de observações atípicas. Como esperado, observamos que os modelos com caudas pesadas são mais robustos, em termos de estimação, à presença de observações atípicas.

Além disso, implementamos o modelo proposto em três conjuntos de dados reais, demonstrando sua utilidade e eficácia em ambientes práticos. Ilustramos como o procedimento desenvolvido pode ser utilizado para avaliar suposições do modelo, identificar observações influentes e obter estimativas robustas dos parâmetros. Aplicamos nossos métodos propostos ao conjunto de dados da Pesquisa Domiciliar (OHS99) da StatsSA, para analisar as rendas semanais, considerando a censura intervalar e dados ausentes, no qual observamos que a distribuição ST-ICR apresenta um melhor ajuste aos dados em comparação com outras distribuições assimétricas (como SN-ICR e SCN-IR) e distribuições simétricas SMN (N-ICR, T-ICR e CN-ICR). Também aplicamos nosso método ao conjunto de dados LNF, previamente analisado por Lachos et al. (2022), o qual avalia a fluidez em nome de letras de estudantes peruanos, por meio da Avaliação da Leitura nos Primeiros Anos (EGRA), estes dados apresentam censura à direita devido à limitação de tempo. Por fim, utilizamos o conjunto de dados de taxa salarial, previamente analisado por Mroz (1987), em que os dados apresentam censura à esquerda.

Esses dois últimos conjuntos de dados mostraram que, baseado nos critérios de seleção de modelos e nos gráficos de resíduos tipo martingale, os modelos com caudas pesadas, como ST-CR e SCN-CR, apresentaram melhor ajuste do que o modelo SN-CR. Além disso, destaca-se que o modelo ST-CR demonstrou um desempenho superior em relação aos modelos simétricos SMN-CR.

Esses resultados demonstram a efetividade e a versatilidade dos métodos propostos para obter estimativas robustas dos parâmetros, destacando a importância de considerar distribuições assimétricas e com caudas pesadas em modelos de regressão censurados.

4.2 PERSPECTIVAS FUTURAS

Para futuros trabalhos, são propostas as seguintes pesquisas:

1. **Desenvolvimento do pacote de software:** Integrar o algoritmo ECME desenvolvido neste estudo como novas funções no pacote existente **SMNCensReg** (GARAY; MASSUIA; LACHOS, 2022). Esto ampliará a utilidade do pacote, ao incluir a classe de distribuições SMSN e possibilitar a análise de dados com censura à direita, à esquerda e intervalar, além de tratar com valores faltantes.

2. **Extensões do modelo proposto:** Realizar análises de diagnóstico para os modelos SMSN-ICR, seguindo Oliveira, Oliveira e Lachos (2023). Estender o modelo proposto para contextos não lineares censurados (SMSN-NLICR), conforme discutido por Garay, Lachos e Abanto-Valle (2011). Adaptar o modelo para trabalhar com dados longitudinais observados de forma irregular, utilizando a classe multivariada de distribuições SMSN, como mencionado em Garay et al. (2017). Desenvolver um modelo que possa lidar com dados faltantes ou censurados nas covariáveis.

Essas extensões são essenciais para ampliar a aplicabilidade do modelo SMSN em cenários mais complexos de análise de dados.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, p. 716–723, 1974.

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*, v. 21, p. 453–473, 2012.

ATKINSON, A. C. Two graphical displays for outlying and influential observations in regression. *Biometrika*, Oxford University Press, v. 68, n. 1, p. 13–20, 1981.

AZZALINI, A. A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, p. 171–178, 1985.

BARROS, M.; GALEA, M.; GONZÁLEZ, M.; LEIVA, V. Influence diagnostics in the tobit censored response model. *Statistical Methods & Applications*, Springer, v. 19, p. 379–397, 2010.

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

BOZDOGAN, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, v. 52, p. 345–370, 1987.

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

BURNHAM, K. P.; ANDERSON, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed.. ed. [S.l.]: Springer-Verlag, 2002.

CAMERON, A. C.; TRIVEDI, P. K. *Microeconometrics: methods and applications*. [S.l.]: Cambridge university press, 2005.

Central Statistical Service. *October Household Survey 1999*. Cape Town: [s.n.], 2000. DataFirst. Disponível em: <https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/64>.

DELYON, B.; LAVIELLE, M.; MOULINES, E. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, v. 27, n. 1, p. 94–128, 1999.

DEMPSTER, A.; LAIRD, N.; RUBIN, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,*, v. 39, p. 1–38, 1977.

FINTEL, D. V. Dealing with earnings bracket responses in household surveys–how sharp are midpoint imputations? *South African Journal of Economics*, Wiley Online Library, v. 75, n. 2, p. 293–312, 2007.

GALARZA, C. M.; KAN, R.; LACHOS., V. H. *MomTrunc: Moments of Folded and Doubly Truncated Multivariate Distributions*. [S.l.], 2022. R package version 6.0. Disponível em: <http://cran.r-project.org/package=MomTrunc>.

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*, v. 42, n. 12, p. 2694–2714, 2015.

GARAY, A. M.; CASTRO, L. M.; LESKOW, J.; LACHOS, V. H. Censored linear regression models for irregularly observed longitudinal data using the multivariate-t distribution. *Statistical Methods in Medical Research*, v. 26, n. 2, p. 542–566, 2017.

GARAY, A. M.; LACHOS, V. H.; ABANTO-VALLE, C. A. Nonlinear regression models based on scale mixtures of skew-normal distributions. *Journal of the Korean Statistical Society*, v. 40, p. 115–124, 2011.

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

GARAY, A. M.; MASSUIA, M. B.; LACHOS, V. *SMNCensReg: Fitting Univariate Censored Regression Model Under the Family of Scale Mixture of Normal Distributions*. [S.l.], 2022. R package version 3.1. Disponível em: <https://CRAN.R-project.org/package=SMNCensReg>.

GARAY, A. W.; MASSUIA, M. B.; LACHOS., V. H. *BayesCR: Bayesian Analysis of Censored Regression Models Under Scale Mixture of Skew Normal Distributions*. [S.l.], 2017. R package version 3.1. Disponível em: <http://cran.r-project.org/package=BayesCR>.

LABRA, F. V.; GARAY, A. M.; LACHOS, V. H.; ORTEGA, E. M. M. Estimation and diagnostics for heteroscedastic nonlinear regression models based on scale mixtures of skew-normal distributions. *Journal of Statistical Planning and Inference*, v. 142, p. 2149–2165, 2012.

LACHOS, V. H.; BAZáN, J. L.; CASTRO, L. M.; PARK, J. The skew-t censored regression model: parameter estimation via an EM-type algorithm. *Commnications for Statistical Applications and Methods*, v. 29, p. 333–351, 2022.

LACHOS, V. H.; GARAY, A.; CABRAL, C. R. Moments of truncated skew-normal/independent distributions. *Brazilian Journal of Probability and Statistics*, v. 34, p. 478–494, 2020.

LIN, T.-I. Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, Springer, v. 20, p. 343–356, 2010.

LIU, C.; RUBIN, D. B. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, v. 81, p. 633–648, 1994.

LOUIS, T. A. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 44, n. 2, p. 226–233, 1982.

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

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

MATTOS, T. d. B.; GARAY, A. M.; LACHOS, V. H. Likelihood-based inference for censored linear regression models with scale mixtures of skew-normal distributions. *Journal of Applied Statistics*, Taylor & Francis, v. 45, n. 11, p. 2039–2066, 2018.

MCCULLAGH, P.; NELDER, J. *Generalized Linear Models, Second Edition*. Chapman & Hall, 1989. (Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series). ISBN 9780412317606. Disponível em: <http://books.google.com/books?id=h9kFH2_FfBkC>.

MEILIJSON, I. A fast improvement to the em algorithm on its own terms. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 51, n. 1, p. 127–138, 1989.

MENG, X.; RUBIN, D. B. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, v. 81, p. 633–648, 1993.

MENG, X.-L.; DYK, D. V. The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 59, n. 3, p. 511–567, 1997.

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

OLIVEIRA, M. S.; OLIVEIRA, D. C.; LACHOS, V. H. Influence diagnostics for skew-t censored linear regression models. *Communications for Statistical Applications and Methods*, Korean Statistical Society, v. 30, n. 6, p. 605–629, 2023.

ORTEGA, E. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostics in generalized log-gamma regression models. *Computational statistics & data analysis*, Elsevier, v. 42, n. 1-2, p. 165–186, 2003.

ORTEGA, E. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostics in generalized log-gamma regression models. *Computational Statistics & Data Analysis*, v. 42, n. 1, p. 165–186, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>.

RTI-FDA. *[Snapshot of School Management Effectiveness: Peru Pilot Study*. [S.l.], 2008. 1–53 p. Disponível em: <https://www.globalreadingnetwork.net/sites/default/files/eddata/SSME_report_FDA_English.pdf>.

RUBIN, D. B. Inference and missing data. *Biometrika*, Oxford University Press, v. 63, n. 3, p. 581–592, 1976.

SCHWARZ, G. et al. Estimating the dimension of a model. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

THERNEAU, T. M.; GRAMBSCH, P. M.; FLEMING, T. R. Martingale-based residuals for survival models. *Biometrika*, Oxford University Press, v. 77, n. 1, p. 147–160, 1990.

APÊNDICE A – DETALHES DO ALGORITMO DO TIPO EM

Apresentamos as expressões necessárias envolvidas no Passo-E do nosso algoritmo ECME, considerando cada uma das distribuições SMSN. Por outro lado, a prova da Proposição 1 também é discutida.

Dessa forma, dado que $Y_i \sim \mathrm{SMSN}\left(\mathbf{x}_i^\top\boldsymbol{\beta} + b_1\Delta, \sigma^2, \lambda; H\right)$, definimos $\mu_i^* = \mathbf{x}_i^\top\boldsymbol{\beta} + b_1\Delta$ conforme discutido na subseção 3.2, e temos o seguinte:

• Para uma observação não censurada "i"

As expressões $\mathcal{E}_{00i}(\boldsymbol{\theta}) = \mathrm{E}\left[U_i \,|\, y_i, \boldsymbol{\theta}\right]$ e $\tau_i = \mathrm{E}\left[U_i^{1/2} W_\Phi\left(\frac{U_i^{1/2}\mu_{T_i}}{M_T}\right) \,|\, y_i, \boldsymbol{\theta}\right]$ em que $W_\Phi(x) = \frac{\phi(x)}{\Phi(x)},\ \text{para } x \in \mathbb{R}$, foram obtidos por Basso et al. (2010) e são apresentados a seguir:

- Se $Y_i \sim \mathrm{SN}\left(\mu_i^*, \sigma^2, \lambda\right)$

$$\mathcal{E}_{00i}(\boldsymbol{\theta}) = 1 \quad \text{e} \quad \tau_i = \frac{\phi\left(A_i^*\right)}{\Phi\left(A_i^*\right)}.$$

- Se $Y_i \sim \mathrm{ST}\left(\mu_i^*, \sigma^2, \lambda, \nu\right)$

$$\mathcal{E}_{00i}(\boldsymbol{\theta}) = \frac{2^2\nu^{\nu/2}\Gamma\left(\frac{\nu+3}{2}\right)\left(\nu + \mathrm{d}^*(y_i)\right)^{-\frac{\nu+3}{2}}}{\psi_{ST}(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \nu)\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\sigma^2}} \mathrm{T}\left(\sqrt{\frac{\nu+3}{\nu + \mathrm{d}^*(y_i)}} A_i^* \,\middle|\, \nu+3\right),$$

$$\tau_i = \frac{2\nu^{\nu/2}\Gamma(\frac{\nu+2}{2})\left(\nu + \mathrm{d}^*(y_i) + A_i^{*2}\right)^{-\frac{\nu+2}{2}}}{\psi_{ST}(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \nu)\Gamma\left(\frac{\nu}{2}\right)\pi\sigma}.$$

- Se $Y_i \sim \mathrm{SCN}\left(\mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}\right)$, com $\boldsymbol{\nu} = (\nu_1, \nu_2)$

$$\begin{aligned}\mathcal{E}_{00i}(\boldsymbol{\theta}) = \frac{2}{\psi_{SCN}(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu})} &\left\{\nu_1\nu_2\phi\left(y_i \,|\, \mu_i^*, \nu_2^{-1}\sigma^2\right)\Phi\left(\nu_2^{1/2}A_i^*\right)\right. \\ &\left. + (1-\nu_1)\phi\left(y_i \,|\, \mu_i^*, \sigma^2\right)\Phi\left(A_i^*\right)\right\} \\ \tau_i = \frac{2}{\psi_{SCN}(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu})} &\left\{\nu_1\nu_2^{1/2}\phi\left(y_i \,|\, \mu_i^*, \nu_2^{-1}\sigma^2\right)\phi\left(\nu_2^{1/2}A_i^*\right)\right. \\ &\left. + (1-\nu_1)\phi\left(y_i \,|\, \mu_i^*, \sigma^2\right)\phi\left(A_i^*\right)\right\},\end{aligned}$$

$$\text{com} \quad A_i^* = \frac{\mu_{T_i}}{M_T} = \frac{\lambda(y_i - \mu_i^*)}{\sigma} \quad \text{e} \quad \mathrm{d}^*(y_i) = \frac{(y_i - \mu_i^*)^2}{\sigma^2}. \tag{A.1}$$

• Para uma observação incompleta "i"

Conforme a Proposição 1, é necessário obter as expressões

$$\begin{aligned}\mathcal{E}_{0ri}(\boldsymbol{\theta}) &= \mathrm{E}[Y_i^r \mathrm{E}[U_i \,|\, Y_i] \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}] \;\text{ e}\\ \mathbf{W}_{\Psi}^k(\boldsymbol{\theta}) &= \mathrm{E}\left[Y_i^k \mathrm{E}\left[U_i^{1/2} W_{\Phi}\left(U_i^{1/2} A_i^*\right) \,|\, Y_i\right] \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}\right]\\ &= \mathrm{E}\left[Y_i^k \tau_i \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}\right]\end{aligned}$$

em que $k \in \{0, 1\}$, dessa forma, para cada uma das distribuições, tem-se o seguinte:

- Se $Y_i \sim \mathrm{SN}\left(\mu_i^*, \sigma^2, \lambda\right)$

$$\begin{aligned}\mathcal{E}_{0ri}(\boldsymbol{\theta}) &= \mathrm{E}[S_{1i}^r],\\ \mathbf{W}_{\Psi}^k(\boldsymbol{\theta}) &= \frac{\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}\left(\frac{\Phi\left(v_{2i} \,|\, \mu_i^*, \sigma_0^2\right) - \Phi\left(v_{1i} \,|\, \mu_i^*, \sigma_0^2\right)}{\Psi_{\mathrm{SN}}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda\right) - \Psi_{\mathrm{SN}}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda\right)}\right) \times \mathrm{E}\left[S_{2i}^r\right],\end{aligned}$$

em que $S_{1i} \sim \mathrm{TSN}\left(\mu_i^*, \sigma^2, \lambda; \lfloor v_{1i}, v_{2i} \rfloor\right)$, $S_{2i} \sim \mathrm{TN}\left(\mu_i^*, \sigma_0^2; \lfloor v_{1i}, v_{2i} \rfloor\right)$ e $\sigma_0^2 = \dfrac{\sigma^2}{1+\lambda^2}$.

- Se $Y_i \sim \mathrm{ST}\left(\mu_i^*, \sigma^2, \lambda, \nu\right)$

$$\begin{aligned}\mathcal{E}_{0ri}(\boldsymbol{\theta}) &= \left(\frac{\Psi_{\mathrm{ST}}\left(v_{2i} \,|\, \mu_i^*, \sigma_1^2, \lambda, \nu+2\right) - \Psi_{\mathrm{ST}}\left(v_{1i} \,|\, \mu_i^*, \sigma_1^2, \lambda, \nu+2\right)}{\Psi_{\mathrm{ST}}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right) - \Psi_{\mathrm{ST}}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)}\right) \times \mathrm{E}[W_{1i}^r],\\ \mathbf{W}_{\Psi}^k(\boldsymbol{\theta}) &= \frac{2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu(1+\lambda^2)\pi}}\left(\frac{\mathrm{T}(v_{2i} \,|\, \mu_i^*, \sigma_2^2, \nu+1) - \mathrm{T}(v_{1i} \,|\, \mu_i^*, \sigma_2^2, \nu+1)}{\Psi_{\mathrm{ST}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu) - \Psi_{\mathrm{ST}}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu)}\right) \times \mathrm{E}[W_{2i}^k],\end{aligned}$$

$W_{1i} \sim \mathrm{TST}\left(\mu_i^*, \sigma_1^2, \lambda, \nu+2; \lfloor v_{1i}, v_{2i} \rfloor\right)$, $W_{2i} \sim \mathrm{TT}\left(\mu_i^*, \sigma_2^2, \nu+1; \lfloor v_{1i}, v_{2i} \rfloor\right)$, $\sigma_1^2 = \dfrac{\nu}{\nu+2}\sigma^2$ e $\sigma_2^2 = \dfrac{\nu}{(\nu+1)(1+\lambda^2)}\sigma^2$.

- Se $Y_i \sim \mathrm{SCN}\left(\mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}\right)$, com $\boldsymbol{\nu} = (\nu_1, \nu_2)$

$$\begin{aligned}\mathcal{E}_{0ri}(\boldsymbol{\theta}) = \frac{1}{\Psi_{\mathrm{SCN}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}) - \Psi_{\mathrm{SCN}}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu})}\Big\{&\nu_1\nu_2\\ &\times \left(\Psi_{\mathrm{SN}}(v_{2i} \,|\, \mu_i^*, \nu_2^{-1}\sigma^2, \lambda) - \Psi_{\mathrm{SN}}(v_{1i} \,|\, \mu_i^*, \nu_2^{-1}\sigma^2, \lambda)\right) \times \mathrm{E}[Z_{1i}^r]\\ &+ (1-\nu_1)\left(\Psi_{\mathrm{SN}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda) - \Psi_{\mathrm{SN}}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda)\right) \times \mathrm{E}[Z_{2i}^r]\Big\}\\ \mathbf{W}_{\Psi}^k(\boldsymbol{\theta}) = \frac{1}{\Psi_{\mathrm{SCN}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}) - \Psi_{SCN}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu})}\Big\{&c_1(\nu)\\ &\times \left(\Phi(v_{2i} \,|\, \mu_i^*, \nu_2^{-1}\sigma_0^2) - \Phi(v_{1i} \,|\, \mu_i^*, \nu_2^{-1}\sigma_0^2)\right) \times \mathrm{E}[Z_{3i}^k]\\ &+ c_2(\nu) \times \left(\Phi(v_{2i} \,|\, \mu_i^*, \sigma_0^2) - \Phi(v_{1i} \,|\, \mu_i^*, \sigma_0^2)\right) \times \mathrm{E}[Z_{4i}^k]\Big\}\end{aligned}$$

com $Z_{1i} \sim \text{TSN}\left(\mu_i^*, \nu_2^{-1}\sigma^2, \lambda; \lfloor v_{1i}, v_{2i} \rfloor\right)$, $Z_{2i} \sim \text{TSN}\left(\mu_i^*, \sigma^2, \lambda; \lfloor v_{1i}, v_{2i} \rfloor\right)$, $Z_{3i} \sim \text{TN}\left(\mu_i^*, \nu_2^{-1}\sigma_0^2; \lfloor v_{1i}, v_{2i} \rfloor\right)$, $Z_{4i} \sim \text{TN}\left(\mu_i^*, \sigma_0^2; \lfloor v_{1i}, v_{2i} \rfloor\right)$, $c_1(\boldsymbol{\nu}) = \dfrac{\sqrt{2}\nu_1\nu_2^{1/2}}{\sqrt{\pi(1+\lambda^2)}}$, $c_2(\boldsymbol{\nu}) = \dfrac{\sqrt{2}(1-\nu_1)}{\sqrt{\pi(1+\lambda^2)}}$ e $\sigma_0^2 = \dfrac{\sigma^2}{1+\lambda^2}$.

Segue abaixo mais detalhes sobre a prova da Proposição 1.

$$
\begin{aligned}
\mathcal{E}_{10i}(\boldsymbol{\theta}^{(k)}) &= \text{E}[U_i T_i \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}] \\
&= \text{E}\left[\text{E}\left[U_i\left(\mu_{T_i} + b_1 + W_\Phi\left(U_i^{1/2}A_i^*\right)U_i^{-1/2}M_T\right)|\, Y_i\right]|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}\right] \\
&= \text{E}\left[\text{E}\left[U_i\left(\frac{\Delta}{\Delta^2+\Gamma}\right)(Y_i - \mu_i^*) + b_1 U_i\right.\right. \\
&\quad \left.\left. + M_T U_i^{1/2} W_\Phi\left(U_i^{1/2}A_i^*\right)|\, Y_i\right]|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}\right] \\
&= \left(\frac{\Delta}{\Delta^2+\Gamma}\right)\left(\mathcal{E}_{01i}(\boldsymbol{\theta}) - \mathcal{E}_{00i}(\boldsymbol{\theta})\mu_i^*\right) + b_1\mathcal{E}_{00i}(\boldsymbol{\theta}) + \sqrt{\frac{\Gamma}{\Delta^2+\Gamma}}\mathbf{W}_\Psi^0(\boldsymbol{\theta}).
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{E}_{20i}(\boldsymbol{\theta}^{(k)}) &= \text{E}[U_i T_i^2 \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}] \\
&= \text{E}\left[\text{E}\left[U_i\left((\mu_{T_i} + b_1)^2 + U_i^{-1}M_T^2\right.\right.\right. \\
&\quad \left.\left.\left. + (\mu_{T_i} + 2b_1)U_i^{-1/2}M_T W_\Phi\left(U_i^{1/2}A_i^*\right)\right)|\, Y_i\right]|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}\right] \\
&= \left(\frac{\Delta}{\Delta^2+\Gamma}\right)^2\left(\mathcal{E}_{02i}(\boldsymbol{\theta}) - 2\mathcal{E}_{01i}(\boldsymbol{\theta})\mu_i^* + \mathcal{E}_{00i}(\boldsymbol{\theta})(\mu_i^*)^2\right) \\
&\quad + 2b_1\left(\frac{\Delta}{\Delta^2+\Gamma}\right)\left(\mathcal{E}_{01i}(\boldsymbol{\theta}) - \mathcal{E}_{00i}(\boldsymbol{\theta})\mu_i^*\right) + b_1^2\mathcal{E}_{00i}(\boldsymbol{\theta}) + \frac{\Gamma}{\Delta^2+\Gamma} \\
&\quad + \sqrt{\frac{\Gamma}{\Delta^2+\Gamma}}\left\{\left(\frac{\Delta}{\Delta^2+\Gamma}\right)\mathbf{W}_\Psi^1(\boldsymbol{\theta}) - \left(\left(\frac{\Delta}{\Delta^2+\Gamma}\right)\mu^* - 2b_1\right)\mathbf{W}_\Psi^0(\boldsymbol{\theta}).\right\}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{E}_{11i}(\boldsymbol{\theta}^{(k)}) &= \text{E}[U_i T_i Y_i \,|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}] \\
&= \text{E}\left[Y_i\text{E}\left[U_i\left(\mu_{T_i} + b_1 + W_\Phi\left(U_i^{1/2}A_i^*\right)U_i^{-1/2}M_{T_i}\right)|\, Y_i\right]|\, v_{1i} \leq Y_i \leq v_{2i}, \boldsymbol{\theta}\right] \\
&= \frac{\Delta}{\Delta^2+\Gamma}\left\{\mathcal{E}_{02i}(\boldsymbol{\theta}) - \mathcal{E}_{01i}(\boldsymbol{\theta})\mu_i^*\right\} + b_1\mathcal{E}_{01i}(\boldsymbol{\theta}) + \sqrt{\frac{\Gamma}{\Delta^2+\Gamma}}\mathbf{W}_\Psi^1(\boldsymbol{\theta}).
\end{aligned}
$$

em que $A_i^* = \dfrac{\lambda(y_i - \mu_i^*)}{\sigma}$. Por outro lado, as expressões $\mathcal{E}_{0ri}(\boldsymbol{\theta})$ e $\mathbf{W}_\Psi^k(\boldsymbol{\theta})$ para as distribuições SMSN são:

- Para SN

$$\begin{aligned}\mathbf{W}_{\Psi}^{k}(\boldsymbol{\theta}) &= \int_{v_{1i}}^{v_{2i}} y_i^k \frac{\phi(A_i^*)}{\Phi(A_i^*)} \frac{2\phi(y_i \,|\, \mu_i^*, \sigma^2)\Phi(A_i^*)}{\Psi_{\mathrm{SN}}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda\right) - \Psi_{\mathrm{SN}}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda\right)} dy_i \\ &= \frac{\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}} \left(\frac{\Phi\left(v_{2i} \,|\, \mu_i^*, \sigma_0^2\right) - \Phi\left(v_{1i} \,|\, \mu_i^*, \sigma_0^2\right)}{\Psi_{\mathrm{SN}}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda\right) - \Psi_{\mathrm{SN}}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda\right)} \right) \times \mathrm{E}\left[S_{2i}^r\right],\end{aligned}$$

- Para ST

$$\begin{aligned}\mathcal{E}_{0ri}(\boldsymbol{\theta}) &= \int_{v_{1i}}^{v_{2i}} \left\{ y_i^r \frac{t\left(y_i \,|\, \mu_i^*, \sigma^2, \nu\right)}{\psi_{\mathrm{ST}}\left(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)} \frac{2(\nu+1)}{(\nu + \mathrm{d}^*(y_i))} \mathrm{T}\left(\sqrt{\frac{\nu+3}{\nu + \mathrm{d}^*(y_i)}} A_i^* \,\middle|\, \nu+3 \right) \right. \\ &\quad \left. \times \frac{\psi_{\mathrm{ST}}\left(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)}{\Psi_{\mathrm{ST}}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right) - \Psi_{\mathrm{ST}}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)} \right\} dy_i \\ &= \left(\frac{\nu+1}{\nu}\right) \frac{1}{\Psi_{\mathrm{ST}}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right) - \Psi_{\mathrm{ST}}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)} \qquad \text{(A.2)} \\ &\quad \times \int_{v_{1i}}^{v_{2i}} y_i^r 2 \left(\frac{\nu}{\nu+1}\right) t\left(y_i \,|\, \mu_i^*, \sigma_1^2, \nu+2\right) \mathrm{T}\left(\sqrt{\frac{\nu+3}{\nu+2+\mathrm{d}_1^*(y_i)}} A_{1i}^* \,\middle|\, \nu+3 \right) dy_i \\ &= \left(\frac{\Psi_{\mathrm{ST}}\left(v_{2i} \,|\, \mu_i^*, \sigma_1^2, \lambda, \nu+2\right) - \Psi_{\mathrm{ST}}\left(v_{1i} \,|\, \mu_i^*, \sigma_1^2, \lambda, \nu+2\right)}{\Psi_{\mathrm{ST}}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right) - \Psi_{\mathrm{ST}}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)} \right) \times \mathrm{E}[W_{1i}^r]\end{aligned}$$

em que a Equação (A.2) é obtida a partir das expressões:

$$\begin{aligned}\left(1 + \frac{\mathrm{d}^*(y_i)}{\nu}\right)^{-1} t\left(y_i \,|\, \mu_i^*, \sigma^2, \nu\right) &= \left(1 + \frac{\mathrm{d}^*(y_i)}{\nu}\right)^{-1} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \left(1 + \frac{\mathrm{d}^*(y_i)}{\nu}\right)^{-(\nu+1)/2} \\ &= \frac{\nu}{\nu+1} t\left(y_i \,|\, \mu_i^*, \sigma_1^2, \nu+2\right) \\ \sqrt{\frac{\nu+3}{\nu+2+\mathrm{d}_1^*(y_i)}} A_{1i}^* &= \sqrt{\frac{(\nu+3)\sigma_1^2}{\sigma_1^2(\nu+2) + (y_i - \mu_i^*)^2}} \left(\lambda \frac{(y_i - \mu_i^*)}{\sigma_1} \right) \\ &= \sqrt{\frac{\nu+3}{\nu + \mathrm{d}^*(y_i)}} A_i^*\end{aligned}$$

com $\mathrm{d}_1^*(y_i) = \frac{(y_i - \mu_i^*)^2}{\sigma_1^2}$, $A_{1i}^* = \frac{\lambda(y_i - \mu_i^*)}{\sigma_1}$ e $\sigma_1^2 = \dfrac{\nu}{\nu+2}\sigma^2$

$$\begin{aligned}
\mathbf{W}_{\Psi}^{k}(\boldsymbol{\theta}) &= \int_{v_{1i}}^{v_{2i}} \left\{ y_i^k \frac{t\left(y_i \,|\, \mu_i^*, \sigma^2, \nu\right)}{\psi_{\mathrm{ST}}\left(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)} \frac{2}{\sqrt{\pi}} \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu+1}{2})} \frac{(\nu + \mathrm{d}^*(y_i))^{(\nu+1)/2}}{(\nu + \mathrm{d}^*(y_i) + A_i^{*2})^{(\nu+2)/2}} \right. \\
&\quad \left. \times \frac{\psi_{\mathrm{ST}}\left(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)}{\Psi_{\mathrm{ST}}\left(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right) - \Psi_{\mathrm{ST}}\left(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu\right)} \right\} dy_i \\
&= \frac{2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu(1+\lambda^2)\pi}} \left(\frac{1}{\Psi_{\mathrm{ST}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu) - \Psi_{\mathrm{ST}}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu)} \right) \quad \text{(A.3)} \\
&\quad \times \int_{v_{1i}}^{v_{2i}} y_i^k t\left(y_i \,|\, \mu_i^*, \sigma_2^2, \nu+1\right) dy_i \\
&= \frac{2\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu(1+\lambda^2)\pi}} \left(\frac{\mathrm{T}(v_{2i} \,|\, \mu_i^*, \sigma_2^2, \nu+1) - \mathrm{T}(v_{1i} \,|\, \mu_i^*, \sigma_2^2, \nu+1)}{\Psi_{\mathrm{ST}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu) - \Psi_{\mathrm{ST}}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \nu)} \right) \times \mathrm{E}[W_{2i}^k],
\end{aligned}$$

em que a Equação (A.3) é derivada de:

$$\begin{aligned}
& t\left(y_i \,|\, \mu_i^*, \sigma^2, \nu\right) \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu+1}{2})} \frac{(\nu + \mathrm{d}^*(y_i))^{(\nu+1)/2}}{(\nu + \mathrm{d}^*(y_i) + A_i^{*2})^{(\nu+2)/2}} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \left(1 + \frac{\mathrm{d}^*(y_i)}{\nu}\right)^{-(\nu+1)/2} \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu+1}{2})} \frac{(\nu + \mathrm{d}^*(y_i))^{(\nu+1)/2}}{(\nu + \mathrm{d}^*(y_i) + A_i^{*2})^{(\nu+2)/2}} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu(1+\lambda^2)}} t\left(y_i \,|\, \mu_i^*, \sigma_2^2, \nu+1\right),
\end{aligned}$$

com $\mathrm{d}_2^*(y_i) = \frac{(y_i - \mu_i^*)^2}{\sigma_2^2}$ e $\sigma_2^2 = \dfrac{\nu}{(\nu+1)(1+\lambda^2)}\sigma^2$.

- Para SCN

$$\begin{aligned}
\mathcal{E}_{0ri}(\boldsymbol{\theta}) &= \int_{v_{1i}}^{v_{2i}} \left\{ y_i^r \frac{2}{\psi_{\mathrm{SCN}}\left(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}\right)} \left(\nu_1\nu_2 \phi(y_i \,|\, \mu_i^*, \nu_2^{-1}\sigma^2) \Phi(\nu_2^{1/2} A_i^*) + (1-\nu_1) \right. \right. \\
&\quad \left. \left. \times \, \phi(y_i \,|\, \mu_i^*, \sigma^2) \Phi(A_i^*) \right) \frac{\psi_{\mathrm{SCN}}\left(y_i \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}\right)}{\Psi_{\mathrm{SCN}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}) - \Psi_{\mathrm{SCN}}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu})} \right\} dy_i \\
&= \frac{1}{\Psi_{\mathrm{SCN}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}) - \Psi_{\mathrm{SCN}}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu})} \Big\{ \nu_1\nu_2 \\
&\quad \times \left(\Psi_{\mathrm{SN}}(v_{2i} \,|\, \mu_i^*, \nu_2^{-1}\sigma^2, \lambda) - \Psi_{\mathrm{SN}}(v_{1i} \,|\, \mu_i^*, \nu_2^{-1}\sigma^2, \lambda) \right) \times \mathrm{E}[Z_{1i}^r] \\
&\quad + (1-\nu_1) \left(\Psi_{\mathrm{SN}}(v_{2i} \,|\, \mu_i^*, \sigma^2, \lambda) - \Psi_{\mathrm{SN}}(v_{1i} \,|\, \mu_i^*, \sigma^2, \lambda) \right) \times \mathrm{E}[Z_{2i}^r] \Big\}
\end{aligned}$$

$$
\begin{aligned}
\mathbf{W}_{\Psi}^{k}(\boldsymbol{\theta}) = \int_{v_{1i}}^{v_{2i}} \Bigg\{ & y_i^k \frac{2}{\psi_{\mathrm{SCN}}\left(y_i|\mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}\right)} \left(\nu_1 \nu_2^{1/2} \phi(y_i|\mu_i^*, \nu_2^{-1}\sigma^2)\phi(\nu_2^{1/2} A_i^*) + (1-\nu_1)\right. \\
& \left. \times\ \phi(y_i|\mu_i^*, \sigma^2)\phi(A_i^*)\right) \frac{\psi_{\mathrm{SCN}}\left(y_i|\mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}\right)}{\Psi_{\mathrm{SCN}}(v_{2i}\,|\,\mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}) - \Psi_{\mathrm{SCN}}(v_{1i}\,|\,\mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu})} \Bigg\}\, dy_i \\
= & \frac{1}{\Psi_{\mathrm{SCN}}(v_{2i}\,|\,\mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu}) - \Psi_{SCN}(v_{1i}\,|\,\mu_i^*, \sigma^2, \lambda, \boldsymbol{\nu})} \Big\{ c_1(\nu) \\
& \times \left(\Phi(v_{2i}\,|\,\mu_i^*, \nu_2^{-1}\sigma_0^2) - \Phi(v_{1i}\,|\,\mu_i^*, \nu_2^{-1}\sigma_0^2)\right) \times \mathrm{E}[Z_{3i}^k] \\
& + c_2(\nu) \times \left(\Phi(v_{2i}\,|\,\mu_i^*, \sigma_0^2) - \Phi(v_{1i}\,|\,\mu_i^*, \sigma_0^2)\right) \times \mathrm{E}[Z_{4i}^k] \Big\}
\end{aligned}
$$

Para este caso, leve em consideração a igualdade:

$$
\phi(y_i\,|\,\mu_i^*, \nu_2^{-1}\sigma^2)\phi(\nu_2^{1/2} A_i^*) = \frac{1}{\sqrt{2\pi\nu_2^{-1}\sigma^2}} \sqrt{\frac{\nu_2^{-1}\sigma^2}{1+\lambda^2}}\, \phi(y_i\,|\,\mu_i^*, \nu_2^{-1}\sigma_0^2)
$$

em que $\sigma_0^2 = \dfrac{\sigma^2}{1+\lambda^2}$.

APÊNDICE B – RESULTADOS COMPLEMENTARES DOS ESTUDOS DE SIMULAÇÃO

Estudo de simulação I

As Figuras 11 - 18 mostram os resultados para viés (Bias) e erro quadrático médio (MSE), considerando os níveis de censura $p \in \{0\%, 8\%, 20\%, 35\%\}$ e dados faltantes $m \in \{0\%, 5\%\}$. Além disso, as Tabelas 10 e 11 apresentam os resultados de Bias, MSE e viés relativo (RBias) para os modelos ST-ICR e SCN-ICR, respectivamente.

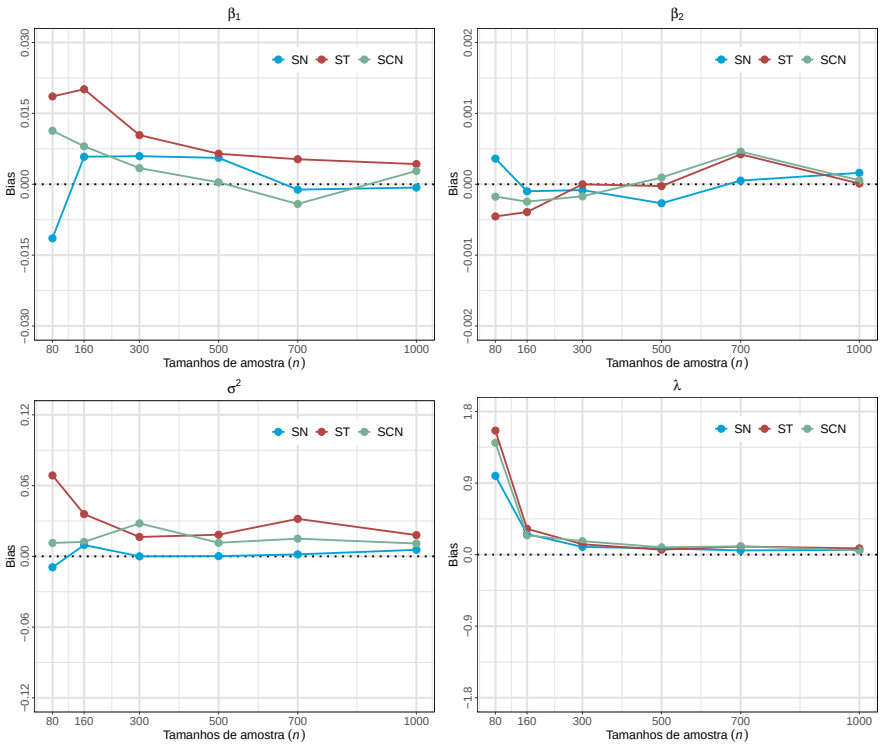

Figura 11 – Estudo de simulação 1. Viés das estimativas de parâmetros dos modelos SMSN-ICR, com níveis de censura e dados faltantes de 0% e 0%, respectivamente.

Estudo de simulação II

As figuras 19 - 21 apresentam os valores médios das mudanças relativas para todos os parâmetros para níveis de censura $p \in \{8\%, 20\%, 35\%\}$, cada um com 5% de dados faltantes.

Tabela 10 – Simulação 1. Resultados do Bias, MSE e RBias das estimativas dos parâmetros do modelo ST-ICR com diferentes tamanhos de amostra (n), níveis de censura intervalar (p) e de dados faltantes (m).

ST								
Nível Cens. (Falt.)		**Medida**	**Tamanhos de amostra (n)**					
			80	**160**	**300**	**500**	**700**	**1000**
0% (0%)	β_1	Bias	0.0186	0.0201	0.0104	0.0065	0.0053	0.0043
		MSE	0.0843	0.0359	0.0172	0.0101	0.0070	0.0055
		RBias	0.0124	0.0134	0.0069	0.0043	0.0035	0.0028
	β_2	Bias	-0.0005	-0.0004	0.0001^*	0.0001^*	0.0004	0.0001^*
		MSE	0.0002	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	-0.0002	-0.0002	0.0001^*	0.0001^*	0.0002	0.0001^*
	σ^2	Bias	0.0686	0.0358	0.0164	0.0183	0.0318	0.0180
		MSE	0.8396	0.1233	0.0643	0.0417	0.0296	0.0202
		RBias	0.0457	0.0239	0.0110	0.0122	0.0212	0.0120
	λ	Bias	1.5603	0.3246	0.1303	0.0622	0.1018	0.0784
		MSE	75.309	1.1203	0.3701	0.2107	0.1659	0.1046
		RBias	0.6501	0.1352	0.0543	0.0259	0.0424	0.0327
8% (5%)	β_1	Bias	0.0052	-0.0032	0.0042	-0.0027	0.0085	0.0027
		MSE	0.0735	0.0342	0.0205	0.0117	0.0078	0.0059
		RBias	0.0035	-0.0021	0.0028	-0.0018	0.0057	0.0018
	β_2	Bias	0.0001	0.0003	-0.0002	0.0002	0.0001^*	0.0001
		MSE	0.0002	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	0.0001^*	0.0001	-0.0001	0.0001	0.0001^*	0.0001^*
	σ^2	Bias	0.0612	0.0185	-0.0079	0.0061	0.0243	0.0209
		MSE	0.2816	0.1230	0.0665	0.0419	0.0323	0.0224
		RBias	0.0408	0.0124	-0.0053	0.0041	0.0162	0.0139
	λ	Bias	1.6499	0.2387	0.0839	0.0547	0.1052	0.0641
		MSE	79.395	0.8318	0.3722	0.1741	0.1805	0.1154
		RBias	0.6875	0.0994	0.0350	0.0228	0.0438	0.0267
15% (5%)	β_1	Bias	0.0152	0.0192	0.0070	0.0062	0.0041	0.0004
		MSE	0.0895	0.0332	0.0217	0.0122	0.0085	0.0066
		RBias	0.0101	0.0128	0.0046	0.0042	0.0027	0.0003
	β_2	Bias	-0.0003	0.0002	-0.0004	0.0001^*	0.0002	0.0002
		MSE	0.0002	0.0001	0.0001	0.0001^*	0.0001^*	0.0001^*
		RBias	-0.0002	0.0001	-0.0002	0.0001^*	0.0001	0.0001
	σ^2	Bias	0.0525	0.0730	0.0195	0.0221	0.0318	0.0035
		MSE	0.3335	0.1510	0.0834	0.0498	0.0343	0.0231
		RBias	0.0350	0.0487	0.0130	0.0147	0.0212	0.0023
	λ	Bias	2.8414	0.4280	0.1578	0.1496	0.0953	0.0696
		MSE	167.53	2.6707	0.4699	0.2946	0.1687	0.1180
		RBias	1.1839	0.1783	0.0657	0.0623	0.0397	0.0290
20% (5%)	β_1	Bias	-0.0040	0.0116	0.0075	0.0131	0.0034	0.0024
		MSE	0.0920	0.0387	0.0230	0.0147	0.0095	0.0077
		RBias	-0.0027	0.0077	0.0050	0.0087	0.0023	0.0016
	β_2	Bias	0.0004	-0.0004	0.0003	-0.0001	0.0004	0.0002
		MSE	0.0002	0.0001	0.0001	0.0001^*	0.0001^*	0.0001^*
		RBias	0.0002	-0.0002	0.0001	0.0001^*	0.0002	0.0001
	σ^2	Bias	0.0422	0.0280	0.0228	0.0393	0.0246	0.0173
		MSE	0.3294	0.1629	0.0827	0.0516	0.0360	0.0266
		RBias	0.0281	0.0186	0.0152	0.0262	0.0164	0.0115
	λ	Bias	3.8671	0.4654	0.2070	0.1200	0.0969	0.0632
		MSE	245.49	11.251	0.6477	0.2629	0.1989	0.1247
		RBias	1.6113	0.1939	0.0863	0.0500	0.0404	0.0263
35% (5%)	β_1	Bias	-0.0025	0.0085	0.0098	0.0040	0.0071	0.0067
		MSE	0.1189	0.0487	0.0258	0.0154	0.0118	0.0086
		RBias	-0.0017	0.0057	0.0066	0.0026	0.0048	0.0044
	β_2	Bias	-0.0007	-0.0005	-0.0006	0.0001^*	0.0001^*	0.0001
		MSE	0.0003	0.0001	0.0001	0.0001^*	0.0001^*	0.0001^*
		RBias	-0.0003	-0.0003	-0.0003	0.0001^*	0.0001^*	0.0001
	σ^2	Bias	0.0235	0.0139	0.0047	0.0207	0.0195	0.0279
		MSE	0.4519	0.1913	0.1062	0.0593	0.0428	0.0298
		RBias	0.0156	0.0093	0.0031	0.0138	0.0130	0.0186
	λ	Bias	4.4308	0.6767	0.1468	0.1244	0.1362	0.1048
		MSE	238.54	10.903	0.6271	0.2898	0.2198	0.1617
		RBias	1.8462	0.2820	0.0612	0.0518	0.0567	0.0437

0.0001^* indica que o número é menor que 0.0001 (< 0.0001).

Tabela 11 – Simulação 1. Resultados do Bias, MSE e RBias das estimativas dos parâmetros do modelo SCN-ICR com diferentes tamanhos de amostra (n), níveis de censura intervalar (p) e de dados faltantes (m).

SCN								
Nível Cens. (Falt.)		**Medida**	**Tamanhos de amostra (n)**					
			80	**160**	**300**	**500**	**700**	**1000**
0% (0%)	β_1	Bias	0.0113	0.0080	0.0034	0.0004	-0.0042	0.0028
		MSE	0.0560	0.0236	0.0119	0.0066	0.0052	0.0040
		RBias	0.0076	0.0054	0.0023	0.0003	-0.0028	0.0019
	β_2	Bias	-0.0002	-0.0002	-0.0002	0.0001	0.0005	0.0001
		MSE	0.0002	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	-0.0001	-0.0001	-0.0001	0.0001^*	0.0002	0.0001^*
	σ^2	Bias	0.0114	0.0123	0.0280	0.0115	0.0150	0.0109
		MSE	0.2361	0.1012	0.0536	0.0291	0.0238	0.0174
		RBias	0.0076	0.0082	0.0186	0.0077	0.0100	0.0073
	λ	Bias	1.4067	0.2432	0.1694	0.0923	0.1038	0.0527
		MSE	91.469	0.9330	0.3810	0.1834	0.1361	0.0956
		RBias	0.5861	0.1013	0.0706	0.0385	0.0433	0.0220
8% (5%)	β_1	Bias	0.0096	0.0158	0.0081	-0.0062	0.0013	0.0028
		MSE	0.0534	0.0248	0.0139	0.0092	0.0053	0.0043
		RBias	0.0064	0.0105	0.0054	-0.0041	0.0009	0.0019
	β_2	Bias	-0.0009	-0.0007	-0.0005	0.0004	0.0001	-0.0001
		MSE	0.0002	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	-0.0004	-0.0004	-0.0002	0.0002	0.0001^*	0.0001^*
	σ^2	Bias	0.0217	0.0177	-0.0023	0.0115	0.0081	0.0144
		MSE	0.2920	0.1293	0.0595	0.0401	0.0229	0.0196
		RBias	0.0145	0.0118	-0.0016	0.0077	0.0054	0.0096
	λ	Bias	2.0195	0.3416	0.1129	0.1118	0.0745	0.0766
		MSE	195.86	1.4924	0.3748	0.2551	0.1504	0.1158
		RBias	0.8414	0.1423	0.0470	0.0466	0.0310	0.0319
15% (5%)	β_1	Bias	-0.0150	-0.0071	0.0099	0.0084	0.0025	0.0014
		MSE	0.0633	0.0259	0.0162	0.0091	0.0064	0.0043
		RBias	-0.0100	-0.0047	0.0066	0.0056	0.0017	0.0009
	β_2	Bias	0.0004	0.0006	-0.0001	-0.0006	0.0001	0.0001^*
		MSE	0.0002	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	0.0002	0.0003	0.0001^*	-0.0003	0.0001^*	0.0001^*
	σ^2	Bias	-0.0071	0.0159	0.0180	0.0024	0.0168	0.0131
		MSE	0.2788	0.1434	0.0696	0.0412	0.0284	0.0197
		RBias	-0.0048	0.0106	0.0120	0.0016	0.0112	0.0088
	λ	Bias	2.1694	0.2745	0.1517	0.0872	0.0934	0.0661
		MSE	113.11	1.4110	0.4331	0.2373	0.1689	0.1209
		RBias	0.9039	0.1144	0.0632	0.0364	0.0389	0.0275
20% (5%)	β_1	Bias	-0.0048	0.0121	0.0026	0.0059	-0.0040	-0.0027
		MSE	0.0730	0.0270	0.0145	0.0089	0.0068	0.0047
		RBias	-0.0032	0.0081	0.0018	0.0040	-0.0027	-0.0018
	β_2	Bias	0.0001	-0.0004	-0.0002	-0.0003	0.0003	0.0001
		MSE	0.0002	0.0001	0.0001^*	0.0001^*	0.0001^*	0.0001^*
		RBias	0.0001	-0.0002	-0.0001	-0.0001	0.0002	0.0001
	σ^2	Bias	-0.0029	0.0222	0.0002	0.0130	0.0057	0.0023
		MSE	0.3022	0.1502	0.0735	0.0416	0.0338	0.0217
		RBias	-0.0019	0.0148	0.0001	0.0087	0.0038	0.0016
	λ	Bias	2.5774	0.3550	0.1475	0.1239	0.0942	0.0637
		MSE	148.77	1.9227	0.5785	0.2596	0.1988	0.1222
		RBias	1.0739	0.1479	0.0614	0.0516	0.0392	0.0266
35% (5%)	β_1	Bias	0.0007	-0.0034	0.0153	-0.0017	0.0013	0.0015
		MSE	0.0898	0.0367	0.0211	0.0120	0.0083	0.0067
		RBias	0.0004	-0.0023	0.0102	-0.0011	0.0008	0.0010
	β_2	Bias	-0.0002	0.0003	-0.0005	0.0001	-0.0002	0.0001
		MSE	0.0002	0.0001	0.0001	0.0001^*	0.0001^*	0.0001^*
		RBias	-0.0001	0.0001	-0.0003	0.0001	-0.0001	0.0001^*
	σ^2	Bias	0.0231	0.0158	0.0141	0.0031	0.0111	0.0081
		MSE	0.3801	0.1817	0.0916	0.0475	0.0363	0.0288
		RBias	0.0154	0.0105	0.0094	0.0021	0.0074	0.0054
	λ	Bias	4.9217	0.6239	0.2389	0.1468	0.1003	0.0828
		MSE	268.12	11.381	0.7544	0.4200	0.2189	0.1682
		RBias	2.0507	0.2600	0.0995	0.0612	0.0418	0.0345

0.0001^* indica que o número é menor que 0.0001 (< 0.0001).

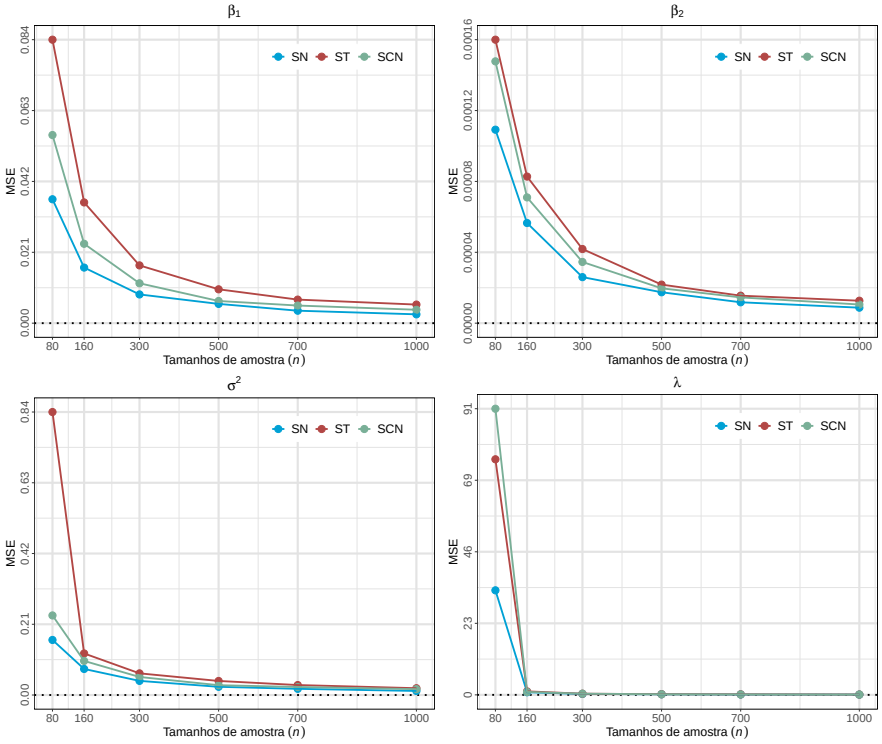

Figura 12 – Estudo de simulação 1. Erro quadrático médio das estimativas de parâmetros dos modelos SMSN-ICR, com níveis de censura de 0% e dados faltantes de 0%, respectivamente.

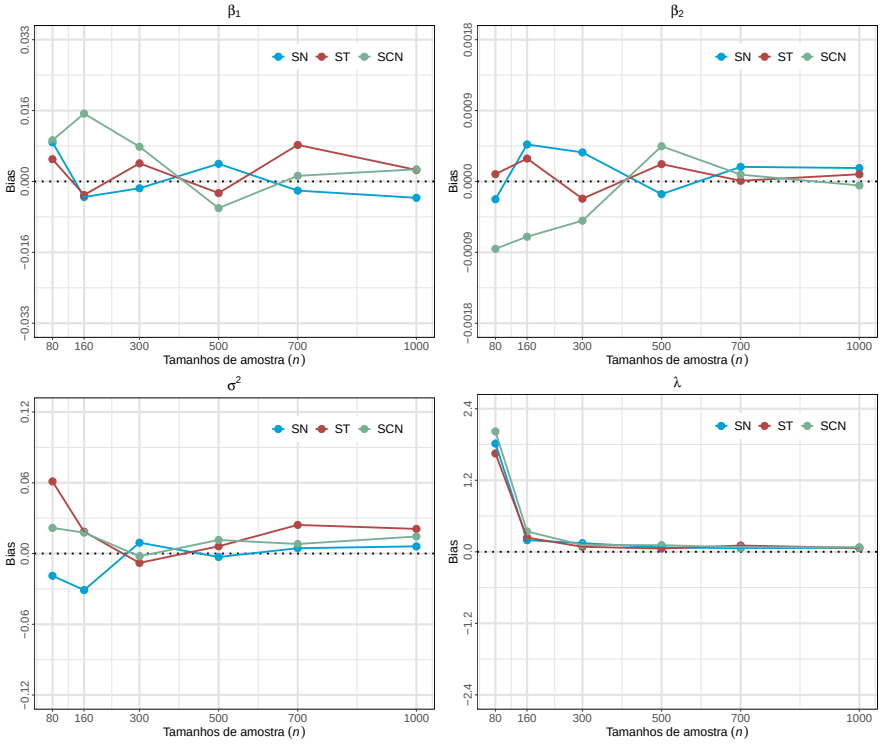

Figura 13 – Estudo de simulação 1. Viés das estimativas de parâmetros dos modelos SMSN-ICR, com níveis de censura e dados faltantes de 8% e 5%, respectivamente.

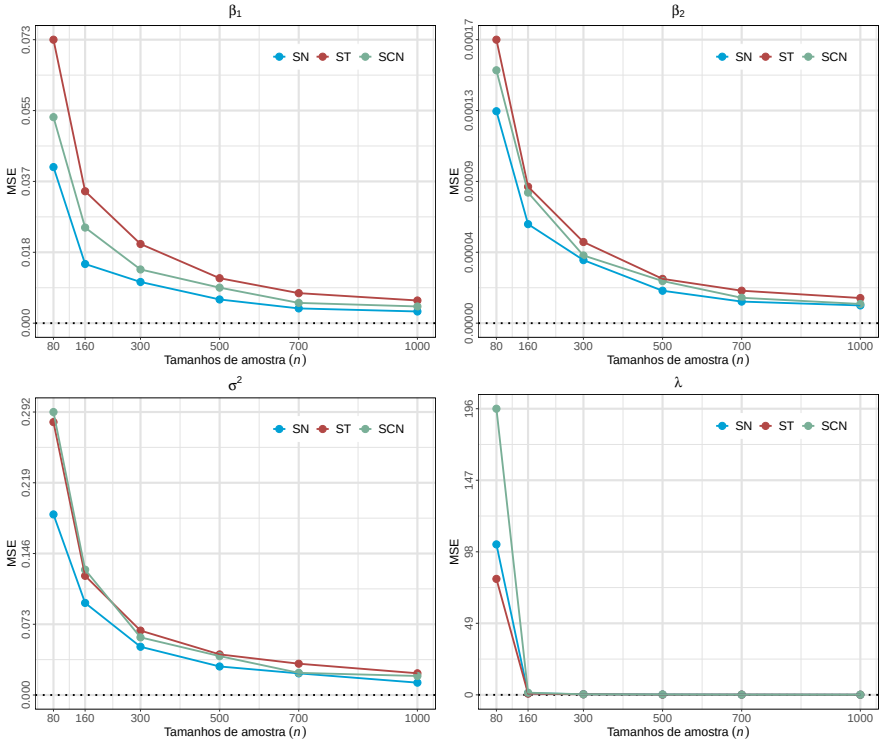

Figura 14 – Estudo de simulação 1. Erro quadrático médio das estimativas de parâmetros dos modelos SMSN-ICR, com níveis de censura de 8% e dados faltantes de 5%, respectivamente.

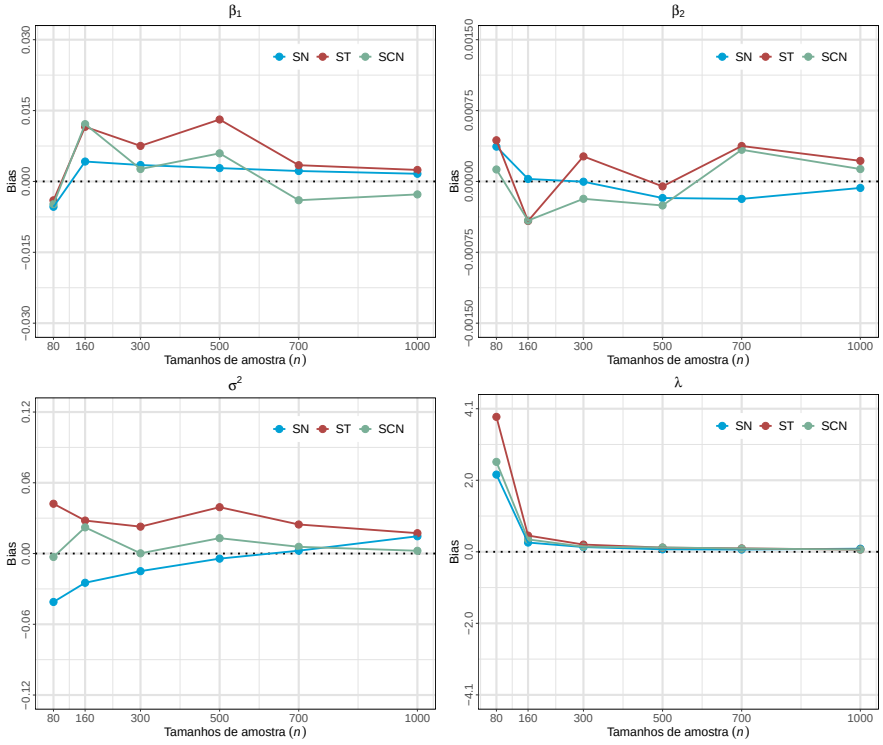

Figura 15 – Estudo de simulação 1. Viés das estimativas de parâmetros dos modelos SMSN-ICR, com níveis de censura e dados faltantes de 20% e 5%, respectivamente.

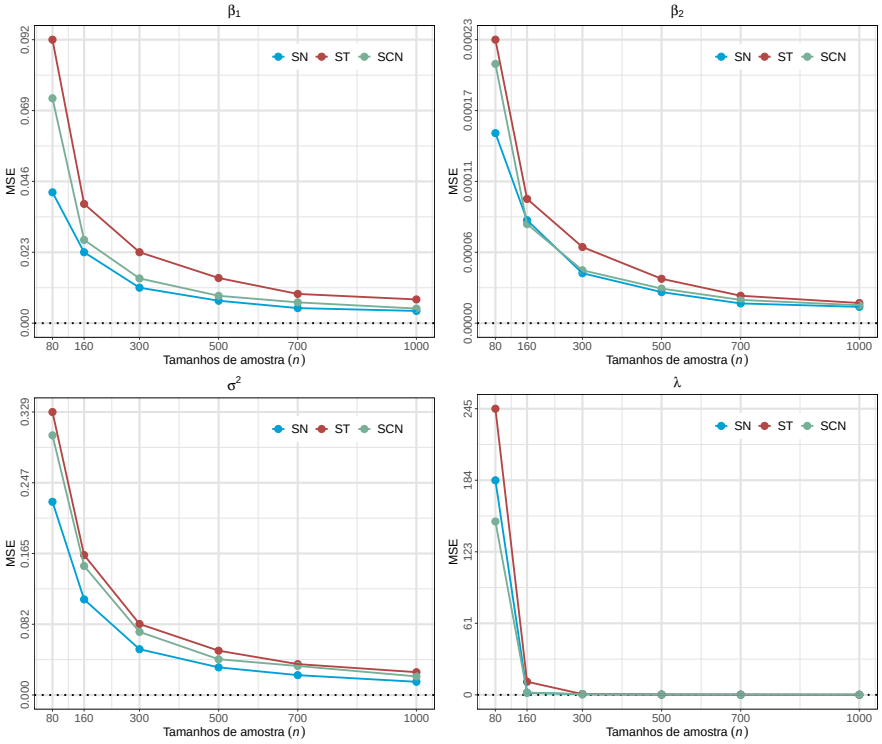

Figura 16 – Estudo de simulação 1. Erro quadrático médio das estimativas de parâmetros dos modelos SMSN-ICR, com níveis de censura de 20% e dados faltantes de 5%, respectivamente.

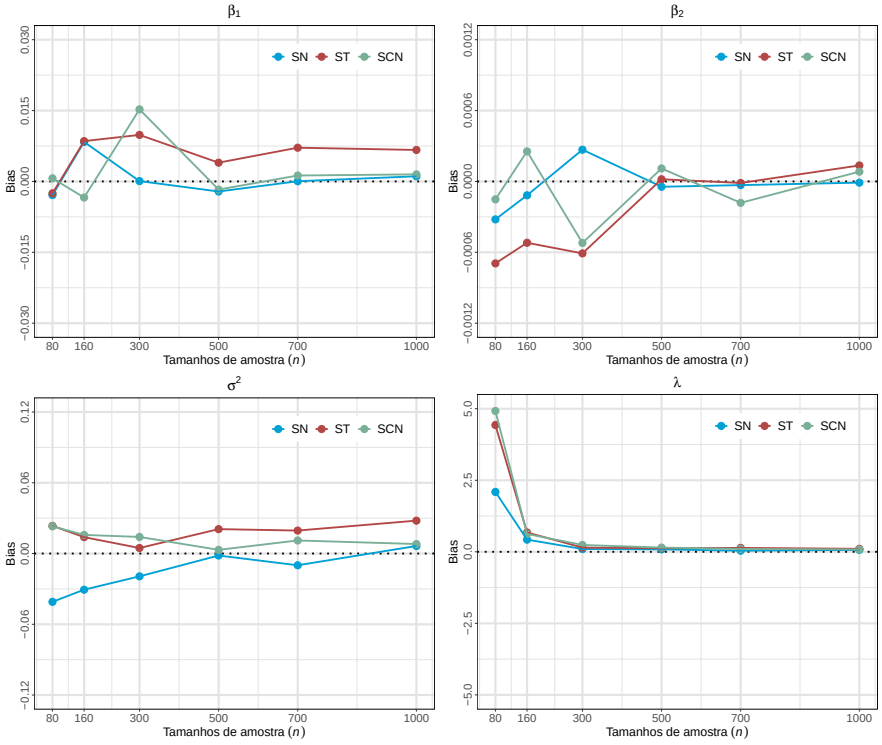

Figura 17 – Estudo de simulação 1. Viés das estimativas de parâmetros dos modelos SMSN-ICR, com níveis de censura e dados faltantes de 35% e 5%, respectivamente

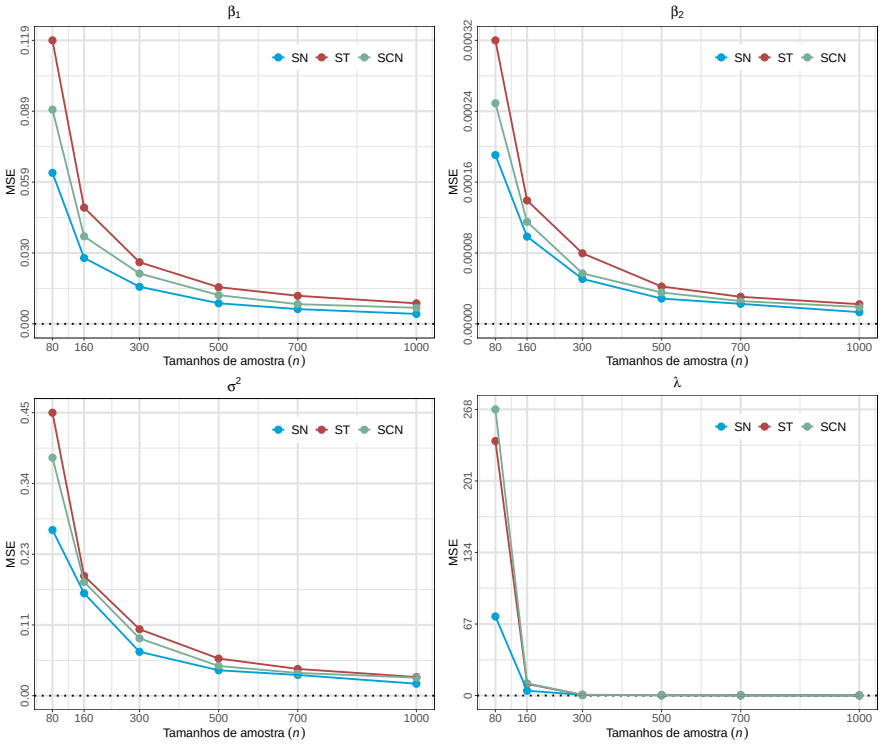

Figura 18 – Estudo de simulação 1. Erro quadrático médio das estimativas de parâmetros dos modelos SMSN-ICR, com níveis de censura de 35% e dados faltantes de 5%, respectivamente.

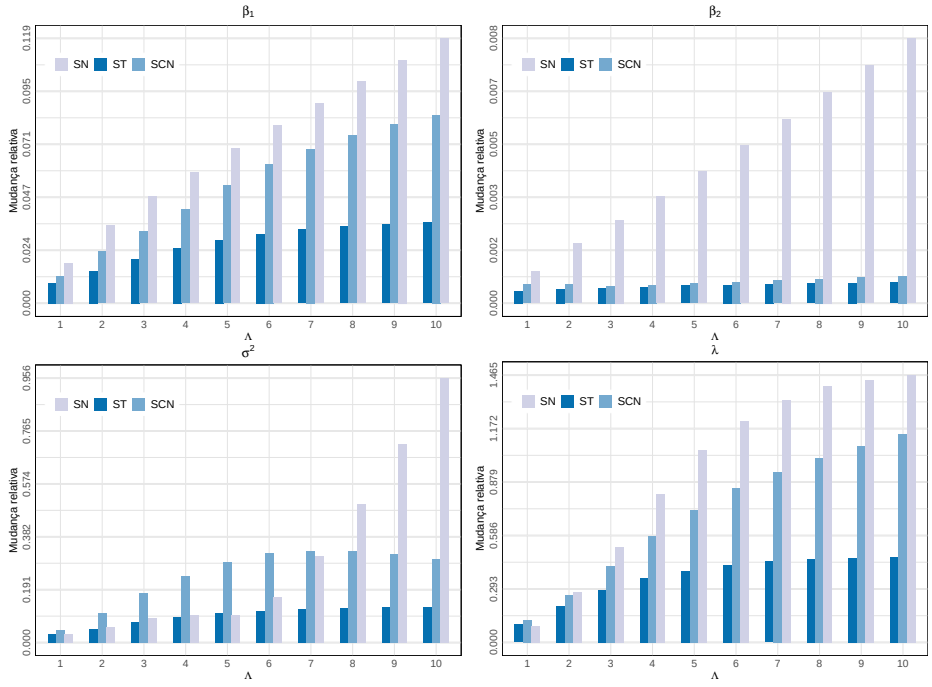

Figura 19 – Estudo de Simulação 2. Mudanças relativas médias das estimativas para diferentes perturbações Λ e nível de censura de 8% e faltantes de 5%.

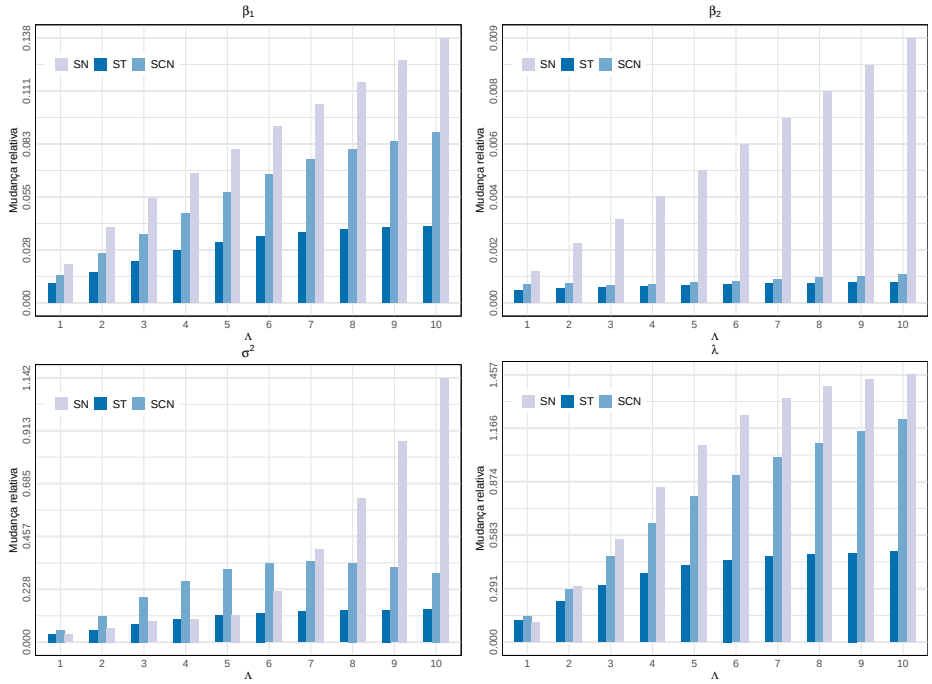

Figura 20 – Estudo de Simulação 2. Mudanças relativas médias das estimativas para diferentes perturbações Λ e nível de censura de 20% e faltantes de 5%.

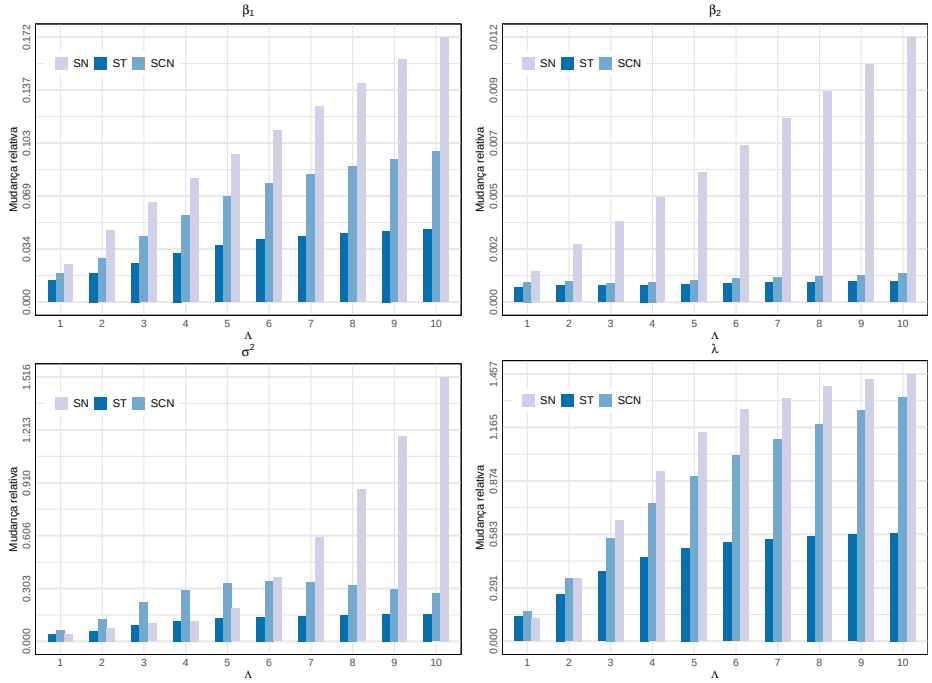

Figura 21 – Estudo de Simulação 2. Mudanças relativas médias das estimativas para diferentes perturbações Λ e nível de censura de 35% e faltantes de 5%.