Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Programa de Pós-Graduação em Matemática

Daniel Camilo Fuentes Guzmán

Modelos de Regressão para Dados Censurados sob a Classe de Distribuições de Misturas de Escala Normal Assimétricas

Juiz de Fora
2018

Daniel Camilo Fuentes Guzmán

Modelos de Regressão para Dados Censurados sob a Classe de Distribuições de Misturas de Escala Normal Assimétricas

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade Federal de Juiz de Fora, na área de concentração em Matemática Aplicada, como requisito parcial para obtenção do título de Mestre em Matemática.

Orientador: Clécio da Silva Ferreira

Coorientadora: Camila Borelli Zeller

Juiz de Fora

2018

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Guzmán, Daniel Camilo Fuentes.

Modelos de Regressão para Dados Censurados sob a Classe de Distribuições de Misturas de Escala Normal Assimétricas / Daniel Camilo Fuentes Guzmán. – 2018.

82 f. : il.

Orientador: Clécio da Silva Ferreira

Coorientadora: Camila Borelli Zeller

Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Matemática, 2018.

1. Modelos de Regressão Para Dados Censurados. 2. Família de Distribuições Assimétricas com Caudas Pesadas. 3. Algoritmo MCEM. I. Ferreira, Clécio da Silva, orient. II. Zeller, Camila Borelli, cooriet. III. Título.

Daniel Camilo Fuentes Guzmán

Modelos de Regressão para Dados Censurados sob a Classe de Distribuições de Misturas de Escala Normal Assimétricas

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade Federal de Juiz de Fora, na área de concentração em Matemática Aplicada, como requisito parcial para obtenção do título de Mestre em Matemática.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Clécio da Silva Ferreira - Orientador
ICE-UFJF

Profa. Dra. Camila Borelli Zeller - Coorientador
ICE-UFJF

Profa. Dra. Larissa Ávila Matos
IMECC-UNICAMP

Prof. Dr. Tiago Maia Magalhães
ICE-UFJF

Dedico este trabalho à minha mãe que, com paciência e esforço, conseguiu me guiar e me dar uma boa vida sem medo de dificuldades com fé e esperança de um futuro melhor; à minhas irmãs Maria Paula e Luisa Paola que são as minhas maiores motivações, a meu sobrinho Juandiego com carinho; à minha Elizabeth com amor e gratidão.

AGRADECIMENTOS

Primeiramente agradeço a Deus por me proporcionar vida, saúde, coragem e bênção a cada passo, a cada decisão que tomo nos rumos da vida, sempre me guiando, sempre me acompanhando.

À minha mãe com muito carinho meu mais profundo agradecimento pelo amor infinito, por ser exemplo de vida e meu maior orgulho, por o apoio incondicional e sempre acreditar em mim mesmo nos dias mais cinzas. Por renunciar a seus próprios sonhos a meu favor, por me dar sentido, direção e fazer de mim uma pessoa em constante evolução, por tudo e muito mais, obrigado.

Sou grato à minha irmã Maria Paula, à minha irmã Luisa Paola e ao meu sobrinho Juandiego por que são meus tesouros mais preciosos, por me amarem e me apoiarem sempre. São os principais motores da minha vida e fazem que minha existência neste justo momento seja única e maravilhosa.

A todos os meus amigos, obrigado pela companhia, as conversas, sorrisos e incentivos em meus muitos momentos de crises existenciais. Vocês são realmente valiosos. Não posso listar todos, mas em qualquer lugar do planeta onde estiverem estarão sempre em meu coração, o pessoal da UT, da UFSC, do POSMEC e da UFJF, os amigos dos roles, do bairro e da república onde morei. Em especial gostaria de agradecer a Debórah Lopez Tavares, Pablo Domingos, Sérgio Corrêa, Patrick Lucas Zagnoli, Leidlaine Medeiros e Felipe Kelly pela gentileza e o sincero carinho que foram fundamentais em muitos momentos de minha experiência em Juiz de Fora.

Gostaria de expressar minha profunda gratidão ao meu orientador Dr Clécio da Silva Ferreira pela infinita paciência, confiança, a constante guia e apoio, por ser uma ótima pessoa em todo sentido, que admiro muito por sua rapidez mental tanto na pesquisa como no futebol e por me permitir ser parte de este trabalho, por me dar sugestões importantes e fazer as devidas correções.

Também gostaria de expressar meus sinceros agradecimentos a minha co-orientadora Camila Borelli Zeller porque é meu modelo a seguir como pessoa e profissional, sempre gentil, sempre disposta, porque demonstra uma tenacidade na

alma que inspira e motiva a sempre a não desistir na perseguição dos objetivos além de qualquer obstáculo, por sua amizade e generosidade, pela paciência e valiosos conselhos, as sugestões cruciais, pelas ajudas e correções. Sem ela, seria impossível terminar esta dissertação.

À minha Elizabeth Ciampi por ser meu maior amor e cúmplice. Seu sorriso da luz aos meus dias e sua companhia conforta minha alma, a Maria Cristina, Igor e Roque por sua grande gentileza, suporte, apoio incondicional, por me integrarem em seu lar sem esperar nada em troca, por fazer da minha vida nestes últimos meses uma aventura inesperada e fantástica. Obrigado pelo imenso carinho.

A todos os professores da banca pelas sugestões e contribuições importantes para poder melhorar e concluir este trabalho, Profa. Dra. Larissa Ávila Matos cujo trabalho de tese e artigos foram constantemente consultados nas distintas etapas do processo da dissertação e ao Prof. Dr. Tiago Maia Magalhães pela valiosa ajuda.

Aos professores da pós-graduação Alexei D., Lucy T., Mateus B., Wilhem P., Mario Jorge D., Lonardo R., Camila Z. e Clécio F. por serem a ponte e guia neste ousado caminho do mundo da Matemática e Estatística, e me mostrarem a disciplina e coragem necessárias para o sucesso nesta profissão. Foi realmente enriquecedor passar por todas as experiências até o ponto final do mestrado.

Aos professores membros do colegiado Dr. William F., Dr. Laércio e Dr. Olímpio M. e especialmente ao professor Dr. Grigori Chapiro junto com a secretaria da pós-graduação Paula pela boa disposição e eficiência que demonstram constantemente ante qualquer situação, tiveram palavras pertinentes nos momentos justos, tomaram decisões nas que sempre tive a fortuna de sair bem livrado fazendo possível minha conclusão do mestrado. Sou muito grato.

Aos meus colegas do mestrado em matemática da UFJF pela amizade, pelos momentos compartilhados e os sorrisos provocados. Vocês são minha moral nas horas em que minha Fé enfraquece. Muito sucesso e tudo de bom para vocês.

Além disso, para mim é muito importante e fundamental agradecer ao Brasil e todas as pessoas que conheci nesta linda nação nos últimos quatro anos, onde tive a fortuna do acolhimento e brinde de incontáveis experiências e oportunidades. Vocês me permitiram estudar e viver dignamente durante esse período; também por

me permitir ser bolsista CAPES e me brindar a possibilidade de sair da pobreza econômica e a imensa injustiça social, por me permitir fugir da desigualdade em que vivia em meu pais de origem. Não terei nunca como agradecer por tanto. Deus os acompanhe sempre.

"O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001"

“De mi debilidad, obtuve una fuerza que nunca me abandonó.”
(Jorge Luis Borgues)

RESUMO

Um problema frequente na análise de regressão é quando a observação da variável resposta é censurada para alguns indivíduos. Isto ocorre em várias situações práticas, por razões como limitações do equipamento de medição ou do desenho experimental. Estes fenômenos podem ser modelados mediante modelos estatísticos e matemáticos. No âmbito dos modelos de regressão censurados, os erros aleatórios são rotineiramente considerados como tendo uma distribuição normal, principalmente por conveniência matemática. No entanto, este método tem sido criticado na literatura por causa de sua sensibilidade a desvios da suposição de normalidade. Nessa dissertação, primeiro estabelecemos uma nova ponte entre o modelo de regressão censurado e a classe de distribuições assimétricas estudadas por Ferreira et al. [13]. As misturas de escala assimétricas das distribuições normais são frequentemente utilizadas para procedimentos estatísticos que envolvem dados assimétricos e caudas pesadas. A principal virtude dos membros dessa família de distribuições é que eles são fáceis de serem simulados e também fornecem algoritmos tipo Esperança-Maximização (EM) para a estimativa de máxima verosimilhança. Neste trabalho, estendemos o algoritmo EM para o algoritmo MCEM para modelos de regressão lineares censurados. O algoritmo do tipo EM foi discutido com ênfase nas distribuições Normal Assimétrica, t-Student Assimétrica, Slash Assimétrica e Normal-Contaminada Assimétrica. Os métodos propostos são verificados através da análise de vários estudos de simulação e aplicação em conjuntos de dados reais.

Palavras-chave: Modelo de regressão para dados censurado; Caudas pesadas; Misturas de Escala Normal Assimétricas; Algoritmo MCEM.

ABSTRACT

A frequent problem in regression analysis is when the observation of the response variable is censored for some subjects. This occurs in several practical situations, for reasons such as limitations of the measuring equipment or the experimental design. These phenomena can be modeled using statistical and mathematical models. In the framework of censored regression models the random errors are routinely assumed to have a normal distribution, mainly for mathematical convenience. However, this method has been criticized in the literature because of its sensitivity to deviations from the normality assumption. In this dissertation, we first establish a new link between the censored regression model and the class of asymmetric distributions studied by Ferreira et al. [13]. Skew scale mixtures of normal distributions are often used for statistical procedures involving asymmetric data and heavy-tailed. The main virtue of the members of this family of distributions is that they are easy to simulate and also provide expectation-maximization (EM) algorithms for maximum likelihood estimation. In this work, we extend the EM algorithm for the MCEM algorithm for linear regression models censored. The EM-type algorithm has been discussed with an emphasis on the Skew-normal, Skew Student-t-normal, Skew slash and Skew-contaminated normal distributions. The proposed methods are verified through the analysis of several simulation studies and applying in real datasets.

Key-words: Censored regression model; Heavy tails; Skew scale mixtures of normal distributions; EM-type algorithm.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SN	Normal Assimétrica (Skew-normal)
TN	Normal Truncada (Truncated Normal)
SMN	Mistura de Escala Normal
STN	Distribuição Skew-t-normal
SSL	Distribuição Skew-Slash
SCN	Distribuição Skew-Normal-Contaminada
SSMN	Misturas de Escala Normal Assimétricas
fdp	Função Densidade de Probabilidade
fda	Função Distribuição Acumulada
MV	Máxima Verosimilhança
EM	Esperança Maximização
MCEM	Monte Carlo EM
CR	Regressão Com Dados Censurados (Censored-Regression)
SE	Erro Padrão
i.i.d.	Distribuições independentes e identicamente distribuídas

LISTA DE SÍMBOLOS

$\forall$	Para todo
$\in$	Pertence
$\sim$	Distribuída como
$\boldsymbol{\Theta}$	Espaço Paramétrico
$\phi(.)$	Função densidade de probabilidade da normal padrão univariada
$\Phi(.)$	Função de distribuição acumulada da normal padrão univariada
$TD_{(a,b)}(\theta)$	Distribuição da variável aleatória $D(\theta)$ truncada no intervalo (a,b)
Γ	Função gama
$\boldsymbol{\beta}$	Parâmetro de locação
σ^2	Parâmetro de escala
λ	Parâmetro de assimetria
ν	Parâmetro de forma, indica os graus de liberdade de uma distribuição STN e SSL e na Distribuição Normal Contaminada representa a porcentagem de *Outliers*
γ	Parâmetro de Escala ou Forma na Normal Contaminada e na SCN
$\mathbf{I}_{(A)}$	Função indicadora do conjunto A
$P_x(a,b)$	fda de uma distribuição $Gama(a,b)$ avaliada em x
$\perp$	Independência de variáveis aleatórias

SUMÁRIO

1 INTRODUÇÃO

Um problema frequente na análise de regressão é quando a observação da variável resposta é censurada para alguns indivíduos. Isto ocorre, em várias situações práticas, por razões como limitações do equipamento de medição ou do desenho experimental. Portanto, o valor verdadeiro exato é registrado apenas se estiver dentro do alcance do intervalo, portanto, as respostas podem ser censuradas à esquerda ou à direita. Variáveis censuradas são comuns nos estudos econométricos, biomédicos, epidemiológicos, de sobrevida e duração. Por exemplo, na econometria, o estudo da participação da força de trabalho de mulheres casadas é geralmente conduzido sob o modelo de regressão normal censurado. Nesse caso, a resposta observada é a taxa salarial, que é tipicamente considerada como censurada abaixo de zero, ou seja, para mulheres trabalhadoras, valores positivos para os salários são registrados, enquanto para mulheres não trabalhadoras os salários observados são zero; ver Mroz [33]. Greene [19] lista vários outros exemplos de variáveis censuradas, por exemplo, o número de detenções após a libertação da prisão (Witte [41]) e as despesas de férias (Melenberg e Soest [32]).

No contexto de modelos de regressão censurados, os erros aleatórios são rotineiramente considerados como tendo uma distribuição normal. No entanto, é bem conhecido que vários fenómenos nem sempre se encaixam sob as hipóteses do modelo normal, produzindo dados com uma distribuição tendo simultaneamente assimetria e caudas pesadas. Daí, a partir de uma perspetiva prática, há uma necessidade de buscar um modelo teórico apropriado que evite transformações de dados, ainda que preserve uma estrutura robusta e conveniente semelhante a um modelo gaussiano.

Muitas extensões do modelo censurado gaussiano clássico foram propostas para ampliar a aplicabilidade da análise de regressão a situações em que a suposição de erro gaussiano pode ser inadequada. Por exemplo, Arellano et al. [2] propuseram o modelo de regressão censurado t-Student, ver também Massuia et al. [26]. Garay et al. [16] propõem modelos de regressão censurados lineares com misturas de escala de distribuições normais. Recentemente, Garay et al. [17] desenvolveram modelos de regressão censurados não-lineares baseados em misturas de escala de distribuições

normais. Eles demonstraram sua robustez contra *outliers* através de extensas simulações. Esses modelos são, sem dúvida, muito flexíveis, mas os problemas relacionados à ocorrência simultânea de assimetria e *outliers* permanecem.

Neste trabalho, propõe-se uma extensão do modelo censurado normal usual, considerando-se Misturas de Escala Normal Assimétricas (SSMN), definidas por Ferreira et al. [13]. A justificativa teórica da proposta baseia-se nos fatos de que a classe SSMN atribui estocasticamente pesos variados a cada sujeito, ou seja, menor peso para *outliers* e, portanto, controla a influência de observações atípicas na inferência geral. Além disso, essa família de distribuições é atraente porque leva em conta simultaneamente a assimetria e as caudas pesadas, além de ter uma representação estocástica que permite a fácil implementação do algoritmo EM.

Cabe mencionar que há outra família de distribuições que leva em conta assimetria e caudas pesadas simultaneamente as Misturas de escala de normais assimétricas (SMSN) propostas por Branco e Dey [5]. Existem algumas diferenças importantes entre as classes de distribuições SSMN e SMSN. Primeiro, os mecanismos para gerar amostras aleatórias são ligeiramente diferentes, o que produz diferentes estruturas de distribuição. Finalmente, essas classes apresentam diferentes coeficientes de assimetria e curtose (Ferreira et al. [14]). No contexto dos modelos de regressão para dados censurados sob as distribuições SMSN podemos citar Mattos [29].

Assim, nesse trabalho são abordados:

(i) Um método de estimação para o modelo de regressão linear baseado nas distribuições **SSMN** para dados censurados;

(ii) Avaliação dos métodos propostos computacionalmente;

(iii) Aplicação desses resultados à análise de um conjunto de dados reais.

A seguir serem apresentados algumas definições e métodos importantes para o desenvolvimento dos capítulos posteriores.

1.1 ESTIMAÇÃO POR MÁXIMA VEROSIMILHANÇA

Os principais resultados do método de estimação por máxima verossimilhança foram apresentados na obra *The Mathematical Foundations of Theoretical Statistics* do estatístico inglês **Ronald Aylmer Fisher** $(1890 - 1962)$ em 1922 (Stigler [38]).

De maneira resumida, vamos descrever o método que será a técnica básica considerada aqui para a estimação paramétrica.

Definição 1.1.1. *Seja* $\mathbf{Y} = (Y_1, \ldots, Y_n)$ *um vetor cujas coordenadas constituem uma amostra aleatória com função densidade de probabilidade (fdp)* f *dependente do vetor de parâmetros* $\boldsymbol{\theta} \in \Theta$*. Supondo que a amostra* $Y_1 = y_1, \ldots, Y_n = y_n$ *foi coletada, definimos a função logaritmo da verosimilhança* ℓ *por*

$$\ell(\boldsymbol{\theta}) = \ln f_{\mathbf{Y}}(y_1, \ldots, y_n; \boldsymbol{\theta}) = \ln \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\theta}). \tag{1.1}$$

Os estimadores de máxima verossimilhança apresentam uma série de propriedades assintóticas como consistência e eficiência que são responsáveis por justificar seu uso frente a outros estimadores e podem ser encontradas em Ritter [36] e Pawitan [35].

Teoricamente, a consistência e a eficiência assintótica podem ser verificadas computacionalmente por meio de medidas de viés e desvios para os conjuntos de dados simulados com diferentes tamanhos e de precisão para dados simulados com diferentes tamanhos amostrais, por exemplo ver Louredo [25].

1.2 ALGORITMOS PARA ESTIMAÇÃO DE MÁXIMA VEROSIMILHANÇA

1.2.1 Algoritmo EM

O Algoritmo EM foi desenvolvido na década de 1970 como uma alternativa aos métodos tradicionais de otimização normalmente utilizados na estimação de máxima verossimilhança (MV), particularmente aqueles dos tipos Newton ou Quasi-Newton (Ver Dempster et al. [7] e Louredo [25] para mais detalhes). A sigla EM significa Expectation-Maximization (Esperança-Maximização).

Em modelos com dados não observados ou incompletos, o algoritmo EM é uma estratégia de otimização iterativa muito comumente usada (Mattos [29]). Esse algoritmo possui muitos recursos atraentes, como estabilidade numérica e simplicidade de implementação, e seus requisitos de memória são bastante razoáveis (Couvreur [6]).

Sejam $\mathbf{y}$ um vetor de dados observados e $\mathbf{u}$ um vetor de dados faltantes ou latentes, o vetor de dados completos é denotado por $\mathbf{y}_c = (\mathbf{y}, \mathbf{u})$. Em Mclachlan & Krishnan [30] é estabelecida a relação probabilística em termos de uma integral entre as fdps $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ e $f_{\mathbf{Y}_C}(\mathbf{y}_C; \boldsymbol{\theta})$ dos dados observados e dos dados completos respetivamente (Louredo [25]). Segundo o mesmo autor, tal relação pode ser reescrita em termos da fdp. condicional $h_{\mathbf{Y}_C|\mathbf{Y}}(\mathbf{y}_C|\mathbf{y}; \boldsymbol{\theta})$ da forma $f_{\mathbf{Y}_C}(\mathbf{y}_C; \boldsymbol{\theta}) = f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) h_{\mathbf{Y}_C|\mathbf{Y}}(\mathbf{y}_C|\mathbf{y}; \boldsymbol{\theta})$.

Há uma relação hierárquica entre as distribuições dos vetores aleatórios $\mathbf{Y}$ e $\mathbf{U}$ cuja fdp é $g_{\mathbf{u}}(\mathbf{u}; \boldsymbol{\theta})$ (Louredo [25]), a qual para $f_{\mathbf{Y}_C}(\mathbf{y}_C; \boldsymbol{\theta})$ resulta na seguinte expressão

$$f_{\mathbf{Y}_C}(\mathbf{y}_C; \boldsymbol{\theta}) = h_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}; \boldsymbol{\theta}) g_{\mathbf{U}}(\mathbf{u}; \boldsymbol{\theta}).$$

Sob a hipótese de independência entre $\mathbf{U}$ e $\mathbf{Y}|\mathbf{U} = \mathbf{u}$ cujas fdps serão denotadas por g e h, respetivamente, temos que as coordenadas de $f_{\mathbf{Y}_C}$ são independentes e iguais a, digamos, f_C (Louredo [25]). Segue daí que podemos escrever a função log-verosimilhança dos dados completos ℓ_C na forma

$$\ell_C(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f_C(\mathbf{y}_{Ci}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln g(\mathbf{u}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \ln h(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\theta}).$$

Então, o algoritmo EM prossegue em duas etapas:

- ◇ **Passo E**: Defina $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = E[(\ell_C(\boldsymbol{\theta}|\mathbf{y_C})|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)})]$.
- ◇ **Passo M**: Maximize $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ com respeito a $\boldsymbol{\theta}$, obtendo $\widehat{\boldsymbol{\theta}}^{(k+1)}$.

Cada iteração do algoritmo EM aumenta a função de verosimilhança $\ell(\boldsymbol{\theta}|\mathbf{y})$ e a sequência $\boldsymbol{\theta}^{(k)}$ converge para um ponto estacionário da verosimilhança observada sob condições de regularidade leve (para mais detalhes ver Wu [42] e Vaida [39]).

Para os casos em que a etapa E não possui forma analítica, Wei e Tanner [40] propuseram o algoritmo Monte Carlo EM (MCEM), em que o passo E é substituído

por uma aproximação de Monte Carlo com base em simulações independentes dos dados faltantes.

1.2.2 Algoritmo MCEM

O passo E do algoritmo EM requer a log-verossimilhança esperada dos dados completos condicionado nos dados observados, quando esta esperança é difícil de ser calculada analiticamente, ela pode ser aproximada via Monte Carlo (Santos [37]).

O algoritmo MCEM consiste portanto em simular a cada iteração sucessivamente os dados faltantes com a distribuição condicional e atualizar os parâmetros desconhecidos do modelo. Assim, Wei and Tanner [40] propuseram que o passo E do algoritmo EM pode ser substituído pelos seguintes passos:

- ◇ **Passo MC**: Sejam $\mathbf{z}_1^{(k+1)}, \ldots, \mathbf{z}_J^{(k+1)}$ amostras i.i.d. de $f_{Z|Y}(z|y, \boldsymbol{\theta}^{(k)})$, em que cada $\mathbf{z}_j^{(k+1)}$ com $j = 1, \ldots, J$ é um vetor de dados latentes. Considere $Y_j = (y, \mathbf{z}_j^{(k+1)})$ o conjunto de dados completos em que os dados faltantes foram substituídos por $\mathbf{z}_j^{(k+1)}$.

- ◇ **Passo E**: Calcule a aproximação $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ dada por

$$\widehat{Q}_J(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \frac{1}{J}\sum_{j=1}^{J} \ell(\boldsymbol{\theta}|Y_j^{(k)}).$$

- ◇ **Passo M**: Obter $\boldsymbol{\theta}^{(k)}$ ao maximizar $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ com respeito a $\boldsymbol{\theta}$ e fazer $k = k + 1$.

  Iterar os passos anteriores até obter convergência para a sequência das estimativas de $\boldsymbol{\theta}^{(k)}$, segundo algum critério de parada.

  Para todos os estudos de simulação feitos neste trabalho foi utilizado o seguinte critério de parada: A norma da diferença entre a estimativa atual menos a estimativa no passo anterior e essa diferença tem que ser menor que um épsilon como é ilustrado a seguir

$$|| \, \widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)} \, || < \epsilon,$$

isto é, a diferença relativa entre duas avaliações sucessivas da log-verossimilhança $\ell(\boldsymbol{\theta}|\mathbf{v})$, dada por

$$|| \ \ell(\boldsymbol{\theta}^{(k+1)}|\mathbf{v}, \rho) - \ell(\boldsymbol{\theta}^{(k)}|\mathbf{v}, \rho) \ ||< \epsilon,$$

em que $(\mathbf{v}, \rho)$ são os dados observados. Em nosso estudo foi escolhido um $\epsilon = 10^{-4}$.

1.3 FAMÍLIA DE DISTRIBUIÇÕES SSMN

A notação que utilizaremos para as distribuições será a correspondente a suas siglas referentes ao idioma inglês, por exemplo SN significa "*Skew-Normal*" correspondendo à normal assimétrica e SMN significa "*Scale Mixtures of Normal*".

As distribuições de probabilidade presentes no modelo proposto a serem tratadas nesta dissertação serão as integrantes da família de Distribuições de Misturas de Escala Normal Assimétricas denotadas pelas siglas SSMN (Skew Scale Mixtures of Normal) conforme definido por Ferreira et al. [13].

1.3.1 Normal Assimétrica (SN)

Sejam $\phi(x; \mu, \sigma^2)$ e $\Phi(x; \mu, \sigma^2)$ a função de densidade de probabilidade (fdp) e a função de distribuição acumulativa (fda), respetivamente, da distribuição $N(\mu, \sigma^2)$ avaliada em x. A variável aleatória Y segue uma distribuição Normal Assimétrica (SN) univariada com parâmetro de locação μ, parâmetro de escala σ^2 e parâmetro de assimetria λ se sua pdf é dada por

$$f(y) = 2\phi(y; \mu, \sigma^2)\Phi\left(\frac{\lambda(y-\mu)}{\sigma}\right), \ y \in \mathbb{R}. \tag{1.2}$$

Para uma variável aleatória com fdp igual a (1.2) usaremos a notação $Y \sim SN(\mu, \sigma^2, \lambda)$.

Quando $\lambda = 0$, a distribuição SN passa a ser uma distribuição normal usual ($Y \sim N(\mu, \sigma^2)$). Sua representação estocástica, que pode ser usada para derivar várias de suas propriedades, é dada por

$$Y \stackrel{d}{=} \mu + \sigma[\delta \mid T_0 \mid + (1-\delta^2)^{1/2} T_1], \tag{1.3}$$

com $\delta = \dfrac{\lambda}{(1+\lambda^2)^{1/2}}$, onde $\mid T_0 \mid$ denota o valor absoluto de T_0, $T_0 \sim N(0,1)$ e $T_1 \sim N(0,1)$ são independentes, e o símbolo $\stackrel{d}{=}$ significa "distribuída como". Da equação (1.3), temos que a esperança e a variância de Y são dadas por

$$E[Y] = \mu + b\sigma\delta, \text{ e } Var[Y] = \sigma^2(1 - b^2\delta^2), \tag{1.4}$$

onde $b = (2/\pi)^{1/2}$.

1.3.2 Misturas de Escala de Normal (SMN)

Em um contexto simétrico, Lange e Sinsheimer [20] forneceram um grupo de distribuições de cauda pesada que tem a distribuição normal como caso particular. Uma variável aleatória Y segue uma distribuição SMN com o parâmetro de localização $\mu \in \mathbb{R}$ e um parâmetro de escala positivo σ^2 se sua fdp assumir a forma

$$f_0 = (y; \mu, \sigma^2, \boldsymbol{\tau}) = \int_0^\infty \phi(y; \mu, k(u)\sigma^2) dH(u, \boldsymbol{\tau}), \tag{1.5}$$

onde $H(u; \boldsymbol{\tau})$ é a fda de uma variável aleatória positiva U indexada pelo vetor de parâmetro $\boldsymbol{\tau}$ e $k(.)$ é uma função estritamente positiva. Para uma variável aleatória com um fdp como na equação (1.5), usamos a notação $Y \sim SMN(\mu, \sigma^2, H; k)$. Além disso, quando $\mu = 0$ e $\sigma^2 = 1$, denotamos $Y \sim SMN(H; k)$.

A representação estocástica da variável aleatória $Y \sim SMN(\mu, \sigma^2, H; k)$ é dada por

$$Y = \mu + k^{1/2}(U)Z, \tag{1.6}$$

onde $Z \sim N(0, \sigma^2)$ e U é uma variável aleatória positiva com fda $H(u; \boldsymbol{\tau})$ e fdp $h(u; \boldsymbol{\tau})$, e independente de Z.

1.3.2.1 Casos Particulares

- Distribuição *t-Student* com $\nu > 0$ graus de liberdade, $Y \sim T(\mu, \sigma^2; \nu)$ tem função de densidade dada por

$$f(y) = \frac{1}{\sigma\sqrt{\nu\pi}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{d}{\nu}\right)^{-(\frac{\nu+1}{2})}, \tag{1.7}$$

onde $d = (y-\mu)^2/\sigma^2$. Neste caso, $k(u) = 1/U$ e $U \sim Gama(\nu/2, \nu/2)$, com fdp

$$h(u;\nu) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} u^{\nu/2-1} e^{-\nu u/2}, \tag{1.8}$$

tal que $E[U^{-m}] = \dfrac{(\nu/2)^m \Gamma(\nu/2-m)}{\Gamma(\nu/2)}$, para $m < \nu/2$. Se $\nu \uparrow \infty$ tem-se que $Y \xrightarrow{D} SN(\mu, \sigma^2, \lambda)$ para maiores detalhes ver Ferreira [10]. Em Lange e Sinsheimer [20] foi mostrado que a distância de Mahalanobis

$$D = \frac{(Y-\mu)^2}{\sigma^2} \sim F_{1,\nu}.$$

- Distribuição *Slash*, denotada como $Y \sim SL(\mu, \sigma^2; \nu)$, com parâmetro de forma $\nu > 0$, apresenta caudas mais pesadas do que a distribuição normal. Além disso, quando $\nu \uparrow \infty$, temos precisamente o caso da normal (Ferreira [10]). Tem fdp dada por

$$f(y) = \frac{\nu}{\sqrt{2\pi}\sigma} \int_0^1 u^{\nu-1/2} e^{-ud/2} du, \tag{1.9}$$

onde $k(U) = 1/U$ e U com densidade

$$h(u;\nu) = \nu u^{\nu-1} \mathbf{I}_{(0,1)}(u), \tag{1.10}$$

tal que $E[U^{-m}] = \dfrac{\nu}{\nu - m}$, para $m < \nu$, onde a notação $\mathbf{I}_{(A)}$ é a função indicadora do conjunto A. A distância de Mahalanobis D tem função distribuição dada por

$$Pr(D \leq r) = Pr(\chi^2 \leq r) - \frac{2^\nu \Gamma(\nu + 1/2)}{r^\nu \sqrt{\pi}} Pr(\chi^2_{2\nu+1} \leq r) \tag{1.11}$$

- Distribuição *Normal Contaminada*, denotada por $Y \sim NC(\mu, \sigma^2; \nu, \gamma)$, com os parâmetros ν representando a porcentagem de *outliers*, enquanto γ é interpretado como fator de escala, $0 \leq \nu \leq 1$, $0 < \gamma \leq 1$ (Little [23]; Ferreira [10]). É uma distribuição utilizada para modelar dados simétricos com presença de pontos aberrantes e possui pdf dada por

$$f(y) = \nu \phi(y|\mu, \sigma^2/\gamma) + (1-\nu)\phi(y|\mu, \sigma^2), \tag{1.12}$$

com $k(U) = 1/U$ e a fdp $h(u;\nu,\gamma)$ dada por

$$h(u;\nu,\gamma) = \nu\mathbf{I}_{(u=\gamma)} + (1-\nu)\mathbf{I}_{(u=1)},\ \boldsymbol{\tau} = (\nu,\gamma)^T. \tag{1.13}$$

Esta distribuição inclui o caso normal quando $\gamma = 1$ e $\nu = 1$. Assim, $E[U^{-m}] = \nu/\gamma^m + 1 - \nu$ e

$$Pr(D \leq r) = \nu Pr(\chi^2 \leq \gamma r) + (1-\nu)Pr(\chi^2 \leq r). \tag{1.14}$$

1.3.3 Misturas de Escalas de Normal Assimétricas (SSMN)

As misturas de escala da distribuição normal fornecem um grupo de distribuições de caudas pesadas usadas para procedimentos e inferência robusta de dados simétricos (Andrews e Mallows [1]; Dempster et al. [8]; Little [23]; Lange et al. [21]; Lange e Sinsheimer [20]; entre outros). A versão assimétrica dessas distribuições foi definida em Ferreira et al. [12], como uma alternativa para o tratamento de dados com distribuições assimétricas e caudas pesadas com a virtude de serem fáceis de gerar por simulação e terem uma representação estocástica que permite uma fácil implementação do algoritmo EM para estimação MV (Estimação por Máxima Verosimilhança).

Definição 1.3.1. *Uma variável aleatória Y segue uma distribuição SSMN com parâmetro de locação $\mu \in \mathbb{R}$, fator de escala σ^2 e parâmetro de assimetria $\lambda \in \mathbb{R}$, se tem sua fdp dada por*

$$f(y) = 2f_0(y;\mu,\sigma^2,\tau)\Phi\left(\frac{\lambda(y-\mu)}{\sigma}\right), \tag{1.15}$$

onde $f_0(\mathbf{y})$ é definida na equação (1.5). Para uma variável aleatória com fdp como em (1.15) (Ver Ferreira et al. [13] para mais detalhes), usaremos a notação $Y \sim SSMN(\mu,\sigma^2,\lambda,H;k)$. Se $\mu = 0$ e $\sigma^2 = 1$, nos referimos a uma distribuição padrão SSMN e a denotamos por $SSMN(\lambda,H;k)$. Se $\lambda = 0$, teremos uma distribuição SMN.

É claro que se $Y \sim SSMN(\mu,\sigma^2,\lambda,H;k)$, então $Z = (Y-\mu)/\sigma^2 \sim SSMN(\lambda,H;k)$ (Ferreira et al. [13]). A seguinte proposição é usada para a implementação do algoritmo EM. A prova pode ser encontrada em Ferreira et al. [12], assim como as demais que serão apresentadas.

Proposição 1.3.1. *Seja* $Y \sim SSMN(\mu, \sigma^2, \lambda, H; k)$*. Então, sua representação hierárquica é dada por*

$$\begin{aligned} Y|U = u &\sim SN(\mu, \sigma^2 k(u), \lambda k(u)^{1/2}), \\ U &\sim H(\boldsymbol{\tau}). \end{aligned} \tag{1.16}$$

Portanto, para gerar uma distribuição SSMN, primeiramente geramos a distribuição U *e depois a distribuição condicional* $Y|U$*, usando a equação (1.3).*

Proposição 1.3.2. *Seja* $Y \sim SSMN(\mu, \sigma^2, \lambda, H; k)$*. Então,*

$$E[Y] = \mu + b\sigma\lambda E_U\left[\frac{k(U)}{(1+\lambda^2 k(U))^{1/2}}\right] \text{ e} \tag{1.17}$$

$$Var[Y] = \sigma^2\left(E_U[k(U)] - b^2\lambda^2 E_U^2\left[\frac{k(U)}{(1+\lambda^2 k(U))^{1/2}}\right]\right). \tag{1.18}$$

Se $k(U) = 1$, então $E[Y] = \mu + b\sigma\delta$ e $Var[Y] = \sigma^2(1 - b^2\delta^2)$ são os mesmos definidos para a distribuição skew-normal, conforme a equação (1.4).

Proposição 1.3.3. *Seja* $Y \sim SSMN(\mu, \sigma^2, \lambda, H; k)$*. Então, a forma quadrática*

$$D_\lambda = \frac{(Y-\mu)^2}{\sigma^2},$$

tem a mesma distribuição que $D = (X-\mu)^2/\sigma^2$*, onde* $X \sim SMN(\mu, \sigma^2, H; k)$*.*

Casos particulares dessa classe de distribuições assimétricas são, por exemplo, a distribuição SN, denotada por $SN(\mu, \sigma^2, \lambda)$ quando $U = 1$; a distribuição de Skew Student-t-normal, denotada por $STN(\mu, \sigma^2, \lambda, \nu)$ quando $U \sim Gamma(\nu/2, \nu/2), \nu > 0$; a distribuição Skew Slash, denotada por $SSL(\mu, \sigma^2, \lambda, \nu)$ quando $U \sim Beta(\nu, 1), \nu > 0$; e a distribuição Skew Normal Contaminada, denotada por $SCN(\mu, \sigma^2, \lambda, \nu, \gamma), 0 < \nu < 1, 0 < \gamma < 1$, quando U tem função de massa de probabilidade $h(u, \boldsymbol{\tau}) = \nu\mathcal{I}_{(u=\gamma)} + (1-\nu)\mathcal{I}_{(u=1)}, \tau = (\nu, \gamma)^T$ (Ferreira et al. [13]).

Referimo-nos a Ferreira et al. [12] para detalhes e propriedades adicionais relacionadas a essa classe de distribuições.

Neste trabalho, propõe-se uma extensão do modelo censurado normal usual, considerando-se as distribuições SSMN, estendendo e complementando os resultados

encontrados em Ferreira et al. [13] para o modelo de regressão assimétrico, denotado por SSMN-RM. O algoritmo EM para obter a estimativa de MV do vetor de parâmetros, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2, \lambda, \boldsymbol{\tau}^T)^T$ do SSMN-RM, pode ser encontrado mesmo em Ferreira et al. [13] generalizando o algoritmo dado em Ferreira et al. [12].

1.4 MOTIVAÇÃO

Nesta seção, o objetivo principal é estudar o efeito de levar em consideração os dados censurados sobre as estimativas dos parâmetros, no contexto de modelo de regressão. Dessa forma geramos 100 amostras provenientes do modelo de regressão Skew-t-normal com $\nu = 5$, definindo o nível de censura à direita em 20%, $\mathbf{x}_i^T = (1, x_i)$, tal que $x_i \sim U(0, 1)$, $i = 1, \ldots, 100$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T = (2, -1)$, $\sigma^2 = 0.3$ e $\lambda = -4$. Para cada conjunto de dados, ajustamos dois modelos, no caso 1, usamos o modelo de regressão Skew-t-normal NAIVE, onde as observações censuradas não são levadas em consideração, ou seja, tais observações são excluídas. No caso 2, ajustamos o modelo de regressão Skew-t-normal para dados censurados. Mais detalhes sobre este modelos serão dados no próximo capítulo.

A Figura 1 mostra os boxplots das estimativas dos parâmetros β_0 e β_1. Note que as estimativas dos coeficientes de regressão para o caso 2 são menos viesadas do que as estimativas obtidas no caso 1.

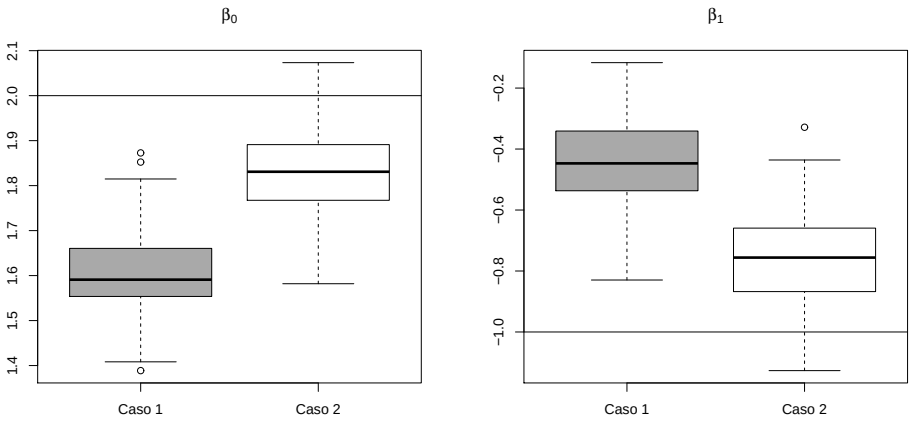

Figura 1 – **Motivação**: Boxplots das estimativas dos parâmetros β_0 e β_1. Modelo de regressão Skew-t-normal para dados censurados e Modelo de regressão Skew-t-normal NAIVE.

A Figura 2 mostra os diagramas de dispersão com as 20 observações censuradas iguais ao ponto de corte ($c = 1,5222$) para uma determinada amostra simulada.

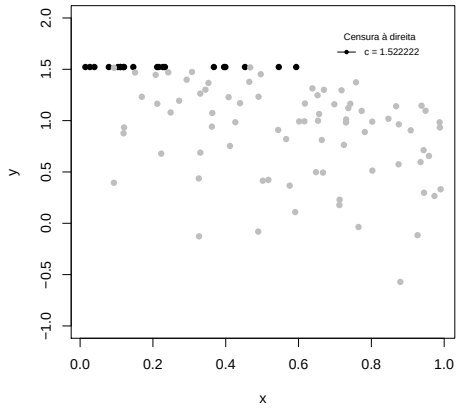

Figura 2 – **Motivação**: Diagrama de dispersão com as 20 observações censuradas iguais ao ponto de corte ($c = 1,5222$).

Por sua vez, a Figura 3 mostra o diagrama de dispersão com os dados reais (antes da inclusão de censura) e os valores imputados para as observações censuradas (denotados por asterisco). Neste contexto, ajustamos três modelos de regressão via mínimos quadrados: Modelo 1 considera os dados reais (sem censura), Modelo 2 considera os valores imputados para as observações que foram censuradas e Modelo 3 não leva em consideração as observações censuradas.

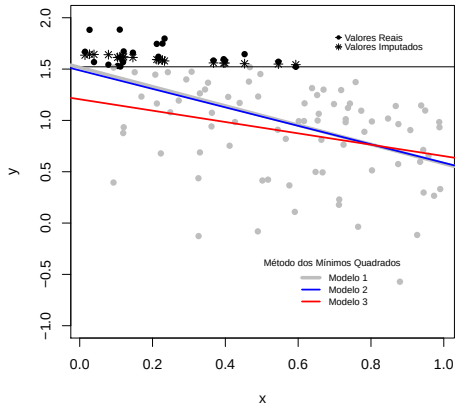

Figura 3 – **Motivação**: Diagrama de dispersão com os dados reais (antes da inclusão de censura) e os valores imputados para as observações censuradas (denotados por asterisco).

Note que o modelo ajustado 2 é o que mais se aproxima do modelo ajustado 1. Portanto, apontamos que é importante considerar o efeito da censura na modelagem dos dados evitando métodos *ad-hoc*.

2 O MODELO SSMN-CR

Nesse capitulo será definido o modelo de regressão linear com variável resposta censurada com erros distribuídos na família das Misturas de Escala Normal Assimétricas. Além disso, mostramos o processo de estimação dos parâmetros dos modelos SSMN-CR ilustrando os passos E e M calculando as estimativas para dados sem censura e com censura. Adicionalmente é explicado o processo de geração de distribuições truncadas TSSMN que será fundamental no passo E quando estamos trabalhando com dados censurados. Na seção final é calculada a matriz de informação empírica que é utilizada para a análise dos erros padrão (SE) na aplicação que será apresentada no capitulo 4.

Considere primeiro um modelo de regressão linear, conforme definido por Ferreira et al. [13], onde as respostas são observadas com erros que são independentes e identicamente distribuídos de acordo com alguma distribuição SSMN. Para ser mais preciso, vamos escrever,

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \xi_i, \quad \xi_i \overset{iid}{\sim} SSMN(0, \sigma^2, \lambda, H; \kappa), \ i = 1, \ldots, n, \tag{2.1}$$

onde os Y_i são respostas ou elementos da variável dependente, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ é um vetor de parâmetros de regressão e $\mathbf{x_i^T} = (x_{i1}, \ldots, x_{ip})$ é um vetor tal que x_{ij} é o valor da $j-$ésima variável explicativa do individuo i . Neste trabalho, estamos interessados na situação em que a variável resposta não é totalmente observada para todos os sujeitos i . Assim, para o $i-$ésimo sujeito e assumindo censura à direita, Y_i é uma variável latente e os dados observados (V_i, ρ_i) tomam a forma,

$$V_i = \begin{cases} c_i, & \text{se} \quad \rho_i = 1, \ \text{(i.e. } Y_i \geq c_i), \\ Y_i, & \text{se} \quad \rho_i = 0. \ \text{(i.e. } Y_i < c_i), \end{cases} \tag{2.2}$$

para algum ponto limite conhecido c_i, $i = 1, \ldots, n$. O indicador de censura $\rho_i = 1$ (ou $\rho_i = 0$) significa que a $i-$ésima observação é censurada (ou não censurada). As extensões de nossos resultados para a censura à esquerda são imediatas: basta transformar a resposta Y_i e o nível de censura c_i para $-Y_i$ e $-c_i$. Chamamos a estrutura definida por (2.1) e (2.2) de modelo SSMN-CR (Modelo de Regressão Linear Censurado SSMN), respetivamente.

Sejam $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_n)$ e $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ a amostra observada de $\mathbf{V} = (V_1, V_2, \ldots, V_n)$, a função log-verosimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2, \lambda)^T$ dos dados observados $(\mathbf{v}, \boldsymbol{\rho})$ é dada por

$$\ell(\boldsymbol{\theta}|\mathbf{v}, \boldsymbol{\rho}) = \sum_{i=1}^{n} \rho_i \log\left[F_{SSMN}\left(\frac{v_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}|\boldsymbol{\theta}, H\right)\right] + \sum_{i=1}^{n}(1-\rho_i)\log[f_{SSMN}(v_i|\boldsymbol{\theta}, H)],$$

onde a função densidade de probabilidade (fdp) é representada por f_{SSMN} e F_{SSMN} representa a função distribuição acumulada (fda) da classe de distribuições SSMN, para mais detalhes; veja Ferreira et al. [12] e Ferreira et al. [13].

2.1 ESTIMAÇÃO MV VIA ALGORITMO MCEM

Para estimação MV do modelo SSMN-CR procederemos de duas maneiras dependendo do tipo dados a estudar, primeiramente para os dados sem censura as estimativas serão obtidas via algoritmo EM como em Ferreira et al. [13]. Para os dados censurados o algoritmo será o MCEM que consistirá em adicionar ao passo E do EM uma simulação por Monte Carlo para geração de amostras de distribuições truncadas que denotaremos por TSSMN. O processo será ilustrado nas seguintes seções.

2.1.1 REPRESENTAÇÃO HIERÁRQUICA

Nesta secção vamos proceder com a metodologia habitual usando a representação estocástica do modelo em termos de distribuições mais tratáveis quem em geral depende de quantidades não observáveis ou dados latentes. Em consequência, usando as equações 1.3 e 1.3.1 teremos que o modelo de regressão (2.1) (SSMN-CR) possui a seguinte representação hierárquica

$$\begin{cases} Y_i|U_i = u_i; T_i = t_i \overset{ind}{\sim} N\left(\mathbf{x}_i^T\boldsymbol{\beta} + \dfrac{\sigma\lambda}{\sqrt{u_i(u_i+\lambda^2)}}t_i, \dfrac{\sigma^2}{u_i+\lambda^2}\right) \\ U_i \overset{iid}{\sim} H(\boldsymbol{\tau}) \\ T_i \overset{iid}{\sim} TN(0,1;(0,+\infty)),\ i = 1;\ldots,n. \end{cases} \tag{2.3}$$

com $U \perp T$ (todos independentes), onde $TN(0,1;(0,+\infty))$ denota a distribuição Normal Truncada padrão univariada (ver $\mid T_0 \mid$ na equação 1.3). Para mais detalhes ver Ferreira et al. [13].

2.1.2 LOG-VEROSIMILHANÇA COMPLETA

Para implementar o algoritmo MCEM, vamos assumir que as variáveis latentes no modelo SSMN-CR, dados pelo vetor de respostas censuradas $\mathbf{Y} = (y_1, \ldots, y_n)^T$, o vetor $\mathbf{t} = (t_1, \ldots, t_n)^T$ e $\mathbf{u} = (u_1, \ldots, u_n)^T$ podem ser observados. Assim, considerando os dados observados $(\mathbf{v}, \boldsymbol{\rho})$ e as variáveis latentes $(\mathbf{Y}, \mathbf{t}, \mathbf{u})$, definimos os dados completos por $y_c = (\mathbf{v}^T, \boldsymbol{\rho}^T, \mathbf{Y}^T, \mathbf{t}^T, \mathbf{u}^T)$. Então, a log-verossimilhança dos dados completos é definida por

$$\begin{aligned} \ell_c(\boldsymbol{\theta}|y_c) \;\propto\; & -n\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[u_i y_i^2 - 2\mu_i u_i y_i + \mu_i^2 u_i + t_i^2 \\ & -2\lambda t_i y_i + 2\lambda\mu_i t_i + \lambda^2 y_i^2 - 2\lambda^2\mu_i y_i + \lambda^2\mu_i^2] \\ & + \sum_{i=1}^{n}\log h(u_i|\boldsymbol{\tau}). \end{aligned} \tag{2.4}$$

Dada a estimativa atual $\widehat{\boldsymbol{\theta}}^{(k)} = (\widehat{\boldsymbol{\beta}}^{(k)T}, \widehat{\sigma}^{2^{(k)}}, \widehat{\lambda}^{(k)}, \widehat{\boldsymbol{\tau}}^{(k)T})^T$ a estimativa de $\boldsymbol{\theta}$ na $k-$ésima iteração, o Passo-E calcula a função

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = E[\ell_c(\boldsymbol{\theta}|\mathbf{y}_c); \mathbf{y}; \widehat{\boldsymbol{\theta}}^{(k)}],$$

tal que

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) \;\propto\; & -n\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[\widehat{uy}_i - 2\mu_i\widehat{uy}_i + \mu_i^2\widehat{u}_i + \widehat{t^2}_i - 2\lambda\widehat{ty}_i \\ & +2\lambda\mu_i\widehat{t}_i + \lambda^2(\widehat{y^2}_i - 2\mu_i\widehat{y}_i + \mu_i^2)] + \sum_{i=1}^{n}E[\log h(u_i|\boldsymbol{\tau})]. \end{aligned} \tag{2.5}$$

2.1.3 PASSO E

Na implementação do algoritmo MCEM para a estimação MV do modelo SSMN-CR, teremos dois casos para o passo E, onde as estimativas são obtidas a partir das propriedades da esperança como segue:

- **Caso** 1 *Sem Censura* ($V_i = y_i$)

$$\begin{aligned}
\widehat{uy}_i &= y_i E[U_i|V_i = y_i, \widehat{\boldsymbol{\theta}}], \\
\widehat{uy^2}_i &= y_i^2 E[U_i|V_i = y_i, \widehat{\boldsymbol{\theta}}], \\
\widehat{u}_i &= E[U_i|V_i = y_i, \widehat{\boldsymbol{\theta}}], \\
\widehat{t^2}_i &= E[T_i^2|V_i = y_i, \widehat{\boldsymbol{\theta}}], \\
\widehat{ty}_i &= y_i E[T_i|V_i = y_i, \widehat{\boldsymbol{\theta}}], \\
\widehat{t}_i &= E[T_i|V_i = y_i, \widehat{\boldsymbol{\theta}}], \\
\widehat{y}_i &= y_i, \\
\widehat{y^2}_i &= y_i^2.
\end{aligned} \tag{2.6}$$

- **Caso** 2 *Com Censura*

$$\begin{aligned}
\widehat{uy}_i &= E[U_i Y_i|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[E(U_i Y_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[Y_i E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}].
\end{aligned} \tag{2.7}$$

$$\begin{aligned}
\widehat{uy^2}_i &= E[U_i Y_i^2|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[E(U_i Y_i^2|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[Y_i^2 E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}].
\end{aligned} \tag{2.8}$$

$$\begin{aligned}
\widehat{u}_i &= E[U_i|V_i, \widehat{\boldsymbol{\theta}}], \\
&= E[U_i|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}].
\end{aligned} \tag{2.9}$$

$$\begin{aligned}
\widehat{ty}_i &= E[T_i Y_i|V_i, \widehat{\boldsymbol{\theta}}], \\
&= E[T_i Y_i|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[E(T_i Y_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[Y_i E(T_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}],
\end{aligned} \tag{2.10}$$

$$\begin{aligned}
\widehat{t}_i &= E[T_i|V_i, \widehat{\boldsymbol{\theta}}], \\
&= E[T_i|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[E(T_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}]
\end{aligned} \tag{2.11}$$

$$\begin{aligned}
\widehat{t^2}_i &= E[T_i^2|V_i, \widehat{\boldsymbol{\theta}}], \\
&= E[T_i^2|Y_i > c, \widehat{\boldsymbol{\theta}}], \\
&= E[E(T_i^2|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}].
\end{aligned} \tag{2.12}$$

$$\begin{aligned} \widehat{y}_i &= E[Y_i|V_i, \widehat{\boldsymbol{\theta}}], \\ &= E[Y_i|Y_i > c, \widehat{\boldsymbol{\theta}}], \end{aligned} \tag{2.13}$$

$$\begin{aligned} \widehat{y^2}_i &= E[Y_i^2|V_i, \widehat{\boldsymbol{\theta}}], \\ &= E[Y_i^2|Y_i > c, \widehat{\boldsymbol{\theta}}]. \end{aligned} \tag{2.14}$$

2.1.3.1 Esperanças Condicionais Para o Caso Sem Censura

Dado que $T_i|Y_i \sim TN(\lambda(y_i - \mu_i), \sigma^2; (0, +\infty))$ e sendo $\widehat{t_i} = E[T_i|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_i]$ e $\widehat{t_i^2} = E[T_i^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_i]$, as esperanças condicionais são obtidas usando os momentos da distribuição normal truncada, dadas por

$$\begin{aligned} \hat{t}_i &= \hat{\lambda}\hat{\eta}_i + \hat{\sigma} W_{\boldsymbol{\Phi}} \tfrac{\hat{\lambda}\hat{\eta}_i}{\hat{\sigma}}, \\ \hat{t}_i^2 &= \hat{\lambda}^2\hat{\eta}_i^2 + \hat{\sigma^2} + \hat{\lambda}\hat{\sigma}\hat{\eta}_i W_{\boldsymbol{\Phi}} \tfrac{\hat{\lambda}\hat{\eta}_i}{\hat{\sigma}}, \end{aligned} \tag{2.15}$$

onde $W_{\boldsymbol{\Phi}}(u) = \frac{\phi_1(u)}{\Phi(u)}$ e $\hat{\eta}_i = y_i - \widehat{\mu}_i$, $i = 1, \ldots, n$.

Para as distribuições da família **SSMN**, os $\widehat{u}_i$ são dados pelas equações (2.17) a (2.21), definidas a seguir para cada distribuição particular considerada.

Proposição 2.1.1. *Seja $Y \sim SSMN(\mu, \sigma^2, \lambda; H)$, então a distribuição condicional de $U|Y = y$ não depende de λ.*

Demonstração: ver [10]

Da Proposição 2.1.1 segue que, sob a distribuição SSMN mais geral considerada aqui, a distribuição condicional $U_i|Y_i$ reduz à distribuição SMN.

Essa peculiaridade simplifica a implementação do algoritmo EM. Assim, para as distribuições discutidas, dado $\boldsymbol{\theta} = (\mu, \sigma^2, \lambda, \boldsymbol{\tau})^T$, tem-se os seguintes resultados (Ferreira [10]).

- Distribuição *Skew-t-normal*, denotada por $Y \sim STN(\mu, \sigma^2, \lambda; \nu,)$, temos que de 1.8 e da proposição 2.1.1

$$f(u|y; \boldsymbol{\theta}) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} u^{\nu/2-1} e^{-\nu u/2} \frac{u^{1/2}}{\sqrt{2\pi}\sigma} e^{-ud/2} \tag{2.16}$$

$$U|Y=y \sim Gama\left(\frac{\nu+1}{2}, \frac{\nu+d}{2}\right),$$

com $d=(y-\mu)^2/\sigma^2$. Portanto, $E[U^\alpha|y,\theta]=\dfrac{\Gamma(\frac{\nu+1}{2}+\alpha)}{\Gamma(\frac{\nu+1}{2})(\frac{\nu+d}{2})^\alpha}$, $\alpha>0$ e

$$\widehat{u}=\frac{\nu+1}{\nu+d}. \tag{2.17}$$

- Distribuição *Skew-Slash*, denotada por $Y \sim SSL(\mu,\sigma^2,\lambda;\nu)$. Então, neste caso, temos que

$$f(u|y;\boldsymbol{\theta})=\nu u^{\nu-1}\mathbf{I}_{(0,1)}(u)\frac{u^{1/2}}{\sqrt{2\pi}\sigma}e^{-ud/2}, \tag{2.18}$$

$$U|Y=y \sim Gama\left(\nu+1/2, d/2\right)\mathbf{I}_{(0,1)}(u).$$

Assim, $E[U^\alpha|y,\theta]=\dfrac{\Gamma(\nu+1/2+\alpha)}{\Gamma(\nu+1/2)(d/2)^\alpha}\dfrac{P_1(\nu+\alpha+1/2,d/2)}{P_1(\nu+1/2,d/2)}$, $\alpha>0$ e

$$\widehat{u}=\frac{(2\nu+1)}{d}\frac{P_1(\nu+3/2,d/2)}{P_1(\nu+1/2,d/2)}, \tag{2.19}$$

onde $P_x(a,b)$ denota a fda de uma distribuição $Gama(a,b)$ avaliada em x.

- Distribuição *Skew-Normal Contaminada*, denotada por $Y \sim SCN(\mu,\sigma^2,\lambda;\nu,\gamma)$. Então, neste caso, temos que

$$f(u|y;\boldsymbol{\theta})=\nu p\mathbf{I}_{(u=\gamma)}+(1-\nu)p\mathbf{I}_{(u=1)}, \tag{2.20}$$

com

$$p=\frac{u^{1/2}\exp(-\frac{du}{2})}{\nu\gamma^{1/2}\exp(-\frac{d\gamma}{2})+(1-\nu)\exp(-\frac{d}{2})}.$$

Portanto, temos que

$$E[U^\alpha|y,\theta]=\frac{1-\nu+\nu\gamma^{\alpha+1/2}\exp{(1-\gamma)d/2}}{1-\nu+\nu\gamma^{1/2}\exp{(1-\gamma)d/2}},$$

$\alpha>0$ e

$$\widehat{u}=\frac{1-\nu+\nu\gamma^{3/2}\exp{(1-\gamma)d/2}}{1-\nu+\nu\gamma^{1/2}\exp{(1-\gamma)d/2}}. \tag{2.21}$$

2.1.3.2 Esperanças Condicionais Para o Caso Com Censura

No caso de dados censurados para obter estimativas dos parâmetros do modelo SSMN-CR, propomos a implementação do algoritmo MCEM (ver Seção 1.2.2), onde é incorporada uma fase de simulação por Monte Carlo na etapa E que vai permitir obter as esperanças condicionais que precisamos. O método de gerar amostras a partir das respectivas distribuições truncadas da família SSMN será ilustrado na Seção 2.1.3.3. As esperanças condicionais para o caso com censura são apresentadas no Apêndice A.

2.1.3.3 Método Para Geração De Distribuições TSSMN

A ideia do método de geração de amostras a partir de distribuições truncadas que é implementado neste trabalho surge da leitura da seção "Aspectos Computacionais"de Mattos [29].

Portanto, nesta seção, descrevemos como obter amostras aleatórias a partir da variável aleatória $Y \sim TSSMN(\mu, \sigma^2, \lambda; H, [a, b])$, isto é, amostras provenientes de distribuições truncadas da família SSMN.

De Ferreira [11], sabemos que se $Y \sim SSMN(\mu, \sigma^2, \lambda; H)$ sua representação estocástica é dada por

$$Y = \mu + \sigma \left[\frac{\lambda \mid T_0 \mid}{[U(U+\lambda^2)]^{\frac{1}{2}}} + \frac{T_1}{(U+\lambda^2)^{\frac{1}{2}}} \right], \tag{2.22}$$

onde $U \sim H(.|\tau),\ T_0 \sim N(0,1),$ e $T_1 \sim N(0,1)$ são independentes. Dessa forma, de 2.22 e assumindo $W = U^{-\frac{1}{2}}[U+\lambda^2]^{\frac{1}{2}} \mid T_0 \mid$, obtemos a seguinte representação hierárquica:

$$\begin{aligned} Y|W=w, U=u &\sim N\left(\mu + \frac{\sigma\lambda w}{u+\lambda^2}, \sigma^2(u+\lambda^2)^{-1}\right), \\ W|U=u &\sim TN\left(0, \frac{u+\lambda^2}{u}; (0,\infty)\right), \\ U &\sim H(\boldsymbol{\tau}), \end{aligned} \tag{2.23}$$

onde U e W são variáveis aleatórias positivas. Então, se $a < Y < b$, temos que

$$\frac{(a-\mu)}{\sigma} < \frac{1}{\sigma}(Y-\mu) < \frac{(b-\mu)}{\sigma}$$

e consequentemente

$$\underbrace{\left[\frac{a-\mu}{\sigma}-\frac{\lambda w}{u+\lambda^2}\right](u+\lambda^2)^{\frac{1}{2}}}_{a_1}<\underbrace{\left[\frac{1}{\sigma}(Y-\mu)-\frac{\lambda w}{u+\lambda^2}\right](u+\lambda^2)^{\frac{1}{2}}}_{T_1}<\underbrace{\left[\frac{b-\mu}{\sigma}-\frac{\lambda w}{u+\lambda^2}\right](u+\lambda^2)^{\frac{1}{2}}}_{b_1}$$

Portanto, o algoritmo para gerar amostras aleatórias dos modelos truncados será o seguinte:

P1) Gere uma amostra aleatória $U_1, U_2, \ldots, U_m$ de $H(\boldsymbol{\tau})$.

P2) Gere uma amostra aleatória $W_1, W_2, \ldots, W_m$ onde $W_i|U = u_i \sim TN\left(0, \frac{u_i+\lambda^2}{u_i}; (0, +\infty)\right)$.

P3) Calcule $a_1(u, w)$ e $b_1(u, w)$.

P4) Temos que $T_1|W = w, U = u \sim TN(0, 1; (a_1, b_1))$.

P5) Usar a representação estocástica (2.22).

2.1.4 SOLUÇÃO DO PASSO M

O passo M então maximiza $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ em relação a $\boldsymbol{\theta}$, obtendo uma nova estimativa $\widehat{\boldsymbol{\theta}}^{(k+1)}$. Dessa forma, conforme o descrito temos:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= [\mathbf{X}^T D(\widehat{\mathbf{u}}+\widehat{\lambda}^2\mathbb{I}_n)\mathbf{X}]^{-1}\mathbf{X}^T[\widehat{\mathbf{uy}}-\widehat{\lambda}\widehat{\mathbf{t}}+\widehat{\lambda}^2\widehat{\mathbf{y}}],\\
\widehat{\lambda} &= \frac{\sum_{i=1}^n[\widehat{ty}_i-\widehat{\mu}_i\widehat{t}_i]}{\sum_{i=1}^n[\widehat{y^2}_i-2\widehat{\mu}_i\widehat{y}_i+\widehat{\mu}_i^2]},\\
\hat{\sigma^2} &= \frac{1}{2n}\sum_{i=1}^n[\widehat{uy}_i-2\widehat{\mu}_i\widehat{uy}_i+\widehat{\mu}_i^2\widehat{u}_i+\widehat{t^2}_i-2\widehat{\lambda}\widehat{ty}_i\\
&\quad +2\widehat{\lambda}\widehat{\mu}_i\widehat{t}_i+\widehat{\lambda}^2(\hat{y_i^2}-2\widehat{\mu}_i\widehat{y}_i+\widehat{\mu}_i^2)].
\end{aligned}$$

Para os parâmetros ν e γ utilizamos o critério de Schwartz que basicamente consiste em criar um *grid* de r valores de ν e/ou γ e avaliar na função $\ell(\widehat{\boldsymbol{\theta}}, \boldsymbol{\tau}_j, \mathbf{v}, \boldsymbol{\rho})$ o valor de ν e/ou γ com a maior verossimilhança. Para os modelos STN-CR e SSL-CR, $\boldsymbol{\tau} = \nu$ e então $j = 1, \ldots, r$; para o modelo SCN-CR, $\boldsymbol{\tau} = (\nu, \gamma)$ e então geramos uma malha ν_j, γ_k, para $j, k = 1, \ldots, r$.

2.2 MATRIZ DE INFORMAÇÃO EMPIRICA

Para calcular a covariância assintótica das estimativas ML dos parâmetros do modelo SSMN-CR, seguimos Lin [22] (Matos [27]). Conforme definido por Meilijson [31], a matriz de informação empírica pode ser aproximada por

$$\mathbf{I_e}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \mathbf{s}(y_i|\boldsymbol{\theta})\mathbf{s}^T(y_i|\boldsymbol{\theta}) - \frac{1}{n}\mathbf{S}(y_i|\boldsymbol{\theta})\mathbf{S}^T(y_i|\boldsymbol{\theta}), \tag{2.24}$$

onde $\mathbf{S}(\mathbf{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{s}(y_i|\boldsymbol{\theta})$ e $\mathbf{s}(y_i|\boldsymbol{\theta})$ é a função escore empírica para o sujeito i. Segundo Louis [24], é possível relacionar a função escore da log-verossimilhança incompleta com a esperança condicional da função log-verossimilhança de dados completos (Matos [27]). Portanto, a função escore individual pode ser determinada como

$$\mathbf{s}(y_i|\boldsymbol{\theta}) = \frac{\partial \log f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = E\left[\frac{\partial l_i(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\theta}} \mid y_i, \boldsymbol{\theta}\right], \tag{2.25}$$

onde $\mathbf{y}_c$ é o vetor de dados completos e $l_i(\boldsymbol{\theta}|\mathbf{y}_c)$ é a verossimilhança completa dos dados formada a partir da $i-$ésima observação. Usando as MV estimativas de $\widehat{\boldsymbol{\theta}}$, dado que $\mathbf{S}(y_i|\widehat{\boldsymbol{\theta}}) = 0$, a matriz de covariância assintótica das estimativas MV pode ser aproximada por (2.25). Assim, a matriz de informação empírica $\mathbf{I}_e$ é reduzida a

$$\mathbf{I}_e(\widehat{\boldsymbol{\theta}}|\mathbf{y}) = \sum_{i=1}^{n} \widehat{\mathbf{s}}_i\widehat{\mathbf{s}}_i^T = \begin{bmatrix} \widehat{\mathbf{s}}_{i,\boldsymbol{\beta}} \\ \widehat{\mathbf{s}}_{i,\sigma^2} \\ \widehat{\mathbf{s}}_{i,\lambda} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{s}}_{i,\boldsymbol{\beta}} & \widehat{\mathbf{s}}_{i,\sigma^2} & \widehat{\mathbf{s}}_{i,\lambda} \end{bmatrix}, \tag{2.26}$$

onde $\widehat{\mathbf{s}}_i = \mathbf{s}(y_i|\widehat{\boldsymbol{\theta}}) = E\left(\frac{\partial l_i(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\theta}}|V_i, c_i, \widehat{\boldsymbol{\theta}}\right)$, com,

$$\widehat{\mathbf{s}}_{i,\boldsymbol{\beta}} = -\frac{1}{2\widehat{\sigma}^2}[-2\widehat{u_i y_i} + 2\widehat{\mu}_i\widehat{u}_i + 2\widehat{\lambda}\widehat{t}_i - 2\widehat{\lambda}^2\widehat{y}_i + 2\widehat{\lambda}^2\widehat{\mu}_i]\mathbf{x}_i,$$

$$\begin{aligned}\widehat{\mathbf{s}}_{i,\sigma^2} = & -\frac{1}{\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}[\widehat{uy^2}_i - 2\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\widehat{uy}_i + \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\widehat{u}_i + \widehat{t}_i^2 \\ & -2\widehat{\lambda}\widehat{ty}_i + 2\widehat{\lambda}\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\widehat{t}_i + \widehat{\lambda}^2\widehat{y}_i^2 - 2\widehat{\lambda}^2\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\widehat{y}_i + \widehat{\lambda}^2\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}],\end{aligned}$$

$$\widehat{\mathbf{s}}_{i,\lambda} = -\frac{1}{\widehat{\sigma}^2}[-\widehat{ty}_i + \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\widehat{t}_i + \widehat{\lambda}\widehat{y^2}_i - 2\widehat{\lambda}\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\widehat{y}_i + \widehat{\lambda}\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}].$$

3 ESTUDOS DE SIMULAÇÃO

Nesse capítulo são apresentadas ilustrações das distribuições da família SSMN-CR (Skew-normal, Skew-t-normal, Skew-slash e Skew-Normal Contaminada) em estudos simulados.

O principal objetivo destes estudos é avaliar o desempenho do algoritmo na obtenção de estimativas de máxima verossimilhança dos modelos, e examinar o comportamento dos modelos sob diferentes configurações.

3.1 ESTUDO 1: DESEMPENHO DOS ESTIMADORES DE MÁXIMA VEROSSIMILHANÇA

Nesta seção, usamos simulações de Monte Carlo para avaliar o desempenho dos estimadores de máxima verossimilhança dos parâmetros do modelo SSMN-CR. O estudo de simulação foi projetado para observar mudanças (alterações) nas estimativas variando os tamanhos amostrais e os níveis de censura à direita.

Os dados foram artificialmente gerados provenientes de modelos SSMN-CR com $\mathbf{x}_i^T = (1, x_i)$, tal que $x_i \sim U(0,1)$, $i = 1, \ldots, n$. Geramos 100 conjuntos de dados provenientes de cada um dos modelos SN-CR, STN-CR, SSL-CR e SCN-CR com a seguinte configuração: $\boldsymbol{\beta} = (\beta_0, \beta 1)^T = (5,1)^T$, $\sigma^2 = 1, \lambda = 3, \nu = 3$ para os modelos STN-CR e SSL-CR, e $\nu = (0.1, 0.1)$ para o modelo SCN-CR. A descrição de cada cenário considerado é dada como segue.

3.1.1 Cenário 1

Uma proporção de censura de 10% e diferentes tamanhos amostrais, isto é, $n = 50, 100, 200, 400$ e 600 foram considerados. O objetivo deste cenário é mostrar o comportamento assintótico dos estimadores de máxima verossimilhança obtidos via o algoritmo MCEM.

3.1.2 Cenário 2

Uma amostra de tamanho $n = 100$ e diferentes proporções de censura, isto é, $0\%, 5\%, 10\%, 20\%$ e 30% foram considerados. Nosso objetivo neste cenário é

estudar o comportamento dos modelos lineares SSMN-CR sob diferentes proporções de censura.

O nível desejado de censura foi obtido da seguinte maneira: as observações foram colocadas em ordem crescente, e um ponto limiar (c_i) foi fixado de tal forma que o numero de observações acima deste ponto correspondem para o nível de censura desejado.

Em ambos os cenários, para cada conjunto de dados provenientes do respetivo modelo SSMN-CR, ajustamos para o mesmo modelo SSMN-CR. Note que para estes cenários, há 20 diferentes configurações de simulação com 100 conjuntos de dados para cada uma. Então, para cada simulação, as estimativas dos parâmetros foram registradas.

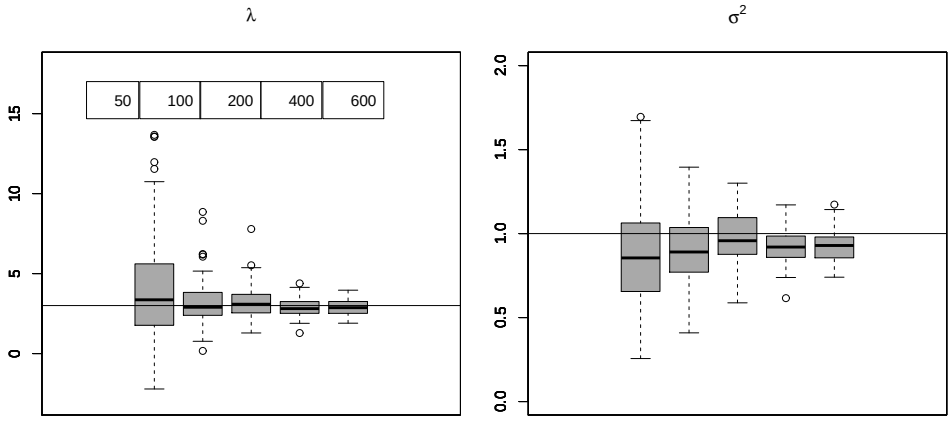

Figura 4 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 1 - Modelo Skew-normal - Parâmetros λ e σ^2.

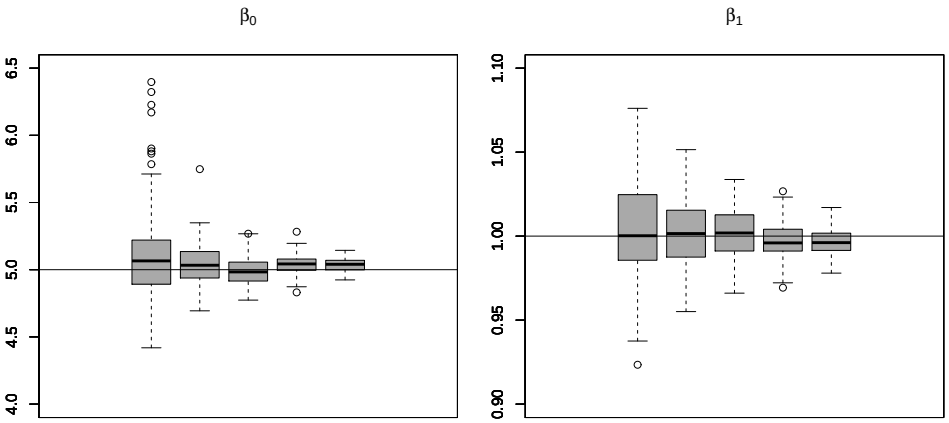

Figura 5 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 1 - Modelo Skew-normal - Parâmetros β_0 e β_1.

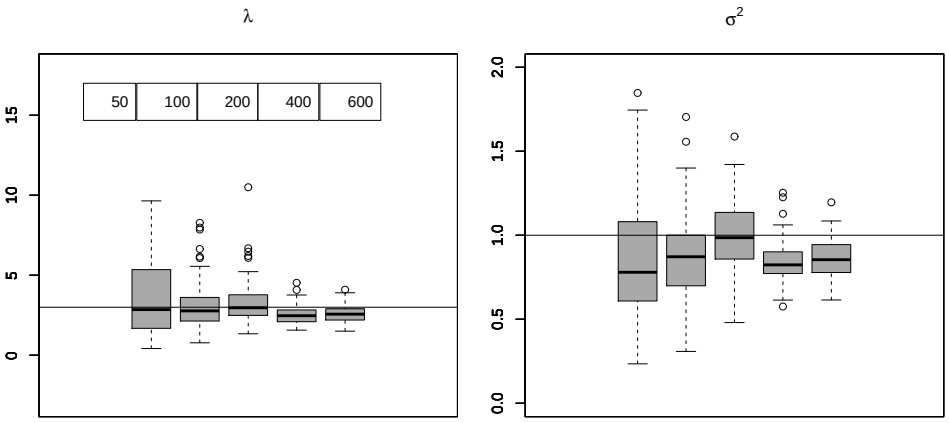

Figura 6 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 1 - Modelo Skew-t-normal - Parâmetros λ e σ^2.

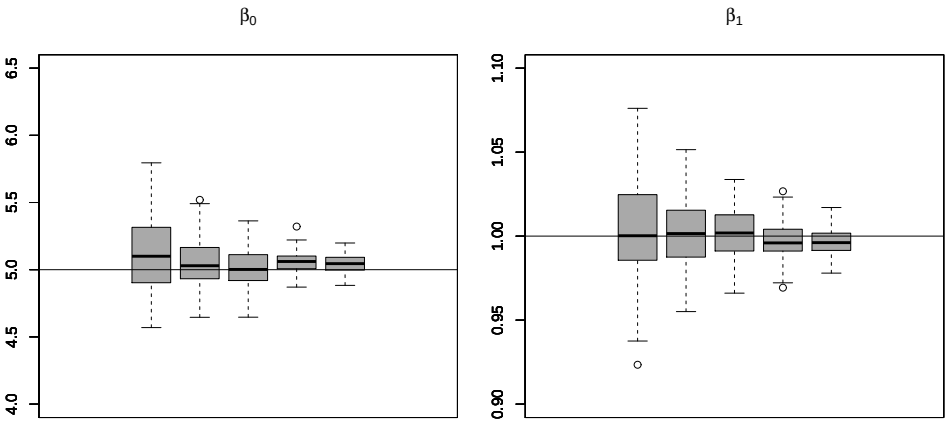

Figura 7 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 1 - Modelo Skew-t-normal - Parâmetros β_0 e β_1.

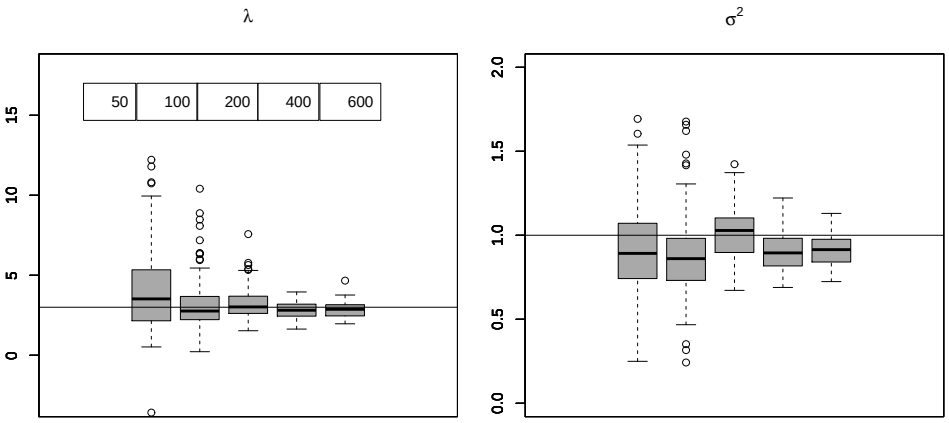

Figura 8 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 1 - Modelo Skew-Slash - Parâmetros λ e σ^2.

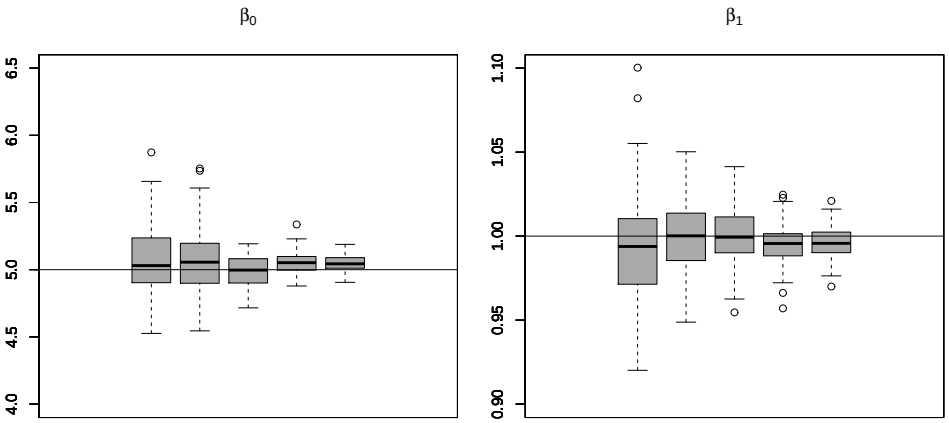

Figura 9 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 1 - Modelo Skew-Slash - Parâmetros β_0 e β_1.

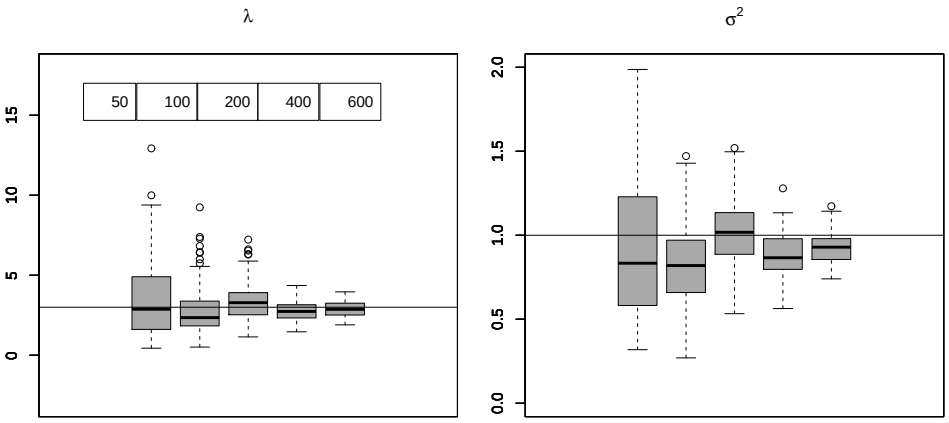

Figura 10 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 1 - Modelo Skew-Normal Contaminada - Parâmetros λ e σ^2.

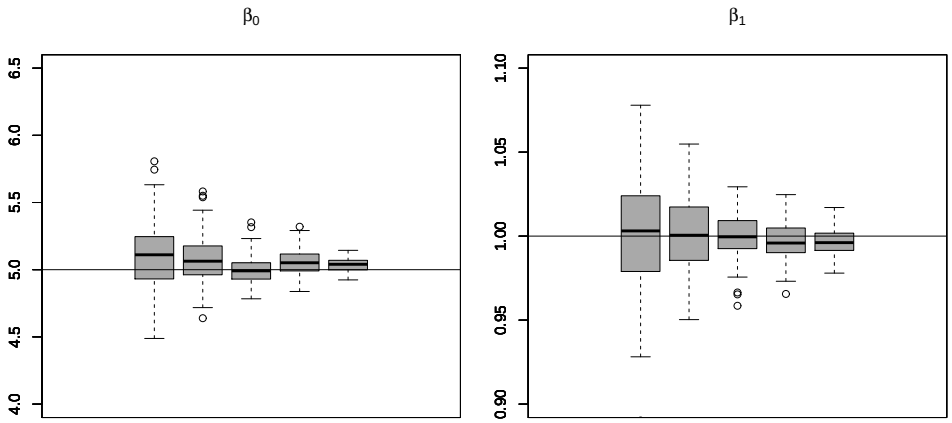

Figura 11 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 1 - Modelo Skew-Normal Contaminada - Parâmetros β_0 e β_1.

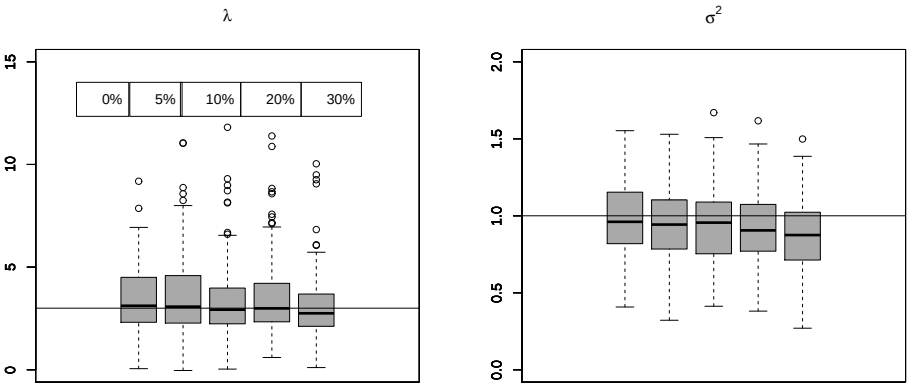

Figura 12 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 2 - Modelo Skew-normal - Parâmetros λ e σ^2.

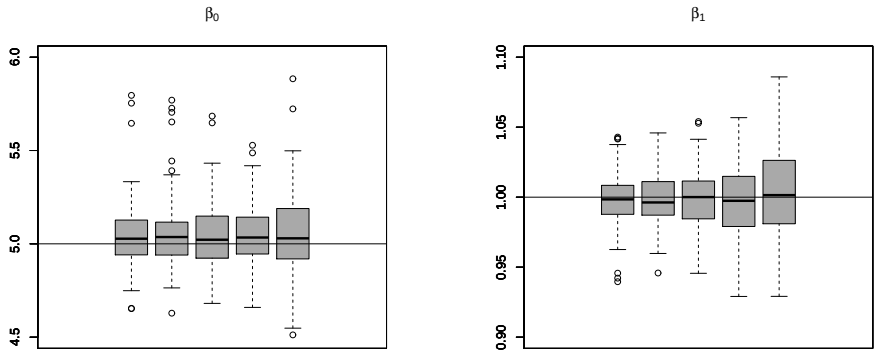

Figura 13 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 2 - Modelo Skew-normal - Parâmetros β_0 e β_1.

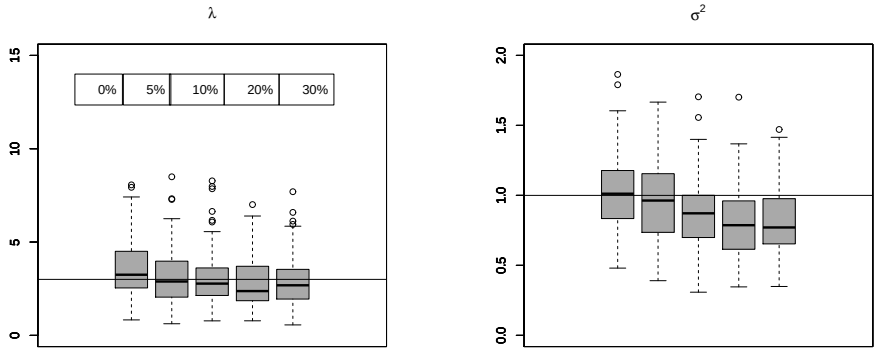

Figura 14 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 2 - Modelo Skew-t-normal - Parâmetros λ e σ^2.

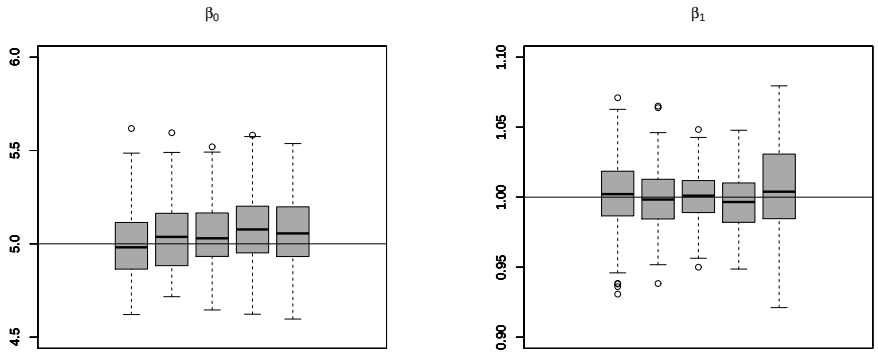

Figura 15 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 2 - Modelo Skew-t-normal - Parâmetros β_0 e β_1.

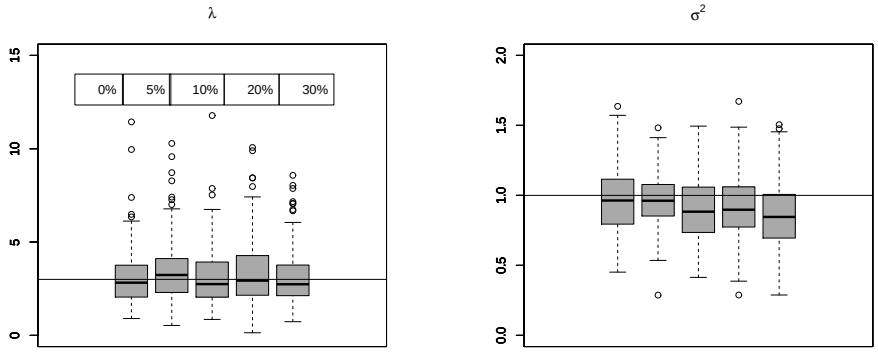

Figura 16 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 2 - Modelo Skew-Slash - Parâmetros λ e σ^2.

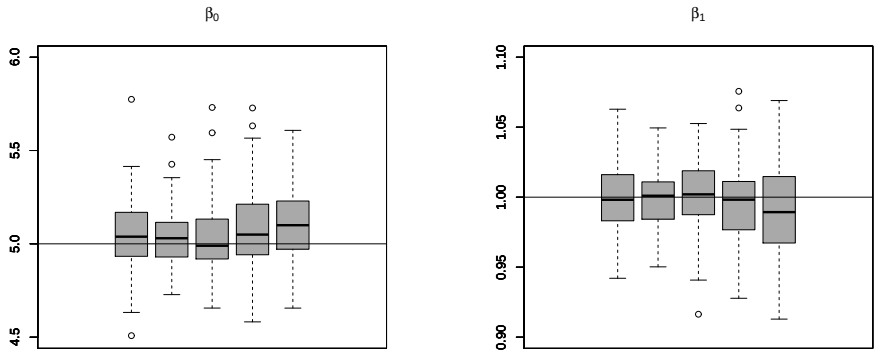

Figura 17 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 2 - Modelo Skew-Slash - Parâmetros β_0 e β_1.

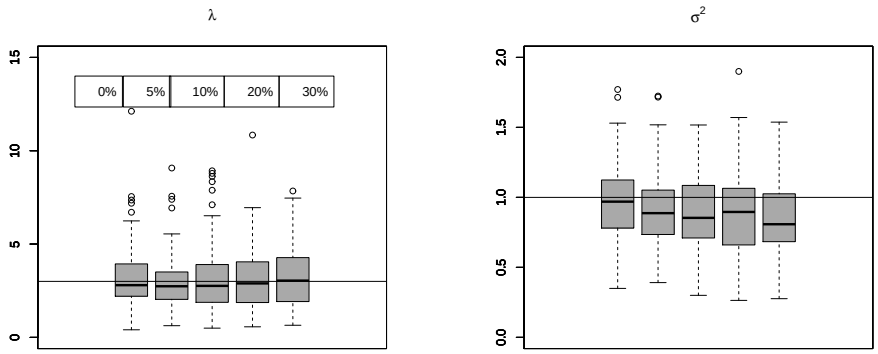

Figura 18 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 2 - Modelo Skew-Normal Contaminada - Parâmetros λ e σ^2.

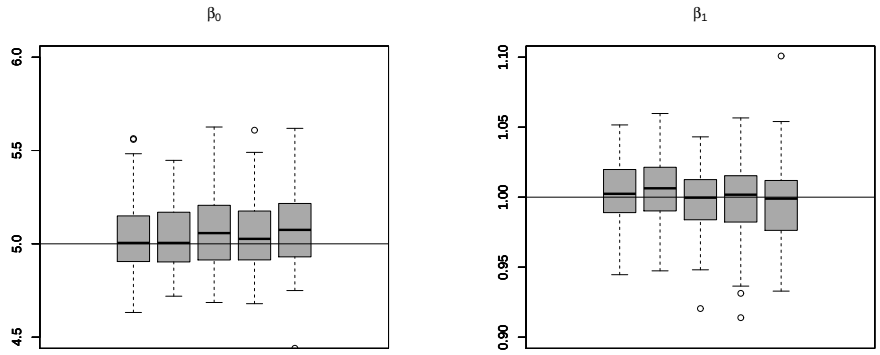

Figura 19 – **Estudo** 1: *Desempenho dos Estimadores MV*. Cenário 2 - Modelo Skew-Normal Contaminada - Parâmetros β_0 e β_1.

As Figuras 4 a 11 mostram os boxplots das estimativas dos parâmetros para os modelos SSMN-CR sob o cenário 1, e as Figuras 12 a 19 correspondem aos boxplots das estimativas dos parâmetros sob o cenário 2. Em geral, para um determinado nível de censura, o viés e a variabilidade das estimativas dos parâmetros decrescem quando o tamanho amostral aumenta. Isso essencialmente concorda com as propriedades assintóticas do estimador de máxima verossimilhança. Além disso, quando o tamanho amostral é fixo, observamos que o aumento do nível de censura corresponde ao aumento do viés e da variabilidade das estimativas dos parâmetros dos modelos analisados.

3.2 ESTUDO 2: RECUPERAÇÃO DOS PARÂMETROS E CRITÉRIOS DE SELEÇÃO

O objetivo principal deste estudo é ilustrar a capacidade dos modelos censurados com distribuições assimétricas de caudas pesadas, especificamente STN-CR e SSL-CR, de ajustar dados com uma estrutura gerada a partir de uma família de distribuições assimétricas diferentes e também investigar os efeitos na inferência paramétrica.

Nesta seção, comparamos a capacidade de alguns critérios clássicos de seleção de modelos para selecionar o modelo apropriado entre os diferentes modelos SSMN-

CR. Como não há um critério universal para a seleção de modelos, escolhemos 2 critérios para comparar os modelos propostos neste trabalho. Estes são: o critério de informação de AKAIKE (AIC) e o critério de informação BAYESIANO (BIC), definidos como seguem: $-2\ell(\widehat{\boldsymbol{\theta}}) + \varphi q_n$, onde $\ell(\boldsymbol{\theta})$ é a log-verossimilhança dos dados observados, φ é o numero de parâmetros livres que devem ser estimados no modelo e o termo de penalidade q_n é uma sequência conveniente de números positivos. Temos que $q_n = 2$ para o AIC e $q_n = \log n$ para o BIC, onde n é o tamanho da amostra (Bai [3]).

Neste estudo, consideramos 100 amostras de tamanho 100 provenientes do modelo SCN-CR com níveis de censura à direita de $0\%, 5\%, 10\%, 20\%$ e 30%, $\boldsymbol{x}_i^T = (1, x_{i1}, x_{i2})$, tais que $X_{i1} \sim U(1,5)$ e $X_{i1} \sim U(0,1)$, e os valores dos parâmetros dados por $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T = (1, -1, 4)^T$, $\sigma^2 = 2, \lambda = -4$ e $\boldsymbol{\nu} = (0.1, 0.1)$. Para cada amostra ajustamos os modelos SN-CR, STN-CR e SSL-CR. Assim, para cada simulação, as estimativas dos parâmetros bem como os critérios AIC e BIC foram registrados.

As figuras 20 a 24 mostram os boxplots das estimativas dos parâmetros para os modelos SN-CR, STN-CR e SSL-CR sob os diversos níveis de censura considerados.

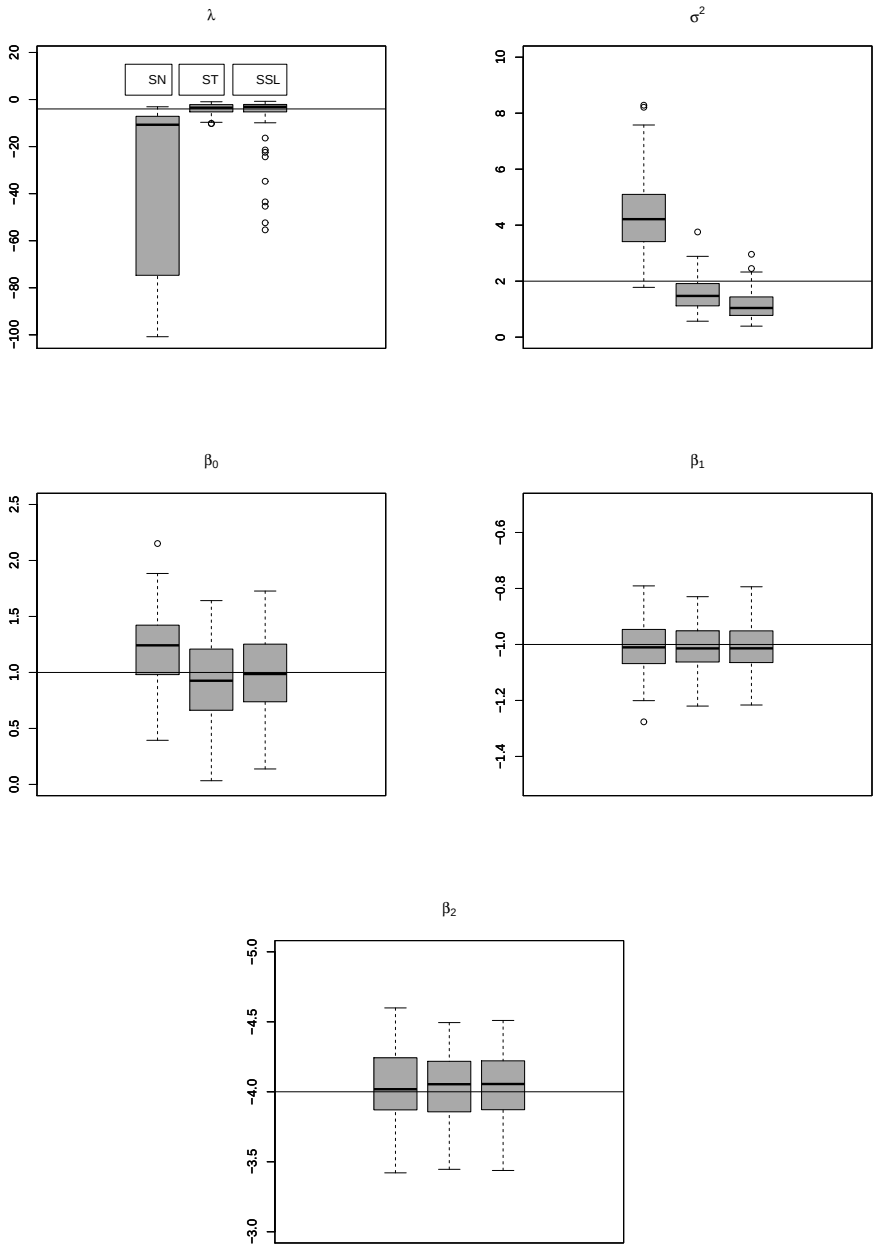

Figura 20 – **Estudo** 2: *Recuperação dos Parâmetros*. Boxplots das estimativas dos parâmetros para os modelos SN-CR, STN-CR e SSL-CR - Sem Censura.

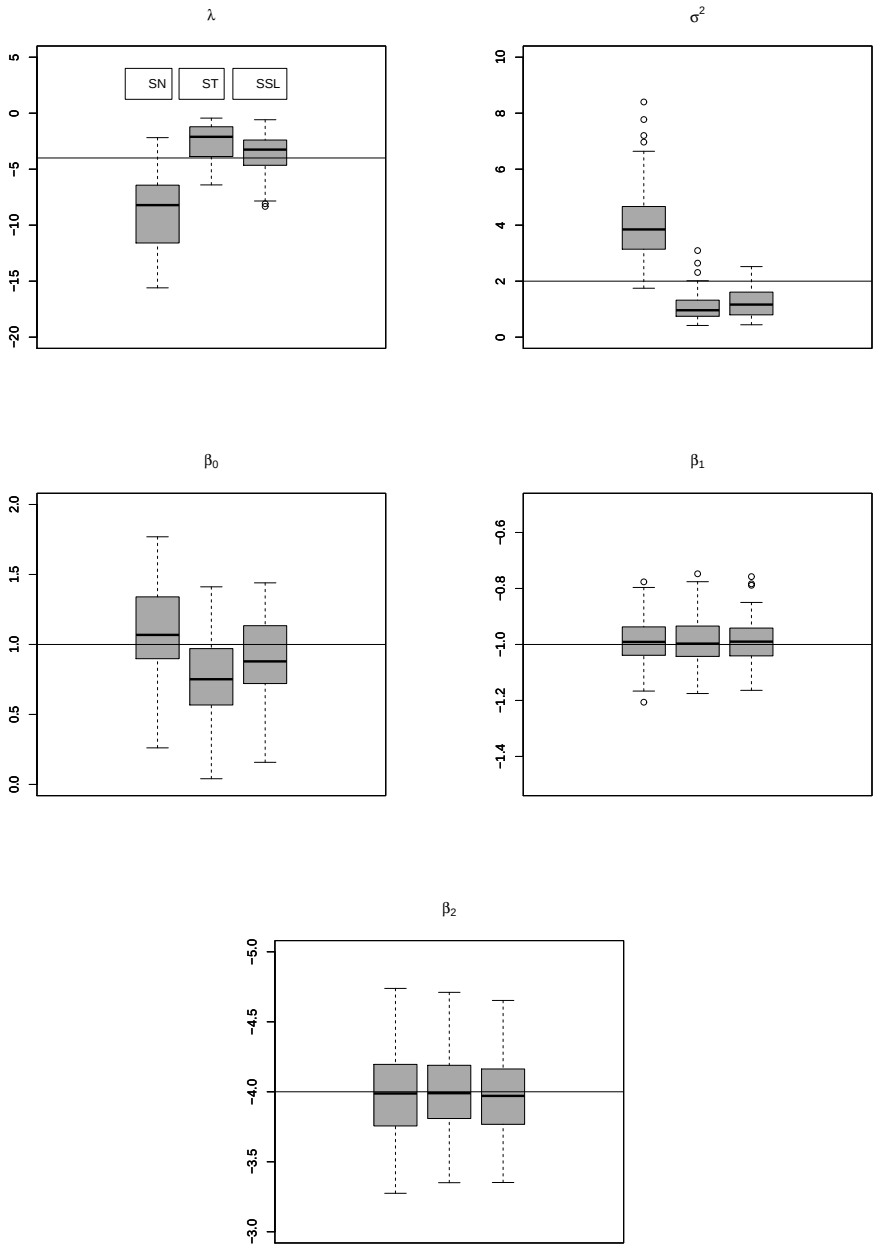

Figura 21 – **Estudo** 2: *Recuperação dos Parâmetros*. Boxplots das estimativas dos parâmetros para os modelos SN-CR, STN-CR e SSL-CR - 5% de Censura.

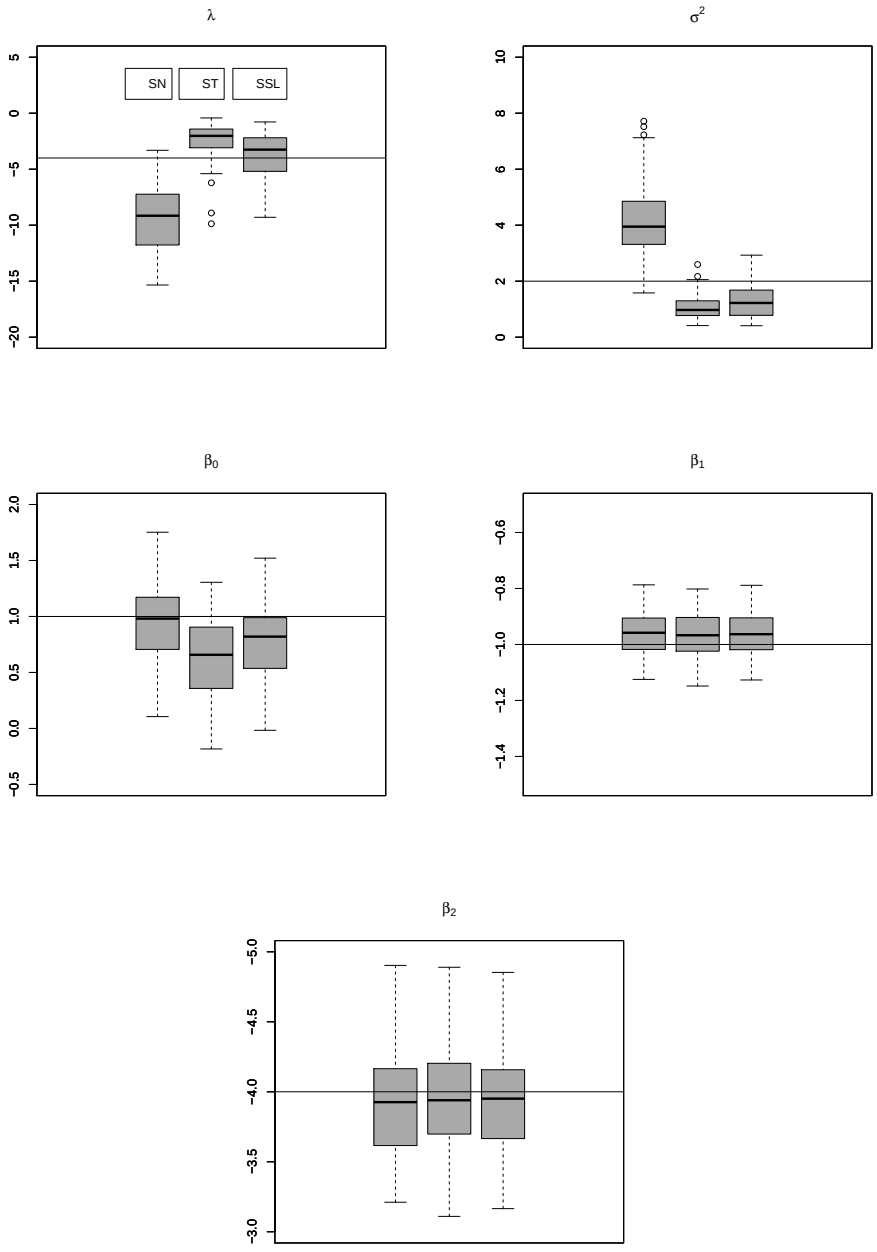

Figura 22 – **Estudo** 2: *Recuperação dos Parâmetros*. Boxplots das estimativas dos parâmetros para os modelos SN-CR, STN-CR e SSL-CR - 10% de Censura.

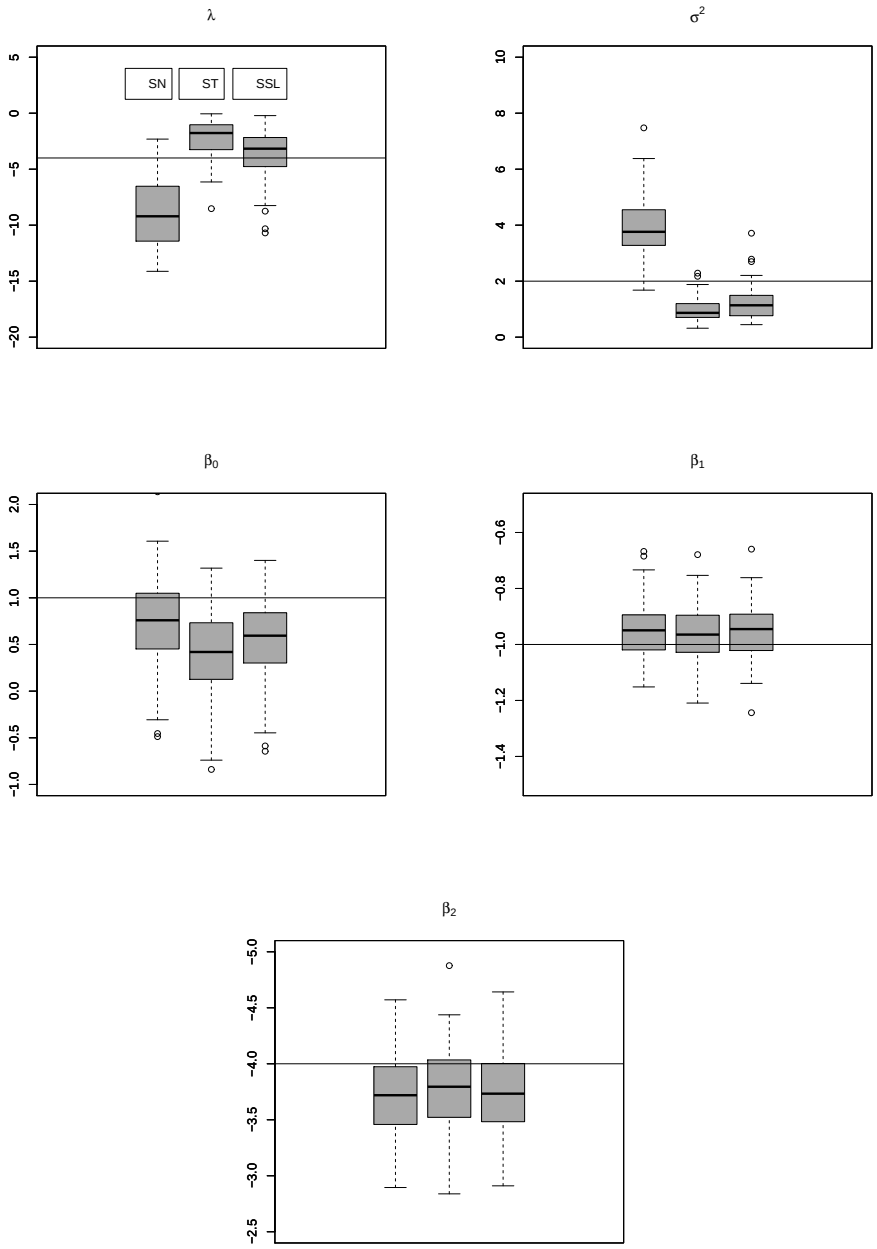

Figura 23 – **Estudo** 2: *Recuperação dos Parâmetros.* Boxplots das estimativas dos parâmetros para os modelos SN-CR, STN-CR e SSL-CR - 20% de Censura.

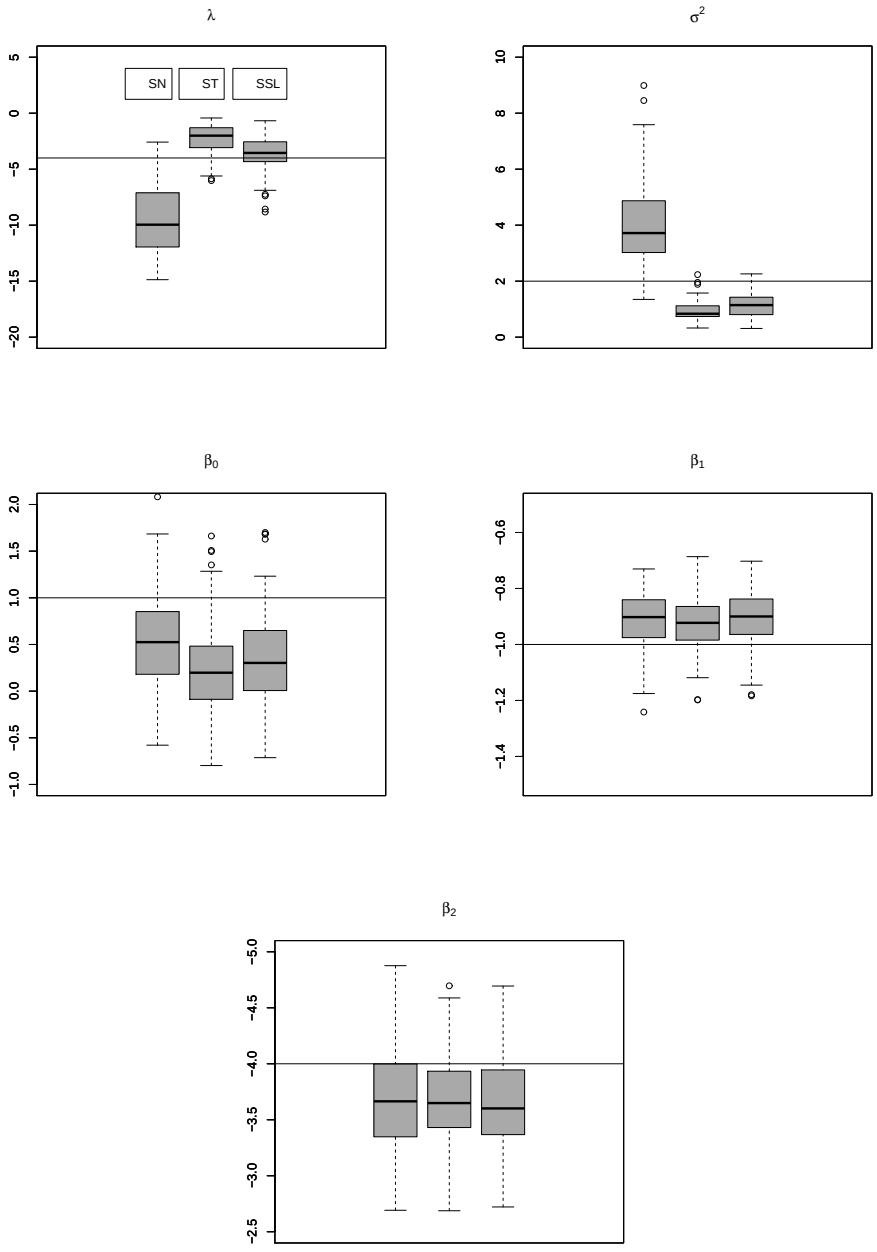

Figura 24 – **Estudo** 2: *Recuperação dos Parâmetros.* Boxplots das estimativas dos parâmetros para os modelos SN-CR, STN-CR e SSL-CR - 30% de Censura.

Nos diferentes níveis de censura considerados, em termos de viés e variabilidade, não foram observados comportamentos ou padrões diferentes nas estimativas dos parâmetros de locação (β_0, β_1 e β_2) entre os modelos ajustados. Já as estimativas dos parâmetros de escala (σ^2 e λ), provenientes dos modelos com distribuições de caudas pesadas, apresentam menores viés e variabilidade em relação ao modelo SN-CR para todos os níveis de censura.

Além disso, é facilmente visto que as estimativas dos parâmetros de escala provenientes dos modelos com distribuições de caudas pesadas são menos sensíveis à variação do nível de censura. Isso indica que estes modelos não são apenas robustos à "misspecification" do modelo (ou seja, para modelar erros de especificação), mas também para diferentes níveis de censura.

A Tabela 1 mostra as porcentagens em que os modelos censurados com distribuições de caudas pesadas, especificamente STN-CR e SSL-CR, são preferíveis ao outro modelo SN-CR ajustado.

Tabela 1 – Porcentagens dos modelos preferidos sob as condições examinadas.

Níveis de Censura	**Condições Examinadas**	**AIC**	**BIC**
0%	SN vs STN	81	75
	SN vs SSL	86	76
5%	SN vs STN	61	71
	SN vs SSL	83	76
10%	SN vs STN	87	80
	SN vs SSL	92	81
20%	SN vs STN	89	79
	SN vs SSL	93	82
30%	SN vs STN	89	80
	SN vs SSL	90	81

Não surpreendentemente, sob os diferentes níveis de censura, todos os critérios favorecem os modelos censurados baseados em distribuições de caudas pesadas.

3.3 ESTUDO 3: IMPUTAÇÃO DE OBSERVAÇÕES CENSURADAS

Neste estudo, estamos interessados em predizer as observações censuradas, denotadas y_i^c. Na implementação do algoritmo MCEM, na k-ésima iteração, as predições das observações censuradas, denotadas por $\tilde{y}_i^{c(k)}$, são calculados como

$$\tilde{y}_i^{c(k)} = E[Y_i|V_i, C_i, \widehat{\boldsymbol{\theta}}^{(k)}], \quad i = 1, \ldots, n, \text{ onde}$$

$$\tilde{y}_i^{c(k)} = \frac{1}{m}\sum_{\ell=1}^{m} y_i^{c(k,c)}.$$

É importante observar que os componentes $y_i^{c(k,c)}$ são obtidos sem esforço computacional do passo E do algoritmo MCEM proposto. Embora possamos obter valores preditos das respostas censuradas a cada iteração do algoritmo, consideramos apenas os valores dessas predições na última iteração do algoritmo MCEM.

Primeiramente, consideramos 100 amostras de tamanho 100 provenientes do modelo SCN-CR com níveis de censura *à* direita $5\%, 10\%, 20\%$ ou 30%, $\mathbf{x}_i^T = (1, x_{i1}, x_{i2})$, tais que $x_{i1} \sim U(1, 5)$ e $x_{i2} \sim U(0, 1)$, e a mesma configuração para os parâmetros do modelo definidos previamente no estudo 2. Para cada amostra, ajustamos os modelos SN-CR, STN-CR e SSL-CR. Assim, para cada simulação as predições das observações censuradas foram registradas.

Com o intuito de investigar o comportamento da predição quando a distribuição do modelo é mal especificada, propomos comparar o desempenho da medição via algoritmo MCEM através de duas medidas de discrepância empíricas, MAE(erro absoluto médio) e MSE (erro quadrático médio); veja Matos et al., [28] para mais detalhes. Estas medidas são dadas por

$$\text{MAE} = \frac{1}{100}\sum_{i,j} \mid y_{i,j} - \tilde{y}_{i,j}^c \mid \tag{3.1}$$

e

$$\text{MSE} = \frac{1}{100}\sum_{i,j} \left(y_{i,j} - \tilde{y}_{i,j}^c\right)^2, \tag{3.2}$$

onde $y_{i,j}$ é o valor original da i-ésima observação na j-ésima simulação e $\tilde{y}_{i,j}^c$ é o valor predito para a i-ésima observação censurada na j-ésima simulação para

$j = 1, \ldots, 100$ e $i = 1, \ldots, n_c$, tal que n_c é o numero de observações censuradas $(5, 10, 20$ ou 30 dependendo da proporção de censura considerada).

A Tabela 2 mostra as medidas MAE e MSE para os modelos SN-CR, STN-CR e SSL-CR com diferentes níveis de censura.

Tabela 2 – Avaliação da acurácia da medição para os modelos SN-CR, STN-CR e SSL-CR com diferentes níveis de censura.

Níveis de Censura	**Medidas**	**SN**	**STN**	**SSL**
5%	MAE	1.5993	1.5778	1.5801
	MSE	0.9037	0.8342	0.8600
10%	MAE	4.1277	3.9365	4.0320
	MSE	3.0452	2.6905	2.8763
20%	MAE	10.2852	9.7301	9.9687
	MSE	9.25091	7.9992	8.5780
30%	MAE	18.3699	17.2280	17.8197
	MSE	18.4904	16.1069	17.3085

Pode-se observar a partir desses resultados que os modelos STN-CR e SSL-CR (baseados em distribuições de caudas pesadas) geram valores preditivos mais próximos dos valores reais. Por fim, nota-se uma perda de acurácia na predição das observações censuradas à medida que aumenta-se o nível de censura.

Finalmente, neste estudo de simulação, também avaliamos o desempenho dos estimadores de máxima verossimilhança dos parâmetros dos modelos SSMN quando preenchemos os valores censurados, isto é, com as observações censuradas devidamente imputadas considerando os valores preditos obtidos via MCEM. Neste contexto, o estudo de simulação foi projetado para observar mudanças nas estimativas variando os níveis de censura à direita $(5\%, 10\%, 20\%, 30\%)$.

Os dados foram artificialmente gerados provenientes dos modelos SSMN-CR com $\mathbf{x}_i^T = (1, x_i)$ tal que $x_i \sim U(0, 1), i = 1, \ldots, 100$. Geramos 100 conjuntos de dados provenientes de cada um dos modelos SN-CR, STN-CR, SSL-CR e SCN-CR com a mesma configuração para os parâmetros dos modelos definidos previamente no estudo 1. Para cada conjunto de dados proveniente do modelo SSMN-CR, ajustamos o modelo SSMN-CR e calculamos $\tilde{y}_i^c$. Então, para cada simulação, as observações censuradas foram imputadas e as estimativas dos parâmetros dos

modelos SSMN para cada conjunto de dados completo (dados sem censura) foram obtidas através do algoritmo EM proposto em Ferreira et al. [13].

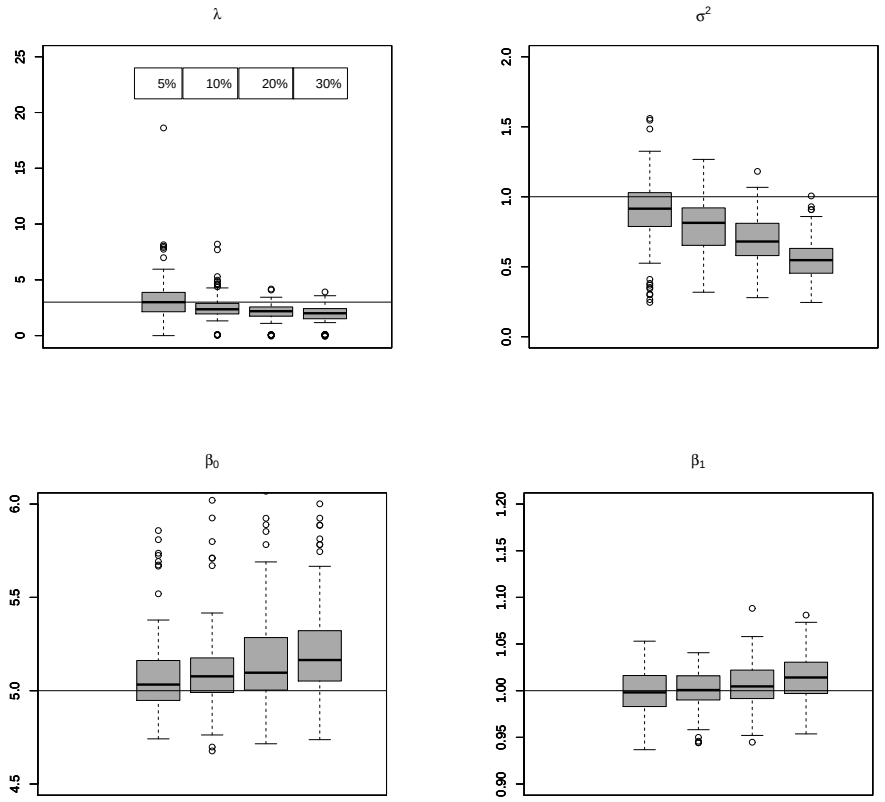

Figura 25 – **Estudo** 3: *Imputação de Observações Censuradas*. Boxplots das estimativas dos parâmetros, quando as observações censuradas são imputadas, para o modelo Skew-normal sob diferentes níveis de censura à direita $(5\%, 10\%, 20\%, 30\%)$.

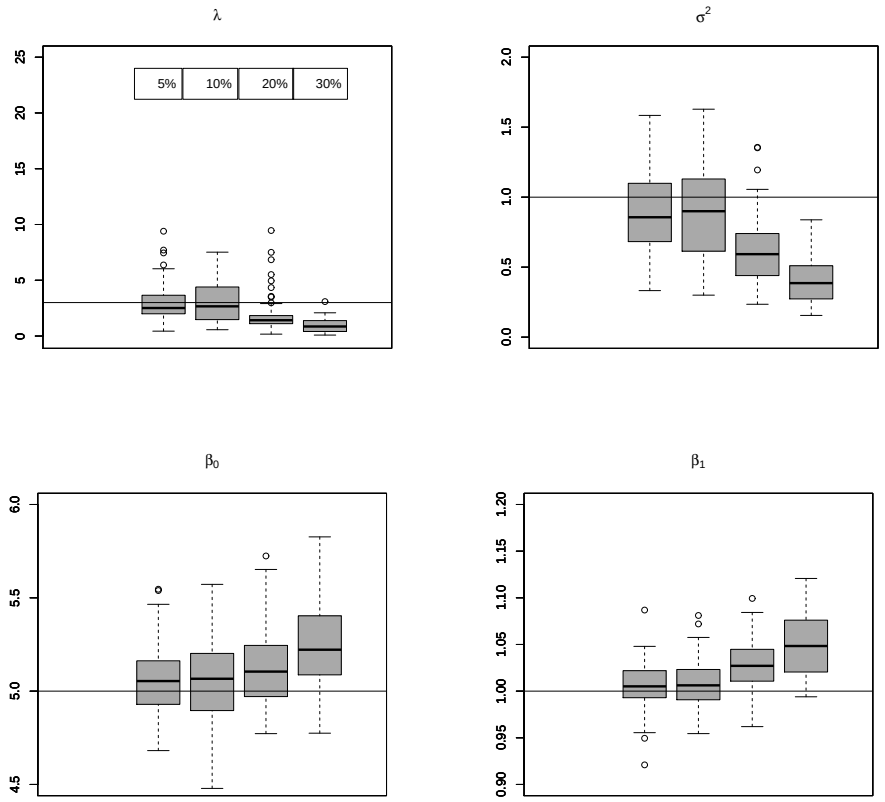

Figura 26 – **Estudo** 3: *Imputação de Observações Censuradas.* Boxplots das estimativas dos parâmetros, quando as observações censuradas são imputadas, para o modelo Skew-t-normal sob diferentes níveis de censura à direita $(5\%, 10\%, 20\%, 30\%)$.

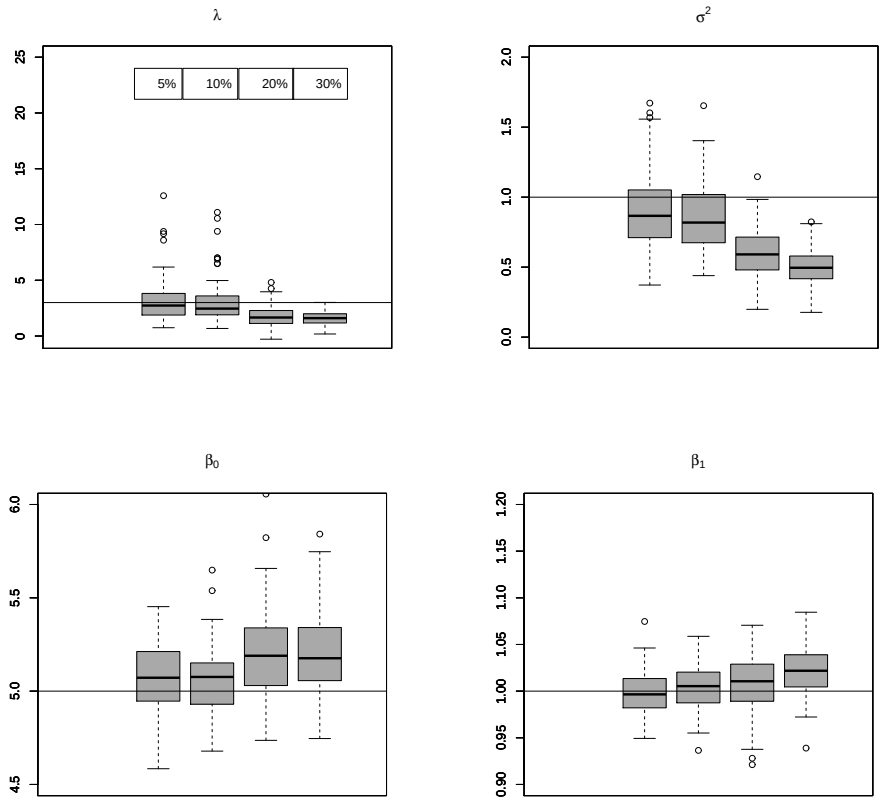

Figura 27 – **Estudo** 3: *Imputação de Observações Censuradas.* Boxplots das estimativas dos parâmetros, quando as observações censuradas são imputadas, para o modelo Skew-Slash sob diferentes níveis de censura à direita $(5\%, 10\%, 20\%, 30\%)$.

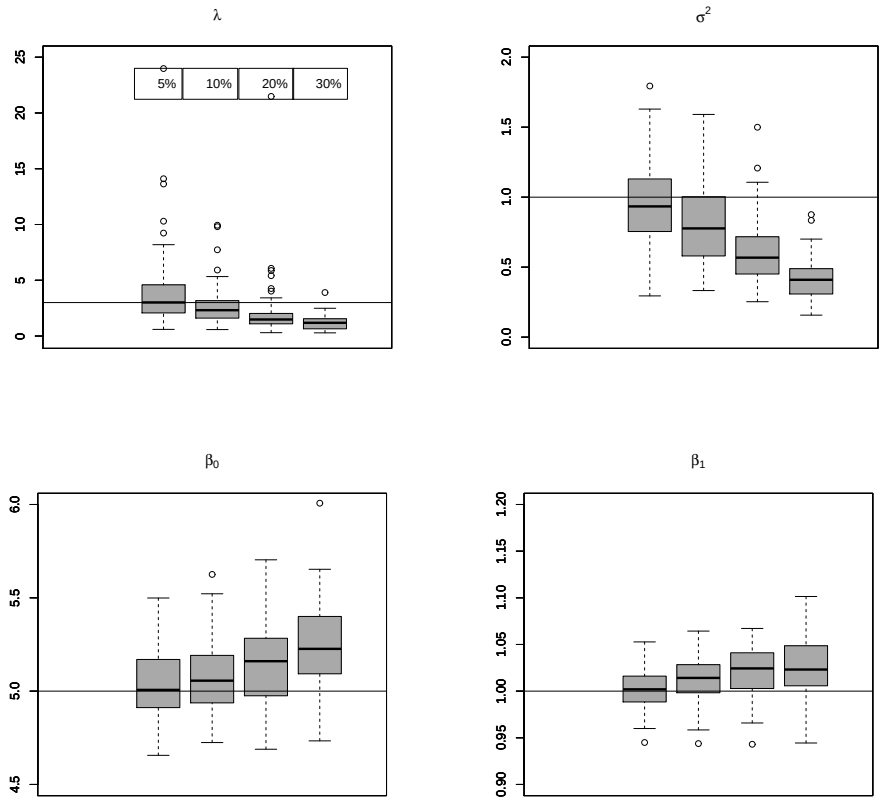

Figura 28 – **Estudo** 3: *Imputação de Observações Censuradas.* Boxplots das estimativas dos parâmetros, quando as observações censuradas são imputadas, para o modelo Skew-Normal Contaminada sob diferentes níveis de censura à direita $(5\%, 10\%, 20\%, 30\%)$.

As Figuras 25 a 28 mostram os boxplots das estimativas dos parâmetros, quando as observações censuradas são imputadas, para os modelos SSMN-CR, sob os diversos níveis de censura considerados. Em geral, para um determinado tamanho amostral, o viés das estimativas dos parâmetros aumenta quando o nível de censura aumenta. Além disso, observamos que o aumento do nível de censura corresponde ao aumento da variabilidade das estimativas dos parâmetros de locação (β_0, β_1) dos modelos analisados. Porém, nota-se que um aumento do nível de censura corresponde ao decréscimo da variabilidade das estimativas dos parâmetros de escala (σ^2, λ) dos modelos analisados.

3.4 ESTUDO 4: INFERÊNCIA DE UM ÚNICO OUTLIER

O objetivo deste estudo de simulação é avaliar a flexibilidade dos modelos SSMN-CR considerando a inferência de uma única observação aberrante na estimativa de máxima verossimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \lambda)^T$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$. Sem perda de generalidade, simulamos um conjunto de dados proveniente do modelo de regressão trigonométrica Skew-normal, tal que

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \operatorname{sin}(6\pi x_i) + \xi_i$$

$x_i \sim U(0,1)$, $\xi \sim SN(0, \sigma^2, \lambda), i = 1, \ldots, 200$, onde $\sigma^2 = 0.1$, $\lambda = -4$ e $\boldsymbol{\beta} = \begin{pmatrix} -1 & 1 & 2 \end{pmatrix}^T$, com nível de censura de 10% à direita e (ponto de corte $c = 0,0705$).

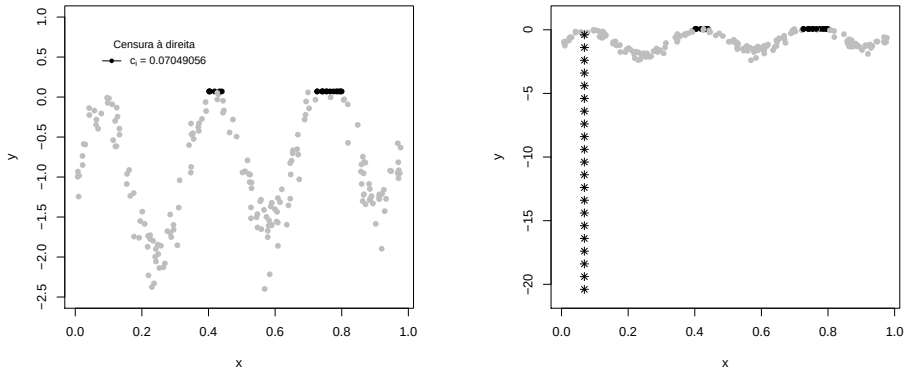

Figura 29 – **Estudo** 4: *Inferência de um Único Outlier.* (**Esquerda**) Diagrama de Dispersão dos Dados provenientes do modelo de regressão trigonométrica Skew-normal com nível de censura de 10% à direita. (**Direita**) Contaminação da Observação 200 ($y_{200} = -0.3955541$) variando δ entre 0 e 20.

Para esta amostra, ajustamos os modelos SN-CR, STN-CR e SSL-CR. Para simplificar, estudamos a influência da variação de δ unidades em uma única observação y_i não censurada na estimativa de máxima verossimilhança de $\boldsymbol{\theta}$, isto é, substituímos a observação y_i pelo valor contaminado $y_1(\delta) = y_i - \delta$. Neste exemplo, contaminamos a observação 200 ($y_{200} = -0.3955541$) variando δ entre 0 e 20. A Figura 29 apresenta o diagrama de dispersão dos dados e ilustra a contaminação da observação 200 (denotados por asterisco).

Seguindo Fagundes et al. [9], a influência de um único outlier nas estimativas pode ser avaliada pela medida MMER, definida abaixo: suponha que $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ é um vetor de parâmetros genérico e $\widehat{\phi}_j(\delta)$ é a estimativa de ϕ_j após a contaminação dos dados. Assim,

$$\text{MMER}(\boldsymbol{\phi}) = \sum_{j=1}^{n_p} \frac{1}{n_p} \mid (\widehat{\theta}_j(\delta) - \widehat{\phi}_j)/\widehat{\phi}_j \mid,$$

onde $\widehat{\phi}_j$ é a estimativa de máxima verossimilhança de θ_j. Por exemplo, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)T}, \theta^{(2)T})^T$, com $\boldsymbol{\theta}^{(1)} = (\beta_0, \beta_1, \beta_2)^T = \boldsymbol{\beta}^T$ e $\boldsymbol{\theta}^{(2)} = (\sigma^2, \lambda)^T$.

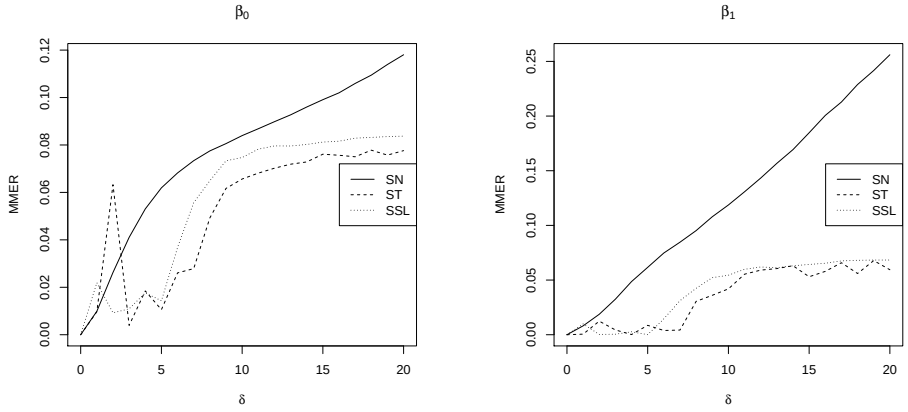

Figura 30 – **Estudo** 4: *Inferência de um Único Outlier*. Medidas MMER'S para diferentes contaminações δ (entre 0 e 20) Sob modelos SN-CR, STN-CR e SSL-CR. Parâmetros β_0 e β_1 .

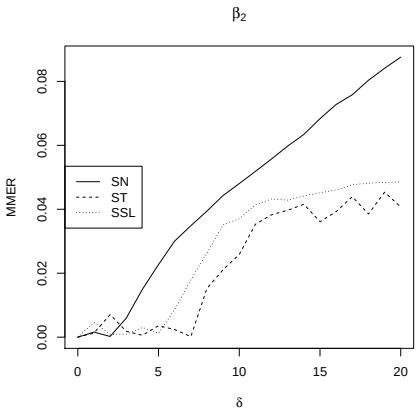

Figura 31 – **Estudo** 4: *Inferência de um Único Outlier*. Medidas MMER'S para diferentes contaminações δ (entre 0 e 20) Sob modelos SN-CR, STN-CR e SSL-CR. Parâmetro β_2.

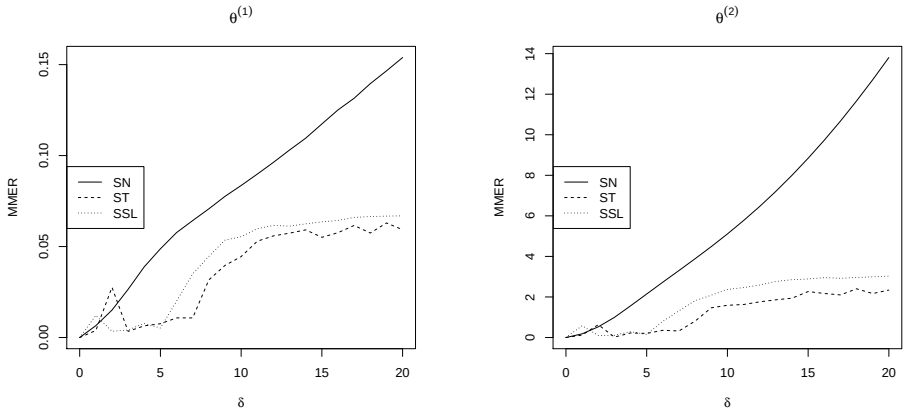

Figura 32 – **Estudo** 4: *Inferência de um Único Outlier*. Medidas MMER'S para diferentes contaminações δ (entre 0 e 20) Sob modelos SN-CR, STN-CR e SSL-CR. Parâmetros $\theta^{(1)}$ e $\theta^{(2)}$.

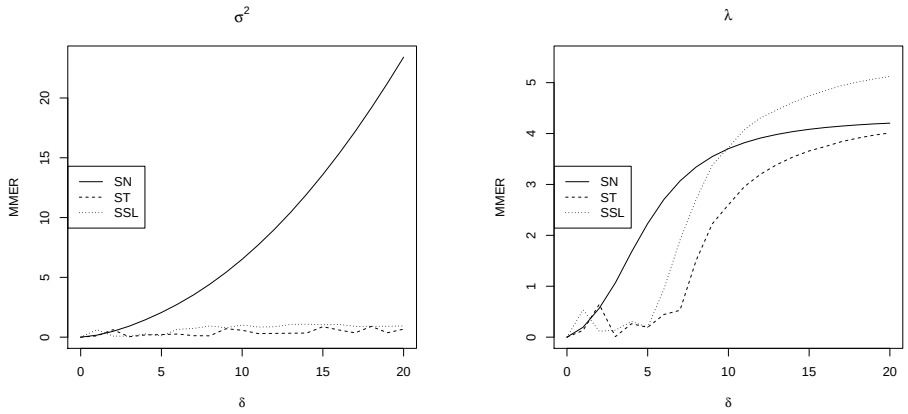

Figura 33 – **Estudo** 4: *Inferência de um Único Outlier*. Medidas MMER'S para diferentes contaminações δ (entre 0 e 20) Sob modelos SN-CR, STN-CR e SSL-CR. Parâmetros σ^2 e λ.

Nas Figuras 30 a 33, apresentamos os resultados das medidas MMER'S para diferentes contaminações δ. Como esperado, as estimativas sob os modelos assimétricos com caudas pesadas (ST-CR e SSL-CR) são menos afetadas pelas variações de δ em relação ao modelo Skew-normal. Além disso, a Figura 34 mostra os valores de BIC para todos os modelos, para cada versão perturbada do conjunto de dados original. Claramente, pode-se ver que à medida que a observação se torna mais atípica, os modelos de caudas pesadas melhor ajustam os dados.

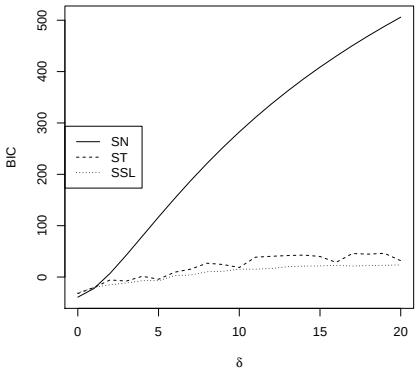

Figura 34 – **Estudo** 4: *Inferência de um Único Outlier.* valores de BIC para os modelos SN-CR, STN-CR e SSL-CR, para cada versão perturbada do conjunto de dados original.

Finalmente, este estudo de simulação também avalia a influencia de uma única observação aberrante na predição de componentes censurados via algoritmo MCEM. Neste exemplo, temos 20 observações censuradas e a Figura 35 apresenta os resultados das medidas MMER'S para diferentes contaminações δ.

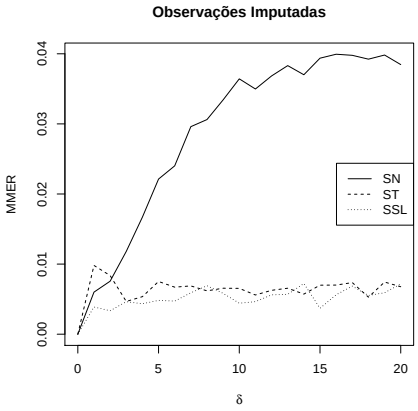

Figura 35 – **Estudo** 4: *Inferência de um único Outlier.* Resultados das medidas MMER'S para diferentes contaminações δ sob 20 observações censuradas.

Observamos que o algoritmo MCEM forneceu uma predição satisfatória desses valores censurados quando distribuições de caudas pesadas são usadas.

4 APLICAÇÃO: CONJUNTO DE DADOS DE TAXA SALARIAL

Nesse capitulo fornecemos uma aplicação dos resultados desenvolvidos no Capítulo 2, usando os dados descritos por Mroz [33]. O conjunto de dados consiste em 753 mulheres brancas casadas com idades entre 30 e 60 anos em 1975, com 428 mulheres que trabalharam em algum momento durante esse ano. A variável de resposta é a taxa salarial, que representa uma medida do salário da dona de casa conhecida como o salário médio por hora. Se os salários forem iguais a zero, essas esposas não trabalharam em 1975. Portanto, essas observações são consideradas censuradas à esquerda em zero, ou seja, um nível de censura de $43,16\%$.

As variáveis envolvidas no estudo estão descritas na Tabela 3. Considere: y_i: como o salário médio por hora (salários); x_1: idade da esposa; x_2: anos de escolaridade da esposa; x_3: o número de crianças menores de seis anos no domicílio; e x_4: o número de crianças com idades entre seis e dezenove anos.

Tabela 3 – Variáveis Envolvidas no Estudo: Conjunto de Dados Taxa Salarial.

y_i	Salário médio por hora (salários)
x_1	Idade da esposa
x_2	Anos de escolaridade da esposa
x_3	Número de crianças menores de seis anos no domicílio
x_4	Número de crianças com idades entre seis e dezenove anos

Esses dados foram analisados por Arellano-Valle et al. [2] usando o modelo de regressão censurada por t-Student; por Garay et al. [16] considerando os modelos SMN-CR, por Massuia et al. [26] para avaliar o desempenho dos modelos da SMSN-CR a partir de uma perspectiva bayesiana e, mais recentemente por Mattos [29], na estimação de modelos de regressão SMSN-CR (Misturas de Escala para a Normal assimétrica com censura). Em nosso caso, revisitamos este conjunto de dados para avaliar o desempenho do algoritmo MCEM proposto para obter estimativas de ML dos modelos SSMN-CR.

A Tabela 4 contém as estimativas MV para os parâmetros dos quatro modelos SSMN-CR, isto é, os modelos SN-CR, STN-CR, SSL-CR e SCN-CR, juntamente com os seus erros padrão correspondentes calculados através da matriz

de informação empírica. Para os modelos STN-CR e SSL-CR, o valor estimado de ν é pequeno, indicando a falta de adequação da suposição de Skew-normal (e normal) para o conjunto de dados de taxas salariais. Além disso, para os modelos assimétricos embora as estimativas de λ sejam pequenas em valor absoluto, o ganho na verossimilhança em relação aos modelos simétricos bem como no AIC e BIC foi considerável (ver [18]).

Tabela 4 – Estimativas dos parâmetros dos modelos SSMN-CR e SE para o conjunto de dados da taxa salarial e critérios de seleção de modelos (os valores em negrito correspondem ao melhor modelo).

	SN-CR		STN-CR		SSL-CR		SCN-CR	
Parâmetro	Estimativa	SE	Estimativa	SE	Estimativa	SE	Estimativa	SE
β_1	-1,0363	0,0103	-7,707e-01	0,4334	-0,8479	0,3431	-1,2089	0,5897
β_2	-0,0229	0,0002	-9,673e-02	0,0066	-0,1049	0,0051	-0,1048	0,0088
β_3	0,1184	0,0005	6,297e-01	0,0177	0,6348	0,0139	0,6403	0,0239
β_4	-0,6660	0,0027	-3,009e+00	0,0934	-3,0627	0,0717	-3,0453	0,1277
β_5	-0,0680	0,0009	-2,846e-01	0,0357	-0,2944	0,0279	-0,2969	0,0476
σ^2	19,2747	0,7216	7,614e+00	0,9532	5,3626	0,5688	10,9944	1,0293
λ	13,2595	8,1281	-1,269e-01	0,0536	0,0555	0,0544	0,0229	0,0982
ν	-	-	2,1	-	1,1	-	0,1	-
γ	-	-	-	-	-	-	0,1	-
Log-Ver.	-1705,596		-1434,391		-1455,742		**−1429, 882**	
AIC	3425,193		2884,782		2927,483		**2877, 763**	
BIC	3457,561		2921,775		2964,476		**2919, 38**	

Note que a Tabela 4 compara o ajuste dos quatro modelos de SSMN-CR usando os critérios de seleção de modelo AIC e BIC definidos na Seção 3.2. Observe que as distribuições SSMN-CR com caudas pesadas têm melhor ajuste do que o modelo SN-CR e o modelo que apresentou o melhor ajuste foi o SCN-CR.

Com o intuito de estudar os desvios do pressuposto para distribuição do erro e a presença de outliers, analisamos a transformação do tipo resíduo de Martingal (MT), denotado por r_{MT_i}, proposto por Barros et al. [4] para modelos censurados. Estes resíduos são definidos por

$$r_{MT_i} = Sinal(r_{M_i})\sqrt{-2[r_{M_i} + \rho_i \log(\rho_i - rM_i)]},\ i = 1, \ldots, n,$$

onde $r_{MT_i} = \rho_i + log(S(y_i; \widehat{\boldsymbol{\theta}}))$ é o resíduo de Martingal, com $\rho_i = 0, 1$ indicando se a observação é censurada ou não, respectivamente, tal que $S(y_i; \widehat{\boldsymbol{\theta}})$ é a estimativa

do MCEM da função de sobrevivência de y - veja mais detalhes em Ortega et al. [34] e Garay et al. [16].

O gráfico de probabilidade normal dos resíduos MT com envelopes simulados para os diversos modelos ajustados é apresentado na Figura 36.

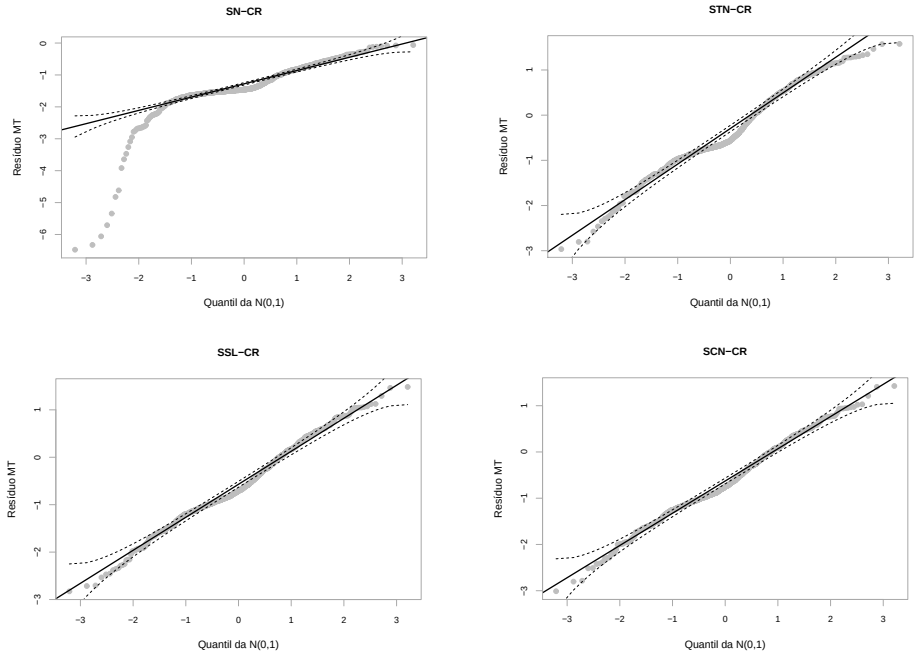

Figura 36 – **Aplicação**: *Conjunto de dados taxa salarial.* Envelopes dos resíduos MT para os modelos SSMN-CR.

Pela Figura 36, temos fortes evidências de que os modelos assimétricos com caudas pesadas são mais apropriados para este conjunto de dados do que o modelo skew-normal. Em particular, o modelo SCN-CR se ajusta melhor a este conjunto de dados.

5 CONCLUSÕES

Neste trabalho, propusemos modelos de regressão linear com respostas censuradas baseadas em distribuições de misturas de escala de normal assimétricas, denotados por SSMN-CR, como substituto da escolha convencional de distribuição normal (ou Normal simétrica) para os erros aleatórios e extensão dos SMN-CR estudados por Garay [15] para modelos lineares para dados censurados.

Para processo de estimação MV, desenvolvemos o algoritmo MCEM, onde foram obtidas soluções analíticas no passo M. Para explorar o desempenho de nossos modelos propostos e do algoritmo, desenvolvemos quatro estudos de simulação.

O primeiro estudo avaliou o desempenho dos estimadores MV dos parâmetros dos modelos SSMN-CR usando simulação de Monte Carlo. Este estudo de simulação foi projetado para observar mudanças nas estimativas variando os tamanhos amostrais e os níveis de censura, mostrando que as estimativas de nosso algoritmo MCEM proposto gozam das propriedades assintóticas dos estimadores de máxima verossimilhança.

O segundo estudo ilustrou a capacidade dos modelos censurados com distribuições de caudas pesadas, especificamente STN-CR e SSL-CR, de ajustar dados com uma estrutura gerada a partir de uma família de distribuições assimétricas diferente e também investigamos os efeitos na inferência paramétrica. Neste estudo sob diferentes níveis de censura, os critérios de seleção de modelos AIC e BIC favoreceram os modelos censurados baseados em distribuições de caudas pesadas.

O terceiro estudo mostrou a predição das observações censuradas. Com o intuito de investigar o comportamento da predição quando a distribuição do modelo é mal especificada, comparamos o desempenho da predição via algoritmo MCEM através de duas medidas de discrepância empírica, MAE (erro absoluto médio) e MSE (erro quadrático médio) e concluímos que os modelos baseados em distribuições de caudas pesadas geram valores preditivos mais próximos dos valores reais.

O quarto estudo avaliou a flexibilidade dos modelos SSMN-CR considerando a influência de uma única observação aberrante na estimativa MV assim

como também avaliou a influência de uma única observação extrema na predição de componentes censurados via algoritmo MCEM. Apresentamos resultados das medidas MMER'S para diferentes contaminações δ. As estimativas sob os modelos assimétricos com caudas pesadas (STN-CR e SSL-CR) foram menos afetadas pelas variações de δ em relação ao modelo skew-normal. Além disso, podemos afirmar que os modelos SSMN-CR apresentam robustez na presença de *outliers*.

Finalmente, também aplicamos os métodos propostos ao conjunto de dados de taxa salárial de Mroz [33], a fim de ilustrar como os procedimentos desenvolvidos podem ser usados para avaliar as premissas do modelo e obter estimativas de parâmetros robustas. Nossa proposta SSMN-CR com modelos de caudas pesadas, como os modelos STN-CR, SSL-CR e SCN-CR, apresenta melhores resultados que o modelo SN-CR. É interessante notar que o modelo SCN-CR ainda apresenta um melhor ajuste geral do que os modelos SSMN-CR.

Naturalmente os modelos de regressão para dados censurados têm sido amplamente estudados na literatura dada sua importância. Este trabalho procura apresentar uma alternativa de modelagem de dados censurados na presença de caudas pesadas e assimetria contribuindo para possíveis aplicações na área.

Na perspetiva de trabalhos futuros há por fazer a análise de diagnóstico assim como estender nossos resultados ao contexto multi-variado.

REFERÊNCIAS

[1] ANDREWS, D. R. E MALLOWS, C. L.; *Scale Mixture of Normal Distributions*. Journal of the Royal Statistical Society, B, 36, 99-102; (1974).

[2] ARELLANO V., R. B., L. M. CASTRO, G. GONZÁLEZ-FARÍAS, and K. A. MUÑOZ-GAJARDO. *Student-t censored regression model: properties and inference*. Statistical Methods & Applications 21, 453–473, (2012).

[3] BAI, Z. D., KRISHNAIAH,P. R., ZHAO, L. C.; *On Rates of Convergence of Efficient detection Criteria In Signal Processing with White Noise*. IEEE Trans. Info. Theory, 35:380-388; (1989).

[4] BARROS, M., M. GALEA, M. GONZÁLEZ, AND V. LEIVA.; *Influence diagnostics in the tobit censored response model*. Statistical Methods & Applications 19, 716–723, (2010).

[5] BRANCO M.D., DEY D.K.; *A General Class of Multivariate Skew Elliptical Distributions*. J. Multivariate Anal. 79, 99–113, (2001).

[6] COUVREUR, C.; *The EM algorithm: a guided tour*. In Proceedings of the 2d IIEE European Workshop on Computationaly Intensive Methos in Control and Signal Processing, (1996).

[7] DEMPSTER, A.; LAIRD, N.; RUBIN, D. B.; *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, Series B , 39 (1), 1–38, (1977).

[8] DEMPSTER, A. P., LAIRD, N. M. E RUBIN, D. B.; *Iteratively Reweighted Least Squares for Linear Regression when Erros are Normal/Independet Distributed*. Em P. R.Krishnaiah (Ed.), Multivariate Analysis V, 35-37. North-Holland, (1980).

[9] FAGUNDES R. A., DE SOUZA R. M. C., CYSNEIROS, F. J. A.; *Robust Regression With Application to Symbolic Interval Data*. Engineering Applications of Artificial Intelligence. 26: 564-573; (2013).

[10] FERREIRA C.S.; *Inferência e Diagnóstico em Modelos Assimétricos*; Tese de Doutorado, IME-USP. (2008).

[11] FERREIRA C.S., BOLFARINE H., LACHOS V. H.; *Likelihood-based inference for multivariate skew scale mixtures of normal distributions*; AStA Adv Stat Anal 100:421. (2016).

[12] FERREIRA C.S., BOLFARINE H., LACHOS V. H.; *Skew scale mixture of normal distributions: properties and estimation*; Statist Method; 8(2):154–171; (2011).

[13] FERREIRA, C. S., LACHOS, V. H. and BOLFARINE, H.; *Inference and diagnostics in skew scale mixtures of normal regression models*, Journal of Statistical Computation and Simulation, 85 (3): 517–537, (2015).

[14] FERREIRA C.S., LACHOS V.H., *Nonlinear regression models under skew scale mixtures of normal distributions*. Statistical Methodology 33, 131–146, (2016).

[15] GARAY, A. W. M.; *Modelos de Regressão Para dados Censurados Sob Distribuições Simétricas*, Tese de Doutorado, USP, (2014).

[16] GARAY, A. M., V. H. LACHOS, H. BOLFARINE, AND C. R. B. CABRAL; *Linear censored regression models with scale mixtures of normal distributions*. Statistical Papers, DOI: 10.1007/s00362–015–0696–9, (2015).

[17] GARAY A. M., LACHOS V. H., LIN T. I.; *Nonlinear censored regression models with heavy-tailed distributions*. Statistics and Its Interface 9: 281-293; (2016).

[18] GARAY, A. M., LACHOS, V. H., BOLFARINE, H., CABRAL, C. R. B.; *Linear censored regression models with scale mixtures of normal distributions*, Statistical Papers, 58, 247-278, (2017).

[19] GREENE W. H.; *Econometric Analysis*, 7th edn. Pearson, Harlow; (2012).

[20] LANGE, K. e SINSHEIMER, J. S.; *Normal/independent Distributions and Their Applications in Robust Regression*. Journal of Computational and Graphical Statistics.2:175–198; (1993).

[21] LANGE, K. L., LITTLE, J. A. e TAYLOR, M. G. J.; *Robust Modeling Using the t Distribution*. Journal of the American Statistical Association, 84, 881-896; (1989).

[22] LIN, T.-I. *Robust mixture modeling using multivariate skew t distributions*. Statistics and Computing 20 (3), 343–356, (2010).

[23] LITTLE, R. J. A.; *Robust Estimation of the Mean and Covariance Matrix form Data With Missing Values*. Applied Statistics, 37, 23-38, (1988).

[24] LOUIS, T. A.; *Finding the observed information matrix when using the EM algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), 226–233; (1982).

[25] LOUREDO, G. M. S.; *Estimação via EM e Diagnóstico em Modelos Misturas Assimétricas com Regressão*, Dissertação de Mestrado, UFJF; (2018).

[26] MASSUIA, M. B., A. M. GARAY, V. H. LACHOS, AND C. R. B. CABRAL; *Bayesian analysis of censored linear regression models with scale mixtures of skew-normal distributions.* Technical Report 3, Universidade Estadual de Campinas, (2015).

[27] MATOS, L. A.; *Estimation and diagnostics in multivariate models for censored data*, State University of Campinas: Doctoral thesis in Statistics, (2016).

[28] MATOS, L. A., CASTRO L. M., LACHOS V. H.; *Censored Mixed-effects Models For Irregulary Observed Repeated Measures With Applications to HIV Viral Loads*. Test, 25, 4, 627-653; (2016).

[29] MATTOS, T. B.; *Robust Estimation in Regression Models for Censored Data*; Dissertation Master in Statistics, UNICAMP, (2016).

[30] MCLACHLAN, G. J.; KRISHNAN, T.; *The EM Algorithm and Extensions*. Hoboken, John Wiley & Sons, (2001).

[31] MEILIJSON, I.; *A fast improvement to the EM algorithm on its own terms*. Journal of the Royal Statistical Society. Series B (Methodological), 127–138, (1989).

[32] MELENBERG B. AND SOEST A. V.; *Parametric and semi-parametric modeling of vacation expenditures*. Journal of applied Econometrics 11:59-76; (1996).

[33] MROZ, T. A.; *The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions*. Econometrica 55, 765–799, (1987).

[34] ORTEGA, E. M. M., H. BOLFARINE, AND G. A. PAULA.; *Influence diagnostics in generalized log-gamma regression models*. Computational Statistics & Data Analysis 42, 165–186; (2003).

[35] PAWITAN, Y.; *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford, Clarendon Press; (2001).

[36] RITTER, G.; *Robust Cluster Analysis and Variable Selection*. Boca Raton, CRC Press; (2015).

[37] SANTOS, C. C.; *Algortimo EM e Variações*. http://www.est.ufmg.br/ cristianocs/MetComput/Aula5.pdf. Notas de Aula, UFMG, (2018).

[38] STIGLER, S. M.; *The Epic Story of Maximum Likelihood*. Statistical Science, 598–620, (2007).

[39] VAIDA, F.; *Parameter convergence for EM and MM algorithms*. Statistica Sinica 15 (3), 831–840; (2005).

[40] WEI, G. C. G., TANNER, M. A.; *A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms*, Journal of the American Statistical Association, 85:411, 699-704, (1990).

[41] WITTE A.; *Estimating an economic model of crime with individual data*. Quaterly Journal of Economics 94:57-84; (1980).

[42] WU, C. F. J.; *On the Convergence Properties of the EM Algorithm*. The Annals of Statistics, 11 (1), 95–103; (1983).

APÊNDICE A – ESPERANÇAS CONDICIONAIS PARA DADOS CENSURADOS NO MODELO SSMN-CR

★ Para $\widehat{uy}_i$ definido como

$$\begin{aligned} \widehat{uy}_i &= E[U_i Y_i | Y_i > c, \widehat{\boldsymbol{\theta}}], \\ &= E[E(U_i Y_i | V_i) | Y_i > c, \widehat{\boldsymbol{\theta}}], \\ &= E[Y_i E(U_i | V_i) | Y_i > c, \widehat{\boldsymbol{\theta}}]. \end{aligned} \tag{A.1}$$

Note que as esperanças condicionais $(U_i|V_i)$ correspondem com as obtidas na Seção 2.1.3.1 e o valor de Y_i que depende da respectiva distribuição truncada SSMN será obtido mediante o método descrito na Seção 2.1.3.3.

⇒ Para o modelo STN-CR, teremos

$$\widehat{uy}_i = E[Y_i E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[y_{ij}\left(\frac{\nu+1}{\gamma+d_{ij}}\right)\right],$$

onde $d_{ij} = \dfrac{y_{ij}-\mu_i}{\sigma^2}$.

⇒ Para o modelo SSL-CR, teremos

$$\widehat{uy}_i = E[Y_i E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[y_{ij}\left(\frac{(2\nu+1)}{d}\frac{P_1(\nu+3/2, d_{ij}/2)}{P_1(\nu+1/2, d_{ij}/2)}\right)\right].$$

⇒ Para o modelo SCN-CR, teremos

$$\widehat{uy}_i = E[Y_i E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[y_{ij}\left(\frac{1-\nu+\nu\gamma^{3/2}\exp(1-\gamma)d_{ij}/2}{1-\nu+\nu\gamma^{1/2}\exp(1-\gamma)d_{ij}/2}\right)\right].$$

Seguindo o mesmo raciocínio, obtemos os demais valores esperados.

★ Para $\widehat{uy^2}_i$ definido como

$$\begin{aligned} \widehat{uy^2}_i &= E[U_i Y_i^2 | Y_i > c, \widehat{\boldsymbol{\theta}}], \\ &= E[E(U_i Y_i^2 | V_i) | Y_i > c, \widehat{\boldsymbol{\theta}}], \\ &= E[Y_i^2 E(U_i | V_i) | Y_i > c, \widehat{\boldsymbol{\theta}}]. \end{aligned} \tag{A.2}$$

⇒ No modelo STN-CR, temos que

$$\widehat{uy^2}_i = E[Y_i^2 E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[y_{ij}^2\left(\frac{\nu+1}{\gamma+d_{ij}}\right)\right].$$

$\Rightarrow$ No modelo SSL-CR, temos que

$$\widehat{uy^2}_i = E[Y_i^2 E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[y_{ij}^2\left(\frac{(2\nu+1)}{d}\frac{P_1(\nu+3/2, d_{ij}/2)}{P_1(\nu+1/2, d_{ij}/2)}\right)\right].$$

$\Rightarrow$ No modelo SCN-CR, temos que

$$\widehat{uy^2}_i = E[Y_i^2 E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[y_{ij}^2\left(\frac{1-\nu+\nu\gamma^{3/2}\exp{(1-\gamma)d_{ij}/2}}{1-\nu+\nu\gamma^{1/2}\exp{(1-\gamma)d_{ij}/2}}\right)\right].$$

★ Para $\widehat{u}_i$ definido como

$$\begin{aligned}\widehat{u}_i &= E[U_i|V_i, \widehat{\boldsymbol{\theta}}],\\ &= E[U_i|Y_i > c, \widehat{\boldsymbol{\theta}}],\\ &= E[E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}].\end{aligned} \tag{A.3}$$

$\Rightarrow$ No modelo STN-CR, temos que

$$\widehat{u}_i = E[E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[\frac{\nu+1}{\gamma+d_{ij}}\right].$$

$\Rightarrow$ No modelo SSL-CR, temos que

$$\widehat{u}_i = E[E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[\frac{(2\nu+1)}{d_{ij}}\frac{P_1(\nu+3/2, d_{ij}/2)}{P_1(\nu+1/2, d_{ij}/2)}\right].$$

$\Rightarrow$ No modelo SCN-CR, temos que

$$\widehat{u}_i = E[E(U_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[\frac{1-\nu+\nu\gamma^{3/2}\exp{(1-\gamma)d_{ij}/2}}{1-\nu+\nu\gamma^{1/2}\exp{(1-\gamma)d_{ij}/2}}\right].$$

★ Para $\widehat{ty}_i$, sabendo que $T_i|Y_i \sim HN(\lambda(y_i-\mu_i), \sigma^2)\ \mathbf{I}_{(0,\infty)}$ e sendo $\widehat{t}_i = E[T_i|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_i]$, teremos

$$\begin{aligned}\widehat{ty}_i &= E[T_iY_i|V_i, \widehat{\boldsymbol{\theta}}],\\ &= E[T_iY_i|Y_i > c, \widehat{\boldsymbol{\theta}}],\\ &= E[E(T_iY_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}],\\ &= E[Y_iE(T_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}]\end{aligned} \tag{A.4}$$

$\Rightarrow$ Portanto, nos modelos STN-CR, SSL-CR e SCN-CR, a esperança condicional é dada por

$$\widehat{ty}_i = E[Y_i E(T_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[y_{ij}\left(\lambda[y_{ij}-\mu_i] + \sigma W_{\Phi}\left(\frac{\lambda(y_{ij}-\mu_i)}{\sigma}\right)\right)\right],$$

onde $W_{\boldsymbol{\Phi}}(u) = \frac{\phi_1(u)}{\Phi(u)}$, $i = 1, \ldots, n$.

★ Para $\widehat{t}_i$ definido como

$$\begin{aligned} \widehat{t}_i &= E[T_i|V_i, \widehat{\boldsymbol{\theta}}], \\ &= E[T_i|Y_i > c, \widehat{\boldsymbol{\theta}}] \\ &= E[E(T_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}]. \end{aligned} \tag{A.5}$$

$\Rightarrow$ Analogamente ao explicado no item anterior, teremos

$$\widehat{t}_i = E[E(T_i|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[\left(\lambda[y_{ij}-\mu_i] + \sigma W_{\Phi}\left(\frac{\lambda(y_{ij}-\mu_i)}{\sigma}\right)\right)\right].$$

★ Para $\widehat{t^2}_i$, sabendo que $T_i|Y_i \sim HN(\lambda(y_i-\mu_i), \sigma^2)\ \mathbf{I}_{(0,\infty)}$ e sendo $\widehat{t}_i^2 = E[T_i^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_i]$, teremos

$$\begin{aligned} \widehat{t^2}_i &= E[T_i^2|V_i, \widehat{\boldsymbol{\theta}}], \\ &= E[T_i^2|Y_i > c, \widehat{\boldsymbol{\theta}}], \\ &= E[E(T_i^2|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}]. \end{aligned} \tag{A.6}$$

$\Rightarrow$ Analogamente ao explicado no item anterior, teremos

$$\widehat{t^2}_i = E[E(T_i^2|V_i)|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r}\left[\lambda^2[y_{ij}-\mu_i]^2 + \sigma^2 + \sigma\lambda[y_{ij}-\mu_i] W_{\Phi}\left(\frac{\lambda(y_{ij}-\mu_i)}{\sigma}\right)\right].$$

★ Para $\widehat{y}_i$ definido como

$$\begin{aligned} \widehat{y}_i &= E[Y_i|V_i, \widehat{\boldsymbol{\theta}}], \\ &= E[Y_i|Y_i > c, \widehat{\boldsymbol{\theta}}]. \end{aligned} \tag{A.7}$$

$\Rightarrow$ No modelo STN-CR, SSL-CR e SCN-CR, temos que

$$\widehat{y}_i = E[Y_i|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r} y_{ij}.$$

★ Para $\widehat{y^2}_i$, definido como

$$\begin{aligned} \widehat{y^2}_i &= E[Y_i^2|V_i, \widehat{\boldsymbol{\theta}}], \\ &= E[Y_i^2|Y_i > c, \widehat{\boldsymbol{\theta}}]. \end{aligned} \tag{A.8}$$

⇒ No modelo STN-CR, SSL-CR e SCN-CR, temos que

$$\widehat{y^2}_i = E[Y_i^2|Y_i > c, \widehat{\boldsymbol{\theta}}] = \frac{1}{r}\sum_{j=1}^{r} y_{ij}^2.$$