Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Departamento de Estatística

Nícolas Oliveira da Silva

Modelos de regressão com ponto de mudança contínuo para dados censurados sob distribuições simétricas

Juiz de Fora
2022

Nícolas Oliveira da Silva

Modelos de regressão com ponto de mudança contínuo para dados censurados sob distribuições simétricas

Monografia apresentada ao Departamento de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientadora: Camila Borelli Zeller

Juiz de Fora

2022

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

Oliveira da Silva, Nícolas.
Modelos de regressão com ponto de mudança contínuo para dados censurados sob distribuições simétricas / Nícolas Oliveira da Silva. – 2022.
62 f. : il.

Orientadora: Camila Borelli Zeller
Trabalho de Conclusão de Curso – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Departamento de Estatística, 2022.

1. Modelos de regressão. 2. Distribuições Simétricas. 3. Pontos de mudança. 4. Algoritmo EM. 5. Dados Censurados. I. Zeller, Camila Borelli, orient. II. Título.

Nícolas Oliveira da Silva

Modelos de regressão com ponto de mudança contínuo para dados censurados sob distribuições simétricas

Monografia apresentada ao Departamento de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do grau de Bacharel em Estatística.

Aprovada em 17 de fevereiro de 2022

BANCA EXAMINADORA

Prof[a]. Dra. Camila Borelli Zeller - Orientadora
Universidade Federal de Juiz de Fora

Professor Dr. Lupércio França Bessegato
Universidade Federal de Juiz de Fora

Professor Dr. Augusto Carvalho de Souza
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Gostaria de agradecer a todos que me apoiaram ao longo desta jornada.

Agradeço aos meus pais, Alberto e Lúcia, e ao meu irmão, Alberto, pelo suporte, carinho e ajuda. Em especial, minha avó Nair, pelo financiamento dos cafezinhos.

Agradeço aos meus amigos e familiares, por todo apoio e tornar essa caminhada mais leve e alegre.

Um agradecimento especial a Laura, minha companheira de todos momentos, que sempre esteve ao meu lado, me apoiando, dando suporte e também os devidos puxões de orelha.

Agradeço meus colegas de curso, pelo compartilhamento de conhecimento e bons momentos.

Gostaria de agradecer à minha orientadora Camila por me aceitar como orientando e ser uma excelente professora e pesquisadora. Agradeço também aos professores Lupércio e Augusto pelas suas contribuições para o trabalho e por aceitarem participar da banca.

Agradeço, também, à UFJF e ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC) do CNPq pelo financiamento do projeto.

RESUMO

Neste trabalho, apresentamos resultados recentes em uma área de pesquisa da Estatística com uma possibilidade enorme de aplicações, que são os modelos de regressão linear para dados censurados. A normalidade dos erros aleatórios é uma suposição rotineira em modelos lineares, que pode ser não realista. Assim, relaxamos a suposição de normalidade considerando que os erros aleatórios seguem uma distribuição mistura de escala normal, por exemplo, a distribuição t-Student. Esta distribuição inclui a distribuição normal como caso especial e fornece flexibilidade em capturar uma ampla variedade de comportamentos não normais, por simplesmente adicionar um parâmetro, denominado grau de liberdade, que controla a curtose. Além disso, consideramos o fato de que o mesmo modelo de regressão linear pode não ser válido para todo um conjunto de dados censurados. Isto é, o modelo pode se alterar após um ponto específico que, em geral, é desconhecido, e denominado ponto de mudança. Neste contexto, a estimação dos parâmetros do modelo será via algoritmo EM, e a seleção de modelos será realizada através dos critérios de informação (SIC e AIC). Dessa forma, o principal objetivo deste trabalho é estudar alguns aspectos de estimação em modelos de regressão linear com ponto de mudança para dados censurados sob distribuições simétricas. Finalmente, exemplos numéricos considerando dados simulados e reais são apresentados para ilustrar o modelo e os resultados inferenciais desenvolvidos.

Palavras-chave: Distribuições simétricas. Modelo de regressão linear. Ponto de mudança. Algoritmo EM. Dados censurados.

SUMÁRIO

1 INTRODUÇÂO

1.1 Motivação

Modelos de regressão lineares são técnicas bastante populares em pesquisa porque apresentam uma estrutura que permite aplicações em diversas áreas científicas, tais como, economia, agricultura, biologia, ciências médicas, entre outras.

Modelos de regressão lineares simples estudam a relação entre uma variável dependente (Y) e uma variável independente (x). Essa relação é representada pelo seguinte modelo estatístico, considerando uma sequência de observações $(x_i, Y_i), i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Usualmente, assume-se que o mesmo modelo de regressão linear é válido para todo um conjunto de dados, mas nem sempre isso é coerente, já que o modelo pode alterar seu comportamento após um ponto específico (tempo ou alguma região do domínio das variáveis preditoras, por exemplo) que, em geral, é desconhecido, e denominado ponto de mudança. O problema de ponto de mudança surgiu, inicialmente, no contexto de controle de qualidade, como demonstrado com os gráficos de Shewhart (1939) [23], e antes da introdução da hipótese de ponto de mudança associado com os modelos de regressão, pesquisadores enfrentavam dificuldades para estabelecer um modelo para alguns conjuntos de dados. Dessa forma, a identificação desse ponto desempenha um importante papel. Por exemplo, em um processo de produção contínuo, é esperado que a qualidade dos produtos se mantenha estável. Entretanto, por muitas razões, o processo pode falhar na produção de produtos com a mesma qualidade. Portanto, deseja-se encontrar se há um ponto em que a partir dele a qualidade do produto começa a se deteriorar.Veja Chen & Gupta (2011) [9] para mais detalhes.

No contexto de modelos de regressão linear, considerando uma sequência de observações $(x_i, Y_i), i = 1, \ldots, n$, o modelo de regressão com ponto de mudança pode ser escrito como

$$Y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, \forall x_i \leq \gamma \\ \alpha_2 + \beta_2 x_i + \varepsilon_i, \forall x_i > \gamma \end{cases}, \tag{1.1}$$

onde $\alpha_1, \alpha_2, \beta_1, \beta_2$ e γ são parâmetros desconhecidos, tal que γ é o ponto de mudança, e os erros $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Tais modelos são divididos em dois tipos. Um tipo em que o modelo é assumido como contínuo no ponto de mudança (também conhecido como mudança gradual ou sem descontinuidade) e outro onde não é (também conhecido como mudança abrupta ou com descontinuidade). A inferência teórica é completamente diferente para cada tipo de modelo. Estes modelos podem ser estendidos para o caso de múltiplos pontos de mudança. Muggeo (2003) [19] adverte que o uso de muitos pontos de mudança é questionável para a

maioria das aplicações práticas e que nestes casos, a modelagem de regressão não-linear ou não paramétrica poderia ser mais apropriada.

Neste trabalho, em contraste com o modelo (1.1), uma restrição de continuidade no ponto de mudança é assumido. Assim, ambos os modelos devem prever o mesmo valor médio no ponto de mudança, o que resulta em uma transição contínua dos dois modelos. Além disso, presume-se que a variável explicativa seja classificada em ordem crescente, $x_i \leq x_{i+1}, i = 1, \ldots, n-1$. Neste contexto, a localização do ponto de mudança não é mais restrita a um x_i observado. Em vez disso, pode ser qualquer valor dentro do intervalo $[a, b]$, onde se encontram os seguintes pontos $a = x_{(1)} = min\{x_1, \ldots, x_n\}$ e $b = x_{(n)} = max\{x_1, \ldots, x_n\}$.

O modelo de regressão linear com um ponto de mudança contínuo pode então ser escrito como

$$Y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, \forall a \leq x_i \leq \gamma \\ \alpha_2 + \beta_2 x_i + \varepsilon_i, \forall \gamma < x_i \leq b \end{cases}, \tag{1.2}$$

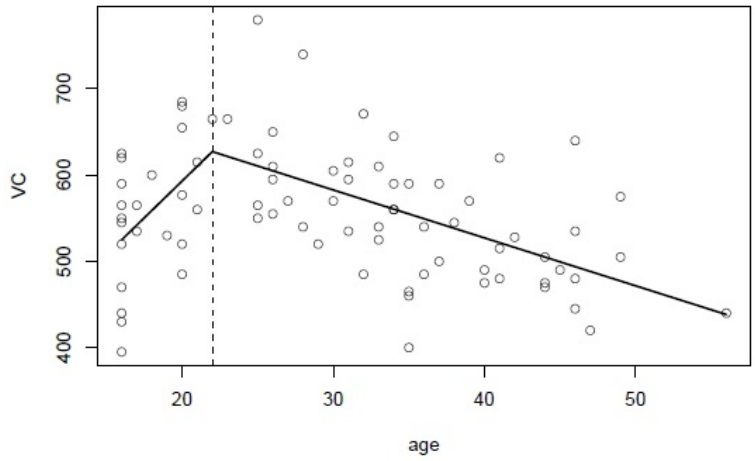
com o contraste de continuidade

$$\alpha_1 + \beta_1 \gamma = \alpha_2 + \beta_2 \gamma.$$

Hofrichter (2007) [13] ilustra o modelo (1.2) com dados provenientes de um estudo de saúde, realizado na Estíria (Aústria). O interesse era avaliar a dependência entre a capacidade pulmonar (VC) e a idade da pessoa. Neste estudo, 79 pessoas do sexo masculino com idades entre 16 anos e 56 anos foram examinadas. Os dados mostram uma tendência clara e crescente para pessoas jovens e uma tendência decrescente para idosos; veja a Figura 1. Logo, teremos duas retas distintas, uma crescente e a outra decrescente, o que indica claramente a existência de um ponto de mudança entre estas duas retas lineares. Assim, os dois modelos devem prever o mesmo valor médio no ponto de mudança, e uma restrição de continuidade é obrigatória. O modelo ajustado com um ponto de mudança contínuo estimado na idade de $\widehat{\gamma} = 22$ está plotado na Figura 1.

A importância do ponto de mudança pode ser notada pelo grande número de artigos que vem sendo publicados em vários periódicos. No mês de Julho de 2020, foi publicado no Statistical Paper uma edição especial sobre pontos de mudança. No artigo do editor Sofronov et al. (2020) [25], é citada a importância do uso dos modelos com ponto de mudança em diversas áreas como biologia, medicina, finanças, economia, entre outras. Esta edição também incluiu os artigos apresentados no Workshop: "Change Point Detection Limit Theorems, Algorithms and Aplicantions in Life Sciences"(Teoremas, algoritmos e aplicações de detecção de limites dos pontos de mudança nas ciências da vida) que relataram diferentes formas de abordar o problema de ponto de mudança em diversos cenários. Por exemplo, pontos de mudança nas sequências de DNA podem indicar a localização dos genes e outros elementos funcionais. Em epidemiologia, as detecções

Figura 1 – Modelo com ponto de mudança contínuo ajustado com a estimativa de ponto de mudança na idade $\widehat{\gamma} = 22$.



oportunas de números crescentes de infecções é uma questão muito importante. Nos cuidados intensivos, os métodos de detecção de ponto de mudança podem contribuir para melhores sistemas de alarme. Atualmente, existem estudos, no contexto de inferência não-paramétrica, desenvolvidos no Instituto de Matemática e Estatística, IME/USP, onde se indentificam os pontos de mudança na curva de casos de Covid-19 e permitem detectar e prever alterações na evolução do contágio da doença. O método indica onde há mudanças na taxa de crescimento ou decrescimento dos dados, o que permite avaliar o efeito de algumas medidas adotadas para conter o número de casos da doença, como por exemplo, a quarentena ou o lockdown. Para mais detalhes, veja Sousa et al. (2021) [26].

Adicionalmente, muitos trabalhos dissertam acerca de diferentes tipos de estruturas de ponto de mudança com descontinuidade em relação à média, à variância e à média e variância simultaneamente, de uma sequência de variáveis aleatórias independentes e normalmente distribuídas. Por exemplo, os problemas de ponto de mudança com descontinuidade na média e/ou na variância, no contexto de normalidade dos dados, já foram examinados por Chen Gupta (1996, 1997, 1999, 2003, 2011) [5], [6], [7], [8] e [9].

Agora vamos voltar nossa atenção para além dos modelos gaussianos, e estudar outros modelos importantes. Inspirados por todos os trabalhos citados acima, propomos os modelos de regressão com ponto de mudança contínuo quando a variável resposta segue uma distribuição na classe misturas de escala normal (SMN),por exemplo, t-Student, e considerando a possibilidade de observações censuradas. A definição e o detalhamento dessa classe de distribuições serão dados no Capítulo 2, na Seção 2.3.

Uma fonte especial de dificuldades na análise estatística é a possibilidade de que os indivíduos (ou unidades experimentais) podem não ter uma observação completa da variável resposta. Tal observação incompleta (ou parcial) da variável resposta é chamada de

censura. As censuras pode ocorrem por uma variedade de razões, incluindo limitações de equipamentos de medição, desenho do experimento e não ocorrência do evento de interesse até o final do estudo. Ressalta-se que, mesmo censurado, todos os resultados devem ser usados na análise estatística. A omissão da censura pode levar à conclusões viciadas. Neste trabalho, utilizamos censura aleatória à direita e à esquerda. A definição mais detalhada de censura é apresentada no Capítulo 2, na Seção 2.1. Existem apenas alguns artigos lidando com a detecção de ponto de mudança no contexto de dados censurados; veja, por exemplo, Husková Neuhaus (2004) [14] e Wang et al. (2007) [27].

Finalmente, no próximo capítulo, apresentamos uma revisão dos principais conceitos que serão tratados neste trabalho, incluindo censura e a classe de distribuições SMN.

Note que todas as siglas que serão utilizadas, neste trabalho, estarão em inglês para uma maior facilidade do leitor em uma busca bibliográfica.

1.2 Objetivos

Neste projeto, apresentamos alguns resultados adicionais para o problema de ponto de mudança sem descontinuidade, no contexto de simetria, em particular para o modelo de regressão censurado sob a classe de distribuições SMN.

A respeito da inferência sobre modelos com ponto de mudança, devem ser considerados os seguintes aspectos: determinar a existência do ponto de mudança, localizar a posição deste ponto, estimar todos os parâmetros de interesse do modelo e realizar análises explicativas. Se a localização do ponto de mudança é conhecida, então a estimação dos parâmetros do modelo é direta, caso contrário, um parâmetro extra (ponto de mudança, γ) deve ser estimado. Inspirados pelo trabalho de Young (2014) [28], o enfoque do trabalho será voltado para a estimação dos parâmetros do modelo via algoritmo tipo-EM (Dempster et al, 1977)[11].

Finalmente, verificaremos a adequação dos modelos baseados nas distribuições simétricas aos dados inspecionando dois critérios de informação: o critério de informação de Akaike (AIC) e o critério bayesiano de informação de Schwarz (BIC); veja Akaike (1974) [1] e Schwarz et al. (1978) [24] para mais detalhes. Cada um desses critérios baseia-se numa penalização da verossimilhança na medida em que o modelo se torna mais complexo, isto é, modelos com um grande número de parâmetros. O modelo que apresentar o menor valor do critério de informação será o modelo selecionado.

Os resultados obtidos serão aplicados em conjuntos de dados reais e/ou simulados. Além disso, será utilizado o programa estatístico R para as programações das metodologias propostas no modelo estudado.

Podemos então relacionar os seguintes objetivos específicos: (i) estimar o ponto de mudança; (ii) implementar e avaliar o algoritmo EM proposto computacionalmente; e (iii)

aplicar esses resultados para analisar dados reais.

2 PRINCIPAIS CONCEITOS

Neste capítulo, apresentamos uma breve revisão do conceito de censura que será tratado neste trabalho. É importante esclarecer a diferença entre as duas principais causas de dados incompletos: a censura e truncamento. Além disso, definimos as distribuições misturas de escala normal (SMN) e destacamos algumas propriedades, tais como momentos, representação estocástica, formas quadráticas e casos particulares que serão úteis no desenvolvimento deste trabalho.

2.1 Censura

Censura é a observação parcial (incompleta) da resposta, por exemplo, o acompanhamento de um paciente foi interrompido por algum motivo, seja porque mudou de cidade, o estudo terminou para a análise dos dados ou o paciente morreu de causa diferente da estudada.

Temos várias formas de censura que serão descritas e exemplificadas abaixo.

Censura tipo I, onde o estudo será concluído após um período de tempo pré-determinado, assim temos um tempo do estudo não aleatório.

Censura tipo II, onde o estudo será concluído após um certo número de acontecimentos do evento de interesse, o que ocasiona em um tempo do estudo aleatório.

Censura aleatória, onde se perde o contato com o paciente, seja por morte por outro motivo diferente do estudado ou interrupção do estudo sem ter ocorrido desfecho.

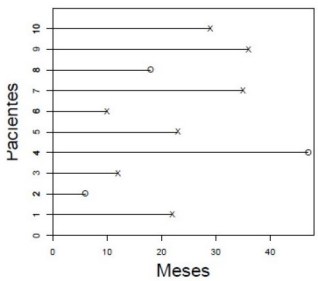
Além disso, temos censuras à direita, à esquerda, intervalar, informativa e não informativa. A Figura 2 apresenta o formato clássico de um conjunto de dados com observações censuradas à direita.

Figura 2 – Exemplo: Dados de 10 pacientes em diálise, onde T_i é o tempo de observação (em meses) e δ_i é a variável indicadora de ocorrência de censura à direita. (Carvalho et al., 2011) [3].

Dados de 10 pacientes Notação Clássica: T_i, δ_i

Paciente (i)	Tempo (T_i)	Status (δ_i)
1	22	1
2	6	0
3	12	1
4	43	0
5	23	1
6	10	1
7	35	1
8	18	0
9	36	1
10	29	1

Graficamente



X indica ocorrência do evento e **O** corresponde à presença de censura.

2.1.1 Censura à direita

- Tempo de ocorrência do desfecho está à direita do tempo registrado.
- Sabe-se que o tempo entre o início do estudo e o evento é maior do que o tempo observado.
- Aproveita-se a informação do tempo durante o qual a pessoa esteve sob observação sem que ocorresse o evento.

Considere, por exemplo, um estudo para analisar o tempo entre o diagnóstico de Aids e o óbito, 193 pacientes foram acompanhados em um ambulatório especializado de 1986 a 2000. Durante esse período, foram observados 92 óbitos. Sabemos que até a data de término do estudo, em dezembro de 2000, ainda permaneciam vivos 101 pacientes. Dizemos, então, que ocorreram 92 eventos e 101 censuras à direita. Veja Carvalho et al. (2011) [3] para mais detalhes e Figura 3 (neste texto).

2.1.2 Censura à esquerda

- Ocorre quando não se conhece o momento da ocorrência do desfecho, mas sabemos que ocorreu antes do tempo observado.
- O evento de interesse aconteceu anterior ao início do estudo.
- Tempo observado é maior que o tempo de falha.

Considere, por exemplo, um estudo para investigar os fatores associados à infecção por leptospirose em uma comunidade de baixa renda, recém-criada (Carvalho et al., 2011) [3]. Os participantes foram incluídos assim que mudaram para a área. Uma vez por ano, os pesquisadores coletam sangue dos participantes para verificar se houve soroconversão. Alguns participantes logo na primeira coleta já estavam soroposotivos. Ou seja, a soroconversão pode ter acontecido em qualquer momento entre o dia da mudança para a comunidade e alguns dias antes do primeiro exame. Só podemos afirmar que o tempo para a soroconversão, que define o tempo de sobrevivência é menor que o tempo para o primeiro exame, como indica o segundo quadro da Figura 3. É importante observar que para haver censura à esquerda, o marcador de tempo de início não pode estar relacionado ao marcador de evento.

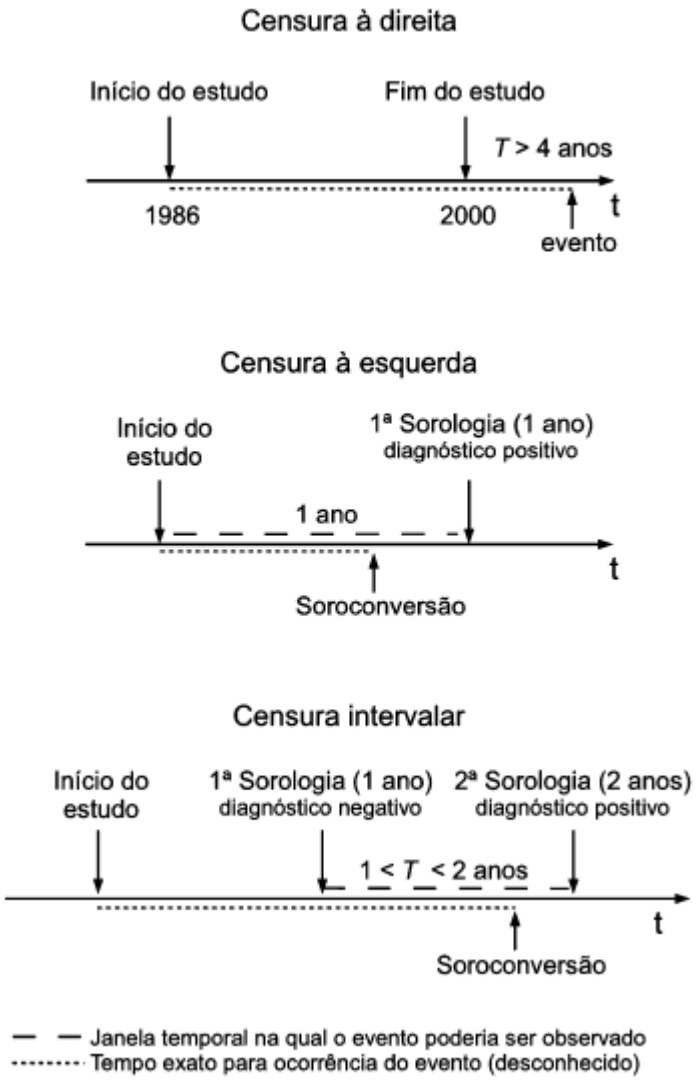
2.1.3 Censura intervalar

- Tipo mais geral de censura, por exemplo, um estudo em que os pacientes são acompanhados em visitas periódicas e conhece-se apenas que o evento de interesse ocorreu em um certo intervalo de tempo.

- Ocorrência do evento entre tempos conhecidos.
- O tempo até a recorrência é maior do que a data do exame negativo e menor que o primeiro exame positivo.

Considere o estudo citado na Seção 2.1.2 e adicionalmente, temos o último quadro da Figura 3 que ilustra outro participante do estudo, no qual a soroconversão aconteceu antes do segundo exame. Então, sabemos que o tempo exato de sobrevivência é maior que o tempo até o primeiro exame e menor do que o tempo até o segundo exame, isto é, o momento em que a soroconversão ocorreu certamente se situa entre o primeiro e o segundo exame.

Figura 3 – Representação gráfica de censura à direita do tempo até o óbito por Aids (Seção 2.1.1) e de censuras à esquerda e intervalar no exemplo do tempo até a soroconversão para leptospirose (Seções 2.1.2 - 2.1.3). A linha pontilhada é o tempo de sobrevivência exato (não observado). A linha tracejada representa a janela temporal onde o evento poderia ocorrer (Carvalho et al., 2011) [3].



2.1.4 Censura informativas e não informativas

- Censura não informativa é quando o motivo da perda de informação não está relacionada com o desfecho do estudo. Por exemplo, data final do estudo não foi definida em função do desfecho.
- Censura informativa é quando o motivo da perda de informação está relacionada com o desfecho do estudo. Por exemplo, abandono do tratamento devido à piora do paciente ou óbito não conhecido por falha no acompanhamento.

Segundo Carvalho et al. (2011) [3], as censuras informativas devem ser evitadas, pois induzem um viés de seleção. Pode-se evitar esse viés predefinindo todos os passos na busca de pacientes e analisando as causas da censura.

2.2 Truncamento

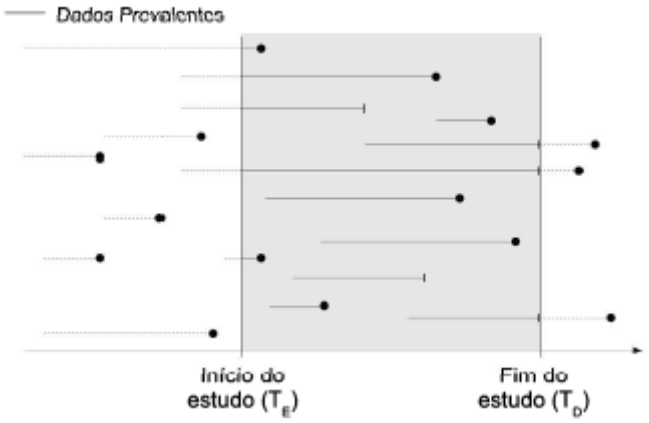
Truncamento ocorre quando os indivíduos, que naturalmente pertenceriam à população estudada, são excluídos do estudo. Em resumo, podemos concluir que no caso da censura, há alguma limitação imposta à mensuração da variável dependente (variável de interesse), impedindo que observemos valores inferiores (censura à esquerda) ou superiores (censura à direita) da variável dependente. As observações permanecem na amostra. Já no caso do truncamento, há alguma limitação imposta à mensuração da variável dependente, impedindo que observemos observações com valores inferiores (truncamento à esquerda) ou superiores (truncamento à direita) da variável dependente. Assim, essas observações não farão parte da amostra

2.2.1 Truncamento à esquerda

- Inclui somente observações em que o desfecho ocorreu após o limite inferior da janela temporal de observação.
- Só ocorre quando a perda de informação está relacionada a indivíduos que foram excluídos porque já tinham experimentado o evento antes do início do estudo e não puderam ser observados.

Considere, por exemplo, o estudo de sobrevivência de pacientes em hemodiálise, os dados referem-se somente àqueles pacientes em tratamento em janeiro de 1998 ou aos que iniciaram o tratamento numa data posterior. Nesse caso, os pacientes diagnosticados antes de janeiro de 1998 e que morream antes daquela data não são incluídos no estudo, caracteziando o que denomina truncamento à esquerda. Para mais detalhes, veja Carvalho et al. (2011) [3]. As informações dos indivíduos que já sofreram o desfecho antes do

Figura 4 – Representação gráfica de truncamento à esquerda. Trajetórias em pontilhada à esqueda do início do estudo não são incluídas na análise e as linhas contínuas são os dados prevalentes a analisar (Carvalho et al., 2011) [3].



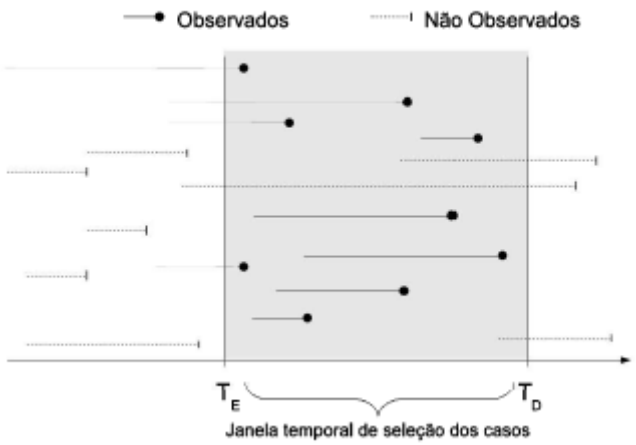
início do estudo não podem ser incluídas no estudo, embora se esteja considerando os sobreviventes. Os pacientes sobreviventes que iniciaram a diálise antes do início do estudo só terão computados os tempos após a data de início do estudo. Observe na Figura 4 a representação gráfica de truncamento à esquerda.

2.2.2 Truncamento à direita

- Ocorre quando o critério de seleção inclui somente os indivíduos que sofreram o evento.
- Data de ocorrência do evento é sempre menor que o limite superior da janela temporal.
- Comum em estudos de sobrevivência que partem do dado do óbito para selecionar as observações.

Tome como exemplo os dados do estudo dos pacientes com Aids citado na Seção 2.1.1. Neste cenário, a amostra é definida pelos eventos na janela temporal, isso significa, por um lado, que não haverá censura à direita, mas por outro lado, que indivíduos que não sofreram o evento, mesmo tendo os fatores de risco presentes, não serão incluídos no estudo. Observe na Figura 5 a representação gráfica de truncamento à direita. Segundo Carvalho et al. (2011) [3], em estudos de doenças com longa duração, o truncamento a direita pode fazer com que o risco de ocorrência do evento seja superestimado.

Figura 5 – Representação gráfica de truncamento à direita. Trajetórias em pontilhada não são observadas, mesmo que cruzem a janela temporal, pois o evento não ocorreu no intervalo definido. Nunca há censura à direita (Carvalho et al., 2011) [3].



2.3 Distribuições misturas de escala normal

Uma variável aleatória Y segue uma distribuição mistura de escala normal com parâmetro de locação $\mu \in \mathbb{R}$ e parâmetro de dispersão σ^2 (positivo) se a sua função de densidade de probabilidade (f.d.p.) é dada por

$$f(y) = \int_0^\infty \phi(y; \mu, \kappa(u)\sigma^2) dH(\mu; \nu),$$

onde $\phi(.; \mu, \sigma^2)$ denota a f.d.p. de uma distribuição normal com média μ e variância σ^2, $\kappa(.)$ é uma função de pesos e U é uma variável aleatória positiva com função de distribuição acumulada (f.d.a.) $H(u; \boldsymbol{\nu})$, tal que $\boldsymbol{\nu}$ é um escalar ou vetor de parâmetros da distribuição de U que controla as caudas das distribuições. Note que o termo $\phi(.; \mu, \sigma^2)$ depende da distância de Mahalanobis

$$d = \frac{(y-\mu)^2}{\sigma^2}, \tag{2.1}$$

útil na identificação de observações aberrantes; veja, por exemplo, Pinheiro et al. (2001) [20]. Esta distribuição será denotada por $SMN(\mu, \sigma^2, \boldsymbol{\nu})$.

A seguir será apresentada a representação estocástica de uma variável aleatória que segue distribuição mistura de escala normal. Tal representação é de suma importância no decorrer deste trabalho, visto que inúmeros resultados podem ser derivados a partir dessa caracterização.

2.3.1 Representação estocástica

Seja Y uma variável aleatória com distribuição $SMN(\mu, \sigma^2, \boldsymbol{\nu})$. Dessa forma, Y pode ser representada estocasticamente como

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

onde Z é uma variável aleatória normal com média zero e variância σ^2,e U é uma variável aleatória positiva com f.d.p. $h(u;\boldsymbol{\nu})$, independente de Z. A representação estocástica dada em (2.2), além de facilitar a implementação do algoritmo EM (Dempster et al., 1977) [11], útil na obtenção dos estimadores de máxima verossimilhança dos parâmetros de interesse, pode ser usada também para derivar muitas propriedades da distribuição de Y. Um exemplo, seria a forma quadrática da distância de Mahalanobis d, definida em (2.1), e que será descrita na Seção 2.4. Além disso, a partir de (2.2) a distribuição mistura de escala normal pode ser reescrita hierarquicamente como apresentada na proposição abaixo.

Proposição 2.3.1. *Sabendo que Y segue uma distribuição mistura de escala normal $SMN(\mu, \sigma^2, \boldsymbol{\nu})$, então este modelo pode ser reescrito hierarquicamente como*

$$Y \mid U = u \ \sim \ N(\mu, k(u)\sigma^2) \quad e \quad U \ \sim \ h(u;\boldsymbol{\nu}).$$

Demonstração. A prova é realizada diretamente a partir da representação estocástica dada em (2.2). □

Tal representação hierárquica será útil para obter algumas propriedades da distribução mistura de escala normal, tais como média e variância, e também realizar inferência estatística nos modelos de regressão linear, como será visto no próximo capítulo. Sendo assim, seguem abaixo algumas propriedades interessantes das distribuções misturas de escala normal. O valor esperado e a variância são dados por

(P1) $E[Y] = \mu$, se $E[\kappa^{1/2}(U)] < \infty$.

(P2) $V[Y] = E[\kappa(U)]\sigma^2$, se $E[\kappa(U)] < \infty$.

As distribuições mistura de escala normal são constituídas por famílias paramétricas de distribuições probabilísticas que preservam a estrutura simétrica das distribuições normais. A família de distribuções normais é um elemento particular desta classe. Outras famílias conhecidas que compõem esta classe de distribuições são: t-Student, slash e normal contaminada, por exemplo. Para uma discussão mais detalhada quanto às distribuções de mistura de escala normal veja, por exemplo, Andrews e Mallows (1974) [2] e Lange e Sinsheimer (1993) [16].

2.3.2 Casos particulares

Casos particulares da classe de distribuições SMN estão detalhados na Tabela 1, incluindo para cada caso as funções $\kappa(.)$ e U que as caracterizam.

A seguir serão descritas com mais detalhes as distribuições normal, t-Student e slash, três importantes membros da classe SMN e que serão utilizadas neste trabalho.

Tabela 1 – Algumas distribuições que compõem a classe SMN.

Distribuição	Notação	$\kappa(.)$	U	
Normal	$N(\mu, \sigma^2)$	1	$Degenerada$	
t-Student	$t(\mu, \sigma^2, \nu)$	$1/u$	$Gamma(\frac{\nu}{2}, \frac{\nu}{2})$	(1)
Slash	$SL(\mu, \sigma^2, \nu)$	$1/u$	$Beta(\nu, 1)$	
Normal Contaminada	$CN(\mu, \sigma^2, \nu, \gamma)$	$1/u$	$Discreta$	(2)

(1) Considerando $Gamma(a, b)$ com média $\frac{a}{b}$.

(2) $Discreta\ com\ f.d.p.\ h(u; \boldsymbol{\nu}) = \nu \mathbb{I}_{(u=\gamma)} + (1-\nu)\mathbb{I}_{(u=1)},\ 0 \leq \nu \leq 1,\ 0 < \gamma \leq 1.$

2.3.2.1 Distribuição normal

A normal é a distribuição pertencente à classe SMN mais utilizada devido a todo o desenvolvimento teórico e aplicado estabelecido no decorrer dos anos. Se $Y \sim N(\mu, \sigma^2)$, então a f.d.p é da forma

$$f(y) = \phi(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{d}{2}\}, \qquad y \in \mathbb{R},$$

onde seu valor esperado e variância são, respectivamente,

$$E(Y) = \mu \qquad \text{e} \qquad Var(Y) = \sigma^2.$$

2.3.2.2 Distribuição t-Student

A variável aleatória Y tem distribuição t-Student com ν graus de liberdade, ou seja, $Y \sim t(\mu, \sigma^2, \nu)$, se sua f.d.p é dada por

$$f(y) = \frac{\Gamma(\frac{1+\nu}{2})}{\Gamma(\frac{\nu}{2})\pi^{1/2}} \nu^{-1/2} \sigma^{-1} \left(1 + \frac{d}{\nu}\right)^{-(1+\nu)/2}, \qquad y \in \mathbb{R}.$$

Dessa forma, temos que

$$E(Y) = \mu \qquad \text{e} \qquad Var(Y) = \sigma^2 \frac{\nu}{\nu - 2},\ \nu > 2.$$

2.3.3 Distribuição slash

A variável aleatória Y tem distribuição slash com ν graus de liberdade, ou seja, $Y \sim SL(\mu, \sigma^2, \nu)$, se sua f.d.p é dada por

$$f(y) = \frac{\nu}{\sqrt{2\pi\sigma^2}} \int_0^1 \qquad y \in \mathbb{R},$$

onde seu valor esperado e variância são, respectivamente,

$$E(Y) = \mu \qquad \text{e} \qquad Var(Y) = \sigma^2 \frac{\nu}{\nu - 1},\ \nu > 1.$$

Em Contreras (2014) [10], a Figura 2.4, na página 21, apresenta graficamente o comportamento das distribuições SMN para diversos valores de μ, σ^2 e ν. É importante ressaltar que a convergência para a distribuição normal nos casos t-Student e slash ocorre quando $\nu \to \infty$.

2.4 Distância de Mahalanobis

Em seguida, descrevemos algumas propriedades da distância de Mahalanobis $d = \dfrac{(Y-\mu)^2}{\sigma^2}$. Mais detalhes sobre propriedades das formas quadráticas podem ser encontradas em Lange & Sinsheimer (1993) [16], por exemplo. Dessa forma, temos que $d \sim \chi^2_1$ para o caso normal e $d \sim F(1,\nu)$ para o caso t-Student. Este resultado é interessante, pois permite avaliar os modelos estatísticos na prática. Substituindo as estimativas de máxima verossimilhança de μ e σ^2 na distância de Mahalanobis d, podemos avaliar os ajustes dos modelos por meio da construção de envelopes. Além disso, mediante gráficos da distância de Mahalanobis e considerando como marca de referência o quantil υ da distribuição da forma quadrática d, podemos identificar "outliers". Por exemplo, para o caso normal, temos que $\upsilon = \chi^2_1(\xi)$, onde $0 < \xi < 1$. No próximo capítulo, estudamos os modelos de regressão lineares sob a classe de distribuições SMN, em particular, daremos destaque aos casos particulares normal e t-Student.

3 MODELOS DE REGRESSÃO LINEARES

Neste capítulo, antes de iniciar os estudos envolvendo os modelos de regressão lineares, apresentamos uma breve descrição do método de estimação de máxima verossimilhança. Neste contexto, utilizamos Casela e Berger (2010) [4] e Montgomery et al. (2012) [18] como principais referências.

3.1 Método de máxima verossimilhança

Seja $\boldsymbol{\theta}$ um vetor de parâmetros de dimensão $p \times 1$. A função de verossimilhança de n variáveis aleatórias $Y_1, \ldots, Y_n$ é definida como sendo a densidade conjunta das n variáveis aleatórias, digamos $f(\mathbf{y}; \boldsymbol{\theta}) = f(y_1, \ldots, y_n; \boldsymbol{\theta})$, que é considerada uma função de $\boldsymbol{\theta}$ fixados os valores observados. Em particular, se $Y_1, \ldots, Y_n$ for uma amostra aleatória, então a função de verossimilhança será $f(y_1; \boldsymbol{\theta}) f(y_2; \boldsymbol{\theta}) \cdots f(y_n; \boldsymbol{\theta})$. Denotamos a função de verossimilhança por

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta})$$

e a função de log-verossimilhança é definida como

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}).$$

Uma vez que a função logarítmica é monótona, o máximo de $\ell(\boldsymbol{\theta}; \mathbf{y})$ coincidirá com o máximo de $L(\boldsymbol{\theta}; \mathbf{y})$, ou seja, a estimativa de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$ é o valor de $\boldsymbol{\theta} \in \Theta$ que maximiza $\ell(\boldsymbol{\theta})$, isto é,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\ \ell(\boldsymbol{\theta}; \mathbf{y}).$$

Para verificar se estamos em um ponto de máximo, temos que avaliar a matriz Hessiana da função de log-verossimilhança, denotada por H. A condição é que a matriz

$$H = \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}}$$

seja negativa definida, isto é, $\mathbf{z}^\top H \mathbf{z} < 0, \forall\ \mathbf{z} \neq \mathbf{0}$. Note que a matriz Hessiana é simétrica, isto é, $H^\top = H$. Para verificarmos que H é negativa definida, tomaremos os determinantes das submatrizes principais (menores principais), denotados por $|H_1|, |H_2|, \ldots, |H_p|$, em que $|H_j|$ denota o determinante da j-ésima submatriz quadrada. Note que se os determinantes dos menores principais de H tiverem sinais alternados, então a matriz H é dita negativa definida.

3.2 Modelo de regressão linear normal

Nesta seção, consideramos os modelos de regressão linear, onde as observações seguem distribuição normal. Em geral, o modelo de regessão linear normal é definido como

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, ..., n, \tag{3.1}$$

onde $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ e consequentemente, $Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. Assim, temos que $E(Y_i) = \beta_0 + \beta_1 x_i$.

Para o modelo de regressão linear normal, a função de log-verossimilhança será dada por:

$$\ell(\beta_0, \beta_1, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

onde $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{x}_i^\top = (1, x_i)$ corresponde a i-ésima linha da matriz $\mathbf{X}_{(n\times 2)}$ e $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. Note que

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top\boldsymbol{\beta}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

Então, considerando $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)^\top$, temos os seguintes estimadores

$$\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}} = \mathbf{0} \implies \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y \ \text{ e } \ \frac{\partial\ell(\boldsymbol{\theta})}{\partial\sigma^2} = 0 \implies \widehat{\sigma^2} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Além disso, temos que

$$\frac{\partial^2\ell(\boldsymbol{\theta};\mathbf{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{\widehat{\sigma^2}}\mathbf{X}^\top\mathbf{X}, \quad \frac{\partial^2\ell(\boldsymbol{\theta};\mathbf{y})}{\partial(\sigma^2)^2}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{n}{2(\hat{\sigma}^2)^2} \quad \text{e} \quad \frac{\partial^2\ell(\boldsymbol{\theta};\mathbf{y})}{\partial\sigma^2\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}.$$

Portanto, a matriz H é negativa definida.

3.3 Modelo de regressão linear t-Student

Nesta seção, consideramos o modelo de regressão linear, definido em (3.1), porém as observações seguem distribuição t-Student. Dessa forma, seguindo, por exemplo, Lange et al. (1989) [15], vamos substituir a usual suposição de normalidade para os erros, pela seguinte suposição mais flexível

$$\varepsilon_i \overset{iid}{\sim} t(0, \sigma^2, \nu),\ i = 1, ..., n,$$

e, consequentemente, $Y_i \overset{ind}{\sim} t(\beta_0 + \beta_1 x_i, \sigma^2, \nu)$. Assim, temos que $E(Y_i) = \beta_0 + \beta_1 x_i$.

A função de log-verossimilhança $\ell(\boldsymbol{\theta})$ para $\boldsymbol{\theta}$=$(\boldsymbol{\beta}^\top, \sigma^2)^\top$, em que $\boldsymbol{\beta}$= $(\beta_0, \beta_1)^\top$, onde assume-se que o parâmetro ν (associado com a variável de mistura U) é fixo (conhecido), é definida por

$$\ell(\boldsymbol{\theta}) = n\ \log\ c(\nu) - \frac{n}{2}\log\ \sigma^2 - \left(\frac{\nu+1}{2}\right)\sum_{i=1}^{n}\log\left\{1 + \frac{d_i(\boldsymbol{\beta}, \sigma^2)}{\nu}\right\}, \tag{3.2}$$

onde

$$d_i = d_i(\boldsymbol{\beta}, \sigma^2) = \frac{\left(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\sigma^2}, i = 1, ..., n,$$

é a distância de Mahalanobis, $\mathbf{x}_i^\top$ corresponde a i-ésima linha da matriz $\mathbf{X}_{(n\times 2)}$, em que $\mathbf{x}_i^\top = (1, x_i)$ e $c(\nu) = \dfrac{\Gamma\left(\frac{\nu+1}{2})\right)}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}$.

A função escore pode ser obtida derivando o logaritmo da função de verossimilhança, com respeito a cada um dos parâmetros desconhecidos. Note que não existem soluções explícitas para o problema de maximização da função de log-verossimilhança dada em (3.2). Neste caso, podemos maximizar numericamente usando, por exemplo, o R. Este *software* contêm rotinas prontas para tratar problemas de maximização (minimização) de qualquer função.

Entretanto, neste trabalho, apresentamos o algoritmo EM para o cálculo dos estimadores de máxima verossimilhança de modo a obtermos soluções analíticas para os estimadores de $\boldsymbol{\beta}$ e σ^2.

3.3.1 Estimação dos Parâmetros via Algoritmo EM

O algoritmo EM (Dempster et al., 1977) [11] é um enfoque amplamente aplicado no cálculo iterativo de estimativas de máxima verossimilhança, sendo bastante útil para problemas com dados incompletos.

Segundo a Proposição 2.3.1, note que o modelo descrito em (3.1)-(3.2) pode ser descrito hierarquicamente como

$$Y_i \mid U_i = u_i \sim N\left(x_i^\top \boldsymbol{\beta}, \frac{\sigma^2}{u_i}\right) \tag{3.3}$$

$$U_i \sim Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right), i = 1, ..., n. \tag{3.4}$$

Neste processo de estimação, considere $\mathbf{y}= (y_1, ..., y_n)^\top$ o vetor de respostas observáveis para n unidades amostrais e $\mathbf{u}= (u_1, ..., u_n)^\top$ denota os dados não observáveis, sendo que os dados aumentados, também chamados dados completos ($\mathbf{y}_c$), correspondem aos dados observados acrescidos dos dados não observáveis. Então, sob representação hierárquica (3.3)-(3.4), segue que a função de log-verossimilhança dos dados completos, $\mathbf{y}_c= (\mathbf{y}^\top, \mathbf{u}^\top)$, é dada por

$$\begin{aligned} \ell_c(\boldsymbol{\theta}; \mathbf{y}_c) &\propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} u_i(y_i - x_i^\top\boldsymbol{\beta})^2 \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top D(\mathbf{u})(\mathbf{y} - \mathbf{X}\beta), \end{aligned}$$

onde D(**u**)= $diag(u_1, ..., u_n)$, tal que a notação $diag()$ representa uma matriz diagonal.

Após manipulações algébricas, a esperança condicional da função de log-verossimilhança completa é dada por $Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)})$, onde

$$\begin{aligned} Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)}) &= E[\ell_c(\boldsymbol{\theta}; \mathbf{y}_c)|\widehat{\boldsymbol{\theta}}^{(t)}, \mathbf{y}] \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\widehat{u}_i^{(t)}(y_i - x_i^\top\boldsymbol{\beta})^2 \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}(\widehat{\mathbf{u}}^{(t)})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \end{aligned}$$

onde $\widehat{\boldsymbol{\theta}}^{(t)} = (\widehat{\boldsymbol{\beta}}^{(t)\top}, \widehat{\sigma^2}^{(t)})$, isto é, a estimativa de $\boldsymbol{\theta}$ na t-ésima iteração. Observe que a função Q é completamente determinada pelo conhecimento de $\widehat{u}_i^{(t)} = E(U_i \mid y_i, \widehat{\theta}^{(t)})$.

Cada iteração do algoritmo EM consiste de dois passos: esperança (passo E) e maximização (passo M). A $(t+1)$- ésima iteração do algoritmo EM é definida como se segue.

PASSO E: Dado $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t)}$, calcule os pesos $\widehat{u}_i^{(t)}$ através das seguinte esperança condicional:

$$E(U_i \mid y_i, \widehat{\theta}^{(t)}) = \widehat{u}_i^{(t)} = \frac{\nu + 1}{\nu + \widehat{d}_i^{(t)}},$$

onde $\widehat{d}_i^{(t)} = d_i(\widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\sigma^2}^{(t)})$. Note que $E(U_i \mid y_i, \widehat{\theta}^{(t)})$ é inversamente proporcional à distância de Mahalanobis. Então, quanto maior o valor de $d(\widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\sigma^2}^{(t)})$, temos um menor valor para $E(U_i \mid y_i, \widehat{\theta}^{(t)})$ e assim o procedimento de estimação tende a dar um menor peso para as observações atípicas no sentido da distância de Mahalanobis. Dessa forma, observamos que quando utilizamos distribuições com caudas mais pesadas que a distribuição normal, o procedimento de estimação no contexto do modelo t-Student acomoda as observações atípicas atribuindo-lhes menor peso no processo de estimação.

PASSO M: Atualize $\widehat{\boldsymbol{\theta}}^{(t+1)}$ maximizando $Q(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(t)})$ sobre $\boldsymbol{\theta}$ obtendo as seguintes expressões fechadas.

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}^\top D(\widehat{\mathbf{u}}^{(t)})\mathbf{X})^{-1}\mathbf{X}^\top D(\widehat{\mathbf{u}}^{(t)})\mathbf{y},$$

$$\widehat{\sigma^2}^{(t+1)} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(t)})^\top D(\widehat{\mathbf{u}}^{(t)})(\mathbf{y} - \mathbf{X}\widehat{\beta}^{(t)}).$$

Deve-se alternar os passos E e M repetidamente até atingir a convergência. Como critério de convergência, utiliza-se $|\ell(\widehat{\boldsymbol{\theta}}^{(t+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(t)})| < \epsilon$. Valores iniciais são necessários para implementar o algoritmo. Eles são obtidos sob a suposição de normalidade. Entretanto, com a finalidade de verificar que a estiamtiva de máxima verossimilhança foi encontrada, recomenda-se rodar o algoritmo EM usando diferentes valores iniciais e verificar se a função escore avaliada em $\widehat{\boldsymbol{\theta}}$ é igual à $\mathbf{0}$ (ou suficientemente pequeno).

Finalmente, de acordo com Lange et al. (1989) [15], a função de verossimilhança perfilada é usada para determinar o valor ótimo de ν como segue: se $\boldsymbol{\theta}= (\boldsymbol{\beta}^\top, \sigma^2)^\top$ for o vetor de parâmetros de interesse e ν é o parâmetro *nuisance*, então a função de verossimilhança perfilada de $\boldsymbol{\theta}$ é $l_P(\boldsymbol{\theta}) = max_\nu \ell(\boldsymbol{\theta}, \nu)$, onde $\ell(\boldsymbol{\theta}, \nu)$ é a log-verossimilhança definida em (3.2).

Note que se $u_i = 1,\ \forall\, i = 1, ..., n$, temos que os resultados expostos acima coincidem com as estimativas de máxima verossimilhança do modelo de regressão linear sob erros normais.

Após uma breve revisão dos principais conceitos e métodos úteis para o desenvolvimento deste trabalho, no próximo capítulo propomos os Modelos de Regressão com Ponto de Mudança Contínuo para Dados Censurados sob Distribuições Simétricas.

4 MODELOS DE REGRESSÃO COM PONTO DE MUDANÇA CONTÍNUO PARA DADOS CENSURADOS SOB DISTRIBUIÇÕES SIMÉTRICAS

Neste capítulo, primeiramente, apresenta-se de forma geral a especificação do modelo proposto com um ponto de mudança contínuo, na estrutura da média (isto é, nos coeficientes de regressão) sob a classe de distribuições simétricas SMN. Em seguida, considera-se dados censurados. É importante ressaltar que essa é a contribuição inovadora deste trabalho. No contexto de censura, é realizado um estudo de inferência com o intuito de estimar os parâmetros do modelo, incluindo o ponto de mudança em que consideramos também como parâmetro desconhecido. Vale ressaltar que, neste trabalho, consideramos o ponto de mudança como "parâmetro verdadeiro" (desconhecido) e este é estimado via máxima verossimilhança.

4.1 Modelo de Regressão Linear com Ponto de Mudança

De acordo com Muggeo (2003) [19] e Young (2014) [28], o modelo de regressão linear com ponto de mudança contínuo pode ser reescrito como,

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+ + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.1}$$

onde $(x_i - \gamma)_+ = (x_i - \gamma) I_{\{x_i > \gamma\}}$, tal que $I_{\{\cdot\}}$ é a função indicadora e γ é o ponto de mudança. Note que se $x_i > \gamma$, então $(x_i - \gamma)_+ = (x_i - \gamma)$ e por sua vez, se $x_i \leq \gamma$, temos que $(x_i - \gamma)_+ = 0$. Portanto, observe que temos uma reta à esquerda do ponto de mudança dada por

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

enquanto que à direita do ponto de mudança, temos

$$Y_i = (\beta_0 - \beta_2 \gamma) + (\beta_1 + \beta_2) x_i + \varepsilon_i,$$

onde β_1 é a inclinação da reta à esquerda do ponto de mudança e $(\beta_1 + \beta_2)$ é a inclinação da reta à direita . Assim, β_2 é o parâmetro que representa a "diferença nas inclinações" dessas retas.

Nesta seção, estendemos os modelos de regressão normal e t-Student, definidos no Capítulo 3, para o contexto de ponto de mudança contínuo sob a classe de distribuições SMN. Esta classe de distribuições têm como casos particulares os modelos normal e t-Student, além de outros; veja a Seção 2.3.2 para mais detalhes. Dessa forma, assumimos

$$\varepsilon_i \overset{iid}{\sim} SMN(0, \sigma^2, \nu), \ i = 1, ..., n, \tag{4.2}$$

e consequentemente

$$Y_i \overset{ind}{\sim} SMN(\beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+, \sigma^2, \nu). \tag{4.3}$$

4.1.1 Estimação dos Parâmetros via Algoritmo EM

De acordo com a Proposição 2.3.1, temos que o modelo definido nas equações (4.1) e (4.2) pode ser escrito hierarquicamente como

$$Y_i \mid U_i = u_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+, \sigma^2/u_i), \quad U_i \sim h(u_i; \nu), \tag{4.4}$$

$i = 1, \ldots, n$.

Neste modelo, para $\boldsymbol{\theta} = (\gamma, \beta_0, \beta_1, \beta_2, \sigma^2)^\top$, onde assume-se que o parâmetro ν (associado com a variável de mistura U) é fixo (conhecido), e sob a representação hierárquica dada em (4.4), segue que a função log-verossimilhança completa associada com $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{u}^\top)^\top$ é dada por

$$\begin{aligned} \ell_c(\boldsymbol{\theta}; \mathbf{y}_c) &\propto -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} u_i (y_i - \beta_0 - \beta_1 x_i - \beta_2 (x_i - \gamma)_+)^2 \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top D(\mathbf{u}) (\mathbf{y} - \mathbf{X}\beta), \end{aligned}$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ e $X = \begin{bmatrix} 1 & x_1 & (x_1 - \gamma)_+ \\ 1 & x_2 & (x_2 - \gamma)_+ \\ \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - \gamma)_+ \end{bmatrix}$.

Após manipulações algébricas, a esperança condicional da função de log-verossimilhança completa, $Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)})$, tem a mesma expressão matricial definida na Seção 3.3.1, porém com $\boldsymbol{\beta}$ e $\mathbf{X}$ dados acima.

Cada iteração do algoritmo EM consiste de dois passos: esperança (passo E) e maximização (passo M). A $(t+1)$- ésima iteração do algoritmo EM é definida como se segue.

PASSO E: Dado $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t)}$, calcule os pesos $\widehat{u}_i^{(t)}$. Neste passo, para o modelo normal, $\widehat{u}_i^{(t)} = 1$, $i = 1, \ldots, n$. Para o modelo t-Student, os pesos são calculados como na Seção 3.3.1. E para o modelo slash, veja o trabalho de Garay et al. (2015) [12], página 7, para mais detalhes.

PASSO1-CM: Considere $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)^\top$, tal que $\theta_1 = \gamma$ e $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}^\top, \sigma^2)^\top$. Neste passo, ocorre a estimação do ponto de mudança, onde assumimos que o ponto de mudança deve ocorrer dentro do domínio de x, e calculamos

$$\widehat{\theta}_1^{(t+1)} = \underset{\theta_1}{\operatorname{argmax}}\ Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)})$$

com θ_2 fixado como $\widehat{\boldsymbol{\theta}}_2^{(t)}$.

PASSO2-CM: Neste passo, calculamos

$$\widehat{\theta}_2^{(t+1)} = \underset{\theta_2}{\operatorname{argmax}}\ Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)})$$

com θ_1 fixado como $\widehat{\boldsymbol{\theta}}_1^{(t)}$. Dessa forma, atualizar $\widehat{\theta}_2^{(t+1)}$ consiste em obter

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}^\top D(\widehat{\mathbf{u}}^{(t)})\mathbf{X})^{-1}\mathbf{X}^\top D(\widehat{\mathbf{u}}^{(t)})\mathbf{y}, \tag{4.5}$$

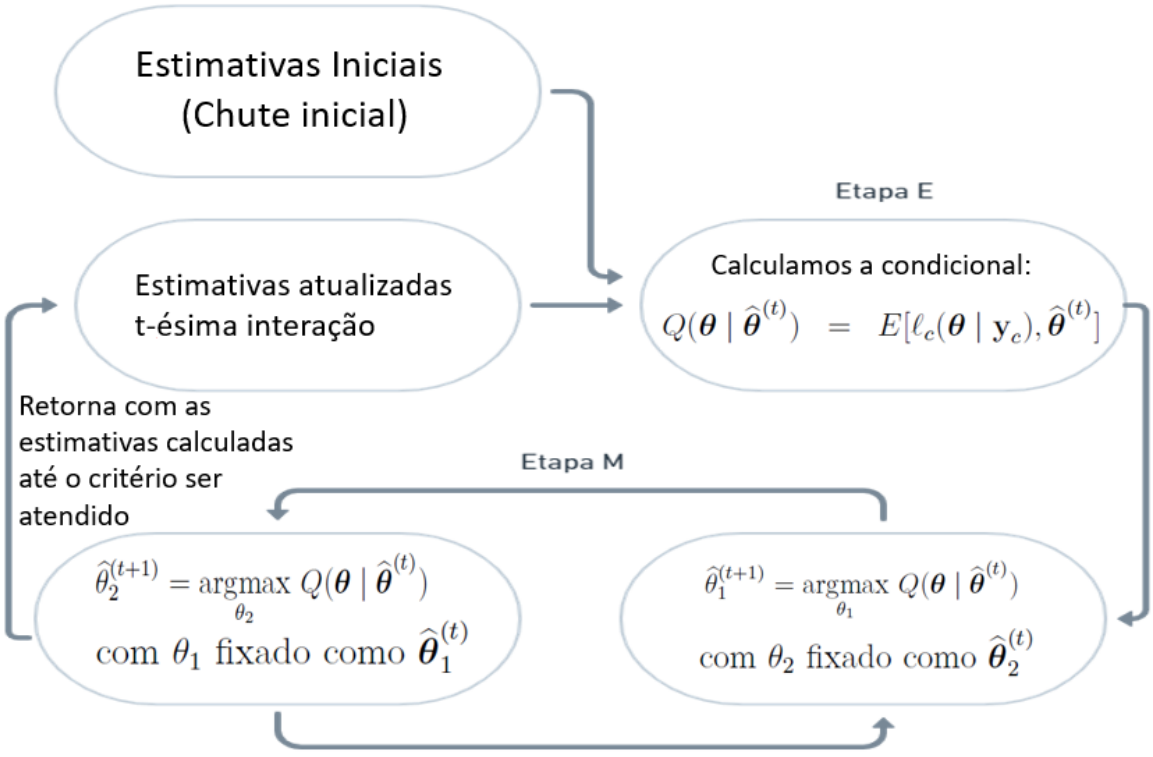
$$\widehat{\sigma^2}^{(t+1)} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(t)})^\top D(\widehat{\mathbf{u}}^{(t)})(\mathbf{y} - \mathbf{X}\widehat{\beta}^{(t)}). \tag{4.6}$$

Note que as expressões dos estimadores na etapa M não se alteram para as diferentes distribuições da classe SMN, apenas a etapa E.

Finalmente, de acordo com Lange et al. (1989) [15], a função de verossimilhança perfilada é usada para determinar o valor ótimo de ν, como descrito na Seção 3.3.1, onde ressaltamos que $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^\top, \sigma^2)^\top$, com $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$. Deve-se alternar os passos E e M repetidamente até atingir a convergência. Além disso, valores iniciais são necessários para implementar este algoritmo. Como valor inicial para o ponto de mudança γ, utilizamos a mediana dos 500 valores gerados aleatoriamente de uma uniforme no intervalo [a,b], domínio correspondente de x. Dado este ponto de mudança, ajustamos o modelo de regressão linear via mínimos quadrados e obtivemos os valores iniciais para os demais parâmetros.

Para facilitar a visualização das etapas do algoritmo EM, segue abaixo o fluxograma.

Figura 6 – Fluxograma das estapas do algoritmo EM



4.2 Dados Censurados

Estamos interessados no caso em que as observações com censura (à esquerda ou à direita) podem ocorrer. Isto é, no cenário de censura à esquerda, as observações são da forma (V_i, φ_i), onde o Y_i é uma variável latente e

$$V_i = \begin{cases} c_i, & \forall \ \ Y_i \leq c_i \ (i.e.\ se\ \varphi_i = 1) \\ Y_i, & \forall \ \ Y_i > c_i, \quad i = 1, ..., n \ (i.e.\ se\ \varphi_i = 0) \end{cases}, \tag{4.7}$$

tal que c_i é o ponto de corte. O indicador de censura $\varphi_i = 1$ (ou $\varphi_i = 0$) significa que a i-ésima observação é censurada (ou não é censurada). Extensões dos resultados para censura à direita são imediatos. Ou seja, é suficiente transformar Y_i e o nível de censura c_i para $-Y_i$ e $-c_i$, respectivamente.

4.2.1 Modelo de Regressão Linear com Ponto de Mudança para Dados Censurados

Nesta seção, definimos o modelo de regressão linear com variável resposta censurada e erros distribuídos na família de distribuições SMN, no contexto de ponto de mudança. Estendemos o modelo de regressão linear definido em (4.1) e (4.2) com a suposição de que a variável resposta não é totalmente observada para todos os sujeitos. Chamamos a estrutura definida por (4.1), (4.2) e (4.7) como o modelo de regressão com ponto de mudança contínuo para dados censurados sob distribuições simétricas.

Neste modelo, a função de log-verossimilhança será dada por

$$\ell(\boldsymbol{\theta}; \mathbf{v}, \boldsymbol{\varphi}) = \sum_{i=1}^{n} \varphi_i \log \left[F_{SMN}\left(\frac{v_i - \mu_i}{\sigma} \right) \right] + \sum_{i=1}^{n} (1 - \varphi_i) \log[f_{SMN}(v_i; \boldsymbol{\theta})],$$

tal que $\mu_i = (\beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+)$, $\mathbf{v} = (v_1, \ldots, v_n)^\top$ é a amostra observada de $\mathbf{V} = (V_1, V_2, \ldots, V_n)^\top$, $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)^\top$, F_{SMN} e f_{SMN} é a função de distribuição acumulada e a função densidade da $SMN(0, 1, \nu)$, respectivamente. Como a função de log-verossimilhança observada envolve expressões complexas, é muito difícil maximizar diretamente $\ell(\boldsymbol{\theta}; \mathbf{v}, \boldsymbol{\varphi})$. Para superar este problema, propomos um algoritmo do tipo EM baseado em uma representação de dados aumentados do modelo de regressão com ponto de mudança contínuo para dados censurados sob distribuições simétricas.

4.2.2 Estimação dos Parâmetros via Algoritmo EM

Com o objetivo de estimar os parâmetros do modelo proposto via algoritmo EM, usaremos a representação hierárquica do modelo proposto em (4.1), (4.2) e (4.7), dada abaixo:

$$Y_i \mid U_i = u_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i + \beta_2 (x_i - \gamma)_+), \sigma^2/u_i), \qquad U_i \sim h(u_i; \nu), \tag{4.8}$$

$i = 1, \ldots, n$.

Se a observação i for censurada, podemos considerar y_i como a realização de uma variável latente não observável $Y_i \sim SMN(\boldsymbol{\mu}_i, \sigma^2, \nu)$. O procedimento chave para o desenvolvimento do algoritmo EM para o modelo em questão é considerar os dados completos $\mathbf{y}_c = (\mathbf{v}^\top, \boldsymbol{\varphi}^\top, \mathbf{y}^\top, \mathbf{u}^\top)^\top$. Assim, considerando a representação (4.8), a função log-verossimilhança completa para $\boldsymbol{\theta}$, associado com $\mathbf{y}_c$, é a mesma definida na Seção 4.1.1. Porém, a esperança condicional da função de log-verossimilhança completa, $Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)})$, é diferente, devido a presença de dados censurados, isto é,

$$Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)}) = -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[\epsilon_{2i}(\widehat{\boldsymbol{\theta}}^{(t)}) - 2\epsilon_{1i}(\widehat{\boldsymbol{\theta}}^{(t)}) \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_{0i}(\widehat{\boldsymbol{\theta}}^{(t)}) (\mathbf{x}_i^\top \boldsymbol{\beta})^2 \right], \quad (4.9)$$

onde $\mathbf{x}_i^\top$ é a i-ésima linha da matriz $\mathbf{X}$ definida na Seção 4.1.1, ou seja, $\mathbf{x}_i^\top = (1, x_i, (x_i - \gamma)_+)^\top$. Adicionalmente, temos que $\epsilon_{si} = E[U_i Y_i^s | v_i, \varphi_i, \widehat{\boldsymbol{\theta}}^{(t)}]$ para $s = 0, 1, 2$.

Cada iteração do algoritmo EM consiste de dois passos: esperança (passo E) e maximização (passo M). A $(t+1)$- ésima iteração do algoritmo EM é definida como se segue.

PASSO E: Dado $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(t)}$, calcule $\epsilon_{2i}(\widehat{\boldsymbol{\theta}}^{(t)})$, $\epsilon_{1i}(\widehat{\boldsymbol{\theta}}^{(t)})$ e $\epsilon_{0i}(\widehat{\boldsymbol{\theta}}^{(t)}), i = 1, \ldots, n$. Para a i-é-sima observação não censurada ($\varphi_i = 0$), temos que $V_i = Y_i$, então $\epsilon_{si} = E[U_i Y_i^s | v_i, \varphi_i, \widehat{\boldsymbol{\theta}}^{(t)}] = y_i^s E[U_i | y_i, \widehat{\boldsymbol{\theta}}^{(t)}]$ para $s = 0, 1, 2$. Note que os pesos $\widehat{u}_i^{(t)} = E[U_i | y_i, \widehat{\boldsymbol{\theta}}^{(t)}]$ são obtidos conforme a distribuição pertencente à classe SMN; veja Seção 4.1.1 para mais detalhes. Para a a i-é-sima observação censurada ($\varphi_i = 1$), temos que $Y_i \leq c_i$ e $V_i = c_i$, então $\epsilon_{si} = E[U_i Y_i^s | v_i, Y_i \leq c_i, \widehat{\boldsymbol{\theta}}^{(t)}]$ para $s = 0, 1, 2$. Estes valores podem ser obtidos para as diferentes distribuições utilizando os resultados da Proposição 1 em Garay et al. (2015) [12]. Para o cálculo desses valores esperados usamos o pacote SMNCensReg disponível no CRAN do R.

PASSO 1-CM: Para $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2^\top)^\top$, onde $\theta_1 = \gamma$ e $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}^\top, \sigma^2)^\top$. Neste passo, ocorre a estimação do ponto de mudança, onde assumimos que o ponto de mudança deve ocorrer do domínio de x, e calculamos

$$\widehat{\theta}_1^{(t+1)} = \underset{\theta_1}{\text{argmax}}\ Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)})$$

com θ_2 fixado como $\widehat{\boldsymbol{\theta}}_2^{(t)}$.

PASSO2-CM: Neste passo, calculamos

$$\widehat{\theta}_2^{(t+1)} = \underset{\theta_2}{\text{argmax}}\ Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(t)})$$

com θ_1 fixado como $\widehat{\boldsymbol{\theta}}_1^{(t)}$. Dessa forma, atualizar $\widehat{\theta}_2^{(t+1)}$ consiste em obter

$$\widehat{\beta}^{(t+1)} = \left(\sum_{i=1}^{n} \boldsymbol{\epsilon}_{0i}(\widehat{\boldsymbol{\theta}}^{(t)}) \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i \boldsymbol{\epsilon}_{1i}(\widehat{\boldsymbol{\theta}}^{(t)})$$

e

$$\widehat{\sigma^2}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\epsilon}_{2i}(\widehat{\boldsymbol{\theta}}^{(t)}) - 2\boldsymbol{\epsilon}_{1i}(\widehat{\boldsymbol{\theta}}^{(t)})\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{\epsilon}_{0i}(\widehat{\boldsymbol{\theta}}^{(t)})\left(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}^{(t)}\right)^2\right].$$

De acordo com Lange et al. (1989) [15], a função de verossimilhança perfilada é usada para determinar o valor ótimo de ν, como descrito na Seção 3.3.1, onde ressaltamos que $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^\top, \sigma^2)^\top$, com $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$. Deve-se alternar os passos E e M repetidamente até atingir a convergência. Além disso, valores iniciais são necessários para implementar este algoritmo; veja mais detalhes na Seção 4.1.1.

Finalmente, no próximo Capítulo, apresentamos aplicações numéricas (dados reais e simulados). Foi utilizado o programa estatístico R. As rotinas em R das metodologias propostas neste Capítulo estão descritas no Apêndice A.

5 APLICAÇÕES NUMÉRICAS

Nesta seção, aplicações a dados reais e simulados são apresentadas a fim de ilustrar o modelo e os resultados inferenciais desenvolvidos. As rotinas em R dos estudos de simulação estão descritas no Apêndice B.

5.1 Estudos de simulação

O primeiro objetivo é verificar se podemos recuperar os valores dos parâmetros reais quando usamos o algoritmo EM proposto, ajustando o modelo de regressão linear sob a classe de distribuições simétricas SMN, com ponto de mudança contínuo, aos dados censurados que foram gerados artificialmente.

Neste caso, geramos 500 amostras de tamanho $n = 25, 50, 100$ e 200 provenientes do modelo de regressão, proposto no Capítulo 4, com níveis de censura à direita 0%, 5%, 10%, 20% e 30%, com ponto de mudança contínuo, considerando γ no percentil 40 dos valores da variável explicativa. A variável explicativa utilizada foi $x_i \sim U(1, 10)$ e os erros $\epsilon_i \overset{iid}{\sim} SMN(0, 1, \nu)$, $i = 1, ..., n$, com os seguintes valores para os parâmetros: $\beta_0 = 2, \beta_1 = -1$, $\beta_2 = 3$ e, $\nu = 2$ para o modelo slash e $\nu = 3$ para o modelo t-Student.

Dessa forma, nesta seção, usamos as simulações para avaliar o desempenho das estimativas de máxima verossimilhança dos parâmetros no modelo proposto. O estudo de simulação foi projetado para observar as mudanças nas estimativas variando os tamanhos das amostras e os níveis de censura à direita. As Figuras 7, 8 e 9 mostram boxplots das estimativas de parâmetros para os modelos normal, t-Student e slash, respectivamente. Adicionalmente, as Figuras 10, 11 e 12 apresentam as frequências dos pontos de mudança estimados para os modelos nos cenários com 5% e 20% de censura.

Note que o algoritmo proposto é eficiente para a estimação dos parâmetros do modelo de regressão linear sob a classe de distribuições simétricas SMN, no contexto de dados censurados, com ponto de mudança contínuo, independente do tamanho da amostra.

Pelas Figuras 7, 8 e 9, em geral, para um determinado nível de censura, o viés e a variabilidade das estimativas dos parâmetros $\beta_0, \beta_1, \beta_2$ e σ^2 diminuem quando o tamanho da amostra aumenta. Isso concorda essencialmente com as propriedades assintóticas do estimador de máxima verossimilhança. Além disso, observe que para um determinado nível de censura, a variabilidade das estimativas do parâmetro γ diminui quando o tamanho da amostra aumenta. Em termos de viés não conseguimos discutir através dos boxplots, pois quando o tamanho da amostra é alterado, consequentemente a localização do ponto de mudança se altera também. As Figuras 10, 11 e 12 mostram que para todos os cenários a moda das distribuições simuladas dos pontos de mudança estimados ocorre na posição definida nos dados simulados à medida que o tamanho da amostra aumenta.

Figura 7 – Boxplots das estimativas, fixando o nível de censura e variando os tamanhos amostrais, do modelo Normal com o ponto de mudança (γ) no percentil 40. Níveis de censura à direita (a) 0%, (b) 5%, (c) 10%, (d) 20% e (e) 30%. Valores verdadeiros identificados nos gráficos com uma linha horizontal.

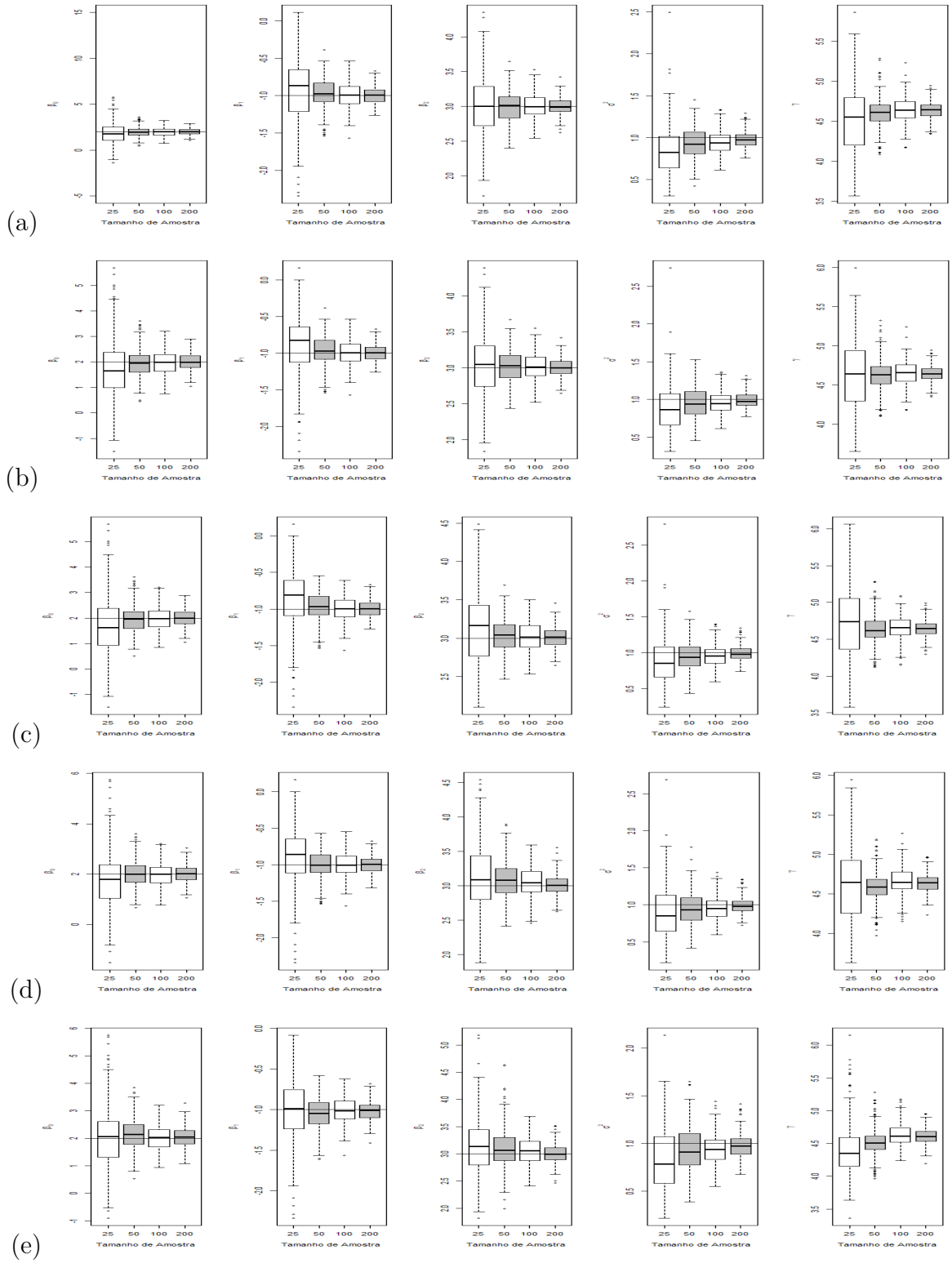


Figura 8 – Boxplots das estimativas, fixando o nível de censura e variando os tamanhos amostrais, do modelo t-Student com o ponto de mudança (γ) no percentil 40. Níveis de censura à direita (a) 0%, (b) 5%, (c) 10%, (d) 20% e (e) 30%. Valores verdadeiros identificados nos gráficos com uma linha horizontal.

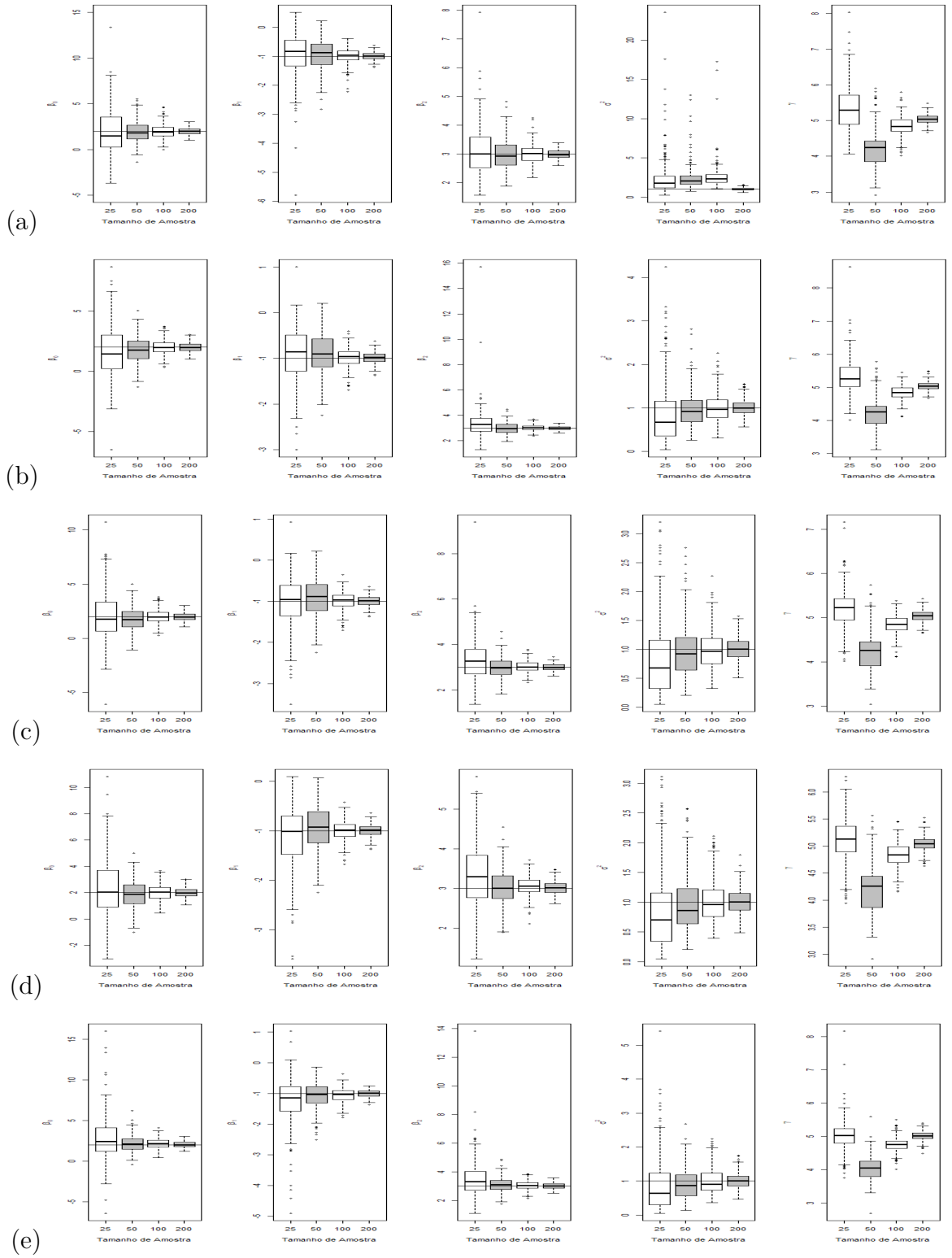


Figura 9 – Boxplots das estimativas, fixando o nível de censura e variando os tamanhos amostrais, do modelo Slash com o ponto de mudança (γ) no percentil 40. Níveis de censura à direita (a) 0%, (b) 5%, (c) 10%, (d) 20% e (e) 30%. Valores verdadeiros identificados nos gráficos com uma linha horizontal.

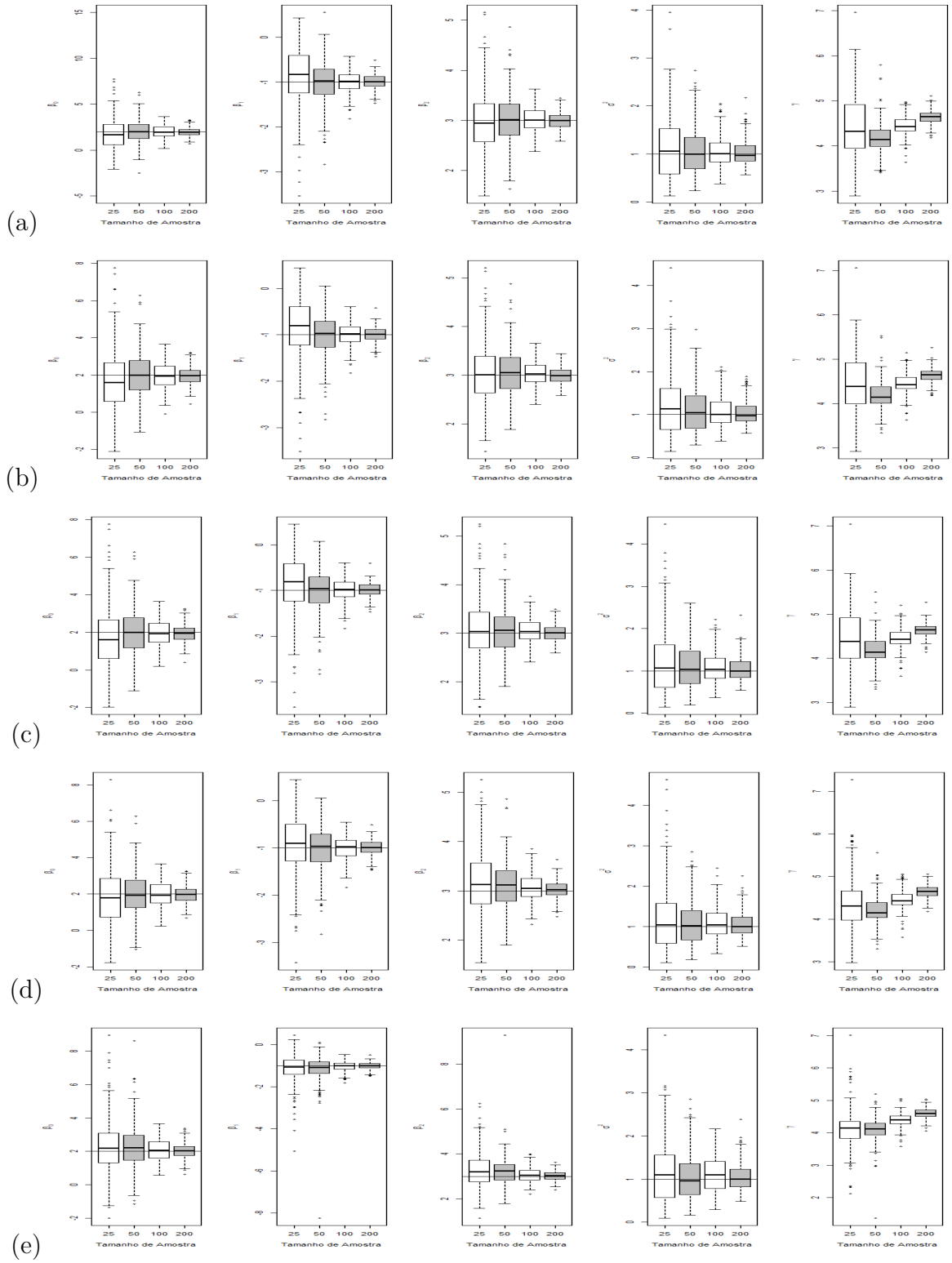
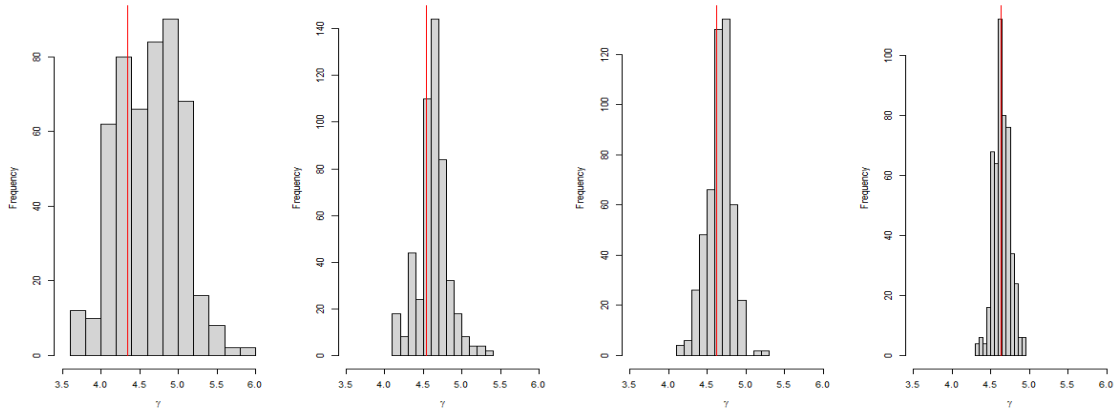


Figura 10 – Frequências dos pontos de mudança estimados para os modelos Normal, fixando o nível de censura e variando os tamanhos amostrais, $n = 25, 50, 100$ e 200, respectivamente, da esquerda para à direita, com o ponto de mudança (γ) no percentil 40. Valores verdadeiros identificados nos gráficos com uma linha vertical. Níveis de censura à direita (a) 5% e (b) 20%.

(a)



(b)

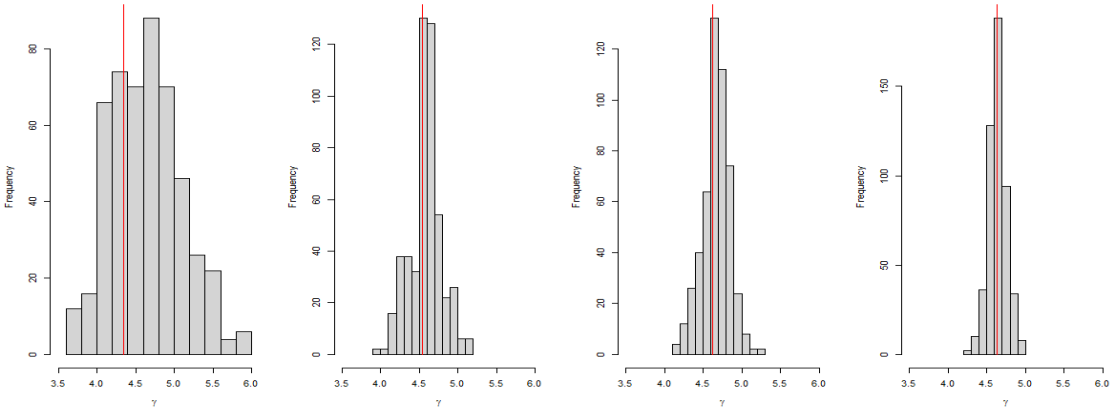


Figura 11 – Frequências dos pontos de mudança estimados para os modelos t-Student, fixando o nível de censura e variando os tamanhos amostrais, $n = 25, 50, 100$ e 200, respectivamente, da esquerda para à direita, com o ponto de mudança (γ) no percentil 40. Valores verdadeiros identificados nos gráficos com uma linha vertical. Níveis de censura à direita (a) 5% e (b) 20%.

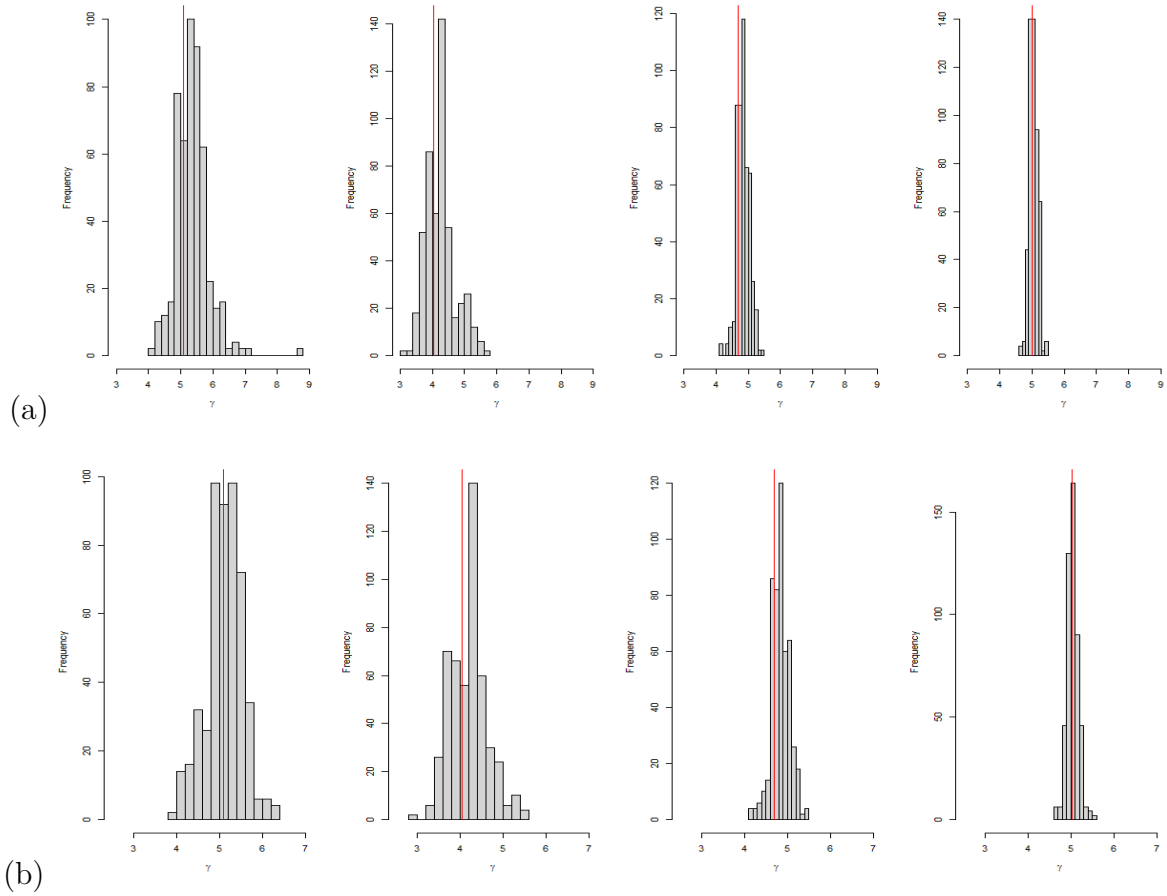
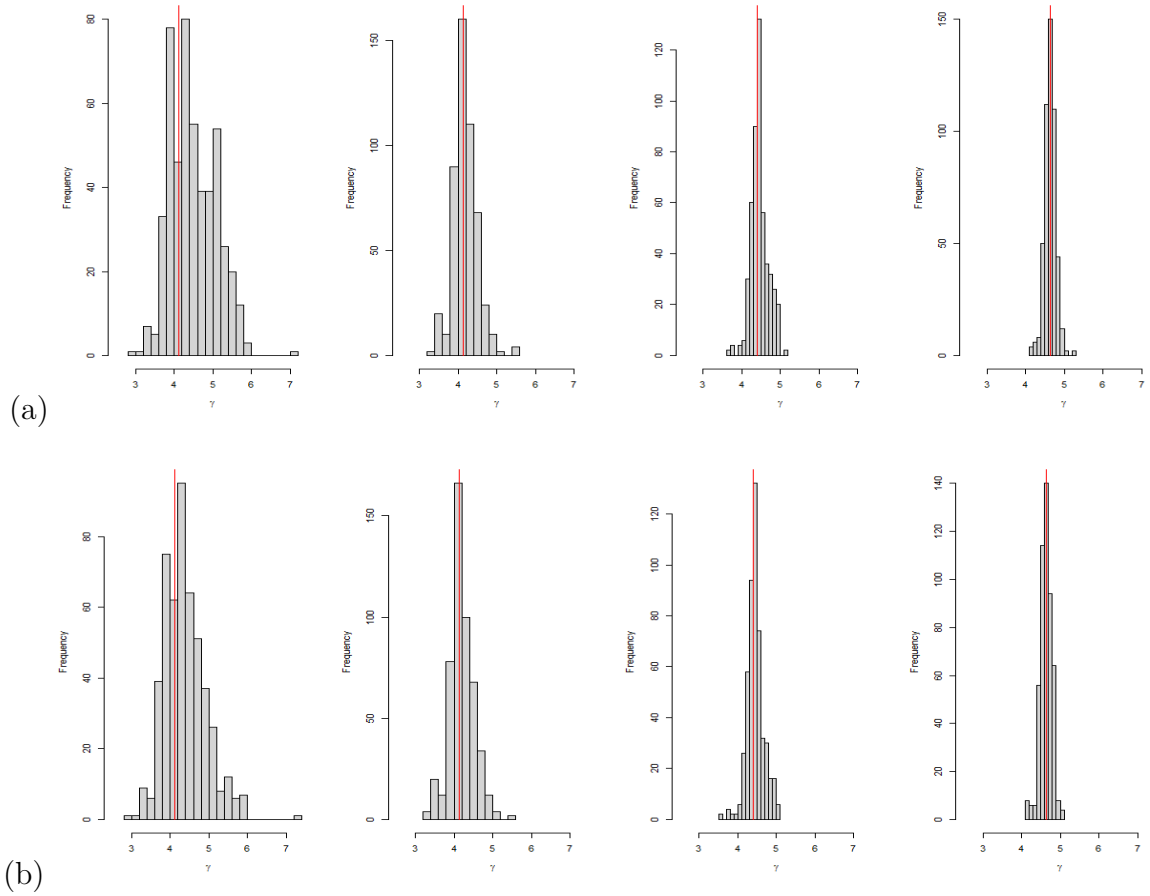


Figura 12 – Frequências dos pontos de mudança estimados para os modelos Slash, fixando o nível de censura e variando os tamanhos amostrais, $n = 25, 50, 100$ e 200, respectivamente, da esquerda para à direita, com o ponto de mudança (γ) no percentil 40. Valores verdadeiros identificados nos gráficos com uma linha vertical. Níveis de censura à direita (a) 5% e (b) 20%.



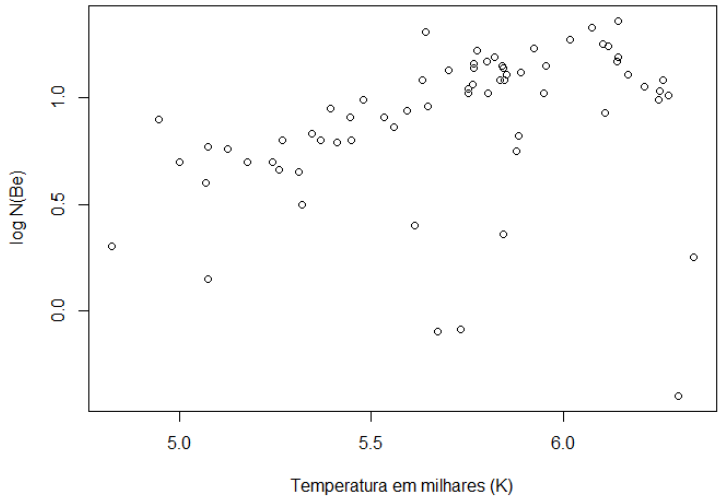
5.2 Aplicação em dados reais

Vamos considerar o banco de dados reais chamado Stellar descrito em Santos et al. (2002) [22]. Estes dados estão disponíveis no pacote do R *astrodatR*. O conjunto de dados consiste em medições de 68 estrelas do tipo solar, tais que

- log N(Be) é a variável resposta que representa o log da abundância do elemento de berílio (Be) em estrelas dimensionado para a abundância do Sol (ou seja, o Sol tem log N(Be) = 0,0).
- Teff/1000 é a variável explicativa que representa a temperatura efetiva da superfície estelar (em Kelvin - K).

De acordo com Santos et al. (2002) [22] devido à limitações de sensibilidades dos aparelhos de medição, alguns objetos podem não ser detectados (1- detectado e 0- não detectado). Dessa forma, neste conjunto de dados, temos 12 observações censuradas à esquerda, ou seja, 12 medições de berílio não detectadas (ou seja, com limitações nas respectivas medições) que representam $19,35\%$ das observações.

Figura 13 – Representação gráfica dos dados, sendo log N(Be) a variável resposta (Santos et al., 2002) [22].



Note que a Figura 13 indica uma mudança na estrutura da média na sequência de observações. Os dados mostram uma tendência clara e crescente para as observações com temperaturas abaixo de 6 K (aproximadamente) e uma tendência decrescente para as observações com temperaturas acima de 6 K (aproximadamente). Logo, teremos duas retas distintas, uma crescente e a outra decrescente, o que indica claramente a existência de um ponto de mudança entre estas duas retas lineares.

As estimativas dos parâmetros para os modelos normal, t-Student e slash com ponto de mudança na posição $\widehat{\gamma}$ estão apresentadas na Tabela 5.2. A fim de compararmos os resultados obtidos, também apresentamos as estimativas dos parâmetros para os modelos normal, t-Student e slash sem ponto de mudança na Tabela 5.2.

Tabela 2 – Estimativas dos parâmetros dos modelos com ponto de mudança.

Parâmetros	β_0	β_1	β_2	σ^2	γ	ν
Modelo Normal	-2.027	0.520	-12.948	0.076	6.234	-
Modelo t-Student	-2.649	0.645	-2.118	0.004	6.081	1.085
Modelo Slash	-2.585	0.635	-2.099	0.002	6.081	0.510

Tabela 3 – Estimativas dos parâmetros dos modelos sem ponto de mudança.

Parâmetros	β_0	β_1	β_2	σ^2	γ	ν
Modelo Normal	-2.023	0.492	-	0.288	-	-
Modelo t-Student	-2.243	0.545	-	0.067	-	3.0
Modelo Slash	-2.227	0.545	-	0.040	-	1.2

Os resultados dos ajustes em termos de $\ell(\widehat{\boldsymbol{\theta}})$, AIC e BIC são fornecidos nas Tabelas 4 e 5. Pelos críterios de informação (menor valor), vemos que o modelo que melhor se ajusta os dados é o t-Student com ponto de mudança.

Tabela 4 – Alguns critérios de informação.

Modelos sem Ponto de Mudança	$\ell(\widehat{\boldsymbol{\theta}})$	AIC	BIC	$\widehat{\gamma}$
Normal	-18.227	42.454	49.112	-
t-Student	-2.127	10.254	16.912	-
Slash	-2.725	11.450	18.108	-

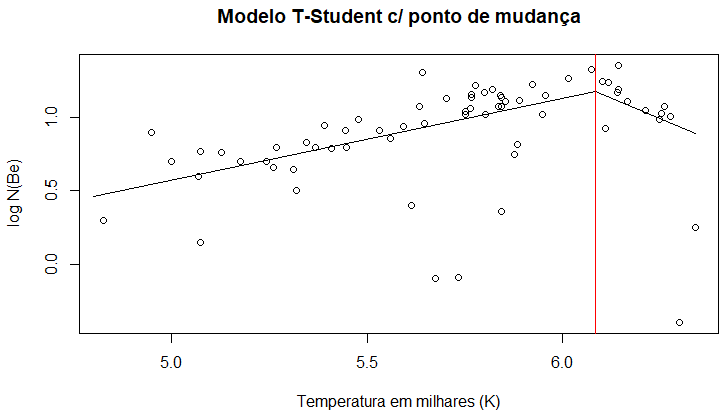
Tabela 5 – Alguns critérios de informação.

Modelos com Ponto de Mudança	$\ell(\widehat{\boldsymbol{\theta}})$	AIC	BIC	$\widehat{\gamma}$
Normal	-8.813	27.626	38.723	6.234
t-Student	18.317	-26.634	-15.536	6.081
Slash	18.282	-26.563	-15.466	6.081

Dessa forma, a Figura 14 mostra o gráfico de dispersão dos dados com o seguinte modelo t-Student ajustado, considerando ponto de mudança contínuo,

$$\widehat{y}_i = \begin{cases} -2.552 + 0.629x_i, & 4.825 \leq x_i \leq 6.084 \\ 10.492 - 9.305x_i, & 6.084 \leq x_i \leq 6.339 \end{cases} \tag{5.1}$$

Figura 14 – Representação gráfica do modelo t-Student com ponto de mudança contínuo.



6 CONCLUSÕES

Neste trabalho, discutimos os modelos de regressão com ponto de mudança contínuo, onde as observações seguem distribuição simétrica na classe SMN, no contexto de dados censurados. Em particular, fixamos nossa atenção às distribuições normal, t-Student e slash, três importantes membros da classe SMN. O algoritmo EM foi desenvolvido, fornecendo uma solução analítica para os estimadores de máxima verossimilhança dos parâmetros dos modelos. O algoritmo EM tem diversas vantagens sobre a maximização direta da função verossimihança uma vez que é facilmente implementável, numericamente estável e bastante acurado, como demonstrado através dos estudos de simulação. Foi utilizado o programa estatístico R para a programação do procedimento de estimação dos modelos ajustados. Todos as rotinas em R estão disponíveis nos Apêndices A e B. Alguns estudos de simulação e aplicação a conjunto de dados reais são apresentados com o propósito de ilustrar os modelos e os resultados inferenciais desenvolvidos aqui. Os resultados desse trabalho podem ser úteis para determinar ponto de mudança em muitas outras aplicações práticas, no contexto de dados censurados.

REFERÊNCIAS

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

[2] Andrews, D. F.; Mallows, C. L. (1974) Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological), 36:99-102.

[3] Carvalho, M. S.; Lima V. A.ndreozzi; Torres C. C.; Pereira D. C., Serrano M. T. B.; Emiko S. S. (2011). Análise de Sobrevivência: Teoria e Aplicações em Saúde. 2ª edição, Rio de Janeiro: Ed. Fiocruz.

[4] Casella, G.; Berger, R.L.(2010) Inferência Estatística - tradução da 2ª edição norte-americana.*Cengage Learning.*

[5] Chen, J.; Gupta, A. K. (1996). Detecting changes of mean in multidimensional normal sequences with application to literature and geology. *Computational Statistics*, 11, 211-221.

[6] Chen, J.; Gupta, A. K. (1997). Testing and locating variance change points with application to stock prices. *Journal of the American Statistical Association*, 92, 739-747.

[7] Chen, J.; Gupta, A. K. (1999). Change point analysis of a gaussian model. *Statistical Papers*, 40, 323-333.

[8] Chen, J.; Gupta, A. K. (2003). Information-theoretic approach for detecting change in the parameters of a normal model. *Mathematical Methods of Statistics*, 12, 116-130.

[9] Chen, J.; Gupta, A. K. (2011). *P*arametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance. Springer Science ; Business Media, New York.

[10] Contreras, C. A. H. (2014). Modelo de Regressão Linear Mistura de Escala Normal com Ponto de Mudança: Estimação e Diagnóstico. Dissertação de Mestrado, Departamento de Estatística, IMECC-UNICAMP.

[11] Dempster, A. P.; Laird, N. M.; Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm.

[12] Garay, A. M.; Lachos V. H.; Bolfarine H.; Cabral C. R. B. (2015). Linear censored regression models with scale mixtures of normal distributions.

[13] Hofrichter, J. (2007). Change Point Detection in Generalized Linear Models. Dissertation zur erlangung des akademischen grades doktor der technischen wissenschaften, Graz University of Technology.

[14] Husková, M.; Neuhaus, G. (2004), Change point analysis for censored data. Journal of Statistical Planning and Inference. Volume 126, Issue 1, Pages 207-223.

[15] Lange, K . L.; Little, R.; Taylor, J. (1989). Robust Statistical modeling using t distribuition. *Journal of the American Statistical Association*, 84, 881-896.

[16] Lange, K. L.; Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics, 2, 175-198.

[17] Meng, X. L.; Rubin, B. D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika, 80:267-278.

[18] Montgomery, D.C.; Peck, E.A.; Vining, G.G.(2012) *Introduction to Linear Regression Analysis*. John Wiley & Sons.

[19] Muggeo, V. M. R. (2003). Estimating regression models with unknown breakpoints. Statistcs in Medicine, 22(19):3055-3071.

[20] Pinheiro, J. C.; Liu, C. H.; Wu, Y. N. (2001). Efficient algotithms for robust estimation in linear mixed-effects models using a multivariate t-distribution. Journal of Computational and Graphical Statistics, 10:249-276.

[21] R CORE TEAM. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. 2020. `https://www.R-project.org/`.

[22] Santos, N.; López, R.G.; Israelian, G.; Mayor, M.; Rebolo, R.; García-Gil, A.; Taoro, M.P.; Randich, S.(2002). Beryllium abundances in stars hosting giant planets, Astron. Astrophys 386, 1028–1038.

[23] Shewhart, W. A. (1939). Statistical Method from the Viewpoint of Quality Control. Dover, Washington.

[24] Schwarz G. (1978). Estimating the dimension of a model. Annals of Statistics, 6:461–464.

[25] Sofronov, G.; Wendler, M.; Liebscher, V. (2020) Editorial for the special issue: Change point detection. Stat Papers 61, 1347–1349.

[26] Sousa, A. R.; Alencar, A.; Leonardi, F. (2021). Participação em banca de Gabriel Tominaga Dielle. A automatização da seleção de nós em regressão por splines e sua aplicação nas curvas de novos casos de COVID19. Trabalho de Conclusão de Curso (Graduação em Abi - Matemática Aplicada e Computacional) - Universidade de São Paulo.

[27] Wang, Z. F., Wu, Y. H.; Zhao, L. C. (2007). Change-point estimation for censored regression model. Science in China Series A: Mathematics, 50, 1, 63-72.

[28] Young, D. S. (2014). Mixtures of regressions with changepoints. Statistics and Computing, 24:265-281.

APÊNDICE A – Rotinas em R desenvolvidas para estimação dos parâmetros do modelo proposto

Os códigos em R a seguir foram elaborados para a obtenção das estimativas, via algoritmo EM, dos parâmetros do modelo proposto no Capítulo 4.

```
#Necessário carregar todas as funções
### chute inicial
inicial<-function(x,y)
{
  n=length(y)
  a=min(x)
  b=max(x)
  gama1=runif(100,a,b)
  gama=median(gama1)

  tabela<-table(x<=gama)
  n1=as.numeric(tabela[2])
  n2=as.numeric(tabela[1])
  w=matrix(1,nrow=n,ncol=1)
  w1=matrix(0,nrow=n1,ncol=1)
  w2=matrix(x[(n1+1):n]-gama,nrow=n2,ncol=1)
  waux=rbind(w1,w2)
  X=matrix(cbind(w,x,waux),n,3)  #matriz de planejamento

  beta=solve(t(X)%*%X)%*%t(X)%*%y

  p1=dim(X)[2]
  n=length(y)
  mi=X%*%beta
  Qbeta=t(y-mi)%*%(y-mi)
  sigma2= as.numeric(Qbeta/(n-p1))

  obj.out <- list(beta=beta,gama=gama,sigma2=sigma2)

} # chute inicial

# calculo da etapa E sem censura
NCensurEsperUY <-  function(y,mu,sigma2,nu,type)
```

```
{
  EUY0 <- EUY1 <- EUY2 <- c()
  d <- (y-mu)^2/sigma2
  n <- length(y)
  if(type=="T")
  {
    EUY0 <- (nu+1)/(nu+d)
    EUY1 <- y*(nu+1)/(nu+d)
    EUY2 <- (y^2)*(nu+1)/(nu+d)
  }
  if(type=="Normal")
  {
    EUY0 <- rep(1,n)
    EUY1 <- y
    EUY2 <- (y^2)
  }

  if(type=="Slash")
  {
    Num  <- GamaInc(nu+1.5,0.5*d)*(0.5*d)^(-nu-1.5)
    Den  <- GamaInc(nu+0.5,0.5*d)*(0.5*d)^(-nu-0.5)
    EUY0 <- Num/Den
    EUY1 <- y*Num/Den
    EUY2 <- (y^2)*Num/Den
  }
  if(type=="NormalC")
  {
    Num  <- 1-nu[1]+nu[1]*(nu[2])^(1.5)*exp(0.5*d*(1-nu[2]))
    Den  <- 1-nu[1]+nu[1]*(nu[2])^(0.5)*exp(0.5*d*(1-nu[2]))
    EUY0 <- Num/Den
    EUY1 <- y*Num/Den
    EUY2 <- (y^2)*Num/Den
  }
  return(list(EUY0=EUY0,EUY1=EUY1,EUY2=EUY2))
}

AcumSlash <- function(y,mu,sigma2,nu)
{
  Acum <- z <- vector(mode = "numeric", length = length(y))
```

```
  for (i in 1:length(y))
  {
  z[i] <- (y[i]-mu)/sqrt(sigma2)
  f1 <- function(u) nu*u^(nu-1)*pnorm(z[i]*sqrt(u))
    Acum[i]<- integrate(f1,0,1)$value
  }
  return(Acum)
}

AcumNormalC <- function(y,mu,sigma2,nu)
{
  Acum <- vector(mode = "numeric", length = length(y))
  for (i in 1:length(y))
  {
    eta  <- nu[1]
    gama <- nu[2]
    Acum[i] <- eta*pnorm(y[i],mu,sqrt(sigma2/gama)) +
    (1-eta)*pnorm(y[i],mu,sqrt(sigma2))
  }
  return(Acum)
}

AcumPearsonVII <- function(y,mu,sigma2,nu,delta)
{
  Acum <- z <- vector(mode = "numeric", length = length(y))
  sigma2a <- sigma2*(delta/nu)
  for (i in 1:length(y))
  {
    z[i] <- (y[i]-mu)/sqrt(sigma2a)
    Acum[i] <- pt(z[i],df=nu)
  }
  return(Acum)
}

dSlash <- function(y,mu,sigma2,nu)
{
  resp <- z <- vector(mode = "numeric", length = length(y))
  for (i in 1:length(y))
  {
```

```
    z[i] <- (y[i]-mu)/sqrt(sigma2)
    f2 <- function(u) nu*u^(nu-0.5)*dnorm(z[i]*sqrt(u))/sqrt(sigma2)
    resp[i] <- integrate(f2,0,1)$value
  }
  return(resp)
}

GamaInc <- function(a,x)
{
  res <- vector(mode = "numeric", length = length(x))
  f <-function(t) exp(-t)*t^(a-1)
  for  (i in 1:length(x))
  {
    res[i] <- integrate(f,0,x[i])$value
  }
  return(res)
}


E_phi <- function(r,a,nu,type=type,cens=cens)
{
  n <- length(a)
  b <- rep(Inf,n)
  b1<- rep(-Inf,n)

  if(setequal(a,b)== TRUE | setequal(a,b1)== TRUE)
  {
    resp <- rep(0,n)
  }
  else
  {
    if(type=="Normal")
    {
      resp <- dnorm(a)
    }
    if(type=="T")
    {
      Aux0 <- gamma(0.5*(nu+2*r))
      Aux1 <- gamma(nu/2)*sqrt(2*pi)
```

```
      Aux2 <- Aux0/Aux1
      Aux3 <- (0.5*nu)^(nu/2)
      Aux4 <- (0.5*(a^2+nu))^(-0.5*(nu+2*r))
      resp <- Aux2*Aux3*Aux4
    }

    if(type=="Slash")
    {
      Aux0 <- nu/sqrt(2*pi)
      Aux1 <- (0.5*a^2)^(-(nu+r))
      Aux2 <- GamaInc(nu+r,0.5*a^2)
      resp <- Aux0*Aux1*Aux2
    }
    if(type=="NormalC")
    {
      Aux0 <- nu[1]*nu[2]^(r)*dnorm(a*sqrt(nu[2]))
      Aux1 <- (1-nu[1])*dnorm(a)
      resp <- Aux0 + Aux1
    }
  }
  return(resp)
}

E_Phi <- function(r,a,nu,type=type)
{
  n <- length(a)
  if(type=="Normal")
  {
    resp <- pnorm(a)
  }
  if(type=="T")
  {
    Aux0 <- gamma(0.5*(nu+(2*r)))
    Aux1 <- gamma(nu/2)
    Aux2 <- Aux0/Aux1
    Aux3 <- (0.5*nu)^(-r)
    Aux4 <- AcumPearsonVII(a,0,1,nu+(2*r),nu)
    resp <- Aux2*Aux3*Aux4
 }
```

```
  if(type=="Slash")
  {
    Aux0 <- nu/(nu+r)
    Aux1 <- AcumSlash(a,0,1,nu+r)
    resp <- Aux0*Aux1
  }
  if(type=="NormalC")
  {
    Aux0 <- nu[2]^(r)*AcumNormalC(a,0,1,nu)
    Aux1 <- (1-nu[2]^(r))*(1-nu[1])*pnorm(a)
    resp <- Aux0 + Aux1
  }
  return(resp)
}



# calculo da etapa E com censura:
#Resultado proveniente da Proposição 1 da Tese Aldo
CensEsperUY1 <- function(mu,sigma2,nu,Lim1,Lim2,type=type,cens=cens)
{
  Lim11 <- (Lim1-mu)/sqrt(sigma2)
  Lim21 <- (Lim2-mu)/sqrt(sigma2)
  n <- length(Lim11)
  if(type=="Normal")
  {
    EU <-  1
    FNIb   <-  pnorm(Lim21)
    FNIa   <-  pnorm(Lim11)
  }
  if(type=="T")
  {
    EU <-  1
    FNIb   <-  pt(Lim21,nu)
    FNIa   <-  pt(Lim11,nu)
  }
   if(type=="Slash")
  {
```

```
    FNIb   <- AcumSlash(Lim21,0,1,nu)
    FNIa   <- AcumSlash(Lim11,0,1,nu)
  }
  if(type=="NormalC")
  {
    EU <-  (nu[1]*nu[2]) + (1-nu[1])
    FNIb  <- AcumNormalC(Lim21,0,1,nu)
    FNIa  <- AcumNormalC(Lim11,0,1,nu)
  }
  if (cens=="1")
  {
    Aux11 <- rep(0,n)
  }else
  {
    Aux11 <- Lim11
  }
  if (cens=="2")
  {
    Aux22 <- rep(0,n)
  }else
  {
    Aux22 <- Lim21
  }
  K <- 1/(FNIb-FNIa)
  EUX0 <- K*(E_Phi(1,Lim21, nu,type)- E_Phi(1,Lim11, nu,type))
  EUX1 <- K*(E_phi(0.5,Lim11,nu,type,cens)- E_phi(0.5,Lim21,nu,type,cens))
  EUX2 <- K*(E_Phi(0,Lim21, nu,type)- E_Phi(0,Lim11, nu,type)  +
  Aux11*E_phi(0.5,Lim11,nu,type,cens) - Aux22*E_phi(0.5,Lim21,nu,type,cens))
  # Neste caso r =2
  EUX20 <- K*(E_Phi(2,Lim21, nu,type)- E_Phi(2,Lim11, nu,type))
  EUX21 <- K*(E_phi(1.5,Lim11,nu,type,cens)- E_phi(1.5,Lim21,nu,type,cens))
  EUY0  <- EUX0
  EUY1  <- mu*EUX0 + sqrt(sigma2)*EUX1
  EUY2  <- EUX0*mu^2 + 2*mu*sqrt(sigma2)*EUX1 + sigma2*EUX2
  EUY20 <- EUX20
  EUY21 <- mu*EUX20 + sqrt(sigma2)*EUX21
  return(list(EUY0=EUY0,EUY1=EUY1,EUY2=EUY2, EUY20=EUY20, EUY21=EUY21))
}
```

```
Qgama<-function(gama,y,x,p,u0,u1,u2)
{
  n=length(y)
  beta=as.matrix(c(p[1],p[2],p[3]),nrow=3)
  sigma2=as.numeric(p[4])

  tabela<-table(x<=gama)
  n1=as.numeric(tabela[2])
  n2=as.numeric(tabela[1])
  w=matrix(1,nrow=n,ncol=1)
  w1=matrix(0,nrow=n1,ncol=1)
  w2=matrix(x[(n1+1):n]-gama,nrow=n2,ncol=1)
  waux=rbind(w1,w2)
  X=matrix(cbind(w,x,waux),n,3)  #matriz de planejamento
  mi=X%*%beta
  Qvbeta=sum(u2-2*u1*mi+mi^2*u0)
  fQgama=(-n/2)*log(sigma2)-0.5*(1/sigma2)*Qvbeta
  return(fQgama)
}  # função Q usada para estimar gama (ponto de mudança)



EMcensura<-function(y,x,nu,type,error,cens,cc)
{

y <- as.vector(y)
        n <- length(y)

# estimação de nu (entrada)
TOLERANCIA<-1e-6
   MAX_NU<-150
   MIN_NU <- 1.01

Lim1  <- Lim2 <- c()

if (cens=="1") # left
   {
     Lim1 <- rep(-Inf,n)
     Lim2 <- y
```

```
   }

if (cens=="2") # right
   {
     Lim1 <- y
     Lim2 <- rep(Inf,n)
   }

chute<-inicial(x,y)
   beta=chute$beta
   gama=chute$gama
   sigma2=chute$sigma2

   p_0=cbind(beta[1],beta[2],beta[3],sigma2,gama,nu)
   crit<-1
   iter<-0

   tabela<-table(x<=gama)
   n1=as.numeric(tabela[2])
   n2=as.numeric(tabela[1])
   n=length(y)
   w=matrix(1,nrow=n,ncol=1)
   w1=matrix(0,nrow=n1,ncol=1)
   w2=matrix(x[(n1+1):n]-gama,nrow=n2,ncol=1)
   waux=rbind(w1,w2)
   X=matrix(cbind(w,x,waux),n,3)  #matriz de planejamento
   mi=X%*%beta


   while(crit>=error)   {
     iter<-iter+1

# inicio da etapa E

NCensEUY <- NCensurEsperUY(y,mu=mi,sigma2=sigma2,nu=nu,type=type)
       u0 <- NCensEUY$EUY0
       u1 <- NCensEUY$EUY1
       u2 <- NCensEUY$EUY2
```

```
if(sum(cc)>0)
       {
         CensEUY <- CensEsperUY1(mu=mi[cc==1],sigma2=sigma2,nu=nu,
         Lim1=Lim1[cc==1],Lim2=Lim2[cc==1],type=type, cens=cens)
         u0[cc==1]<- CensEUY$EUY0
         u1[cc==1]<- CensEUY$EUY1
         u2[cc==1]<- CensEUY$EUY2
       }


# fim da etapa E

# inicio da etapa M

a=min(x)
b=max(x)-0.1
p=c(beta[1],beta[2],beta[3],sigma2)
gama<-optim(gama,Qgama,gr = NULL,y,x,p,u0,u1,u2,method="L-BFGS-B",lower = a,
upper = b,control=list(fnscale=-1))$par

tabela<-table(x<=gama)
n1=as.numeric(tabela[2])
n2=as.numeric(tabela[1])
w=matrix(1,nrow=n,ncol=1)
w1=matrix(0,nrow=n1,ncol=1)
w2=matrix(x[(n1+1):n]-gama,nrow=n2,ncol=1)
waux=rbind(w1,w2)
X=matrix(cbind(w,x,waux),n,3)  #matriz de planejamento
mi=X%*%beta

      suma1 <- t(t(X)%*%u1)
      #u0.matriz <- Diagonal(n,as.numeric(sqrt(u0)))
      u0.matriz<-matrix(0,ncol=n, nrow=n)
      diag(u0.matriz)=c(as.numeric(sqrt(u0)))
      xnovo <- as.matrix(u0.matriz%*%X)
      suma2 <- t(xnovo)%*%xnovo

      beta <- matrix(t(solve(suma2)%*%t(suma1)),3,1)
```

```
      sigma2 <- sum(u2-2*u1*mi+mi^2*u0)/n



      auxf0 <- (y-X%*%(beta))/sqrt(sigma2)
      #auxf <- (Lim1-X%*%t(beta))/sqrt(sigma2)
      #auxf1 <- (Lim2-X%*%t(beta))/sqrt(sigma2)

## calculo das logs e estimação de nu

if(type=="Normal"){

if(sum(cc)>0)
      {
        if(cens=="1")
        {
          logver <- sum(log(dnorm(auxf0[cc==0])/sqrt(sigma2)))+
          sum(log(pnorm(auxf0[cc==1])))
        }
        if(cens=="2")
        {
          logver <- sum(log(dnorm(auxf0[cc==0])/sqrt(sigma2)))+
          sum(log(1-pnorm(auxf0[cc==1])))
        }
      }else{
        logver <- sum(log(dnorm(auxf0[cc==0])/sqrt(sigma2)))
      }
}


if (type == "T")
{
if(sum(cc)>0)
      {
        ft1 <- function(nu){sum(log(dt(auxf0[cc==0],df=nu)/sqrt(sigma2)))+
        sum(log(pt(auxf0[cc==1],df=nu))) }
        ft2 <- function(nu){sum(log(dt(auxf0[cc==0],df=nu)/sqrt(sigma2)))+
        sum(log(pt(-auxf0[cc==1],df=nu))) }
         }else{
```

```
        ft1 <- ft2  <- function(nu){sum(log(dt(auxf0[cc==0],df=nu)/sqrt(sigma2)))}
      }

      if (cens=="1")
      {
        nu <- optimize(f=ft1, interval=c(MIN_NU,MAX_NU),lower = MIN_NU,
        upper=MAX_NU,maximum=TRUE,tol=TOLERANCIA)$maximum
        logver=ft1(nu)
      }
      if (cens=="2")
      {
        nu <- optimize(f=ft2, interval=c(MIN_NU,MAX_NU),lower = MIN_NU,
        upper=MAX_NU,maximum=TRUE,tol=TOLERANCIA)$maximum
        logver=ft2(nu)
      }
}

if (type == "Slash")
{
if(sum(cc)>0)
      {
        fs1 <- function(nu){sum(log(dSlash(auxf0[cc==0],0,1,nu)/sqrt(sigma2)))+
        sum(log(AcumSlash(auxf0[cc==1],0,1,nu)))}
        fs2 <- function(nu){sum(log(dSlash(auxf0[cc==0],0,1,nu)/sqrt(sigma2)))+
        sum(log(AcumSlash(-auxf0[cc==1],0,1,nu)))}

      }
      else
      {
        fs1 <- function(nu){sum(log(dSlash(auxf0[cc==0],0,1,nu)/sqrt(sigma2)))}
        fs2 <- function(nu){sum(log(dSlash(auxf0[cc==0],0,1,nu)/sqrt(sigma2)))}

      }
      if (cens=="1")
      {
        nu <- optimize(fs1, c(1.1,30), tol = TOLERANCIA, maximum = TRUE)$maximum
        logver=fs1(nu)
      }
      if (cens=="2")
```

```
      {
        nu <- optimize(fs2, c(1.1,30), tol = TOLERANCIA, maximum = TRUE)$maximum
        logver=fs2(nu)
      }
}



# fim da etapa M


p<-cbind(beta[1],beta[2],beta[3],sigma2,gama,nu)
#print(p)
crit<-sum((p-p_0)^2)
p_0=p
} # o algoritmo EM
obj.out <- list(p=p,logver=logver)
return(obj.out)
  }

#gera a slash assimétrica, também conhecida como skew-slash.
#Entetanto, com lambda =0, temos a slash simétrica
rssl <- function(n, mu=0, sigma2=1, lambda=0, nu=30)
{
 y <- rep(0,n)
 u <- rbeta(n=n,shape1=nu,shape2=1)
 deltinha <- lambda/sqrt(1+lambda^2)
  Delta <-  sqrt(sigma2)*deltinha
  tau <- sigma2*(1-deltinha^2)

  T0 <- rnorm(n)
  T1 <- rnorm(n)
  T2 <- abs(T0)*u^(-1/2)
  y <-  mu + Delta*T2 + u^(-1/2)*sqrt(tau)*T1
return(y)
}

criterios<-function(logver,n,p){
AIC <- (-2)*logver + 2*(p+2)
```

```
BIC <- (-2)*logver + (p+2)*log(n)
EDC <- (-2)*logver + (p+2)*0.2*sqrt(n)
obj.out <- list(AIC=AIC, BIC=BIC, EDC=EDC)
return(obj.out)
}
```

APÊNDICE B – Rotinas para os estudos de simulação

```
#### IMPORTANTE
##Nos estudos de simulação, trocamos as últimas linhas da função EMcensura por
#  obj.out <- as.matrix(p)
## ao invés de
# obj.out <- list(p=p,logver=logver)
#  Com esta estratégia, temos como saída somente o vetor de estimativas
# dos parâmetros, facilitando o cálculo das médias e dos desvios padrão
# amostrais dessas estimativas.

mont=500
n=200 # tamanho da amostra
x=sort(runif(n,1,10))  #variavel explicativa

###### Gerando normal #####
estinormal_n200_g40_c00=matrix(0,mont,6)
estinormal_n200_g40_c05=matrix(0,mont,6)
estinormal_n200_g40_c10=matrix(0,mont,6)
estinormal_n200_g40_c20=matrix(0,mont,6)
estinormal_n200_g40_c30=matrix(0,mont,6)



####Gerando Normal
for (j in 1:mont){

  # Geração da amostra
  print(j)

#caso normal

nu=30000 #nu=3 psra t-Student e nu=2 para Slash
rs=rnorm(n)  # gerando a normal
#rs=rt(n,df=nu)  #gerando a t de student
#rs=rssl(n,mu=0, sigma2=1, lambda=0, nu=nu)  #gerando a slash
sigma=1   #parametro de escala sigma2
beta0=2
beta1=-1
```

```
beta2=3
beta=matrix(c(beta0,beta1,beta2),3,1)
gama=quantile(x,0.40) #ponto de mudança
tabela<-table(x<=gama)
n1=as.numeric(tabela[2])
n2=as.numeric(tabela[1])
w=matrix(1,nrow=n,ncol=1)
w1=matrix(0,nrow=n1,ncol=1)
w2=matrix(x[(n1+1):n]-gama,nrow=n2,ncol=1)
waux=rbind(w1,w2)
X=matrix(cbind(w,x,waux),n,3)  #matriz de planejamento
mi=X%*%beta
y=mi+sigma*rs #"dados observados"

############################ gerando dados com censura
# incluindo censura

 #level censored: 0
aa            <- sort(y,decreasing=TRUE)
cutof         <- aa[ceiling(0*n)]
cc            <- matrix(1,n,1)*(y>=cutof)  # right censoring
#cc            <- matrix(1,n,1)*(y<cutof) #Left censoring
y[cc==1]      <- cutof

#cens="1" #left
cens="2" # right

#level censored: 5

yc5=y
aa5           <- sort(yc5,decreasing=TRUE)
cutof5        <- aa5[ceiling(0.05*n)]
cc5           <- matrix(1,n,1)*(yc5>=cutof5)  # right censoring
#cc5           <- matrix(1,n,1)*(yc5<cutof5) #Left censoring
yc5[cc5==1]     <- cutof5

#level censored: 10

yc10=y
```

```
aa10            <- sort(yc10,decreasing=TRUE)
cutof10         <- aa10[ceiling(0.10*n)]
cc10           <- matrix(1,n,1)*(yc10>=cutof10)  # right censoring
#cc10            <- matrix(1,n,1)*(yc10<cutof10) #Left censoring
yc10[cc10==1]      <- cutof10

#level censored: 20

yc20=y
aa20            <- sort(yc20,decreasing=TRUE)
cutof20         <- aa20[ceiling(0.20*n)]
cc20            <- matrix(1,n,1)*(yc20>=cutof20)  # right censoring
#cc20            <- matrix(1,n,1)*(yc20<cutof20) #Left censoring
yc20[cc20==1]      <- cutof20


#level censored: 30

yc30=y
aa30            <- sort(yc30,decreasing=TRUE)
cutof30         <- aa30[ceiling(0.30*n)]
cc30            <- matrix(1,n,1)*(yc30>=cutof30)  # right censoring
#cc30            <- matrix(1,n,1)*(yc30<cutof30) #Left censoring
yc30[cc30==1]      <- cutof30




type="Normal"

error=10^(-6)
estinormal_n200_g40_c00[j,]<-EMcensura(y,x,nu,type,error,cens,cc)
estinormal_n200_g40_c05[j,]<-EMcensura(yc5,x,nu,type,error,cens,cc5)
estinormal_n200_g40_c10[j,]<-EMcensura(yc10,x,nu,type,error,cens,cc10)
estinormal_n200_g40_c20[j,]<-EMcensura(yc20,x,nu,type,error,cens,cc20)
estinormal_n200_g40_c30[j,]<-EMcensura(yc30,x,nu,type,error,cens,cc30)
```

```
}


grupo1=matrix(c(1),n1,1)
grupo2=matrix(c(2),n2,1)
grupo=rbind(grupo1,grupo2)
plot(x,yc30,pch=c(1,2)[unclass=grupo])


# encontrando as posições com censura
#aux=1:n
#pos=aux[cc==1]
#p0<-length(estinormal$p)
#criteriosnormal<-criterios(estinormal$logver,n,p=p0)

####### Média e desvios padrão das estimativas dos parâmetros
####### do modelo Normal ######

apply(estinormal_n200_g40_c00, 2 , mean)
apply(estinormal_n200_g40_c00, 2 , sd)
apply(estinormal_n200_g40_c05, 2 , mean)
apply(estinormal_n200_g40_c05, 2 , sd)
apply(estinormal_n200_g40_c10, 2 , mean)
apply(estinormal_n200_g40_c10, 2 , sd)
apply(estinormal_n200_g40_c20, 2 , mean)
apply(estinormal_n200_g40_c20, 2 , sd)
apply(estinormal_n200_g40_c30, 2 , mean)
apply(estinormal_n200_g40_c30, 2 , sd)
```