Monique Bettio Massuia

Modelos para dados censurados sob a classe de distribuições misturas de escala skew-normal

CAMPINAS

2015

UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística
e Computação Científica

MONIQUE BETTIO MASSUIA

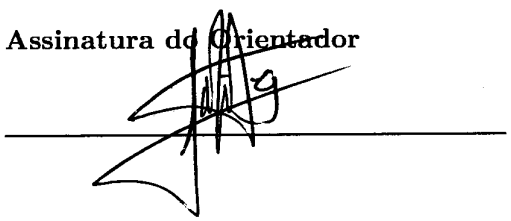
MODELOS PARA DADOS CENSURADOS SOB A CLASSE DE DISTRIBUIÇÕES MISTURAS DE ESCALA SKEW-NORMAL

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em estatística.

Orientador: Víctor Hugo Lachos Dávila

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA MONIQUE BETTIO MASSUIA, E ORIENTADA PELO PROF. DR. VÍCTOR HUGO LACHOS DÁVILA.

Assinatura do Orientador



CAMPINAS

2015

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

M389m

Massuia, Monique Bettio, 1989-
Modelos para dados censurados sob a classe de distribuições misturas de escala skew-normal / Monique Bettio Massuia. – Campinas, SP : [s.n.], 2015.

Orientador: Víctor Hugo Lachos Dávila.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

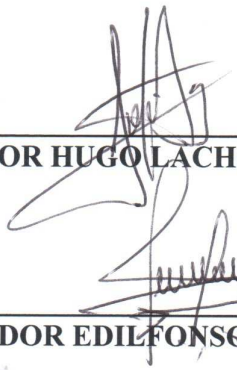
1. Modelos lineares (Estatística). 2. Análise de regressão. 3. Distribuição normal assimétrica. 4. Algoritmos de esperança-maximização. I. Lachos Dávila, Víctor Hugo,1973-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Censored regression models under the class of scale mixture of skew-normal distributions
Palavras-chave em inglês:
Linear models (Statistics)
Regression analysis
Skew-normal distributions
Expectation-maximization algorithms
Área de concentração: Estatística
Titulação: Mestra em Estatística
Banca examinadora:
Víctor Hugo Lachos Dávila [Orientador]
Filidor Edilfonso Vilca Labra
Luis Mauricio Castro Cepero
Data de defesa: 06-03-2015
Programa de Pós-Graduação: Estatística

Dissertação de Mestrado defendida em 06 de março de 2015 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.



Prof(a). Dr(a). VÍCTOR HUGO LACHOS DÁVILA

Prof(a). Dr(a). FILIDOR EDILFONSO VILCA LABRA

Prof(a). Dr(a). LUIS MAURICIO CASTRO CEPERO

Abstract

This work aims to present the linear regression model with censored response variable under the class of scale mixture of skew-normal distributions (SMSN), generalizing the well known Tobit model as providing a more robust alternative to the normal distribution.

A study based on classic inference is developed to investigate these censored models under two special cases of this family of distributions, normal and Student's-t, using the EM algorithm for obtaining maximum likelihood estimates and developing methods of diagnostic based on global and local influence as suggested by Cook (1986) and Poom & Poon (1999). Under a Bayesian approach, the censored regression model was studied under some special cases of SMSN class, such as normal, Student's-t, skew-normal, skew-t and skew-slash. In these cases, the Gibbs sampler was the main tool used to make inference about the model parameters.

We also present some simulation studies for evaluating the developed methodologies that, finally, are applied on two real data sets. The packages `SMNCensReg`,`CensRegMod` and `BayesCR` implemented for the software `R` give computational support to this work.

Keywords: Linear regression Models; Censored response variable; Gibbs sampler; EM algorithm; Local influence; Scale mixture of skew-normal distributions

Resumo

Este trabalho tem como objetivo principal apresentar os modelos de regressão lineares com respostas censuradas sob a classe de distribuições de mistura de escala skew-normal (SMSN),

visando generalizar o clássico modelo Tobit ao oferecer alternativas mais robustas à distribuição Normal.

Um estudo de inferência clássico é desenvolvido para os modelos em questão sob dois casos especiais desta família de distribuições, a normal e a t de Student, utilizando o algoritmo EM para obter as estimativas de máxima verossimilhança dos parâmetros dos modelos e desenvolvendo métodos de diagnóstico de influência global e local com base na metodologia proposta por Cook (1986) e Poom & Poon (1999). Sob o enfoque Bayesiano, o modelo de regressão para respostas censuradas é estudado sob alguns casos especiais da classe SMSN, como a normal, a t de Student, a skew-normal, a skew-t e a skew-slash. Neste caso, o amostrador de Gibbs é a principal ferramenta utilizada para a inferência sobre os parâmetros do modelo.

Apresentamos também alguns estudos de simulação para avaliar a metodologia desenvolvida que, por fim, é aplicada em dois conjuntos de dados reais. Os pacotes `SMNCensReg`,`CensRegMod` e `BayesCR` implementados em `R` dão suporte computacional para este trabalho.

Palavras-chave: Modelos de regressão linear; Variável resposta censurada; Amostrador de Gibbs; Algoritmo EM; Influência Local; Distribuições misturas da escala skew normal.

Este trabalho foi financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) através do processo número 2012/18702-9.

Sumário

Ao meu esposo Adriano, com todo o meu amor.

Agradecimentos

Ao meu esposo, Adriano Azinheira Massuia, o grande incentivador de minha graduação e mestrado. Obrigada por seu amor, seu companheirismo, sua paciência e por suas ideias valiosas que salvaram este trabalho inúmeras vezes.

Ao meu orientador, Victor Hugo Lachos, pelos seus ensinamentos e compreensão e em especial pela confiança depositada em mim.

Ao meu amigo Aldo Medina, por todo o suporte na construção desta dissertação e amizade ao longo destes anos.

À Fapesp, pelo apoio financeiro.

Lista de Ilustrações

Lista de Tabelas

Capítulo 1

Preliminares

1.1 Motivação

O problema de estimação dos parâmetros de um modelo de regressão onde a variável resposta é censurada surge em diversos campos de estudo, como em econometria, engenharia e testes clínicos, dentre outros. No caso em que a variável de interesse é o tempo até o acontecimento de um evento existem diversas técnicas de modelagem na área de análise de sobrevivência, porém, quando este não é o caso, a aplicação destas técnicas pode não ser adequada, principalmente se a variável de interesse puder assumir valores negativos.

O modelo de regressão mais conhecido e utilizado para modelar variáveis que podem assumir valores negativos e/ou censurados é o modelo Tobit (veja Tobin, 1958), onde a principal hipótese assumida é a de que os erros aleatórios seguem uma distribuição normal. Barros *et al.* (2010) faz um estudo de inferência e diagnóstico para este modelo. No entanto, alguns conjuntos de dados não são compatíveis com a suposição de normalidade, seja pela falta de simetria ou pela presença de valores atípicos. Neste sentido, Arellano-Valle *et al.* (2012) propõe a distribuição t de Student como alternativa à normal no modelo Tobit e, generalizando este trabalho, Garay (2014) propõe utilizar a família de distribuições de mistura de escala normal dando atenção especial a alguns de seus casos particulares, como a normal, a t de Student, a normal contaminada e a slash. Dado que estas distribuições são todas simétricas, buscamos neste trabalho apresentar alternativas à

distribuição normal no modelo Tobit que sejam capazes de incorporar parâmetros de curtose e/ou de assimetria: a chamada família de mistura de escala skew-normal (veja Branco & Dey, 2001). Esta família engloba distribuições como a skew-normal, skew-t, skew-slash e também suas versões simétricas, desta forma, ao adotar uma classe de distribuições mais genérica, conseguimos lidar tanto com conjuntos de dados assimétricos e que contêm observações atípicas quanto com conjuntos de dados bem comportados.

Neste primeiro capítulo faremos uma introdução sobre certos aspectos teóricos relevantes para o desenvolvimento dos capítulos seguintes, como o conceito de censuras e truncamento, o algoritmo EM, algoritmos MCMC e alguns métodos para comparação de modelos e diagnóstico.

1.2 Introdução

Começaremos este capítulo introduzindo algumas notações que serão utilizadas ao longo deste trabalho. Em geral, adotaremos a convenção tradicional denotando uma variável ou vetor aleatório por uma letra maiúscula e sua realização pela letra minúscula correspondente. Vetores e matrizes (aleatórios ou não) são representados por letras em negrito. $\boldsymbol{X}^\top$ é a transposição de $\boldsymbol{X}$.

Sejam X e Y duas variáveis aleatórias, a notação $X \perp Y$ indica que são independentes e $X \overset{\mathrm{d}}{=} Y$ indica que têm mesma distribuição. Denotaremos por $f(x)$ a densidade de X, por $F(x)$ sua função de distribuição acumulada, por $f(x, y)$ a densidade conjunto de (X, Y) e por $f(x|y)$ a densidade condicional de $X|Y = y$. Quando $\boldsymbol{\theta}$, o vetor de parâmetros que indexa a distribuição de X, for relevante (e considerado um valor fixo), a densidade de X será denotada por $f(x; \boldsymbol{\theta})$ e sua função de distribuição por $F(x; \boldsymbol{\theta})$. O símbolo $\sim$, como em $X \sim f(x)$, significa que X é distribuída conforme $f(x)$. O símbolo $\overset{\text{iid}}{\sim}$, como em $X, Y \overset{\text{iid}}{\sim} f(x)$, denota que X e Y são independentes e identicamente distribuídas segundo $f(x)$. $\mathbb{E}_X[h(X)]$ e $\mathrm{Var}_X[h(X)]$ denotam respectivamente a esperança e a variância da função $h(X)$ em relação à densidade da variável aleatória X. Embora sejam um tanto abusivas, estas notações facilitarão o desenvolvimento matemático deste trabalho.

Denotamos por $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a distribuição normal $p-$variada com vetor de locação μ e matriz de variância-covariância $\boldsymbol{\Sigma}$, com densidade $\phi_p(\,\cdot\,;\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ e função de distribuição acumulada

$\Phi_p(\,\cdot\,;\boldsymbol{\mu},\boldsymbol{\Sigma})$. $\mathcal{T}_p(\,\cdot\,;\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ denota a função de distribuição acumulada de uma t de Student $p-$variada, com vetor de parâmetros de locação $\boldsymbol{\mu}$, matriz de escala $\boldsymbol{\Sigma}$ e grau de liberdade ν; a respectiva densidade é denotada por $t_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\nu)$. Se o subescrito p for omitido, então estas funções referem-se à versão univariada destas distribuições, e neste caso, se forem também omitidos os parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, estamos nos referindo à sua versão padrão (com parâmetro de locação 0 e de escala 1). $G(\alpha,\beta)$ denota a distribuição gama com esperança α/β e $IG(\alpha,\beta)$ denota a distribuição gama inversa com esperança $\beta/(\alpha-1)$.

A notação $\mathbb{1}_{\mathbb{A}}(x)$ denota a função indicadora em x no conjunto $\mathbb{A}$, isto é, $\mathbb{1}_{\mathbb{A}}(x)=1$ se $x\in\mathbb{A}$ e $\mathbb{1}_{\mathbb{A}}(x)=0$ caso contrário. $\Gamma(\cdot)$ denota a função gama. Abreviaremos "função de distribuição de probabilidade" por "pdf" e "função de distribuição acumulada" por "cdf".

1.3 Censura e truncamento

1.3.1 Variáveis censuradas e truncadas

Em diversos campos da ciência nos deparamos com situações em que a variável de interesse não pode ser completamente observada para todos os indivíduos do experimento, mas, ao invés disso observa-se somente um intervalo em que esta variável está contida, caracterizando o que chamamos de *censuras*.

Existem três tipos de censura, o mais comum é a censura à direita, que ocorre quando o intervalo observado é do tipo $[a,\infty)$ para alguma constante finita a conhecida, isto é, quando sabemos que o verdadeiro valor da variável de interesse é maior do que o valor observado a. Este tipo de censura ocorre com muita frequência quando a variável de interesse é o tempo até a ocorrência de um evento, como nos estudos clínicos sobre o tempo de sobrevida ou remissão de pacientes ou em estudos sobre o tempo de vida útil de equipamentos eletrônicos. Nestes casos, as censuras correspondem aos indivíduos que não experimentaram o evento de interesse antes do término do estudo. Um outro exemplo de censura à direita que não envolve a variável "tempo" ocorre quando um instrumento de medição tem uma capacidade máxima fixa e não fornece a quantia de interesse quando esta é ultrapassada.

O segundo tipo de censura é a censura à esquerda, quando o verdadeiro valor da variável de interesse é menor do que o valor observado a. Neste caso, o intervalo observado é do tipo $(-\infty, a]$, onde a é uma constante finita e conhecida. Este tipo de censura ocorre, por exemplo, em testes para detectar o vírus HIV, onde a carga viral de um indivíduo portador não pode ser mensurada se for menor do que um determinado ponto limítrofe.

O último tipo de censura é a intervalar, que ocorre quando só é possível observar um intervalo finito do tipo $[a, b]$ no qual o verdadeiro valor da variável está contida, com $|a| < \infty$, $|b| < \infty$ e $a < b$. Esta censura é menos comum que os outros dois tipos e costuma aparecer em experimentos nos quais não há vigilância contínua das unidades experimentais e o interesse é estudar o tempo até a ocorrência de um evento, de forma que existe a possibilidade de que o evento de interesse ocorra entre uma inspeção e outra.

O truncamento ocorre quando algumas observações (seja na variável resposta ou nas regressoras) não estão disponíveis. Ao contrário das censuras, onde a perda de informação é parcial, no truncamento simplesmente não há qualquer registro para a variável em questão. Um exemplo de dados truncados é retratado em Colosimo & Giolo (2006), onde é usado um banco de dados da previdência social para estudar a expectativa de vida dos moradores de uma certa localidade - neste caso, somente moradores que atingiram a idade da aposentadoria fazem parte da amostra e indivíduos mais jovens são automaticamente excluídos do estudo. Outros exemplos de truncamento podem ser encontrados em Nelson (1990) e Kalbfleisch & Lawless (1992).

Neste trabalho daremos enfoque a modelos para dados com censuras à direita e à esquerda, porém os resultados são expansíveis para conjuntos de dados com censura intervalar.

1.3.2 Distribuições truncadas

Seja X uma variável aleatória com densidade $f(\cdot)$, função de distribuição acumulada $F(\cdot)$ e suporte χ. Se esta variável é sujeita a censura, então observar o intervalo $\mathbb{A} = \lfloor a, b \rfloor \subset \chi$ como produto desta censura quer dizer que obtivemos uma nova informação que deve ser incorporada à

função de densidade de $f(\cdot)$, a de que $X \in \mathbb{A}$. Este processo gera o que chamamos de distribuição truncada. Aqui, a notação $\lfloor \cdot\,, \cdot \rfloor$ expressa um intervalo cujo cada extremo pode ser tanto aberto quanto fechado.

Desta forma, denotando por $Tf(\,\cdot\,;\,\mathbb{A}\,)$ a versão truncada da distribuição $f(\cdot)$ no intervalo $\mathbb{A}$, temos a seguinte relação: se $X \sim f(\cdot)$, então $X|X \in \mathbb{A} \sim Tf(\,\cdot\,;\,\mathbb{A}\,)$ e

$$Tf(x\,;\,\mathbb{A}) = \frac{f(x)}{F(b) - F(a)} \mathbb{1}_{\{a \leq x \leq b\}}.$$

Note que se X for censurada à direita, então $b = \infty$ e $Tf(x\,;\,\mathbb{A}) = \frac{f(x)}{1-F(a)} \mathbb{1}_{\{x \geq a\}}$ e, se for censurada à esquerda, então $a = -\infty$ e $Tf(x\,;\,\mathbb{A}) = \frac{f(x)}{F(b)} \mathbb{1}_{\{x \leq b\}}$.

1.4 O algoritmo EM

Na área de inferência estatística clássica é bastante comum lidar com problemas de maximização de funções a fim de estimar os parâmetros desconhecidos do modelo. Em muitos destes casos não há forma analítica fechada para tais estimadores e métodos iterativos de maximização tornam-se uma boa alternativa. Dentre estes métodos, o algoritmo EM ("*Expectation-Maximization*", Dempster *et al.* (1977)) é uma alternativa relativamente simples pois não requer o cálculo de segundas derivadas e pode ser empregado sempre que existir uma representação dos dados em termos de uma variável latente.

Cada iteração do algoritmo consiste em duas etapas. Na primeira etapa, E ou "*Expectation*", os dados observados e a estimativa atual do parâmetro são utilizados para encontrar a distribuição dos dados latentes, enquanto na segunda etapa, M ou "*Maximization*", uma re-estimação do parâmetro é feita sob a hipótese de que a distribuição da variável latente encontrada no passo anterior é de fato sua distribuição verdadeira.

Denote por $\theta \in \Theta$ o vetor de parâmetros de interesse, por $\boldsymbol{Y}$ o vetor com os dados observáveis e por $\ell(\theta; \boldsymbol{y})$ a função de log-verossimilhança assumida pelo modelo estatístico. Suponha que cada componente do vetor aleatório observável $\boldsymbol{Y}$ possa ser escrita como uma função do vetor de variáveis latentes $\boldsymbol{Z}$. Mais especificamente, assumimos que existe uma função $h : \mathcal{Z} \to \mathcal{Y}$ tal que $h(\boldsymbol{Z}) = \boldsymbol{Y}$, de forma que, uma vez observado $\boldsymbol{Y} = \boldsymbol{y}$, a única coisa que sabemos sobre $\boldsymbol{Z}$ é que este

vetor aleatório está restrito ao espaço determinado por $\mathcal{Z}(\boldsymbol{y}) \subseteq \mathcal{Z}$, o subespaço de $\mathcal{Z}$ determinado pela equação $h(\boldsymbol{Z}) = \boldsymbol{y}$.

A estratégia deste algoritmo é considerar os chamados "dados aumentados", isto é, tomar o vetor de variáveis latentes $\boldsymbol{Z}$ como se este fosse observável. Desta forma conhecemos a função de distribuição deste vetor, $f(\boldsymbol{z};\boldsymbol{\theta})$, e temos à disposição a função de log-verossimilhança completa, $\ell_c(\boldsymbol{\theta};\boldsymbol{z}) \equiv \log f(\boldsymbol{z};\boldsymbol{\theta})$. Pode-se também calcular a distribuição condicional de $\boldsymbol{Z}$ dado $\boldsymbol{Y} = \boldsymbol{y}$, $f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})$, da seguinte forma:

$$\begin{aligned} f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}) &= \frac{f(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{z};\boldsymbol{\theta})f(\boldsymbol{z};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} \\ &= \frac{f(\boldsymbol{z};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})}, \quad \boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y}), \end{aligned} \tag{1.4.1}$$

sendo que a segunda igualdade vem do Teorema de Bayes e a terceira igualdade vem do fato de que o vetor $\boldsymbol{Y}$ fica completamente determinado se temos a informação de que $\boldsymbol{Z} = \boldsymbol{z}$, de forma que $f(\boldsymbol{y}|\boldsymbol{z};\boldsymbol{\theta})$ é uma distribuição que atribui probabilidade 1 em $\boldsymbol{Y} = h(\boldsymbol{z})$ e 0 em qualquer outro lugar.

Seja $\boldsymbol{\theta}^{(t)}$ a estimativa de $\boldsymbol{\theta}$ na iteração t do algoritmo, $t = 0, 1, 2, \ldots$, onde $\boldsymbol{\theta}^{(0)}$ é um valor inicial. A iteração $(t+1)$ do algoritmo consiste em dois passos:

Etapa E: Cálculo da função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y}}\left[\ell_c(\boldsymbol{\theta};\boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}\right]$.

Etapa M: Atualizar a estimativa, fazendo $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

Estes dois passos são repetidos até que alguma medida de convergência seja satisfeita, como por exemplo $||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}|| < \epsilon$, para algum $\epsilon > 0$ suficientemente pequeno.

O principal ganho do algoritmo EM é trocar uma única maximização complicada da função de verossimilhança por várias maximizações simples da função $Q(\,\cdot\,|\boldsymbol{\theta}^{(t)})$. No entanto, é preciso garantir que as estimativas encontradas na convergência do algoritmo sejam as mesmas que maximizam a função de verossimilhança. Neste sentido, o algoritmo EM possui uma propriedade interessante chamada "ascendência", que diz que o valor da verossimilhança avaliada em uma nova

atualização da estimativa não deve ser menor do que se avaliada em estimativas anteriores. Esta propriedade é formulada no Teorema 1.

Teorema 1. *Se $\ell(\theta;\boldsymbol{y})$ é a função de verossimilhança do modelo e $\boldsymbol{\theta}^{(t)}$ denota a estimativa de $\boldsymbol{\theta}$ na $t-$ésima iteração do algoritmo EM, temos que, $\forall t \in \mathbb{Z}^+$:*

$$\ell(\boldsymbol{\theta}^{(t)};\boldsymbol{y}) \;\geq\; \ell(\boldsymbol{\theta}^{(t+1)};\boldsymbol{y}) \tag{1.4.2}$$

Prova:

Aplicando o log em ambos os lados da equação 1.4.1, temos:

$$\begin{aligned}
\log\left(f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})\right) &= \log\left(f(\boldsymbol{z};\boldsymbol{\theta})\right) - \log\left(f(\boldsymbol{y};\boldsymbol{\theta})\right) \quad \Rightarrow \\
\log\left(f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})\right) &= \ell_c\left(\boldsymbol{\theta}|\boldsymbol{z}\right) - \ell\left(\boldsymbol{\theta}|\boldsymbol{y}\right) \quad \Rightarrow \\
\ell\left(\boldsymbol{\theta}|\boldsymbol{y}\right) - \ell_c\left(\boldsymbol{\theta}|\boldsymbol{z}\right) &= -\log\left(f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})\right). \qquad (1.4.3)
\end{aligned}$$

Tomando a esperança em ambos os lados da Equação (1.4.3) com relação à distribuição $f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}^{(t)})$, temos:

$$\ell\left(\boldsymbol{\theta}|\boldsymbol{y}\right) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \;=\; -H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \tag{1.4.4}$$

onde $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y}}\left[\log f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}) \mid \boldsymbol{y},\boldsymbol{\theta}^{(t)}\right]$. Agora,

$$\begin{aligned}
H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}}\left[\log f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}^{(t)}) \mid \boldsymbol{y},\boldsymbol{\theta}^{(t)}\right] - \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}}\left[\log f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}) \mid \boldsymbol{y},\boldsymbol{\theta}^{(t)}\right] \\
&= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}}\left[-\log\left(\frac{f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}^{(t)})}\right) \,\middle|\, \boldsymbol{y},\boldsymbol{\theta}^{(t)}\right] \\
&\geq -\log\left(\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}}\left[\frac{f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}^{(t)})} \,\middle|\, \boldsymbol{y},\boldsymbol{\theta}^{(t)}\right]\right), \text{ pela Desigualdade de Jensen} \\
&= -\log\int_{\mathcal{Z}(\boldsymbol{y})} \frac{f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}^{(t)})}\, f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}^{(t)})\, d\boldsymbol{z} \\
&= -\log\int_{\mathcal{Z}(\boldsymbol{y})} f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})\, d\boldsymbol{z} \\
&\geq 0,
\end{aligned}$$

e, portanto, $H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \geq H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. Desta forma, substituindo este resultado na Equação (1.4.4):

$$\begin{aligned} \ell\left(\boldsymbol{\theta}|\boldsymbol{y}\right) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= -H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\ &\geq -H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = \ell\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{y}\right) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}), \end{aligned}$$

logo,

$$\ell\left(\boldsymbol{\theta}|\boldsymbol{y}\right) - \ell\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{y}\right) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}). \tag{1.4.5}$$

Em outras palavras, a Equação (1.4.5) nos diz que ao escolhermos um valor de $\boldsymbol{\theta}$ que aumente o valor da função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ além da constante $Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$, estaremos aumentando também o valor da função $\ell\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ além da constante $\ell\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{y}\right)$ em pelo menos o mesmo tanto. Desta forma acabamos de provar que o algoritmo EM é correto no sentido de que, ao maximizar a função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ em cada iteração, maximiza também a função de verossimilhança e, na convergência, deve encontrar as estimativas de máxima verossimilhança do parâmetro $\boldsymbol{\theta}$ (supondo que não há problemas graves de multimodalidade que possam "desviar" a convergência para um máximo local).

Para finalizar a demonstração do Teorema 1, note que, ao atualizarmos o valor da estimativa com $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, temos por definição que $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \ \forall\, \boldsymbol{\theta} \in \Theta$, inclusive para $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Logo, $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ e a igualdade só é satisfeita se $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$, comprovando a propriedade ascendente do algoritmo dada a discussão do parágrafo anterior.

■

1.5 Algoritmos MCMC

O objetivo maior dos algoritmos MCMC (Monte Carlo em Cadeias de Markov) é o de gerar uma cadeia de observações de uma densidade da qual não seja possível amostrar diretamente. Dentre esta classe de algoritmos, o de Metropolis-Hastings, inicialmente desenvolvido por Metropolis *et al.* (1953) e posteriormente generalizado por Hastings (1970), talvez seja um dos mais conhecidos e utilizados na estatística, principalmente devido ao impacto de um dos seus casos especiais, o amostrador de Gibbs, na área de inferência Bayesiana (Tanner & Wong (1987) e Gelfand & Smith

(1990)).

Na teoria de cadeias de Markov com espaço de estados contínuas, os problemas geralmente começam com um kernel de transição $P(x, A)$, com $x \in \mathbb{R}^d$ e $A \in \mathscr{B}$, onde $\mathscr{B}$ é uma σ-álgebra de Borel em $\mathbb{R}^d$. Este kernel de transição é simplesmente uma função de distribuição condicional que representa a probabilidade de a cadeia sair do ponto x e ir para um ponto do conjunto A. Suponha que possamos escrever:

$$P(x, dy) = p(x, y)dy + r(x)\mathbb{1}_{x \in dy},$$

para alguma função $p(\cdot, \cdot)$, onde $p(x, x) = 0$ e $r(x) = 1 - \int_{\mathbb{R}^d} p(x, z)dz$ representa a probabilidade de a cadeia permanecer no estado x. Se a função $p(\cdot, \cdot)$ satisfizer a condição de reversibilidade, isto é, se existir uma função $\pi(\cdot)$ com domínio em $\mathbb{R}^d$ tal que:

$$\pi(x)p(x, y) = \pi(y)p(y, x) \quad \forall x, y \in \mathbb{R}^d, \tag{1.5.1}$$

então dizemos que $\pi(\cdot)$ é a função de densidade invariante de $P(x, \cdot)$ (Tierney (1994)). O resultado interessante desta propriedade, que será explorado pelos algoritmos MCMC, é que, se a cadeia for irredutível e aperiódica, $\pi(\cdot)$ também é a distribuição estacionária de $P(x, \cdot)$, o que quer dizer que após um número suficientemente grande de passos (mudanças de estado), este kernel converge para $\pi(\cdot)$.

A engenhosidade da classe de algoritmos MCMC está em definir $\pi(\cdot)$ como a distribuição alvo da qual se quer amostrar e então construir um kernel apropriado que satisfaça a condição de reversibilidade em (1.5.1) a partir de uma densidade da qual seja simples gerar observações. Desta forma, após um número suficientemente grande de iterações, a cadeia de valores obtida através deste método se aproximará de um conjunto de amostras da distribuição $\pi(\cdot)$. A seguir apresentaremos o funcionamento do algoritmo de Metropolis-Hastings e de seu caso particular mais famoso, o amostrador de Gibbs.

1.5.1 Metropolis-Hastings

Dentro do contexto geral dos algoritmos MCMC, entender o funcionamento do algoritmo de Metropolis-Hastings quer dizer conhecer o processo utilizado por ele para a construção do kernel de transição. Seja então $\pi(x)$ a função de densidade (absolutamente contínua) da variável/vetor aleatório X do qual queremos amostrar, com $X \in \chi$. Suponha que saibamos gerar observações de uma densidade $q(x, y)$ cujo domínio seja igual a χ^2 e tal que $\int_\chi q(x, y)dy = 1$. No contexto de cadeias de Markov, devemos interpretar $q(x, y)$ como a densidade da qual um novo valor y é gerado quando a cadeia encontra-se no estado x. Geralmente esta densidade, vista como um kernel de transição, não satisfaz a propriedade de reversibilidade com relação a $\pi(x)$, isto é, provavelmente existirá um par de valores (x, y) em χ^2 tal que $\pi(x)q(x, y) > \pi(y)q(y, x)$. Informalmente isto que dizer que uma cadeia com o kernel de transição $q(\cdot, \cdot)$ move-se muito mais vezes do estado x para o estado y do que é necessário para convergir para a densidade $\pi(\cdot)$. O que o algoritmo de Metropolis-Hastings faz então é construir um kernel apropriado corrigindo a densidade $q(x, y)$ através da introdução de uma probabilidade $\alpha(x, y)$, a qual, uma vez gerado um novo "candidato" y através de $q(x, y)$, diz se a cadeia deve mover-se para este candidato ou se deve continuar no estado atual x.

As iterações do algoritmo consistem nos seguintes passos:

Inicialização:

Passo 1: Escolha um valor inicial $x^{(0)}$.

Passo 2: Defina uma distribuição proposta $q(\cdot, \cdot)$.

***Na iteração t*, $t = 1, 2, 3 \ldots$:**

Passo 1: Gere y de $q(x^{(t-1)}, \cdot\,)$.

Passo 2: Calcule a probabilidade:

$$\alpha(x^{(t-1)}, y) = min \left\{ 1 \,;\, \frac{\pi(y)\, q(y, x^{(t-1)})}{\pi(x^{(t-1)})\, q(x^{(t-1)}, y)} \right\}.$$

Passo 3: Faça $X^{(t)} = y$ com probabilidade $\alpha(x^{(t-1)}, y)$ ou $X^{(t)} = x^{(t-1)}$ com probabilidade $1 - \alpha(x^{(t-1)}, y)$.

Uma segunda versão bastante interessante do algoritmo de Metropolis-Hastings aparece quando a distribuição alvo $\pi(\cdot)$ é multivariada, isto é, quando queremos amostrar da distribuição conjunta do vetor aleatório $X = (X_1, X_2, \ldots, X_d)$, com $d > 1$, e ao invés disso, é mais fácil amostrar das chamadas "condicionais completas", as distribuições de $X_i|X_{-i}$, onde $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)^\top$, isto é, o vetor X_{-i} contem todos os elementos originais de X, exceto por X_i, para $i = 1, 2, \ldots, d$. Neste caso, ao invés de aplicar o algoritmo simultaneamente a todos os elementos do vetor, constrói-se uma versão "componente-a-componente" do algoritmo, aplicando-o em sequência a cada elemento. Esta versão foi originalmente discutida por Metropolis *et al.* (1953) e simplifica a escolha da distribuição proposta.

Denote por $\pi_i(\cdot|x_{-i})$ a distribuição condicional de $X_i|X_{-i}$ e suponha que existam d kernels de transição $P(x_i, A|x_{-i})$ cuja distribuição invariante seja $\pi_i(\cdot|x_{-i})$, $i = 1, 2, \ldots, d$. Cada um destes kernels representa a probabilidade condicional de a cadeia de Markov definida pela sequência de observações da variável X_i sair do estado atual x_i e ir para algum ponto do conjunto A, dado que $X_{-i} = x_{-i}$. Neste cenário, se cada um dos d componentes do vetor X for movido em uma ordem fixa (ou seja, se fizermos a atualização componente-a-componente), o processo resultante será uma cadeia de Markov cujo kernel de transição corresponde ao produto dos d kernels iniciais. Este processo tem como distribuição invariante $\pi(x)$ e, se a cadeia for irredutível e aperiódica, $\pi(x)$ é também sua distribuição estacionária. É este resultado que serve como base para a construção desta versão do algoritmo.

Cada iteração da versão "componente-a-componente" do Metropolis-Hasting consiste em d atualizações. Denotando por $x_i^{(t)}$ o valor atribuído a X_i na $t-$ésima iteração, o passo i da iteração t consiste em gerar um candidato y_i de uma distribuição proposta $q_i(x_i^{(t-1)}, \cdot\,|x_{-i}^{(t)})$ e aceitá-lo conforme uma probabilidade $\alpha(x_i^{(t-1)}, y_i\,|x_{-i}^{(t)})$, onde $x_{-i}^{(t)} = (x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, \ldots, x_d^{(t-1)})^\top$ denota o valor atribuído ao vetor X_{-i} após completar o passo $i-1$ da iteração atual. De forma mais precisa, os passos são:

***Inicialização**:*

Passo 1: Escolha um valor inicial $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_d^{(0)})^\top$

Passo 2: Defina d distribuições propostas $q_i(\cdot, \cdot \mid x_{-i})$, para $i = 1, 2, \ldots, d$

***Na iteração t*, $t = 1, 2, 3 \ldots$, *para* $i = 1, 2, \ldots, d$**:

Passo 1: Gere y_i de $q(x_i^{(t-1)}, \cdot \mid x_{-i}^{(t)})$.

Passo 2: Calcule a probabilidade:

$$\alpha_i(x_i^{(t-1)}, y_i \mid x_{-i}^{(t)}) = min\left\{1\,;\ \frac{\pi_i(y_i|x_{-i}^{(t)})\ q_i(y_i, x_i^{(t-1)}|x_{-i}^{(t)})}{\pi_i(x_i^{(t-1)}|x_{-i}^{(t)})\ q_i(x_i^{(t-1)}, y_i|x_{-i}^{(t)})}\right\}.$$

Passo 3: Faça $X_i^{(t)} = y_i$ com probabilidade $\alpha_i(x_i^{(t-1)}, y_i \mid x_{-i}^{(t)})$ ou $X_i^{(t)} = x_i^{(t-1)}$ com probabilidade $1 - \alpha_i(x_i^{(t-1)}, y_i \mid x_{-i}^{(t)})$.

Gelman (1992) sugere um método para avaliar a convergência do kernel de transição da cadeia para a distribuição estacionária. Este método requer que pelo menos duas cadeias sejam rodadas paralelamente, de preferência a partir de valores iniciais distintos, pois compara a variância inter e intra cadeias. Desta forma, seja x_{ij} a $i-$ésima observação proveniente da $j-$ésima cadeia, com $i = 1, 2, \ldots, M$ e $j = 1, 2, \ldots, J$. Denote por $\bar{x}_{\cdot j} = \frac{1}{M}\sum_{i=1}^{M} x_{ij}$ e $\bar{x}_{\cdot\cdot} = \frac{1}{JM}\sum_{i,j} x_{ij}$. Calculamos as variâncias inter (W) e intra (B) cadeias da seguinte forma:

$$W = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{M-1}\sum_{i=1}^{M}(\bar{x}_{ij} - \bar{x}_{\cdot j})^2\right] \quad \text{e} \quad B = \frac{M}{J-1}\sum_{j=1}^{J}(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2.$$

A variância de X pode então ser estimada por:

$$\widehat{\text{Var}}[X] = \frac{M-1}{M}W + \frac{1}{M}B.$$

Esta estimativa em geral superestima a variância real de X e, ao mesmo tempo, W tende a subestimar a variância inter cadeia. No entanto, conforme $M \to \infty$, ambas as quantidades convergem para $\text{Var}[X]$, de forma que podemos usar a razão entre $\widehat{\text{Var}}[X]$ e W para monitorar a convergência da cadeia: quanto mais próxima de 1 for a razão, mais indícios temos de que o kernel de transição convergiu para a distribuição estacionária. Neste sentido, a estatística de Gelman-Rubin é definida por:

$$\widehat{R} = \sqrt{\frac{\widehat{\text{Var}}[X]}{W}}.$$

Espera-se que $\widehat{R}$ seja próximo de 1 para assegurar a convergência das cadeias.

Sobre a sequência de amostras retornadas, uma prática comum e bastante eficiente é eliminar os valores resultantes das primeiras iterações do algoritmo, realizadas antes de a convergência ser atingida (conhecida como "*burn-in*"). Além disso, existe uma dependência entre observações sucessivas, fruto da origem markoviana do método. O que se faz, no sentido de eliminar ou minimizar esta correlação, é guardar as observações espaçadas utilizando um passo constante, cujo tamanho pode ser facilmente determinado através da construção de gráficos de auto-correlação para a amostra gerada (processo conhecido como "*thinning*").

Para a demonstração de que o kernel construído no algoritmo de Metropolis-Hastings converge para a distribuição alvo, veja Hastings (1970). Chib & Greenberg (1995) discute problemas de implementação deste algoritmo, como a escolha da distribuição proposta.

1.5.2 Amostrador de Gibbs

O amostrador de Gibbs é um caso particular da versão "componente-a-componente" do algoritmo de Metropolis-Hastings. Sua principal aplicação na estatística é na área da inferência Bayesiana quando deseja-se amostrar da distribuição a posteriori conjunta de um vetor aleatório de interesse $X = (X_1, X_2, \ldots, X_d)$ e as distribuições condicionais completas de $X_i|X_{-i}$, para $i = 1, 2, \ldots, d$, são conhecidas e fáceis de se obter amostras, seja diretamente ou através de algum método iterativo como o próprio Metropolis-Hastings.

Este algoritmo MCMC surge do Metropolis-Hastings quando a distribuição proposta para atualizar o valor de X_i é tomada como a própria condicional completa de $X_i|X_{-i}$, $\pi_i(\cdot|x_{-i})$, para $i = 1, 2, \ldots, d$. Ao fazer esta escolha, a probabilidade de aceitação de um valor candidato é 1, o que pode ser visto facilmente se substituirmos $q_i(y_i, x_i^{(t-1)}|x_{-i}^{(t)})$ por $\pi_i(x_i^{(t-1)}|x_{-i}^{(t)})$ e $q_i(x_i^{(t-1)}, y_i|x_{-i}^{(t)})$ por $\pi_i(y_i|x_{-i}^{(t)})$ na equação que define $\alpha_i(x_i^{(t-1)}, y_i \mid x_{-i}^{(t)})$ nos passos do Metropolis-Hastings "componente-a-componente". Desta forma, as iterações do amostrador de Gibbs consistem em:

***Inicialização**:*

Passo 1: Escolha um valor inicial $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_d^{(0)})^\top$.

Na iteração $\boldsymbol{t}$, $\boldsymbol{t = 1, 2, 3 \ldots}$, ***para*** $\boldsymbol{i = 1, 2, \ldots, d}$:

Passo 1: Gere uma nova amostra $x_i^{(t)}$ para a variável aleatória X_i a partir de $\pi(x_i^{(t-1)} | \, x_{-i}^{(t)})$.

Sendo o amostrador de Gibbs um caso particular do Metropolis-Hastings, aqui ocorrem os mesmos problemas acerca do número de iterações até a convergência e correlação entre amostras sucessivas discutidos na Subseção 1.5.1, portanto as observações feitas anteriormente sobre tais assuntos são pertinentes para este algoritmo.

1.6 Critérios para comparação de modelos

Existe uma grande variedade de metodologias para comparar a adequação de diferentes modelos a um determinado conjunto de dados, geralmente balanceando-se a qualidade do ajuste promovido e a complexidade do modelo. Apresentaremos aqui algumas das principais metodologias utilizadas: o AIC, o BIC e o EDC, para estudos frequentistas, e o EAIC, EBIC, LPML, DIC e WAIC, para estudos Bayesianos.

1.6.1 Critérios frequentistas

Alguns dos principais critérios de comparação de modelos utilizados em estudos frequentistas são o AIC ("*Akaike information criterion*"), o BIC ("*Bayesian information criterion*") e o EDC ("*efficient determination criterion*"). Suponha que temos n observações de um determinado modelo, cujo vetor de parâmetros $\boldsymbol{\theta} \in \mathbb{R}^q$ possui estimativa igual a $\hat{\boldsymbol{\theta}}$. Seja $\ell(\boldsymbol{\theta})$ a função de log-verossimilhança deste modelo. Estes três critérios podem ser expressos por

$$-2\ell(\hat{\boldsymbol{\theta}}) + qc_n,$$

onde q é o número de parâmetros livres a serem estimados no modelo e c_n é uma sequência (conveniente) de números positivos. O AIC é definido tomando-se $c_n = 2$ e o BIC, fazendo-se $c_n = \log(n)$. Para o EDC, c_n é escolhido de modo a satisfazer as condições $c_n/n \to 0$ e

$c_n/\log(n) \to 0$ quando $n \to \infty$. Neste trabalho utilizaremos $c_n = 0.2\sqrt{n}$, conforme sugerido por Bai *et al.* (1989).

1.6.2 Critérios Bayesianos

Existem várias propostas de critérios de comparação de modelos Bayesianos, úteis quando é preciso escolher entre modelos distintos para o mesmo conjunto de dados (veja Ando (2010)). Considere um modelo com vetor de parâmetros $\boldsymbol{\theta}$ e seja $\mathbf{z} = \{z_1, \ldots, z_n\}$ um conjunto com n observações da variável de interesse. Um dos critérios mais conhecidos é o LPML ("*log pseudo marginal likelihood*"), derivado a partir da estatística CPO ("*conditional predictive ordinate*"). Para a $i-$ésima observação, $i = 1, 2, \ldots, n$, a estatística CPO_i é definida como a densidade preditiva $f(z_i|\mathbf{z}_{-i})$, representando uma medida do quão provável seria obter uma futura observação igual a z_i dado a amostra $\mathbf{z}_{-i}$, onde $\mathbf{z}_{-i}$ é o conjunto obtido ao excluir-se z_i de $\mathbf{z}$. Desta forma, é fácil enxergar esta estatística como um método de identificação de observações aberrantes, um dos seus usos mais conhecidos na inferência Bayesiana. Para maiores detalhes sobre o CPO, veja Gelfand *et al.* (1992). Pode-se mostrar que o CPO pode ser escrito da seguinte forma:

$$\text{CPO}_i = \int f(z_i|\boldsymbol{\theta})\, f(\boldsymbol{\theta}|\mathbf{z}_{-i})\, d\boldsymbol{\theta} = \left(\int \frac{f(\boldsymbol{\theta}|\mathbf{z})}{f(z_i|\boldsymbol{\theta})}\, d\boldsymbol{\theta} \right)^{-1} = \left(\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Z}} \left[f(z_i|\boldsymbol{\theta})^{-1} \middle|\ \mathbf{z} \right] \right)^{-1}. \tag{1.6.1}$$

Para a maioria dos modelos a estatística CPO_i não possui uma forma analítica fechada. No entanto pode-se obter uma aproximação desta estatística usando uma amostra MCMC de $f(\boldsymbol{\theta}|\mathbf{z})$, a distribuição a posteriori de $\boldsymbol{\theta}$: $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_Q\}$ (após o processo de *burn-in* e *thinning*). É dado em Dey *et al.* (1997) que:

$$\widehat{\text{CPO}_i} = \left(\frac{1}{Q} \sum_{j=1}^{Q} \frac{1}{f(z_i|\boldsymbol{\theta}_j)} \right)^{-1}. \tag{1.6.2}$$

O critério LPML é uma sumarização destas n estatísticas, definido por:

$$\text{LPML} = \sum_{i=1}^{n} \log(\widehat{\text{CPO}_i}).$$

Quanto maior o valor de LMPL, melhor é a adequação do modelo proposto ao conjunto de dados.

Os critérios DIC, EAIC, EBIC e WAIC ponderam a qualidade do ajuste do modelo e sua complexidade. Para defini-los vamos primeiramente definir uma medida de qualidade de ajuste, o *deviance*, dada por:

$$\mathrm{D}(\boldsymbol{\theta}) = -2\log\left(\prod_{i=1}^{n} f(z_i|\boldsymbol{\theta})\right).$$

O DIC ("*deviance information criterion*", Spiegelhalter *et al.* (2002)) pondera o *deviance* com uma medida relacionada com a complexidade do modelo, o número de parâmetros efetivos, definida por:

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

onde o primeiro termo é a esperança a posteriori do *deviance*, dada por

$$\overline{\mathrm{D}}(\boldsymbol{\theta}) = -2\sum_{i=1}^{n} \mathrm{E}\left[\log f(z_i|\boldsymbol{\theta})|\mathbf{z}\right],$$

e o segundo termo é o *deviance* avaliado em alguma estimativa pontual $\tilde{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$. A média a posteriori é uma escolha natural para $\tilde{\boldsymbol{\theta}}$, mas existem outras alternativas, como a moda ou a mediana a posteriori. Finalmente, definimos o DIC por:

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

Quanto menor o valor do DIC, mais adequado é o modelo ao conjunto de dados. O cálculo da integral $\overline{\mathrm{D}}(\boldsymbol{\theta})$ pode ser bastante complexo, por isso, pode-se usar também aqui uma amostra MCMC $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_Q\}$ de $f(\boldsymbol{\theta}|\mathbf{z})$ para estimar o valor do DIC. Desta forma, aproximamos $\overline{\mathrm{D}}(\boldsymbol{\theta})$ pela média amostral a posteriori dos desvios $D(\boldsymbol{\theta})$:

$$\widehat{\overline{\mathrm{D}}}(\boldsymbol{\theta}) = -\frac{2}{Q}\sum_{j=1}^{Q}\log\left(\prod_{i=1}^{n}\pi(z_i|\boldsymbol{\theta}_j)\right). \tag{1.6.3}$$

Assim, uma aproximação do DIC é dada por:

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

Mais recentemente Watanabe (2010) introduziu outro critério para seleção de modelos, o WAIC ("*Watanabe-Akaike information criterion*"). Para defini-lo, vamos primeiramente definir a *log-*

densidade preditiva, dada por:

$$p(\mathbf{z}) = \sum_{i=1}^{n} \log \int f(z_i|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{z})\, d\boldsymbol{\theta}.$$

Basicamente, o WAIC calcula $p(\mathbf{z})$ e adiciona uma correção relacionada ao número efetivo de parâmetros do modelo, a fim de compensar um possível superajuste. Gelman *et al.* (2014) sugere duas maneiras de calcular esta correção. A primeira é muito parecida com a utilizada em ρ_{DIC} e é dada por

$$\rho_{\text{WAIC}_1} = 2\, p(\mathbf{z}) + \overline{\text{D}}(\boldsymbol{\theta}).$$

A segunda é definida por

$$\rho_{\text{WAIC}_2} = \sum_{i=1}^{n} Var\left[\log f(z_i|\boldsymbol{\theta})|\boldsymbol{z}\right].$$

Finalmente, as duas versões do critério WAIC são definidas da seguinte forma:

$$\text{WAIC}_k = 2\, \rho_{\text{WAIC}_k} - 2\, p(\mathbf{z}), \qquad k = 1, 2. \tag{1.6.4}$$

Quanto menor for o valor do WAIC, mais adequado é o modelo ao conjunto de dados. É importante notar que na definição original de Watanabe o critério WAIC foi definido somente por $-p(\mathbf{z})/n$ adicionada a uma correção. Aqui, seguindo a sugestão dada por Gelman *et al.* (2014), este termo foi multiplicado por -2 de forma a ficar na mesma escala que os critérios DIC, EAIC e EBIC.

Novamente, computar o WAIC envolve envolve calcular integrais que geralmente não são diretas ou são computacionalmente custosas. Desta forma, é possível aproximar o valor deste critério como foi feito para o DIC: utilizando uma amostra MCMC da densidade a posteriori de $\boldsymbol{\theta}$, $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_Q\}$. Primeiramente, aproximamos $p(\mathbf{z})$ da seguinte forma:

$$\widehat{p(\mathbf{z})} = \sum_{i=1}^{n} \log\left(\frac{1}{Q}\sum_{j=1}^{Q} f(z_i|\boldsymbol{\theta}_j)\right),$$

e depois, considerando a aproximação de $\overline{\text{D}}(\boldsymbol{\theta})$ dada na Equação (1.6.3), a primeira versão do WAIC pode ser aproximada por

$$\widehat{\text{WAIC}}_1 = 2\, \widehat{p(\mathbf{z})} + 2\, \widehat{\overline{\text{D}}}(\boldsymbol{\theta}).$$

A aproximação da segunda versão do WAIC, $\widehat{\text{WAIC}}_2$, pode ser calculada considerando a variância amostral $V_{j=1}^{Q}(x) = \frac{1}{Q-1}\sum_{j=1}^{Q}(x-\bar{x})$ como uma estimativa da variância, onde $\bar{x} = \frac{1}{Q}\sum_{j=1}^{Q} x_j$, e utilizando a amostra MCMC da densidade à posteriori de $\boldsymbol{\theta}$ para aproximar o valor de $f(z_i|\boldsymbol{\theta})$, $i = 1, 2, \ldots, n$, fazendo:

$$\widehat{f(z_i|\boldsymbol{\theta})} = \frac{1}{Q}\sum_{j=1}^{Q} f(z_i|\boldsymbol{\theta}_j).$$

Outros critérios podem ser usados para comparar a adequação de modelos distintos a um mesmo conjunto de dados: o EAIC ("*expected Akaike information criterion*"), veja Brooks (2002), e o EBIC ("*expected Bayesian information criterion*"), veja Carlin & Louis (2001). Estes critérios são definidos por:

$$EAIC = \overline{\mathrm{D}}(\boldsymbol{\theta}) + 2q \quad \text{e} \quad EBIC = \overline{\mathrm{D}}(\boldsymbol{\theta}) + q\log(n),$$

onde q é o número de parâmetros livres a serem estimados no modelos. Quanto menor o valor do EAIC ou do EBIC, mais adequado é o modelo ao conjunto de dados. Substituindo $\overline{\mathrm{D}}(\boldsymbol{\theta})$ por $\widehat{\overline{\mathrm{D}}}(\boldsymbol{\theta})$, pode-se obter uma estimativa destes critérios.

Quando a intenção não é comparar modelos distintos mas sim investigar se um determinado modelo é suficientemente adequado para um conjunto de dados, usamos uma medida baseada na distribuição preditiva à posteriori, o p-valor Bayesiano preditivo (Gelman *et al.*, 2004). Usando alguma estatística pré-fixada como medida de discrepância, $T(\mathbf{z}, \boldsymbol{\theta})$, pode-se determinar se seu valor observado é extremo em relação à distribuição preditiva à posteriori com o auxílio de amostras simuladas desta distribuição.

Neste trabalho seguiremos a sugestão de Gelman *et al.* (2004) e usaremos uma função da log-verossimilhança como medida de discrepância, dada por:

$$T(\mathbf{z}, \boldsymbol{\theta}) = -2\sum_{i=1}^{n} \log\left[f(z_i \mid \boldsymbol{\theta})\right]. \tag{1.6.5}$$

O p-valor Bayesiano preditivo, denotado por $\mathrm{p_B}$, é definido como a porcentagem de vezes em que $T(\mathbf{z}_{pr}, \boldsymbol{\theta})$ é maior do que $T(\mathbf{z}, \boldsymbol{\theta})$ em L conjuntos de dados simulados, isto é, $p_B = \mathbb{P}\left(T(\mathbf{z}_{pr}, \boldsymbol{\theta}\right) \geq T(\mathbf{z}, \boldsymbol{\theta}) \mid \boldsymbol{Z} = \mathbf{z})$, onde $\mathbf{z}_{pr}$ é a amostra simulada da distribuição preditiva à posteriori. Se o modelo for adequado, p_B deve ser próximo de 0.5 sendo que valores muito alto ou muito baixos indicam uma má especificação do modelo.

1.7 Detecção de observações influentes em estudos Bayesianos

Nesta Seção iremos abordar algumas medidas Bayesianas de diagnóstico para detectar observações extremas, que exercem uma influência desproporcionalmente grande nos resultados do ajuste. A abordagem que utilizaremos é uma das mais conhecidas, o método de deleção de casos.

Um modo bastante comum em estudos Bayesianos para quantificar a influência de uma determinada observação é calcular alguma medida de divergência entre a distribuição à posteriori obtida com o conjunto de dados completo e excluindo-se dele tal observação. Se obtivermos uma medida significativamente grande, então esta observação é considerada influente. Desta forma, seja $\mathbf{z} = \{z_1, z_2, \ldots, z_n\}$ a amostra observada e I um subconjunto de $\{1, 2, \ldots, n\}$. Considere o conjunto $\mathbf{z}_I = \{z_i;\ i \in I\}$, denotando por $\mathbf{z}_{-I}$ seu complementar em $\mathbf{z}$. A fim de estabelecer uma medida de divergência apropriada, denote por $f(\boldsymbol{\theta}|\mathbf{z}; M_0)$ e $f(\boldsymbol{\theta}|\mathbf{z}; M_1)$ as distribuições a posteriori de $\boldsymbol{\theta}$ sob os modelos M_0 e M_1 respectivamente. Para comparar o quão parecidos são estes dois modelos em termos da inferência final sobre $\boldsymbol{\theta}$, definimos a função de perturbação entre M_0 e M_1 como:

$$m(\boldsymbol{\theta}; M_0, M_1) = \frac{f(\boldsymbol{\theta}|\mathbf{z}; M_1)}{f(\boldsymbol{\theta}|\mathbf{z}; M_0)}.$$

A partir desta função de perturbação Csiszár (1967) define a medida "*q-divergente*" entre os modelos M_0 e M_1 da seguinte forma:

$$d_q\left(m(\boldsymbol{\theta}; M_0, M_1)\right) = \mathbb{E}_{M_0}\left[q(m(\boldsymbol{\theta}; M_0, M_1))\right] \tag{1.7.1}$$

onde $q(\cdot)$ é uma função convexa tal que $q(1) = 0$.

Se tomarmos M_0 como o modelo que considera toda a amostra $\mathbf{z}$ para o cálculo da posteriori e M_1 como o modelo que considera somente $\mathbf{z}_{-I}$ para este fim, então podemos definir a função de perturbação do conjunto I na distribuição a posteriori como:

$$m_I(\boldsymbol{\theta}) = m(\boldsymbol{\theta}; M_0, M_1) = \frac{f(\boldsymbol{\theta}|\mathbf{z}_{-I})}{f(\boldsymbol{\theta}|\mathbf{z})}. \tag{1.7.2}$$

A medida q-divergente pode também ser definida para a função de perturbação do conjunto I da seguinte forma:

$$d_q(\mathbf{z}, \mathbf{z}_{-I}) = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Z}}[q(m_I(\boldsymbol{\theta}))]. \tag{1.7.3}$$

Algumas medidas de divergência bastante conhecidas são obtidas considerando-se diferentes funções $q(\cdot)$, por exemplo, se $q(x) = -\log(x)$ obtemos a divergência de Kullback-Leibler (K-L), ao considerar $q(x) = (x-1)\log(x)$ estaremos trabalhando com a distância J e a distância L_1 é obtida fazendo-se $q(x) = |x-1|$.

Para a maioria dos modelos é bastante complicado calcular a esperança da Equação (1.7.3), logo, é conveniente considerar uma aproximação MCMC das medidas de influência citadas. Neste sentido, enunciamos a seguinte proposição:

Proposição 1. *Seja $m_I(\boldsymbol{\theta})$ a função de perturbação do conjunto I conforme definida na Equação (1.7.2) e denote por CPO_I a estatística CPO para o conjunto de observações $\mathbf{z}_I$, isto é, $CPO_I = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Z}}\left[f(z_I|\boldsymbol{\theta})^{-1}\Big|\ \mathbf{z}\right]^{-1} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{Z}}\left[\prod_{j\in I} f(z_j|\boldsymbol{\theta})^{-1}\Big|\ \mathbf{z}\right]^{-1}$. Pode-se escrever:*

$$m_I(\boldsymbol{\theta}) = \frac{CPO_I}{f(\mathbf{z}_I|\boldsymbol{\theta})}.$$

Prova:

Segundo a definição dada na Equação (1.7.2),

$$\begin{aligned}
m_I(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}|\mathbf{z}_{-I}) f(\boldsymbol{\theta}|\mathbf{z})^{-1} \\
&= \frac{f(\boldsymbol{\theta}) f(\mathbf{z}_{-I}|\boldsymbol{\theta})}{f(\mathbf{z}_{-I})} \frac{f(\mathbf{z})}{f(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta})} \\
&= \frac{f(\mathbf{z})}{f(\mathbf{z}_I|\boldsymbol{\theta})} \left(\int f(\mathbf{z}_{-I}|\boldsymbol{\theta}) f(\boldsymbol{\theta})\, d\boldsymbol{\theta} \right)^{-1} \qquad (1.7.4) \\
&= \frac{f(\mathbf{z})}{f(\mathbf{z}_I|\boldsymbol{\theta})} \left(\int f(\mathbf{z}|\boldsymbol{\theta}) f(\mathbf{z}_I|\boldsymbol{\theta})^{-1} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} \right)^{-1} \qquad (1.7.5) \\
&= f(\mathbf{z}_I|\boldsymbol{\theta})^{-1} \left(\int \frac{f(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta})}{f(\mathbf{z})} f(\mathbf{z}_I|\boldsymbol{\theta})^{-1}\, d\boldsymbol{\theta} \right)^{-1} \\
&= f(\mathbf{z}_I|\boldsymbol{\theta})^{-1} \left(\int f(\mathbf{z}_I|\boldsymbol{\theta})^{-1} f(\boldsymbol{\theta}|\mathbf{z})\, d\boldsymbol{\theta} \right)^{-1} \\
&= \frac{CPO_I}{f(\mathbf{z}_I|\boldsymbol{\theta})},
\end{aligned}$$

onde a Equação (2.2.5) foi obtida através das igualdades $f(\mathbf{z}_{-I}|\boldsymbol{\theta})/f(\mathbf{z}|\boldsymbol{\theta}) = f(\mathbf{z}_I|\boldsymbol{\theta})^{-1}$ e $f(\mathbf{z}_{-I}) = \int f(\mathbf{z}_{-I}|\boldsymbol{\theta})f(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ e, multiplicando-se o integrando por $f(\mathbf{z}_I|\boldsymbol{\theta})/f(\mathbf{z}_I|\boldsymbol{\theta})$, obtém-se a Equação (2.2.6).

■

Com a Proposição 1 encontramos uma maneira de aproximar $m_i(\boldsymbol{\theta})$ para uma dada observação i, $i = 1, 2, \ldots, n$, e com isso obter também uma aproximação das medidas de influência citadas para tal observação. Basta-nos aproximar a função $m_i(\boldsymbol{\theta})$ através de uma amostra MCMC da distribuição a posteriori de $\boldsymbol{\theta}$, $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_Q\}$ da seguinte forma:

$$\widehat{m}_i(\boldsymbol{\theta}) = \widehat{CPO}_i \left(\frac{1}{Q} \sum_{j=1}^{Q} f(z_i|\boldsymbol{\theta}_j) \right)^{-1},$$

onde $\widehat{CPO}_i$ é exibido na Equação (1.6.2). Em seguida, a aproximação da medida de influência é dada por:

$$\widehat{d}_q(\mathbf{z}, \mathbf{z}_{-i}) = \frac{1}{Q} \sum_{j=1}^{Q} q(\widehat{m}_i(\boldsymbol{\theta}_j)).$$

Note que é preciso estabelecer um ponto limiar a partir do qual estas medidas classificam uma dada observação como influente. Neste sentido, Peng & Dey (1995) and Vidal & Castro (2010) fazem analogia a um experimento clássico para determinar este ponto limiar, o lançamento de uma moeda: suponha que uma moeda com probabilidade $p \in [0, 1]$ de cara é arremessada. A variável aleatória X representa o resultado do lançamento, sendo que $X = 1$ significa que o experimento resultou em cara e $X = 0$, em coroa. Sob o ponto da inferência frequentista, seja M_0 o modelo no qual não é feita nenhuma suposição a respeito de p e M_1 o modelo no qual a moeda é considerada justa, ou seja, no qual consideramos $p = 0.5$. A distribuição associada a X sob o modelo M_0 é $f_0(x; p) = p^x(1-p)^{1-x}$ e, sob o modelo M_1, é $f_1(x; p) = 0.5$, com $x \in \{0, 1\}$ em ambos os casos. Da Equação (1.7.1), a medida $q-$divergente entre estes dois modelos é dada por:

$$d_q^*(M_0, M_1) = \frac{q(2p) + q(2(1-p))}{2}. \tag{1.7.6}$$

Note que $d_q^*(M_0, M_1)$ aumenta conforme p se afasta de 0.5, é simétrica em torno de 0.5 e atinge seu ponto de mínimo também em $p = 0.5$ (onde os modelos são iguais). Se estabelecermos o

critério de que uma estimativa para p maior do que 0.8 (ou menor do que 0.2) sugere fortes evidências de que a moeda não é justa, então $d_q^*(0.8)$ seria um ponto limiar acima do qual uma observação seria considerada influente pela medida $q-$divergente. Calculando $d_q^*(0.8)$ para as medidas particulares apontadas, temos $d_{KL}^*(0.8) \approx .0.2231$ para K-L, $d_J^*(0.80) \approx 0.4159$ para a distância J e $d_{L_1}^*(0.8) = 0.6$ para a distância L_1.

1.8 Apresentação dos próximos capítulos

Inciaremos o Capítulo 2 apresentando a classe de distribuições de mistura de escala, dando atenção especial à distribuição t de Student. Definiremos então os modelos de regressão para respostas censuradas, começando pelo clássico modelo Tobit e estendendo seus resultados para erros com distribuição t de Student. Em seguida, faremos um estudo de inferência Bayesiano através da construção do amostrador de Gibbs e também um estudo frequentista, desenvolvendo o algoritmo EM para a estimação paramétrica e técnicas de diagnósticos baseadas em influência global e local. Estudos de simulação serão desenvolvidos para avaliar a qualidade das estimativas EM e a robustez dos modelos. Finalmente, as técnicas desenvolvidas no capítulo serão aplicadas a dois conjuntos de dados reais.

No Capítulo 3 apresentaremos a classe de distribuições de misturas da skew-escala normal (SMSN), originalmente introduzida por Branco & Dey (2001) e posteriormente estudada por Kim (2008b), Basso (2009), entre outros. Em seguida, desenvolveremos um estudo de inferência Bayesiano para os modelos de regressão com respostas censuradas sob esta classe de distribuições através da construção do amostrador de Gibbs para a estimação paramétrica. Além disso, apresentaremos dois estudos de simulação para avaliar a qualidade do método de estimação e a robustez dos modelos propostos. Finalmente, as técnicas desenvolvidas serão aplicadas em um conjunto de dados reais.

No Capítulo 4, finalizaremos esta trabalho com as conclusões finais, a apresentação da produção técnica derivada desta dissertação e algumas perspectivas para trabalhos futuros.

É importante destacar que todas as aplicações e estudos de simulação apresentados neste trabalho foram desenvolvidos no *software* `R`.

Capítulo 2

Modelos Normal e t de Student para dados censurados

2.1 Introdução

Problemas ou experimentos cuja variável de interesse está sujeita à censura surgem em diversos campos de estudo. Quando o interesse principal é medir o efeito de determinados fatores nesta variável, o modelo de regressão linear Tobit (Tobin (1958)) é um dos métodos mais conhecidos e utilizados. Por exemplo, Tan *et al.* (2009) aplica este modelo aos dados de Schmee & Hahn (1979), "*Insulation life data with censoring times*", na tentativa de medir a influência da temperatura no tempo de vida útil de motores elétricos.

No entanto, mesmo em conjuntos de dados relativamente simétricos, a suposição de normalidade pode não ser válida devido à presença de valores extremos. Nestes casos, o ajuste do modelo Tobit não é adequado e uma distribuição com "caudas mais pesadas", capaz de comportar a presença destes valores, deveria ser levada em consideração.

Neste segundo capítulo definiremos o modelo Tobit, denotado neste trabalho por N-CR ("*normal censored regression*") e apresentaremos a família de distribuições de mistura de escala normal, introduzindo a distribuição t de Student como alternativa à normal para os erros aleatórios do

N-CR, dando origem ao modelo t-CR ("*t de Student censored regression*"). Faremos um estudo de inferência clássica e Bayesiana para estes modelos, apresentando algoritmos iterativos para estimar seus parâmetros e desenvolvendo medidas de diagnóstico. Por fim, aplicaremos ambos os modelos a um conjunto de dados reais e apresentaremos alguns estudos de simulação que visam comparar a performance dos modelos normal e t de Student para dados censurados.

2.2 A família de mistura de escala normal (SMN)

Andrews & Mallows (1974) introduziram uma família de distribuições simétricas chamada "distribuições de mistura de escala normal". O ganho desta classe é incorporar um parâmetro de forma (relacionado à curtose e não à assimetria) à densidade normal e, com isso, obter distribuições mais maleáveis e com caudas mais pesadas do que esta, bastante úteis em inferência robusta para dados simétricos. A seguir, definiremos esta família de distribuições e desenvolveremos algumas de suas propriedades.

Definição 1. *Dizemos que a variável aleatória X tem distribuição pertencente à família de mistura de escala normal, com parâmetro de locação $\mu \in \mathbb{R}$, de escala $\sigma^2 > 0$ e de forma $\boldsymbol{\nu}$ se ela pode ser escrita da seguinte forma:*

$$X = \mu + \kappa(U)^{1/2} Z, \quad Z \sim N(0, \sigma^2), \quad U \perp Z, \tag{2.2.1}$$

onde $\kappa(\cdot)$ é uma fução real positiva e U uma variável aleatória positiva, cuja função de distribuição acumulada, $H(\,\cdot\,; \boldsymbol{\nu})$, é indexada pelo vetor de parâmetros $\boldsymbol{\nu}$. Denotamos $X \sim SMN(\mu, \sigma^2\,;\, H)$.

É fácil ver na Equação (2.2.1) que $X|\kappa(U) = \kappa(u) \sim N(\mu, \kappa(u)\sigma^2)$, portanto, a densidade de X é dada por:

$$f_{SMN}(x) = \int_0^\infty \phi(x\,;\, \mu, \kappa(u)\sigma^2)\, dH(u). \tag{2.2.2}$$

Alguns dos casos especiais mais conhecidos desta classe de distribuição são alcançados quando $\kappa(U) = 1/U$. Fixando a função $\kappa(\cdot)$ desta forma, se tomarmos na Equação (2.2.1):

i) U degenerada em 1, isto é, $\mathbb{P}(U = 1) = 1$, então $X \sim N(\mu, \sigma^2)$

ii) $U \sim Beta(\nu, 1)$, temos X distribuída de acordo com uma Slash, com parâmetros de locação μ, escala σ^2 e forma ν.

iii) U discreta, assumindo o valor ν_1 com probabilidade ν_2 ou o valor 1, com probabilidade $1-\nu_2$. Neste caso, X segue uma normal contaminada com parâmetro de locação μ, de escala σ^2 e de forma $\boldsymbol{\nu} = (\nu_1, \nu_2)$.

iv) $U \sim G(\nu_1/2, \nu_2/2)$, então X segue a distribuição Pearson VII com locação μ, escala σ^2 e forma $\boldsymbol{\nu} = (\nu_1, \nu_2)$. Note que, se $\nu_1 = \nu_2 := \nu$, então a distribuição de X se reduz à t de Student com ν graus de liberdade e mesmos parâmetros de locação e escala. Se $\nu = 1$, temos a Cauchy.

A seguir apresentaremos algumas a distribuição t de Student, no intuito de definir o modelo t-CR.

2.2.1 A distribuição t de Student

Como discutido anteriormente, a suposição do modelo N-CR de que os erros ϵ_i em (2.3.2), para $i = 1, 2, \ldots, n$, são normalmente distribuídos pode prejudicar o processo de inferência se houver valores extremos no conjunto de dados. Diante deste problema, introduziremos aqui a distribuição t de Student visando a substituição da distribuição Normal no modelo de regressão para dados censurados, na tentativa de, ao utilizar uma distribuição mais robusta, conseguir comportar a existência de valores considerados extremos sob a distribuição Normal. Little (1999) e Lange *et al.* (1989), por exemplo, utilizaram a distribuição t de Student para modelagem robusta. A seguir, apresentamos a densidade da distribuição t de Student e desenvolvemos algumas de suas propriedades.

Definição 2. *Dizemos que a variável aleatória X tem distribuição t de Student, com parâmetro de locação μ, de escala σ^2 e ν graus de liberdade, denotada por $t(\mu, \sigma^2, \nu)$, se sua densidade é dada por:*

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \left(1 + \frac{d(x)^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}, \quad x \in \mathbb{R}, \tag{2.2.3}$$

onde $d(x) = \frac{(x-\mu)}{\sigma}$.

Na Figura 2.1 mostramos a densidade da t de Student para diferentes valores dos parâmetros σ^2 e ν. É importante destacar no primeiro gráfico como o valor de ν influencia na curtose da distribuição, sendo que valores menores de ν geram caudas mais pesadas. A distribuição Normal é o caso limite da t de Student quando $\nu \to \infty$.

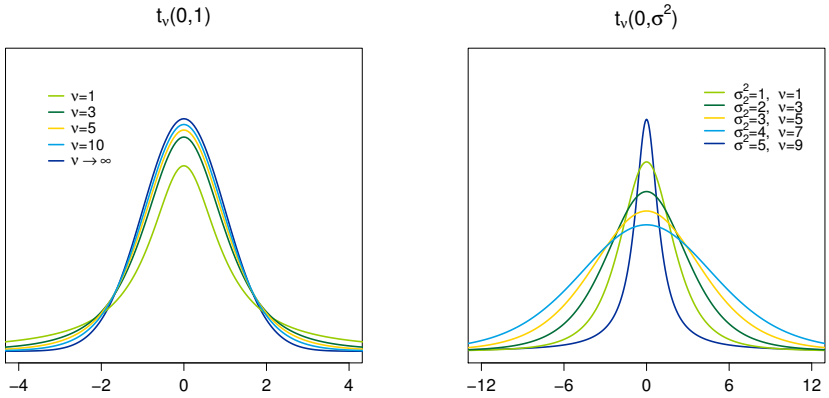


Figura 2.1: Densidade de t de Student para $\mu = 0$ e valores variados para σ^2 e ν.

Proposição 2. *Seja $X \sim t(\mu, \sigma^2, \nu)$. Então, a distribuição de X pertence à família de mistura de escala normal e esta variável aleatória pode ser escrita como:*

$$X = \mu + U^{-1/2} Z, \tag{2.2.4}$$

onde $U \sim G(\nu/2, \nu/2)$ é independente de $Z \sim N(\mu, \sigma^2)$. Além disso, $\mathbb{E}[X] = \mu$ e, se $\nu > 2$, $Var[X] = \sigma^2 \frac{\nu}{\nu-2}$.

Prova:

Seja X uma variável aleatória como na Equação (2.2.4). Então, dado que a densidade condicional de $X|U = u$ é $N(\mu, \sigma^2/u)$, podemos utilizar a relação $f(x, u) = f(x|u)f(u)$ para calcular a

densidade marginal de X:

$$\begin{aligned} f(x) &= \int_0^\infty \phi(x\,;\,\mu,\sigma^2)\, Gama(u\,;\,\nu/2,\nu/2)\, du \\ &= \frac{(\nu/2)^{\nu/2}}{\sqrt{2\pi}\sigma\,\Gamma(\nu/2)} \int_0^\infty u^{\frac{\nu+1}{2}-1} \exp\left\{-\frac{u}{2}(d(x)^2+\nu)\right\} \\ &= \frac{(\nu/2)^{\nu/2}}{\sqrt{2\pi}\sigma\,\Gamma(\nu/2)}\,\Gamma\left(\frac{\nu+1}{2}\right)\left(\frac{d(x)^2+\nu}{2}\right)^{-\frac{\nu+1}{2}} \qquad (2.2.5) \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\sigma\,\Gamma(\nu/2)}\left(1+\frac{d(x)^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad (2.2.6) \end{aligned}$$

onde $Gama(\,\cdot\,;\,\alpha,\beta)$ denota a função de densidade de uma $G(\alpha,\beta)$, $d(x)$ é como na Definição 2 e a Equação (2.2.5) foi obtida "completando" a densidade de uma $G(\frac{\nu+1}{2}, \frac{d(x)^2+\nu}{2})$ em u na integral, de forma que o resultado desta integral fosse 1. Em (2.2.6) vemos que a densidade de X corresponde a uma $t(\mu,\sigma^2,\nu)$, conforme a Definição 2.2.3. Agora,

$$\begin{aligned} \mathbb{E}_X[X] &= \mathbb{E}_U[\,\mathbb{E}_{X|U}[X|U]\,] \\ &= \mathbb{E}_U[\mu] = \mu. \end{aligned}$$

Além disso,

$$\begin{aligned} \mathrm{Var}_X[X] &= \mathbb{E}_U[\,\mathrm{Var}_{X|U}[X|U]\,] + \mathrm{Var}_U[\,\mathbb{E}_{X|U}[X|U]\,] \\ &= \mathbb{E}_U[\sigma^2 U^{-1}] + \mathrm{Var}_U[\mu] \\ &= \sigma^2\frac{\nu}{\nu-2}, \quad \nu>2, \qquad (2.2.7) \end{aligned}$$

aqui, a Equação (2.2.7) foi obtida usando o seguinte resultado (cuja demonstração será omitida): se $U \sim G(\alpha,\beta)$, então a distribuição de U^{-1} é uma Inversa Gama com parâmetros de forma α e de escala β, cuja esperança é $\frac{\beta}{\alpha-1}$.

■

2.3 Definição dos modelos N-CR e t-CR

Supunha que queiramos ajustar um modelo de regressão à variável de interesse Y, de acordo com:

$$Y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \epsilon_i, \qquad i=1,2,\ldots,n, \qquad (2.3.1)$$

em que, para $i = 1, 2, \ldots, n$, Y_i é a resposta para o indivíduo i, ϵ_i a variável aleatória representando o erro da regressão, $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^\top$ um vetor $p \times 1$ de variáveis explicativas para este indivíduo e $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_p]^\top$ o vetor de parâmetros da regressão.

O modelo acima supõe que a variável Y pode ser observada para todos os indivíduos no estudo. Porém neste trabalho estamos interessados no caso em que Y é sujeita à censura, isto é, no caso em que existem indivíduos na amostra para o qual o valor assumido por Y não é conhecido, somente observa-se o intervalo $\mathbb{A}$ em que ele está contido. Se a variável resposta é censurada à direita, $\mathbb{A} = [a, \infty)$ e, se é censurada à esqueda, $\mathbb{A} = (-\infty, a]$, para a uma constante representando o valor limítrofe observado para Y. Neste contexto, a variável aleatória passível de observação e que será modelada é V, uma função de Y dada por:

$$V_i = \begin{cases} a_i & \text{se} \quad Y_i \geq a_i \\ Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i & \text{se} \quad Y_i < a_i, \end{cases} \tag{2.3.2}$$

para $i = 1, 2, \ldots, n$ e $\{a_i\}_{i=1}^n$ o conjunto de valores limiares fixos e conhecidos. Note que, a partir do conhecimento deste conjunto de valores limiares, ao observar V_i observa-se também a variável indicadora C_i, que assume o valor 1 quando a resposta do indivíduo i é censurada ($V_i = a_i$) e 0, quando não é ($V_i \neq a_i$), com $i = 1, 2, \ldots, n$.

As Equações (2.3.1) e (2.3.2) definem o modelo de regressão para dados censurados. Se assumirmos na Equação (2.3.1) que $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ obtemos o modelo N-CR, e, se consideramos $\epsilon_i \overset{\text{iid}}{\sim} t_v(0, \sigma^2)$, o modelo t-CR. Para mais detalhes sobre o modelo N-CR veja Barros *et al.* (2010), onde é desenvolvido um estudo de inferência e diagnóstico.

Dado uma amostra observada $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$ de $\boldsymbol{V} = (V_1, V_2, \ldots, V_n)^\top$, a função de log-verossimilhança para estes modelos assumindo censura à esquerda é dada por:

$$\ell(\boldsymbol{\theta}; \boldsymbol{v}) = \sum_{i=1}^{n} \log \left[F_{SMN}\left(v_i; \boldsymbol{\theta}\right) \right] \mathbb{1}_{(-\infty, a_i]}(v_i) + \sum_{i=1}^{n} \log \left[f_{SMN}(v_i; \boldsymbol{\theta}) \right] \mathbb{1}_{(a_i, \infty)}(v_i), \tag{2.3.3}$$

onde $f_{SMN}(\cdot; \boldsymbol{\theta})$ e $F_{SMN}(\cdot; \boldsymbol{\theta})$ representam a função de densidade e de distribuição acumulada de Y. No caso do N-CR, estas funções referem-se a uma variável aleatória normal com média $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ e

variância σ^2, no caso do t-CR, a uma variável aleatória que segue uma distribuição $t(\boldsymbol{x}_i^\top\boldsymbol{\beta}, \sigma^2, \nu)$. Se considerássemos censuras à direita, a função de log-verossimilhança seria bastante análoga:

$$\ell(\boldsymbol{\theta}; \boldsymbol{v}) = \sum_{i=1}^{n} \log\left[1 - F_{SMN}\left(v_i; \boldsymbol{\theta}\right)\right] \mathbb{1}_{[a_i,\infty)}(v_i) + \sum_{i=1}^{n} \log\left[f_{SMN}(v_i; \boldsymbol{\theta})\right] \mathbb{1}_{(-\infty,a_i)}(v_i), \tag{2.3.4}$$

Nas próximas duas Seções será desenvolvido um estudo de inferência e diagnóstico para estes modelos sob as óticas Bayesianas e frequentistas, respectivamente.

2.4 Inferência Bayesiana para os modelos N-CR e t-CR

2.4.1 Construção do amostrador de Gibbs

Nesta Seção desenvolveremos o amostrador de Gibbs para fazer inferência para os modelos N-CR e t-CR sob um ponto de vista Baysiano. Os modelos serão então aplicados em um conjunto de dados reais e estudos de simulação serão realizados para avaliar a qualidade do processo de estimação e a robustez dos dois modelos.

O primeiro passo para fazer inferência Bayesiana para os modelos N-CR e t-CR, definidos na Seção 2.3, é definir uma densidade à priori para o vetor de parâmetros $\boldsymbol{\theta}$. Para o N-CR, $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$ e para o t-CR $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \nu)^\top$. Aqui, usaremos a suposição à priori de independência entre os parâmetros do modelos, portanto a densidade conjunta à priori do vetor $\boldsymbol{\theta}$ será o produto das densidades à priori de cada um de seus elementos. Para ambos os modelos assumiremos à priori que $\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma_0})$, com hiperparâmetros fixos e conhecidos $\boldsymbol{\mu}_0 \in \mathbb{R}^p$ e $\boldsymbol{\Sigma_0}$ uma matriz $p \times p$ diagonal e definida positiva, e que σ^2 tem como densidade à priori uma $GI(a_{\sigma^2}, b_{\sigma^2})$, onde $a_{\sigma^2} > 0$ e $b_{\sigma^2} > 0$ são hiperparâmetros fixos e conhecidos.

No caso do modelo t-CR é preciso ainda definir uma priori para ν. Neste sentido, existe um grande número de sugestões como a clássica exponencial, a exponencial truncada, Geweke (1993), a priori de Jeffreys, Fonseca *et al.* (2008) e a exponencial hierárquica, Cabral & Madruga (2012). Esta discussão é sumarizada em um estudo de simulação feito em Garay *et al.* (2013) que guia nossa escolha pela exponencial truncada hierárquica, ou seja, neste trabalho assumimos

$\nu \sim Texp(\gamma\ ;\ (2,\infty)\ )$ e $\gamma \sim Unif(c,d)$, onde $Texp(\gamma\ ;\ \mathbb{A})$ denota uma distribuição exponencial de parâmetro γ truncada no intervalo $\mathbb{A}$ e, $Unif[c,d]$, uma distribuição uniforme no intervalo $[c,d]$. Aqui, o truncamento da exponencial serve principalmente para assegurar que todos os valores gerados da amostra à posteriori de ν caiam em um intervalo para o qual o segundo momento da t de Student exista.

No contexto Bayesiano, estimativas pontuais são obtidas como características associadas à distribuição à posteriori deste parâmetro, como a esperança ou a moda. Devido à forma complexa destas distribuições a posteriori muitas vezes é bastante custoso (algébrica ou computacionalmente) obter estes estimadores de forma direta ou por integração numérica. Atualmente, algoritmos do tipo MCMC são bastante utilizados a fim de amostrar da distribuição a posteriori e fazer inferência para o problema. Em nosso caso, o algoritmo do tipo MCMC a ser utilizado é o amostrador de Gibbs (veja a Seção 1.5 para detalhes sobre o algoritmo) e, para o modelo t-CR, ele será construído com base na representação estocástica dada na Equação (2.2.4). O artifício utilizado na elaboração do amostrado de Gibbs é o "aumento dos dados", isto é, vamos supor que o vetor de variáveis respostas sujeitas à censura $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ e, no caso do t-CR, também o vetor de variáveis latentes $\boldsymbol{U} = (U_1, U_2, \ldots, U_n)$, podem ser completamente observados - isto nos permitirá calcular as distribuições condicionais completas para $\boldsymbol{\theta}$ e $\boldsymbol{Y}$ e, sob o t-CR, também de γ e $\boldsymbol{U}$ e amostrar destas distribuições.

Considere $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$ e $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^\top$ os vetores de observações de V_i e C_i, $i = 1, 2, \ldots, n$, respectivamente. Seja $\vartheta^{(0)}$ um valor inicial para ϑ e $\vartheta^{(k)}$ o valor de ϑ na iteração k do algoritmo. Calculando-se as condicionais completas, a k-ésima iteração do amostrador de Gibbs sob o modelo t-CR é da seguinte forma:

Passo 1: Para $i = 1, 2, \ldots, n$, se $c_i = 0$ tome $y_i^{(k)} = v_i$. Caso contrário, se $c_i = 1$, gere $y_i^{(k)}$ (independentemente) de uma normal truncada:

$$NT(\mathbf{x_i}^\top \boldsymbol{\beta}^{(k-1)}, \sigma^{2(k-1)}/u_i^{(k-1)}\ ;\ \mathbb{A}),$$

onde $\mathbb{A} = [v_i, \infty)$ se a variável resposta for censurada à direita ou $\mathbb{A} = (\infty, v_i]$, se for censurada

à esquerda.

Passo 2: Gere $\sigma^{2(k)}$ de uma gama inversa:

$$GI\left(a_{\sigma^2}+\frac{n}{2}\ ,\ b_{\sigma^2}+\frac{1}{2}\sum_{i=1}^{n}u_i^{(k-1)}(y_i^{(k)}-\boldsymbol{x}_i^{\top}\boldsymbol{\beta}^{(k-1)})^2\right).$$

Passo 3: Para $i=1,2,\ldots,n$, gere $u_i^{(k)}$ independentemente de uma gama:

$$G\left(\frac{\nu^{(k-1)}+1}{2}\ \ ,\ \ \frac{1}{2}\left(\nu^{(k-1)}+\frac{(y_i^{(k)}-\boldsymbol{x}_i^{\top}\boldsymbol{\beta}^{(k-1)})^2}{\sigma^{2(k)}}\right)\right).$$

Passo 4: Gere $\boldsymbol{\beta}^{(k)}$ de uma normal $p-$variada $N_p(\boldsymbol{\mu}^{*(k)},\boldsymbol{\Sigma}^{*(k)})$, onde

$$\boldsymbol{\mu}^{*(k)}=\boldsymbol{\Sigma}^{*(k)}\left(\boldsymbol{\Sigma_0^{-1}\mu_0}+\frac{\boldsymbol{x}^{*(k)\top}\boldsymbol{y}^{*(k)}}{\sigma^{2(k)}}\right),\quad \boldsymbol{\Sigma}^{*(k)}=\left(\frac{\boldsymbol{x}^{*(k)\top}\boldsymbol{x}^{*(k)}}{\sigma^{2(k)}}+\boldsymbol{\Sigma_0^{-1}}\right)^{-1},$$

$\boldsymbol{x}^{*(k)}$ é uma matriz $n\times p$ cuja i-ésima linha é formada pelo vetor $\sqrt{u_i^{(k)}}\boldsymbol{x}_i^{\top}$ e $\boldsymbol{y}^{*(k)}$ é o vetor $n\times 1$ cuja i-ésima componente é dada por $\sqrt{u_i^{(k)}}y_i^{(k)}$, $i=1,2,\ldots,n$.

Passo 5: Gere $\gamma^{(k)}$ de uma gama truncada: $TG(2,\nu^{(k-1)}\ ;\ [a_\gamma,b_\gamma])$.

Passo 6: $\nu^{(k)}$ deve ser gerado de $f(\nu|\boldsymbol{y}^{(k)},\boldsymbol{u}^{(k)},\boldsymbol{\beta}^{(k)},\sigma^{2(k)},\gamma^{(k)})$, que é proporcional a:

$$\left(\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\right)^n\left(\prod_{i=1}^{n}u_i^{(k)}\right)^{\frac{\nu}{2}-1}\exp\left\{-\nu\left(\gamma^{(k)}+\frac{1}{2}\sum_{i=1}^{n}u_i^{(k)}\right)\right\}\mathbb{1}_{(2,\infty)}(\nu),$$

em que utilizamos um passo de Metropolis-Hasting para amostrar desta distribuição (veja detalhes deste algoritmo na Seção 1.5). Dada a observação $\nu^{(k-1)}$ obtida na iteração $k-1$ do amostrado de Gibbs, gere um candidato ν^* da distribuição candidata $g(\nu)$, uma distribuição normal truncada:

$$g(\nu)\equiv \mathrm{TN}\left(\omega_{\nu^{(k-1)}},\varsigma_{\nu^{(k-1)}};(2;\infty)\right),$$

onde o parâmetro de locação desta distribuição candidata é dado por $\omega_{\nu^{(k-1)}}=\nu^{(k-1)}-\frac{q_1(\nu^{(k-1)})}{q_2(\nu^{(k-1)})}$ e o de escala por $\varsigma_{\nu^{(k-1)}}=-\frac{1}{q_2(\nu^{(k-1)})}$, onde:

$$\begin{aligned}q_1(\nu) &= \frac{d}{d\nu}\log f(\nu\mid\boldsymbol{y}^{(k)},\boldsymbol{u}^{(k)},\boldsymbol{\beta}^{(k)},\sigma^{2(k)}.\gamma^{(k)}),\\ q_2(\nu) &= \frac{d^2}{d\nu^2}\log f(\nu\mid\boldsymbol{y}^{(k)},\boldsymbol{u}^{(k)},\boldsymbol{\beta}^{(k)},\sigma^{2(k)}.\gamma^{(k)}),\end{aligned}$$

respectivamente a primeira e segunda derivadas da condicional completa de ν. As escolhas dos parâmetros de locação e de escala da distribuição candidata foram baseadas no trabalho de Abanto-Valle *et al.* (2013) e o truncamento foi feito no intervalo $(2, \infty)$ para garantir a existência dos primeiros dois momentos da t de Student utilizando o candidato gerado como graus de liberdade.

O amostrado de Gibbs construído sob o modelo N-CR é como o anterior, porém considerando $u_i = 1$ em todos os passos e iterações e descartando-se os passos 3, 5 e 6.

2.4.2 Aplicação I

Nesta Subseção aplicaremos os métodos de inferência desenvolvidos para os modelos N-CR e t-CR utilizando os dados disponibilizados em Tan *et al.* (2009) "*Insulation life data with censoring times*", onde é feito um teste acelerado sobre o tempo de vida útil do isolamento térmico de motores elétricos (classe B, originalmente projetada para suportar temperaturas até 130^oC). Um total de 40 motores foram envolvidos no experimento, sendo que grupos de 10 foram testados sob as quatro diferentes temperaturas: 150^oC, 170^oC, 190^oC e 220^oC. Os testes foram interrompidos em diferentes tempos para cada nível de temperatura, de forma que os motores que não sofreram danos até o término do estudo representam observações censuradas à direita, sendo estes 23 de um total de 40.

Para este conjunto de dados vamos ajustar os modelos N-CR e t-CR e comparar a performance de ambos, utilizando para isto o pacote `BayesCR` (veja sua descrição na Subseção 4.1.3). Seguindo a sugestão dada em Tan *et al.* (2009), a variável resposta (censurada à direita) será o logaritmo na base 10 do tempo de vida útil (em horas) e a variável explicativa é uma função decrescente da temperatura, $(100\,(\text{temperatura} + 273,2)^{-1})$. Conforme os modelos N-CR e t-CR descritos nas subseções anteriores, denotaremos por β_1 o intercepto do modelo e β_2 o coeficiente de regressão correspondente à variável explicativa.

Usando o amostrador de Gibbs geramos cadeias de tamanho 200000, aplicamos um *burn-in* de 50000 e um *thinning* de tamanho 30, a fim de eliminar possíveis correlações. Como resultado, obtivemos 5000 observações das distribuições a posteriori. A estatística de Gelman-Rubin (veja

Seção 1.5), $\hat{R}$, foi calculada para avaliar a convergência da cadeia. Os resultados estão mostrados na Tabela 2.1

	Modelos							
	N-CR				**t-CR**			
Parâmetros	Média	SD	HPD (95%)	$\hat{R}$	Média	SD	HPD (95%)	$\hat{R}$
β_1	-6.498	1.674	(-9.677;-3.125)	1.00067	-6.372	1.515	(-9.428;-3.486)	1.00039
β_2	4.602	0.774	(3.122; 6.143)	1.00072	4.528	0.703	(3.189; 5.937)	1.00035
σ^2	0.197	0.090	(0.075; 0.378)	1.00048	0.148	0.082	(0.030; 0.296)	1.00234
ν	—	—	—	—	14.457	18.465	(2.101; 59.157)	1.00087

Tabela 2.1: *Insulation life data with censoring times.* Média a posteriori, desvio padrão à posteriori (SD), intervalo HPD (95%) e estatística de Gelman-Rubin ($\hat{R}$) sob os modelos N-CR e t-CR.

Como o valor estimado para β_1 é positivo nos dois modelos, podemos concluir que o logaritmo na base 10 do tempo de vida útil do isolamento (e, consequentemente, o próprio tempo de vida) diminui conforme aumenta-se a temperatura do teste (note que a variável explicativa é função decrescente da temperatura). Desta forma, para uma dada temperatura t, o tempo médio de vida é estimado em $10^{-6.498+1674/(t+273.2)}$ pelo modelo N-CR e em $10^{-6.372+1515/(t+273.2)}$ pelo modelo t-CR. Note que o valor alto estimado para ν sob o t-CR indica que o modelo N-CR (seu caso limite quando $\nu \to \infty$) pode ser adequado para este conjunto de dados.

Na Tabela 2.2 mostramos a comparação entre a adequação dos dois modelos ajustados através dos critérios descritos na Subseção 1.6.2. Todos os critérios utilizados dão preferência para o ajuste promovido pelo N-CR, embora seus valores não apresentem uma diferença tão grande comparados com os calculados sob o t-CR. Os p-valores Bayesianos não indicam má especificação de nenhum dos dois modelos.

No intuito de identificar observações que exercem alguma influência "desproporcional" na inferência final, calculamos as medidas de divergência de Kullback-Leibler, a distância J e a distância L_1 (veja Subseção 1.7), apresentadas nos gráficos da Figura 2.2. Estas medidas não identificaram nenhuma observação influente no conjunto de dados, para qualquer um dos dois modelos ajustados.

Modelo	LPML	DIC	EAIC	EBIC	WAIC_1	WAIC_2	p_B
N-CR	**-23.164**	**46.132**	**44.642**	**48.019**	**45.914**	**46.288**	0.405
t-CR	-23.422	46.912	46.327	51.393	46.483	46.812	0.530

Tabela 2.2: *Insulation life data with censoring times.* Comparação entre os modelos N-CR e t-CR.

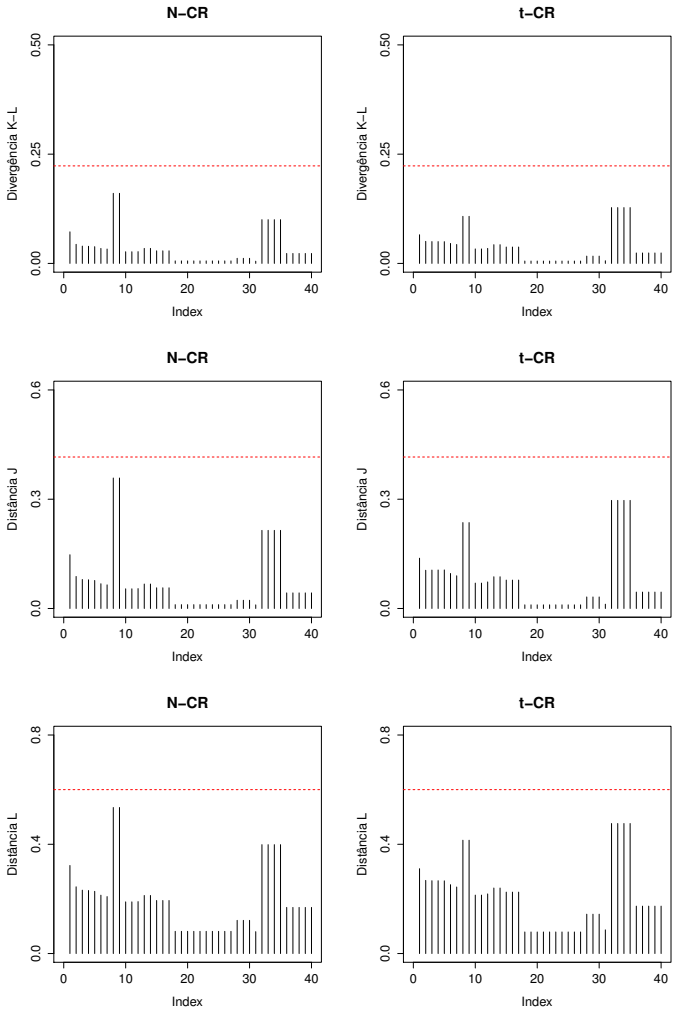


Figura 2.2: *Insulation life data with censoring times.* Divergência de Kullback-Leibler e distâncias J e L_1 sob os modelos N-CR e t-CR.

A avaliação da qualidade das estimativas pontuais obtidas com o amostrador de Gibbs e da robustez dos modelos N-CR e t-CR será feita juntamente com a dos modelos assimétricos no próximo capítulo (Seções 3.5), através de um estudo de simulação.

2.5 Inferência clássica para os modelos N-CR e t-CR

Nesta Seção desenvolveremos o algoritmo EM para fazer inferência para os modelos N-CR e t-CR sob um ponto de vista frequentista (veja Seção 1.4 para mais detalhes sobre este algoritmo). Discutiremos diagnóstico de influência para estes modelos com base em técnicas de deleção de casos e influência local. A performance dos dois modelos e a qualidade das estimativas obtidas com o algoritmo EM serão então avaliados através de estudos de simulação. Por fim, as técnicas aqui desenvolvidas serão aplicadas em um conjunto de dados reais. No intuito de simplificar a notação, nesta Seção denotaremos respectivamente por $\mathbb{E}_{\boldsymbol{\theta}}[Y]$ e por $Var_{\boldsymbol{\theta}}[Y]$, a esperança e variância de Y calculadas sob a suposição de que a distribuição de Y é indexada pelo vetor de parâmetros $\boldsymbol{\theta}$, quando esta for uma informação relevante.

Os modelos N-CR e t-CR já foram definidos anteriormente na Seção 2.3, portanto, começaremos esta Seção enunciando alguns resultados que serão bastante úteis na construção do algoritmo EM para os modelos N-CR e t-CR. Os dois lemas seguintes encontram-se demonstrados em Genç (2013) (veja também Kim, 2008a) e apresentam os dois primeiros momentos das distribuições Normal e t de Student truncadas (veja a Subseção 1.3.2 para a definição deste tipo de distribuições).

Lema 1. *Seja* $Y \sim \mathrm{TN}(\mu, \sigma^2; (a, b))$*, então:*

$$\begin{aligned}
\mathbb{E}_Y[Y] &= \mu + \sigma \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}, \\
\mathbb{E}_Y[Y^2] &= \mu^2 + \frac{2\mu\sigma\left[\phi(\alpha) - \phi(\beta)\right] + \sigma^2\left[\phi(\beta) - \phi(\alpha) + \alpha\phi(\alpha) - \beta\phi(\beta)\right]}{\Phi(\beta) - \Phi(\alpha)},
\end{aligned}$$

onde $\alpha = \frac{a-\mu}{\sigma}$ *e* $\beta = \frac{b-\mu}{\sigma}$.

Lema 2. *Se* $Y \sim \mathrm{Tt}(\mu, \sigma^2, \nu; (a,b))$, *então:*

$$\begin{aligned}
\mathbb{E}_Y[Y] &= \mu + G(\nu)\left\{(\nu+\alpha^2)^{-(\nu-1)/2} - (\nu+\beta^2)^{-(\nu-1)/2}\right\}\sigma,\ \nu > 1,\\
\mathbb{E}_Y[Y^2] &= \mu^2 + \sigma^2\left\{A(\nu) + G(\nu)\left[\alpha(\nu+\alpha^2)^{-(\nu-1)/2} - \beta(\nu+\beta^2)^{-(\nu-1)/2}\right]\right\}\\
&\quad +2\mu\sigma G(\nu)\left\{(\nu+\alpha^2)^{-(\nu-1)/2} - (\nu+\beta^2)^{-(\nu-1)/2}\right\}, \nu > 2,
\end{aligned}$$

onde $A(\nu) = \left(\dfrac{\nu}{\nu-2}\right)\dfrac{\mathcal{T}(\beta^*;0,1,\nu) - \mathcal{T}(\alpha^*;0,1,\nu)}{\mathcal{T}(\beta;0,1,\nu) - \mathcal{T}(\alpha;0,1,\nu)}$, $G(\nu) = \dfrac{\Gamma((\nu-1)/2)\nu^{\nu/2}}{2(\mathcal{T}(\beta;0,1,\nu) - \mathcal{T}(\alpha;0,1,\nu))\Gamma(\nu/2)\Gamma(1/2)}$, $\alpha = \frac{a-\mu}{\sigma}$, $\beta = \frac{b-\mu}{\sigma}$, $\alpha^* = \frac{\alpha}{\sqrt{(\nu-2)/\nu}}$, $\beta^* = \frac{\beta}{\sqrt{(\nu-2)/\nu}}$.

O resultado seguinte será útil na implementação do algoritmo EM e na estimação da matriz de variância-covariância dos estimadores dos parâmetros da regressão.

Lema 3. *Seja* $Y \sim \mathrm{Tt}(\mu, \sigma^2, \nu; (a,b))$ *e* $d^2(Y) = (Y-\mu)^2/\sigma^2$. *Então, para* $k = 0, 1, 2$ *e para* $r = 1, 2$*:*

$$\begin{aligned}
\mathbb{E}_Y\left[\left(\frac{\nu+1}{\nu+d^2(Y)}\right)^r Y^k\right] &= c(\nu,r)\mathbb{E}_X[X^k]\left[\mathcal{T}\left(\frac{b-\mu}{\sigma^*};0,1,\nu+2r\right) - \mathcal{T}\left(\frac{a-\mu}{\sigma^*};0,1,\nu+2r\right)\right]\\
&\quad \times\left[\mathcal{T}\left(\frac{b-\mu}{\sigma};0,1,\nu\right) - \mathcal{T}\left(\frac{a-\mu}{\sigma};0,1,\nu\right)\right]^{-1},
\end{aligned}$$

onde

$$X \sim \mathrm{Tt}(\mu, \sigma^{*2}, \nu+2r; (a,b)), \quad com \quad \sigma^{*2} = \frac{\nu}{(\nu+2r)}\sigma^2,$$

e

$$c(\nu,r) = \left(\frac{\nu+1}{\nu}\right)^r \frac{\Gamma((\nu+1)/2)\Gamma((\nu+2r)/2)}{\Gamma(\nu/2)\Gamma((\nu+2r+1)/2)}.$$

Prova:

Após alguma manipulação algébrica, pode-se ver que:

$$\left(\frac{\nu+1}{\nu+d^2(Y)}\right)^r t(x;\mu,\sigma^2,\nu) = c(\nu,r)\, t(x;\mu,\sigma^{*2},\nu+2r),$$

e isto implica que:

$$\mathbb{E}_Y\left[\left(\frac{\nu+1}{\nu+d^2(Y)}\right)^r Y^k\right] = \frac{c(\nu,r)P(W\in(a,b))}{P(Z\in(a,b))}\int_{(a,b)} w^k \frac{t(w;\mu,\sigma^{*2},\nu+2r)}{P(W\in(a,b))}\,dw,$$

onde $Z \sim \mathrm{t}(\mu,\sigma;\nu)$ e $W \sim \mathrm{t}(\mu,\sigma^{*2},\nu+2r)$. Daqui, a obtenção do resultado é direta utilizando os dois primeiros momentos da distribuição t de Student truncada, dados no Lema 2.

■

2.5.1 Construção do algoritmo EM

Considerando os modelos N-CR e t-CR deifnidos anteriormente, construiremos um algoritmo do tipo EM para estimar o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$ destes modelos. Note que, no caso do modelo t-CR, estamos considerando os graus de liberdade ν como uma constante fixa e não como um parâmetro a ser estimado e é claro que esta escolha precisa ser justificada com algum argumento teórico. Nesta direção, o trabalho de Fernandez & Steel (1999) é crucial, pois nele são discutidos alguns problemas que podem surgir na inferência como um todo devido à estimação dos graus de liberdade, em especial para a distribuição t de Student. Isto é devido à função de verossimilhança não ser limitada perto da fronteira do espaço paramétrico, o que torna questionável o esquema de estimação por máxima verossimilhança desenvolvido em Lange & Sinsheimer (1993) por não fornece informação suficiente para concluir se a estimativa do vetor de parâmetros obtida é de fato o ponto de máximo global ou é simplesmente um ponto de máximo local. Além disso, Lucas (1997) mostra que as estimativas paramétricas se comportam de forma mais robusta na presença de observações atípicas quando a hipótese de graus de liberdade fixos é feita. Dito isso, os desenvolvimentos desta Seção considerarão ν uma constante fixa, e na aplicação em dados reais, o valor mais apropriado para ν será escolhido através de um procedimento baseado nos critérios de seleção AIC e BIC, (veja Lange *et al.*, 1989; Meza *et al.*, 2012).

Para desenvolver o algoritmo EM iremos utilizar novamente a representação estocástica da variável resposta censurada Y dada na Definição 1, de forma que:

$$Y_i | U_i = u_i \sim \mathrm{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, u_i^{-1}\sigma^2), \tag{2.5.1}$$

onde $U_i = 1$ com probabilidade 1 sob o modelo N-CR e $U_i \sim G(\nu/2, \nu/2)$ sob o modelo t-CR, para $i = 1, 2, \ldots, n$.

Assim como no estudo de inferência Bayesiano feito Seção 2.4, também no caso frequentista a chave para o desenvolvimento do algoritmo de estimação será considerar os "dados aumentados", isto é, considerar que o vetor de variáveis sujeita à censura $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ e, no caso do modelo t-CR, também o vetor de variáveis latentes $\boldsymbol{U} = (U_1, U_2, \ldots, U_n)^\top$ poderiam ser de fato observados. Sob este esquema, usamos a Equação (2.5.1) para obter a função de log-verossimilhança

completa, dada por:

$$\ell_c(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 + \frac{n}{2}\sum_{i=1}^{n}\log u_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}u_i(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2 + \sum_{i=1}^{n}\log h(u_i), \quad (2.5.2)$$

onde $h(u_i)$ é a densidade de U_i, isto é, $\mathbb{1}_{\{1\}}(u_i)$ se estamos trabalhando sob o N-CR ou a densidade $G(\nu/2,\nu/2)$, se sob o t-CR, para $i = 1, 2, \ldots, n$.

No que segue, o sobrescrito (k) indica a estimativa do respectivo parâmetro na $k-$ésima iteração do algoritmo. No passo E do algoritmo EM, obtemos a função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\ell_c(\boldsymbol{\theta};\boldsymbol{Y},\boldsymbol{U})|\boldsymbol{v}],$$

que fica completamente determinada através do cálculo das seguintes esperanças:

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s|\boldsymbol{v}], \;\; s = 0, 1, 2, \quad (2.5.3)$$

já que $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log U_i|\boldsymbol{v}]$ e $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log h(U_i)|\boldsymbol{v}]$ dependem somente de ν, que é suposto conhecido. Desta forma, como o interesse é maximizar a função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ em relação a $\boldsymbol{\theta}$, podemos omitir os termos constantes em $\boldsymbol{\theta}$ e escrever tal função de uma forma mais sintética:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\mathcal{E}_{2i}(\boldsymbol{\theta}^{(k)}) - 2\mathcal{E}_{1i}(\boldsymbol{\theta}^{(k)})\mathbf{x}_i^\top\boldsymbol{\beta} + \mathcal{E}_{0i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_i^\top\boldsymbol{\beta})^2\right]. \quad (2.5.4)$$

Os dois seguintes Lemas têm como intuito apresentar o formato das esperanças $\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)})$ sob os modelos N-CR e t-CR.

Lema 4. *Suponha que $Y \sim t(\mu,\sigma^2,\nu)$, de forma que vale a representação estocástica apresentada na Definição 1, com $U \sim G(\nu/2,\nu/2)$. Então:*

$$\mathbb{E}[U|Y=y] = \frac{\nu+1}{\nu+d^2(y)}, \quad (2.5.5)$$

$$\mathbb{E}[U^2|Y=y] = \frac{(\nu+1)(\nu+3)}{(\nu+d^2(y))^2}, \quad (2.5.6)$$

$$Var[U|Y=y] = \frac{2(\nu+1)}{\nu+d^2(y)}, \quad (2.5.7)$$

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

Prova:

Para $m = 1, 2$, temos que:

$$\begin{aligned}
\mathbb{E}[U^m|y] &= \frac{1}{f(y)} \int u f(y|u) f(u)\, du \\
&= \frac{(\nu/2)^{(\nu/2)}}{\sqrt{2\pi\sigma^2} f(y) \Gamma(\nu/2)} \int u^{\frac{\nu+1+2m}{2}-1} \exp\left\{ -\frac{u}{2} \left[\frac{(y-\mu)^2}{\sigma^2} + \frac{\nu}{2} \right] \right\} du \\
&= \frac{(\nu/2)^{(\nu+1)/2}}{\Gamma\left(\frac{\nu+1}{2}\right)} \left[1 + \frac{d^2(y)}{\nu} \right]^{(\nu+1)/2} \int u^{\frac{\nu+1+2m}{2}-1} \exp\left\{ -\frac{u}{2} \left[d^2(y) + \nu \right] \right\} du.
\end{aligned}$$

Defina:

$$a = \frac{\nu + 1 + 2m}{2} \text{ e } b = \frac{1}{2} \left[\nu + d^2(y) \right],$$

então:

$$\begin{aligned}
\mathbb{E}[U^m|y] &= \frac{(\nu/2)^{(\nu+1)/2}}{\Gamma\left(\frac{\nu+1}{2}\right)} \left(\frac{2b}{\nu} \right)^{(\nu+1)/2} \int u^{a-1} \exp\{-b\, u\}\, du \\
&= \frac{(\nu/2)^{(\nu+1)/2}}{\Gamma\left(\frac{\nu+1}{2}\right)} \left(\frac{2b}{\nu} \right)^{(\nu+1)/2} \frac{\Gamma(a)}{b^a} \\
&= \frac{(\nu/2)^{-m}}{\Gamma(\frac{\nu+1}{2})} \Gamma\left(\frac{\nu+1}{2} + m \right) \left(1 + \frac{d^2(y)}{\nu} \right)^{-m}.
\end{aligned}$$

Simplificando a última equação para $m = 1$ e $m = 2$ obtemos as expressões para $\mathbb{E}[U|Y = y]$ e $\mathbb{E}[U^2|Y = y]$. A $Var[U|Y = y]$ pode então ser facilmente calculada através da relação: $Var[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X]$.

■

Lema 5. *Considere o modelo de regressão para respostas censuradas definidos em (2.3.1) e (2.3.2) e suponha que Y_i admite a representação estocástica dada na Definição1, de forma que, sob os modelos N-CR e t-CR, vale o resultado em (2.5.1). Seja $\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)})$ conforme definido em (2.5.3), então, para $s = 0, 1, 2$:*

- Se a $i-$ésima observação não é censurada*:*

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = v_i^s, \qquad \textit{sob o modelo N-CR,} \tag{2.5.8}$$

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = \frac{v_i^s(\nu+1)}{\nu + d^2(\boldsymbol{\theta}^{(k)}, v_i)}, \qquad \textit{sob o modelo t-CR.} \tag{2.5.9}$$

- Se a $i-$ésima observação é censurada*:*

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[Y_i^s | Y_i \in \mathbb{A}_i\right], \qquad \textit{sob o modelo N-CR,} \tag{2.5.10}$$

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\frac{(\nu+1)Y_i^s}{\nu + d^2(\boldsymbol{\theta}^{(k)}, Y_i)} | Y_i \in \mathbb{A}_i\right], \qquad \textit{sob o modelo t-CR,} \tag{2.5.11}$$

onde $\mathbb{A}_i = [v_i, \infty)$ *se estamos trabalhando com censuras à direita e* $\mathbb{A}_i = (\infty, v_i]$*, se com censuras à esquerda. As esperanças em (2.5.10) e (2.5.11) podem ser obtidas através dos Lemas 1 e 3, respectivamente.*

Prova:

Primeiro, suponha que estamos trabalhando sob o modelo N-CR, de forma que a distribuição de Y_i seja uma $N(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)}, \sigma^{2(k)})$ e que $U_i = 1$ com probabilidade 1 na Definição 1, $i = 1, 2, \ldots, n$. Desta forma, se a $i-$ésima observação não é censurada, $Y_i = V_i$ e $\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Y_i^s | Y_i = v_i] = v_i^s$, para $s = 1, 2$. Agora, se a $i-$ésima observação é censurada, $Y_i \in \mathbb{A}_i$, onde $\mathbb{A}_i$ é como especificado no Lema 5, portanto $\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Y_i^s | Y_i \in \mathbb{A}_i]$. Esta esperança corresponde ao $s-$ésimo momento de uma distribuição $NT(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)}, \sigma^{2(k)}; \mathbb{A}_i)$, que pode ser facilmente calculado através do Lema 1.

Suponha agora que estamos trabalhando sob o modelo t-CR, então $Y_i \sim t(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)} \sigma^{2(k)}, \nu)$ e $U_i \sim G(\nu/2, \nu/2)$, $i = 1, 2, \ldots, n$. Se a $i-$ésima observação não é censurada, $Y_i = V_i$ e:

$$\begin{aligned} \mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | Y_i = v_i] \\ &= v_i^s \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i | Y_i = v_i] \\ &= v_i^s \frac{\nu+1}{\nu + d^2(\boldsymbol{\theta}^{(k)}, v_i)}, \end{aligned} \tag{2.5.12}$$

onde $d(\boldsymbol{\theta}^{(k)}, v_i) = \frac{v_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(k)}}{\sigma^{(k)}}$ e a Equação (2.5.12) foi obtida utilizando o Lema 4.

Por outro lado, se a observação i for censurada, temos que $Y_i \in \mathbb{A}_i$, onde $\mathbb{A}_i$ é como especificado

no Lema 5. Desta forma:

$$\begin{aligned}
\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | Y_i \in \mathbb{A}_i] \\
&= \int\int u_i y_i^s f(u_i|Y_i = y_i, Y_i \in \mathbb{A}_i) f(y_i|Y_i = y_i, Y_i \in \mathbb{A}_i) du_i dy_i \\
&= \int y_i^s \left[\int u_i f(u_i|y_i)\, du_i \right] f(y_i|Y_i \in \mathbb{A}_i)\, dy_i \qquad (2.5.13) \\
&= \int \frac{(\nu+1)y_i^s}{\nu + d^2(\boldsymbol{\theta}^{(k)}, y_i)} f(y_i|Y_i \in \mathbb{A}_i)\, dy_i \qquad (2.5.14) \\
&= \mathbb{E}_{\boldsymbol{\theta}^{(k)}} \left[\frac{(\nu+1)Y_i^s}{\nu + d^2(\boldsymbol{\theta}^{(k)}, Y_i)} | Y_i \in \mathbb{A}_i \right], \qquad (2.5.15)
\end{aligned}$$

onde a Equação (2.5.14) vem do fato de que, se Y_i estivesse disponível, seria uma realização de uma distribuição $\mathrm{t}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2, \nu)$ e a integral mais interna da Equação (2.5.13) seria igual à esperança $\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i | Y_i = y_i]$, dada no Lema 4. Finalmente, a esperança em (2.5.15) pode ser calculada facilmente através do Lema 3, já que a distribuição de $Y_i | Y_i \in \mathbb{A}_i$ é uma $\mathrm{Tt}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2, \nu; \mathbb{A}_i)$.

■

Após determinarmos estas esperanças, maximizar a função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, é relativamente simples, bastando-nos igualar a 0 as primeiras derivadas parciais desta função em relação a $\boldsymbol{\beta}$ e a σ^2 (e verificar que de fato os argumentos encontrados são pontos de máximo através da análise das derivadas segundas desta função), de forma que os argumentos encontrados neste processo de maximização consistem no passo M do algoritmo. Desta forma, considerando a função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ dada na Equação (2.5.4), temos que a iteração $(k+1)$ do algoritmo EM pode ser sumarizada da seguinte forma:

Passo E: Dado $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, calcule $\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)})$ para $i = 1, 2, \ldots, n$ e para $s = 0, 1, 2$, utilizando o Lema 5.

Passo M: Atualize o valor de $\boldsymbol{\theta}^{(k)}$ maximizando $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ em relação a $\boldsymbol{\theta}$, que leva às seguintes expressões:

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \left(\sum_{i=1}^n \mathcal{E}_{0i}(\boldsymbol{\theta}^{(k)}) \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i=1}^n \mathbf{x}_i \mathcal{E}_{1i}(\boldsymbol{\theta}^{(k)}), \qquad (2.5.16) \\
\sigma^{2(k+1)} &= \frac{1}{n} \sum_{i=1}^n \left[\mathcal{E}_{2i}(\boldsymbol{\theta}^{(k)}) - 2\mathcal{E}_{1i}(\boldsymbol{\theta}^{(k)}) \mathbf{x}_i^\top \boldsymbol{\beta}^{(k+1)} + \mathcal{E}_{0i}(\boldsymbol{\theta}^{(k)}) (\mathbf{x}_i^\top \boldsymbol{\beta}^{(k+1)})^2 \right]. \qquad (2.5.17)
\end{aligned}$$

Este processo deve ser repetido até que alguma distância envolvendo duas avaliações sucessivas da log-verossimilhança seja pequena o suficiente, como por exemplo, $||\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})||$ ou $||\ell(\boldsymbol{\theta}^{(k+1)})/\ell(\boldsymbol{\theta}^{(k)}) - 1||$.

Em aplicações, é necessário especificar um valor para os graus de liberdade ν antes de aplicar o algoritmo desenvolvido nesta Subseção. Neste sentido, a sugestão que daremos neste trabalho é baseada nos critérios frequentistas de seleção de modelos: fixe alguns valores para ν, faça a estimação paramétrica e considere como resultado para análise final o cenário que fornecer maior valor para a função de verossimilhança. Um exemplo deste método será desenvolvido na Subseção 2.5.6.

2.5.2 Aproximação da variância dos estimadores dos parâmetros da regressão

No processo de inferência é importante também avaliar a variabilidade dos estimadores obtidos, neste caso, via algoritmo EM. Neste trabalho, aproximaremos a matriz de variância-covariância dos estimadores de $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$ sob os modelos N-CR e t-CR pela inversa da matriz de informação empírica, definida por:

$$\boldsymbol{I}_e(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{w}(v_i;\boldsymbol{\theta})\, \boldsymbol{w}(v_i;\boldsymbol{\theta})^\top \tag{2.5.18}$$

$$= -\frac{1}{n}\boldsymbol{W}(\boldsymbol{v};\boldsymbol{\theta})\, \boldsymbol{W}(\boldsymbol{v};\boldsymbol{\theta})^\top, \tag{2.5.19}$$

onde $\boldsymbol{W}(\boldsymbol{v};\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{w}(v_i;\boldsymbol{\theta})$ e o vetor score para cada observação, $\boldsymbol{w}(v_i;\boldsymbol{\theta})$, $i = 1, 2, \ldots, n$, pode ser obtido da seguinte forma:

$$\boldsymbol{w}(v_i;\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}; v_i)}{\partial \boldsymbol{\theta}} = \mathbb{E}\left[\frac{\partial \ell_c(\boldsymbol{\theta}; y_i, u_i)}{\partial \boldsymbol{\theta}} \mid v_i; \boldsymbol{\theta}\right], \tag{2.5.20}$$

onde $\ell_c(\boldsymbol{\theta}; y_i, u_i)$ é a função de log-verossimilhança completa, dada na Equação (2.5.2). O resultado da Equação (2.5.20), exibido em Louis (1982), pode ser provado observando que $\ell(\boldsymbol{\theta};\boldsymbol{v}) = \log f(\boldsymbol{v};\boldsymbol{\theta}) = \log \int_{\mathcal{R}} f(\boldsymbol{y};\boldsymbol{\theta}) d\boldsymbol{y}$, onde $\mathcal{R} = \{\boldsymbol{y} : \boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{v}\}$, isto é, $\mathcal{R}$ é o conjunto contido em $\mathbb{R}$ com os valores possíveis para o vetor de variáveis latentes $\boldsymbol{Y}$ capazes de gerar o vetor de observações $\boldsymbol{v}$

(veja a Seção 1.4). Desta forma, o vetor score é dado por:

$$\begin{aligned}
\boldsymbol{w}(\boldsymbol{v};\boldsymbol{\theta}) &= \frac{\partial \ell(\boldsymbol{\theta};\boldsymbol{v})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \log \int_{\mathcal{R}} f(\boldsymbol{y};\boldsymbol{\theta})\, d\boldsymbol{y} \\
&= \frac{\int_{\mathcal{R}} f^{'}(\boldsymbol{y};\boldsymbol{\theta})\, d\boldsymbol{y}}{\int_{\mathcal{R}} f(\boldsymbol{y};\boldsymbol{\theta})\, d\boldsymbol{y}} \qquad (2.5.21) \\
&= \frac{\int_{\mathcal{R}} \frac{f^{'}(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} f(\boldsymbol{y};\boldsymbol{\theta})\, d\boldsymbol{y}}{f(\boldsymbol{v};\boldsymbol{\theta})} \qquad (2.5.22) \\
&= \int_{\mathcal{R}} \frac{\partial \ell_c(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}} \frac{f(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{v};\boldsymbol{\theta})}\, d\boldsymbol{y} \\
&= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{V}} \left[\frac{\partial \ell_c(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}} \mid \boldsymbol{v} \right], \qquad (2.5.23)
\end{aligned}$$

onde $f^{'}(\boldsymbol{y};\boldsymbol{\theta})$ representa a primeira derivada de $f(\boldsymbol{y};\boldsymbol{\theta})$ com relação a $\boldsymbol{\theta}$, a Equação (2.5.22) foi obtida multiplicando-se e dividindo o integrando do numerador em (2.5.21) por $f(\boldsymbol{y};\boldsymbol{\theta})$ e, finalmente, a Equação (2.5.23) vem de (1.4.1).

O vetor score para a $i-$ésima observação, $i = 1, 2, \ldots, n$, é decomposto em

$$\boldsymbol{w}(v_i;\boldsymbol{\theta}) = \left(\boldsymbol{w}_{\boldsymbol{\beta}}(v_i;\boldsymbol{\theta}), w_{\sigma 2}(v_i;\boldsymbol{\theta}) \right),$$

e cada um destes elementos é dado por:

$$\begin{aligned}
\boldsymbol{w}_{\boldsymbol{\beta}}(v_i;\boldsymbol{\theta}) &= \mathbb{E}\left[\frac{\partial \ell_c(\boldsymbol{\theta};Y_i,U_i)}{\partial \boldsymbol{\beta}} | V_i = v_i \right] \\
&= \frac{\boldsymbol{x}_i}{\sigma^2} \mathcal{E}_{1i}(\boldsymbol{\theta}) - \frac{\boldsymbol{x}_i \boldsymbol{x}_i^{\top}}{\sigma^2} \boldsymbol{\beta} \mathcal{E}_{0i}(\boldsymbol{\theta}), \qquad (2.5.24)
\end{aligned}$$

e

$$\begin{aligned}
\boldsymbol{w}_{\sigma^2}(v_i;\boldsymbol{\theta}) &= \mathbb{E}\left[\frac{\partial \ell_c(\boldsymbol{\theta};Y_i,U_i)}{\partial \sigma^2} | V_i = v_i \right] \\
&= \frac{-1}{2\sigma^2} + \frac{1}{2\sigma^4} \left(\mathcal{E}_{2i}(\boldsymbol{\theta}) - 2\boldsymbol{x}_i^{\top} \boldsymbol{\beta} \mathcal{E}_{1i}(\boldsymbol{\theta}) + (\boldsymbol{x}_i^{\top} \boldsymbol{\beta})^2 \mathcal{E}_{0i}(\boldsymbol{\theta}) \right), \qquad (2.5.25)
\end{aligned}$$

onde as esperanças $\mathcal{E}_{si}(\boldsymbol{\theta}) = \mathbb{E}[Y_i^s U_i | V_i = v_i]$, para $s = 0, 1, 2$, são dadas para os modelos N-CR e t-CR no Lema 5.

Substituindo $\boldsymbol{\theta}$ por sua estimativa de máxima verossimilhança, $\hat{\boldsymbol{\theta}}$, a matriz de informação empírica observada pode ser calculada através das Equações (2.5.18), (2.5.24) e (2.5.25). Finalmente, a matriz de variância-covariância do estimador de máxima verossimilhança de $\boldsymbol{\theta}$ é estimada como $\boldsymbol{I}_e^{-1}(\hat{\boldsymbol{\theta}})$.

2.5.3 Análise de diagnóstico

Técnicas de diagnóstico de influência consistem em avaliar a sensibilidade das estimativas paramétricas quando ocorre uma perturbação no conjunto de dados ou em pressupostos do próprio modelo. Existem duas principais abordagens para a detecção de observações influentes, a primeira delas é a "deleção de casos" (Cook, 1977), na qual o efeito ou a influência de uma dada observação na inferência final do modelo é medida através da comparação entre as estimativas dos parâmetros obtidas com o conjunto de dados completo e eliminando-se dele tal observação, utilizando-se, para isso, alguma métrica adequada. A segunda abordagem é através do método da influência local Cook (1986), que avalia as alterações nos resultados da análise devido a uma perturbação nas hipóteses do modelo ou no conjunto de dados.

Utilizando os resultados da Subseção 2.5.1, introduziremos nesta Subseção medidas de influência para os modelos N-CR e t-CR baseando-nos na função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ previamente definida.

Medidas de influência via deleção de casos

A deleção de casos é um método clássico para estudar a influência da $i-$ésima observação do conjunto de dados na inferência final. No que segue, uma quantidade com o subescrito "$-i$ " denota a quantidade original sem o $i-$ésimo caso, por exemplo, $\boldsymbol{y}_{-1} = (y_2, y_3, \ldots, y_n)^\top$. A função de log-verossimilhança completa calculada após a exclusão da $i-$ésima observação do conjunto de dados será denotada por $\ell_c(\boldsymbol{\theta}; \boldsymbol{y}_{-i}, \boldsymbol{u}_{-i})$. Seja $\widehat{\boldsymbol{\theta}}_{-i} = (\widehat{\boldsymbol{\beta}}_{-i}^\top, \widehat{\sigma^2}_{-i})^\top$ o argumento que maximiza a função $Q_{-i}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \mathbb{E}_{\widehat{\boldsymbol{\theta}}}[\ell_c(\boldsymbol{\theta}; \boldsymbol{y}_{-i}, \boldsymbol{u}_{-i})|\boldsymbol{v}_{-i}]$, onde $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\sigma^2})^\top$ é a estimativa EM de $\boldsymbol{\theta}$.

Para acessar a influência da $i-$ésima observação nas estimativas de máxima verossimilhança de $\boldsymbol{\theta}$ iremos comparar a diferença entre $\widehat{\boldsymbol{\theta}}_{-i}$ e $\widehat{\boldsymbol{\theta}}$, de modo que se $\widehat{\boldsymbol{\theta}}_{-i}$ é distante de $\widehat{\boldsymbol{\theta}}$ (em algum sentido),

então o $i-$ésimo caso é considerado influente no processo de estimação e requer atenção especial. Como a obtenção dos valores $\widehat{\boldsymbol{\theta}}_{-i}$ é necessária para todo $i \in \{1, 2, \ldots, n\}$, o esforço computacional necessário pode ser bastante grande já que tais estimativas são alcançadas através de um algoritmo iterativo. Para tentar driblar este problema, apresentamos $\widehat{\boldsymbol{\theta}}^*_{-i}$, uma pseudo-aproximação de $\widehat{\boldsymbol{\theta}}_{-i}$ que pode ser obtida em um único passo (veja Cook & Weisberg, 1982):

$$\widehat{\boldsymbol{\theta}}^*_{-i} = \widehat{\boldsymbol{\theta}} + \{-\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{-i}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}), \tag{2.5.26}$$

onde

$$\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \text{ e } \dot{Q}_{-i}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \frac{\partial Q_{-i}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \tag{2.5.27}$$

são respectivamente a matriz Hessiana e o vetor gradiente da função $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ avaliados em $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, respectivamente. Em particular, a matriz Hessiana é um elemento essencial no método desenvolvido por Zhu *et al.* (2001) para obter as medidas de diagnóstico baseadas na deleção de casos e em influência local para um determinado esquema de perturbação (veja também Zhu *et al.*, 2009). As seguintes fórmulas podem ser obtidas facilmente a partir da relação (2.5.4).

$$\begin{aligned}
\dot{Q}_{-i\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial Q_{-i}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \frac{1}{\widehat{\sigma^2}}E_{(1)-i}, \\
\dot{Q}_{-i\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial Q_{-i}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{2\widehat{\sigma^2}}E_{(2)-i},
\end{aligned}$$

onde

$$E_{(1)-i} = \sum_{j\neq i}\left[\mathbf{x}_j\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\mathbf{x}_j\mathbf{x}_j^\top\widehat{\boldsymbol{\beta}}\right] \text{ e} \tag{2.5.28}$$

$$E_{(2)-i} = \sum_{j\neq i}\left[1 - \frac{1}{\widehat{\sigma^2}}\left(\mathcal{E}_{2j}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1j}(\boldsymbol{\theta})\mathbf{x}_j^\top\widehat{\boldsymbol{\beta}} + \mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})(\mathbf{x}_j^\top\widehat{\boldsymbol{\beta}})^2\right)\right]. \tag{2.5.29}$$

As derivadas parciais de segunda ordem de $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ avaliadas em $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\sigma^2})^\top$ são:

$$\ddot{Q}_\beta(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^n\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i\mathbf{x}_i^\top, \tag{2.5.30}$$

$$\begin{aligned}
\ddot{Q}_{\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\} &= \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2\partial\sigma^2}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \\
&= \frac{1}{2\widehat{\sigma^4}}\sum_{i=1}^n\left[1 - \frac{2}{\widehat{\sigma^2}}\left(\mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})^2\right)\right],
\end{aligned} \tag{2.5.31}$$

$$\ddot{Q}_{\beta\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\sigma^2}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{\widehat{\sigma^4}}\sum_{i=1}^n\left[\mathbf{x}_i\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}\right]. \tag{2.5.32}$$

Utilizando as expressões dadas em (2.5.16) e (2.5.17) com a substituição de $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)^\top}, \sigma^{2(k)})^\top$ por $\widehat{\boldsymbol{\theta}}$, pode-se mostrar que $\ddot{Q}_{\beta\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}$ é um vetor $p-$dimensional nulo. Isto significa que a matriz Hessiana é bloco-diagonal da forma:

$$\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \text{bloco diag}\{\ddot{Q}_{\beta}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}), \ddot{Q}_{\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\},$$

onde $\ddot{Q}_{\beta}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ e $\ddot{Q}_{\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ são dados em (2.5.30) e (2.5.31), respectivamente.

Com estes resultados, podemos aplicar a aproximação dada em (2.5.26) e obter fórmulas concisas para as aproximações de $\widehat{\boldsymbol{\theta}}_{-i} = (\widehat{\boldsymbol{\beta}}_{-i}^\top, \widehat{\sigma^2}_{-i})^\top$, $i = 1, 2, \ldots, n,$. O Teorema seguinte nos mostra esta relação entre as estimativas paramétricas obtidas com o conjunto de dados completo e excluindo-se dele a $i-$ésima observação. Sua prova será omitida pois trata-se simplesmente de manipulações algébricas.

Teorema 2. *Para os modelos N-CR e t-CR, uma aproximação para as estimativas de $\boldsymbol{\beta}$ e σ^2 obtidas após a exclusão do $i-$ésimo caso da amostra é dada por:*

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^*_{-i} &= \widehat{\boldsymbol{\beta}} + \left(\sum_{i=1}^{n} \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} E_{(1)-i}, \\
\widehat{\sigma^2}^*_{-i} &= \widehat{\sigma^2} + \left[1 - \frac{2}{\widehat{\sigma^2}}\left(\mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})^2\right)\right]^{-1} E_{(2)-i},
\end{aligned}$$

onde $\widehat{\boldsymbol{\beta}}$ e $\widehat{\sigma^2}$ são os estimadores EM de $\boldsymbol{\beta}$ e σ^2, $E_{(1)-i}$ e $E_{(2)-i}$ são dados nas Equações (2.5.28) e (2.5.29) (respectivamente) e as esperanças $\mathcal{E}_{si}$ são dadas para os modelos N-CR e t-CR no Lema 5, para $s = 0, 1, 2$.

Dado o resultado apresentado no Teorema 2, resta-nos agora escolher métricas adequadas para comparar as estimativas $\widehat{\boldsymbol{\theta}}_{-i}$ e $\widehat{\boldsymbol{\theta}}$, a fim de acessar a influência que a $i-$ésima observação exerce sobre a inferência final. Baseando-nos na métrica proposta por Zhu *et al.* (2001), iremos considerar primeiramente a *distância generalizada de Cook*, definida por:

$$GD_i = (\widehat{\boldsymbol{\theta}}_{-i} - \widehat{\boldsymbol{\theta}})^\top\{-\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}(\widehat{\boldsymbol{\theta}}_{-i} - \widehat{\boldsymbol{\theta}}), \quad i = 1, 2, \ldots, n. \tag{2.5.33}$$

Substituindo (2.5.26) em (2.5.33), obtemos a seguinte aproximação para esta métrica:

$$GD_i^* = \dot{Q}_{-i}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})^\top\{-\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{-i}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}), \quad i = 1, 2, \ldots, n.$$

Uma vez em que $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ é uma matriz bloco-diagonal, GD_i^* pode ser decomposta na soma:

$$GD_i^* = GD_i^*(\boldsymbol{\beta}) + GD_i^*(\sigma^2),$$

onde

$$\begin{aligned}
GD_i^*(\boldsymbol{\beta}) &= \dot{Q}_{-i\beta}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})^\top\{-\ddot{Q}_\beta(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{-i\beta}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) \\
&= \frac{1}{\widehat{\sigma^2}}E_{(1)-i}^\top\left[\sum_{i=1}^n \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i\mathbf{x}_i^\top\right]^{-1} E_{(1)-i} \text{ e} \\
GD_i^*(\sigma^2) &= \dot{Q}_{-i\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})^\top\{-\ddot{Q}_{\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{-i\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) \\
&= \frac{1}{2\widehat{\sigma^2}}\sum_{i=1}^n\left[1 - \frac{2}{\widehat{\sigma^2}}\left(\mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})^2\right)\right]^{-1}
\end{aligned}$$

são medidas da influência exercida pela $i-$ésima observação nas estimativas dos parâmetros $\boldsymbol{\beta}$ e σ^2, respectivamente, funcionando como versões da distância generalizada de Cook para cada um destes parâmetros.

Outra métrica que pode ser utilizada para nosso objetivo principal, que é comparar $\widehat{\boldsymbol{\theta}}_{-i}$ e $\widehat{\boldsymbol{\theta}}$, é a chamada "*distância-Q*". Esta medida, bastante análoga à "distância pela verossimilhança" LD_i (Cook & Weisberg, 1982), é definida como:

$$QD_i = 2\{Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - Q(\widehat{\boldsymbol{\theta}}_{-i}|\widehat{\boldsymbol{\theta}})\}. \tag{2.5.34}$$

Pode-se calcular uma aproximação da distância-Q através da substituição de (2.5.26) em (2.5.34), resultando em:

$$QD_i^* = 2\{Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - Q(\widehat{\boldsymbol{\theta}}_{-i}^*|\widehat{\boldsymbol{\theta}})\}.$$

Note que é necessário especificar um ponto limite acima do qual estas métricas classificam uma observação como influente. Neste sentido, faremos aqui uma adaptação da sugestão dada por Barros *et al.* (2010) e usaremos $2(p+1)/n$ como ponto limítrofe para GD_i, $2p/n$ para $GD_i(\boldsymbol{\beta})$ e $2/n$ para $GD_i(\sigma^2)$, para $i = 1, 2, \ldots, n$, onde p é a dimensão do vetor $\boldsymbol{\beta}$.

Medidas de influência local

Em geral os métodos de diagnóstico visam a verificação de possíveis afastamentos das suposições feitas pelo o modelo, bem como a identificação da existência de observações extremas com alguma interferência indesejada na inferência final. O método de influência local, introduzido por Cook (1986) consiste em verificar a existência de pontos que, sob pequenas modificações no modelo ou no próprio conjunto de dados, causam variações "desproporcionais" nos resultados do ajuste. Para este propósito vamos considerar dois esquemas diferentes de perturbação: a ponderação de casos e a perturbação na escala.

Considere um vetor de perturbações $\boldsymbol{\omega} = (\omega_1, ..., \omega_g)^\top$ variando em uma região aberta $\boldsymbol{\Omega} \subset \mathbb{R}^g$. Seja $\ell_c(\boldsymbol{\theta}, \boldsymbol{\omega}; \boldsymbol{y}, \boldsymbol{u})$ a função de log-verossimilhança completa do modelo perturbado. Assumiremos que existe um $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ tal que $\ell_c(\boldsymbol{\theta}, \boldsymbol{\omega}_0; \mathbf{y}, \boldsymbol{u}) = \ell_c(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{u})$, para todo $\boldsymbol{\theta}$. Defina:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) &= \mathrm{E}_{\widehat{\boldsymbol{\theta}}}[\ell_c(\boldsymbol{\theta}, \boldsymbol{\omega}; \mathbf{Y}, \boldsymbol{U})|\boldsymbol{v}] \quad \text{e} \\ \widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}) &= \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = (\widehat{\boldsymbol{\beta}}(\boldsymbol{\omega})^\top, \widehat{\sigma^2}(\boldsymbol{\omega}))^\top. \end{aligned}$$

O gráfico de influência é então definido como $\boldsymbol{\alpha}(\boldsymbol{\omega}) = (\boldsymbol{\omega}^\top, f_Q(\boldsymbol{\omega}))^\top$, onde $f_Q(\boldsymbol{\omega})$ é a chamada *função de afastamento*, dada por:

$$f_Q(\boldsymbol{\omega}) = 2\left[Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})|\widehat{\boldsymbol{\theta}})\right].$$

Seguindo o trabalho de Cook (1986) e Zhu & Lee (2001), a curvatura normal $C_{f_Q,\mathbf{d}}$ do gráfico $\boldsymbol{\alpha}(\boldsymbol{\omega})$ no ponto $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ na direção de um vetor unitário $\mathbf{d}$ pode ser utilizada para sumarizar o comportamento local da função de afastamento (para detalhes sobre a definição de curvatura normal de superfícies, veja do Carmo (2006)). Sejam

$$\boldsymbol{\Delta}_{\boldsymbol{\omega}} = \frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \text{ e } \ddot{Q}_{\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}.$$

Então, pode-se mostrar que

$$C_{f_Q,\mathbf{d}} = -2\mathbf{d}^\top \ddot{Q}_{\boldsymbol{\omega}_0} \mathbf{d} = 2\mathbf{d}^\top \boldsymbol{\Delta}_{\boldsymbol{\omega}_0}^\top \left\{-\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\right\}^{-1} \boldsymbol{\Delta}_{\boldsymbol{\omega}_0} \mathbf{d},$$

onde $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ é como definido em (2.5.27).

Seguindo o mesmo procedimento adotado por Cook (1986), a informação proveniente da matriz simétrica $-\ddot{Q}_{\boldsymbol{\omega}_0}$ é bastante útil para detectar observações influentes. Primeiro, considere a decomposição espectral desta matriz:

$$-2\ddot{Q}_{\boldsymbol{\omega}_0} = \sum_{k=1}^{g} \zeta_k \boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k^\top,$$

onde $\{(\zeta_k, \boldsymbol{\varepsilon}_k), k = 1, \ldots, g\}$ são pares de auto-valor e auto-vetor de $-2\ddot{Q}_{\boldsymbol{\omega}_0}$ com $\zeta_1 \geq \ldots \geq \zeta_r > \zeta_{r+1} = \ldots = 0$ e auto-vetores ortonormais $\boldsymbol{\varepsilon}_k, k = 1, \ldots, g$. Zhu & Lee (2001) propuseram inspecionar os auto-vetores correspondentes a auto-valores não nulos para capturar mais informação segundo o seguinte método: seja

$$\tilde{\zeta}_k = \frac{\zeta_k}{\zeta_1 + \ldots + \zeta_r}, \quad \boldsymbol{\varepsilon}_k^2 = (\varepsilon_{k1}^2, \ldots, \varepsilon_{kg}^2)^\top \quad \text{e} \quad M(0) = \sum_{k=1}^{r} \tilde{\zeta}_k \boldsymbol{\varepsilon}_k^2,$$

e seja $M(0)_l = \sum_{k=1}^{r} \tilde{\zeta}_k \varepsilon_{kl}^2$ a $l-$ésima componente do vetor $M(0)$. A detecção de observações influentes é baseada na inspeção visual do gráfico de $M(0)_l, l = 1, \ldots, g$ plotado contra o índice l. O $l-$ésimo caso é então considerado influente se $M(0)_l$ é maior do que um valor limítrofe adequado.

Utilizar a curvatura normal em sua forma original para avaliar a influência de uma determinada observação pode gerar alguns problemas, uma vez em que $C_{f_Q,\mathbf{d}}$ pode assumir qualquer valor na reta e não é invariante a mudanças de escala uniformes. Desta forma, com base no trabalho de Poom & Poon (1999) e de Zhu & Lee (2001), utilizaremos a curvatura normal conforme, dada por:

$$B_{f_Q,\mathbf{d}} = \frac{C_{f_Q,\mathbf{d}}}{\text{tr}[-2\ddot{Q}_{\boldsymbol{\omega}_0}]},$$

cujo cálculo é bastante simples e também possui a propriedade de que $0 \leq B_{f_Q,\mathbf{d}} \leq 1$. Seja $\mathbf{d}_l$ um vetor de zeros, exceto pela $l-$ésima componente que é igual a 1. Zhu & Lee (2001) mostraram que $M(0)_l = B_{f_Q,\mathbf{d}_l}$ para todo l. Desta forma, pode-se obter $M(0)_l$ via $B_{f_Q,\mathbf{d}_l}$.

Até o momento não foi dada nenhuma sugestão na literatura sobre uma regra geral para determinar um ponto limítrofe para $M(0)_l$, acima do qual a observação correspondente seria considerada

influente. Denote então por $\overline{M(0)}$ e por $SM(0)$ a média e o desvio padrão de $\{M(0)_l;\ l = 1, \ldots, g\}$, respectivamente. Utilizando o fato de que os vetores $\boldsymbol{\varepsilon}_k$ são ortonormais, não é difícil provar que $\overline{M(0)} = 1/g$. Poom & Poon (1999) propuseram utilizar $2\overline{M(0)}$ como um ponto limítrofe para $M(0)$, enquanto Zhu & Lee (2001) propuseram utilizar $\overline{M(0)} + 2SM(0)$ como tal, a fim de levar em consideração também a variância de $\{M(0)_l;\ l = 1, \ldots, g\}$. Há inúmeras escolhas de funções de $M(0)$ que servem como ponto limítrofe, de fato, segundo Lee & Xu (2004), a escolha de $\overline{M(0)}$ como tal função é subjetiva, por isso eles propuseram utilizar $\overline{M(0)} + c^* SM(0)$, onde c^* é uma constante apropriada cuja escolha depende da aplicação. Neste trabalho adotaremos esta sugestão utilizando $c^* = 3.5$.

A seguir, apresentaremos dois esquemas de perturbação. O primeiro é a ponderação de casos, que é apropriada para detectar observações com contribuições atípicas para a função de log-verossimilhança e, por isso, podem exercer influência significativa nas estimativas de máxima verossimilhança. O segundo esquema é a perturbação na escala, uma perturbação feita no parâmetro σ^2 que pode revelar o quão sensível são as estimativas do modelo acerca da hipótese de homocedasticidade e identificar as observações que mais contribuem para esta sensibilidade.

Nos dois esquemas de perturbação consideraremos $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$, isto é, o vetor de perturbação terá uma componente para cada observação no conjunto de dados. Observe que, uma vez em que a curvatura normal conforme pode ser calculada através somente de $\boldsymbol{\omega}_0$ e da matriz $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$, não é necessário especificar valores para cada componente de $\boldsymbol{\omega}$ para realizar a análise de diagnóstico, desde que a função de log-verossimilhança completa perturbada seja suave o suficiente para que as derivadas requeridas sejam bem definidas.

Para cada um dos dois esquemas de perturbação propostos, especificaremos o vetor $\boldsymbol{\omega}_0$ e avaliaremos o formato da matriz $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$, que pode ser particionada na forma:

$$\boldsymbol{\Delta}_{\boldsymbol{\omega}_0} = (\boldsymbol{\Delta}_\beta^\top, \boldsymbol{\Delta}_{\sigma^2}^\top)^\top,$$

onde

$$\boldsymbol{\Delta}_\beta = \frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}_0)} \in \mathbb{R}^{p\times g} \quad \text{e} \quad \boldsymbol{\Delta}_{\sigma^2} = \frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial \sigma^2 \partial \boldsymbol{\omega}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}_0)} \in \mathbb{R}^{1\times g}.$$

Ponderação de casos

Neste esquema de perturbação temos que $\boldsymbol{\omega}_0 = (1, \ldots, 1)^\top = \mathbf{1}_n$. Além disso, é possível mostrar que a influência local sob ponderação de casos é equivalente ao método de deleção de casos discutido anteriormente, portanto $\widehat{\boldsymbol{\beta}}(\boldsymbol{\omega}_0) = \widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}(\boldsymbol{\omega}_0) = \widehat{\sigma^2}$. A função $Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$, a versão perturbada de $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$, é dada por:

$$Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \mathbb{E}_{\widehat{\boldsymbol{\theta}}}[\ell_c(\boldsymbol{\theta}, \boldsymbol{\omega}; \boldsymbol{Y}, \boldsymbol{U})|\boldsymbol{v}] = \sum_{i=1}^{n} \omega_i \mathrm{E}_{\widehat{\boldsymbol{\theta}}}[\ell_{ci}(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{U})|\boldsymbol{v}] = \sum_{i=1}^{n} \omega_i Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}),$$

onde $Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ representa o termo da função $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ relativo à $i-$ésima observação, isto é:

$$Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^\top\boldsymbol{\beta} + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})(\mathbf{x}_i^\top\boldsymbol{\beta})^2\right].$$

Desta forma, sob este esquema de perturbação, os componentes de $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$ são dados por:

$$\begin{aligned}
\boldsymbol{\Delta}_{\beta} &= \frac{1}{\widehat{\sigma^2}}\left[\mathbf{X}^\top \mathrm{diag}\{\mathcal{E}_1(\widehat{\boldsymbol{\theta}})\} - \mathbf{A}\right], \\
\boldsymbol{\Delta}_{\sigma^2} &= -\frac{1}{2\widehat{\sigma^2}}\left\{\mathbf{1}_n^\top - \frac{1}{\widehat{\sigma^2}}\mathbf{B}^\top\right\},
\end{aligned}$$

onde $\mathbf{A}$ é uma matriz com n colunas igual a $\mathbf{X}^\top\mathrm{diag}\{\mathcal{E}_0(\widehat{\boldsymbol{\theta}})\}\mathbf{X}^\top\widehat{\boldsymbol{\beta}}$, com $\mathcal{E}_j(\widehat{\boldsymbol{\theta}}) = (\mathcal{E}_{j1}(\widehat{\boldsymbol{\theta}}), \ldots, \mathcal{E}_{jn}(\widehat{\boldsymbol{\theta}}))^\top$, para $j = 1, 2$, $\mathbf{X}$ é uma matriz com linhas dada pelos vetores $\mathbf{x}_i^\top$ (isto é, a matriz de desenho) e $\mathbf{B}$ é um vetor $n-$dimensional com coordenadas $B_i = \mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}})^2$, para $i = 1, 2, \ldots, n$.

Perturbação na escala

A fim de identificar observações que influenciam de forma significativa nas estimativas paramétricas quando a hipótese de homocedasticidade não é válida, consideramos a perturbação $\sigma^2(\omega_i) = \omega_i^{-1}\sigma^2$, para $i = 1, 2, \ldots, n.$, isto é, sob esta perturbação cada erro aleatório do modelo dado na Equação (2.3.1) segue uma distribuição Normal ou t de Student (dependendo do modelo assumido) com parâmetro de escala $\sigma^2(\omega_i)$, gerando um cenário homocedástico. É claro que o

vetor relacionado ao modelo não perturbado é dado por $\boldsymbol{\omega}_0 = \mathbf{1}_n$. Além disso, $Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$ é como na Equação (2.5.4), fazendo-se a substituição de σ^2 por $\sigma^2(\omega_i)$ e de $\boldsymbol{\theta}^{(k)}$ por $\widehat{\boldsymbol{\theta}}$.

A matriz $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$ tem os seguintes elementos:

$$\begin{aligned}\boldsymbol{\Delta}_{\beta} &= \frac{1}{\widehat{\sigma^2}}\left[\mathbf{X}^{\top}\mathrm{diag}\{\mathcal{E}_1(\widehat{\boldsymbol{\theta}})\} - \mathbf{A}\right], \\ \boldsymbol{\Delta}_{\sigma^2} &= \frac{1}{2\widehat{\sigma^4}}\mathbf{B}^{\top},\end{aligned}$$

onde $\mathbf{A}$ e $\mathbf{B}$ são como definidos no esquema de ponderação de casos.

2.5.4 Estudo de simulação I: Robustez das estimativas EM

O objetivo deste estudo de simulação é comparar a sensibilidade dos estimadores EM obtidos sob os modelos N-CR e t-CR quando ocorre uma perturbação na variável resposta, gerando observações atípicas. Para isso, geramos 1000 conjuntos de dados de tamanho 100 cada sob o modelo N-CR especificado nas Equações (2.3.1) e (2.3.2), com $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ e fixando $\boldsymbol{\beta}^{\top} = (\beta_1, \beta_2) = (1, 4)$, $\sigma^2 = 2$ e $\boldsymbol{x}_i^{\top} = (1, x_i)$, com x_i gerado aleatoriamente de uma distribuição uniforme no intervalo $(2, 20)$, para $i = 1, 2, \ldots, n$. É importante destacar que estes valores foram fixados para as 1000 simulações. Após gerado, cada conjunto teve sua variável resposta censurada à direita a um nível de 8%.

Para avaliar o quanto as estimativas EM são afetadas pela presença de observações atípicas substituímos (em todos os 1000 conjuntos de dados) a observação y_{50} por $y_{50}(\vartheta) = y_{50} - \vartheta$, com $\vartheta = 1, 2, \ldots, 10$. Para cada conjunto de dados obtido (incluindo os originais e aqueles com os 10 padrões de perturbação) foram ajustados os modelos N-CR e t-CR, com $\nu \in \{3, 6, 8, 10, 12\}$. Estamos interessados em avaliar a mudança relativa média sofrida pelas estimativas como uma função de ϑ. Para cada um dos 1000 conjuntos de dados originais, a mudança relativa sofrida pelo parâmetro θ sob uma perturbação ϑ é definida como:

$$RC(\theta) = |(\widehat{\theta}(\vartheta) - \widehat{\theta})/\widehat{\theta}|. \tag{2.5.35}$$

onde $\widehat{\theta}(\vartheta)$ é a estimativa EM do parâmetro θ obtida com o conjunto de dados perturbado em ϑ e $\widehat{\theta}$, com o conjunto de dados original.

A Tabela 2.3 e a Figura 2.3 descrevem os valores médios das mudanças relativas sofridas pelos parâmetros β_1, β_2 e σ^2 em função de ϑ, sob cada modelo ajustado. O que observamos é que, para todos os parâmetros, as mudanças relativas médias aumentam bruscamente sob o N-CR conforme cresce o valor de ϑ, enquanto que, sob o modelo t-CR, este aumento é bastante sutil (para todos os valores de ν). Este cenário mostra a maior sensibilidade das estimativas do modelo N-CR na presença de observações atípicas e a robustez do modelo t-CR para lidar com este tipo de problema. Além disso, pode-se notar que as estimativas dos parâmetros β_1 e σ^2 são as mais afetadas pela presença do "*outlier*", enquanto a de β_2 se mantém mais estável, com mudanças relativas médias de menor amplitude.

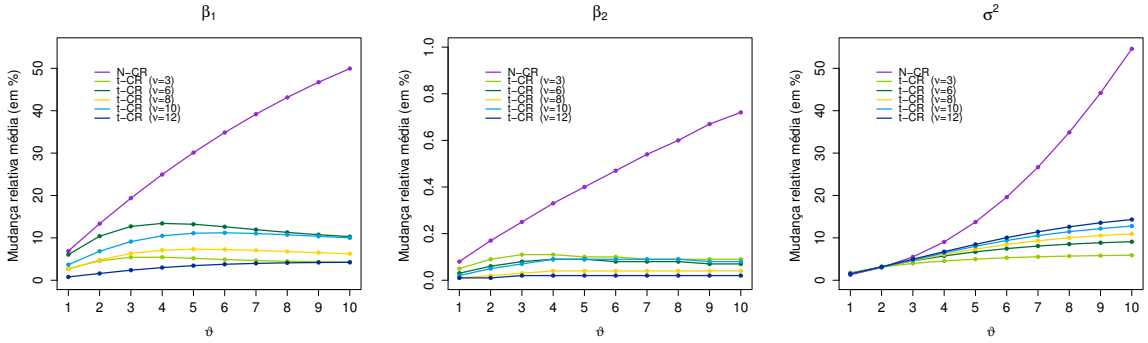


Figura 2.3: *Estudo de simulação I.* Mudança relativa média nas estimações para diferentes contaminações ϑ.

2.5.5 Estudo de simulação II: Desvios padrões dos estimadores EM

Neste estudo de simulação avaliaremos a qualidade do método sugerido na Subseção 2.5.2 para aproximar a variância dos estimadores de máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$ e σ^2 sob os modelos N-CR e t-CR, sob diferentes níveis de censura. Para isso, fixamos $\beta_1 = 2$, $\beta_2 = 1$, $\sigma^2 = 1$ e a matriz de desenho X igual a $(\mathbf{1}_{100}^\top, \boldsymbol{t} \otimes \mathbf{1}_{10})$, onde $t = (1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8)$ e $\mathbf{1}_p$ é um vetor $p-$dimensional com elementos iguais a 1, e geramos geramos 1000 conjuntos de dados de tamanho 100 segundo o modelo em (2.3.1) sob a distribuição Normal e outros 1000 também de tamanho 100 sob a distribuição t-Student com graus de liberdade igual a 4. Após gerado, cada

		ϑ									
	Modelos	1	2	3	4	5	6	7	8	9	10
β_1	N-CR	6.89	13.35	19.37	24.95	30.11	34.85	39.19	43.14	46.72	49.94
	t-CR ($\nu = 3$)	2.71	4.60	5.40	5.43	5.17	4.88	4.65	4.47	4.34	4.25
	t-CR ($\nu = 6$)	6.05	10.39	12.70	13.41	13.21	12.62	11.94	11.29	10.74	10.29
	t-CR ($\nu = 8$)	2.60	4.82	6.32	7.09	7.32	7.25	7.04	6.78	6.51	6.25
	t-CR ($\nu = 10$)	3.67	6.84	9.13	10.48	11.09	11.20	11.03	10.73	10.37	10.00
	t-CR ($\nu = 12$)	0.78	1.60	2.37	2.99	3.45	3.77	3.98	4.11	4.19	4.23
β_2	N-CR	0.08	0.17	0.25	0.33	0.40	0.47	0.54	0.60	0.67	0.72
	t-CR ($\nu = 3$)	0.05	0.09	0.11	0.11	0.10	0.10	0.09	0.09	0.09	0.09
	t-CR ($\nu = 6$)	0.03	0.06	0.08	0.09	0.09	0.08	0.08	0.08	0.07	0.07
	t-CR ($\nu = 8$)	0.01	0.02	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	t-CR ($\nu = 10$)	0.02	0.05	0.07	0.08	0.09	0.09	0.09	0.09	0.08	0.08
	t-CR ($\nu = 12$)	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
σ^2	N-CR	1.29	3.00	5.50	9.04	13.73	19.61	26.68	34.88	44.19	54.61
	t-CR ($\nu = 3$)	1.78	3.10	3.97	4.55	4.98	5.31	5.54	5.70	5.82	5.90
	t-CR ($\nu = 6$)	1.60	3.17	4.57	5.74	6.70	7.47	8.07	8.52	8.85	9.09
	t-CR ($\nu = 8$)	1.55	3.12	4.63	6.07	7.35	8.44	9.32	10.01	10.54	10.95
	t-CR ($\nu = 10$)	1.52	3.09	4.75	6.44	8.01	9.38	10.53	11.46	12.19	12.78
	t-CR ($\nu = 12$)	1.49	3.18	4.97	6.74	8.46	10.05	11.43	12.60	13.56	14.33

Tabela 2.3: Estudo de simulação I. Mudança relativa média nas estimações dos parâmetros β_1, β_2 e σ^2 para diferentes contaminações ϑ (em %).

conjunto de 100 observações da variável resposta foi censurado à direita em níveis de $5\%, 10\%, 20\%$ e 50%.

Para cada conjunto de dados foi ajustado o modelo adequado (N-CR ou t-CR), calculando as estimativas EM dos parâmetros envolvidos, a aproximação dos desvios padrões para β_1 e β_2 segundo os desenvolvimentos da Subseção 2.5.2 (cujo valor médio para as 1000 amostras é denotado por IM SE), e também um intervalo de 95% de confiança para tais parâmetros, utilizando a hipótese de normalidade assintótica. Em seguida, calculamos o desvio padrão observado para as 1000 estimativas de β_1 e de β_2 (denotado por SE), a fim de compará-lo com o IM SE calculado. Também foi calculada a cobertura média dos intervalos calculados (supondo normalidade assintó-

tica dos estimadores), denotada por COV, isto é, a porcentagem dos 1000 intervalos de confiança assintóticos que continham o verdadeiro valor do parâmetro. Os resultados estão apresentados na Tabela 2.4, sugerindo que, de fato, o método da Subseção 2.5.2 é apropriado sob os dois modelos e para os quatro níveis de censura considerados. Nota-se, no entanto, que a porcentagem de cobertura é um pouco prejudicada para os intervalos contruídos para σ^2, principalmente sob o t-CR, o que é compreensível já que a normalidade assintótica pode não ser válida para o estimador deste parâmetro.

		N-CR			t-CR		
Nível de censura (%)	Medida	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\sigma}^2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\sigma}^2$
5	IM SE	0.3526	0.1784	0.1519	0.4051	0.2041	0.1867
	SE	0.3463	0.1715	0.1502	0.4191	0.2120	0.2840
	COV	95.7%	96.1%	92.5%	93.6%	93.7%	80.0%
10	IM SE	0.3538	0.1796	0.1574	0.4048	0.2042	0.1905
	SE	0.3474	0.1728	0.1563	0.4204	0.2125	0.2842
	COV	95.7%	96.2%	91.7%	93.2%	93.6%	80.3%
20	IM SE	0.3582	0.1838	0.1701	0.4075	0.2067	0.2026
	SE	0.3529	0.1783	0.1724	0.4215	0.2131	0.2941
	COV	95.1%	95.8%	92.0%	93.6%	93.9%	82.1%
50	IM SE	0.3912	0.2154	0.2295	0.4505	0.2471	0.2732
	SE	0.3879	0.2095	0.2181	0.4722	0.2603	0.3916
	COV	94.7%	94.8%	91.3%	93.2%	93.1%	82.2%

Tabela 2.4: Estudo de simulação II. Valores médios (em 1000 amostras) do desvio padrão computado via matriz de informação empírica (IM SE), desvio padrão observado para os estimadores (SE), e cobertura média dos intervalos assintóticos de 95% de confiança (COV).

2.5.6 Aplicação II

Nesta Subseção ajustaremos os modelos N-CR e t-CR via algoritmo EM ao conjunto de dados descrito em Mroz (1987), "*Wage Rate*". Também faremos um estudo de diagnóstico destes modelos

com base na metodologia descrita nas Seções 2.5.3. O respaldo computacional desta aplicação é feito pelos pacotes `CensRegMod` e `SMNCensReg`, descritos na Subseção 4.1.3.

O conjunto de dados "*Wage Rate*" descreve os ganhos por hora (em dólares) de 753 mulheres brancas e casadas com idades entre 30 e 60 anos no ano de 1975, sendo que 325 destas mulheres não trabalharam neste ano e seus ganhos são tidos como 0. Algumas características pessoais e familiares destas mulheres também foram gravadas e funcionarão como variáveis explicativas para o nosso modelo.

Para ajustar o modelo de regressão para dados censurados teremos como objetivo modelar o "ganho potencial" destas mulheres, de forma que se um indivíduo trabalhou no ano de 1975 seu ganho potencial é sua própria renda, porém, se este não for o caso, seu ganho potencial é um valor negativo, representando o quanto deixou de ganhar por não ter trabalhado. Assim, o ganho potencial é uma variável aleatória sujeita a censura à esquerda, pois só conseguimos observar seu valor real se o indivíduo exerceu alguma atividade remunerada durante o ano de 1975, caso contrário somente sabemos que seu ganho potencial pertence ao intervalo $(-\infty, 0]$.

Dito isto, modelaremos a variável censurada à esquerda Y_i, definida como o ganho potencial do indivíduo i em função das variáveis explicativas:

- x_2: número de filhos (em casa) com menos de 6 anos,
- x_3: número de filhos (em casa) entre 6 e 19 anos,
- x_4: idade,
- x_5: anos de estudo,
- x_6: número de horas trabalhadas pelo marido em 1975,
- x_7: renda (por hora, em dólares) do marido em 1975,
- x_8: taxa de impostos pagos pela mulher,
- x_9: número de anos trabalhados (antes de 1975),

de forma que o vetor de covariáveis para o indivíduo i é dado por $\mathbf{x}_i^\top = (1, x_{2i}, x_{3i}, \ldots, x_{9i})$, com $i = 1, 2, \ldots, 753$.

O algoritmo EM desenvolvido na Subseção 2.5.1 foi aplicado para ajustar os modelos N-CR e t-CR a este conjunto de dados. Para o modelo t-CR, os graus de liberdade foram fixados em $\nu = 2.3$, escolha justificada através da Figura 2.4, que mostra os valores assumidos pela função de log-verossimilhança avaliada nas estimativas EM obtidas fixando ν nos valores $2, 2.1, 2.2, \ldots, 10$.

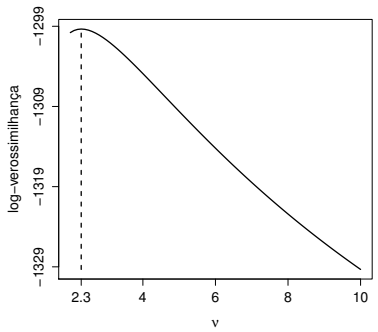


Figura 2.4: *Wage rate data*. Gráfico da log-verossimilhança perfilada para ν.

Na Tabela 2.5 estão apresentadas as estimativas paramétricas obtidas nos ajustes dos modelos N-CR e t-CR via algoritmo EM, bem como a aproximação dos desvios padrões para os estimadores dos efeitos fixos, denotada por SD (veja Subseção 2.5.2). Nesta tabela podemos notar que os desvios padrões estimados sob o t-CR são sempre menores do que sob o N-CR, indicando que o modelo t-CR produz estimativas mais precisas. Além disso, o pequeno valor de ν escolhido para os graus da liberdade da t de Student reflete um cenário em que o modelo N-CR pode não ser o mais adequado para este conjunto de dados, o que é comprovado pelos critérios de seleção de modelos apresentados na Tabela 2.6.

Para avaliar a adequação dos dois modelos ajustados, realizaremos uma análise de resíduos baseada em gráficos de envelope (veja Atkinson, 1985). Aqui, optamos por utilizar os resíduos "deviance" para gerar os envelopes pois, ao contrário de resíduos mais clássicos (como os de Pearson), estes são capazes de incorporar informação sobre as censuras e, segundo Ortega *et al.* (2003),

	Modelos			
	N-CR		t-CR	
Parâmetro	Estimativa	SD	Estimativa	SD
β_1	30.0152	3.6592	35.4547	2.6735
β_2	-2.0625	0.4054	-1.7787	0.2558
β_3	0.3536	0.1515	0.2893	0.0936
β_4	-0.1473	0.0265	-0.1306	0.0177
β_5	0.5378	0.0850	0.3822	0.0513
β_6	-0.0024	0.0003	-0.0026	0.0002
β_7	-0.5486	0.0648	-0.6521	0.0531
β_8	-32.3373	3.4749	-35.7824	2.4793
β_9	0.1753	0.0270	0.1381	0.0160
σ^2	16.8390	0.5759	4.3183	0.4356

Tabela 2.5: *Wage rates data.* Resultados dos ajustes dos modelos N-CR e t-CR via algoritmo EM.

Modelo	AIC	BIC	EDC
N-CR	2820.161	2866.401	2855.042
t-CR	**2660.688**	**2671.553**	**2659.058**

Tabela 2.6: Wage rate data. Comparação entre os modelos N-CR e t-CR.

podem ser utilizados para avaliar a qualidade do ajuste de um determinado modelo.

Primeiramente, defina o *resíduo de martingale* (Ortega *et al.*, 2003) para o $i-$ésimo indivíduo como:

$$m_i = (1 - c_i) + \log\left(1 - F\left(\frac{v_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\right),$$

onde $F(\cdot)$ é a cdf da versão padrão da densidade dos erros aleatórios (em nosso caso, $\mathrm{N}(0,1)$ para o N-CR ou $\mathrm{t}_\nu(0,1)$ para o t-CR) e c_i é uma função indicadora de censura, isto é, $c_i = 1$ se a $i-$ésima

observação é censurada ou $c_i = 0$, se não é. Finalmente o *resíduo deviance* para este indivíduo é definido como:

$$d_i = \text{sign}(m_i) \left[-2 \left\{ (1 - c_i) \log(1 - c_i - m_i) + m_i \right\} \right]^{1/2}.$$

Os gráficos de envelope gerados estão apresentados na Figura 2.5, onde podemos ver claramente que o modelo t-CR se ajusta melhor aos dados do que o N-CR, apesar de ainda apresentar indícios de má especificação (provavelmente relacionado à presença de assimetria dos dados).

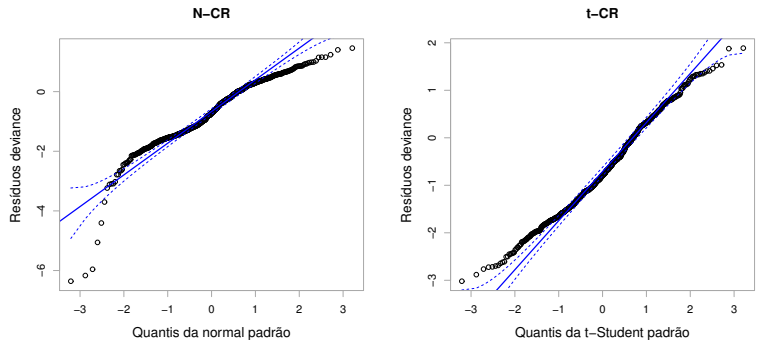


Figura 2.5: *Wage rates data*. Gráficos de envelope baseados nos resíduos deviance para os ajustes dos modelos N-CR e t-CR.

A robustez dos modelos N-CR e t-CR podem ser avaliadas através da influência que uma única observação atípica exerce sobre as estimativas EM do vetor de parâmetros $\boldsymbol{\beta}$. Em particular, podemos investigar a mudança relativa sofrida nas estimativas de $\boldsymbol{\beta}$ ao adicionar-se δ unidades em uma única observação y_i, substituindo-se y_i por $y_i(\delta) = y_i + \delta$. Aqui, utilizaremos a mesma definição de mudança relativa dada na Equação (2.5.35). A Figura 2.6 mostra as mudanças relativas sofridas pelas estimativas de $\beta_1, \beta_2, \ldots, \beta_9$ sob ambos os modelos ao contaminarmos a observação número 44 (não censurada) com valores de δ variando entre 0 e 40 em passos de tamanho 0.5. Como esperado, as estimativas do modelo t-CR são menos afetadas pelas contaminações e a robustez deste modelo em comparação com o N-CR fica mais clara à medida que aumenta-se o valor de δ.

A fim de verificar se existem obervações no conjunto de dados capazes de influenciar (de modo desproporcional) os ajustes dos dois modelos, prosseguimos com a análise de diagnóstico conforme

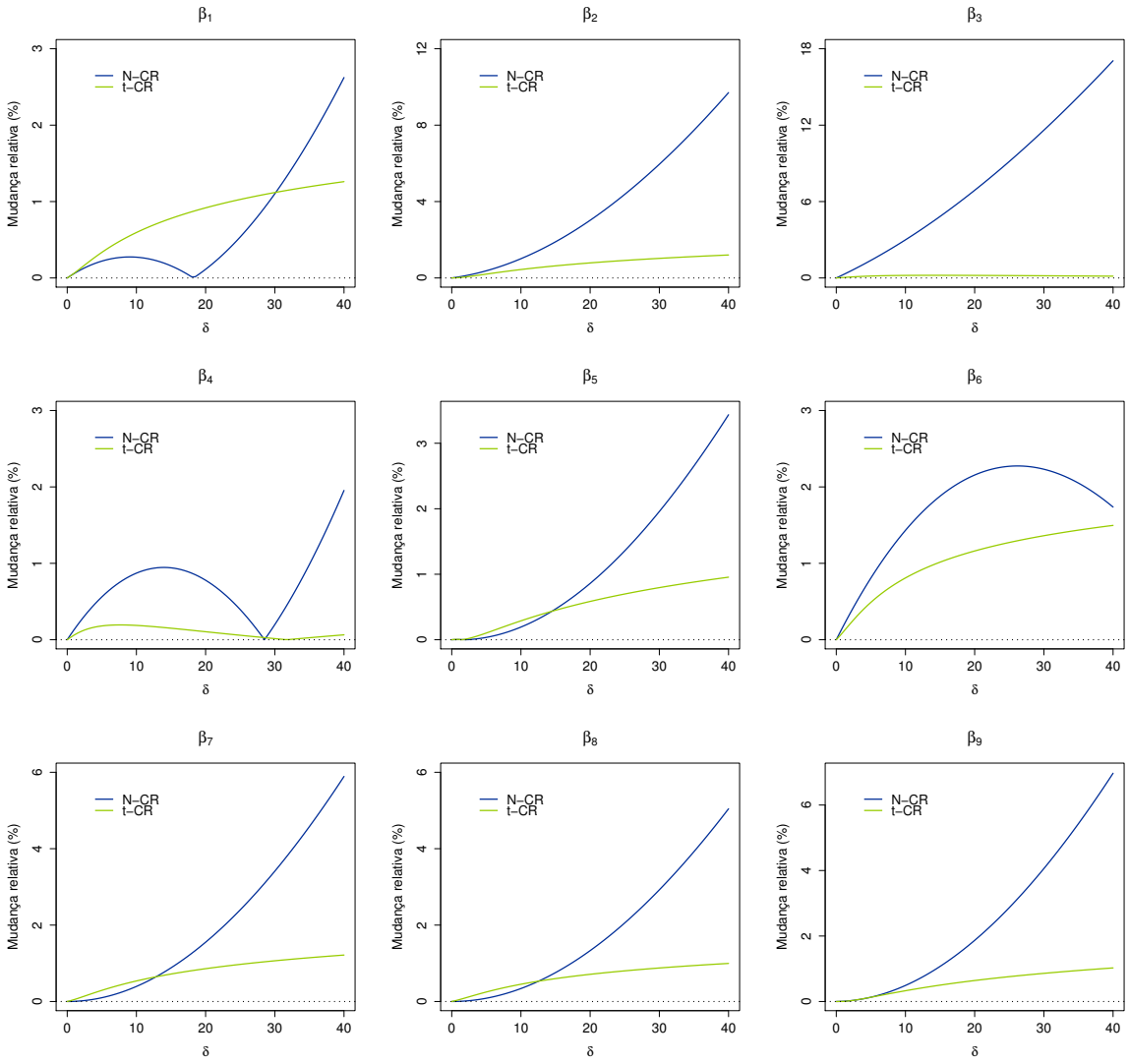


Figura 2.6: *Wage rates data*. Mudanças relativas nas estimativas EM de $\boldsymbol{\beta}$ sob os modelos N-CR e t-CR, para diferentes contaminações δ.

o desenvolvimento feito na Subseção 2.5.3. A Figura 2.7 mostra as distâncias generalizadas de Cook, assim como o ponto limiar acima do qual uma observação é considerada influente (linha vermelha), para os subconjuntos de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$, $\boldsymbol{\beta}$ e σ^2, sob os modelos N-CR e t-CR. No mesmo sentido, a Figura 2.8 mostra os gráficos para as medidas de influência local baseadas nas quantias $M(0)$ sob ponderação de casos e perturbação na escala para os modelos N-CR e t-CR, utilizando o critério $M(0)_i > \overline{M(0)} + 3.5SM(0)$ para classificar a $i-$ésima observação

como influente (linha vermelha). Em todos estes cenários podemos observar uma quantidade muito menor de observações influentes sob o t-CR do que sob o N-CR, refletindo mais uma vez a maior robustez do t-CR.

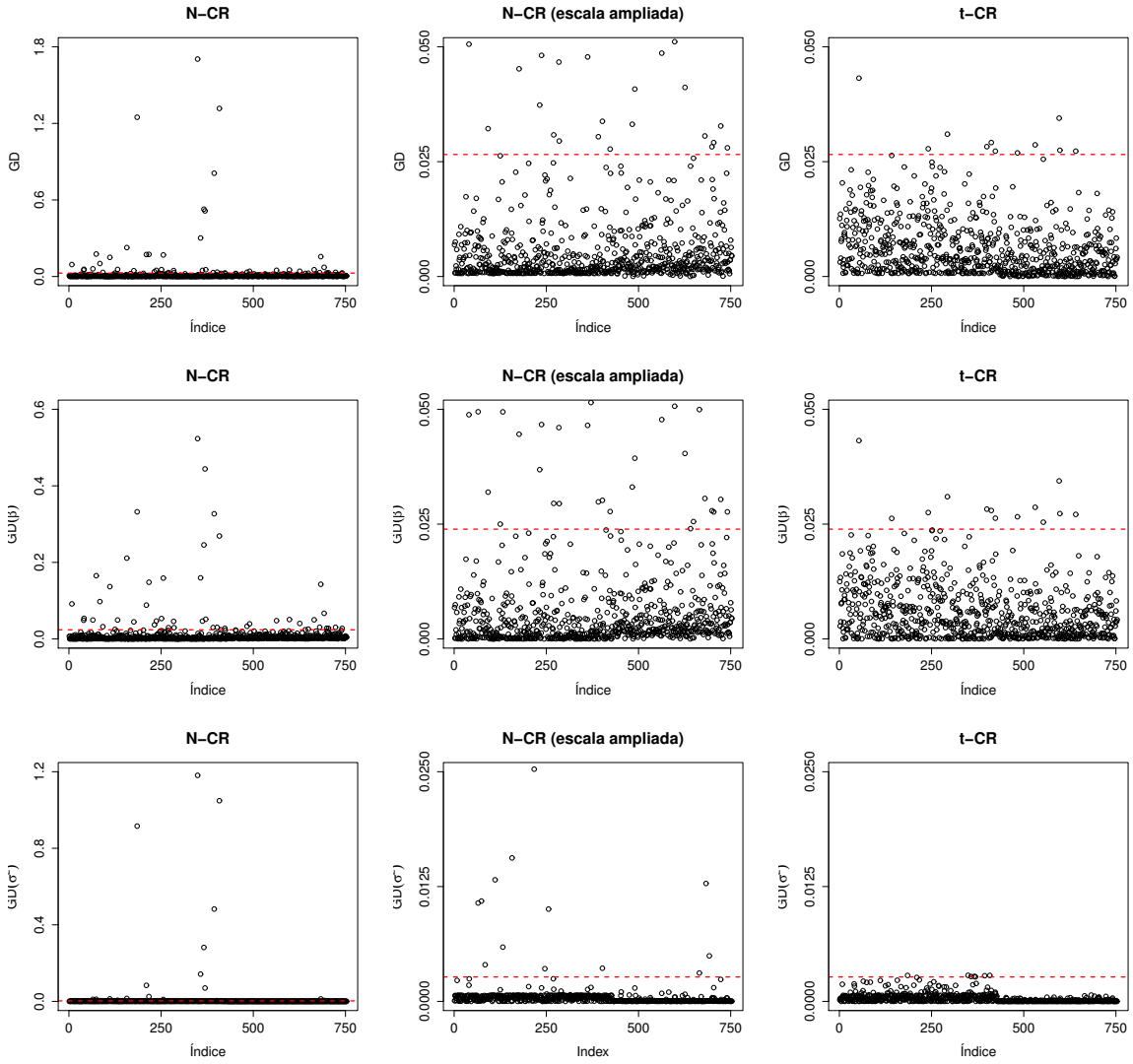


Figura 2.7: *Wage rate data.* Na primeira linha, distância generalizada de Cook GD_i sob os modelos N-CR e t-CR. Na segunda linha, GD_i para o subconjunto de parâmetros $\boldsymbol{\beta}$ sob os modelos N-CR e t-CR. Na terceira linha, GD_i para σ^2 sob os modelos N-CR e t-CR.

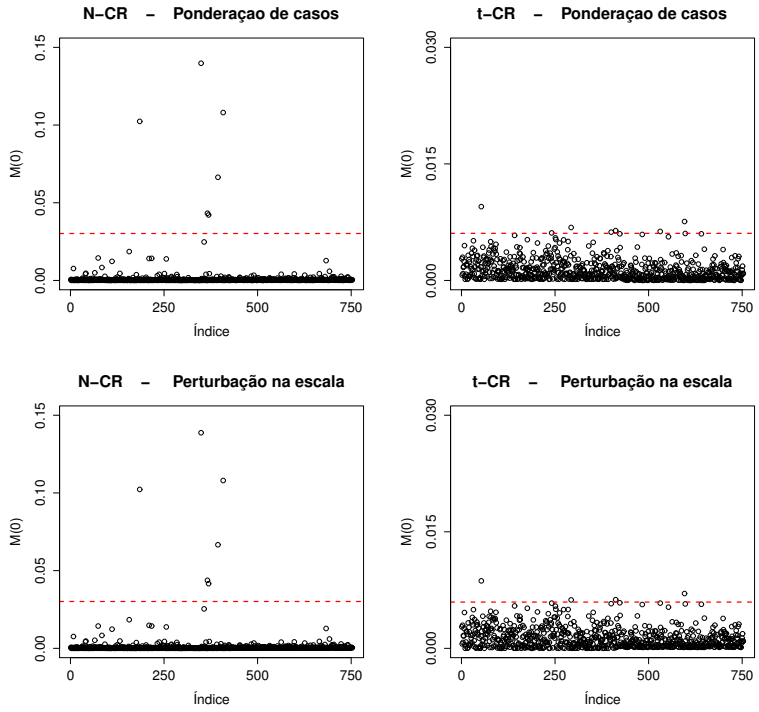


Figura 2.8: *Wage rate data.* Gráficos de $M(0)_i$, $i = 1, 2, \ldots, n$, sob os esquemas de perturbação: ponderação de casos (primeira linha) e perturbação na escala (segunda linha), para os modelos N-CR e t-CR.

Capítulo 3

Modelos para dados censurados sob a família de misturas de escala skew-normal

3.1 Introdução

Conforme discutido no início do Capítulo 2, o modelo Tobit, frequentemente usado para modelar conjuntos de dados sujeitos à censura, pode não ser adequado se a variável resposta apresentar indícios de desvio da normalidade. Neste sentido, o capítulo anterior apresentou a t de Student generalizada como alternativa à normal, visando sanar problemas relacionados à existência de valores extremos nos dados, já que esta possui caudas mais pesadas do que a distribuição normal.

No entanto, quando os dados apresentam assimetria além de possíveis problemas relacionados a valores extremos, a distribuição t de Student pode ser tão inaqueda quanto a normal. Neste sentido, apresentaremos neste capítulo a classe de distribuições de mistura de escala skew-normal (SMSN) proposta por Branco & Dey (2001), que inclui distribuições como a skew-normal, skew-t, skew-slash e suas versões simétricas. Esta classe, ao incorporar parâmetros de forma e de assimetria à distribuição normal, consegue lidar ao mesmo tempo com assimetria e com valores extremos.

Neste capítulo será apresentada a classe SMSN e suas principais propriedades com o intuito de fazer um estudo de inferência Bayesiana para o modelo de regressão com respostas censuradas sob esta família de distribuições. Em seguida, faremos a aplicação deste modelo a dados reais e também desenvolveremos alguns estudos de simulação para comparar a performance dos diferentes modelos na presença de valores extremos e de assimetria, variando o nível de censura dos dados.

3.2 Distribuições de mistura de escala skew-normal (classe SMSN)

Antes de introduzir a classe SMSN, definiremos a distribuição skew-normal (a versão assimétrica da normal) conforme feito por Azzalini (1985).

Definição 3. *Uma variável aleatória X tem distribuição skew-normal com parâmetro de locação $\mu \in \mathbb{R}$, de escala $\sigma^2 > 0$ e de forma $\lambda \in \mathbb{R}$, denotada por $X \sim SN(\mu, \sigma^2, \lambda)$, se sua densidade é dada por:*

$$f(x) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\frac{\lambda(x-\mu)}{\sigma}\right), \qquad x \in \mathbb{R}, \tag{3.2.1}$$

aqui o parâmetro de forma λ controla a assimetria da distribuição, de forma que valores positivos de λ indicam assimetria à direita e valores negativos indicam assimetria à esquerda, além disso, o grau de assimetria da distribuição aumenta conforme aumenta-se o valor absoluto de λ. Se $\mu = 0$ e $\sigma^2 = 1$, dizemos que X tem distribuição skew-normal padrão com parâmetro de forma λ.

Na Figura 3.1 mostramos a densidade da skew-normal padrão para diferentes valores de λ.

Algumas das principais propriedades da distribuição skew-normal estão listadas na Proposição abaixo.

Proposição 3. *Seja $X \sim SN(\mu, \sigma^2, \lambda)$ e $Z \sim N(\mu, \sigma^2)$ então:*

I) Se $\lambda = 0$, então $X \overset{\mathrm{d}}{=} Z$

II) Conforme $\lambda \to \infty$, a distribuição de X tende a uma $TN(\mu, \sigma^2, [\mu, \infty))$. Se $\lambda \to -\infty$, X tende a uma $TN(\mu, \sigma^2, (-\infty, \mu])$.

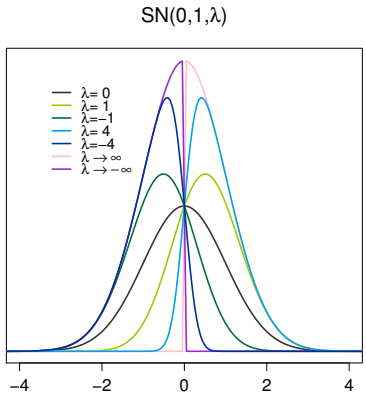


Figura 3.1: Densidade de skew-normal para $\mu = 0$, $\sigma = 1$ e valores variados para λ.

III) $\frac{X-\mu}{\sigma} \sim SN(0, 1, \lambda)$

IV) $-X \sim SN(-\mu, \sigma^2, -\lambda)$

V) A densidade de X *é log-côncava, e portanto, é unimodal (vista como função de* x*)*

VI) $1 - F(x; \mu, \sigma_2, \lambda) = F(x; -\mu, \sigma^2, -\lambda)$,

VII) $|X| \stackrel{\mathrm{d}}{=} |Z|$

VIII) Se $X \sim SN(0, 1, \lambda)$ *então* $X^2 \sim \chi^2_1$.

Prova:

Não desenvolveremos a prova completa destas propriedades, porém faremos um esboço. As propriedades de $I)$ a $IV)$ seguem diretamente da densidade em (3.2.1), sendo que para as propriedades $III)$ e $IV)$ é simples utilizar o método do Jacobiano para encontrar a distribuição de $\frac{X-\mu}{\sigma}$ e de $-X$. A propriedade $V)$ é provada utilizando os seguintes resultados:

- O produto de duas funções log-côncavas é uma função log-côncava.
- Se uma densidade é log-côncava, então a respectiva função de distribuição acumulada também é log-côncava (veja a prova em Bagnoli & Bergstrom (2005)).

- Uma função não negativa $g(\cdot)$ duas vezes difenciável e com domínio em um intervalo contínuo é log-côncava se e somente se:

$$g(x)g^{''}(x) \leq g^{'}(x)^2, \quad \forall x \ \ t.q. \ \ g(x) > 0, \tag{3.2.2}$$

onde $g^{'}(x)$ e $g^{''}(x)$ denotam respectivamente a primeira e segunda derivada de $g(x)$ em relação a x. Desta forma, para provar o resultado basta-nos provar que a densidade da normal com locação μ e escala σ^2 é log-côncava, o que é feito utilizando o terceiro resultado acima. As propriedades de $VI)$ a $IX)$ são provadas em Azzalini (1985).

■

Uma discussão mais detalhada das propriedades da distribuição skew-normal podem ser vistas em Bayes (2005) e Basso (2009).

Proposição 4. *Seja* $X \sim SN(\mu, \sigma^2, \lambda)$ *e* $T_0, T_1 \overset{iid}{\sim} N(0,1)$. *Então* X *admite a seguinte representação:*

$$X = \mu + \sigma\left(\delta|T_0| + (1-\delta^2)^{1/2}T_1\right), \tag{3.2.3}$$

onde $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

Prova:

Seja Y como em (3.2.3) e tome $\mu = 0$ e $\sigma^2 = 1$. Note que $Y|\ |T_0| = t \sim N(\delta t, (1-\delta^2))$ com $|T_0| \sim TN(0,1;[0,\infty))$. Então, pelo Lema A.2 apresentado em Basso (2009), temos que:

$$\begin{aligned} f(y) &= \int_0^\infty \phi(x; \delta t, (1-\delta^2)) 2\phi(t) dt \\ &= 2\int_0^\infty \phi(x;0,1)\phi(t; \delta x, (1-\delta^2)) dt \\ &= 2\phi(x;0,1)\Phi(\frac{\delta x}{\sqrt{1-\delta^2}}), \end{aligned}$$

isto é, $Y \sim SN(0,1,\lambda)$ com $\lambda = \frac{\delta}{\sqrt{1-\delta^2}}$. Agora, para μ e σ^2 geral, basta usar a propriedade $III)$ da Proposição 3 e concluir que $X = \mu + \sigma Y \sim SN(\mu, \sigma^2, \lambda)$.

■

Proposição 5. *A função geradora de momentos da variável aleatória* $Y \sim SN(\mu, \sigma^2, \lambda)$ *é dada por:*

$$M_Y(t) = 2e^{\mu t + \frac{t^2\sigma^2}{2}}\Phi(\delta\sigma t).$$

Prova:

Provemos primeiramente o resultado para a skew-normal padrão. Assim, considere $Z \sim SN(0, 1, \lambda)$. Utilizando o Lema A.1 apresentado em Basso (2009), com $\mathbf{a} = \lambda t, \mathbf{B} = \lambda, \boldsymbol{\mu} = \boldsymbol{\eta} = 0$ e $\boldsymbol{\Sigma} = \boldsymbol{\Omega} = 1$

$$\begin{aligned}
M_Z(t) &= E_Z[e^{Zt}] = 2\int_{-\infty}^{\infty} e^{zt}\phi(z)\Phi(\lambda z)dz \\
&= 2\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2-2zt)}\Phi(\lambda z)dz \\
&= 2e^{t^2/2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}\Phi(\lambda z)dz \\
&= 2e^{t^2/2}\int_{-\infty}^{\infty} \phi(x)\Phi(\lambda x + \lambda t)dx, \quad x = z - t, dx = dz \\
&= 2e^{t^2/2}E[\Phi(\lambda x + \lambda t)] \\
&= 2e^{t^2/2}E_X[\Phi(\lambda X + \lambda t)] \\
&= 2e^{t^2/2}\Phi(\frac{\lambda t - 0}{\sqrt{1+\lambda^2}}) \\
&= 2e^{t^2/2}\Phi(\delta t), \quad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}.
\end{aligned}$$

Agora, dado que $Y = \mu + \sigma Z$, temos que a função geradora de Y é dada por:

$$\begin{aligned}
M_Y(t) &= E_Z[e^{(\mu+\sigma Z)t}] = E_Z[e^{\mu t}e^{\sigma t Z}] = e^{\mu t}M_Z(\sigma t) \\
&= 2e^{\mu t + \frac{t^2\sigma^2}{2}}\Phi(\delta\sigma t).
\end{aligned}$$

■

Corolário 1. *Se* $Y \sim SN(\mu, \sigma^2, \lambda)$*, então:*

$$E[Y] = \mu + \sigma\delta\sqrt{\frac{2}{\pi}}, \quad Var[Y] = \sigma^2\left(1 - \frac{2}{\pi}\delta^2\right).$$

Uma vez definida a skew-normal, podemos agora definir a classe de distribuições de mistura de escala skew-normal, objetivo principal desta Seção.

Definição 4. *Dizemos que uma variável aleatória Y possui densidade pertencente à família de distribuições de mistura de escala skew-normal se ela pode ser escrita da seguinte forma:*

$$Y = \mu + \kappa(U)^{1/2} Z, \quad U \perp Z, \tag{3.2.4}$$

em que μ é um parâmetro de locação, $Z \sim SN(0, \sigma^2, \lambda)$, $\kappa(\cdot)$ é uma função positiva e U é uma variável aleatória com função de distribuição $H(\cdot\,; \boldsymbol{\nu})$ e densidade $h(\cdot\,; \boldsymbol{\nu})$ e $\boldsymbol{\nu}$ é um escalar ou vetor de parâmetros indexando a distribuição de U, o qual pode ser conhecido ou desconhecido (Lange et al.*, 1989). Enquanto o parâmetro λ está relacionado com o grau de assimetria da distribuição, $\boldsymbol{\nu}$ controla a curtose. Note que, sob a restrição $\lambda = 0$, a família SMSN é reduzida à família SMN. Denotamos $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$.*

Existe uma segunda representação estocástica para $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$ baseada na anterior e na Proposição 4. Esta representação, dada a seguir, é muito conveniente para derivar algumas propriedades desta classe de distribuições, assim como para desenvolver um estudo de inferência Bayesiana para os modelos de regressão com erros SMSN, como feito por exemplo em Basso (2009) no contexto de misturas de distribuições SMSN e em Cancho *et al.* (2011), no contexto de modelos de regressão não lineares.

$$Y = \mu + \Delta T + \kappa(U)^{1/2} \tau^{1/2} T_1, \tag{3.2.5}$$

onde $\Delta = \sigma\delta$, $\tau = (1 - \delta^2)\sigma^2$, $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$, $T = \kappa(U)^{1/2}|T_0|$, T_0, $T_1 \overset{\text{iid}}{\sim} N(0,1)$ e $|\cdot|$ denota valor absoluto.

A variável aleatória U pode ser discreta ou contínua e sua distribuição determina diretamente a distribuição de Y. Neste capítulo focaremos em três membros da classe de distribuições SMSN: a skew-normal, denotada por $SN(\mu, \sigma^2, \lambda)$, a skew-t, denotada por $St(\mu, \sigma^2, \lambda, \nu)$ e a skew-slash, denotada por $SSL(\mu, \sigma^2, \lambda, \nu)$. Existem outros exemplos de distribuições pertencentes a esta família, tal como a skew-normal-contaminada, a skew-Cauchy, a skew-Perason VII e todas as versões simétricas das mesmas. Usando a representação dada na Definição 4 é fácil ver que

$Y|U = u \sim SN(\mu, \kappa(u)\sigma^2, \lambda)$, portanto, integrando a densidade conjunta de (Y, U) em relação a U obtemos a densidade marginal de Y, dada por:

$$f(y) = 2\int_0^\infty \phi(y; \mu, \kappa(u)\sigma^2)\Phi\left(\frac{\lambda(y-\mu)}{\sigma\kappa(u)^{1/2}}\right) dH(u), \tag{3.2.6}$$

e considerando a representação estocástica dada em (3.2.5), temos que:

$$\begin{aligned} Y|T = t, U = u &\sim N(\mu + \Delta t\ ,\ \kappa(u)\tau), \\ T|U = u &\sim TN(0, \kappa(u)\ ;\ [0, \infty)). \end{aligned} \tag{3.2.7}$$

Desta forma, uma outra maneira de escrever a densidade da variável aleatória Y é:

$$\begin{aligned} f(y) &= \int\int f(y|t,u)f(t|u)f(u)\ dt\ du \\ &= 2\int_0^\infty \int_0^\infty \phi(y\ ;\ \mu + \Delta t\ ,\ \kappa(u)\tau)\ \phi(t\ ;\ 0, \kappa(u))\ dt\ dH(u). \end{aligned} \tag{3.2.8}$$

A seguinte proposição apresenta o formato das cdf da família de distribuições SMSN.

Proposição 6. *Seja $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$. Então, a cdf da variável aleatória Y pode ser escrita de duas formas distintas:*

$$a)\quad F(y) = 2\int_0^\infty \int_0^\infty \phi(t\ ;\ 0, 1)\Phi_1\Big(\frac{y-\mu}{\sigma\kappa(u)^{1/2}}\sqrt{1+\lambda^2} - \lambda t\ ;\ 0, 1\Big)\ dt\ dH(u), \tag{3.2.9}$$

$$b)\quad F(y) = 2\int_0^\infty \Phi_2\Big(\boldsymbol{y}(u)^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}\Big)\ dH(u), \tag{3.2.10}$$

onde $\boldsymbol{y}(u)^* = (\kappa(u)^{-1/2}y, 0)^\top$, $\boldsymbol{\mu}^* = (\mu, 0)^\top$, $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & -\delta\sigma \\ -\delta\sigma & 1 \end{pmatrix}$.

Prova:

Seja $Y \sim SMSN(\mu, \sigma^2, \lambda\,;\, H)$. Dada a pdf de Y na Equação (3.2.8), temos que:

$$
\begin{aligned}
F(y) &= \int_{-\infty}^{y} f(z)dz \\
&= \int_{-\infty}^{y}\int_0^\infty\int_0^\infty 2\phi(z\,;\,\mu+\Delta t, \kappa(u)\tau)\phi(t\,;\,0,\kappa(u))\,dt\,dH(u)\,dz \\
&= \int_0^\infty\int_0^\infty 2\left[\int_{-\infty}^{y}\phi(z\,;\,\mu+\Delta t,\kappa(u)\tau)\,dz\right]\phi(t\,;\,0,\kappa(u))\,dt\,dH(u) \qquad (3.2.11)\\
&= \int_0^\infty\int_0^\infty 2\left[\int_{-\infty}^{y}\phi(z\,;\,\mu+\Delta\kappa(u)^{1/2}x,\kappa(u)\tau)\,dz\right]\phi(x\,;\,0,1)\,dx\,dH(u) \qquad (3.2.12)\\
&= 2\int_0^\infty\int_0^\infty \Phi\left(\frac{y-\mu-\Delta\kappa(u)^{1/2}x}{\kappa(u)^{1/2}\tau^{1/2}}\,;\,0,1\right)\phi(x\,;\,0,1)\,dx\,dH(u) \\
&= 2\int_0^\infty\int_0^\infty \Phi\left(\frac{y-\mu}{\kappa(u)^{1/2}\sigma\sqrt{1-\delta^2}}-\frac{\delta}{\sqrt{1-\delta^2}}x\,;\,0,1\right)\phi(x\,;\,0,1)\,dx\,dH(u) \\
&= 2\int_0^\infty\int_0^\infty \Phi\left(\frac{(y-\mu)\sqrt{1+\lambda^2}}{\kappa(u)^{1/2}\sigma}-\lambda x\,;\,0,1\right)\phi(x\,;\,0,1)\,dx\,dH(u), \qquad (3.2.13)
\end{aligned}
$$

onde a Equação (3.2.13) é consequência das relações $\Delta = \sigma\delta$, $\tau = \sigma^2(1-\delta^2)$ e $\delta = \lambda/\sqrt{1+\lambda^2}$. Desta forma, está provada a parte (a) da Proposição 6. A parte (b) é obtida se continuarmos desenvolvendo a Equação (3.2.11) escrevendo a expressão completa para $\phi(\cdot)$:

$$
\begin{aligned}
F(y) &= 2\int_0^\infty\int_0^\infty\int_{-\infty}^{y}\frac{1}{2\pi\kappa(u)\tau^{1/2}}\times \\
&\quad \exp\left\{-\frac{1}{2\kappa(u)\tau}\left[(z-\mu)^2+(\Delta^2+\tau)t^2-2(z-\mu)\Delta t\right]\right\}\,dz\,dt\,dH(u) \\
&= 2\int_0^\infty\int_0^\infty\int_{-\infty}^{y}\frac{1}{2\pi\kappa(u)\sigma\sqrt{1-\delta^2}}\times \\
&\quad \exp\left\{-\frac{1}{2(1-\delta^2)}\left[\frac{(z-\mu)^2}{\kappa(u)\sigma^2}+\frac{t^2}{\kappa(u)}-\frac{2\delta}{\kappa(u)\sigma}(z-\mu)t\right]\right\}\,dz\,dt\,dH(u) \\
&= 2\int_0^\infty\int_0^\infty\int_{-\infty}^{y}\frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}}\times \\
&\quad \exp\left\{-\frac{1}{2}\begin{pmatrix} z-\mu \\ t\end{pmatrix}^\top\boldsymbol{\Sigma}^{-1}\begin{pmatrix} z-\mu \\ t\end{pmatrix}\right\}\,dz\,dt\,dH(u) \qquad (3.2.14)
\end{aligned}
$$

Note que, se considerarmos o vetor de variáveis aleatórias $(X, W) \sim N_2\left((\mu, 0)\,,\,\boldsymbol{\Sigma}\right)$, então a integração em z e t na Equação (3.2.14) representa $\mathbb{P}(X \leq y\,,\,W \geq 0)$, que, por simetria, é

equivalente a $\mathbb{P}(X \leq y\,,\, W \leq 0)$. Desta forma, podemos reescrever esta Equação da seguinte forma:

$$F(y) = 2\int_0^{\infty} \Phi_2\left(\boldsymbol{y}^*\,;\,\boldsymbol{\mu}^*, \boldsymbol{\Sigma}\right) dH(u),$$

onde $\boldsymbol{y}^*$ e $\boldsymbol{\mu}^*$ foram definidos anteriormente, provando a parte (b) da Proposição 6.

■

A partir deste momento consideraremos $\kappa(U) = U^{-1}$ na representação estocástica dada em (3.2.5). Esta escolha, feita também em Basso (2009), nos possibilitará desenvolver algumas propriedades interessantes da família de distribuições SMSN, como a função geradora de momentos, a esperança e a variância, apresentadas nas proposições seguintes.

Proposição 7. *Seja $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$, então a função geradora de momentos (fgm) de Y é dada por:*

$$M_Y(t) = 2\int_0^{\infty} e^{t\mu + \frac{1}{2}u^{-1}t^2\sigma^2}\Phi(u^{-1/2}\delta\sigma t)dH(u), \quad t \in \mathbb{R}. \tag{3.2.15}$$

Prova:

Da Definição 4 temos que $Y|U = u \sim SN(\mu, u^{-1}\sigma^2, \lambda)$. Considerando então propriedades da esperança condicional e a Proposição 5, temos que:

$$\begin{aligned} M_Y(t) &= E_Y[e^{Yt}] = E_U[E_{Y|U}[e^{tY}|U]] \\ &= E_U[2e^{t\mu + \frac{t^2\sigma^2}{2}}\Phi(\delta\sigma t)] \\ &= 2\int_0^{\infty} e^{t\mu + \frac{1}{2}u^{-1}t^2\sigma^2}\Phi(u^{-1/2}\delta\sigma t)dH(u). \end{aligned}$$

■

Proposição 8. *Seja $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$. Então,*

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

ii) Se $E[U^{-1}] < \infty$, $Var_Y[Y] = k_2\sigma^2 - \frac{2}{\pi}k_1^2\Delta^2$,

onde $k_m = E_Y[U^{-m/2}]$ para $m = 1, 2$, $\Delta = \sigma\delta$ e $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

Prova:

Da Definição 4, do Corolário 1 e da suposição de independência entre U e Z, temos que:

$$\begin{aligned}
E_Y[Y] &= \mu + E_{(Z,U)}[U^{-1/2}Z] \\
&= \mu + E_U[U^{-1/2}]E_Z[Z] \\
&= \mu + k_1\sqrt{\frac{2}{\pi}}\delta, \\
Var_Y[Y] &= Var_{(Z,U)}[U^{-1/2}Z] \\
&= E_{(Z,U)}[U^{-1}Z^2] - \left(E_{(Z,U)}[U^{-1/2}Z]\right)^2 \\
&= E_U[U^{-1}]E_Z[Z^2] - (k_1 E_Z[Z])^2 \\
&= k_2\sigma^2 - k_1^2\left(\frac{2}{\pi}\right)\sigma^2\delta^2.
\end{aligned}$$

■

Apresentaremos agora alguns casos particulares da classe SMSN, com os quais trabalharemos no restante deste capítulo. Para cada um dos casos desenvolveremos as respectivas pdf, cdf e a esperança $k_m = \mathbb{E}_U[U^{-m/2}]$, $m \in \mathbb{N}$, útil na implementação do amostrador de Gibbs para o modelo de regressão censurado sob esta família de distribuições.

- *Skew-normal:* Este caso é obtido quando U é uma variável aleatória degenerada em 1, isto é, quando $P(U = 1) = 1$, desta forma, $k_m = 1$. A densidade de $Y \sim SN(\mu, \sigma^2, \lambda)$ é definida em (3.2.1) e, usando a parte (b) da Proposição 6, sua cdf é dada por:

$$F(y) = 2\Phi_2\Big(\boldsymbol{y}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}\Big), \tag{3.2.16}$$

  onde $\boldsymbol{y}^* = (y, 0)^\top$, $\boldsymbol{\mu}^* = (\mu, 0)^\top$ e $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & -\delta\sigma \\ -\delta\sigma & 1 \end{pmatrix}$.

- *Skew-t:* A distribuição skew-t surge quando consideramos $U \sim G(\nu/2, \nu/2)$ na Definição 4, de forma que $k_m = (\nu/2)^{(m/2)}\Gamma(\frac{\nu-m}{2})\Gamma(\nu/2)^{-1}$. A densidade de $Y \sim St(\mu, \sigma^2, \lambda)$ é dada por:

$$f(y|\mu,\sigma^2,\lambda;\nu) = \frac{2\,\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{d(y)^2}{\nu}\right)^{-\frac{\nu+1}{2}} \mathcal{T}_1\left(\lambda\, d(y)\sqrt{\frac{\nu+1}{\nu+d(y)^2}}\,;\,0,1,\nu+1\right), \quad y\in\mathbb{R}, \tag{3.2.17}$$

onde $d(y) = (y-\mu)/\sigma$. A demonstração deste resultado encontra-se no Apêndice A. Um caso particular da distribuição skew-t é a skew–Cauchy, obtida quando $\nu = 1$. Além disso, quando $\nu \to \infty$, obtemos a skew-normal como caso limite.

Utilizando a parte (b) da Proposição 6 podemos simplificar a expressão da cdf da skew-t:

$$F(y) = 2\,\mathcal{T}_2\left(\boldsymbol{y}^*\,;\,\boldsymbol{\mu}^*,\boldsymbol{\Sigma},\nu\right), \tag{3.2.18}$$

onde $\boldsymbol{y}^*$, $\boldsymbol{\mu}^*$ e $\boldsymbol{\Sigma}$ são como definidos no caso da skew-normal. Este resultado também está demonstrado no Apêndice A. A distribuição skew-t tem como caso particular a skew-Cauchy, quando $\nu = 1$, e tem como caso limite a skew-normal, alcançada conforme $\nu \to \infty$. Aplicações da skew-t podem ser encontradas em Lin *et al.* (2007) e Azzalini & Genton (2008).

- *Skew-slash*: Aqui, consideramos $U \sim Beta(\nu, 1)$ na Definição 4, com $\nu > 0$, assim, $k_m = \frac{\nu}{\nu - m/2}$. A densidade de Y é dada por:

$$f(y|\mu,\sigma^2,\lambda;\nu) = 2\nu\int_0^1 u^{\nu-1}\phi(y;\mu,u^{-1}\sigma^2)\Phi(u^{1/2}\frac{\lambda(y-\mu)}{\sigma})\,du, \quad y\in\mathbb{R}. \tag{3.2.19}$$

A cdf da skew-slash não possui uma forma fechada, no entanto, utilizando a parte (b) da Proposição 6, podemos escrevê-la em termos de uma integral que pode ser aproximada por métodos numéricos:

$$F(y) = \int_0^\infty 2\nu\Phi_2\Big(\boldsymbol{y}(u)^*;\boldsymbol{\mu}^*,\boldsymbol{\Sigma}\Big)u^{\nu-1}\,du, \tag{3.2.20}$$

onde $\boldsymbol{y}(u)^*$, $\boldsymbol{\mu}^*$ e $\boldsymbol{\Sigma}$ são como na Proposição 6 (com $\kappa(u) = 1/u$). A skew-slash tem caudas mais pesadas do que a skew-normal, tendendo a esta distribuição quando $\nu \to \infty$. Aplicações desta distribuição podem ser encontradas em Wang & Genton (2006).

A Figura 3.2 compara as distribuições skew-normal, skew-t e skew-slash para diferentes valores de λ e ν. No gráfico da esquerda consideramos $\lambda = 2$ e $\nu = 3$, no gráfico do centro, $\lambda = 0$ e $\nu = 4$ e no gráfico da direita, $\lambda = -2$ e $\nu = 5$. Nestes gráficos podemos ver claramente que as distribuições

skew-t e skew-slash possuem caudas mais pesadas do que a skew-normal. É importante destacar que todos os três gráficos têm o eixo y com a mesma escala.

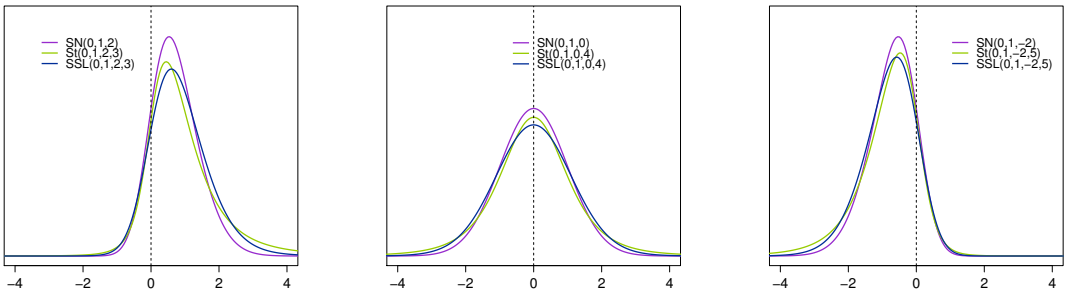


Figura 3.2: Densidades skew-normal, skew-t e skew-slash para diferentes valores de λ e ν.

Métodos para estimar os parâmetros das distribuições SMSN podem ser encontrados em Basso *et al.* (2010), a partir do método dos momentos, e em Garay (2009), a partir do algoritmo EM. Nestas duas referências a estimação é feita supondo ν um valor conhecido, porém é possível adaptar o algoritmo EM de Garay (2009) para estimar este parâmetro: em cada iteração, após calcular o valor das estimativas de μ, σ^2 e λ, a estimativa de ν é escolhida como o argumento que maximiza a função de log-verossimilhança vista somente como função de ν, substituindo os outros parâmetros por suas estimativas atuais. Esta versão do algoritmo EM é conhecida como ECM ("*expectation conditional maximization*").

3.3 Definição e inferência Bayesiana para os modelos SMSN-CR

O modelo de regressão para dados censurados sob a classe de distribuições de mistura de escala skew-normal, denotado por SMSN-CR, é definido como nas Equações (2.3.1) e (2.3.2), porém fazendo-se a suposição de que $\epsilon_i \overset{\text{iid}}{\sim} SMSN(\eta\Delta, \sigma^2, \lambda; H)$ para $i = 1, 2, \ldots, n$ na Equação (2.3.1), onde $\eta = -\sqrt{\frac{2}{\pi}}k_1$. O parâmetro de locação dos erros aleatórios é diferente de 0 e foi escolhido com base na Proposição 8, uma vez que, com esta adaptação, $Y_i \sim SMSN(\boldsymbol{x}_i^\top\boldsymbol{\beta} + \eta\Delta, \sigma^2, \lambda; H)$ e

$\mathbb{E}_Y[Y] = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

Nesta Seção iremos fazer inferência Bayesiana para o modelo de regressão com respostas censuradas sob três distribuições da família SMSN: a skew-normal, a skew-t e a skew-slash. Para isto, seguindo a sugestão de Cancho *et al.* (2011), consideraremos uma reparametrização da classe SMSN baseada na representação 3.2.5, o que simplificará a construção do algoritmo mencionado.

Seja então $\boldsymbol{\omega} = (\boldsymbol{\beta}^\top, \Delta, \tau, \nu)^\top$ o vetor de parâmetros com o qual trabalharemos. É importante destacar que existe uma correspondência um a um entre este vetor e o vetor de parâmetros original $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \sigma^2, \lambda, \nu\right)^\top$, uma vez que $\Delta = \sigma\sqrt{\dfrac{\lambda}{\lambda^2+1}} \in \mathbb{R}$ e $\tau = \dfrac{\sigma^2}{\lambda^2+1} > 0$, sendo possível obter σ^2 e λ a partir de Δ e τ fazendo-se $\sigma^2 = \tau + \Delta^2$ e $\lambda = \Delta/\sqrt{\tau}$. Desta forma, as amostras a posteriori de $\boldsymbol{\omega}$ podem ser utilizadas para fazer inferência sobre $\boldsymbol{\theta}$, bastando para isso aplicar a transformação anteriormente mencionada.

Seguindo novamente o trabalho de Cancho *et al.* (2011), assumiremos à priori que $\boldsymbol{\beta} \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \Sigma_0)$, $\Delta \sim \mathrm{N}(\mu_\Delta, \sigma_\Delta^2)$ e $\tau \sim IG(a_\tau, b_\tau)$, onde os hiperparâmetros fixos e conhecidos são $\boldsymbol{\mu}_0$, um vetor $p \times 1$, Σ_0, uma matriz $p \times p$ definida positiva, $\mu_\Delta \in \mathbb{R}$, $\sigma_\Delta^2 > 0$, $a_\tau > 0$ e $b_\tau > 0$. Estas escolhas de prioris são feitas para garantir conjugação.

Para os modelos de regressão censurados skew-t (St-CR) e skew-slash (SSL-CR) precisamos especificar também a distribuição a priori de ν, o parâmetro que indexa a distribuição de U (veja a representação dada na Definição 4). Para isso utilizaremos a sugestão dada em Cabral & Madruga (2012): $\nu \sim \mathrm{Texp}(\gamma\,;\,\mathbb{A})$ e $\gamma \sim \mathrm{Unif}(c, d)$, onde c e d são hiperparâmetros conhecidos. Aqui, $\mathrm{Texp}(\gamma\,;\,\mathbb{A})$ denota a distribuição exponencial com parâmetro de escala $1/\gamma > 0$ truncada no intervalo $\mathbb{A}$ e $\mathrm{Unif}(c, d)$ denota a distribuição uniforme no intervalo (c, d). Para garantir a existência dos primeiro e segundo momentos da distribuição assumida para os erros da regressão, neste trabalho tomamos $\mathbb{A} = (2, \infty)$ e $(c, d) = (0.02, 0.49)$ para o modelo St-CR e $\mathbb{A} = (1, \infty)$ e $(c, d) = (0.02, 0.9)$ para o modelo SSL-CR. Assumimos também independência à priori entre os parâmetros, portanto a distribuição à priori do vetor $\boldsymbol{\omega}$ é:

$$f(\boldsymbol{\omega}) = f(\boldsymbol{\beta})\, f(\Delta)\, f(\tau)\, f(\nu)\,. \tag{3.3.1}$$

Embora a hipótese de independência possa não ser realista para alguns conjuntos de parâmetros, ela leva a propriedades interessantes para as distribuições à posteriori, como a conjugação, e facilita o desenvolvimento de um algoritmo para amostrar destas distribuições, além disso, se esta hipótese realmente não for verdadeira, ela será corrigida à posteriori e não irá prejudicar o processo de inferência.

3.3.1 Construção do amostrador de Gibbs

No contexto Bayesiano, estimativas pontuais são obtidas como características associadas à distribuição à posteriori, como a esperança ou a moda. Dada a forma matemática complexa destas quantidades é bastante complicado aproximá-las através de técnicas como a integração numérica. Portanto, utilizaremos o amostrador de Gibbs para gerar uma amostra da distribuição à posteriori do vetor de parâmetros e fazer estimativas pontuais baseando-nos nesta amostra. Para desenvolver este algoritmo faremos uso do "aumento de dados", isto é, vamos supor que o vetor de variáveis sujeitas à censura $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ e os de variáveis latentes $\boldsymbol{U} = (U_1, U_2, \ldots, U_n)^\top$ e $\boldsymbol{T} = (T_1, T_2, \ldots, T_n)$ (veja a representação em (3.2.5)), poderiam ser completamente observados e então calcular as distribuições condicionais completas para cada parâmetro do modelo e para cada variável latente.

A representação estocástica da classe SMSN de distribuições dada em (3.2.5) nos permite escrever:

$$\begin{aligned} Y_i|U_i = u_i, T_i = t_i &\sim \mathrm{N}(\mathbf{x}_i^\top \boldsymbol{\beta} + \Delta t_i, u_i^{-1}\tau), \\ T_i|U_i = u_i &\sim \mathrm{TN}(\eta, u_i^{-1}\,;\,(\eta, \infty)), \\ U_i &\sim H(\cdot|\boldsymbol{\nu}), \end{aligned}$$

para $i = 1, 2, \ldots, n$. Considere $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$ o vetor de observações de V_i, $i = 1, 2, \ldots, n$. Seja $\vartheta_{(0)}$ um valor inicial para ϑ e $\vartheta_{(m)}$ o valor de ϑ na iteração m do algoritmo. Calculando-se as condicionais completas, a m-ésima iteração do amostrador de Gibbs sob as distribuições skew-normal, skew-t e skew-slash é da seguinte forma:

Passo 1: Para $i = 1, 2, \ldots, n$, se $i-$ésima observação não for censurada, faça $y_{i(m)} = v_i$, caso contrário gere $y_{i(m)}$ (independentemente) de $f(y_i \mid v_i, t_i, u_i, \boldsymbol{\beta}, \Delta, \tau)$, que é uma distribuição normal truncada:

$$\mathrm{TN}(\boldsymbol{x}_i^\top \boldsymbol{\beta}_{(m-1)} + \Delta_{(m-1)} t_{i(m-1)}, \tau_{(m-1)}/u_{i(m-1)} \; ; \; \mathbb{A}),$$

onde $\mathbb{A} = [v_i, \infty)$ se a variável resposta for censurada à direita ou $\mathbb{A} = (\infty, v_i]$, se for censurada à esquerda.

Passo 2: Para $i = 1, 2, \ldots, n$, gere $t_{i(m)}$ independentemente de $f(t_i \mid v_i, y_i, u_i, \boldsymbol{\beta}, \Delta, \tau)$, que é

$$\mathrm{TN}(\mu_{t_i(m-1)}, \sigma_{t_i(m-1)} \; ; \; [\eta, \infty)),$$

onde $\mu_{t_i(m-1)} = \frac{\Delta_{(m-1)}}{\Delta^2_{(m-1)} + \tau_{(m-1)}} \left(y_{i(m)} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{(m-1)} + \frac{\eta \tau_{(m-1)}}{\Delta_{(m-1)}} \right)$ e $\sigma^2_{t_i(m-1)} = \frac{\tau_{(m-1)}}{u_{i(m-1)}(\Delta^2_{(m-1)} + \tau_{(m-1)})}$.

Passo 3: Gere $\boldsymbol{\beta}_{(m)}$ de $f(\boldsymbol{\beta} \mid \boldsymbol{v}, \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u}, \Delta, \tau, \boldsymbol{\nu})$, que é $\mathrm{N}_p \left(\boldsymbol{\mu}^*_(m-1), \Sigma^{*(m-1)} \right)$, onde

$$\begin{aligned}
\boldsymbol{\mu}^*_{(m-1)} &= \Sigma^*_{(m-1)} \left(\Sigma_0^{-1} \boldsymbol{\mu}_0 + \frac{\boldsymbol{X}^{*\top}_{(m-1)} \boldsymbol{y}^*_{(m-1)}}{\tau_{(m-1)}} - \frac{\Delta_{(m-1)} \boldsymbol{X}^{*\top}_{(m-1)} \boldsymbol{t}^*_{(m-1)}}{\tau_{(m-1)}} \right), \\
\Sigma^*_{(m-1)} &= \left(\frac{\boldsymbol{X}^{*\top}_{(m-1)} \boldsymbol{X}^*_{(m-1)}}{\tau_{(m-1)}} + \Sigma_0^{-1} \right)^{-1},
\end{aligned}$$

$\boldsymbol{t}^*_{(m-1)}$ é o vetor composto por $t^*_{i(m-1)} = \sqrt{u_{i(m-1)}} t_{i(m)}$, $i = 1, 2, \ldots, n$, $\boldsymbol{y}^*_{(m-1)}$ é o vetor com elementos $y^*_{i(m-1)} = \sqrt{u_{i(m-1)}} y_{i(m)}$, $i = 1, 2, \ldots, n$ e $\boldsymbol{X}^*_{(m-1)}$ é a matriz composta por linhas dadas pelos vetores $\boldsymbol{x}^*_{i(m-1)} = (\sqrt{u_{i(m-1)}} x_{i1}, \ldots, \sqrt{u_{i(m-1)}} x_{ip})^\top$ para $i = 1, 2, \ldots, n$.

Passo 4: Gere $\Delta_{(m)}$ de $f(\Delta \mid \boldsymbol{v}, \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{\beta}, \tau, \boldsymbol{\nu})$, que é $\mathrm{N} \left(\mu^*_{\Delta(m-1)}, \sigma^{2*}_{\Delta(m-1)} \right)$ com

$$\begin{aligned}
\mu^*_{\Delta(m-1)} &= \sigma^{2*}_{\Delta(m-1)} \left(\frac{\mu_\Delta}{\sigma^2_\Delta} + \frac{1}{\tau_{(m-1)}} \sum_{i=1}^n u_{i(m-1)} t_{i(m)} (y_{i(m)} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{(m)}) \right), \\
\sigma^{2*}_{\Delta(m-1)} &= \left(\frac{1}{\tau_{(m-1)}} \sum_{i=1}^n u_{i(m-1)} t^2_{i(m)} + \frac{1}{\sigma^2_\Delta} \right)^{-1}.
\end{aligned}$$

Passo 5: Gere $\tau_{(m)}$ de $f(\tau \mid \boldsymbol{v}, \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{\beta}, \Delta, \boldsymbol{\nu})$, que é uma gama inversa:

$$IG \left(a_\tau + \frac{n}{2} \, , \; b_\tau + \frac{1}{2} \sum_{i=1}^n u_{i(m-1)} (y_{i(m)} - \mathbf{X}_i^\top \boldsymbol{\beta}_{(m)} - \Delta_{(m)} t_{i(m)})^2 \right).$$

Passo 6: Para o caso skew-normal, faça $u_{i(m)} = 1$, $i = 1, 2, \ldots, n$. Para a skew-t e skew-slash gere $u_{i(m)}$ para $i = 1, 2, \ldots, n$ (independentemente) de $f(u_i \mid v_i, y_i, t_i, \boldsymbol{\beta}, \Delta, \tau, \boldsymbol{\nu})$, que é:

(a) sob a distribuição skew-t,

$$\mathrm{G}\left(\frac{\nu_{(m-1)}}{2} + 1, \frac{\nu_{(m-1)} + A_{i(m)}}{2}\right),$$

onde $A_{i(m)} = \left(y_{i(m)} - \mathbf{X}_i^\top \boldsymbol{\beta}_{(m)} - \Delta_{(m)} t_{i(m)}\right)^2 / \tau_{(m)} + (t_{i(m)} - \eta)^2$.

(b) sob a distribuição skew-slash,

$$\mathrm{TG}\left(\nu_{(m-1)} + 1, \frac{A_{i(m)}}{2}\ ;\ [0, 1]\right),$$

uma distribuição gama truncada em $[0, 1]$.

Passo 7: Para os casos skew-slash ou skew-t, precisamos ainda gerar $\nu_{(m)}$ e $\gamma_{(m)}$, através do seguinte procedimento:

(a) sob a distribuição skew-t,

i. Gere $\gamma_{(m)}$ de $f(\gamma|\nu)$, que é $\mathrm{TG}(2, \nu_{(m-1)}; [0.02, 0.49])$.

ii. Utilizando um passo de Metropolis-Hastings, gere $\nu_{(m)}$ de:

$$f(\nu \mid \boldsymbol{v}, \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{\beta}, \Delta, \tau, \gamma) \propto \left(\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\right)^n \exp\left\{-\nu\left(\frac{1}{2}\sum_{i=1}^n u_{i(m)} + \gamma_{(m)}\right)\right\} \prod_{i=1}^n u_{i(m)}^{\frac{\nu}{2}-1} \mathbb{1}_{(2,\infty)}(\nu). \quad (3.3.2)$$

O passo de Metropolis-Hastings é da seguinte forma: dada a observação $\nu_{(m-1)}$ obtida na iteração $m-1$ do amostrado de Gibbs, gere um candidato ν^* da distribuição candidata $g(\nu)$, uma distribuição normal truncada:

$$g(\nu) \equiv \mathrm{TN}\left(\omega_{\nu_{(m-1)}}, \varsigma_{\nu_{(m-1)}}; (2; \infty)\right),$$

onde o parâmetro de locação desta distribuição candidata é dado por $\omega_{\nu_{(m-1)}} = \nu_{(m-1)} - \frac{q_1(\nu_{(m-1)})}{q_2(\nu_{(m-1)})}$ e o de escala por $\varsigma_{\nu_{(m-1)}} = -\frac{1}{q_2(\nu_{(m-1)})}$, onde:

$$\begin{aligned} q_1(\nu) &= \frac{d}{d\nu} \log f(\nu \mid \boldsymbol{y}_{(m)}, \boldsymbol{t}_{(m)}, \boldsymbol{u}_{(m)}, \boldsymbol{\beta}_{(m)}, \Delta_{(m)}, \tau_{(m)}, \gamma_{(m)}), \\ q_2(\nu) &= \frac{d^2}{d\nu^2} \log f(\nu \mid \boldsymbol{y}_{(m)}, \boldsymbol{t}_{(m)}, \boldsymbol{u}_{(m)}, \boldsymbol{\beta}_{(m)}, \Delta_{(m)}, \tau_{(m)}, \gamma_{(m)}), \end{aligned}$$

respectivamente a primeira e segunda derivadas da condicional completa de ν. As escolhas dos parâmetros de locação e escala da distribuição candidata foram baseadas no trabalho de Abanto-Valle *et al.* (2013) e o truncamento foi feito no intervalo $(2, \infty)$ para garantir a existência dos primeiros dois momentos da t de Student utilizando o candidato gerado como graus de liberdade. Assim, $\nu_{(m)}$ é tomado como o valor candidato gerado ν^* com probabilidade:

$$\alpha(\nu_{(m-1)}) = \min\left\{\frac{f(\nu^*)g(\nu_{(m-1)})}{f(\nu_{(m-1)})g(\nu^*)}\ ;\ 1\right\},$$

ou então $\nu_{(m)} = \nu_{(m-1)}$ com probabilidade $1 - \alpha(\nu_{(m-1)})$, onde a função $f(\cdot)$ é a mesma que em (3.3.2).

(b) sob a distribuição skew-slash,

i. Gere $\gamma_{(m)}$ de $f(\gamma|\nu)$, que é uma $\text{TG}(2, \nu_{(m-1)}; [0.02, 0.9])$.

ii. Gere $\nu_{(m)}$ de $f(\nu \mid \boldsymbol{v}, \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{\beta}, \Delta, \tau, \lambda)$, que é:

$$\text{TG}\left(n+1, \gamma_{(m)} - \sum_{i=1}^{n} \log\left(u_{i(m)}\right)\ ;\ (1, \infty)\right).$$

3.4 Estudo de simulação III: perfomance dos modelos assimétricos sob perturbações

O objetivo deste estudo de simulação é comparar a performance dos modelos SN-CR, St-CR e SSL-CR na presença de observações atípicas no conjunto de dados (em relação à variável resposta). Para isso, geramos observações de um modelo de regressão para dados censurados skew-normal, conforme as Equações (2.3.1) e (2.3.2), supondo que $n = 100$, $\epsilon_i \overset{\text{iid}}{\sim} SN(\eta\Delta, \sigma^2, \lambda)$, com $\eta = -\sqrt{\frac{2}{\pi}}$, $\Delta = \sigma\sqrt{\frac{\lambda}{1+\lambda^2}}$, $\sigma^2 = 2$ e $\lambda = -4$ e que $\boldsymbol{x}_i^\top = (1, x_i)$, onde x_i foram gerados de forma independente a partir de uma uniforme em $(1, 3)$, para $i = 1, 2, \ldots, n$; supomos também que $\boldsymbol{\beta}^\top = (\beta_1, \beta_2) = (10, 15)$. Após gerado, o conjunto de dados teve sua variável resposta censurada à esquerda a um nível de 10% do total de observações. Perturbamos então as observações #3 ($y_3 = 43.22178$), #66 ($y_{66} = 51.17056$) e #92 ($y_{92} = 31.82169$), escolhidas aleatoriamente

dentre as não censuradas, aumentando aos valores de y em $\Lambda\%$ de seus valores originais, para $\Lambda = 10, 20, 30, \ldots, 150$. Desta forma, se y denota o valor original da variável resposta, seu valor perturbado y^* é dado por:

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

Para cada um dos 15 padrões de perturbação e para o conjunto de dados não perturbado, foram ajustados os modelos SN-CR, St-CR e SSL-CR, computando em cada ajuste os valores dos critérios LPML, DIC, EAIC, EBIC e WAIC_2 (veja Seção 1.6.2) e também a mudança relativa na estimação pontual de β_j, $j = 1, 2$, dada por:

$$RC(\beta_j)_k = \left| \frac{\hat{\beta}_{j0} - \hat{\beta}_{jk}}{\hat{\beta}_{j0}} \right|, \qquad j = 1, 2 \quad k = 1, 2, \ldots, 15,$$

onde $\hat{\beta}_{j0}$ representa a estimativa pontual de β_j utilizando o conjunto de dados não perturbado e $\hat{\beta}_{jk}$, utilizando o conjunto de dados com o $k-$ésimo padrão de perturbação.

A Figura 3.3 mostra os resultados em relação às mudanças relativas. Para β_1 vemos que para perturbações menores do que 100% as mudanças relativas não são muito significativas (menores do que 5%) e não obedecem um padrão, porém conforme Λ torna-se maior do que 100 vemos que a mudança relativa para β_1 sob os modelos St-CR e SSL-CR parecem se estabilizar perto dos 5%, enquanto que sob o modelo SN-CR ela apresenta um padrão crescente, atingindo 10% para $\Lambda = 150$. Em relação a β_2, pode-se observar que para pequenas perturbações ($\Lambda \in \{10, 20, 30\}$) os três modelos se comportam de formas muito parecidas, porém conforme Λ aumenta o modelo SN-CR perde performance em relação aos outros dois modelos, que se comportam de forma muito parecida entre si. Isto confirma que o modelo SN-CR é menos robusto do que o St-CR e o SSL-CR para lidar com observações atípicas.

O cenário descrito acima é confirmado nos critérios mostrados na Figura 3.4, onde o modelo SN-CR se mostra tão bom quanto os outros (ou até preferível) para pequenas perturbações, o que é esperado já que os dados são gerados de um modelo skew-normal, porém conforme aumentamos as perturbações os modelos St-CR e SSL-CR mostram-se superiores ao N-CR (e com performances bastante parecidas entre si).

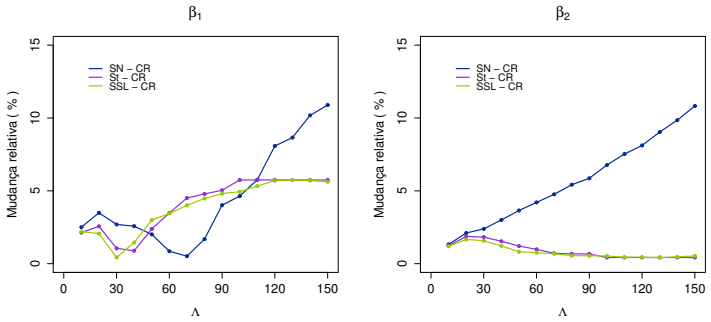


Figura 3.3: Estudo de simulação III. Mudança relativa (em %) para β_1 e β_2 para os modelos SN-CR, SSL-CR e St-CR, sob diferentes níveis de perturbação Λ.

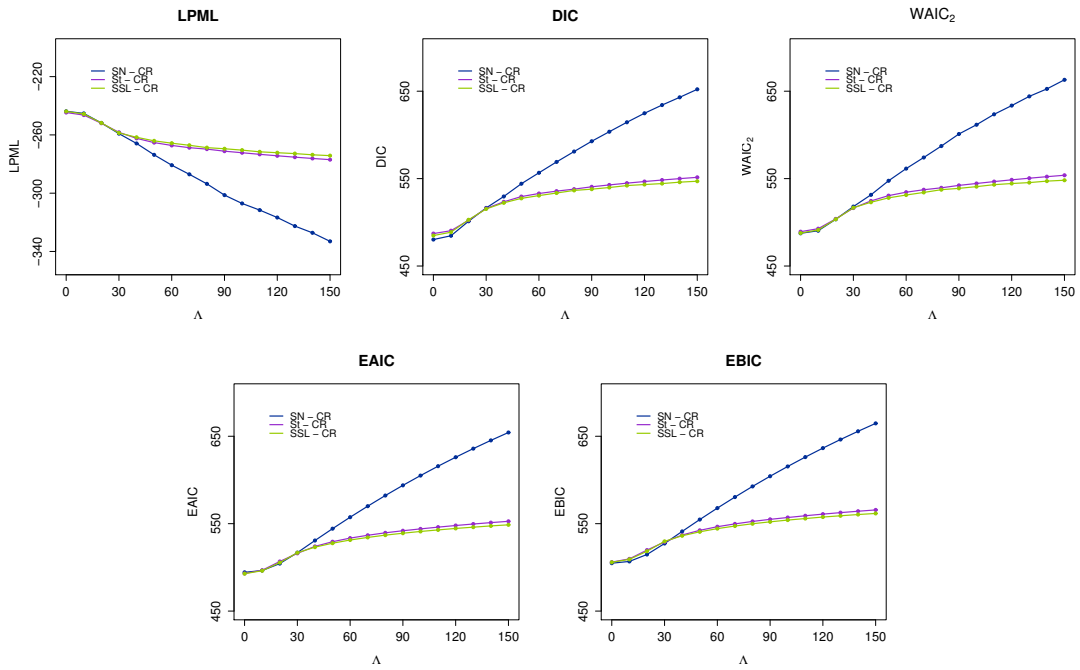


Figura 3.4: Estudo de simulação III. Critérios LPML, DIC, WAIC, EAIC e EBIC para os modelos SN-CR, SSL-CR e St-CR, sob diferentes níveis de perturbação Λ.

Com estes resultados confirmamos a hipótese de que os modelos St-CR e SSL-CR são mais robustos para lidar com observações atípicas do que o SN-CR.

3.5 Estudo de simulação IV: qualidade das estimativas dos modelos SMSN-CR

O foco principal deste estudo de simulação é investigar o impacto na inferência final dos modelos SMSN-CR (sob um ponto de vista Bayesiano) quando a hipótese de normalidade é inapropriada, para diferentes níveis de censura nos dados. Para isto, geramos uma variável elatória Y conforme o modelo geral de regressão para dados censurados definidos em (2.3.1) e (2.3.2) usando n=200, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-10,\ 2)^\top$, $\boldsymbol{X}$ uma matriz 200×2 cuja primeira coluna tem todos os elementos iguais a 1 e os elementos da segunda coluna são gerados de forma independente a partir de uma distribuição uniforme em $(0, 4)$, além disso, os erros ϵ_i, $i = 1, 2, \ldots, n$, foram gerados de forma independente de uma distribuição normal inversa gaussiana com parâmetros $\alpha = \alpha_0 = 5$, $\beta = \beta_0 = 4.9$, $\delta = \delta_0 = 2$ e $\mu = \mu_0 = -\frac{\delta\beta}{\sqrt{\alpha^2-\beta^2}}$ (aqui usamos a notação para os parâmetros da normal inversa gaussiana conforme definida em Barndorff-Nielsen (1997), onde pode-se encontrar maiores detalhes sobre esta distribuição. Note que o parâmetro de escala foi definido de forma a garantir que $\mathbb{E}_{Y_i}[Y_i] = \boldsymbol{x}_i^\top\boldsymbol{\beta}$, $i = 1, 2, \ldots, n)$. A distribuição dos erros, conforme foi gerada, apresenta assimetria e valores atípicos em relação à distribuição normal e sua densidade está graficada na Figura 3.5.

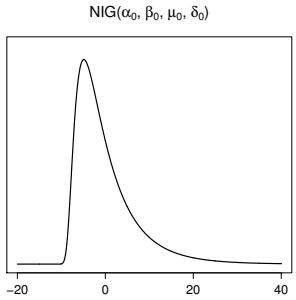


Figura 3.5: Estudo de simulação IV. Densidade da normal inversa gaussiana com parâmetros $\alpha = \alpha_0 = 5$, $\beta = \beta_0 = 4.9$, $\delta = \delta_0 = 2$ e $\mu = \mu_0 = -\frac{\delta\beta}{\sqrt{\alpha^2-\beta^2}}$.

Geramos um total de 150 conjuntos de dados segundo o esquema apresentado anteriormente, sendo que, após gerado, cada conjunto teve sua variável resposta censurada segundo 4 níveis de

censura: $10\%, 25\%, 40\%$ e 50% do total de observações. Desta forma, cada um dos 150 conjunto de dados gerou novos quatro conjuntos, sendo o original descartado. Os algoritmos de Gibbs desenvolvidos nas Seções 2.4 e 3.3 foram aplicados em cada um destes conjuntos de dados a fim de ajustar a eles os modelos N-CR, t-CR, SN-CR, St-CR e SSL-CR, sendo que a especificação das distribuições à priori foi feita conforme a discussão nas mesmas Seções, com $\boldsymbol{\mu}_0 = \mathbf{0}_2$, $\boldsymbol{\Sigma}_0 = 100\,\mathbf{I}_2$, $\mu_\Delta = 0$, $\sigma^2_\Delta = 100$,$a_\tau = 2.1$, $b_\tau = 3$, $c = 0.02$ e $d = 0.49$ para o modelo t-CR e St-CR e $c = 0.02$ e $d = 0.9$ para o modelo SSL-CR. Aqui, $\mathbf{0}_2$ denota um vetor de tamanho 2 com todos os componentes iguais a zero e $\mathbf{I}_2$ denota a matriz identidade com dimensão 2×2. Foram rodadas 60000 iterações do Gibbs, com um *burn-in* de 18000 e um *thinning* de 3, gerando cadeias finais de tamanho 14000.

Em cada um dos ajustes as estimativas pontuais dos parâmetros foram gravadas (média da cadeia MCMC gerada, após *burn-in* e *thinning*), de forma que pudemos calcular o erro médio absoluto (MAE) e o erro médio quadrático (MSE) para as estimativas dos coeficientes de regressão de cada modelo sob cada um dos quatro níveis de censura. Definimos o MAE e o MSE para o parâmetro β_j, $j = 1, 2$, como:

$$MAE = \frac{1}{150}\sum_{i=1}^{150}|\hat{\beta}_j^{(i)} - \beta_j| \quad , \quad MSE = \frac{1}{150}\sum_{i=1}^{150}(\hat{\beta}_j^{(i)} - \beta_j)^2$$

onde $\hat{\beta}_j^{(i)}$ é a estimativa de β_j no $i-$ésimo conjunto de dados simulado, para $j = 1, 2$ e $i = 1, 2, \ldots, 150$.

A Figura 3.6 apresenta o MAE e o MSE para as estimativas de β_1 e β_2 para os cinco modelos ajustados e para os quatro níveis de censura. A Figura 3.7 sumariza via *box-plot* as 150 estimativas pontuais obtidas para β_1 and β_2, comparando-as com o verdadeiro valor destes parâmetros para os diferentes modelos ajustados e diferentes níveis de censura.

Na Figura 3.6 observamos que as estimativas do intercepto β_1 é a que mais sofre impacto quando muda-se o modelo ajustado, sendo que os modelos simétricos (N-CR e t-CR) são os que apresentam maiores valores do MAE e MSE, enquanto os modelos St-CR e SSL-CR possuem performances muito parecidas e significativamente melhor do que o SN-CR. Enquanto isso, os valores do MAE e MSE para o parâmetro β_2 são pequenos sob todos os modelos e não apresentam diferenças significativas. É importante notar que todos os modelos perdem performance conforme aumentamos

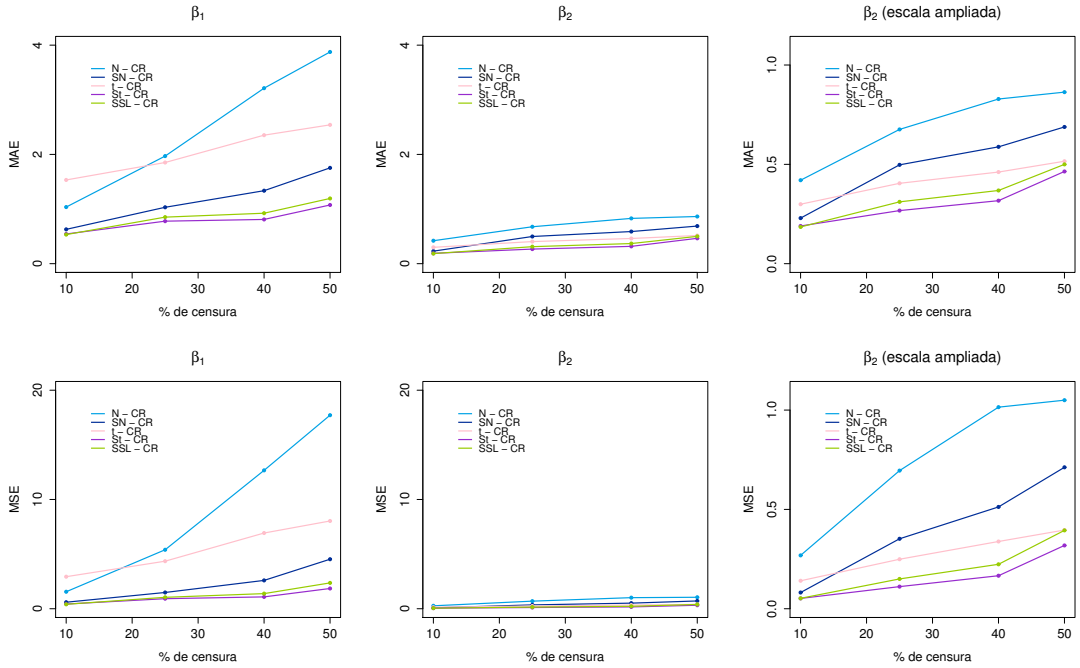


Figura 3.6: Estudo de simulação IV. MAE e MSE das estimativas pontuais de β_1 e β_2 para cinco modelos SMSN-CR e quatro diferentes níveis de censura.

o nível de censura, porém este aumento parece afetar com menor intensidade a qualidade das estimativas obtidas sob os modelos St-CR e SSL-CR.

A Figura 3.7 nos mostra um cenário parecido: comparando os modelos simétricos com suas versões assimétricas vemos uma melhora na qualidade das estimativas, o que também acontece quando comparamos modelos menos robustos com mais robustos (N-CR com t-CR e SN-CR com St-CR e SSL-CR). Novamente, os modelos St-CR e SSL-CR se comportam de forma muito semelhante e possuem estimativas mais precisas do que os outros modelos. Nestes *box-plots* notamos mais uma vez que todos os modelos são prejudicados pelo aumento do nível de censura.

Com este estudo mostramos como as inferências finais para um modelo de regressão para dados censurados podem ser prejudicadas quando há desvio da normalidade e um modelo adequado, capaz de acomodar assimetria e/ou observações atípicas, não for escolhido.

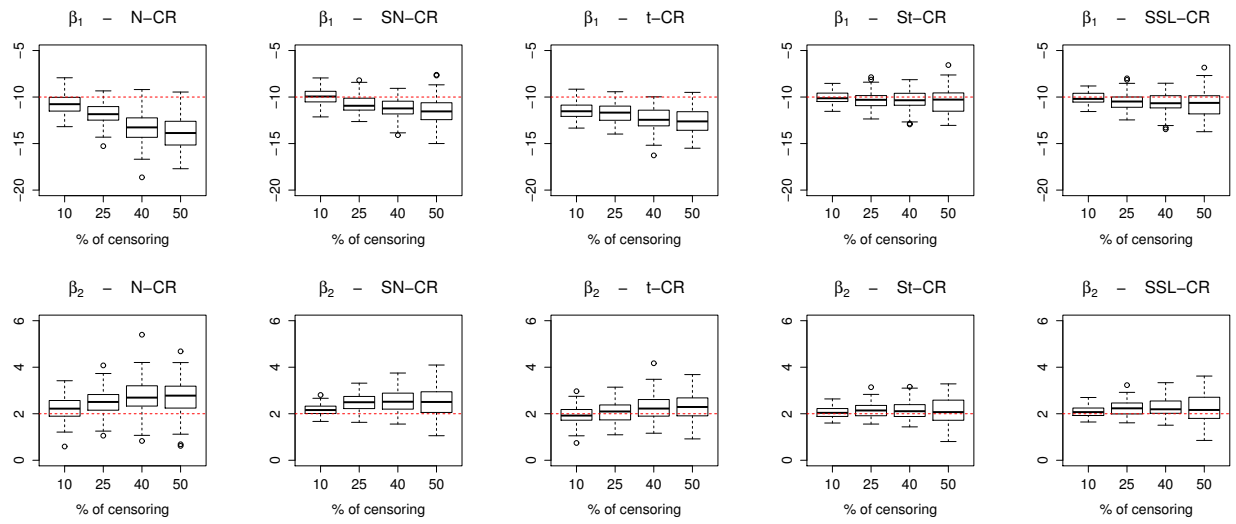


Figura 3.7: Estudo de simulação IV. *Box-plot* para as 150 estimativas pontuais de β_1 e β_2 para os modelos N-CR, SN-CR, t-CR, St-CR e SSL-CR e diferentes níveis de censura, em comparação com o valor verdadeiro dos parâmetros (linha vermelha).

3.6 Aplicação III

Nesta Seção ajustaremos o modelo de regressão para dados censurados sob as distribuições normal, t de Student, skew-normal, skew-t e skew-slash ao conjunto de dados "*Wage Rate*", descrito na Subseção 2.5.6, utilizando os algoritmos desenvolvidos nas Seções 2.4 e 3.3. Também faremos um estudo de diagnóstico destes modelos com base na metodologia descrita nas Seções 1.6.2 e 1.7. Nesta aplicação será utilizado o pacote `BayesCR` (veja sua descrição na Subseção 4.1.3) como respaldo computacional.

Modelaremos a variável censurada à esquerda Y_i, definida como o ganho potencial do indivíduo i em função das variáveis explicativas:

- x_2: idade,
- x_3: anos de estudo,

- x_4: número de filhos (em casa) com menos de 6 anos,
- x_5: número de filhos (em casa) entre 6 e 19 anos,

de forma que o vetor de covariáveis para o indivíduo i é dado por $\mathbf{x}_i^\top = (1, x_{2i}, x_{3i}, x_{4i}, x_{5i})$, com $i = 1, 2, \ldots, 753$.

Para o processo de estimação utilizaremos as densidades à priori definidas na Seção 3.3 com $\boldsymbol{\mu}_0 = \mathbf{0}_2$, $\boldsymbol{\Sigma}_0 = 100\,\mathbf{I}_2$, $\mu_\Delta = 0$, $\sigma^2_\Delta = 100$,$a_\tau = 2.1$, $b_\tau = 3$, $c = 0.02$, $d = 0.49$ (St-CR) e $d = 0.9$ (SSL-CR). O amostrador de Gibbs foi utilizado para gerar duas cadeias MCMC da distribuição à posteriori do vetor de parâmetros, de forma que pudemos analisar a convergência destas cadeias utilizando a estatística de Gelman-Rubin (veja Seção 1.5). Cada cadeia MCMC tem tamanho inicial 400,000 e, considerando um *burn-in* de 100,000 e um *thinning* de 30, tamanho final igual a 10,000.

A Tabela 3.1 mostra a média a posteriori (Média), o desvio padrão (SD), o intervalo HPD (95%) e a estatística de Gelman-Rubin ($\hat{R}$) para os parâmetros de cada um dos modelos ajustados. Pode-se notar que a mais impactante diferença entre os modelos ajustados é a significância do intercepto: somente sob o modelo SSL-CR o intervalo HPD para β_1 não contém o 0. Um outro aspecto a se observar é os valores pequenos das estimativas pontuais de ν sob os modelos t-CR, St-CR e SSL-CR, indicando que o modelo N-CR ou SN-CR podem não ser adequados para este conjunto de dados, já que a distribuição t de Student tende à normal e a skew-t e skew-slash tendem à skew-normal conforme $\nu \to \infty$. Sobre a interpretação dos parâmetros, todos os modelos concordam que o ganho potencial de uma mulher aumenta conforme mais anos de escolaridade ela tem e diminui conforme aumenta sua idade e/ou o número de filhos.

A Tabela 3.2 compara os ajustes dos cinco modelos considerados usando os critérios discutidos na Seção 1.6.2. Note que os modelos com caudas pesadas têm uma perfomance significativamente melhor (comparando-se o N-CR com o t-CR e também o SN-CR com o St-CR e SSL-CR), além disso os modelos assimétricos são também mais adequados do que suas versões simétricas (comparando-se o N-CR com o SN-CR e o t-CR com o St-CR). De fato o $p-$valor Bayesiano indica que os modelos simétricos e/ou não robustos (N-CR, SN-CR e t-CR) não são uma boa escolha

	Modelos							
	N-CR				**T-CR**			
Parâmetro	Média	SD	HPD (95%)	$\hat{R}$	Média	SD	HPD (95%)	$\hat{R}$
β_1	-2.752	1.748	(-6.133; 0.665)	1.000003	-1.184	1.433	(-3.937; 1.669)	1.000005
β_2	-0.106	0.028	(-0.161;-0.051)	1.000007	-0.111	0.023	(-0.155;-0.066)	1.000001
β_3	0.731	0.084	(0.569; 0.896)	0.999999	0.655	0.073	(0.514; 0.794)	1.000001
β_4	-3.056	0.448	(-3.923;-2.188)	1.000000	-3.166	0.398	(-3.951;-2.404)	0.999999
β_5	-0.215	0.153	(-0.521; 0.077)	1.000003	-0.294	0.130	(-0.548;-0.037)	1.000001
σ^2	21.325	1.5999	(18.222;24.483)	1.000010	11.644	1.019	(9.773;13.739)	1.000082
ν	—	—	—		5.351	0.626	(4.557; 6.578)	1.000451
	SN-CR				**ST-CR**			
Parâmetro	Média	SD	HPD (95%)	$\hat{R}$	Média	SD	HPD (95%)	$\hat{R}$
β_1	-1.034	1.632	(-4.178;2.206)	1.000004	-3.058	1.516	(-5.856; 0.083)	1.000025
β_2	-0.120	0.026	(-0.170;-0.070)	0.999999	-0.088	0.024	(-0.133;-0.040)	1.000011
β_3	0.675	0.081	(0.519; 0.836)	0.999999	0.673	0.068	(0.540; 0.806)	1.000012
β_4	-3.243	0.442	(-4.112;-2.389)	1.000005	-2.809	0.387	(-3.569;-2.065)	1.000011
β_5	-0.259	0.146	(-0.542;0.030)	1.000001	-0.267	0.128	(-0.510;-0.011)	1.000007
σ^2	33.708	3.270	(27.143; 39.833)	1.000229	22.562	4.495	(13.774;31.283)	0.999999
λ	1.803	0.380	(1.159; 2.576)	1.000663	-1.422	0.377	(-2.141;-0.656)	1.000060
ν	—	—	—	4.877	0.255	(4.656; 5.369)	1.006467	
	SSL-CR							
Parâmetro	Média	SD	HPD (95%)	$\hat{R}$				
β_1	-4.127	1.485	(-7.097; -1.349)	1.000003				
β_2	-0.079	0.023	(-0.124; -0.036)	1.000013				
β_3	0.669	0.065	(0.542; 0.796)	1.000006				
β_4	-2.688	0.366	(-3.406; -1.979)	0.999998				
β_5	-0.265	0.122	(-0.505; -0.030)	1.000003				
σ^2	13.424	2.369	(8.938; 18.123)	1.000026				
λ	-1.940	0.397	(-2.728; -1.183)	1.000036				
ν	1.063	0.064	(1.001; 1.191)	1.000144				

Tabela 3.1: *Wage rate data.* Média e desvio padrão a posteriori, intervalo HPD (95%) e estatística de Gelman-Rubin sob os modelos N-CR, t-CR, SN-CR, St-CR e SSL-CR.

para este conjunto de dados. Dentre todos os modelos, o SSL-CR mostra-se o mais adequado.

Modelo	LPML	DIC	EAIC	EBIC	$WAIC_1$	$WAIC_2$	p_B
N-CR	-1489.290	2975.017	2975.381	3003.126	2978.080	2978.651	0.3693
t-CR	-1447.537	2893.862	2895.804	2928.172	2894.622	2894.943	0.8181
SN-CR	-1479.075	2955.640	2955.402	2987.770	2958.067	2958.144	0.6098
St-CR	-1441.834	2881.913	2884.199	2921.192	2883.431	2883.766	0.5293
SSL-CR	**-1432.518**	**2863.778**	**2864.841**	**2901.834**	**2864.796**	**2865.119**	0.5425

Tabela 3.2: Wage rate data. Comparação entre os modelos SMSN-CR.

Considerando a amostra MCMC da distribuição à posteriori do vetor de parâmetros dos cinco modelos ajustados para os dados *wage rate*, computamos as medidas $q-$divergentes descritas na Seção 1.7 a fim de identificar possíveis observações influentes (utilizando $p = 0.8$ na Equação (1.7.6) para calcular o ponto limite a partir do qual uma observação é considerada influente). As Figuras 3.8, 3.9 e 3.10 mostram respectivamente as medidas de Kullback-Lieber, distância J e distância L_1 sob os cinco modelos SMSN-CR ajustados.

Note que as observações #185, #349 e #408 foram consideradas influentes sob os modelos N-CR e SN-CR segundo todas as medidas calculadas, enquanto a observação #394 o foi sob estes dois modelos somente segundo a distância J e sob o modelo N-CR segundo a medida K-L (embora ela tenha ficado muito próxima do ponto limite sob o SN-CR). O que é notável é que nenhuma das observações citadas foram consideradas influentes quando os modelos de caudas mais pesadas foram ajustados, o t-CR, St-CR e SSL-CR, mostrando a robustez destes em relação ao N-CR e SN-CR na presença de observações atípicas.

A fim de avaliar o real impacto das observações #185, #349, #394 e #408 na inferência sobre os coeficientes de regressão, em geral os parâmetros que mais influenciam na interpretação prática do problema, comparamos os modelos N-CR e SSL-CR (respectivamente o menos e o mais adequado para o conjunto de dados segundo a Tabela 3.2) em relação à mudança relativa que a estimativa pontual (média à posteriori) destes parâmetros sofre quando cada uma destas observações é excluída do conjunto de dados, calculamos também esta mudança relativa quando

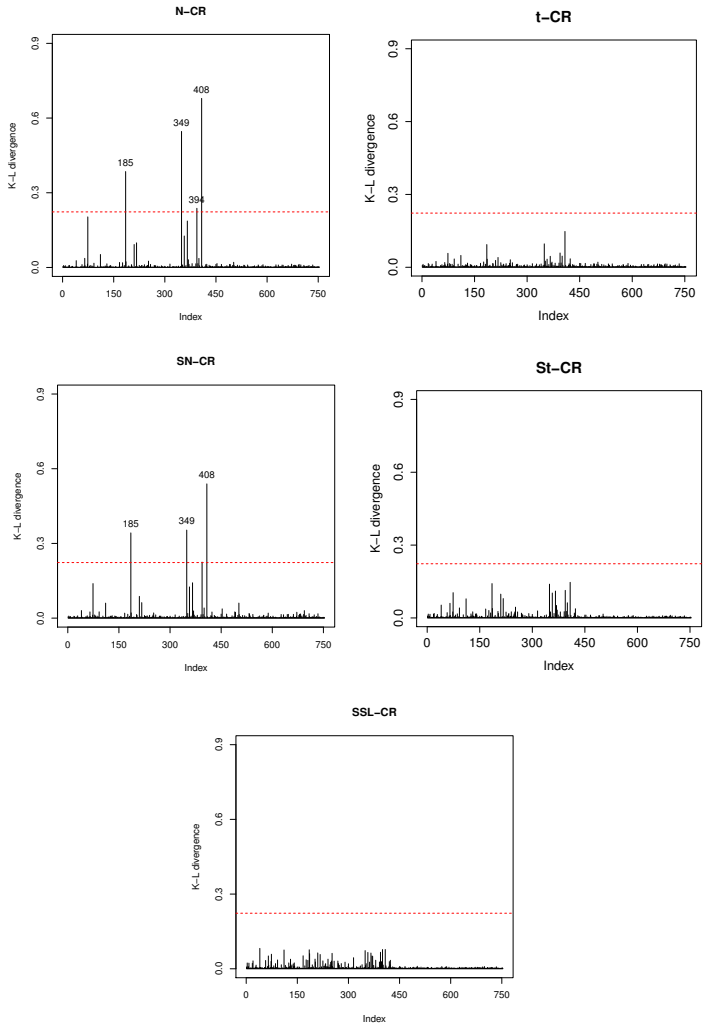


Figura 3.8: Wage rate data. Divergência de Kullback-Lieber para os modelos SMSN-CR.

todas estas observações são excluídas. Assim, defina a mudança relativa na estimatica do parâmetro θ por $\mathbf{RC}_\theta = \left|(\hat{\theta} - \hat{\theta}_{(I)})/\hat{\theta}\right| \times 100$, onde $\hat{\theta}$ denota a estimativa pontual de θ utilizando toda a amostra e $\hat{\theta}_{(I)}$, excluindo-se o conjunto I de observações. O resultado deste procedimento encontra-se na Tabela 3.3, onde o símbolo $*$ indica parâmetros que não eram significativos no ajuste original e passaram a ser quando determinada observação foi removida. Nesta Tabela podemos observar que o intercepto β_1 é o mais impactado por estas observações quando comparado com os outros coeficientes de regressão. Todas as mudanças relativas sob o modelo SSL-CR são menores do que o N-CR, além disso nenhum parâmetro teve sua significância estatística alterada sob o ajuste do

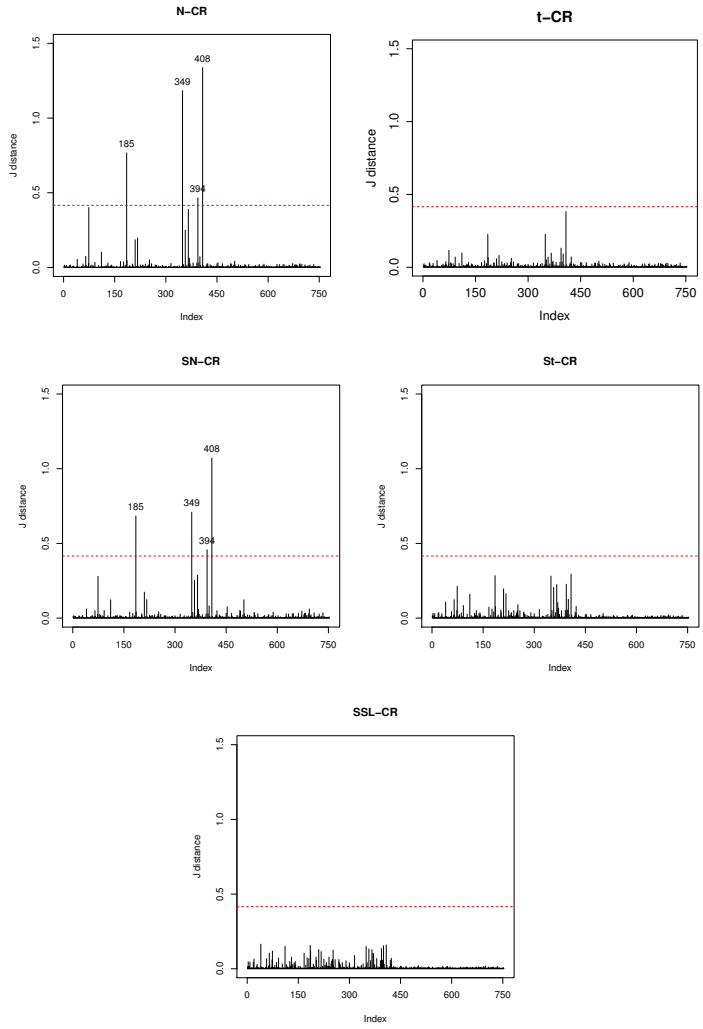


Figura 3.9: Wage rate data. Distância J para os modelos SMSN-CR.

modelo SSL-CR, enquanto sob o N-CR o parâmetro β_5, que não era considerado significativo no ajuste original, passou a ser quando removemos a observação #185 e também quando todas as observações influentes foram removidas. Estes fatos confirmar mais uma vez que o SSL-CR é mais robusto do que o N-CR, conforme já era esperado.

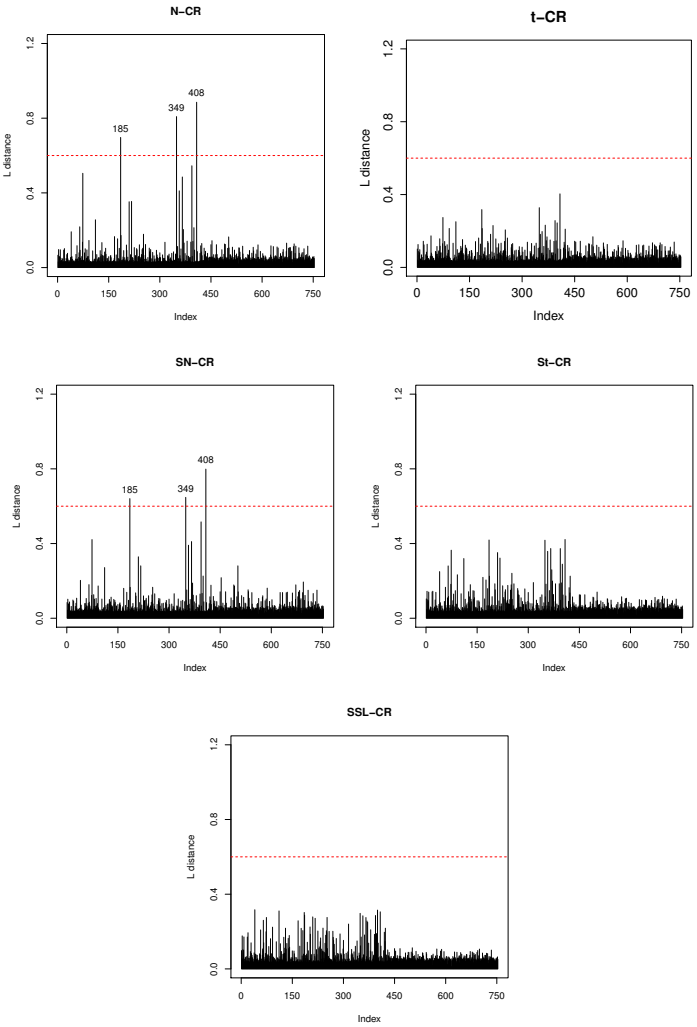


Figura 3.10: Wage rate data. Distância L_1 para os modelos SMSN-CR

	Modelos									
	N-CR					**SSL-CR**				
Conjunto I	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
{#185}	2.43	3.63	0.86	3.30	3.56∗	2.46	2.27	0.22	1.11	1.91
{#349}	22.59	10.19	1.01	0.44	19.85	0.37	0.06	0.19	0.62	0.59
{#394}	8.46	3.14	4.09	2.07	5.65	1.16	1.52	0.49	0.32	2.64
{#408}	7.16	0.36	0.80	3.48	19.47	2.06	1.42	0.14	0.39	1.65
{#185, #349, #394, #408}	33.15	1.02	7.36	0.46	35.53∗	7.79	6.37	0.20	3.23	3.57

Tabela 3.3: Wage rate data. Mudança relativa (em %) para os coeficientes de regressão sob os modelos N-CR e SSL-CR.

Capítulo 4

Considerações finais

4.1 Produção técnica

Nesta Seção, descreveremos a produção técnica derivada desta dissertação de mestrado.

4.1.1 Artigos aceitos para publicação

- "*Influence diagnostics for Student-t censored linear regression models*"
  Autores: Monique Bettio Massuia, Celso Rômulo Barbosa Cabral, Larissa Ávila Matos e Victor Hugo Lachos Dávila.
  Periódico: Statistics (Taylor & Francis)
  DOI: 10.1080/02331888.2014.958489

Este artigo é referente à Subseção 2.5 desta dissertação e apresenta um estudo de inferência frequentista para o modelo t-CR com base no algoritmo EM e em técnicas de diagnósticos através de influência global e local.

4.1.2 Artigos submetidos

- "*Bayesian Analysis of Censored Linear Regression Models with Scale Mixtures of Skew-Normal Distributions*"

Autores: Monique Bettio Massuia, Aldo Medina Garay, Victor Hugo Lachos Dávila e Celso Rômulo Cabral.

Este artigo é referente ao Capítulo 3 desta dissertação e apresenta um estudo de inferência Bayesiano para os modelos SMSN-CR, utilizando o amostrador de Gibbs para realizar os ajustes e baseando-se nas medidas q-divergentes para realizar diagnóstico de influência.

4.1.3 Pacotes para o *software* `R`

`CensRegMod`

Este pacote foi desenvolvido para dar suporte computacional aos desenvolvimentos da Subseção 2.5 e ao artigo Massuia *et al.* (2014), estimando os parâmetros dos modelos t-CR e N-CR via algoritmo EM, aproximando os erros padrões dos estimadores dos coeficientes de regressão através do método mostrado na Subseção 2.5.2 e calculando as medidas de diagnóstico local e global conforme foi feito na Subseção 2.5.3. Este pacote calcula também os critérios AIC, BIC e EDC para seleção de modelos e encontra-se disponível para download gratuito no site: `http://cran.r-project.org/web/packages/CensRegMod/index.html`.

Descrição

O comando principal a ser utilizado neste pacote é da seguinte forma:

Códigos em `R`

```
em.cens(cc, x, y, nu, dist, diagnostic, typediag)
```

com os seguintes argumentos:

- *cc*: vetor de indicadores de censura, cujo $i-$ésimo componente é igual a 1 se a observações correspondente for censurada ou igual a 0, se não for.
- x: matriz de desenho.
- y: vetor com as observações da variável resposta.

- *nu*: valor inicial para os graus de liberdade (ou `NULL`, se for o modelos N-CR a ser ajustado).
- *dist*: `Normal`, se o modelo a ser ajustado for o N-CR ou `T`, se for o t-CR.
- *diagnostic*: `TRUE` se quiser que as medidas de diagnóstico apresentadas na Subseção 2.5.3 sejam computadas ou `FALSE`, caso contrário.
- *typediag*: Caso `diagnostic=TRUE`, se `typediag=1` calcula-se a distância generalizada de Cook (e suas decomposições para os subconjuntos de parâmetros $\boldsymbol{\beta}$ e σ^2), se `typediag=2`, calcula-se as medidas de influência local sob ponderação de casos e, se `typediag=3`, sob a perturbação na escala.

Exemplo

Códigos em `R`

```
> library(CensRegMod)
> data(wage.rates)
> attach(wage.rates)
> N_CR = em.cens(cc,-x,-y,dist="Normal",diagnostic=TRUE,typediag=1)
> t_CR = em.cens(cc,-x,-y,nu=5,dist="T")
```

Observação: Ao contrário do desenvolvimento da Seção 2.5, este pacote estima o valor do parâmetro ν no caso do modelo t-CR, tomando como estimativa em cada iteração o argumento que maximiza a função verossimilhança, já avaliada nas estimativas calculadas para $\boldsymbol{\beta}$ e σ^2. No entanto, a análise de diagnóstico é feita considerando ν um valor fixo, tomado como o valor estimado ao final do algoritmo EM. É importante notar também que, embora este pacote tenha sido desenvolvido para lidar com censuras à direita, pode-se fazer uma adaptação para o ajuste do modelo censurado à esquerda, passando como argumento para a função em.cens o vetor $-\boldsymbol{y}$ e a matriz $-\boldsymbol{x}$ ao invés de seus valores originais.

SMNCensReg

Este pacote foi desenvolvido em conjunto com Aldo Medina Garay e dá suporte computacional à Subseção 2.5, assim como à sua tese de doutorado (veja Garay, 2014). Este pacote ajusta modelos de regressão para respostas censuradas sob as seguintes distribuições da classe SMN (mistura de escala normal) a saber: Normal, t de Student, Pearson VII, Slash e Normal Contaminada. Os erros padrões para os estimadores dos coeficientes de regressão são estimados através do método descrito na Subseção 2.5.2. Também faz o gráfico de envelope com base nos resíduos *deviance* para a análise de diagnóstico e calcula os critérios AIC, BIC e EDC para seleção de modelos. Disponível para download gratuito no site `http://cran.r-project.org/web/packages/SMNCensReg/index.html`.

Descrição

O comando principal a ser utilizado neste pacote é da seguinte forma:

Códigos em `R`

```
CensReg.SMN(cc,x,y,LS,nu,delta,cens,dist,show.envelope,error,iter.max)
```

com os seguintes argumentos:

- *cc*: vetor de indicadores de censura, cujo $i-$ésimo componente é igual a 1 se a observações correspondente for censurada ou igual a 0, se não for.
- x: matriz de desenho.
- y: vetor com as observações da variável resposta no caso de censuras à direita ou à esquerda. No caso de censura intervalar, `y` é o vetor com os limites inferiores dos intervalos observados.
- LS: no caso de censura intervalar, contém os limites superiores dos intervalos observados.
- *nu*: valor inicial para os graus de liberdade da t de Student, Pearson VII ou Slash. Um vetor bidimensional com os valores iniciais dos parâmetros da Normal Contaminada.`NULL`, para a distribuição Normal.
- *delta*: valor inicial para o segundo parâmetro da distribuição Pearson VII (ou `NULL`, para as outras distribuições).

- *cens*: `left`, se a variável resposta for censurada à esquerda, `right`, se for à direita e `interval`, se a censura for intervalar.
- *dist*: `Normal`, se a distribuição assumida para os erros do modelo for Normal, `T`, se for t de Student, `PearsonVII`, se for Pearson VII, `Slash`, se for Slash e, finalmente, `NormalC`, se for Normal contaminada.
- *show.envelope*: `TRUE`, caso queira que o gráfico de envelope seja mostrado ou `FALSE`, caso contrário.
- *error*: precisão para que o critério de convergência seja atingido. O padrão do pacote é 0.0001.
- *iter.max*: Número máximo de iterações. O padrão do pacote é 300.

Exemplo

Códigos em `R`

```
> library(SMNCensReg)
> data(wage.rates)
> attach(wage.rates)
> y = wage.rates$wage
> x = cbind(wage.rates$age,wage.rates$educ,wage.rates$kidslt6,wage.rates$kidsge6)
> cc = c(rep(0,428),rep(1,325))
> N_CR = CensReg.SMN(cc,x,y,cens="left",dist="Normal")
> t_CR = CensReg.SMN(cc,x,y,nu=3,cens="left",dist="T",show.envelope="TRUE")
```

`BayesCR`

Este pacote foi desenvolvido em conjunto com Aldo Medina Garay e dá suporte computacional ao estudo de inferência Bayesiana para as Subseções 3.3 e 2.4.1, assim como à sua tese de doutorado (veja Garay, 2014). Com este pacote é possível ajustar, via amostrador de Gibbs, modelos de regressão para dados censurados (à direita ou à esquerda) sob as seguintes distribuiçõe da classe

SMSN (mistura de escala skew-normal): Normal, Skew-Normal, t de Student, Skew t de Student, Slash, Skew-Slash e Normal Contaminada. O pacote também calcula os seguintes critérios de seleção de modelos: LPML, DIC, EAIC, EBIC, WAIC_1 e WAIC_2, além das medidas de divergência de Kullback-Lieber, e as distância J, L_1 e Chi. Além disso, o pacote pode ser utilizado para gerar observações das distribuições SMSN consideradas.

Descrição

O comando que ajusta os modelos SMSN-CR via amostrador de Gibbs é da seguinte forma:

Códigos em `R`

```
Bayes.CR(cc,x,y,cens,dist,influence,criteria,spacing,prior,hyper,n.thin,burnin,
n.iter,n.chains,chain)
```

---

com os seguintes argumentos:

- *cc*: vetor de indicadores de censura, cujo $i-$ésimo componente é igual a 1 se a observações correspondente for censurada ou igual a 0, se não for.
- x: matriz de desenho.
- y: vetor com as observações da variável resposta.
- *dist*: `Normal`, se a distribuição assumida para os erros do modelo for Normal, `SN`, se for Skew-Normal, `T`, se for t de Student, `ST`, se for Skew-t, `Slash`, se for Slash, `SSL` se for Skew-Slash e, finalmente, `NormalC`, se for Normal contaminada.
- *influence*: `TRUE`, caso queira que sejam computadas a divergência de Kullback-Lieber e as distância J, L_1 e Chi ou `FALSE` caso contrário.
- *criteria*: `TRUE`, caso queira que sejam computados os critérios LPML, DIC, EAIC, EBIC, WAIC_1 e WAIC_2 ou `FALSE` caso contrário.

- *spacing*: Somente deve ser fornecido se `influence=TRUE` ou `criteria=TRUE`, especificando o *lag* entre observações da cadeia final a ser utilizado para o cálculo das medidas de influência e/ou dos critérios de seleção de modelos.

- *prior*: Distribuição à priori a ser considerada para os graus de liberdade no caso da distribuição t de Student, sendo `Exp` para a priori exponencial, `Jeffreys` para a de Jeffreys, `Unif` para a uniforme ou `Hierar` para a priori hierárquica, utilizada nesta dissertação (veja Garay (2014)).

- *hyper*: valor do hiperparâmetro da priori exponencial ou `NULL`, se foram utilizadas outras prioris ou modelos.

- *n.thin*: "lag" a ser considerado para a cadeia final de observações.

- *burnin*: "burn-in" a ser considerado para a cadeia final de observações.

- *n.iter*: número de iterações para cada cadeia do amostrador de Gibbs.

- *n.chains*: número de cadeias paralelas a serem geradas pelo amostrador de Gibbs.

- *chain*: `TRUE`, caso as cadeias finais devam ser armazenadas para análise ou `FALSE`, caso contrário.

Além disso, é possível gerar observações das distribuições consideradas a partir da seguinte função:

Códigos em `R`

```
rSMSN(n,mu,sigma2,lambda,nu,dist)
```

Com os seguintes argumentos:

- *n*: número de observações a serem geradas.

- *mu*: parâmetro de locação.

- *sgiam2*: parâmetro de escala.

- *lambda*: parâmetro de forma relativo à assimetria.
- *nu*: graus de liberdade para as distribuições t de Student, Slash e suas versões assimétricas.
- *dist*: distribuição da qual se quer gerar. `Normal`, para Normal, `SN`, para Skew-Normal, `T`, para t de Student, `ST`,para Skew-t, `Slash`, para Slash, `SSL` para Skew-Slash e, finalmente, `NormalC`, para Normal contaminada.

Exemplo

Códigos em `R`

```
> library(BayesCR)
> data(wage.rates)
> attach(wage.rates)
> y = wage.rates$wage
> x = cbind(wage.rates$age,wage.rates$educ,wage.rates$kidslt6,wage.rates$kidsge6)
> cc = c(rep(0,428),rep(1,325))
> t_CR = Bayes.CR(cc,x,y,cens="left",dist="T",influence=FALSE,criteria=FALSE,
prior="Hierar, n.thin=10, burnin=10000,n.iter=100000,n.chains=1,chain=FALSE)
```

4.2 Trabalhos futuros

Os desenvolvimentos realizados nesta dissertação abrem perspectivas para diversos trabalhos futuros, como:

- Desenvolver um estudo de inferência e diagnóstico frequentista para os modelos SMSN-CR com base no algoritmo EM-SAEM (que utiliza aproximações MCMC para as esperanças calculadas na etapa E do algoritmo).
- Estender os resultados apresentados para modelos não lineares com respostas censuradas.
- Estender os desenvolvimentos desta dissertação para o caso multivariado.

4.3 Conclusão

Neste trabalho consideramos a classe de distribuições de mistura de escala skew-normal como alternativa para a hipótese convencional de normalidade atribuída aos erros dos modelos de regressão lineares para respostas censuradas, generalizando os trabalhos de Barros *et al.* (2010), que faz análise de diagnóstico para o modelo Tobit, de Arellano-Valle *et al.* (2012), que desenvolve um estudo de inferência frequentistica para o modelo t-Student e de Garay (2014), que considera o modelo de regressão com respostas censuradas sob a classe de distribuições de mistura de escala normal.

Sob a perspectiva frequentista, demos atenção especial aos modelos N-CR e t-CR no capítulo 2, desenvolvendo o algoritmo EM para a estimação paramétrica destes modelos e também um estudo de diagnóstico baseados em medidas de influência local e global. Os pacotes `CensRegMod` e `SMNCensReg` dão suporte computacional a este problema e estão disponíveis para download no repositório `CRAN`. Foram feitos dois estudos de simulação, o primeiro compara a robustez das estimativas EM obtidas sob os modelos N-CR e t-CR quando uma perturbação é feita no conjunto de dados, mostrando que as estimativas do modelo t-CR são menos sensíveis à perturbação do que as do modelo N-CR. O segundo estudo de simulação avaliou a consistência do método utilizado para estimar o desvio padrão dos estimadores EM dos parâmetros de regressão, que mostrou-se bastante adequado. Os modelos foram então ajustados ao conjunto de dados de Mroz (1987) utilizando os pacotes citados e, como esperado, o modelo t-CR mostrou-se bem mais adequado do que o N-CR para acomodar observações atípicas.

Sob a perspectiva Bayesiana, consideramos o modelo de regressão linear para dados censurados sob diversas distribuições da família SMSN: a normal, normal assimétria, t de Student, t de Student assimétrica, e, finalmente, slash assimétrica. A estimação paramétrica foi feita com base no amostrador de Gibbs e a análise de diagnóstico, com base nas medidas q-divergentes. O pacote `BayesCR` dá respaldo computacional ao problema e foi utilizado nas aplicações I e III, assim como nos estudos de simulação III e IV, que compararam a qualidade das estimativas dos modelos citados na presença de observações atípicas e/ou assimetria, mostrando o impacto sofrido pelas estimativas dos modelos quando é utilizada uma distribuição que não consegue acomodar

tais características presentes no conjunto de dados.

Referências Bibliográficas

Abanto-Valle, C. A., Lachos, V. H. & Dey, D. K. (2013). Bayesian estimation of a skew-t stochastic volatility model. *Methodology: Computing in Applied Probability (revision invited).*

Ando, T. (2010). *Bayesian Model Selection and Statistical Modeling.* Chapman & Hall/CRC.

Andrews, D. R. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, **36**, 99–102.

Arellano-Valle, R. B., Castro, L. M., Farías, G. G. & Gajardo, K. A. M. (2012). Student-t censored regression model: properties and inference. *Statistical Methods and Applications*, **21**(4), 453–473.

Atkinson, A. C. (1985). *Plots, transformations and regression. An introduction to graphical methods of diagnostic regression analysis.* Oxford Statistical Science Series.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. & Genton, M. G. (2008). Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, **76**, 1490–1507.

Bagnoli, M. & Bergstrom, T. (2005). Log-concave probability and its applications. *Economic Theory*, **24**(2), 445–469.

Bai, Z. D., krishnaiah, P. R. & Zhao, L. C. (1989). On rates of convergence of efficient detection criteria in signal processing with white noise. *IEEE Transactions on Information Theory*, **35**, 380–388.

Barndorff-Nielsen, O. E. (1997). Normal inverse gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics*, **24**, 1–13.

Barros, M., Galea, M., González, M. & Leiva, V. (2010). Influence diagnostics in the tobit censored response model. *Statistical Methods & Applications*, **19**, 716–723.

Basso, R. M. (2009). *Misturas Finitas de Misturas de Escala Skew-Normal*. Dissertação do mestrado, Instituto de Matemática, Estatística e Computação Científica, IMECC - UNICAMP.

Basso, R. M., Lachos, V. H., Cabral, C. R. B. & Ghosh, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics and Data Analysis*, **54**, 2926–2941.

Bayes, C. L. (2005). *Inferência Bayesiana no modelo normal assimétrico*. Dissertação do mestrado, Instituto de Matemática e Estatística, IME-USP.

Branco, M. D. & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99–113.

Brooks, S. P. (2002). Discussion on the paper by spiegelhalter, best, carlin, and van der linde (2002). *JRSSB*, **64**(4), 616–618.

Cabral, C. R. B. & Madruga, V. H. (2012). Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the randon-effects population. *Journal of Statistical Planning and Inference*, **142**, 181–200.

Cancho, V. G., Dey, D. K., Lachos, V. H. & Andrade, M. G. (2011). Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: Estimation and case influence diagnostics. *Computational Statistics and Data Analysis*, **55**, 588–602.

Carlin, B. P. & Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, Boca Raton, second edition.

Chib, S. & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, **49**, 327–335.

Colosimo, E. & Giolo, S. (2006). *Análise de sobrevivência aplicada*. ABE - Projeto Fisher.

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

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

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

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, **2**, 299–318.

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

Dey, D. K., Chen, M. H. & Chang, H. (1997). Bayesian approach for the nonlinear random effects models. *Biometrics*, **53**, 1239–1252.

do Carmo, M. P. (2006). *Geometria diferencial de curvas e superfícies*. Sociedade Brasileira de Matemática, second edition.

Fernandez, C. & Steel, M. J. F. (1999). Multivariate student-t regression models: pitfalls and inference. *Biometrika*, **86**, 153–157.

Fonseca, T. C. O., Ferreira, M. A. R. & Migon, H. S. (2008). Objective bayesian analysis for the student-t regression model. *Biometrika*, **95**, 325–333.

Garay, A. M. (2009). *Modelos não lineares sob a classe de distribuições misturas da escala skew-normal*. Dissertação de mestrado, Instituto de Matemática, Estatística e Computação Científica, IMECC - UNICAMP.

Garay, A. M. (2014). *Modelos de regressão para dados censurados sob distribuições simétricas*. Tese de doutorado, Instituto de Matemática e Estatística, IME - USP.

Garay, A. M., Lachos, V. H., Bolfarine, H. & Cabral, C. R. B. (2013). Bayesian analysis censored linear regression models with scale mixtures of normal distributions. Technical Report 14, Universidade Estadual de Campinas.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

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

Gelman, A. (1992). Iterative and non-iterative simulation algorithm. *Computing Science and Statistics*, **7**, 457–511.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian data analysis*. Chapman & Hall/CRC.

Gelman, A., Hwang, J. & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, **24**, 997–1016.

Genç, A. I. (2013). Moments of truncated normal/independent distributions. *Statistical Papers*, **54**, 741–754.

Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, **8**, S19–S40.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *JBiometrika*, **57**, 97–109.

Kalbfleisch, J. & Lawless, J. (1992). Some useful statistical methods for truncated data. *Journal of Quality and Technology*, **24(3)**, 145–152.

Kim, H. J. (2008a). Moments of truncated student-t distribution. *Journal of the Korean Statistical Society*, **37**, 81–87.

Kim, H. M. (2008b). A note on scale mixtures of skew normal distribution. *Statistics and Probability Letters*, **78**, 1694–1701.

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.

Lange, K. L., Little, J. A. & Taylor, M. G. J. (1989). Robust modeling using the t distribution. *Journal of the American Statistical Association*, **84**, 881–896.

Lee, S. Y. & Xu, L. (2004). R influence analysis of nonlinear mixed-effects models. *Computational Statistics and Data Analysis*, **45**, 321–341.

Lin, T. I., Lee, J. C. & Hsieh, W. J. (2007). Robust mixture modelling using the skew t distribution. *Statistics and Computing*, **17**, 81–92.

Little, R. J. A. (1999). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, **37**, 23–38.

Louis, T. A. (1982). Finding the observed information matrix when using the em. *Journal of the Royal Statistical Society*, **44**, 226–233.

Lucas, A. (1997). Robustness of the student-t based m-estimator. *Communications in Statistics, Theory and Methods*, **26**, 1165–1182.

Massuia, M. B., Cabral, C. R. B., Matos, L. A. & Lachos, V. H. (2014). Influence diagnostics for student-t censored linear regression models. *Statistics*, **DOI: 10.1080/02331888.2014.958489**.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Meza, C., Osorio, F. & la Cruz, R. D. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, **22**, 121–139.

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

Nelson, W. (1990). Hazard plotting of left truncated life data. *Journal of Quality and Technology*, **22(3)**, 230–238.

Ortega, E. M., Bolfarine, H. & Paula, G. A. (2003). Influence diagnostics in generalized log-gamma regression models. *Computational Statistics and Data Analysis*, **42**, 165–186.

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

Poom, W. Y. & Poon, Y. S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society*, **61**, 51–61.

Schmee, J. & Hahn, G. J. (1979). A simple method for regression analysis with censored data. *Technometrics*, **21**, 417–432.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). A bayesian measure of model complexity and fit (with discussion). *Journal of Royal Statistical Society*, **64**, 583–639.

Tan, M., Tian, G. L. & Ng, K. W. (2009). *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*. Chapman & Hall/CRC Biostatistics Series, Boca Raton, NY.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the Maerican Statistical Association*, **82**, 528–549.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701–1762.

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

Vidal, I. & Castro, L. M. (2010). Influential observations in the independent Student-t measurement error model with weak nondifferential error. *Chilean Journal of Statistics*, **1**, 17–34.

Wang, J. & Genton, M. G. (2006). The multivariate skew-slash distribution. *Journal of Statistical Planning and Inference*, **136**, 209–220.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, **11**, 3571–3594.

Zhu, H. & Lee, S. (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society*, **63**, 111–126.

Zhu, H., Lee, S., Wei, B. & Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, **88**, 727–737.

Zhu, H., Ibtahim, J. G. & Shi, X. (2009). Diagnostic measures for generalized linear models with missing covariates. *Scandinavian Journal of Statistics*, **36**, 686–712.

Apêndice A

Desenvolvimento da pdf e cdf da skew-t

Nesta Seção derivaremos a forma fechada da pdf e cdf da distribuição skew-t. Para isso, considere o seguinte Lema:

Lema 6. *Seja* $U \sim G(\alpha, \beta)$, $\alpha > 0$ *and* $\beta > 0$. *Então, para qualquer vetor fixo* $\boldsymbol{w} \in \mathbb{R}^p$, *temos que:*

$$\mathbb{E}_U \left[\boldsymbol{\Phi}_p \left(\sqrt{U}\boldsymbol{w} \; ; \; \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) \right] = \mathcal{T}_p \left(\sqrt{\frac{\alpha}{\beta}}\boldsymbol{w} \; ; \; \boldsymbol{\mu}, \boldsymbol{\Sigma}, 2\alpha \right).$$

Prova:

Considere o vetor aleatório $\boldsymbol{V} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ independente de $U \sim G(\alpha, \beta)$, então:

$$\begin{aligned}
\mathbb{E}_U\left[\boldsymbol{\Phi}_p\left(\sqrt{U}\boldsymbol{w}\,;\,\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\right] &= \mathbb{E}_U\left[\mathbb{P}(\boldsymbol{V} \le \boldsymbol{w}\sqrt{U})\right] \\
&= \mathbb{E}_U\left[\mathbb{P}\left(\frac{\boldsymbol{V}}{\sqrt{U}} \le \boldsymbol{w}\right)\right] \\
&= \mathbb{E}_U\left[\mathbb{P}\left(\frac{\boldsymbol{V}}{\sqrt{2\beta U}} \le \sqrt{\frac{1}{2\beta}}\boldsymbol{w}\right)\right] \\
&= \mathbb{E}_U\left[\mathbb{P}\left(\frac{\boldsymbol{V}}{\sqrt{\beta/\alpha U}} \le \sqrt{\frac{\alpha}{\beta}}\boldsymbol{w}\right)\right] \\
&= \mathbb{E}_U\left[\mathbb{P}\left(X \le \sqrt{\frac{\alpha}{\beta}}\boldsymbol{w}\right)\right] \qquad \text{with} \quad X \sim t_p(\boldsymbol{\mu},\boldsymbol{\Sigma},2\alpha) \quad \text{(A.0.1)} \\
&= \mathcal{T}_p\left(\frac{\alpha}{\beta}\boldsymbol{w}\,;\,\boldsymbol{\mu},\boldsymbol{\Sigma},2\alpha\right),
\end{aligned}$$

aqui, a Equação (A.0.1) foi obtida usando os seguintes resultados (os quais não serão demonstrados):

- Se $U \sim G(\alpha, \beta)$, então, para qualquer constante $c > 0$, $c\,U \sim G(\alpha, \beta/c)$.
- Se $U \sim G(\alpha, 1/2)$, então U tem a mesma distribuição que $Y \sim \mathcal{X}_{2\alpha}$.
- Se $\boldsymbol{V} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ é independente de $U \sim \chi_\nu$, então $X = \frac{\boldsymbol{V}}{\sqrt{U/\nu}} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$,

onde $\mathcal{X}_\nu$ denota a distribuição chi-quadrado com ν graus de liberdade.

■

Agora, considerando o formato geral da pdf da classe SMSN dada na Equação (3.2.6), a densidade da skew-t é dada por:

$$\begin{aligned}
f(y) &= 2\int_0^\infty \phi(y\,;\,\mu, u^{-1}\sigma^2)\,\Phi\left(\frac{\lambda(y-\mu)}{u^{-1/2}\sigma}\,;\,0,1\right)\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}u^{\nu/2-1}\exp\left\{-\frac{\nu}{2}u\right\}\,du \\
&= \frac{\sqrt{2}\,(\nu/2)^{\nu/2}}{\sqrt{\pi}\sigma\Gamma(\nu/2)}\int_0^\infty u^{\frac{\nu-1}{2}}\exp\left\{-u\left(\frac{\nu}{2}+\frac{d(y)^2}{2}\right)\right\}\Phi\left(\lambda\,d(y)\,\sqrt{u}\right)du \\
&= \frac{\sqrt{2}(\nu/2)^{\nu/2}}{\sqrt{\pi}\sigma\Gamma(\nu/2)}\Gamma\left(\frac{\nu+1}{2}\right)\left(\frac{\nu+d(y)^2}{2}\right)^{-\frac{\nu+1}{2}}\mathbb{E}_X\left[\Phi\left(\lambda\,d(y)\,\sqrt{X}\,;\,0,1\right)\right],
\end{aligned}$$

onde $d(y) = \frac{y-\mu}{\sigma}$ e $X \sim Gamma\left(\frac{\nu+1}{2}, \frac{\nu+d(y)^2}{2}\right)$. Utilizando o Lema 6, temos que:

$$f(y) = \frac{2\,\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\sqrt{\pi\nu}\sigma}\left(1+\frac{d(y)^2}{\nu}\right)^{-\frac{\nu+1}{2}} \mathcal{T}_1\left(\lambda\, d(y)\sqrt{\frac{\nu+1}{\nu+d(y)^2}}\,;\,\nu+1\right).$$

A partir da parte (b) da Proposição 6, a cdf da skew-t fica:

$$F(y) = 2\mathbb{E}_U\left[\Phi_2\left(\boldsymbol{y}(U)^*\,;\,\boldsymbol{\mu}^*, \boldsymbol{\Sigma}\right)\right],$$

onde $U \sim Gamma(\nu/2, \nu/2)$, $\boldsymbol{y}(u)^*$ $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são definidas na parte (b) da Proposição (6). Desta forma, pelo Lema 6, temos que:

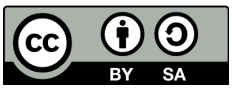
$$F(y) = 2\,\mathcal{T}_2\left(\begin{pmatrix} y \\ 0 \end{pmatrix}\,;\,\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \nu\right).$$

Apêndice B

Licença

Copyright (c) 2015 de Monique Bettio Massuia.

Exceto quando indicado o contrário, esta obra está licenciada sob a licença Creative Commons Atribuição-CompartilhaIgual 3.0 Não Adaptada. Para ver uma cópia desta licença, visite `http://creativecommons.org/licenses/by-sa/3.0/`.



A marca e o logotipo da UNICAMP são propriedade da Universidade Estadual de Campinas. Maiores informações sobre encontram-se disponíveis em `http://www.unicamp.br/unicamp/a-unicamp/logotipo/normas%20oficiais-para-uso-do-logotipo`.

B.1 Sobre a licença dessa obra

A licença Creative Commons Atribuição-CompartilhaIgual 3.0 Não Adaptada utilizada nessa obra diz que:

1. Você tem a liberdade de:

    - Compartilhar — copiar, distribuir e transmitir a obra;
    - Remixar — criar obras derivadas;

- fazer uso comercial da obra.

2. Sob as seguintes condições:

   - Atribuição — Você deve creditar a obra da forma especificada pelo autor ou licenciante (mas não de maneira que sugira que estes concedem qualquer aval a você ou ao seu uso da obra).
   - Compartilhamento pela mesma licença — Se você alterar, transformar ou criar em cima desta obra, você poderá distribuir a obra resultante apenas sob a mesma licença, ou sob uma licença similar à presente.