UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Diagnóstico de influência em modelos de regressão para dados censurados utilizando distribuições de caudas pesadas

Isabel Cristina Gomes Moura
Orientadora: Professora Lourdes Coral Contreras Montenegro
Coorientador: Professor Víctor Hugo Lachos Dávila

Belo Horizonte
Março, 2016

Isabel Cristina Gomes Moura

Diagnóstico de influência em modelos de regressão para dados censurados utilizando distribuições de caudas pesadas

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais como parte dos requisitos para a obtenção do grau de Doutor em Estatística.

Orientadora: Professora Lourdes Coral Contreras Montenegro
Coorientador: Professor Víctor Hugo Lachos Dávila

Belo Horizonte
Março, 2016

Para minha amiga Emília Sakurai†.
Carinho, gratidão e saudades eternas!

"Pode uma mulher esquecer-se daquele que amamenta?
Não ter ternura pelo fruto de suas entranhas?
E mesmo que ela o esquecesse,
eu não te esqueceria nunca."
Isaías 49, 15.

"Eis por que sinto alegria nas fraquezas, nas afrontas, nas necessidades,
nas perseguições, no profundo desgosto sofrido por amor de Cristo.
Porque quando me sinto fraco, então é que sou forte."
II Coríntios 12, 10.

Agradecimentos

Agradeço a Deus por iluminar meus caminhos, minhas escolhas e colocar tantas pessoas especiais em minha vida.

Aos meus pais, Abel e Delma, por sempre me apoiarem na jornada dos estudos e por serem meus exemplos de amor. Ao meu irmão Anderson pelo carinho, apoio, amor e palavras sábias nos momentos certos.

Ao meu esposo Jonatas por viver comigo este amor que dá sabor à vida. Depois que o conheci o fardo do doutorado ficou mais leve. "*I was born to love you*".

À amiga Emília Sakurai[†] pelo carinho, doces palavras, apoio e por ter acreditado em mim mais do que eu mesma. Grande exemplo de ser humano.

À minha orientadora, Professora Lourdes Montenegro, por sua paciência e sabedoria em me conduzir, e por compreender os momentos difíceis pelo qual passei. Ao meu coorientador, Professor Víctor Hugo Lachos Dávila, por suas preciosas contribuições e pronto apoio. Ao Professor Aldo Garay, sempre muito atencioso quando foi solicitado e ao Professor Marcos Prates por suas contribuições, seu tempo e suas palavras no dia da defesa.

Aos demais membros da banca examinadora, Professores Fábio Nogueira Demarqui, Vinicius Diniz Mayrink e Camila Borelli Zeller, por suas correções e sugestões para a tese.

À Professora Glaura Franco, coordenadora da pós-graduação, por sua compreensão e paciência. Às funcionárias Rogéria, Rose, Cristina, Ana Maria e Maysa pela simpatia e presteza.

Aos meus amigos de doutorado Cristiano, Rívert, Zezinho e Fábio, pela companhia, amizade e momentos de aprendizado. Às amigas de pós-graduação Graziele, Márcia, Letícia, Vanessa e Aline, pelos bons momentos de descontração e amizade.

Aos amigos de toda a vida Fabrícia, Isabella, Deisi, Elaine (minhas damas!), Alysson, Allan, Juliana, Aline Viana e Marcelo, simplesmente por fazerem parte da minha vida. À Alessandra pelo carinho e confiança. Ao Grupo de Pesquisa em Economia da Saúde, por todo o aprendizado e paciência. Ao grupo PNAUM e à Juliana Álvares, pela oportunidade e confiança em meu trabalho.

À Faculdade Ciências Médicas por me permitir fazer o que eu amo, em especial ao Professor Antônio Vieira Machado, Professora Maria da Glória Rodrigues Machado e Professor Eduardo Back Sternick, pela compreensão, confiança, palavras de incentivo e por acreditarem em meu trabalho.

Ao Helian Nunes, pela doçura, competência, compreensão e amizade. À Guiomar pela eficiência, competência e carinho.

Ao pequeno Maurício Diogo por me ensinar que nos momentos de felicidade o tempo pára. À Dona Agrecina por me ensinar que espontaneidade não é um defeito.

A todos os professores que fizeram parte da minha jornada acadêmica, desde o jardim de infância até hoje, e me fizeram me apaixonar pelo aprendizado. A todas as pessoas que um dia me dirigiram um sorriso sincero.

A todas as pessoas que citei acima, saibam que eu não teria conseguido terminar este doutorado sem a participação de vocês.

Muito obrigada!

Resumo

A ampla utilização dos modelos de regressão para descrever fenômenos em diversas áreas do conhecimento motivam as pesquisas estatísticas a aperfeiçoar a formulação destas técnicas. Uma etapa importante da modelagem que tem recebido atenção especial na literatura estatística é a análise de influência, definida como o estudo da dependência dos resultados fornecidos pelo modelo a pequenas perturbações em sua elaboração. O objetivo deste trabalho é construir medidas de influência global e local, considerando variável resposta censurada, para os modelos regressão linear e não linear utilizando distribuições da família Normal/Independente, e para modelos de regressão linear para dados longitudinais utilizando distribuição t de Student multivariada e estrutura de correlação *damped exponential*. Especificamente o foco é comparar os resultados obtidos na análise de influência feita via modelo Normal com os obtidos utilizando-se as distribuições de caudas pesadas. Os resultados obtidos via estudos de simulação e aplicações mostraram que os modelos de caudas pesadas são menos influenciados por observações discrepantes que o modelo Normal. Os achados deste estudo comprovam que além de gerarem resultados mais robustos na estimação, os modelos de caudas pesadas fornecem resultados mais estáveis, na presença de observações atípicas, que o modelo Normal.

Palavras-chave: Dados censurados, Diagnóstico de influência, Distribuições Normais/Independentes, Modelos de regressão.

Abstract

The wide use of regression models in various fields of knowledge motivate statistical research to improve the development of these techniques. An important stage of modeling that has received attention in the statistical literature is the influence analysis, defined as the study of the dependence of the results provided by the model to small perturbations in their formulation. The objective of this thesis is to build global and local influence measures, considering a censored response variable, for linear and nonlinear regression models using the Normal/Independent family of distributions, and to linear regression models for longitudinal data using the multivariate Student-t distribution and the damped exponential correlation structure. Specifically, the focus is to compare the results obtained from the analysis of the influence made by the Normal model with those obtained using the heavy-tailed distributions. The results obtained through simulation studies and applications have shown the heavy-tailed models are less influenced by outliers than the Normal model. The findings of this study show that besides generating more robust results in the estimation, the heavy-tailed models provides more stable results in the presence of atypical observations than the Normal model.

Keywords: Censored data, Influence diagnostics, Normal/Independent distributions, Regression models.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A análise de regressão é uma das ferramentas estatísticas mais utilizadas para descrever a relação entre variáveis. A finalidade deste método é estimar como as variações em uma ou mais variáveis preditoras afetam a variável resposta. As primeiras notas de desenvolvimento formal deste método datam do final do século XIX e início do século XX, com os trabalhos de Francis Galton (Galton, 1894) e Karl Pearson (Pearson, 1896, 1922, 1930). A investigação de Galton foi motivada por um problema biológico, coincidentemente ou não, ele era primo do famoso Charles Darwin (Stanton, 2001).

A suposição de normalidade dos erros aleatórios em modelos de regressão é bastante utilizada na literatura. A modelagem para este tipo de situação é facilitada por todo o desenvolvimento matemático já consolidado em relação à distribuição Normal e está disponível em grande parte dos *softwares* estatísticos. Entretanto, esta suposição pode ser inadequada para conjuntos de dados susceptíveis a observações atípicas. Uma alternativa para estes casos seria a aplicação de uma transformação aos dados, como a de Box-Cox (Box e Cox, 1964), por exemplo. Este tipo de solução pode levar a perda de interpretação dos parâmetros, entre outras dificuldades, de modo que seria mais apropriado propor um modelo teórico adequado para dados com essa característica. Neste contexto, a utilização de distribuições de caudas pesadas surge como uma possibilidade. Alguns autores têm adotado as distribuições da família Normal/Independente (NI) (Lange e Sinsheimer, 1993). Esta é uma classe de distribuições simétricas que engloba a distribuição Normal e também distribuições de caudas pesadas como t de Student, Slash e Normal Contaminada, entre outras (veja Liu (1996); Rosa et al. (2003); Lachos et al. (2011)).

Apesar do frequente uso dos modelos de regressão nas diversas áreas do conhecimento científico, ainda há uma considerável distância entre a base teórica e a aplicação prática destas técnicas. Isso se deve, entre oturos motivos, à necessidade de se fazer suposições, nem sempre válidas na prática, para a correta aplicação dos

modelos. A análise de resíduos é uma das abordagens utilizadas para identificar possíveis incompatibilidades entre o modelo assumido e os dados utilizados. Os primeiros métodos propostos nesta linha são da década de 1960 (Anscombe, 1961; Srikantan, 1961; Cox e Snell, 1968). Na década de 1970 houve grande expansão da utilização destes métodos e, por volta de 1975, a análise dos resíduos já era considerada fundamental em qualquer modelo de regressão (Cook e Weisberg, 1982).

Seguindo a linha de pesquisa de métodos para se estudar discordâncias entre os modelos e a realidade surgiram as primeiras abordagens de análise de influência no final da década de 1970 (Cook, 1977; Cook e Weisberg, 1982). Estes estudos representaram uma nova forma de compreender a análise de regressão: o modelo proposto poderia apresentar alguma deficiência inerente e não tratável, que poderia torná-lo inadequado para representar o fenômeno estudado. A motivação dos estudos de influência é avaliar a estabilidade dos resultados fornecidos diante de pequenas mudanças na formulação do modelo. Para isso, são introduzidas pequenas perturbações nas definições de um modelo e avaliada a influência das mesmas sobre o resultado de uma análise.

A análise de influência apresentada neste trabalho será baseada no estudo do gráfico de influência definido a seguir (Cook, 1986). Vamos assumir que a perturbação será inserida no modelo por um vetor $\boldsymbol{\omega}$ de dimensão $q \times 1$ restrito a um aberto $\boldsymbol{\Omega} \in \mathbb{R}^q$. Existe um vetor $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ que define o modelo não perturbado como um caso particular do modelo perturbado. Sejam $\boldsymbol{\theta}$ um vetor de dimensão $p \times 1$ de parâmetros desconhecidos e $Q(\boldsymbol{\theta})$ a esperança da função de log-verossimilhança do modelo não perturbado. Sejam $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ as estimativas de máxima verossimilhança obtidas dos modelos proposto e perturbado, respectivamente. O gráfico de influência é então definido como uma representação gráfica de $\boldsymbol{\alpha}(\boldsymbol{\omega}) = \left(\boldsymbol{\omega}^\top, f_Q(\boldsymbol{\omega})\right)^\top$, em que $f_Q(\boldsymbol{\omega}) = 2\left[Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})|\widehat{\boldsymbol{\theta}})\right]$.

Para construir uma análise de influência é necessário eleger o esquema de perturbação de acordo com aspecto particular do modelo a ser perturbado. Por exemplo, pode-se perturbar cada observação do banco de dados, a variável resposta, uma variável preditora contínua, os coeficientes, o parâmetro de escala, entre outros. O próximo passo é escolher a medida de influência a ser utilizada. Há dois tipos de medida de influência: global e local, que caracterizam o comportamento do gráfico de influência sobre todo o $\boldsymbol{\Omega}$ e ao redor de um ponto particular $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$, respectivamente.

Na linha de influência global, o esquema de perturbação mais utilizado é a exclusão de um caso particular ou de um conjunto de casos de um banco de dados. A perturbação para este caso é aplicada aos dados e temos um vetor de perturbação $\boldsymbol{\omega}$ com entradas iguais a 0 para as observações excluídas e iguais a 1 para as observações remanescentes. A medida de influência utilizada foi proposta por Cook (1977)

e se baseia na diferença entre as estimativas fornecidas pelo modelo com todos os dados e com uma observação excluída utilizando a função de log-verossimilhança. Zhu et al. (2001) propuseram uma adaptação desta medida utilizando a esperança da função de log-verossimilhança, o que tornava o método aplicável a uma série de outros problemas estatísticos, como aqueles que apresentam dados faltantes ou censuras.

O uso do método de exclusão de casos é bastante comum porque permite um aprofundamento no estudo de observações atípicas (ou pontos aberrantes) identificados na análise de resíduos. Essas observações podem exercer um peso desproporcional nas estimativas, testes e outras inferências produzidas pelo modelo. A deleção de uma única observação pode, no entanto, levar ao chamado "*masking effect*", que consiste em deixar de detectar casos conjuntamente influentes. Outros trabalhos nessa linha foram feitos por Andrews e Pregibon (1978); Atkinson (1982); Johnson e Geisser (1983).

Na abordagem da influência local teremos pesos arbitrários $\boldsymbol{\omega}$ atribuídos às características que se deseja perturbar no modelo. Nesta escola as medidas de influência são obtidas estudando-se a curvatura do gráfico de influência $\boldsymbol{\alpha}(\boldsymbol{\omega})$ na direção de algum vetor unitário ao redor de um ponto particular $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$. Ao contrário do método de influência global anteriormente descrito, o desenvolvimento computacional é mais simples neste caso. Essa ideia foi proposta por Cook (1986) utilizando a função de log-verossimilhança e aprimorada por trabalhos como os de Poon e Poon (1999); Zhu e Lee (2001). Na literatura de influência local os esquemas de perturbação mais frequentes são os de ponderação de casos, no parâmetro de escala, em uma variável preditora contínua, sobre a variável resposta e sobre os coeficientes do modelo.

É importante ressaltar que essas técnicas não objetivam excluir os dados, afinal casos influentes não são necessariamente indesejáveis. Eles podem fornecer informações mais importantes que os outros casos. O foco principal destes métodos é identificar as possíveis observações influentes e investigar a proveniência das mesmas. Se forem resultados de erros de medida, de digitação ou de condições experimentais inapropriadas, devem ser corrigidas, se possível, ou excluídas. Se forem observações genuínas, pode-se propor um modelo mais robusto que acomode o efeito das mesmas.

Alguns estudos podem apresentar variável resposta com medidas incompletas. Por exemplo, pode ser o resultado de um teste diagnóstico susceptível a um limite (Vaida e Liu, 2009b; Lachos et al., 2011), ou pode ser o tempo até determinado evento em uma análise de sobrevivência (Heuchenne e Keilegom, 2007). Variáveis com estas características são chamadas de censuradas, ou seja, suas medidas são apenas parcialmente conhecidas. É importante ressaltar a diferença entre observações censuradas e truncadas, que reside na causa da perda de informação: no caso de censura, a causa deve ser aleatória, já no caso de truncamento, provavelmente

será uma limitação do desenho do estudo. Por exemplo, em uma amostra censurada, temos a informação, mesmo que incompleta, de todos os indivíduos da amostra, já em uma amostra truncada, alguns indivíduos não teriam suas medidas disponíveis. A modelagem na presença de censura requer o cuidado de incorporar este efeito de informação incompleta à análise.

Desta forma, neste trabalho temos a proposta de construir medidas de diagnóstico de influência global e local para modelos de regressão com dados censurados e observações atípicas, utilizando distribuições de caudas pesadas.

1.1 Objetivos

O objetivo principal desta tese é construir a análise de influência, pelas abordagens global e local, para modelos de regressão com respostas censuradas utilizando distribuições de caudas pesadas. O foco é a comparação dos resultados obtidos pela análise de influência utilizando a distribuição Normal com algumas distribuições de caudas pesadas. A análise de influência será feita segundo as propostas de Zhu et al. (2001); Zhu e Lee (2001) para os modelos de regressão linear e não linear utilizando distribuições da família NI, e para o modelo de regressão linear para dados longitudinais utilizando a distribuição t de Student multivariada e uma estrutura de correlação *damped exponential* (DEC) (Munoz et al., 1992). Esta formulação incorpora ao modelo a estrutura de dependência entre as observações de um mesmo indivíduo e engloba as estruturas de correlação simétrica, de modelos auto-regressivos de primeira ordem e de modelos média móvel de ordem 1, entre outros casos. Além disso, esta estrutura permite a modelagem de dados longitudinais com medidas feitas em intervalos irregulares de tempo e com dados faltantes para alguns indivíduos (Wang, 2013).

1.2 Organização da tese

Este texto é composto por duas partes: uma referente a revisão de literatura e outra composta pelos resultados obtidos por esta tese.

A Parte I - "Revisão de Literatura" é constituída de quatro Capítulos que se destinam a apresentação das distribuições de caudas pesadas (Capítulo 2); dos modelos de regressão para dados censurados e erros com distribuições de caudas pesadas (Capítulo 3); da estimação dos modelos de regressão de interesse deste trabalho (Capítulo 4) e dos métodos para se obter as medidas de diagnóstico de influência global e local (Capítulo 5).

Os resultados deste trabalho são apresentados em três Capítulos da Parte II - "Diagnóstico de influência em modelos de regressão para dados censurados e erros

seguindo distribuições de caudas pesadas". O Capítulo 6 aborda o modelo de regressão linear com distribuição NI. O Capítulo 7 trata o modelo de regressão não linear com distribuição NI e o Capítulo 8 traz o modelo de regressão para dados longitudinais, distribuição t de Student multivariada e estrutura de correlação DEC. Para os três capítulos desta parte serão apresentadas as medidas de influência global e local, bem como estudos de simulação e aplicação a dados reais.

Esta tese é finalizada com o Capítulo 9, no qual são expostas as conclusões e propostas de trabalhos futuros.

PARTE I

Revisão de Literatura

CAPÍTULO 2

Distribuições de caudas pesadas

Este capítulo se destina a definir e apresentar as distribuições de caudas pesadas de interesse desta tese. A Seção 2.1 aborda as distribuições da família NI e a Seção 2.2 a distribuição t de Student multivariada.

2.1 Distribuições Normais/Independentes

A família de distribuições NI (Lange e Sinsheimer, 1993) inclui distribuições simétricas com ou sem caudas pesadas. Esta classe de distribuições tem sido aplicada no desenvolvimento de métodos de inferência robusta, por exemplo, para modelar dados com observações atípicas.

Uma variável aleatória (va) Y pertence a família NI se é definida como a mistura de uma va positiva U, independente de uma va Z com distribuição Normal de média 0 e variância σ^2, da seguinte forma

$$Y = \mu + \frac{Z}{\sqrt{U}}, \tag{2.1}$$

em que $\mu \in \mathbb{R}$ é uma constante conhecida. A variável com distribuição NI será denotada por $Y \sim \text{NI}(\mu, \sigma^2, \nu)$, em que ν é o parâmetro que caracteriza a distribuição de U. Como consequência de (2.1), temos que $Y|U = u \sim \text{N}(\mu, u^{-1}\sigma^2)$, $\mathbb{E}(Y) = \mu$ e $\text{Var}(Y) = \mathbb{E}(U^{-1})\sigma^2$. Para obter a função de densidade de probabilidade (fdp) marginal de Y, é necessário integrar em U a densidade conjunta de Y e U. A fdp de Y é então dada por

$$f_{NI}(y) = \int_0^\infty \frac{e^{-\frac{u}{2\sigma^2}(y-\mu)^2}\sqrt{u}}{\sqrt{2\pi\sigma^2}} d\mathcal{F}_U(u), \quad y \in \mathbb{R},$$

em que $\mathcal{F}_U(u)$ é a função de distribuição acumulada (fda) de U.

A distribuição Normal é um caso particular desta família e ocorre quando $\mathcal{F}_U$

é uma distribuição degenerada em 1 ($U = 1$ com probabilidade 1). Algumas das distribuições mais utilizadas desta classe são apresentadas a seguir.

2.1.1 Distribuição Pearson Tipo VII

Seja $U \sim \text{Gama}\left(\gamma - \frac{1}{2}, \frac{1}{2}\right)$, $\gamma > 0$, em que Gama(a, b) denota a distribuição Gama com média a/b. A fdp de Y é dada por

$$f_{PTVII}(y|\mu, \sigma^2, \gamma) = \frac{\Gamma(\gamma)}{\sqrt{\pi\sigma^2}\Gamma\left(\gamma - \frac{1}{2}\right)} \left(1 + \frac{(y-\mu)^2}{\sigma^2}\right)^{-\gamma}, \; y \in \mathbb{R},$$

em que $\Gamma(.)$ é a função Gama. Como casos particulares temos a distribuição t de Student quando $\gamma = \frac{\nu+1}{2}$ e $\sigma^2 = \sigma^2\nu$ e a distribuição Normal se $\gamma \to \infty$ (Sun, 2010). A notação utilizada neste caso será $Y \sim \text{PTVII}(\mu, \sigma^2, \gamma)$, e para a distribuição t de Student, $Y \sim \text{t}(\mu, \sigma^2, \nu)$.

2.1.2 Distribuição Slash

Esta distribuição ocorre quando $U \sim \text{Beta}(\nu, 1)$, $\nu > 0$. A fdp de Y é

$$f_{SL}(y|\mu, \sigma^2, \nu) = \frac{\nu}{\sqrt{2\pi\sigma^2}} \int_0^1 u^{\nu - 1/2} exp\left\{-\frac{u}{2\sigma^2}(y-\mu)^2\right\} du, \; y \in \mathbb{R}.$$

Para esta distribuição $\mathbb{E}(Y) = \mu$ e $\text{Var}(Y) = \left(\frac{\nu}{\nu-1}\right)\sigma^2$. Quando $\nu \to \infty$ tem-se a distribuição Normal (Wang e Genton, 2006). A notação utilizada será $Y \sim \text{SL}(\mu, \sigma^2, \nu)$.

2.1.3 Distribuição Normal Contaminada

A distribuição Normal contaminada se caracteriza por uma variável de mistura discreta que pode assumir dois valores: $\lambda \in (0, 1)$ com probabilidade ν e 1 com probabilidade $1 - \nu$. A fdp tem a seguinte forma

$$f_{NC}(y|\mu, \sigma^2, \nu, \lambda) = \nu\phi(y|\mu, \sigma^2/\lambda) + (1-\nu)\phi(y|\mu, \sigma^2), \; y \in \mathbb{R},$$

em que $\phi(.|\mu, \sigma^2)$ denota a fdp da distribuição Normal com média μ e variância σ^2. O parâmetro ν pode ser interpretado com a proporção de *outliers* e o parâmetro λ como um fator de escala. Neste caso tem-se $\mathbb{E}(Y) = \mu$ e $\text{Var}(Y) = \left(\frac{\nu}{\lambda} + (1-\nu)\right)\sigma^2$. A notação utilizada para esta distribuição será $Y \sim \text{NC}(\mu, \sigma^2, \nu, \lambda)$.

2.2 Distribuição t de Student multivariada

A distribuição t de Student q-variada é definida como a distribuição do vetor aleatório $\mathbf{Y} \in \mathbb{R}^q$ definido como (Arellano-Valle e Bolfarine, 1995; Ho et al., 2012; Matos et al., 2013; Garay et al., 2014)

$$\mathbf{Y} = \boldsymbol{\mu} + \frac{\mathbf{Z}}{\sqrt{U}}, \tag{2.2}$$

em que $\boldsymbol{\mu} \in \mathbb{R}^q$ é um vetor de locação, $\mathbf{Z} \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ é independente da va $U \sim \mathrm{Gama}(\nu/2, \nu/2)$, ν é o número de graus de liberdade da distribuição e $\boldsymbol{\Sigma}_{q\times q}$ é a matriz de escala. A fdp de $\mathbf{Y}$ é dada por

$$f_{t_q}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{q+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{q/2}}\nu^{-q/2}|\boldsymbol{\Sigma}|^{-1/2}\left(1 + \frac{\kappa(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\nu}\right)^{-(q+v)/2}, \ \mathbf{y} \in \mathbb{R}^q,$$

em que $|\mathbf{A}|$ denota o determinante da matriz $\mathbf{A}$ e $\kappa(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ a distância de Mahalanobis entre $\mathbf{y}$ e $\boldsymbol{\mu}$. Se $\nu > 1$, $\boldsymbol{\mu}$ é a média e se $\nu > 2$, $\nu(\nu-2)^{-1}\boldsymbol{\Sigma}$ é a matriz de variâncias e covariâncias da distribuição. A distribuição do vetor aleatório definido em (2.2) será denotada por $\mathbf{Y} \sim \mathrm{t}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$.

A distribuição t de Student q-variada truncada é a distribuição de $\mathbf{Y}$ restrita ao hiperplano $\mathbb{A} = \{\mathbf{y}, \mathbf{a} \in \mathbb{R}^q : \mathbf{y} \leq \mathbf{a}\}$. A notação utilizada neste caso será $\mathbf{Y}|\mathbf{Y} \in \mathbb{A} \sim \mathrm{Tt}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$, com a seguinte fdp

$$f_{Tt_q}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A}) = \frac{f_{t_q}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}{\mathcal{F}_{t_q}(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}\mathbb{I}_{\{\mathbb{A}\}}(\mathbf{y}),$$

em que $\mathcal{F}_{t_q}(.|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denota a fda da distribuição t de Student q-variada de parâmetros $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ e $\mathbb{I}_{\{\mathbb{A}\}}(.)$ é a função indicadora do conjunto $\mathbb{A}$.

CAPÍTULO 3

Modelos de regressão para dados censurados utilizando distribuições de caudas pesadas

Neste capítulo são apresentados os modelos de regressão utilizados nesta tese. As Seções 3.1 e 3.2 abordam a definição dos modelos de regressão linear e não linear censurados, respectivamente, utilizando distribuições da família NI para representar o comportamento dos erros aleatórios. A Seção 3.3 se destina a apresentação do modelo de regressão linear censurado para dados longitudinais com erros seguindo a distribuição t de Student multivariada.

3.1 Modelo de regressão linear censurado NI

O modelo de regressão linear é representado pela seguinte equação (Massuia et al., 2015; Garay et al., 2015a)

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim NI(0, \sigma^2, \nu), \quad i = 1, ..., n, \tag{3.1}$$

em que Y_i é a resposta do i-ésimo indivíduo, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ é o vetor de coeficientes da regressão e $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^\top$ o vetor de variáveis preditoras do i-ésimo indivíduo. Sem perda de generalidade, o desenvolvimento apresentado neste trabalho será feito no contexto de censura à direita; extensões para os casos de censura à esquerda e intervalar são imediatas. Neste contexto, a variável resposta pode assumir os seguintes valores

$$Y_i^\star = \begin{cases} \delta_i, & \text{se } Y_i \geq \delta_i, \\ Y_i, & \text{se } Y_i < \delta_i, \end{cases}$$

em que $Y_i^\star$ é a resposta observada, Y_i é o valor real e δ_i representa um ponto de corte conhecido para $i = 1, ...n$. Este modelo será denotado por RLCNI.

3.2 Modelo de regressão não linear censurado NI

O modelo de regressão não linear é representado pela seguinte equação (Garay et al., 2016)

$$Y_i \;\; = \;\; \eta(\mathbf{X}_i^\top, \boldsymbol{\beta}) + \epsilon_i, \;\; \epsilon_i \sim NI(0, \sigma^2, \nu), \;\; i = 1, ..., n, \tag{3.2}$$

em que $\eta(\mathbf{X}_i^\top, \boldsymbol{\beta})$ é uma função não linear de $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$, duas vezes continuamente diferenciável em relação a este vetor de parâmetros, Y_i é a variável resposta e $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^\top$ é o vetor de variáveis preditoras do i-ésimo indivíduo.

Assim como no caso do modelo de regressão linear, sem perda de generalidade, será considerado o contexto de censura à direita, de modo que as observações podem assumir os seguintes valores

$$Y_i^\star = \begin{cases} \delta_i, & \text{se } Y_i \geq \delta_i, \\ Y_i, & \text{se } Y_i < \delta_i, \end{cases}$$

em que $Y_i^\star$ é a resposta observada, Y_i é o valor real e δ_i representa um ponto de corte conhecido para $i = 1, ...n$. Este modelo será denotado por RNLCNI.

3.3 Modelo de regressão linear censurado multivariado t de Student

O modelo de regressão linear para dados longitudinais é representado pela seguinte equação (Garay et al., 2014)

$$\mathbf{Y}_i \;\; = \;\; \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \;\; \boldsymbol{\epsilon}_i \sim t_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i, \nu), \;\; i = 1, ..., m, \tag{3.3}$$

em que m é o número de indivíduos, $\mathbf{Y}_i$ é o vetor de respostas, de dimensão $n_i \times 1$, referente à i-ésima unidade amostral medida nos tempos $\mathbf{t}_i = (t_{i1}, ..., t_{in_i})^\top$. $\mathbf{X}_i$ é a matriz de desenho, de dimensão $n_i \times p$, associada ao vetor de efeitos fixos $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$.

Por se tratar de dados longitudinais cada unidade amostral terá várias medidas realizadas em tempos diferentes, de modo que é preciso incorporar a dependência do tempo à formulação do modelo. Para isso definimos a matriz de dispersão $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{E}_i$, para a qual será utilizada uma estrutura de correlação *damped exponential* (DEC) (Munoz et al., 1992). Esta estrutura é bastante flexível e permite acomodar de forma

adequada o efeito de observações irregularmente medidas no tempo. A estrutura DEC é definida como

$$\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{E}_i(\boldsymbol{\phi}, \mathbf{t}_i) = \sigma^2 \left(\phi_1^{|t_{ij} - t_{ik}|^{\phi_2}}\right),\ i = 1, ..., m \text{ e } j, k = 1, ..., n_i,$$

onde, na matriz $\mathbf{E}_i$ o parâmetro ϕ_1 é interpretado como a autocorrelação entre as observações em dois pontos do tempo e o parâmetro ϕ_2 é a taxa de decaimento da função de autocorrelação. Combinações destes dois parâmetros levam a diferentes estrutura de correlação. Por exemplo, para um valor positivo de ϕ_1, se $\phi_2 = 0$, $\mathbf{E}_i$ é uma estrutura de correlação simétrica; se $0 < \phi_2 < 1$, $\mathbf{E}_i$ é uma estrutura de correlação com taxa de decaimento entre a estrutura simétrica e um modelo autoregressivo de ordem 1 (AR(1)); se $\phi_2 = 1$, $\mathbf{E}_i$ é uma estrutura de correlação do modelo AR(1); $\phi_2 > 1$, $\mathbf{E}_i$ é uma estrutura de correlação com taxa de decaimento mais rápida que a do modelo AR1 e se $\phi_2 \to \infty$, $\mathbf{E}_i$ é uma estrutura de correlação de um modelo média móvel de ordem 1 (MA(1)) (Munoz et al., 1992).

Para este estudo será assumido como espaço paramétrico para ϕ_1 e ϕ_2 $\{(\phi_1, \phi_2) : 0 < \phi_1 < 1,\ \phi_2 > 0\}$, para evitar problemas computacionais no processo de estimação da estrutura de correlação DEC (Garay, 2014).

Sem perda de generalidade, para este modelo será assumido o contexto de censura à esquerda. Vamos assumir que observamos para o i-ésimo indivíduo $(\mathbf{V}_i^\top, \mathbf{C}_i^\top)^\top$, em que $\mathbf{V}_i$ é o vetor de respostas não censuradas e $\mathbf{C}_i$ é o indicador de censura. Desta forma, tem-se $y_{ij} \leq V_{ij}$ se $C_{ij} = 1$ e $y_{ij} = V_{ij}$ se $C_{ij} = 0$ (Vaida e Liu, 2009b). Este modelo será denotado por RLCMT.

CAPÍTULO 4

Estimação dos modelos de regressão para dados censurados utilizando distribuições de caudas pesadas

Neste capítulo são apresentados os processos de estimação dos modelos de regressão para dados censurados utilizando distribuições de caudas pesadas apresentados no Capítulo 3. A Seção 4.1 define o algoritmo EM. As Seções 4.2, 4.3 e 4.4 abordam a estimação dos modelos RLCNI, RNLCNI e RLCMT, respectivamente.

4.1 O algoritmo EM

É um método de cálculo iterativo de estimativas de máxima verossimilhança de um vetor de parâmetros $\boldsymbol{\theta}$ para dados com observações incompletas, proposto por Dempster et al. (1977). O nome EM vem do fato de que cada iteração do algoritmo consiste na obtenção da esperança (passo E) seguida da maximização (passo M). A estimação dos modelos de regressão para dados censurados abordados nesta tese será feita através deste método.

Para descrever este algoritmo no contexto de dados com observações censuradas, denotemos por $\mathbf{Y}_{obs}$ o vetor de dados observados (não censurados) e $\mathbf{Y}_{cens}$ o vetor de dados censurados. A combinação destes dois vetores gera o vetor de dados completos (ou similarmente, aumentados), definido por $\mathbf{Y}_c = (\mathbf{Y}_{obs}^\top, \mathbf{Y}_{cens}^\top)^\top$. Denotamos por $f(\mathbf{Y}_c|\boldsymbol{\theta})$ e $l(\boldsymbol{\theta}|\mathbf{Y}_c)$ as funções de verossimilhança e log-verossimilhança dos dados completos. Quantidades com o sobrescrito "(r)" indicam estimativas obtidas na r-ésima iteração do algoritmo. A $(r+1)$-ésima iteração do algoritmo será composta pelo passo E, que consiste em utilizar $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{(r)}$ para calcular a função Q

$$Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)}\right) = \mathbb{E}\left[l(\boldsymbol{\theta}|\mathbf{Y}_c)|\mathbf{Y}_{obs}, \widehat{\boldsymbol{\theta}}^{(r)}\right].$$

Em seguida, no passo M se obtém $\widehat{\boldsymbol{\theta}}^{(r+1)}$ maximizando $Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)}\right)$, de tal forma que

$$Q\left(\widehat{\boldsymbol{\theta}}^{(r+1)}|\widehat{\boldsymbol{\theta}}^{(r)}\right) > Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)}\right).$$

As iterações devem ser repetidas até que a convergência seja atingida. Há vários critérios de convergência que podem ser utilizados, por exemplo, $||\widehat{\boldsymbol{\theta}}^{(r+1)} - \widehat{\boldsymbol{\theta}}^{(r)}|| < \zeta$, em que $||\mathbf{b}||$ denota a norma euclidiana do vetor $\mathbf{b}$ e $\zeta > 0$.

Apesar de produzir boas estimativas e ser adequado a uma grande variedade de problemas, a aplicação do algoritmo EM pode sofrer dificuldades nas situações em que surgem expressões sem solução analítica nos passos E, M ou em ambos. Visando solucionar estes e outros casos, várias propostas de extensões do algoritmo foram construídas. Alguns exemplos são o ECM (Meng e Rubin, 1993), que trabalha com a maximização condicional e o ECME (Liu e Rubin, 1994), que consiste em uma extensão do ECM computacionalmente mais rápida, entre outras.

Para os três modelos abordados nesta tese assumiu-se ν fixo e conhecido como alguns autores já trataram em seus tratalhos, por exemplo, Lange et al. (1989) e Meza et al. (2012). Lange et al. (1989) concluíram que há um aumento na variância do modelo quando ν é estimado, comparado ao modelo no qual ν é fixo e conhecido. Meza et al. (2012) propuseram ajustar o modelo para vários valores de ν e escolher o valor que maximiza a função de log-verossimilhança. Nesta tese os valores de ν foram obtidos utilizando-se a proposta de Meza (2012).

4.2 Estimação do modelo RLCNI

A estimação do modelo RLCNI foi desenvolvida por Garay et al. (2015a) e nesta Seção será apresentada uma breve descrição deste método.

Seja $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$ o vetor de parâmetros desconhecidos do modelo RLCNI. Vamos assumir que há c observações censuradas na amostra, de modo que o vetor de respostas observadas $\mathbf{Y}^\star$ apresenta um conjunto de c valores censurados e outro de $n-c$ valores não censurados. A função de log-verossimilhança é construída somando-se as contribuições oriundas das observações não censuradas e das observações censuradas, como se segue

$$l(\boldsymbol{\theta}|\mathbf{y}^\star) = \sum_{i=1}^{c} log\left[\mathcal{F}_{NI}\left(\frac{\mathbf{X}_i^\top\boldsymbol{\beta} - \delta_i}{\sigma}\middle|0,1,\nu\right)\right] + \sum_{i=c+1}^{n} log\left[f_{NI}\left(y_i|\mathbf{X}_i^\top\boldsymbol{\beta},\sigma^2,\nu\right)\right],$$

em que $f_{NI}(b|\mu,\sigma^2,\nu)$ e $\mathcal{F}_{NI}(b|\mu,\sigma^2,\nu)$ denotam a fdp e a fda de uma variável NI de parâmetros μ, σ^2, ν, aplicadas no ponto b.

A estimação do vetor de parâmetros é construída via algoritmo EM. Neste con-

texto podemos considerar as observações censuradas são realizações de uma variável latente $\mathbf{Y}_L$ com distribuição NI de parâmetros $\mathbf{X}_i^\top\boldsymbol{\beta}$, σ^2, e ν. O vetor de dados completos é $\mathbf{Y}_c = (\mathbf{Y}^{\star\top}, \mathbf{Y}_L^\top, \mathbf{U}^\top)^\top$, e podemos então reescrever a função de log-verossimilhança como

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}_c) &= log\left(\prod_{i=1}^{n}\frac{\sqrt{u_i}}{\sqrt{2\pi\sigma^2}}exp\left\{-\frac{u_i}{2\sigma^2}(y_i-\mathbf{X}_i^\top\boldsymbol{\beta})^2\right\}f_U(u_i|\nu)\right) \\ &= \frac{n}{2}\sum_{i=1}^{n}log(u_i) - \frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) + \sum_{i=1}^{n}log\, f_U(u_i|\nu) \\ &\quad - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(u_iy_i^2 - 2u_iy_i\mathbf{X}_i^\top\boldsymbol{\beta} + u_i(\mathbf{X}_i^\top\boldsymbol{\beta})^\top(\mathbf{X}_i^\top\boldsymbol{\beta})\right), \end{aligned} \tag{4.1}$$

em $f_U(u|\nu)$ é a fdp da va U. No passo E, calcula-se a função Q, definida como

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) &= \mathbb{E}_{\boldsymbol{\theta}^{(r)}}\left[l(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{y}^\star\right], \\ &= \chi - \frac{n}{2}log(\sigma^{2(r)}) - \frac{1}{2\sigma^{2(r)}}\times \sum_{i=1}^{n}\Big[\mathcal{E}_{2i}(\boldsymbol{\theta}^{(k)}) \\ &\quad -2(\mathbf{X}_i^\top\boldsymbol{\beta}^{(r)})\mathcal{E}_{1i}(\boldsymbol{\theta}^{(k)}) + (\mathbf{X}_i^\top\boldsymbol{\beta}^{(r)})^\top(\mathbf{X}_i^\top\boldsymbol{\beta}^{(r)})\mathcal{E}_{0i}(\boldsymbol{\theta}^{(k)})\Big], \end{aligned}$$

em que $\mathbb{E}_{\boldsymbol{\theta}^{(r)}}$ denota a esperança utilizando a estimativa $\boldsymbol{\theta}^{(r)}$ para $\boldsymbol{\theta}$ e m que χ é uma constante independente de $\boldsymbol{\theta}$.

A expressão acima depende do cálculo das seguintes esperanças

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(r)}) = \mathbb{E}_{\boldsymbol{\theta}^{(r)}}\left[U_iY_i^s|y_i^\star\right], \quad s=0,1,2, \quad i=1,...,n, \tag{4.2}$$

em que s são as potências assumidas pela variável Y_i na equação (4.1).

A seguir será apresentada a metodologia de obtenção das esperanças da expressão (4.2).

Esperanças condicionais - observações não censuradas

Para uma observação não censurada tem-se $Y_i^\star = Y_i \sim \text{NI}(\mathbf{X}_i^\top\boldsymbol{\beta}, \sigma^2, \nu)$ e

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(r)}) = y_i^s\mathbb{E}_{\boldsymbol{\theta}^{(r)}}(U_i|y_i).$$

As expressões para $\mathbb{E}_{\boldsymbol{\theta}^{(r)}}(U_i|y_i)$ para algumas distribuições da família NI foram obtidas Garay et al. (2015a) utilizando os resultados de Osorio et al. (2007). São apresentadas a seguir as expressões para as distribuições utilizadas neste trabalho.

Distribuição Pearson Tipo VII: Se $Y_i \sim \text{PTVII}(\mathbf{X}_i^\top \boldsymbol{\beta}, \sigma^2, \nu, \gamma)$, então

$$\mathbb{E}_{\boldsymbol{\theta}^{(r)}}(U_i|y_i) = \frac{\nu+1}{\gamma + d^2\left(\boldsymbol{\theta}^{(r)}, y_i\right)}. \tag{4.3}$$

Distribuição Slash: Para $Y_i \sim \text{SL}(\mathbf{X}_i^\top \boldsymbol{\beta}, \sigma^2, \nu)$, tem-se

$$\mathbb{E}_{\boldsymbol{\theta}^{(r)}}(U_i|y_i) = \frac{\Gamma^*\left(\nu+1,5, d^2\left(\boldsymbol{\theta}^{(r)}, y_i\right)/2\right)}{\Gamma^*\left(\nu+0,5, d^2\left(\boldsymbol{\theta}^{(r)}, y_i\right)/2\right)}. \tag{4.4}$$

em que $\Gamma^*(a,x) = \int_0^x e^{-t}\, t^{a-1}\, dt$ denota a função Gama incompleta.

Distribuição Normal Contaminada: No caso de $Y_i \sim \text{NC}(\mathbf{X}_i^\top \boldsymbol{\beta}, \sigma^2, \nu, \lambda)$, tem-se

$$\mathbb{E}_{\boldsymbol{\theta}^{(r)}}(U_i|y_i) = \frac{1-\nu+\nu\lambda^{1,5} exp\left\{0,5(1-\lambda)d^2\left(\boldsymbol{\theta}^{(r)}, y_i\right)\right\}}{1-\nu+\nu\lambda^{0,5} exp\left\{0,5(1-\lambda)d^2\left(\boldsymbol{\theta}^{(r)}, y_i\right)\right\}}. \tag{4.5}$$

Nas expressões acima, $d\left(\boldsymbol{\theta}^{(r)}, y_i\right) = \left(y_i - \mathbf{X}_i^\top \boldsymbol{\beta}^{(r)}\right)/\sigma^{(r)}$.

Esperanças condicionais - observações censuradas

Para uma observação censurada $Y_i = \delta_i$ se $Y_i \geq \delta_i$, isto é, $Y_i \in (\delta_i, \infty)$, $i = 1, ..., n$, de modo que

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(r)}) = \mathbb{E}_{\boldsymbol{\theta}^{(r)}}(U_i Y_i^s | Y_i \geq \delta_i). \tag{4.6}$$

As expressões para as esperanças condicionais em (4.6) foram desenvolvidas por Garay et al. (2015a), partindo da Proposição 1, estabelecida e provada por estes autores.

Proposição 1. *Seja $X \sim NI(0,1,\nu)$, U a variável de mistura e $\mathcal{F}_U(.)$ a fda de U. Então, para $\alpha < \beta$,*

$$\begin{aligned}
\mathbb{E}\left[U^s | X \in (\alpha,\beta)\right] &= G(\alpha,\beta)\left[\mathbb{E}_\Phi(s,\beta) - \mathbb{E}_\Phi(s,\alpha)\right]; \\
\mathbb{E}\left[U^s X | X \in (\alpha,\beta)\right] &= G(\alpha,\beta)\left[\mathbb{E}_\phi(s-1/2,\alpha) - \mathbb{E}_\phi(s-1/2,\beta)\right]; \\
\mathbb{E}\left[U^s X^2 | X \in (\alpha,\beta)\right] &= G(\alpha,\beta)\left[\mathbb{E}_\Phi(s-1,\beta) - \mathbb{E}_\Phi(s-1,\alpha)\right. \\
&\quad \left. +\alpha\mathbb{E}_\phi(s-1/2,\alpha) - \beta\mathbb{E}_\phi(s-1/2,\beta)\right],
\end{aligned}$$

em que

$$\begin{aligned} G(\alpha,\beta) &= \left(\mathcal{F}_{NI}(\beta|0,1,\nu)-\mathcal{F}_{NI}(\alpha|0,1,\nu)\right)^{-1}; \\ \mathbb{E}_{\phi}(s,k) &= \int_0^{\infty} u^s\phi(ku^{1/2})d\mathcal{F}_U(u|\nu); \\ \mathbb{E}_{\Phi}(s,k) &= \int_0^{\infty} u^s\Phi(ku^{1/2})d\mathcal{F}_U(u|\nu), \end{aligned}$$

em que $\phi(.)$ e $\Phi(.)$ denotam, respectivamente, a fdp e a fda de uma distribuição Normal Padrão.

A Proposição 1 fornece formas de calcular as esperanças condicionais das observações censuradas para distribuições NI particulares, de parâmetros $\mu = 0$ e $\sigma^2 = 1$. O Corolário 1, também apresentado por Garay et al. (2015a), apresenta expressões para calcular as esperanças condicionais das observações censuradas para o caso geral das distribuições NI, com parâmetros μ, σ^2 e ν quaisquer.

Corolário 1. *Seja $Y \sim NI(\mu,\sigma^2,\nu)$, com variável de mistura U e $\mathcal{A} = (\alpha,\beta)$. Então, para $s \geq 1$,*

$$\begin{aligned} \mathbb{E}\left[U^s|Y\in\mathcal{A}\right] &= \mathbb{E}(U^s|X\in\mathcal{A}^*); \\ \mathbb{E}\left[U^sY|Y\in\mathcal{A}\right] &= \mu\mathbb{E}(U^s|X\in\mathcal{A}^*)+\sigma\mathbb{E}(U^sX|X\in\mathcal{A}^*); \\ \mathbb{E}\left[U^sY^2|Y\in\mathcal{A}\right] &= \mu^2\mathbb{E}(U^s|X\in\mathcal{A}^*)+2\mu\sigma\mathbb{E}(U^sX|X\in\mathcal{A}^*) \\ &\quad +\sigma^2\mathbb{E}(U^sX^2|X\in\mathcal{A}^*), \end{aligned}$$

em que $X \sim NI(0,1,\nu)$ e $\mathcal{A}^* = (\alpha^*,\beta^*)$, com $\alpha^* = (\alpha-\mu)/\sigma$ e $\beta^* = (\beta-\mu)/\sigma$.

Utilizando a Proposição 1 e o Corolário 1, Garay et al. (2015a) obtiveram as fórmulas para $\mathbb{E}_{\phi}(s,k)$ e $\mathbb{E}_{\Phi}(s,k)$ para algumas distribuições da família NI. As fórmulas para as distribuições de interesse desta tese estão apresentadas a seguir.

Distribuição Pearson Tipo VII: Se $Y_i \sim \text{PTVII}(0,1,\nu,\gamma)$, então

$$\mathbb{E}_{\Phi}(s,k) = \frac{\Gamma\left(\frac{\nu+2s}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\gamma}{2}\right)^{-s}\mathcal{F}_{PTVII}(k|0,1,\nu+2s,\gamma), \tag{4.7}$$

$$\mathbb{E}_{\phi}(s,k) = \frac{\Gamma\left(\frac{\nu+2s}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi}}\left(\frac{\gamma}{2}\right)^{\frac{\nu}{2}}\left(\frac{k^2+\gamma}{2}\right)^{-\frac{\nu+2s}{2}}. \tag{4.8}$$

Distribuição Slash: Para $Y_i \sim \text{SL}(0,1,\nu)$, tem-se

$$\mathbb{E}_{\Phi}(s,k) = \left(\frac{\nu}{\nu+s}\right)\mathcal{F}_{SL}(k|0,1,\nu+s), \tag{4.9}$$

$$\mathbb{E}_{\phi}(s,k) = \frac{\nu}{\sqrt{2\pi}}\left(\frac{k^2}{2}\right)^{-(\nu+s)}\Gamma^*\left(\nu+s,\frac{k^2}{2}\right). \tag{4.10}$$

Distribuição Normal Contaminada: No caso de $Y_i \sim \text{NC}(0, 1, \nu, \lambda)$, tem-se

$$\mathbb{E}_\Phi(s,k) = \lambda^s \mathcal{F}_{CN}(k|0,1,\nu,\lambda) + (1-\lambda^s)(1-\nu)\Phi(k), \tag{4.11}$$

$$\mathbb{E}_\phi(s,k) = \nu\lambda^s \phi(k\sqrt{\lambda}) + (1-\nu)\phi(k\sqrt{\lambda}). \tag{4.12}$$

Com as expressões de (4.3) a (4.5) e de (4.7) a (4.12), o cálculo das esperanças condicionais está completo e o passo E finalizado.

No passo M deve-se maximizar $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ para se obter as estimativas da $(r+1)$-ésima iteração dos parâmetros $\boldsymbol{\beta}$ e σ^2, através das seguintes expressões

$$\begin{aligned}
\boldsymbol{\beta}^{(r+1)} &= \sum_{i=1}^{n} \mathbf{X}_i \mathcal{E}_{1i}(\boldsymbol{\theta}^{(r)}) \left[\sum_{i=1}^{n} \mathcal{E}_{0i}(\boldsymbol{\theta}^{(r)}) \mathbf{X}_i^\top \mathbf{X}_i\right]^{-1}, \text{ e} \\
\sigma^{2^{(r+1)}} &= \frac{1}{n}\sum_{i=1}^{n} \Big[\mathcal{E}_{2i}(\boldsymbol{\theta}^{(r)}) - 2\mathcal{E}_{1i}(\boldsymbol{\theta}^{(r)}) \mathbf{X}_i^\top \boldsymbol{\beta}^{(r+1)} \\
&\quad + \mathcal{E}_{0i}(\boldsymbol{\theta}^{(r)}) \left(\mathbf{X}_i^\top \boldsymbol{\beta}^{(r+1)}\right)^\top \left(\mathbf{X}_i^\top \boldsymbol{\beta}^{(r+1)}\right)\Big].
\end{aligned}$$

em que os $\mathcal{E}_{si}(\boldsymbol{\theta})$ estão definidos na equação (4.2) Para mais detalhes sobre o desenvolvimento e expressões para a estimação do modelo, veja Garay et al. (2015a). Ressalta-se que a opção pela utilização do valor de ν fixo e conhecido se deu por ser a estimação apenas um passo para a obtenção do objetivo deste trabalho.

4.3 Estimação do modelo RNLCNI

A estimação do modelo RNLCNI foi desenvolvida por Garay et al. (2016) e nesta Seção apresentaremos uma descrição sucinta deste método.

O vetor de parâmetros desconhecidos para o modelo RNLCNI será $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$. Assume-se que há c observações censuradas na variável resposta observada $\mathbf{Y}^\star$, de modo que a função de log-verossimilhança é expressa por

$$\begin{aligned}
l(\boldsymbol{\theta}|\mathbf{y}^\star) &= \sum_{i=1}^{c} log\left[\mathcal{F}_{NI}\left(\frac{\eta(\mathbf{X}_i^\top, \boldsymbol{\beta}) - \delta_i}{\sigma}\middle| 0, 1, \nu\right)\right] \\
&\quad + \sum_{i=c+1}^{n} log\left[f_{NI}\left(y_i|\eta(\mathbf{X}_i^\top, \boldsymbol{\beta}), \sigma^2, \nu\right)\right].
\end{aligned}$$

Para o desenvolvimento do algoritmo EM as observações censuradas podem ser consideradas realizações de uma variável latente $\mathbf{Y}_L \sim \text{NI}(\eta(\mathbf{X}_i^\top, \boldsymbol{\beta}), \sigma^2, \nu)$. O vetor de dados completos é então $\mathbf{Y}_c = (\mathbf{Y}^{\star\top}, \mathbf{Y}_L^\top, \mathbf{U}^\top)^\top$, de modo que a função de log-

verossimilhança dos dados completos é

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}_c) &= \frac{1}{2}\sum_{i=1}^{n} log(u_i) - \frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) + \sum_{i=1}^{n} log\, f_U(u_i|\nu) \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(u_i y_i^2 - 2u_i y_i \eta(\mathbf{X}_i^\top, \boldsymbol{\beta}) + u_i \eta(\mathbf{X}_i^\top, \boldsymbol{\beta})^\top \eta(\mathbf{X}_i^\top, \boldsymbol{\beta})\right). \end{aligned} \tag{4.13}$$

No passo E se calcula a função Q, dada por $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathbb{E}_{\boldsymbol{\theta}^{(r)}}\left[l(\boldsymbol{\theta}|\mathbf{Y}_c)|\mathbf{y}^\star\right]$, que depende das seguintes esperanças

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(r)}) = \mathbb{E}_{\boldsymbol{\theta}^{(r)}}\left[U_i Y_i^s | y_i^\star\right], \quad s = 0, 1, 2, \quad i = 1, ..., n, \tag{4.14}$$

em que s denota as potências de Y_i na equação (4.13). A função Q pode ser então expressa como

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) &= \chi - \frac{n}{2}log(\sigma^{2^{(r)}}) - \frac{1}{2\sigma^{2^{(r)}}}\sum_{i=1}^{n}\Big[\mathcal{E}_{2i}(\boldsymbol{\theta}^{(r)}) - 2\eta(\mathbf{X}_i^\top, \boldsymbol{\beta}^{(r)})\mathcal{E}_{1i}(\boldsymbol{\theta}^{(r)}) \\ &+ \eta(\mathbf{X}_i^\top, \boldsymbol{\beta}^{(r)})^\top \eta(\mathbf{X}_i^\top \boldsymbol{\beta}^{(r)})\mathcal{E}_{0i}(\boldsymbol{\theta}^{(r)})\Big], \end{aligned}$$

em que χ é uma constante que não depende de $\boldsymbol{\theta}$.

Para uma observação não censurada, $Y_i \sim \text{NI}(\eta(\mathbf{X}_i^\top, \boldsymbol{\beta}), \sigma^2, \nu)$, de modo que $\mathcal{E}_{si}(\boldsymbol{\theta}^{(r)}) = y_i^s \mathbb{E}_{\boldsymbol{\theta}^{(r)}}(U_i|y_i)$. Para calcular as esperanças condicionais (4.14) é preciso obter expressões de $\mathbb{E}_{\boldsymbol{\theta}^{(r)}}(U_i|y_i)$ para as distribuições da família NI. Para as distribuições t de Student, Slash e Normal Contaminada, basta utilizar as expressões (4.3) a (4.5), com $d\left(\boldsymbol{\theta}^{(r)}, y_i\right) = \left(y_i - \eta(\mathbf{X}_i^\top, \boldsymbol{\beta}^{(r)}),\right)/\sigma^{(r)}$.

Para uma observação censurada tem-se $Y_i = \delta_i$ se $Y_i \geq \delta_i$, isto é, $Y_i \in (\delta_i, \infty)$, $i = 1, ..., n$. As expressões para as esperanças condicionais (4.14) para as distribuições t de Student, Slash e Normal Contaminada foram obtidas por Garay et al. (2015a). Basta utilizar as fórmulas de (4.7) a (4.12).

O passo M do algoritmo consiste em maximizar $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ em relação aos parâmetros $\boldsymbol{\beta}$ e σ^2, através das seguintes expressões

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{(r+1)} &= \underset{\boldsymbol{\beta}}{\text{argmin}}(\boldsymbol{\tau}^{(r)} - \boldsymbol{\mu}^{(r)})^\top \widehat{\boldsymbol{\mathcal{E}}}_0^{(r)}(\boldsymbol{\tau}^{(r)} - \boldsymbol{\mu}^{(r)}), \\ \widehat{\sigma^2}^{(r+1)} &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{E}_{2i}(\boldsymbol{\theta}^{(r)}) - 2\mu_i^{(r)}\mathcal{E}_{1i}(\boldsymbol{\theta}^{(r)}) + \mu_i^{(r)\top}\mu_i\mathcal{E}_{0i}(\boldsymbol{\theta}^{(r)})\right], \end{aligned}$$

em que $\widehat{\boldsymbol{\mathcal{E}}}_0^{(r)} = \text{diag}(\widehat{\mathcal{E}}_{01}^{(r)}, ..., \widehat{\mathcal{E}}_{0n}^{(r)})^\top$, $\boldsymbol{\mu}^{(r)} = (\mu_1^{(r)}, ..., \mu_n^{(r)})^\top$, $\mu_i^{(r)} = \eta(\mathbf{X}_i, \boldsymbol{\beta}^{(r)})$, e $\boldsymbol{\tau}^{(r)} = \left(\tau_1^{(r)}, ..., \tau_n^{(r)}\right)^\top$ representa a resposta observada corrigida $\tau_i^{(r)} = \mathcal{E}_{1i}(\boldsymbol{\theta}^{(r)})/\mathcal{E}_{0i}(\boldsymbol{\theta}^{(r)})$. Para mais detalhes e cálculos, veja Garay et al. (2016). Assim como no caso linear, foi adotado ν fixo e conhecido.

4.4 Estimação do modelo RLCMT

A estimação do modelo RLCMT foi proposta por Garay et al. (2014) e neste texto iremos apresentar uma breve descrição deste processo.

O vetor de parâmetros desconhecidos para este modelo é $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\phi}^\top)^\top$, com $\boldsymbol{\phi} = (\phi_1, \phi_2)^\top$. O algoritmo ECM foi utilizado para a estimação do modelo por ser uma extensão do algoritmo EM utilizada quando a implementação do passo M é dificultada por alguma característica do modelo. No caso do modelo de interesse deste Capítulo, a estrutura de correlação DEC e a presença de censura desempenham esse papel.

Vamos assumir que existam c observações censuradas da variável resposta da i-ésima unidade amostral. Neste caso, a amostra observada $\mathbf{y}_i$ pode ser vista como $\mathbf{y}_i = (\mathbf{y}_i^{o^\top}, \mathbf{y}_i^{c^\top})^\top$, de modo que $C_{ij} = 0$ para os componentes de $\mathbf{y}_i^o$ e $C_{ij} = 1$ para os componentes de $\mathbf{y}_i^c$. A dimensão do vetor $\mathbf{y}_i^o$ é n_i^o, e a dimensão de $\mathbf{y}_i^c$ é n_i^c. Temos ainda $\mathbf{u} = (u_0, ..., u_m)$, $\mathbf{V}_i = vec(\mathbf{V}_i^o, \mathbf{V}_i^c)$ e $\mathbf{C}_i = vec(\mathbf{C}_i^o, \mathbf{C}_i^c)$, em que $vec(.)$ é a função que empilha vetores ou coluna de matrizes. A matriz de dispersão é

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_i^{oo} & \boldsymbol{\Sigma}_i^{oc} \\ \boldsymbol{\Sigma}_i^{co} & \boldsymbol{\Sigma}_i^{cc} \end{pmatrix}.$$

Garay et al. (2014) utilizaram o resultado $\mathbf{y}_i^o \sim t_{n_i^o}(\mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}, \nu)$ e $\mathbf{y}_i^c|\mathbf{y}_i^o \sim t_{n_i^c}(\boldsymbol{\mu}_i^{co}, \mathbf{S}_i^{co}, \nu + n_i^o)$, obtido por Arellano-Valle e Bolfarine (1995), em que

$$\mu_i^{co} = \mathbf{X}_i^c\boldsymbol{\beta} + \boldsymbol{\Sigma}_i^{co}\boldsymbol{\Sigma}_i^{oo-1}(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta}),\ \mathbf{S}_i^{co} = \left(\frac{\nu + \kappa(\mathbf{y}_i^o, \boldsymbol{\theta}^o)}{\nu + n_i^o}\right)\boldsymbol{\Sigma}_i^{cc.o}, \tag{4.15}$$

com $\boldsymbol{\Sigma}_i^{cc.o} = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co}\boldsymbol{\Sigma}_i^{oo-1}\boldsymbol{\Sigma}_i^{oc}$ e $\kappa(\mathbf{y}_i^o, \boldsymbol{\theta}^o) = (\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta})^\top\boldsymbol{\Sigma}_i^{oo-1}(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta})$. A função de log-verossimilhança para a i-ésima unidade amostral é dada por

$$\begin{aligned} L_i(\boldsymbol{\theta}|\mathbf{y}) &= f_{t_{n_i^c}}(\mathbf{y}_i^c \leq \mathbf{V}_i^c|\mathbf{y}_i^o = \mathbf{V}_i^o, \boldsymbol{\theta}) f_{t_{n_i^o}}(\mathbf{y}_i^o = \mathbf{V}_i^o|\boldsymbol{\theta}), \\ &= \mathcal{F}_{t_{n_i^c}}(\mathbf{V}_i^c|\boldsymbol{\mu}_i^{co}, \mathbf{S}_i^{co}, \nu + n_i^o)\ f_{t_{n_i^o}}(\mathbf{V}_i^o|\mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}, \nu). \end{aligned}$$

Para o desenvolvimento do algoritmo as observações censuradas pertencentes ao vetor $\mathbf{y}_i$ e o vetor $\mathbf{u}$ são considerados dados perdidos. O vetor de dados completos é $\mathbf{Y}_c = (\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{y}^\top, \mathbf{u}^\top)^\top$. A função de log-verossimilhança em relação ao vetor de dados completos é $l(\boldsymbol{\theta}|\mathbf{y}_c) = \sum_i^m l_i(\boldsymbol{\theta}|\mathbf{y}_c) = \sum_{i=1}^m log\left[L_i(\boldsymbol{\theta}|\mathbf{y}_c)\right]$, em que

$$\begin{aligned} l_i(\boldsymbol{\theta}|\mathbf{y}_c) &= -\frac{1}{2}\left[n_i log(\sigma^2) + log(|\mathbf{E}_i|) + \frac{u_i}{\sigma^2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top\mathbf{E}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right] \\ &\quad + log(f_U(u_i|\nu)) + \chi, \end{aligned}$$

onde χ é uma constante independente do vetor de parâmetros $\boldsymbol{\theta}$ e $f_U(u_i|\nu) \sim$

Gama$(\nu/2, \nu/2)$.

No passo E, calcula-se a função Q

$$Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)}\right) = \mathbb{E}_{\boldsymbol{\theta}^{(r)}}\left[l(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{y}\right] = \sum_{i=1}^{m} Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)}),$$

em que

$$Q_i\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)}\right) = -\frac{n_i}{2}log(\widehat{\sigma}^{2(r)}) - \frac{1}{2}log\left(|\widehat{\mathbf{E}}_i^{(r)}|\right) - \frac{1}{2\sigma^2}A_i\left(\widehat{\boldsymbol{\beta}}^{(r)}, \widehat{\boldsymbol{\phi}}^{(r)}\right),$$

com

$$\begin{aligned} A_i\left(\widehat{\boldsymbol{\beta}}^{(r)}, \widehat{\boldsymbol{\phi}}^{(r)}\right) = & \Big[tr\left(\mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}^{(k)})(\widehat{\mathbf{E}}_i^{(r)})^{-1}\right) - 2\widehat{\boldsymbol{\beta}}^{(r)}\mathbf{X}_i^{\top}(\widehat{\mathbf{E}}_i^{(r)})^{-1}\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}^{(r)}) \\ & + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}^{(r)})\widehat{\boldsymbol{\beta}}^{(r)}\mathbf{X}_i^{\top}(\widehat{\mathbf{E}}_i^{(r)})^{-1}\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(r)}\Big], \end{aligned}$$

onde $tr(\mathbf{A})$ denota o traço da matriz A e $\mathcal{E}_{ji}(\widehat{\boldsymbol{\theta}}^{(k)})$, $j = 0, 1, 2$ denotam as esperanças condicionais

$$\begin{aligned} \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}^{(r)}) &= \mathbb{E}\left[u_i|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(r)}\right], \\ \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}^{(r)}) &= \mathbb{E}\left[u_i\mathbf{Y}_i|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(r)}\right], \text{ e} \\ \mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}^{(r)}) &= \mathbb{E}\left[u_i\mathbf{Y}_i\mathbf{Y}_i^{\top}|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(r)}\right]. \end{aligned}$$

A esperanças condicionais citadas acima foram obtidas por Garay et al. (2014) considerando três situações diferentes, em relação às medidas da i-ésima unidade amostral:

i. Todas as respostas do i-ésimo indivíduo são censuradas. Neste caso tem-se expressões:

$$\begin{aligned} \mathcal{E}_{0i}(\boldsymbol{\theta}^{(r)}) &= \frac{\mathcal{F}_{t_{n_i}}\left(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i^{(r)}, \widehat{\boldsymbol{\Sigma}}_i^{*(r)}, \nu+2\right)}{\mathcal{F}_{t_{n_i}}\left(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i^{(r)}, \widehat{\boldsymbol{\Sigma}}_i^{(r)}, \nu\right)}, \\ \mathcal{E}_{1i}(\boldsymbol{\theta}^{(r)}) &= \frac{\mathcal{F}_{t_{n_i}}\left(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i^{(r)}, \widehat{\boldsymbol{\Sigma}}_i^{*(r)}, \nu+2\right)}{\mathcal{F}_{t_{n_i}}\left(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i^{(r)}, \widehat{\boldsymbol{\Sigma}}_i^{(r)}, \nu\right)}\mathbb{E}\left(\mathbf{W}_i\right), \text{ e} \\ \mathcal{E}_{2i}(\boldsymbol{\theta}^{(r)}) &= \frac{\mathcal{F}_{t_{n_i}}\left(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i^{(r)}, \widehat{\boldsymbol{\Sigma}}_i^{*(r)}, \nu+2\right)}{\mathcal{F}_{t_{n_i}}\left(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i^{(r)}, \widehat{\boldsymbol{\Sigma}}_i^{(r)}, \nu\right)}\mathbb{E}\left(\mathbf{W}_i\mathbf{W}_i^{\top}\right), \end{aligned}$$

em que $\mathbf{W}_i \sim \mathrm{Tt}_{n_i}\left(\widehat{\boldsymbol{\mu}}_i^{(r)}, \widehat{\boldsymbol{\Sigma}}_i^{*(r)}, \nu+2; \mathbb{A}_i\right)$, $\widehat{\boldsymbol{\mu}}_i^{(r)} = \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(r)}$, $\widehat{\boldsymbol{\Sigma}}_i^{*(r)} = \nu(\nu+2)^{-1}\widehat{\boldsymbol{\Sigma}}_i^{(r)}$, $\widehat{\boldsymbol{\Sigma}}_i^{(r)} = \widehat{\sigma^2}^{(r)}\widehat{\mathbf{E}}_i^{(r)}$ e $\mathbb{A}_i = \{\mathbf{W}_i, \mathbf{V}_i \in \mathbb{R}^{n_i}; \mathbf{w}_i \leq \mathbf{V}_i\}$.

ii. Todas as respostas do i-ésimo indivíduo são não censuradas:

$$\begin{aligned}
\mathcal{E}_{0i}(\boldsymbol{\theta}^{(r)}) &= \left(\frac{\nu+n_i}{\nu+\kappa(\mathbf{y}_i)}\right), \\
\mathcal{E}_{1i}(\boldsymbol{\theta}^{(r)}) &= \left(\frac{\nu+n_i}{\nu+\kappa(\mathbf{y}_i)}\right)\mathbf{y}_i, \text{ e} \\
\mathcal{E}_{2i}(\boldsymbol{\theta}^{(r)}) &= \left(\frac{\nu+n_i}{\nu+\kappa(\mathbf{y}_i)}\right)\mathbf{y}_i\mathbf{y}_i^\top,
\end{aligned}$$

em que $\kappa(\mathbf{y}_i) = (\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(r)})^\top \left(\widehat{\boldsymbol{\Sigma}}_i^{(r)}\right)^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(r)})$.

iii. O i-ésimo indivíduo tem respostas censuradas e não censuradas: Neste caso temos

$$\begin{aligned}
\mathcal{E}_{0i}(\boldsymbol{\theta}^{(r)}) &= \left(\frac{n_i^o+\nu}{\nu+\kappa(\mathbf{y}_i^o)}\right)\frac{\mathcal{F}_{t_{n_i}}\left(\mathbf{V}_i|\boldsymbol{\mu}_i^{co},\tilde{\mathbf{S}}_i^{co},\nu+n_i^o+2\right)}{\mathcal{F}_{t_{n_i}}\left(\mathbf{V}_i|\boldsymbol{\mu}_i^{co},\mathbf{S}_i^{co},\nu+n_i^o\right)}, \\
\mathcal{E}_{1i}(\boldsymbol{\theta}^{(r)}) &= vec\left(\mathbf{y}_i^o,\hat{u}_i,\widehat{\mathbf{w}}_i^c\right), \\
\mathcal{E}_{2i}(\boldsymbol{\theta}^{(r)}) &= \begin{pmatrix} \mathbf{y}_i^o\mathbf{y}_i^{o\top}\hat{u}_i & \hat{u}_i\mathbf{y}_i^o\widehat{\mathbf{w}}_i^{c\top} \\ \hat{u}_i\widehat{\mathbf{w}}_i^c\mathbf{y}_i^{o\top} & \hat{u}_i\widehat{\mathbf{w}}_i^{2c} \end{pmatrix},
\end{aligned}$$

em que $\tilde{S}_i^{co} = \left(\frac{\nu+\kappa(\mathbf{y}_i^o)}{\nu+2+n_i^o}\right)\boldsymbol{\Sigma}_i^{cc.o}$, $\widehat{\mathbf{w}}_i^c = \mathbb{E}(\mathbf{W}_i)$ e $\widehat{\mathbf{w}}_i^{2c} = \mathbb{E}\left(\mathbf{W}_i\mathbf{W}_i^\top\right)$, com $\mathbf{W}_i \sim \mathrm{Tt}_{n_i^c}\left(\boldsymbol{\mu}_i^{co},\tilde{S}_i^{co},\nu+n_i^o+2;\mathbb{A}_i^c\right)$ e $\boldsymbol{\Sigma}_i^{cc.o}$, $\boldsymbol{\mu}_i^{co}$ e $\mathbf{S}_i^{co}$ como mostrados na equação (4.15).

No passo M do algoritmo a função $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ é maximizada em relação aos parâmetros $\boldsymbol{\beta}$, σ^2 e $\boldsymbol{\phi}$, através das seguintes expressões

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(r+1)} &= \left(\sum_{i=1}^m \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}^{(r)})\mathbf{X}_i^\top(\widehat{\mathbf{E}}_i^{(r)})^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^n \mathbf{X}_i^\top(\widehat{\mathbf{E}}_i^{(r)})^{-1}\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}^{(r)}) \\
\widehat{\sigma^2}^{(r+1)} &= \frac{1}{N}\sum_{i=1}^m A_i\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(r)}),\widehat{\boldsymbol{\phi}}(\boldsymbol{\theta}^{(r)})\right), \text{ e} \\
\widehat{\boldsymbol{\phi}}^{(r+1)} &= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left\{-\frac{1}{2}\sum_{i=1}^m\left[log\left(|\widehat{\mathbf{E}}_i^{(r)}|\right)+A_i\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(r)}),\widehat{\boldsymbol{\phi}}(\boldsymbol{\theta}^{(r)})\right)\right]\right\},
\end{aligned}$$

em que $N = \sum_{i=1}^n n_i$. Para mais detalhes, veja Garay et al. (2014). A obtenção dos momentos em relação à variável aleatória $\mathbf{W}$ foram obtidos por Ho et al. (2012).

CAPÍTULO 5

Diagnóstico de influência

Neste capítulo são descritas as técnicas para a obtenção de medidas de influência utilizadas no desenvolvimento desta tese. As Seções 5.1 e 5.2 descrevem a forma de se calcular as medidas de influência global e local, respectivamente.

5.1 Influência global

Para a análise da influência global de uma observação utilizamos o método de exclusão de casos, que consiste em avaliar a diferença das estimativas fornecidas pelo modelo com todas as observações com as fornecidas pelo modelo com uma observação excluída. Quando se falar da deleção de uma observação fica subentendido que tudo o que for estabelecido vale para a deleção de um conjunto de observações. O trabalho pioneiro desta linha de pesquisa foi o de Cook (1977). Zhu et al. (2001) propuseram uma medida de exclusão de casos apropriada para modelos com variáveis latentes e/ou dados incompletos utilizando a esperança condicional da função de log-verossimilhança baseada nos dados completos do modelo. Nesta Seção descreveremos este método.

Para definirmos notações, sejam $\mathbf{y}_o$ e $\mathbf{y}_m$ os vetores de dados observados e faltantes, respectivamente. Toda quantidade com o subscrito "$[i]$" denota a original com a i-ésima observação y_i excluída. Sejam $\mathbf{y}_{c[i]}$ e $l\left(\boldsymbol{\theta}|\mathbf{y}_{c[i]}\right)$ o vetor de dados completos e a função de log-verossimilhança baseada neste vetor. Seja $Q_{[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \mathbb{E}\left[l(\widehat{\boldsymbol{\theta}}|\mathbf{y}_{c[i]})|\mathbf{y}_o\right]$ a função Q do modelo sem a i-ésima observação, $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_{[i]}$ as estimativas fornecidas para os modelos com todas as observações e com a i-ésima observação excluída, respectivamente. Se as estimativas $\widehat{\boldsymbol{\theta}}$ e $\widehat{\boldsymbol{\theta}}_{[i]}$ forem muito diferentes segundo alguma métrica, teremos evidências de que a i-ésima observação é influente.

O cálculo de $\widehat{\boldsymbol{\theta}}_{[i]}$ exige que a estimação do modelo seja feita n vezes (para uma amostra de tamanho n), o que demanda um alto custo computacional. Para evitar

este esforço, pode-se utilizar a aproximação de um passo, definida por

$$\widehat{\boldsymbol{\theta}}^{1}_{[i]} = \widehat{\boldsymbol{\theta}} + \left[-\mathcal{H}_Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})\right]^{-1} \mathcal{G}_{Q[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}), \tag{5.1}$$

em que

$$\mathcal{H}_Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \quad \text{e} \quad \mathcal{G}_{Q[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \frac{\partial Q_{[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}},$$

denotam a matriz Hessiana e o vetor gradiente da função Q, respectivamente. A utilização desta aproximação é justificada pelo seguinte Teorema proposto e demonstrado por Zhu et al. (2001).

Teorema 1. *Assumindo que* $\mathcal{H}_Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = O_p(n)$, $\mathcal{G}_{Q[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) - \mathcal{G}_Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = O_p(1)$ *e* $\mathcal{H}_{Q[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) - \mathcal{H}_Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = O_p(1)$, *tem-se que*

$$\widehat{\boldsymbol{\theta}}_{[i]} = \widehat{\boldsymbol{\theta}} + \left[-\mathcal{H}_Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})\right]^{-1} \mathcal{G}_{Q[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) + O_p(n^{-2}) = \widehat{\boldsymbol{\theta}}^{1}_{[i]} + O_p(n^{-2}).$$

Pode-se calcular a distância de Cook, generalizada para modelos com variáveis latentes e/ou dados incompletos, utilizando a expressão

$$GD_i = \left(\widehat{\boldsymbol{\theta}}_{[i]} - \widehat{\boldsymbol{\theta}}\right)^\top \left[-\mathcal{H}_Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})\right] \left(\widehat{\boldsymbol{\theta}}_{[i]} - \widehat{\boldsymbol{\theta}}\right), \; i = 1, ..., n. \tag{5.2}$$

E reescrevendo a expressão (5.2) utilizando a fórmula (5.1), tem-se

$$GD^1_i = \mathcal{G}_{[Qi]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})^\top \left[-\mathcal{H}_Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})\right]^{-1} \mathcal{G}_{Q[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}). \tag{5.3}$$

De modo que para se calcular a medida de exclusão de casos é preciso obter somente a matriz Hessiana, o vetor gradiente da função Q com a i-ésima observação excluída e o vetor de parâmetros estimados pelo modelo.

Uma observação deve ser considerada influente se GD^1_i for maior que $(p+1)/n$, em que p é o número de colunas da matriz X e n é o tamanho da amostra (Massuia et al., 2015).

5.2 Influência local

Para este tipo de análise de influência tem-se um vetor de perturbação $\boldsymbol{\omega} = (\omega_1, ..., \omega_g)^\top$ variando em um aberto $\boldsymbol{\Omega} \subset \mathbb{R}^g$, aplicado a alguma característica do modelo. A análise de influência local consiste na comparação do modelo proposto (original) e do modelo perturbado na vizinhança de um ponto particular do espaço paramétrico de $\boldsymbol{\omega}$. Esta comparação foi feita primeiramente por Cook na década de 1980 (veja Cook e Weisberg (1982); Cook (1986)). Os autores Zhu e Lee (2001)

trabalharam com o método proposto por Cook utilizando a esperança condicional da função de log-verossimilhança baseada nos dados completos, extendendo a aplicação da técnica para modelos com variáveis latentes e/ou dados faltantes. Este é o método que será utilizado neste texto, descrito a seguir.

Seja $l(\boldsymbol{\theta}, \boldsymbol{\omega}|\mathbf{y}_c)$ a função de log-verossimilhança baseada nos dados completos do modelo perturbado. Vamos assumir que existe um vetor $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ tal que $l(\boldsymbol{\theta}, \boldsymbol{\omega}_0|\mathbf{y}_c) = l(\boldsymbol{\theta}|\mathbf{Y}_c)\ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, ou seja, um vetor que define o modelo proposto como um caso particular do modelo perturbado. O gráfico de influência é definido como $\boldsymbol{\alpha}(\boldsymbol{\omega}) = \left(\boldsymbol{\omega}^\top, f_Q(\boldsymbol{\omega})\right)^\top$, em que $f_Q(\boldsymbol{\omega})$ é a função Q-deslocada

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

Na equação acima, $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ é a estimativa de $\boldsymbol{\theta}$ que maximiza a função Q do modelo perturbado, $Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})$. A curvatura normal $C_{f_Q,\mathbf{d}}$ de $\boldsymbol{\alpha}(\boldsymbol{\omega})$ em $\boldsymbol{\omega}_0$ na direção de um vetor unitário $\mathbf{d}$ será utilizada para descrever o comportamento local de $f_Q(\boldsymbol{\omega})$. Sejam

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

As quantidades apresentadas acima, quando escritas com o subscrito $\boldsymbol{\omega}_0$, são avaliadas em $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. A curvatura normal $C_{f_Q,\mathbf{d}}$ é calculada por

$$C_{f_Q,\mathbf{d}} = -2\mathbf{d}^\top \mathcal{H}_{Q,\boldsymbol{\omega}_0} \mathbf{d} = 2\mathbf{d}^\top \boldsymbol{\Delta}_{\boldsymbol{\omega}_0}^\top \left[-\mathcal{H}_{Q,\boldsymbol{\theta}}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})^{-1}\right] \boldsymbol{\Delta}_{\boldsymbol{\omega}_0} \mathbf{d}. \tag{5.4}$$

A função de curvatura normal apresentada em (5.4) pode assumir qualquer valor, de forma que fica difícil estabelecer valores de referência para julgar se uma observação é ou não influente. Baseados no trabalho de Poon e Poon (1999), Zhu e Lee (2001) utilizaram a curvatura normal conformalizada, dada por

$$B_{f_q,\mathbf{d}} = \frac{C_{f_Q,\mathbf{d}}}{\text{tr}\left[-2\mathcal{H}_{Q,\boldsymbol{\omega}_0}\right]}. \tag{5.5}$$

A função $B_{f_q,\mathbf{d}}$ é uma função um-a-um da curvatura normal $C_{f_Q,\mathbf{d}}$, assume valores entre 0 e 1 e é invariante a reparametrizações de $\boldsymbol{\omega}$ e $\boldsymbol{\theta}$. Para avaliar a influência local devemos utilizar a informação obtida da matriz $-\mathcal{H}_{Q,\boldsymbol{\omega}_0} = \boldsymbol{\Delta}_{\boldsymbol{\omega}_0}^\top \left[-\mathcal{H}_{Q,\boldsymbol{\theta}}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})^{-1}\right] \boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$. A análise desta matriz parte da decomposição espectral

$$-2\mathcal{H}_{Q,\boldsymbol{\omega}_0} = \sum_{i=1}^{g} \tau_i \mathbf{e}_i \mathbf{e}_i^T,$$

em que $\{(\tau_i, \mathbf{e}_i),\ i = 1, .., g\}$ são pares de autovalores e autovetores de $-2\mathcal{H}_{Q,\boldsymbol{\omega}_0}$,

com $\tau_1 \geq ... \geq \tau_p > \tau_{p+1} = ... = \tau_g = 0$ e autovetores $\mathbf{e}_i$, $i = 1, ..., g$.

Como $\text{tr}(-2\mathcal{H}_{Q,\boldsymbol{\omega}_0}) = \sum_{i=1}^{p} \tau_i$, pode ser visto que

$$C_{f_Q,\mathbf{u}_j} = \sum_{i=1}^{p} \tau_i \mathbf{e}_{ij}^2, \text{ e } B_{f_Q,\mathbf{u}_j} = \sum_{i=1}^{p} \tilde{\tau}_i \mathbf{e}_{ij}^2.$$

em que $\tilde{\tau}_i = \tau_i / \sum_{k=1}^{p} \tau_k$. Um autovetor $\mathbf{e}_i$ é chamado de m_0-influente se $B_{f_Q,\mathbf{e}_i} \geq m_0/p$. A soma ponderada dos autovetores m_0-influentes é dada por

$$M(m_0) = \sum_{i:\tilde{\tau}_i \geq m_0/p} \tilde{\tau}_i \mathbf{e}_i^2.$$

Zhu e Lee (2001) mostraram que $M(0)_j = B_{f_Q,\mathbf{u}_j}\ \forall\ j$ e que a média de $M(0)$ é igual a $1/g$. A análise da influência local pode então ser feita através de um gráfico de $M(0)_j,\ j = 1, ..., g$ contra os índices j.

Segundo o trabalho de Lee e Xu (2004), uma observação deve ser considerada localmente influente se $M(0)_j \geq \overline{M(0)} + \varsigma \times s(M(0))$, em que $\overline{M(0)}$ e $s(M(0))$ são a média e o desvio-padrão das medidas $M(0)$ e ς é uma constante selecionada. Zhu e Lee (2001) utilizaram $\varsigma = 2$, Russo et al. (2009) $\varsigma = 3$ e Zeller et al. (2010) $\varsigma = 4$.

A utilização deste método depende da matriz hessiana da função Q e da matriz $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$, sendo que a segunda matriz dependerá do esquema de perturbação aplicado ao modelo.

PARTE II

Diagnóstico de influência em modelos de regressão para dados censurados com erros seguindo distribuições de caudas pesadas

CAPÍTULO 6

Diagnóstico de influência em modelos de regressão linear censurados utilizando distribuições NI

Neste capítulo são apresentadas as medidas de influência global e local propostas para a análise de diagnóstico de modelos RLCNI. As medidas de influência para o modelo RLCNI estão na Seção 6.1. A Seção 6.2 mostra os resultados dos estudos de simulação sobre as medidas de diagnóstico propostas e a Seção 6.3 uma aplicação a dados reais.

6.1 Diagnóstico de influência

Nesta seção são apresentadas as medidas de influência global e local para o modelo (3.1), segundo as metodologias de Zhu et al. (2001); Zhu e Lee (2001).

6.1.1 Influência global

A análise de influência global será avaliada através da distância generalizada de Cook (GD). O método foi descrito na Seção 5.1 do Capítulo 5. A medida de influência neste caso depende do vetor gradiente da função Q, sem a i-ésima observação, cujas entradas são (Massuia et al., 2015)

$$\begin{aligned}
\mathcal{G}_{Q,\beta,[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= \frac{1}{\widehat{\sigma^2}}\sum_{i\neq j}\left[\mathbf{X}_j\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\mathbf{X}_j\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}}\right], \text{ e} \\
\mathcal{G}_{Q,\sigma^2,[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= -\frac{1}{2\widehat{\sigma^2}}\sum_{i\neq j}\left[1 - \frac{1}{\widehat{\sigma^2}}\left(\mathcal{E}_{2j}(\widehat{\boldsymbol{\theta}}) - 2\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}}\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}})\right.\right. \\
&\quad \left.\left. +\mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\left(\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}}\right)^\top\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}}\right)\right],
\end{aligned}$$

e da matriz hessiana da função Q, composta pelas seguintes entradas (Massuia et al., 2015)

$$\begin{aligned}
\mathcal{H}_{Q,\beta}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= -\frac{1}{\widehat{\sigma^2}}\sum_{j=1}^{n}\mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\mathbf{X}_j\mathbf{X}_j^\top, \\
\mathcal{H}_{Q,\sigma^2}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= \frac{1}{2\widehat{\sigma^4}}\sum_{j=1}^{n}\left[1-\frac{2}{\widehat{\sigma^2}}\left(\mathcal{E}_{2j}(\widehat{\boldsymbol{\theta}})-2\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}}\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}})\right.\right. \\
&\quad \left.\left.+\mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\left(\mathbf{X}_j^T\widehat{\boldsymbol{\beta}}\right)^\top\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}}\right)\right], \text{ e} \\
\mathcal{H}_{Q,\beta\sigma^2}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= -\frac{1}{\widehat{\sigma^4}}\sum_{j=1}^{n}\left[\mathbf{X}_j\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}})-\mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\mathbf{X}_j\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}}\right].
\end{aligned}$$

Com essas duas quantidades calculadas, deve-se substituí-las na expressão (5.3) para se obter as medidas de influência global.

6.1.2 Influência local

Seguindo o método descrito na Seção 5.2 do Capítulo 5, para se obter as medidas de influência local é preciso calcular a matriz hessiana da função Q e a matriz $\boldsymbol{\Delta_\omega}$ para cada esquema de perturbação de interesse. A matriz hessiana de Q está apresentada na Subseção 6.1.1. A seguir serão apresentadas as entradas da matriz $\boldsymbol{\Delta_\omega}$ para o modelo RLCNI sob os esquemas de perturbação ponderação de casos, sobre o parâmetro de escala, em uma variável preditora contínua e sobre os coeficientes do modelo.

Esquemas de perturbação

Nesta seção é apresentada a construção da matriz $\boldsymbol{\Delta_\omega}$ sobre os esquemas de perturbação de interesse. Para cada esquema de perturbação, as entradas da matriz correspondem a

$$\boldsymbol{\Delta}_\beta = \frac{\partial^2 Q(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\omega}^\top} \quad \text{e} \quad \boldsymbol{\Delta}_{\sigma^2} = \frac{\partial^2 Q(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2\partial\boldsymbol{\omega}^\top},$$

de modo que $\boldsymbol{\Delta_\omega} = \left(\boldsymbol{\Delta}_\beta^\top, \boldsymbol{\Delta}_{\sigma^2}^\top\right)^\top$.

Perturbação Ponderação de casos: Neste contexto os pesos são atribuídos aos valores esperados da função de log-verossimilhança dos dados completos do modelo. Para este esquema, $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)^\top$ e $\boldsymbol{\omega}_0 = (1, ..., 1)^\top = \mathbf{1}_n^\top$. A matriz $\boldsymbol{\Delta_\omega}_0$ será

formada pelos seguintes componentes (Massuia et al., 2015)

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

em que $diag(\mathbf{W})$ denota a diagonal da matriz $\mathbf{W}$, $\mathbf{A}$ é a matriz definida por $\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}\mathcal{E}_0(\widehat{\theta})^\top$, $\mathcal{E}_i(\widehat{\boldsymbol{\theta}}) = (\mathcal{E}_{i1}(\widehat{\boldsymbol{\theta}}),\ldots,\mathcal{E}_{in}(\widehat{\boldsymbol{\theta}}))^\top$, $i = 0,1,2$, $\mathbf{X}$ é a matriz de desenho e $\mathbf{B}$ é um vetor de dimensão n com entradas $B_i = \mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}} + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})(\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}})^\top(\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}})$, $i = 1,...,n$.

Este esquema de perturbação permite identificar observações com influência desproporcional sobre o processo de estimação (Osorio, 2006).

Perturbação sobre o parâmetro de escala: Neste caso perturba-se σ^2 ao substituí-lo por $\sigma^2(\omega_i) = \omega_i^{-1}\sigma^2$, $i = 1,...,n$, $\omega_i > 0\ \forall\ i$, na função Q. Sob esse esquema, $\boldsymbol{\omega} = (\omega_1,...,\omega_n)^\top$ e $\boldsymbol{\omega}_0 = \mathbf{1}_n^\top$. Deste modo, temos (Massuia et al., 2015)

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

em que $\mathbf{A}$ e $\mathbf{B}$ são as quantidades definidas no esquema de perturbação ponderação de casos apresentado acima. Este esquema de perturbação revela observações com uma influência importante sobre a estimação específica do parâmetro de escala (Osorio, 2006).

Perturbação sobre uma variável preditora: A perturbação, neste caso, é inserida em uma variável preditora contínua $\mathbf{X}_{(,t)}(\boldsymbol{\omega}) = \mathbf{X}_{(,t)} + \boldsymbol{\omega}^\top$, em que $\mathbf{X}_{(,t)} \in \mathbb{R}^n$ é a t-ésima coluna da matriz $\mathbf{X}$ e $\boldsymbol{\omega} \in \mathbb{R}^n$. Então, cada linha da matriz de desenho será $\mathbf{X}_i(\boldsymbol{\omega})^\top = (X_{i1},...,X_{it}+\omega_i,...,X_{ip}) = \mathbf{X}_i^\top + \omega_i\mathbf{c}_t^\top$, em que $\mathbf{c}_t$ denota um vetor de dimensão $p\times 1$ cuja p-ésima entrada é igual a 1 e as demais iguais a 0. Neste caso $\boldsymbol{\omega}_0 = \mathbf{0}_n^\top$. Para estudar a influência local deve-se substituir $\mathbf{X}_i$ por $\mathbf{X}_i(\boldsymbol{\omega}) = (\mathbf{X}_i + \omega_i\mathbf{c_t})^\top$ na função Q. Desta forma, tem-se

$$\boldsymbol{\Delta}_{\boldsymbol{\beta}} = \frac{1}{\widehat{\sigma^2}}\left[\mathbf{c}_t\mathcal{E}_1(\widehat{\boldsymbol{\theta}}) - 2\mathbf{c}_t\widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top diag\left(\mathcal{E}_0(\widehat{\boldsymbol{\theta}})\right) - 2\mathbf{c}_t\widehat{\boldsymbol{\beta}}^\top\mathbf{c}_t\boldsymbol{\omega}^\top\ diag\left(\mathcal{E}_0(\widehat{\boldsymbol{\theta}})\right)\right],$$

$$\begin{aligned}\boldsymbol{\Delta}_{\sigma^2} &= \frac{1}{\widehat{\sigma^4}}\left[\mathbf{c}_t^\top\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top diag\left(\mathcal{E}_0(\widehat{\boldsymbol{\theta}})\right)+\mathbf{c}_t^\top\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top\mathbf{c}_t\boldsymbol{\omega}^\top diag\left(\mathcal{E}_0(\widehat{\boldsymbol{\theta}})\right)\right.\\ &\quad\left.-\mathbf{c}_t^\top\widehat{\boldsymbol{\beta}}\mathcal{E}_1^\top(\widehat{\boldsymbol{\theta}})\right].\end{aligned}$$

Este esquema de perturbação pode mostrar observações cujos valores de uma variável preditora contínua influenciam consideravelmente a estimação dos parâmetros.

Perturbação sobre os coeficientes: A perturbação nos $\boldsymbol{\beta}$'s é inserida substituindo $\boldsymbol{\beta}$ por $\boldsymbol{\beta}(\omega)=\boldsymbol{\beta}\omega_i$, $i=1,...,n$, $\boldsymbol{\omega}\in\mathbb{R}^n$ na função Q. Tem-se, neste caso, $\boldsymbol{\omega}_0=\mathbf{1}_n^\top$. Então

$$\boldsymbol{\Delta}_{\boldsymbol{\beta}} = \frac{1}{\widehat{\sigma^2}}\left[\mathbf{X}^\top diag\left(\mathcal{E}_1(\widehat{\boldsymbol{\theta}})\right)-2\mathbf{X}^\top diag\left(\mathcal{E}_0(\widehat{\boldsymbol{\theta}})\right)\mathbf{X}\widehat{\boldsymbol{\beta}}\boldsymbol{\omega}^\top\right],\text{ e}$$

$$\boldsymbol{\Delta}_{\sigma^2} = \frac{1}{\widehat{\sigma^4}}\left[\boldsymbol{\beta}^\top\mathbf{X}^\top diag\left(\mathcal{E}_0(\widehat{\boldsymbol{\theta}})\right)\mathbf{X}\boldsymbol{\beta}\boldsymbol{\omega}^\top-\boldsymbol{\beta}^\top\mathbf{X}^\top diag\left(\mathcal{E}_1(\widehat{\boldsymbol{\theta}})\right)\right].$$

Este esquema de perturbação permite a identificação de observações que influenciam a estimação específica dos coeficientes do modelo.

6.2 Estudos de simulação

O modelo utilizado nesta seção será

$$Y_i=\beta_0+\beta_1X_i+\epsilon_i,\ ,i=1,...,n, \tag{6.1}$$

com $n=100$, $\epsilon_i\sim \text{N}(0,\sigma^2)$, $X_i\sim\text{Unif}(10,20)$, $i=1,...,100$ e $\beta_0=37$, $\beta_1=0,6$, e $\sigma^2=1$. O percentual de censura utilizado foi 20%, para os estudos II e III. Os modelos utilizados foram o Normal, t de Student ($\nu=4,947$), Slash ($\nu=1,690$) e Normal Contaminada ($\nu=(0,1;0,1)$) A escolha dos valores de ν foram feitas utilizando a proposta de Meza et al. (2012). Para gerar as observações censuradas à direita utilizou-se a proposta de Tsuyuguchi (2012). O modelo (6.1) foi gerado e o ponto de corte δ_i foi definido como o r-ésimo valor do vetor $\mathbf{Y}$ ordenado, ou seja, $\delta_i=Y_{(r)},\ \forall i=1,...,r$ em que $r=n-n\times pc$ é o número de observações censuradas e pc é o percentual de censura. Desta forma, $Y_{(r)}$ se torna uma observação censurada e todos os valores de $\mathbf{Y}$ maiores ou iguais a ele assumem o mesmo valor que $Y_{(r)}$.

6.2.1 Estudo I: Avaliação do efeito do percentual de censura sobre as medidas de diagnóstico

Um estudo de Monte Carlo com 1.000 réplicas do modelo (6.1) foi realizado, considerando-se as distribuições Normal, t de Student (T), Slash (SL) e Normal Contaminada (NC) na estimação. Foram obtidos os valores médios e os desvios-padrão de Monte Carlo para as medidas de influência dadas pela distância de Cook e pelos esquemas de perturbação ponderação de casos, sobre o parâmetro de escala, sobre uma variável preditora contínua e sobre os coeficientes do modelo. O objetivo deste estudo foi verificar o impacto do nível de censura sobre as medidas de diagnóstico. Os percentuais de censura avaliados foram 10%, 20% e 40%. Os resultados estão apresentados na Tabela 6.1. As médias das medidas de influência local são todas iguais a 0,01, por definição (considerando que a média é $1/g$, em que g é a dimensão do vetor de perturbação). Para os esquemas de perturbação ponderação de casos e sobre o parâmetro de escala, observa-se um aumento dos desvios-padrão das medidas de influência com o aumento do percentual de censura. Em relação à distância de Cook (GD) e aos esquemas de perturbação sobre uma variável preditora e sobre os coeficientes, os desvios-padrão e as médias (no caso de GD) são razoavelmente estáveis à mudança do percentual de censura, em geral alterando-se somente a partir da terceira casa decimal. Diante do comportamento estável das medidas de influência, optamos por utilizar o nível de 20% de censura nos próximos dois estudos de simulação.

No estudo de Garay et al. (2015a) foi realizado um estudo de Monte Carlo para verificar a influência dos percentuais de censura sobre os erros padrão das estimativas fornecidas pelo modelo de regressão linear censurado, e eles concluíram que os erros padrão não dependiam do percentual de censura. Como utilizamos a estimação proposta por este estudo, os resultados apresentados nesta seção podem ser corroborados por este achado do estudo de referência.

Tabela 6.1: Estudo de simulação I. Estatísticas das medidas de influência segundo as distribuições, percentuais de censura e esquemas de perturbação distância generalizada de Cook (GD), ponderação de casos (PC), escala (ES), variável preditora (VP) e coeficientes (CO).

Modelos	Esquemas	10% censura Média (DP[a])	20% censura Média (DP)	40% censura Média (DP)
Normal	GD	0,0276 (0,0505)	0,0257 (0,0541)	0,0203 (0,0565)
	PC	0,0100 (0,0181)	0,0100 (0,0196)	0,0100 (0,0230)
	ES	0,0100 (0,0191)	0,0100 (0,0197)	0,0100 (0,0221)
	VP	0,0100 ($8{,}7e^{-4}$)	0,0100 ($8{,}0e^{-4}$)	0,0100 ($8{,}8e^{-4}$)
	CO	0,0100 (0,0072)	0,0100 (0,0061)	0,0100 (0,0061)
t de Student	GD	0,0215 (0,0246)	0,0211 (0,0249)	0,0205 (0,0240)
	PC	0,0100 (0,0110)	0,0100 (0,0120)	0,0100 (0,0156)
	ES	0,0100 (0,0127)	0,0100 (0,0132)	0,0100 (0,0155)
	VP	0,0100 (0,0032)	0,0100 (0,0031)	0,0100 (0,0027)
	CO	0,0100 (0,0086)	0,0100 (0,0081)	0,0100 (0,0074)
Slash	GD	0,0253 (0,0372)	0,0239 (0,0365)	0,0282 (0,0357)
	PC	0,0100 (0,0146)	0,0100 (0,0144)	0,0100 (0,0185)
	ES	0,0100 (0,0162)	0,0100 (0,0167)	0,0100 (0,0189)
	VP	0,0100 (0,0014)	0,0100 (0,0014)	0,0100 (0,0012)
	CO	0,0100 (0,0075)	0,0100 (0,0071)	0,0100 (0,0065)
Normal Contaminada	GD	0,0261 (0,0410)	0,0257 (0,0495)	0,0285 (0,0562)
	PC	0,0100 (0,0154)	0,0100 (0,0165)	0,0100 (0,0193)
	ES	0,0100 (0,0170)	0,0100 (0,0173)	0,0100 (0,0192)
	VP	0,0100 (0,0012)	0,0100 (0,0011)	0,0100 (0,0011)
	CO	0,0100 (0,0073)	0,0100 (0,0070)	0,0100 (0,0074)

[a] DP é o desvio-padrão de Monte Carlo.

6.2.2 Estudo II: Análise de sensibilidade de ς

Na análise de influência local, Lee e Xu (2004) propuseram a utilização de um valor de referência $M + \varsigma s$, em que M e s representam a média e desvio-padrão das medidas de influência, e ς uma constante conhecida. Zhu e Lee (2001), Russo et al. (2009) e Zeller et al. (2010) utilizaram, respectivamente, os seguintes valores para ς: 2, 3 e 4. O objetivo deste estudo é realizar uma análise da escolha do valor de ς.

Para esta análise o modelo (6.1) foi gerado utilizando as distribuições Normal, T, SL e NC.

Um estudo de Monte Carlo com 1.000 réplicas dos modelos propostos foi realizado e foram obtidos o desvio-padrão médio e os coeficientes de variação das medidas de influência local para cada cenário. Foram obtidos também os percentis médios referentes à cada valor de referência relacionados a $\varsigma = (1, 2, 3, 4)$ e os histogramas das medidas de influência para uma amostra. A Tabela 6.2 apresenta os resultados deste estudo e as Figuras 6.1 a 6.4 apresentam os histogramas das medidas de influência para uma amostra. As distribuições das medidas de influência são assimétricas à direita para os esquemas de perturbação ponderação de casos e sobre o parâmetro de escala, casos em que as medidas apresentam maior variabilidade, todas com coeficiente de variação (CV) maiores que 100%. As medidas resultantes dos esquemas

de perturbação sobre uma variável preditora e sobre os coeficientes apresentaram menor variabilidade.

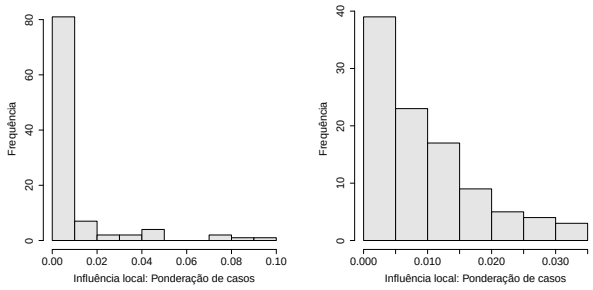
Os resultados deste estudo de simulação sugerem a utilização de $\varsigma = 2$ para detectar o grupo de aproximadamente 5% de observações maiores, para todos os esquemas de perturbação. Para se detectar o grupo de aproximadamente 1% das observações maiores, sugere-se $\varsigma = 4$ para os esquemas ponderação de casos e escala, e $\varsigma = 3$ para os esquemas variável preditora e coeficientes.

A utilização da taxa de 5% é mais sensível e pode gerar mais falsos positivos que a taxa de 1%. Em contrapartida, a taxa de 1% é mais específica, mas pode gerar mais falsos negativos. Nesta tese utilizaremos $\varsigma = 2$ (taxa de 5%), uma vez que desejamos mostrar que o modelo Normal é mais sensível que os modelos de caudas pesadas a observações atípicas.

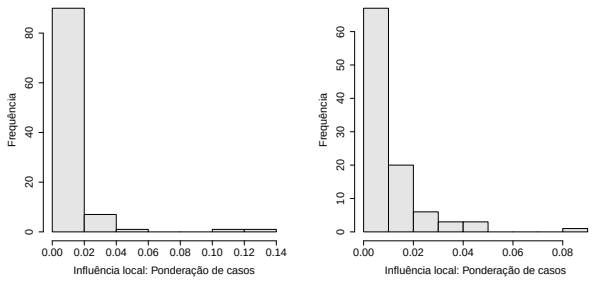
Tabela 6.2: Estudo de simulação II. Estatísticas das medidas de influência local segundo as distribuições e os esquemas de perturbação de interesse.

Modelos	Estatísticas	Ponderação de casos Estimativa (DP[a])	Escala Estimativa (DP)	Variável Preditora Estimativa (DP)	Coeficientes Estimativa (DP)
Normal	PL_1[b]	92,08 (2,52)	91,35 (2,60)	81,68 (2,01)	82,00 (1,0e^{-5})
	PL_2	95,80 (1,54)	95,66 (1,53)	97,95 (1,09)	95,00 (1,0e^{-6})
	PL_3	97,54 (0,98)	97,50 (0,98)	99,94 (0,24)	100,00 (-)
	PL_4	98,46 (0,66)	98,52 (0,64)	100,00 (-)	100,00 (-)
	DP	0,03817	0,01989	0,00084	0,00676
	CV[c]	381,7%	198,9%	8,4%	67,6%
t de Student	PL_1	87,20 (3,67)	86,55 (3,77)	84,79 (3,76)	78,73 (2,06)
	PL_2	94,27 (1,56)	94,42 (1,59)	99,89 (0,43)	97,22 (1,25)
	PL_3	97,89 (1,25)	98,10 (1,30)	100,00 (-)	99,99 (0,09)
	PL_4	99,22 (0,81)	99,29 (0,76)	100,00 (-)	100,00 (-)
	DP	0,01138	0,01305	0,00293	0,00733
	CV	113,8%	130,5%	29,3%	73,3%
Slash	PL_1	89,37 (3,40)	88,76 (3,48)	84,79 (4,59)	80,83 (1,84)
	PL_2	95,44 (1,61)	95,49 (1,58)	97,26 (2,26)	95,94 (0,83)
	PL_3	97,70 (0,92)	97,83 (0,98)	99,29 (0,99)	99,89 (0,49)
	PL_4	98,79 (0,75)	98,91 (0,76)	99,76 (0,50)	99,97 (0,18)
	DP	0,01378	0,01503	0,00654	0,00692
	CV	137,8%	150,3%	65,4%	69,2%
Normal Contaminada	PL_1	89,41 (3,03)	88,86 (3,11)	83,46 (4,75)	80,83 (2,03)
	PL_2	95,70 (1,43)	95,86 (1,34)	96,09 (1,73)	96,69 (0,98)
	PL_3	97,75 (0,86)	97,88 (0,84)	98,84 (1,01)	99,84 (0,58)
	PL_4	98,72 (0,61)	98,85 (0,62)	99,62 (0,56)	99,94 (0,29)
	DP	0,01350	0,01473	0,00969	0,00710
	CV	135,0%	147,3%	96,9%	71,0%

[a] DP é o desvio-padrão de Monte Carlo.
[b] PL_i, $i = 1, 2, 3, 4$ representam os percentis médios (em %) referentes aos limites L_i, $i = 1, 2, 3, 4$, em que $L_i = M + is$, onde i representa os valores de ς testados.
[c] CV é o coeficiente de variação, em percentual.

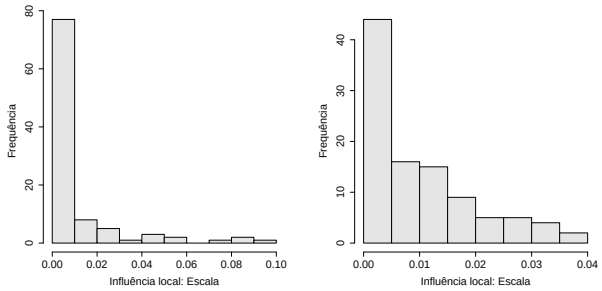


(a) Normal **(b)** t de Student

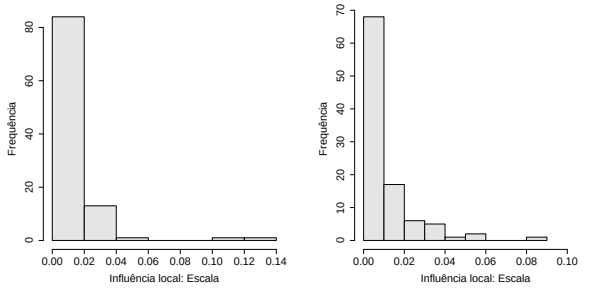


(c) Slash **(d)** Normal Contaminada

Figura 6.1: Estudo de simulação II. Medidas de influência para perturbação ponderação de casos - Modelos Normal, t de Student, Slash e Normal Contaminada.

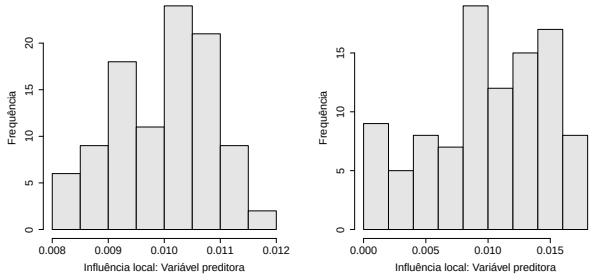


(a) Normal **(b)** t de Student

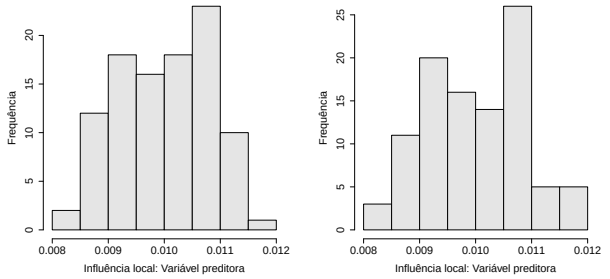


(c) Slash **(d)** Normal Contaminada

Figura 6.2: Estudo de simulação II. Medidas de influência para perturbação no parâmetro de escala - Modelos Normal, t de Student, Slash e Normal Contaminada.

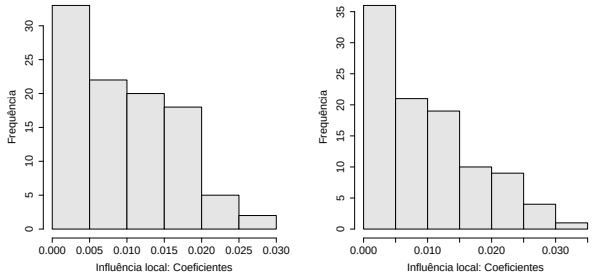


(a) Normal **(b)** t de Student

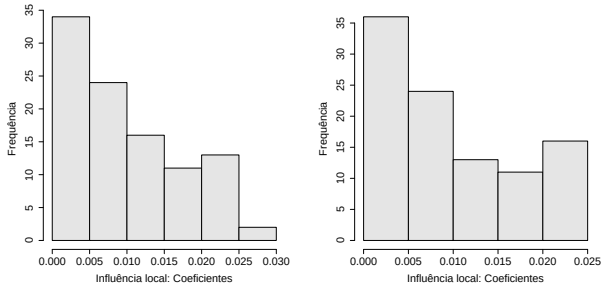


(c) Slash **(d)** Normal Contaminada

Figura 6.3: Estudo de simulação II. Medidas de influência para perturbação em uma variável preditora - Modelos Normal, t de Student, Slash e Normal Contaminada.



(a) Normal **(b)** t de Student



(c) Slash **(d)** Normal Contaminada

Figura 6.4: Estudo de simulação II. Medidas de influência para perturbação nos coeficientes - Modelos Normal, t de Student, Slash e Normal Contaminada.

6.2.3 Estudo III: Estimação das medidas de influência propostas

Neste estudo o modelo (6.1) foi gerado considerando-se $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ para $i = 2, 3, ...99$, $\epsilon_1 = -5$ e $\epsilon_{100} = 5$. A definição dos erros determinou a perturbação sobre os casos #1 e #100, e dessa forma não prejudicou a simetria da distribuição dos erros. O objetivo foi verificar se a metodologia proposta consegue identificar corretamente as observações influentes e se os modelos de caudas pesadas são menos influenciados por elas que o modelo Normal.

Um estudo de Monte Carlo com 1.000 réplicas dos modelos propostos foi realizado para avaliar o percentual de réplicas em que as observações contaminadas foram influentes, e a média e o desvio-padrão das medidas de influência.

Foi observada uma diferença considerável entre o modelo Normal e os modelos T, SL e NC para todas as medidas de diagnóstico avaliadas (Tabela 6.3). A observação #1 foi classificada como influente em várias réplicas para os esquemas de perturbação ponderação de casos e no parâmetro de escala, para todas as distribuições. Apesar disso, o valor médio das medidas de influência desta observação para os modelos T, SL e NC foram mais próximos dos valores de referência, enquanto para o modelo Normal as medidas dessa observação são bem maiores que a referência, caracterizando "salto". Esta observação foi influente no caso do esquema de perturbação em uma variável preditora apenas para o modelo Normal. Isto pode ser também observado nos gráficos das medidas de influência (Figuras de 6.5 a 6.9). Optamos por identificar nos gráficos somente as observações suspeitas (#1 e #100), uma vez que o objetivo do estudo é verificar se o método consegue identificar corretamente as observações contaminadas. Os demais casos que excederam os limites de influência podem ser considerados falsos positivos, uma vez que a utilização da constante $\varsigma = 2$ permite uma taxa de até 5% de falsos positivos.

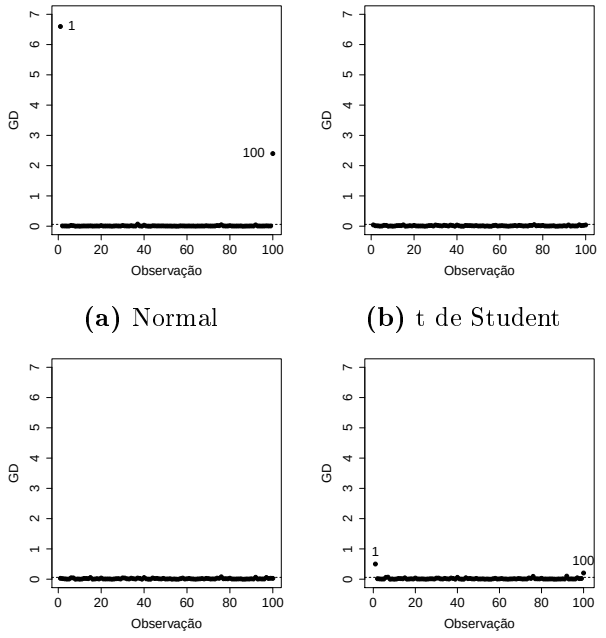
A observação #100 foi identificada como influente em poucos casos. Isso se deve ao fato de este caso representar um *outlier* à direita, sendo, portanto, censurado em grande parte das réplicas.

Os resultados deste estudo sugerem que as observações contaminadas exerceram forte influência sobre a estimação construída via modelo Normal, nas estimações global e do parâmetro de escala. No esquema de perturbação sobre uma variável preditora apenas a observação #1 foi influente para o modelo Normal e na perturbação sobre os coeficientes nenhuma observação foi considerada influente. Isso se deve ao fato de termos inserido a perturbação nos erros do modelo, e não na variável preditora. Conclui-se então que a influência exercida pelas observações contaminadas foi substancialmente menor para os modelos de caudas pesadas.

Tabela 6.3: Estudo de simulação III. Análise de influência via estudo de Monte Carlo para as observações #1 e #100 por distribuição e medida de diagnóstico: GD - distância generalizada de Cook, PC - ponderação de casos, ES - parâmetro de escala, VP - variável preditora e CO - coeficientes. Modelo linear.

Diag	Estatística	Normal		t de Student		Slash		Normal Cont.	
		#1	#100	#1	#100	#1	#100	#1	#100
GD	% Inf[a]	100,0%	100,0%	94,8%	82,1%	5,0%	0,0%	50,1%	8,0%
	M[b]	1,799	0,336	0,155	0,022	0,094	0,047	0,074	0,050
	DP[c]	(0,425)	(0,076)	(0,070)	(0,003)	(0,037)	(0,008)	(0,056)	(0,007)
	Ref[d]	0,060		0,060		0,060		0,060	
PC	% Inf	100,0%	0,0%	99,2%	0,0%	41,6%	37,2%	39,1%	1,1%
	M	0,492	0,032	0,071	0,005	0,031	0,030	0,032	0,022
	DP	(0,073)	(0,008)	(0,026)	(0,001)	(0,010)	(0,007)	(0,026)	(0,003)
	M (DP) Ref	0,109 (0,014)		0,032 (0,004)		0,032 (0,002)		0,035 (0,003)	
ES	% Inf	100,0%	0,0%	99,9%	0,0%	61,0%	46,7%	41,5%	1,0%
	M	0,485	0,003	0,077	0,019	0,037	0,034	0,038	0,005
	DP	(0,073)	($7{,}2e^{-4}$)	(0,023)	(0,003)	(0,008)	(0,007)	(0,024)	(0,001)
	M (DP) Ref	0,108 (0,014)		0,036 (0,003)		0,034 (0,002)		0,037 (0,002)	
VP	% Inf	100,0%	0,0%	0,0%	0,0%	0%	0%	0%	0%
	M	0,013	0,008	$5{,}0e^{-4}$	0,010	$2{,}8e^{-4}$	$8{,}1e^{-4}$	$4{,}0e^{-4}$	0,010
	DP	($1{,}7e^{-4}$)	($8{,}0e^{-5}$)	($3{,}4e^{-4}$)	($2{,}2e^{-4}$)	($1{,}9e^{-4}$)	($5{,}9e^{-4}$)	($3{,}3e^{-4}$)	($3{,}2e^{-4}$)
	M (DP) Ref	0,012 ($9{,}7e^{-5}$)		0,016 (0,001)		0,015 ($9{,}4e^{-4}$)		0,014 ($7{,}4e^{-4}$)	
CO	% Inf	0,0%	0,0%	0,0%	0,0%	0%	0%	0%	0%
	M	0,006	0,011	$7{,}3e^{-4}$	0,003	$4{,}3e^{-4}$	0,016	$1{,}2e^{-4}$	0,011
	DP	($2{,}3e^{-7}$)	($5{,}9e^{-7}$)	($5{,}6e^{-4}$)	(0,001)	($2{,}6e^{-4}$)	($4{,}7e^{-4}$)	($8{,}5e^{-5}$)	($6{,}8e^{-4}$)
	M (DP) Ref	0,026 ($1{,}0e^{-6}$)		0,026 ($5{,}2e^{-4}$)		0,025 ($3{,}0e^{-4}$)		0,022 ($2{,}5e^{-4}$)	

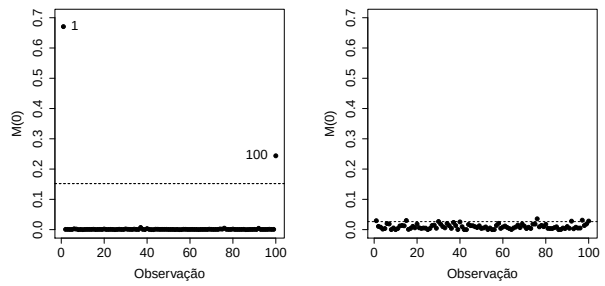
[a] % Inf: percentual de réplicas de Monte Carlo em que a observação foi considerada influente (maior que o valor de referência).
[b] M é a média das medidas de influência.
[c] DP é o desvio-padrão das medidas de influência.
[d] Ref é o valor de referência para considerar uma observação influente.



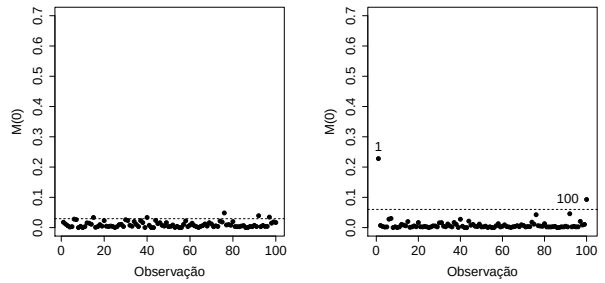
(a) Normal **(b)** t de Student

(c) Slash **(d)** Normal Contaminada

Figura 6.5: Estudo de simulação III. Distância generalizada de Cook. Modelo linear.

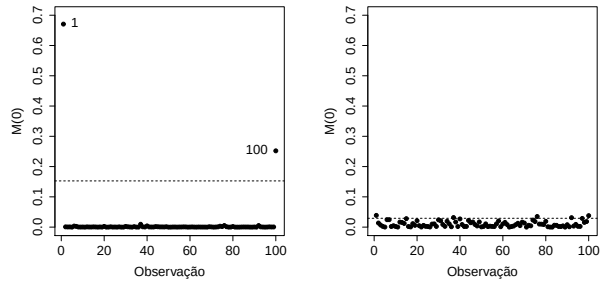


(a) Normal **(b)** t de Student

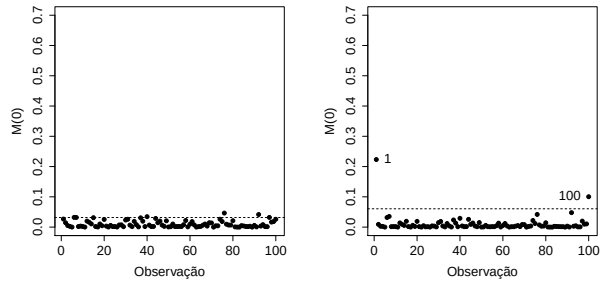


(c) Slash **(d)** Normal Contaminada

Figura 6.6: Estudo de simulação III. Perturbação ponderação de casos. Modelo linear.

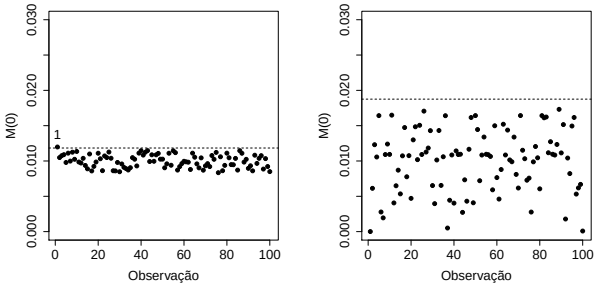


(a) Normal **(b)** t de Student

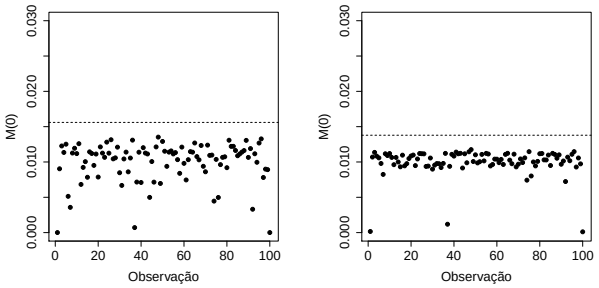


(c) Slash **(d)** Normal Contaminada

Figura 6.7: Estudo de simulação III. Perturbação sobre o parâmetro de escala. Modelo linear.

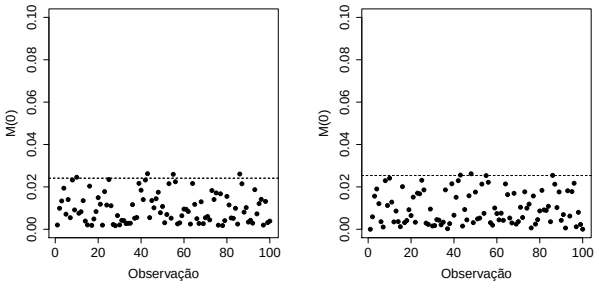


(a) Normal **(b)** t de Student

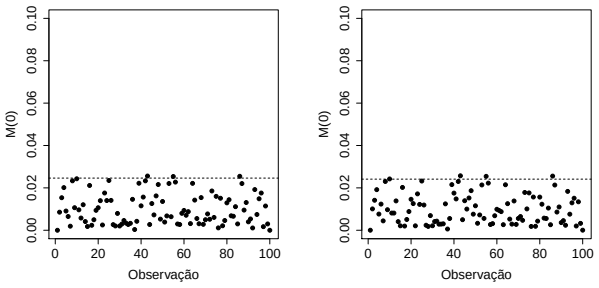


(c) Slash **(d)** Normal Contaminada

Figura 6.8: Estudo de simulação III. Perturbação sobre uma variável preditora contínua. Modelo linear.



(a) Normal **(b)** t de Student



(c) Slash **(d)** Normal Contaminada

Figura 6.9: Estudo de simulação III. Perturbação sobre os coeficientes. Modelo linear.

6.3 Aplicação

A metodologia de análise de diagnóstico proposta foi aplicada ao banco de dados "`wage.rates`", apresentado por Mroz (1987) e analisado previamente por vários autores. Garay et al. (2015a), por exemplo, trabalharam em um contexto de estimação, e Massuia et al. (2015) no contexto de diagnóstico, porém avaliando somente as distribuições Normal e t de Student.

Este banco de dados faz parte do estudo "*Panel Study of Income Dynamics*" conduzido pela Universidade de Michigan, e utilizou uma amostra de 753 mulheres brancas casadas com idades entre 30 e 60 anos, com dados referentes ao ano de 1975. A variável resposta foi o rendimento médio por hora, em dólares, destas mulheres. Como houve 325 mulheres que não trabalharam durante o período de análise, tendo rendimento médio igual a zero, consideraremos estas observações como censuradas à esquerda, uma vez que não trabalhar poderia ser interpretado como prejuízo para a família, ou seja, uma esposa com rendimento negativo. As variáveis explicativas utilizadas são a idade e o número de anos de estudo da mulher, e o número de crianças até 6 anos e de 6 a 18 anos no domicílio.

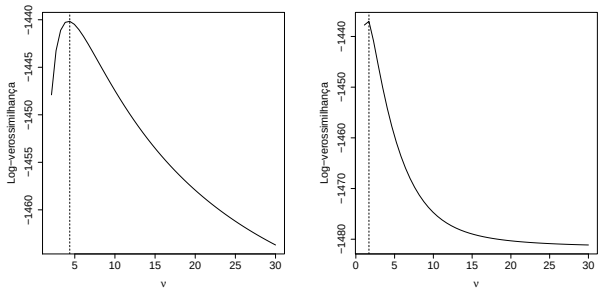
O seguinte modelo foi proposto para ajustar estes dados

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i,\ i = 1, ..., 753, \tag{6.2}$$

em que Y_i é o rendimento médio anual da i-ésima unidade amostral, as variáveis $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ e $\mathbf{X}_4$ representam a idade, o número de anos de estudo, o número de crianças até 6 anos no domicílio e o número de crianças de 6 a 18 anos, respectivamente, e $\epsilon_i \sim \text{NI}(0, \sigma^2, \nu)$. O nível de censura deste banco de dados foi 43,2%. Os modelos Normal, t de Student ($\nu = 4,377$), Slash ($\nu = 1,690$) e Normal Contaminada ($\nu = (0,1,0,1)$) foram utilizados. Os valores de ν foram obtidos através da proposta de Meza et al. (2012) (Figura 6.10 e Tabela 6.4). A estimação foi feita utilizando-se o pacote `SMNCensReg` (Garay et al., 2015b) do *software* R versão 2.2.1 (R Core Team, 2015).

Tabela 6.4: Dados "`wage.rates`". Valores de log-verossimilhanças de acordo com os valores de $\nu = (\nu_1, \nu_2)$ testados para o modelo Normal Contaminada.

	ν_2				
ν_1	0,1	0,2	0,3	0,4	0,5
0,1	-1.432,1	-1.437,9	-1.445,9	-1.454,2	-1,461,9
0,2	-1.437,6	-1.440,0	-1.446,4	-1.453,7	-1.460,9
0,3	-1.445,2	-1.444,5	-1.449,3	-1.455,5	-1.461,9
0,4	-1.453,1	-1.449,7	-1.453,2	-1.458,3	-1.463,7
0,5	-1.460,4	-1.455,2	-1.457,6	-1.461,6	-1.466,1



(a) t de Student **(b)** Slash

Figura 6.10: Dados "`wage.rates`". Valores de ν *vs* log-verossimilhança para os modelos t de Student e Slash.

A Tabela 6.5 mostra os resultados da estimação e a Figura 6.11 apresenta os gráficos de envelopes, considerando os resíduos martingal transformados (Garay et al., 2016), definidos como

$$r_{MT_i} = \text{sinal}(r_{M_i})\sqrt{-2\left[r_{M_i} + \delta_i log(\delta_i - r_{M_i})\right]},\ i = 1, ..., n, \tag{6.3}$$

em que $r_{M_i} = \delta_i + log(S(y_i, \widehat{\boldsymbol{\theta}}))$, $S(y_i, \widehat{\boldsymbol{\theta}}) = \mathbb{P}_{\widehat{\boldsymbol{\theta}}}(Y_i > y_i)$ é a função de sobrevivência e $\delta_i,\ , i = 1, ..., n$ é o indicador de censura.

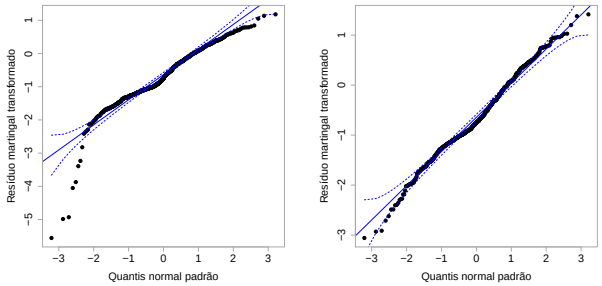
Os gráficos de envelope mostram melhor adequação dos modelos NI (t de Student, Slash e Normal Contaminada), em relação ao modelo Normal. Os modelos NI, em geral, geraram estimativas mais precisas dos coeficientes do modelo (menores valores de erros padrão), e o modelo Normal Contaminada apresentou a menor variância. Segundo os critérios de comparação de modelos (AIC, BIC e log-verossimilhança), o modelo Normal Contaminada apresentou o melhor ajuste.

Tabela 6.5: Dados "`wage.rates`". Estimativas EM dos parâmetros do modelo.

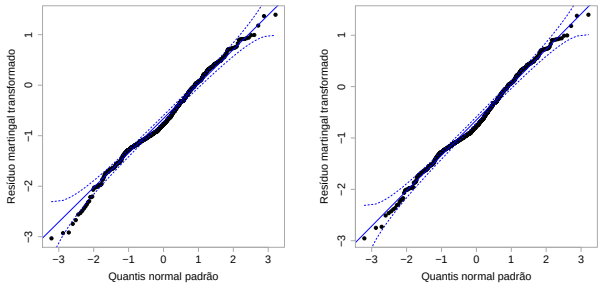
Parâmetros	Normal Estimativa (EP [a])	t de Student Estimativa (EP)	Slash Estimativa (EP)	Normal Contaminada Estimativa (EP)
β_0	-2,751 (1,890)	-1,069 (1,446)	-1,276 (1,463)	-1,290 (1,445)
β_1	-0,105 (0,028)	-0,111 (0,023)	-0,108 (0,024)	-0,106 (0,023)
β_2	0,728 (0,083)	0,648 (0,065)	0,648 (0,064)	0,647 (0,063)
β_3	-3,026 (0,420)	-3,158 (0,385)	-3,074 (0,379)	-3,065 (0,375)
β_4	-0,214 (0,149)	-0,296 (0,122)	-0,290 (0,122)	-0,300 (0,122)
σ^2	20,940(0,811)	10,771(1,012)	7,604 (0,681)	11,169 (0,994)
Variância	20,940	19,835	18,623	12,174
ν	-	4,377	1,690	(0,1; 0,1)
Log-verossimilhança	-1.481,66	-1.440,17	-1.437,02	-1.432,09
AIC	2.975,31	2.894,34	2.888,03	2.880,17
BIC	3.003,06	2.926,70	2.920,40	2.917,16

[a] EP é o erro padrão das estimativas.

A Figura 6.12 apresenta os resíduos martingal transformados, obtidos pela estimação via modelo Normal, com a identificação dos casos em que estes resíduos superaram os limites de três desvios-padrão para mais ou para menos. Foram as



(a) Normal **(b)** t de Student



(c) Slash **(d)** Normal Contaminada

Figura 6.11: Dados "`wage.rates`". Gráficos de envelopes para os resíduos martingal transformado segundo as distribuições Normal, t de Student, Slash e Normal Contaminada.

seguintes observações: #65, #74, #185, #210, #217, #349, #357, #366, #369, #394, #408, #453 e #502.

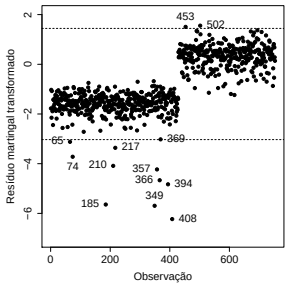


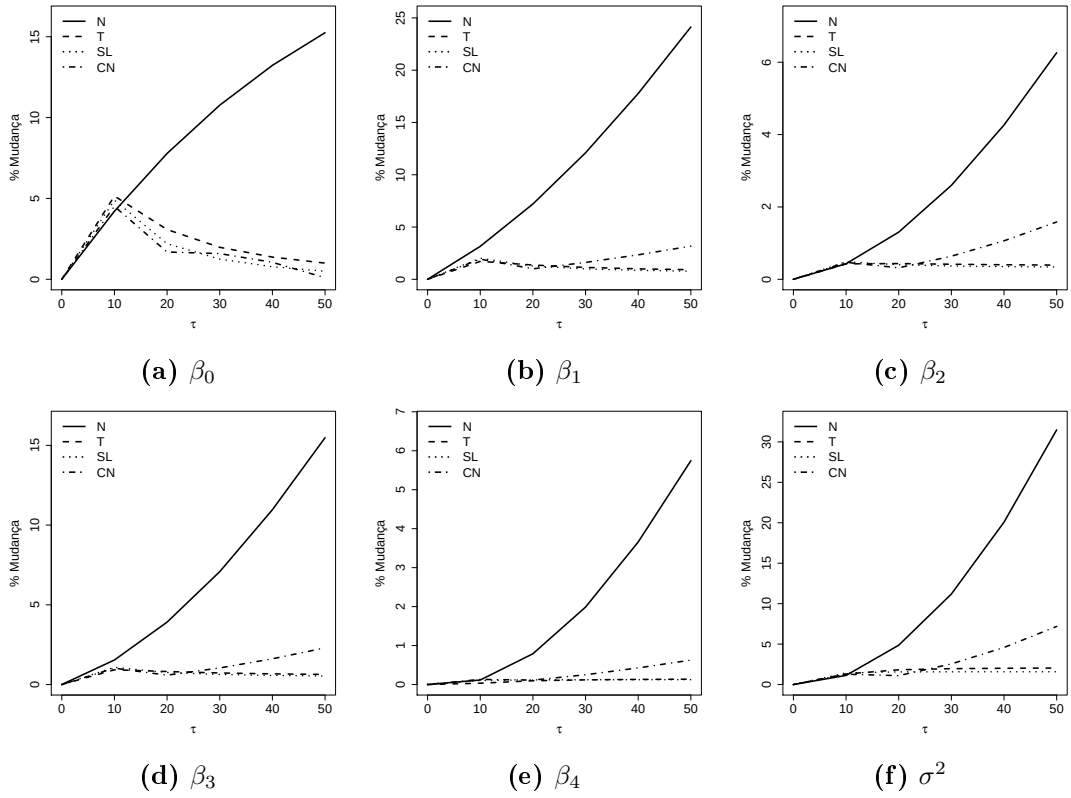
Figura 6.12: Dados "`wage.rates`". Resíduos martingal transformados para o modelo Normal.

Nota: Limites definidos como média ± 3 desvios-padrão.

A robustez das estimativas dos modelos propostos foi avaliada estudando a influência de uma observação contaminada nas estimativas geradas pelo algoritmo EM para os coeficientes $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top$ e para o σ^2. Para isso foi substituída a observação original y_{100}, escolhida arbitrariamente, sem perda de generalidade, pela observação contaminada $y_{100} + \tau$, para $\tau = (0, 10, 20, 30, 40, 50)$, e foi calculada a mudança relativa absoluta nas estimativas, através da seguinte fórmula

$$MR_i = \left| \frac{\hat{\theta}_i^c - \hat{\theta}_i}{\hat{\theta}_i} \right| \times 100, \tag{6.4}$$

em que $\hat{\theta}_i$ e $\hat{\theta}_i^c$ denotam as estimativas com o modelo proposto e contaminado, respectivamente. A Figura 6.13 apresenta os gráficos das mudanças relativas para cada um dos parâmetros do modelo (6.2), por nível de contaminação. Observa-se que os modelos NI foram menos influenciados pela contaminação, fornecendo estimativas mais estáveis de todos os parâmetros avaliados.

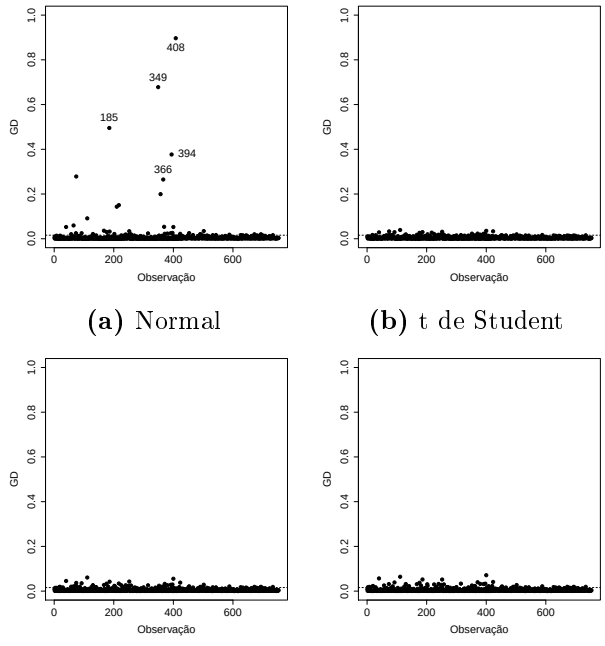


(a) β_0 **(b)** β_1 **(c)** β_2

(d) β_3 **(e)** β_4 **(f)** σ^2

Figura 6.13: Dados "`wage.rates`". Mudanças relativas nas estimativas por nível de contaminação.

As medidas de influência global e local propostas foram aplicadas ao modelo ajustado. A Figura 6.14 mostra os gráficos da distância generalizada de Cook (GD), segundo os modelos propostos e a Tabela 6.6 apresenta as médias e desvio-padrão das medidas de influência obtidas, por modelo. Pode ser observado que os modelos NI foram menos sensíveis a observações influentes que o modelo Normal, no caso do GD. Todas as observações destacadas no Gráfico 6.12 foram influentes para o modelo Normal. Algumas destas observações foram influentes para os modelos NI, porém, com medidas de influência bem próximas ao valor de referência. Houve ainda outras observações que excederam o limite para todos os modelos, devido ao fato de estarmos utilizando um critério para definir o valor de referência que permite até 5% de falsos positivos.



(a) Normal **(b)** t de Student

(c) Slash **(d)** Normal Contaminada

Figura 6.14: Dados "`wage.rates`". Distância de Cook generalizada.

Nota: Observações acima do limite de referência (além das identificadas nos gráficos).
Normal: #65, #74, #210, #217, #357, #369, #453 e #502. **t de Student**: #74, #217 e #369.
Slash: #65, #74, #217, #369 e #453. **Normal Contaminada**: #408, #453 e #502.

Tabela 6.6: Dados "`wage.rates`". Estatísticas descritivas das medidas de influência.

Medidas	Normal Média (DP[a])	t de Student Média (DP)	Slash Média (DP)	Normal Contaminada Média (DP)
Distância de Cook (GD)	0,009 (0,050)	0,005 (0,005)	0,005 (0,007)	0,005 (0,008)
Ponderação de casos	M[b] (7,343e^{-3})	M (1,499e^{-3})	M (1,757e^{-3})	M (1,891e^{-3})
Escala	M (7,228e^{-3})	M (1,509e^{-3})	M (1,752e^{-3})	M (1,871e^{-3})
Variável preditora	M (3,054e^{-3})	M (1,185e^{-3})	M (1,344e^{-3})	M (1,399e^{-3})
Coeficientes	M (8,820e^{-4})	M (9,679e^{-4})	M (9,068e^{-4})	M (9,022e^{-4})

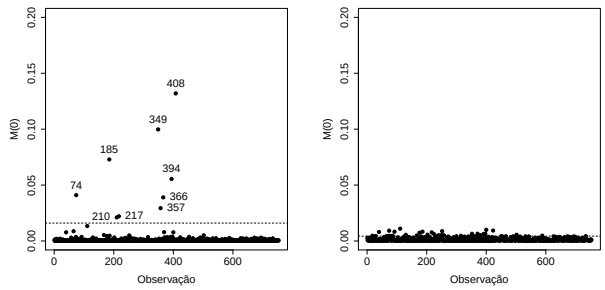
[a] DP é o desvio-padrão das estimativas.
[b] M é a média das medidas de influência local (iguais, por definição), neste caso 1,328e^{-3}.

Os gráficos das medidas de influência local por modelos estão apresentados nas Figuras de 6.15 a 6.18. O comportamento das medidas para os esquemas de perturbação ponderação de casos, sobre o parâmetro de escala e sobre uma variável preditora é parecido com aquele observado na distância de Cook. A maioria das observações suspeitas identificadas na Figura 6.12 apresentam-se como influentes considerando o modelo Normal, e algumas delas excedem os limites de referência para os modelos NI, porém, com medidas de influência próximas ao valor de referência. As medidas de influência local apresentam a mesma média, por definição, porém o desvio-padrão é menor para os modelos NI. Para o esquema de perturbação sobre os coeficientes do modelo, apenas uma das observações foi classificada como influente para o modelo Normal.

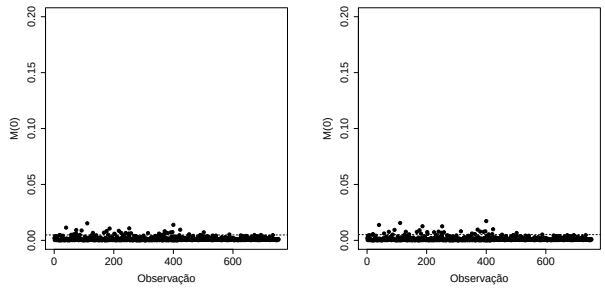
Apesar de algumas observações ultrapassarem o limite de referência no caso dos modelos NI, observa-se que os desvios-padrão das medidas são bem menores. Isso se deve ao fato de estarmos utilizando um valor de referência mais sensível ($\varsigma = 2$). Nestes casos as medidas oscilam em torno do valor limite, enquanto no caso do modelo Normal elas diferem muito do valor limite, como observado nos estudos de simulação.

A análise deste conjunto de dados revelou algumas observações influentes para a estimação realizada via modelo Normal. Estas observações influenciaram o processo de estimação geral do modelo, e especificamente a estimação do parâmetro de escala. Algumas destas observações parecem ter valores atípicos para a variável preditora "número de anos de estudos da mãe", por exemplo, as observações #453 e #502 que apresentaram 17 e 16 anos de estudo, respectivamente. O efeito destas observações foi notavelmente menor sobre a estimação via modelos NI.

A análise apresentada neste capítulo extende a proposta de Massuia et al. (2015) ao utilizar outras distribuições de caudas pesadas além da t de Student e abordar os esquemas de perturbação sobre uma variável preditora contínua e sobre os coeficientes.



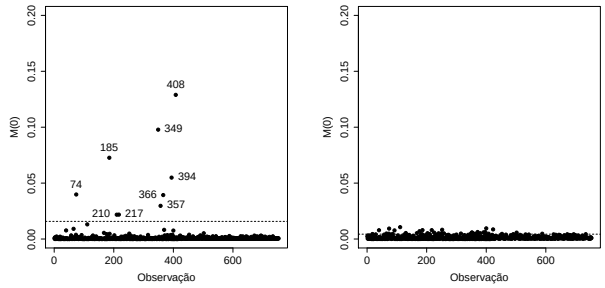
(a) Normal **(b)** t de Student



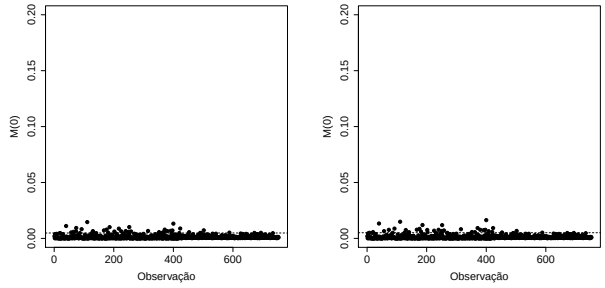
(c) Slash **(d)** Normal Contaminada

Figura 6.15: Dados "`wage.rates`". Perturbação ponderação de casos.

Nota: Observações acima do limite de referência (além das identificadas nos gráficos).
Normal: No gráfico. **t de Student**: #65, #74, #217, #369 #394, #453 e #502.
Slash: #65, #74, #217, #369, #453 e #502. **Normal Contaminada**: #408, #453 e #502.



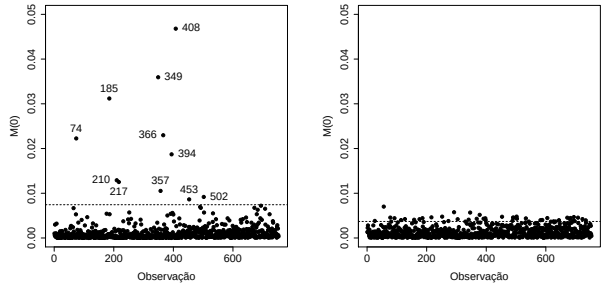
(a) Normal **(b)** t de Student



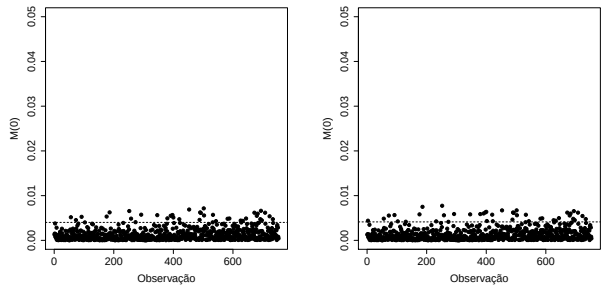
(c) Slash **(d)** Normal Contaminada

Figura 6.16: Dados "`wage.rates`". Perturbação sobre o parâmetro de escala.

Nota: Observações acima do limite de referência (além das identificadas nos gráficos).
Normal: No gráfico. **t de Student**: #65, #74, #210, #217, #349, #357, #366, #369 #394, #408, #453 e #502.
Slash: #65, #74, #217, #369, #453 e #502. **Normal Contaminada**: #408, #453 e #502.



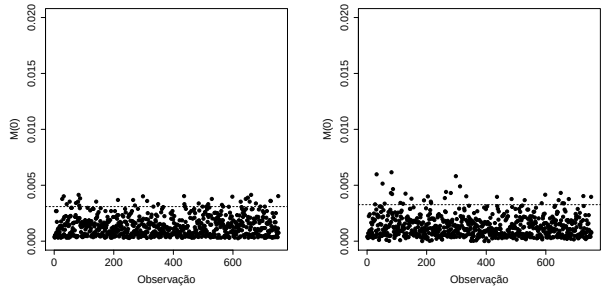
(a) Normal **(b)** t de Student



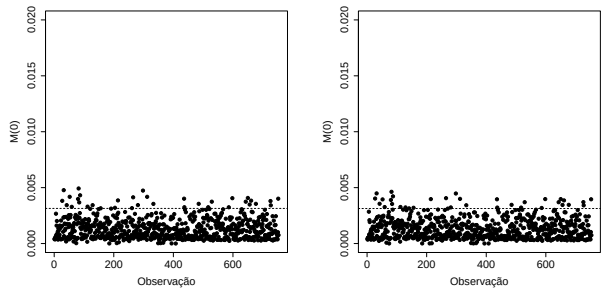
(c) Slash **(d)** Normal Contaminada

Figura 6.17: Dados "`wage.rates`". Perturbação sobre a variável preditora "Nº de anos de estudos da mãe".

Nota: Observações acima do limite de referência (além das identificadas nos gráficos).
Normal: No gráfico. **t de Student**: #453 e #502.
Slash: #453 e #502. **Normal Contaminada**: #453 e #502.



(a) Normal **(b)** t de Student



(c) Slash **(d)** Normal Contaminada

Figura 6.18: Dados "`wage.rates`". Perturbação sobre os coeficientes.

Nota: Observações acima do limite de referência (além das identificadas nos gráficos).
Normal: #349. **t de Student**: Nenhuma.
Slash: Nenhuma. **Normal Contaminada**: Nenhuma.

CAPÍTULO 7

Diagnóstico de influência em modelos de regressão não linear censurados utilizando distribuições NI

Neste capítulo são apresentadas as medidas de influência global e local propostas para a análise de diagnóstico de modelos RNLCNI. A obtenção das medidas de influência para este modelo estão na Seção 7.1. A Seção 7.2 mostra os resultados dos estudos de simulação sobre as medidas de diagnóstico propostas e a Seção 7.3 uma aplicação a dados reais.

7.1 Diagnóstico de influência

Nesta seção são apresentadas as medidas de influência global e local para o modelo (3.2), segundo as metodologias de Zhu et al. (2001); Zhu e Lee (2001).

7.1.1 Influência global

Da mesma forma que o modelo linear, a análise de influência global será avaliada através da distância generalizada de Cook (Seção 5.1 do Capítulo 5). A medida de influência neste caso depende do vetor gradiente da função Q, sem a i-ésima observação, cujas entradas são

$$\begin{aligned}
\mathcal{G}_{Q,\beta,[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= \frac{1}{\widehat{\sigma^2}}\sum_{i\neq j}\left[\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}})d\mu_j - \mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\mu_j d\mu_j\right], \text{ e} \\
\mathcal{G}_{Q,\sigma^2,[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= -\frac{n}{2\widehat{\sigma^2}} + \frac{1}{2\widehat{\sigma^4}}\sum_{i\neq j}\left[\mathcal{E}_{2j}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}})\mu_j + \mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\mu_j^2\right].
\end{aligned}$$

A distância generalizada de Cook depende também da matriz hessiana da função

Q, para este modelo, composta pelas seguintes entradas

$$\begin{aligned}
\mathcal{H}_{Q,\beta}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= \frac{1}{\widehat{\sigma^2}}\sum_{i=1}^{n}\left[\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})D\mu_i - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\left(d\mu_i d\mu_i^\top + \mu_i D\mu_i\right)\right], \\
\mathcal{H}_{Q,\beta,\sigma^2}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= -\frac{1}{\widehat{\sigma^4}}\sum_{i=1}^{n}\left[\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})d\mu_i - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mu_i d\mu_i\right], \text{ e} \\
\mathcal{H}_{Q,\sigma^2}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) &= \frac{n}{2\widehat{\sigma^4}} - \frac{1}{\widehat{\sigma^6}}\sum_{i=1}^{n}\left[\mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\mu_i + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mu_i^2\right],
\end{aligned}$$

em que $\mu_i = \eta(\mathbf{X}_i, \widehat{\boldsymbol{\beta}})$, $d\mu_i = \dfrac{\partial\eta(\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}$, e $D\mu_i = \dfrac{\partial^2\eta(\mathbf{X}_i,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}$.

Com essas duas quantidades calculadas, deve-se substituí-las na expressão (5.3) para se obter as medidas de influência global.

7.1.2 Influência local

Assim como no caso do modelo linear, para se obter as medidas de influência local é preciso calcular a matriz hessiana da função Q e a matriz $\boldsymbol{\Delta_\omega}$ para cada esquema de perturbação de interesse. A matriz hessiana de Q, para este modelo, foi apresentada na Subseção 7.1.1. A seguir serão apresentadas as entradas da matriz $\boldsymbol{\Delta_\omega}$ para o modelo RNLCNI sob os esquemas de perturbação ponderação de casos, sobre o parâmetro de escala, em uma variável preditora contínua e sobre os coeficientes do modelo.

Esquemas de perturbação

Nesta seção é apresentada a construção da matriz $\boldsymbol{\Delta_\omega}$ sobre os esquemas de perturbação de interesse. Para cada esquema de perturbação, as entradas da matriz correspondem a

$$\boldsymbol{\Delta}_\beta = \frac{\partial^2 Q(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\omega}^\top} \quad \text{e} \quad \boldsymbol{\Delta}_{\sigma^2} = \frac{\partial^2 Q(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2\partial\boldsymbol{\omega}^\top},$$

de modo que $\boldsymbol{\Delta_\omega} = \left(\boldsymbol{\Delta}_\beta^\top, \boldsymbol{\Delta}_{\sigma^2}^\top\right)^\top$.

Para os esquemas de perturbação sobre uma variável preditora e sobre os coeficientes do modelo, a contaminação é feita dentro da função não linear $\eta(\mathbf{X}_i, \boldsymbol{\beta})$. Desta forma deve-se calcular as entradas da matriz $\boldsymbol{\Delta_\omega}$ para cada função não linear. Estas expressões estão apresentadas no Apêndice A para as funções não lineares utilizadas no estudo de simulação (Seção 7.2) e na aplicação a dados reais (Seção 7.3).

Já os esquemas de perturbação ponderação de casos e sobre o parâmetro de escala possuem expressões gerais, independentes da função não linear, e estão apresentadas abaixo.

Perturbação Ponderação de casos: Assim como no caso linear, os pesos são atribuídos aos valores esperados da função de log-verossimilhança dos dados completos do modelo. Para este esquema, $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)^\top$ e $\boldsymbol{\omega}_0 = (1, ..., 1)^\top = \mathbf{1}_n^\top$. A matriz $\boldsymbol{\Delta}_{\boldsymbol{\omega}}$ será formada pelos seguintes componentes

$$\begin{aligned}\boldsymbol{\Delta}_{\boldsymbol{\beta}} &= \frac{1}{\widehat{\sigma^2}}\left[d\boldsymbol{\mu}^\top diag\left(\mathcal{E}_1(\widehat{\boldsymbol{\theta}})\right) - d\boldsymbol{\mu}^\top \mathcal{E}_0(\widehat{\boldsymbol{\theta}})\boldsymbol{\mu}^\top\right], \text{ e}\\ \boldsymbol{\Delta}_{\sigma^2} &= -\frac{1}{2\widehat{\sigma^2}}\left[\mathbf{1}_n^\top - \frac{1}{\widehat{\sigma^2}}\mathbf{B}^\top\right],\end{aligned}$$

em que $\boldsymbol{\mu}$ é um vetor de dimensão n com entradas $\mu_i = \eta(\mathbf{X}_i, \widehat{\boldsymbol{\beta}})$, $d\boldsymbol{\mu}$ é uma matriz de dimensão $n \times p$ de derivadas de primeira ordem de $\boldsymbol{\mu}$ em relação ao vetor $\boldsymbol{\beta}$ e $\mathbf{B}$ é um vetor de dimensão n com entradas $B_i = \mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\mu_i + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mu_i^2$.

Perturbação sobre o parâmetro de escala: A perturbação sobre σ^2 é inserida ao substituí-lo por $\sigma^2(\omega_i) = \omega_i^{-1}\sigma^2$, $i = 1, ..., n$, $\omega_i > 0\ \forall\ i$ na função Q. Sob esse esquema, $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)^\top$ e $\boldsymbol{\omega}_0 = \mathbf{1}_n^\top$. Assim,

$$\begin{aligned}\boldsymbol{\Delta}_{\boldsymbol{\beta}} &= \frac{1}{\widehat{\sigma^2}}\left[d\boldsymbol{\mu}^\top diag\left(\mathcal{E}_1(\widehat{\boldsymbol{\theta}})\right) - d\boldsymbol{\mu}^\top \mathcal{E}_0(\widehat{\boldsymbol{\theta}})\boldsymbol{\mu}^\top\right], \text{ e}\\ \boldsymbol{\Delta}_{\sigma^2} &= \frac{1}{2\widehat{\sigma^4}}\mathbf{B}^\top.\end{aligned}$$

7.2 Estudo de simulação

Nesta seção apresentamos os resultados de um estudo de Monte Carlo desenvolvido para comparar o desempenho do modelo não linear censurado na presença de *outliers* utilizando as distribuições Normal, t de Student, Slash e Normal Contaminada. Foi simulado o modelo de curva de crescimento não linear

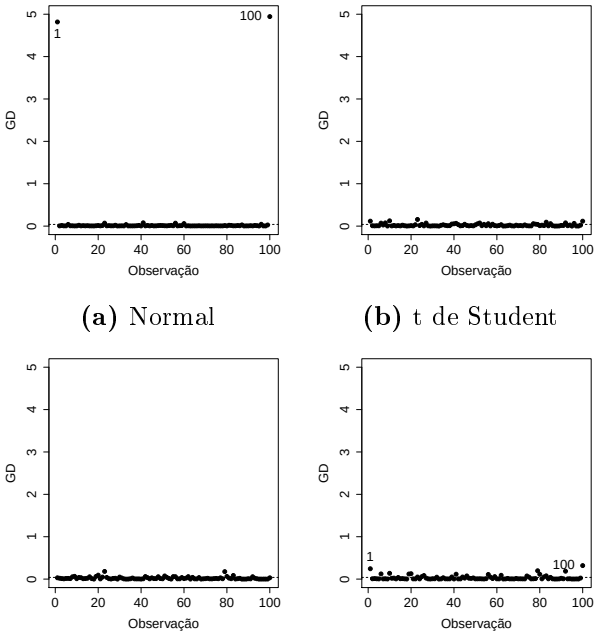
$$Y_i = \frac{\beta_1}{1 + exp(\beta_2 + \beta_3 x_i))} + \epsilon_i,\ i = 1, ..., n, \tag{7.1}$$

com $n = 100$, $\epsilon_i \sim N(0, \sigma^2)$ para $i = 2, 3, ...99$, $\epsilon_1 = -5$ e $\epsilon_{100} = 5$, $x_i \sim \text{Unif}(10; 20)$, $\beta_1 = 330$, $\beta_2 = 6,5$, $\beta_3 = -0,7$ e $\sigma^2 = 1$. A definição dos erros determinou a perturbação sobre as observações #1 e #100. O percentual de censura adotado foi 20%. As distribuições analisadas foram a Normal, Student-t ($\nu = 4,864$), Slash ($\nu = 1,174$) e Normal Contaminada ($\nu = (0,1,0,1)$). A escolha dos valores de ν foi feita utilizando os valores que maximizavam a função de log-verossimilhança (Meza

et al., 2012). Para gerar as observações censuradas à direita utilizou-se a proposta de Tsuyuguchi (2012). O modelo (7.1) foi gerado e o ponto de corte δ_i foi definido como o r-ésimo valor do vetor $\mathbf{Y}$ ordenado, ou seja, $\delta_i = Y_{(r)}, \ \forall i = 1, ..., r$ em que $r = n - n \times pc$ é o número de observações censuradas e pc é o percentual de censura. Desta forma, $Y_{(r)}$ se torna uma observação censurada e todos os valores de $\mathbf{Y}$ maiores ou iguais a ele assumem o mesmo valor que $Y_{(r)}$. Os valores iniciais foram obtidos através da função *nls* do pacote *stats*, do *software* R 3.2.2 (R Core Team, 2015).

Um estudo de Monte Carlo com 1.000 réplicas do modelo (7.1) foi realizado e foram obtidos o percentual de réplicas em que as observações contaminadas foram influentes, e ainda a média e o desvio-padrão das medidas de influência propostas. Os resultados estão apresentados na Tabela 7.1.

As observações contaminadas foram influentes em todas as réplicas de Monte Carlo, considerando os modelos Normal, t de Student e Normal Contaminada, para a distância de Cook. Estas observações, porém, não foram influentes para o modelo Slash em nenhuma das réplicas. Apesar de serem classificadas como influentes, no caso dos modelos t de Student e Normal Contaminada, as medidas de influência foram bem menores que as obtidas via modelo Normal. A Figura 7.1 apresenta a distância generalizada de Cook para os quatro modelos analisados, em uma das réplicas. Pode ser visto que o modelo Normal foi altamente influenciado pelas observações contaminadas, enquanto os modelos NI (t de Student, Slash e Normal Contaminada) acomodaram melhor o efeito destas observações.



(a) Normal **(b)** t de Student

(c) Slash **(d)** Normal Contaminada

Figura 7.1: Estudo de simulação. Distância generalizada de Cook. Modelo não linear.

Tabela 7.1: Estudo de simulação. Análise de influência via estudo de Monte Carlo para as observações #1 e #100 por distribuição e medida de diagnóstico: GD - distância generalizada de Cook, PC - ponderação de casos, ES - parâmetro de escala, VP - variável preditora e CO - coeficientes. Modelo não linear.

Diag	Estatística	Normal		t de Student		Slash		Normal Cont.	
		#1	#100	#1	#100	#1	#100	#1	#100
GD	% Inf[a]	100%	100%	100%	100%	0%	0%	100%	100%
	M[b]	4,712	5,277	0,117	0,116	0,031	0,030	0,402	0,411
	DP[c]	(0,537)	(0,503)	($4{,}0e^{-4}$)	($4{,}0e^{-4}$)	(0,001)	(0,001)	(0,124)	(0,116)
	Ref[d]	0,040		0,040		0,040		0,040	
PC	% Inf	100%	100%	100%	100%	0%	0%	100%	100%
	M	0,402	0,444	0,051	0,052	0,022	0,021	0,113	0,115
	DP	(0,028)	(0,025)	(0,001)	(0,001)	($9{,}2e^{-4}$)	($9{,}1e^{-4}$)	(0,026)	(0,024)
	M (DP) Ref	0,129 (0,003)		0,032 (0,001)		0,031 (0,002)		0,047 (0,005)	
ES	% Inf	100%	100%	100%	100%	0%	0%	100%	100%
	M	0,429	0,475	0,057	0,056	0,020	0,019	0,132	0,135
	DP	(0,030)	(0,026)	(0,003)	(0,002)	(0,001)	(0,001)	(0,033)	(0,030)
	M (DP) Ref	0,137 (0,003)		0,037 (0,002)		0,036 (0,003)		0,054 (0,006)	
VP	% Inf	0%	0%	0%	0%	0%	0%	0%	0%
	M	0,008	0,007	$1{,}8e^{-5}$	$1{,}7e^{-5}$	$5{,}8e^{-6}$	$5{,}5e^{-6}$	$9{,}4e^{-5}$	$9{,}1e^{-5}$
	DP	($3{,}1e^{-6}$)	($2{,}9e^{-6}$)	($6{,}9e^{-6}$)	($6{,}5e^{-6}$)	($2{,}0e^{-6}$)	($1{,}9e^{-6}$)	($5{,}5e^{-7}$)	($5{,}2e^{-7}$)
	M (DP) Ref	0,014 ($5{,}3e^{-6}$)		0,017 ($2{,}8e^{-4}$)		0,017 ($2{,}8e^{-4}$)		0,015 ($2{,}2e^{-4}$)	
CO	% Inf	100%	100%	0%	0%	0%	0%	0%	0%
	M	0,121	0,114	0,001	0,001	$3{,}7e^{-4}$	$3{,}3e^{-4}$	0,004	0,004
	DP	(0,007)	(0,006)	($2{,}0e^{-4}$)	($1{,}8e^{-4}$)	($6{,}6e^{-5}$)	($5{,}9e^{-5}$)	($6{,}5e^{-4}$)	($5{,}4e^{-4}$)
	M (DP) Ref	0,042 (0,002)		0,017 ($3{,}0e^{-4}$)		0,017 ($2{,}9e^{-4}$)		0,019 ($4{,}6e^{-4}$)	

[a] % Inf: percentual de réplicas de Monte Carlo em que a observação foi considerada influente (maior que o valor de referência).
[b] M é a média das medidas de influência.
[c] DP é o desvio-padrão das medidas de influência.
[d] Ref é o valor de referência para considerar uma observação influente.

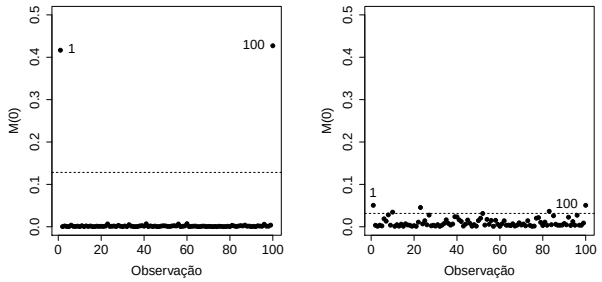
Ao analisar as medidas de influência local (Tabela 7.1), no caso dos esquemas de perturbação ponderação de casos e sobre o parâmetro de escala, as observações contaminadas foram influentes para todas as réplicas considerando os modelos Normal, t de Student e Normal Contaminada, porém, os modelos NI apresentaram medidas de influência menores que o modelo Normal. Em relação ao modelo Slash, estes casos não foram influentes em nenhuma das réplicas.

No esquema de perturbação sobre uma variável preditora, as observações não foram influentes para nenhum dos modelos. Isso pode ser devido ao fato de a perturbação ter sido inserida nos resíduos, e não sobre a variável preditora. No caso da perturbação sobre os coeficientes, as observações contaminadas foram influentes na estimação via modelo Normal para todas as réplicas, enquanto para os modelos NI elas não foram classificadas como influentes em nenhuma das réplicas.

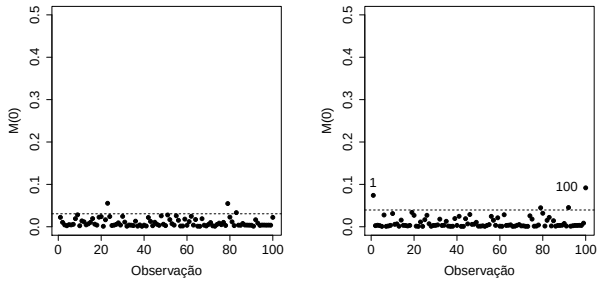
As Figuras 7.2 a 7.5 trazem a representação gráfica das medidas de influência local para uma réplica. A forma como a contaminação foi inserida, sobre os resíduos, afetou a estimação global do modelo e as estimações específicas dos coeficientes e do parâmetro de escala, no caso do modelo Normal. O efeito destas observações foi

notadamente inferior sobre os modelos NI.

Os resultados deste estudo sugerem que as medidas de influência propostas conseguem identificar corretamente as observações contaminadas e os pontos específicos da formulação dos modelos onde elas o influenciam. Assim como na análise do modelo de regressão linear, optamos por identificar nos gráficos de influência somente as observações suspeitas, para atender ao objetivo do estudo. As demais observações que ultrapassaram os limites de referência se devem à forma como definimos os limites de referência, permitindo até 5% de falsos positivos.

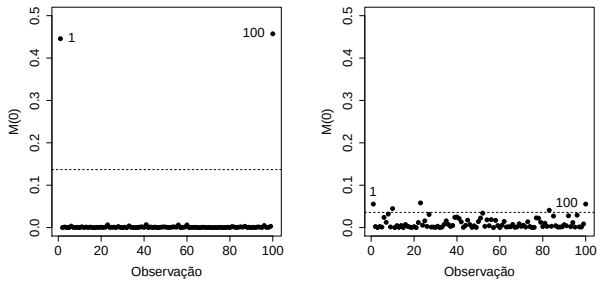


(a) Normal **(b)** t de Student

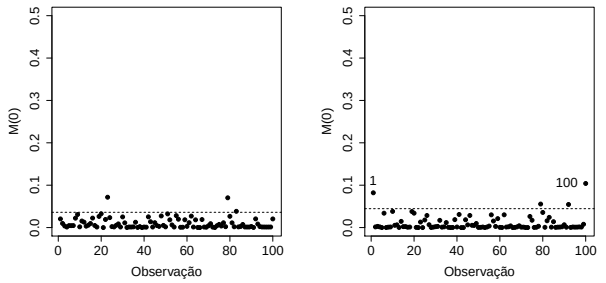


(c) Slash **(d)** Normal Contaminada

Figura 7.2: Estudo de simulação. Perturbação ponderação de casos. Modelo não linear.

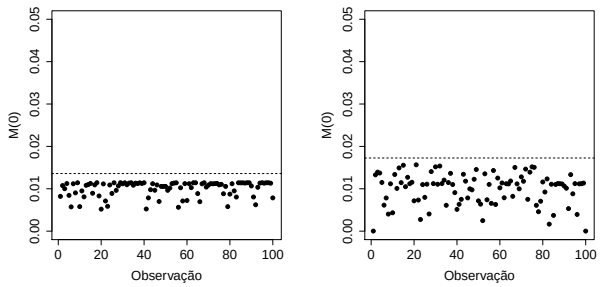


(a) Normal **(b)** t de Student

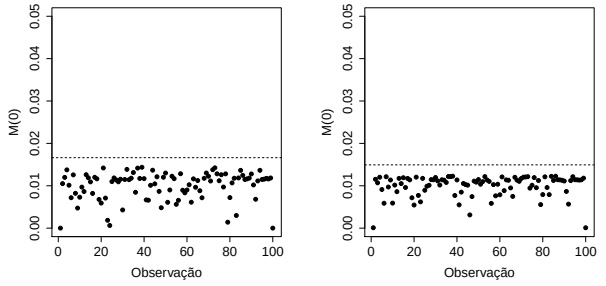


(c) Slash **(d)** Normal Contaminada

Figura 7.3: Estudo de simulação. Perturbação sobre o parâmetro de escala. Modelo não linear.

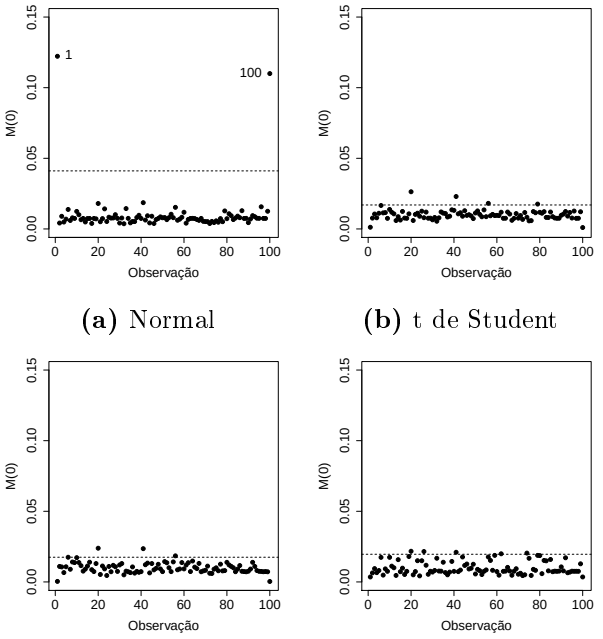


(a) Normal **(b)** t de Student



(c) Slash **(d)** Normal Contaminada

Figura 7.4: Estudo de simulação. Perturbação sobre uma variável preditora contínua. Modelo não linear.



(a) Normal **(b)** t de Student

(c) Slash **(d)** Normal Contaminada

Figura 7.5: Estudo de simulação. Perturbação sobre os coeficientes. Modelo não linear.

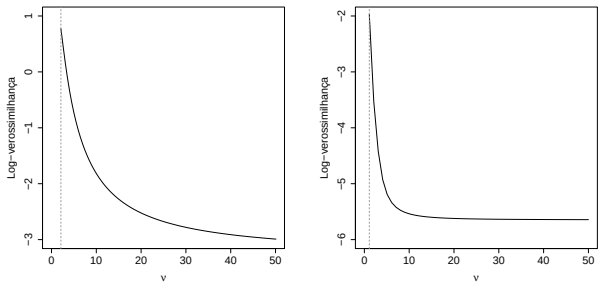
7.3 Aplicação

Nesta seção as medidas de influência propostas são aplicadas aos dados de um teste de deformação de metais (Nelson, 2004). Os dados consistem no número de ciclos até a deformação de 26 espécimes de metais submetidos a um pseudo-stress. A variável número de ciclos apresenta 15,4% de observações censuradas à direita porque alguns metais não deformaram ao fim do período do estudo.

O seguinte modelo foi proposto para ajustar estes dados

$$y_i = \beta_1 \exp(\beta_2 x_i) + \epsilon_i,\ i = 1, ..., 26,$$

em que $y_i = log_{10}(\text{Ciclos})$, $x_i = 1/(\text{Pseudo-stress})$ e $\epsilon_i \sim NI(0, \sigma^2, \nu)$. As distribuições Normal, t de Student ($\nu = 2, 101$), Slash ($\nu = 1, 101$) e Normal Contaminada ($\nu = (0, 4, 0, 05)$) foram utilizadas para ajustar estes dados. Os valores de ν foram obtidos como os que maximizavam a função de log-verossimilhança, assim como foi feito por Meza et al. (2012) (ver Figura 7.6 e Tabela 7.2).



(a) t de Student **(b)** Slash

Figura 7.6: Dados de deformação de metais. Valores de ν *vs* log-verossimilhança para os modelos t de Student e Slash.

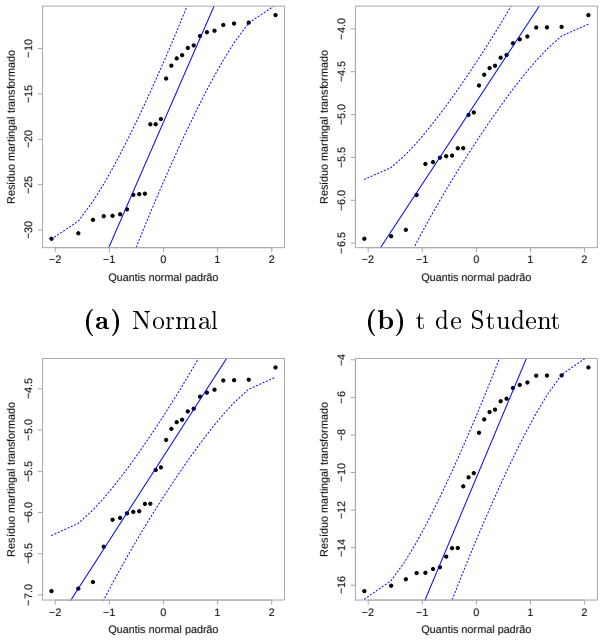
Tabela 7.2: Dados de deformação de metais. Valores de log-verossimilhanças de acordo com os valores de $\nu = (\nu_1, \nu_2)$ testados para o modelo Normal Contaminada.

	ν_2					
ν_1	0,05	0,1	0,2	0,3	0,4	0,5
0,1	-2,16	-2,41	-3,13	-3,88	-4,55	-5,05
0,2	-0,95	-1,60	-2,74	-3,54	-4,21	-4,76
0,3	-0,31	-1,25	-2,60	-3,47	-4,13	-4,67
0,4	-0,10	-1,22	-2,66	-3,55	-4,18	-4,68
0,5	-0,26	-1,44	-2,87	-3,72	-4,30	-4,75

A Tabela 7.3 apresenta os resultados da estimação. O modelo Normal Contaminada apresentou as estimativas mais precisas (com menores erros padrão), porém, o modelo t de Student apresentou os melhores resultados de AIC, BIC e log-verossimilhança. A melhor adequação do modelo t de Student é comprovada pelos envelopes dos resíduos Martingal transformados (Fórmula 6.3 e Figura 7.7).

Tabela 7.3: Dados de deformação de metais. Estimativas e erros padrão (EP - em parênteses) para os parâmetros do modelo, segundo as distribuições Normal, t de Student Slash e Normal Contaminada.

Parâmetro	Normal Estimativa (EP)	t de Student Estimativa (EP)	Slash Estimativa (SE)	Normal Contaminada Estimativa (EP)
β_1	2,445 (0,312)	2,396 (0,180)	2,396 (0,210)	2,385 (0,164)
β_2	62,049 (30,536)	65,325 (17,624)	65,166 (21,075)	65,820 (16,295)
σ^2	0,077	0,016	0,013	0,008
ν	-	2,101	1,101	(0,4; 0,05)
AIC	17,294	6,450	11,936	10,200
BIC	21,068	11,483	16,969	16,490
Log-verossimilhança	-5,6468	0,7749	-1,9681	-0,0999



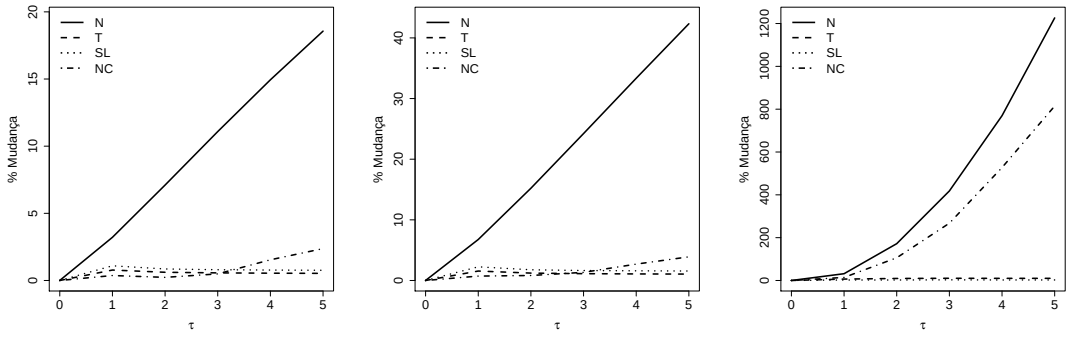
(a) Normal **(b)** t de Student

(c) Slash **(d)** Normal Contaminada

Figura 7.7: Dados de deformação de metais. Envelopes dos resíduos martingal transformados para os modelos Normal, t de Student, Slash e Normal Contaminada.

A robustez dos modelos propostos foi avaliada através da análise de influência de uma observação contaminada sobre as estimativas fornecidas para os parâmetros do modelo. Para isso, a observação y_{10} foi escolhida arbitrariamente, sem perda de generalidade, e substituída por $y_{10} + \tau$, τ entre 0 e 5, e foram calculadas as mudanças relativas absolutas (Fórmula 6.4), em percentual, das estimativas. A

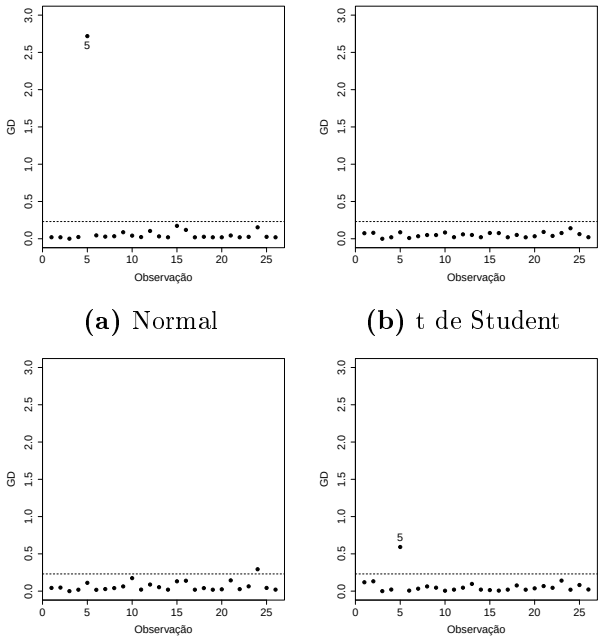
Figura 7.8 mostra que os modelos NI (t de Student, Slash e Normal Contaminada) foram menos influenciados pela observação contaminada que o modelo Normal. Em relação à estimação de σ^2, o modelo Normal Contaminada apresentou um resultado parecido com o do modelo Normal.



(a) β_1 **(b)** β_2 **(c)** σ^2

Figura 7.8: Dados de deformação de metais. Mudanças relativas absolutas por nível de contaminação.

Em seguida foi realizado o diagnóstico de influência aplicando-se as medidas propostas na Seção 7.1. A análise da distância de Cook (Figura 7.9 e Tabela 7.4) revela que a observação #5 foi altamente influente no caso do modelo Normal. Os modelos NI acomodaram melhor o efeito desta observação, apesar de ela ser classificada como influente também para o modelo Normal Contaminada.



(a) Normal **(b)** t de Student

(c) Slash **(d)** Normal Contaminada

Figura 7.9: Dados de deformação de metais. Distância generalizada de Cook.

No diagnóstico de influência local, um comportamento similar ao da distância

Tabela 7.4: Dados de deformação de metais. Estatísticas descritivas das medidas de influência.

Medidas	Normal Média (DP[a])	t de Student Média (DP)	Slash Média (DP)	Normal Contaminada Média (DP)
Distância de Cook (GD)	0,149 (0,526)	0,052 (0,033)	0,065 (0,066)	0,067 (0,117)
Ponderação de casos	M[b] (0,117)	M (0,030)	M (0,040)	M (0,054)
Escala	M (0,139)	M (0,035)	M (0,048)	M (0,070)
Variável preditora	M (0,028)	M (0,037)	M (0,034)	M (0,038)
Coeficientes	M (0,073)	M (0,039)	M (0,043)	M (0,039)

[a] DP é o desvio-padrão das estimativas.
[b] M é a média das medidas de influência, no caso da influência local, é igual a 0,038.

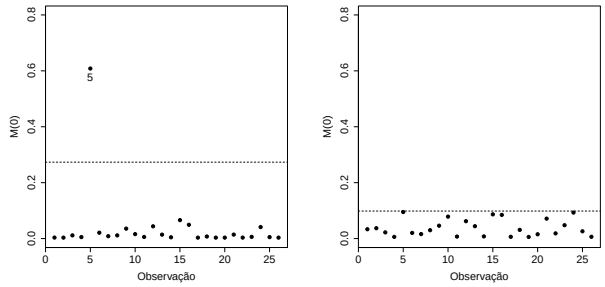
de Cook foi observado para os esquemas de perturbação ponderação de casos (Figura 7.10) e sobre o parâmetro de escala (Figura 7.11). No caso do esquema de perturbação sobre uma variável preditora contínua, a observação #5 não foi influente para nenhum dos modelos e no esquema de perturbação sobre os coeficientes, esta observação foi influente para todos os modelos, porém, com menores medidas de influência sobre os modelos NI (t de Student, Slash e Normal Contaminada).

O impacto da observação #5 sobre as estimativas fornecidas pelos modelos foi avaliado através das mudanças relativas absolutas, utilizando a fórmula (6.4) aplicando a estimativa do modelo sem o caso #5 para θ^c. A Tabela 7.5 apresenta os resultados, onde se confirmam as informações de maior influência exercida sobre o modelo Normal, comparado aos modelos NI. A exclusão desta observação diminui de forma considerável o valor das estimativas do parâmetro σ^2 em todos os modelos.

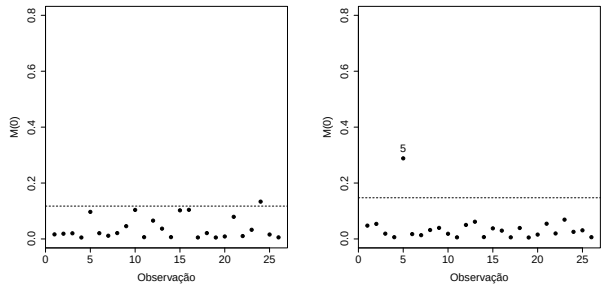
Os resultados desta análise sugerem que a observação #5 exerceu grande influência na estimação geral realizada via modelo Normal, e especificamente sobre as estimações dos coeficientes e do parâmetro de escala. Esta observação foi influente na estimação dos coeficientes, considerando os modelos NI, porém, com influência menor que a exercida sobre o modelo Normal.

Tabela 7.5: Dados de deformação de metais. Mudanças relativas absolutas sobre os parâmetros estimados pelo modelo com todas as observações e sem a observação #5.

Parâmetros	Normal	t de Student	Slash	Normal Contaminada
β_1	3,23%	0,22%	0,33%	0,26%
β_2	6,30%	0,42%	0,71%	0,26%
σ^2	50,0%	25,0%	26,9%	40,0%

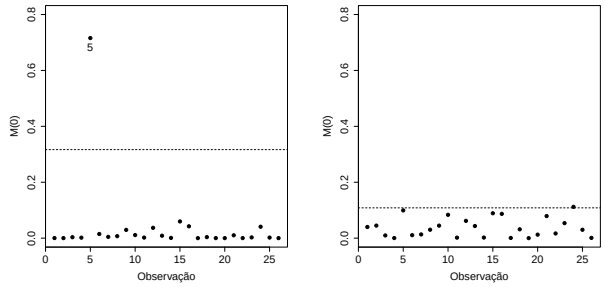


(a) Normal **(b)** t de Student

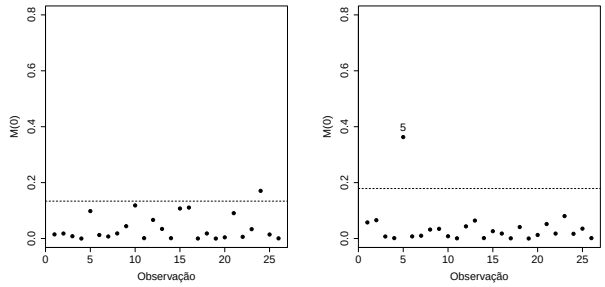


(c) Slash **(d)** Normal Contaminada

Figura 7.10: Dados de deformação de metais. Perturbação ponderação de casos.

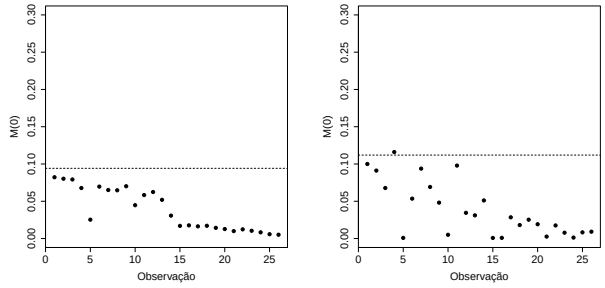


(a) Normal **(b)** t de Student

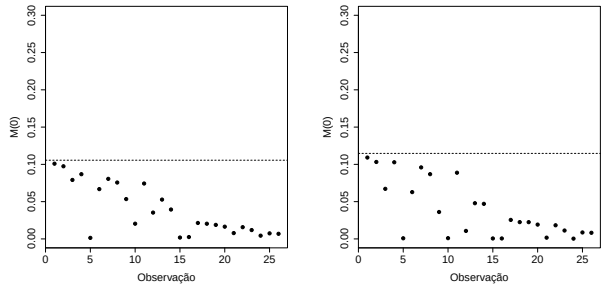


(c) Slash **(d)** Normal Contaminada

Figura 7.11: Dados de deformação de metais. Perturbação sobre o parâmetro de escala.

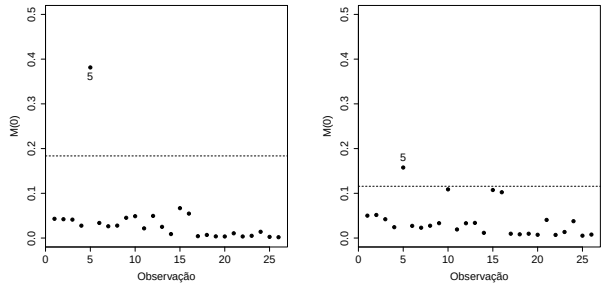


(a) Normal **(b)** t de Student

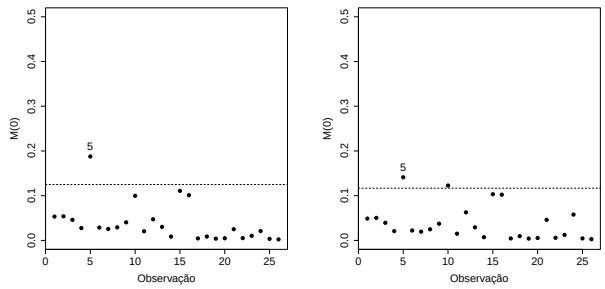


(c) Slash **(d)** Normal Contaminada

Figura 7.12: Dados de deformação de metais. Perturbação sobre uma variável preditora.



(a) Normal **(b)** t de Student



(c) Slash **(d)** Normal Contaminada

Figura 7.13: Dados de deformação de metais. Perturbação sobre os coeficientes do modelo.

CAPÍTULO 8

Diagnóstico de influência em modelos de regressão linear para dados longitudinais censurados utilizando a distribuição t de Student

Neste capítulo são apresentadas as medidas de influência global e local propostas para a análise de diagnóstico de modelos RLCMT. A obtenção das medidas de influência para este modelo estão na Seção 8.1. A Seção 8.2 mostra os resultados dos estudos de simulação sobre as medidas de diagnóstico propostas e a Seção 8.3 uma aplicação a dados reais.

8.1 Diagnóstico de influência

Nesta seção são apresentadas as medidas de influência global e local para o modelo (3.3), segundo as metodologias de Zhu et al. (2001); Zhu e Lee (2001).

8.1.1 Influência global

A análise de influência global será avaliada através da distância generalizada de Cook (GD), descrita na Seção 5.1 do Capítulo 5. A medida de influência neste caso depende do vetor gradiente da função Q, sem a i-ésima observação, cujas entradas são

$$\mathcal{G}_{Q,\beta,[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \frac{1}{\widehat{\sigma^2}}\sum_{i\neq j}\left[\mathbf{X}_j^\top\widehat{\mathbf{E}}_j^{-1}\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\mathbf{X}_j^\top\widehat{\mathbf{E}}_j^{-1}\mathbf{X}_j\widehat{\boldsymbol{\beta}}\right],$$

$$\mathcal{G}_{Q,\sigma^2,[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \sum_{i\neq j}\left\{-\frac{n_j}{2\widehat{\sigma^2}}+\frac{1}{2\widehat{\sigma^4}}\left[tr\left(\mathcal{E}_{2j}(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{E}}_j^{-1}\right)-2\widehat{\boldsymbol{\beta}}^\top\mathbf{X}_j^\top\widehat{\mathbf{E}}_j^{-1}\mathcal{E}_{1j}(\widehat{\boldsymbol{\theta}})\right.\right.$$
$$\left.\left.+\mathcal{E}_{0j}(\widehat{\boldsymbol{\theta}})\widehat{\boldsymbol{\beta}}^\top\mathbf{X}_j^\top\widehat{\mathbf{E}}_j^{-1}\mathbf{X}_j\widehat{\boldsymbol{\beta}}\right]\right\},\ \text{e}$$

$$\mathcal{G}_{Q,\phi_k,[i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \sum_{i\neq j}\left\{-\frac{1}{2}tr\left[\widehat{\mathbf{E}}_j^{-1}d\phi_k\right]-\frac{1}{2\widehat{\sigma^2}}\left[tr\left(-B_j\widehat{\mathbf{E}}_j^{-1}d\phi_k\widehat{\mathbf{E}}_j^{-1}\right)\right]\right\},\ k=1,2,$$

em que $d\phi_k=\frac{\partial\mathbf{E}_i}{\partial\phi_k}$ e $B_i=\mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}})-2\mathbf{X}_i\widehat{\boldsymbol{\beta}}\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})^\top+\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{X}_i\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top\mathbf{X}_i^\top$. A matriz hessiana da função Q é composta pelas seguintes entradas

$$\mathcal{H}_{Q,\boldsymbol{\beta}}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^N\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1}\mathbf{X}_i,$$

$$\mathcal{H}_{Q,\sigma^2}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^N\left\{\frac{n_i}{2\widehat{\sigma^4}}-\frac{1}{\widehat{\sigma^6}}\left[tr\left(B_i\widehat{\mathbf{E}}_i^{-1}\right)\right]\right\},$$

$$\mathcal{H}_{Q,\boldsymbol{\beta},\sigma^2}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \frac{1}{\widehat{\sigma^4}}\sum_{i=1}^N\left[-\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1}\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})-\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1}\mathbf{X}_i\widehat{\boldsymbol{\beta}}\right],$$

$$\mathcal{H}_{Q,\phi_k}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^N\left\{-\frac{1}{2}tr\left[-\widehat{\mathbf{E}}_i^{-1}d\phi_k\widehat{\mathbf{E}}_i^{-1}d\phi_k+\widehat{\mathbf{E}}_i^{-1}D\phi_k\right]\right.$$
$$\left.-\frac{1}{2\widehat{\sigma^2}}tr\left[2B_i\widehat{\mathbf{E}}_i^{-1}d\phi_k\widehat{\mathbf{E}}_i^{-1}d\phi_k\widehat{\mathbf{E}}_i^{-1}-B_i\widehat{\mathbf{E}}_i^{-1}D\phi_k\widehat{\mathbf{E}}_i^{-1}\right]\right\},\ k=1,2,$$

$$\mathcal{H}_{Q,\boldsymbol{\beta},\phi_k}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^N\left\{\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1}d\phi_k\widehat{\mathbf{E}}_i^{-1}\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\right.$$
$$\left.-\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1}d\phi_k\widehat{\mathbf{E}}_i^{-1}\mathbf{X}_i\widehat{\boldsymbol{\beta}}\right\},\ k=1,2,$$

$$\mathcal{H}_{Q,\sigma^2,\phi_k}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \frac{1}{2\widehat{\sigma^4}}\sum_{i=1}^N\left\{tr\left[-\mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}\widehat{\mathbf{E}}_i^{-1}d\phi_k\widehat{\mathbf{E}}_i^{-1})\right]+2\boldsymbol{\beta}^\top\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1}d\phi_k\widehat{\mathbf{E}}_i^{-1}\mathcal{E}1i(\widehat{\boldsymbol{\theta}})\right.$$
$$\left.-\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\boldsymbol{\beta}^\top\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1}d\phi_k\widehat{\mathbf{E}}_i^{-1}\mathbf{X}_i\widehat{\boldsymbol{\beta}}\right\},\ k=1,2,\ \text{e}$$

$$\mathcal{H}_{Q,\phi_1,\phi_2}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^N\left\{-\frac{1}{2}tr\left[-\widehat{\mathbf{E}}_i^{-1}d\phi_1\widehat{\mathbf{E}}_i^{-1}d\phi_2+\widehat{\mathbf{E}}_i^{-1}D\phi_1\phi_2\right]\right.$$
$$\left.-\frac{1}{2\widehat{\sigma^2}}tr\left[2B_i\widehat{\mathbf{E}}_i^{-1}d\phi_1\widehat{\mathbf{E}}_i^{-1}d\phi_2\widehat{\mathbf{E}}_i^{-1}-B_i\widehat{\mathbf{E}}_i^{-1}D\phi_1\phi_2\widehat{\mathbf{E}}_i^{-1}\right]\right\},$$

em que $D\phi_k=\frac{\partial^2\mathbf{E}_i}{\partial\phi_k\partial\phi_k}$, k=1,2, e $D\phi_1\phi_2=\frac{\partial^2\mathbf{E}_i}{\partial\phi_1\partial\phi_2}$. As expressões para as derivadas de primeira e segunda ordem de $\phi_k,\ k=1,2$, estão apresentadas no Apêndice B. Com essas duas quantidades calculadas, deve-se substituí-las na expressão (5.3) para se obter as medidas de influência global.

8.1.2 Influência local

Seguindo o método descrito na Seção 5.2 da parte I, para obter as medidas de influência local é preciso calcular a matriz hessiana da função Q e a matriz $\mathbf{\Delta_\omega}$ para cada esquema de perturbação de interesse. A matriz hessiana de Q está apresentada na Subseção 8.1.1. A seguir serão apresentadas as entradas da matriz $\mathbf{\Delta_\omega}$ para o modelo RLCMT sob os esquemas de perturbação ponderação de casos, sobre o parâmetro de escala, em uma variável preditora contínua e sobre os coeficientes do modelo.

Esquemas de perturbação

Nesta seção é apresentada a construção da matriz $\mathbf{\Delta_\omega}$ sobre os esquemas de perturbação de interesse. Para cada esquema de perturbação, as entradas da matriz correspondem a

$$\mathbf{\Delta}_\beta = \frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^\top}, \quad \mathbf{\Delta}_{\sigma^2} = \frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial \sigma^2 \partial \boldsymbol{\omega}^\top} \quad \text{e} \quad \mathbf{\Delta}_\phi = \frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\omega}^\top}.$$

Deste modo, $\mathbf{\Delta_\omega} = \left(\mathbf{\Delta}_\beta^\top, \mathbf{\Delta}_{\sigma^2}^\top, \mathbf{\Delta}_\phi^\top\right)^\top$.

Perturbação ponderação de casos: Neste contexto os pesos são atribuídos aos valores esperados da função de log-verossimilhança dos dados completos do modelo. Para este esquema, $\boldsymbol{\omega} = (\omega_1, ..., \omega_m)^\top$ e $\boldsymbol{\omega}_0 = (1, ..., 1)^\top = \mathbf{1}_m^\top$. A matriz $\mathbf{\Delta_\omega}$ será formada pelos seguintes componentes

$$\begin{aligned}
\mathbf{\Delta}_\beta &= \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{m} \left[\mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathbf{X}_i \widehat{\boldsymbol{\beta}} \right], \\
\mathbf{\Delta}_{\sigma^2} &= \sum_{i=1}^{m} \left\{ -\frac{n_i}{2\widehat{\sigma^2}} + \frac{1}{2\widehat{\sigma^4}} tr\left(B_i \widehat{\mathbf{E}}_i^{-1} \right) \right\}, \text{ e} \\
\mathbf{\Delta}_{\phi_j} &= \sum_{i=1}^{m} \left\{ -\frac{1}{2} tr\left(\widehat{\mathbf{E}}_i^{-1} d\phi_j \right) - \frac{1}{2\widehat{\sigma^2}} tr\left(-B_i \widehat{\mathbf{E}}_i^{-1} d\phi_j \widehat{\mathbf{E}}_i^{-1} \right) \right\}, \ j = 1, 2,
\end{aligned}$$

em que $B_i = \mathcal{E}_{2i}(\widehat{\boldsymbol{\theta}}) - 2\mathbf{X}_i \widehat{\boldsymbol{\beta}} \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})^\top + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \mathbf{X}_i \widehat{\boldsymbol{\beta}} \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_i^\top$.

Perturbação sobre o parâmetro de escala: Neste caso perturba-se σ^2 ao substituí-lo por $\sigma^2(\omega_i) = \omega_i^{-1}\sigma^2$, $i = 1, ..., m$, na função Q. Sob esse esquema,

$\boldsymbol{\omega} = (\omega_1, ..., \omega_m)^\top$ e $\boldsymbol{\omega}_0 = \mathbf{1}_m^\top$. Deste modo, temos

$$\begin{aligned}
\boldsymbol{\Delta}_\beta &= \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{m} \left\{ \mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathbf{X}_i \widehat{\boldsymbol{\beta}} \right\}, \\
\boldsymbol{\Delta}_{\sigma^2} &= \sum_{i=1}^{m} \left\{ \frac{1}{2\widehat{\sigma^4}} tr\left(B_i \widehat{\mathbf{E}}_i^{-1} \right) \right\}, \text{ e} \\
\boldsymbol{\Delta}_{\phi_j} &= \sum_{i=1}^{m} \left\{ -\frac{1}{2\widehat{\sigma^2}} tr\left(-B_i \widehat{\mathbf{E}}_i^{-1} d\phi_j \widehat{\mathbf{E}}_i^{-1} \right) \right\}, \ j = 1, 2.
\end{aligned}$$

Perturbação sobre uma variável preditora: A perturbação, neste caso, é inserida em uma variável preditora contínua substituindo $\mathbf{X}_i$ por $\mathbf{X}_{i\omega_i} = \mathbf{X}_i + \omega_i \mathbf{c}_t^\top$, $i = 1, ..., m$, em que $\mathbf{c}_t$ denota um vetor de dimensão $p \times 1$ cuja p-ésima entrada é igual a 1 e as demais iguais a 0, na função Q. Neste caso, a dimensão de $\boldsymbol{\omega}$ será $\sum_{i=1}^{m} n_i = N$ e $\boldsymbol{\omega}_0 = \mathbf{0}_N$. Desta forma, tem-se

$$\begin{aligned}
\boldsymbol{\Delta}_\beta &= \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{N} \Big(-\mathbf{c}_t \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \widehat{\mathbf{E}}_i^{-1} + 2\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathbf{1}_{n_i} \mathbf{c}_t^\top \widehat{\boldsymbol{\beta}} \mathbf{1}_p^\top \\
&\quad + 2\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \mathbf{c}_t \omega_i^\top \widehat{\mathbf{E}}_i^{-1} \mathbf{1}_{n_i} \mathbf{c}_t^\top \widehat{\boldsymbol{\beta}} \mathbf{1}_{n_i}^\top \Big), \\
\boldsymbol{\Delta}_{\sigma^2} &= \frac{1}{\widehat{\sigma^4}} \sum_{i=1}^{N} \Big(\widehat{\boldsymbol{\beta}}^\top \mathbf{c}_t \mathbf{1}_{n_i}^\top \widehat{\mathbf{E}}_i^{-1} \mathcal{E}_{1i}(\boldsymbol{\theta}) \mathbf{1}_{n_i}^\top + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \widehat{\boldsymbol{\theta}}^\top \mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathbf{1}_{n_i} \mathbf{c}_t^\top \widehat{\boldsymbol{\beta}} \mathbf{1}_{n_i}^\top \\
&\quad + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \widehat{\boldsymbol{\beta}}^\top \mathbf{c}_t \omega_i^\top \widehat{\mathbf{E}}_i^{-1} \mathbf{1}_{n_i} \mathbf{c}_t^\top \widehat{\boldsymbol{\beta}} \mathbf{1}_{n_i}^\top \Big), \text{ e} \\
\boldsymbol{\Delta}_{\phi_j} &= -\frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{N} \Big(\widehat{\boldsymbol{\beta}}^\top \mathbf{c}_t \mathbf{1}_{n_i}^\top \widehat{\mathbf{E}}_i^{-1} d\phi_j \widehat{\mathbf{E}}_i^{-1} \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \mathbf{1}_{n_i}^\top - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} d\phi_j \widehat{\mathbf{E}}_i^{-1} \mathbf{1}_{n_i} \mathbf{c}_t^\top \widehat{\boldsymbol{\beta}} \mathbf{1}_{n_i}^\top \\
&\quad - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \widehat{\boldsymbol{\beta}}^\top \mathbf{c}_t \omega_i^\top \widehat{\mathbf{E}}_i^{-1} d\phi_j \widehat{\mathbf{E}}_i^{-1} \mathbf{1}_{n_i} \mathbf{c}_t^\top \widehat{\boldsymbol{\beta}} \mathbf{1}_{n_i}^T \Big), \ j = 1, 2.
\end{aligned}$$

Perturbação sobre os coeficientes: A perturbação nos $\boldsymbol{\beta}$'s é inserida substituindo $\boldsymbol{\beta}$ por $\boldsymbol{\beta}(\omega) = \boldsymbol{\beta}\omega_i$, $i = 1, ..., m$, $\boldsymbol{\omega} \in \mathbb{R}^m$ na função Q. Tem-se, neste caso, $\boldsymbol{\omega}_0 = \mathbf{1}_m^\top$. Então

$$\begin{aligned}
\boldsymbol{\Delta}_\beta &= \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{m} \left[\mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) - 2\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \omega_i \mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathbf{X}_i \widehat{\boldsymbol{\beta}} \right], \\
\boldsymbol{\Delta}_{\sigma^2} &= \frac{1}{\widehat{\sigma^4}} \sum_{i=1}^{m} \left[-\widehat{\boldsymbol{\beta}}^\top \mathbf{X}_i^\top \widehat{\mathbf{E}}_i^{-1} \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \widehat{\boldsymbol{\beta}}^\top \mathbf{X}_i^\top \omega_i \widehat{\mathbf{E}}_i^{-1} \mathbf{X}_i \widehat{\boldsymbol{\beta}} \right], \text{ e}
\end{aligned}$$

$$\boldsymbol{\Delta}_{\phi_j} = -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^{m} tr\left[\left(\mathbf{X}_i\widehat{\boldsymbol{\beta}}\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\mathbf{X}_i\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top\mathbf{X}_i^\top\omega_i\right)\widehat{\mathbf{E}}_i^{-1}d\phi_j\widehat{\mathbf{E}}_i^{-1}\right]$$

8.2 Estudo de simulação

O modelo utilizado nesta seção será

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i,\ i = 1, ..., m, \tag{8.1}$$

com $m = 100$, em que $\mathbf{Y}_i = (y_{i1}, ..., y_{in_i})^\top$ é um vetor de dimensão $n_i \times 1$ de respostas do i-ésimo indivíduo, medidas nos tempos $\mathbf{t}_i = (1, 3, 5, 7, 10, 12, 15)^\top$ e $\mathbf{X}_i$ é a matriz de desenho de dimensão $n_i \times p$ associada ao vetor de efeitos fixos $\boldsymbol{\beta} = (\beta_1, ..., \beta_7)^\top$. Para incorporar ao modelo a autocorrelação das observações repetidas, assumimos a estrutura de correlação DEC para os erros aleatórios. Assim, a matriz $\boldsymbol{\Sigma}_i$ será $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{E}_i$, com $\mathbf{E}_i = \sigma^2\left(\phi_1^{|t_{ij}-t_{ik}|^{\phi_2}}\right)$, $i = 1, ..., m$ e $j, k = 1, ..., n_i$. As estruturas de erros correlacionados (EC), AR(1), MA(1), simetria composta (SC) e independente (Ind) foram testadas. Foram utilizados os seguintes valores para os parâmetros: β=(2, 3, 4, 5, 6, 7, 8), σ^2=1, ϕ_1=0,8 e ϕ_2=1. Os erros aleatórios foram gerados de uma distribuição Normal e foram perturbadas as observações #1 e #700 ao serem substituídas pelos valores 0 e 15, respectivamente. Os modelos Normal e t de Student ($\nu = 7$) foram utilizados para o ajuste do modelo. O valor de ν foi obtido como o valor que maximizava a log-verossimilhança (Meza et al., 2012). Para gerar as observações censuradas à esquerda utilizou-se a proposta de Tsuyuguchi (2012). O modelo (8.1) foi gerado e o ponto de corte δ_i foi definido como o r-ésimo valor do vetor $\mathbf{Y}$ ordenado, ou seja, $\delta_i = Y_{(r)},\ i = 1; ..., r$ em que $r = N \times pc$ é o número de observações censuradas, pc é o percentual de censura desejado e $N = m \times p$ (p é a dimensão de t_i). Desta forma, $Y_{(r)}$ se torna uma observação censurada e todos os valores de $\mathbf{Y}$ menores ou iguais a ele vão assumir o mesmo valor que $Y_{(r)}$. O percentual de censura utilizado foi 20%.

A Tabela 8.1 apresenta os resultados do estudo de Monte Carlo de 100 réplicas para a estrutura de erros correlacionados, considerando as distribuições Normal e t de Student. A observação #100 foi influente na grande maioria das réplicas para o modelo Normal, considerando a distância de Cook e os esquemas de perturbação ponderação de casos, sobre o parâmetro de escala e sobre os coeficientes do modelo. A observação #1 foi classificada com influente em poucas réplicas, sempre com medida de influência próxima ao valor de referência, porque devido a forma de contaminação (seria uma observação na cauda à esquerda), este caso foi censurado na maior parte das réplicas. A distribuição t de Student foi menos influenciada por

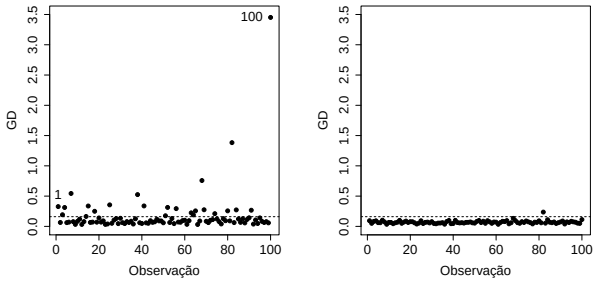
estas observações e, nos casos em que alguma delas foi considerada influente, foi com valores de medida de influência próximos ao limite de referência, sem caracterizar saltos como os observados na distribuição Normal. A Figura 8.1 mostra os gráficos para uma réplica, corroborando os resultados descritos acima.

O modelo foi estimado considerando as outras estruturas de correlação de interesse: AR(1), MA(1), simetria composta e independente. Os resultados da análise de influência foram similares ao obtido via estrutura de erros correlacionados e podem ser vistos no Apêndice B. Como no caso dos modelos de regressão lineares e não lineares, optamos por identificar somente as observações suspeitas de serem influentes.

Tabela 8.1: Estudo de simulação. Análise de influência via estudo de Monte Carlo para as observações #1 e #100 por distribuição e medida de diagnóstico: GD - distância generalizada de Cook, PC - ponderação de casos, ES - parâmetro de escala, VP - variável preditora e CO - coeficientes. Modelo para dados longitudinais com estrutura de correlação EC.

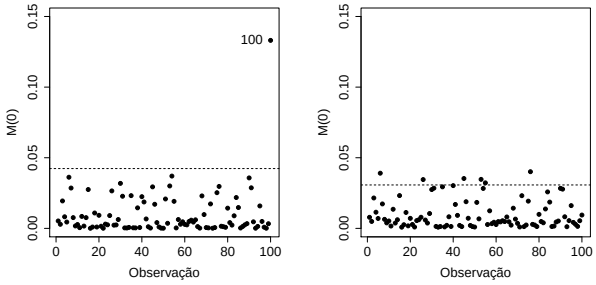
Medidas	**Estatísticas**	**Normal**		**t de Student**	
		#1	#100	#1	#100
GD	% Inf[a]	8%	100%	0%	4%
	M[b]	0,1559	2,9692	0,0712	0,1229
	DP[c]	0,5860	0,9816	0,0226	0,0166
	Ref[d]	0,16		0,16	
PC	% Inf	0%	100%	10%	0%
	M	0,0102	0,1781	0,0124	0,0134
	DP	0,0112	0,0559	0,0116	0,0061
	M (DP) Ref	0,0512 (0,0088)		0,0312 (0,0019)	
ES	% Inf	0%	100%	10%	0%
	M	0,0102	0,1776	0,0122	0,0149
	DP	0,0111	0,0548	0,0110	0,0058
	M (DP) Ref	0,0510 (0,0087)		0,0301 (0,0019)	
VP	% Inf	0%	0%	2%	8%
	M	0,0009	0,0013	0,0013	0,0016
	DP	0,0008	0,0012	0,0008	0,0011
	M (DP) Ref	0,0059 (0,0009)		0,0033 (0,0003)	
CO	% Inf	6%	46%	10%	0%
	M	0,0112	0,0337	0,0108	0,0021
	DP	0,0150	0,0234	0,0092	0,0013
	M (DP) Ref	0,0354 (0,0069)		0,0256 (0,0017)	

[a] % Inf: percentual de réplicas de Monte Carlo em que a observação foi considerada influente (maior que o valor de referência).
[b] M é a média das medidas de influência.
[c] DP é o desvio-padrão das medidas de influência.
[d] Ref é o valor de referência para considerar uma observação influente.



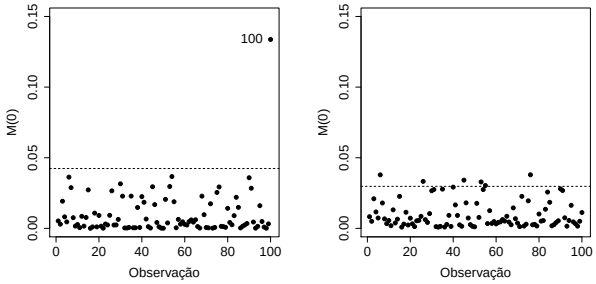
(a) GD - Normal

(b) GD - t de Student



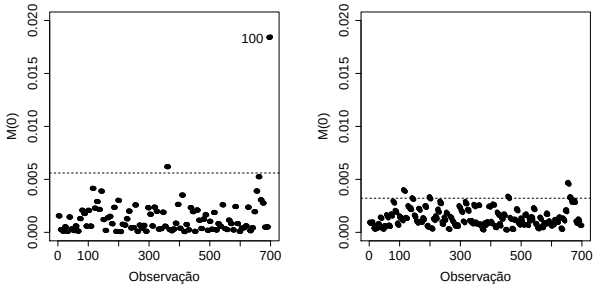
(c) PC - Normal

(d) PC - t de Student



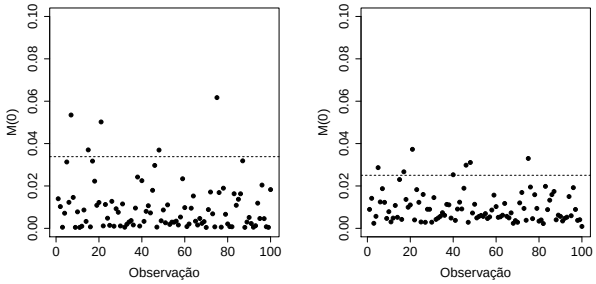
(e) ES - Normal

(f) ES - t de Student



(g) VP - Normal

(h) VP - t de Student



(i) CO - Normal

(j) CO - t de Student

Figura 8.1: Estudo de simulação. Medidas de influência considerando a estrutura de erros correlacionados, para as distribuições Normal e t de Student. Modelo longitudinal.

8.3 Aplicação

Nesta seção as medidas de diagnóstico propostas foram aplicadas ao banco de dados UTI. Este banco de dados consiste em um estudo longitudinal de interrupção de tratamento de 72 adolescentes infectados com o vírus HIV, provenientes de quatro instituições dos Estados Unidos. A medida de carga viral de HIV foi obtida para cada indivíduo em oito tempos diferentes após a interrupção do tratamento: $t_1 = 0$, $t_2 = 1$, $t_3 = 3$, $t_4 = 6$, $t_5 = 9$, $t_6 = 12$, $t_7 = 18$ e $t_8 = 24$ meses. Alguns pacientes não apresentaram todas as medidas, de modo que o número de medidas por indivíduo n_i pode variar de 1 a 8. A medida de carga viral é sujeita a censura se é menor que um limite inferior de detecção (50 cópias/ml). Este banco de dados contém 362 observações, das quais 26 (7,18%) foram censuradas à esquerda e está disponível no pacote `lmec` (Vaida e Liu, 2009a) do *software* R (R Core Team, 2015).

O seguinte modelo será utilizado para ajustar estes dados

$$\mathbf{Y}_i \;=\; \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \;\; \boldsymbol{\epsilon}_i \sim t_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i, \nu), \;\; i = 1, ..., m, \tag{8.2}$$

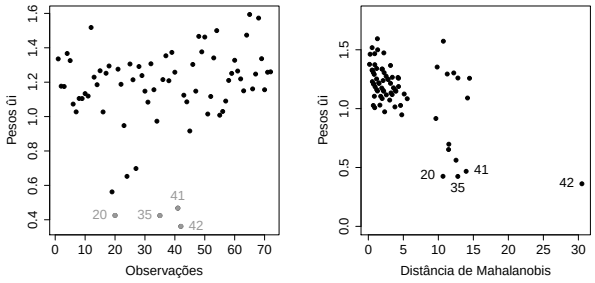
em que $\mathbf{Y}_i = (y_{i1}, ..., y_{in_i})^\top$ é o vetor de dimensão $n_i \times 1$ de carga viral de HIV do i-ésimo indivíduo, medidas nos tempos $\mathbf{t}_i = (t_{i1}, ..., t_{in_i})^\top$ e $\mathbf{X}_i$ é a matriz de desenho de dimensão $n_i \times p$ associada ao vetor de efeitos fixos $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$. Para acomodar o efeito das observações repetidas, assumimos a estrutura de correlação DEC para os erros aleatórios. Assim, a matriz $\boldsymbol{\Sigma}_i$ será $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{E}_i$, com $\mathbf{E}_i = \sigma^2\left(\phi_1^{|t_{ij}-t_{ik}|^{\phi_2}}\right)$, $i = 1, ..., m$ e $j, k = 1, ..., n_i$. Neste texto trabalhamos com os modelos de erros correlacionados (EC), AR(1), MA(1), simetria composta (SC) e independente (Ind). Ajustamos o modelo considerando erros aleatórios com distribuição Normal e t de Student ($\nu = 10$), como utilizado por Garay et al. (2014) para este banco de dados.

As estimativas fornecidas pelo processo de estimação estão apresentadas na Tabela 8.2. Observa-se resultados melhores dos critérios de comparação para o modelo t de Student (menores AIC e BIC e maiores log-verossimilhanças). Entre as estruturas de correlação, os critérios indicam melhor ajuste os modelos EC e SC. Os envelopes dos resíduos martingal transformado (Figura 8.3) mostram maior adequação do modelo t de Student EC.

Análises anteriores deste conjunto de dados apontaram como possíveis observações influentes os indivíduos #20, #35, #41 e #42 (Lachos et al., 2011; Garay et al., 2014). A Figura 8.2 mostra que estes pontos de destacam de fato, considerando a estimação realizada via modelo t de Student com estrutura de correlação EC. Para este estudo, construímos as análises de influência global e local para os modelos Normal e t de Student, considerando as estruturas de correlação EC, AR(1), MA(1), SC e Ind.

Tabela 8.2: Dados "UTI". Estimativas e erros padrão (em parênteses) para os parâmetros do modelo, segundo as distribuições Normal e t de Student e as diferentes estruturas de correlação: EC, AR(1), MA(1), SC, e Ind.

Parâmetros	EC	AR(1)	MA(1)	SC	Ind
			Normal		
β_1	3,6194 (0,0157)	3,6334 (0,0163)	3,6195 (0,0151)	3,6187 (0,0157)	3,6160 (0,0153)
β_2	4,1832 (0,0165)	4,2095 (0,0169)	4,1825 (0,0167)	4,1815 (0,0165)	4,1527 (0,0172)
β_3	4,2566 (0,0170)	4,2503 (0,0182)	4,2384 (0,0182)	4,2565 (0,0170)	4,2382 (0,0184)
β_4	4,3736 (0,0171)	4,3225 (0,0190)	4,3730 (0,0185)	4,3755 (0,0171)	4,3727 (0,0187)
β_5	4,5790 (0,0196)	4,4683 (0,0238)	4,3652 (0,0245)	4,5817 (0,0196)	4,3650 (0,0248)
β_6	4,5819 (0,0222)	4,3782 (0,0304)	4,2328 (0,0309)	4,5848 (0,0221)	4,2327 (0,0314)
β_7	4,6878 (0,0275)	4,3751 (0,0463)	4,3260 (0,0439)	4,6929 (0,0271)	4,3259 (0,0445)
β_8	4,8062 (0,0418)	4,5762 (0,0843)	4,5621 (0,0807)	4,8093 (0,0408)	4,5621 (0,0818)
σ^2	1,1059	1,1498	1,0487	1,1086	1,0631
ϕ_1	0,7029	0,8252	0,4069	0,6920	-
ϕ_2	0,0287	1,0000	∞	-	-
Log-veros.	-409,57	-460,32	-510,12	-409,70	-517,60
AIC	841,15	940,54	1.040,24	839,39	1.053,19
BIC	883,95	979,55	1.079,16	878,31	1.088,22
			t de Student ($\nu = 10$)		
β_1	3,6338 (0,0153)	3,6406 (0,0155)	3,6570 (0,0122)	3,6313 (0,0157)	3,6536 (0,0119)
β_2	4,2747 (0,0171)	4,3018 (0,0172)	4,2711 (0,0147)	4,2665 (0,0176)	4,2439 (0,0145)
β_3	4,3300 (0,0177)	4,3307 (0,0187)	4,3256 (0,0159)	4,3271 (0,0181)	4,3206 (0,0154)
β_4	4,4664 (0,0179)	4,4292 (0,0195)	4,4803 (0,0162)	4,4712 (0,0183)	4,4772 (0,0157)
β_5	4,6283 (0,0208)	4,5477 (0,0249)	4,5307 (0,0214)	4,6369 (0,0210)	4,5309 (0,0208)
β_6	4,6139 (0,0240)	4,4437 (0,0318)	4,3929 (0,0266)	4,6259 (0,0239)	4,3934 (0,0262)
β_7	4,6924 (0,0310)	4,4648 (0,0472)	4,5043 (0,0379)	4,7126 (0,0297)	4,5058 (0,0368)
β_8	4,7893 (0,0492)	4,6475 (0,0863)	4,6914 (0,0692)	4,8020 (0,0452)	4,6925 (0,0680)
σ^2	1,0071	1,0285	0,8008	1,0319	0,7926
ϕ_1	0,6892	0,7758	0,2738	0,6591	-
ϕ_2	0,1000	1,0000	∞	-	-
Log-veros.	-369,31	-421,21	-476,79	-369,85	-483,52
AIC	760,61	862,41	973,58	759,70	985,04
BIC	803,42	901,33	1.012,49	798,62	1.020,06



(a) Pesos "$\hat{u}_i$". **(b)** DM *vs* pesos "$\hat{u}_i$".

Figura 8.2: Dados "UTI". Esperanças condicionais $\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})$, também chamadas de pesos "$\hat{u}_i$", e distância de Mahalanobis (DM) *vs* Pesos "$\hat{u}_i$". Modelo t de Student com a estrutura de correlação EC.

A análise de influência global está apresentada na Figura 8.4. Para o modelo Normal estas observações influentes citadas acima apresentaram distância generalizada de Cook bem diferentes das demais observações, caracterizando saltos, para todas as estruturas de correlação. No caso do modelo t de Student, apesar de algumas observações ultrapassarem os limites de referência, as medidas de influência

destas observações foram mais parecidas com as medidas de influência das demais observações (ver Tabela 8.3).

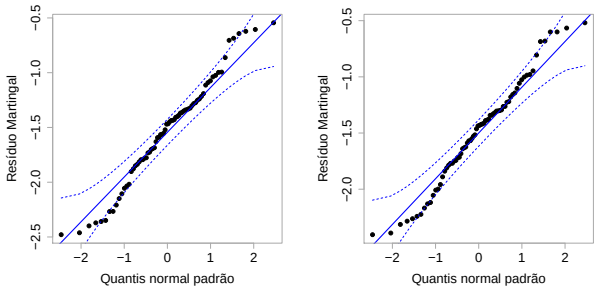
Tabela 8.3: Dados "`UTI`". Estatísticas descritivas das medidas de influência, segundo as estruturas de correlação EC, AR(1), MA(1), SC e Ind, e as distribuições Normal e t de Student.

Estatísticas	GD		PC		ES		VP		CO	
	N	t	N	t	N	t	N	t	N	t
					Estrutura de correlação EC					
Mínimo	0,0098	0,0093	$6{,}9e^{-4}$	$7{,}1e^{-4}$	$1{,}8e^{-4}$	$3{,}4e^{-4}$	$2{,}7e^{-4}$	$2{,}1e^{-4}$	0,0016	0,0015
Média	0,1976	0,1829	0,0139	0,0139	0,0139	0,0139	0,0028	0,0028	0,0139	0,0139
Desvio-padrão	0,3837	0,2480	0,0270	0,0188	0,0292	0,0209	0,0026	0,0015	0,0224	0,0103
Máximo	2,5320	1,1520	0,1780	0,0875	0,1920	0,0931	0,0163	0,0103	0,1672	0,0489
					Estrutura de correlação AR(1)					
Mínimo	0,0108	0,0089	$6{,}4e^{-4}$	$8{,}0e^{-4}$	$2{,}1e^{-4}$	$6{,}9e^{-4}$	$6{,}2e^{-5}$	$8{,}1e^{-5}$	$5{,}6e^{-4}$	$8{,}3e^{-4}$
Média	0,2347	0,1549	0,0139	0,0139	0,0139	0,0139	0,0028	0,0028	0,0139	0,0139
Desvio-padrão	0,8691	0,2289	0,0514	0,0205	0,0493	0,0218	0,0032	0,0026	0,0348	0,0167
Máximo	7,2670	1,4280	0,4300	0,1280	0,4056	0,1147	0,0212	0,0104	0,2931	0,0722
					Estrutura de correlação MA(1)					
Mínimo	0,0101	0,0112	0,0009	0,0016	0,0002	0,0008	$5{,}9e^{-5}$	$5{,}3e^{-5}$	0,0006	0,0006
Média	0,1575	0,0991	0,0139	0,0139	0,0139	0,0139	0,0028	0,0028	0,0139	0,0139
Desvio-padrão	0,3730	0,0702	0,0329	0,0098	0,0343	0,0121	0,0019	0,0022	0,0189	0,0110
Máximo	2,5050	0,3101	0,2209	0,0434	0,2334	0,0484	0,0166	0,0091	0,1410	0,0446
					Estrutura de correlação SC					
Mínimo	0,0098	0,0093	0,0008	0,0009	0,0002	0,0003	$5{,}1e^{-5}$	$1{,}9e^{-4}$	0,0029	0,0023
Média	0,1696	0,1381	0,0139	0,0139	0,0139	0,0139	0,0028	0,0028	0,0139	0,0139
Desvio-padrão	0,2965	0,1836	0,0243	0,0185	0,0263	0,0203	0,0022	0,0014	0,0169	0,0086
Máximo	1,8970	0,7567	0,1553	0,0761	0,1696	0,0838	0,0150	0,0120	0,1098	0,0392
					Estrutura de correlação Ind					
Mínimo	0,0100	0,0113	0,0009	0,0016	0,0002	0,0006	$5{,}5e^{-5}$	$5{,}1e^{-5}$	0,0006	0,0006
Média	0,1623	0,0988	0,0139	0,0139	0,0139	0,0139	0,0028	0,0028	0,0139	0,0139
Desvio-padrão	0,4080	0,0684	0,0349	0,0096	0,0363	0,0122	0,0019	0,0022	0,0189	0,0105
Máximo	2,7373	0,3072	0,2342	0,0432	0,2478	0,0530	0,0119	0,0090	0,1410	0,0437

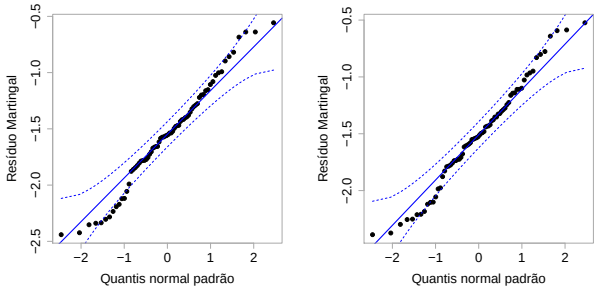
A análise de influência local foi realizada considerando os esquemas de perturbação ponderação de casos (PC), sobre o parâmetro de escala (ES), sobre uma variável preditora contínua (VP) e sobre os coeficientes (CO). As estatísticas descritivas das medidas de influência para estes esquemas, segundo as diferentes estruturas de correlação e os modelos Normal e t de Student estão apresentadas na Tabela 8.3. Para os esquemas PC, ES e CO, as médias das medidas de influência são todas iguais a 0,0139, correspondendo ao inverso da dimensão do vetor de perturbação $\boldsymbol{\omega}$, 72. No caso do esquema de perturbação sobre uma variável preditora, a dimensão do vetor de perturbação é igual à 362, de modo que a média das medidas de influência é diferente para este esquema.

A representação gráfica das medidas de influência para os quatro esquema de perturbação avaliados, segundo as diferentes estruturas de correlação e os modelos Normal e t de Student estão apresentadas nas Figuras de 8.5 a 8.8. O comportamento das observações #20, #35, #41 e #42 são parecidos ao observado no caso da distância de Cook. Para o modelo Normal, estas observações, em geral, apresentam

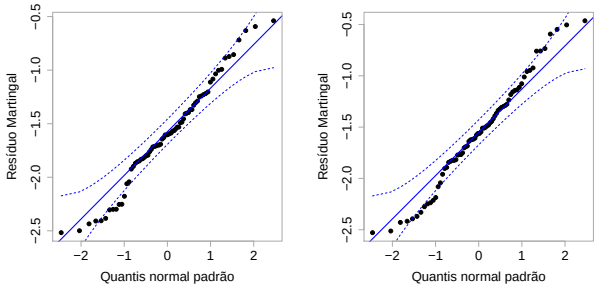
medidas de influência bem diferentes das demais. No caso do modelo t de Student as medidas de influência destas observações são parecidas com as das demais observações. Deste modo, mostramos que o modelo t de Student consegue acomodar melhor o efeito da influência destas observações, de modo que elas afetam menos as estimativas fornecidas por este modelo.



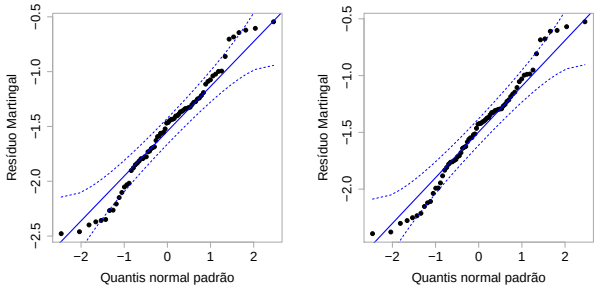
(a) Normal - EC **(b)** t de Student - EC



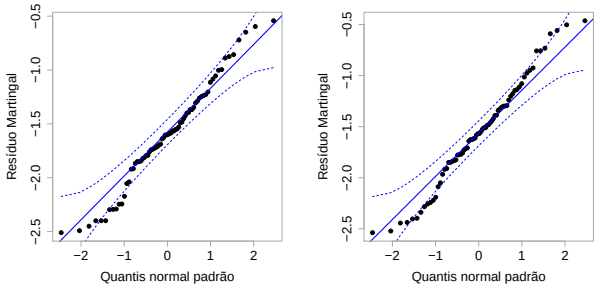
(c) Normal - AR(1) **(d)** t de Student - AR(1)



(e) Normal - MA(1) **(f)** t de Student - MA(1)

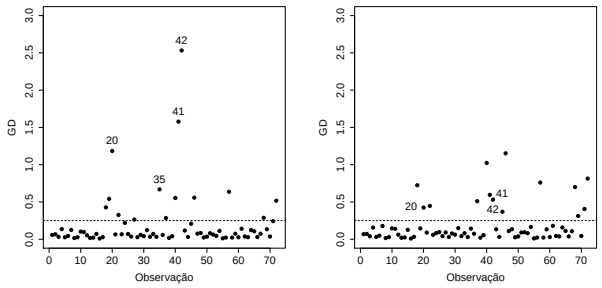


(g) Normal - SC **(h)** t de Student - SC



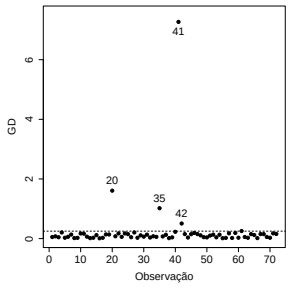
(i) Normal - Ind **(j)** t de Student - Ind

Figura 8.3: Dados "`UTI`". Envelopes para os resíduos martingal transformados segundo as estruturas de correlação EC, AR(1), MA(1), SC e Ind, para as distribuições Normal e t de Student.

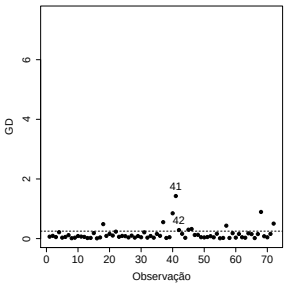


(a) Normal - EC

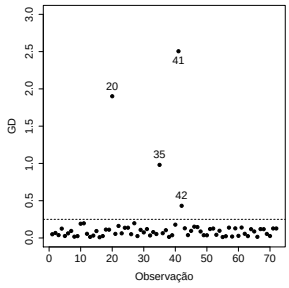
(b) t de Student - EC



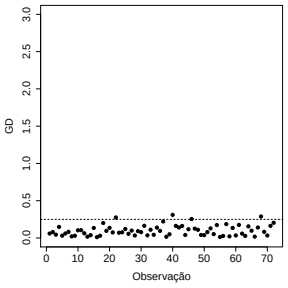
(c) Normal - AR(1)



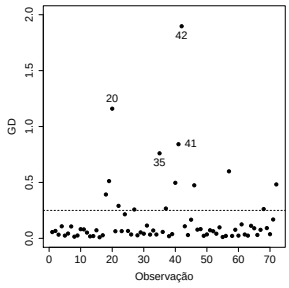
(d) t de Student - AR(1)



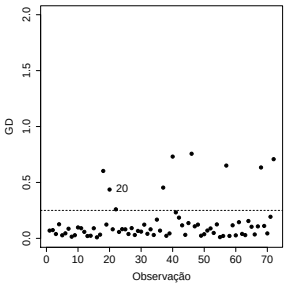
(e) Normal - MA(1)



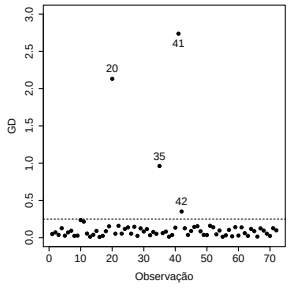
(f) t de Student - MA(1)



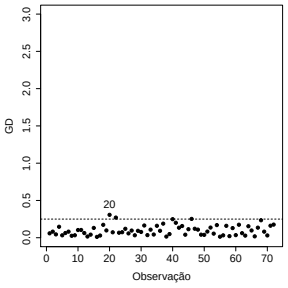
(g) Normal - SC



(h) t de Student - SC

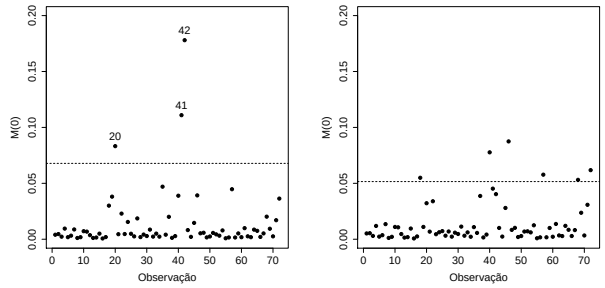


(i) Normal - Ind

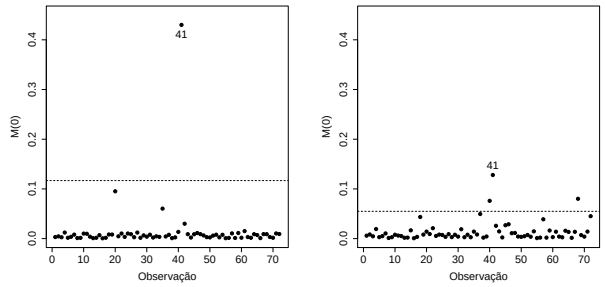


(j) t de Student - Ind

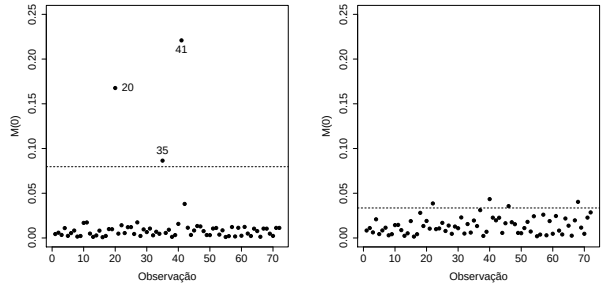
Figura 8.4: Dados "UTI". Distância generalizada de Cook segundo as estruturas de correlação EC, AR(1), MA(1), SC e Ind, para as distribuições Normal e t de Student.



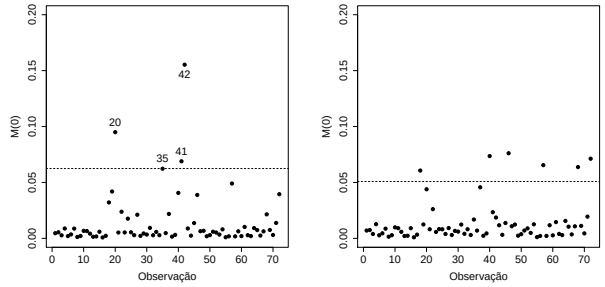
(a) Normal - EC **(b)** t de Student - EC



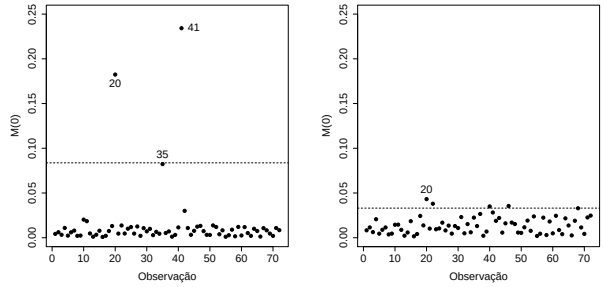
(c) Normal - AR(1) **(d)** t de Student - AR(1)



(e) Normal - MA(1) **(f)** t de Student - MA(1)

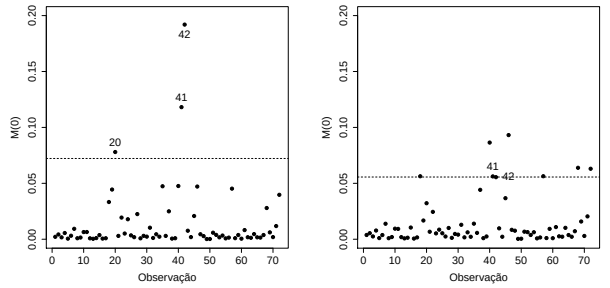


(g) Normal - SC **(h)** t de Student - SC



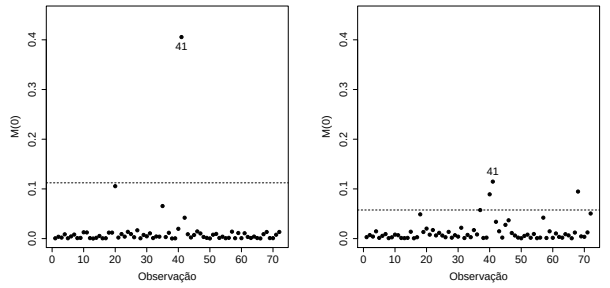
(i) Normal - Ind **(j)** t de Student - Ind

Figura 8.5: Dados "`UTI`". Perturbação ponderação de casos segundo as estruturas de correlação EC, AR(1), MA(1), SC e Ind, para as distribuições Normal e t de Student.



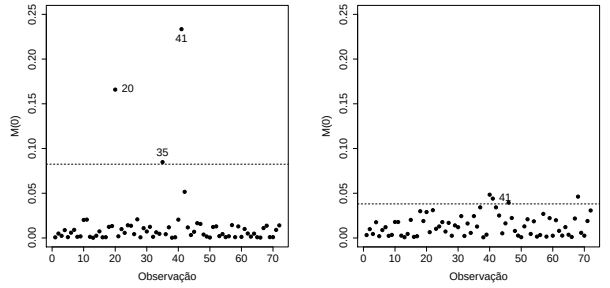
(a) Normal - EC

(b) t de Student - EC



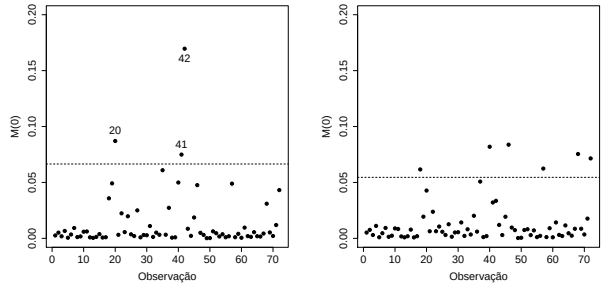
(c) Normal - AR(1)

(d) t de Student - AR(1)



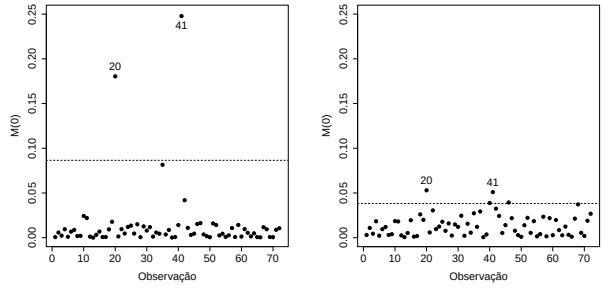
(e) Normal - MA(1)

(f) t de Student - MA(1)



(g) Normal - SC

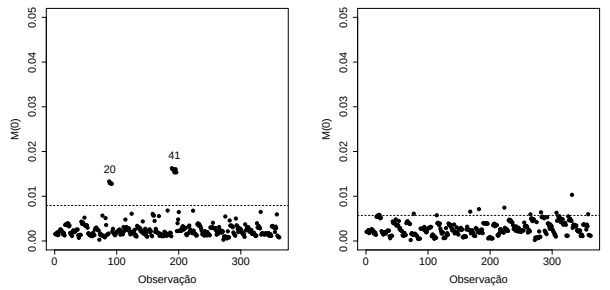
(h) t de Student - SC



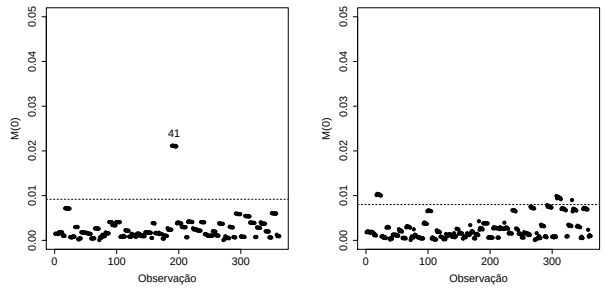
(i) Normal - Ind

(j) t de Student - Ind

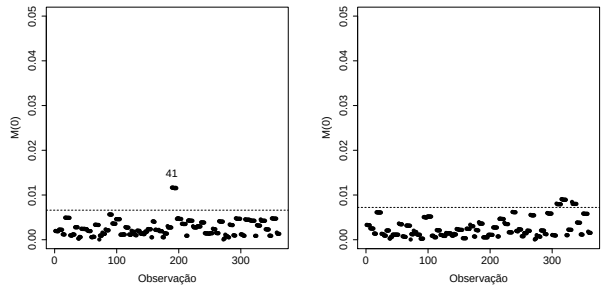
Figura 8.6: Dados "UTI". Perturbação sobre o parâmetro de escala segundo as estruturas de correlação EC, AR(1), MA(1), SC e Ind, para as distribuições Normal e t de Student.



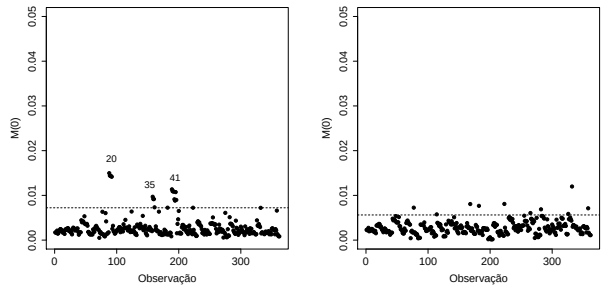
(a) Normal - EC **(b)** t de Student - EC



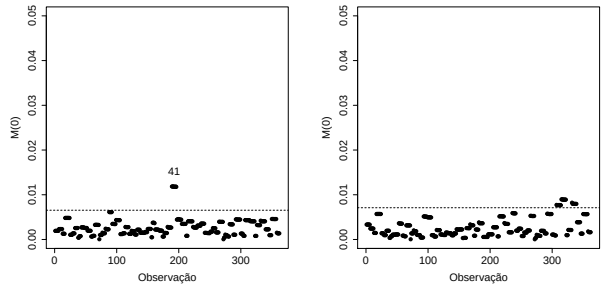
(c) Normal - AR(1) **(d)** t de Student - AR(1)



(e) Normal - MA(1) **(f)** t de Student - MA(1)

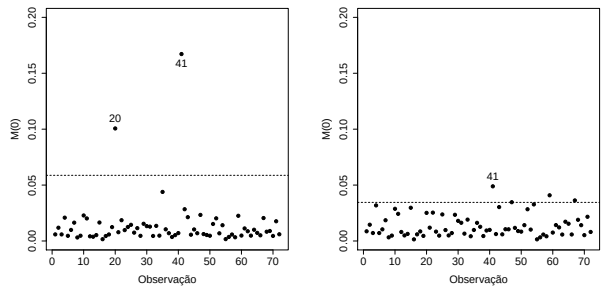


(g) Normal - SC **(h)** t de Student - SC



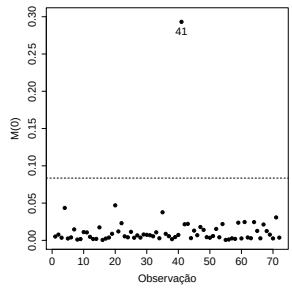
(i) Normal - Ind **(j)** t de Student - Ind

Figura 8.7: Dados "UTI". Perturbação sobre uma variável preditora contínua segundo as estruturas de correlação EC, AR(1), MA(1), SC e Ind, para as distribuições Normal e t de Student.

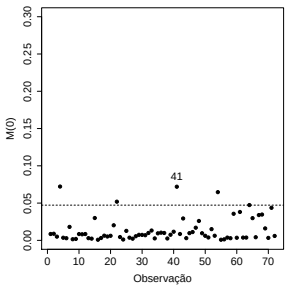


(a) Normal - EC

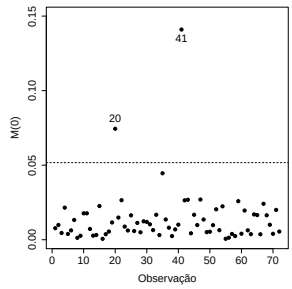
(b) t de Student - EC



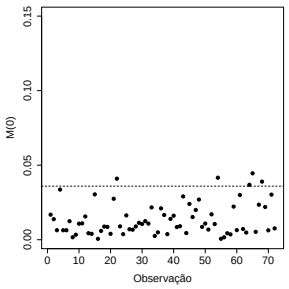
(c) Normal - AR(1)



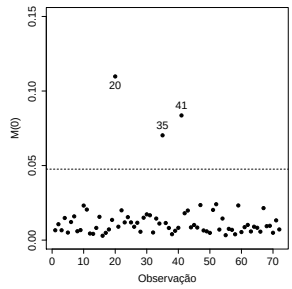
(d) t de Student - AR(1)



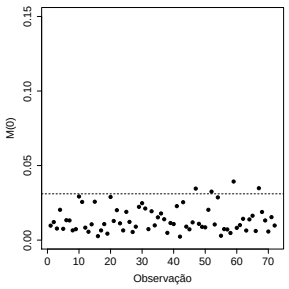
(e) Normal - MA(1)



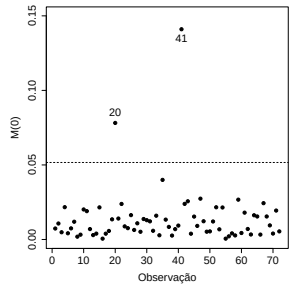
(f) t de Student - MA(1)



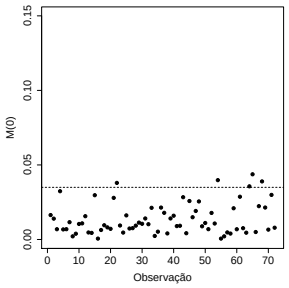
(g) Normal - SC



(h) t de Student - SC



(i) Normal - Ind



(j) t de Student - Ind

Figura 8.8: Dados "UTI". Perturbação sobre os coeficientes do modelo segundo as estruturas de correlação EC, AR(1), MA(1), SC e Ind, para as distribuições Normal e t de Student.

CAPÍTULO 9

Conclusões e trabalhos futuros

9.1 Conclusões

A análise de influência aprofunda o entendimento de um modelo de regressão. Por ser realizada após o ajuste, esta etapa proporciona um refinamento da técnica, identificando possíveis observações influentes em pontos específicos da formulação do modelo. A importância desta análise é justificada pela variabilidade característica dos fenômenos aleatórios.

Nesta tese foram propostas medidas de influência global e local para modelos de regressão linear (para dados longitudinais e para respostas univariadas) e não linear (com respostas univariadas), para dados censurados. Foram utilizadas distribuições Normal e de caudas pesadas, como as da família NI (t de Student, Slash e Normal Contaminada). Os estudos de simulação mostraram que as técnicas propostas conseguem captar corretamente possíveis *outliers* e que as distribuições de caudas pesadas acomodaram melhor o efeito destas observações atípicas. Em especial, o desempenho das distribuições Slash e t de Student foi superior. A técnica proposta foi também aplicada a conjuntos de dados reais, analisados previamente por outros pesquisadores, e confirmando os resultados da literatura.

A análise da sensibilidade da constante ς que define os limites de referência da análise de influência local é uma importante contribuição desta tese, assim como a constatação de que as medidas de influência global e local podem ser aplicadas com sucesso a modelos complexos. Como limitação do estudo, os resultados apresentados para o modelo longitudinal são válidos para dados que tenham pelo menos 6 medidas para algum indivíduo da amostra, podendo haver outros indivíduos com números menores de medidas. Há um erro numérico no método em dados com menos de 6 repetições para todos os indivíduos.

A utilização da análise de influência permitiu a proposição de modelos mais adequados às características dos dados para as quais o modelo Normal foi ineficaz, por

exemplo, o ajuste de modelos com *outliers*. Eliminou a necessidade de transformações que visam o alcance da normalidade e/ou exclusões de observações influentes, que podem ser oriundas de dados genuínos. Deste modo, houve um ganho em se complexificar o modelo utilizando distribuições de caudas pesadas. Este trabalho representa mais um pequeno passo na compreensão da utilização de modelos teóricos para representar fenômenos reais complexos.

9.2 Trabalhos futuros

Em trabalhos futuros podem ser trabalhadas as medidas de diagnóstico para modelos de efeitos mistos considerando misturas de escala skew-normal (Lachos et al., 2010). Outra possível extensão seria a modelos semiparamétricos censurados (Ibacache-Pulgar et al., 2013). Podem ainda ser avaliados outros esquemas de perturbação, por exemplo, na variável resposta.

Lista de referências

Andrews DF, Pregibon D. Finding the outliers that matter. Journal of the Royal Statistical Society, Series B. 1978; 40: 85−93.

Anscombe FJ. Examination of residuals. Fourth Berkeley Symposium, University of California Press. 1961; 1: 1−36.

Arellano-Valle RB, Bolfarine H. On some characterizations of the t-distribution. Statistics and Probability Letters. 1995; 25: 79−85.

Atkinson AC. Regression diagnostics, transformations and constructed variables. Journal of the Royal Statistical Society, Series B. 1982; 44: 1−36.

Box GEP, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society, Series B. 1964; 26: 211−252.

Cook RD. Detection of influential observation in linear regression. Technometrics. 1977; 19: 15−18.

Cook RD. Assessment of local influence. Journal of the Royal Statistical Society, Series B. 1986; 48: 133−169.

Cook RD, Weisberg S. Residual and Influence in Regression. Londres: Chapman and Hall; 1982.

Cox DR, Snell EJ. A general definition of residuals. Journal of the Royal Statistical Society, Series B. 1968; 30: 248−275.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39: 1−38.

Galton F. Natural inheritance. Nova Yorque: Macmillan and Company; 1894.

Garay AM. Modelos de regressão para dados censurados sob distribuições simétricas. Tese de doutorado. Universidade de São Paulo. São Paulo; 2014.

Garay AM, Castro LM, Leskow J, Lachos, VH. Censored linear regression models for irregularly observed longitudinal data using the multivariate-t distribution. Statistical Methods in Medical Research. 2014. Doi: 10.1177/0962280214551191

Garay AM, Lachos VH, Bolfarine H, Cabral CRB. Linear censored regression models with scale mixtures of normal distributions. Statistical Papers. 2015(a). Doi: 10.1007/s00362-015-0696-9.

Garay AM, Lachos VH, Lin T. Nonlinear censored regression models with heavy-tailed distributions. Statistics and Its Interface. 2016; 9: 281−293.

Garay AM, Lachos VH, Massuia MB. SMNCensReg: Fitting univariate censored regression model under the scale mixture of normal distributions. R package version 3.0. 2015(b). http://CRAN.R-project.org/package=SMNCensReg.

Heuchenne C, Keilegom IV. Nonlinear regression with censored data. Technometrics. 2007; 49: 34−44.

Ho HJ, Lin T, Chen H, Wang W. Some results on the truncated multivariate t distribution. Journal of Statistical Planning and Inference. 2012; 142: 25−40.

Ibacache-Pulgar G, Paula GA, Cysneiros FJA. Semiparametric additive models under symmetric distributions. Test. 2013; 22: 103−121.

Johnson W, Geisser S. A predictive view of the detection and characterization of influential observations in regression analysis. Journal of the American Statistical Association. 1983; 78: 137−144.

Lachos VH, Ghosh P, Arellano-Valle RB. Likelihood based inference for skew-normal/independent linear mixed model. Statistica Sinica. 2010; 20: 303−322.

Lachos VH, Bandyopadhyay D, Dey DK. Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. Biometrics. 2011; 67: 1594−1604.

Lange KL, Little RJA, Taylor JMG. Robust statistical modeling using t distribution. Journal of the American Statistical Association. 1989; 84: 881−896.

Lange KL, Sinsheimer JS. Normal/independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics. 1993; 2: 175−198.

Lee S, Xu L. Infuence analyses of nonlinear mixed-effects models. Computational Statistics and Data Analysis. 2004; 45: 321−341.

Liu C. Bayesian robust multivariate linear regression with incomplete data. Journal of the American Statistical Association. 1996; 91: 1219−1227.

Liu C, Rubin DB. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika. 1994; 80: 267−278.

Massuia MB, Cabral CRB, Matos LA, Lachos VH. Influence diagnostics for student-t censored linear regression models. Statistics. 2015; 49: 1074−1094.

Matos LA, Prates MO, Chen MH, Lachos VH. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. Statistica Sinica. 2013; 23: 1323−1342.

Meng X, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993; 80: 268−278.

Meza C, Osorio F, De la Cruz R. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. Statistics and Computing. 2012; 22: 121−139.

Mroz TA. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Econometrica. 1987; 55: 765−799.

Munoz A, Carey V, Schouten JP, Segal M, Rosner B. A parametric family of correlation structures for the analysis of longitudinal data. Biometrics. 1992; 48: 733−742.

Nelson WB. Acelerated testing: statistical models, test plans and data analysis. John Wiley; 2004.

Osorio F. Diagnóstico de infuência em modelos elípticos com efeitos mistos. Tese de doutorado. Universidade de São Paulo. São Paulo; 2006.

Osorio F, Paula GA, Galea M. Assessment of local influence in elliptical linear models with longitudinal structure. Computational Statistics and Data Analysis. 2007; 51: 4354−4368.

Pearson K. Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia. Philosophical Transactions of the Royal Society of London. 1896; 187: 253−318.

Pearson K. Francis Galton: a centenary appreciation. Cambridge University Press; 1922.

Pearson K. The life, letters and labors of Francis Galton. Cambridge University Press; 1930.

Poon WY, Poon YS. Conformal normal curvature and assessment of local influence. Journal of the Royal Statistical Society, Series B. 1999; 61: 51−61.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2015. (http://www.R-project.org/)

Rosa GJM, Padovani CR, Gianola D. Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. Biometrical Journal. 2003; 45: 573−590.

Russo CM, Paula GA, Aoki R. Influence diagnostics in nonlinear mixed-effects elliptical models. Computational Statistics and Data Analysis. 2009; 53: 4143−4156.

Srikantan KS. Testing for a single outlier in a regression model. Sankhyã A. 1961; 23: 251−260.

Stanton JM. Galton, Pearson, and the peas: a brief history of linear regression for statistics instructors (online). Journal of Statistics Education. 2001.

Sun J, Kabán A, Garibaldi JM. Robust mixture clustering using Pearson type VII distribution. Pattern Recognition Letters. 2010; 31: 2447−2454.

Tsuyuguchi AB. Testes de bondade de ajuste para a distribuição Birnbaum-Saunders. Dissertação de mestrado. Universidade Federal de Campina Grande; 2012.

Vaida F, Liu L. lmec: Linear Mixed-Effects Models with Censored Responses. R package version 1.0. 2009(a). http://CRAN.R-project.org/package=lmec.

Vaida F, Liu L. Fast implementation for normal mixed effects models with censored response. Journal of Computational and Graphical Statistics. 2009(b); 18: 797−817.

Wang W. Multivariate t linear mixed models for irregularly observed multiple repeated measures with missing outcomes. Biometrical Journal. 2013; 55: 554−571.

Wang J, Genton MG. The multivariate skew-slash distribution. Journal of Statistical Planning and Inference. 2006; 136: 209−220.

Zeller CB, Labra FV, Lachos VH, Balakrishnan N. Influence analyses of skewnormal/independent linear mixed models. Computational Statistics and Data Analysis. 2010; 54: 1266−1280.

Zhu H, Lee S. Local influence for incomplete-data models. Journal of the Royal Statistical Society, Series B. 2001; 63: 111−126.

Zhu H, Lee S, Wei B, Zhou J. Case-deletion measures for models with incomplete data. Biometrika. 2001; 88: 727−737.

APÊNDICE A

Resultados complementares referentes ao Capítulo 7

A.1 Simulações

A função não linear utilizada no estudo de simulação foi

$$\mu_i = \eta(\boldsymbol{\beta}, x_i) \quad = \quad \frac{\beta_1}{1 + exp(\beta_2 + \beta_3 x_i)},\ i = 1, ..., n.$$

As entradas do vetor de derivadas parciais de primeira ordem desta função são

$$\begin{aligned}
\frac{\partial \mu_i}{\partial \beta_1} &= \frac{1}{\psi(\boldsymbol{\beta}, x_i)}, \\
\frac{\partial \mu_i}{\partial \beta_2} &= -\frac{\hat{\beta}_1 \xi(\boldsymbol{\beta}, x_i)}{[\psi(\boldsymbol{\beta}, x_i)]^2}, \text{ e} \\
\frac{\partial \mu_i}{\partial \beta_3} &= -\frac{x_i \hat{\beta}_1 \xi(\boldsymbol{\beta}, x_i)}{[\psi(\boldsymbol{\beta}, x_i)]^2}.
\end{aligned}$$

E as entradas das derivadas parciais de segunda ordem de μ_i são

$$\begin{aligned}
\frac{\partial^2 \mu_i}{\partial \beta_1 \beta_1} &= 0, \\
\frac{\partial^2 \mu_i}{\partial \beta_1 \beta_2} &= -\frac{\xi(\boldsymbol{\beta}, x_i)}{[\psi(\boldsymbol{\beta}, x_i)]^2}, \\
\frac{\partial^2 \mu_i}{\partial \beta_1 \beta_3} &= -\frac{x_i \xi(\boldsymbol{\beta}, x_i)}{[\psi(\boldsymbol{\beta}, x_i)]^2},
\end{aligned}$$

$$\frac{\partial^2 \mu_i}{\partial \beta_2 \beta_2} = \frac{1}{[\psi(\boldsymbol{\beta}, x_i)]^3} \left\{ \hat{\beta}_1 \xi(\boldsymbol{\beta}, x_i) \left[2\xi(\boldsymbol{\beta}, x_i) - \psi(\boldsymbol{\beta}, x_i) \right] \right\}$$

$$\frac{\partial^2 \mu_i}{\partial \beta_2 \beta_3} = \frac{1}{[\psi(\boldsymbol{\beta}, x_i)]^3} \left\{ \hat{\beta}_1 x_i \xi(\boldsymbol{\beta}, x_i) \left[2\xi(\boldsymbol{\beta}, x_i) - \psi(\boldsymbol{\beta}, x_i) \right] \right\}, \text{ e}$$

$$\frac{\partial^2 \mu_i}{\partial \beta_3 \beta_3} = \frac{1}{[\psi(\boldsymbol{\beta}, x_i)]^3} \left\{ \hat{\beta}_1 x_i^2 \xi(\boldsymbol{\beta}, x_i) \left[2\xi(\boldsymbol{\beta}, x_i) - \psi(\boldsymbol{\beta}, x_i) \right] \right\}$$

em que $\xi(\boldsymbol{\beta}, x_i) = exp(\hat{\beta}_2 + \hat{\beta}_3 x_i)$ e $\psi(\boldsymbol{\beta}, x_i) = 1 + exp(\hat{\beta}_2 + \hat{\beta}_3 x_i)$.

A seguir as expressões de $\boldsymbol{\Delta_\omega}$ para os esquemas de perturbação sobre uma variável preditora contínua e sobre um coeficiente do modelo.

Perturbação sobre uma variável preditora: É inserida através da substituição $x_i(\omega_i) = x_i + \omega_i$ na função Q perturbada, dentro da função não linear $\eta(\boldsymbol{\beta}, x_i) = \frac{\beta_1}{1+exp(\beta_2+\beta_3 x_i)}$. Neste caso, $\boldsymbol{\omega} \in \mathbb{R}^n$ e $\boldsymbol{\omega}_0 = \mathbf{0}_n^\top$. Desta forma, tem-se

$$\boldsymbol{\Delta}_{\beta_1} = \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{n} \left\{ \frac{-\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_3 \xi(\boldsymbol{\beta}, x_i, \omega_i)}{[\psi(\boldsymbol{\beta}, x_i, \omega_i)]^2} + \frac{2\hat{\beta}_1 \hat{\beta}_3 \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \xi(\boldsymbol{\beta}, x_i, \omega_i)}{[\psi(\boldsymbol{\beta}, x_i, \omega_i)]^3} \right\},$$

$$\boldsymbol{\Delta}_{\beta_2} = \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{n} \left\{ \frac{\hat{\beta}_1 \hat{\beta}_3 \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \xi(\boldsymbol{\beta}, x_i, \omega_i) \left[2\xi(\boldsymbol{\beta}, x_i, \omega_i) - \psi(\boldsymbol{\beta}, x_i, \omega_i) \right]}{[\psi(\boldsymbol{\beta}, x_i, \omega_i)]^3} \right.$$

$$\left. + \frac{\hat{\beta}_1^2 \hat{\beta}_3 \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \xi(\boldsymbol{\beta}, x_i, \omega_i) \left\{ 3\xi(\boldsymbol{\beta}, x_i, \omega_i) - \psi(\boldsymbol{\beta}, x_i, \omega_i) \right\}}{[\psi(\boldsymbol{\beta}, x_i, \omega_i)]^4} \right\},$$

$$\boldsymbol{\Delta}_{\beta_3} = \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{n} \left\{ \frac{1}{[\psi(\boldsymbol{\beta}, x_i, \omega_i)]^4} \left[\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1^2 \xi(\boldsymbol{\beta}, x_i, \omega_i) \right.\right.$$

$$\left. \times \left[\psi(\boldsymbol{\beta}, x_i, \omega_i) \left(\hat{\beta}_3 (x_i + \omega_i) + 1 \right) - 3\hat{\beta}_3 (x_i + \omega_i) \xi(\boldsymbol{\beta}, x_i, \omega_i) \right] \right]$$

$$- \frac{1}{\psi(\boldsymbol{\beta}, x_i, \omega_i)} \left\{ \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1 \xi(\boldsymbol{\beta}, x_i, \omega_i) \left[\psi(\boldsymbol{\beta}, x_i, \omega_i) \left(\hat{\beta}_3 (x_i + \omega_i) + 1 \right) \right.\right.$$

$$\left.\left.\left. - 2\hat{\beta}_3 (x_i + \omega_i) \xi(\boldsymbol{\beta}, x_i, \omega_i) \right] \right\} \right\}, \text{ e}$$

$$\boldsymbol{\Delta}_{\sigma^2} = \frac{1}{\widehat{\sigma^4}} \sum_{i=1}^{n} \left\{ \frac{\hat{\beta}_1 \hat{\beta}_3 \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \xi(\boldsymbol{\beta}, x_i, \omega_i)}{[\psi(\boldsymbol{\beta}, x_i, \omega_i)]^2} - \frac{\hat{\beta}_1^2 \hat{\beta}_3 \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \xi(\boldsymbol{\beta}, x_i, \omega_i)}{[\psi(\boldsymbol{\beta}, x_i, \omega_i)]^3} \right\},$$

em que $\xi(\boldsymbol{\beta}, x_i, \omega_i) = exp(\hat{\beta}_2 + \hat{\beta}_3 (x_i + \omega_i))$ e $\psi(\boldsymbol{\beta}, x_i, \omega_i) = 1 + exp(\hat{\beta}_2 + \hat{\beta}_3 (x_i + \omega_i))$.

Perturbação sobre um coeficiente: A perturbação foi inserida substituindo-se especificamente o coeficiente β_1 por $\beta_1(\omega_i) = \beta_1\omega_i$ na função Q perturbada. Deste modo, $\boldsymbol{\omega} \in \mathbb{R}^n$ e $\boldsymbol{\omega}_0 = \mathbf{1}_n^\top$. Neste caso,

$$\begin{aligned}
\boldsymbol{\Delta}_{\beta_1} &= -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^{n}\left[\frac{2\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\hat{\beta}_1\omega_i}{[\psi(\boldsymbol{\beta},x_i)]^2} - \frac{\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})}{\psi(\boldsymbol{\beta},x_i)}\right], \\
\boldsymbol{\Delta}_{\beta_2} &= \frac{1}{\widehat{\sigma^2}}\sum_{i=1}^{n}\left[\frac{2\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\hat{\beta}_1^2\omega_i\xi(\boldsymbol{\beta},x_i)}{[\psi(\boldsymbol{\beta},x_i)]^3} - \frac{\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\hat{\beta}_1\xi(\boldsymbol{\beta},x_i)}{[\psi(\boldsymbol{\beta},x_i)]^2}\right], \\
\boldsymbol{\Delta}_{\beta_3} &= -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^{n}\left[\frac{\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\hat{\beta}_1x_i\xi(\boldsymbol{\beta},x_i)}{[\psi(\boldsymbol{\beta},x_i)]^2} - \frac{2\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\hat{\beta}_1^2\omega_ix_i\xi(\boldsymbol{\beta},x_i)}{[\psi(\boldsymbol{\beta},x_i)]^3} -\right], \text{ e} \\
\boldsymbol{\Delta}_{\sigma^2} &= \frac{1}{\widehat{\sigma^4}}\sum_{i=1}^{n}\left[\frac{\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}})\hat{\beta}_1^2\omega_i}{[\psi(\boldsymbol{\beta},x_i)]^2} - \frac{\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}})\hat{\beta}_1}{\psi(\boldsymbol{\beta},x_i)}\right],
\end{aligned}$$

em que $\xi(\boldsymbol{\beta},x_i) = exp(\hat{\beta}_2 + \hat{\beta}_3x_i))$ e $\psi(\boldsymbol{\beta},x_i) = 1 + exp(\hat{\beta}_2 + \hat{\beta}_3x_i)$.

A.2 Aplicação

A função não linear utilizada na aplicação aos dados de deformação de metais foi

$$\mu_i = \eta(\boldsymbol{\beta},x_i) = \beta_1 exp(\beta_2x_i),\ i = 1,...,n.$$

As entradas do vetor de derivadas parciais de primeira ordem desta função são

$$\begin{aligned}
\frac{\partial\mu_i}{\partial\beta_1} &= exp(\beta_2x_i), \text{ e} \\
\frac{\partial\mu_i}{\partial\beta_2} &= \beta_1x_iexp(\beta_2x_i).
\end{aligned}$$

E as entradas das derivadas parciais de segunda ordem de μ_i são

$$\begin{aligned}
\frac{\partial^2\mu_i}{\partial\beta_1\beta_1} &= 0, \\
\frac{\partial^2\mu_i}{\partial\beta_1\beta_2} &= x_iexp(\beta_2x_i), \text{ e} \\
\frac{\partial^2\mu_i}{\partial\beta_2\beta_2} &= \beta_1x_i^2exp(\beta_2x_i).
\end{aligned}$$

A seguir as expressões de $\boldsymbol{\Delta}_{\boldsymbol{\omega}}$ para os esquemas de perturbação sobre uma va-

riável preditora contínua e sobre os coeficientes do modelo para a aplicação.

Perturbação sobre uma variável preditora: É inserida através da substituição $x_i(\omega_i) = x_i + \omega_i$ na função Q perturbada, dentro da função não linear $\eta(\boldsymbol{\beta}, x_i) = \beta_1 exp(\beta_2 x_i)$. Neste caso, $\boldsymbol{\omega} \in \mathbb{R}^n$ e $\boldsymbol{\omega}_0 = \mathbf{0}_n^\top$. Desta forma, tem-se

$$\begin{aligned}
\boldsymbol{\Delta}_{\beta_1} &= \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{n} \left\{ \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_2 e^{\hat{\beta}_2(x_i+\omega_i)} - 2\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1 \hat{\beta}_2 e^{2\hat{\beta}_2(x_i+\omega_i)} \right\}, \\
\boldsymbol{\Delta}_{\beta_2} &= \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{n} \left\{ \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1 e^{\hat{\beta}_2(x_i+\omega_i)} \left(1 + \hat{\beta}_2(x_i+\omega_i)\right) \right. \\
&\quad \left. - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1^2 e^{2\hat{\beta}_2(x_i+\omega_i)} \left(1 + 2\hat{\beta}_2(x_i+\omega_i)\right) \right\} \text{ e} \\
\boldsymbol{\Delta}_{\sigma^2} &= \frac{1}{2\widehat{\sigma^4}} \sum_{i=1}^{n} \left\{ -\mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1 \hat{\beta}_2 e^{\hat{\beta}_2(x_i+\omega_i)} + \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1 \hat{\beta}_2 e^{2\hat{\beta}_2(x_i+\omega_i)} \right\}.
\end{aligned}$$

Perturbação sobre os coeficientes: A perturbação foi inserida substituindo-se os β_i's, $i = 1, 2$, por $\beta_i(\omega_i) = \beta_i \omega_i$ na função Q perturbada. Deste modo, $\boldsymbol{\omega} \in \mathbb{R}^n$ e $\boldsymbol{\omega}_0 = \mathbf{1}_n^\top$. Neste caso,

$$\begin{aligned}
\boldsymbol{\Delta}_{\beta_1} &= \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{n} \left\{ \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) e^{\hat{\beta}_2 x_i \omega_i} \left(1 + \hat{\beta}_2 x_i \omega_i\right) - 2\mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1 \omega_i e^{2\hat{\beta}_2 x_i \omega_i} \left(1 + 2\hat{\beta}_2 x_i \omega_i\right) \right\}, \\
\boldsymbol{\Delta}_{\beta_2} &= \frac{1}{\widehat{\sigma^2}} \sum_{i=1}^{n} \left\{ \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1 x_i \omega_i e^{\hat{\beta}_2 x_i \omega_i} \left(2 + \hat{\beta}_2 x_i \omega_i\right) \right. \\
&\quad \left. - \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1^2 \omega_i^2 x_i e^{2\hat{\beta}_2 x_i \omega_i} \left(3 + 2\hat{\beta}_2 x_i \omega_i\right) \right\} \text{ e} \\
\boldsymbol{\Delta}_{\sigma^2} &= \frac{1}{\widehat{\sigma^4}} \sum_{i=1}^{n} \left\{ \mathcal{E}_{0i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1^2 \omega_i e^{2\hat{\beta}_2 x_i \omega_i} \left(1 + 2\hat{\beta}_2 x_i \omega_i\right) - \mathcal{E}_{1i}(\widehat{\boldsymbol{\theta}}) \hat{\beta}_1 e^{\hat{\beta}_2 x_i \omega_i} \left(1 + \hat{\beta}_2 x_i \omega_i\right) \right\}.
\end{aligned}$$

APÊNDICE B

Resultados complementares referentes ao Capítulo 8

Neste capítulo são apresentados os resultados omitidos no texto do Capítulo 8. A Seção B.1 mostra as derivadas de primeira e segunda ordem de $\mathbf{E}_i$ em relação ao ϕ_k, k=1,2. As demais Seções referem-se aos estudos de simulação realizados considerando as estruturas de correlação AR(1), MA(1), simetria composta e independente.

B.1 Derivadas de ϕ_k, k=1,2

Nesta seção apresentamos as derivadas $d\phi_k = \frac{\partial \mathbf{E}_i}{\partial \phi_k}$, $D\phi_i = \frac{partial^2 \mathbf{E}_i}{\partial \phi_k \partial \phi_k}$, k=1,2, e $D\phi_1\phi_2 = \frac{\partial^2 \mathbf{E}_i}{\partial \phi_1 \partial \phi_2}$ segundo as estruturas de correlação de interesse deste trabalho.

Estrutura de erros correlacionados: As derivadas são:

$$d\phi_1 = |t_{ij} - t_{ik}|^{\phi_2} \phi_1^{|t_{ij}-t_{ik}|^{\phi_2}-1},$$

$$d\phi_2 = log(\phi_1) log(|t_{ij} - t_{ik}|) |t_{ij} - t_{ik}|^{\phi_2} \phi_1^{|t_{ij}-t_{ik}|^{\phi_2}},$$

$$D\phi_1 = |t_{ij} - t_{ik}|^{\phi_2} \left(|t_{ij} - t_{ik}|^{\phi_2} - 1\right) \phi_1^{|t_{ij}-t_{ik}|^{\phi_2}-2},$$

$$D\phi_2 = |t_{ij} - t_{ik}|^{\phi_2} \left(log(|t_{ij} - t_{ik}|)\right)^2 \phi_1^{|t_{ij}-t_{ik}|^{\phi_2}} log(\phi_1) \left[log(\phi_1)|t_{ij} - t_{ik}|^{\phi_2} + 1\right], \text{ e}$$

$$D\phi_1\phi_2 = |t_{ij} - t_{ik}|^{\phi_2} log(|t_{ij} - t_{ik}|) \phi_1^{|t_{ij}-t_{ik}|^{\phi_2}-1} \left[log(\phi_1)|t_{ij} - t_{ik}|^{\phi_2} + 1\right].$$

Estrutura de correlação AR(1): As derivadas são:

$$d\phi_1 = |t_{ij} - t_{ik}|\phi_1^{|t_{ij}-t_{ik}|-1}, \text{ e}$$

$$D\phi_1 = |t_{ij} - t_{ik}|\left(|t_{ij} - t_{ik}| - 1\right)\phi_1^{|t_{ij}-t_{ik}|-2},$$

Estrutura de correlação simetria composta: As derivadas são:

$$d\phi_1 = \begin{cases} 0, & \text{se } j = k, \\ 1, & \text{se } j \neq k, \end{cases}, \text{ e}$$

$$D\phi_1 = 0,$$

Estruturas de correlação independente e MA(1): As derivadas são $d\phi_1 = 0$ e $D\phi_1 = 0$.

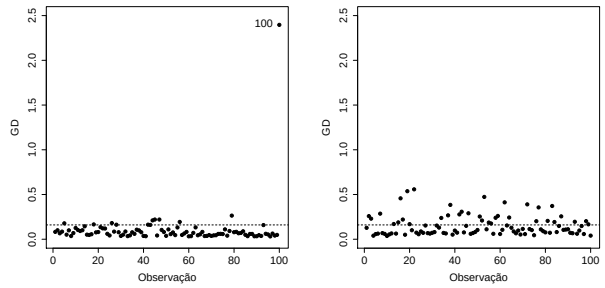
B.2 Simulação: estrutura de correlação AR(1)

A Tabela B.1 apresenta os resultados da análise de influência de um estudo de Monte Carlo com 100 réplicas do modelo (8.1) da Seção 8.2. A Figura B.1 mostra a representação gráfica das medidas de influência para uma réplica.

Tabela B.1: Estudo de simulação. Análise de influência via estudo de Monte Carlo para as observações #1 e #100 por distribuição e medida de diagnóstico: GD - distância generalizada de Cook, PC - ponderação de casos, ES - parâmetro de escala, VP - variável preditora e CO - coeficientes. Modelo para dados longitudinais com estrutura de correlação AR(1).

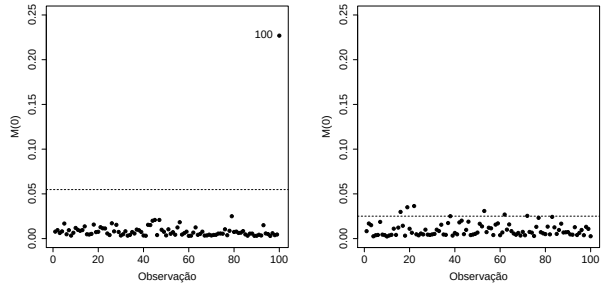
Medidas	**Estatísticas**	**Normal**		**t de Student**	
		#1	#100	#1	#100
GD	% Inf[a]	11%	100%	27%	0%
	M[b]	0,0902	2,5031	0,1318	0,0250
	DP[c]	0,0847	0,5295	0,0973	0,0217
	Ref[d]	0,16		0,16	
PC	% Inf	1%	100%	6%	0%
	M	0,0079	0,2163	0,0096	0,0018
	DP	0,0075	0,0352	0,0071	0,0015
	M (DP) Ref	0,0545 (0,0066)		0,0240 (0,0014)	
ES	% Inf	1%	100%	6%	0%
	M	0,0082	0,2031	0,0096	0,0050
	DP	0,0079	0,0301	0,0056	0,0012
	M (DP) Ref	0,0522 (0,0056)		0,0210 (0,0013)	
VP	% Inf	5%	1%	6%	0%
	M	0,0014	0,0013	0,0015	0,0015
	DP	0,0006	0,0004	0,0012	0,0013
	M (DP) Ref	0,0029 (0,0002)		0,0036 (0,0002)	
CO	% Inf	5%	47%	4%	0%
	M	0,0096	0,0289	0,0091	0,0004
	DP	0,0074	0,0155	0,0065	0,0003
	M (DP) Ref	0,0254 (0,0017)		0,0245 (0,0011)	

[a] % Inf: percentual de réplicas de Monte Carlo em que a observação foi considerada influente.
[b] M é a média das medidas de influência.
[c] DP é o desvio-padrão das medidas de influência.
[d] Ref é o valor de referência para considerar uma observação influente.



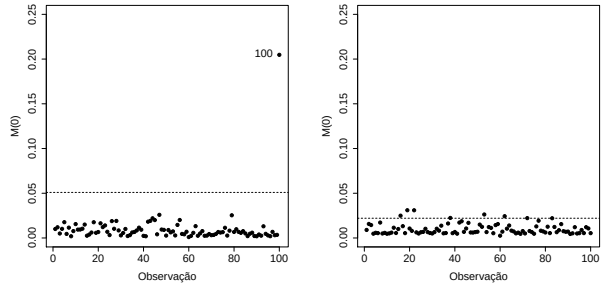
(a) GD - Normal

(b) GD - t de Student



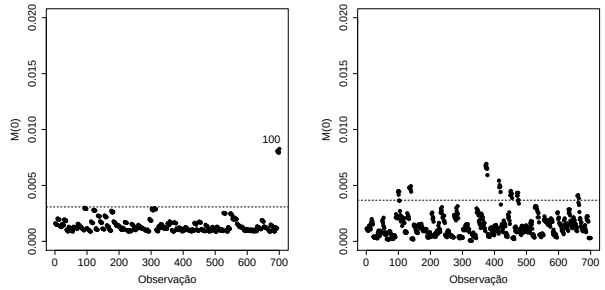
(c) PC - Normal

(d) PC - t de Student



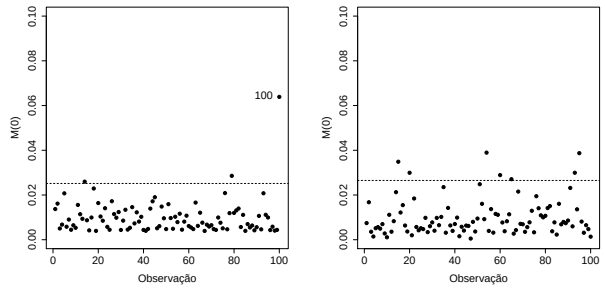
(e) ES - Normal

(f) ES - t de Student



(g) VP - Normal

(h) VP - t de Student



(i) CO - Normal

(j) CO - t de Student

Figura B.1: Estudo de simulação. Medidas de influência considerando a estrutura de correlação AR(1), para as distribuições Normal e t de Student. Modelo longitudinal.

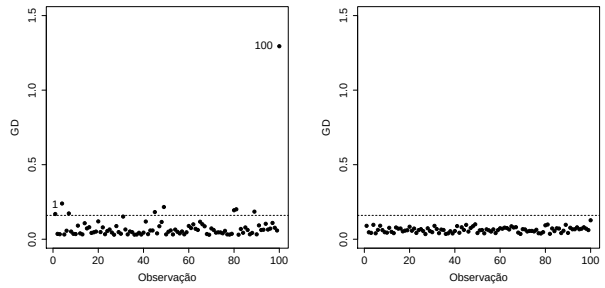
B.3 Simulação: Estrutura de correlação MA(1)

A Tabela B.2 apresenta os resultados da análise de influência de um estudo de Monte Carlo com 100 réplicas do modelo (8.1) da Seção 8.2. A Figura B.2 mostra a representação gráfica das medidas de influência para uma réplica.

Tabela B.2: Estudo de simulação. Análise de influência via estudo de Monte Carlo para as observações #1 e #100 por distribuição e medida de diagnóstico: GD - distância generalizada de Cook, PC - ponderação de casos, ES - parâmetro de escala, VP - variável preditora e CO - coeficientes. Modelo para dados longitudinais com estrutura de correlação MA(1).

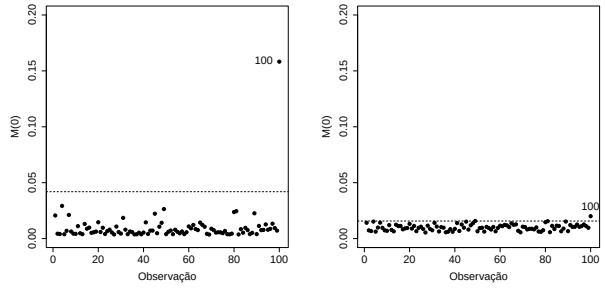
Medidas	**Estatísticas**	**Normal**		**t de Student**	
		#1	#100	#1	#100
GD	% Inf[a]	5%	100%	0%	0%
	M[b]	0,0594	1,826	0,0581	0,1308
	DP[c]	0,0396	0,391	0,0156	0,0032
	Ref[d]	0,16		0,16	
PC	% Inf	0%	100%	0%	100%
	M	0,0067	0,2041	0,0091	0,0206
	DP	0,0045	0,0374	0,0025	0,0005
	M (DP) Ref	0,0523 (0,0065)		0,0160 (0,0003)	
ES	% Inf	0%	100%	1%	100%
	M	0,0067	0,1840	0,0088	0,0278
	DP	0,0059	0,0315	0,0043	0,0007
	M (DP) Ref	0,0494 (0,0053)		0,0200 (0,0005)	
VP	% Inf	3%	3%	2%	2%
	M	0,0014	0,0015	0,0016	0,0015
	DP	0,0002	0,0003	0,0004	0,0003
	M (DP) Ref	0,0021 ($9{,}7e^{-5}$)		0,0022 ($4{,}0e^{-5}$)	
CO	% Inf	5%	65%	5%	0%
	M	0,0095	0,0307	0,0105	0,0029
	DP	0,0051	0,0159	0,0030	0,0014
	M (DP) Ref	0,0221 (0,0015)		0,0160 (0,0004)	

[a] % Inf: percentual de réplicas de Monte Carlo em que a observação foi considerada influente.
[b] M é a média das medidas de influência.
[c] DP é o desvio-padrão das medidas de influência.
[d] Ref é o valor de referência para considerar uma observação influente.



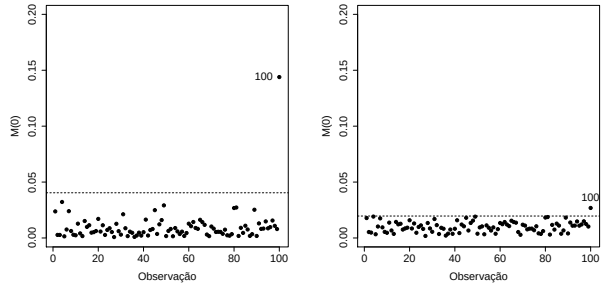
(a) GD - Normal

(b) GD - t de Student



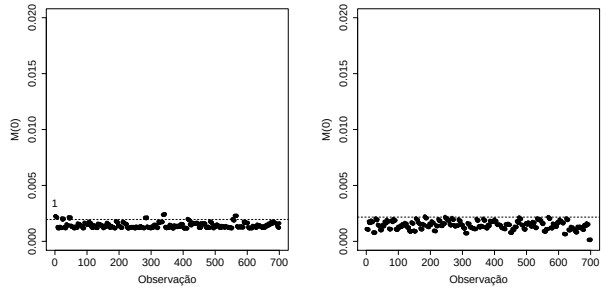
(c) PC - Normal

(d) PC - t de Student



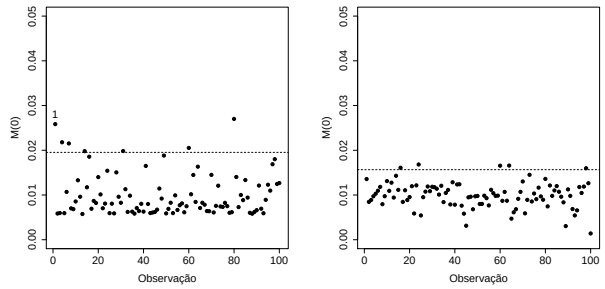
(e) ES - Normal

(f) ES - t de Student



(g) VP - Normal

(h) VP - t de Student



(i) CO - Normal

(j) CO - t de Student

Figura B.2: Estudo de simulação. Medidas de influência considerando a estrutura de correlação MA(1), para as distribuições Normal e t de Student. Modelo longitudinal.

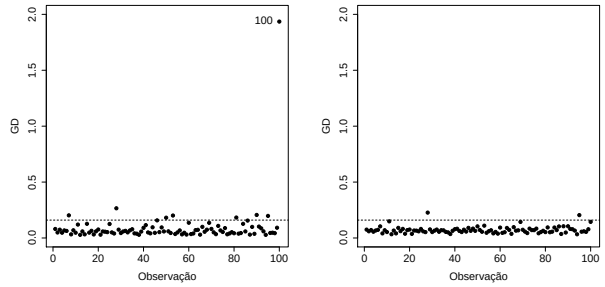
B.4 Simulação: Estrutura de correlação simetria composta (SC)

A Tabela B.3 apresenta os resultados da análise de influência de um estudo de Monte Carlo com 100 réplicas do modelo (8.1) da Seção 8.2. A Figura B.3 mostra a representação gráfica das medidas de influência para uma réplica.

Tabela B.3: Estudo de simulação. Análise de influência via estudo de Monte Carlo para as observações #1 e #100 por distribuição e medida de diagnóstico: GD - distância generalizada de Cook, PC - ponderação de casos, ES - parâmetro de escala, VP - variável preditora e CO - coeficientes. Modelo para dados longitudinais com estrutura de correlação simetria composta.

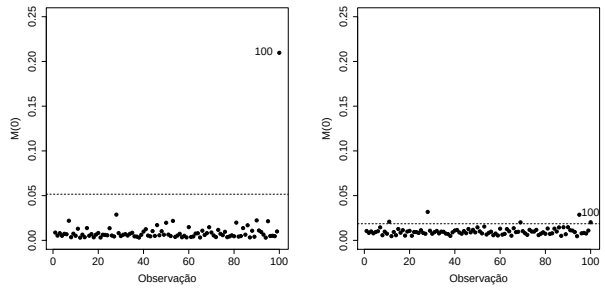
Medidas	**Estatísticas**	**Normal**		**t de Student**	
		#1	#100	#1	#100
GD	% Inf[a]	7%	100%	0%	62%
	M[b]	0,0733	3,1110	0,0669	0,1654
	DP[c]	0,0715	0,9249	0,0226	0,0164
	Ref[d]	0,16		0,16	
PC	% Inf	1%	100%	2%	97%
	M	0,0070	0,2910	0,0094	0,0233
	DP	0,0067	0,0613	0,0032	0,0022
	M (DP) Ref	0,0681 (0,0119)		0,0184 (0,0012)	
ES	% Inf	1%	100%	3%	100%
	M	0,0074	0,2607	0,0095	0,0299
	DP	0,0076	0,0537	0,0045	0,0019
	M (DP) Ref	0,0626 (0,0102)		0,0204 (0,0008)	
VP	% Inf	11%	8%	12%	6%
	M	0,0021	0,0017	0,0022	0,0018
	DP	0,0006	0,0008	0,0007	0,0007
	M (DP) Ref	0,0030 (0,0001)		0,0029 (0,0001)	
CO	% Inf	4%	59%	4%	0%
	M	0,0094	0,0306	0,0096	0,0023
	DP	0,0061	0,0140	0,0048	0,0015
	M (DP) Ref	0,0240 (0,0014)		0,0205 (0,0009)	

[a] % Inf: percentual de réplicas de Monte Carlo em que a observação foi considerada influente.
[b] M é a média das medidas de influência.
[c] DP é o desvio-padrão das medidas de influência.
[d] Ref é o valor de referência para considerar uma observação influente.



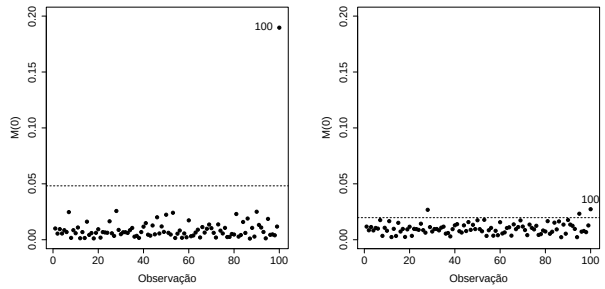
(a) GD - Normal

(b) GD - t de Student



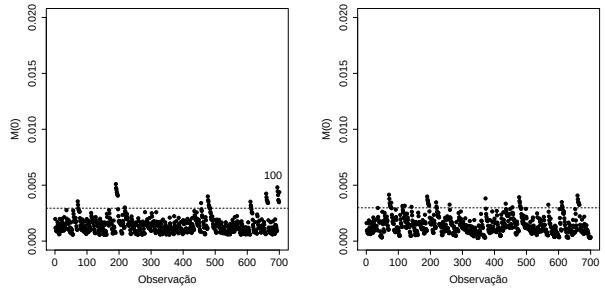
(c) PC - Normal

(d) PC - t de Student



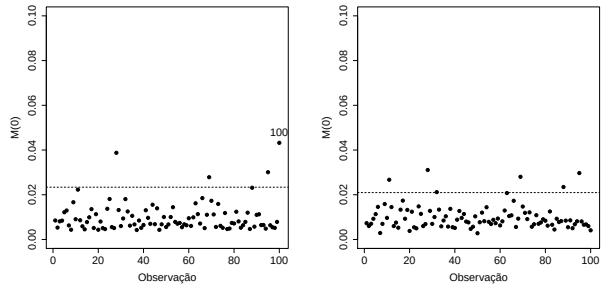
(e) ES - Normal

(f) ES - t de Student



(g) VP - Normal

(h) VP - t de Student



(i) CO - Normal

(j) CO - t de Student

Figura B.3: Estudo de simulação. Medidas de influência considerando a estrutura de correlação simetria composta, para as distribuições Normal e t de Student. Modelo longitudinal.

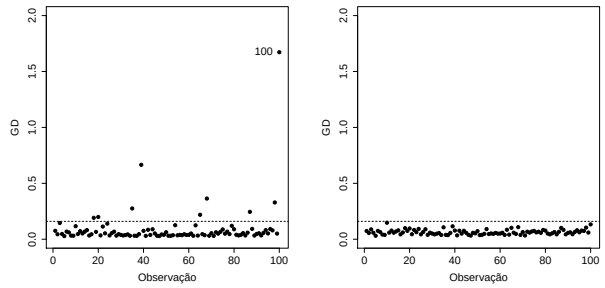
B.5 Simulação: Estrutura de correlação Independente (Ind)

A Tabela B.4 apresenta os resultados da análise de influência de um estudo de Monte Carlo com 100 réplicas do modelo (8.1) da Seção 8.2. A Figura B.4 mostra a representação gráfica das medidas de influência para uma réplica.

Tabela B.4: Estudo de simulação. Análise de influência via estudo de Monte Carlo para as observações #1 e #100 por distribuição e medida de diagnóstico: GD - distância generalizada de Cook, PC - ponderação de casos, ES - parâmetro de escala, VP - variável preditora e CO - coeficientes. Modelo para dados longitudinais com estrutura de correlação independente.

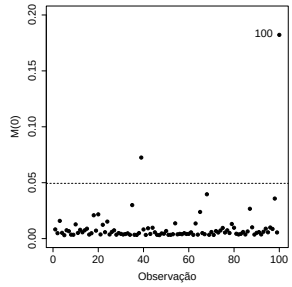
Medidas	**Estatísticas**	**Normal**		**t de Student**	
		#1	#100	#1	#100
GD	% Inf[a]	5%	100%	0%	0%
	M[b]	0,0716	1,8258	0,0613	0,1307
	DP[c]	0,0819	0,3697	0,0165	0,0037
	Ref[d]	0,16		0,16	
PC	% Inf	1%	100%	3%	100%
	M	0,0079	0,2050	0,0097	0,0206
	DP	0,0085	0,0354	0,0026	0,0006
	M (DP) Ref	0,0522 (0,0063)		0,0159 (0,0004)	
ES	% Inf	1%	100%	3%	100%
	M	0,0082	0,1845	0,0096	0,0278
	DP	0,0095	0,0299	0,0046	0,0008
	M (DP) Ref	0,0493 (0,0052)		0,0199 (0,0006)	
VP	% Inf	4%	6%	3%	0%
	M	0,0015	0,0015	0,0015	0,0014
	DP	0,0003	0,0005	0,0004	0,0004
	M (DP) Ref	0,0021 (0,0001)		0,0022 ($4{,}7e^{-5}$)	
CO	% Inf	3%	75%	2%	0%
	M	0,0101	0,0308	0,0104	0,0029
	DP	0,0056	0,0149	0,0032	0,0014
	M (DP) Ref	0,0220 (0,0018)		0,0161 (0,0003)	

[a] % Inf: percentual de réplicas de Monte Carlo em que a observação foi considerada influente.
[b] M é a média das medidas de influência.
[c] DP é o desvio-padrão das medidas de influência.
[d] Ref é o valor de referência para considerar uma observação influente.

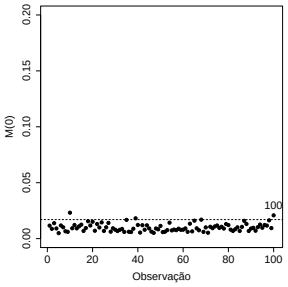


(a) GD - Normal

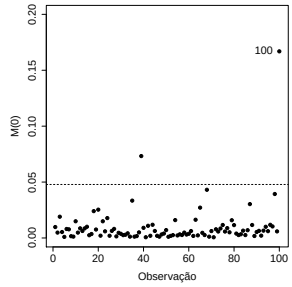
(b) GD - t de Student



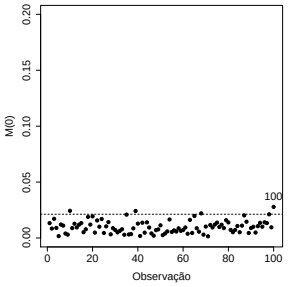
(c) PC - Normal



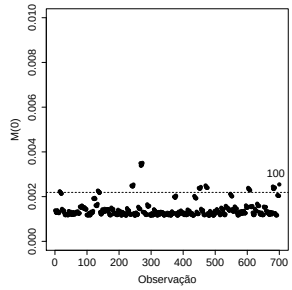
(d) PC - t de Student



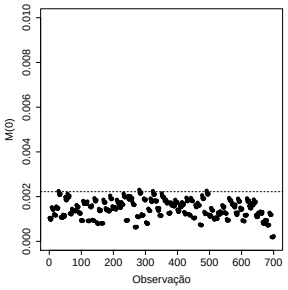
(e) ES - Normal



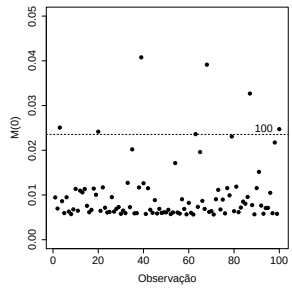
(f) ES - t de Student



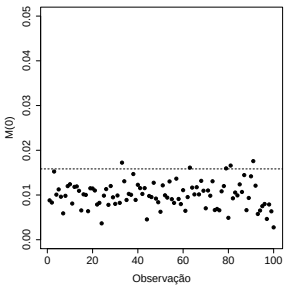
(g) VP - Normal



(h) VP - t de Student



(i) CO - Normal



(j) CO - t de Student

Figura B.4: Estudo de simulação. Medidas de influência considerando a estrutura de correlação independente, para as distribuições Normal e t de Student. Modelo longitudinal.