Sandra Santos dos Reis

Diagnóstico em modelos de regressão linear e não-linear com erros simétricos

CAMPINAS
2013

Universidade Estadual de Campinas

Instituto de Matemática, Estatística
e Computação Científica

Sandra Santos dos Reis

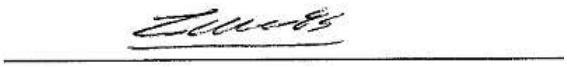
Diagnóstico em modelos de regressão linear e não-linear com erros simétricos

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

Orientador: Mauricio Enrique Zevallos Herencia

Este exemplar corresponde à versão final da dissertação defendida pela aluna Sandra Santos dos Reis, e orientada pelo Prof. Dr. Mauricio Enrique Zevallos Herencia.

Assinatura do Orientador



Campinas

2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

R277d

Reis, Sandra Santos dos, 1983-
Diagnóstico em modelos de regressão linear e não-linear com erros simétricos / Sandra Santos dos Reis. – Campinas, SP : [s.n.], 2013.

Orientador: Mauricio Enrique Zevallos Herencia.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

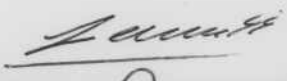
1. Estimativa de parâmetro. 2. Modelos lineares (Estatística). 3. Modelos não lineares (Estatística). 4. Observações influentes. I. Zevallos Herencia, Mauricio Enrique,1966-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

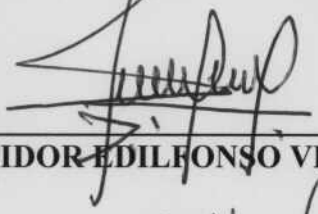
Título em outro idioma: Diagnostic in linear and nonlinear regression models with symmetrical errors
Palavras-chave em inglês:
Parameter estimation
Linear models (Statistics)
Nonlinear models (Statistics)
Influential observations
Área de concentração: Estatística
Titulação: Mestra em Estatística
Banca examinadora:
Mauricio Enrique Zevallos Herencia [Orientador]
Filidor Edilfonso Vilca Labra
Edwin Moises Marcos Ortega
Data de defesa: 16-12-2013
Programa de Pós-Graduação: Estatística

Dissertação de Mestrado defendida em 16 de dezembro de 2013 e aprovada

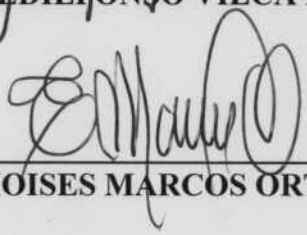
Pela Banca Examinadora composta pelos Profs. Drs.



Prof(a). Dr(a). MAURICIO ENRIQUE ZEVALLOS HERENCIA



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



Prof(a). Dr(a). EDWIN MOISES MARCOS ORTEGA

Abstract

We discuss the detection of influential observations in symmetrical linear and nonlinear regression models. First a simulation study is conducted to evaluate the performance of three estimation methods on data generated by four situations: without influential observations with outliers in the response variable, with influential observations average leverage and influential observations with high leverage. Two methods of maximum likelihood and robust method are analyzed. We considered linear and nonlinear regression models with logistic-II and Student-t errors. Secondly detection of influential observations by generalized Cook's distance , the statistic PeÃśa and Andrews - Pregibon statistic is discussed. In particular the convenience of using the methodology to characterize a threshold observation as influential or not influential , as well as the effect of parameter estimation in the construction of thresholds is discussed. These measures were applied to sets of real and simulated data considering the fit of some symmetrical regression models with an adaptation estimation method of Fisher scoring.

Keywords: Parameter estimation, Linear models (Statistic), Nonlinear models (Statistic), Influential observations.

Resumo

Neste trabalho discutimos a detecção de observações influentes em modelos simétricos lineares e não lineares. Em primeiro lugar é realizado um estudo de simulação para avaliar o desempenho de três métodos de estimação em dados gerados por quatro situações: sem observações influentes, com outliers na variável resposta, com observações influentes de média alavancagem e com observações influentes de alta alavancagem. São analisados dois métodos de máxima verossimilhança e um método robusto. Foram considerados modelos de regressão linear e não linear com erros logísticos tipo II e t-Student. Em segundo lugar é discutida detecção de observações influentes mediante a distância de Cook generalizada, a estatística de Peña e a estatística de Andrews-Pregibon. Em particular é discutida a conveniência de utilizar a metodologia de limiares para caracterizar uma observação como influente ou não influente, assim como o efeito da estimação de parâmetros na construção de limiares. Estas medidas foram aplicadas a conjuntos de dados reais e simulados considerando o ajuste de alguns modelos simétricos com uma adaptação no método de estimação scoring de Fisher.

Palavras-chave: Estimativa de parêmetro, Modelos lineares (Estatística), Modelos não lineares (Estatística), Observações influentes.

Sumário

Dedicatória **xi**

Agradecimentos **xv**

1 Introdução **1**

1.1 Organização da Dissertação 3

2 Modelos de regressão com erros simétricos **5**

2.1 Modelo 6

2.2 Verossimilhança, função escore e matriz de informação 6

2.3 Estimação 8

2.3.1 Método 1 9

2.3.2 Método 2 10

2.3.3 Método 3 11

3 Estudo de Simulação **13**

3.1 Modelo linear 14

3.1.1 Resultados da estimação sob erros logístico tipo II 14

3.1.2 Resultados da estimação sob erros t-Student 19

3.2 Modelos não-lineares 29

3.2.1 Modelo de decaimento exponencial 29

3.2.2 Modelo de regressão assintótica 42

4 Influência em Modelos Simétricos **57**

4.1 Estatísticas de influência 57

4.1.1 Distância de Cook generalizada 58

4.1.2 Estatística de Peña 59

4.1.3 Estatística de Andrews-Pregibon 60

4.2 Metodologia para identificação de pontos influentes 62

5 Aplicações **65**

5.1 Modelos lineares 65

5.1.1 Simulação 65

5.1.2 Dados HRD 73

5.1.3 Refinaria 80

5.2 Modelos não-lineares 84

5.2.1 Simulação 84

5.2.2 Dados de ultra-sons 88

5.2.3 Dados da produção de cebola 93

5.2.4 Coelhos europeus 98

6 Conclusões e considerações finais **105**

Referências Bibliográficas **107**

A Informações adicionais **111**

A.1 Modelo de exclusão de casos 111

A.2 Uma aproximação para distância de Cook generalizada 113

A.3 Uma aproximação para estatística de Peña 113

A.4 Matriz de informação de Fisher 116

Aos meus pais, Maria da Paz e Elias,
e aos meus irmãos, Simone, Alessandro e Rafael.

Para minha mãe, Maria da Paz

"Maria, Maria,
É um dom,
Uma certa magia
Uma força que nos alerta
Uma mulher que merece
Viver e amar
Como outra qualquer
Do planeta

(...)
Mas é preciso ter força,
É preciso ter raça
É preciso ter gana sempre
Quem traz no corpo a marca
Maria, Maria,
Mistura a dor e a alegria

Mas é preciso ter manha,
É preciso ter graça
É preciso ter sonho sempre
Quem traz na pele essa marca
Possui a estranha mania
De ter fé na vida."

Milton Nascimento/Fernando Brant

Para minha irmã, Simone

"Se lembra da fogueira
se lembra dos balões
se lembra dos luares, dos sertões
A roupa no varal, feriado nacional
e as estrelas salpicadas nas canções
(...)

(...)
Se lembra do futuro
que a gente combinou
Eu era tão criança e ainda sou
Querendo acreditar que o dia vai raiar
Só porque uma cantiga anunciou
(...)
"

Chico Buarque

Para meu bem, José Roberto

"Ainda bem
Que agora encontrei você
Eu realmente não sei
O que eu fiz pra merecer
Você

Porque ninguém
Dava nada por mim
Quem dava eu não tava a fim
Até desacreditei
De mim

O meu coração
Já estava acostumado
Com a solidão quem diria
Que ao meu lado você iria ficar

(...)
Agora você chegou
Tudo se transformou
Você que me faz feliz
Você que me faz cantar
Assim
(...)"

Marisa Monte/Arnaldo Antunes

Agradecimentos

Agradeço primeiramente a Deus, por ser meu refúgio e minha fortaleza sempre.

À minha mãe e minha irmã, pelo amor, carinho, apoio, incentivo e pela compreensão que depositaram em mim durante essa trajetória. A ajuda e a força que elas me deram foram de extrema importância para que eu pudesse dar continuidade a este trabalho.

Aos meus tios, tias, primos e primas que sempre torceram por mim. A cada encontro me receberam com grande alegria e satisfação. Me sinto abençoada por tê-los em minha vida.

Ao meu namorado, José Roberto, pelo carinho, cuidado, ternura, incentivo e grande companherismo a cada instante. A convivência com ele só me trouxe entusiasmo e um grande bem-estar.

Aos amigos que fiz durante o mestrado, em especial a Carlos Trucios, Larissa Matos, Marcos Cascone, Diana Galvis e Julián, José Alejandro e Panlly, e Diego Bernardini pelos momentos divertidíssimos e de grande descontração.

Aos amigos da UFBA, em especial a Silvânia, Carlos André e Gabriel, pelo apoio e incentivo durante essa jornada.

Ao professor Mauricio Zevallos, pela orientação e dedicação durante a elaboração deste trabalho.

Aos participantes da banca examinadora, pelas sugestões.

Aos professores do Departamento de Estatística do IMECC-UNICAMP, pelos ensinamentos.

A CAPES, pelo apoio financeiro.

Lista de Figuras

3.1 *Gráfico dos dados de uma simulação com $n = 50$ para cada situação sob erros logístico tipo II. (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada por dois estágios.* . . . 15

3.2 *Gráfico dos dados de uma simulação com $n = 100$ para cada situação sob erros logístico tipo II. (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada por dois estágios.* . . . 16

3.3 *Gráfico dos dados de uma simulação com $n = 200$ para cada situação sob erros logístico tipo II. (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada por dois estágios.* . . . 17

3.4 *Gráfico dos dados de uma simulação do modelo linear sob erros t-Student da **Situação (1)** de acordo com o tamanho da amostra e grau de liberdade.* (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada pelo Método 2. 20

3.5 *Gráfico dos dados de uma simulação do modelo linear sob erros t-Student da **Situação (2)** de acordo com o tamanho da amostra e grau de liberdade.* (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada pelo Método 2. 21

3.6 *Gráfico dos dados de uma simulação do modelo linear sob erros t-Student da **Situação (3)** de acordo com o tamanho da amostra e grau de liberdade.* (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada pelo Método 2. 22

3.7 *Gráfico dos dados de uma simulação do modelo linear sob erros t-Student da **Situação (4)** de acordo com o tamanho da amostra e grau de liberdade.* (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada pelo Método 2. 23

3.8 *Gráfico dos dados de uma simulação do modelo (3.2.1) com $n = 50$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2.* 30

3.9 *Gráfico dos dados de uma simulação do modelo (3.2.1) com $n = 100$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2.* 31

3.10 *Gráfico dos dados de uma simulação do modelo (3.2.1) com $n = 200$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2.* 32
3.11 *Gráfico dos dados de uma simulação do modelo (3.2.1) sob erros t-Student da* ***Situação (1)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada por MV com ν fixo e (· · · · · · ·) curva ajustada pelo Método 2. 35
3.12 *Gráfico dos dados de uma simulação do modelo (3.2.1) sob erros t-Student da* ***Situação (2)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada por MV com ν fixo e (· · · · · · ·) curva ajustada pelo Método 2. 36
3.13 *Gráfico dos dados de uma simulação do modelo (3.2.1) sob erros t-Student da* ***Situação (3)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada por MV com ν fixo e (· · · · · · ·) curva ajustada pelo Método 2. 37
3.14 *Gráfico dos dados de uma simulação do modelo (3.2.1) sob erros t-Student da* ***Situação (4)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada por MV com ν fixo e (· · · · · · ·) curva ajustada pelo Método 2. 38
3.15 *Gráfico dos dados de uma simulação do modelo (3.2.2) com $n = 50$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2.* 44
3.16 *Gráfico dos dados de uma simulação do modelo (3.2.2) com $n = 100$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2.* 45
3.17 *Gráfico dos dados de uma simulação do modelo (3.2.2) com $n = 200$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada por Método 2.* 46
3.18 *Gráfico dos dados de uma simulação do modelo (3.2.2) sob erros t-Student da* ***Situação (1)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2. 49
3.19 *Gráfico dos dados de uma simulação do modelo (3.2.2) sob erros t-Student da* ***Situação (2)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2. 50

3.20 *Gráfico dos dados de uma simulação do modelo (3.2.2) sob erros t-Student da* ***Situação (3)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (− − −) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (− · − · −) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2. 51
3.21 *Gráfico dos dados de uma simulação do modelo (3.2.2) sob erros t-Student da* ***Situação (4)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (− − −) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (− · − · −) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2. 52

5.1 *Estatística de influência para o ajuste dos dados simulados da Situação (1).* 66
5.2 *Estatística de influência para o ajuste dos dados simulados da Situação (2).* 66
5.3 *Estatística de influência para o ajuste dos dados simulados da Situação (3).* 67
5.4 *Estatística de influência para o ajuste dos dados simulados da Situação (4).* 67
5.5 *Em a gráfico dos dados simulados e em b-d gráfico dos dados com o ajuste do modelo normal, Logístico-II e t-Student, respectivamente.* 69
5.6 *Estatísticas de influência para o ajuste do modelo (5.1.1) sob erros normais.* 70
5.7 *Estatísticas de influência para o ajuste do modelo (5.1.1) sob erros logísticos tipo II.* . . . 71
5.8 *Estatísticas de influência para o ajuste do modelo (5.1.1) sob erros t-Student.* 72
5.9 *Em (a) gráfico dos dados HRD e em (b) - (d) gráfico dos dados com o ajuste do modelo normal, Logístico-II e exponencial potência, respectivamente.* 75
5.10 *Gráficos normais de probabilidades com envelope para os resíduos padronizados para o ajuste do modelo normal (a), logístico-II (b) e exponencial potência (c) ajustados aos dados HRD.* 76
5.11 *Análise de influência para os dados HRD sob erros normais. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.* 77
5.12 *Análise de influência para os dados HRD sob erros Logísticos tipo II . As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.* 78
5.13 *Análise de influência para os dados HRD sob erros exponencial potência. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.* 79
5.14 *Gráficos normais de probabilidades com envelope para os resíduos padronizados para o ajuste do modelo normal (a), logístico-II (b) e exponencial potência (c) ajustados aos dados Refinaria.* 81
5.15 *Análise de influência para os dados Refinaria sob erros normais. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.* 82
5.16 *Análise de influência para os dados refinaria sob erros Logísticos tipo II. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.* 83

5.17 *Análise de influência para os dados refinaria sob erros exponencial potência. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.* 84
5.18 *Estatística de influência para o ajuste dos dados simulados da Situação (1).* 86
5.19 *Estatística de influência para o ajuste dos dados simulados da Situação (2).* 86
5.20 *Estatística de influência para o ajuste dos dados simulados da Situação (3).* 87
5.21 *Estatística de influência para o ajuste dos dados simulados da Situação (4).* 87
5.22 *Em (a) gráfico dos dados de ultra-sons contaminados e em (b) - (d) gráfico dos dados com o ajuste do modelo normal, Logístico-II e t-Student, respectivamente.* 89
5.23 *Análise de influência para os dados de ultra-sons contaminados sob erros normais.* 90
5.24 *Análise de influência para os dados de ultra-sons contaminados sob erros Logísticos tipo II.* . . . 91
5.25 *Análise de influência para os dados de ultra-sons contaminados sob erros t-Student.* 92
5.26 *Em (a) gráfico dos dados da produção de cebola e em (b) - (d) gráfico dos dados com o ajuste do modelo normal, Logístico-II e t-Student, respectivamente.* 94
5.27 *Gráficos normais de probabilidades com envelope para os resíduos padronizados para o ajuste do modelo normal (a), logístico-II (b) e t-Student (c) ajustados aos dados da produção de cebola.* 95
5.28 *Análise de influência para os dados da produção de cebola sob erros normais.* 96
5.29 *Análise de influência para os dados da produção de cebola sob erros Logísticos tipo II.* 97
5.30 *Análise de influência para os dados da produção de cebola sob erros t-Student.* 98
5.31 *Em (a) gráfico dos dados dos coelhos europeus e em (b) - (d) gráfico dos dados com o ajuste do modelo normal, Logístico-II e t-Student, respectivamente.* 99
5.32 *Gráficos normais de probabilidades com envelope para os resíduos padronizados para o ajuste do modelo normal (a), logístico-II (b) e t-Student (c) ajustados aos dados dos coelhos europeus.* 100
5.33 *Análise de influência para o ajuste dos dados de coelhos europeus sob erros normais.* 101
5.34 *Análise de influência para o ajuste dos dados de coelhos europeus sob erros Logísticos-II.* 102
5.35 *Análise de influência para o ajuste dos dados de coelhos europeus sob erros t-Student.* 103

Lista de Tabelas

2.1 *Expressões para $g(u)$, $W_g(u)$ e $W'g(u)$ para algumas distribuições simétricas.* 7
2.2 *Valores de d_g, $f_g(u)$ e ξ para algumas distribuições simétricas.* 9

3.1 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros logísticos tipo II. $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$, $\sigma^2 = 296,09$.* ***Situação (1).*** 18
3.2 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros logísticos tipo II. $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$, $\sigma^2 = 296,09$.* ***Situação (2).*** 18
3.3 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros logísticos tipo II. $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$, $\sigma^2 = 296,09$.* ***Situação (3).*** 19
3.4 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros logísticos tipo II. $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$, $\sigma^2 = 296,09$.* ***Situação (4).*** 19
3.5 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros t-Student. $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$.* ***Situação (1).*** 24
3.6 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros t-Student. $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$.* ***Situação (2).*** 25
3.7 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros t-Student. $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$.* ***Situação (3).*** 26
3.8 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo t-Student. $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$.* ***Situação (4).*** 27
3.9 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros logísticos tipo II . $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$, $\phi = 17$ e $\sigma^2 = 56$.* ***Situação (1)*** 33
3.10 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros logísticos tipo II. $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$, $\phi = 17$ e $\sigma^2 = 56$.* ***Situação (2)*** 33
3.11 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros logísticos tipo II. $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$, $\phi = 17$ e $\sigma^2 = 56$.* ***Situação (3)*** 34
3.12 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros logísticos tipo II. $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$, $\phi = 17$ e $\sigma^2 = 56$.* ***Situação (4)*** 34
3.13 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros t-student. $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$ e $\phi = 17$.* ***Situação (1).*** 39

3.14 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros t-Student.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$ *e* $\phi = 17$. ***Situação (2)*** . . . 40
3.15 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros t-Student.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$ *e* $\phi = 17$. ***Situação (3)*** . . . 41
3.16 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros t-Student.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$ *e* $\phi = 17$. ***Situação (4)***. . . . 42
3.17 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros logísticos tipo II .* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$, $\phi = 20$ *e* $\sigma^2 = 65,8$. ***Situação (1)***. . . 47
3.18 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros logísticos tipo II.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$ *e* $\sigma^2 = 65,8$. ***Situação (2)***. . . 47
3.19 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros logísticos tipo II.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$, $\phi = 20$ *e* $\sigma^2 = 65,8$. ***Situação (3)***. . . 48
3.20 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros logísticos tipo II.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$, $\phi = 20$ *e* $\sigma^2 = 65,8$. ***Situação (4)***. . . 48
3.21 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros t-student.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$. ***Situação (1)***. . . . 53
3.22 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros t-student.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$. ***Situação (2)***. . . . 54
3.23 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros t-student.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$. ***Situação (3)***. . . . 55
3.24 *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros t-student.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$. ***Situação (4)***. . . . 56

5.1 *Estimativas de máxima verossimilhança (erro-padrão) para alguns modelos simétricos ajustados aos dados simulados.* $\boldsymbol{\theta} = (\beta_0 = 110, \beta_1 = -4, \phi = 90)^{\top}$. . . . 68
5.2 *Estatística de Andrews-Pregibon para os dados simulados.* . . . 73
5.3 *Mudanças (em percentuais) nas estimativas após exclusão das observações (15,16,17).* 73
5.4 *Estimativas de máxima verossimilhança (erro padrão aproximado) para alguns modelos simétricos lineares ajustados aos dados HRD.* . . . 74
5.5 *Estatística de Andrews-Pregibon para os dados HRD.* . . . 74
5.6 *Mudanças (em percentuais) nas estimativas após exclusão de observações.* . . . 80
5.7 *Estimativas de máxima verossimilhança (erro padrão aproximado) para alguns modelos simétricos lineares ajustados aos dados Refinaria.* . . . 80
5.8 *Estatística de Andrews-Pregibon para os dados Refinaria.* . . . 85
5.9 *Mudanças (em percentuais) nas estimativas após exclusão de algumas observações.* . . . 85
5.10 *Estimativas de máxima verossimilhança (erro-padrão aproximado) para alguns modelos simétricos ajustados aos dados de ultra-sons contaminados.* . . . 89
5.11 *Mudanças (em percentuais) nas estimativas após exclusão de algumas observações.* . 92
5.12 *Estimativas de máxima verossimilhança (erro-padrão) para alguns modelos simétricos ajustados aos dados produção de cebola.* . . . 93
5.13 *Mudanças (em percentuais) nas estimativas após exclusão da observação 38.* . . . 97

5.14 *Estimativas de máxima verossimilhança (erro-padrão aproximado) para alguns modelos simétricos ajustados aos dados dos coelhos europeus.* 99

Capítulo 1

Introdução

Os modelos de regressão são amplamente utilizados em aplicações práticas com o objetivo de responder a questões sobre a dependência de uma variável resposta em uma ou mais variáveis explicativas. Geralmente, esses modelos especificam algumas suposições ao seu componente aleatório, os erros. As suposições mais comuns são a homocedasticidade e normalidade. A suposição de normalidade sempre foi muito utilizada para os modelos de regressão com resposta contínua, mas nem sempre essa suposição é especificada. A modelagem sob a suposição de normalidade dos erros é vunerável a observações atípicas. Nos últimos anos, diversos resultados de natureza teórica e aplicada surgiram como alternativa à modelagem com erros normais como, por exemplo, o uso de distribuições simétricas (veja Fang et al. (1990) e Fang and Anderson (1990)). Temos também as contibuições de Lange et al. (1989) que discutiram o uso da distribuição t-Student em modelos de regressão linear e não-linear onde os graus de liberdade são usados para controlar a curtose. Taylor (1992) propõe o ajuste do modelo de regressão linear sob erros expnencial potência com um parâmetro extra de forma. Arellano-Valle (1994) apresenta resultados sobre regressão usando a distribuição t-Student. Estes resultados foram estendidos para os modelos lineares simétricos por Ferrari and Uribe-Opazo (2001). No trabalho de Cordeiro et al. (2000) foi a correção de viés dos estimadores de máxima verossimilhança para modelos não-lineares simétricos.

Mesmo em modelagens com erros não normais, é importante examinar o ajuste do modelo, pois se o modelo não estiver bém ajustado, o mesmo pode fornecer conclusões errôneas. Alguns elementos do conjunto de dados podem controlar propriedades importantes no modelo e nestas situações, são considerados influentes. Uma observação influente produz alterações relevantes no resultado da análise quando é extraída ou submetida a uma pequena perturbação. Uma técnica que pode ajudar na detecção de observações influentes é o diagnóstico de influência. O principal objetivo do diagnóstico de influência é avaliar a qualidade do ajuste do modelo após pequenas modificações. Essa abordagem envolve modificações que podem ser uma simples exclusão de observações ou perturbações em direções específicas.

Diversas medidas de influência podem ser encontradas na literatura (veja, por exemplo, Belsley et al. (1980), Cook and Weisberg (1982), Atkinson (1985), Chalterjee and Hadi (1988). Uma medida bastante popular, chamada distância de Cook, foi proposta por Cook (1977), e consiste em avaliar o efeito da exclusão de observações nas estimativas dos parâmetros, em modelos de regressão linear normal. No entanto, Cook e Weisberg (1982) definiu uma medida mais geral para

avaliar a influência, chamada afastamento da verossimilhança, que verifica o efeito de pequenas perturbações sobre a verossimilhança.

Peña (2005) propôs uma medida alternativa à distância de Cook para analisar influência em modelos de regressão linear. Ao invés de avaliar o efeito da exclusão de observações nas estimativas dos parâmetros, o objetivo é verificar como uma observação é influenciada pelo restante da amostra. Ou seja, verifica-se o quão sensível é a predição de uma determinada observação à exclusão de cada observação na amostra. Propriedades foram provadas para esta medida, dentre ela a capacidade de encontrar grupos de observações influente similares, que são mais difíceis de serem detectados pela distância de Cook.

A princípio, as estatísticas de influência foram desenvolvidas para regressão linear normal. Entretanto, muitas técnicas de influência foram estendidas para outros modelos. Pregibon (1981) estendeu os resultados de Cook (1977) aos modelos lineares generalizados. Diversos autores aplicaram o método influência local em vários modelos de regressão, como Beckman et al. (1987), Lawrance (1988), Thomas and Cook (1990), Tsai and Wu (1992), Paula (1993), Kim (1995) entre outros. Recentemente, métodos de diagnóstico foram estendidos aos modelos de regressão linear simétricos, por exemplo, Galea et al. (1997), Liu (2000) e Galea et al. (2003). Galea et al. (2000) desenvolveu medidas de influência em modelos elípticos lineares. Galea et al. (2005) desenvolveram um resíduo padronizado e propôs algumas medidas de influência local sob diferentes esquemas de perturbação em modelos de regressão não-linear simétricos. Vanegas and Cysneiros (2010) desenvolveram procedimentos de diagnóstico com base no modelo de exclusão de casos para modelos de regressão não-linear simétricos, que complementam os estudos de Galea et al. (2005).

As contribuições desta dissertação são as seguintes. Primeiro, a extensão da estatística proposta por Peña (2005) aos modelos simétricos lineares e não-lineares, que até então não foi abordada nessa classe de modelos. Em segundo lugar, um estudo acerca da qualidade das estimativas obtidas em três métodos de estimação considerando dados gerados por quatro situações: sem observações influentes, com outliers na variável resposta, com observações influentes de média alavancagem e com observações influentes de alta alavancagem. São analisados dois métodos de máxima verossimilhança e um método robusto. Foram considerados modelos de regressão linear e não-linear com erros logísticos tipo II e t-Student. Em terceiro lugar é discutida detecção de observações influentes mediante a distância de Cook generalizada, a estatística de Peña e a estatística de Andrews and Pregibon (1978). Em particular é discutida a conveniência de utilizar a metodologia de limiares simulados para caracterizar uma observação como influente ou não influente, assim como o efeito da estimação de parâmetros na construção de limiares. Estas medidas foram aplicadas a conjuntos de dados simulados e reais considerando o ajuste de alguns modelos simétricos com uma adaptação no método de estimação scoring de Fisher. Apresentamos, também, aproximações para a distância de Cook generalizada e para a estatística de Peña baseadas no modelo de exclusão de casos que fornece uma forma mais simples e com menos custo computacional para o cálculo das estatísticas. Além disso, apresentamos a matriz de informação de Fisher para o modelo t-Student e exponencial potência.

1.1 Organização da Dissertação

No capítulo 2 definiremos a classe de modelos linear e não-linear simétricos, assim como a função log-verossimilhança, função escore e matriz de informação de Fisher e a estimação dos parâmetros.

No capítulo 3 apresentaremos um estudo de simulação para avaliar o desempenho de três métodos de estimação em dados gerados por quatro situações: sem observações influentes, com outliers na variável resposta, com observações influentes de média alavancagem e com observações influentes de alta alavancagem. Foram considerados modelos de regressão linear e não-linear com erros logísticos tipo II e t-Student.

No capítulo 4 d as estatísticas de influência distância de Cook generalizada, estatística de Peña e a estatística de Andrews. Também, será descrita a metodologia de limiares para identificar observações influentes baseada em simulações de Monte Carlo. No capítulo 5, apresentamos as aplicações com dados simulados e dados reais, através do ajuste de alguns modelos lineares e não-lineares simétricos.

O capítulo 6 apresenta as conclusões e considerações finais a respeito dos resultados encontrados. O apêndice A aborda conceitos a respeito do modelo de exclusão de casos que fornecem uma estrutura teórica para o desenvolvimento de aproximações para a distância de Cook generalizada e a estatística de Peña, além de apresentar os cálculos para obtenção da matriz de informação de Fisher no modelo t-Student e exponencial potência.

Capítulo 2

Modelos de regressão com erros simétricos

Em muitas situações da modelagem estatística os dados apresentam observações discrepantes, por esse motivo há necessidade de estudar os efeitos destas observações nas estimativas do modelo. Inferência estatística baseada na distribuição normal é conhecida por ser vunerável a observações aberrantes. Como alternativa, métodos robustos foram desenvolvidos com o objetivo de acomodar observações extremas (Rousseeuw and Leroy, 1987 e Maronna et al., 2006). Com a mesma finalidade, podem ser considerados modelos em que a distribuição dos erros apresentam caudas mais pesadas do que as da normal.

Os modelos simétricos de regressão, que incluem todas as distribuições contínuas simétricas para os erros, tem recebido uma crescente atenção nos últimos anos. Por exemplo, Lange et al. (1989) discutem o uso da distribuição t-Student em modelos de regressão linear e não-linear onde os graus de liberdade são usados para controlar a curtose. Taylor (1992) propõe o ajuste do modelo de regressão linear sob erros expnencial potência com um parâmetro extra de forma. Arellano-Valle (1994) apresenta resultados sobre regressão usando a distribuição t-Student. Estes resultados foram estendidos para os modelos lineares simétricos por Ferrari and Uribe-Opazo (2001). No trabalho de Cordeiro et al. (2000) foi a correção de viés dos estimadores de máxima verossimilhança para modelos não-lineares simétricos. Esses resultados foram estendidos para os modelos não-lineares simétricos heterocedásticos. Vanegas e Cysneiros (2010) avaliam métodos de diagnóstico em modelos simétricos não-lineares. Citamos aqui apenas algumas contribuições para a classe de modelos simétricos.

Neste capítulo apresentamos a classe de modelos lineares e não-lineares simétricos. Na seção 2.2 descrevemos a função verossimilhança, função escore e a matriz de informação. Na seção 2.3 apresentamos alguns métodos de estimação. E na seção 3 apresentamos um estudo de simulação com um modelo linear e três modelos não-lineares simétricos comparando os métodos de estimação cujo objetivo é avaliar qualidade da estimação na presença de *outliers* e pontos influentes.

2.1 Modelo

Sejam $\epsilon_1, \ldots, \epsilon_n$ são variáveis aleatórias independentes com função de densidade

$$f_{\epsilon_i}(\epsilon) = \frac{1}{\sqrt{\phi}} g\left(\epsilon^2/\phi\right), \quad \epsilon \in \mathbb{R}, \tag{2.1.1}$$

em que $g(\cdot)$ é denominada função geradora de densidades, com $g(u) > 0$, para $u > 0$ e $\int_0^\infty g(u)du < \infty$. Usamos a notação $\epsilon_i \sim S(0, \phi, g)$ para dizer que ϵ_i foi gerado a partir de uma distribuição simétrica. O modelo não-linear simétrico é definido por

$$y_i = \mu_i(\boldsymbol{\beta}; \mathbf{x}_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.1.2}$$

em que $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ o vetor de respostas observadas e $\mu_i(\boldsymbol{\beta}; \mathbf{x}_i)$ é uma função não-linear contínua e diferenciável com respeito à $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ tal que a matriz de derivadas $\mathbf{D}_\beta = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$ tenha posto p $(p < n)$ para todo $\boldsymbol{\beta} \in \Omega_\beta \subset \mathbb{R}^p$, com Ω_β um conjunto compacto com pontos interiores. Além disso, considere $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ e $\mathbf{x}_i$ um vetor de variáveis explanatórias. No caso linear, tem-se que $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ com $\mathbf{D}_\beta = \mathbf{X}$ uma matriz $n \times p$ de posto completo cuja i-ésima linha é denotada por $\mathbf{x}_i^{\mathrm{T}} = (1, x_{i1}, \ldots, x_{ip-1})$.

A densidade de y_i é dada por

$$f_{y_i}(y_i) = \frac{1}{\sqrt{\phi}} g\left(u_i\right), \quad y_i \in \mathbb{R}, \tag{2.1.3}$$

com $u_i = (y_i - \mu_i)^2/\phi$ e $y_i \sim S(\mu_i, \phi, g)$. A função característica de y_i pode ser expressa como $\varsigma_{y_i}(t) = E\left(\mathrm{e}^{ity_i}\right) = \mathrm{e}^{it\mu}\varphi\left(t^2\phi\right)$, $t \in \mathbb{R}$ para alguma função $\varphi(\cdot)$, com $\varphi(u) \in \mathbb{R}$ para $u > 0$. Quando existem, $E(y_i) = \mu_i$ e $\mathrm{Var}(y_i) = \xi\phi$, onde $\xi > 0$ é uma constante dada por $\xi = -2\varphi'(0)$ com $\varphi'(0) = \left\{\frac{\partial\varphi(u)}{\partial u}\right\}_{u=0}$ (Fang et al., 1990, p. 43). Na classe dos modelos simétricas estão inclusos os modelos que apresentam erros com distribuição normal, t-Student, exponencial potência, logística, entre outros.

2.2 Verossimilhança, função escore e matriz de informação

Para o modelo definido por (2.1.1) e (2.1.2), a função verossimilhança de $\boldsymbol{\theta}$ baseada em uma amostra $y_i, \ldots, y_n$ independente é dada por

$$L\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n} l(y_i; \boldsymbol{\theta}) = -\frac{n}{2}\log\phi + \sum_{i=1}^{n} \log\left[g\left(u_i\right)\right]. \tag{2.2.1}$$

Em alguns modelos simétricos, como o normal e logístico tipo II, o vetor de parâmetros desconhecidos é $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \phi)^{\mathrm{T}}$. Outros modelos podem ter um ou mais parâmetros, além de $\boldsymbol{\beta}$ e ϕ, chamados de parâmetros extras. Os modelos t-Student e exponencial potência, por exemplo, tem como parâmetro extra o parâmetro de forma (ν), na família t-Student também é chamado de grau de liberdade. Estes quatro modelos receberam especial atenção neste trabalho.

Para obter a função escore e a matriz de informação de Fisher é preciso derivar $L(\boldsymbol{\theta})$ com respeito aos parâmetros desconhecidos e em então calcular alguns momentos dessas derivadas.

Ao considerar $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \phi)^{\mathrm{T}}$ a função escore será definida por $\mathbf{U}(\boldsymbol{\theta}) = \left(\mathbf{U}_{\beta}^{\mathrm{T}}(\boldsymbol{\theta}), \mathrm{U}_{\phi}(\boldsymbol{\theta})\right)^{\mathrm{T}}$ com

$$\mathbf{U}^{\mathrm{T}}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{U}_{\beta}(\boldsymbol{\theta}) \\ \mathrm{U}_{\phi}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \phi^{-1}\mathbf{D}_{\beta}\mathbf{D}(\mathbf{v})(\mathbf{y}-\boldsymbol{\mu}) \\ (2\phi)^{-1}\left[\phi^{-1}Q_{\mathbf{V}}(\boldsymbol{\beta}) - n\right] \end{bmatrix}, \tag{2.2.2}$$

onde $Q_{\mathbf{V}}(\boldsymbol{\beta}) = (\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{D}(\mathbf{v})(\mathbf{y}-\boldsymbol{\mu})$ e $\mathbf{D}(\mathbf{v}) = \mathrm{diag}\{\upsilon_1, \ldots, \upsilon_n\}$ com $\upsilon_i = -2W_g(u_i) = -2\frac{g'(u_i)}{g(u_i)}$, $g'(u_i) = \partial g(u_i)/\partial u_i$. Expressões para $g(u)$, $W_g(u)$ e $W_g'(u)$, com $u > 0$, para algumas distribuições simétricas são apresentadas na Tabela 2.1 (veja Cysneiros and Paula, 2005).

Tabela 2.1: *Expressões para $g(u)$, $W_g(u)$ e $W'g(u)$ para algumas distribuições simétricas.*

Distribuição	$g(u)$	$W_g(u)$	$W_g'(u)$
Normal	$\frac{1}{\sqrt{2\pi}}\exp(-u/2)$	$-\frac{1}{2}$	0
Logística-II	$\frac{\exp(-\sqrt{u})}{[1+\exp(-\sqrt{u})]^2}$	$\frac{\exp(-\sqrt{u})-1}{2\sqrt{u}[1+\exp(-\sqrt{u})]}$	$\frac{2\exp(-\sqrt{u})\sqrt{(u)}+\exp(-2\sqrt{u})-1}{-4u^{3/2}[1+\exp(-\sqrt{u})]^2}$
t-Student	$\frac{\nu^{\nu/2}}{\mathrm{Beta}(1/2,\nu/2)}(\nu+u)^{-\frac{\nu+1}{2}}, \quad \nu>0$	$-\frac{\nu+1}{2(\nu+u)}$	$\frac{\nu+1}{2(\nu+u)^2}$
Exp. potência	$\frac{\exp\left[-\frac{1}{2}u^{1/(1+\nu)}\right]}{\Gamma(1+\frac{1+\nu}{2})2^{1+(1+\nu)/2}}, \quad -1<\nu\leq 1$	$-\frac{1}{2(1+\nu)u^{\nu/(\nu+1)}}$	$\frac{\nu}{(1+\nu)^2 2u^{(2\nu+1)/(1+\nu)}}$

No caso dos modelos t-Student e exponencial potência podemos considerar $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\tau}^{\mathrm{T}})^{\mathrm{T}}$, com $\boldsymbol{\tau} = (\phi, \nu)^{\mathrm{T}}$. Desta forma, tem-se $\mathbf{U}(\boldsymbol{\theta}) = \left(\mathbf{U}_{\beta}^{\mathrm{T}}(\boldsymbol{\theta}), \mathbf{U}_{\tau}^{\mathrm{T}}(\boldsymbol{\theta})\right)^{\mathrm{T}}$, em que

$$\mathbf{U}_{\tau}(\boldsymbol{\theta}) = \left(\mathrm{U}_{\phi}(\boldsymbol{\theta}), \mathrm{U}_{\nu}(\boldsymbol{\theta})\right)^{\mathrm{T}}.$$

Assim, calculamos a função escore para os graus de liberdade para ,o modelo t-Student,

$$\mathrm{U}_{\nu}(\boldsymbol{\theta}) = \frac{n}{2}\left[1 - R(\nu)\right] - \frac{1}{2}\sum_{i=1}^{n}\log\left(1+\frac{u_i}{\nu}\right) + \frac{\nu+1}{\nu+u_i},$$

e para o modelo exponencial potência

$$\mathrm{U}_{\nu}(\boldsymbol{\theta}) = n\frac{C'(\nu)}{C(\nu)} + \frac{1}{2(\nu+1)^2}\sum_{i=1}^{n} u_i^{\frac{1}{\nu+1}}\log(u_i),$$

onde

$$R(\nu) = \Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right)$$

e

$$\frac{C'(\nu)}{C(\nu)} = -\frac{1}{2}\left[\Psi\left(1+\frac{1+\nu}{2}\right) + \log(2)\right],$$

em que $C'(\nu) = \partial C(\nu)/\partial\nu$, com $C(\nu) = 1/\left[\Gamma(1+(1+\nu)/2)2^{1+(1+\nu)/2}\right]$ e $\Psi(\cdot)$ denota a função

digama.

A matriz de informação observada é dada por

$$\mathbf{J}(\boldsymbol{\theta}) = -\frac{\partial \mathbf{U}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}}.$$

Quando $\theta = (\boldsymbol{\theta}^{\mathrm{T}}, \phi)^{\mathrm{T}}$

$$-\mathbf{J}(\boldsymbol{\theta}) = -\begin{bmatrix} \mathbf{J}_{\beta\beta} & \mathbf{J}_{\beta\phi} \\ \mathbf{J}_{\phi\beta} & \mathbf{J}_{\phi\phi} \end{bmatrix}$$

e quando $\theta = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\tau}^{\mathrm{T}})^{\mathrm{T}}$, como nos modelos t-Student e exponencial potência, é dada por

$$-\mathbf{J}(\boldsymbol{\theta}) = -\begin{bmatrix} \mathbf{J}_{\beta\beta} & \mathbf{J}_{\beta\tau} \\ \mathbf{J}_{\tau\beta} & \mathbf{J}_{\tau\tau} \end{bmatrix}.$$

A matriz de informação de Fisher, para uma família de densidades simétricas, é um bloco diagonal (Lange et. al, 1989). Um bloco contém os componentes da média e no outro bloco os componentes de escala e de forma, para os modelos t-Student e exponencial potência. Para os modelos que não possuem parâmetros extras ou nos casos em que estes são fixados, o segundo bloco corresponde apenas ao componente de escala, ou seja, a matriz de informação de Fisher é dada por

$$\mathbf{K}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{K}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{\phi\phi} \end{bmatrix}$$

em que

$$\mathbf{K}_{\beta\beta} = \frac{4d_g}{\phi}\mathbf{D}_{\beta}^{\mathrm{T}}\mathbf{D}_{\beta}, \quad \text{onde} \quad d_g = E(W_g^2(Z^2)Z^2) \quad \text{e} \quad Z \sim S(0,1,g)$$

e

$$\mathbf{K}_{\phi\phi} = \frac{n}{4d_g\phi^2}(4f_g - 1),$$

com $f_g = E(W_g^2(Z^2)Z^4)$. A forma da matriz de informação de Fisher para os modelos t-Student e exponencial potência é apresentada no apêndice A.4. Valores de d_g, f_g e ξ para algumas distribuições simétricas estão na Tabela 2.2 (veja Cysneiros and Paula, 2005).

2.3 Estimação

Para o ajuste dos modelos simétricos de regressão propomos três métodos de estimação. O primeiro considera a estimação por máxima verossimilhança de todos parâmetros, simultaneamente, pelo método BFGS (algoritmo de métrica variável) publicado, independentemente, por Broyden, Fletcher, Goldfarb e Shanno, em 1970. No segundo, as estimativas dos parâmetros de regressão ($\boldsymbol{\beta}$) são obtidas por um método robusto e fixadas para obtenção das estimativas dos demais parâmetros do modelo ($\boldsymbol{\tau}$) com o método de máxima verossimilhança com o BFGS. No terceiro método a estimação dos parâmetros de posição e escala é por máxima verossimilhança usando *scoring* de

Tabela 2.2: *Valores de d_g, $f_g(u)$ e ξ para algumas distribuições simétricas.*

Distribuição	d_g	f_g	ξ
Normal	$\frac{1}{4}$	$\frac{3}{4}$	1
Logística-II	$\frac{1}{12}$	0,60749	$\frac{\pi^2}{3}$
t-Student	$\frac{\nu+1}{4(\nu+3)}$	$\frac{3(\nu+1)}{4(\nu+3)}$	$\frac{\nu}{(\nu-2)}$, $\nu > 2$
Exp. potência	$\frac{\Gamma[(3-\nu)/2]}{4(2^{\nu-1})(1+\nu)^2\Gamma(1+\nu/2)}$	$\frac{\nu+3}{4(\nu+1)}$	$2^{(1+\nu)}\frac{\Gamma[3(\nu+1)/2]}{\Gamma[(\nu+1)/2]}$

Fisher como processo iterativo. Neste último, o parêmtro extra é fixado no valor que fornece a maior verossimilhança em uma grade de valores.

2.3.1 Método 1

O estimador de máxima verossimilhança de $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, pode ser obtido resolvendo $\mathbf{U}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. Para a classe de modelos simétricos, geralmente, este sistema só pode ser resolvido por meio de algoritmos iterativos.

Note que o modelo de regressão linear normal é um caso particular de (2.1.2) e (2.1.3). Neste caso, as estimativas de máxima verossimilhança assumem expressões em forma fechada, a saber $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$ e $\hat{\phi} = n^{-1}Q_{\mathbf{I}}(\hat{\boldsymbol{\beta}})$, com $\mathbf{I}$ sendo a matriz identidade. Entretanto, para os demais modelos lineares simétricos e, principalmente, para os modelos não-lineares simétricos as estimativas são obtidas por métodos iterativos. Para estes modelos o método BFGS pode ser usado como algoritmo iterativo.

No método BFGS os parâmetros são estimados conjuntamente qualquer que seja o modelo e a distribuição associada aos erros. Para estimação de modelos sob erros normais e logístico tipo II, por exemplo, nos referimos a esse método como estimação com BFGS. Nos modelos t-Student e exponencial potência o chamamos de estimação completa, pois os parâmetros da regressão, escala e de forma são estimados simultaneamente.

O método BFGS é um dos mais populares métodos quasi-Newton. Neste algoritmo a matriz hessiana da função verossimilhança, $\mathbf{J}$, é aproximada por uma outra matriz simétrica positiva-definida $\mathbf{B}^{(a)}$ tal que

$$\lim_{a\to\infty} \mathbf{B}^{(a)} = \mathbf{J}.$$

A matrix $\mathbf{B}^{(a)}$ é atualizada após cada passo, o que proporciona o acréscimo de conhecimento ganho durante a iteração. A atualização é feita utilizando a forma recursiva

$$\mathbf{B}^{(a+1)} = \mathbf{B}^{(a)} + \frac{\boldsymbol{\gamma}^{(a)}(\boldsymbol{\gamma}^{(a)})^\top}{(\boldsymbol{\gamma}^{(a)})^\top\boldsymbol{\delta}^{(a)}} - \frac{\mathbf{B}^{(a)}\boldsymbol{\delta}^{(a)}(\boldsymbol{\delta}^{(a)})^\top\mathbf{B}^{(a)}}{(\boldsymbol{\delta}^{(a)})^\top\mathbf{B}^{(a)}\boldsymbol{\delta}^{(a)}}, \quad a = 0, 1, 2, \ldots, \tag{2.3.1}$$

em que $\boldsymbol{\gamma}^{(a)} = \mathbf{U}(\boldsymbol{\theta}^{(a+1)}) - \mathbf{U}(\boldsymbol{\theta}^{(a)})$ e $\boldsymbol{\delta}^{(a)} = \boldsymbol{\theta}^{(a+1)} - \boldsymbol{\theta}^{(a)}$. Assim, o processo iterativo é dado por

$$\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \lambda^{(a)}(\mathbf{B}^{(a)})^{-1}\mathbf{U}(\boldsymbol{\theta}^{(a)}),$$

em que o termo $\lambda^{(a)}$ é um escalar determinado por algum procedimento de busca linear a partir de $\boldsymbol{\theta}^{(a)}$ na direção $\mathbf{B}^{-1}(\boldsymbol{\theta}^{(a)})\mathbf{U}(\boldsymbol{\theta}^{(a)})$. É comum considerar a matriz inicial, $\mathbf{B}^{(0)}$, como sendo a matriz identidade de mesma ordem, pois ela é positiva definida e simétrica, o que resulta em aproximações $\mathbf{B}^{(a)}$ positivas definidas e simétricas.

2.3.2 Método 2

Outra forma de obter o ajuste dos modelos simétricos é por meio da maximização do perfil da verossimilhança de $\boldsymbol{\tau} = (\phi, \nu)^{\mathrm{T}}$, $L_{\boldsymbol{\beta}}(\boldsymbol{\tau})$, com $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$, em que $\tilde{\boldsymbol{\beta}}$ é o estimador L_1. Portanto, as estimativas de $\boldsymbol{\beta}$ são obtidas resolvendo

$$\tilde{\boldsymbol{\beta}} = \arg\min \frac{1}{n}\sum_{i=1}^{n} |y_i - \mu(\mathbf{x}_i^{\mathrm{T}}; \boldsymbol{\beta})| \quad \text{e}$$

$$\hat{\boldsymbol{\tau}} = \arg\max L_{\tilde{\boldsymbol{\beta}}}(\boldsymbol{\tau}).$$

Assim como na modelagem usando distribuições com caudas mais pesadas que as da distribuição normal, um critério para modelagem usando a regressão L_1 é dar menor peso para grandes desvios, logo é mais resistente aos efeitos de *outliers* e muitas vezes é usada como uma técnica para regressão robusta.

Se os erros do modelo são independentes e identicamente distribuídos a partir de uma distribuição exponencial dupla, então $\tilde{\boldsymbol{\beta}}$ é a estimativa de máxima verossimilhança de $\boldsymbol{\beta}$ e portanto condições regulares não satisfeitas, neste caso. Em geral a regressão L_1 estima a mediana de $y|x$, não a média condicional.

No caso linear, se os erros são *iid* a partir de uma distribuição F_0 com densidade f_0, a mediana populacional é $\zeta_\rho = F^{-1}(\rho)$ com $\rho = 1/2$, e a mediana amostral é $\hat{\zeta}_{1/2} = \hat{F}^{-1}(1/2)$. Assumindo a versão padronizada de f_0, $f(u) = (1/\sqrt{\phi})f_0(u/\sqrt{\phi})$, considere $\mathbf{Q}_n = n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ e suponha que em grandes amostras $\mathbf{Q}_n \to \mathbf{Q}_0$, uma matriz fixa, então

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N_p(0, \omega\mathbf{Q}_0^{-1}), \tag{2.3.2}$$

onde $\omega = \phi/4\left\{f_0\left[F_0^{-1}(1/2)\right]\right\}^2$. Por exemplo, se f é a densidade da normal padrão, $f_0\left[F_0^{-1}(1/2)\right] = 1/\sqrt{2\pi} = 0,399$ e $\sqrt{\omega} = 0,5\sqrt{\phi}/0,399 = 1,26\sqrt{\phi}$, então no caso normal os erros padrões dos estimadores L_1 são 26% maiores do que os erros padrões dos estimadores de mínimos quadrados. Se f é densidade da logística-II padronizada $f_0\left[F_0^{-1}(1/2)\right] = 0,25$ e $\sqrt{\omega} = 0,5\sqrt{\phi}/0,25 = 2\sqrt{\phi}$, considerando o resultado assintótico para os estimadores de máxima verossimilhança do modelo logístico-II, isto é,

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, 3\phi(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}),$$

então neste caso os erros padrões dos estimadores L_1 são 16% maiores do que os erros padrões dos

estimadores de máxima verossimilhança.

No caso não-linear, o erro padrão das estimativas é obtido utilizando uma abordagem bootstrap com base na linearização final do modelo avaliado nos parâmetros estimados. Mais detalhes sobre estimação quantilica em modelos de regressão linear e não-linear podem ser encontrados em Koenker (2005), por exemplo.

2.3.3 Método 3

Neste método as estimativas de máxima verossimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top}, \phi)^{\top}$ são obtidas pelo processo iterativo *scoring* de Fisher. Quando o modelo simétrico apresentar outros parâmetros, além de $\boldsymbol{\beta}$ e ϕ, é criada uma grade de valores para o parâmetro adicional. Para cada valor é calculado $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\top}, \hat{\phi})^{\top}$ e escolhido como estimativa do parâmetro extra, aquele valor com maior verossimilhança. Como é o caso dos modelos sob erros t-Student e exponencial potência, em que $\boldsymbol{\beta}$ e ϕ são estimados com $\nu = \nu^*$, em que ν^* produz o maior valor para $L_{\nu}(\boldsymbol{\beta}, \phi)$.

Taylor (1992) e Lange et. al (1989) mostram que os estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e de ϕ são assintoticamente não correlacionados para as distribuições t-Student e exponencial potência, quando o parâmetro de forma é fixado. Como consequencia deste resultado os erros-padrão assintóticos dos coeficientes não são afetados pela estimação do parâmetro de escala ou do parâmetro de forma.

O método *scoring* de Fisher é utilizado nos casos em que a função verossimilhança é uma função geral, digamos $L(\boldsymbol{\theta})$. Neste algoritmo a matriz hessiana da função verossimilhança, $\mathbf{J}$, é aproximada por seu valor esperado, $\mathbf{K}$, a matriz de informação de Fisher. Neste caso, o processo iterativo é dado por

$$\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \mathbf{K}^{-1}(\boldsymbol{\theta}^{(a)})\mathbf{U}(\boldsymbol{\theta}^{(a)}), \quad a = 0, 1, 2, \ldots.$$

No contexto dos modelos simétricos com $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top}, \phi)^{\top}$, o processo iterativo assume a forma

$$\boldsymbol{\beta}^{(a+1)} = \boldsymbol{\beta}^{(a)} + (4d_g)^{-1}[\mathbf{D}_{\beta}^{\top(a)}\mathbf{D}_{\beta}^{(a)}]^{-1}\mathbf{D}_{\beta}^{(a)}\mathbf{D}(\mathbf{v}^{(a)})[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(a)})]$$

e

$$\phi^{(a+1)} = \frac{1}{n}\left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(a+1)})\right]^{\top}\mathbf{D}(\mathbf{v}^{(a+1)})\left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(a+1)})\right], \quad a = 0, 1, 2, \ldots.$$

Sob certas condições de regularidade (Cox and Hinkley, 1974), $\hat{\boldsymbol{\beta}}$, o estimador de máxima verossimilhança de $\boldsymbol{\beta}$, é um estimador consistente e

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N_p(0, \mathbf{V}_{\beta\beta}^{-1}), \quad \text{em que} \quad \mathbf{V}_{\beta\beta} = \lim_{n\to\infty}\frac{1}{n}\mathbf{K}_{\beta\beta}. \tag{2.3.3}$$

Assim, $\mathbf{K}_{\hat{\beta}\hat{\beta}}^{-1} = \frac{\hat{\phi}}{d_g}\left(\mathbf{D}_{\hat{\beta}}^{\top}\mathbf{D}_{\hat{\beta}}\right)^{-1}$ é um estimador consistente para matriz de variância e covariância assintótica de $\hat{\boldsymbol{\beta}}$. No caso linear $\mathbf{K}_{\hat{\beta}\hat{\beta}}^{-1} = \frac{\hat{\phi}}{4\hat{d}_g}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}$. De modo similar o estimador de máxima verossimilhança $\hat{\phi}$ é um estimador consistente de ϕ, e

$$\sqrt{n}(\hat{\phi}-\phi) \xrightarrow{d} N(0, \mathbf{V}_{\phi\phi}^{-1}), \quad \text{em que} \quad \mathbf{V}_{\phi\phi} = \lim_{n\to\infty} \frac{1}{n}\mathbf{K}_{\phi\phi}. \tag{2.3.4}$$

Capítulo 3

Estudo de Simulação

Neste capítulo apresentamos um estudo de simulação para avaliar a qualidade dos três métodos de estimação quando os dados apresentam *outliers* e pontos influentes.

Foram considerados um modelo de regressão linear simples e dois modelos não-lineares. Além disso, para os erros dos modelos consideramos as distribuições:

1. Logística-II;

2. t-Student com $\nu = 2,5; 5; 10$;

As amostras foram geradas com 50, 100 e 200 observações. Para cada modelo, tipo de erro e tamanho de amostra específicos foram consideradas 500 replicações. Além disso, foram gerados dados sem pontos influentes, com observações atípicas na variável resposta, com observação atípicas de média e alta alavangem.

Foi construido um programa em `R` para calcular as estimativas. Para o Método 1 de estimação utilizamos o comando `optim` sob o método BFGS. Para o segundo método foram utilizadas as funções `rq` e `nlrq` do pacote `quantreg`. Para estimação com o Método 2 no modelo lineare não-linear, respectivamente, e em seguida o optim. E para o Método 3 utilizamos a função `elliptical` da `library elliptical` desenvolvida por Cysneiros, Paula e Galea (2005). Com as 500 réplicas de cada modelo, obtemos o ajuste e calculamos a média e o desvio padrão das estimativas. Além disso, obtemos as estimativas da variância dos erros com base nas estimativas do parâmetro de escala e do parâmetro de forma.

3.1 Modelo linear

Para análise de estimação no caso linear consideramos o modelo

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{3.1.1}$$

com $(\beta_0, \beta_1) = (110, -4)$, $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ fixo para todas as réplicas e gerado a partir da distribuição $N(13, 16)$ e $\epsilon_i \overset{iid}{\sim} S(\mu, \phi, g)$, $i = 1, \ldots, n$, com $\mu = 0$ e $\phi = 90$.
Para cada distribuição, com $n = 50, 100, 200$, foram gerados:

- ***Situação (1)*** - dados sem pontos influentes;
- ***Situação (2)*** - dados com três *outliers* de 5 desvios padrões na variável resposta, a saber, $y_i^* = \mu_i + 5\sigma$, $i = 15, 16, 17$;
- ***Situação (3)*** - dados com três *outliers* de média alavancagem, a saber, $y_i^* = \mu_i + 5\sigma$ com $\mathbf{x}_i^* = 30$, $i = 15, 16, 17$ e
- ***Situação (4)*** - dados com três *outliers* de alta alavancagem, a saber, $y_i^* = \mu_i + 5\sigma$ com $\mathbf{x}_i^* = 45$, $i = 15, 16, 17$.

onde $\sigma = \sqrt{\phi\xi}$ é o desvio padrão dos erros. Valores de ξ para algumas distribuições simétricas estão na Tabela 2.2.

Para o modelo t-Student, os seguintes métodos de estimação foram considerados:

1. Estimação conjunta de $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \phi, \nu)^{\mathrm{T}}$ por máxima verossimilhança (Método 1);
2. Estimação de $\boldsymbol{\beta}$ usando regressão quantílica (L1) e a estimação de (ϕ, ν) por MV considerando $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ (Método 2);
3. Estimação de $(\boldsymbol{\beta}, \phi)$ por MV com $\nu = \nu^*$, o grau de liberdade que produz o maior valor para $L_\nu(\boldsymbol{\beta}, \phi)$. Neste método, usa-se o *scoring* de Fisher (Método 3). Para encontrar ν^* foram avaliados 50 valores para ν no conjunto $[2, 30]$.

Nos tópicos que seguem apresentamos os resultados da estimação em cada modelo de acordo com situação em que os dados foram gerados.

3.1.1 Resultados da estimação sob erros logístico tipo II

Nesta parte, apresentamos os resultados do ajuste do modelo linear (3.1.1) sob erros logísticos tipo II. A seguir apresentamos os gráficos correspondentes à análise de uma replicação (ou simulação) específica. Adicionamos a cada gráfico a reta considerando os verdadeiros valores dos parâmetros de regressão e as retas obtidas com estimativas de cada método de estimação.

A Figura 3.1 mostra o gráfico dos dados de um determinada simulação para uma amostra de tamanho 50 segundo a situação em que foram gerados. Na situação (1) e (2) as retas ajustadas dos respectivos processos de estimação seguem a mesma orientação e estão bem próximas da

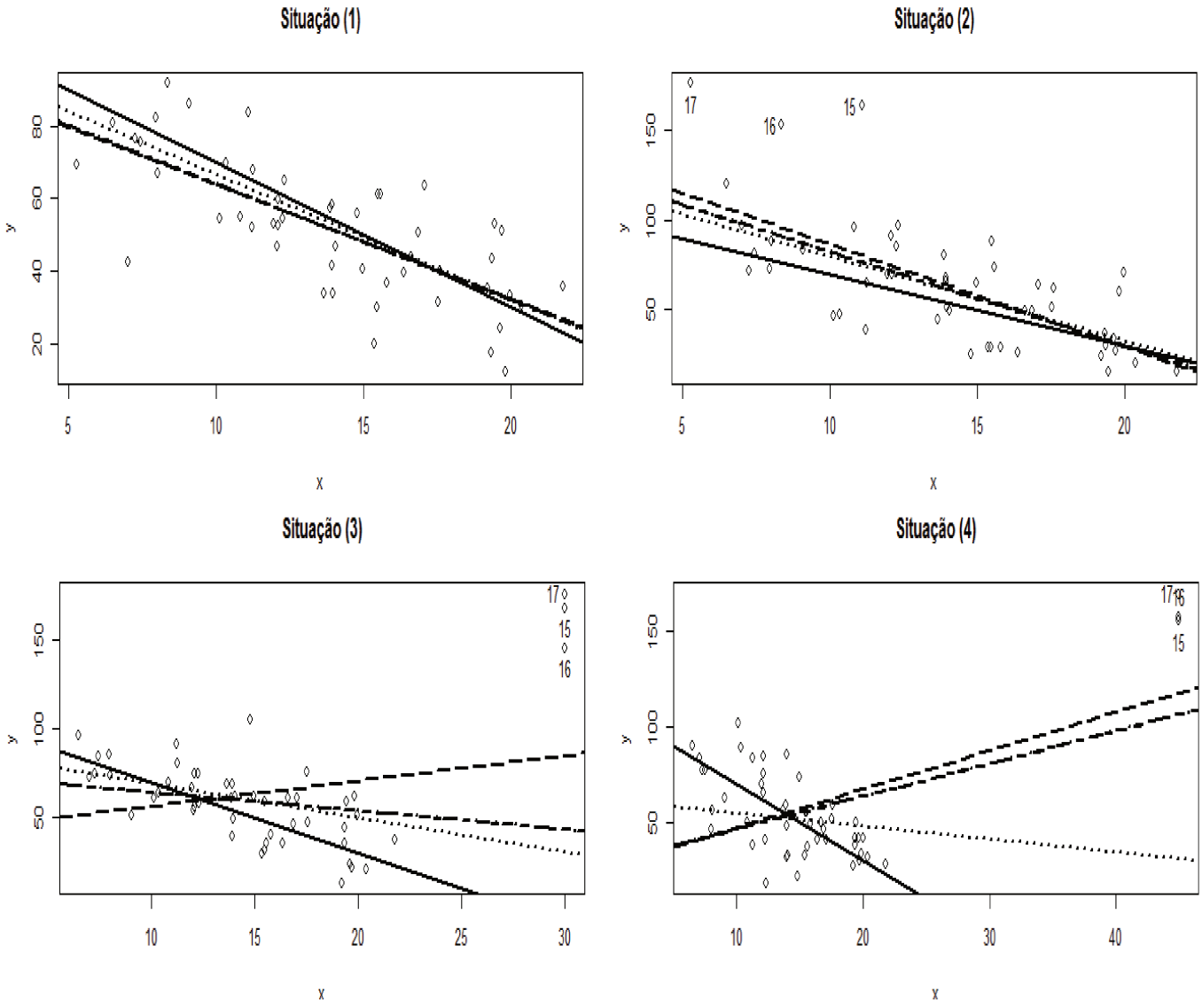


Figura 3.1: *Gráfico dos dados de uma simulação com $n = 50$ para cada situação sob erros logístico tipo II. (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (·····) reta ajustada por dois estágios.*

reta verdadeira, sendo que a reta obtida considerando estimativas com o Método 2 está ainda mais próxima da reta verdadeira. Já nas situações (3) e (4), temos o efeito da alavancagem e verificamos que o ajuste de mínimos quadrados teve maior influência dos pontos atípicos, seguido pelo ajustes considerando MV completa sob erros logísticos tipo II e pelo ajuste com o Método 2.

Nas Figuras 3.2 e 3.3 temos os gráficos referentes às amostras de tamanho 100 e 200, respectivamente. Percebemos que as retas ajustadas estão ainda mais próximas da reta verdadeira na Situação (1) e na Situação (2), mesmo sob a presença de *outliers*. Nas situações (3) e (4) o efeito de alavancagem é menor do que quando consideramos uma amostra de tamnho 50, exceto para o

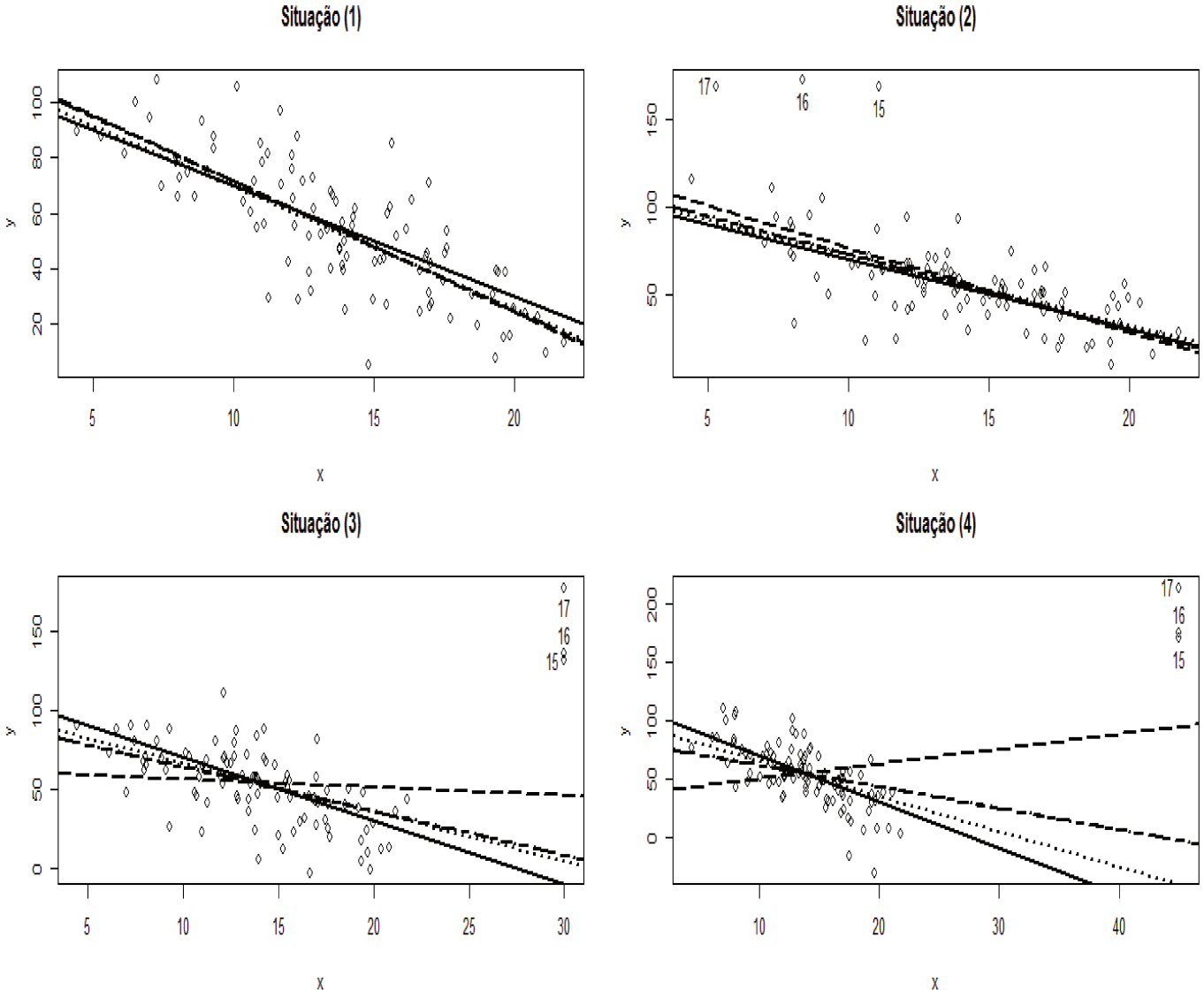


Figura 3.2: *Gráfico dos dados de uma simulação com $n = 100$ para cada situação sob erros logístico tipo II. (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada por dois estágios.*

ajuste de mínimos quadrados que está mais longe da reta verdadeira.

Na Tabela 3.1 mostra a média e desvio padrão das estimativas obtidas pelo ajuste do modelo linear sob erros logísticos tipo II para Situação (1). De modo geral, em média, as estimativas dos parâmetros da regressão ($\boldsymbol{\beta}$) são próximas dos valores verdadeiros e os resultados melhoram a medida que a amostra aumenta. Já a estimativa do parâmetro de escala (ϕ) está mais próxima do valor verdadeiro quando n=50, assim como a estimativa média da varância dos erros (σ^2). Comparando os métodos de estimação, vemos que a média das estimativas são similares, entretanto o desvio padrão é menor para os métodos 1 e 3.

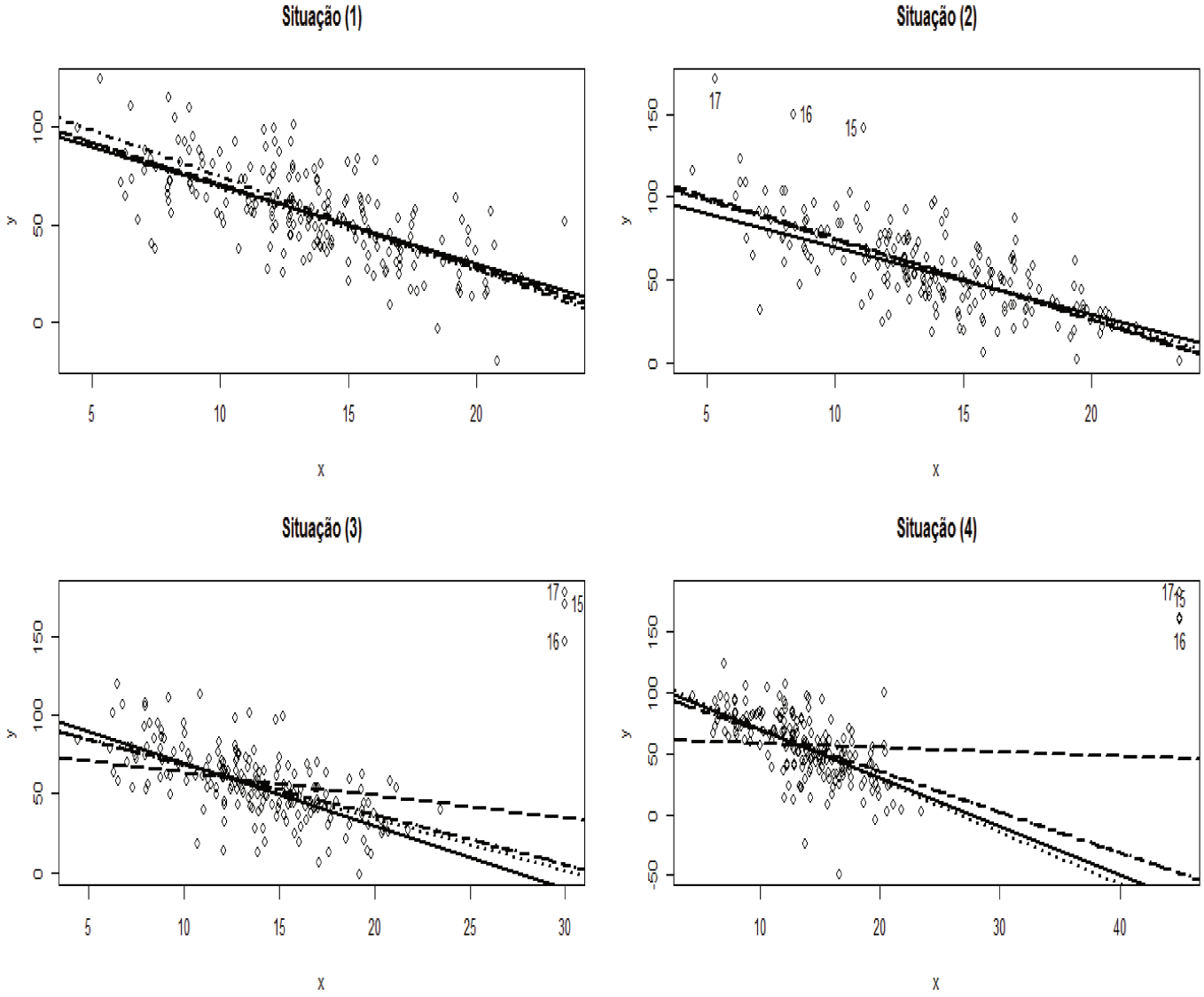


Figura 3.3: *Gráfico dos dados de uma simulação com $n = 200$ para cada situação sob erros logístico tipo II. (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (· · · · · · ·) reta ajustada por dois estágios.*

A Tabela 3.2 mostra a média e desvio padrão das estimativas referente à Situação(2). Percebemos que, em média, as estimativas dos parâmetros da regressão são próximas dos valores verdadeiros, enquanto que o parâmetro de escala foi superestimado em amostras com 50 e 100 observações, consequentemente a varâcia dos erros também foi superestimada. No método robusto as estimativas dos parâmetros de regressão são mais próximas dos parâmetros verdadeiros.

Nas Tabelas 3.3 e 3.4 temos o resumo dos resultados referente ás simulações da Situação (3) e da Situação(4), respectivamente. Observamos que nas duas situações a média das estimativas estão bem afastadas dos valores verdadeiros e que o parâmetro de escala foi superestimado quando

Tabela 3.1: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros logísticos tipo II.* $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$, $\sigma^2 = 296,09$. ***Situação (1)***.

n	Estimação	Média (desvio padrão) $\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\sigma}^2$
50	Método 1	109,998 (7,958)	-3,982 (0,553)	91,860 (19,127)	302,208 (62,925)
	Método 2	109,777 (9,476)	-3,968 (0,659)	92,686 (19,289)	304,926 (63,457)
	Método 3	109,998 (7,958)	-3,982 (0,553)	91,863 (19,127)	302,216 (62,925)
100	Método 1	110,443 (6,068)	-4,040 (0,435)	77,267 (14,832)	254,198 (48,795)
	Método 2	110,475 (6,976)	-4,045 (0,494)	77,617 (14,916)	255,350 (49,072)
	Método 3	110,442 (6,068)	-4,039 (0,435)	77,269 (14,832)	254,206 (48,797)
200	Método 1	110,084 (4,281)	-4,007 (0,303)	97,972 (6,257)	322,316 (20,584)
	Método 2	110,031 (4,956)	-4,002 (0,349)	98,196 (6,265)	323,050 (20,611)
	Método 3	110,085 (4,282)	-4,007 (0,303)	97,975 (6,257)	322,325 (20,586)

Tabela 3.2: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros logísticos tipo II.* $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$, $\sigma^2 = 296,09$. ***Situação (2)***.

n	Estimação	Média (desvio padrão) $\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\sigma}^2$
50	Método 1	124,549 (8,025)	-4,864 (0,553)	148,506 (16,393)	488,565 (53,930)
	Método 2	119,441 (8,890)	-4,565 (0,624)	148,982 (16,466)	490,130 (54,172)
	Método 3	124,550 (8,025)	-4,864 (0,553)	148,510 (16,394)	488,578 (53,933)
100	Método 1	117,013 (6,271)	-4,427 (0,438)	115,779 (11,885)	374,085 (32,294)
	Método 2	115,142 (7,257)	-4,315 (0,510)	116,172 (11,960)	375,231 (32,448)
	Método 3	117,013 (6,271)	-4,427 (0,438)	115,782 (11,885)	374,095 (32,292)
200	Método 1	113,775 (4,405)	-4,233 (0,317)	88,717 (4,670)	291,866 (15,362)
	Método 2	113,112 (4,992)	-4,191 (0,357)	88,894 (4,677)	292,449 (15,385)
	Método 3	113,776 (4,404)	-4,233 (0,317)	88,715 (4,670)	291,862 (15,364)

$n = 50$. Para amostras maiores os resultados melhoram para os parâmetros da regressão, mas o parâmetro de escala ainda é superestimado. Na estimação com o Método 2 a média das estimativas dos parâmetros de regressão são melhores do aquelas obtidas nos outros métodos de estimação.

Em geral, os métodos de estimação apresentaram resultados similares. No ajuste dos dados gerados sem observações influentes, os três métodos de estimação forneceram boas estimativas cujos valores estão próximos dos valores reais dos parâmetros. Na presença de observações extremas na variável resposta, observamos que as estimativas dos parâmetros da regressão foram pouco influenciados pelos pontos atípicos, mas o parâmetro de escala foi superestimado, assim como a variância dos erros. Para as situções com pontos atípicos de média ou alta alavancagem as estimativas de todos os parâmetros sofreram forte influência das observações atípicas em amostras com 50 observações. Em amostras maiores as estimativas de $\boldsymbol{\beta}$ foram pouco influenciadas, mas as estimativas de ϕ superestimam o parâmetro.

Tabela 3.3: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros logísticos tipo II.* $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$, $\sigma^2 = 296,09$. ***Situação (3)***.

n	Estimação	Média (desvio padrão)			
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\sigma}^2$
50	Método 1	69,531 (8,634)	-0,940 (0,561)	248,177 (29,460)	816,470 (96,920)
	Método 2	82,425 (10,083)	-1,906 (0,656)	242,872 (29,305)	799,017 (96,409)
	Método 3	69,529 (8,634)	-0,940 (0,561)	248,182 (29,461)	816,486 (96,924)
100	Método 1	91,057 (5,950)	-2,550 (0,402)	158,266 (13,885)	520,675 (45,681)
	Método 2	95,998 (6,410)	-2,939 (0,433)	157,758 (13,925)	519,004 (45,811)
	Método 3	91,056 (5,950)	-2,550 (0,402)	158,271 (13,885)	520,691 (45,680)
200	Método 1	101,382 (4,218)	-3,322 (0,297)	121,882 (8,644)	400,977 (28,437)
	Método 2	103,794 (4,581)	-3,511 (0,327)	121,941 (8,674)	401,170 (28,538)
	Método 3	101,381 (4,218)	-3,322 (0,297)	121,886 (8,645)	400,988 (28,439)

Tabela 3.4: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros logísticos tipo II.* $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$, $\sigma^2 = 296,09$. ***Situação (4)***.

n	Estimação	Média (desvio padrão)			
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\sigma}^2$
50	Método 1	34,177 (6,182)	1,514 (0,363)	305,411 (51,308)	1004,761 (168,798)
	Método 2	63,110 (12,421)	-0,590 (0,834)	294,342 (49,395)	968,348 (162,502)
	Método 3	34,175 (6,182)	1,514 (0,363)	305,414 (51,308)	1004,772 (168,795)
100	Método 1	91,057 (5,950)	-2,550 (0,402)	158,266 (13,885)	520,675 (45,681)
	Método 2	95,998 (6,410)	-2,939 (0,433)	157,758 (13,925)	519,004 (45,811)
	Método 3	91,056 (5,950)	-2,550 (0,402)	158,271 (13,885)	520,691 (45,680)
200	Método 1	89,842 (6,301)	-2,471 (0,456)	125,068 (19,354)	411,458 (63,671)
	Método 2	93,790 (8,241)	-2,772 (0,600)	124,643 (19,276)	410,059 (63,416)
	Método 3	89,841 (6,300)	-2,471 (0,456)	125,071 (19,355)	411,468 (63,675)

3.1.2 Resultados da estimação sob erros t-Student

Aqui apresentamos os resutados para o ajuste do modelo linear sob erros t-Student. Inicialmente, ilustramos, por meio de gráficos, alguns exemplos dos dados gerados para este modelo. A cada gráfico adicionamos a reta considerando o verdadeiro valor dos parâmetros de regressão e a reta ajustada correspondente a cada método de estimação.

A situação (1) corresponde à dados sem observações atípicas e como podemos observar a as retas ajustadas coincidem com a reta real, como mostra a Figura 3.4. Na Figura 3.5 estão os gráficos de alguns dados simulados para a Situação (2). Neste caso, observamos que o ajuste dos métodos propostos estão ptóximos da reta verdadeira.

Nos exemplos considerando os dados gerados para Situação (3) somente a reta ajustada por mínimos quadrados apresentou orientação bem diferente da reta verdadeira, enquanto que as demais retas ajustadas estão bem próximas da reta verdadeira (veja Figura 3.6).

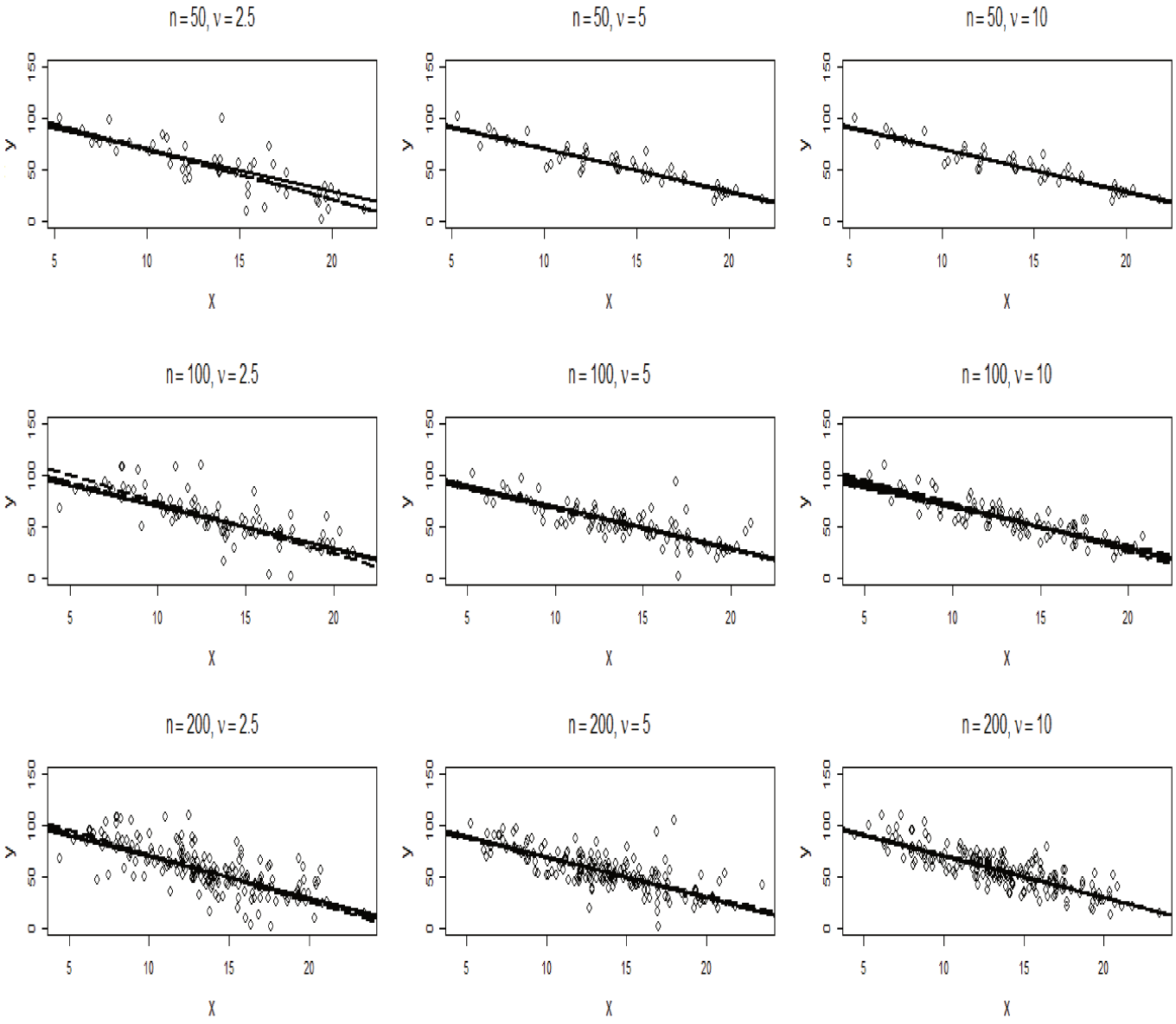


Figura 3.4: *Gráfico dos dados de uma simulação do modelo linear sob erros t-Student da* ***Situação (1)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (·····) reta ajustada pelo Método 2.

Na Figura 3.7 temos os gráficos dos dados gerados na situação (4). Para $n = 50$ o ajuste considerando estimação pelo Método 1 é similar ao ajuste por mínimos quadrados, como podemos ver as retas se coincidem, enquanto que os ajustes pelo Método 2 e pelo Método 3estão mais próximos da reta verdadeira. Para os outros tamanhos de amostra verificamos que a reta pelo método mínimos quadrados está mais afastado da reta verdadeira, mas os outros métodos apresentam resultado melhor com ajuste das reta mais próximo da reta verdadeira.

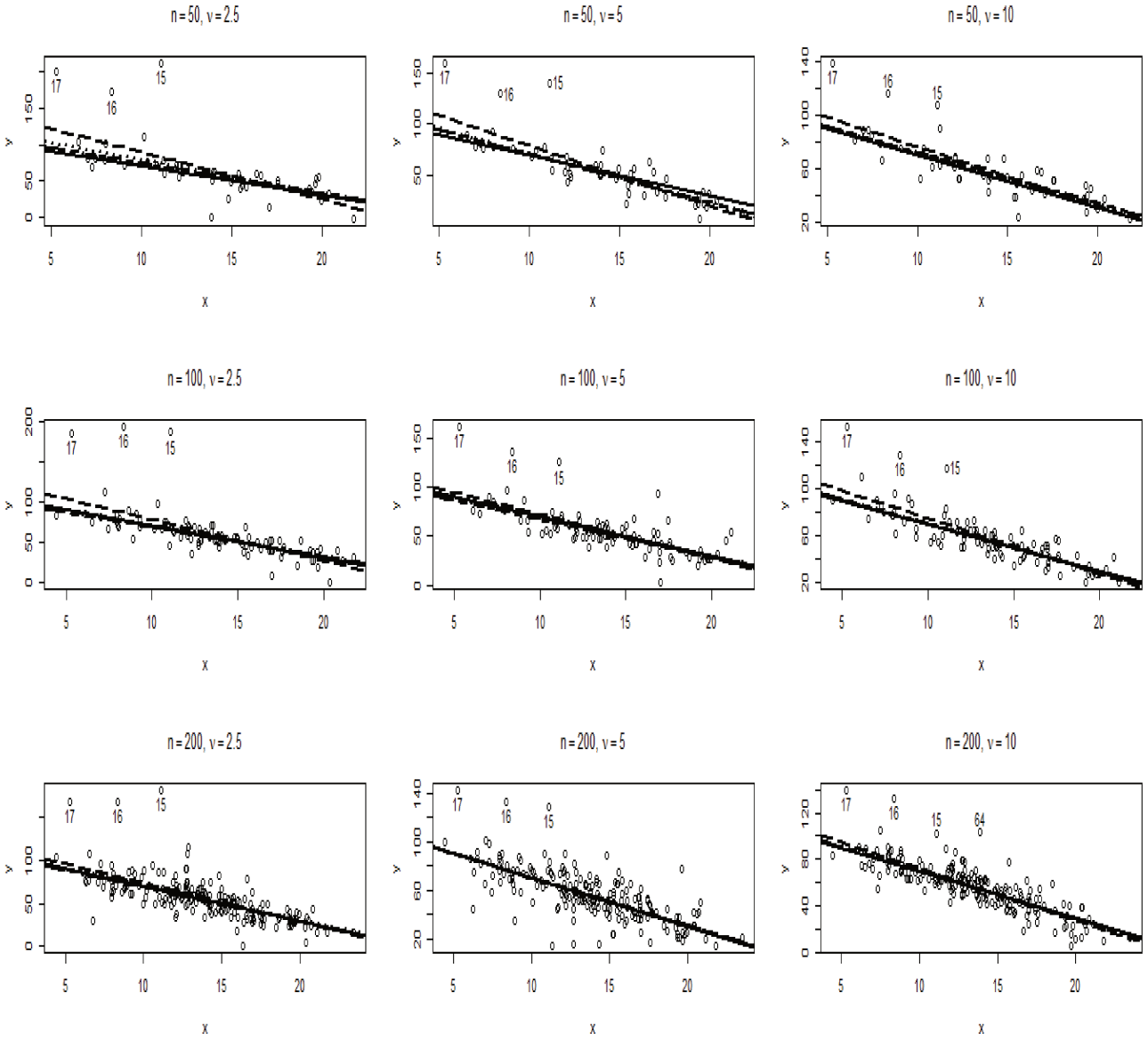


Figura 3.5: *Gráfico dos dados de uma simulação do modelo linear sob erros t-Student da* ***Situação (2)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (·····) reta ajustada pelo Método 2.

Na Tabela 3.5 apresentamos o resultado das estimativas obtidas sem pontos influentes. Observamos que, em média, as estimativas para os parâmetros de regressão estão próximas dos valores verdadeiros, com resultados similares nos três métodos. Para o parâmetro ϕ a média das estimativas são boas, mas com desvio padrão grande. Para o parâmetro ν a média das estimativas é maior que valor verdadeiro, mas a mediana está mais próxima do valor teórico. Portanto, as estimativas do grau de liberdade tem distribuição assimétrica com desvio padrão grande nos três

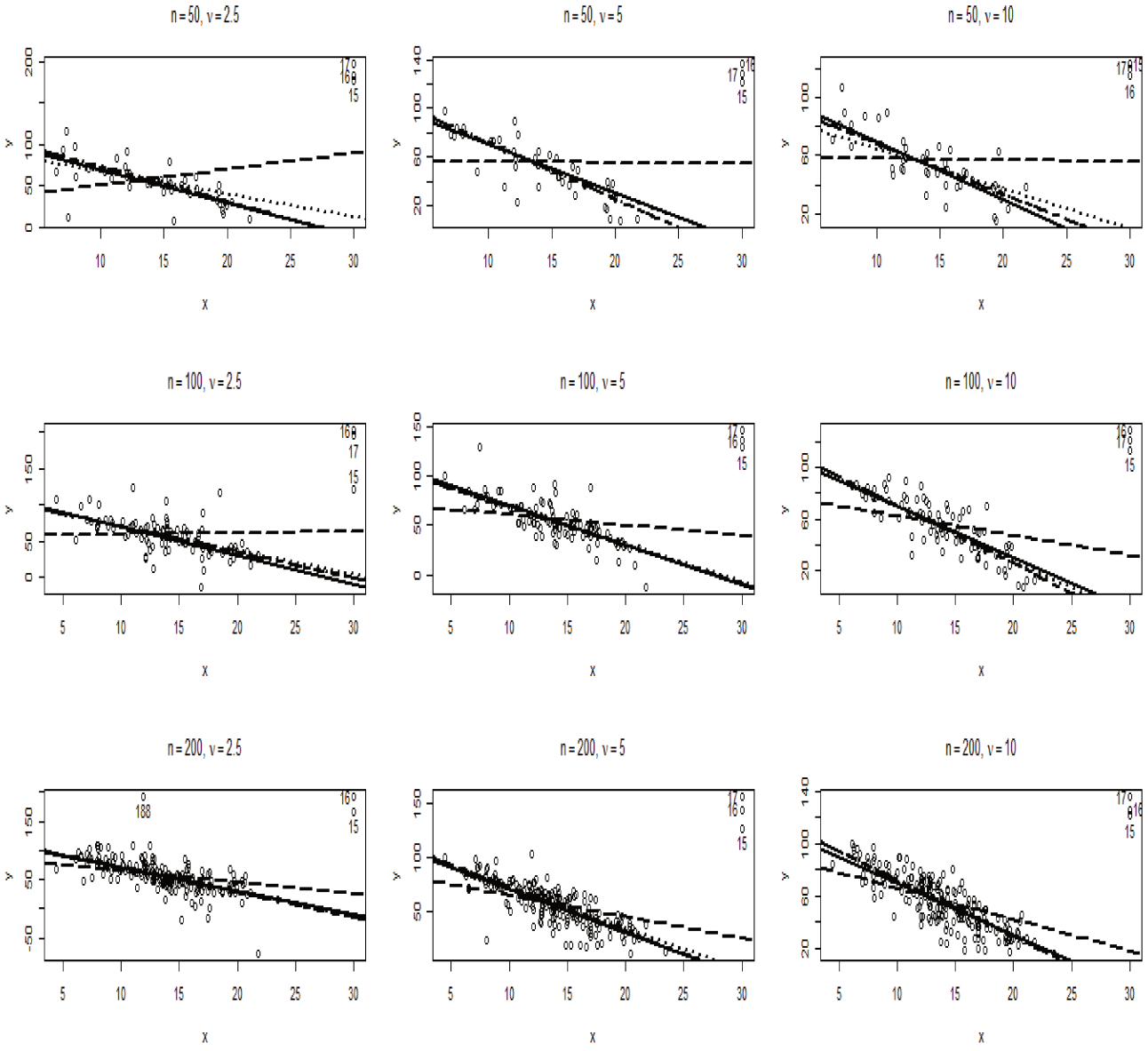


Figura 3.6: *Gráfico dos dados de uma simulação do modelo linear sob erros t-Student da* ***Situação (3)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) reta verdadeira, (− − −) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (− · − · −) reta ajustada pelo Método 3 e (·····) reta ajustada pelo Método 2.

métodos quando $n = 50$. Para as amostras maiores as estimativas estão mais próximas dos valores verdadeiros e apresentam resultados similares nos três métodos.

Ainda na Tabela 3.5 A média das estimativas da variância quando $\nu = 2,5$ está mais próxima do valor verdadeiro no método 1 e 2, mas para estes métodos o desvio padrão obtido são maiores do que o Método 3. Mesmo para as amostras maiores a média das estimativas nos métodos 1 e 2 se afastam do valor verdadeiro, mas no Método 3 a estimativa média se aproxima um pouco da

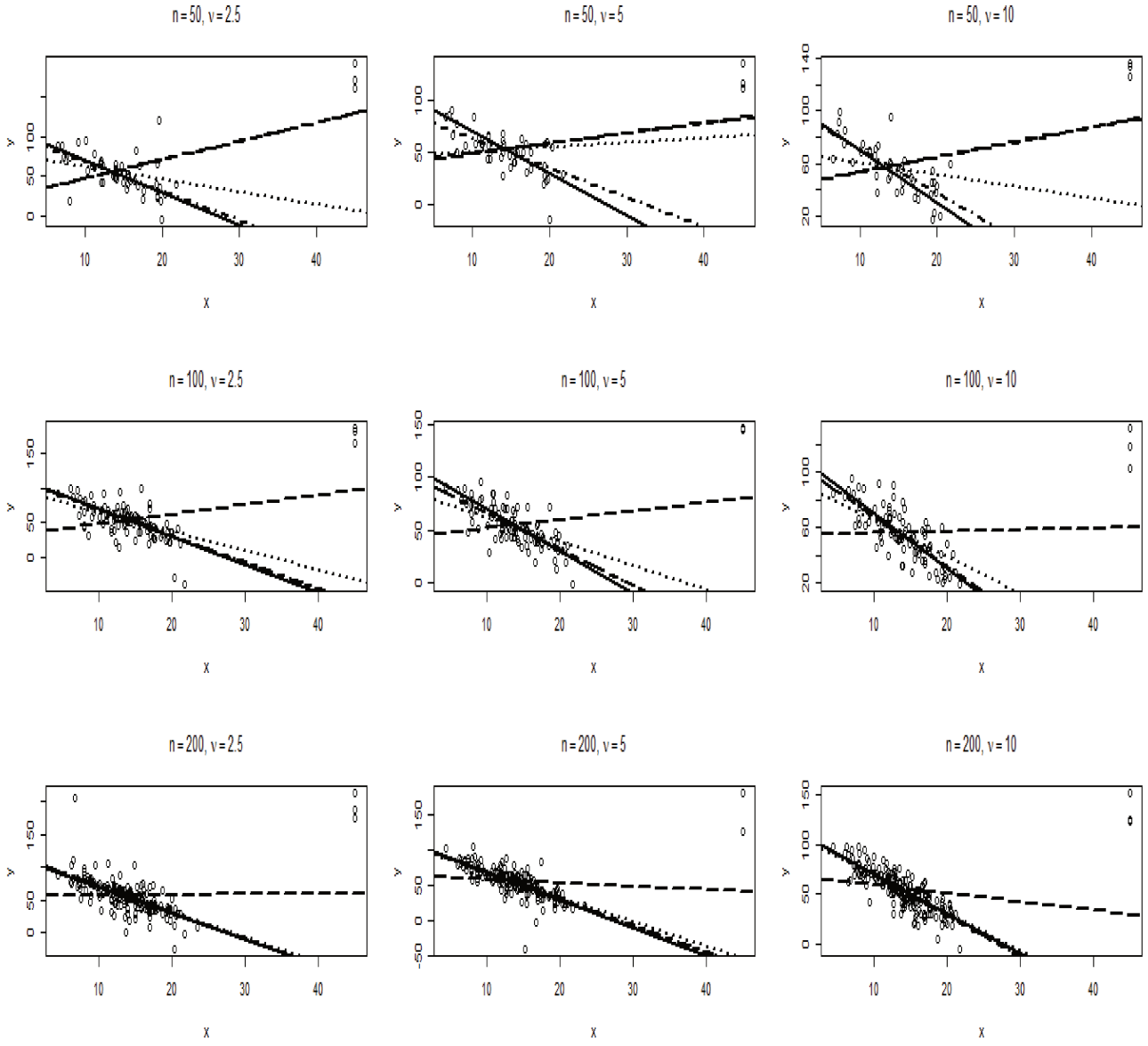


Figura 3.7: *Gráfico dos dados de uma simulação do modelo linear sob erros t-Student da* ***Situação (4)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) reta verdadeira, (— — —) reta ajustada por mínimos quadrados, (- - - -) reta ajustada pelo Método 1, (— · — · —) reta ajustada pelo Método 3 e (·····) reta ajustada pelo Método 2.

variância teórica.

Quando $\nu = 5, 10$ e $n = 50$, verificamos que a média das estimativas da variância foi melhor no Método 3. Para as amostras maiores a estimativa média nos três métodos são similares e está próxima da variância teórica (Veja a Tabela 3.5).

A Tabela 3.6 mostra a média e desvio padrão das estimativas para situação (2). Neste caso, percebemos que o ajuste do modelo linear t-Student sugere uma distribuição com cauda mais pesada

Tabela 3.5: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros t-Student.* $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$. ***Situação (1)***.

ν (σ^2)	n	Estimação	Média (desvio padrão)				
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\nu}^a$	$\hat{\sigma}^2$
		Método 1	110,22 (5,82)	-4,01 (0,40)	101,19 (37,54)	6,6; 3,1 (8,4)	412,8 (879,5)
	50	Método 2	110,05 (6,53)	-4,00 (0,44)	100,32 (36,77)	5,8; 3,0 (7,3)	500,5 (1244,6)
		Método 3	110,21 (5,82)	-4,01 (0,40)	101,23 (37,53)	6,5; 3,1 (8,3)	285,2 (162,6)
		Método 1	110,13 (4,23)	-4,00 (0,29)	96,70 (27,38)	3,7; 2,9 (3,6)	540,5 (1262,0)
2,5 (450)	100	Método 2	110,27 (4,73)	-4,02 (0,33)	96,53 (27,16)	3,6; 2,9 (3,3)	526,5 (1273,7)
		Método 3	110,13 (4,23)	-4,01 (0,29)	96,70 (27,43)	3,7; 2,8 (3,5)	332,6 (171,8)
		Método 1	110,01 (3,13)	-4,00 (0,22)	92,70 (16,81)	2,9; 2,7 (0,8)	971,1 (6194,1)
	200	Método 2	109,99 (3,49)	-4,00 (0,25)	92,65 (16,71)	2,9; 2,7 (0,8)	680,1 (2000,2)
		Método 3	110,01 (3,13)	-4,00 (0,22)	92,79 (16,83)	2,9; 2,8 (0,8)	387,4 (184,2)
		Método 1	109,42 (5,41)	-3,96 (0,37)	90,74 (29,72)	12,4; 6,3 (11,1)	178,8 (267,4)
	50	Método 2	109,19 (6,25)	-3,95 (0,43)	89,92 (29,65)	10,9; 5,7 (10,2)	195,5 (414,0)
		Método 3	109,43 (5,41)	-3,96 (0,37)	90,67 (29,62)	12,1; 6,4 (10,8)	155,2 (75,3)
		Método 1	109,83 (3,81)	-3,98 (0,26)	92,17 (22,60)	9,2; 5,7 (8,6)	154,4 (47,9)
5 (150)	100	Método 2	109,89 (4,29)	-3,99 (0,30)	91,30 (22,30)	8,4; 5,4 (7,7)	157,2 (51,6)
		Método 3	109,83 (3,81)	-3,98 (0,26)	92,05 (22,45)	9,0; 5,7 (8,2)	154,9 (50,6)
		Método 1	109,89 (2,80)	-3,99 (0,20)	91,32 (16,43)	6,9; 5,3 (5,0)	150,0 (31,0)
	200	Método 2	109,78 (3,14)	-3,98 (0,23)	90,94 (16,25)	6,6; 5,3 (4,4)	151,0 (31,7)
		Método 3	109,88 (2,80)	-3,99 (0,20)	91,34 (16,35)	6,8; 5,3 (4,6)	150,1 (33,6)
		Método 1	109,95 (5,01)	-4,00 (0,34)	85,85 (24,88)	18,5; 21,26 (11,5)	119,4 (106,9)
	50	Método 2	109,92 (5,89)	-3,99 (0,41)	85,32 (25,01)	16,6; 12,38 (11,4)	117,2 (42,2)
		Método 3	109,95 (5,01)	-4,00 (0,34)	85,76 (24,81)	18,2; 19,32 (11,4)	113,7 (37,5)
		Método 1	109,95 (3,60)	-3,99 (0,25)	88,28 (19,41)	17,0; 12,1 (11,0)	113,4 (23,3)
10 (112,5)	100	Método 2	109,95 (4,30)	-3,99 (0,30)	87,84 (19,45)	15,9; 10,4 (10,8)	114,8 (24,0)
		Método 3	109,95 (3,60)	-3,99 (0,25)	88,19 (19,31)	16,8; 12,2 (10,9)	113,5 (24,5)
		Método 1	110,10 (2,56)	-4,005 (0,18)	89,48 (14,06)	15,0; 11,2 (9,4)	112,1 (14,9)
	200	Método 2	110,15 (3,13)	-4,007 (0,22)	89,10 (13,99)	14,1; 10,6 (8,9)	112,7 (15,0)
		Método 3	110,10 (2,55)	-4,005 (0,18)	89,40 (13,92)	14,8; 11,1 (9,2)	112,1 (14,9)

[a]Média; mediana (desvio padrão) de ν.

do que a distribuição imposta para simulação. Resultados já esperados, haja vista que ajustes de modelos com caudas pesadas tendem acomodar pontos atípicos diminuindo o efeito destes pontos na estimativa dos regressores. Existem poucas diferenças entre os métodos de estimação, assim, podemos dizer que estes métodos apresentam resultados bem similares com boas estimativas para os parâmetros da regressão, superestimação do parâmetro de escala quando $n = 50$ e $\nu = 2,5$ e subestimação dos graus de liberdade.

Na Situação 2 a estimativa média do garu de liberdade está bem próxima de 2 quando $\nu = 2,5$

Tabela 3.6: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros t-Student.* $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$. ***Situação (2)***.

ν (σ^2)	n	Estimação	Média (desvio padrão)				
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\nu}$	$\hat{\sigma^2}$
		Método 1	112,87 (6,72)	-4,17 (0,45)	120,90 (35,45)	2,00 (0,03)	-
	50	Método 2	116,31 (7,57)	-4,37 (0,51)	123,73 (36,69)	2,00 (0,03)	-
		Método 3	112,87 (6,72)	-4,17 (0,45)	120,89 (35,42)	2,00 (0,03)	-
		Método 1	110,97 (4,68)	-4,06 (0,33)	96,02 (21,24)	2,04 (0,10)	-
2,5 (450)	100	Método 2	112,66 (5,01)	-4,16 (0,35)	96,71 (21,50)	2,04 (0,11)	-
		Método 3	110,97 (4,68)	-4,05 (0,32)	95,94 (21,30)	2,04 (0,11)	-
		Método 1	110,48 (3,21)	-4,03 (0,23)	89,99 (15,37)	2,16 (0,22)	-
	200	Método 2	111,31 (3,63)	-4,08 (0,26)	90,18 (15,36)	2,16 (0,22)	-
		Método 3	110,48 (3,21)	-4,03 (0,23)	89,88 (15,45)	2,16 (0,24)	-
		Método 1	112,94 (6,40)	-4,17 (0,43)	86,55 (28,18)	2,36 (1,82)	-
	50	Método 2	115,23 (7,03)	-4,30 (0,47)	87,81 (27,19)	2,26 (0,43)	-
		Método 3	112,92 (6,39)	-4,17 (0,43)	86,31 (27,95)	2,30 (1,32)	-
		Método 1	111,16 (4,35)	-4,07 (0,30)	81,58 (19,64)	2,71 (0,51)	328,5 (821,6)
5 (150)	100	Método 2	112,05 (4,76)	-4,12 (0,33)	81,79 (19,60)	2,70 (0,50)	336,2 (955,1)
		Método 3	111,18 (4,35)	-4,07 (0,30)	81,64 (19,66)	2,71 (0,52)	315,89 (140,1)
		Método 1	110,75 (2,98)	-4,04 (0,21)	82,62 (13,71)	3,35 (0,57)	211,1 (134,9)
	200	Método 2	111,01 (3,32)	-4,06 (0,24)	82,60 (13,71)	3,34 (0,60)	212,1 (171,5)
		Método 3	110,75 (2,98)	-4,04 (0,21)	82,70 (13,70)	3,36 (0,58)	210,09 (67,6)
		Método 1	112,80 (6,04)	-4,16 (0,39)	79,75 (26,83)	2,55 (0,77)	809,6 (1563,8)
	50	Método 2	114,62 (6,65)	-4,27 (0,44)	81,24 (27,61)	2,56 (0,82)	894,2 (2209,7)
		Método 3	112,80 (6,05)	-4,16 (0,39)	79,79 (26,91)	2,55 (0,77)	411,7 (156,3)
		Método 1	111,80 (4,26)	-4,10 (0,30)	77,21 (18,58)	3,35 (0,95)	246,6 (264,7)
10 (112,5)	100	Método 2	112,51 (4,73)	-4,14 (0,32)	77,35 (18,46)	3,33 (0,91)	249,1 (250,1)
		Método 3	111,80 (4,26)	-4,10 (0,30)	77,22 (18,59)	3,35 (0,95)	224,1 (71,06)
		Método 1	110,92 (2,77)	-4,05 (0,19)	78,96 (12,44)	4,31 (0,90)	154,3 (21,2)
	200	Método 2	111,19 (3,26)	-4,07 (0,23)	78,96 (12,50)	4,28 (0,88)	155,2 (21,2)
		Método 3	110,92 (2,77)	-4,05 (0,19)	79,03 (12,45)	4,32 (0,90)	154,1 (20,9)

e quando $\nu = 5$ nas amostras com 50, consequentemente a média da variância tende ao infinito. Nestes casos, na Tabela 3.6 não apresentamos a média das estimativas da variância. Nos casos em que $\nu = 5, 10$ com 50, 100 e 200 observações, verificamos que a variância foi superestimada nos três métodos.

A Tabela 3.7 mostra o resumo das estimativas para a situação (3). De modo geral, os resultados entre os métodos são próximos, entretanto os métdos 1 e 3 apresentam, em média, estimativas para **fi** e ϕ melhores que na estimação pelo Método 2. Observamos, também, que para $n = 50$, em média, as estimativas de β_0 são subestimadas, enquanto que o parâmetro ϕ é superestimado e as estimativas de β_1 se mantém próximas dos valores verdadeiros. Quando a amostra aumenta, a

Tabela 3.7: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo linear sob erros t-Student.* $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$. ***Situação (3)***.

ν	n	Estimação	Média (desvio padrão)			
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\nu}$
		Método 1	104,193 (6,694)	-3,570 (0,456)	151,767 (41,295)	2,000 (0,000)
	50	Método 2	92,520 (8,541)	-2,694 (0,586)	172,887 (47,425)	2,000 (0,000)
		Método 3	104,192 (6,694)	-3,570 (0,456)	151,790 (41,301)	2,000 (0,000)
2, 5		Método 1	107,565 (5,036)	-3,814 (0,350)	107,848 (19,463)	2,001 (0,010)
	100	Método 2	101,955 (5,650)	-3,383 (0,392)	111,836 (20,276)	2,002 (0,016)
		Método 3	107,565 (5,036)	-3,814 (0,350)	107,858 (19,467)	2,001 (0,012)
		Método 1	109,128 (3,071)	-3,933 (0,217)	89,141 (13,543)	2,023 (0,075)
	200	Método 2	106,601 (3,486)	-3,735 (0,244)	89,959 (13,709)	2,025 (0,077)
		Método 3	109,129 (3,071)	-3,933 (0,217)	89,092 (13,509)	2,021 (0,080)
		Método 1	105,430 (6,456)	-3,659 (0,445)	107,538 (24,605)	2,000 (0,000)
	50	Método 2	94,999 (7,720)	-2,872 (0,529)	121,217 (27,891)	2,002 (0,012)
		Método 3	105,429 (6,456)	-3,659 (0,445)	107,552 (24,608)	2,000 (0,000)
5		Método 1	107,975 (4,166)	-3,846 (0,293)	75,125 (12,739)	2,023 (0,063)
	100	Método 2	103,124 (4,673)	-3,471 (0,330)	77,901 (13,738)	2,037 (0,086)
		Método 3	107,978 (4,164)	-3,846 (0,293)	75,034 (12,721)	2,019 (0,071)
		Método 1	109,413 (2,955)	-3,958 (0,205)	73,454 (10,732)	2,427 (0,235)
	200	Método 2	107,218 (3,333)	-3,785 (0,228)	74,161 (10,829)	2,436 (0,235)
		Método 3	109,408 (2,957)	-3,958 (0,205)	73,505 (10,900)	2,431 (0,251)
		Método 1	104,712 (6,178)	-3,615 (0,421)	90,982 (18,206)	2,000 (0,000)
	50	Método 2	94,613 (7,802)	-2,847 (0,531)	102,475 (21,092)	2,001 (0,008)
		Método 3	104,711 (6,178)	-3,615 (0,421)	90,993 (18,208)	2,000 (0,000)
10		Método 1	107,819 (4,138)	-3,836 (0,285)	70,080 (12,111)	2,090 (0,126)
	100	Método 2	103,017 (4,471)	-3,466 (0,306)	72,538 (12,849)	2,111 (0,140)
		Método 3	107,823 (4,139)	-3,836 (0,286)	70,030 (12,280)	2,088 (0,148)
		Método 1	108,973 (2,797)	-3,923 (0,195)	66,910 (8,964)	2,684 (0,192)
	200	Método 2	106,833 (3,156)	-3,755 (0,219)	67,525 (9,114)	2,693 (0,195)
		Método 3	108,972 (2,799)	-3,923 (0,195)	66,914 (8,942)	2,686 (0,209)

média das estimavas melhoram, inclusive do parâmetro de escala, mas quando $\nu = 5, 10$ este último é subestimado. Além disso, a média das estimativas do parâmetro ν foram subestimadas nos três métodos de estimação. Nesta situação, também observamos que a média das estimativas dos graus de liberdade estão em torno de 2, portanto a estimativa média da variância será superestimada.

Na Tabela 3.8 está o resumo dos resultados para simulação de dados sob a Situação (4). Verificamos que para $n = 50$ a média das estimativas obtidas com a estimação pelo Método 3 é melhor do que a média nos outros métodos. Neste caso, a estimação pelo Método 2 e pelo Método 1 apresentam resultados ruins, principalmente, para as estimativas de ϕ, que é superestimado. Para amostras maiores os resultados são melhoram e apresentam boas estimativas para estimação

Tabela 3.8: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo t-Student.* $\beta_0 = 110$, $\beta_1 = -4$, $\phi = 90$. ***Situação (4)***.

ν	n	Estimação	Média (desvio padrão)			
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\nu}$
		Método 1	42,211 (35,805)	0,938 (2,589)	766,135 (386,631)	22,545 (12,324)
	50	Método 2	63,565 (16,563)	-0,649 (1,142)	350,724 (170,943)	2,746 (3,660)
		Método 3	102,332 (8,939)	-3,428 (0,636)	150,568 (67,317)	2,111 (1,755)
2,5		Método 1	106,201 (5,334)	-3,721 (0,375)	116,674 (21,384)	2,000 (0,005)
	100	Método 2	93,127 (6,587)	-2,747 (0,466)	132,691 (24,773)	2,004 (0,023)
		Método 3	106,201 (5,334)	-3,721 (0,375)	116,674 (21,355)	2,000 (0,000)
		Método 1	109,269 (3,098)	-3,946 (0,223)	90,445 (13,238)	2,007 (0,043)
	200	Método 2	103,800 (3,696)	-3,534 (0,267)	93,257 (13,849)	2,010 (0,046)
		Método 3	109,270 (3,097)	-3,946 (0,223)	90,453 (13,287)	2,007 (0,047)
		Método 1	51,840 (27,162)	0,223 (1,966)	530,181 (239,724)	23,697 (11,598)
	50	Método 2	75,371 (10,538)	-1,482 (0,729)	190,195 (72,247)	2,284 (1,899)
		Método 3	103,119 (7,964)	-3,499 (0,541)	106,433 (33,050)	2,056 (1,242)
5		Método 1	106,482 (4,832)	-3,736 (0,338)	87,728 (12,759)	2,007 (0,027)
	100	Método 2	94,319 (5,848)	-2,832 (0,411)	99,656 (15,742)	2,040 (0,085)
		Método 3	106,484 (4,830)	-3,737 (0,338)	87,653 (12,695)	2,003 (0,030)
		Método 1	108,658 (3,230)	-3,898 (0,228)	72,251 (12,993)	2,273 (0,194)
	200	Método 2	103,721 (3,815)	-3,526 (0,270)	74,957 (13,779)	2,314 (0,204)
		Método 3	108,657 (3,230)	-3,898 (0,228)	72,280 (13,161)	2,276 (0,216)
		Método 1	53,292 (24,828)	0,116 (1,795)	476,967 (211,465)	24,257 (11,300)
	50	Método 2	77,123 (9,195)	-1,622 (0,624)	156,570 (48,416)	2,091 (0,369)
		Método 3	103,217 (6,453)	-3,511 (0,435)	91,829 (19,791)	2,000 (0,000)
10		Método 1	104,726 (3,826)	-3,603 (0,261)	70,478 (14,126)	2,023 (0,055)
	100	Método 2	92,952 (4,397)	-2,729 (0,294)	80,009 (17,018)	2,069 (0,104)
		Método 3	104,732 (3,824)	-3,604 (0,261)	70,302 (13,988)	2,014 (0,060)
		Método 1	109,185 (3,087)	-3,937 (0,225)	64,651 (9,473)	2,471 (0,179)
	200	Método 2	104,511 (3,631)	-3,583 (0,264)	66,981 (9,994)	2,517 (0,185)
		Método 3	109,184 (3,086)	-3,937 (0,225)	64,704 (9,428)	2,477 (0,191)

pelo Método 1 e pelo Método 3. Nesta situação, também observamos que os graus de liberdade foram subestimados, exceto para o Método 1 que apresentou, em média, valores para $\hat{\nu} > 20$.

Os resultados da estimação com erros t-Student considerando dados sem observações influentes, as estimativas nos três métodos são satisfatórias com valores próximos aos que foram assumidos para os parâmetros, exceto para o grau de liberdade cujas estimativas superestimam o parâmetro. Na presença de observações atípicas, os resultados entre os métodos são similares com boas estimativas para os parâmetros da regressão, mas com superestimação do parâmetro de escala quando $n = 50$ e $\nu = 2,5$ e subestimação dos graus de liberdade nos três métodos, para os três tamanhos de amostra considerados. Já para dados com observações influentes com média e alta alavancagem

as estimativas obtidas pelo Método 1 e pelo Método 2 são ruins quando amostra tem 50 observações. Nas duas situações o parâmetro de escala foi superestimado e só observamos resultados melhores para as amostras maiores. A estimação pelo método pelo Método 3 apresentou resultados melhores do que os outros métodos. De modo geral, verificamos que os graus de liberdade são subestimados quando as observações apresentam observações atípicas ou influentes com média e alta alavancagem.

3.2 Modelos não-lineares

Nesta parte, apresentamos um estudo de simulaçao considerando modelos não-lineares. Temos aqui os resultados de cada método de estimação no modelo de decaimento exponencial e no modelo de regressão assintótico . Assim como foi feito para o caso linear, no caso não-linear consideramos, também, 500 réplicas de cada modelo segundo o tamanho da amostra ($n = 50, 100, 200$), distribuição (t-student e logística-II) e a situação em que os dados foram gerados: sem observação influente, Situação (1), com observações atípicas na variável resposta, Situação (2), pontos atípicos com média alavancagem, Situação (3) e pontos atípicos alta com alta alavancagem, Situação (4).

3.2.1 Modelo de decaimento exponencial

O modelo de decaimento exponencial é dado por

$$y_i = \frac{\exp(-\beta_1 x_i)}{\beta_2 + \beta_3 x_i} + \epsilon_i. \tag{3.2.1}$$

Este modelo foi sugerido pelo NIST (*National Institute of Standards and Technology*) (1979) para o conjunto de dados `Chwirut2` no pacote `NISTnls` do `R`. Os valores para os parâmetros da regressão foram $\beta_1 = 0,1$, $\beta_2 = 0,005$ e $\beta_3 = 0,01$, que se baseiam nas estimativas encontradas para este conjunto. Os erros, $\epsilon_i \sim S(0, \phi, g)$, são independentes e identicamente distribuidos para $i = 1, \ldots, n$, com $\phi = 17$. A variável explanatória $\boldsymbol{x}$, foi gerada apartir da distribuição uniforme no intervalo $(0, 6)$ e seus valores foram fixados durante as simulações. Esse caso consideramos como sendo a ***Situação (1)***, dados sem observções influentes.

Nas demais situações foram inseridas observações atípicas nos casos $i = 20, 21, 22$. Para a ***Situação (2)*** temos $y_i^* = \mu_i + 5\sigma_{y_i}$, na ***Situação (3)*** temos $y_i^* = \mu_i + 5\sigma_{y_i}$ com $x_i^* = 11$ e na ***Situação (4)*** temos $y_i^* = \mu_i + 5\sigma_{y_i}$ com $x_i^* = 16$.

As distribuições assumidas para ϵ_i foram Logística-II e t-Student ($\nu = 2, 5; 5$ e 10) e isso foi feito para cada situação descrita anteriormente. Para obter as estimativas em cada réplica, ajustamos o modelo (3.2.1) de acordo com a ditribuição em que os dados foram gerados. A seguir apresentamos os resultados.

Resultados na estimação sob erros logístico tipo II

Nesta parte apresentamos os resultados do modelo (3.2.1) sob erros logísticos tipo II. Primeiro, ilustramos os dados gerados por uma determinada replicação. Em seguida mostramos o resumo das estimativas para esse modelo.

Na Figura 3.8 estão os gráficos dos dados para amostras de tamanho 50 de uma replicação específica. Observamos um bom ajuste na Situação (1) para qualquer uma das curvas obtidas nos métodos de estimação, uma leve alteração nos ajustes para Situação (2) e nas situações (3) e (4) observamos que alguns ajustes seguem uma configuração diferente da maioria dos dados. Na Figura 3.9 mostramos os gráficos dos dados com 100 observações. Verificamos que independente do processo de estimação o ajuste foi satisfatório, exceto na Situação (4) em que o ajuste por mínimos quadrados sofre influência dos pontos atípicos, mas estes não afetam os demais ajustes.

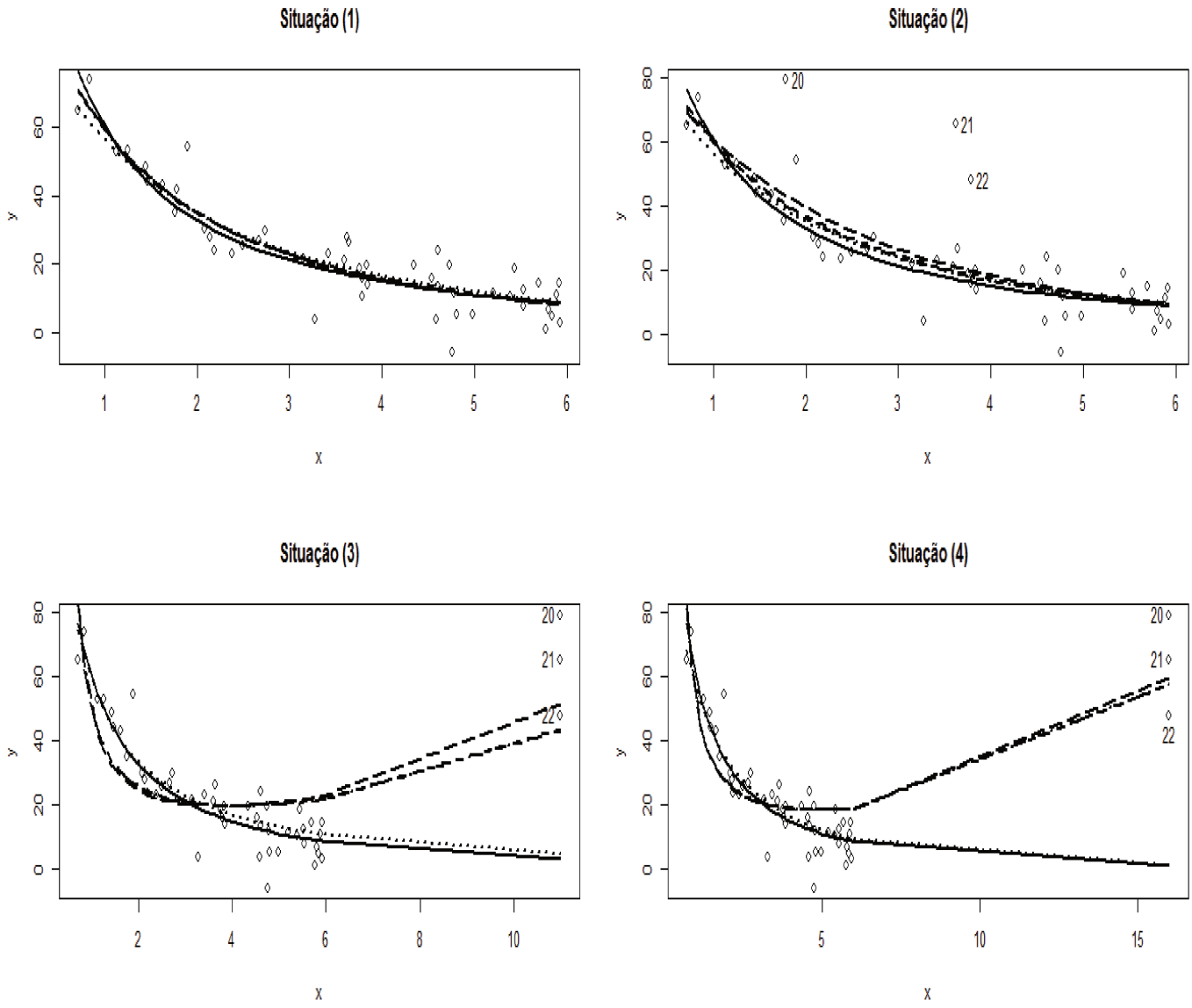


Figura 3.8: *Gráfico dos dados de uma simulação do modelo (3.2.1) com $n = 50$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada pelo Método 2.*

Para amostras com 200 observações os poucos pontos inseridos como observações discrepantes não causaram mudanças nas curvas ajustadas(ver Figura 3.10).

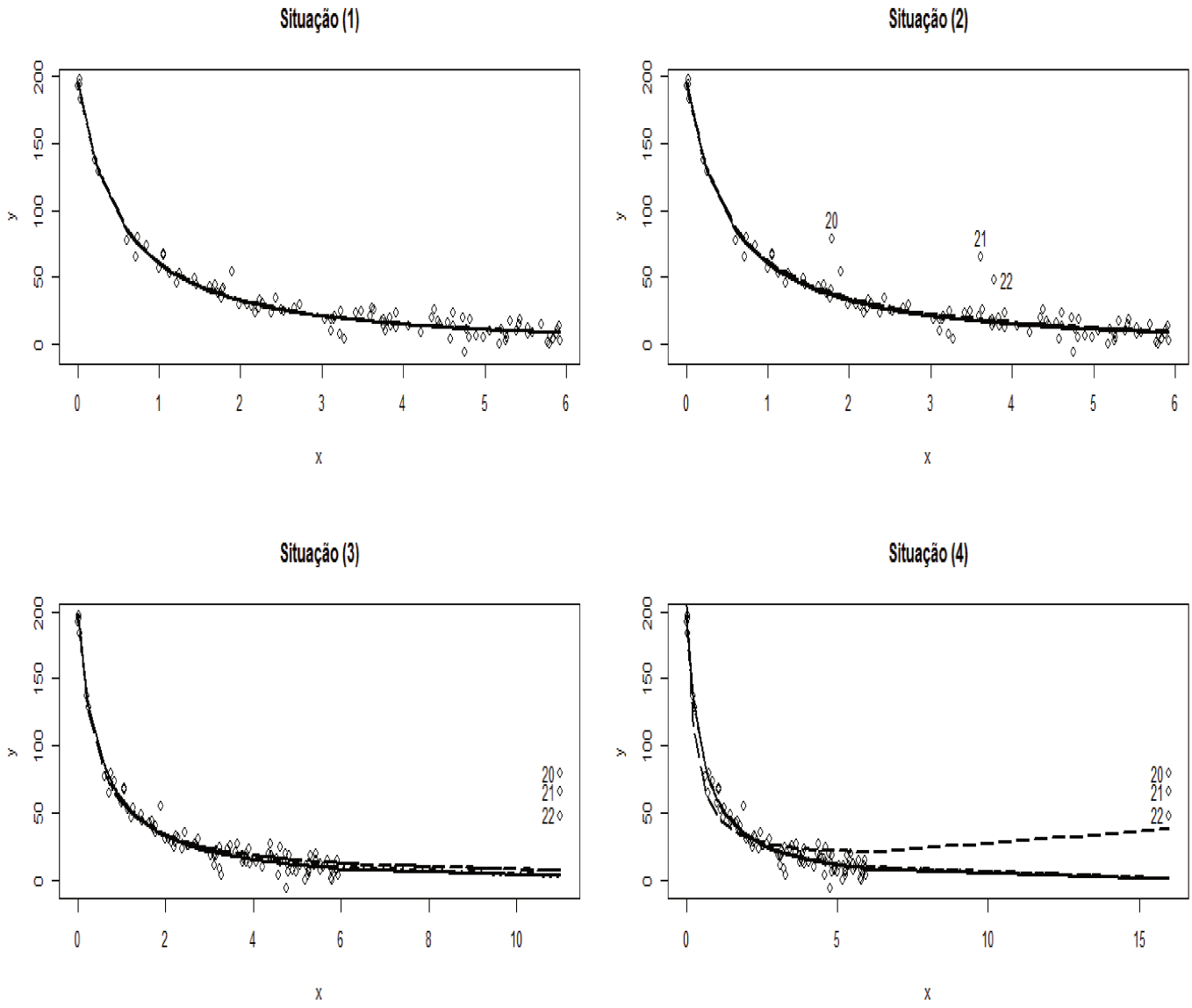


Figura 3.9: *Gráfico dos dados de uma simulação do modelo (3.2.1) com $n = 100$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada pelo Método 2.*

Na Tabela 3.9 temos média e desvio padrão das estimativas do ajuste aos dados sem observações influentes, Situação (1). Verificamos que foram obtidas boas estimativas para os parâmetros de regressão, com valores próximos aos verdadeiros e resultados similares nos três métodos. As estimativas média de ϕ e da variância foram melhores com a estimação pelo Método 3, sendo que nas amostras maiores o Método 2 também apresenta estimativas médias próximas aos valores verdadeiros.

Os resultados da estimação com dados sob a presença de *outliers* na variável reposta, Situação (2), estão na Tabela 3.10. A estimativa média de $\boldsymbol{\beta}$ foi satisfatória, para todos os tamanhos de

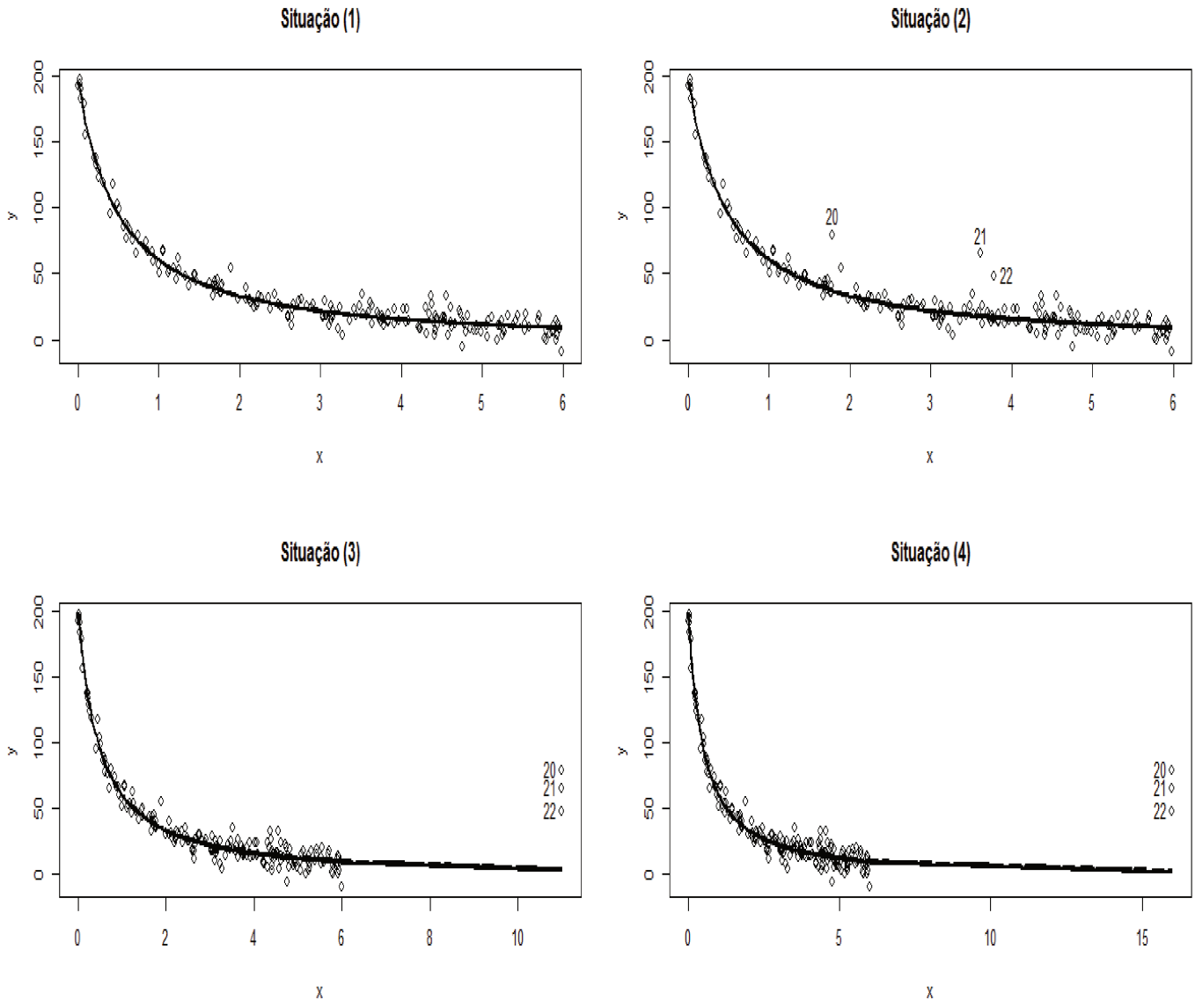


Figura 3.10: *Gráfico dos dados de uma simulação do modelo (3.2.1) com $n = 200$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada pelo Método 2.*

amostra considerados. Mas, em média, o parâmetro ϕ e a variância foram superestimados em amostras de tamanho 50, pricipalmente na estimação com os métodos 1 e 2. Nas amostras maiores as estimatiivas do parâmetro de escala estão mais próximas dos valores verdadeiros para estimação com os métodos 2 e 3, mas com Método 1 ainda há superestimação do parâmetro. Em geral, as estimativas com o Método 1 são melhores.

Para Situação (3) observamos que em $n = 50$ as estimativas são ruins (Tabela 3.11) para todos os parâmetros. Para os demais tamanhos de amostra vemos que na estimação pelo Método 3 foram obtidos resultados melhores dos que os outros métodos de estimação. Na Situação (4), as

Tabela 3.9: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros logísticos tipo II .* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$, $\phi = 17$ *e* $\sigma^2 = 56$. ***Situação (1)***

n	Estimação	Média (desvio padrão)				
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\sigma^2}$
50	Método 1	0,095 (0,07)	0,004 (2,5×10^{-3})	0,011 (3,5×10^{-3})	42,71 (18,20)	140,51 (59,87)
	Método 2	0,102 (0,08)	0,005 (2,8×10^{-3})	0,011 (3,8×10^{-3})	21,34 (15,66)	70,20 (51,53)
	Método 3	0,099 (0,07)	0,005 (2,5×10^{-3})	0,011 (3,4×10^{-3})	15,71 (4,07)	51,68 (13,38)
100	Método 1	0,099 (0,03)	0,005 (1,0×10^{-4})	0,010 (7,7×10^{-4})	50,98 (12,19)	167,70 (40,10)
	Método 2	0,100 (0,03)	0,005 (1,2×10^{-4})	0,010 (8,6×10^{-4})	16,47 (2,82)	54,14 (9,28)
	Método 3	0,099 (0,03)	0,005 (1,0×10^{-4})	0,010 (7,4×10^{-4})	16,35 (2,80)	53,78 (9,22)
200	Método 1	0,101 (0,02)	0,005 (9,5×10^{-5})	0,010 (5,3×10^{-4})	51,40 (12,21)	169,09 (40,16)
	Método 2	0,100 (0,02)	0,005 (9,7×10^{-5})	0,010 (5,7×10^{-4})	16,68 (2,16)	54,86 (7,11)
	Método 3	0,100 (0,02)	0,005 (8,5×10^{-5})	0,010 (5,0×10^{-4})	16,63 (2,15)	54,69 (7,08)

Tabela 3.10: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros logísticos tipo II.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$, $\phi = 17$ *e* $\sigma^2 = 56$. ***Situação (2)***

n	Estimação	Média (desvio padrão)				
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\sigma^2}$
50	Método 1	0,130 (0,12)	0,006 (2,81×10^{-3})	0,010 (4,25×10^{-3})	86,80 (46,00)	261,14 (126,02)
	Método 2	0,107 (0,14)	0,005 (2,76×10^{-3})	0,010 (4,08×10^{-3})	134,71 (229,30)	442,66 (1013,96)
	Método 3	0,102 (0,07)	0,005 (2,33×10^{-3})	0,010 (3,19×10^{-3})	29,53 (5,84)	97,14 (19,22)
100	Método 1	0,087 (0,03)	0,005 (1,2×10^{-4})	0,010 (7,5×10^{-4})	85,87 (23,25)	274,60 (73,96)
	Método 2	0,096 (0,03)	0,005 (1,2×10^{-4})	0,010 (8,6×10^{-4})	51,91 (24,56)	169,59 (66,18)
	Método 3	0,092 (0,03)	0,005 (1,0×10^{-4})	0,010 (7,3×10^{-4})	22,55 (3,41)	74,36 (10,76)
200	Método 1	0,088 (0,02)	0,005 (8,8×10^{-5})	0,010 (5,1×10^{-4})	74,32 (9,76)	242,75 (28,70)
	Método 2	0,097 (0,02)	0,005 (9,6×10^{-5})	0,010 (5,6×10^{-4})	19,66 (2,36)	64,75 (7,69)
	Método 3	0,096 (0,02)	0,005 (8,6×10^{-5})	0,010 (5,0×10^{-4})	19,60 (2,35)	64,48 (7,72)

estimativas também são ruins, em todos os métodos, para amostras com 50 observações. Nas amostras maiores as estimativas melhoram, principalmente, na estimação pelo Método 3.

Nesta parte, verificamos que inserção de observações atípicas pouco influenciaram as estimativas do parâmetros de regressão, mas o parâmetro de escla foi superestimado em amostras com 50 observações. Na presença de observações influentes com média e alta alavancagem as estimativas de todos os parâmetros foram bastante influenciadas, quando $n = 50$. Em amostras maiores os resultados foram melhores com estimativas mais próximas dos verdadeiros valores dos parâmetros. Verificamos que as estimativas obtidas com o Método 3 foram melhores do que a dos outros métodos.

Tabela 3.11: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros logísticos tipo II.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$, $\phi = 17$ *e* $\sigma^2 = 56$. ***Situação (3)***

n	Estimação	Média (desvio padrão)				
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\sigma}^2$
50	Método 1	-0,256 (0,09)	-0,017 ($6,9\times10^{-3}$)	0,043 ($1,03\times10^{-2}$)	52,00 (26,59)	167,57 (79,19)
	Método 2	0,0239 (0,12)	0,001 ($6,0\times10^{-3}$)	0,016 ($8,96\times10^{-3}$)	225,86 (47,13)	738,41 (128,57)
	Método 3	-0,258 (0,09)	-0,017 ($6,7\times10^{-3}$)	0,044 ($9,98\times10^{-3}$)	44,50 (8,30)	149,03 (26,28)
100	Método 1	0,043 (0,06)	0,005 ($1,5\times10^{-4}$)	0,012 ($1,6\times10^{-3}$)	144,62 (22,92)	477,82 (69,16)
	Método 2	0,091 (0,03)	0,005 ($1,2\times10^{-4}$)	0,010 ($9,2\times10^{-4}$)	130,07 (21,62)	426,30 (42,58)
	Método 3	0,084 (0,03)	0,005 ($1,0\times10^{-4}$)	0,010 ($8,2\times10^{-4}$)	26,95 (3,75)	88,87 (11,79)
200	Método 1	0,069 (0,02)	0,005 ($9,2\times10^{-5}$)	0,012 ($6,1\times10^{-4}$)	102,11 (16,54)	337,92 (40,10)
	Método 2	0,097 (0,02)	0,005 ($9,7\times10^{-5}$)	0,010 ($5,8\times10^{-4}$)	24,67 (11,35)	75,56 (25,12)
	Método 3	0,094 (0,02)	0,005 ($8,6\times10^{-5}$)	0,010 ($5,2\times10^{-4}$)	21,57 (2,47)	71,04 (8,05)

Tabela 3.12: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros logísticos tipo II.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$, $\phi = 17$ *e* $\sigma^2 = 56$. ***Situação (4)***

n	Estimação	Média (desvio padrão)				
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\sigma}^2$
50	Método 1	-0,213 (0,01)	-0,0122 ($2,5\times10^{-3}$)	0,036 ($3,2\times10^{-3}$)	35,85 (16,98)	114,72 (55,48)
	Método 2	-0,025 (0,19)	-0,0027 ($8,9\times10^{-3}$)	0,021 ($1,3\times10^{-2}$)	198,84 (347,16)	584,52 (301,70)
	Método 3	-0,213 (0,01)	-0,0122 ($2,5\times10^{-3}$)	0,036 ($3,2\times10^{-3}$)	32,23 (6,81)	104,05 (21,25)
100	Método 1	0,050 (0,08)	0,005 ($1,54\times10^{-4}$)	0,012 ($2,7\times10^{-3}$)	143,87 (25,28)	477,94 (65,56)
	Método 2	0,094 (0,03)	0,005 ($1,17\times10^{-4}$)	0,010 ($9,5\times10^{-4}$)	138,25 (21,65)	449,36 (42,04)
	Método 3	0,088 (0,03)	0,005 ($1,02\times10^{-4}$)	0,010 ($9,6\times10^{-4}$)	27,39 (3,77)	90,15 (11,81)
200	Método 1	0,078 (0,03)	0,005 ($1,0\times10^{-4}$)	0,011 ($6,8\times10^{-4}$)	101,68 (23,54)	335,72 (68,50)
	Método 2	0,098 (0,02)	0,005 ($9,7\times10^{-5}$)	0,010 ($5,8\times10^{-4}$)	26,81 (13,97)	79,00 (32,44)
	Método 3	0,096 (0,02)	0,005 ($8,6\times10^{-5}$)	0,010 ($5,2\times10^{-4}$)	21,74 (2,48)	71,60 (8,08)

Resultados da estimação sob erros t-Student

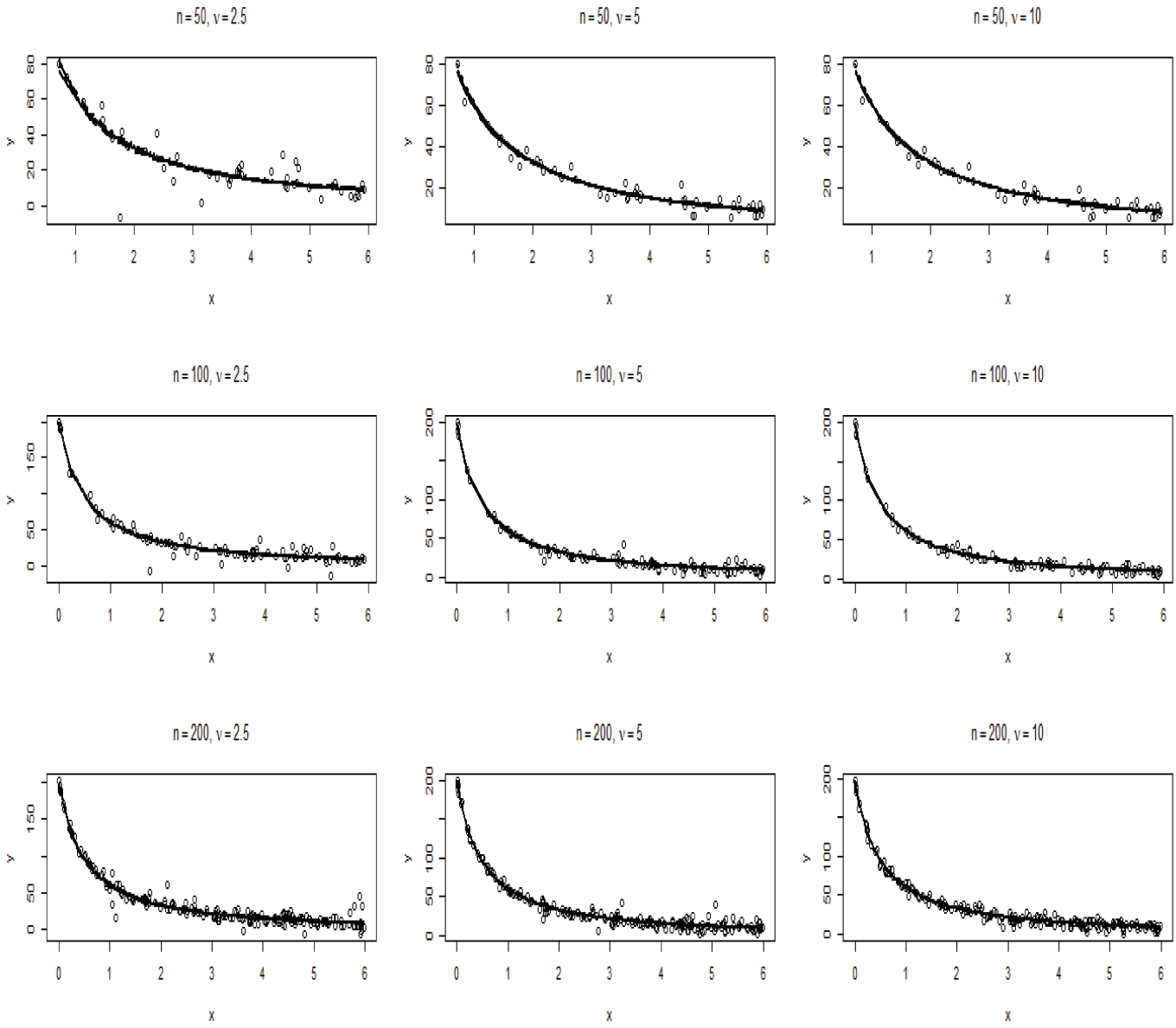


Figura 3.11: *Gráfico dos dados de uma simulação do modelo (3.2.1) sob erros t-Student da* ***Situação (1)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada por MV com ν fixo e (· · · · · · ·) curva ajustada pelo Método 2.

A Figura 3.11 mostra os gráficos dos dados gerados para Situação (1) considerando o modelo (3.2.1) sob erros t-Student para exemplos com diferentes tamanhos de amostra e graus de liberdade. Observamos que para, esses exemplos, qualquer um dos ajustes considerados adequam aos dados.

A Figura 3.12 mostra os gáficos dos dados gerados para Situação (2) segundo tamanho da amostra e valor dos graus de liberdade. Verificamos que o ajuste da curva em qualquer um dos

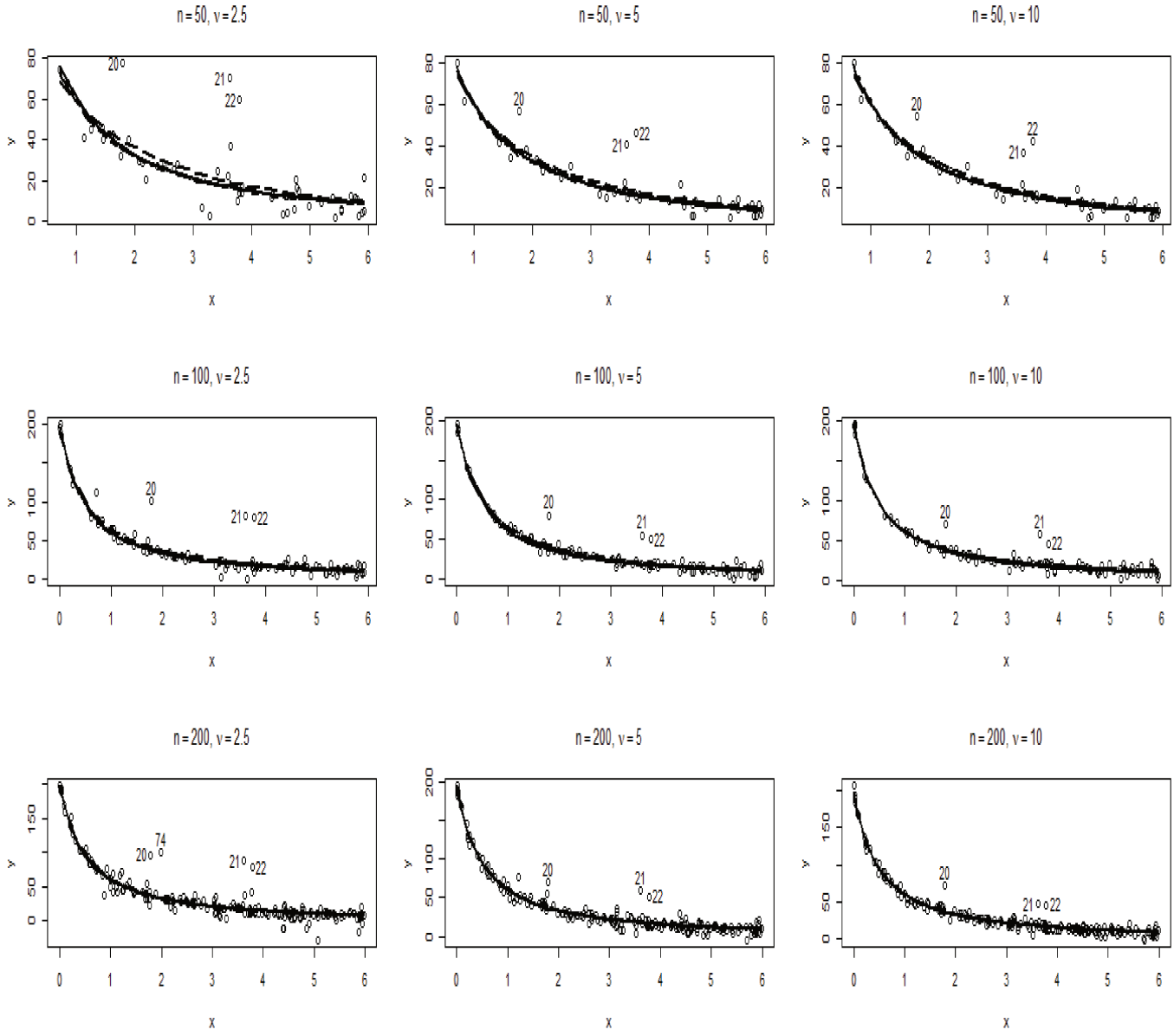


Figura 3.12: *Gráfico dos dados de uma simulação do modelo (3.2.1) sob erros t-Student da* ***Situação (2)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada por MV com ν fixo e (· · · · · · ·) curva ajustada pelo Método 2.

métodos é satisfatório para estes exemplos.

Na Figura 3.13 observamos que para $n = 50$ ajuste foi melhor quando usada a estimação Método 2. Com $n = 100$ ou 200 vemos que qualquer um dos métodos fornecem um bom ajuste.

Como ilustração dos dados que foram gerados para Situação (4), na Figura 3.14 os gráficos dos dados segundo tamanho de amostra e valor do grau de liberdade. Em amostras de tamanho 50 os

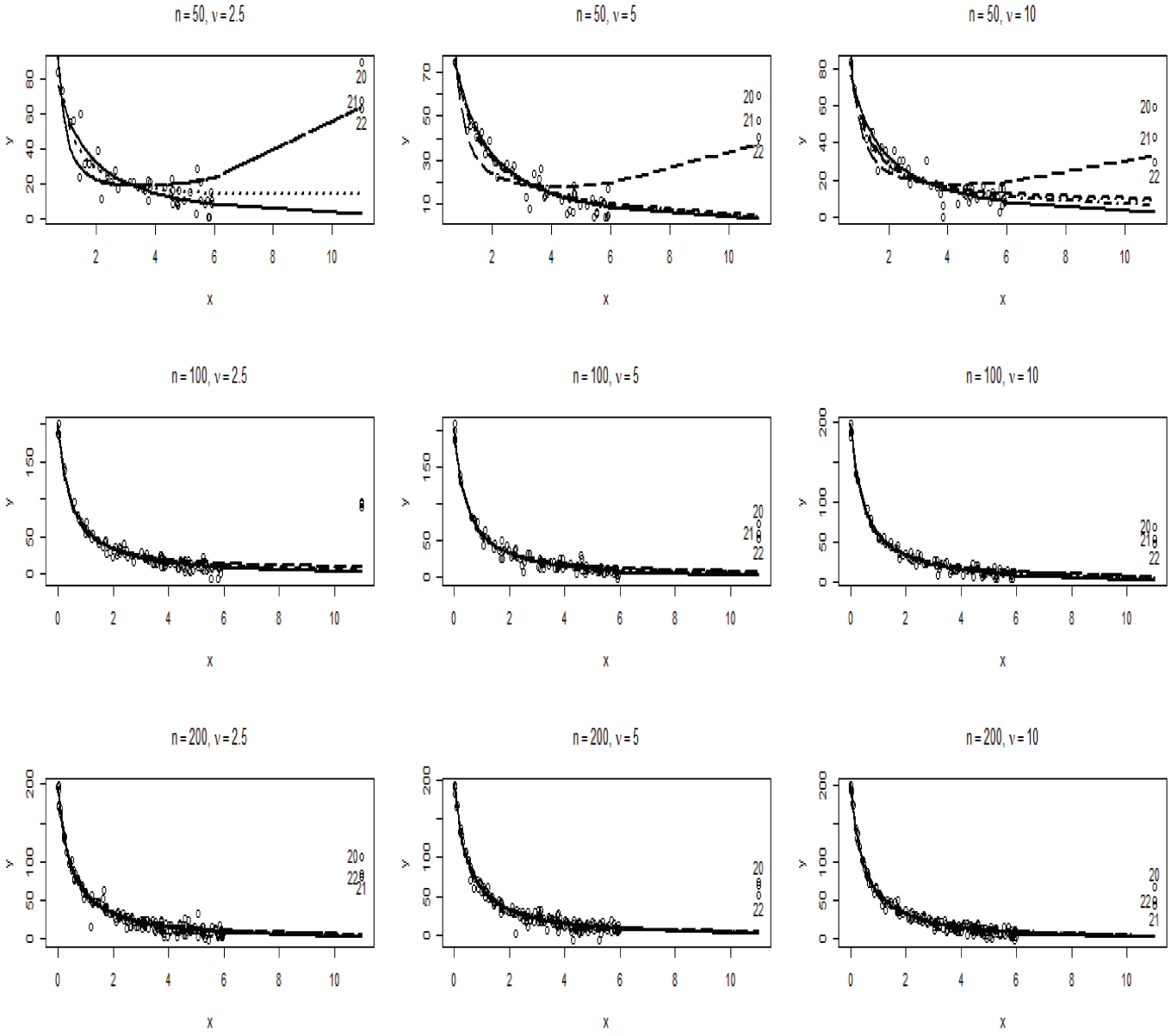


Figura 3.13: *Gráfico dos dados de uma simulação do modelo (3.2.1) sob erros t-Student da* ***Situação (3)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (− − −) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (− · − · −) curva ajustada por MV com ν fixo e (· · · · · · ·) curva ajustada pelo Método 2.

ajustes apresentam efeito dos *outliers* com alta alavancagem. Em amostras de tamanho 100 vemos que apenas o ajuste mínimos quadrados é afetado pelos pontos atípicos quando $\nu = 2,5;5$, com $\nu = 10$ os ajustes foram adequados, assim como, $n = 200$.

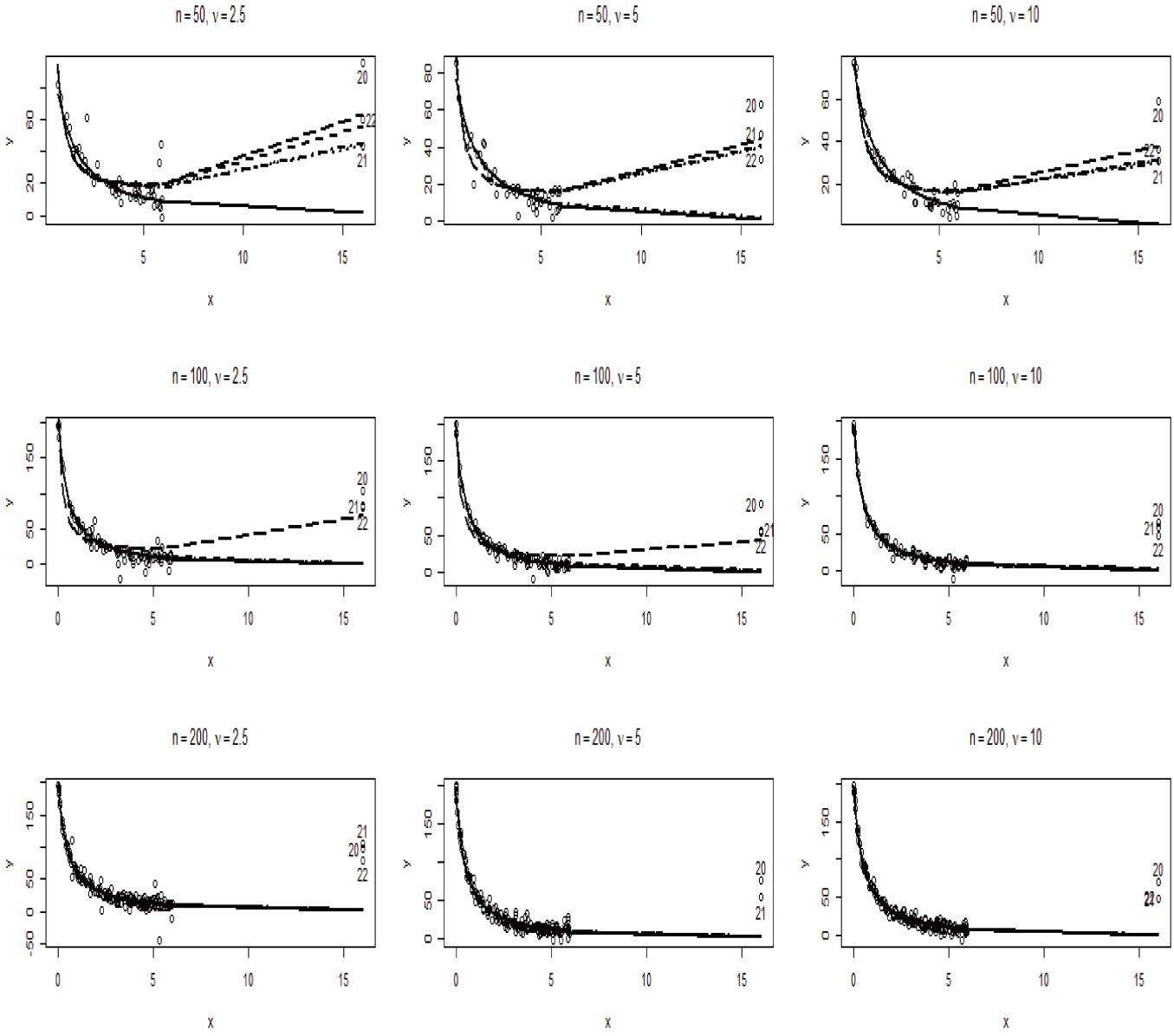


Figura 3.14: *Gráfico dos dados de uma simulação do modelo (3.2.1) sob erros t-Student da* ***Situação (4)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada por MV com ν fixo e (· · · · · · ·) curva ajustada pelo Método 2.

A Tabela 3.13 mostra a média e o desvio padrão das estimativas considerando os dados da Situação (1). No geral, as estimativas obtidas para $\hat{\boldsymbol{\beta}}$ foram boas, com valores médios próximos aos verdadeiros parâmetros em qualquer um dos métodos de estimação. A estimativa média do parâmetro de escala no Método 1 aumenta com a amostra, enquanto que nos métodos 2 e 3 a estimativa média se aproxima do verdadeiro parâmetro quando a amostra aumenta. Para as

Tabela 3.13: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros t-student.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$ *e* $\phi = 17$. ***Situação (1)***.

ν (σ^2)	n	Estimação	Média (desvio padrão)					
			$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\nu}^a$	$\hat{\sigma}^2$
		Método 1	0,09 (0,06)	0,004 (0,0023)	0,01 (0,003)	47,5 (70,6)	10,5; 7,5 (7,6)	58,9 (236,8)
	50	Método 2	0,10 (0,06)	0,005 (0,0024)	0,01 (0,003)	18,3 (7,0)	6,2; 3,5 (7,4)	79,4 (145,3)
		Método 3	0,10 (0,05)	0,005 (0,0022)	0,01 (0,003)	23,4 (7,9)	7,7; 5,6 (6,9)	40,4 (20,0)
2,5 (85)		Método 1	0,10 (0,03)	0,005 (0,0001)	0,01 (0,001)	54,5 (29,5)	5,1; 3,9 (4,1)	152,0 (182,2)
	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	17,7 (5,0)	3,7; 3,0 (3,6)	127,6 (306,1)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	17,8 (5,1)	4,1; 3,1 (4,1)	63,8 (34,3)
		Método 1	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	57,6 (22,9)	4,2; 3,1 (2,7)	199,5 (171,6)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,01 (0,001)	17,2 (3,3)	2,9; 2,7 (0,7)	131,4 (319,4)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,01 (0,001)	17,2 (3,3)	3,0; 2,8 (0,8)	74,0 (35,0)
		Método 1	0,01 (0,05)	0,005 (0,0018)	0,01 (0,003)	24,4 (7,8)	14,0; 9,2 (9,9)	34,7 (31,4)
	50	Método 2	0,10 (0,06)	0,005 (0,0021)	0,01 (0,003)	16,5 (5,4)	10,8; 5,6 (10,0)	35,6 (57,0)
		Método 3	0,10 (0,05)	0,005 (0,0018)	0,01 (0,003)	18,1 (5,0)	13,3; 7,2 (10,4)	25,3 (8,7)
			23,6					
5 (28,3)		Método 1	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	26,7 (6,1)	9,3; 5,7 (8,5)	50,8 (44,8)
	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	16,9 (4,3)	8,4; 5,2 (7,8)	29,6 (9,7)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	17,0 (4,3)	9,2; 5,6 (8,6)	29,1 (9,8)
		Método 1	0,10 (0,01)	0,005 (0,0001)	0,01 (0,001)	27,5 (4,7)	6,2; 5,3 (3,0)	48,3 (21,3)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,01 (0,001)	16,8 (3,0)	5,8; 5,0 (2,7)	28,9 (5,8)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,01 (0,001)	16,9 (3,1)	6,0; 5,0 (2,3)	28,6 (5,9)
		Método 1	0,10 (0,05)	0,005 (0,0018)	0,01 (0,003)	19,2 (4,9)	19,3; 23,6 (10,7)	23,7 (8,1)
	50	Método 2	0,10 (0,06)	0,005 (0,0021)	0,01 (0,003)	15,6 (4,9)	16,3; 12,4 (11,5)	24,8 (55,3)
		Método 3	0,10 (0,05)	0,005 (0,0018)	0,01 (0,003)	16,2 (4,4)	19,0; 23,6 (11,0)	20,0 (5,0)
10 (21,3)		Método 1	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	20,6 (3,8)	16,9; 13,3 (10,9)	27,8 (10,8)
	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	16,3 (3,7)	15,6; 10,6 (10,9)	21,6 (5,2)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,01 (0,001)	16,4 (3,6)	16,8; 13,3 (10,9)	21,2 (5,0)
		Método 1	0,10 (0,01)	0,005 (0,0001)	0,01 (0,001)	28,7 (3,3)	14,3; 11,7 (7,7)	36,7 (8,1)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,01 (0,001)	15,3 (2,4)	13,2; 9,8 (7,5)	19,6 (2,5)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,01 (0,001)	15,5 (2,3)	13,8; 11,3(7,4)	19,6 (2,5)

[a]Média; mediana (desvio padrão) ν.

estimativas médias dos graus de liberdade, observamos que o Método 2 apresenta valores mais próximos aos valores dos parâmetros, principalmente os valores medianos das estimativas.

Ainda na Tabela 3.13, quando$\nu = 2,5$, a média das estimativas da variância foi melhor no Método 2. Para as amostras maiores a estimativa média da variância nos métodos 1 e 2 aumenta com amostra, enquanto que no Método 3 se aproxima da verdadeira variância.

Quando $\nu = 5, 10$, na Tabela 3.13, a estimativa média da variância está mais próxima da variância real no Método 3 qualquer que seja o tamanho da amostra, enquanto que o Método 2

Tabela 3.14: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros t-Student.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$ *e* $\phi = 17$. ***Situação (2)***

ν	n	Estimação	Média (desvio padrão)				
			$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\nu}$
		Método 1	0,11 (0,07)	0,005 (0,002)	0,010 (0,003)	169,05 (39,58)	3,01 (0,11)
	50	Método 2	0,10 (0,05)	0,005 (0,002)	0,010 (0,003)	22,32 (6,26)	2,00 (0,01)
		Método 3	0,08 (0,06)	0,003 (0,008)	0,012 (0,007)	39,37 (15,66)	4,00 (1,1)
2,5		Método 1	0,08 (0,03)	0,005 (0,0001)	0,010 (0,001)	121,62 (40,55)	3,00 (0,00)
	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	17,68 (4,01)	2,02 (0,07)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	17,57 (3,95)	2,02 (0,10)
		Método 1	0,09 (0,02)	0,005 (0,0001)	0,010 (0,001)	95,73 (28,99)	3,00 (0,00)
	200	Método 2	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	16,66 (3,01)	2,12 (0,19)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	16,56 (3,10)	2,11 (0,24)
		Método 1	0,10 (0,05)	0,005 (0,0017)	0,010 (0,002)	65,43 (8,97)	3,01 (0,22)
	50	Método 2	0,10 (0,05)	0,005 (0,0020)	0,010 (0,003)	15,50 (4,83)	2,10 (0,23)
		Método 3	0,10 (0,04)	0,005 (0,0018)	0,010 (0,002)	27,00 (7,39)	5,27 (1,40)
5		Método 1	0,08 (0,02)	0,005 (0,0001)	0,010 (0,001)	47,38 (7,82)	2,67 (0,54)
	100	Método 2	0,09 (0,02)	0,005 (0,0001)	0,010 (0,001)	15,10 (4,06)	2,59 (0,45)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	15,11 (4,16)	2,62 (0,51)
		Método 1	0,09 (0,01)	0,005 (0,0001)	0,010 (0,0003)	37,22 (4,96)	3,37 (0,61)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,010 (0,0004)	15,32 (2,73)	3,26 (0,53)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,0003)	15,37 (2,77)	3,31 (0,57)
		Método 1	0,10 (0,04)	0,005 (0,0016)	0,010 (0,002)	49,42 (5,65)	3,01 (0,14)
	50	Método 2	0,09 (0,05)	0,005 (0,0019)	0,010 (0,003)	14,41 (4,85)	2,31 (0,41)
		Método 3	0,09 (0,04)	0,005 (0,0017)	0,010 (0,002)	22,62 (5,35)	5,37 (1,36)
10		Método 1	0,09 (0,02)	0,005 (0,0001)	0,010 (0,001)	35,38 (3,68)	3,29 (0,82)
	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	14,03 (3,34)	3,08 (0,59)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	14,10 (3,35)	3,15 (0,61)
		Método 1	0,09 (0,01)	0,005 (0,0001)	0,010 (0,001)	28,42 (2,75)	4,25 (0,89)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	14,56 (2,32)	4,09 (0,67)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	14,58 (2,32)	4,14 (0,70)

tem valores próximos para as amostras maiores.

Na Tabela 3.14 apresenta um resumo do ajuste com os dados da Situação (2). As estimativas dos parâmetros de regressão obtidas com os três métodos apresentam bons resultados. Entretanto, o parâmetro de escala foi superestimado no Método 1. Para amostras com 100 e 200 observações, em $\nu = 5, 10$, O grau de liberdade foi subestimado nos três métodos e o parâmetro de escala subestimado Método 2 e pelo Método 3.

A média e desvio padrão das estimativas para os dados Situação (3) estão na Tabela 3.15. Verificamos que, em média, as estimativas de todos os parâmetros foram ruins para amostras de tamanho 50. Neste caso, os resultados foram melhores para a estimação Método 2. Para os outros

Tabela 3.15: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros t-Student.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$ *e* $\phi = 17$. ***Situação (3)***

ν	n	Estimação	Média (desvio padrão) $\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\nu}$
		Método 1	-0,23 (0,15)	-0,017 (0,011)	0,043 (0,017)	157,95 (49,65)	17,89 (12,34)
	50	Método 2	0,05 (0,09)	0,003 (0,004)	0,013 (0,006)	23,69 (14,72)	2,36 (2,97)
		Método 3	-0,14 (0,21)	-0,014 (0,058)	0,038 (0,081)	88,76 (78,84)	14,71 (12,78)
2,5		Método 1	0,03 (0,05)	0,005 (0,0001)	0,012 (0,002)	189,16 (32,48)	3,00 (0,01)
	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	18,27 (3,67)	2,00 (0,01)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	18,18 (3,65)	2,00 (0,02)
		Método 1	0,07 (0,03)	0,005 (0,0001)	0,011 (0,001)	131,52 (25,69)	3,00 (0,01)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	15,83 (2,50)	2,04 (0,10)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	15,77 (2,56)	2,04 (0,14)
		Método 1	-0,01 (0,06)	0,002 (0,0034)	0,016 (0,005)	95,68 (17,57)	3,22 (1,88)
	50	Método 2	0,04 (0,05)	0,003 (0,0023)	0,013 (0,003)	17,30 (4,36)	2,00 (0,01)
		Método 3	0,06 (0,04)	0,004 (0,0019)	0,012 (0,002)	27,13 (7,18)	4,03 (0,66)
5		Método 1	0,04 (0,02)	0,005 (0,0001)	0,012 (0,001)	91,97 (6,55)	3,00 (0,01)
	100	Método 2	0,09 (0,02)	0,005 (0,0001)	0,010 (0,001)	12,81 (2,84)	2,10 (0,13)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	12,75 (2,95)	2,08 (0,20)
		Método 1	0,08 (0,02)	0,005 (0,0001)	0,010 (0,001)	63,37 (4,71)	3,00 (0,01)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	13,96 (2,33)	2,61 (0,30)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	14,03 (2,38)	2,64 (0,35)
		Método 1	0,01 (0,03)	0,002 (0,0018)	0,015 (0,002)	82,22 (9,78)	3,07 (0,19)
	50	Método 2	0,03 (0,06)	0,003 (0,0024)	0,014 (0,003)	14,46 (3,53)	2,00 (0,04)
		Método 3	0,06 (0,04)	0,004 (0,0019)	0,012 (0,002)	22,23 (4,95)	4,12 (0,62)
10		Método 1	0,04 (0,02)	0,005 (0,0001)	0,011 (0,001)	77,56 (3,52)	3,00 (0,01)
	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	12,28 (2,36)	2,22 (0,21)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	12,34 (2,46)	2,26 (0,30)
		Método 1	0,08 (0,01)	0,005 (0,0001)	0,010 (0,001)	51,93 (2,70)	3,00 (0,00)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	12,62 (1,93)	2,90 (0,29)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	12,65 (1,94)	2,92 (0,33)

tamanho de amostra vemos que melhores resultados foram obtidos com a estimação Método 2 e pelo Método 3. Nas amostras maiores o grau de liberdade foram subestimados em $\nu = 5, 10$. Para a Situação (4) as estimativasde todos os parâmetros, também, foram ruins em amostras com 50 observações (Tabela 3.16), em que o parâmetro de escala foi superestimado. Obtem-se estimativas melhores em amostras maiores com a estimação Método 2 e pelo Método 3, entretanto verifcamos uma subestimação de do parâmetro de escala e do grau de liberdade.

O resultado da estimação com erros t-Student no modelo (3.2.1) mostrou que as estimativas dos parâmetros de regressão são pouco influenciadas por *outliers* na variável resposta, mas são fortemente influenciadas por observações atípicas com média e alta alavancagem em amostras

Tabela 3.16: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.1) sob erros t-Student.* $\beta_1 = 0,1$, $\beta_2 = 0,005$, $\beta_3 = 0,01$ *e* $\phi = 17$. ***Situação (4).***

ν	n	Estimação	Média (desvio padrão) $\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\nu}$
		Método 1	-0,23 (0,01)	-0,014 (0,0023)	0,038 (0,003)	108,73 (48,89)	11,22 (6,63)
	50	Método 2	-0,15 (0,14)	-0,010 (0,0086)	0,032 (0,012)	51,75 (21,92)	7,47 (8,55)
		Método 3	-0,23 (0,06)	-0,020 (0,0914)	0,046 (0,128)	86,27 (50,99)	10,61 (8,73)
		Método 1	0,02 (0,11)	0,005 (0,0002)	0,013 (0,005)	185,61 (34,54)	2,00 (0,02)
2,5	100	Método 2	0,10 (0,03)	0,005 (0,0001)	0,010 (0,001)	17,89 (3,55)	2,00 (0,01)
		Método 3	0,11 (0,02)	0,005 (0,0001)	0,010 (0,001)	17,84 (3,54)	2,00 (0,02)
		Método 1	0,07 (0,02)	0,005 (0,0001)	0,011 (0,001)	132,17 (29,65)	2,04 (0,11)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	15,77 (2,41)	2,04 (0,10)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	15,74 (2,42)	2,04 (0,11)
		Método 1	-0,19 (0,02)	-0,011 (0,0027)	0,033 (0,002)	68,94 (14,55)	13,51 (10,70)
	50	Método 2	-0,08 (0,28)	-0,005 (0,0078)	0,024 (0,010)	32,26 (39,46)	6,69 (8,42)
		Método 3	-0,19 (0,03)	-0,010 (0,0036)	0,032 (0,004)	50,30 (14,58)	13,07 (10,83)
		Método 1	0,05 (0,04)	0,005 (0,0001)	0,011 (0,002)	97,78 (10,67)	3.00 (0,00)
5	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	13,19 (2,70)	2,08 (0,13)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	13,25 (2,73)	2,11 (0,17)
		Método 1	0,08 (0,02)	0,005 (0,0001)	0,010 (0,001)	62,93 (6,54)	2,64 (0,32)
	200	Método 2	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	13,85 (2,30)	2,61 (0,30)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	13,89 (2,32)	2,64 (0,35)
		Método 1	-0,18 (0,01)	-0,010 (0,0026)	0,032 (0,003)	57,81 (9,78)	12,70 (10,51)
	50	Método 2	-0,08 (0,10)	-0,004 (0,0074)	0,024 (0,010)	24,30 (11,25)	5,12 (4,64)
		Método 3	-0,19 (0,03)	-0,010 (0,0025)	0,030 (0,004)	42,22 (12,25)	12,27 (9,58)
		Método 1	0,06 (0,02)	0,005 (0,0001)	0,011 (0,001)	81,86 (8,33)	2,24 (0,24)
10	100	Método 2	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	12,28 (2,56)	2,21 (0,22)
		Método 3	0,10 (0,02)	0,005 (0,0001)	0,010 (0,001)	12,35 (2,59)	2,25 (0,28)
		Método 1	0,08 (0,01)	0,005 (0,0001)	0,010 (0,001)	52,17 (4,58)	2,97 (0,30)
	200	Método 2	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	12,95 (1,84)	2,94 (0,29)
		Método 3	0,10 (0,01)	0,005 (0,0001)	0,010 (0,001)	12,99 (1,83)	2,92 (0,33)

de tamanho 50. O parâmetro de escala foi superestimado na presença de *outliers* e observações influentes, enquanto que o grau de liberdade foi subestimado. Os melhores resultados foram observados para o Método 2 e pelo Método 3.

3.2.2 Modelo de regressão assintótica

O modelo de regressão assintótica é dado por

$$y_i = \alpha_1 + (\alpha_2 - \alpha_1)\exp[-\exp(\alpha_3)x_i] + \epsilon_i. \tag{3.2.2}$$

Na função de regressão α_1 é o parâmetro que representa a assíntota horizontal do lado direito (valores muito grandes de x_i), α_2 é o valor da função quando $x_i = 0$ e α_3 é o logaritmo natural da constante de velocidade. O valor dado a cada parâmetro da regressão foi $\alpha_1 = 100$, $\alpha_2 = -8.5$ e $\alpha_3 = -3.2$. Os erros, $\epsilon_i \sim S(0, \phi, g)$, são independentes e identicamente distribuidos para $i = 1, \ldots, n$, com $\phi = 20$. A variável explicativa $\boldsymbol{x}$, foi gerada apartir da distibuição uniforme no intervalo $(0, 150)$ e seus valores foram mantidos fixos durante as simulações. Os dados gerados a partir desse modelo constituem os dados da ***Situação (1)***, ou seja, sem observações influentes.

Nas demais situações foram inseridas observações atípicas nos casos em que $i = 2, 26, 22$. Para a ***Situação (2)*** temos que $y_i^* = \mu_i - 5\sigma_{y_i}$, dados com *outliers* na variável resposta. Na ***Situação (3)***, temos dados com *outliers* com média alavacangem, $y_i^* = \mu_i + 5\sigma_{y_i}$ com $x_i^* = 250$. E na ***Situação (4)***, temos dados com *outliers* com alta alavacangem, $y_i^* = \mu_i + 5\sigma_{y_i}$ com $x_i^* = 350$.

Para distribuição dos erros também consideramos a distribuição logística-II, t-Student ($\nu = 2, 5; 5$ e 10) e exponencial potência ($\nu = -0, 5; 0$ e $0, 5$). As estimativas obtidas em cada réplica resultam do ajuste do modelo (3.2.2) de acordo com a ditribuição que foi assumida para gerar os dados. Nos tópicos seguintes apresentamos os resultados.

Resultados da estimação sob erros logístico tipo II

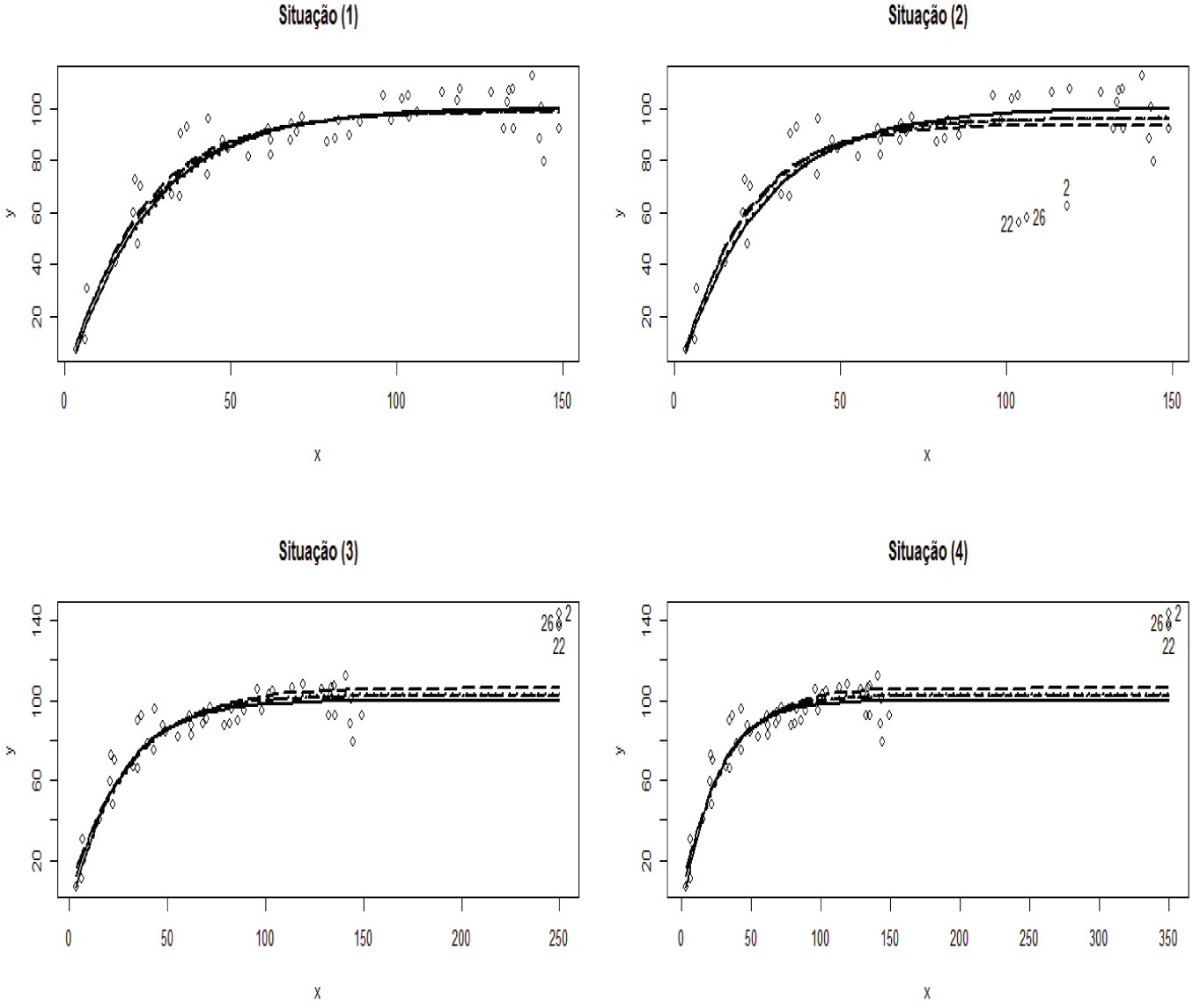


Figura 3.15: *Gráfico dos dados de uma simulação do modelo (3.2.2) com $n = 50$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (· · · · · · ·) curva ajustada pelo Método 2.*

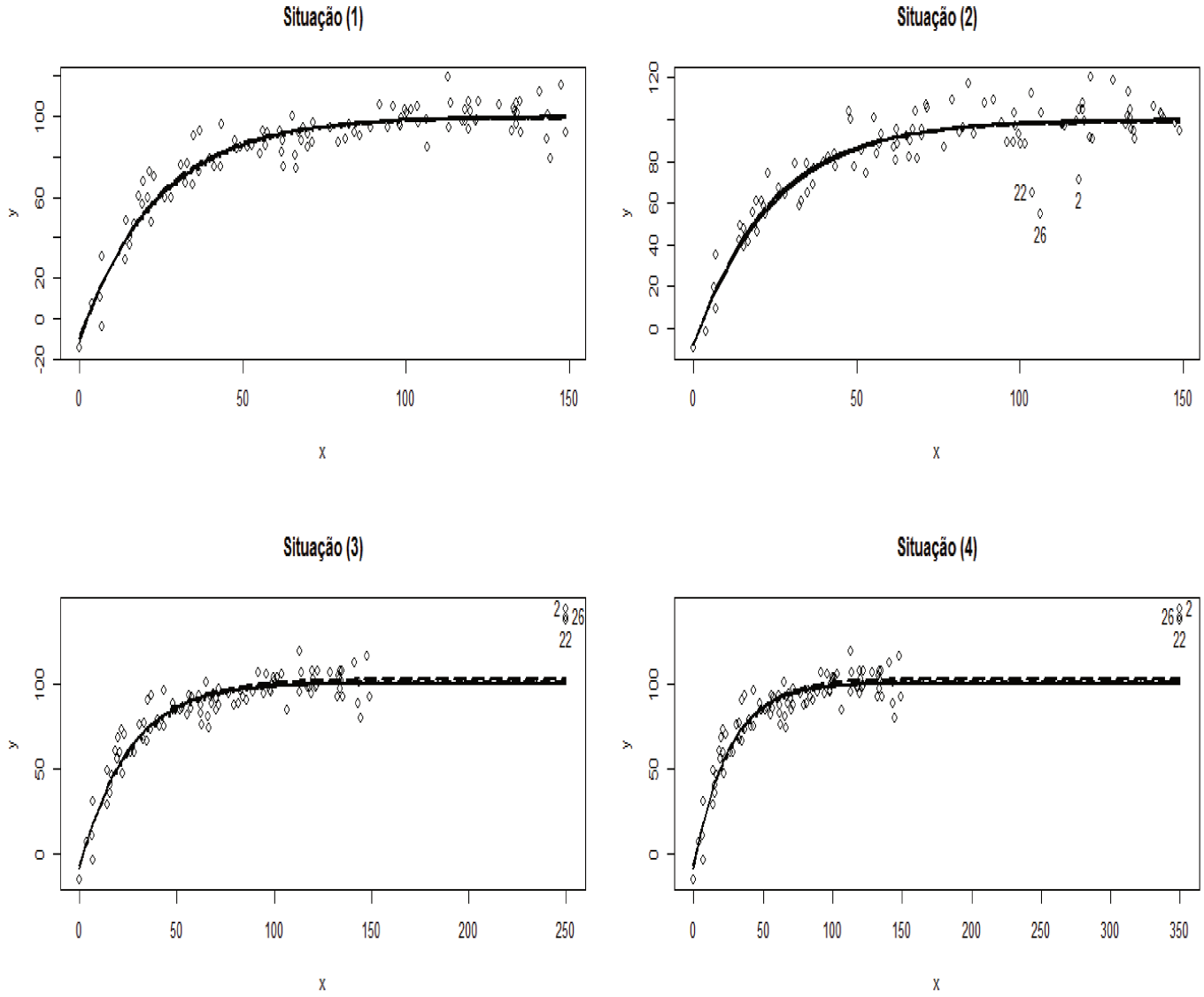


Figura 3.16: *Gráfico dos dados de uma simulação do modelo (3.2.2) com $n = 100$ para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada pelo Método 2.*

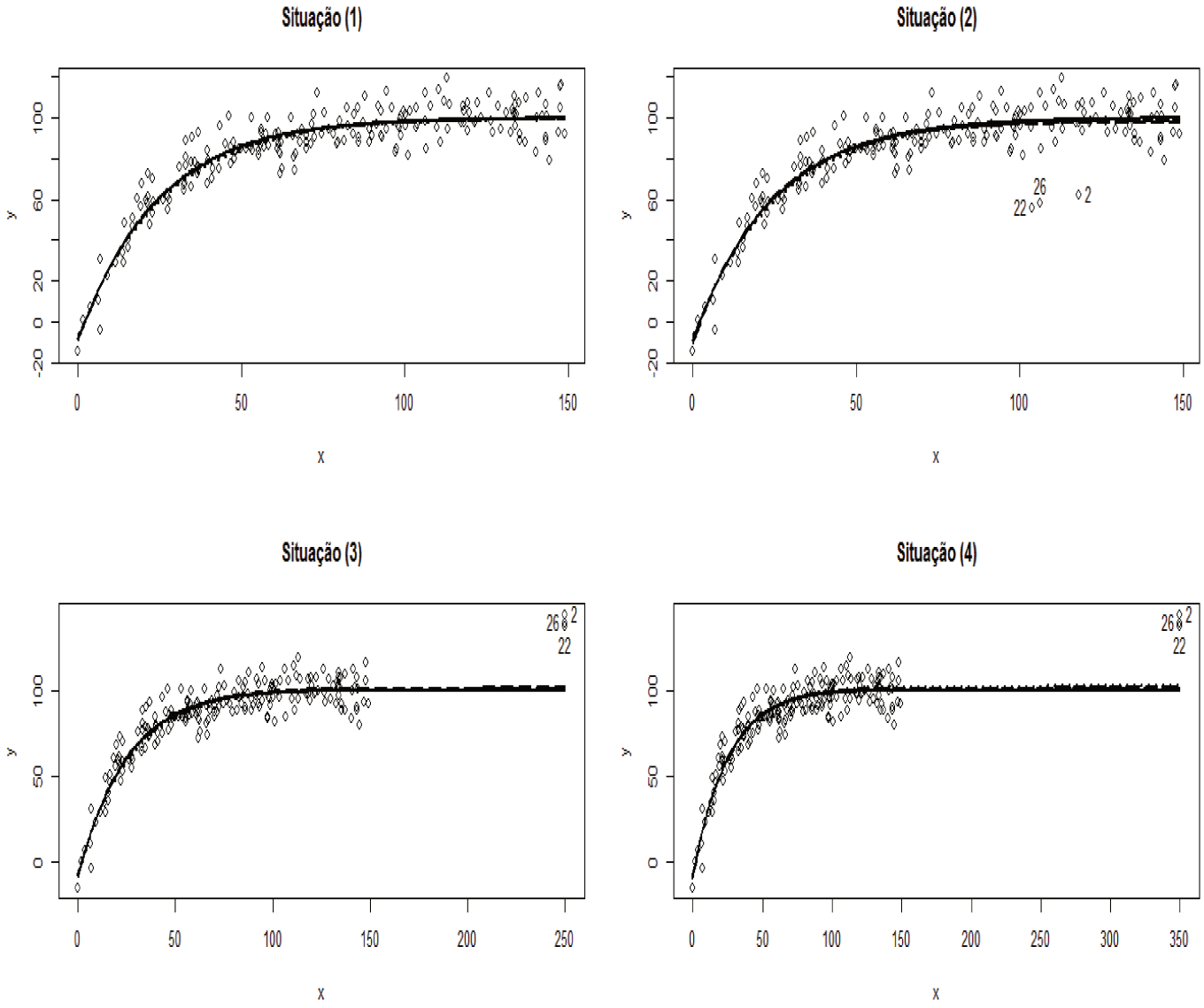


Figura 3.17: *Gráfico dos dados de uma simulação do modelo (3.2.2) com* $n = 200$ *para cada situação sob erros logístico tipo II. (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada por Método 2.*

Na Figura 3.15 os gráficos dos dados para amostras de tamanho 50. Vemos que métodos considerados fornecem bons ajustes aos dados, mesmo na de *outliers* com média e alta alavancagem. As Figuras 3.16 e 3.17 mostram os gráficos dos gerados com amostras de tamanho 100 e 200 respectivamente. Percebemos que os ajustes se adequam aos dados ilustrados.

Tabela 3.17: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros logísticos tipo II .* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$, $\phi = 20$ *e* $\sigma^2 = 65,8$. ***Situação (1).***

n	Estimação	Média (desvio padrão)				
		$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\phi}$	$\hat{\sigma}^2$
50	Método 1	100,070 (1,891)	-8,585 (7,143)	-3,203 (0,103)	18,532 (4,833)	60,969 (15,901)
	Método 2	100,141 (2,086)	-8,672 (7,782)	-3,207 (0,115)	30,720 (22,655)	101,064 (74,531)
	Método 3	100,069 (1,890)	-8,598 (7,134)	-3,203 (0,103)	18,531 (4,832)	60,966 (5,897)
100	Método 1	100,021 (1,355)	-8,673 (4,966)	-3,198 (0,071)	19,170 (3,296)	63,067 (10,844)
	Método 2	100,050 (1,521)	-8,496 (5,701)	-3,200 (0,082)	19,510 (4,741)	64,186 (15,597)
	Método 3	100,021 (1,355)	-8,677 (4,960)	-3,198 (0,071)	19,167 (3,298)	63,057 (10,850)
200	Método 1	100,025 (0,964)	-8,595 (3,828)	-3,200 (0,052)	19,628 (2,966)	64,574 (9,759)
	Método 2	100,066 (1,047)	-8,527 (4,435)	-3,201 (0,060)	19,616 (2,577)	64,533 (8,478)
	Método 3	100,024 (0,963)	-8,601 (3,813)	-3,200 (0,051)	19,558 (2,568)	64,342 (8,447)

Tabela 3.18: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros logísticos tipo II.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$ *e* $\sigma^2 = 65,8$. ***Situação (2).***

n	Estimação	Média (desvio padrão)				
		$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\phi}$	$\hat{\sigma}^2$
50	Método 1	98,010 (2,001)	-9,873 (7,333)	-3,147 (0,107)	35,581 (11,062)	117,057 (36,393)
	Método 2	98,867 (2,200)	-9,444 (7,948)	-3,170 (0,119)	147,835 (31,668)	486,358 (104,184)
	Método 3	97,971 (1,936)	-9,858 (7,350)	-3,145 (0,105)	34,546 (6,964)	113,653 (22,910)
100	Método 1	99,032 (1,326)	-8,464 (4,317)	-3,180 (0,068)	26,895 (2,670)	88,480 (8,785)
	Método 2	99,497 (1,531)	-8,310 (5,054)	-3,193 (0,080)	27,005 (2,667)	271,823 (25,387)
	Método 3	99,030 (1,327)	-8,500 (4,342)	-3,179 (0,069)	26,855 (2,657)	88,348 (8,740)
200	Método 1	99,568 (0,962)	-8,869 (3,833)	-3,187 (0,051)	23,121 (2,834)	76,359 (13,801)
	Método 2	99,765 (1,044)	-8,712 (4,462)	-3,192 (0,059)	23,201 (2,896)	76,095 (9,159)
	Método 3	99,561 (0,963)	-8,918 (3,831)	-3,187 (0,051)	23,112 (2,835)	75,870 (9,149)

Na Tabela 3.17 apresenta o resumo das estimativas para os dados gerados segundo a Situação (1). Observamos que, em média, as estimativas são próximas dos verdadeiros valores dos parâmetros. Quando $n = 50$, na estimação Método 2, o parâmetro de escala (ϕ) e a variância (σ^2) foram superestimados, mas a medida que amostra aumenta as estimativas médias se aproximam dos valores reais. Nos métodos 1 e 3 as estimativas médias de ϕ e σ^2 estão próximas dos valores verdadeiros em todos os tamanhos de amostra considerados.

A Tabela 3.18 temos os resultados referentes a Situação (2). As esimativas dos parâmetros de regressão, em média, estão próximos dos verdadeiros valores. Mas, quando $n = 50$, o parâmetro de escala e a variância foram superestimadas, principalmente, no Método 2. Para amostras maiores as estimativas são melhores, apresentando valores mais próximos dos verdadeiros em todos os métodos de estimação.

Sob efeito de pontos atípicos com média alavancagem, as estimativas dos parâmetros de regressão foram pouco influenciados pelas observações extremas, apresentando estimativas médias

Tabela 3.19: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros logísticos tipo II.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$, $\phi = 20$ *e* $\sigma^2 = 65,8$. ***Situação (3)***.

n	Estimação	Média (desvio padrão)				
		$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\phi}$	$\hat{\sigma}^2$
	Método 1	103,306 (2,428)	-4,944 (7,522)	-3,323 (0,128)	33,889 (11,049)	112,628 (33,682)
50	Método 2	102,025 (2,513)	-6,601 (8,276)	-3,274 (0,131)	132,630 (29,307)	432,501 (69,585)
	Método 3	103,317 (2,408)	-4,991 (7,531)	-3,323 (0,127)	33,096 (6,607)	110,034 (20,973)
	Método 1	101,400 (1,492)	-7,188 (5,082)	-3,248 (0,077)	26,294 (6,648)	86,098 (12,670)
100	Método 2	100,957 (1,641)	-7,504 (5,719)	-3,233 (0,087)	71,480 (25,013)	238,057 (56,856)
	Método 3	101,392 (1,481)	-7,193 (5,035)	-3,248 (0,077)	26,050 (4,008)	86,059 (12,631)
	Método 1	100,673 (1,004)	-7,765 (3,851)	-3,225 (0,054)	23,122 (5,006)	75,427 (9,177)
200	Método 2	100,482 (1,076)	-7,964 (4,462)	-3,217 (0,062)	22,927 (2,816)	75,606 (9,167)
	Método 3	100,665 (1,000)	-7,787 (3,834)	-3,224 (0,053)	22,865 (2,811)	75,397 (9,152)

Tabela 3.20: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros logísticos tipo II.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$, $\phi = 20$ *e* $\sigma^2 = 65,8$. ***Situação (4)***.

n	Estimação	Média (desvio padrão)				
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$\hat{\sigma}^2$
	Método 1	103,836 (5,225)	-4,460 (9,000)	-3,341 (0,211)	34,061 (11,336)	112,482 (33,536)
50	Método 2	102,066 (2,582)	-6,549 (8,348)	-3,276 (0,135)	132,213 (29,678)	431,866 (69,738)
	Método 3	103,804 (5,255)	-4,564 (8,893)	-3,339 (0,207)	33,148 (6,818)	110,032 (21,005)
	Método 1	101,398 (1,486)	-7,176 (5,080)	-3,248 (0,077)	26,072 (4,030)	86,104 (12,667)
100	Método 2	100,957 (1,644)	-7,503 (5,713)	-3,233 (0,087)	71,452 (25,015)	237,974 (56,884)
	Método 3	101,396 (1,484)	-7,181 (5,041)	-3,248 (0,077)	26,049 (4,008)	86,054 (12,631)
	Método 1	100,670 (1,002)	-7,764 (3,839)	-3,225 (0,053)	22,968 (3,646)	75,419 (9,173)
200	Método 2	100,482 (1,076)	-7,962 (4,461)	-3,217 (0,062)	22,927 (2,816)	75,604 (9,166)
	Método 3	100,667 (1,001)	-7,782 (3,835)	-3,224 (0,053)	22,864 (2,811)	75,395 (9,152)

com relação aos verdadeiros valores (Tabela 3.19). Mas, em amostras de tamanho 50 ϕ e σ^2, em média, foram superestimados, principalmente no Método 2. Nas amostras maiores os resultados melhoram, apresentando estimativas mais próximas dos valores verdadeiros. A mesma observação pode ser considerada ao analisarmos os resultados na Tabela 3.20 referente à Situação (4).

Resultados da estimação sob erros t-Student

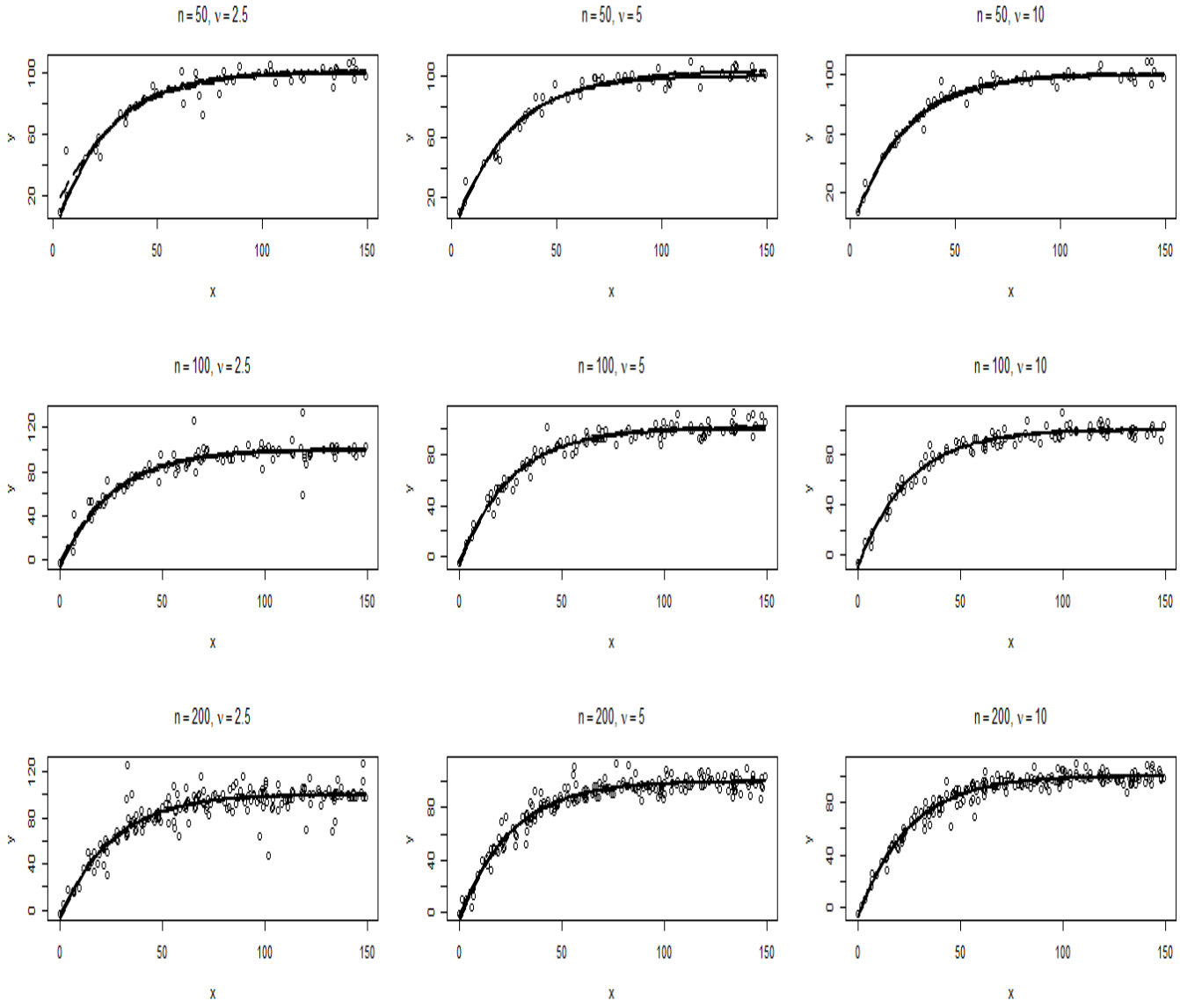


Figura 3.18: *Gráfico dos dados de uma simulação do modelo (3.2.2) sob erros t-Student da* ***Situação (1)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada pelo Método 2.

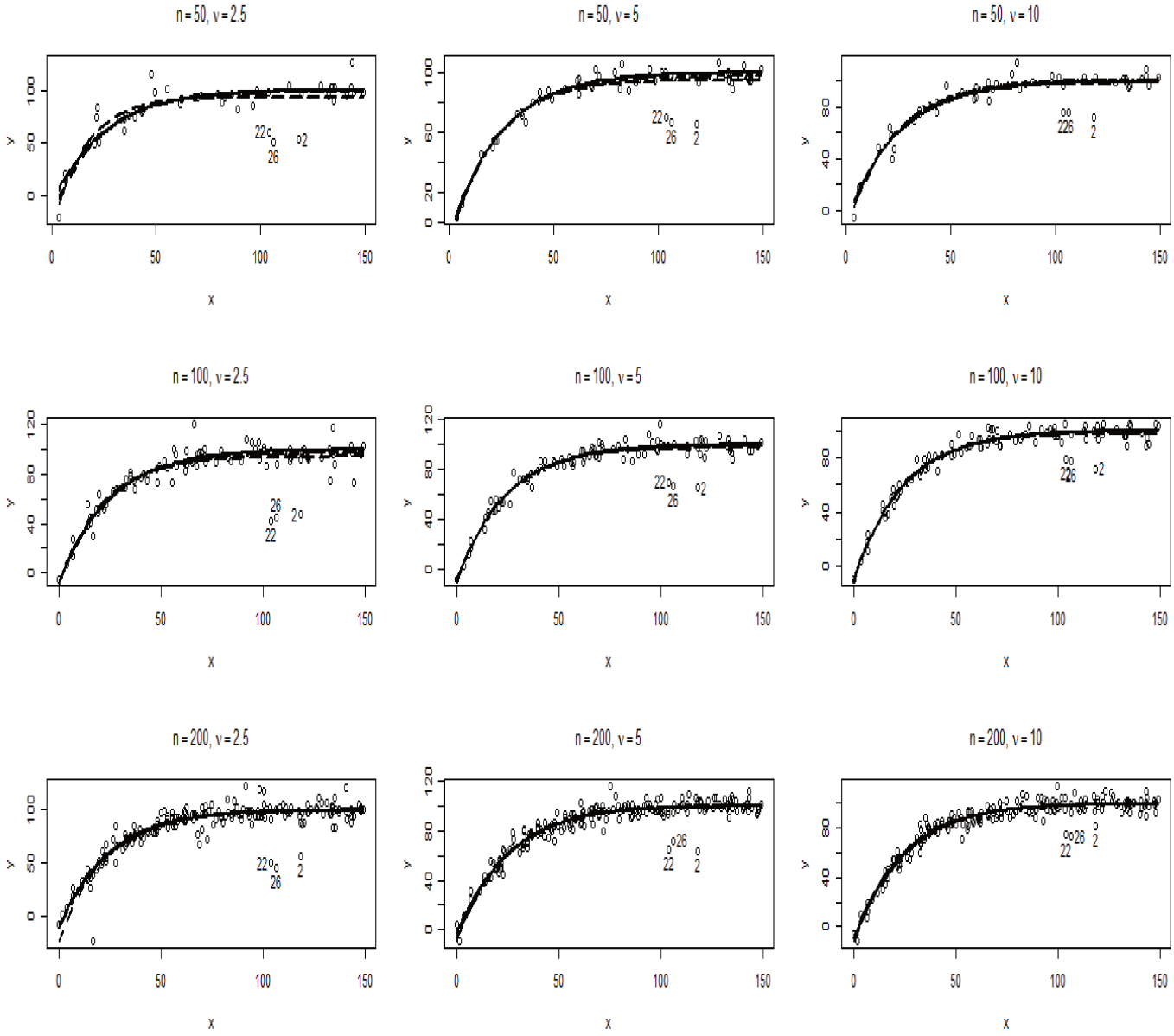


Figura 3.19: *Gráfico dos dados de uma simulação do modelo (3.2.2) sob erros t-Student da* ***Situação (2)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada pelo Método 2.

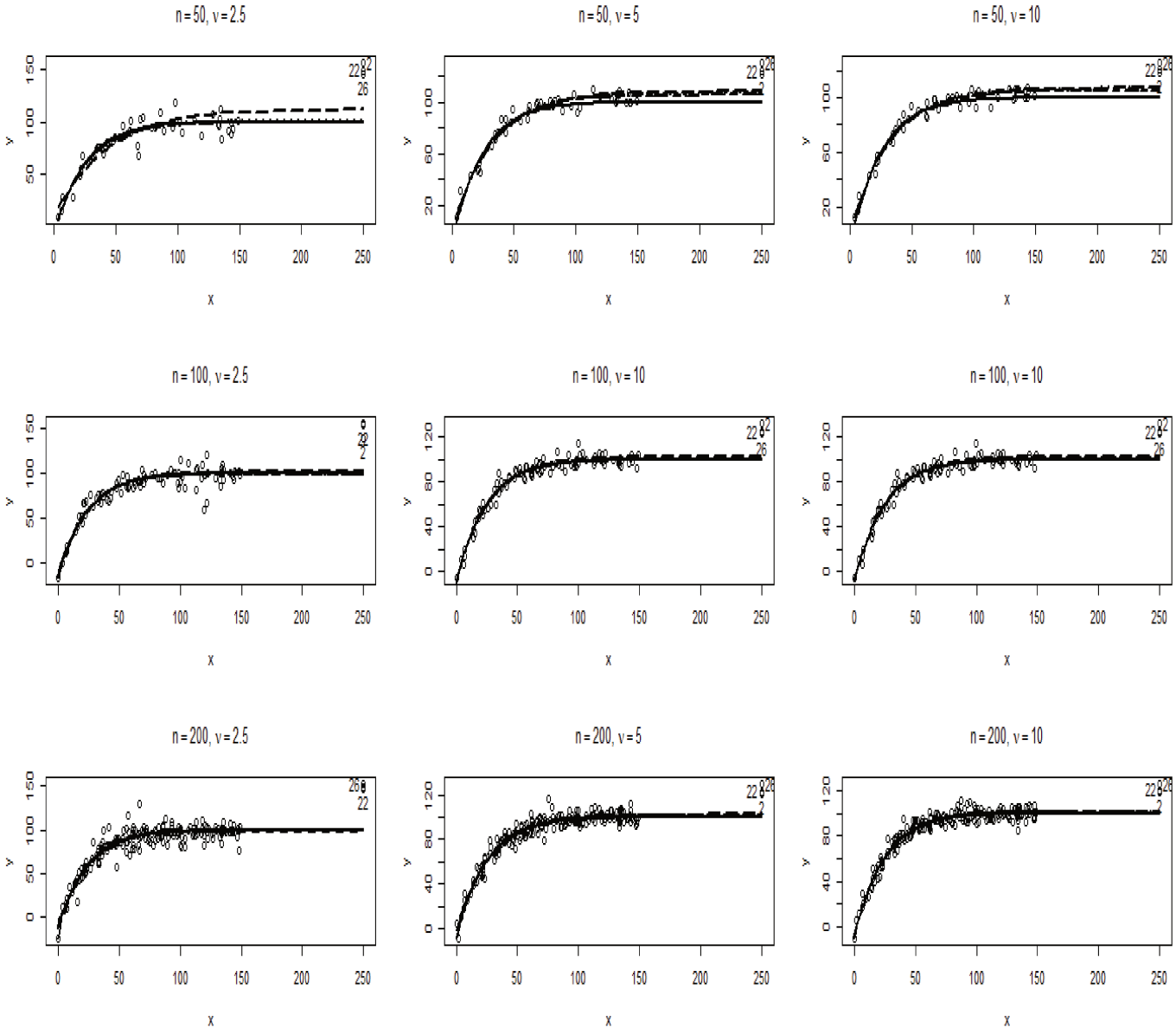


Figura 3.20: *Gráfico dos dados de uma simulação do modelo (3.2.2) sob erros t-Student da* ***Situação (3)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada pelo Método 2.

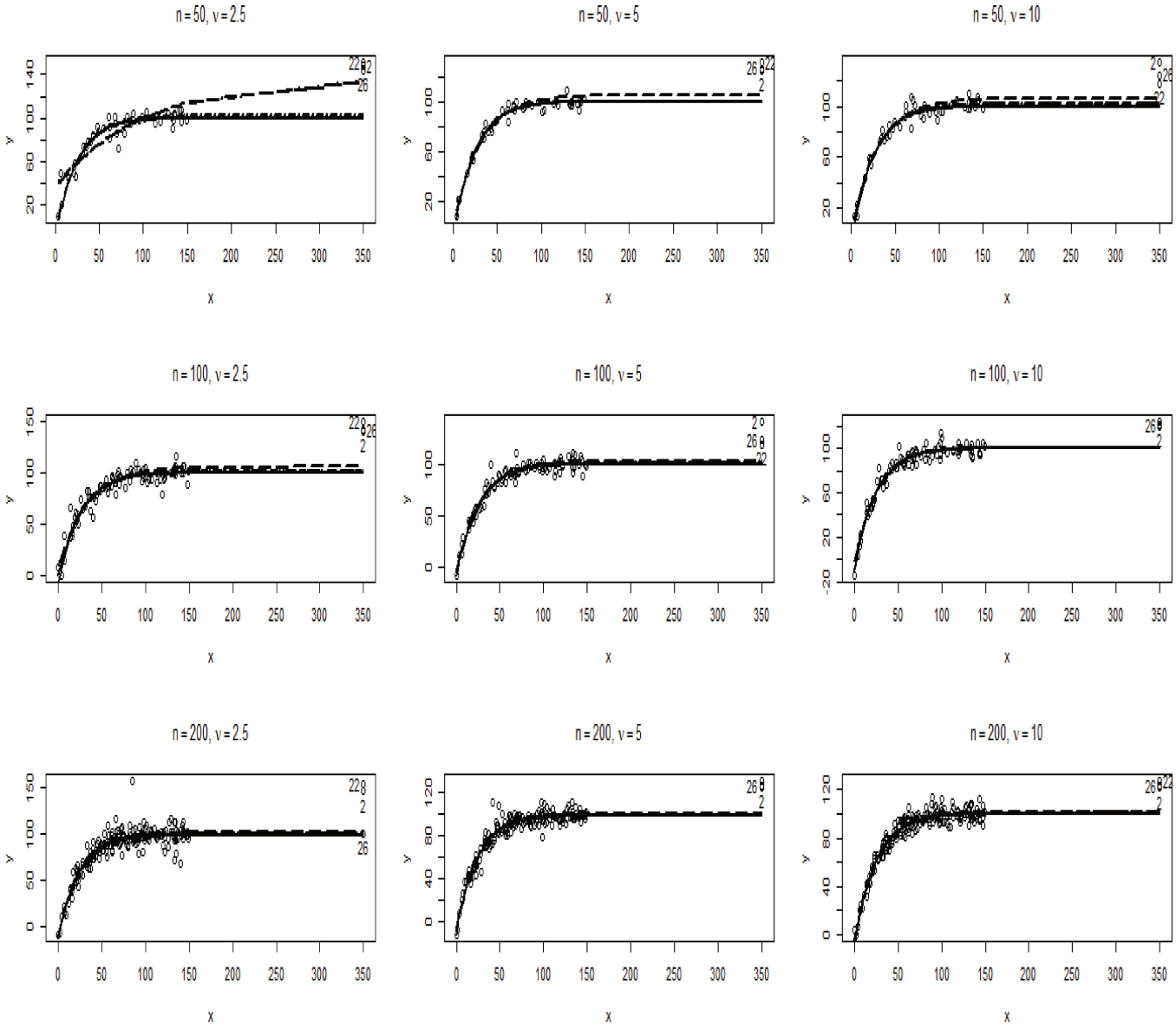


Figura 3.21: *Gráfico dos dados de uma simulação do modelo (3.2.2) sob erros t-Student da* ***Situação (4)*** *de acordo com o tamanho da amostra e grau de liberdade.* (——) curva verdadeira, (— — —) curva ajustada por mínimos quadrados, (- - - -) curva ajustada pelo Método 1, (— · — · —) curva ajustada pelo Método 3 e (·····) curva ajustada pelo Método 2.

A Figura 3.18 ilustra os gráficos dos dados sem observações influentes. Observamos que as curvas ajustadas se adequam aos dados exemplificados. Os gráficos dos dados da Situação (2) estão na Figura 3.19. As Figuras 3.20 e 3.21 mostram os gráficos dos dados gerados considerando Situação (3) e a Situação (4), respectivamente. Para amostras geradas com 50 observações e grau de liberdade $2, 5$, nas respectivas figuras, o ajuste por mínimos quadrados foi afetado pelos pontos atípicos, enquanto que os os outros são adequados aos dados. Para $n = 100, 200$ os justes não

sofrem influência dos pontos extremos.

Tabela 3.21: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros t-student.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$. ***Situação (1)***.

ν (σ^2)	n	Estimação	Média (desvio padrão)					
			$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\phi}$	$\hat{\nu}^a$	$\hat{\sigma}^2$
		Método 1	100,06 (1,36)	-9,15 (6,01)	-3,20 (0,08)	22,6 (9,1)	6,5; 3,1 (8,4)	89,3 (144,0)
	50	Método 2	100,11 (1,45)	-8,52 (5,89)	-3,21 (0,08)	22,0 (8,6)	5,4; 3,0 (6,7)	102,6 (190,6)
		Método 3	100,10 (1,38)	-8,62 (5,95)	-3,20 (0,08)	22,4 (8,9)	6,4; 3,1 (8,2)	66,0 (41,3)
		Método 1	99,87 (0,95)	-10,68 (5,84)	-3,18 (0,07)	21,3 (6,4)	3,6; 2,9 (3,3)	146,4 (444,2)
2,5 (100)	100	Método 2	100,06 (0,96)	-8,55 (3,98)	-3,20 (0,06)	20,9 (6,0)	3,4; 2,8 (3,0)	125,9 (238,3)
		Método 3	100,04 (0,90)	-8,54 (3,65)	-3,20 (0,05)	21,0 (6,2)	3,6; 2,8 (3,3)	75,2 (41,1)
		Método 1	99,80 (0,74)	-11,89 (7,54)	-3,17 (0,07)	21,2 (4,1)	3,0; 2,8 (0,8)	134,1 (289,6)
	200	Método 2	100,04 (0,73)	-8,47 (3,12)	-3,20 (0,04)	20,6 (3,9)	2,9; 2,8 (0,8)	145,4 (467,1)
		Método 3	100,04 (0,65)	-8,45 (2,83)	-3,20 (0,04)	20,6 (4,0)	2,9; 2,8 (0,8)	81,1 (37,5)
		Método 1	99,99 (1,22)	-8,81 (4,79)	-3,20 (0,07)	20,1 (6,7)	13,5; 7,4 (11,2)	39,9 (119,5)
	50	Método 2	100,06 (1,36)	-8,79 (5,20)	-3,20 (0,08)	19,7 (6,6)	11,4; 11,4 (10,5)	35,5 (25,4)
		Método 3	99,995 (1,23)	-8,75 (4,83)	-3,20 (0,07)	20,0 (6,6)	13,3; 7,2 (11,2)	32,9 (16,3)
		Método 1	99,92 (0,89)	-9,67 (3,99)	-3,19 (0,05)	20,3 (5,1)	9,6; 5,6 (8,9)	34,1 (10,7)
5 (33,3)	100	Método 2	100,08 (0,99)	-8,43 (3,60)	-3,20 (0,05)	20,0 (5,0)	8,7; 5,2 (8,2)	34,9 (12,8)
		Método 3	100,03 (0,88)	-8,46 (3,23)	-3,20 (0,05)	20,2 (5,1)	9,5; 5,6 (8,8)	34,0 (11,0)
		Método 1	99,90 (0,65)	-10,23 (3,85)	-3,18 (0,05)	20,3 (3,6)	7,0; 5,3 (5,0)	33,2 (6,3)
	200	Método 2	100,03 (0,70)	-8,44 (2,79)	-3,20 (0,04)	20,0 (3,6)	6,6; 5,1 (4,6)	33,3 (6,4)
		Método 3	100,03 (0,62)	-8,50 (2,60)	-3,20 (0,04)	20,1 (3,7)	7,0; 5,3 (5,0)	33,0 (6,1)
		Método 1	100,00 (1,25)	-8,72 (4,69)	-3,20 (0,07)	19,1 (5,7)	19,7; 29,8 (11,5)	25,6 (14,4)
	50	Método 2	100,04 (1,41)	-8,79 (5,52)	-3,20 (0,08)	18,8 (5,6)	16,7; 12,7 (11,5)	26,4 (13,3)
		Método 3	100,00 (1,25)	-8,72 (4,69)	-3,20 (0,07)	19,0 (5,7)	19,5; 29,8 (11,4)	24,8 (7,9)
		Método 1	99,93 (0,84)	-9,57 (3,44)	-3,19 (0,05)	19,6 (4,2)	17,7; 15,4 (10,9)	24,7 (4,4)
10 (25)	100	Método 2	100,03 (0,99)	-8,42 (3,52)	-3,20 (0,05)	19,4 (4,2)	16,2; 11,2 (10,8)	25,0 (4,6)
		Método 3	100,04 (0,84)	-8,34 (2,97)	-3,20 (0,05)	19,5 (4,2)	17,5; 14,1 (10,9)	24,6 (4,4)
		Método 1	99,92 (0,62)	-9,81 (3,31)	-3,19 (0,04)	19,9 (3,1)	15,2; 12,0 (9,4)	24,9 (3,3)
	200	Método 2	100,01 (0,69)	-8,54 (2,75)	-3,20 (0,04)	19,7 (3,1)	14,1; 10,7 (9,0)	25,0 (3,3)
		Método 3	100,03 (0,60)	-8,43 (2,42)	-3,20 (0,03)	19,8 (3,1)	15,0; 11,6 (9,3)	24,8 (3,3)

[a]Média; mediana (desvio padrão) de $\hat{\nu}$.

A Tabela 3.21 mostra a média e desvio padrão das estimativas considerando dados sem observações influentes. Observamos que a média das estimativas dos parâmetros de regressão e escala são próximas dos valores verdadeiros e os graus de liberdade foram superestimado em amostras com 50 observações.

Ainda na Tabela 3.21, quando$\nu = 2,5$, a média das estimativas da variância foi melhor no Método 2, com valor próximo ao verdadeiro. Para as amostras maiores a estimativa média da variância nos métodos 1 e 2 aumenta com amostra e se afasta do valor real, enquanto que no Método 3 se aproxima da verdadeira variância.

Quando $\nu = 5, 10$, na Tabela 3.21, a estimativa média da variância está mais próxima da variância real no Método 3 qualquer que seja o tamanho da amostra, enquanto que o Método 2 tem valores próximos nas amostras maiores.

Na presença de pontos atípicos na variável resposta, em geral, os graus de liberdade foram subestimados. Entretanto, observamos que a média das estimativas para os parâmetros da regressão

Tabela 3.22: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros t-student.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$. ***Situação (2).***

ν	n	Estimação	Média (desvio padrão)				
			$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\phi}$	$\hat{\nu}$
		Método 1	99,535 (1,458)	-9,653 (6,035)	-3,182 (0,082)	25,705 (7,233)	2,002 (0,024)
	50	Método 2	98,961 (1,579)	-9,783 (6,639)	-3,167 (0,093)	26,048 (7,284)	2,002 (0,016)
		Método 3	99,553 (1,474)	-9,598 (6,374)	-3,183 (0,085)	25,730 (7,331)	2,006 (0,049)
		Método 1	99,610 (1,064)	-11,307 (7,818)	-3,169 (0,078)	21,584 (5,774)	2,041 (0,107)
2,5	100	Método 2	99,486 (1,084)	-8,945 (3,998)	-3,184 (0,057)	21,210 (5,275)	2,038 (0,102)
		Método 3	99,803 (0,994)	-8,713 (3,682)	-3,194 (0,053)	21,096 (5,270)	2,040 (0,113)
		Método 1	99,608 (0,855)	-14,086 (13,665)	-3,154 (0,094)	20,402 (4,182)	2,148 (0,207)
	200	Método 2	99,770 (0,733)	-8,830 (2,946)	-3,192 (0,040)	19,537 (3,290)	2,132 (0,194)
		Método 3	99,919 (0,656)	-8,778 (2,741)	-3,197 (0,036)	19,483 (3,283)	2,133 (0,211)
		Método 1	99,646 (1,350)	-8,935 (5,109)	-3,192 (0,074)	18,346 (5,766)	2,180 (0,346)
	50	Método 2	99,278 (1,486)	-8,934 (5,420)	-3,184 (0,082)	18,493 (5,790)	2,171 (0,330)
		Método 3	99,647 (1,350)	-8,903 (5,144)	-3,192 (0,074)	18,443 (5,780)	2,199 (0,356)
		Método 1	99,625 (0,913)	-10,017 (3,998)	-3,178 (0,055)	18,260 (4,495)	2,676 (0,553)
5	100	Método 2	99,589 (0,980)	-8,756 (3,536)	-3,189 (0,054)	18,067 (4,403)	2,640 (0,534)
		Método 3	99,740 (0,892)	-8,738 (3,191)	-3,192 (0,048)	18,066 (4,410)	2,665 (0,550)
		Método 1	99,767 (0,691)	-10,070 (3,792)	-3,181 (0,045)	18,537 (3,278)	3,377 (0,622)
	200	Método 2	99,589 (0,980)	-8,756 (3,536)	-3,189 (0,054)	18,067 (4,403)	2,640 (0,534)
		Método 3	99,892 (0,661)	-8,486 (2,506)	-3,197 (0,036)	18,332 (3,110)	3,358 (0,609)
		Método 1	99,552 (1,343)	-8,721 (4,857)	-3,189 (0,072)	16,638 (5,381)	2,418 (0,572)
	50	Método 2	99,212 (1,447)	-8,909 (5,267)	-3,179 (0,078)	16,775 (5,371)	2,404 (0,561)
		Método 3	99,552 (1,344)	-8,710 (4,898)	-3,190 (0,072)	16,678 (5,325)	2,427 (0,570)
		Método 1	99,651 (0,901)	-9,689 (3,642)	-3,179 (0,050)	16,962 (4,023)	3,290 (0,778)
10	100	Método 2	99,617 (1,043)	-8,760 (3,639)	-3,187 (0,055)	16,828 (4,006)	3,229 (0,734)
		Método 3	99,743 (0,901)	-8,686 (3,131)	-3,190 (0,046)	16,876 (4,016)	3,288 (0,779)
		Método 1	99,744 (0,607)	-10,136 (3,441)	-3,180 (0,040)	17,374 (2,731)	4,232 (0,874)
	200	Método 2	99,814 (0,660)	-8,764 (2,762)	-3,193 (0,037)	17,231 (2,703)	4,176 (0,845)
		Método 3	99,859 (0,585)	-8,669 (2,474)	-3,195 (0,033)	17,267 (2,708)	4,229 (0,879)

e escala são mais próximas dos valores verdadeiros (Tabela 3.22). Além disso, verificamos que os resultados obtidos nos três métodos de estimação similares.

Mesmo sob a presença de *outliers* com média alavangem as estimativas dos parâmetros de regressão e escala apresentam média próximas dos valores verdadeiros, mas a média das estimativas do grau de liberdade são menores do que os valores verdadeiros (subestimação), principalmente, quando valores teóricos desse parâmetro são 5 e 10 (veja Tabela 3.23). Também observamos uma similaridade entre os métodos de estimação.

Para Situação (4) podemos tirar as mesmas conlusões que foram consideradas para Situação (3). Isto é, a média das estimativas dos parâmetros de regressão e escala foram pouco influenciadas

Tabela 3.23: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros t-student.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$. ***Situação (3).***

ν	n	Estimação	Média (desvio padrão)				
			$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\phi}$	$\hat{\nu}$
		Método 1	100,678 (1,441)	-8,261 (5,998)	-3,221 (0,086)	26,051 (7,132)	2,006 (0,042)
	50	Método 2	101,651 (1,722)	-6,770 (6,653)	-3,262 (0,102)	26,741 (7,431)	2,009 (0,064)
		Método 3	100,895 (4,052)	-7,689 (6,581)	-3,231 (0,132)	26,291 (8,450)	2,020 (0,174)
2,5		Método 1	100,104 (1,052)	-10,793 (8,187)	-3,187 (0,082)	21,694 (5,293)	2,041 (0,108)
	100	Método 2	100,732 (1,142)	-7,575 (4,273)	-3,229 (0,065)	21,425 (4,919)	2,040 (0,110)
		Método 3	100,321 (0,990)	-8,035 (3,867)	-3,214 (0,057)	21,237 (4,895)	2,040 (0,122)
		Método 1	99,821 (0,823)	-13,121 (12,045)	-3,165 (0,085)	20,381 (3,885)	2,162 (0,215)
	200	Método 2	100,309 (0,770)	-8,249 (3,060)	-3,211 (0,045)	19,729 (3,133)	2,150 (0,204)
		Método 3	100,120 (0,665)	-8,427 (2,776)	-3,204 (0,040)	19,669 (3,156)	2,152 (0,221)
		Método 1	100,708 (1,468)	-8,054 (5,094)	-3,226 (0,077)	18,738 (6,064)	2,278 (0,508)
	50	Método 2	101,309 (1,637)	-7,415 (5,665)	-3,249 (0,088)	19,142 (6,230)	2,295 (0,554)
		Método 3	100,716 (1,465)	-7,984 (5,104)	-3,226 (0,076)	18,828 (6,061)	2,292 (0,509)
5		Método 1	100,247 (0,981)	-9,805 (4,443)	-3,194 (0,060)	18,565 (4,556)	2,801 (0,609)
	100	Método 2	100,656 (1,051)	-7,752 (3,616)	-3,224 (0,058)	18,400 (4,456)	2,778 (0,597)
		Método 3	100,412 (0,962)	-8,068 (3,323)	-3,214 (0,052)	18,394 (4,506)	2,808 (0,623)
		Método 1	100,061 (0,695)	-9,993 (3,817)	-3,190 (0,045)	18,935 (3,296)	3,493 (0,684)
	200	Método 2	100,287 (0,725)	-7,993 (2,933)	-3,212 (0,041)	18,731 (3,155)	3,459 (0,669)
		Método 3	100,210 (0,665)	-8,161 (2,629)	-3,209 (0,037)	18,758 (3,147)	3,488 (0,681)
		Método 1	100,835 (1,511)	-7,722 (4,957)	-3,229 (0,076)	17,572 (6,741)	2,965 (3,216)
	50	Método 2	101,261 (1,594)	-7,288 (5,403)	-3,244 (0,082)	17,762 (6,415)	2,738 (1,738)
		Método 3	100,842 (1,512)	-7,650 (4,999)	-3,230 (0,077)	17,576 (6,695)	2,967 (3,189)
10		Método 1	100,320 (0,933)	-9,497 (3,647)	-3,198 (0,051)	17,335 (4,130)	3,549 (1,622)
	100	Método 2	100,633 (1,036)	-8,017 (3,519)	-3,218 (0,054)	17,311 (4,120)	3,522 (1,548)
		Método 3	100,453 (0,939)	-8,152 (3,127)	-3,213 (0,048)	17,317 (4,142)	3,577 (1,595)
		Método 1	100,112 (0,631)	-9,866 (3,530)	-3,192 (0,041)	17,649 (2,910)	4,517 (1,559)
	200	Método 2	100,275 (0,675)	-8,214 (2,846)	-3,209 (0,039)	17,531 (2,883)	4,479 (1,549)
		Método 3	100,247 (0,614)	-8,233 (2,555)	-3,208 (0,034)	17,593 (2,896)	4,558 (1,597)

pelos pontos atípicos, mas para os graus de liberdade a média das estimativas foram bem menores do que o valores nominais 5 e 10. Em geral, a média das estimativas nos três métodos são similares.

No ajuste do modelo (3.2.2) com erros logísticos-II as estimativas de $\boldsymbol{\beta}$ foram pouco influenciadas pelos *outliers* e observações atípicas com média e alta alavancagem, mas o parâmetro de escala foi superestimado no Método 2 quando $n = 50, 100$. Verificamos que os melhores resultados foram observados na estimação pelo Método 3. No ajuste com erros t-student observamos que os métodos forneceram resultados similares e as estimativas de $\boldsymbol{\beta}$ e ϕ foram pouco influenciadas pelos *outliers* e pelas observações com média e alta alavancagem, mas o grau de liberdade foi subestimados nestas situações.

Tabela 3.24: *Média (desvio padrão) das estimativas obtidas em 500 réplicas do modelo não-linear (3.2.2) sob erros t-student.* $\alpha_1 = 100$, $\alpha_2 = -8,5$, $\alpha_3 = -3,2$ *e* $\phi = 20$. ***Situação (4).***

ν	n	Estimação	Média (desvio padrão)				
			$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\phi}$	$\hat{\nu}$
		Método 1	100,785 (1,795)	-8,763 (6,950)	-3,220 (0,094)	27,073 (9,491)	2,061 (1,252)
	50	Método 2	101,699 (1,844)	-7,203 (6,064)	-3,257 (0,097)	27,352 (7,433)	2,007 (0,046)
		Método 3	101,152 (1,962)	-7,555 (5,914)	-3,236 (0,100)	26,601 (8,605)	2,694 (0,134)
2,5		Método 1	100,027 (1,088)	-11,347 (9,052)	-3,180 (0,088)	21,631 (5,443)	2,042 (0,110)
	100	Método 2	100,682 (1,112)	-7,823 (4,230)	-3,225 (0,063)	21,331 (5,008)	2,043 (0,113)
		Método 3	100,261 (0,994)	-8,219 (3,776)	-3,210 (0,055)	21,146 (4,990)	2,043 (0,124)
		Método 1	99,847 (0,801)	-13,035 (12,337)	-3,167 (0,087)	20,330 (3,669)	2,157 (0,214)
	200	Método 2	100,358 (0,754)	-7,946 (3,118)	-3,216 (0,046)	19,694 (3,044)	2,147 (0,202)
		Método 3	100,145 (0,645)	-8,270 (2,851)	-3,207 (0,040)	19,600 (3,054)	2,143 (0,220)
		Método 1	100,721 (1,421)	-8,077 (5,149)	-3,226 (0,076)	18,619 (5,972)	2,252 (0,433)
	50	Método 2	101,335 (1,576)	-7,296 (5,834)	-3,251 (0,088)	19,025 (6,133)	2,272 (0,499)
		Método 3	100,730 (1,417)	-7,971 (5,152)	-3,227 (0,076)	18,668 (5,955)	2,261 (0,437)
5		Método 1	100,205 (0,963)	-9,738 (4,207)	-3,194 (0,057)	18,599 (4,472)	2,807 (0,615)
	100	Método 2	100,598 (1,050)	-7,819 (3,548)	-3,222 (0,056)	18,476 (4,375)	2,786 (0,602)
		Método 3	100,357 (0,945)	-8,148 (3,269)	-3,212 (0,050)	18,462 (4,422)	2,813 (0,626)
		Método 1	100,040 (0,713)	-10,103 (3,970)	-3,189 (0,047)	18,727 (3,256)	3,463 (0,672)
	200	Método 2	100,263 (0,751)	-8,064 (2,912)	-3,211 (0,042)	18,522 (3,118)	3,429 (0,652)
		Método 3	100,193 (0,673)	-8,203 (2,622)	-3,208 (0,037)	18,536 (3,127)	3,454 (0,671)
		Método 1	100,893 (1,581)	-7,684 (4,986)	-3,231 (0,080)	17,551 (6,602)	2,971 (3,196)
	50	Método 2	101,288 (1,650)	-7,394 (5,574)	-3,244 (0,086)	17,682 (6,294)	2,736 (1,675)
		Método 3	100,901 (1,580)	-7,596 (5,032)	-3,232 (0,081)	17,564 (6,559)	2,974 (3,167)
10		Método 1	100,307 (0,952)	-9,354 (3,640)	-3,199 (0,052)	17,255 (4,127)	3,540 (1,563)
	100	Método 2	100,616 (1,059)	-7,959 (3,533)	-3,219 (0,056)	17,229 (4,145)	3,516 (1,524)
		Método 3	100,436 (0,949)	-8,058 (3,097)	-3,214 (0,048)	17,237 (4,163)	3,571 (1,554)
		Método 1	100,119 (0,630)	-9,895 (3,500)	-3,192 (0,041)	17,778 (2,884)	4,543 (1,529)
	200	Método 2	100,276 (0,677)	-8,248 (2,801)	-3,209 (0,038)	17,663 (2,859)	4,510 (1,541)
		Método 3	100,254 (0,614)	-8,264 (2,519)	-3,208 (0,034)	17,720 (2,880)	4,587 (1,589)

Capítulo 4

Influência em Modelos Simétricos

A identificação de observações que afetam as estimativas dos parâmetros é um componente principal para validação de modelos. A presença de tais observações pode tornar inadequadas as conclusões com base num ajuste que nao incorpore o carater atipico destas. Uma abordagem para identificação de observações influentes é avaliar o efeito da exclusão de uma observação na estimativa dos parâmetros do modelo. Sob esse aspecto, uma medida bastante popular foi proposta por Cook (1977) para modelos de regressão linear normal, a distância de Cook. Mas podemos destacar outros métodos em análise de influência que foram desenvolvidos para o modelo linear normal e que foram estendidos para outros modelos. Assim, Andrews e Pregibon (1978) propuseram uma medida para identificar uma ou mais observações influentes em regressão linear. Cook e Weisberg (1982) definiram uma medida mais geral chamada afastamento da verossimilhança. Belsley et al. (1980), Atkinson (1985) e Chatterjee e Hadi (1988) examinaram essa área com aplicações em regressão linear e outros modelos. Peña (2005) introduziu uma nova forma para analisar influência no modelo linear normal. Os resultados de Cook (1977) foram estendidos para os modelos lineares generalizados por Pregibon (1981). Galea et al. (2000) desenvolveu medidas de influência em modelos elípticos lineares. Vanegas e Cysneiros (2010) estenderam os resultados de Cook e Weisberg (1982) e Cook (1977) para o modelos simétricos não-lineares.

Neste trabalho propomos uma extensão da medida proposta por Peña (2005) para os modelos simétricos lineares e não-lineares. Como medidas comparativas utilizamos a distância de Cook generalizada e a estatística de Andrews e Pregibon (1978). Em particular é discutida a conveniência de utilizar a metodologia de limiares para caracterizar uma observação como influente ou não, assim como o efeito da estimação de parâmetros na construção de limiares.

Na seção 4.1 apresentamos as estatísticas de influência distância de Cook generalizada, estatística de Peña e estatística de Andrew-Pregibon. A metodologia de detecção de observações influentes para cada uma dessas estatísticas descrevemos na seção 4.2. Como aplicação, ilustramos exemplos com dados simulados e dados reais na seção 5.

4.1 Estatísticas de influência

Considere o modelo de regressão com erros simétricos

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{4.1.1}$$

em que $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ o vetor de respostas observadas, $\epsilon_i \sim S(0, \phi, g)$ e $\mu_i = \mu(\boldsymbol{\beta}; \mathbf{x}_i)$ é uma função (linear ou não-linear) contínua e diferenciável com respeito à $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ tal que a matriz de derivadas $\mathbf{D}_\beta = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$ tenha posto p ($p < n$) para todo $\boldsymbol{\beta} \in \Omega_\beta \subset \mathbb{R}^p$, com Ω_β um conjunto compacto com pontos interiores. Além disso, considere $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ e $\mathbf{x}_i$ um vetor de variáveis explanatórias. No caso linear, tem-se que $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ com $\mathbf{D}_\beta = \mathbf{X}$ uma matriz $n \times p$ de posto completo cuja i-ésima linha é denotada por $\mathbf{x}_i^{\mathrm{T}} = (1, x_{i1}, \ldots, x_{i(p-1)})$.

Para o modelo definido por (4.1.1), a função de log-verossimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)^{\mathrm{T}}$ baseada em uma amostra, $y_i, \ldots, y_n$, independente é dada por

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} l(y_i; \boldsymbol{\theta}) = -\frac{n}{2}\log\phi + \sum_{i=1}^{n} \log\left[g\left(u_i\right)\right]. \tag{4.1.2}$$

Denotamos por $\hat{\boldsymbol{\theta}}$ o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ e por $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^{\mathrm{T}} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$ o vetor dos preditores. Seja $\hat{\boldsymbol{\theta}}_{(i)} = (\hat{\boldsymbol{\beta}}_{(i)}, \phi_{(i)})^{\mathrm{T}}$ o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ quando a i-ésima observação é excluída e $\hat{\boldsymbol{y}}_{(i)} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{(i)})$.

A seguir apresentamos as estatísticas de influência distância de Cook generalizada, estatística de Peña e estatística de Andrew-Pregibon.

4.1.1 Distância de Cook generalizada

Uma medida muito popular foi proposta por Cook (1977), conhecida como distância de Cook (D_i). A princípio foi desenvolvida considerando os modelos de regressão linear normal, tendo como base a curva de influência amostral para $\boldsymbol{\beta}$. Entretanto, podemos estender está medida para avaliar influência em outros parâmetros, além de $\boldsymbol{\beta}$. Isto pode ser feito considerando a distância de Cook generalizada,

$$D(\hat{\boldsymbol{\theta}}_{(i)}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{C}(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}),$$

onde $\mathbf{C}$ é uma matriz positiva definida assintoticamente equivalente à matriz de informação de Fisher de $\boldsymbol{\theta}$. Para $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\phi})^{\mathrm{T}}$, podemos avaliar influência de observações em um determinado subconjunto de $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ ou ϕ. Neste caso, a distância de Cook generalizada pode ser expressa da seguinte forma para $\boldsymbol{\beta}$ e ϕ, respectivamente,

$$(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^{\mathrm{T}} \left\{(\mathbf{I}_p, 0)\mathbf{C}^{-1}(\mathbf{I}_p, 0)\right\}^{-1} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})$$

e

$$(\hat{\phi}_{(i)} - \hat{\phi})^{\mathrm{T}} \left\{(\mathbf{0}_p, 1)\mathbf{C}^{-1}(\mathbf{0}_p, 1)\right\}^{-1} (\hat{\phi}_{(i)} - \hat{\phi}),$$

onde $\mathbf{I}_p$ e $\mathbf{0}$ são a matriz de identidade e a matriz de zeros, respectivamente. Diversas escolhas para $\mathbf{C}$ são proposta por Cook e Weisberg (1982, pg. 124), neste trabalho escolhemos $\mathbf{C} = \mathbf{K}(\hat{\boldsymbol{\theta}})$, a matriz de informação de Fisher. Neste caso, $D(\hat{\boldsymbol{\theta}}_{(i)})$ pode ser decomposto em duas partes,

$$D(\hat{\boldsymbol{\theta}}_{(i)}) = D(\hat{\boldsymbol{\beta}}_{(i)}) + D(\hat{\phi}_{(i)}), \tag{4.1.3}$$

onde

$$D(\hat{\boldsymbol{\beta}}_{(i)}) = (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^{\mathrm{T}} \mathbf{K}_{\hat{\beta}\hat{\beta}} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}) \tag{4.1.4}$$

$$D(\hat{\phi}_{(i)}) = (\hat{\phi}_{(i)} - \hat{\phi})^2 \mathrm{K}_{\hat{\phi}\hat{\phi}}. \tag{4.1.5}$$

Para os modelos de regressão linear normal, um critério para a comparação da magnitude dos valores de $D(\hat{\boldsymbol{\beta}}(i))/p$ consiste em compará-los com $F_{(p,n-p;0,5)}$, o percentil 50% da distribuição F com p graus de liberdade no numerador e $n-p$ no denominador. Se $D(\hat{\boldsymbol{\beta}}(i))/p > F_{(p,n-p;0,5)}$, então a i-ésima obsrvação pode ser considerada influente no modelo.

4.1.2 Estatística de Peña

Como uma alternativa à distância de Cook, Peña (2005) propôs avaliar, para cada observação, a mudança na predição quando cada um dos outros pontos na amostra é excluído. Isto pode ser feito considerando o vetor

$$\mathbf{s}_i = (\hat{y}_i - \hat{y}_{i(1)}, \ldots, \hat{y}_i - \hat{y}_{i(n)})^{\mathrm{T}}. \tag{4.1.6}$$

O vetor em (4.1.6) sugere avaliar quão sensível é a previsão da i-ésima observação à exclusão de cada observação na amostra. A nova estatística foi definida pelo autor como

$$\mathrm{S}_i = \frac{\mathbf{s}_i^{\mathrm{T}} \mathbf{s}_i}{p\widehat{\mathrm{Var}}(\hat{y}_i)} = \frac{\sum_{j=1}^{n} (\hat{y}_i - \hat{y}_{i(j)})^2}{p\widehat{\mathrm{Var}}(\hat{y}_i)}, \tag{4.1.7}$$

em que p é o número de parâmetros da função de regressão e $\widehat{\mathrm{Var}}(\hat{y}_i)$ é a estimativa da variância do i-ésimo valor predito.

Para os modelos de regressão linear normal, Peña mostrou que S_i pode obtida por

$$\mathrm{S}_i = \frac{1}{ps^2 h_{ii}} \sum_{j=1}^{n} \frac{h_{ij}^2 e_j^2}{(1-h_{ii})^2}, \tag{4.1.8}$$

em que $e_j = y_j - \hat{y}_j$ e $s^2 = \mathbf{e}'\mathbf{e}/(n-p)$, para $\mathbf{e} = (e_1, \ldots, e_n)^{\mathrm{T}}$. Além disso, foram provadas três propriedades de (4.1.8). A primeira é que sob a hipótese de dados sem *outliers* e com baixa alavancagem o valor esperado de S_i é $1/p$. A segunda propriedade é que em amostras grandes e com muitos preditores, a distribuição de S_i será aproxidamente normal. E a terceira é que a estatística (4.1.8) consegue descriminar os pontos atípicos dos pontos bons, quando a amostra está contaminada com um grupo de *outliers* semelhantes com alta alavancagem.

Para estender a medida (4.1.7) aos modelos simétricos lineares e não-lineares é preciso obter $\widehat{\mathrm{Var}}(\hat{y}_i)$. Em particular, temos o resultado assintótico

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \mathbf{K}_{\beta\beta}^{-1}), \quad \mathbf{K}_{\beta\beta}^{-1} = \frac{\phi}{4d_g} (\mathbf{D}_\beta^{\mathrm{T}} \mathbf{D}_\beta)^{-1},$$

mantido sob certas condições de regularidade (Cox and Hinkley, 1974). No caso linear, temos que $\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}) = \mathbf{X}\hat{\boldsymbol{\beta}}$ e $\mathbf{D}_\beta = \mathbf{X}$. Portanto, assintoticamente,

$$\text{Var}[\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})] = \text{Var}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{X}\text{Var}(\hat{\boldsymbol{\beta}})\mathbf{X}^{\text{T}} = \frac{\phi}{4d_g}\mathbf{H},$$

com $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}$. Então, a estimativa da variância do i-ésimo valor predito, no caso linear, é dada por $\widehat{\text{Var}}(\hat{y}_i) = \frac{\hat{\phi}}{4d_g}h_{ii}$, em que h_{ii} é o i-ésimo elemento da diagonal de $\mathbf{H}$.

A matriz de variância e covariância dos valores ajustados para regressão não-linear, $\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$, pode ser obtida de forma semelhante, utilizando a expansão linear de Taylor

$$\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\mu}(\boldsymbol{\beta}) + \mathbf{D}_\beta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Então,

$$\begin{aligned} \text{Var}[\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})] &\approx \text{Var}[\mathbf{D}_\beta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \mathbf{D}_\beta\text{Var}(\hat{\boldsymbol{\beta}})\mathbf{D}_\beta^{\text{T}} \\ &\approx \frac{\phi}{4d_g}\mathbf{H}, \quad \text{com} \quad \mathbf{H} = \mathbf{D}_\beta(\mathbf{D}_\beta^{\text{T}}\mathbf{D}_\beta)^{-1}\mathbf{D}_\beta^{\text{T}}. \end{aligned}$$

Portanto, a estimativa da variância do i-ésimo valor predito, no caso não-linear, é dada por $\widehat{\text{Var}}(\hat{y}_i) = \frac{\hat{\phi}}{4d_g}\hat{h}_{ii}$, em que $\hat{h}_{ii}$ é o i-ésimo elemento da diagonal de $\hat{\mathbf{H}} = \mathbf{D}_{\hat{\beta}}(\mathbf{D}_{\hat{\beta}}^{\text{T}}\mathbf{D}_{\hat{\beta}})^{-1}\mathbf{D}_{\hat{\beta}}^{\text{T}}$.

Com base na segunda propriedade Penã (2005) proprôs valores de referência para identificar observações influentes. A i-ésima observação será influente se S_i não satisfaz a condição

$$\text{med}(\mathbf{S}) - 4,5\text{med}(|\mathbf{S} - \text{med}(\mathbf{S})|) \leq \mathbf{S}_i \leq \text{med}(\mathbf{S}) - 4,5\text{med}(|\mathbf{S} - \text{med}(\mathbf{S})|),$$

onde med($\mathbf{S}$) é a mediana dos valores de S_i. Desta forma,

$$\text{med}(\mathbf{S}) \pm 4,5\text{med}(|\mathbf{S} - \text{med}(\mathbf{S})|) \tag{4.1.9}$$

são os valores de referência sugeridos por Peña.

Peña verificou que esta estatística é muito útil para identificar *outliers* com média ou alta alavancagem, mas não é eficiente para detectar *outliers* com baixa alavancagem. Além disso, grupos de *outliers* com baixa alavancagem aumentam a variabilidade dos valores da estatística, mas não os separam das demais observações.

4.1.3 Estatística de Andrews-Pregibon

O método proposto por Andrews e Pregibon (1978) examina um modelo linear na presença de uma ou mais observações atípicas que tem grande influência na estimativa dos parâmetros. Os autores levaram em conta dois métodos de diagnóstico que até então eram analisados separadamente. O primeiro, considera que a exclusão de um caso correspondente à um *outlier* na variável resposta tende a diminuir significativamente a soma dos quadros dos resíduos. O segundo, considera que a

influência de uma linha de $\mathbf{X}$ é, pelo menos em parte, refletida pela mudança no $|\mathbf{X}^{\mathrm{T}}\mathbf{X}|$ quando uma linha é excluída.

Assim, Andrews e Pregibon sugeriram combinar as duas idéias calculando a mudança em $SQR \times |\mathbf{X}^{\mathrm{T}}\mathbf{X}|$ resultante da exclusão de uma ou mais observações, com $SQR = \mathbf{e}'\mathbf{e}$, com $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Para avaliar a mudança relativa devido a exclusão de k observações indexadas por $\mathbf{I} = \{i_1, \ldots, i_k\}$, os autores propuseram a razão

$$R_{\mathbf{I}} = \frac{SQR_{(\mathbf{I})}}{SQR} \times \frac{\left|\mathbf{X}^{\mathrm{T}}_{(\mathbf{I})}\mathbf{X}_{(\mathbf{I})}\right|}{|\mathbf{X}^{\mathrm{T}}\mathbf{X}|}. \tag{4.1.10}$$

Considerando as estimativas de mínimos quadrados, a razão (4.1.10) pode ser reescrita como

$$R_{\mathbf{I}} = \frac{\left|\mathbf{X}^{*\mathrm{T}}_{(\mathbf{I})}\mathbf{X}^{*}_{(\mathbf{I})}\right|}{|\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}|}, \tag{4.1.11}$$

onde $\mathbf{X}^* = (\mathbf{X} : \mathbf{Y})$ é a matriz de variáveis explanatórias aumentada com $\mathbf{Y}$. Portanto, $R_{\mathbf{I}}$ é uma medida adimensional e $1 - \sqrt{R_{\mathbf{I}}}$ corresponde à proporção do volume gerado por $\mathbf{X}^*$ atribuível aos casos indexados por $\mathbf{I}$. Desta forma, pequenos valores de $R_{\mathbf{I}}$ estão associados a observações atípicas e/ou influentes. Portanto, é nessário obter o menor valor observado para $R_{\mathbf{I}}$, $R^0_k = \min \binom{n}{k} R_{\mathbf{I}}$, para $k = 1, \ldots, k_{\max}$, com $k_{\max}$ escolhido pelo analista.

A determinação de $R_{\mathbf{I}}$ para todos os subconjuntos de k casos requer um esforço computacional muito grande. Andrews e Pregibon (1978) discutem estratégias para abordar este problema e sugerem calcular todos os $\binom{m}{k}$ valores de $R_{\mathbf{I}}$ correspondentes as $m = 2k_{\max}$ observações com o menor W^*_{ll}, um elemento da diagonal de $\mathbf{W}^* = \mathbf{I}_n - \mathbf{X}^*(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)^{-1}\mathbf{X}^{*\mathrm{T}}$.

Para evitar que $\mathbf{W}^*$ seja uma matriz singular, é conveniente usar a decomposição $\mathbf{X}^* = \mathbf{Q}^*\mathbf{T}^*$, com $\mathbf{Q}^*$ uma matriz ortogonal e $\mathbf{T}^*$ uma matriz triangular. Assim, $\mathbf{W}^* = \mathbf{I}_n - \mathbf{Q}^*\mathbf{Q}^{*\mathrm{T}}$.

Andrews e Pregibon conseguiram desenvolver uma teoria da distribuição de $R_{\mathbf{I}}$ quando $\boldsymbol{y}$ é normal e $\mathbf{X}$ é fixo. Embora útil apenas para n de tamanho modesto (≤ 30), eles fornecem alguns níveis de significância para encontrar conjuntos de valores discrepantes.

Em amostras com grande número de observações o foco é a detecção sem a construção de níveis de significância. Nestes casos, os autores sugerem uma análise gráfica comparando R^0_k com outros pequenos valores de $R_{\mathbf{I}}$. Sob essa sugestão, é necessário examinar o gráfico de $(\log[R_{\mathbf{I}}/R^0_k], k)$ para $k = 1, \ldots, k_{\max}$ e buscar o menor ponto que estiver isolado no gráfico.

Considerando a sugestão de Andrews e Pregibon, temos

$$I^k_{\mathbf{I}} = \log(R_{(\mathbf{I})}/R^0_k),$$

em que $R_{(\mathbf{I})}$ corresponde aos valores ordenados de $R_{\mathbf{I}}$.Observe que $I^k_{(\mathbf{I})} \geq 0$, sendo igual a 0 quando $R_{(\mathbf{I})} = R^0_k$. Para identificar o conjunto de observações mais discrepante utilizamos a medida influência relativa

$$D^k_{\mathbf{I}} = I^k_{(\mathbf{I}+1)} - I^k_{(\mathbf{I})},$$

em que $I^k_{(\mathbf{I})}$ corresponde aos valores ordenados de $I^k_{\mathbf{I}}$. O primeiro valor de $D^k_{\mathbf{I}}$, D^0_k, é a medida

de influência relativa obtida para R_k^0. Assim, obtemos os valores de D_k^0, para $k = 1, \ldots, k_{\max}$. O mínimo mais extremo dentre todos os mínimos observados é aquele que apresentar o maior valor para D_k^0 e portanto, o subconjunto de observações correspondente à esse valor são observações atípicas e/ou influentes. Neste caso, é razoável deixar de lado as observações correspondentes e re-examinar os pontos restantes.

4.2 Metodologia para identificação de pontos influentes

Após o cálculo das estatísticas de influência, deve-se avaliar a magnitude de suas medidas para caracterizá-las como influentes através de alguma referência específica. Nesta seção, serão discutidos os procedimentos para o cálculo de valores de referência, construídos a partir de simulações de Monte Carlo, e o consequente diagnóstico de influência.

Na literatura, para as estatísticas de influência discutidas anteriormente, os valores de referência foram obtidos sob a hipótese de erros com distribuição normal. Porém, ao estendermos para outros modelos simétricos esses valores de referencia podem ser inadequados.

Como alternativa, nós propomos o uso de limites construídos por meio de simulações de Monte Carlo para investigar a significância das estatísticas de influência. Assim, através das estimativas obtidas com o ajuste de um determinado modelo, para a t-ésima replicação ($t = 1, \ldots, T$) é gerada uma amostra $y_1^{(1)}, \ldots, y_n^{(T)}$, da qual se calculam as estatísticas de influência. Com as medidas das estatísticas de influência obtidas em cada replicação podemos construir bandas de confiança baseadas nos percentis das distribuições das mesmas.

Em resumo, considerando T replicações, os limites simulados, para a estatística de Peña ou para distância de Cook generalizada, são calculados da seguinte maneira:

1. Estime $\boldsymbol{\theta}$ considerando as observações $y_1, \ldots, y_n$ e o correspondente modelo (linear ou não-linear) com erros simétricos;

2. Para $t = 1, \ldots, T$:
   (i) Simule $y_1^{(t)}, \ldots, y_n^{(T)}$ considerando $\hat{\boldsymbol{\theta}}$ e $\mathbf{X}$;
   (ii) Calcule (4.1.7) e (4.1.3) para $i = 1, \ldots, n$ a partir dos dados em (i) e salve os respectivos resultados na t-ésima linha das matrizes $\mathbf{S}_{T\times n}$ e $\mathbf{D}_{T\times n}$;

3. Calcule $S_{\max}^{(t)} = \max_i\{S_{t1}, \ldots, S_{tn}\}$ e $D_{\max}^{(t)} = \max_i\{D_{t1}, \ldots, D_{tn}\}$ a partir das estatísticas em (ii), para $t = 1, \ldots, T$.

4. Para cada estatística, calcule M_1, o percentil 95% de $\{S_{\max}^{(1)}, \ldots, S_{\max}^{(T)}\}$ e de $\{D_{\max}^{(1)}, \ldots, D_{\max}^{(T)}\}$;

5. Para cada estatística, calcule M_{2_i}, o percentil 95% de $\{S_{1i}, \ldots, S_{Ti}\}$ e de $\{D_{ti}, \ldots, D_{Ti}\}$, para $i = 1, \ldots, n$.

Nos passos 4 e 5 temos duas metodologias para detecção de observações influentes: M_1 que considera um nivel de confiança para os valores extremos das estatísticas nas amostras e $M_2 = \{M_{2_i} : i = 1, \ldots, n\}$ que considera um nível de confiança para os valores das estatísticas em cada observação. No passo 2.(i), avaliamos a qualidade de cada metodologia considerando $\mathbf{X}$ fixo

nos valores da amostra e $\mathbf{X}$ variável, gerado a partir de uma distribuição conhecida. Usamos o subescrito f apenas para nos referirmos aos limites simulados que foram calculados considerando $\mathbf{X}$ fixo, a saber, M_{1_f} e M_{2_f}, para as respectivas metodologias.

Capítulo 5

Aplicações

Neste capítulo ilustramos o uso das estatísticas de influência e as técnicas de limiares descritas no Capítulo 4 através de exemplos referentes a modelos lineares e não-lineares.

Primeiro apresentaremos exemplos considerando modelos de regressão linear simples com erros simétricos. Nesses exemplos temos aplicações com dados simulados e dados reais. Para aplicação com dados simulados, apresentaremos alguns resultados para as estatísticas de influência calculadas com os dados do estudo de simulação. Em seguida avaliamos a qualidade das metodologias de influência em resultados particulares do estudo de simulação. Por último, consideramos a aplicação em dois conjuntos de dados reais.

Por fim, apresentamos exemplos com modelos não-lineares. Neste caso, também consideramos alguns resultados do estudo de simulação e três exemplos com dados reais. Os ajustes e os diagnósticos de influência foram obtidos usando a estimação por máxima verossimilhança com *scoring* de Fisher.

5.1 Modelos lineares

5.1.1 Simulação

Resultados das estatísticas de influência para o estudo de simulação

Considerando o estudo de simulação apresentado na seção 3.1, para cada uma das 500 replicações calculamos a estatística de Peña e a distâcia de Cook generalizada. Dos ajustes que foram obtidos no estudo de simulação, apresentamos nesta seção o resultado da análise de influência para os ajustes dos modelos lineares sob erros logísticos tipo II.

A Figura 5.1 mostra os gráficos boxplot para estatística de Peña e distância de Cook generalizada considerando os resultados das 500 replicações para Situação (1), dados sem observações influentes. Observamos que o valor mediano de Si é aproximadamente o mesmo para todas as observações e para a maioria dos dados os valores da estatística estão concentrados entre de 0,2 e 1. Com relação à distância de Cook generalizada, observamos que há uma diferença de variabilidade de $D(\hat{\boldsymbol{\beta}}_{(i)})$ entre as observações, de forma que para algumas observações essa estatística apresenta maior variabilidade do que outras, mas com valores entre 0 e 0,2. Para $D(\hat{\phi}_{(i)})$ observamos que os

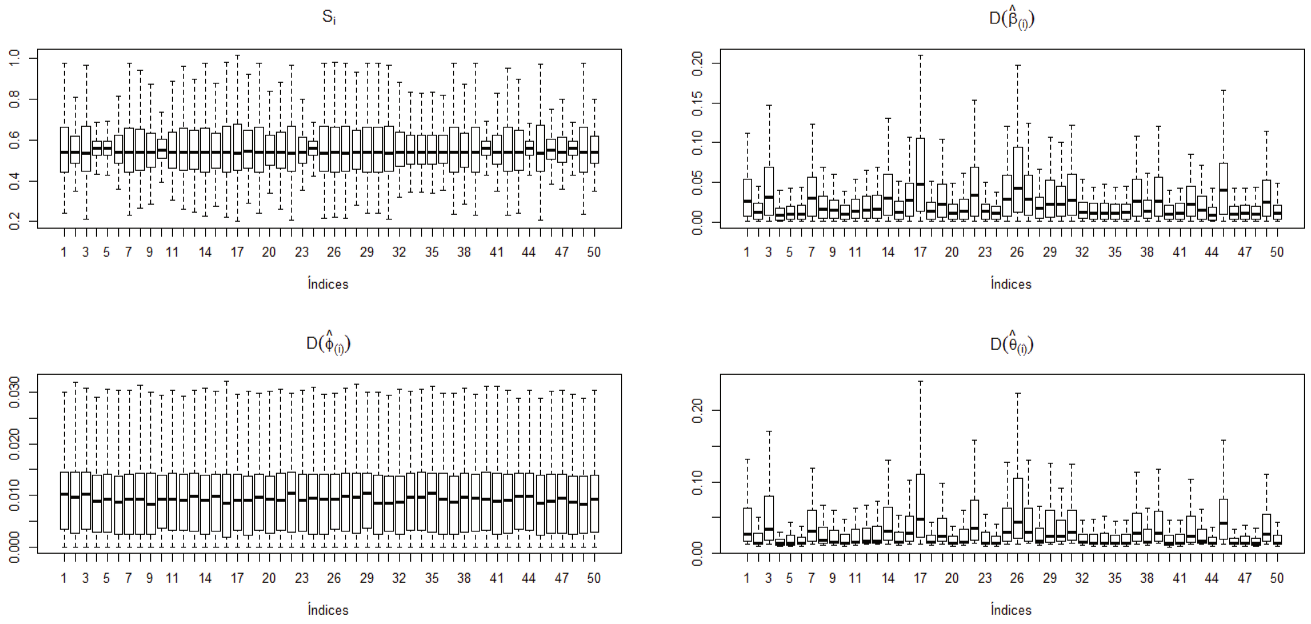


Figura 5.1: *Estatística de influência para o ajuste dos dados simulados da Situação (1).*

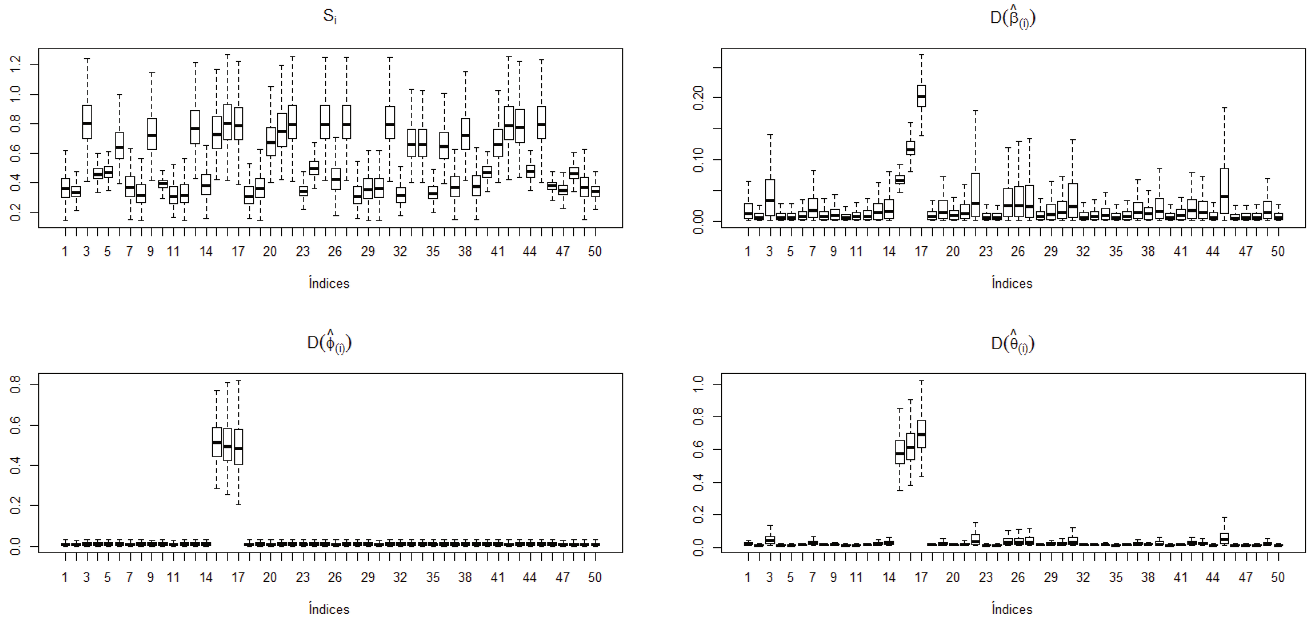


Figura 5.2: *Estatística de influência para o ajuste dos dados simulados da Situação (2).*

valores são bem pequenos e a variabilidade é a mesma entre as observações.

A Figura 5.2 mostra os gráficos das estatística para Situação (2). Observamos que a presença de *outliers* causou uma variação dos valores da estatística de Peña entre as observações, mas nenhuma observação se destaca das demais. Em $D(\hat{\boldsymbol{\beta}}_{(i)})$, observamos que o caso 17 se destaca mais "mascarando"o efeito das observações 15 e 16 que se destacam de forma moderada das demais

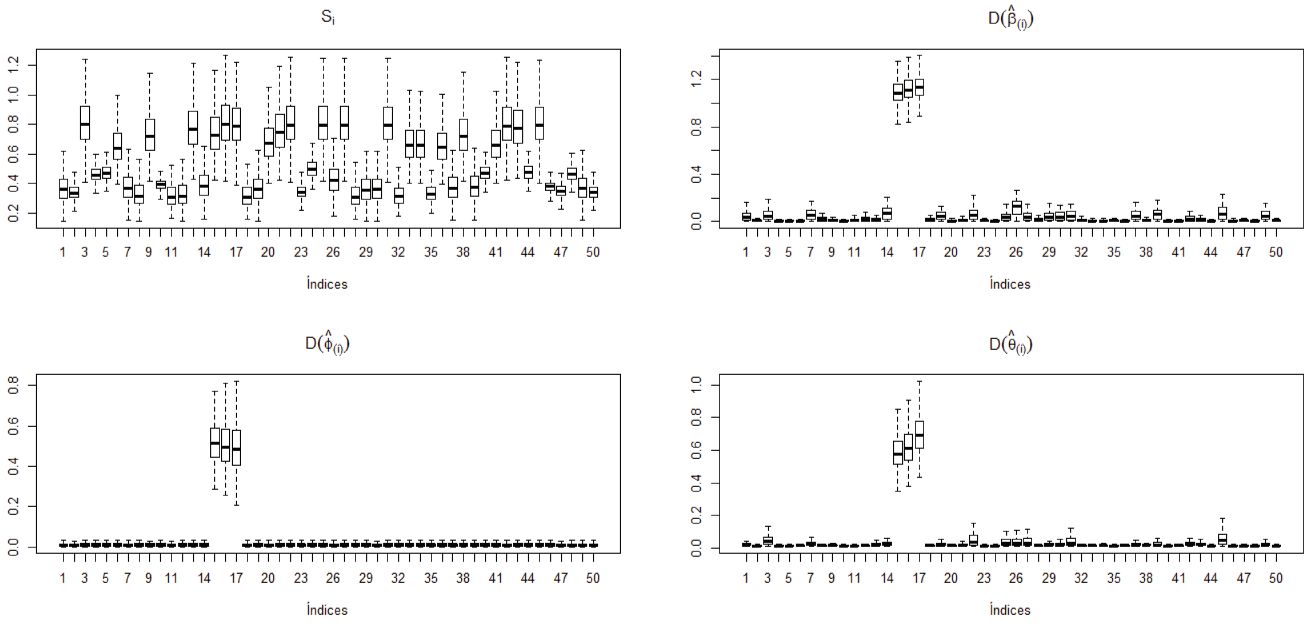


Figura 5.3: *Estatística de influência para o ajuste dos dados simulados da Situação (3).*

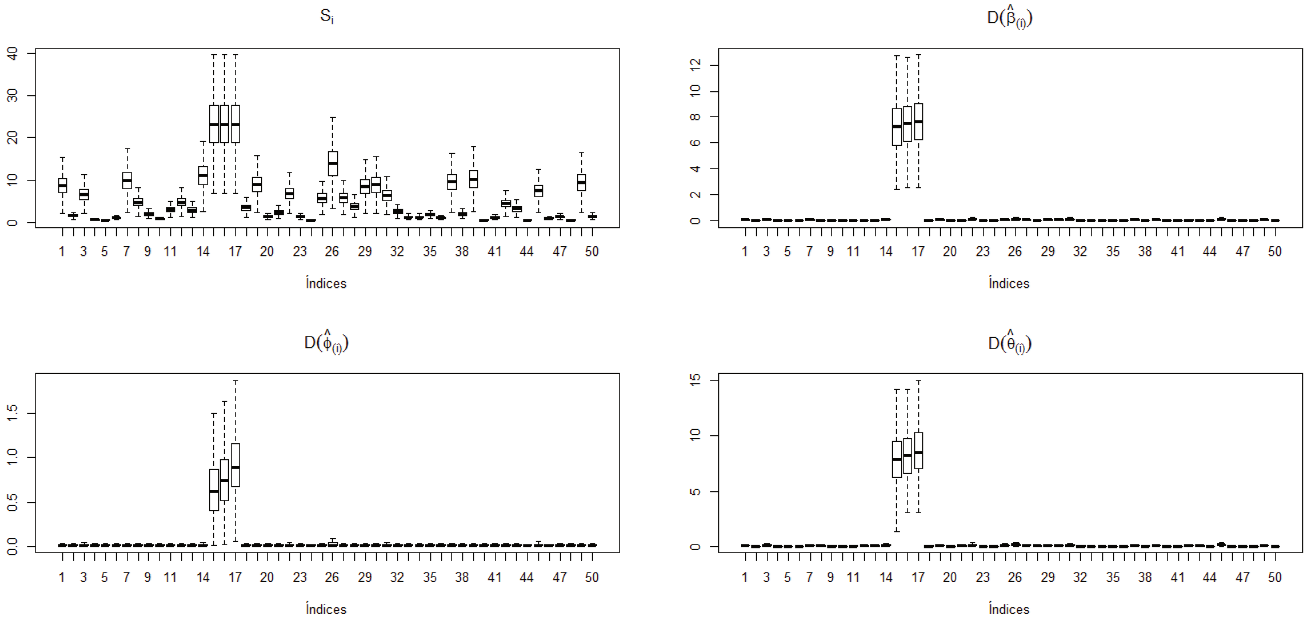


Figura 5.4: *Estatística de influência para o ajuste dos dados simulados da Situação (4).*

observações. Em $D(\hat{\phi}_{(i)})$, veirficamos que os três *outliers* tem forte influência sobre a estimativa de ϕ, o boxplot dos três pontos atípicos se sobrepõem aos demais. O mesmo comentário se aplica aos resultados de $D(\hat{\boldsymbol{\theta}}_{(i)})$. Portanto, a distância de Cook generalizada detecta os pontos discrepantes, mas a estatística de Peña não.

A análise de influência para os dados gerados na Situação (3) estão na Figura 5.3. Observamos

que a inserção de *ouliers* com média alavancagem causou uma grande variação nos valores de S_i e os boxplots correspondentes às observações discrepante se distanciam muito pouco dos demais. Para distância de Cook generalizada verificamos claramente a separação das observações influentes dos pontos restantes.

A Figura 5.4 mostra os resultados da análise de influência para a Situação (4). Para estatística S_i os boxplots das obsevações influentes estão afastados dos demais, portanto a estatística de Peña é capaz de identificar os pontos atípicos. Para distância de Cook generalizada é ainda mais perceptível separação dos boxplots correspondentes às observações influentes.

Um caso particular do estudo de simulação

Para avaliar a qualidade das metodologias para detecção observações influentes, ilustramos um exemplo do estudo de simulação em que o conjunto de dados foi gerado considerando um modelo linear sob a presença de três observações atípicas com alta alavancagem, Situação (4). Lembrando que, para obter os dados, utilizamos o modelo

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, 50, \tag{5.1.1}$$

em que $\beta_0 = 110$, $\beta_1 = -4$, x_i gerado a partir da distribuição $N(13, 4)$. Para este exemplo, consideramos o caso em que $\epsilon_i \sim \text{LogisII}(\mu = 0, \phi = 90)$ e as observações $i = 15, 16, 17$ foram substituindas por $y_i^* = y_i + 5\sigma_{y_i}$ e $x^* = 45$. Aos dados foram ajustados modelos sob erros normais, logístico-II e t-Student.

Tabela 5.1: *Estimativas de máxima verossimilhança (erro-padrão) para alguns modelos simétricos ajustados aos dados simulados.* $\boldsymbol{\theta} = (\beta_0 = 110, \beta_1 = -4, \phi = 90)^{\text{T}}$.

Distribuição	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$L(\hat{\boldsymbol{\theta}})$	**AIC**
Normal	20,999 (8,060)	2,185 (0,444)	678,666 (135,733)	-233,950	471,9003
Logis II	27,947 (8,030)	1,659 (0,443)	224,538 (53,109)	-234,965	473,9302
t-Student(2)	91,538 (4,125)	-2,887 (0,227)	106,634 (33,721)	-221,786	447,5727

A Tabela 5.1 mostra as estimativas de máxima verossimilhança obtidas em cada ajuste. No ajuste com o modelo normal as estimativas foram fortemente influenciadas pelas observações atípicas com alta alavancagem. No modelo logístico-II as estimativas foram melhores do que o modelo normal com erro-padrão menor, mesmo assim, as estimativas estão bem afastadas dos valores verdadeiros. No ajuste sob erros t-Student, o maior valor para verossimilhança foi obtido com $\nu = 2$, o que configura uma incerteza a respeito da variância dos erros, ainda que as estimativas obtidas estejam mais próximas dos valores reais.

A Figura 5.5 mostra o gráfico dos dados gerados em a, assim como o ajuste do modelo normal, logístico-II e t-Student em b-d, respectivamente. Em destaque temos os pontos inseridos como *outliers* de alta alavancagem. Os gráficos em b e c mostram que os ajustes com modelo normal e com o logístico-II não estão em conformidade com configuração da maioria dos dados, verificamos

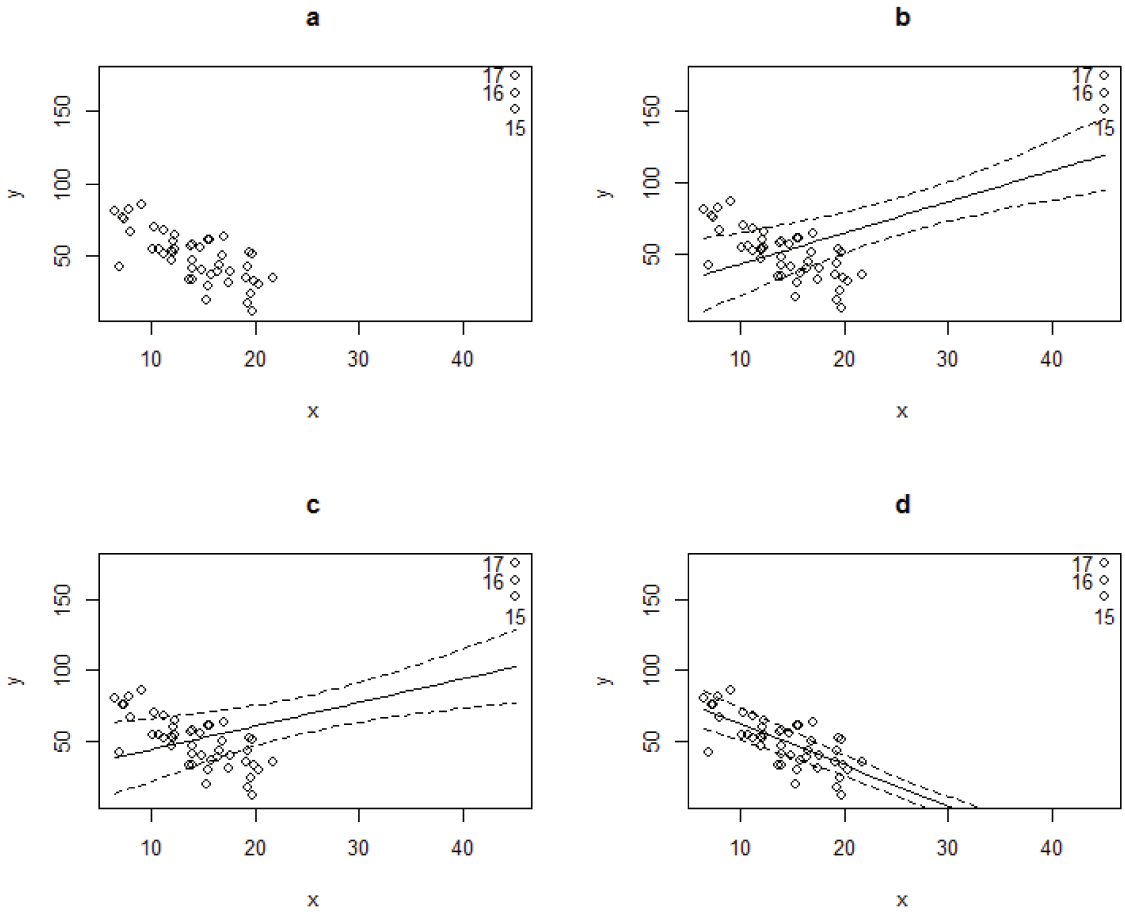


Figura 5.5: *Em a gráfico dos dados simulados e em b-d gráfico dos dados com o ajuste do modelo normal, Logístico-II e t-Student, respectivamente.*

que as observações atípicas são altamente influentes para os respectivos modelos. Entretanto, no ajuste do modelo t-Student, em d, o efeito das observações atípicas foi reduzido de forma que o ajuste segue a configuração da maioria dos dados.

Para cada ajuste foram calculadas a estatística de Peña e a distância de Cook generalizada. Apresentamos os resultados nas Figuras a seguir. As linhas tracejadas e contínuas em vermelho são os limites simulados M_{1_f} e M_{2_f}, respectivamente. As linhas tracejadas e contínuas em verde são os limites simulados M_1 e M_2, respectivamente, com $\mathbf{X}$ gerado a partir da distribuição normal e a linha contínua em azul limites simulados M_1 e M_2 com $\hat{\boldsymbol{\theta}}$ obtido sem as observações influentes. A linha contínua em preto é o limite de Peña.

A Figura 5.6 mostra os resultados das estatísticas de influência para o ajuste do modelo (5.1.1) sob erros normais. Observamos que as linhas em azul coincidem com as linhas em verde, portanto os limites simulados considerando as estimativas originais e os limites simulados considerando as estimativas sem as observações influentes são similares, com isso, podemos dizer que ao gerarmos limites com $\mathbf{X}$ variável estamos retomando a escala de valores das estatísticas sem observações

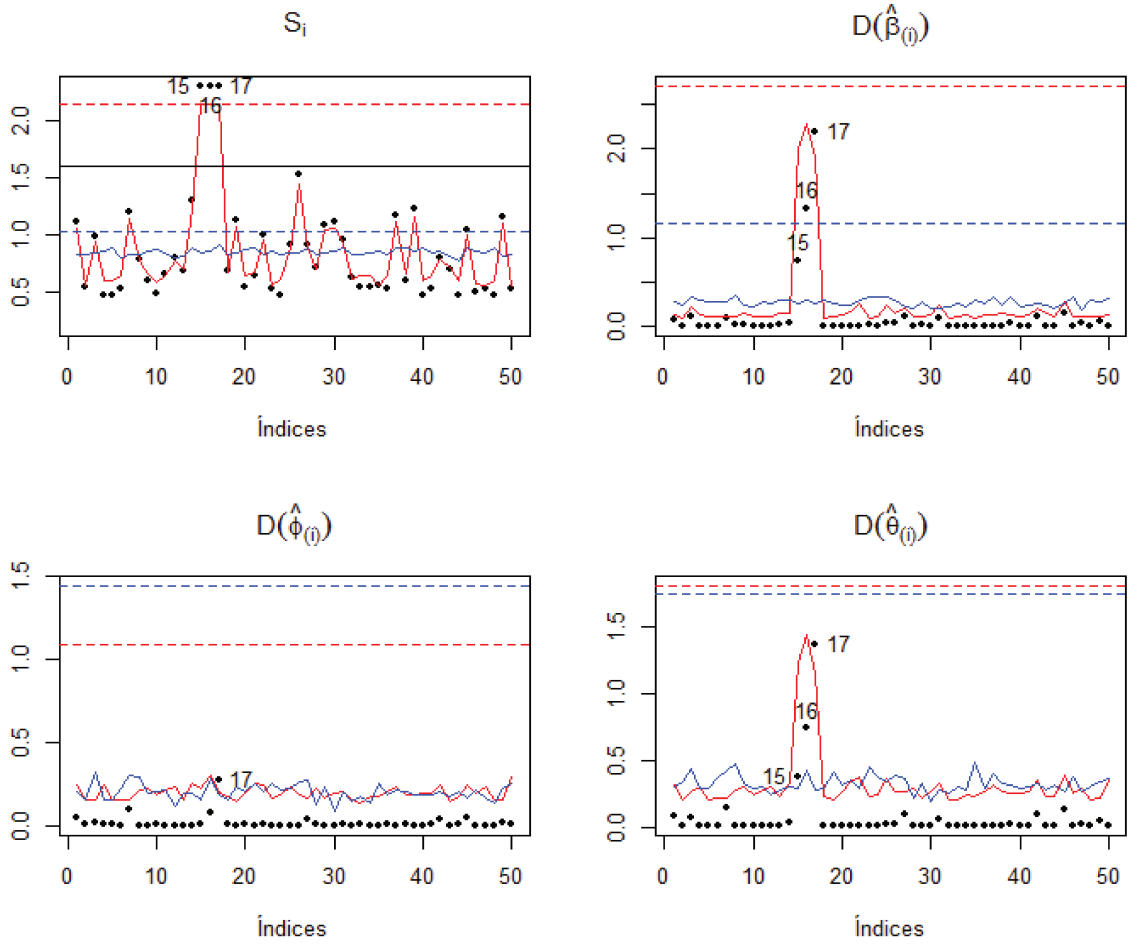


Figura 5.6: *Estatísticas de influência para o ajuste do modelo (5.1.1) sob erros normais.*

influentes. Enquanto, que os limites simulados M_{1_f} e M_{2_f} são influenciados pelas observações com alta alavancagem, pois os limites são maiores para esses pontos. Devido à grande variabilidade da estatística de Peña sob o ajuste do modelo normal, os limites simulados detectam mais observações do que deveriam, enquanto que os limiares de Peña detectam exatamente as observações influentes. Para a distância de Cook generalizada o limite simulado M_2 foi melhor do que o limite M_1, pois conseguiu identificar as observações geradas como atípicas.

Na Figura 5.7 estão os resultados das estatísticas de influência para o ajuste do modelo (5.1.1) sob erros logísticos tipo II. Novamente observamos que os limites simulados M_1 e M_2 estão convertidos a uma escala de valores das estatísticas sem observações influentes, pois as linhas azuis coincidem com as linhas verdes. Os limites simulados M_{1_f} e M_{2_f}, indicados pelas linhas vermelhas, são influenciados pela variável explanatória, exceto em $D(\hat{\phi}_{(i)})$, em que os limites de ambos os tipos são similares. A estatística de Peña apresentou uma variabilidade muito grande e os limites simulados são menores e detecam observações além daquelas geradas como atípicas. Somente o limite de Peña detectou exatamente os pontos gerados como atípicos. Entretanto, para a distância de Cook generalizada todos os limites simulados conseguem detectar as observações influentes.

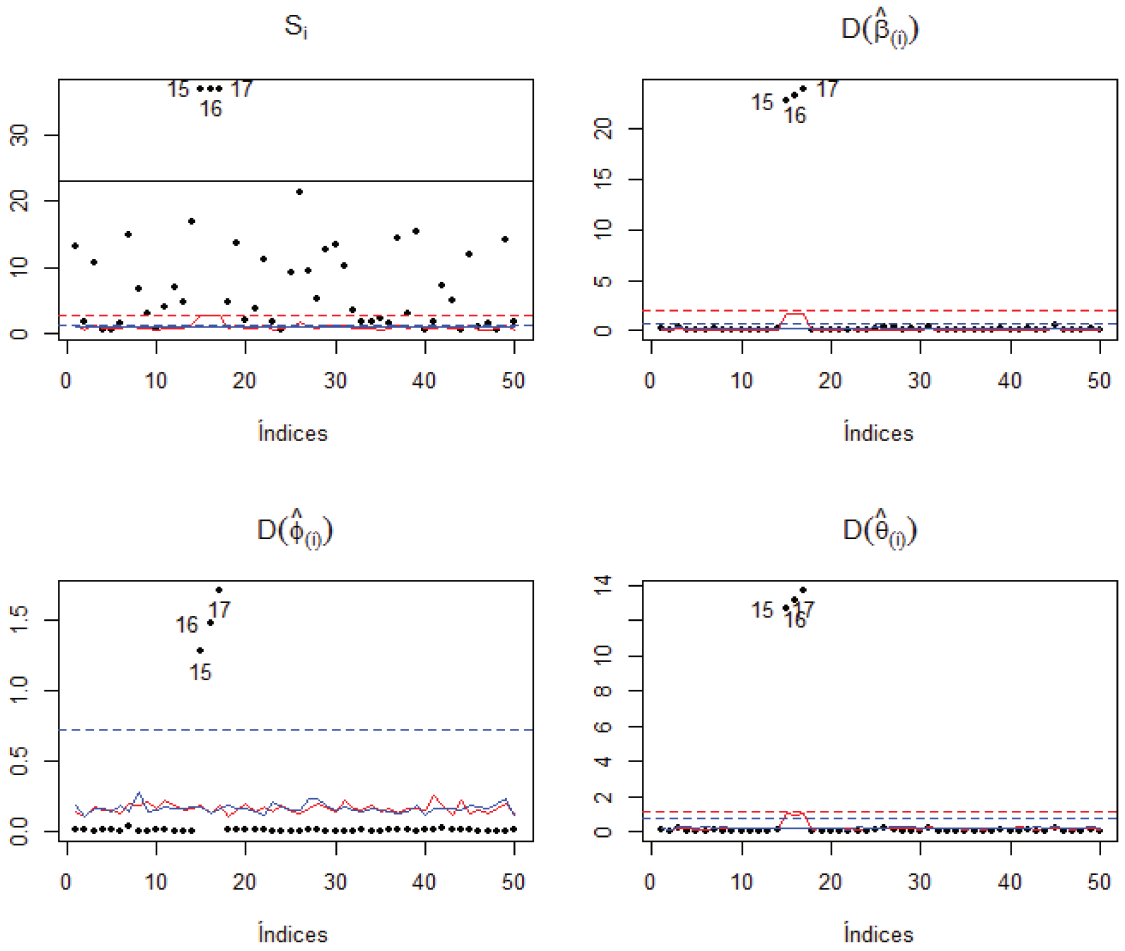


Figura 5.7: *Estatísticas de influência para o ajuste do modelo (5.1.1) sob erros logísticos tipo II.*

Para o ajuste do modelo t-Student, mostramos a análise de influência na Figura 5.8. Verificamos que os limites simulados M_1 e M_2 considerando as estimativas originais diferem dos limites M_1 e M_2 considerando as estimativas obtidas sem as observações influentes, isto, porque nas estimativas originais o grau de liberdade é igual a 2 e nas estimativas sem observações influentes o grau de liberdade aumentou para 30. Para o cálculo de todos os limites simulados o grau de liberdade obtido no ajuste foi fixado em todas as simulações, uma vez que este parâmetro é superestimado sob a suposição de dados sem observações influentes. Para a estatística de Peña os limites simulados M_{1_f} e M_{2_f} não identificaram observações influentes, o limite M_1 (linha verde tracejada) identica observações além daquelas geradas como atípicas, o limite M_2 (linha verde contínua) foi menor do que M_1, identificando mais observações, mesmo as que não são influentes, e o limite de Peña identificou as observações (15, 16, 17, 26). Para a distância de Cook generalizada nenhuma observação foi identificada pelos limites simulados M_1 e M_{1_f} não detectam observações influentes.

A Tabela 5.2 mostra os resultados para estatística de Andrews-Pregibon aplicada aos dados simulados. Para esse conjunto de dados escolhemos $k_{\max} = 4$, assim, obtemos os valores de $R_{\mathbf{I}}$ para o subconjunto com $m = 8$ observações correspondentes aos menores elementos da diagonal de

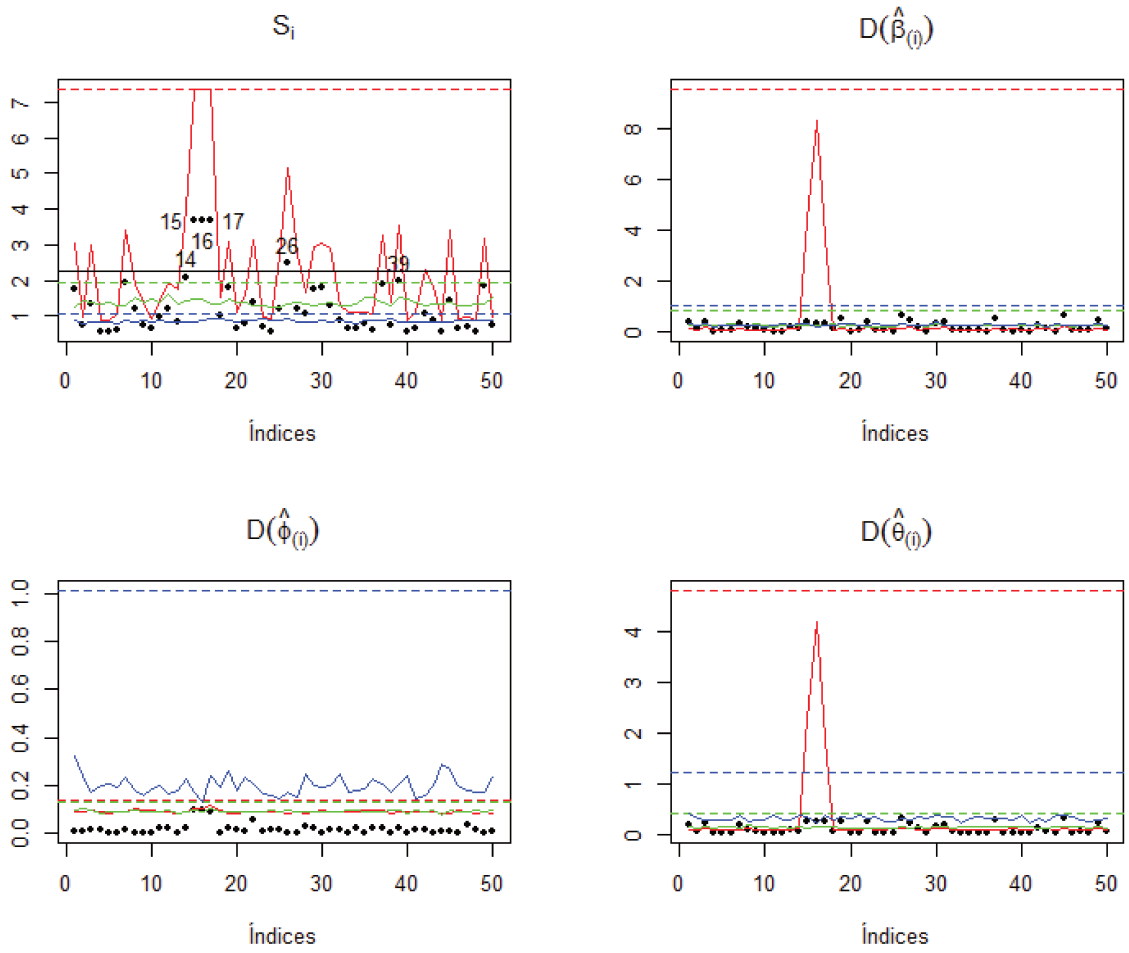


Figura 5.8: *Estatísticas de influência para o ajuste do modelo (5.1.1) sob erros t-Student.*

$\mathbf{W}^*$. Além disso, temos os valores de D_k^0, uma medida relativa das menores razões, para detectar o conjunto de observações mais desviante. O valor de D_k^0 foi maior quando $k = 3$, ou seja, nos ajustes considerados, o conjunto de observações mais desviante se refere aos casos 15, 16 e 17. Verificamos, também, que para o ajuste com o modelo t-Student os valores de R_k^0 foram menores e os valores D_k^0 foram maiores, pois a SQR é maior para esse modelo do que para os outros e portanto a mudança em SQR após a exclusão das observações 15, 16 e 17 é maior.

A Tabela 5.3 mostra a mudança percentual após a exclusão das observações 15, 16 e 17. Observamos que no ajuste dos modelos normal e logístico tipo II a mundança percentual para as estimativas de $\boldsymbol{\beta}$ superou 100% e para as estimativas de ϕ foi aproximadamente de 79% . Entretanto, para o ajuste do modelo t-Student as mudanças foram muito menores.

Para este exemplo, verificamos que os ajustes com o modelo normal e com logístico tipo II foram os mais influenciados pelos pontos gerados como atípicos. Nestes ajustes, Os limites simulados, M_1 e M_2, retomam a escala de valores das estatísticas de influência sem observações influentes. Entretanto nos limites M_{1_f} e M_{2_f}, as observações discrepantes na variável explanatória afetam os resultados de forma que esses limites falham, principalmente, no ajuste do modelo normal.

Tabela 5.2: *Estatística de Andrews-Pregibon para os dados simulados.*

Distribuição	k	R_k^0	D_k^0	Observações
Normal	1	0,646754	0,053735	17
	2	0,330378	0,079540	16, 17
	3	0,042279	1,941147	15, 16, 17
	4	0,037038	0,002713	7, 15, 16, 17
Logis II	1	0,855363	0,025890	17
	2	0,464577	0,099604	16, 17
	3	0,041082	2,305507	15, 16, 17
	4	0,036006	0,003473	7, 15, 16, 17
t-Student	1	0,493547	0,059863	17
	2	0,168211	0,094048	16, 17
	3	0,011234	2,601028	15, 16, 17
	4	0,009866	0,005159	7, 15, 16, 17

Tabela 5.3: *Mudanças (em percentuais) nas estimativas após exclusão das observações (15,16,17).*

Distribuição	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$
Normal	339,04	236,94	78,84
Logis II	236,88	287,12	78,73
t-Student(2)	4,75	-11,06	25,40

Observamos também que para esses dados, a estatística de Peña apresentou uma alta variabilidade e os limites simulados M_1 e M_2 são muito menores do que a faixa de valores obtida para a estatística. Somente os limiares de Peña foram capazes identificar os pontos de alavanca no ajuste com modelo normal e logístico tipo II. O ajuste obtido com o modelo t-Student com 2 graus de liberdade reduziu o efeito dos pontos atípicos. Neste caso é esperado que menos pontos sejam detectados pelas estatísticas de influência. Entretanto, a estatística de Peña apresentou uma alta variabilidade e ainda ocorre uma separação dos pontos gerados como atípicos dos demais pontos da amostra e o limite simulado M_1 falha detectando pontos além daqueles gerados como influentes. Para distância de Cook generalizada o limite M_1 não detectou nenhum ponto influente.

Diante dos resultados com dados simulados, na aplicação com dados reais consideramos o limites simulados M_1 (linha verde tracejada) e M_2 (linha verde contínua) como metodologia para detecção de observações influentes.

5.1.2 Dados HRD

Este conjunto de dados corresponde à informações sobre a constelação CYG OB1, que consiste de 47 estrelas na direção de *Cygnus*. A base de dados contém duas variáveis, sendo x o logaritmo da temperatura efetiva na superfície da estrela e y o logaritmo a intensidade da luz. Estes dados foram apresentados por Rousseeuw and Leroy (1987), sendo analisados por diversos autores inclusive por

Peña (2005) como um interessante problema de "mascaramento".

O seguinte modelo foi ajustado aos dados

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, 47, \tag{5.1.2}$$

Tabela 5.4: *Estimativas de máxima verossimilhança (erro padrão aproximado) para alguns modelos simétricos lineares ajustados aos dados HRD.*

Distribuição	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$	$L(\hat{\boldsymbol{\theta}})$	**AIC**
Normal	6,793 (1,209)	-0,413 (0,280)	0,305 (0,063)	-38,804	81,608
Logística II	6,873 (1,262)	-0,426 (0,292)	0,110 (0,027)	-40,678	85,355
Exp. potência(-0,3)	6,830 (1,114)	-0,430 (0,258)	0,465 (0,080)	-36,817	77,712

Tabela 5.5: *Estatística de Andrews-Pregibon para os dados HRD.*

Distribuição	k	R_k^0	D_k^0	**Observações**
Normal	1	0,744442	0,031159	34
	2	0,513321	0,036222	30, 34
	3	0,301440	0,043258	20, 30, 34
	4	0,103231	0,867752	11, 20, 30, 34
	5	0,056798	0,403001	7, 11, 20, 30, 34
	6	0,041240	0,196715	7, 11, 14, 20, 30, 34
Logístico II	1	0,744359	0,032447	34
	2	0,528073	0,041907	30, 34
	3	0,326984	0,056996	20, 30, 34
	4	0,104739	0,966270	11, 20, 30, 34
	5	0,056770	0,423359	7, 11, 20, 30, 34
	6	0,041389	0,193096	7, 11, 14, 20, 30, 34
Exp, potência	1	0,744442	0,031694	34
	2	0,515382	0,037041	30, 34
	3	0,307478	0,046731	20, 30, 34
	4	0,104130	0,878874	11, 20, 30, 34
	5	0,057033	0,404290	7, 11, 20, 30, 34
	6	0,041735	0,192292	7, 11, 14, 20, 30, 34

Na Tabela 5.4 temos as estimativas de máxima verossimilhança dos parâmetros obtidos com ajuste do modelo sob erros normais, logísticos tipo II, t-Student e exponencial potência. De modo geral, as estimativas dos parâmetros de regressão são similares e indicam o ajuste de uma reta decrescente. Mas, O ajuste do modelo exponencial potência apresentou o menor erro padrão para

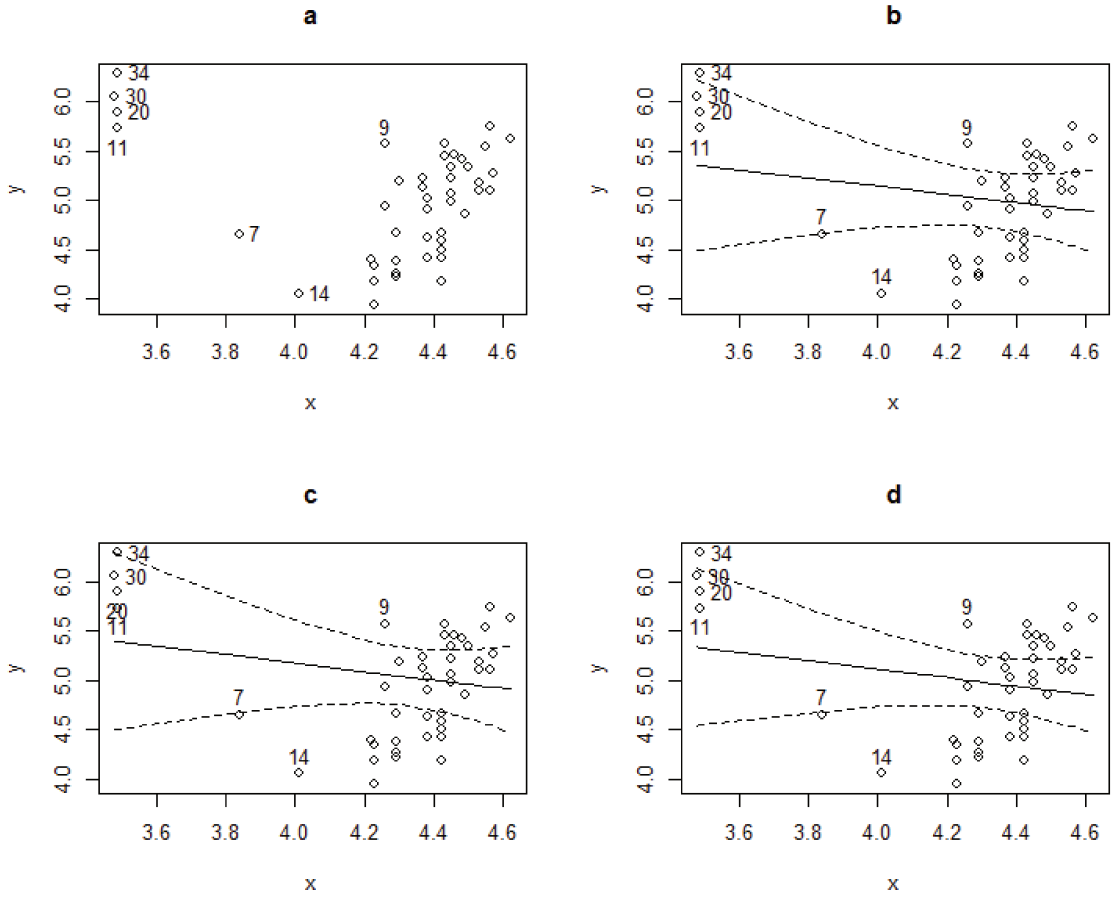


Figura 5.9: *Em (a) gráfico dos dados HRD e em (b) - (d) gráfico dos dados com o ajuste do modelo normal, Logístico-II e exponencial potência, respectivamente.*

as estimativas, com o maior valor para a verossimilhança. Além disso verificamos que o valor obtido para o parâmetro de forma ($\nu = -0,3$) indica uma distribuição platicúrtica para os erros.

Na Figura 5.9a observamos que quatro pontos dos dados (11, 20, 30, 34) são claramente observações atípicas e outras duas observações (7,14) parecem estar longe da linha de regressão principal. Os resultados mostram que os pontos atípicos, destacados no gráfico dos dados, são potencialmente influentes em todos os ajustes indicando uma relação contrário àquela que é apontada pela maioria dos dados, como pode ser visto na Figura 5.9b-d.

Na Figura 5.10 os gráficos normais de probabilidades com envelope para os resíduos padronizados evidenciam falta ajuste para os três modelos considerados, uma vez que algumas observações estam foram do envelope.

O gráfico da estatística de Peña e da distância de Cook generalizada para o ajuste considerando modelo normal é apresentado na Figura 5.11. Para S_i os limites de Peña identificam corretamente 5 casos, mas destacam 6 observações a mais (3, 9, 17, 19, 29 e 35), pelos limites simulados, M_1 e M_2, foram identificadas as observações potencialmente influentes (7, 11, 14, 20, 30 e 34). Para

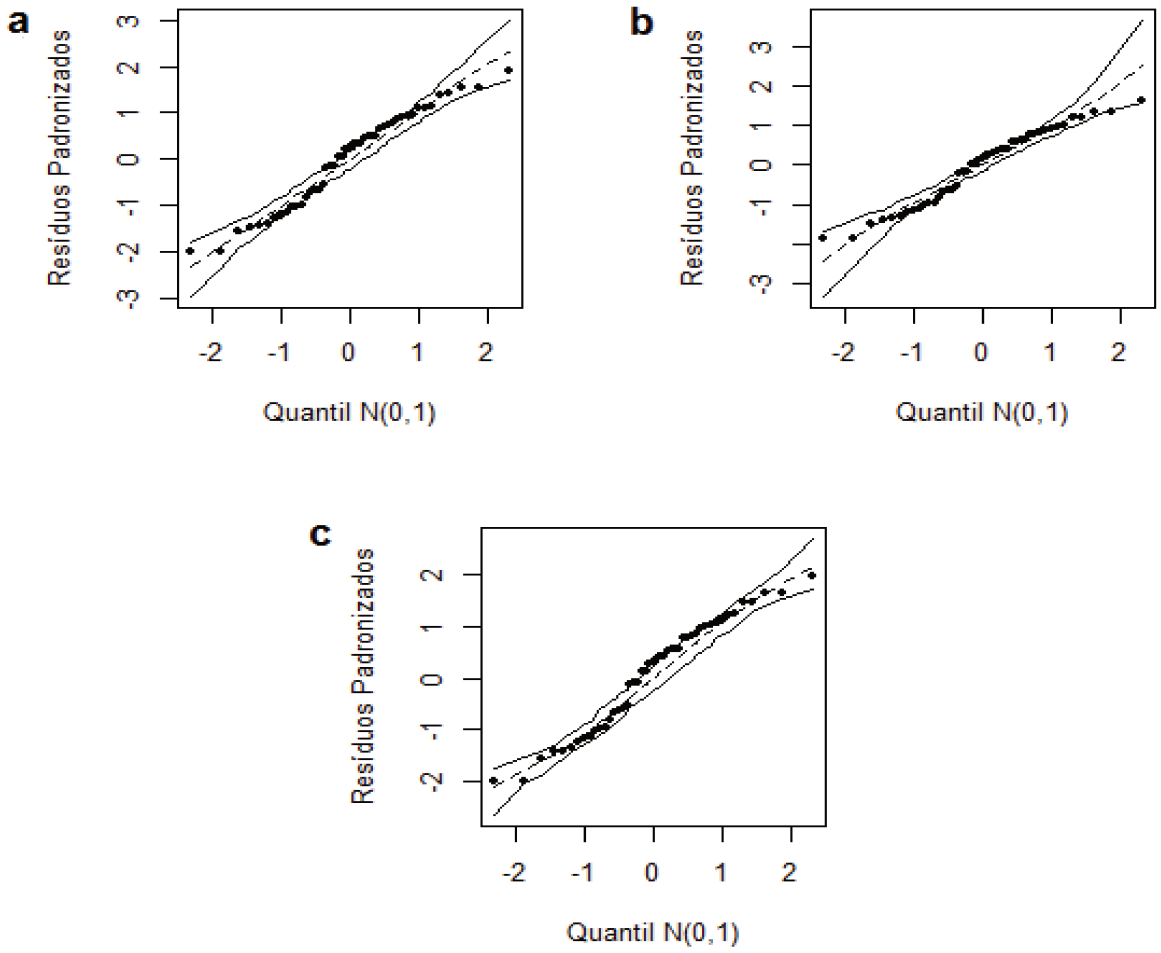


Figura 5.10: *Gráficos normais de probabilidades com envelope para os resíduos padronizados para o ajuste do modelo normal (a), logístico-II (b) e exponencial potência (c) ajustados aos dados HRD.*

distância de Cook generalizada o limite simulado M_1 não detectou nenhuma observação, enquanto que o limite simulado M_2 detectou as (20, 30, 34) como influentes para as estimativas de $\boldsymbol{\beta}$, as observações (14, 17) para a estimativa de ϕ e as observações (14, 30, 34) para as estimativas de $\boldsymbol{\beta}$ e ϕ, simultaneamente.

Na Figura 5.12 temos o resultado das medidas de influência para o ajuste do modelo logístico tipo II. Verificamos que para S_i o limite simulado M_1 está bem próximo do limite de Peña e portanto, outras observações são identificadas além daquelas que são potencialmente influentes (11,20,30,34). Para distância de Cook generalizada o limite simulado M_1 identificou 3 observações influentes (20, 30 e 34) para as estimativas de $\boldsymbol{\beta}$ e duas observações para estimativas de $\boldsymbol{\beta}$ e ϕ, simultaneamente. Enquanto, que o limite simulado M_2 identificou as observações (7,11,14,20,30,34) como influentes para as estimativas de $\boldsymbol{\beta}$ e as observações (11,20,30,34) para as estimativas de $\boldsymbol{\beta}$ e ϕ, simultaneamente.

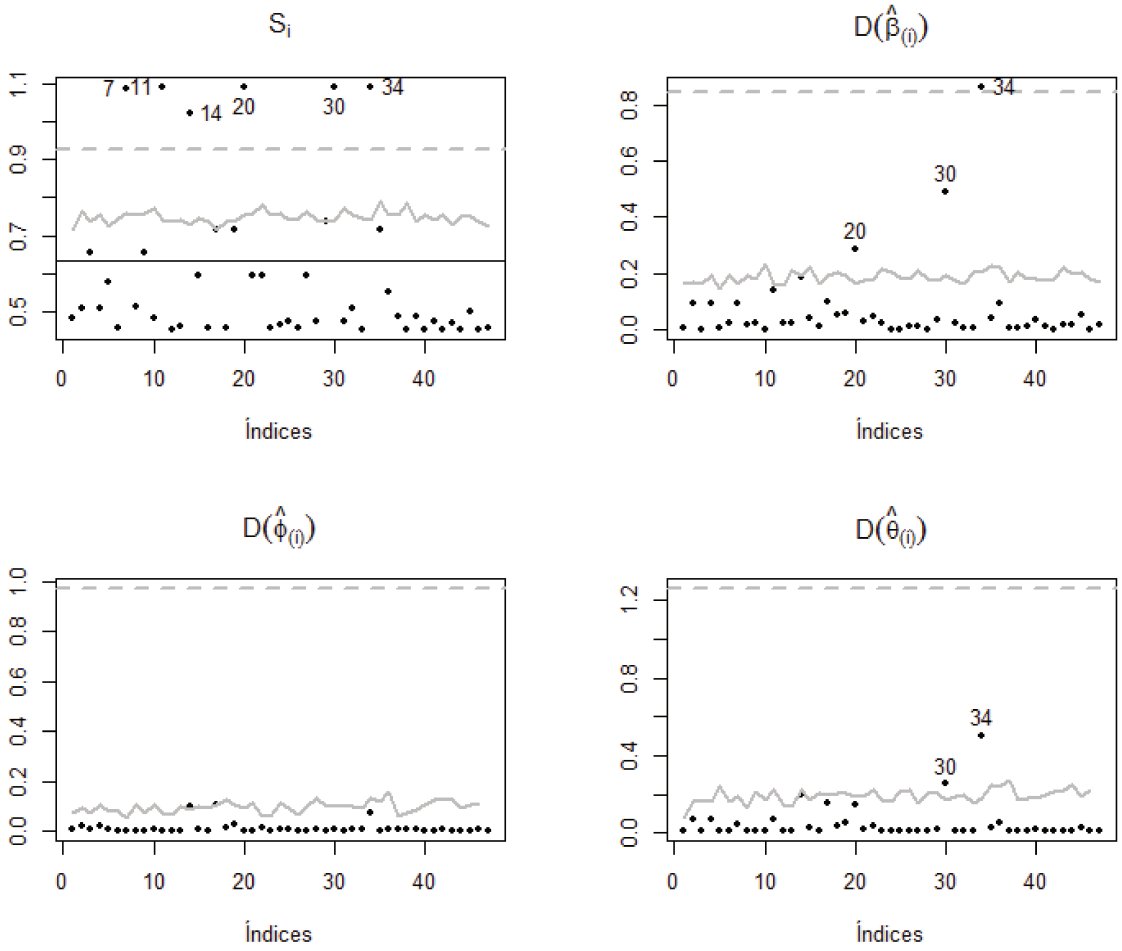


Figura 5.11: *Análise de influência para os dados HRD sob erros normais. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.*

O diagnóstico de influência para o ajuste do modelo exponencial potência está na Figura 5.13. Os valores de S_i são menores para esse ajuste e os limites simulados não detectam as observações potencialmente influentes (11, 20, 30, 34), mas o limite de Peña detecta outras observações além daquelas que são realmente influentes. Para distância de Cook generalizada, o limite simulado M_1 não detectou as observações. Mas, o limite simulado M_2 identificou a observação 34 em $D(\hat{\boldsymbol{\beta}}_{(i)})$, (14, 17) em $D(\hat{\phi}_{(i)})$ e as observações (14, 17, 34) em $D(\hat{\boldsymbol{\theta}}_{(i)})$.

Para estatística de Andrews-Pregibon escolhemos $k_{\max} = 6$, os resultados estão na Tabela 5.5. Observamos as razões mínimas pouco mudaram de um ajuste para outro. Vemos que a medida relativa de R_k para este exemplo foi maior quando $k = 4$, indicando que as observações 11, 20, 30 e 34 são *outliers* pontencialmente influentes para as estimativas dos modelos simétricos considerados.

As mudanças percentuais após omissão de algumas observações estão na Tabela 5.6. Para os três ajustes considerados as mudanças percentuais nas estimativas de $\boldsymbol{\beta}$ ultrapassam 100% e nas estimativas de ϕ chegam a ser superior a 50%. A eliminação das observações 7, 11, 14, 20, 30 e 34

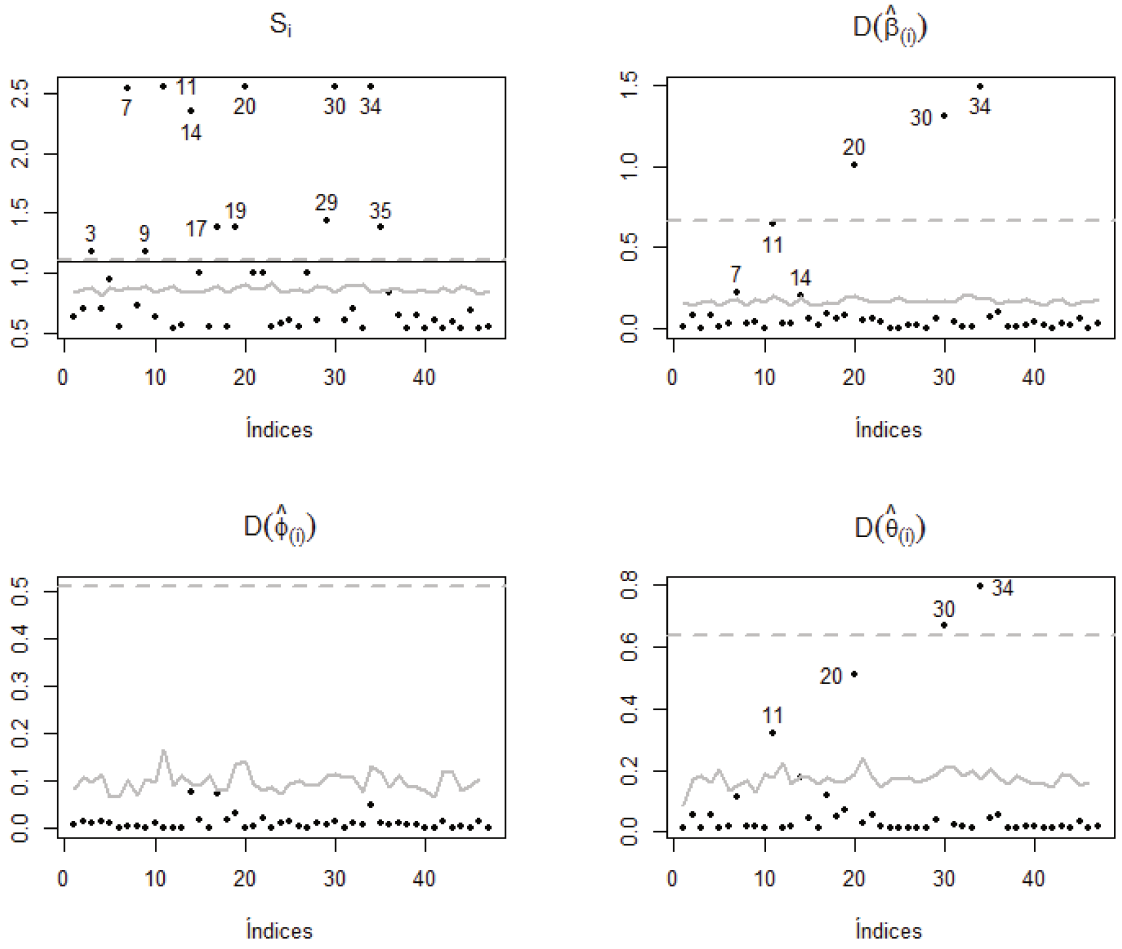


Figura 5.12: *Análise de influência para os dados HRD sob erros Logísticos tipo II . As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.*

produz maiores mudanças nas estimativas do modelo Logis II do que nas estimativas dos modelos normal e exponencial potência.

Este exemplo mostra que, apesar de alguns modelos simétricos conseguirem diminuir o efeito de *outliers* na variável reposta, eles não conseguem lidar com valores grosseiramente aberrantes nas variáveis explanatórias (chamados de alavancagem) em amostras com menos de 50 observações. Vimos que, nesta situação, o ajuste com o modelo exponencial potência indica uma distribuição com caudas mais leve que a normal, sendo vunerável a todo tipo *outlier*. Ajustamos também o modelo t-Student cujos resultados foram omitidos, pois para este exemplo o grau de liberdade encontrado foi igual a 30, uma distribuição aproximadamente normal com estimativas similares às dos demais ajustes.

Para os três ajustes as estimativas foram fortemente influenciadas pelos pontos de alta alavancagem (11,20,30,34) e pelos pontos discrepantes (7,9,14). No diagnóstico de influência para o ajuste do modelo normal os limites M_1 e M_2 foram melhores do que o limite de Peña, na detecção

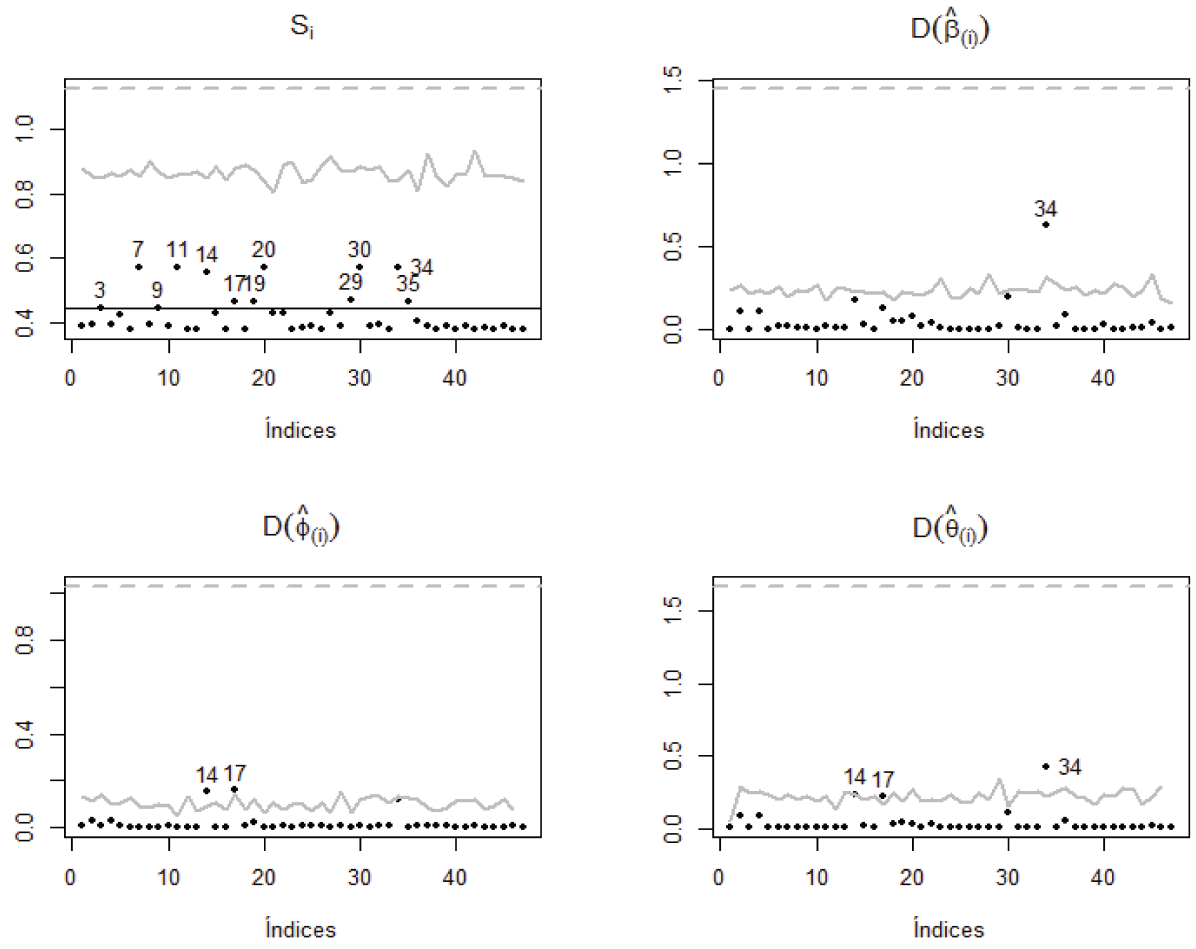


Figura 5.13: *Análise de influência para os dados HRD sob erros exponencial potência. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.*

dos pontos (7,11,14,20,30,34) sob a avaliação da estatística de Peña. Para a distância de Cook generalizada alguns pontos foram identificados, mas somente pelo limite M_2.

Para o ajuste do modelo logístico tipo II a estatística de Peña apresentou uma alta variabilidade de forma que nem os limites simulados e nem o limite de Peña foram eficazes na detecção das observações 7, 11, 14, 20, 30 e 34, pois estes limites também detectam outras observações. Para distância de Cook referente às estimativas de $\boldsymbol{\beta}$ o limite M_2 é mais eficaz que o limite M_1, pois detecta as observações 7, 11, 14, 20, 30 e 34, enquanto que o últimos detecta apenas as observações 20, 30 e 34.

No ajuste do modelo exponencial potência as observações atípicas não foram identificadas pelas medidas da estatística de Peña. Para distância de Cook generalizada nem todas as observações atípicas foram detectadas. Observamos apenas identificação das observações 14 e 34 e também a observação 17 que não parece estar longe da configuração dos dados, segundo a Figura 5.9.

Tabela 5.6: *Mudanças (em percentuais) nas estimativas após exclusão de observações.*

Observações excluídas	**Distribuição**	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\phi}$
11, 20, 30, 34	Normal	159,71	-595,19	48,56
	Logis II	181,13	-661,66	52,23
	Exp. potência(-0,3)	139,90	-505,33	46,49
7, 11, 14, 20, 30, 34	Normal	220,81	-822,00	55,29
	Logis II	235,50	-859,37	59,48
	Exp. potência(-0,3)	192,25	-693,83	50,93
7, 9, 11, 20, 30, 34	Normal	225,12	-837,03	63,82
	Logis II	226,14	-824,90	64,35
	Exp. potência(-0,3)	220,18	-792,67	62,82

5.1.3 Refinaria

Os dados foram apresentados por Wood (1973) e obtidos num estudo de um processo produção em uma unidade de refinaria envolvendo quatro variáveis independentes. $\mathbf{x}_1$, $\mathbf{x}_2$ e $\mathbf{x}_3$ representam composições do alimento enquanto $\mathbf{x}_4$ é o registo de uma combinação de condições do processo. A variável dependente, y, é o índice de octano no produto produzido. O seguinte modelo linear foi considerado

$$y_i = \beta_0 + \sum_{a=1}^{4} x_{ia}\beta_a + \epsilon_i, i = 1, \ldots, 82.$$

Tabela 5.7: *Estimativas de máxima verossimilhança (erro padrão aproximado) para alguns modelos simétricos lineares ajustados aos dados Refinaria.*

Distribuição	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\phi}$	$L(\hat{\boldsymbol{\theta}})$	**AIC**
Normal	95,939 (1,172)	-0,094 (0,005)	-0,126 (0,031)	-0,024 (0,013)	1,919 (0,311)	0,179 (0,028)	-45,725	101,450
Logística II	95,958 (1,181)	-0,093 (0,005)	-0,119 (0,031)	-0,026 (0,013)	1,912 (0,313)	0,060 (0,011)	-47,384	104,7689
Exp, potência(-0,3)	95,774 (1,116)	-0,095 (0,005)	-0,140 (0,029)	-0,021 (0,013)	1,973 (0,296)	0,291 (0,037)	-45,006	100,0122

A Tabela 5.7 mostra as estimativas de máxima verossimilhança para o ajuste do modelo sob erros normais, logísticos tipo II e Exponêncial potência. As estimativas obtidas para os parâmetros da regressão nos respectivos ajustes são similares. Para o ajuste do modelo exponencial potência, encontramos o maior valor para verossimilhança quando $\nu = -0,3$. Na Figura 5.14 os gráficos normais de probabilidades com envelope não apresentam nemhum comportamento não usual.

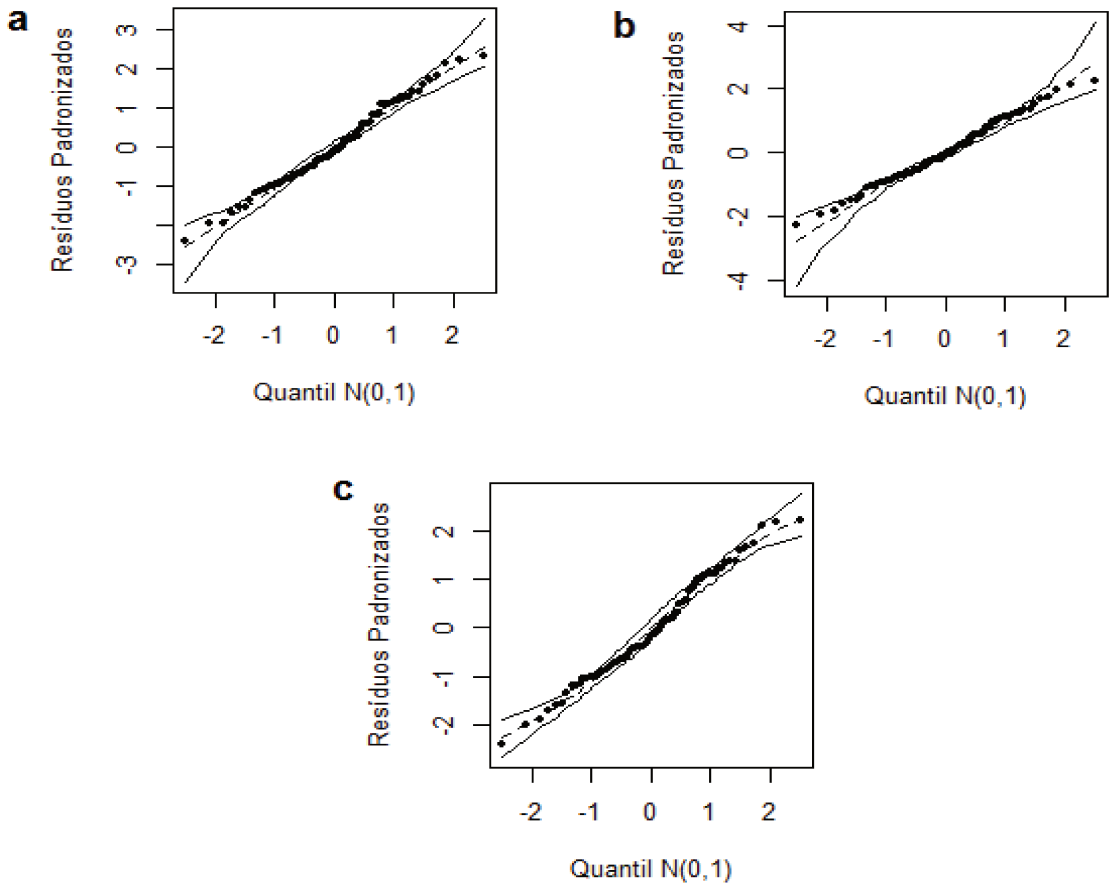


Figura 5.14: *Gráficos normais de probabilidades com envelope para os resíduos padronizados para o ajuste do modelo normal (a), logístico-II (b) e exponencial potência (c) ajustados aos dados Refinaria.*

Na análise de influência considerando o ajuste do modelo normal observamos que os limites construídos para estatística de Peña, assim como a estatística de Peña não indentificaram observações influentes. Já para a distância de Cook generalizada foram identificadas as observações 73 e 77 como influentes para as estimativas de $\boldsymbol{\beta}$ pelo limite simulado M_2 (veja Figura 5.15). Para o ajuste do modelo logístico tipo II as observações 44, 75 e 77 foram detectadas como observações influentes para as estimativas de $\boldsymbol{\beta}$ segundo o limite simulado M_2 (Figura 5.16).

A Figura 5.17 mostra a nálise de influência para o ajuste do modelo exponencial potência. Para a estatística de Peña não foram identificadas observações influentes por qualquer um dos limites. Para a distância de Cook generalizada foram identificadas as observações 73 e 82 como influentes para as estimativas de $\boldsymbol{\beta}$ e as observações 21, 61 e 82 para as estimativas de ϕ segundo o limite simulado M_2.

Para estatística de Andrews-Pregibon consideramos $k_{\max} = 5$. Portanto, $m = 10$ observações suspeitas foram examinadas. Nos três ajustes vemos que as razões mínimas são similares e os maiores valores de D_k^0 ocorreram quando até $k = 3$ observações foram omitidas dos ajustes. Assim,

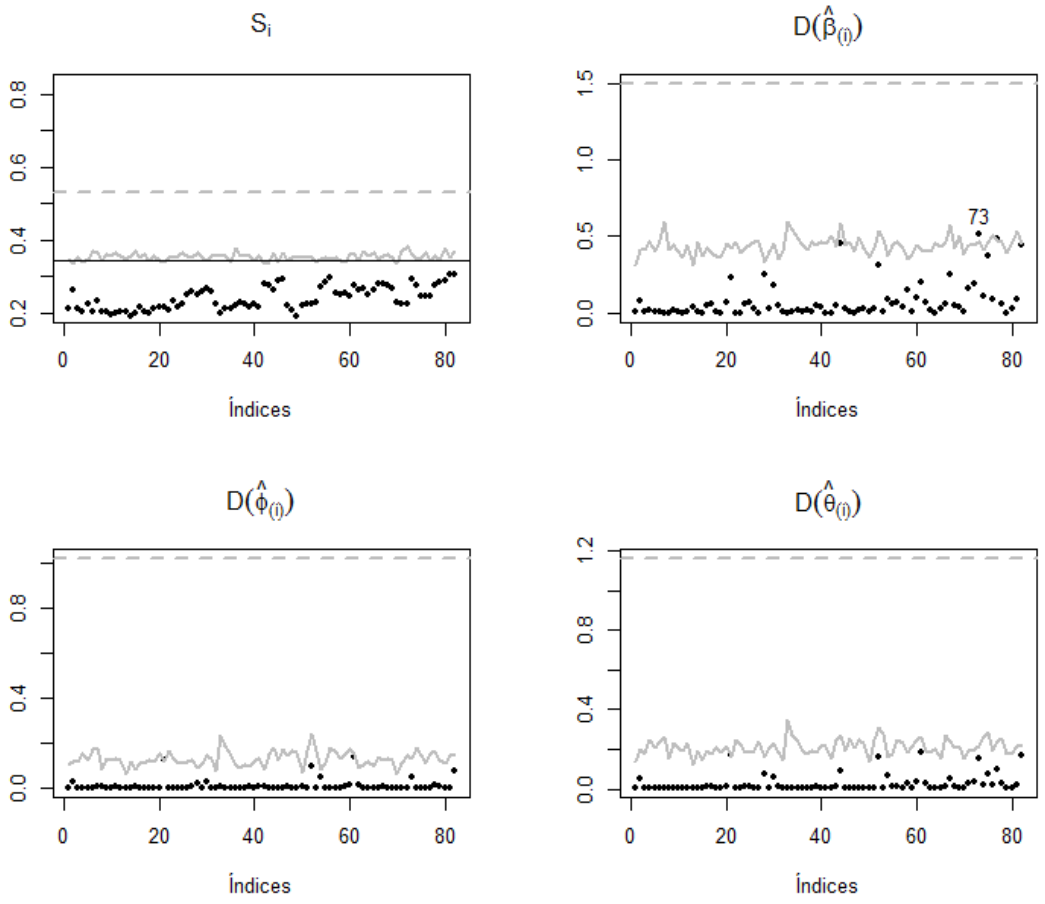


Figura 5.15: *Análise de influência para os dados Refinaria sob erros normais. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.*

as observações mais desviantes são 75, 76 e 77. Estas observações também foram identificadas por Wood (1973) após análise de resíduos e análise de influência usando variáveis *dummy*. Andrews e Pregibon, também, detectaram essas observações após análise gráfica dos pontos $(I^k_{\mathbf{I}}, k)$, verificando que o ponto mais isolado no gráfico é o que corresponde a $k = 3$.

Verificamos que as observações identificadas pelos limites simulados da distância de Cook generalizada não são as mesmas que foram identificadas pela estatística de Andrews-Pregibon. Uma observação feita por Andrews e Pregibon foi que diagnósticos influência baseados na examinação de um caso por vez não conseguem identificar as observações que foram identificadas no método proposto por eles. Além disso, Wood (1973) notou que um estudo das condições em que as observações influentes (75-77) foram tomadas resultou em uma economia considerável para o processo de produção sob investigação.

Apresentamos na Tabela 5.9 a mudança percentual nas estimativas após a exclusão de algumas oservações. Observamos que as maiores mudanças foram produzidas com a exclusão das observações 73, 75, 76 e 77 no ajuste do modelo logístico tipo II e as menores mudanças foram obtidas

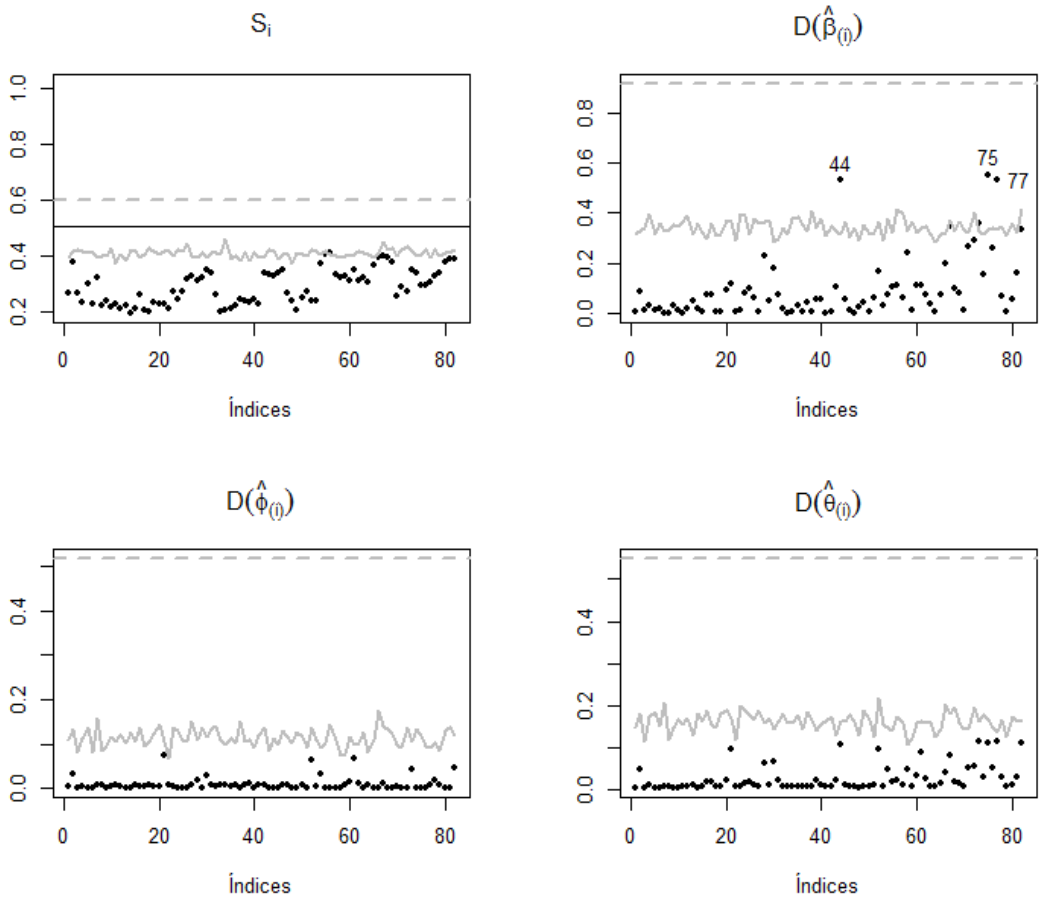


Figura 5.16: *Análise de influência para os dados refinaria sob erros Logísticos tipo II. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.*

com o ajuste do modelo exponencial potência.

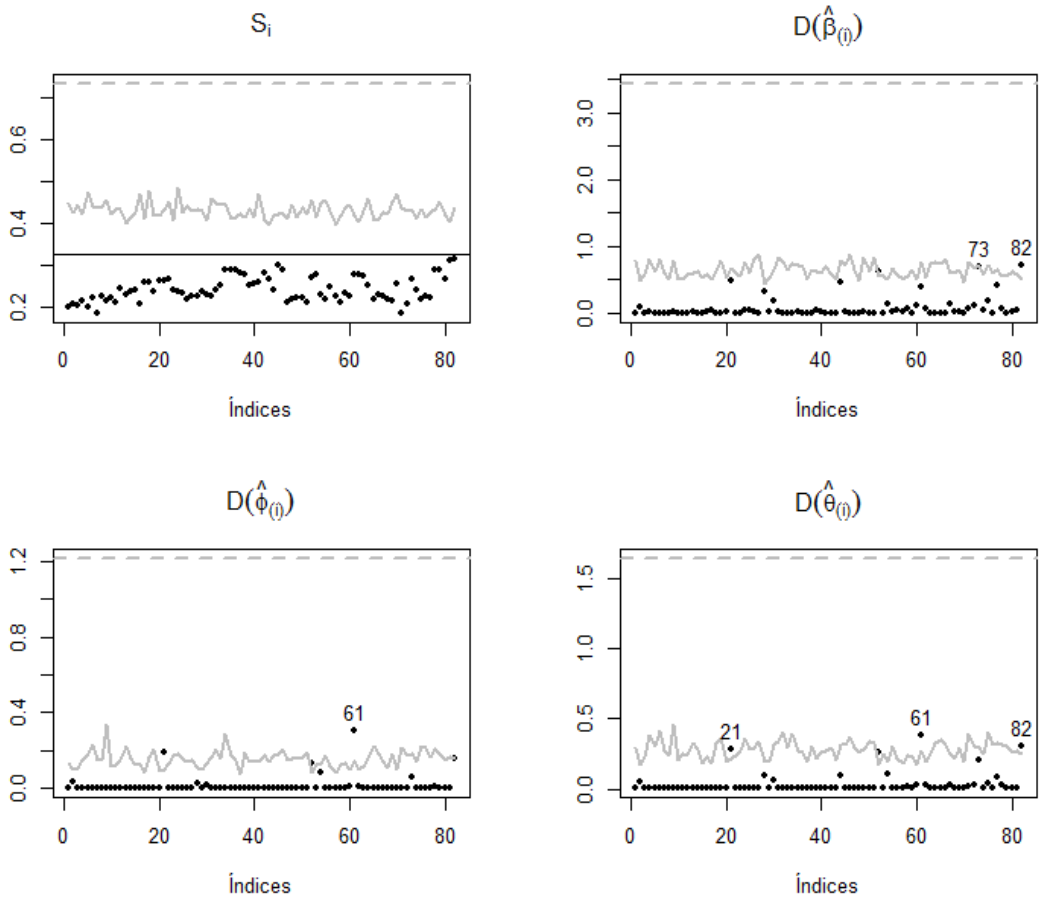


Figura 5.17: *Análise de influência para os dados refinaria sob erros exponencial potência. As linhas cinzas tracejadas são os limites simulados M_1, as linhas cinzas contínuas são os limites simulados M_2 e a linha preta o limite de Peña.*

5.2 Modelos não-lineares

Nesta seção apresentamos aplicação dos métodos de diagnóstico de influência aos modelos não-lineares. Começamos com um exemplo obtido pelo estudo de simulação. Em seguida apresentamos dois exemplos baseados em dados reais que apresentam observações atípicas. Aqui verificamos o desempenho das estatísticas de influência, bem como a metodologia para detecção de observações influentes sob o ajuste de alguns modelos simétricos não-lineares.

5.2.1 Simulação

Resultados das estatísticas de influência para o estudo de simulação

Nesta parte mostramos os resultados do diagnóstico de influência para o ajuste dos dados do estudo de simulação em 3.2.1 considerando modelo

Tabela 5.8: *Estatística de Andrews-Pregibon para os dados Refinaria.*

Distribuição	k	R_k^0	D_k^0	**Observações**
	1	0,629585	0,135515	75
	2	0,373253	0,215014	75, 76
Normal	3	0,221259	0,220034	75, 76, 77
	4	0,153432	0,063114	73, 75, 76, 77
	5	0,111238	0,044304	44, 73, 75, 76, 77
	1	0,629411	0,136491	75
	2	0,373199	0,214892	75, 76
Logis II	3	0,221865	0,217456	75, 76, 77
	4	0,153509	0,064135	73, 75, 76, 77
	5	0,111314	0,044361	44, 73, 75, 76, 77
	1	0,6292146	0,137310	75
	2	0,373079	0,215153	75, 76
Exp. potência	3	0,221699	0,218257	75, 76, 77
	4	0,153775	0,062693	44, 75, 76, 77
	5	0,111477	0,045116	44, 73, 75, 76, 77

Tabela 5.9: *Mudanças (em percentuais) nas estimativas após exclusão de algumas observações.*

Observações	**Distribuição**	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\phi}$
	Normal	0,84	-4,84	-0,22	-21,20	7,26	0,37
75-77	Logis II	1,17	-6,67	-5,00	-31,41	8,99	1,03
	Exp. potência(-0,3)	0,47	-2,86	-2,20	-11,41	4,40	0,95
	Normal	1,32	-8,41	-20,89	-48,49	4,02	5,59
73, 75-77	Logis II	1,56	-9,56	-22,70	-53,66	5,97	7,51
	Exp. potência(-0,3)	0,95	-6,62	-16,78	-40,10	1,66	4,02
	Normal	1,11	-6,82	-20,43	-23,80	10,08	0,84
44, 75-77	Logis II	1,39	-8,15	-18,30	-31,65	12,06	1,97
	Exp. potência(-0,3)	0,84	-5,60	-20,59	-17,09	7,37	0,26

$$y_i = \frac{\exp\left(-\beta_1 x_i\right)}{\beta_2 + \beta_3 x_i} + \epsilon_i. \tag{5.2.1}$$

sob erros logísticos tipo II. Lembrando que, os valores para os parâmetros da regressão foram $\beta_1 = 0,1$, $\beta_2 = 0,005$ e $\beta_3 = 0,01$, que se baseiam nas estimativas encontradas para este conjunto. Os erros, $\epsilon_i \sim S(0, \phi, g)$, são independentes e identicamente distribuidos para $i = 1, \ldots, n$, com $\phi = 17$. A variável explanatória $\boldsymbol{x}$, foi gerada apartir da distribuição uniforme no intervalo $(0, 6)$ e seus valores foram fixados durante as simulações. Esse caso consideramos como sendo a ***Situação (1)***, dados sem observções influentes.

Nas demais situações foram inseridas observações atípicas nos casos $i = 20, 21, 22$. Para a

Situação (2) temos $y_i^* = \mu_i + 5\sigma_{\mu_i}$, na ***Situação (3)*** temos $y_i^* = \mu_i + 5\sigma_{y_i}$ com $x_i^* = 11$ e na ***Situação (4)*** temos $y_i^* = \mu_i + 5\sigma_{y_i}$ com $x_i^* = 16$.

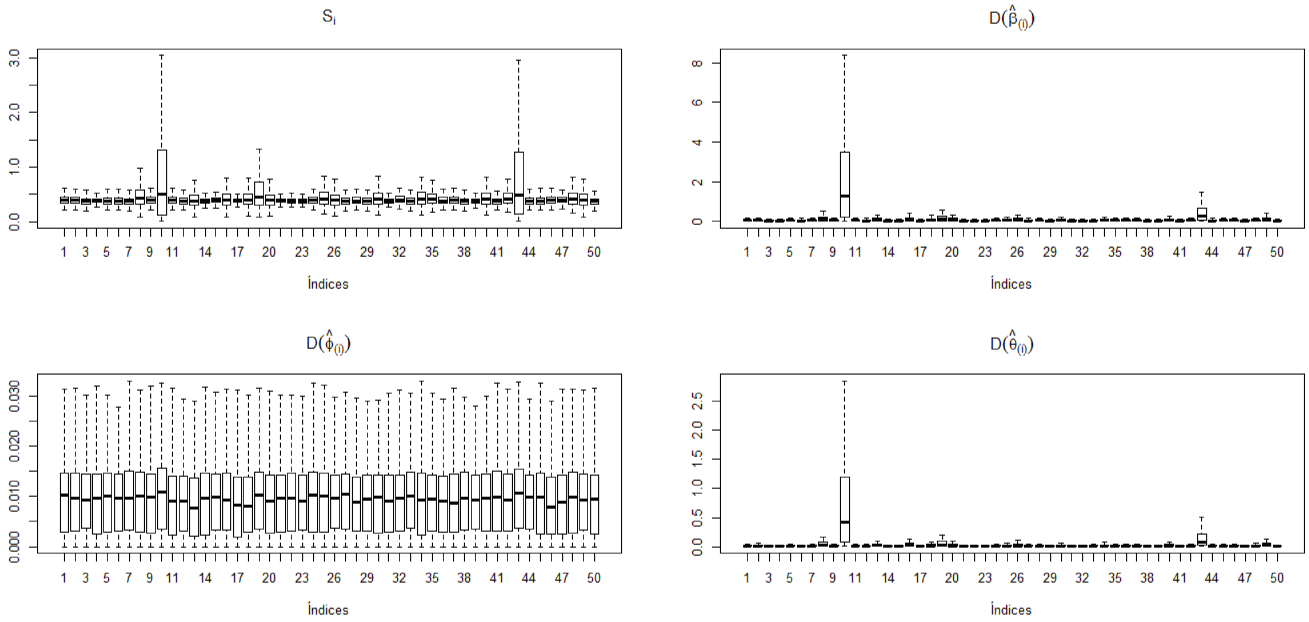


Figura 5.18: *Estatística de influência para o ajuste dos dados simulados da Situação (1).*

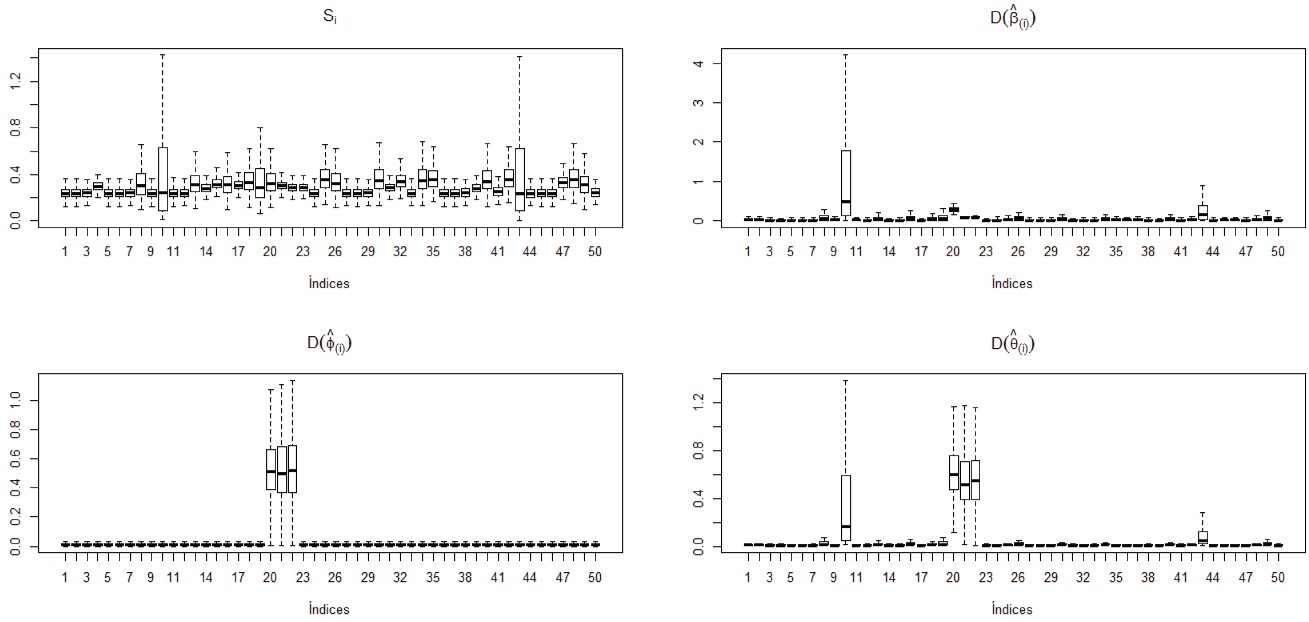


Figura 5.19: *Estatística de influência para o ajuste dos dados simulados da Situação (2).*

Na Figura 5.18 estão os gráficos de diagnóstico de influência para o estudos de simulação considerando dados sem observações influentes. No gráfico da estatística de Peña, os casos 10 e 43

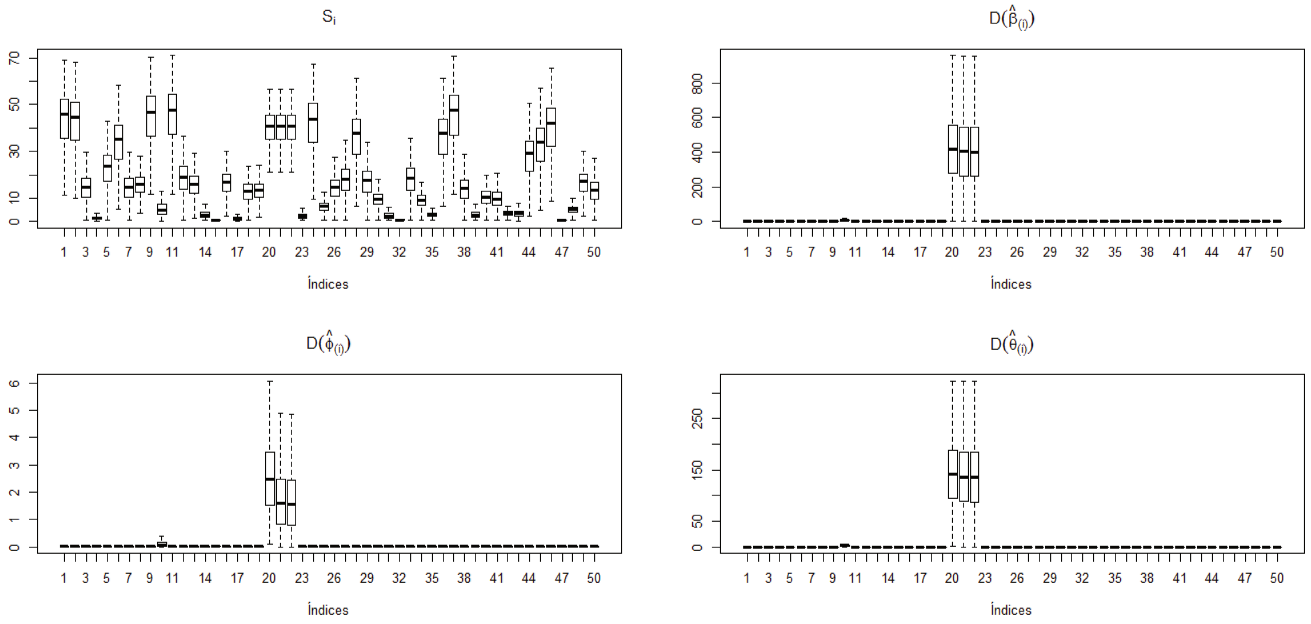


Figura 5.20: *Estatística de influência para o ajuste dos dados simulados da Situação (3).*

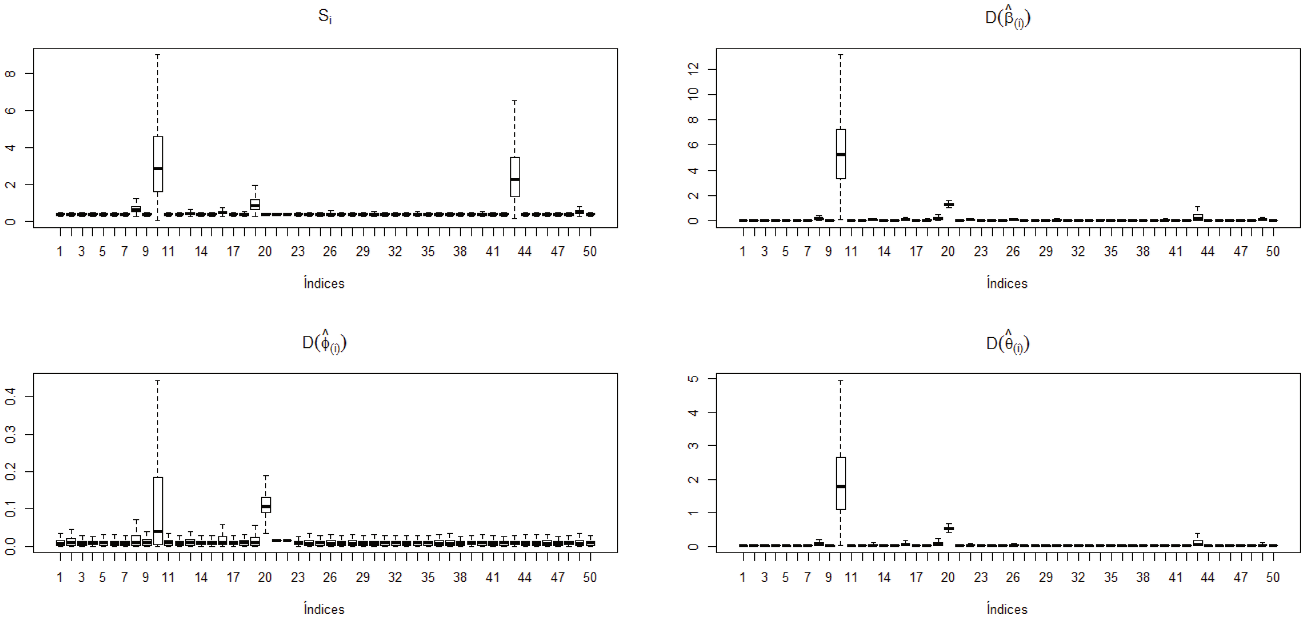


Figura 5.21: *Estatística de influência para o ajuste dos dados simulados da Situação (4).*

apresentaram variabilidade muito maior que as demais. Observamos que a observação 10, também, apresentou uma grande variabilidade para as medidas de $D(\hat{\boldsymbol{\theta}}_{(i)})$. Mas, para $D(\hat{\phi}_{(i)})$ não há efeito de alavanca, pois a variabilidade da estatística é a mesma para todas as observações.

Considerando o diagnóstico de influência para os ajustes dos dados com *outliers* na variável resposta, observamos em S_i e $D(\hat{\boldsymbol{\beta}}_{(i)})$ que as observações atípicas que foram inseridas nos dados

não aparecem nos gráficos, aparecem apenas as observações que são pontos de alavanca. Na análise de influência para as estimativas de ϕ observamos claramente a separação das observações atípicas (Figura 5.19).

Para os dados sob a presença de *outliers* com média alavancagem (veja Figura 5.20), os resultados obtidos para estatísticade de Peña mostram uma grande variação entre as observações e os pontos atípicos não se destacam dos demais. Já para a distância de Cook generalizada vemos claramente a separação dos pontos influentes das observações na amostra .

O diagnóstico de influência para os ajustes de dados com *outliers* de alta alavancagem está na Figura 5.21. Observamos que os pontos de alavanca do modelo "mascaram"o efeito das observações atípicas, que são substanciamente influentes para as estimativas dos parâmetro. Tanto para a estatística de Peña quanto para a distãncia de Cook generalizada não observamos a separação dos casos influentes dos pontos restantes.

As observações 10 e 43 são pontos de alavanca e correspondem aos casos em que os valores da variável explanatória são menores que 1, mostrando a dificuldade de predição nesses casos. De acordo com o gráfico das estatísticas de influência, podemos verificar que os limites simulados construídos com modelo de decaimento exponencial podem atingir valores muito altos para as medidas da estatísticitica de Peña e da distância de Cook referente às estimativas de $\boldsymbol{\beta}$.

5.2.2 Dados de ultra-sons

Estes dados são o resultado de um estudo do NIST (*National Institute of Standards and Technology*) envolvendo calibração ultra-sônica. Os dados consistem de 54 observações sobre duas variáveis. A variável resposta (y) é a resposta ultra-sônica e a variável preditora (x) é a distância do metal . Os dados podem ser encontrados em www.itl.nist.gov/div898/strd/nls/data/LINKS/DATA/Chwirut2.dat. No site, o modelo de regressão não-linear sugerido foi

$$y_i = \frac{\exp(-\beta_1 x_i)}{\beta_2 + \beta_3 x_i} + \epsilon_i. \tag{5.2.2}$$

com ajuste de mínimos quadrados. Castillo et al. (2009) examinaram os dados considerando análise de sensibilidade local para estimação com mínimos quadrados, mini-max e LAV. Para verificar a eficácia dos métodos propostos na detecção de *outliers*, os autores plantaram duas observações discrepantes no conjunto de dados: a observação 55 (com $y = 11$, $x = 1,75$) e a observação 56 (com $y = 25$, $x = 5,8$). Nós, também, usamos os dados contaminados para avaliar a qualidade das metodologias de detecção de observações influentes considerando o ajuste de modelos simétricos não-lineares.

Aos dados contaminados ajustamos o modelo normal, logístico tipo II e t-Student. As estimativas de máxima verossimilhança em cada ajuste estão na Tabela 5.10. Para o parâmetros de forma do modelo t-Student, o valor que produziu a maior verossimilhança foi igual a 2. Observamos que o modelo t-Student apresentou os menores erros-padrão aproximados e também o menor valor para o AIC.

Na Figura 5.22 estão o gráfico dos dados de ultra-sons contaminados e os gráficos dos dados com as bandas de 95% de confiança para curva ajustada dos respectivos modelos. Observamos que as bandas de confiança obtida com os dados contaminados estão bem próximas das bandas de

Tabela 5.10: *Estimativas de máxima verossimilhança (erro-padrão aproximado) para alguns modelos simétricos ajustados aos dados de ultra-sons contaminados.*

Distribuição	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	$L(\hat{\boldsymbol{\theta}})$	**AIC**
Normal	0,1101 (0,0479)	0,0041 (0,0010)	0,0148 (0,0022)	21,0878 (3,9852)	-164,8240	335,6480
Logis II	0,1421 (0,0379)	0,0048 (0,0007)	0,0133 (0,0016)	3,7301 (0,8337)	-154,4945	314,9890
t-Student(2)	0,1402 (0,0227)	0,0047 (0,0004)	0,0135 (0,0010)	2,4180 (0,7225)	-142,1192	290,2383

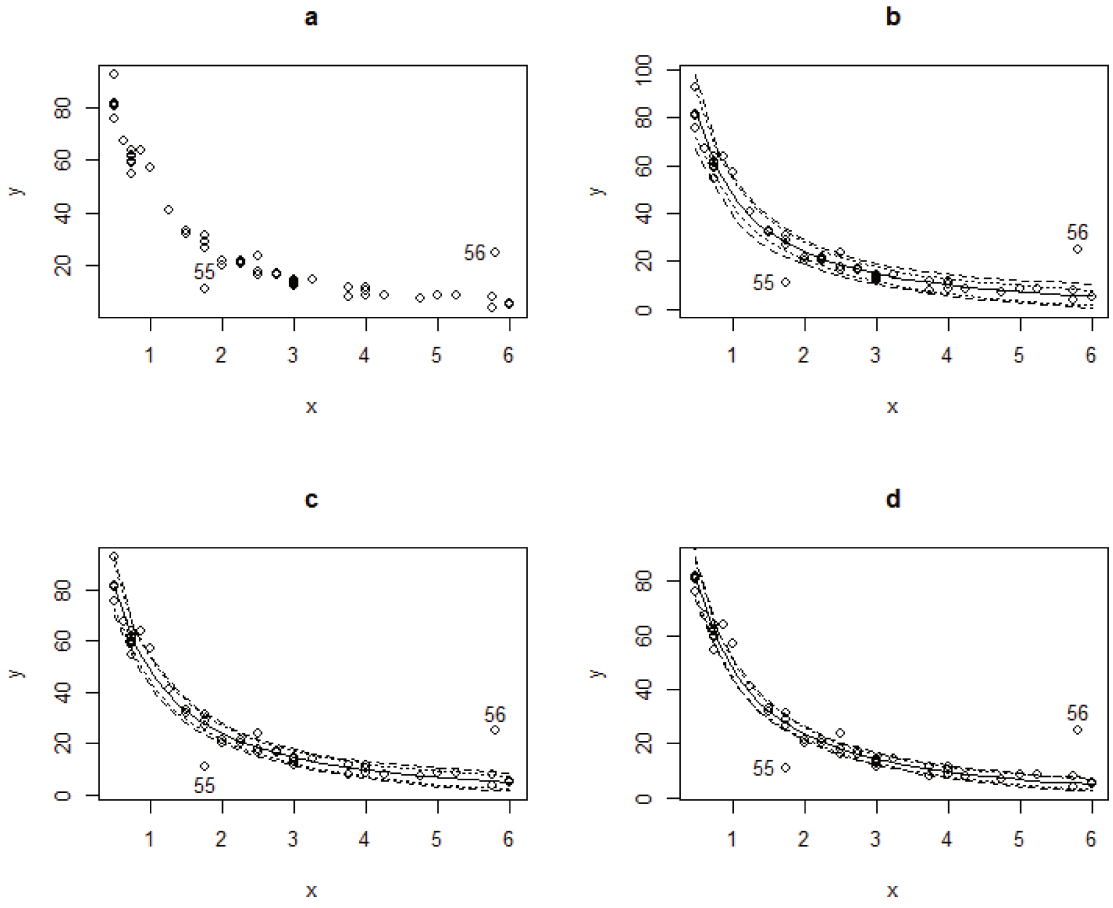


Figura 5.22: *Em (a) gráfico dos dados de ultra-sons contaminados e em (b) - (d) gráfico dos dados com o ajuste do modelo normal, Logístico-II e t-Student, respectivamente.*

confiança com obtidas com os dados originais quando nos reportamos ao ajuste do modelo normal e logístico-II e no ajuste do modelo t-Student as bandas de confiança se coincidem. Verificamos também que as observações discrepantes (55, 56) estão fora das bandas de confiança nos três

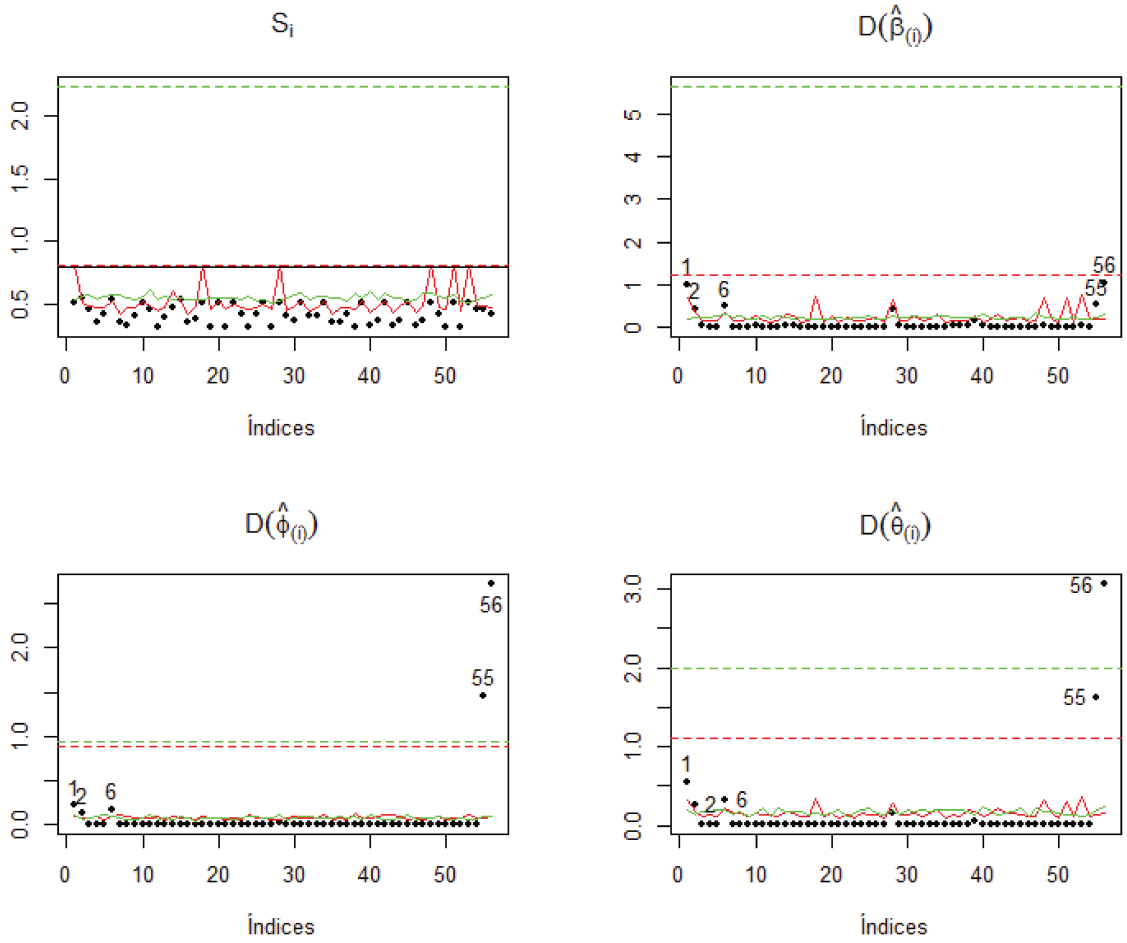


Figura 5.23: *Análise de influência para os dados de ultra-sons contaminados sob erros normais.*

ajustes.

Para cada ajuste foram calculadas a estatística de Peña e a distância de Cook generalizada. Apresentamos os resultados nas Figuras a seguir. As linhas tracejadas e contínuas em vermelho são os limites simulados M_{1_f} e M_{2_f}, respectivamente. As linhas tracejadas e contínuas em verde são os limites simulados M_1 e M_2, respectivamente, com $\mathbf{x}$ gerado a partir da distribuição uniforme no intervalo $(\min(\mathbf{x}), \max(\mathbf{x}))$. A linha contínua em preto é o limite de Peña.

A Figura 5.23 mostra o diagnóstico de influência para ajuste do modelo sob erros normais. Para a estatística de Peña observamos que nenhuma observação foi detectada pelo limite de Peña ou pelos limites simulados M_1 e M_{1_f}. Em $D(\hat{\boldsymbol{\beta}}_{(i)})$ nenhuma observação foi detectada pelos limites M_1 e M_{1_f}, mas as observações 1, 2, e 6 e as observações discrepantes (55 e 56) foram detectadas pelos limites M_2 e M_{2_f}. Em $D(\hat{\phi}_{(i)})$ as observações 55 e 56 foram detectadas pelos limites M_1 e M_{1_f}, mas os limites M_2 e M_{2_f} detectaram também as observações 1, 2 e 6, que corresponde aos casos em x é igual a $1/2$, 1 e $0,875$. Em $D(\hat{\boldsymbol{\theta}}_{(i)})$ a observação 55 foi detectada pelo limite M_1, o limite M_{1_f} detectou apenas observações discrepantes e os limites M_2 e M_{2_f} detectaram as observações 1, 2 e 6, além das observações discrepantes.

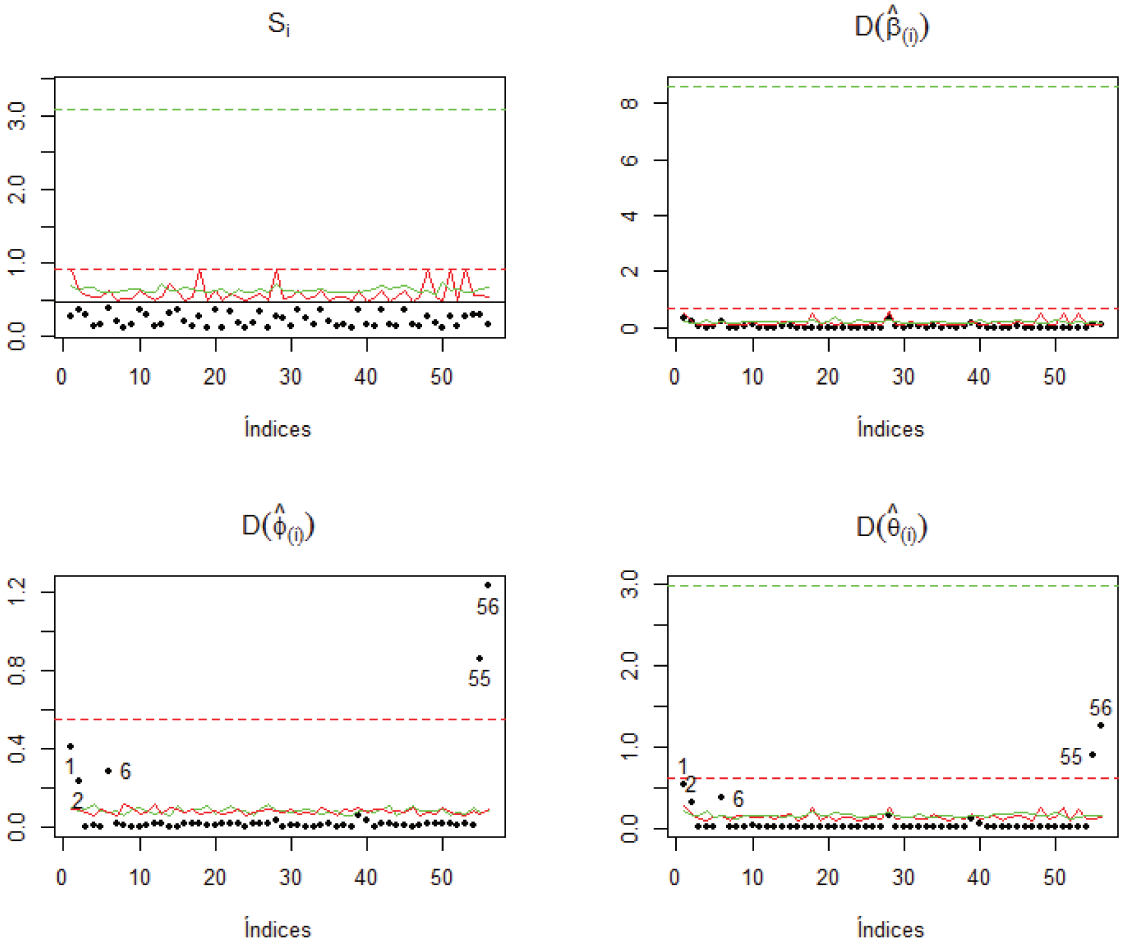


Figura 5.24: *Análise de influência para os dados de ultra-sons contaminados sob erros Logísticos tipo II.*

Para o ajuste com erros logísticos tipo II a análise de influência é apresentada na Figura 5.24. Observamos que para a estatística de Peña e a distância de Cook referente às estimativas de $\boldsymbol{\beta}$ não foram detectadas as observações discrepantes. Mas, para a distância de Cook referente às estimativas de ϕ as observações discrepantes foram detectadas pelos limites M_1 e M_{1_f}, que neste caso são similares e os limites M_2 e M_{2_f} identificaram mais observações (1, 2 e 6). Para $D(\hat{\boldsymbol{\theta}}_{(i)})$ as observações discrepantes foram detectadas pelo limite M_{1_f}, mas os limites M_2 e M_{2_f} identificam outras observações além das observações discrepantes.

No diagnóstico de influência para o ajuste do modelo t-Student (Figura 5.25), as medidas das estatísticas de influência foram menores do aquelas apresentadas para os outros ajustes e nenhuma observação foi detectada pelos limites simulados ou pelo limite de Peña no caso da estatística de Peña.

As mudanças percentuais após a exclusão de algumas observações são mostradas na Tabela 5.11. Observamos que as maiores mudanças foram produzidas no ajuste do modelo normal, principalmente, com a exclusão das observações discrepantes (55 e 56). As menores mudanças foram obtidas com ajuste do modelo t-Student com 2 graus de liberdade, exceto com a exclusão das

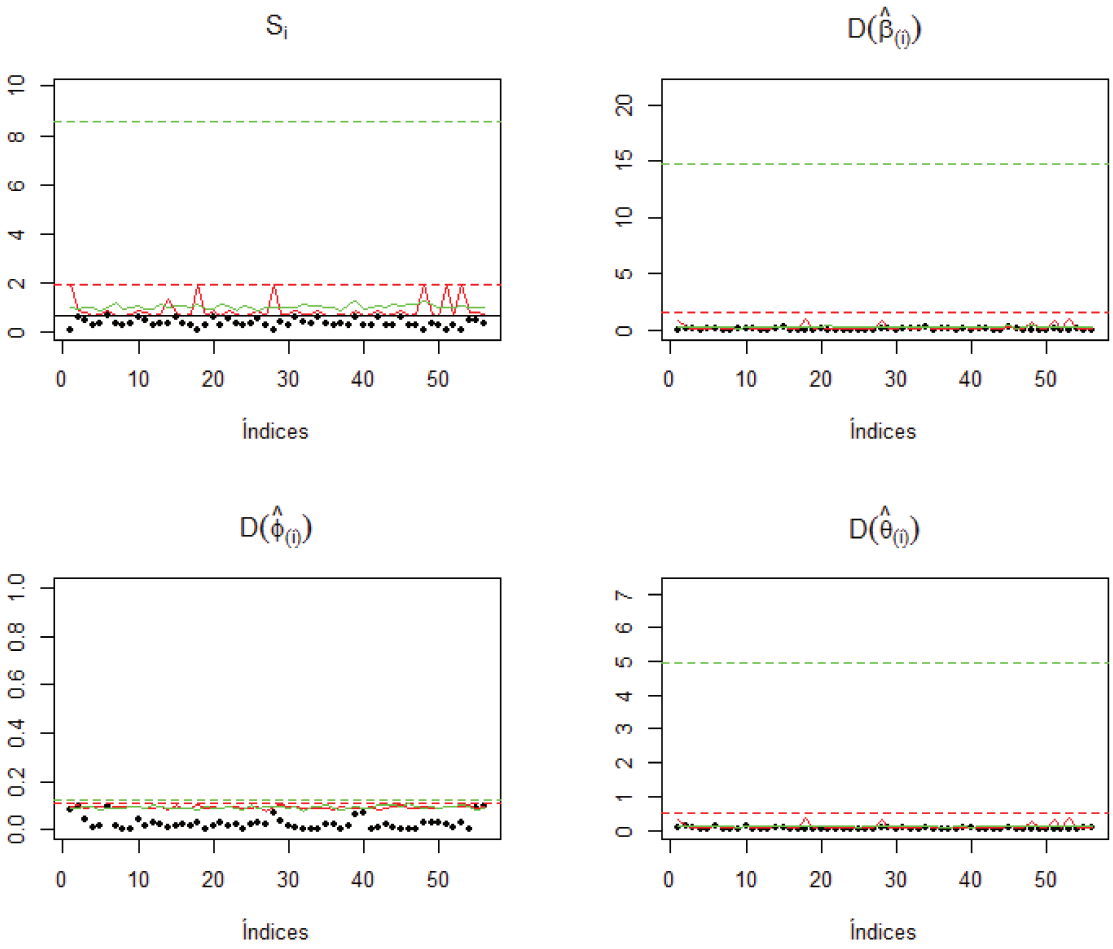


Figura 5.25: *Análise de influência para os dados de ultra-sons contaminados sob erros t-Student.*

Tabela 5.11: *Mudanças (em percentuais) nas estimativas após exclusão de algumas observações.*

Observações	**Distribuição**	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$
1, 2, 6	Normal	26,06	10,31	11,17	25,86
	Logis II	10,04	4,31	5,77	35,71
	t-Sudent(2)	4,35	2,56	2,41	24,62
55, 56	Normal	51,35	26,72	17,74	54,95
	Logis II	9,66	5,74	4,80	42,79
	t-Student(2)	0,71	0,38	0,36	17,40
1, 2, 6, 55, 56	Normal	28,88	21,90	9,81	81,46
	Logis II	1,20	0,61	1,17	69,34
	t-Student(2)	3,71	2,22	2,06	37,60

observações 1, 2, 6, 55 e 56 em que as menores mudanças foram obtidas com o ajuste do modelo logístico tipo II.

Na análise de sensibilidade de Castillo et al. (2009) a inserção das observações 55 e 56 causou uma mudança substancial para o ajuste de mínimos quadrados e no ajuste com método mini-max as observações 1, 55 e 56 foram detectadas como observações mais influentes. Nos ajustes com o modelo normal e logístico tipo II, verificamos que as observações 55 e 56 influenciam mais as estimativas do parâmetro de escala (ϕ) do que as estimativas de $\boldsymbol{\beta}$, sendo detectadas pelos limites simulados M_1 e M_{1_f}. As observações 1, 2 e 6 correspondem aos casos em que x é igual a 1/2, 1 e 0,875, respectivamente, mas estes pontos não exercem forte influence sobre as estimativas dos parâmetros. O ajuste do modelo t-Student reduziu o efeito das observações atípicas e por isso as não foram detectadas pelos limites simulados. Nos ajustes que foram vuneráveis aos casos atípicos, os limites simulados referentes aos quantis de 95% das estatísticas nas observações além de identificar os casos discrepantes, identificam também observação que não são tão influentes para o ajuste.

5.2.3 Dados da produção de cebola

Este conjunto de dados foi descrito por Ratkowsky (1983). A variável resposta, y, é a produção de cebola (g/planta) e a variável explanatória, x, é a densidade da planta (planta/m^2), em uma amostra com 42 observações. Ratkowsky ajustou o modelo

$$log(y_i) = -log(\alpha + \beta x_i) + \epsilon_i, \tag{5.2.3}$$

aos dados, por mínimos quadrados. Nós reanalisamos os dados usando duas distribuições para os erros, t-Student e Logística tipo II, além da distribuição normal.

Tabela 5.12: *Estimativas de máxima verossimilhança (erro-padrão) para alguns modelos simétricos ajustados aos dados produção de cebola.*

Distribuição	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\phi}$	$L(\hat{\boldsymbol{\theta}})$	**AIC**
Normal	0,003462 (0,000476)	0,000129 (0,000009)	0,021713 (0,004738)	20,83	-37,66
Logis II	0,003780 (0,000384)	0,000121 (0,000007)	0,004783 (0,001234)	26,80	-49,60
t-Student(3,5)	0,003886 (0,000343)	0,000118 (0,000006)	0,007933 (0,002359)	28,96	-53,92

As estimativas de máxima verossimilhança são apresentadas na Tabela 5.12. Observamos que as estimativas são altamente significantes em todos os modelos, mas as estimativas do ajuste do modelo t-Student com 3,5 graus de liberdade apresentam erros-padrão menores e também o menor valor para o AIC.

A Figura 5.26 mostra o gráfico dos dados da pridução de cebola e os gráficos dos dados com o ajuste e bandas de confiança dos respectivos modelos. Observação 38 aparece claramente nos gráficos como um *outlier*. Nas análises feitas por Ratkowsky esta observação foi omitida. Castillo et al. (2009) incluiram esta observação para uma análise de sensibilidade.

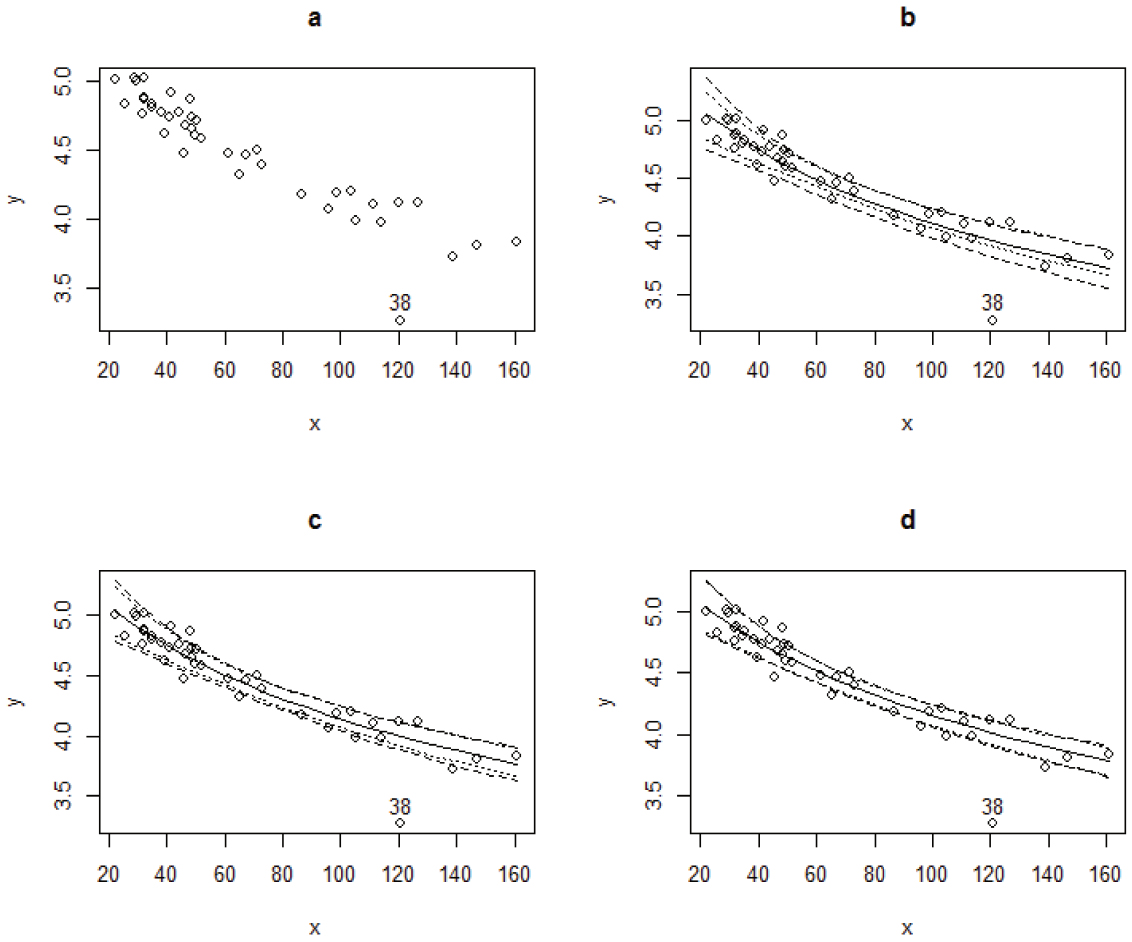


Figura 5.26: *Em (a) gráfico dos dados da produção de cebola e em (b) - (d) gráfico dos dados com o ajuste do modelo normal, Logístico-II e t-Student, respectivamente.*

Na Figura 5.27 estão os gráficos normais de probabilidades para os resíduos padronizados. Esta figura indica claramente que o ajuste com o modelo t-Student é mais adequado para modelar os dados do que o modelo normal e o logístico-II, uma vez que não há observações que estejam fora do envelope. Além disso, existe evidência de falta de ajuste para o modelo normal.

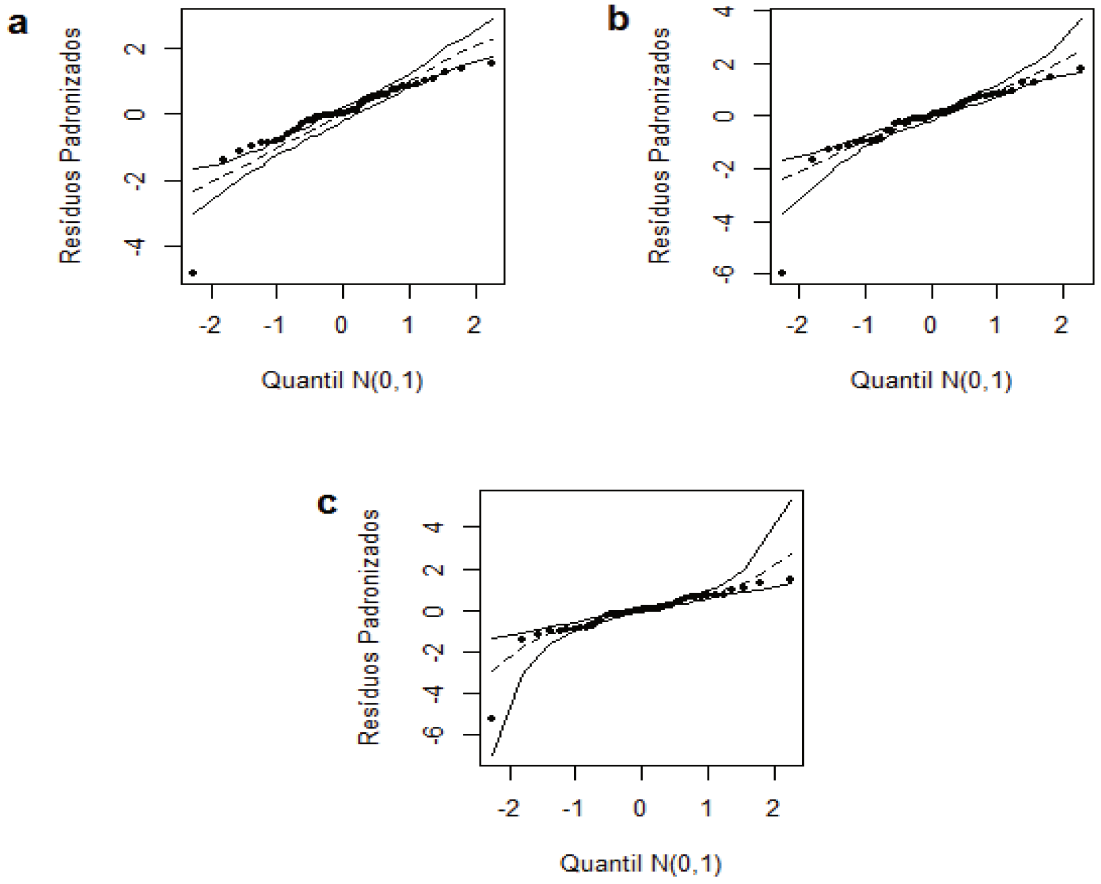


Figura 5.27: *Gráficos normais de probabilidades com envelope para os resíduos padronizados para o ajuste do modelo normal (a), logístico-II (b) e t-Student (c) ajustados aos dados da produção de cebola.*

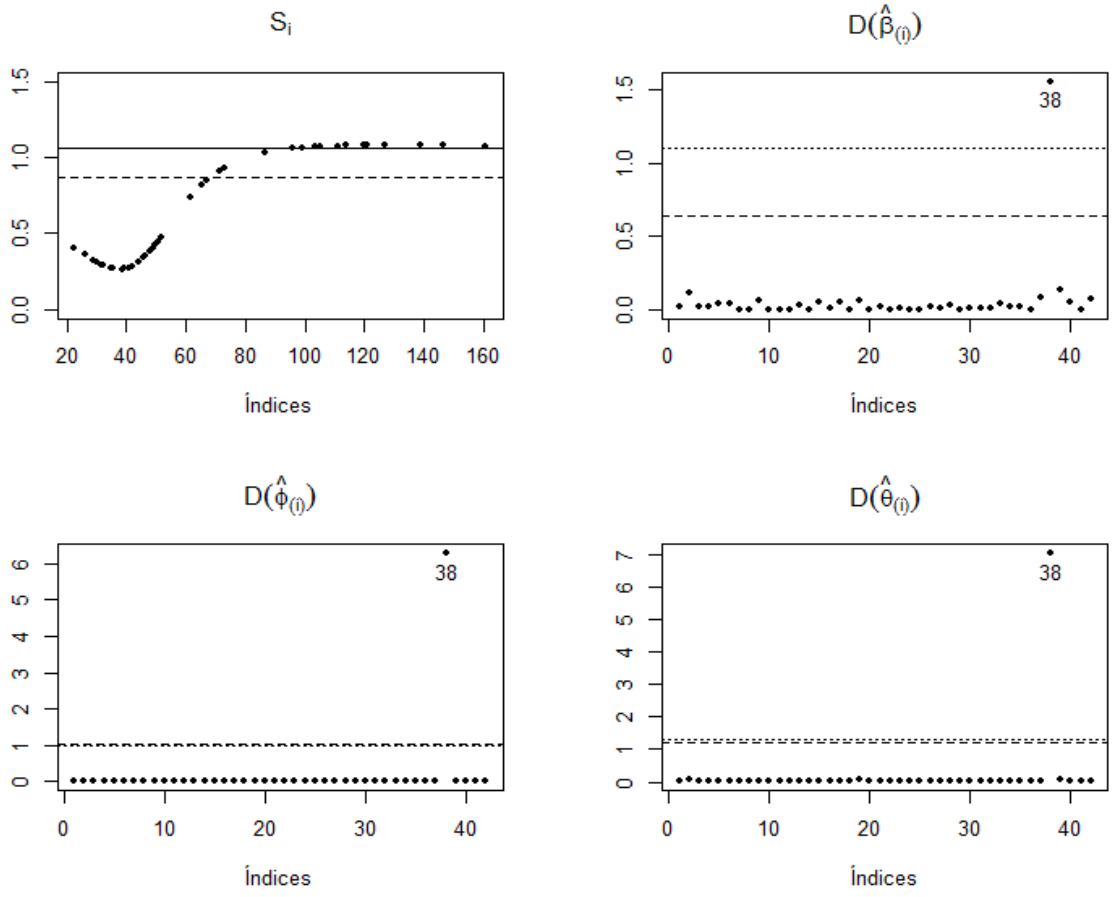


Figura 5.28: *Análise de influência para os dados da produção de cebola sob erros normais.*

Para cada ajuste foram calculadas a estatística de Peña e a distância de Cook generalizada. Apresentamos os resultados nas Figuras a seguir. Para cada estatística, a linha tracejada é o limite simulado M_{1_f}, a linha pontilhada o limite simulado M_1, com $\mathbf{x}$ gerado a partir da distribuição uniforme no intervalo [min($\mathbf{x}$), max($\mathbf{x}$)]. A linha preta é o limite de Peña.

No diagnóstico de influência para o ajuste do modelo normal, observamos que a estatística de Peña segue uma tendência crescente. Isso porque, os pares de observações estão ordenados segundo os valores da variável explanatória. O limite simulado M_1 coincidiu com o limite de Peña e algumas observações ultrapassaram o limite M_{1_f}. Para distância de Cook generalizada a observação 38 foi detectada tanto pelo limite M_1 quanto pelo limite M_{1_f} (Veja Figura 5.28). Para o ajuste sob erros logísticos tipo II, a observação 38 é influente para as estimativas de ϕ, pois foi detectada tanto pelo limite M_1 e M_{1_f}. Pela estatística de Peña nenhuma observação foi dtectada (Figura 5.29). No ajuste do modelo t-student nenhuma observação foi detectada como influente, segundo as medidas da estatística de Peña e distância de Cook generalizada (Figura 5.30).

A Tabela 5.13 mostra as mundanças nas estimativas após a exclusão da observação 38. As menores mudanças foram obtidas com o ajuste do modelo t-Student.

Neste exemplo, os limites simulados, para distância de Cook generalizada, detectaram a ob-

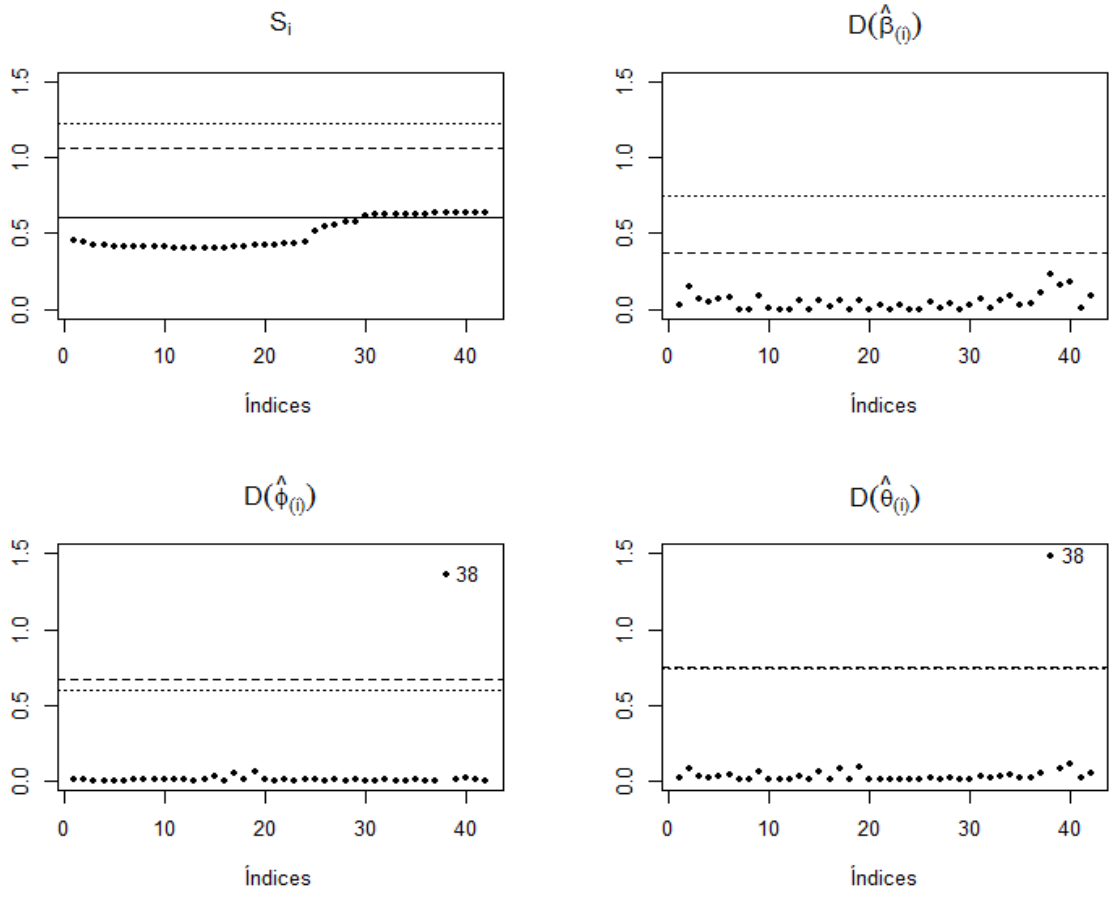


Figura 5.29: *Análise de influência para os dados da produção de cebola sob erros Logísticos tipo II.*

Tabela 5.13: *Mudanças (em percentuais) nas estimativas após exclusão da observação 38.*

Distribuição	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\phi}$
Normal	11,79	8,37	54,67
Logis II	3,37	2,75	30,14
t-Student(3,5)	1,28	1,09	13,28

servação 38 como sendo influente para as estimativas de $\hat{\boldsymbol{\beta}}$ e ϕ, sob o ajuste do modelo normal, e influente para as estimativa de ϕ, sob o ajuste do modelo logístico-II. No ajuste do modelo t-Student com 3,5 graus de liberdade, o efeito da observação 38 foi reduzido e os limites simulados não detectaram observações influentes.

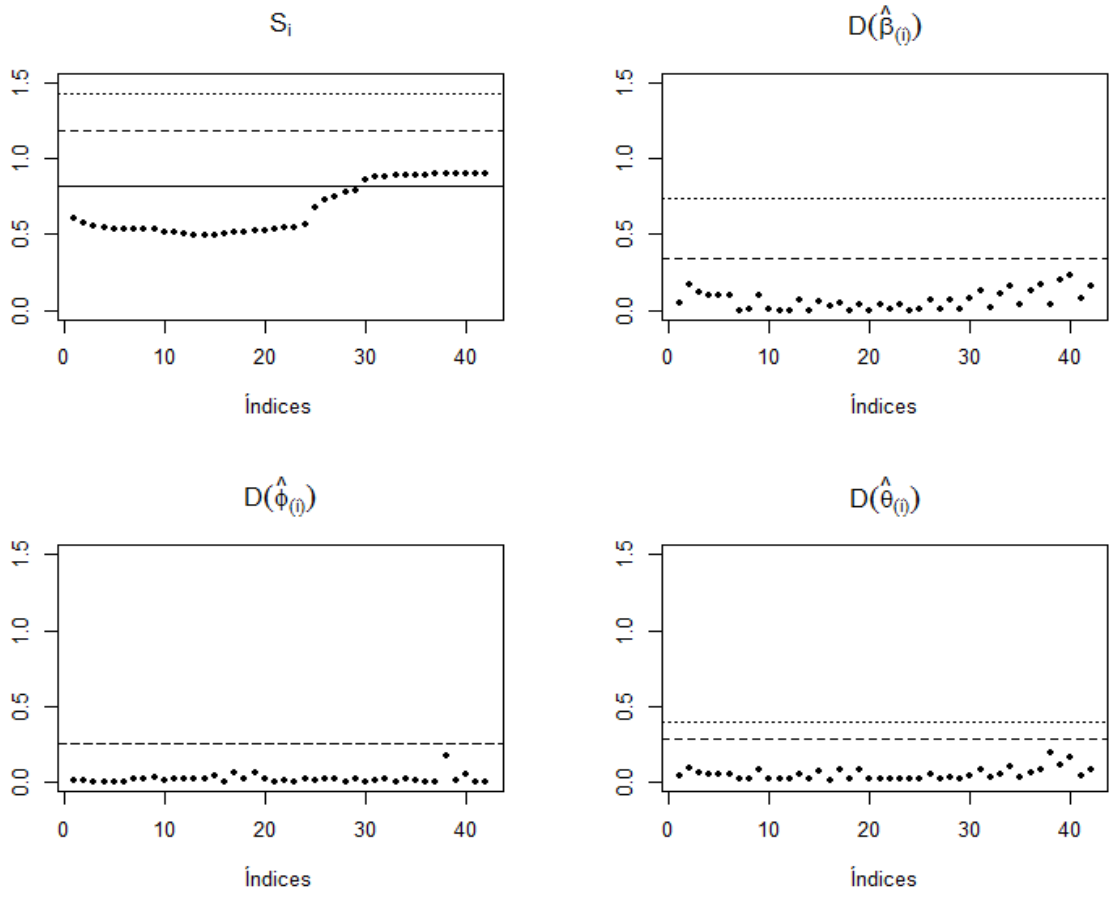


Figura 5.30: *Análise de influência para os dados da produção de cebola sob erros t-Student.*

5.2.4 Coelhos europeus

Este conjunto de dados foi analisado por Ratkowsky (1983) como um exemplo para ajuste de modelos de rendimento. Outros autores reanalizaram este conjunto de dados como Galea et al. (2005), que discutiu a aplicação dos procedimentos de diagnóstico de influência local em modelos com erros normais e em outras distribuições simétricas e Vanegas e Cysneiros (2010) que avaliaram procedimentos de diagnóstico com enfoque em influência global. Os dados sobre coelhos europeus contém informações sobre o peso da lente do olho do coelho, y (mg), e a idade do animal, x (dias). O modelo regressão não-linear sugerido aos dados foi

$$\log(y_i) = \alpha - \beta(x_i + \gamma)^{-1} + \epsilon_i, \quad i = 1, \ldots, 71. \tag{5.2.4}$$

Aos dados foram ajustados o modelo normal, logístico tipo II e t-Student. Para determinar os graus de liberdade do modelo t-Student encontramos o maior valor para verossimilhança quando $\nu = 4$. As estimativas de máxima verossimilhança para cada ajuste estão na Tabela 5.14. Em geral, as estimativas são parecidas, embora os erros-padrão aproximados das estimatimativas no

Tabela 5.14: *Estimativas de máxima verossimilhança (erro-padrão aproximado) para alguns modelos simétricos ajustados aos dados dos coelhos europeus.*

Distribuição	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\phi}$	$L(\hat{\boldsymbol{\theta}})$	**AIC**
Normal	5,6404 (0,0196)	130,7827 (5,6199)	37,6739 (2,2795)	0,0038 (0,0006)	97,0551	-188,1102
Logis II	5,6336 (0,0179)	127,4894 (5,0197)	35,9468 (2,0260)	0,0011 (0,0002)	99,0283	-192,0567
t-Student(4)	5,6319 (0,0168)	126,5226 (4,6802)	35,3795 (1,8840)	0,0021 (0,0005)	99,4151	-192,8302

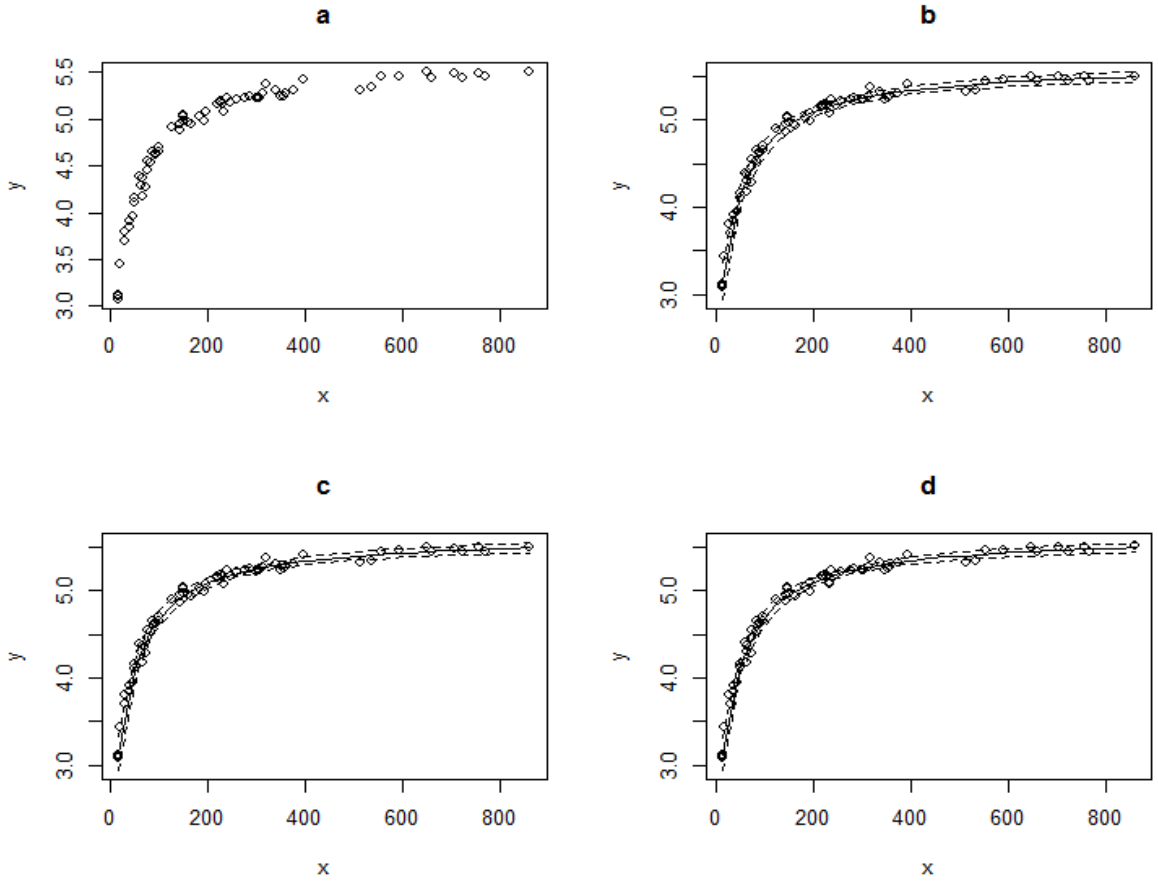


Figura 5.31: *Em (a) gráfico dos dados dos coelhos europeus e em (b) - (d) gráfico dos dados com o ajuste do modelo normal, Logístico-II e t-Student, respectivamente.*

modelo t-Student e logístico-II são menores do que os erros-padrão do modelo normal.

O gráfico dos dados é apresentado na Figura 5.31a e os gráficos dos dados com os ajustes dos respectivos modelos e as bandas de confiança estão em Figura 5.31b-d. Parece não haver

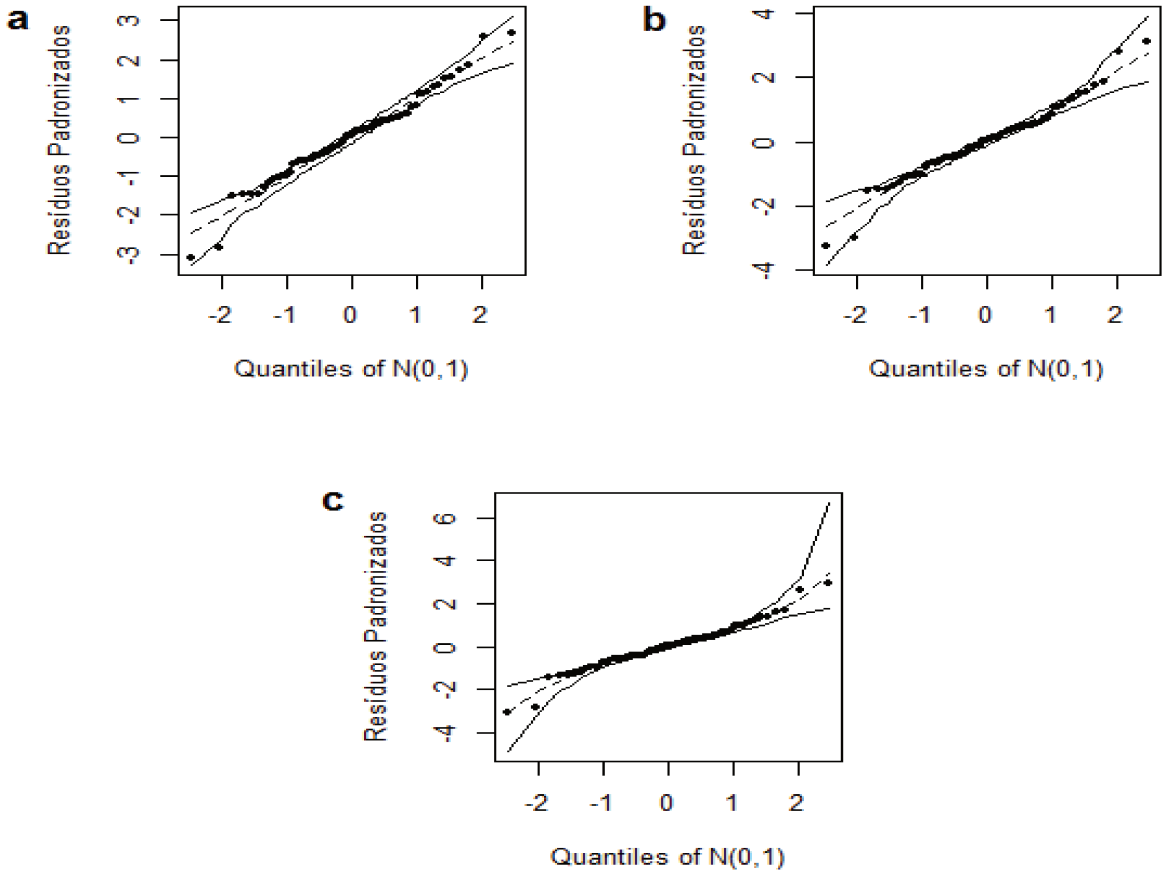


Figura 5.32: *Gráficos normais de probabilidades com envelope para os resíduos padronizados para o ajuste do modelo normal (a), logístico-II (b) e t-Student (c) ajustados aos dados dos coelhos europeus.*

observações atípicas para esses dados, entretanto existe a suspeita de *outliers* sob o ajuste de mínimos quadrados (Wei, 1998, Exemplo 6.8).

A Figura 5.32 mostra os gráficos normais de probabilidades dos resíduos padronizados. Observamos que para o modelo normal existe evidência de falta de ajustes, enquanto que o modelo t-Student parece ser mais adequado para modelar os dados.

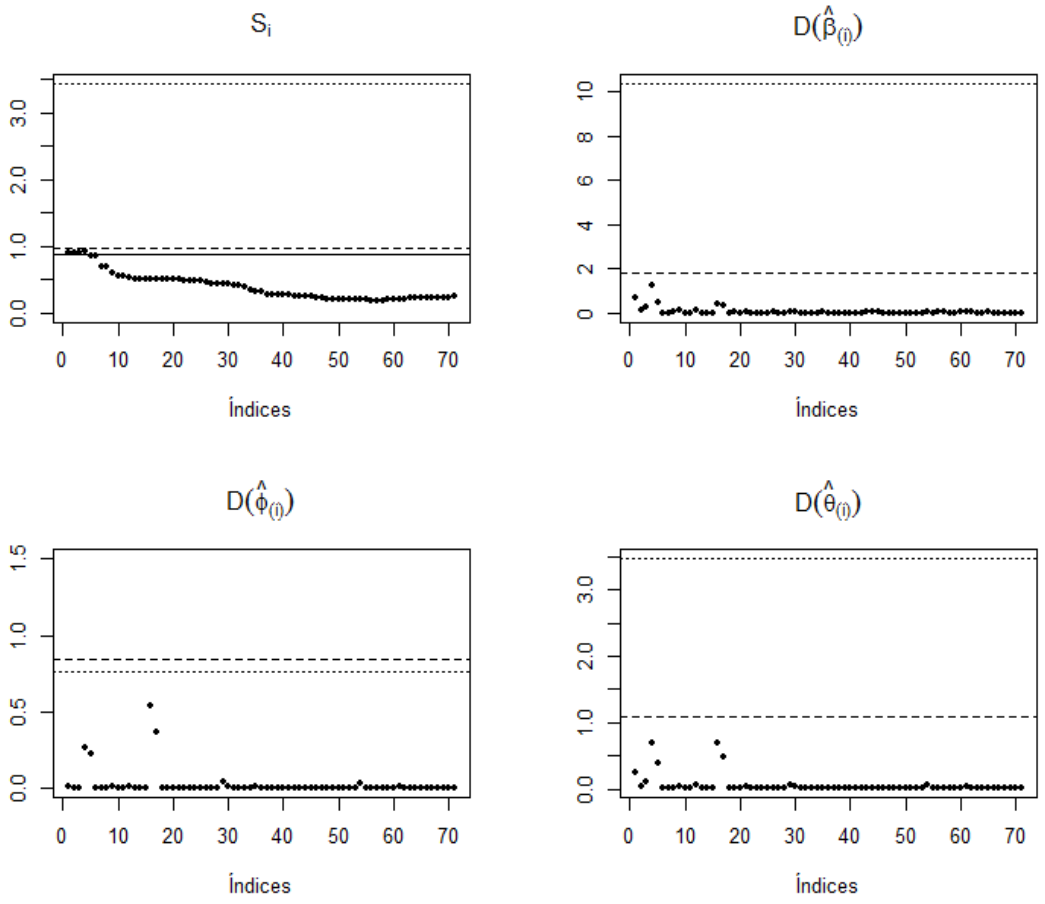


Figura 5.33: *Análise de influência para o ajuste dos dados de coelhos europeus sob erros normais.*

Para cada ajuste calculamos a estatística de Peña e a distância de Cook generalizada e as dispomos em gráficos. Como critério para identificar observações influentes calculamos os limites simulados M_1 (linha pontilhada) e M_{1_f} (linha tracejada) para cada estatística, além do limite de Peña (linha preta) que consta no gráfico da estatística de Peña. Para obter o limite simulado M_1 geramos valores para a variável explatória a partir da distribuição uniforme no intervalo $[\min(\mathbf{x}), \max(\mathbf{x})]$ e para obter M_{1_f} a variável explanatória foi fixada nos valores dos dados originais.

Observamos que não foram detectadas observações influentes em nenhum dos três ajustes segundo os limites simulados. Considerando as medidas da distância de Cook generalizada e sob a suposição de normalidade, Vanegas e Cysneiros (2010) notaram que as observações 1, 3, 4, 5, 16 e 17 aparecem como observações atípicas e com alta influência em $\hat{\boldsymbol{\beta}}$ e $\hat{\phi}$, mas a metodologia de detecção utilizada por eles difere das que estamos usando aqui.

Verificamos que o limite simulado M_1 em S_i e $D(\hat{\boldsymbol{\beta}}_{(i)})$ é muito maior que o limite M_{1_f}. De fato, o modelo (5.2.4) pode apresentar pontos de alavanca nos casos em que a variável explanatória apresenta valores menores do que o restante da amostra e consequentemente estes pontos apresentarão valores elevados para as estatísticas. O limite M_1 é obtido considerando $\mathbf{x}$ variável, o

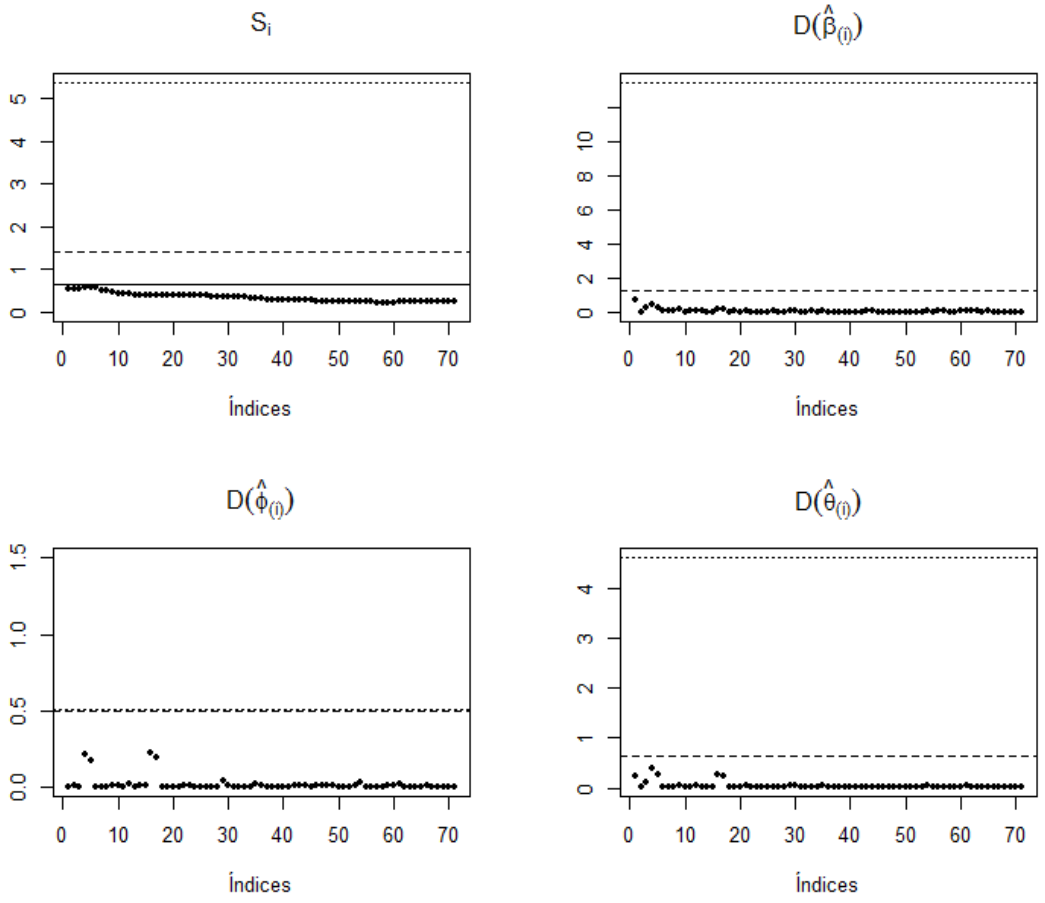


Figura 5.34: *Análise de influência para o ajuste dos dados de coelhos europeus sob erros Logísticos-II.*

que aumenta as chances de $\mathbf{x}$ apresentar valores pequenos. O limite M_{1_f} é obtido com $\mathbf{x}$ fixo nos valores dos dados originais, mas também apresentou valores grandes. Situações como essa pode dificultar a detecção de observações influentes via limites simulados.

As observações 4, 5, 16 e 17 são *outliers* segundo análise de resíduos padronizados feita por Galea et al. (2005). As observações 1 e 3 correspondem aos dados dos coelhos mais jovens e sob o ajuste do modelo normal são pontos de alavanca (Galea et al., 2005) mostrando a dificuldade de predição nesses casos. Para os ajustes do modelo logístico-II e t-Student, o efeito desses pontos diminui. Mas em geral, pontos de alavanca como esses podem apresentar resíduos pequenos (Laurent e Cook, 1992).

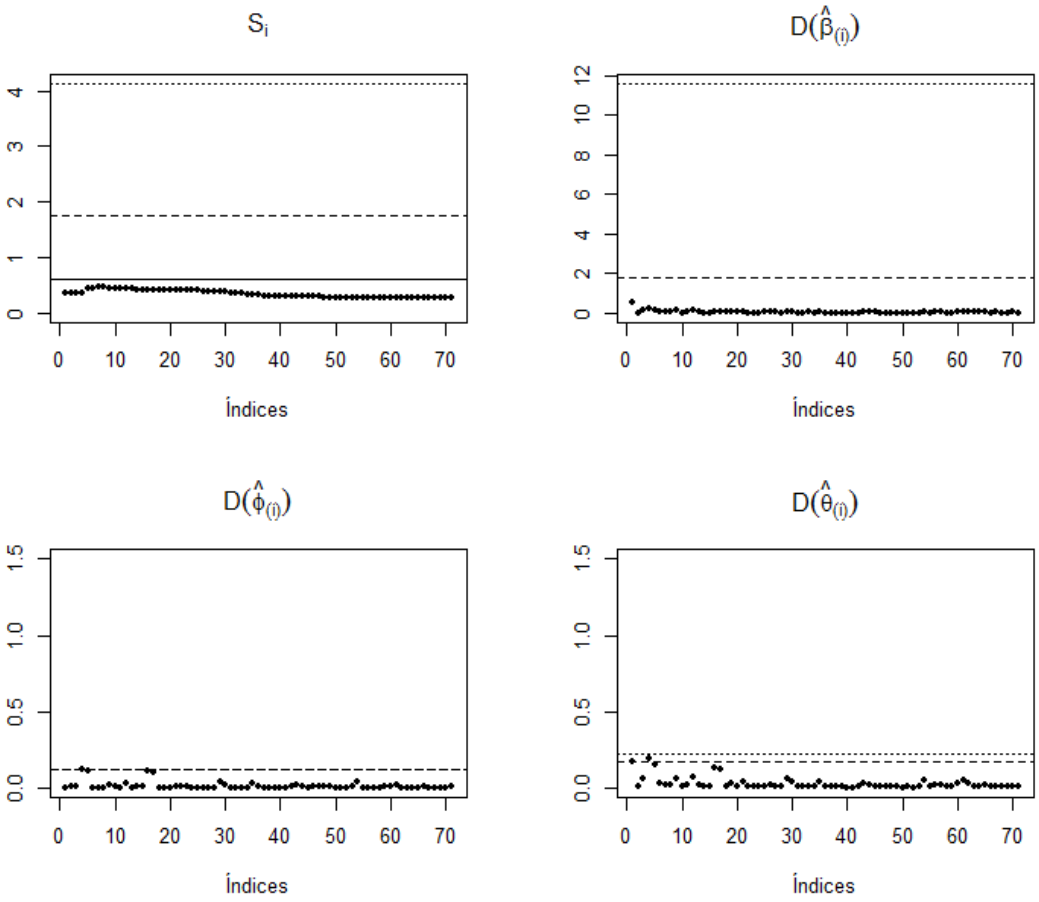


Figura 5.35: *Análise de influência para o ajuste dos dados de coelhos europeus sob erros t-Student.*

Capítulo 6

Conclusões e considerações finais

As principais contribuições desta dissertação são o dignóstico de influência em modelos lineares e não-lineares simétricos baseado no método proposto por Peña (2005) e o estudo acerca da utilização da metodologia de limiares simulados para caracterizar observações com influentes ou não, assim como o efeito da estimação dos parâmetro na construção dos limiares.

Foram simulados modelos lineares e não-lineares com erros t-Student e logístico-II e foram avaliados os desempenhos de três métodos de estimação em dados gerados por quatro situações: sem observações influentes, com *outliers* na variável resposta, com observações influentes de média alavancagem e com observações influentes de média alavancagem. Foram analisados o método de máxima verossimilhança com BFGS, o método de máxima verossimilhança com *scoring* de Fisher e um método robusto.

Os três métodos apresentam boas estimativas quando os dados não apresentam observações influentes no modelo logístico-II, mas no ajuste com o modelo t-Student verificamos que as estimativas para o grau de liberdade não são satisfatórias. No modelo t-Student, em geral, as estimativas dos parâmetros de regressão são boas, mas o parâmetro de escala é superestimado e o grau de liberdade foram subestimados no ajuste dos dados simulados sob a presença de observações influentes. No modelo logístico-II também verificamos a superestimação do parâmetro de escala sob a presença de observações influentes.

Para o diagnóstico de influência além da estatística de Peña consideramos a distância de Cook generalizada e a estatística de Andrews-Pregibon que foram aplicadas a dados simulados com incorporação de observações influentes e dados reais sob suspeita de observações influentes, juntamente com a técnica de limiares obtidos por simulação. Consideramos limites simulados obtidos com a variável explicativa fixa, com a variável explicativa gerada a partir de uma distribuição conhecida, sob controle dos valores originais da variável.

No ajuste de modelos lineares simétricos, verificamos que os limites simulados com a variável explicativa gerada em cada simulaçao foram mais eficazes na detecção de observações influentes quando comparados com os limites considerando a variável explicativa fixa. No ajuste de modelos não-lineres simétricos os limites simulados foram eficazes na detecção de observações com grandes desvios residuais. Sob o ajuste do modelo logístico-II a estatística de Peña apresentou uma grande variabilidade e somente o limite de Peña conseguiu detectar observações atípicas neste caso. Para obter os limites simulados considerando o ajuste do modelo t-Student, é importante fixar o grau

de liberdade na estimativa obtida com os dados originais, uma vez que os graus de liberdade são superestimados mesmo sem a presença de observações influentes.

Para finalizar, descrevemos os trbalhos futuros a seguir:

1. Aplicar a metodologia de influência proposta aos modelos de regressão linear e não-linear mistos com erros simétricos e assimétricos;

2. Aplicar a metodologia de influência aos modelos de regresão linar e não-linear multivariados.

Referências Bibliográficas

Andrews, D. F. and Pregibon, D. a. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):pp. 85–93. 2

Arellano-Valle, R. (1994). *Distrbuições Elípticas: Propriedades, Inferência e Aplicações a Modelos de Regressão. Tese de doutorado.* IME-USP, São Paulo, SP. 1, 5

Atkinson, A. C. (1985). *Plots, Transformation an Regression.* Clarendon Press, Oxford. 1

Beckman, R. J., Nachtsheim, C. J., and Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, 29(4):413–426. 2

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley, New York. 1

Castillo, E., Hadi, A., and Minguez, R. (2009). Diagnostics for non-linear regression. *Journal of Statistical Computation and Simulation*, 79(9):1109–1128. 88

Chalterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression.* John Wiley, New York. 1

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18. 1

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman & Hall, New York. 1

Cordeiro, G. M., Ferrari, S., Uribe-Opazo, M. A., and Vasconcellos, K. L. (2000). Corrected maximum-likelihood estimation in a class of symmetric nonlinear regression models. *Statistics & Probability Letters*, 46(4):317–328. 1, 5

Cox, D. and Hinkley, D. (1974). *Quantile regression.* Cambridge University Press, Cambridge. 11

Cysneiros, F. J. A. and Paula, G. A. (2005). Restricted methods in symmetrical linear regression models. *Computational Statistics & Data Analysis*, 49(3):689 – 708. 7, 8

Fang, K. and Anderson, T. (1990). *Statistical Inference in Elliptical Contoured and Related Distributions.* Allerton Press, New York. 1

Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric multivariate and related distributions*. Chapman & Hall, London. 1

Ferrari, S. and Uribe-Opazo, M. (2001). Corrected likelihood ratio tests in a classe symmetric linear regression models. *Brazilian Journal of Probability and Statistics*, 15:49–67. 1, 5

Galea, M., Paula, G. A., and B., H. (1997). Local influence in elliptical linear regression models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(1):71–79. 2

Galea, M., Paula, G. A., and Cysneiros, F. J. A. (2005). On diagnostics in symmetrical nonlinear models. *Statistics & Probability Letters*, 73(4):459–467. 2

Galea, M., Paula, G. A., and Uribe-Opazo, M. (2003). On influence diagnostic in univariate elliptical linear regression models. *Statistical Papers*, 44(1):23–45. 2

Galea, M., Riquelme, M., and Paula, G. (2000). Diagnostic methods in elliptical lnear regression models. *Brazilian Journal of Probability and Statistics*, 14:167–184. 2, 57

Kim, M. G. (1995). Local influence in multivariate regression. *Communications in Statistics - Theory and Methods*, 24(5):1271–1278. 2

Koenker, R. (2005). *Quantile regression*. Cambridge University Press, Cambridge. 11

Lange, K., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896. 1, 5

Lawrance, A. J. (1988). Regression transformation diagnostics using local influence. *Journal of the American Statistical Association*, 83(404):1067–1072. 2

Liu, S. (2000). On local influence for elliptical linear models. *Statistical Papers*, 41(2):211–224. 2

Maronna, R., Martin, R., and Yohai, V. (2006). *Robust statistics: Theory and methods*. John Wiley, New York. 5

Paula, G. A. (1993). Assessing local influence in restricted regression models. *Computational Statistics & Data Analysis*, 16(1):63 – 79. 2

Peña, D. (2005). A new statistic for influence in linear regression. *Technometrics*, 47(1):1–12. 2

Pregibon, D. (1981). Logistics regression diagnostics. *The annals of Statistics*, 9(4):705–724. 2

Ratkowsky, D. (1983). *Nonlinear Regression Modeling: A United Practical Approach*. Marcel Dekker, New York. 93

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley, New York. 73

Taylor, J. M. G. (1992). Properties of modelling the error distribution with an extra shape parameter. *Computational Statistics & Data Analysis*, 13(1):33 – 46. 1, 5

Thomas, W. and Cook, R. D. (1990). Assessing influence on predictions from generalized linear models. *Technometrics*, 32(1):59–65. 2

Tsai, C.-L. and Wu, X. (1992). Assessing local influence in linear regression models with first-order autoregressive or heteroscedastic error structure. *Statistics & Probability Letters*, 14(3):247–252. 2

Vanegas, L. H. and Cysneiros, F. J. A. (2010). Assessment of diagnostic procedures in symmetrical nonlinear regression models. *Computational Statistics and Data Analysis*, 54(4):1002 – 1016. 2

Wood, F. S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics*, 15(4):677–695. 80

Apêndice A

Informações adicionais

A.1 Modelo de exclusão de casos

Considere o modelo de exclusão de casos

$$y_j = \mu(\mathbf{x}_j^{\mathrm{T}}; \boldsymbol{\beta}) + \epsilon_j, \quad j \neq i, \tag{A.1.1}$$

onde $\mu(\cdot)$ é uma função (linear ou não-linear) contínua e diferenciável com respeito à $\boldsymbol{\beta}$, $\mathbf{x}_j^{\mathrm{T}}$ um vetor de variáveis explanatórias, $\epsilon_i \sim S(0, \phi, g)$, erros independentes e identicamente distribuídos e $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \phi)^{\mathrm{T}}$ o vetor de parâmetros desconhecidos. Definimos a estimativa de máxima verossimilhança de $\boldsymbol{\theta}$ no modelo (A.1.1) como $\hat{\boldsymbol{\theta}}_{(i)}$. Mas o cálculo de $\hat{\boldsymbol{\theta}}_{(i)}$, $i = 1, \ldots, n$, pode ser computacionalmentte caro, quando n é grande.

Seja $L_{(i)}(\boldsymbol{\theta})$ a função log-verrossimilança e $\mathbf{U}_{(i)}(\boldsymbol{\theta})$ a função escore referente ao modelo (A.1.1). Podemos obter a aproximação em um passo, $\hat{\boldsymbol{\theta}}^1_{(i)}$, de $\hat{\boldsymbol{\theta}}_{(i)}$ maximizando a expansão de $L_{(i)}(\boldsymbol{\theta})$ em uma série de Taylor de segunda ordem em torno de $\hat{\boldsymbol{\theta}}$. Em outras palavras, a aproximação quadrática de $L_{(i)}(\boldsymbol{\theta})$ é maximizada em

$$\hat{\boldsymbol{\theta}}^1_{(i)} = \hat{\boldsymbol{\theta}} + \hat{\mathbf{J}}^{-1}_{(i)} \hat{\mathbf{U}}_{(i)} \tag{A.1.2}$$

onde

$$\hat{\mathbf{J}}_{(i)} \stackrel{\text{def}}{=} \mathbf{J}_{(i)}(\hat{\boldsymbol{\theta}}) = - \left. \frac{\partial^2 L_{(i)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \quad \text{e} \quad \hat{\mathbf{U}}_{(i)} \stackrel{\text{def}}{=} \mathbf{U}_{(i)}(\hat{\boldsymbol{\theta}}).$$

Isso é equivalente a um único passo dos algoritmos tipo Newton usando $\hat{\boldsymbol{\theta}}$ como valor inicial (Cook e Weisberg, 1982, pg. 182).

A função $\mathbf{U}_{(i)}(\hat{\boldsymbol{\theta}})$ é facilmente obtida considerando que

$$\mathbf{U}(\hat{\boldsymbol{\theta}}) = \mathbf{U}^i(\hat{\boldsymbol{\theta}}) + \sum_{j \neq i}^{n} \mathbf{U}^j(\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

e

$$\mathbf{U}_{(i)}(\hat{\boldsymbol{\theta}}_{(i)}) = \sum_{j \neq i}^{n} \mathbf{U}^j(\hat{\boldsymbol{\theta}}_{(i)}) = \mathbf{0},$$

onde $\mathbf{U}^i(\boldsymbol{\theta}) = \partial l(y_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Logo,

$$\mathbf{U}_{(i)}(\hat{\boldsymbol{\theta}}) = -\mathbf{U}^i(\hat{\boldsymbol{\theta}}). \tag{A.1.3}$$

A matriz $\hat{\mathbf{J}}_{(i)}^{-1}$ pode ser substituida pela inversa da matriz de informação de Fisher avaliado em $\hat{\boldsymbol{\theta}}$, $\mathbf{K}^{(i)-1}(\hat{\boldsymbol{\theta}})$. Com essa substituição é possível obter uma expressão para $\hat{\boldsymbol{\theta}}^1_{(i)}$. Considerando $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\top}, \phi)^{\top}$, temos que

$$[\mathbf{K}^{(i)}(\hat{\boldsymbol{\theta}})]^{-1} = \begin{pmatrix} \mathbf{K}^{(i)-1}_{\hat{\beta}\hat{\beta}} & \mathbf{0} \\ \mathbf{0} & \mathrm{K}^{(i)-1}_{\hat{\phi}\hat{\phi}} \end{pmatrix},$$

em que $\mathbf{K}^{(i)}_{\hat{\beta}\hat{\beta}} = \frac{4d_g}{\hat{\phi}}\mathbf{D}^{\top}_{\hat{\beta}}\Delta_i\mathbf{D}_{\hat{\beta}}$, $\mathrm{K}^{(i)}_{\hat{\phi}\hat{\phi}} = \frac{n-1}{4\hat{\phi}^2}(4f_g - 1)$, $\Delta_i = \mathrm{diag}\{\delta_1, \ldots, \delta_n\}$, $\delta_i = 0$, $\delta_j = 1$, para todo $j \neq i$, e $\mathbf{D}_{\hat{\beta}} = \left.\frac{\partial\boldsymbol{\mu}}{\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$.

Assim, sabendo que

$$\begin{aligned}
\mathbf{K}^{(i)-1}_{\hat{\beta}\hat{\beta}} &= \frac{\hat{\phi}}{4d_g}\left(\mathbf{D}^{\top}_{\hat{\beta}}\Delta_i\mathbf{D}_{\hat{\beta}}\right)^{-1} \\
&= \frac{\hat{\phi}}{4d_g}\left[\mathbf{D}^{\top}_{\hat{\beta}}\mathbf{D}_{\hat{\beta}} - \hat{\mathbf{d}}_i\hat{\mathbf{d}}_i^{\top}\right]^{-1} \\
&= \frac{\hat{\phi}}{4d_g}\left[\left(\mathbf{D}^{\top}_{\beta}\mathbf{D}_{\hat{\beta}}\right)^{-1} + \frac{\left(\mathbf{D}^{\top}_{\hat{\beta}}\mathbf{D}_{\hat{\beta}}\right)^{-1}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i^{\top}\left(\mathbf{D}^{\top}_{\hat{\beta}}\mathbf{D}_{\hat{\beta}}\right)^{-1}}{1-\hat{h}_{ii}}\right],
\end{aligned} \tag{A.1.4}$$

$$\hat{\mathbf{U}}^i_{\beta} = \frac{\hat{\upsilon}_i\hat{z}_i\mathbf{x}_i}{\sqrt{\hat{\phi}}} \quad \text{e} \quad \hat{\mathbf{U}}^i_{\phi} = (2\hat{\phi})^{-1}(\hat{\upsilon}_i\hat{z}_i^2 - 1).$$

Portanto, substituindo (A.1.4) e $\mathbf{U}_{(i)}(\hat{\boldsymbol{\theta}})$ em (A.1.2), obtemos

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^1_{(i)} = \begin{bmatrix} \hat{\boldsymbol{\beta}}^1_{(i)} \\ \hat{\phi}^1_{(i)} \end{bmatrix} &= \begin{bmatrix} \hat{\boldsymbol{\beta}} - \hat{\mathbf{K}}^{(i)-1}_{\beta\beta}\hat{\mathbf{U}}^i_{\beta} \\ \hat{\phi} - \hat{\mathbf{K}}^{(i)-1}_{\phi\phi}\hat{\mathbf{U}}^i_{\phi} \end{bmatrix} \\
&= \begin{bmatrix} \hat{\boldsymbol{\beta}} - \hat{\phi}^{1/2}(4d_g)^{-1}\hat{\upsilon}_i\hat{z}_i\left(\mathbf{D}^{\top}_{\beta}\mathbf{D}_{\hat{\beta}}\right)^{-1}\hat{\mathbf{d}}_i/(1-\hat{h}_{ii}) \\ \hat{\phi} - 2\hat{\phi}(\hat{\upsilon}_i\hat{z}_i^2 - 1)/(n-1)(4f_g - 1) \end{bmatrix}.
\end{aligned} \tag{A.1.5}$$

onde $\hat{\mathbf{d}}_i^{\top}$ é a i-ésima linha de $\mathbf{D}_{\hat{\beta}}$ e $\hat{h}_{ii} = \hat{\mathbf{d}}_i^{\top}\left(\mathbf{D}^{\top}_{\hat{\beta}}\mathbf{D}_{\hat{\beta}}\right)^{-1}\hat{\mathbf{d}}_i$. Os resultados aqui apresentados podem ser encontrados em Vanegas e Cysneiros (2010).

Para um modelo de regressão linear simétrico, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, segue que $\mathbf{D}_{\beta} = \mathbf{X}$ de modo que

$$\hat{\boldsymbol{\theta}}^{1}_{(i)} = \begin{bmatrix} \hat{\boldsymbol{\beta}}^{1}_{(i)} \\ \hat{\phi}^{1}_{(i)} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\beta}} - \hat{\phi}^{1/2}(4d_g)^{-1}\hat{\upsilon}_i\hat{z}_i \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{x}_i/(1-h_{ii}) \\ \hat{\phi} - 2\hat{\phi}(\hat{\upsilon}_i\hat{z}_i^2 - 1)/(n-1)(4f_g - 1) \end{bmatrix}. \tag{A.1.6}$$

onde $h_{ii} = \mathbf{x}_i^{\top}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{x}_i$ e $\mathbf{x}_i^{\top}$ é a i-ésima linha da matriz $\mathbf{X}$. Por meio dos resultados em (A.1.5) e (A.1.6) podemos obter medidas como distância de Cook generalizda e estatística de Peña para avaliar influência em modelos simétricos não-lineares e lineares, respectivamente, sob a abordagem do modelo de exclusão de casos.

A.2 Uma aproximação para distância de Cook generalizada

Sabemos que a distância de Cook generalizada para os subconjuntos de parâmetros $\boldsymbol{\beta}$ e ϕ pode assume a forma

$$\begin{aligned} D(\hat{\boldsymbol{\beta}}_{(i)}) &= (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^{\top}\mathbf{K}_{\hat{\beta}\hat{\beta}}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}) \\ D(\hat{\phi}_{(i)}) &= (\hat{\phi}_{(i)} - \hat{\phi})^2 K_{\hat{\phi}\hat{\phi}}. \end{aligned}$$

Vanegas e Cysneiros (2010), usando as equações em (A.1.5), obtiveram uma aproximação de $D(\hat{\boldsymbol{\beta}}_{(i)})$ e $D(\hat{\phi}_{(i)})$, respectivamente

$$D(\hat{\boldsymbol{\beta}}^{1}_{(i)}) = \frac{(4d_g)^{-1}\hat{\upsilon}_i^2\hat{z}_i^2\hat{h}_{ii}}{(1-\hat{h}_{ii})^2} \tag{A.2.1}$$

e

$$D(\hat{\boldsymbol{\phi}}^{1}_{(i)}) = \frac{n(\hat{\upsilon}_i\hat{z}_i^2 - 1)^2}{(n-1)^2(4f_g - 1)} \tag{A.2.2}$$

As expressões (A.2.1) e (A.2.2) extendem a distância de Cook, estudada nos modelos de regressão normal, aos modelos não-lineares simétricos. Além disso, podemos verificar que a influência da i-ésima observação em $\hat{\boldsymbol{\theta}}$ depende de υ_i que varia de acordo com a distribuição do erro do modelo. Para os modelos lineares simétricos podemos utilizar as equações em (A.1.6) e de forma análoga obter as respectivas aproximações.

A.3 Uma aproximação para estatística de Peña

A estatística de Peña (2005) é dada por

$$\mathrm{S}_i = \frac{\sum_{j=1}^{n}(\hat{y}_i - \hat{y}_{i(j)})^2}{p\widehat{\mathrm{Var}}(\hat{y}_i)}. \tag{A.3.1}$$

em que p é o número de parâmetros da função de regressão e $\widehat{\mathrm{Var}}(\hat{y}_i)$ é a estimativa da variância do i-ésimo valor predito.

Para os modelos lineares sob erros com distribuição normal, Peña (2005) mostrou que (A.3.1) pode ser reescrita como

$$\mathrm{S}_i = \frac{1}{ps^2 h_{ii}} \sum_{j=1}^{n} \frac{h_{ij}^2 e_j^2}{(1-h_{ii})^2}, \tag{A.3.2}$$

em que $e_j = y_j - \hat{y}_j$ e $s^2 = \mathbf{e}'\mathbf{e}/(n-p)$, para $\mathbf{e} = (e_1, \ldots, e_n)^{\mathrm{T}}$.

Afim de extender a abordagem tratada por Peña(2005) aos modelos simétricos lineares, consideramos as expressões em (A.1.5) e que o estimador para variância de $\hat{y}_i$ é igual a $\frac{\hat{\phi}}{4d_g} h_{ii}$ obtemos

$$\begin{aligned} \mathrm{S}_i^1 &= \frac{1}{p} \sum_{j=1}^{n} \frac{h_{ij}^2}{h_{jj} h_{ii}} \frac{(4d_g)^{-1} \upsilon_j^2 \hat{z}_j^2 h_{jj}}{(1-h_{jj})^2} \\ &= \frac{1}{p} \sum_{j=1}^{n} \rho_{ij}^2 D(\boldsymbol{\beta}_{(j)}^1) \end{aligned} \tag{A.3.3}$$

em que $\rho_{ij} = h_{ij}/\sqrt{h_{jj} h_{ii}}$.

Prova:

Considerando que um estimador consistente para a matriz de variância-covariância assintótica de $\hat{\boldsymbol{\beta}}$ é $\mathbf{K}_{\hat{\beta}\hat{\beta}}^{-1} = \frac{\hat{\phi}}{4d_g}(\mathbf{X}^T\mathbf{X})^{-1}$, temos que um estimador para matriz de variância-covariância de $\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}) = \mathbf{X}\hat{\boldsymbol{\beta}}$ é $\frac{\hat{\phi}}{4d_g}\mathbf{H}$, onde $\mathbf{H} = \mathbf{X}^T(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}^T$. Portanto, um estimador para variância de $\hat{y}_i$ é igual a $\frac{\hat{\phi}}{4d_g} h_{ii}$.

Lembrando que $\hat{y}_{i(j)} = \mu_i(\hat{\boldsymbol{\beta}}_{(j)}) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(j)}$, temos que

$$y_i - \hat{y}_{i(j)} = \mathbf{x}_i^T \left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(j)}\right]. \tag{A.3.4}$$

Segundo a expressão em (A.1.5)

$$\hat{\boldsymbol{\beta}}_{(j)}^1 = \hat{\boldsymbol{\beta}} - \frac{\hat{\phi}^{1/2} \hat{\upsilon}_j \hat{z}_j \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_j}{4d_g(1-h_{jj})}.$$

Substituindo $\hat{\boldsymbol{\beta}}_{(j)}$ por $\hat{\boldsymbol{\beta}}_{(j)}^1$ em (A.3.4), obtemos

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(j)} &= \hat{\boldsymbol{\beta}} - \left[\hat{\boldsymbol{\beta}} - \frac{\hat{\phi}^{1/2}\hat{\upsilon}_j\hat{z}_j\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_j}{1-h_{jj}}\right] \\ &= \frac{\hat{\phi}^{1/2}\rho(\hat{z}_j)\hat{z}_j\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_j}{4d_g(1-h_{jj})}. \end{aligned} \tag{A.3.5}$$

Usando (A.3.5) em (A.3.4), segue que

$$y_i - \hat{y}_{i(j)} = \frac{\hat{\phi}^{1/2}\hat{\upsilon}_j\hat{z}_j h_{ij}}{4d_g(1-h_{jj})}.$$

Portanto, uma aproximação para S_i é definida por

$$\begin{aligned}\mathrm{S}_i^1 &= \frac{4d_g}{p\hat{\phi}h_{ii}}\sum_{j=1}^{n}\frac{\hat{\phi}(4d_g)^{-2}\hat{\upsilon}_j^2\hat{z}_j^2h_{ij}^2}{(1-h_{jj})^2} \\ &= \frac{1}{p}\sum_{j=1}^{n}\frac{h_{ij}^2}{h_{jj}h_{ii}}\frac{(4d_g)^{-1}\hat{\upsilon}_j^2\hat{z}_j^2h_{jj}}{(1-h_{jj})^2} \\ &= \frac{1}{p}\sum_{j=1}^{n}\delta_{ij}^2 D(\hat{\boldsymbol{\beta}}_{(j)}^1)\end{aligned}$$

A última igualdade em (A.3.3) mostra que uma aproximação de S_i pode ser obtida usando uma combinação linear de $D(\boldsymbol{\beta}_{(j)}^1)$ definida em (A.2.1). No caso normal linear S_i^1 é exatamente igual a S_i, pois $\upsilon_j = 1$ para todo $j = 1, \ldots, n$.

A.4 Matriz de informação de Fisher

Para modelos da família t-Student e exponencial potência, tem-se as respectivas densidades

$$f(y_i) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\phi\nu}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{u_i}{\nu}\right)^{\frac{\nu+1}{2}}, \quad 0<\nu\leq\infty$$

e

$$f(y_i) = C_k\phi^{-1/2}\exp\left\{-\frac{1}{2}u_i^{\frac{1}{1+\nu}}\right\}, \quad -1<\nu<1,$$

em que $u_i = (y_i-\mu_i)^2/\phi$, $\mu_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ é o parâmetro de localização com $\mathbf{x}_i^{\mathrm{T}} = (1, x_{i1}, \ldots, x_{ip})$ e $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$, $\phi>0$ o parâmetro de escala e $C_k^{-1} = \Gamma\left(1+\frac{1+\nu}{2}\right)2^{1+\frac{1+\nu}{2}}$.

Além disso, considere

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} l(y_i;\boldsymbol{\theta}) = \sum_{i=1}^{n}\log\left(f_{Y_i}(y_i)\right)$$

a função log-verossimilhança.

Para o modelo t-Student

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right) = \frac{\nu+1}{\phi(\nu+3)}\mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{A.4.1}$$

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\phi}\right) = E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\nu}\right) = \mathbf{0} \tag{A.4.2}$$

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\phi^2}\right) = n\frac{\nu}{2\phi^2(\nu+3)} \tag{A.4.3}$$

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\nu^2}\right) = -\frac{n}{2}\left[\frac{1}{2}\Psi'\left(\frac{\nu+1}{2}\right)-\frac{1}{2}\Psi'\left(\frac{\nu}{2}\right)+\frac{1-\nu}{\nu(\nu+1)}+\frac{\nu+2}{\nu(\nu+3)}\right] \tag{A.4.4}$$

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\nu\partial\phi}\right) = n\frac{\phi^{-1}}{(\nu+3)(\nu+1)}. \tag{A.4.5}$$

onde $\Psi'(\cdot)$ é a função trigama. A prova para as expressões (A.4.1) a (A.4.5) pode ser encontrada em Lange et al. (1989). Neste caso, a inversa da matriz de informação de Fisher é dada por

$$\mathbf{K}_{\theta\theta}^{-1} = \begin{bmatrix} \mathbf{K}_{\beta\beta}^{-1} & \mathbf{0}_{p\times 2} \\ \mathbf{0}_{2\times p} & \mathbf{K}_{\tau\tau}^{-1} \end{bmatrix} \tag{A.4.6}$$

com $\boldsymbol{\tau} = (\phi,\nu)^{\mathrm{T}}$,

$$\mathbf{K}_{\beta\beta}^{-1} = \frac{\phi(\nu+3)}{\nu+1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \quad \text{e}$$

$$\mathbf{K}_{\tau\tau} = \begin{bmatrix} E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\phi^2}\right) & E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\phi\partial\nu}\right) \\ E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\nu\partial\phi}\right) & E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\nu^2}\right) \end{bmatrix}.$$

Para o modelo expnencial potência

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right) = \frac{2^{-\nu}\Gamma\left(1+\frac{1-\nu}{2}\right)}{\phi(1+\nu)\Gamma\left(\frac{3+\nu}{2}\right)}\mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{A.4.7}$$

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\phi}\right) = E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\nu}\right) = \mathbf{0} \tag{A.4.8}$$

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\phi^2}\right) = \frac{n}{2\phi^2(1+\nu)} \tag{A.4.9}$$

$$\begin{aligned} E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\nu^2}\right) &= \frac{n}{2}\left[\frac{1}{2}\Psi'\left(\frac{3+\nu}{2}\right) + \frac{\left(\log(2)+\Psi\left(\frac{3+\nu}{2}\right)\right)^2}{4(1+\nu)}\right] + \\ &+ \frac{n}{2}\left[\frac{\Psi'\left(\frac{3+\nu}{2}\right)}{4(1+\nu)} + \frac{\log(2)+\Psi\left(\frac{3+\nu}{2}\right)}{2(1+\nu)}\right] \end{aligned} \tag{A.4.10}$$

$$E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\nu\partial\phi}\right) = n\frac{1-\log(2)-\Psi\left(\frac{3+\nu}{2}\right)}{2\phi(1+\nu)}. \tag{A.4.11}$$

em que $\Psi(\cdot)$ denota a função digama. Os resultados de (A.4.7) a A.4.9 podem ser encontrados em Taylor (1992). Considerando que

$$\frac{\partial^2 l}{\partial\phi\partial\nu} = \frac{(2\phi)^{-1}}{(1+\nu)^2}\left[\frac{u_i^{\frac{1}{1+\nu}}\log(u_i)}{1+\nu} - u_i^{\frac{1}{1+\nu}}\right], \tag{A.4.12}$$

$$\frac{\partial^2 l}{\partial\nu^2} = -\frac{1}{2}\left[\frac{1}{2}\Psi'\left(1+\frac{1+\nu}{2}\right) + \frac{u_i^{\frac{1}{1+\nu}}\left[\log(u_i)\right]^2}{(1+\nu)^4} + \frac{2u_i^{\frac{1}{1+\nu}}\log(u_i)}{(1+\nu)^3}\right] \tag{A.4.13}$$

e substituindo u_i por $|z_i|^2$, com $z_i = (y_i-\mu_i)/\sqrt{\phi}$, podemos obter o resultado em (A.4.10) e (A.4.11) calculando

$$E\left(|Z|^{\frac{2}{1+\nu}}\right) \tag{A.4.14}$$

$$E\left(|Z|^{\frac{2}{1+\nu}}\log(|Z|^2)\right) = 2E\left(|Z|^{\frac{2}{1+\nu}}\log(|Z|)\right) \quad \text{e} \tag{A.4.15}$$

$$E\left(|Z|^{\frac{2}{1+\nu}}\left[\log(|Z|^2)\right]^2\right) = 4E\left(|Z|^{\frac{2}{1+\nu}}\left[\log(|Z|)\right]^2\right). \tag{A.4.16}$$

Para $Z \sim EP(\mu = 0, \phi = 1)$

$$\begin{aligned} E\left(|Z|^r\right) &= \int_{-\infty}^{+\infty} |z|^r C_k \exp\left\{-\frac{|z|^{\frac{2}{1+\nu}}}{2}\right\} dz \\ &= 2^{\gamma}(1+\nu)C_k \int_0^{+\infty} \mathrm{v}^{\gamma-1}\exp\left\{-\mathrm{v}\right\} d\mathrm{v} \\ &= \frac{2^{r\frac{1+\nu}{2}}\Gamma(\gamma)}{\Gamma\left(\frac{1+\nu}{2}\right)}, \end{aligned} \tag{A.4.17}$$

$$\begin{aligned} E\left(|Z|^r\log(|Z|)\right) &= \int_{-\infty}^{+\infty} |z|^r \log(|z|) C_k \exp\left\{-\frac{|z|^{\frac{2}{1+\nu}}}{2}\right\} dz \\ &= 2^{\gamma-1}(1+\nu)^2 C_k \int_0^{+\infty} \mathrm{v}^{\gamma-1}\left[\log(2)+\log(\mathrm{v})\right]\exp\left\{-\mathrm{v}\right\} d\mathrm{v} \\ &= \frac{(1+\nu)2^{r\frac{1+\nu}{2}-1}}{\Gamma\left(\frac{1+\nu}{2}\right)}\left[\Gamma(\gamma)\log(2)+\Gamma'(\gamma)\right], \end{aligned} \tag{A.4.18}$$

$$\begin{aligned} E\left(|Z|^r\left[\log(|Z|)\right]^2\right) &= \int_{-\infty}^{+\infty} |z|^r \left[\log(|z|)^2\right] C_k \exp\left\{-\frac{|z|^{\frac{2}{1+\nu}}}{2}\right\} dz \\ &= 2^{\gamma-2}(1+\nu)^3 C_k \int_0^{+\infty} \mathrm{v}^{\gamma-1}\left[\log^2(2)+2\log(2)\log(\mathrm{v})+\log^2(\mathrm{v})\right]\exp\left\{-\mathrm{v}\right\} d\mathrm{v} \\ &= \frac{(1+\nu)^2 2^{r\frac{1+\nu}{2}-2}}{\Gamma\left(\frac{1+\nu}{2}\right)}\left[\Gamma(\gamma)\log^2(2)+2\log(2)\Gamma'(\gamma)+\Gamma''(\gamma)\right], \end{aligned} \tag{A.4.19}$$

com $\gamma = (r+1)\frac{1+\nu}{2}$, $\Gamma'(\gamma) = \Psi(\gamma)\Gamma(\gamma)$ e $\Gamma''(\gamma) = \left[\Psi'(\gamma)+\Psi^2(\gamma)\right]\Gamma(\gamma)$. Assim, as esperanças de (A.4.14) a (A.4.16) tornam-se

$$E\left(|Z|^{\frac{2}{1+\nu}}\right) = 1+\nu \tag{A.4.20}$$

$$E\left(|Z|^{\frac{2}{1+\nu}}\log(|Z|^2)\right) = (1+\nu)^2\left[\log(2)+\Psi\left(\frac{3+\nu}{2}\right)\right] \quad \text{e} \tag{A.4.21}$$

$$E\left(|Z|^{\frac{2}{1+\nu}}\left[\log(|Z|^2)\right]^2\right) = (1+\nu)^3\left\{\left[\log(2)+\Psi\left(\frac{3+\nu}{2}\right)\right]^2+\Psi'\left(\frac{3+\nu}{2}\right)\right\}. \tag{A.4.22}$$

Portanto, com as expressões de (A.4.20) a (A.4.22) obtemos (A.4.10) e (A.4.11).
Neste caso, a inversa da matriz de informação de Fisher é dada por

$$\mathbf{K}_{\theta\theta}^{-1} = \begin{bmatrix} \mathbf{K}_{\beta\beta}^{-1} & \mathbf{0}_{p\times 2} \\ \mathbf{0}_{2\times p} & \mathbf{K}_{\tau\tau}^{-1} \end{bmatrix} \tag{A.4.23}$$

com $\boldsymbol{\tau} = (\phi,\nu)^{\top}$,

$$\mathbf{K}_{\beta\beta}^{-1} = \frac{\phi(1+\nu)\Gamma\left(\frac{3+\nu}{2}\right)}{2^{-\nu}\Gamma\left(1+\frac{1-\nu}{2}\right)}(\mathbf{X}^{\top}\mathbf{X})^{-1} \quad \text{e}$$

$$\mathbf{K}_{\tau\tau} = \begin{bmatrix} E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\phi^2}\right) & E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\phi\partial\nu}\right) \\ E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\nu\partial\phi}\right) & E\left(-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\nu^2}\right) \end{bmatrix}.$$