
Influência local com procura "forward" em
modelos de regressão linear.

Juan Pablo Mamani Bustamante

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Influência local com procura "forward" em modelos de regressão linear.

Juan Pablo Mamani Bustamante

Orientadora: Profa. Dra. Reiko Aoki

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

USP/UFSCar – São Carlos
Março de 2015

**Ficha catalográfica elaborada pelo DePT da
Biblioteca Comunitária da UFSCar**

M263iL Mamani Bustamante, Juan Pablo.
Influência local com procura "*forward*" em modelos de regressão linear / Juan Pablo Mamani Bustamante. -- São Carlos : UFSCar, 2015.
120 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2015.

1. Estatística. 2. Influência local. 3. Procura *forward*. 4. Curvatura normal conformal. 5. Modelos de regressão. I. Título.

CDD: 519.5 (20^a)

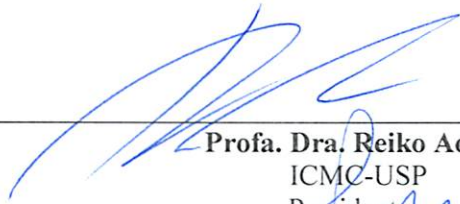
JUAN PABLO MAMANI BUSTAMANTE

INFLUÊNCIA LOCAL COM PROCURA "FORWARD" EM MODELOS DE REGRESSÃO LINEAR

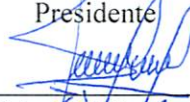
Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovado em 25 de fevereiro de 2015.

COMISSÃO JULGADORA:



Prof. Dra. Reiko Aoki
ICMC-USP
Presidente



Prof. Dr. Filidor Edilfonso Vilca Labra
UNICAMP
Membro



Prof. Dr. Cristian Marcelo Villegas Lobos
ESALQ-USP
Membro

*A Deus e à minha linda
família.*

Agradecimentos

Em primeiro lugar agradeço a Deus por me devolver a saúde e librar-me há muito tempo.

À minha família, meu pai Feliciano Mamani, minha mãe Juana Bustamante e minha irmã Silvia por querer-me muito.

À Professora Reiko Aoki, minha querida orientadora, por me entender e quem me ensino muitas coisas da estatística, pelas orientações na elaboração deste trabalho e também pela paciência, apoio e amizade durante todo o mestrado.

Ao meu amigo Marcelo Hartmann, que ensino muitas coisas no mestrado tanto da vida como no acadêmico, e pela amizade.

Ao meu amigo Vinícius Siqueira, pela ajuda, piadas, risos e aguentar-me durante todo o mestrado.

À minha querida amiga Marina Mitie Gishifu Osio quem foi quase como uma mãe na estadia no Brasil desde que a conheci, pela grande amizade e respeito.

À minha querida amiga Cristel por compartilhar e trocar conhecimento, pelas piadas, alegrias e risos durante todo o mestrado.

À minha querida amiga Lorena e a Miguel por ajudar-me e ensinar-me muitas vezes naqueles dias em que não entendi nada do que estava fazendo.

Ao meu amigo Luciano porque se por ele não estaria aqui terminando os estudos porque ele foi um parceiro, amigo e meu camarada para vir a São Carlos.

Aos meus amigos Efrain Candia e Sayda Chahuasonco porque eles foram como meus pais quando eu comecei trabalhar na universidade em Madre de Dios-Perú.

Ao meu amigo Norbil e a Patricia por dar-me moradia muitas vezes em sua casa aqui em São Carlos.

Aos meus caros amigos Thales Ricarte, Amélia Fernandes e John Garavito por compartilhar e trocar conhecimento, pelas ensinança da estatística e alegrias. Seria bom se eles mudassem mas esta bom como eles são.

Aos amigos Renato, Evandro, Victor, Felipe, Vanessa Rufino, Vanessa Masitéli, Arthur, Ricardo e Amanda, à turma da estatística, e a tantos outros que não mencionei aqui mas que me ajudaram de alguma forma no meu aprendizado.

Aos professores que participaram da banca do exame de qualificação e/ou da defesa de mestrado: Cibele Maria Russo (ICMC-USP), Cristian Marcelo Villegas Lobos (ESALQ-USP) e Filidor Edilfonso Vilca Labra (UNICAMP), que muito contribuíram com as correções e sugestões desse trabalho.

A todos os professores que contribuíram na minha formação e aos funcionários do (ICMC-USP) e (DEs-UFSCar).

Agradeço aos amigos Jose (Trujillo) e Carlos (Puno) pela convivência que nós tivemos na republica.

Agradeço aos amigos Carlos Franklin Taco, Jhon Bernedo, George Lucas e Paulo Seminario pelos jogos que compartilhamos.

Agradeço a Maria Baldassari uma amiga espiritual quem estive no momento exato em que mais precisava de um conselho.

Por ultimo agradeço à todas as pessoas, amigos de longe e de perto, que contribuíram no desenvolvimento do trabalho.

Resumo

A identificação de observações influentes e/ou aberrantes de um conjunto de dados é conhecida como uma parte das análises de diagnóstico. Esta técnica de diagnóstico têm como uma das finalidades verificar a robustez de um modelo estatístico, pois a não identificação dos dados influentes pode afetar a análise ou obter resultados incorretos.

As metodologias comumente utilizadas para o diagnóstico de observações influentes em modelos de regressão são métodos de influência global (Belsey et al., 1980). Cook (1986) introduziu um método geral para avaliar a influência local de pequenas perturbações no modelo estatístico ou nos dados, usando diferentes tipos de perturbações. Como complemento às técnicas de detecção de observações discrepantes, é proposto o método procura “forward”, por Atkinson e Riani (2000), que é uma metodologia para detectar observações atípicas mascaradas.

Neste trabalho, propomos o uso da influência local com procura “forward” na obtenção de observações mascaradas influentes considerando modelos de regressão linear.

Palavras-chave: *Método de diagnóstico, modelo de regressão, influência local, gráfico de influência, curvatura normal, curvatura normal conformal, procura “forward”.*

Abstract

The identification of influential and/or atypical observations in a data set is known as a part of the diagnostic analysis. One of the purposes of the diagnostic analysis is to verify the robustness of a statistical model, as the non-identification of influential observations can affect the analysis or may cause the obtainment of incorrect results.

The most commonly used methodology for the diagnostic of influential observations in regression models are the global influence (Belsey et al., 1980). Cook (1986) introduced a general method to evaluate the local influence of small perturbations in the statistical model or in the data set using different perturbation schemes. As a complement to the techniques of detection atypical observations, it is proposed the forward search procedure by Atkinson e Riani (2000), which is a methodology to detect the masked atypical observations in a data set.

In this work we propose the use of the local influence approach together with the forward search to obtain the masked influential observations in linear regression models.

Key words: *Diagnostic methods, regression model, local influence, influence graphs, normal curvature, conformal normal curvature, forward search.*

Sumário

1	Introdução	1
1.1	Revisão Bibliográfica	3
1.2	Diagnóstico	7
1.2.1	Análise de Resíduos	7
1.2.2	Análise da suposição de normalidade	7
1.2.3	Análise de sensibilidade	8
1.2.4	Análise da suposição de correlação nula	8
2	Métodos de diagnóstico	9
2.1	Medidas de influência global	11
2.2	Influência local de Cook (1986)	15
2.2.1	Ponderação de casos	17
2.2.2	Perturbação na variável explanatória	18
2.2.3	Perturbação na variável resposta	20
2.2.4	Perturbação na variância do erro	21
2.3	Curvatura Normal Conformal de Poon e Poon (1999)	22
2.4	Procura “Forward” de Atkinson e Riani (2000)	27
2.4.1	Descrição do método	28
3	Aplicações	33
3.1	Dados de ganso	33
3.1.1	Influência local	37
3.1.2	Curvatura normal conformal	39
3.1.3	Procura “Forward”	40
3.2	Dados de rato	43
3.2.1	Influência local	45

3.2.2	Curvatura normal conformal	47
3.2.3	Procura “Forward”	49
3.3	Dados de Stack Loss	52
3.3.1	Influência local	55
3.3.2	Curvatura normal conformal	56
3.3.3	Procura “forward”	58
4	Influência local com procura “Forward”	63
5	Aplicações da metodologia de influência local com procura “forward”	67
5.1	Dados de ganso	67
5.1.1	Ponderação de casos (σ^2 conhecido)	68
5.1.2	Ponderação de casos (σ^2 desconhecido)	69
5.1.3	Perturbação na covariável (σ^2 desconhecido)	70
5.1.4	Perturbação na resposta (σ^2 desconhecido)	72
5.1.5	Perturbação na variância (σ^2 desconhecido)	74
5.2	Dados de rato	75
5.2.1	Ponderação de casos (σ^2 conhecido)	75
5.2.2	Ponderação de casos (σ^2 desconhecido)	76
5.2.3	Perturbação na covariável (σ^2 desconhecido)	77
5.2.4	Perturbação na resposta (σ^2 desconhecido)	78
5.2.5	Perturbação na variância (σ^2 desconhecido)	80
5.3	Dados de Stack Loss	81
5.3.1	Ponderação de casos (σ^2 conhecido)	81
5.3.2	Ponderação de casos (σ^2 desconhecido)	81
5.3.3	Perturbação na covariável (σ^2 desconhecido)	83
5.3.4	Perturbação na resposta (σ^2 desconhecido)	84
5.3.5	Perturbação na variância (σ^2 desconhecido)	85
6	Conclusões	87
6.1	Conclusões	87
6.2	Trabalhos futuros	89
	Referências Bibliográficas	89
A	Conjunto de dados	95
A.1	GANSO	95
A.2	RATO	97
A.3	STACK LOSS	98

B	Influência local Gráficos	99
B.1	Conjunto de dados de ganso	100
B.1.1	Ponderação de casos - quando σ^2 é conhecido	100
B.1.2	Ponderação de casos - quando σ^2 não é conhecido	101
B.1.3	Perturbação na covariável - quando σ^2 é conhecido	102
B.1.4	Perturbação na covariável - quando σ^2 não é conhecido	103
B.1.5	Perturbação na variável resposta - quando σ^2 é conhecido	104
B.1.6	Perturbação na variável resposta - quando σ^2 não é conhecido	105
B.1.7	Perturbação na variância do erro - quando σ^2 é não conhecido	106
B.2	Conjunto de dados de rato	107
B.2.1	Ponderação de casos - quando σ^2 é conhecido	107
B.2.2	Ponderação de casos - quando σ^2 não é conhecido	108
B.2.3	Perturbação na covariável - quando σ^2 é conhecido	109
B.2.4	Perturbação na covariável - quando σ^2 não é conhecido	110
B.2.5	Perturbação na variável resposta - quando σ^2 é conhecido	111
B.2.6	Perturbação na variável resposta - quando σ^2 não é conhecido	112
B.2.7	Perturbação na variância do erro - quando σ^2 é não conhecido	113
B.3	Conjunto de dados de “Stack Loss”	114
B.3.1	Ponderação de casos - quando σ^2 é conhecido	114
B.3.2	Ponderação de casos - quando σ^2 não é conhecido	115
B.3.3	Perturbação na covariável - quando σ^2 é conhecido	116
B.3.4	Perturbação na covariável - quando σ^2 não é conhecido	117
B.3.5	Perturbação na variável resposta - quando σ^2 é conhecido	118
B.3.6	Perturbação na variável resposta - quando σ^2 não é conhecido	119
B.3.7	Perturbação na variância do erro - quando σ^2 é não conhecido	120

Lista de Figuras

2.1	A idéia do gráfico da função ou gráfico da superfície.	16
2.2	A divisão em dois grupos.	28
3.1	Gráfico de dispersão dos dados de ganso.	35
3.2	Diagrama de dispersão com a reta ajustada dos dados de ganso.	35
3.3	Gráfico de medida de influência global dos dados de ganso: (a) resíduos estudentizados, (b) pontos de alavanca, (c) distância de Cook e (d) DFFitS.	36
3.4	Reta ajustada com os dados completos dos dados de ganso (a), sem a observação 29 (b), sem as observações 28 e 29 (c) e sem as observações 28, 29 e 41 (d).	38
3.5	Autovalores normalizados dos dados de ganso.	40
3.6	Contribuição agregada dos dados de ganso.	40
3.7	Gráfico dos resíduos escalonados ao quadrado utilizando a procura “forward” dos dados de ganso.	41
3.8	Gráfico das estimativas dos coeficientes utilizando a procura “forward” dos dados de ganso.	42
3.9	Gráfico das estimativas dos coeficientes escalonados (2.33) utilizando a procura “forward” dos dados de ganso.	42
3.10	Gráfico da matriz de dispersão dos dados de rato.	44
3.11	Gráfico de medida de influência global dos dados de rato: (a) resíduos estudentizados, (b) pontos de alavanca, (c) distância de Cook e (d) DFFitS.	45
3.12	Autovalores normalizados dos dados de rato.	48
3.13	Contribuição agregada dos dados de rato.	49
3.14	Gráfico dos resíduos escalonados ao quadrado segundo a procura “forward” dos dados de rato.	50

3.15	Gráfico das estimativas dos coeficientes escalonado 2.33 segundo a procura “forward” dos dados de rato.	50
3.16	Gráfico das estimativas dos parâmetros segundo a procura “forward” dos dados de rato.	51
3.17	Gráfico dos resíduos escalonados após eliminar o dado 5 segundo a procura “forward” dos dados de rato.	51
3.18	Gráfico da matriz de dispersão dos dados de Stack Lock.	53
3.19	Gráficos de medidas de influência global dos dados de Stack Loss (a) Gráfico de resíduos estudentizados (b) Gráficos dos pontos de alavanca (c) Gráfico da distância de Cook (d) Gráfico de DFFITS.	54
3.20	Autovalores normalizados dos dados de Stack Loss.	57
3.21	Contribuição agregada dos dados de Stack Loss.	58
3.22	Gráfico dos resíduos escalonados segundo a procura “forward” no modelo de segunda ordem dos dados de Stack Loss.	59
3.23	Gráfico das estimativas dos parâmetros segundo a procura “forward” no modelo de segunda ordem dos dados de Stack Loss.	59
3.24	Gráfico das estimativas dos parâmetros escalonados segundo a procura “forward” no modelo de segunda ordem dos dados de Stack Loss.	60
3.25	Gráfico resíduos escalonados segundo a procura “forward” dos dados de Stack Loss.	60
3.26	Gráfico das estimativas dos coeficientes segundo a procura forward dos dados de Stack Loss.	61
3.27	Gráfico das estimativas dos coeficientes escalonados 2.33 segundo a procura “forward” dos dados de Stack Loss.	61
5.1	Gráfico do valor absoluto do autovetor, $lmax = Lmax$, associado ao maior autovalor $Cmax$ dos dados de ganço.	68
5.2	Estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de ganço.	68
5.3	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de ganço.	69
5.4	Estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de ganço.	70
5.5	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de ganço.	70
5.6	Gráficos das estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de ganço.	71

5.7	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor dos dados de ganso no passo $m = 2$ do processo de influência local com procura “forward”.	71
5.8	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de ganso.	72
5.9	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor dos dados de ganso para $m = 2$ da primeira rodada.	73
5.10	Gráficos das estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de ganso.	73
5.11	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de ganso.	74
5.12	Gráficos das estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de ganso.	74
5.13	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de rato.	75
5.14	Gráficos das estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de rato.	76
5.15	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de rato.	76
5.16	Gráficos das estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de rato.	77
5.17	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de rato, com as observações. Os valores representados de 1 a 19, 20 a 38 e 39 a 57, no vetor $lmax$, correspondem às covariáveis x_1 , x_2 e x_3 no processo de influência local com procura “forward”.	77
5.18	Gráficos das estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de rato.	78
5.19	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de rato.	78
5.20	Gráfico do valor absoluto do autovetor $lmax$ associado ao maior autovalor para $m = 2$ no processo de procura “forward” dos dados de rato.	79
5.21	Gráficos das estimativas dos coeficientes, coeficientes escalonados e $Cmax$ dos dados de rato.	79
5.22	Gráfico do valor absoluto do autovetor $lmax = Lmax$ associado ao maior autovalor $Cmax$ dos dados de rato.	80

5.23	Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de rato.	80
5.24	Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss.	81
5.25	Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.	82
5.26	Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss.	82
5.27	Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.	83
5.28	Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss, com as observações. Os valores representados de 1 a 21, 22 a 42, 43 a 63 e 64 a 84, no vetor l_{max} , correspondem às covariáveis x_1 , x_2 , x_1^2 e x_1x_2 no processo de influência local com procura “forward”.	83
5.29	Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.	84
5.30	Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss.	84
5.31	Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.	85
5.32	Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss.	85
5.33	Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.	86
B.1	Gráfico na ponderação de casos (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável no conjunto dos dados de ganso.	100
B.2	Gráfico na ponderação de casos (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável dos dados de ganso.	101
B.3	Gráfico na perturbação na covariável (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de ganso. . .	102
B.4	Gráfico na perturbação na covariável (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de ganso. . .	103

- B.5 Gráfico na perturbação na resposta (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável dos dados de ganso. 104
- B.6 Gráfico na perturbação na resposta (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável dos dados de ganso. 105
- B.7 Gráfico na perturbação na variância (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável no conjunto dos dados de ganso. 106
- B.8 Gráfico na ponderação de casos (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis dos dados de rato. 107
- B.9 Gráfico na ponderação de casos (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis dos dados de rato. 108
- B.10 Gráfico na perturbação na covariável (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de rato. As observações 1 a 19, 20 a 38 e 39 a 57 respectivamente, a covariável x_1 , x_2 e x_3 109
- B.11 Gráfico na perturbação na covariável (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de rato. As observações 1 a 19, 20 a 38 e 39 a 57 respectivamente, a covariável x_1 , x_2 e x_3 110
- B.12 Gráfico na perturbação na resposta (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis dos dados de rato. 111
- B.13 Gráfico na perturbação na resposta (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis dos dados de rato. 112
- B.14 Gráfico na perturbação na variância (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis do conjunto de dados de rato. 113
- B.15 Gráfico na ponderação de casos (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis no conjunto de dados Stack Loss. 114

B.16	Gráfico na ponderação de casos (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis no conjunto dos dados de Stack Loss.	115
B.17	Gráfico na perturbação na covariável (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de Stack Loss. As observações de 1 a 21, 22 a 42, 43 a 63, 64 a 84 referem-se respectivamente, a covariáveis x_1, x_2, x_3 e x_4	116
B.18	Gráfico na perturbação na covariável (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de Stack Loss. As observações de 1 a 21, 22 a 42, 43 a 63, 64 a 84 referem-se respectivamente, a covariáveis x_1, x_2, x_3 e x_4	117
B.19	Gráfico na perturbação na resposta (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis dos dados de Stack Loss. . . .	118
B.20	Gráfico na perturbação na resposta (σ^2 desconhecido) de autovetor máximo dos dados de Stack Loss com: (a) resíduo (b) observação (c) resposta y e as covariáveis.	119
B.21	Gráfico na perturbação na variância (σ^2 desconhecido) de autovetor máximo dos dados de Stack Loss com: (a) resíduo (b) observação (c) resposta y e as covariáveis.	120

Lista de Tabelas

3.1	Estatísticas descritivas dos dados de ganso.	34
3.2	Medida de influência global dos dados de ganso.	36
3.3	Valores do C_{max} para os dados de ganso.	37
3.4	Estimativa dos parâmetros dos dados de ganso com a mudança relativa entre parênteses.	39
3.5	Medidas de influência para os dados de ganso utilizando a curvatura normal conformal para o esquema de ponderação de casos.	39
3.6	Estatísticas descritivas dos dados de rato.	43
3.7	Estimativa dos parâmetros dos dados de rato.	44
3.8	Medida de influência segundo as medidas de influência DF_{Beta} , DF_{Fit} , Distância de Cook e os pontos de alavanca descritos na Secção 2.1 dos dados de rato.	44
3.9	Estimativa dos parâmetros dos dados de rato após retirar a observação 3.	45
3.10	Valores do C_{max} para os dados de rato.	46
3.11	Estimativa dos parâmetros dos dados de rato com a mudança relativa entre parênteses.	47
3.12	Medidas de influência utilizando a curvatura normal conformal para o esquema de ponderação de casos dos dados de rato.	48
3.13	Estimativa dos parâmetros dos dados de rato com a mudança relativa entre parênteses.	52
3.14	Estatísticas descritivas dos dados de Stack Loss.	53
3.15	Medida de influência global dos dados de Stack Loss.	54
3.16	Estimativa dos parâmetros dos dados de Stack Loss com a mudança relativa entre parênteses.	55
3.17	Valores do C_{max} para os dados de Stack Loss.	56

3.18	Estimativa dos parâmetros dos dados de Stack Loss com a mudança relativa entre parênteses.	57
3.19	Medidas de influência utilizando a curvatura normal conformal para o esquema de ponderação de casos dos dados de Stack Loss.	58
3.20	Estimativas dos parâmetros dos dados de Stack Loss tirando algumas observações	62
5.1	Estimativa dos parâmetros dos dados de ganso com a mudança relativa entre parênteses.	72
5.2	Estimativa dos parâmetros dos dados de Stack Loss com a mudança relativa entre parênteses.	86
A.2	Dados rato	97
A.3	Dados Stack Loss	98

Introdução

Os modelos estatísticos são instrumentos extremamente úteis para compreender as características essenciais de um conjunto de dados, além disso, quase sempre são descrições aproximadas de processos mais complexos e devido a essa imprecisão deve ser realizada uma avaliação do modelo adotado, pois, podem existir anomalias no ajuste do modelo, enfatizados especialmente na presença de dados discrepantes e/ou influentes.

Segundo Samprit e Ali (2006), em modelos de regressão, dados atípicos na variável resposta são as observações com grandes resíduos padronizados (na direção \mathbf{Y}). Uma vez que os resíduos padronizados tem uma distribuição aproximadamente normal com média zero e variância 1, observações com resíduos padronizados maiores do que 2 ou 3, são chamados de dados atípicos (outliers). Eles podem ser identificados utilizando métodos de análise formal (como a influência global ver, Hawkins (1980), Belsey et al. (1980), Cook e Weisberg (1982), Atkinson (1985), Hadi e Simonoff (1993), Barnett e Lewis (1994) e Hadi e Velleman (1997)) ou através de gráficos dos resíduos adequadamente escolhidos.

Observações atípicas também podem ocorrer nas variáveis preditoras e podem afetar os resultados da regressão. Os pontos de alavanca podem ser utilizados neste caso (no espaço \mathbf{X}) e possuem várias propriedades interessantes (ver Dodge e Hadi (1999) e Chatterjee e Hadi (1988)).

Em qualquer análise, os pontos de alavanca alto devem ser marcados e examinados para ver se são influentes. Um gráfico dos pontos de alavanca (por exemplo, gráfico de dispersão, ou “box plot”) podem revelar pontos de alavanca alto, caso existam.

No ajuste de um modelo a um determinado conjunto de dados, gostaríamos de garantir que o ajuste não é excessivamente determinado por uma ou algumas observações. Por exemplo, no modelo de regressão linear se o ponto extremo fosse retirado, poderia resultar mudanças nas estimativas dos parâmetros (poderia resultar em uma linha muito diferente). Quando temos várias variáveis, não é possível detectar tal situação graficamente. Queremos, no entanto, saber da existência de tais pontos (pontos atípicos). Deve-se salientar que olhar para os resíduos, neste caso, pode ser de pouca ajuda, pois, o resíduo para este ponto às vezes é próximo de zero, ou seja ele não tem um resíduo grande, mas, pode ser um ponto muito influente. Uma observação é uma observação influente, se a sua exclusão, isoladamente ou em combinação com outros (2 ou mais), provocam alterações substanciais no modelo ajustado (mudanças nos coeficientes estimados, nos valores ajustados, testes-t, etc). A exclusão de alguns pontos em geral, pode causar alterações no ajuste, ou seja, estamos interessados em detectar os pontos cuja exclusão possa causar grandes mudanças (pontos que exercer uma influência indevida) (Samprit e Ali, 2006).

A distância de Cook (1977), originalmente desenvolvida para modelos lineares normais para avaliar o impacto da retirada de uma observação particular nas estimativas da regressão, foi rapidamente assimilada e estendida para diversas classes de modelos.

Alguns métodos estão disponíveis para a realização de tal avaliação em diferentes contextos de regressão linear normal, e muitos trabalhos lidam com o esquema de influência global como em Belsey et al. (1980), Cook e Weisberg (1982), Atkinson (1985). Para detectar essas anomalias (dados atípicos), podem ser usadas as estatísticas de diagnóstico global como, por exemplo, a medida de alavanca, $DF\beta$ que é importante quando o coeficiente de regressão tem um significado prático e mede a alteração no vetor estimado ao retirar o i -ésimo ponto da análise, $DFFitS$ que mede a alteração provocada no valor ajustado pela retirada da i -ésima observação e a distância de Cook que é também uma medida de afastamento do vetor de estimativas provocada pela retirada da i -ésima observação e é muito semelhante ao $DFFitS$, mas considera o resíduo estudentizado internamente. As quatro estatísticas introduzidas acima para a medição dos pontos de alavanca, a medida de influência sobre os parâmetros, medida de influência geral e medida de influência geral utilizando o resíduo estudentizado internamente, respectivamente, resulta num vetor de tamanho n . Os valores interessantes

para fins de verificação do modelo podem ser caracterizadas por ter h_{ii} e/ou resíduos grandes, ser inconsistentes e/ou ser influentes (Ver Secção 2.1).

Cook (1986) apresenta um método geral para avaliar a influência local de pequenas perturbações no modelo estatístico ou nos dados. Hoje em dia, existem inúmeras aplicações sobre o método proposto por Cook (1986). Ele sugere que examinemos a direção de maior curvatura normal, pois, a partir dela identificamos observações potencialmente influentes sob o esquema de perturbação ao qual sujeitamos o modelo. Mais tarde, Poon e Poon (1999) introduziram a curvatura normal conformal baseando-se no trabalho de Cook (1986). Poon e Poon (1999) propuseram a construção de uma curvatura normal conformal, com a mesma finalidade. Esta curvatura fornece uma medida de influência local que varia entre 0 e 1, com o objetivo de ter um ponto de corte ou uma faixa limite para detectar dados atípicos.

Modelos cuidadosamente construídos em combinação com métodos de diagnóstico apropriados fornecem uma base útil para a análise estatística, no entanto, um simples diagnóstico de eliminação de casos pode ser computacionalmente intensivo e sofrer do problema de mascaramento. Assim, como complemento às técnicas de diagnóstico existentes, foi proposto o método procura “forward” por Atkinson e Riani (2000). O método tem por finalidade descrever explicitamente na evolução, através de gráficos, o impacto que cada observação tem sobre o ajuste de modelos e respectivas análises de diagnósticos.

Neste trabalho, considerando o modelo de regressão linear, vamos propor o uso da influência local em conjunto com a procura “forward” para fazer o diagnóstico de influência.

1.1 Revisão Bibliográfica

Existe uma vasta literatura de métodos e técnicas de análise de diagnóstico em modelos estatísticos.

Hawkins et al. (1984) descreve que os dados atípicos em uma regressão linear múltipla com p variáveis preditoras podem ser identificados extraindo todos os subconjuntos de tamanho p dos casos restantes e ajustando o modelo. Cada um destes modelos produz um subconjunto residual elementar para o caso em questão, e um resumo estatístico adequado pode ser utilizado como uma estimativa da tendência dos casos atípicos. É proposto duas estatísticas de resumo: uma mediana não ponderada, que é de influência limitada, e uma mediana ponderada, que é mais eficiente, mas menos robusto. A

carga de processamento computacional do processo é reduzido usando amostras aleatórias no lugar de todos os subconjuntos de tamanho p . Apesar de poder localizar e testar a significância de um dado atípico, vários valores atípicos continuam a ser um problema. A dificuldade em identifica-los, mesmo em amostras aleatórias simples, tem sido reconhecida como problema de mascaramento.

Pamula et al. (2011) utiliza uma metodologia PLDOF (Fator Atípico de Poda Baseada em uma Distancia Local) de agrupamento baseado na captura de dados atípicos, todos os pontos são ordenados baseados em uma pontuação para os dados atípicos. É usado o algoritmo das k -medias para o agrupamento, (divide o conjunto de dados em agrupamentos (clusters)). Os pontos localizados perto do centro do agrupamento (cluster) não são prováveis candidatos para serem dados atípicos e tais pontos podem ser retirados de cada agrupamento. O número de cálculos computacionais de PLDOF é menor e o método proposto é mais preciso do que o método LDOF (Fator Atípico Baseada em uma Distancia Local).

Um dos trabalhos que tratam da identificação de dados atípicos múltiplos pode ser visto em Hadi (1992). Ele propõe um procedimento para detectar dados atípicos múltiplos em dados multivariados. Primeiro, as n observações são ordenadas de menor a maior, usando uma medida robusta de dados atípicos apropriadamente escolhida, então, os dados são divididos em dois subconjuntos iniciais: um subconjunto básico que contém $p + 1$ observações (em que, p é o número de covariáveis), e um subconjunto que contém as $n - p - 1$ observações restantes. Depois, são calculadas as distâncias relativas de cada ponto do conjunto de dados ao centro do subconjunto básico. Finalmente, as n observações são ordenadas em ordem ascendente, e o conjunto de dados é dividido em dois subconjuntos: um subconjunto básico com as primeiras $p + 2$ observações e outro subconjunto que contém as $n - p - 2$ observações restantes. Esse processo é repetido até que um critério de parada escolhido seja atendido. Em particular, Hadi (1992) utilizou a distância de Mahalanobis para detectar os dados atípicos. A distância clássica de Mahalanobis não é claramente efetiva na identificação de dados atípicos múltiplos já que sofre de problemas de mascaramento (Hadi, 1992).

O **problema de mascaramento** ocorre quando um subconjunto anômalo não é detectada devido à presença de um outro, geralmente um subconjunto adjacente. Arrastamento ocorre quando as observações “boas” são incorretamente identificadas como valores atípicos por causa da presença de um outro subconjunto de observações. Métodos que são menos suscetíveis aos problemas de mascaramento e arrastamento, (resíduos padronizados e pontos de alavanca) são dadas em Hadi e Simonoff (1993).

Hadi e Simonoff (1993) propõem um procedimento para a detecção de dados atípicos, considerando o problema de identificação e teste de hipótese de dados atípicos múltiplos. Os métodos de identificação de dados atípicos disponíveis muitas vezes não conseguem detectar dados atípicos múltiplos porque são afetados pelas próprias observações que deveriam ser identificados, “o problema de mascaramento”. Hadi e Simonoff (1993) introduzem dois procedimentos que são desenvolvidos, descritos e testados para a detecção simultânea de dados atípicos múltiplos que parecem ser menos sensíveis a este problema. Em ambos os procedimentos, os dados são separados em um subconjunto limpo ou livre de observações atípicas e um conjunto de pontos de dados que contém os dados atípicos em potencial. Depois os potenciais dados atípicos são testados para ver quão extremo são em relação ao subconjunto limpo ou livre de observações atípicas.

Rousseeuw e Van Zomeren (1990) propõem calcular distâncias baseadas em estimativas muito robustas de locação e escala. Essas distâncias robustas são mais adequadas para expor os dados atípicos e evitar o efeito de mascaramento.

Atkinson e Mulira (1993) usam a distância de Mahalanobis e constroem sequencialmente um subconjunto livre de dados atípicos, a partir de um pequeno subconjunto aleatório. O gráfico de “estalactite” fornece um resumo convincente de suspeitas de valores atípicos conforme o tamanho do subconjunto aumenta. Combinado com gráficos de probabilidade e procedimentos de reamostragem, o gráfico de “estalactite”, em particular na sua forma normalizada, leva a identificação de valores extremos multivariados, mesmo na presença de mascaramento apreciável.

Cerioli e Riani (1999) usam a técnica do algoritmo de procura “forward” que ordena as observações de acordo com um modelo de auto-correlação especificado. Isso leva à identificação de dados atípicos. Em particular, o foco da análise é sobre modelos de previsão espacial. Eles mostraram que o diagnóstico padrão (de exclusão) para predição são afetadas por mascaramento e problemas de arrastamento quando vários valores atípicos estão presentes. A eficácia do método sugerido na detecção de vários valores atípicos mascarados é, mais geralmente, na ordenação de dados espaciais (pode ser concebida como uma realização de um processo estocástico). Cerioli e Riani (1999) apresentam exemplos que revelam claramente o poder do seu método em relação a procedimento de diagnóstico padrão (de exclusão).

Atkinson e Riani (2000) desenvolveram uma metodologia baseada nos artigos de Hadi (1992) e Hadi e Simonoff (1993), o método procura “forward” para a detecção de dados atípicos. Considerando modelos de regressão, o conjunto de dados é dividido em subconjuntos de tamanho m e ajustado o modelo aos dados de cada subconjunto. É

selecionado um dos subconjuntos de tamanho m (conjunto com m dados) livre de dados atípicos. Depois são selecionados subconjuntos de tamanho $m + 1$ (conjunto com $m + 1$ dados) e ajustado o modelo estatístico aos dados do subconjunto e assim sucessivamente até chegar a n . O propósito do método é observar as características do modelo (resíduos, estimativas dos parâmetros etc) e os dados na evolução do método, ver quais deles se destacam, são dados atípicos, quais sofrem do problema de mascaramento, etc.

A eliminação de pontos é o método mais tradicional de sensibilidade que consiste em avaliar o impacto da retirada de uma observação nas estimativas dos parâmetros de um modelo (Secção 2.1). Há uma vasta literatura sobre o assunto, veja por exemplo, Cook (1977); Belsey et al. (1980); Chatterjee e Hadi (1988); Cook e Weisberg (1982) e Weisberg (2005). Contudo, o método de eliminação de pontos não verifica a influência conjunta das observações nas estimativas dos parâmetros. Nesse sentido o método de influência local tem se constituído numa ferramenta importante na análise da influência conjunta das observações nas estimativas dos parâmetros do modelo. Cook (1986) propõe uma metodologia de influência local baseada em pequenas perturbações no modelo ou no conjunto de dados utilizando a curvatura normal. Poon e Poon (1999) propõe o uso da curvatura normal conformal para o mesmo propósito de Cook (1986) com o objetivo de ter um ponto de corte para o valor obtido.

O diagnóstico compreende questões como o análise de resíduos, análise da suposição de normalidade, análise de sensibilidade e análise da suposição de correlação nula que serao mencionados de forma sucinta na seguinte secção.

1.2 Diagnóstico

Modelos estatísticos são utilizados como aproximações de processos mais complexos e são construídos sobre um conjunto de suposições, é importante avaliar se tais aproximações são aceitáveis. Isto pode ser concretizado por meio de técnicas de diagnóstico, que englobam a avaliação do ajuste e a análise de sensibilidade. No primeiro caso, o objetivo é avaliar se as suposições adotadas são compatíveis com os dados; no segundo, o objetivo é verificar se o modelo é sensível a perturbações nos dados ou no modelo. Se esta variação for “substancial” no sentido de mudar as conclusões, diz-se que o modelo não é robusto. Neste caso, ou as conclusões devem ser tomadas (se tomadas) de forma cautelosa, ou então deve-se optar por outro modelo, segundo Singer e de Andrade (1986) e Samprit e Ali (2006).

1.2.1 Análise de Resíduos

Resíduos são utilizados para avaliar a validade de determinadas suposições de modelos estatísticos. No caso de modelos de regressão linear clássico, podemos utilizá-los para verificar homocedasticidade, existência de pontos discrepantes, normalidade e independência dos erros.

1.2.2 Análise da suposição de normalidade

Na teoria clássica de regressão, tanto intervalos de confiança quanto testes de hipóteses sobre os parâmetros de modelos lineares são baseados na suposição de normalidade dos erros. A verificação da plausibilidade dessa suposição é fundamental para a validade dos procedimentos inferenciais (exatos). Como os resíduos são essencialmente preditores dos erros do modelo, nada mais natural do que utilizá-los com essa finalidade. Nesse sentido, gráficos do tipo QQ (quantis-quantis), em que dispomos os resíduos studentizados ordenados (quantis observados) no eixo das abscisas e os quantis obtidos da distribuição normal padrão (quantis teóricos) no eixo das ordenadas, são ferramentas úteis. Quando a distribuição dos erros é gaussiana, espera-se que esses resíduos estejam dispostos numa vizinhança da reta com inclinação de 45 graus. Esse tipo de gráfico também pode ser útil para detectar a presença de observações discrepantes, para avaliar se a distribuição dos erros possui caudas mais pesadas que a distribuição normal, para avaliar se os erros são heterocedásticos.

1.2.3 Análise de sensibilidade

A análise de sensibilidade visa avaliar o comportamento do ajuste de um modelo sujeito a algum tipo de perturbação, ou seja, sob alguma mudança nas hipóteses ou nos dados. Como cada observação não tem a mesma influência em todas as características do ajuste do modelo, é natural que se defina aquela na qual se quer focar a análise. Se o objetivo for fazer previsões, então é razoável medir a influência das observações nos valores preditos e não nos parâmetros de locação, como mencionam Chatterjee e Hadi (1986) e Chatterjee e Hadi (1988).

Existem medidas de influência baseadas nos resíduos, na curva de influência, na verossimilhança, no volume dos elipsóides de confiança, em um subconjunto do vetor de parâmetro de locação (influência parcial), nos pontos remotos do espaço vetorial gerado pelas colunas da matriz de especificação \mathbf{X} , entre outras.

Dentre as abordagens mais utilizadas na prática, para medir influência em modelos lineares, destacam-se aquelas baseadas em influência local considerada em Cook (1986) e aquelas obtidas por intermédio da eliminação de observações (influência global).

1.2.4 Análise da suposição de correlação nula

Em geral, a suposição de que os erros do modelo linear são não-correlacionados deve ser questionada com base no procedimento de coleta de dados. Se existir autocorrelação entre os resíduos, uma maneira de contornar esse problema, é modificar os componentes aleatórios do modelo para incorporar uma possível autocorrelação nos erros. Por exemplo para testar a hipótese de que os erros são não-correlacionados pode-se utilizar a estatística de Durbin Watson.

No próximo capítulo apresentaremos os métodos de diagnóstico que são a motivação para o nosso trabalho.

Métodos de diagnóstico

Nesta dissertação vamos considerar o método de influência local em conjunto com a procura “forward”. Nos estudos de modelagem estatística, a análise de diagnóstico é uma etapa muito importante e utiliza um conjunto de ferramentas para avaliar a qualidade do ajuste do modelo proposto aos dados e ainda para verificar a coerência das suposições iniciais. Métodos de estimação utilizando a função de verossimilhança podem ser sensíveis a observações aberrantes, especialmente no modelo normal, e o diagnóstico de influência inclui técnicas que permitem identificar observações que podem influenciar desproporcionalmente as estimativas dos parâmetros.

O diagnóstico de influência é usado para investigar vários aspectos do modelo ajustado, permitindo a validação das suposições do modelo proposto. Esse diagnóstico inclui, basicamente, dois métodos de análises de influência: local e global. A técnica de influência local, em particular, identifica observações influentes por meio de pequenas perturbações nos dados ou no modelo, enquanto que a influência global utiliza alguma medida como $DF\beta$, $DFFitS$ e $D-Cook$ (Cook, 1977; Belsey et al., 1980) para analisar, por exemplo, as mudanças nos modelos ajustados quando é induzida a exclusão de um subconjunto de observações. (Veja Secção 2.1).

Para realizar uma análise de diagnóstico, a técnica de influência local tem se constituído uma ferramenta muito importante com ampla utilização. Por exemplo, recentemente, foi aplicada por Souza (2006) no modelo de regressão logística, Soares da

Silva Gomes (2005) faz uma análise de influência local para a distribuição Dirichlet, Nobre (2004) propõe um modelo misto para modelar a estrutura de correlação intra-unidade experimental de um conjunto de dados de escova de dente e faz o uso do método de diagnóstico de influência local, Aoki et al. (2007) e Russo (2006) aplicam diagnóstico de influência local em modelos com erros de medição.

De forma geral, a influência local consiste em analisar, por meio de uma medida adequada de influência, a robustez das estimativas dos parâmetros ajustados quando pequenas perturbações são introduzidas no modelo ou nos dados.

2.1 Medidas de influência global

No modelo de regressão clássico temos:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \quad (2.1)$$

sendo \mathbf{Y} o vetor de dimensão $n \times 1$ da variável resposta, $\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, o componente sistemático, \mathbf{X} a matriz do modelo de dimensões $n \times p$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$ o vetor dos parâmetros, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$, os erros aleatórios com $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$. Nesse caso, temos que $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 I)$.

As estatísticas mais utilizadas para a verificação de pontos atípicos são:

- Medida de ponto de alavanca: h_{ii}
- Medida de inconsistência: $r_{se(i)}$
- Medidas de influência geral: $DFFitS_{(i)}$ e $D_{(i)}$ (distância de Cook)
- Medidas de influência sobre o parâmetro β_j : $DFBetaS_{j(i)}$ para β_j .

Assim, em geral, pode-se classificar uma observação como:

- Ponto inconsistente: ponto com $r_{se(i)}$ alto, isto é, tal que $|r_{se(i)}| \geq t_{\{\frac{\gamma}{2n}; n-p-1\}}$, com nível de significância igual a $100\gamma\%$.
- Ponto de alavanca: ponto com h_{ii} alto, isto é, tal que $h_{ii} \geq \frac{2p}{n}$. Pode ser classificado como bom, quando consistente, ou ruim, quando inconsistente.
- Observação atípica (outlier): ponto inconsistente com ponto de alavanca baixo, ou seja, com $r_{se(i)}$ alto e h_{ii} baixo.
- Ponto influente: ponto com $DFFitS_{(i)}$, $D_{(i)}$ ou $DFBetaS_{j(i)}$ alto.
Observação: $DFFitS_{(i)}$ é considerado alto se $DFFitS_{(i)} \geq 2\sqrt{\frac{p}{n}}$.
 $DFBetaS_{j(i)}$ é considerado alto se $DFBetaS_{j(i)} \geq 2/\sqrt{n}$.

Observação importante: O R considera a i -ésima observação influente

se $|DFBetaS_{(i)}| > 1$, se $|DFFitS_{(i)}| > 3\sqrt{p/(n-p)}$, se $D_{(i)} > F_{\{50\%; p; n-p\}}$, ou se $h_{ii} > 3p/n$.

Notação:

- Resíduos ordinários $r_i = y_i - \hat{\mu}_i$;

- $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ com os elementos da diagonal dados por:

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i,$$

em que, $\mathbf{x}_i^T = (1, x_{i1}, \dots, x_{i(p-1)})$;

-

$$s^2 = QMRes = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{(n-p)},$$

quadrado médio do resíduo como a estimativa de σ^2 ;

- **Resíduos escalonados:** Pode ser definida usando a estimativa de σ^2 .

$$\frac{r_i}{\hat{\sigma}^2} = \frac{r_i}{s^2} \quad (2.2)$$

- $s_{(i)}^2$ representa o quadrado médio do resíduo para o modelo ajustado, excluindo a i -ésima observação;

- Resíduos Studentizados externamente:

$$r_{se(i)} = \frac{r_i}{s_{(i)} \sqrt{(1 - h_{ii})}};$$

- Resíduos Studentizados internamente:

$$r_{si(i)} = \frac{r_i}{s \sqrt{(1 - h_{ii})}} = \frac{r_i}{\sqrt{(1 - h_{ii})} QMRes}.$$

A seguir são descritas as estatísticas citadas.

- a) **Elementos da diagonal da matriz de projeção \mathbf{H}** (h_{ii} , ponto de alavanca) - A distância de uma observação em relação às demais no espaço \mathbf{X} é medida pelo h (medida de ponto de alavanca).

Sendo \mathbf{H} uma matriz de projeção, tem-se $\mathbf{H} = \mathbf{H}^2$ e, portanto,

$$h_{ii} = \sum_{j=1}^n h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

concluindo-se que $0 \leq h_{ii} \leq 1$ e $\sum_{j=1}^n h_{ij} = 1$. Além disso,

$$r(\mathbf{H}) = \text{tr} [\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] = \text{tr} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}] = \text{tr}(\mathbf{I}_p) = \sum_{i=1}^n h_{ii} = p,$$

e, portanto, o valor médio de h_{ii} é p/n .

No processo de ajuste, como $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$, tem-se

$$\hat{\mu}_i = \sum_{j=1}^n h_{ij} y_j = h_{i1} y_1 + h_{i2} y_2 + \cdots + h_{ii} y_i + \cdots + h_{in} y_n,$$

com $1 \leq i \leq n$. Portanto, o valor ajustado $\hat{\mu}_i$ é a média ponderada dos valores observados e o peso de ponderação é o valor de h_{ij} . Assim, o elemento da diagonal de \mathbf{H} é o peso com que a observação y_i participa do processo de obtenção do valor ajustado $\hat{\mu}_i$. Valores de $h_{ii} \geq 2p/n$, segundo Belsey et al. (1980, p. 17) indicam observações que merecem uma análise mais apurada.

- b) **DFBeta** é importante quando o coeficiente de regressão tem um significado prático. Mede a alteração no vetor estimado $\hat{\boldsymbol{\beta}}$ ao retirar a i -ésima observação da análise. É definido como:

$$DFBeta_{(i)} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{r_i}{(1 - h_{ii})} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i.$$

Não tem interpretação simples ou ainda considerando que $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{C}\mathbf{Y}$ em que $\mathbf{C} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ é uma matriz $p \times n$, tem-se:

$$DFBeta_{(i)} = \frac{r_i}{(1 - h_{ii})} \mathbf{c}_i, \quad i = 1, \dots, n,$$

sendo \mathbf{c}_i^T a i -ésima linha de \mathbf{C}^T . Então,

$$DFBeta_{j(i)} = \frac{r_i}{(1 - h_{ii})} c_{ji}, \quad i = 1, \dots, n, \quad j = 0, \dots, p - 1.$$

Cook e Weisberg (1982) propuseram curvas empíricas para o estudo dessa medida. Como $\text{Cov}(\hat{\boldsymbol{\beta}}) = \text{CVar}(\mathbf{Y})\mathbf{C}^T$, a versão Studentizada de $DFBeta_{j(i)}$ reduz-se a

$$DFBetaS_{j(i)} = \frac{c_{ji}}{s_{(i)} \sqrt{\sum_{i=1}^n c_{ji}^2}} \frac{r_i}{(1 - h_{ii})}, \quad i = 1, \dots, n, \quad j = 0, \dots, p - 1.$$

- c) **DFFitS** mede a alteração provocada no valor ajustado pela retirada da observação i .

É dada pela mudança de ajuste, definido como

$$DFFit_{(i)} = \hat{y}_i - \hat{y}_{i(i)} = \mathbf{x}_i^T (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$

$$DFFitS_{(i)} = \frac{DFFit_{(i)}}{\sqrt{h_{ii}s_{(i)}^2}} = \frac{\mathbf{x}_i^T (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{\sqrt{h_{ii}s_{(i)}^2}} = \frac{1}{\sqrt{h_{ii}s_{(i)}^2}} \frac{r_i}{(1 - h_{ii})} \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$$

ou, ainda,

$$DFFitS_{(i)} = \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{\frac{1}{2}} \frac{r_i}{s_{(i)}(1 - h_{ii})^{\frac{1}{2}}} = \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{\frac{1}{2}} r_{se(i)}$$

sendo o quociente $\frac{h_{ii}}{1 - h_{ii}}$, chamado potencial de influência, uma medida da distância do ponto \mathbf{X} em relação às demais observações. Belsey et al. (1980, p. 28) sugerem que valores absolutos excedendo $2\sqrt{p/n}$ podem identificar observações influentes.

- d) **Distância de Cook** é também uma medida de afastamento do vetor de estimativas provocado pela retirada da observação i . É uma expressão muito semelhante ao *DFFitS* mas que usa como estimativa da variância residual aquela obtida com todas as n observações, ou ainda, considera o resíduo Studentizado internamente. É dada por

$$D_{(i)} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T (\mathbf{X}^T \mathbf{X}) (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{ps^2} = \frac{h_{ii}}{(1 - h_{ii})^2 ps^2} \frac{r_i^2}{ps^2}$$

$$D_{(i)} = \left[\frac{r_i}{(1 - h_{ii})^{\frac{1}{2}} s} \right]^2 \frac{h_{ii}}{p(1 - h_{ii})}$$

ou, ainda,

$$D_{(i)} = \frac{h_{ii} r_{si}^2}{p(1 - h_{ii})}.$$

2.2 Influência local de Cook (1986)

Cook (1986) propõe o diagnóstico de influência local, em que ao invés de eliminar observações ele apresenta um método bastante geral para avaliar a influência conjunta das observações sob pequenas perturbações no modelo ou nos dados (ou de uma observação). Essa metodologia, denominada influência local teve grande aceitação sobre os pesquisadores e usuários de modelos de regressão e outros modelos estatísticos. A popularidade da influência local se dá pelo fato de poder ser aplicada a qualquer problema em que se conheça a função de verossimilhança. Nessa técnica, a ideia principal consiste em efetuar pequenas perturbações nos dados ou no modelo e verificar se os resultados são alterados de forma significativa. A proposta de Cook (1986) é descrita e utilizada em muitos trabalhos envolvendo modelos estatísticos, como por exemplo, em Fung e Kwan (1997), Billor e Loynes (1993) e Aoki (2001).

Sejam:

- $\boldsymbol{\theta}_{p \times 1}$ um vetor de parâmetros desconhecidos;
- $L(\boldsymbol{\theta})$ a função de verossimilhança;
- $L(\boldsymbol{\theta}|\boldsymbol{\omega})$ a função de verossimilhança perturbada;
- $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$;
- $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \log L(\boldsymbol{\theta}|\boldsymbol{\omega})$;
- $\boldsymbol{\omega} = (w_1, w_2, \dots, w_q)^T$ o vetor de perturbações;
- $\boldsymbol{\omega}_o$ o vetor de não perturbação tal que $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}_o) = \ell(\boldsymbol{\theta})$;
- $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ em $L(\boldsymbol{\theta})$ e $L(\boldsymbol{\theta}|\boldsymbol{\omega})$, respectivamente.

Para verificar a influência das perturbações nas estimativas de $\boldsymbol{\theta}$, Cook (1986) propôs o afastamento da verossimilhança:

$$LD(\boldsymbol{\omega}) = 2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\}.$$

Como a análise de $LD(\boldsymbol{\omega})$ para todos os possíveis valores de $\boldsymbol{\omega}$ é inviável, Cook (1986) propôs o estudo do comportamento local em torno de $\boldsymbol{\omega}_o$ considerando uma

superfície formada pelos elementos do vetor $\alpha(\boldsymbol{\omega})$, denotado por gráfico de influência, em que

$$\alpha(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\omega} \\ LD(\boldsymbol{\omega}) \end{pmatrix} \quad (2.3)$$

e a ideia básica foi analisar como $\alpha(\boldsymbol{\omega})$ desvia-se do plano tangente em $\boldsymbol{\omega}_o$ e também como a função se comporta em torno de $\boldsymbol{\omega}_o$. Maiores detalhes podem ser vistos em Cook (1986).

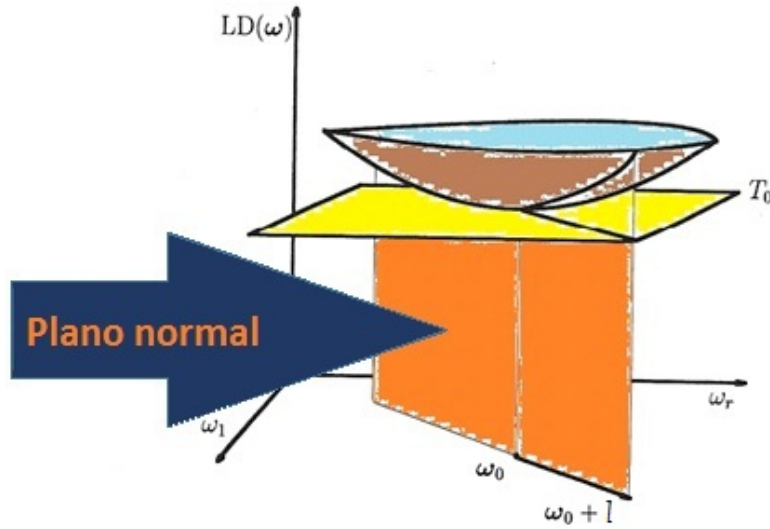


Figura 2.1: A ideia do gráfico da função ou gráfico da superfície.

O método busca analisar o gráfico de $LD(\boldsymbol{\omega}_o + a\mathbf{l})$ após ter selecionado uma direção unitária \mathbf{l} e $a \in \mathbb{R}$. $LD(\boldsymbol{\omega}_o + a\mathbf{l})$ apresenta um mínimo local (em $a = 0$) e neste caso o gráfico, cuja curvatura chamaremos $C_{\mathbf{l}}$, pode ser visto como o círculo (numa vizinhança de $\boldsymbol{\omega}_0$) de melhor ajuste em $\boldsymbol{\omega}_0$. A maior curvatura contém as observações que mais influenciam em $LD(\boldsymbol{\omega})$. Cook (1986) mostrou que a expressão para a curvatura $C_{\mathbf{l}}$ em $\boldsymbol{\omega}_0$, pode ser escrita como:

$$C_{\mathbf{l}} = 2|\mathbf{l}^T \boldsymbol{\Delta}^T \ddot{L}^{-1} \boldsymbol{\Delta} \mathbf{l}|, \quad (2.4)$$

em que $\|\mathbf{l}\| = 1$, $-\ddot{L}$ é a matriz de informação observada, em que

$$\ddot{L} = \left. \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad e \quad (2.5)$$

$$\boldsymbol{\Delta} = \left. \frac{\partial^2 \ell(\boldsymbol{\theta} | \boldsymbol{\omega})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0}. \quad (2.6)$$

Por exemplo, sob o esquema de ponderação de casos, tratado mais adiante, uma possibilidade é considerar a matriz $\Delta^T \ddot{L}^{-1} \Delta$ e determinar o autovetor \mathbf{l}_{max} a correspondente ao maior autovalor C_{max} , em que, o gráfico de \mathbf{l}_{max} com as respectivas observações podem revelar observações influentes.

A seguir, vamos descrever alguns tipos de perturbações que serão utilizados nesta dissertação.

2.2.1 Ponderação de casos

Seja o modelo de regressão

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.7)$$

em que, \mathbf{Y} é o vetor ($n \times 1$) da variável resposta; \mathbf{X} é a matriz ($n \times p$) de covariáveis; $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$ o vetor ($p \times 1$) de parâmetros desconhecidos e $\boldsymbol{\epsilon}$ o vetor ($n \times 1$) de erro aleatório, com $\epsilon_i \stackrel{ind.}{\sim} N(0, \sigma^2)$, $i = 1, \dots, n$.

Vamos considerar os casos em que, σ^2 é conhecida e também desconhecida (em que $\hat{\sigma}^2$ considere-se o valor do estimador de máxima verossimilhança de σ^2 , em qualquer dos dois casos).

- Quando σ^2 é conhecida

Seja $\boldsymbol{\omega}_{n \times 1}$ um vetor de perturbação do modelo em 2.7, assumindo que σ^2 é conhecida, a parte relevante da log-verossimilhança para o modelo perturbado é dada por $\ell(\boldsymbol{\beta}|\boldsymbol{\omega}) = -\frac{1}{2\sigma^2} \sum_{i=1}^n w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$, em que, w_i e y_i são os componentes de $\boldsymbol{\omega}$ e \mathbf{Y} , respectivamente, e \mathbf{x}_i^T é a i -ésima linha da matriz \mathbf{X} . Diferenciando $\ell(\boldsymbol{\beta}|\boldsymbol{\omega})$ em relação a, $\boldsymbol{\beta}$ e $\boldsymbol{\omega}$ e avaliando em $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1}$, obtemos $\Delta = \mathbf{X}^T D(\mathbf{e})/\sigma^2$, em que, $\mathbf{e} = (e_1, \dots, e_n)^T$ é o vetor $n \times 1$ de resíduos comuns quando $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1}$ e $D(\mathbf{e}) = \text{diag}(e_1, \dots, e_n)$, em que, diag representa a matriz diagonal. Uma vez que $\ddot{L} = \left. \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = -\mathbf{X}^T \mathbf{X}/\sigma^2$, C_l é a curvatura do autovetor \mathbf{l} associado a algum autovalor da matriz $\ddot{F} = \Delta^T \ddot{L}^{-1} \Delta$, em que

$$\begin{aligned} C_l &= 2 \left| \mathbf{l}^T \Delta^T \ddot{L}^{-1} \Delta \mathbf{l} \right| \\ &= 2 \mathbf{l}^T D(\mathbf{e}) P_{\mathbf{X}} D(\mathbf{e}) \mathbf{l} / \sigma^2, \end{aligned} \quad (2.8)$$

em que, $P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

- Quando σ^2 não é conhecida

Seja $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2)^T$, $\mathbf{1}_n^T = (1, \dots, 1)_n$ e calculando a derivada da verossimilhança perturbada em relação a σ^2 e $\boldsymbol{\omega}^T$ temos que:

$$\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \sigma^2 \partial \boldsymbol{\omega}^T} &= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\frac{\partial}{\partial \sigma^2} \sum_{i=1}^n w_i \ell_i(\boldsymbol{\theta}) \right] \\
&= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\frac{\partial}{\partial \sigma^2} \sum_{i=1}^n w_i \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}} \right\} \right] \\
&= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\frac{\partial}{\partial \sigma^2} \sum_{i=1}^n w_i \log \left\{ (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}} \right\} \right] \\
&= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\frac{\partial}{\partial \sigma^2} \sum_{i=1}^n \left\{ -\frac{1}{2} w_i \log(2\pi) - \frac{1}{2} w_i \log \sigma^2 - \frac{1}{2} \frac{w_i}{\sigma^2} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\} \right] \\
&= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\sum_{i=1}^n \left\{ -\frac{1}{2} \frac{w_i}{\sigma^2} + \frac{1}{2} \frac{w_i}{\sigma^4} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\} \right] \\
&= \frac{1}{2} \left(\frac{\mathbf{e}_{sq}^T}{\sigma^4} - \frac{\mathbf{1}_n^T}{\sigma^2} \right) \Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\
&= \frac{1}{2} \left(\frac{\mathbf{e}_{sq}^T}{\hat{\sigma}^4} - \frac{\mathbf{1}_n^T}{\hat{\sigma}^2} \right). \tag{2.9}
\end{aligned}$$

em que, $\hat{\sigma}^2$ é o estimador de máxima verossimilhança de σ^2 e \mathbf{e}_{sq} é um vetor $n \times 1$ com elementos $\mathbf{e}_{sq} = \mathbf{e} \odot \mathbf{e} = (e_1^2, \dots, e_n^2)^T$, $i = 1, \dots, n$.

Um cálculo semelhante produz:

$$\Delta = \begin{pmatrix} \frac{\mathbf{X}^T D(\mathbf{e})}{\hat{\sigma}^2} \\ \frac{\mathbf{e}_{sq}^T}{2\hat{\sigma}^4} - \frac{\mathbf{1}_n^T}{2\hat{\sigma}^2} \end{pmatrix}, \tag{2.10}$$

Neste caso,

$$\ddot{L} = - \begin{pmatrix} \mathbf{X}^T \mathbf{X} / \hat{\sigma}^2 & \mathbf{0} \\ \mathbf{0} & n / (2\hat{\sigma}^4) \end{pmatrix}. \tag{2.11}$$

2.2.2 Perturbação na variável explanatória

Seja s_k , $k = 0, 1, \dots, p-1$, os fatores de escala para as diferentes unidades de medida associadas com as colunas de \mathbf{X} . Então, a log-verossimilhança é construída de

(2.7) com \mathbf{X} substituído por

$$\mathbf{X}_\omega = \mathbf{X} + WS$$

em que, W é uma matriz $n \times p$ de perturbações e $S = \text{diag}(s_0, s_1, \dots, s_{p-1})$, com $s_0 = 0$. Os elementos da diagonal s_k de S convertem as perturbações genéricas w_{jk} a tamanhos apropriados e de modo que as unidades $w_{jk}s_k$ é compatível com o jk -ésimo elemento de \mathbf{X} .

- Quando σ^2 é conhecida

Neste caso, Δ é uma matriz $p \times np$ particionada em $\Delta = (\Delta_0, \Delta_1, \dots, \Delta_{p-1})$, com os elementos Δ_k que é uma matriz $(p \times n)$, dada por $\frac{\partial^2 \ell(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial \beta_i \partial w_{jk}}$, $i = 0, 1, \dots, p-1$ e $j = 1, 2, \dots, n$. Então,

$$\Delta_k = s_k(d_k e^T - \hat{\beta}_k \mathbf{X}^T) / \sigma^2, \quad (2.12)$$

para $k = 0, 1, \dots, p-1$, em que, d_k é um vetor $p \times 1$ com 1 na k -ésima posição e zeros em outros lugares.

- Quando σ^2 não é conhecida

Seja $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2)^T$, então a parte relevante da log-verossimilhança perturbada, em que, $\boldsymbol{\omega}_i^T$ é a i -ésima linha da matriz W , é dada por:

$$\begin{aligned} \ell(\boldsymbol{\theta}|\boldsymbol{\omega}) &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\mathbf{x}_i^T + \boldsymbol{\omega}_i^T S)\boldsymbol{\beta})^2 \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \boldsymbol{\omega}_i^T S \boldsymbol{\beta})^2 \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \boldsymbol{\omega}_i^T S \boldsymbol{\beta})^2 \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - [w_{i0}s_0\beta_0 + w_{i1}s_1\beta_1 + \dots + \\ &\quad + w_{i(k-1)}s_{k-1}\beta_{k-1} + w_{ik}s_k\beta_k + w_{i(k+1)}s_{k+1}\beta_{k+1} + \dots + \\ &\quad + w_{i(p-1)}s_{p-1}\beta_{p-1}])^2. \end{aligned}$$

Derivando em relação a β_k , temos que:

$$\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \beta_k} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \boldsymbol{\omega}_i^T S \boldsymbol{\beta})(x_{ik} + w_{ik}s_k).$$

Derivando agora, em relação a w_{ik} , e colocado no ponto $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ temos:

$$\begin{aligned} \frac{\partial}{\partial w_{ik}} \left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \beta_k} \right) &= \frac{1}{\hat{\sigma}^2} \left(-s_k \hat{\beta}_k x_{ik} + e_i s_k \right), \\ &= \frac{s_k}{\hat{\sigma}^2} \left(e_i - \hat{\beta}_k x_{ik} \right). \end{aligned} \quad (2.13)$$

Derivando em relação a β_m , para $m = 0, 1, \dots, p-1$, quando $m \neq k$, temos:

$$\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \beta_m} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \boldsymbol{\omega}_i^T S \boldsymbol{\beta}) (x_{im} + w_{im} s_m).$$

Derivando agora em relação a w_{ik} , temos:

$$\begin{aligned} \frac{\partial}{\partial w_{ik}} \left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \beta_m} \right) &= \frac{1}{\sigma^2} (-s_k \beta_k x_{im}) \\ &= \frac{s_k}{\sigma^2} (-\beta_k x_{im}). \end{aligned} \quad (2.14)$$

Derivando em relação a σ^2 , temos:

$$\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \boldsymbol{\omega}_i^T S \boldsymbol{\beta})^2$$

Derivando agora em relação a w_{ik} e colocado no ponto $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, temos:

$$\frac{\partial}{\partial w_{ik}} \left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \sigma^2} \right) = \frac{s_k}{\sigma^4} (\beta_k e_i). \quad (2.15)$$

Então,

$$\boldsymbol{\Delta}_k = \begin{pmatrix} s_k (d_k e^T - \hat{\beta}_k \mathbf{X}^T) / \hat{\sigma}^2 \\ -s_k (\hat{\beta}_k e^T) / \hat{\sigma}^4 \end{pmatrix}_{(p+1) \times n}. \quad (2.16)$$

2.2.3 Perturbação na variável resposta

Seja s_y , o fator de escala para a unidade de medida associadas com a variável \mathbf{Y} . Então a log-verossimilhança perturbada é construído de (2.7) com \mathbf{Y} substituído por

$$\mathbf{Y}_\omega = \mathbf{Y} + s_y \boldsymbol{\omega}$$

em que, $\boldsymbol{\omega}$ é um vetor $n \times 1$ de perturbações. O elemento s_y converte as perturbações genéricas w_j a tamanhos apropriados e de modo que as unidades $s_y w_j$ seja compatível com o j -ésimo elemento de \mathbf{Y} .

- Quando σ^2 é conhecida

Neste caso, $\boldsymbol{\Delta}$ é a matriz $p \times n$ em que, os elementos são dados por $\frac{\partial^2 \ell(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial \beta_i \partial w_j}$, $i = 0, 1, \dots, p-1$; $j = 1, 2, \dots, n$. Então

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}^T} &= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i + s_y w_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\} \right] \\ &= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\left\{ \frac{1}{\sigma^2} \sum_{i=1}^n (y_i + s_y w_i - \mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i \right\} \right] \\ &= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\frac{1}{\sigma^2} \mathbf{X}^T (\mathbf{Y} + s_y \boldsymbol{\omega} - \mathbf{X} \boldsymbol{\beta}) \right] \\ \boldsymbol{\Delta} &= \frac{s_y}{\sigma^2} \mathbf{X}^T \end{aligned} \quad (2.17)$$

- Quando σ^2 não é conhecida

Seja $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2)^T$ e calculando a derivada da verossimilhança perturbada, temos que:

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \sigma^2 \partial \boldsymbol{\omega}^T} &= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\frac{\partial}{\partial \sigma^2} \sum_{i=1}^n \left\{ -\frac{1}{2\sigma^2} (y_i + s_y w_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\} \right] \\ &= \frac{\partial}{\partial \boldsymbol{\omega}^T} \left[\sum_{i=1}^n \left\{ +\frac{1}{2\sigma^4} (y_i + s_y w_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\} \right] \\ &= \frac{s_y}{\sigma^4} ((\mathbf{y} - \mathbf{X}^T \boldsymbol{\beta})^T + s_y \boldsymbol{\omega}^T) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0} \\ &= \frac{s_y}{\sigma^4} (\mathbf{e}^T) \end{aligned} \quad (2.18)$$

$$\boldsymbol{\Delta} = \begin{pmatrix} \frac{1}{\sigma^2} s_y \mathbf{X}^T \\ \frac{1}{\sigma^4} s_y (\mathbf{e}^T) \end{pmatrix}. \quad (2.19)$$

2.2.4 Perturbação na variância do erro

Neste esquema de perturbação faremos:

$$\sigma_{\boldsymbol{\omega}_i}^2 = \frac{\sigma^2}{w_i}, i = 1, \dots, n.$$

Seja $\boldsymbol{\omega}_{n \times 1}$ um vetor de perturbação do modelo em (2.7), assumindo que σ^2 é não conhecida, a parte relevante da log-verossimilhança para o modelo perturbado é dada por $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = -\sum_{i=1}^n \left\{ \frac{1}{2} \log \frac{\sigma^2}{w_i} + \frac{w_i}{2\sigma^2} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\}$, $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2)^T$ em que, w_i e y_i são os componentes de $\boldsymbol{\omega}$ e \mathbf{Y} , respectivamente, e \mathbf{x}_i^T é a i -ésima linha da matriz \mathbf{X} .

Diferenciando $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ em relação a, $\boldsymbol{\beta}$ e $\boldsymbol{\omega}$ e avaliando em $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\sigma^2 = \hat{\sigma}^2$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1}$, obtemos

$$\Delta_1 = \mathbf{X}^T D(\mathbf{e}) / \sigma^2.$$

Diferenciando $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ em relação a, σ^2 e $\boldsymbol{\omega}$ e avaliando em $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\sigma^2 = \hat{\sigma}^2$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1}$, obtemos

$$\Delta_2 = \frac{\mathbf{e}_{sq}^T}{2\hat{\sigma}^4}.$$

Obtemos

$$\Delta = \begin{pmatrix} \Delta_1 \\ \Delta_2 \end{pmatrix}$$

$$\Delta = \begin{pmatrix} \mathbf{X}^T D(\mathbf{e}) / \sigma^2 \\ \frac{\mathbf{e}_{sq}^T}{2\hat{\sigma}^4} \end{pmatrix}$$

em que, $\mathbf{e} = (e_1, \dots, e_n)^T$ é o vetor $n \times 1$ de resíduos comuns quando $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1}$ e $D(\mathbf{e}) = \text{diag}(e_1, \dots, e_n)$, em que, diag representa a matriz diagonal.

Em que, $\hat{\sigma}^2$ é o estimador de máxima verossimilhança de σ^2 e \mathbf{e}_{sq} é um vetor $n \times 1$ com elementos $\mathbf{e}_{sq} = \mathbf{e} \odot \mathbf{e} = (e_1^2, \dots, e_n^2)^T$, $i = 1, \dots, n$.

2.3 Curvatura Normal Conformal de Poon e Poon (1999)

Poon e Poon (1999) estudaram a influência local proposta por Cook (1986) mediante a curvatura normal conformal. O enfoque de influência local é um método de diagnóstico que estuda o comportamento local de uma superfície chamada gráfico de influência. Mais precisamente, Cook (1986) sugere que examinemos a direção em que a curvatura normal é máxima, pois a partir dela identificamos observações potencialmente influentes sob o esquema de perturbação no qual o modelo está sujeito. Porém, Poon e Poon (1999) chamam a atenção para o fato de que a curvatura normal pode assumir qualquer valor real e não é invariante sob uma mudança uniforme de escala, ocasionando perda de objetividade no julgamento da grandeza da curvatura. Com o objetivo de solucionar essa problemática e, conseqüentemente, aperfeiçoar o método de influência local, propõe fazer o uso da curvatura normal conformal que está relacionada

com a curvatura normal, mas assume valores em um intervalo limitado da reta real e é invariante sob uma classe de reparametrizações. A curvatura normal conformal combinada com suas propriedades nos dão suporte para construirmos valores de referência que permitem julgarmos, a grandeza dessa curvatura, de forma objetiva, flexível e teria maior aplicabilidade ao método.

Seja \mathbf{I} a primeira forma fundamental de um mapa da superfície e $\mathbf{\Pi}$ a segunda forma fundamental de um mapa da superfície representados por

$$\mathbf{I}_{ij} = \delta_{ij} + \frac{\partial f}{\partial w_i} \frac{\partial f}{\partial w_j} \quad (2.20)$$

e

$$\mathbf{\Pi}_{ij} = \frac{1}{(1 + |\nabla_f|^2)^{1/2}} \frac{\partial^2 f}{\partial w_i \partial w_j}$$

em que, $f = LD(\boldsymbol{\omega})$, δ_{ij} é igual a 1 se $i = j$ e é zero caso contrario, e $|\nabla_f|$ representa a norma do vetor gradiente de f . Essas duas formas são avaliadas nos vetores \mathbf{v} e \mathbf{w} por $\mathbf{I}(\mathbf{v}, \mathbf{w}) = \mathbf{v}^T \mathbf{I} \mathbf{w}$ e $\mathbf{\Pi}(\mathbf{v}, \mathbf{w}) = \mathbf{v}^T \mathbf{\Pi} \mathbf{w}$. Uma reta $\boldsymbol{\omega}(a) = \boldsymbol{\omega}_0 + a\mathbf{l}$ é definido em Ω passando por $\boldsymbol{\omega}_0$, em que $a \in \mathbb{R}$, $\boldsymbol{\omega}_0$ é o vetor de não perturbação e $\mathbf{l} \in \mathbb{R}^q$. Então a curvatura normal do gráfico $\boldsymbol{\alpha}$ (2.3) na direção \mathbf{l} no ponto $\boldsymbol{\omega}_0$ é dada por:

$$C_{\mathbf{l}} = C(\mathbf{l}, \mathbf{l}) = \frac{\mathbf{\Pi}(\mathbf{l}, \mathbf{l})}{\mathbf{I}(\mathbf{l}, \mathbf{l})} = \frac{\mathbf{l}^T \mathbf{H}_f \mathbf{l}}{\mathbf{l}^T (\mathbf{I}_n + \nabla_f \nabla_f^T) \mathbf{l} (1 + |\nabla_f|^2)^{1/2}} \Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} \quad (2.21)$$

em que, \mathbf{I}_n é a matriz identidade $n \times n$ e

$$\mathbf{H}_f = \frac{\partial^2 f}{\partial w_i \partial w_j}$$

é a matriz hessiana. Cook (1986) propõe usar a curvatura normal para estudar características do gráfico de influência e deduziu que se $\mathbf{l}^t \mathbf{l} = 1$, então a equação (2.21) é reduzida a

$$C_{\mathbf{l}} = \mathbf{l}^T \mathbf{H}_f \mathbf{l} |_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} \quad (2.22)$$

da equação (2.22) temos que

$$C_{\mathbf{l}} = -2(\mathbf{l}^T \ddot{F} \mathbf{l}) |_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} \quad (2.23)$$

em que, \ddot{F} é uma matriz $q \times q$ com elementos dados por $\partial^2 \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}) / \partial w_i \partial w_j$. Seja $\boldsymbol{\Delta}$ uma matriz $p \times q$ como definido em (2.6) e \ddot{L} uma matriz $p \times p$ como definido em (2.5) e avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Então a equação (2.23), pode ser escrita como (Cook,

1986):

$$C_l = -2\{\mathbf{l}^T \Delta^T (\ddot{L})^{-1} \Delta \mathbf{l}\} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0}. \quad (2.24)$$

A matriz $-\ddot{F}$ é positiva semidefinida em $\boldsymbol{\omega}_0$ e a função de afastamento pela verossimilhança apresenta o mínimo em $\boldsymbol{\omega}_0$. Seja $C_{max} = \max_l(C_l)$ o qual corresponde ao maior autovalor de $-\ddot{F}$, e seja \mathbf{l}_{max} o autovetor correspondente. Cook (1986) sugere que um valor grande de C_{max} (aproximadamente uma curvatura maior que 2) é uma indicação de um problema local sério (sensibilidade local), e se o i -ésimo elemento em \mathbf{l}_{max} for relativamente grande uma especial atenção deve ser dada ao elemento perturbado por w_i . Apesar da metodologia ter sido mostrada muito útil, muitas questões foram levantadas em Cook (1986) e para tentar resolver estas questões Poon e Poon (1999) introduziram a curvatura normal conformal.

Curvatura Normal Conformal e suas propriedades

A curvatura normal conformal no ponto $\boldsymbol{\omega}_0$ de um gráfico $\boldsymbol{\alpha}$ na direção \mathbf{l} é dado por:

$$B_l = \frac{\boldsymbol{\Pi}(\mathbf{l}, \mathbf{l})}{\mathbf{I}(\mathbf{l}, \mathbf{l}) \{tr(\boldsymbol{\Pi}^2)\}^{\frac{1}{2}}} \Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}. \quad (2.25)$$

Dado que a representação da matriz $\boldsymbol{\Pi}$ é simétrica, quando os autovalores da segunda forma fundamental são λ_i , $i = 1, \dots, n$, $tr(\boldsymbol{\Pi}^2) = \sum_{i=1}^n \lambda_i^2$. Definindo $\|\mathbf{H}_f\| = \sqrt{tr(\mathbf{H}_f^2)}$ então da equação (2.21) e (2.25) temos que

$$B_l = \frac{1}{\mathbf{l}^T (\mathbf{I}_n + \nabla_f \nabla_f^T) \mathbf{l}} \frac{\mathbf{l}^T \mathbf{H}_f \mathbf{l}}{\|\mathbf{H}_f\|} \Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}. \quad (2.26)$$

Da equação (2.22)-(2.24) podemos deduzir que a curvatura normal conformal do gráfico na direção \mathbf{l} no ponto crítico $\boldsymbol{\omega}_0$ é

$$B_l = \frac{\mathbf{l}^T \mathbf{H}_f \mathbf{l}}{\|\mathbf{H}_f\|} \Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = - \frac{\mathbf{l}^T \ddot{F} \mathbf{l}}{\sqrt{tr(\ddot{F}^2)}} \Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{\mathbf{l}^T \Delta^T (\ddot{L})^{-1} \Delta \mathbf{l}}{\sqrt{tr(\Delta^T (\ddot{L})^{-1} \Delta)^2}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0} \quad (2.27)$$

ou seja, com as expressões necessárias para calcular a curvatura normal podemos obter a curvatura normal conformal, sem precisar obter novas expressões.

Poon e Poon (1999) mostraram que:

1. Para qualquer direção \mathbf{l} , B_l satisfaz $0 \leq |B_l| \leq 1$

Se $\mathbf{l}_i : 1 \leq i \leq n$ é uma coleção de vetores ortonormais de $\mathbf{\Pi}$, então $\sum_i B_{i_i}^2 = 1$. B_{i_i} é igual ao autovalor normalizado $\hat{\lambda}_i$ o qual é

$$\hat{\lambda}_i = \frac{\lambda_i}{(\sum_{k=1}^n \lambda_k^2)^{1/2}}. \quad (2.28)$$

2. Cook (1986) sugere inspecionar o autovetor \mathbf{l}_{max} com a máxima curvatura normal C_{max} independente do tamanho; as duas curvaturas são medidas equivalentes.
3. Um autovetor \mathbf{l} é q influente se $|B_{i_i}| \geq q/\sqrt{n}$.
4. Influência do autovetor individual.

Seja \mathbf{e}_t o vetor coluna cuja t -ésima entrada é 1 e todas as outras são 0. Chamamos \mathbf{e}_t o vetor de perturbação básica do espaço de perturbação. Seja $\{\mathbf{l}_i : 1 \leq i \leq n\}$ uma coleção de autovetores ortonormais com autovalores normalizados e $\hat{\lambda}_i$ dado na equação (2.28). Quando $\mathbf{l}_i = \sum_{t=1}^n a_{it}\mathbf{e}_t$, temos $\sum_{t=1}^n a_{it}^2 = 1$. Isto significa que, para qualquer i fixo, se a contribuição de todos os a_{it} for uniforme, então $|a_{it}| = 1/\sqrt{n}$. Com isso, pode-se construir, inclusive, o valor de referência. Este método pode ser aplicado para estudar \mathbf{l}_{max} (Poon e Poon, 1999).

5. Contribuição agregada dos vetores básicos de perturbação.

Definindo-se $u_i = |\hat{\lambda}_i|$ os valores absolutos dos autovalores próprios normalizados

$$u_{max} = u_1 \geq \dots \geq u_k \geq q/\sqrt{n} > u_{k+1} \dots u_n \geq 0$$

e usando a_{ij} para denotar o j -ésimo elemento do autovetor normalizado correspondente a u_i tal que $\mathbf{l}_i = \sum_{t=1}^n a_{it}\mathbf{e}_t$, então a contribuição agregada do t -ésimo vetor básico de perturbação para todos os autovetores q -influyente (q -ésimo influyente) é dado por $m(q)_t = \sqrt{\sum_{i=1}^k \mu_i a_{it}^2}$.

Como $\sum_{t=1}^n m(q)_t^2 = \sum_t (\sum_{i=1}^k \mu_i a_{it}^2) = \sum_{i=1}^k \mu_i (\sum_{t=1}^n a_{it}^2) = \sum_{i=1}^k \mu_i$, se a contribuição de todos os vetores básicos de perturbação forem uniforme, então cada um é igual a

$$\bar{m}(q) = \sqrt{\frac{1}{n} \sum_{i=1}^k \mu_i} \quad (2.29)$$

ou seja, se a contribuição de todos os vetores de perturbação básicos for uniforme, então cada contribuição deve ser igual a $\bar{m}(q)$. Isto significa que, ao analisarmos a contribuição de $\bar{m}(q)$, determinamos a significância da contribuição individual dos vetores de perturbação básicos.

Existem dois casos, a primeira é fazer com que q seja suficientemente grande para que a contribuição individual dos vetores de perturbação básicos seja considerada apenas para \mathbf{l}_{max} . A outra é fazer $q = 0$ de modo que todos os autovalores sejam incluídos na análise.

6. Quando $q = 0$, a contribuição $m(q)_t$ é chamada de contribuição total e é dada por

$$m_t = m(0)_t = \sqrt{\sum_{i=1}^n \mu_i a_{it}^2}.$$

7. Poon e Poon (1999) sugerem comparar a contribuição total $m(0)_t$ com

$$\bar{m} = \bar{m}(0) = \sqrt{\frac{1}{n} \sum_{i=1}^n \mu_i} = \sqrt{\sum_{i=1}^n |\lambda_i| / n} \sqrt{\sum_{i=1}^n \lambda_i^2}$$

ou seja, se a contribuição de todos os parâmetros básicos de perturbação for uniforme, então cada contribuição total deve ser igual a $\bar{m}(0)$.

8. A contribuição total m_t e a curvatura normal conformal B_{e_t} do vetor de perturbação básico são altamente relacionadas. Se todos os autovalores forem não-negativos, então B_{e_t} será igual ao quadrado da contribuição total do t -ésimo vetor de perturbação básico. Resumindo, se a matriz Hessiana \mathbf{H}_f é definida semipositiva e todos os autovalores são não-negativos então $m_t^2 = m_t^2(0) = B_{e_t}$ para todo t .

Se a contribuição de todos os B_{e_t} s é uniforme, então cada um é igual a

$$b = \frac{tr(\mathbf{\Pi})}{n\sqrt{tr(\mathbf{\Pi}^2)}}. \quad (2.30)$$

9. Comparamos B_{e_t} contra $2b = 2\frac{tr(\mathbf{\Pi})}{n\sqrt{tr(\mathbf{\Pi}^2)}} = 2\frac{tr(\ddot{F})}{n\sqrt{tr(\ddot{F}^2)}}$.

(observações influentes considerando a contribuição total).

10. Comparamos $m(q)_t$ contra $\sqrt{2}\bar{m}(q) = \sqrt{2}\sqrt{\frac{1}{n}\sum_{i=1}^k \mu_i}$

(observações influentes considerando a contribuição agregada dos autovetores q -influentes).

2.4 Procura “Forward” de Atkinson e Riani (2000)

No ajuste de modelos de regressão, deseja-se apontar a estrutura de dependência existente entre a variável resposta do estudo e as possíveis fontes de variação (as variáveis explicativas). A identificação da estrutura de dependência existente nos dados compreende a seleção das variáveis explicativas e estimação dos parâmetros do modelo proposto.

Em modelos de regressão linear o processo de seleção de variáveis explicativas e estimação dos parâmetros é muito sensível à presença de dados atípicos (Atkinson e Riani, 2000; Montgomery et al., 2011). Tais valores podem vir a camuflar (mascarar) a significância e homocedasticidade, feitas à componente do erro.

A etapa de verificação da adequabilidade das suposições feitas para o ajuste do modelo proposto é conhecida como análise de diagnóstico. Para essa análise, algumas técnicas de identificação de dados atípicos já são frequentemente utilizadas, como análise de resíduos, identificação de pontos de alavanca e análise da distância de Cook (1977). O método procura “forward” é inserido na análise de diagnóstico como um complemento às técnicas já existentes, em que através de processos gráficos, pode-se esclarecer o comportamento do modelo frente à presença de observações discrepantes.

A aplicação do método procura “forward” em modelos de regressão linear segue os mesmos passos descritos em Hadi (1992), ou seja, escolha do grupo de dados livre de dados atípicos, *grupo limpo*, que ao longo do processo será acrescido de observações provenientes do restante do conjunto de dados. O que é importante no nosso processo é que o subconjunto inicial, (grupo limpo) seja livre de dados atípicos ou contém valores atípicos desmascarados que são imediatamente removidos pelo processo procura “forward”. A procura é muitas vezes capaz de se recuperar de um início que não é muito robusta. Um exemplo, considerando modelos de regressão, é dada por Atkinson e Mulira (1993) e por Cerioli e Riani (1999).

No contexto de modelo de regressão linear, a escolha do grupo limpo será dada pelo ajuste do modelo a vários candidatos a grupo limpo: aquele que obtiver melhor ajuste será escolhido como ponto inicial do algoritmo.

O critério de distância para a entrada das observações restantes também sofre uma modificação em relação à distância descrita em Hadi (1992). Na análise descritiva a medida das observações candidatas a entrar no grupo limpo é o resíduo obtido a partir do modelo ajustado ao grupo limpo. Assim a observação com menor resíduo será incluída no grupo e a entrada da observação no grupo limpo segue até o momento em que contenha todas as observações do conjunto de dados brutos.

O método tem por objetivo produzir não apenas informações a respeito da detecção dos dados atípicos, mas também o efeito que cada observação tem no aspecto inferencial do modelo.

2.4.1 Descrição do método

A maioria dos métodos de detecção de dados atípicos procura dividir os dados em duas partes, uma referente aos dados limpos e a outra com possíveis dados atípicos. Os dados limpos são usados na estimação dos parâmetros. O método proposto por Atkinson e Riani (2000) segue o mesmo raciocínio, tendo como diferencial a estimação dos parâmetros, que é atualizada a cada passo do processo.

Segundo Paula (2004, Capítulo 1 p. 30) a remoção de pontos talvez seja a técnica mais conhecida para avaliar o impacto de uma observação particular nas estimativas de regressão, o que levaria a um artifício de classificação do conjunto de dados em dois grupos: o de dados atípicos, referente aos pontos excluídos, e o grupo dos dados limpos. Livros sobre diagnóstico de regressão, como Cook e Weisberg (1982), incluem fórmulas de diagnóstico nos casos de deleção múltipla, em que um pequeno número de dois ou três potenciais dados atípicos são excluídos de uma vez, mas a vasta combinação de casos a serem avaliados leva a uma tarefa onerosa de revisão dos ajustes.



Figura 2.2: A divisão em dois grupos.

Muitos métodos de detecção de dados atípicos múltiplos utilizam técnicas robustas para a classificação das observações em dados limpos e dados atípicos. Atkinson e Riani (2000) optam pela utilização do método de regressão robusta, cujas técnicas são

complementares às técnicas clássicas de mínimos quadrados, pois oferecem respostas similares a regressão por mínimos quadrados, quando existe uma relação linear entre a resposta e as variáveis explicativas, assumindo que os erros são normalmente distribuídos. Porém, as técnicas robustas diferem significativamente do ajuste por mínimos quadrados quando os erros não são normalmente distribuídos ou quando os dados contêm dados atípicos. Seu uso torna-se justificável porque, quanto maior o número de variáveis de um modelo, mais difícil a identificação de dados atípicos com o uso das técnicas de regressão clássica.

Devido ao fato de que uma simples observação atípica pode distorcer significativamente os resultados obtidos por meio da estimação por mínimos quadrados, o uso de métodos robustos, que utilizam estimadores resistentes a um certo percentual de dados atípicos, poderá fornecer resultados significativamente mais confiáveis. De modo, a formalizar o quão resistente a dados atípicos é um estimador, foi proposto o conceito de ponto de ruptura.

Segundo Donoho e Huber (1983) e Montgomery et al. (2001) o ponto de ruptura é uma medida global de robustez e numa amostra finita no contexto de regressão é a menor fração de dados atípicos que poderia causar perturbação ao estimador. A fração mínima para o ponto de ruptura é de $1/n$, em que, apenas com uma observação atípica o estimador do parâmetro já sofre viés dessa observação. O ponto de ruptura do estimador de mínimos quadrados é $1/n$. Um estimador robusto adequado seria aquele que possui o ponto de ruptura igual a 0.5, ou seja, um estimador que resiste a um percentual de até 50% dos dados atípicos. Um valor de ruptura maior que 0.5 não faria sentido, pois, se mais da metade das observações fossem dados atípicos não seria mais possível diferenciar quais seriam os dados atípicos.

Dentre os estimadores robustos tem-se o estimador LMS (Least Median of Squares) que fornece estimativas para os parâmetros que não são afetadas pela presença de dados atípicos.

O algoritmo procura “forward” na regressão linear é composto de três passos: o primeiro é concentrado na escolha do grupo limpo; o segundo é referente à forma de progresso do método e o terceiro é relacionado ao monitoramento das estatísticas durante o algoritmo.

O método começa com a escolha do grupo limpo, pois segundo Atkinson e Riani (2000) o importante no procedimento é que o conjunto inicial seja livre de dados atípicos (para garantir, o grupo limpo é escolhido através de estimativas robustas, como por exemplo, a mínima mediana dos quadrados). Rousseeuw (1984) propõe tomar de forma aleatória subconjuntos de tamanho k . Se $\binom{n}{k}$ (combinatória) for muito grande,

Atkinson e Riani (2000) recomendam tomar 1000 subconjuntos de amostras ou um grande número de candidatos, ajustar o modelo de regressão robusta aos grupos e tomar como o conjunto limpo de dados atípicos aquele que produzir o menor resíduo mediano quadrático. Assim, o grupo limpo tem tamanho $m_0 = k$, o vetor de parâmetros estimados desse grupo é denotado por $\hat{\theta}_{m_0}^*$ e o estimador de mínimos quadrados ao fim do processo será $\hat{\theta}_n^* = \hat{\theta}$. Na ausência de dados atípicos a seguinte relação ocorre:

$$E(\hat{\theta}_{m_0}^*) = E(\hat{\theta}) = \theta, \quad (2.31)$$

ou seja, ambos os estimadores serão não viesados.

Após selecionar o grupo seguinte com o menor LMS, o grupo passa a ter tamanho $m = m_0 + 1$. Além disso, pode acontecer que 2 ou mais unidades podem entrar ao grupo como uma ou mais deixar o grupo, crescendo de uma em uma unidade até n . O estimador da procura “forward”, $\hat{\theta}_{PF}$ é definido como a coleção de estimadores dos parâmetros produzidos a cada passo do processo

$$\hat{\theta}_{PF} = (\hat{\theta}_{m_0}^*, \dots, \hat{\theta}_n^*). \quad (2.32)$$

O método evita a inclusão de dados atípicos no conjunto inicial e produz uma ordenação natural dos dados de acordo com o modelo especificado. Tem-se que, na escolha do *grupo limpo*, é usado um método robusto, e ao mesmo tempo estimadores de mínimos quadrados. A introdução de observações atípicas é sinalizada por picos nas curvas que está sendo monitorado. As curvas presentes nos gráficos do método procura “forward” exibem o valor da estatística passo a passo no processo. Como por exemplo as estimativas dos coeficientes escalonados definidos como

$$\hat{\theta}_{escalonado}^* = \frac{signal(\bar{\theta})}{\bar{\theta}} \hat{\theta}_m^*, \quad (2.33)$$

em que, $\bar{\theta} = \sum_{m=m_0}^n \frac{\hat{\theta}_m^*}{n-m_0+1}$.

O método procura “forward” aplicado na regressão é robusto não devido à escolha de um particular estimador com alto ponto de ruptura, mas pela inclusão progressiva de unidades no grupo limpo que, no primeiro passo, são livres de dados atípicos. Como bônus do método, as observações podem ser naturalmente ordenadas de acordo com o modelo especificado. Além disso, o método possibilita analisar o efeito inferencial de unidades atípicas na análise estatística.

Segundo Atkinson e Riani (2000), o interesse está na evolução dos parâmetros e gráficos ou figuras das estimativas dos coeficientes, estimativas dos coeficientes escalonados, resíduos escalonados (Secção 2.2), resíduos escalonados ao quadrado etc., quando m cresce de $m_0 + 1$ a n .

Aplicações

Como motivação para o desenvolvimento, neste capítulo vamos considerar alguns conjuntos de dados e aplicaremos o método de influência global, local e procura “forward” em modelos de regressão linear.

Depois, no capítulo seguinte vamos propor o uso da influência local juntamente com a procura “forward”. Todos os conjuntos de dados utilizados neste capítulo e na aplicação encontram-se no Apêndice A.

3.1 Dados de ganso

Como um primeiro exemplo numérico, considere os dados de gansos para um observador como apresentado no Apêndice A.1 e relatado em Weisberg (2005, p.113).

Estudos aéreos às vezes dependem de métodos visuais para estimar o número de animais em uma área. Por exemplo, para estudar gansos em sua área de movimentação no verão no oeste da Baía de Hudson, no Canadá, foram utilizados aviões de pequeno porte que voam sobre uma faixa, e quando um bando de gansos eram flagrados, uma pessoa com experiência estimava o número de gansos no bando. Para investigar a confiabilidade deste método de contagem, um experimento foi conduzido em que um avião transportando dois observadores sobrevoou 45 bandos, e cada observador fez

uma estimativa do número de aves em cada bando independentemente, além disso, uma fotografia do rebanho foi feita de modo que uma segunda forma de contagem mais exata do número de aves do bando pudesse ser feita. Apresentamos as estatísticas descritivas para estes dados na Tabela 3.1 e o gráfico de dispersão na Figura 3.1, em que, x e y representam, respectivamente, o número de gansos estimado pelo observador 1 e número de gansos contados a partir de fotografias aéreas.

Tabela 3.1: Estatísticas descritivas dos dados de ganso.

Estatísticas descritivas	x	y
Mínimo	9.00	9.00
Máximo	500.00	409.00
Mediana	40.00	57.00
Média	71.00	89.31
Desvio padrão da média	12.85	13.10
Desvio padrão	86.22	87.85
Coefficiente de variação	1.21	0.98

Esses dados foram coletados em um esforço para determinar o quão bem os tamanhos dos rebanhos podem ser estimados visualmente durante um censo da população, e foram determinantes para a decisão de basear as contagens reais do censo em fotografias aéreas.

Weisberg (2005) ajustou um modelo linear simples para estes dados, de modo que:

- Y (variável resposta): o número de gansos por bando, contados a partir da fotografia aérea;
- X (covariável): o número de gansos estimado pelo observador 1 do experimento.

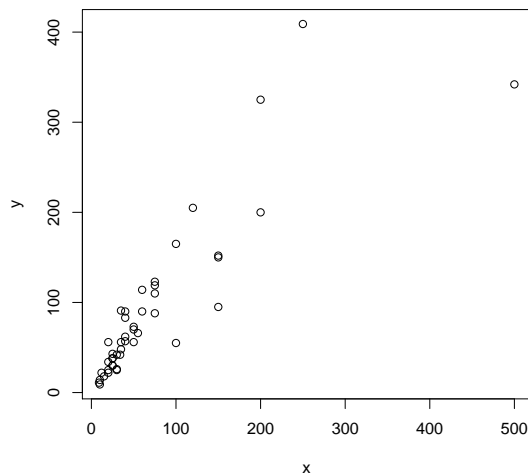


Figura 3.1: Gráfico de dispersão dos dados de ganso.

O modelo linear simples definido em (2.7), fica por tanto dado por:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \dots, 45,$$

com a suposição de que $\varepsilon_i \sim N(0, \sigma^2)$, com os ε_i , mutuamente independentes.

As estimativas dos parâmetros foram: $\hat{\alpha} = 26,650$ e $\hat{\beta} = 0,883$, resultando na Figura 3.2.

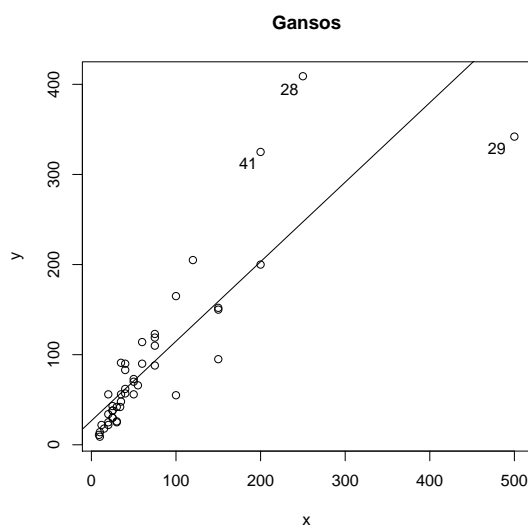


Figura 3.2: Diagrama de dispersão com a reta ajustada dos dados de ganso.

Assim, podemos esperar que a introdução de uma pequena perturbação revele alguma espécie de sensibilidade no modelo.

Primeiro, obtivemos algumas medidas de influência global.

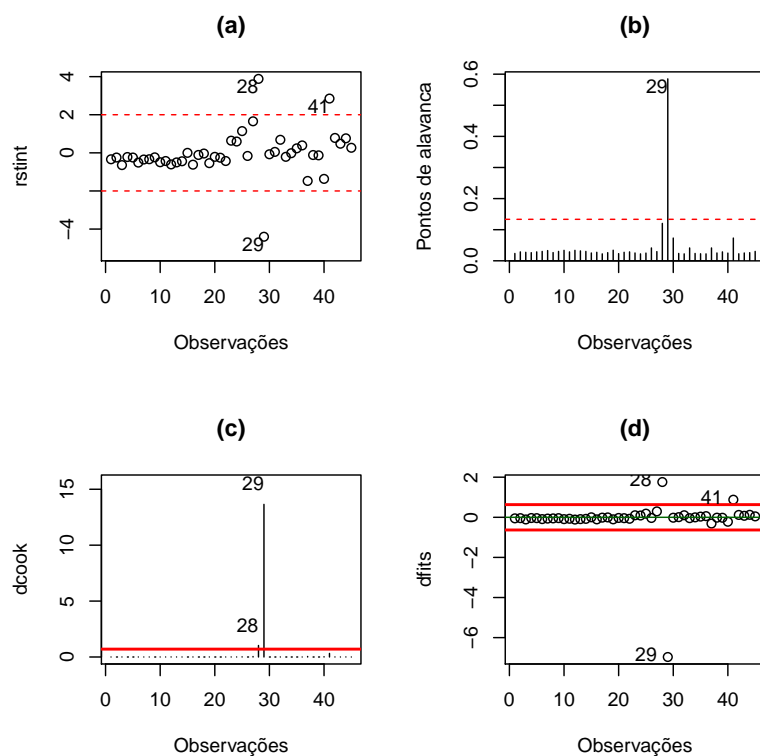


Figura 3.3: Gráfico de medida de influência global dos dados de ganso: (a) resíduos estudentizados, (b) pontos de alavanca, (c) distância de Cook e (d) DFFitS.

Cada uma das medidas de influência estão descritas na Secção 2.1 e como pode ser observado na Figura (3.3) em (a) e (d), resíduos estudentizados e DFFitS respectivamente, os dados 28, 29 e 41 estão afastados do conjunto de dados. Em (b) e (c) os dados 28 e 29 se destacam, considerando o ponto de alavanca e a distância de Cook. Estas três observações correspondem às três observações com o maior numero de gansos no bando e eles se destacam das demais observações como pode ser visto na Figura 3.2.

Tabela 3.2: Medida de influência global dos dados de ganso.

	DFBeta-intercepto	DFBeta- x	DFFit	D-Cook	Ponto de alavanca
28	-0.44	1.59_*	1.76_*	1.03_*	0.12
29	3.33_*	-6.83_*	-6.97_*	13.65_*	0.58_*
41	-0.10	0.73	0.88_*	0.32	0.07

_* indica que é dado atípico segundo a medida de influência

Na Tabela 3.2 temos 3 observações (28, 29 e 41), que podem ser dados influentes considerando as medidas de influência DFBetaS, DFFit e distância de Cook e além disso, temos um ponto de alavanca que é o 29.

3.1.1 Influência local

Nesta Secção, considerando o conjunto de dados dos gansos foram aplicados as perturbações descritas na Secção 2.2. Os valores dos Cmax obtidos para cada caso podem ser encontrados na Tabela 3.3.

Tabela 3.3: Valores do Cmax para os dados de ganso.

Esquema de perturbação	σ^2	Valor da curvatura máxima
Ponderação de casos	σ_C^2	14.37
	σ_{NC}^2	14.50
Perturbação na covariável	σ_C^2	8.19
	σ_{NC}^2	15.66
Perturbação na variável resposta	σ_C^2	8.19
	σ_{NC}^2	16.38
Pertubação na variância do erro	σ_{NC}^2	14.53

σ_C^2 : σ^2 conhecido; σ_{NC}^2 : σ^2 não conhecido

Para ver o comportamento, foram feitos os gráficos do autovetor, lmax, correspondente ao maior autovalor, Cmax, para cada tipo de perturbação. Os gráficos podem ser encontrados no Apêndice B.

Considerando a ponderação de casos com (σ^2 conhecido e desconhecido), nas Figuras B.1 e B.2, podemos ver que as observações 29, 28 e 41 se destacam como possíveis dados influentes e os gráficos correspondentes a cada caso (σ^2 conhecido e desconhecido) são muito semelhantes.

Os dados 29, 28 e 41 correspondem a observações com os maiores valores de x e y , e também os maiores valores dos resíduos. Ao retirar a observação 29, como esperado, existe uma grande variação nas estimativas dos parâmetros, como pode ser visto na Tabela 3.4

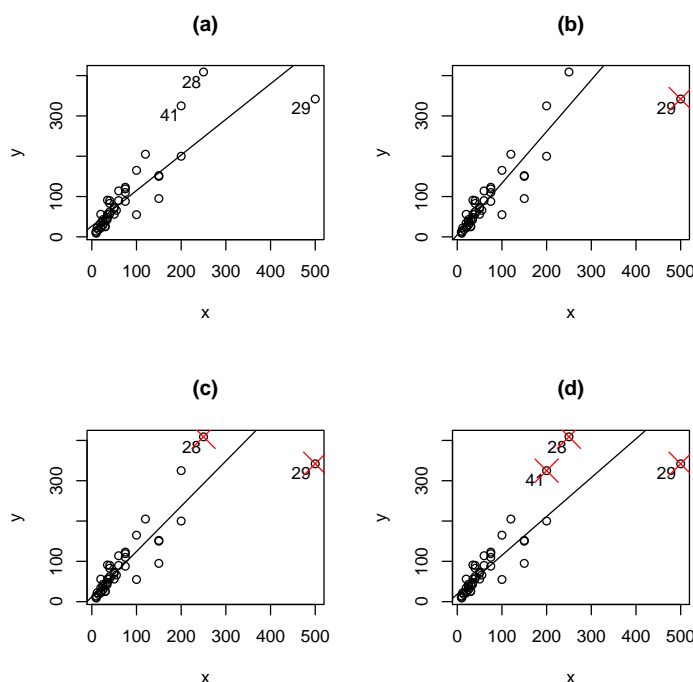


Figura 3.4: Retas ajustadas com os dados completos dos dados de ganso (a), sem a observação 29 (b), sem as observações 28 e 29 (c) e sem as observações 28, 29 e 41 (d).

Para a perturbação na covariável, foi utilizado o valor $s_x = 86.22$ (desvio padrão de x) em que, $S = \text{diag}(0, s_x)$. Na Figura B.3 (σ^2 conhecido) podemos ver claramente que o dado 29 se destaca das demais observações enquanto que na Figura B.4 (σ^2 desconhecido) os dados 29, 28 e 41 se destacam como possíveis pontos influentes.

Para perturbarmos a variável resposta, foi utilizado o valor $s_y = 87.85$ (desvio padrão de y). Nas Figura B.5 (σ^2 conhecido) e B.6 (σ^2 desconhecido) observamos que os possíveis pontos influentes são respectivamente as observações 29 e (28, 29 e 41).

Considerando agora a perturbação na variância do erro (σ^2 não conhecido) na Figura B.7 os casos que se destacaram foram as observações 28, 29 e 41.

Ao retirarmos a observação 29, (28 e 29) e (29 e 41) obtivemos; respectivamente, os seguintes valores para o p-valor: 0.493, 0.090 e 0.272. Como pode ser visto na Tabela 3.4, estes valores são maiores do que 0.05, ou seja, o intercepto passa a não ser significativo. Como era de se esperar as observações 28, 29 e 41 influenciam bastante nas estimativas dos parâmetros. As retas ajustadas para alguns casos podem ser vistos na Figura 3.4. Pelo gráfico, observa-se que a retirada da observação 29 é o que causa uma maior mudança na reta ajustada comparada com as outras 3 situações.

Na Tabela 3.4 foram obtidos as estimativas dos parâmetros após a retirada dos pontos 28, 29 e 41 e as possíveis combinações.

Tabela 3.4: Estimativa dos parâmetros dos dados de ganso com a mudança relativa entre parênteses.

Observação	$\hat{\beta}_0$	p_{valor}	$\hat{\beta}_1$	p_{valor}
todos	26.65	0.00	0.88	0.00
sem 28	29.71 (0.11)	0.00	0.78 (0.11)	0.00
sem 29	5.14 (0.81)	0.49	1.28 (0.45)	0.00
sem 41	27.41 (0.03)	0.00	0.83 (0.06)	0.00
sem 28 e 29	12.18 (0.54)	0.09	1.12 (0.27)	0.00
sem 28 e 41	30.85 (0.16)	0.00	0.71 (0.19)	0.00
sem 29 e 41	8.07 (0.70)	0.27	1.20 (0.36)	0.00
sem 28,29 e 41	18.57 (0.30)	0.00	0.96 (0.09)	0.00

() mudança relativa

3.1.2 Curvatura normal conformal

Consideramos o esquema de ponderação de casos apresentado na seção 2.2.1 em que, σ^2 é conhecido. A Figura 3.5 mostra os autovalores normalizados e os valores de q . Para q variando de 1 a 6, temos apenas um autovalor acima do valor q/\sqrt{n} . Na Tabela 3.5 foi apresentado os resultados para $q = 6$ (para os outros valores de q os resultados são iguais) e $q = 0$ (B_{E_j}). Em Poon e Poon (1999) foi sugerido o ponto de corte $2b$, com b dado na equação (2.30). Neste caso $2b = 0.047$. Para $m_j(q)$ foi usado o **limitante** $\bar{m}_j(q)\sqrt{2}$. Em todos os casos as observações 28, 29 e 41 aparecem como pontos influentes.

Tabela 3.5: Medidas de influência para os dados de ganso utilizando a curvatura normal conformal para o esquema de ponderação de casos.

	Número de autovetores influentes	Média	Limitante	Valores para os seguintes casos		
				28	29	41
$m_j(6)$	1	0.149	0.211	0.476	0.823	0.271
B_{E_j}	45	0.023	0.047	0.232	0.684	0.080

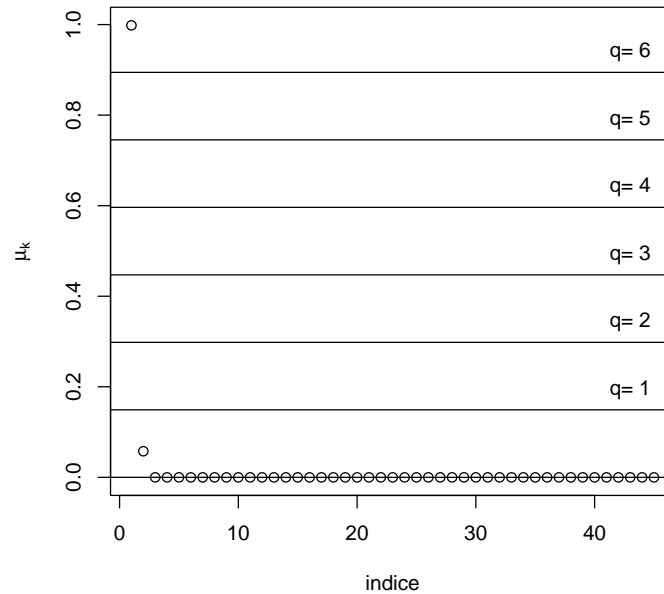


Figura 3.5: Autovalores normalizados dos dados de ganso.

Na Figura 3.6 foi feito o gráfico das observações com B_{E_j} e $q = 6$ que equivale ao gráfico de $lmax$ (Cook, 1986), (neste caso, para os outros valores de q (1,2,3,4 e 5) o resultado é igual).

Concluimos que os pontos 28, 29 e 41 são possíveis pontos influentes.

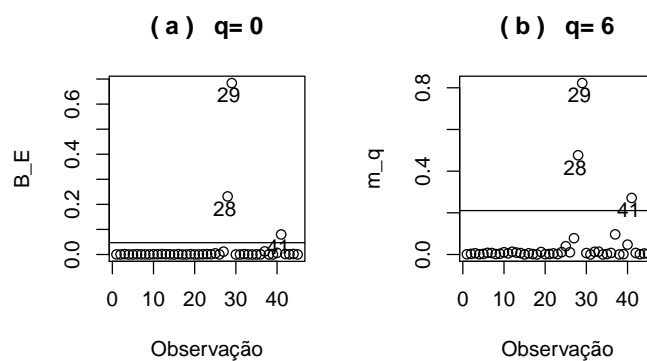


Figura 3.6: Contribuição agregada dos dados de ganso.

3.1.3 Procura “Forward”

Nesta subsecção considerando os dados dos gansos, utilizamos a metodologia de procura “forward” de Atkinson e Riani (2000) e obtivemos as Figuras (3.7), (3.8) e

(3.9). Na Figura 3.7, observa-se que o dado 29 se destaca no início e depois no final ele se destaca juntamente com as observações 28, 29 e 41.

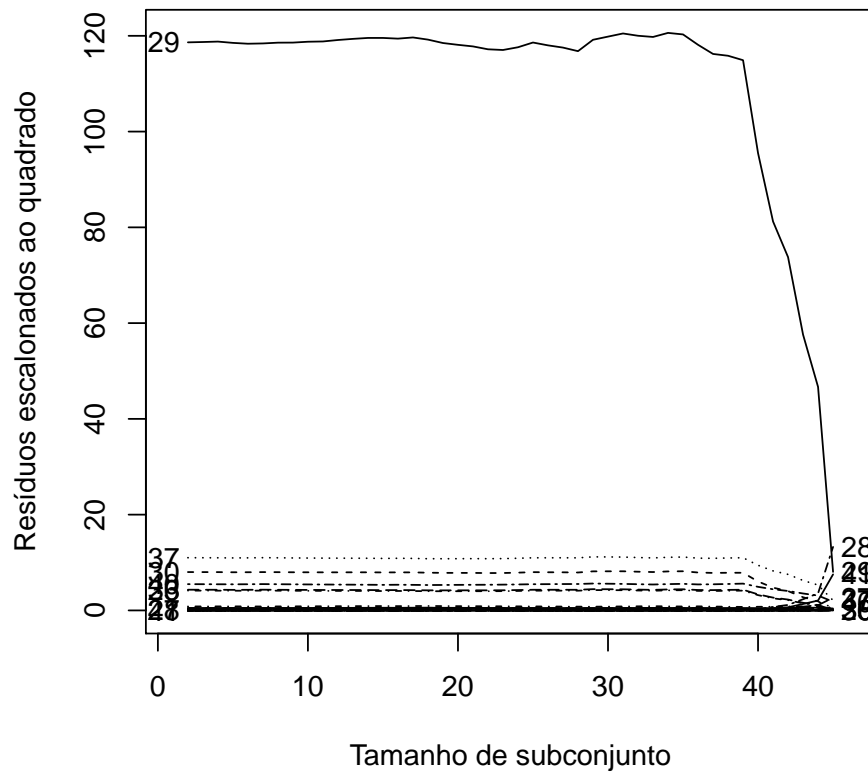


Figura 3.7: Gráfico dos resíduos escalonados ao quadrado utilizando a procura “forward” dos dados de ganso.

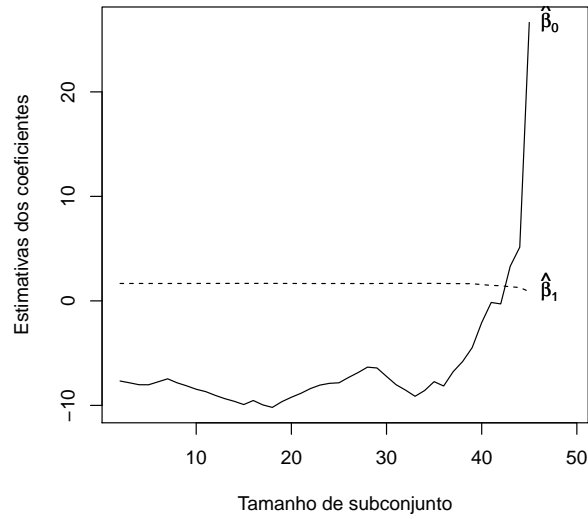


Figura 3.8: Gráfico das estimativas dos coeficientes utilizando a procura “forward” dos dados de ganso.

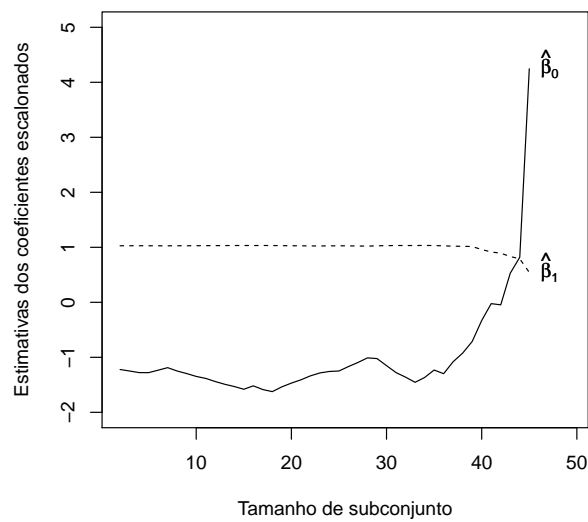


Figura 3.9: Gráfico das estimativas dos coeficientes escalonados (2.33) utilizando a procura “forward” dos dados de ganso.

Nas Figuras, 3.8 e 3.9, podemos observar que na ultima iteração da procura “forward” houve pico, ao ingressar o dado 29 e observou-se que o intercepto aumentou drasticamente.

Neste conjunto de dados não foi detectado nenhum dado mascarado utilizando a procura “forward”.

3.2 Dados de rato

Como um segundo conjunto de dados foram utilizados os dados de ratos e o modelo de regressão linear múltipla como foi feito em Weisberg (2005, p. 200).

Um experimento foi feito para investigar a quantidade de um determinado medicamento presente no fígado de um rato. Dezenove ratos foram selecionados aleatoriamente, pesados e colocados sob anestesia com éter. A dose efetiva que um animal recebeu foi determinada como aproximadamente 40 mg de fármaco por kg de peso corporal devido ao fato de que fígados grandes absorvem mais o fármaco do que fígados menores.

O peso do fígado é conhecido por estar fortemente relacionado com o peso corporal.

Depois de um determinado período de tempo, cada um dos ratos foi sacrificado, o fígado pesado, e foi determinada a porcentagem da dose no fígado. Observou-se a relação entre a porcentagem da dose no fígado (y) e o peso corporal ($BodyWt = x_1$), peso do fígado ($LiverWt = x_2$) e a dose relativa ($Dose = x_3$). O modelo linear simples definido em (2.7), fica por tanto dado por:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i, \quad i = 1, 2, \dots, 19,$$

com a suposição de que $\varepsilon_i \sim N(0, \sigma^2)$, com os ε_i , mutuamente independentes.

Os dados também podem ser obtidos no programa R (no pacote “alr3” ou pode ser visto no Apêndice A.2). Apresentamos as estatísticas descritivas para estes dados na Tabela 3.6 e o gráfico da matriz de dispersão na Figura 3.10.

Tabela 3.6: Estatísticas descritivas dos dados de rato.

Estatísticas descritivas	x_1	x_2	x_3	y
Mínimo	146.00	5.20	0.73	0.21
Máximo	200.00	10.00	1.00	0.56
Mediana	176.00	7.90	0.88	0.33
Média	171.53	7.81	0.86	0.34
Desvio padrão da média	3.78	0.28	0.02	0.02
Desvio padrão	16.49	1.22	0.09	0.09
Coefficiente de variação	0.10	0.16	0.10	0.26

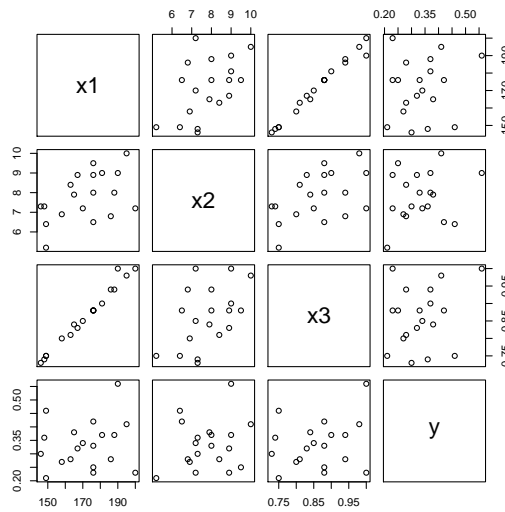


Figura 3.10: Gráfico da matriz de dispersão dos dados de rato.

Fazendo o ajuste do modelo, temos as seguintes estimativas dos parâmetros

Tabela 3.7: Estimativa dos parâmetros dos dados de rato.

	Estimativa	Desvio padrão	Estatística t	P-valor
(Intercepto)	0.2659	0.1946	1.37	0.1919
(peso corporal) x_1	-0.0212	0.0080	-2.66	0.0177
(peso do fígado) x_2	0.0143	0.0172	0.83	0.4193
(dose relativa) x_3	4.1781	1.5226	2.74	0.0151

Considerando as medidas de influência global foram obtidas a Figura 3.11 e a Tabela 3.8.

Tabela 3.8: Medida de influência segundo as medidas de influência DFBeta, DFFit, Distância de Cook e os pontos de alavanca descritos na Seção 2.1 dos dados de rato.

	DFBeta intercepto	DFBeta x_1	DFBeta x_2	DFBeta x_3	DFFit	D-Cook	Ponto de alavanca
1	-0.04	0.31	-0.70_*	-0.24	0.89	0.17	0.18
2	0.14	-0.10	-0.48_*	0.13	-0.61	0.09	0.18
3	-0.23	-1.67_*	0.30	1.75_*	1.90_*	0.93_*	0.85_*
5	0.52_*	-0.40	0.55_*	0.27	-0.91	0.20	0.39
13	-0.77_*	0.14	0.77_*	-0.12	-1.10	0.27	0.32
19	0.86_*	-0.25	-0.29	0.17	1.00	0.20	0.18

_* indica que é dado atípico segundo a medida de influência

Podemos observar que o dado 3 é possivelmente um dado atípico. A distância de Cook, foi dada por $D_3 = 0,93$ e $h_{33} = 0,85$. Retirando-se a observação 3 obtivemos a

Tabela 3.9, comparando com a Tabela 3.7 em que, pode ser visto que houve mudanças significativas no ajuste do modelo.

Tabela 3.9: Estimativa dos parâmetros dos dados de rato após retirar a observação 3.

	Estimativa	Desvio do erro padrão	Estatística t	P-valor
(Intercepto)	0.3114	0.2051	1.52	0.1512
(peso corporal) x_1	-0.0078	0.0187	-0.42	0.6838
(peso do fígado) x_2	0.0090	0.0187	0.48	0.6374
(dose relativa) x_3	1.4849	3.7131	0.40	0.6953

Observa-se que os p -valores são maiores que 0.05 o que implicaria que as variáveis não são significativas para o nosso modelo de forma individual.

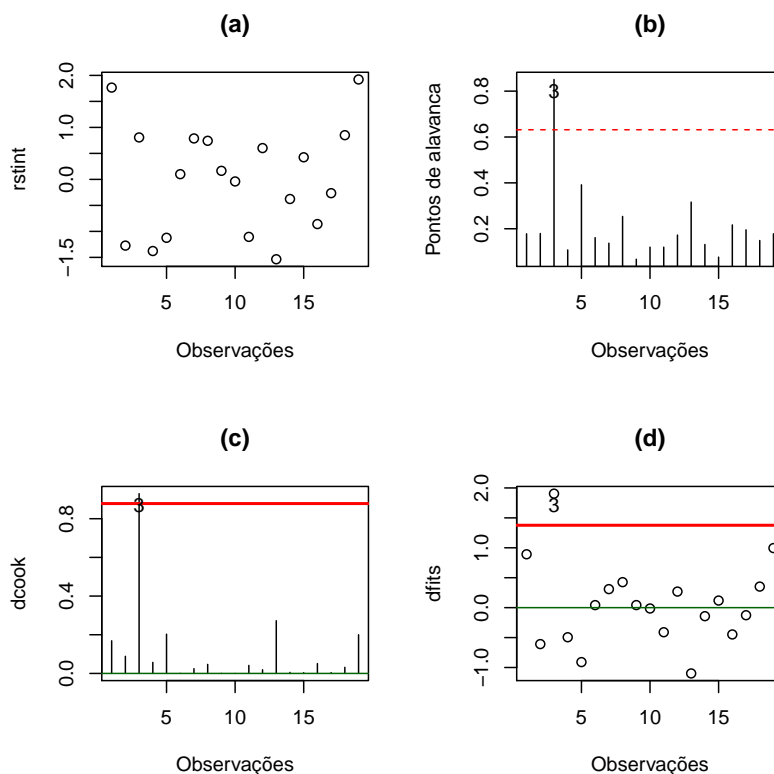


Figura 3.11: Gráfico de medida de influência global dos dados de rato: (a) resíduos estudentizados, (b) pontos de alavanca, (c) distância de Cook e (d) DFFitS.

3.2.1 Influência local

Considerando este conjunto de dados, foi feito a análise de influência local para cada tipo de perturbação descrita na Seção 2.2, os valores do C_{max} podem ser encontrados na Tabela 3.10.

Tabela 3.10: Valores do C_{max} para os dados de rato.

Esquema de perturbação	σ^2	Valor da curvatura máxima
Ponderação de casos	σ_C^2	3.58
	σ_{NC}^2	3.86
Perturbação na covariável	σ_C^2	324.60
	σ_{NC}^2	491.89
Perturbação na variável resposta	σ_C^2	3.14
	σ_{NC}^2	6.57
Perturbação na variância do erro	σ_{NC}^2	4.01

$\sigma_C^2: \sigma^2$ conhecido; $\sigma_{NC}^2: \sigma^2$ não conhecido

Foram construídos os gráficos do autovetor, l_{max} , correspondente ao maior autovalor, C_{max} , para cada tipo de perturbação. Os gráficos podem ser encontrados no Apêndice B.

Considerando a ponderação de casos (σ^2 conhecido e desconhecido) podemos observar que os pontos 1, 13 e 19 se destacam das demais observações, enquanto que na medida de influência global foi detectada a observação 3.

Para perturbar a covariável foi utilizado o valor $S = \text{diag}(s_0, s_1, s_2, s_3)$ em que, $s_1 = \text{desvio padrão}(x_1)$, $s_2 = \text{desvio padrão}(x_2)$ e $s_3 = \text{desvio padrão}(x_3)$. Podemos ver na Figura B.10 e B.11 (σ^2 conhecido e desconhecido) que os dados (1 e 3) e (1 e 19) aparecem, respectivamente, como possíveis dados atípicos.

Na perturbação da variável resposta foi utilizado o valor $s_y = 0.088$ (desvio de y). Nas Figuras B.12 (σ^2 conhecido) e B.13 (σ^2 desconhecido), respectivamente, a observação 3 e as observações 1 e 19 se destacam das demais como possíveis pontos influentes.

Os pontos influentes detectados são diferentes para os casos com variância conhecida e desconhecida.

Ao perturbar a variância do erro observamos na Figura B.14 que os possíveis pontos influentes são 1, 13 e 19.

Na Tabela 3.11 foram obtidas as estimativas dos parâmetros ao retirar as observações 1, 13 e 19 e as possíveis combinações. Foi considerado a retirada das observações 1 e 3, retirado as observações 1 e 3 e as observações 1 e 19 também que foram detectadas na perturbação de covariável (σ^2 conhecido e desconhecido), em que, foi obtido os maiores valores do C_{max} , houve uma mudança significativa na estimativa dos parâmetros.

Tabela 3.11: Estimativa dos parâmetros dos dados de rato com a mudança relativa entre parênteses.

	completo	sem 1	sem 13	sem 19	sem 1 3	sem 1 13	sem 1 19	sem 13 19	sem 1 13 19
$\hat{\beta}_0$	0.27 (0.00)	0.27 (0.03)	0.41 (0.54)	0.12 (0.56)	0.30 (0.13)	0.39 (0.45)	0.11 (0.58)	0.24 (0.10)	0.18 (0.30)
p_{valor}	0.19	0.15	0.07	0.55	0.14	0.07	0.50	0.28	0.35
$\hat{\beta}_1$	-0.02 (0.00)	-0.02 (0.14)	-0.02 (0.05)	-0.02 (0.10)	-0.01 (0.29)	-0.02 (0.14)	-0.02 (0.05)	-0.02 (0.00)	-0.02 (0.05)
p_{valor}	0.02	0.01	0.01	0.02	0.41	0.01	0.00	0.01	0.00
$\hat{\beta}_2$	0.01 (0.00)	0.02 (0.79)	0.00 (0.86)	0.02 (0.36)	0.02 (0.57)	0.01 (0.00)	0.03 (1.21)	0.01 (0.36)	0.02 (0.79)
p_{valor}	0.42	0.15	0.92	0.25	0.27	0.47	0.04	0.60	0.16
$\hat{\beta}_3$	4.18 (0.00)	4.52 (0.08)	4.35 (0.04)	3.94 (0.06)	2.85 (0.32)	4.61 (0.10)	4.30 (0.03)	4.10 (0.02)	4.37 (0.04)
p_{valor}	0.01	0.01	0.01	0.01	0.44	0.00	0.00	0.01	0.00

() mudança relativa

A observação 1 é de um rato que recebeu a dose 0.88, e entre os que receberam esta dose foi o que teve a maior retenção da dose e o menor peso do fígado. O rato 13 teve a menor retenção da dose no fígado, no entanto, não é o de menor peso e o rato 19 é o que teve a segunda maior retenção da dose, no entanto, é um dos que tem os menores pesos.

3.2.2 Curvatura normal conformal

Consideramos o esquema de ponderação de casos quando σ^2 é conhecido, apresentado na seção 2.2.1. A Figura 3.12 mostra os autovalores normalizados e os valores de q . Para $q = 3$ temos apenas um autovalor acima do valor de q/\sqrt{n} . No entanto para $q = 2$ temos dois autovalores acima do valor q/\sqrt{n} e para $q = 1$ temos três autovalores acima do valor de q/\sqrt{n} . A Tabela 3.12 mostra que as observações que podem exercer influência para todos os valores de q são as observações 1, 13 e 19. Considerando a contribuição agregada dos autovalores ($q = 1$ e 2) vemos claramente que a observação 5 se destaca como influente. Considerando o enfoque de Cook (1986), somente as observações 1, 13 e 19 haviam sido detectadas neste esquema de perturbação.

Tabela 3.12: Medidas de influência utilizando a curvatura normal conformal para o esquema de ponderação de casos dos dados de rato.

	Número de autovetores influentes	Média	Limitante	Valores para os seguintes casos			
				1	5	13	19
$m_j(3)$	1	0.207	0.293	0.389	0.127	0.536	0.474
$m_j(2)$	2	0.259	0.367	0.511	0.411	0.543	0.523
$m_j(1)$	3	0.292	0.413	0.511	0.417	0.543	0.559
B_{E_j}	19	0.089	0.178	0.264	0.174	0.295	0.312

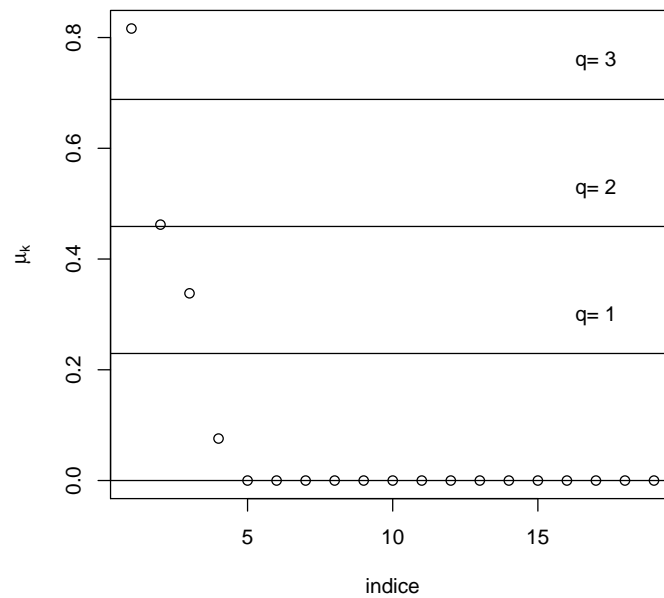


Figura 3.12: Autovalores normalizados dos dados de rato.

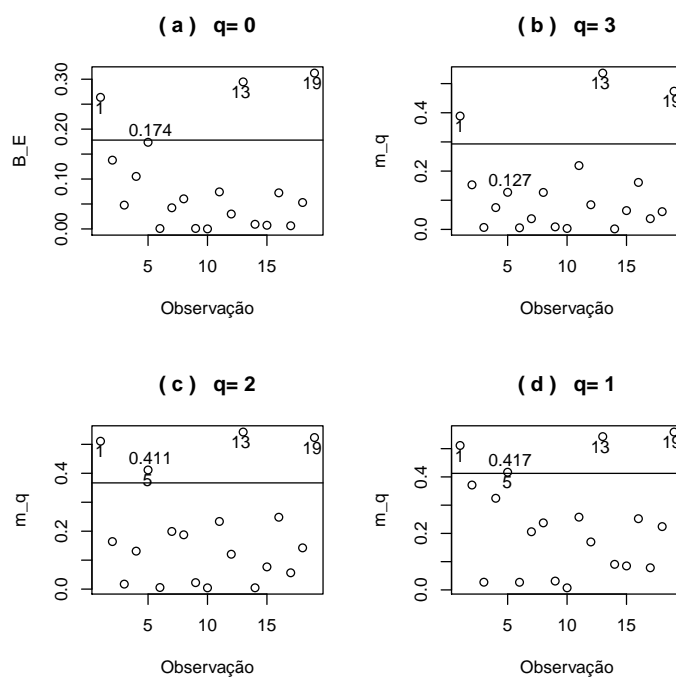


Figura 3.13: Contribuição agregada dos dados de rato.

3.2.3 Procura “Forward”

Considerando o modelo de regressão utilizando os dados de rato aplicamos a metodologia procura “forward” e obtivemos o gráfico 3.14.

Na Figura 3.14 podemos observar que o dado 5 pode ser um dado influente mascarado. Na Figura 3.17 observa-se que tirando a observação 5, a observação 3 fica mascarada.

O rato 5 é o rato que possui o maior peso corporal e o que recebeu a maior dose relativa no conjunto de dados

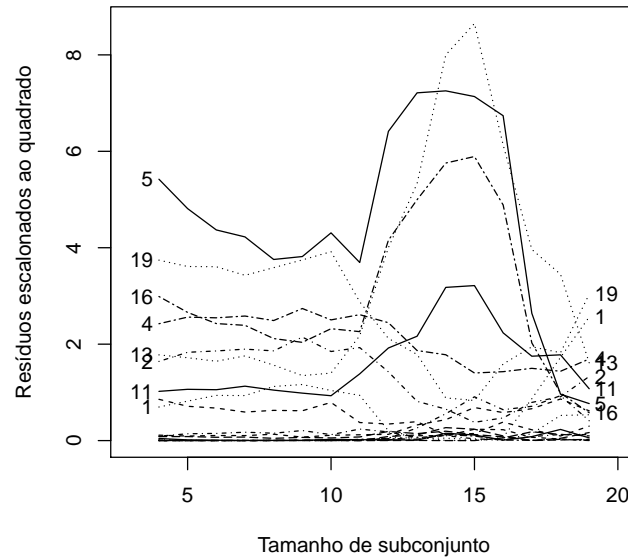


Figura 3.14: Gráfico dos resíduos escalonados ao quadrado segundo a procura “forward” dos dados de rato.

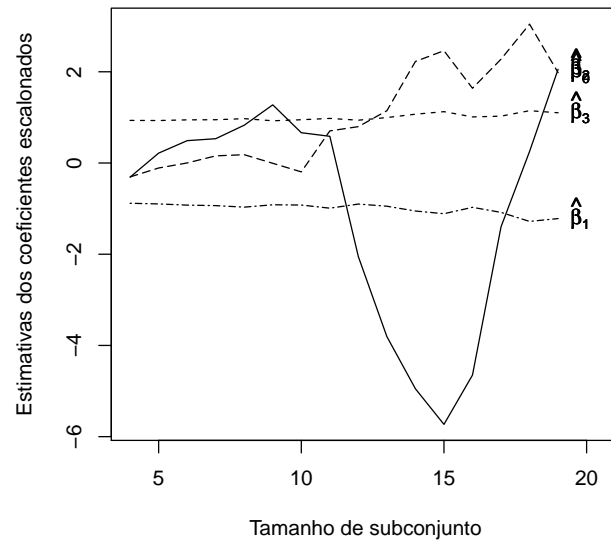


Figura 3.15: Gráfico das estimativas dos coeficientes escalonado 2.33 segundo a procura “forward” dos dados de rato.

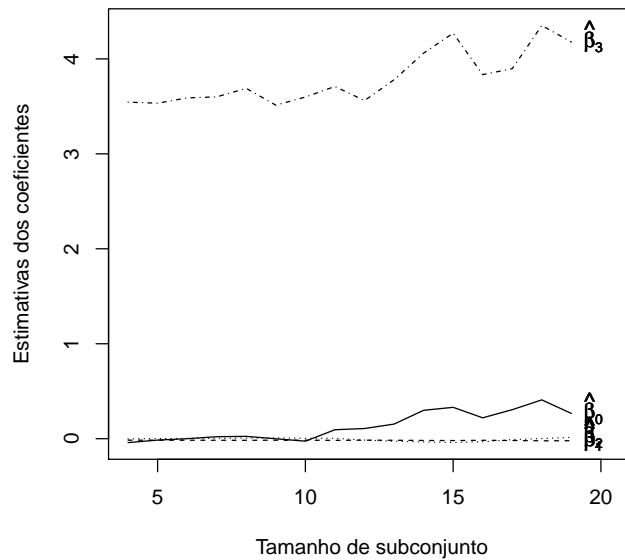


Figura 3.16: Gráfico das estimativas dos parâmetros segundo a procura “forward” dos dados de rato.

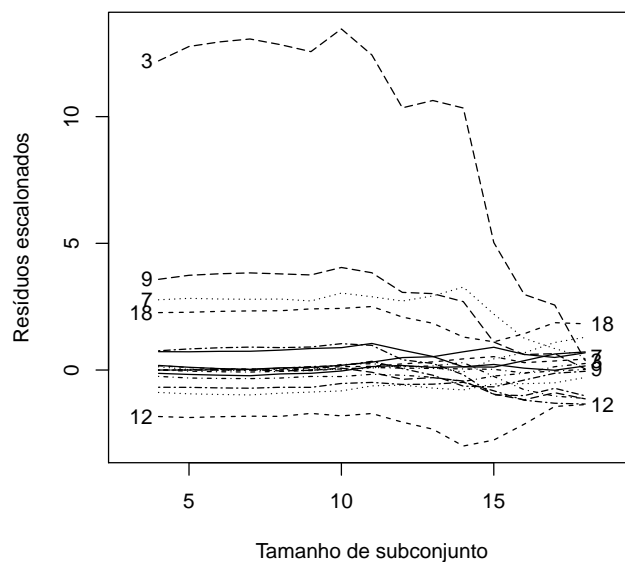


Figura 3.17: Gráfico dos resíduos escalonados após eliminar o dado 5 segundo a procura “forward” dos dados de rato.

Na Tabela 3.13 foram obtidas as estimativas dos parâmetros após a retirada das observações 5 e 3.

Tabela 3.13: Estimativa dos parâmetros dos dados de rato com a mudança relativa entre parênteses.

	completo	sem 3	sem 5	sem 3 5
$\hat{\beta}_0$	0.27 (0.00)	0.31 (0.17)	0.16 (0.38)	0.21 (0.21)
p_{valor}	0.19	0.15	0.45	0.35
$\hat{\beta}_1$	-0.02 (0.00)	-0.01 (0.62)	-0.02 (0.14)	-0.00 (0.95)
p_{valor}	0.02	0.68	0.05	0.94
$\hat{\beta}_2$	0.01 (0.00)	0.01 (0.36)	0.00 (0.64)	-0.00 (1.21)
p_{valor}	0.42	0.64	0.80	0.90
$\hat{\beta}_3$	4.18 (0.00)	1.49 (0.64)	3.76 (0.10)	0.44 (0.89)
p_{valor}	0.01	0.69	0.03	0.91

() mudança relativa

3.3 Dados de Stack Loss

Nesta Secção vamos considerar os dados de Stack Loss tomados de Brownlee (1965, p. 454), Apêndice A.3. Este conjunto de dados contém observações de 21 dias de operações de uma planta para a oxidação de amônia, como uma etapa na produção de ácido nítrico. As variáveis são:

y : perda de pilha (stack loss); 10 vezes a porcentagem de amoníaco entrante que escapa a partir da coluna de absorção.

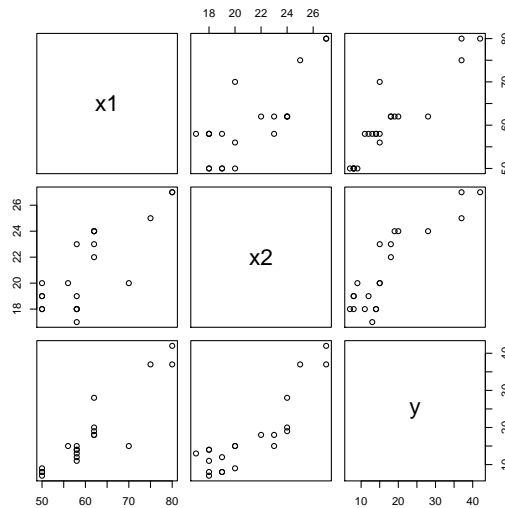
x_1 : fluxo de ar;

x_2 : arrefecimento da temperatura de entrada da água;

Apresentamos as estatísticas descritivas para este conjunto de dados na Tabela 3.14 e o gráfico de dispersão na Figura 3.18.

Tabela 3.14: Estatísticas descritivas dos dados de Stack Loss.

Estatísticas descritivas	x_1	x_2	y
Mínimo	50.00	17.00	7.00
Máximo	80.00	27.00	42.00
Mediana	58.00	20.00	15.00
Média	60.43	21.10	17.52
Desvio padrão da média	2.00	0.69	2.22
Desvio padrão	9.17	3.16	10.17
Coefficiente de variação	0.15	0.15	0.58

**Figura 3.18:** Gráfico da matriz de dispersão dos dados de Stack Lock.

O modelo ajustado a este conjunto de dados segundo em Atkinson e Riani (2000) foi:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{1i} x_{2i} + \varepsilon_i, \quad i = 1, 2, \dots, 21,$$

(de segunda ordem) com a suposição de que $\varepsilon_i \sim N(0, \sigma^2)$, com os ε_i , mutuamente independentes, e as estimativas dos parâmetros obtidas foram $\hat{\beta}_0 = 34.15$, $\hat{\beta}_1 = 1.38$, $\hat{\beta}_2 = -8.47$, $\hat{\beta}_3 = -0.03$ e $\hat{\beta}_4 = 0.16$. Obtivemos os gráficos das medidas de influência global que se encontra na Figura 3.19.

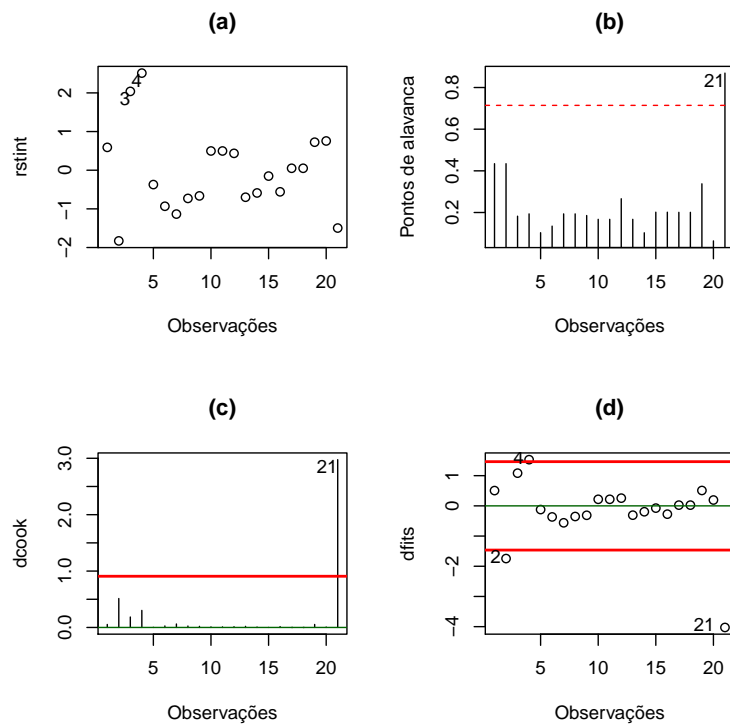


Figura 3.19: Gráficos de medidas de influência global dos dados de Stack Loss (a) Gráfico de resíduos estudentizados (b) Gráficos dos pontos de alavanca (c) Gráfico da distância de Cook (d) Gráfico de DFFITS.

Tabela 3.15: Medida de influência global dos dados de Stack Loss.

	DFBeta intercepto	DFBeta x_1	DFBeta x_2	DFBeta x_1^2	DFBeta $x_1 * x_2$	D FFit	D Cook	Ponto de alavanca
2	-0.86*	0.42	0.41	0.02	-0.42	-1.74*	0.51	0.43
4	-0.67*	0.72*	-0.08	-0.50*	0.17	1.53*	0.30	0.19
21	2.21*	1.47*	-3.13*	-2.85*	3.29*	-4.03*	2.98*	0.87*

* indica que é dado atípico segundo a medida de influência

Analisando a Figura (3.19) observa-se que em (a) aparecem como dados atípicos as observações 3 e 4, segundo os gráficos (b) e (c) temos o dado 21 como dado atípico e só na medida de DFFitS é que aparece as observações 2, 4 e 21 como dados atípicos.

Observamos na Tabela (3.16) que ao retirar as observações 3, 4 e 21 as estimativas dos parâmetros mudam significativamente.

Tabela 3.16: Estimativa dos parâmetros dos dados de Stack Loss com a mudança relativa entre parênteses.

	completo	sem 3	sem 4	sem 19	sem 21	sem 3 4	sem 3 21	sem 4 21	sem 3 4 21
$\hat{\beta}_0$	34.15 (0.00)	37.55 (0.10)	52.01 (0.52)	33.01 (0.03)	-35.72 (2.05)	56.52 (0.66)	14.41 (0.58)	-45.02 (2.32)	-3.10 (1.09)
p_{valor}	0.32	0.22	0.07	0.34	0.52	0.01	0.82	0.22	0.93
$\hat{\beta}_1$	1.38 (0.00)	1.53 (0.11)	0.74 (0.46)	1.81 (0.32)	-0.17 (1.13)	0.88 (0.36)	1.01 (0.27)	-1.60 (2.16)	-0.57 (1.41)
p_{valor}	0.23	0.14	0.42	0.17	0.91	0.21	0.53	0.12	0.57
$\hat{\beta}_2$	-8.47 (0.00)	-9.12 (0.08)	-8.24 (0.03)	-9.66 (0.14)	2.38 (1.28)	-8.92 (0.05)	-5.51 (0.35)	7.40 (1.87)	0.67 (1.08)
p_{valor}	0.03	0.01	0.01	0.03	0.76	0.00	0.55	0.17	0.90
$\hat{\beta}_3$	-0.03 (0.00)	-0.04 (0.12)	-0.03 (0.21)	-0.04 (0.21)	0.01 (1.38)	-0.03 (0.09)	-0.02 (0.35)	0.04 (2.24)	0.01 (1.35)
p_{valor}	0.07	0.03	0.07	0.06	0.71	0.01	0.59	0.09	0.64
$\hat{\beta}_4$	0.16 (0.00)	0.17 (0.07)	0.15 (0.05)	0.18 (0.12)	-0.03 (1.17)	0.16 (0.02)	0.11 (0.32)	-0.12 (1.74)	-0.00 (1.02)
p_{valor}	0.02	0.01	0.01	0.02	0.84	0.00	0.49	0.20	0.98

() mudança relativa

3.3.1 Influência local

Para ver o comportamento da sensibilidade, a pequenas perturbações foram feitos os gráficos do autovetor, \mathbf{l}_{max} , correspondente ao maior autovalor, C_{max} , considerando as perturbações definidas na Secção 2.2.

Os valores do C_{max} para cada tipo de perturbação podem ser encontradas na Tabela 3.17.

Considerando o esquema de ponderação de casos obtivemos as Figuras B.15 e B.16 (σ^2 conhecido e desconhecido) e observa-se, respectivamente, que os dados (2, 3, 4 e 7) e (3 e 4) podem ser considerados como possíveis dados influentes. Neste caso, o dado 21 não aparece como um possível dado influente.

Considerando a ponderação de casos (σ^2 desconhecido)

Para perturbação na covariável foi utilizado $S = \text{diag}(s_0, s_1, s_2, s_3, s_4) = \text{diag}(0, \text{desvio padrão}(x_1), \text{desvio padrão}(x_2), \text{desvio padrão}(x_1^2), \text{desvio padrão}(x_1 * x_2))$. Nas Figuras B.17 (σ^2 conhecido) e B.18 (σ^2 desconhecido) podemos ver, respectivamente, que os pontos (3 e 4) e (2, 3 e 4) são possíveis pontos influentes.

Tabela 3.17: Valores do Cmax para os dados de Stack Loss.

Esquema de perturbação	σ^2	Valor da curvatura máxima
Ponderação de casos	$\hat{\sigma}_C^2$	4.12
	$\hat{\sigma}_{NC}^2$	5.73
Perturbação na covariável	$\hat{\sigma}_C^2$	8242.26
	$\hat{\sigma}_{NC}^2$	11749.33
Perturbação na variável resposta	$\hat{\sigma}_C^2$	26.85
	$\hat{\sigma}_{NC}^2$	71.95
Perturbação na variância do erro	$\hat{\sigma}_{NC}^2$	6.14

$\hat{\sigma}^2_C: \hat{\sigma}^2$ conhecido; $\hat{\sigma}^2_{NC}: \hat{\sigma}^2$ não conhecido

Para perturbarmos a variável resposta foi utilizado o valor $s_y = 10.17$ (desvio padrão de y). Considerando a Figura B.19 (σ^2 conhecido) e a Figura B.20 (σ^2 desconhecido), observamos que os possíveis dados influentes para este tipo de perturbação, são respectivamente, os dados 19 e (3 e 4).

Considerando agora, a perturbação na variância do erro com σ^2 não conhecido as observações 3 e 4 foram detectadas como possíveis pontos influentes.

3.3.2 Curvatura normal conformal

Consideramos o esquema de ponderação de casos (σ^2 conhecido) apresentado na Secção 2.2.1. A Figura 3.20 mostra os autovalores normalizados e os valores de q . Para $q = 3$ temos apenas um autovalor acima do valor de q/\sqrt{n} . No entanto para $q = 1$ e 2 temos dois autovalores acima do valor q/\sqrt{n} .

A Tabela 3.19 mostra as observações que exercem influência são o 2, 3 e 4. Retirado as observações 2, 3 e 4 há mudanças significativas nas estimativas dos parâmetros.

Tabela 3.18: Estimativa dos parâmetros dos dados de Stack Loss com a mudança relativa entre parênteses.

	completo	sem 2	sem 3	sem 4	sem 6	sem 7
$\hat{\beta}_0$	34.15 (0.00)	73.25 (1.15)			64.48 (0.89)	
p_{valor}	0.32	0.00			0.01	
$\hat{\beta}_1$	1.38 (0.00)	0.57 (0.58)			0.87 (0.37)	
p_{valor}	0.23	0.36			0.22	
$\hat{\beta}_2$	-8.47 (0.00)	-9.72 (0.15)			-9.80 (0.16)	
p_{valor}	0.03	0.00			0.00	
$\hat{\beta}_3$	-0.03 (0.00)	-0.03 (0.12)			-0.03 (0.00)	
p_{valor}	0.07	0.01			0.01	
$\hat{\beta}_4$	0.16 (0.00)	0.18 (0.11)			0.18 (0.13)	
p_{valor}	0.02	0.00			0.00	

() mudança relativa

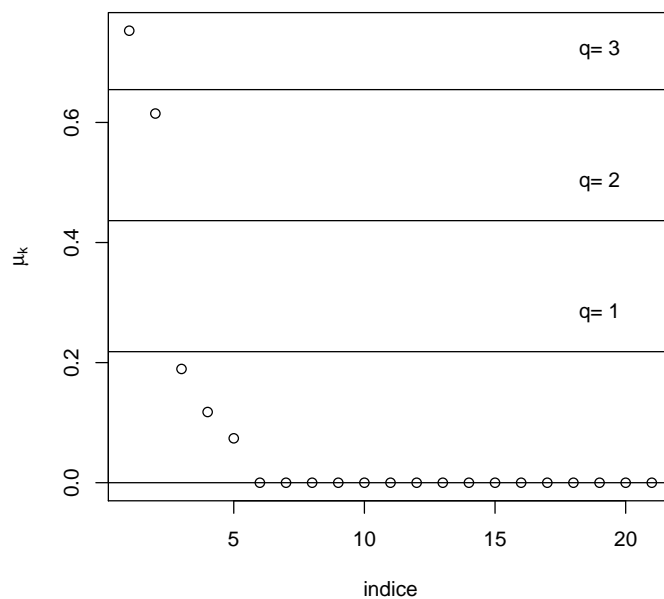


Figura 3.20: Autovalores normalizados dos dados de Stack Loss.

Tabela 3.19: Medidas de influência utilizando a curvatura normal conformal para o esquema de ponderação de casos dos dados de Stack Loss.

	Número de autovetores influentes	Média	Limitante	Valores para os seguintes casos		
				2	3	4
$m_j(3)$	1	0.189	0.268	0.329	0.385	0.577
$m_j(2)$	2	0.255	0.361	0.605	0.514	0.687
B_{E_j}	21	0.083	0.167	0.394	0.298	0.472

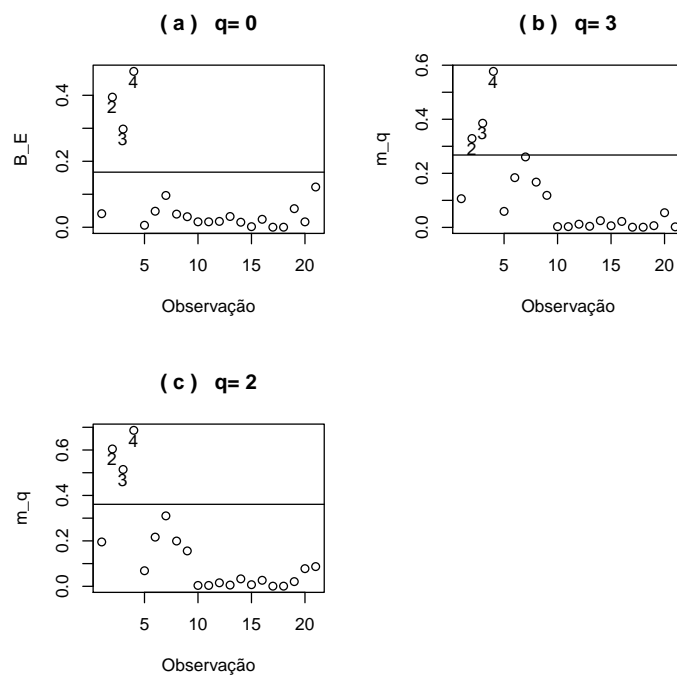


Figura 3.21: Contribuição agregada dos dados de Stack Loss.

3.3.3 Procura “forward”

Considerando a metodologia procura “forward” e os dados de “Stack Loss” obtivemos o gráfico 3.22 dos resíduos escalonados. Neste gráfico as observações 1, 2, 3 e 4 inicialmente têm grandes resíduos, mas revela que há mascaramento da observação 1 e possivelmente da observação 2. A observação 1 não aparecem em nenhuma das análise.

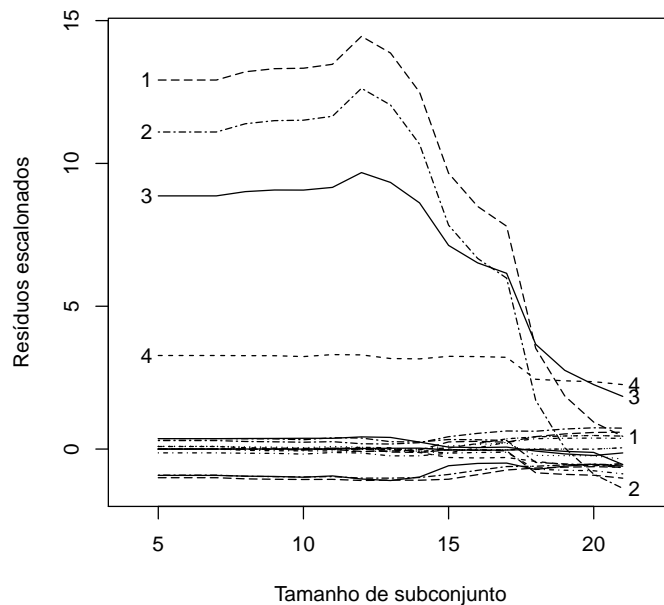


Figura 3.22: Gráfico dos resíduos escalonados segundo a procura “forward” no modelo de segunda ordem dos dados de Stack Loss.

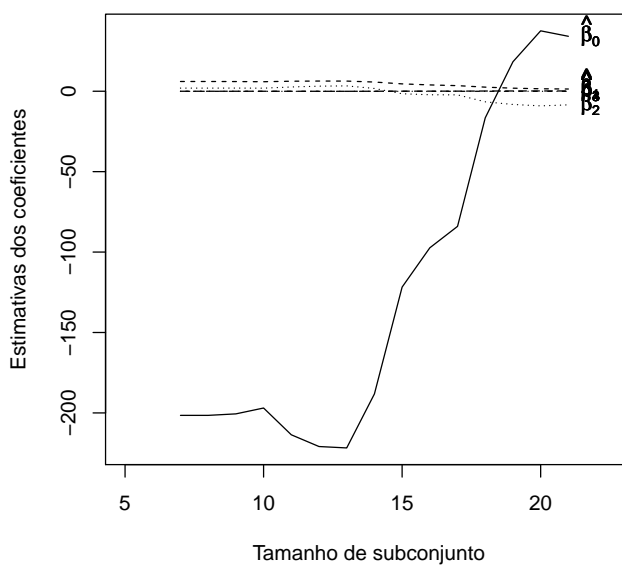


Figura 3.23: Gráfico das estimativas dos parâmetros segundo a procura “forward” no modelo de segunda ordem dos dados de Stack Loss.

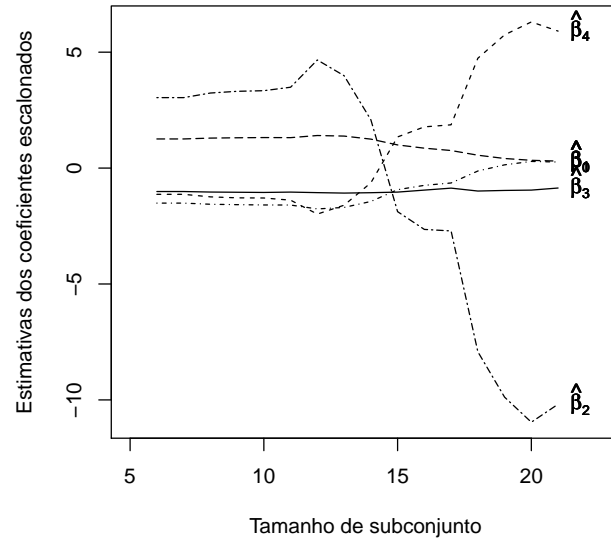


Figura 3.24: Gráfico das estimativas dos parâmetros escalonados segundo a procura “forward” no modelo de segunda ordem dos dados de Stack Loss.

Observando a Figura 3.23, nota-se que as estimativas do intercepto são bastante alterados ao longo do processo.

Retirando a observação 1 obtivemos os seguintes gráficos

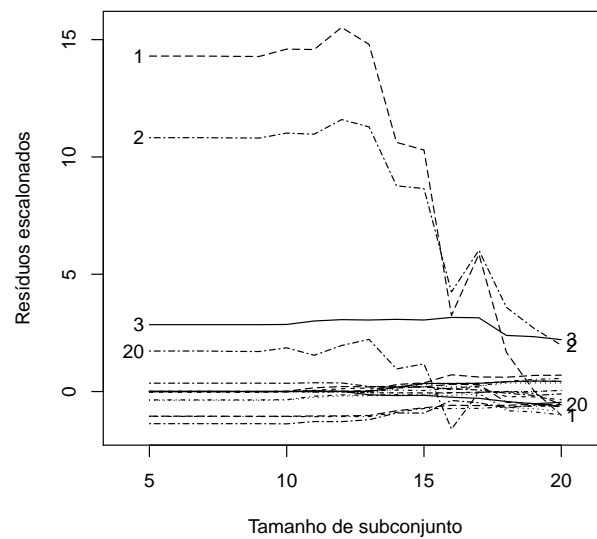


Figura 3.25: Gráfico resíduos escalonados segundo a procura “forward” dos dados de Stack Loss.

Observamos na Figura 3.25 que a observação 2 (que passa a ser 1 após a retirada da observação 1) passa a ser um ponto mascarado.

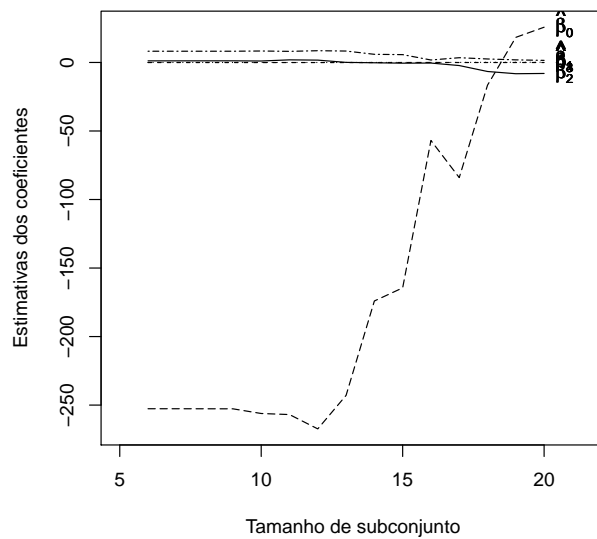


Figura 3.26: Gráfico das estimativas dos coeficientes segundo a procura forward dos dados de Stack Loss.

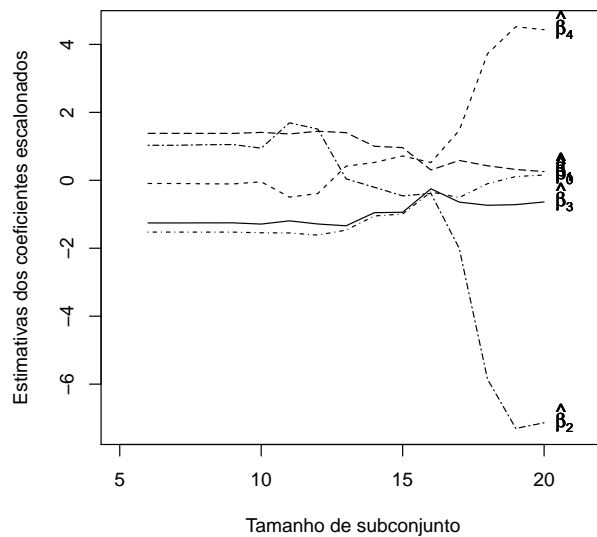


Figura 3.27: Gráfico das estimativas dos coeficientes escalonados 2.33 segundo a procura “forward” dos dados de Stack Loss.

Tabela 3.20: Estimativas dos parâmetros dos dados de Stack Loss tirando algumas observações

	Completo	sem 1	sem 1 e 2
(Intercepto)	34.15	25.72	110.22*
x_1	1.38	1.51	0.15
x_2	-8.47*	-8.03	-12.44**
x_1^2	-0.03	-0.03	-0.04*
$x_1 * x_2$	0.16*	0.15*	0.23**
R^2	0.94	0.92	0.92
Num. obs.	21	20	19

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ com (desvio padrão)

Na Tabela 3.20 foram retiradas as observações 1 e 2 e obtidas as estimativas dos parâmetros. Podemos ver que há mudanças significativas ao retirar estas observações.

Entretanto, a observação 1 não havia sido detectada em nenhum das análises anteriores.

Influência local com procura “Forward”

Na comparação de metodologias, neste capítulo vamos propor o uso da metodologia procura “forward” conjuntamente com o método de influência local e aplicar em modelos de regressão linear.

Na metodologia de influência global a verificação das mudanças decorrentes nas estimativas dos parâmetros de regressão ao retirar uma observação (feito para cada caso) ocasiona um alto custo computacional na detecção de dados atípicos e/ou influentes. Além disso, podem ainda haver dados que não foram detectadas pelos métodos de influência global devido à presença de algum outro dado (s) (problema de mascaramento), ou seja, a existência de dados atípicos mascarados que não foram detectadas pela medida de influência global. Para isso, Atkinson e Riani (2000) propuseram a metodologia de procura “forward”. Como pode ser visto no Capítulo anterior, por exemplo nos dados do rato, usando a metodologia de influencia global foi detectado a observação 3 como sendo um dado atípico. No entanto, ao considerarmos a metodologia de procura “forward”, observamos que existia um dado atípico influente mascarado que é a observação 5.

A metodologia de influência local de Cook (1986) é uma medida de influência local que introduz pequenas perturbações nos dados ou no modelo para avaliar se o modelo é sensível a essas perturbações. Diferentemente da metodologia de influência global, esta metodologia permite avaliar a influência conjunta de todas as observações. No entanto,

neste caso também pode haver dados influentes mascarados que não são detectados com o uso da influência local. Desta forma, a nossa proposta é utilizar a influência local de Cook (1986) e propor uma nova metodologia do tipo procura “forward”.

Nesta nova metodologia, no primeiro passo começamos ajustando modelos de regressão linear aos pequenos subconjuntos formados por $m = m_0$ elementos. Obtemos as estimativas de máxima verossimilhança dos parâmetros em cada um dos subconjuntos.

Considerando agora o conjunto de dados completo (com as n observações), obtemos os autovetores associado ao maior auto valor (segundo o enfoque de Cook (1986)). Depois, ordenamos estes auto vetores e finalmente, obtemos as medianas destes autovetores (autovetores com n elementos) associados ao maior auto valor para cada um dos subconjuntos de dados (subconjuntos com m observações). Ordenamos e tomamos o subconjunto referente ao menor valor mediano do autovetor correspondente ao maior auto valor.

No segundo passo, subconjuntos de tamanho $m + 1$ são considerados de tal forma que uma ou mais observações podem entrar ou sair do conjunto anterior.

Da mesma forma como no passo anterior, são obtidas as estimativas dos parâmetros para cada subconjunto de tamanho $m + 1$, depois considerando o conjunto de dados todo, são obtidos os autovetores associados ao maior auto valor e finalmente estes autovetores são ordenados. As medianas de cada um desses autovetores são obtidas e é escolhido o subconjunto que tiver o menor valor mediano.

No terceiro passo é obtido o subconjunto de tamanho $m + 2$ da mesma forma como foram obtidos os subconjuntos de tamanhos m e $m + 1$. Assim sucessivamente até chegarmos ao conjunto de dados completo com as n observações.

Denotando por \mathbf{l}_m^* o autovetor selecionado a cada passo da iteração de influência local de Cook (1986) com procura “forward” (*ILCPF*) e \mathbf{l}_{ILCPF} a coleção desses autovetores associados ao maior autovalor produzidos a cada passo do processo, temos:

$$\mathbf{l}_{ILCPF} = (\mathbf{l}_{m_0}^*, \dots, \mathbf{l}_n^*). \quad (4.1)$$

Se houver algum pico no gráfico de \mathbf{l}_{ILCPF} durante o desenvolvimento do processo de influência local com a procura “forward” é indício de que foi incluso uma possível observação influente.

De forma resumida, a ideia é considerar a metodologia de procura “forward” conforme Atkinson e Riani (2000) e utilizar como critério de escolha do grupo limpo a metodologia de influência local de Cook (1986).

Desta forma, seja:

$$LD_i(\omega) = 2\{\ell_i(\hat{\theta}) - \ell_i(\hat{\theta}_\omega)\} \quad i = 1, \dots, n, \quad (4.2)$$

em que, $\ell(\theta) = \sum_{i=1}^n \log L_i(\theta) = \sum_{i=1}^n \ell_i(\theta)$; $\hat{\theta}$ e $\hat{\theta}_\omega$ os estimadores de máxima verossimilhança de θ em $L(\theta)$ e $L(\theta|\omega)$ respectivamente. $L(\theta)$ e $L(\theta|\omega)$ como definido na Secção 2.3

A metodologia consiste em obter o subconjunto de tamanho m livre de pontos discrepantes ou influentes, estimar os parâmetros e obter o $LD_i(\omega)$ definido em (4.2) considerando todos os indivíduos. Ordenar, e tomar o valor mediano e considerar como o subconjunto seguinte de tamanho $(m + 1)$ aquele que tiver o menor valor mediano.

Desta forma, vamos considerar um subconjunto de observações que sejam resistentes a pequenas perturbações no conjunto de dados ou no modelo. No entanto, observe que ao considerar a direção \mathbf{l}_{max} correspondente ao C_{max} , por exemplo, se o i -ésimo elemento de \mathbf{l}_{max} for relativamente grande é porque a perturbação no peso ω_i do i -ésimo caso pode levar a mudanças substanciais no resultado da análise, ou seja, vamos ordenar então o vetor \mathbf{l}_{max} e considerar o subconjunto que obtiver o menor valor mediano considerando o conjunto todo. Vamos utilizar esta metodologia em modelos de regressão linear. Utilizando os resultados obtidos na Secção 2.2 (influência local e as perturbações) aplicamos esta metodologia aos conjuntos de dados descritos na Secção 3.

Aplicações da metodologia de influência local com procura “forward”

No processo da metodologia de influência local com procura “forward”, se $\binom{n}{m}$ (a combinatória) for muito grande, foi considerado subconjuntos aleatórios de 5000 amostras (a semente no programa R foi a mesma, ou seja, 5000) e ajustado o modelo de regressão robusta aos grupos conforme descrito no capítulo anterior.

Nos gráficos conforme o número de amostras sorteadas aumentavam, as curvas resultantes dos autovetores ficavam mais suaves.

5.1 Dados de ganso

Considerando os dados de ganso foi aplicado a metodologia de influência local com procura “forward” descrita no Capítulo 4, para cada tipo de perturbação descrita na Secção 2.2.

5.1.1 Ponderação de casos (σ^2 conhecido)

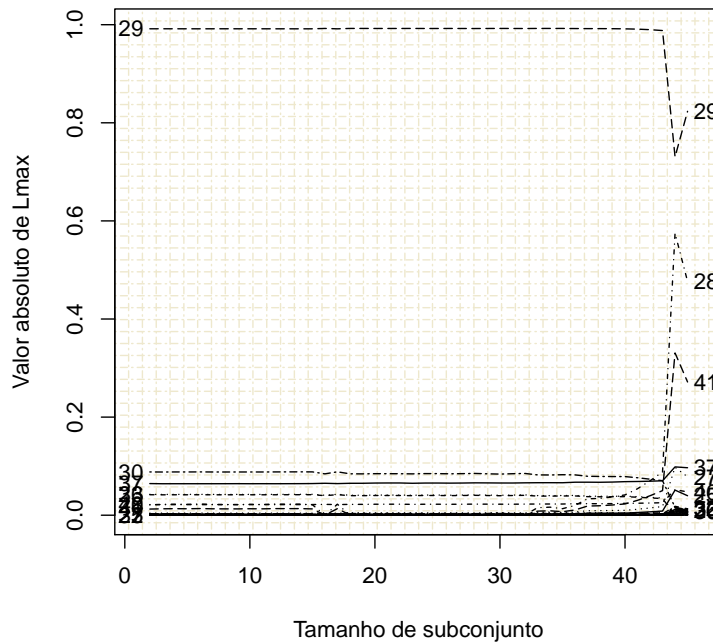


Figura 5.1: Gráfico do valor absoluto do autovetor, $l_{max} = L_{max}$, associado ao maior autovalor C_{max} dos dados de ganso.

Na Figura 5.1 foi construído o gráfico do valor absoluto do autovetor correspondente ao maior autovalor com os tamanhos dos subconjuntos (m). Podemos ver que a observação 29 se destaca do início ao final do processo de procura “forward”. Neste caso, não houve dados mascarados.

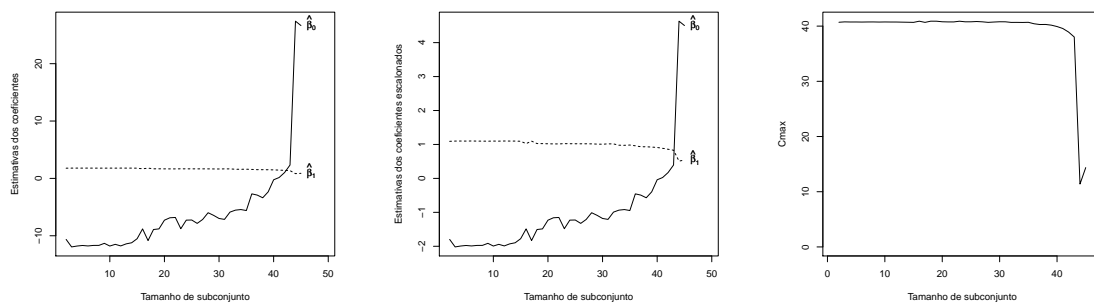


Figura 5.2: Estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de ganso.

Na Figura 5.2, vemos que quando são inseridas as observações 29 e 30 no antepenúltimo passo, há um pico nos gráficos.

5.1.2 Ponderação de casos (σ^2 desconhecido)

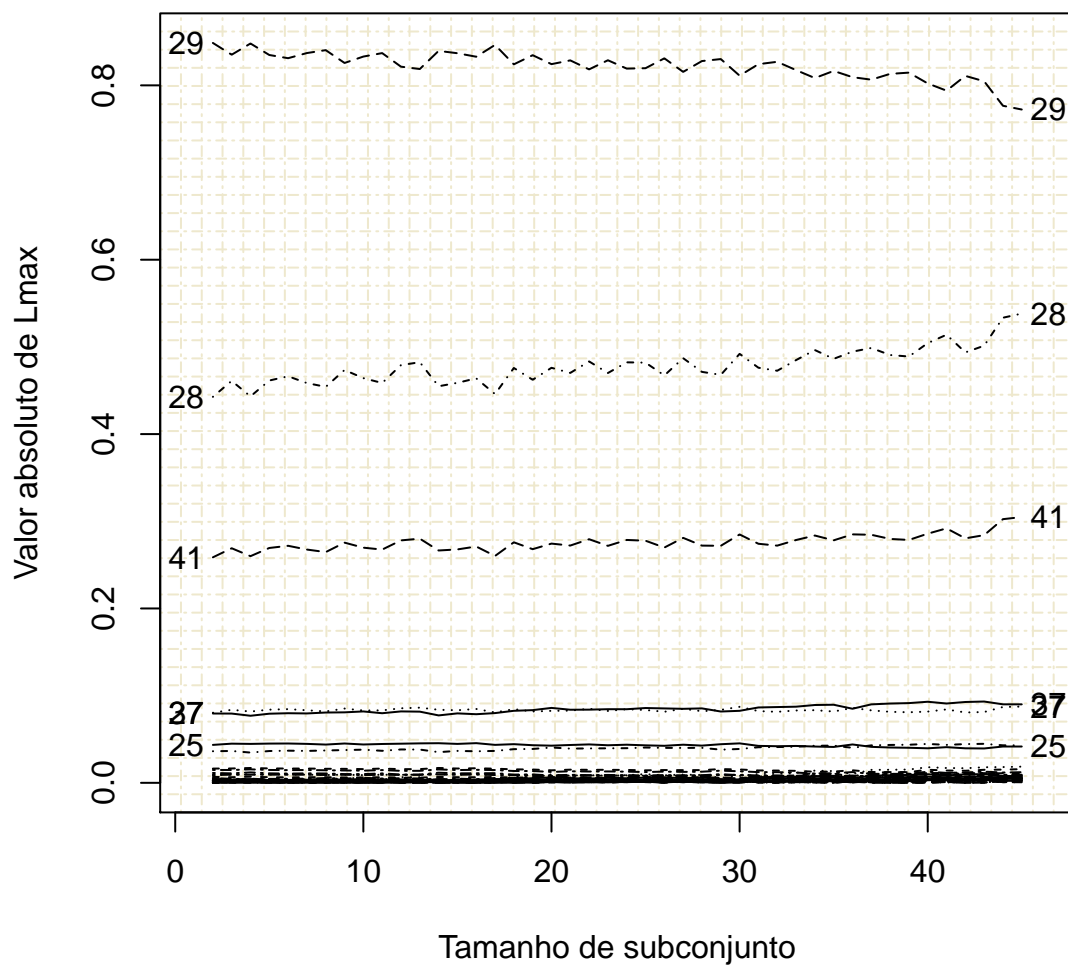


Figura 5.3: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de ganso.

Neste caso também, não há dado mascarado influente.

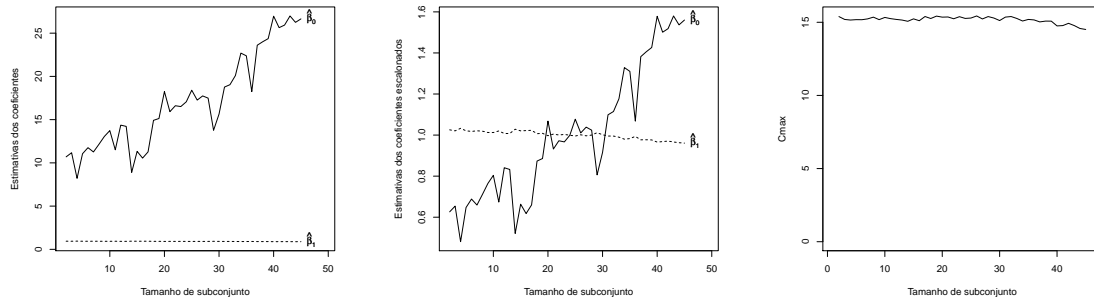


Figura 5.4: Estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de gancho.

5.1.3 Perturbação na covariável (σ^2 desconhecido)

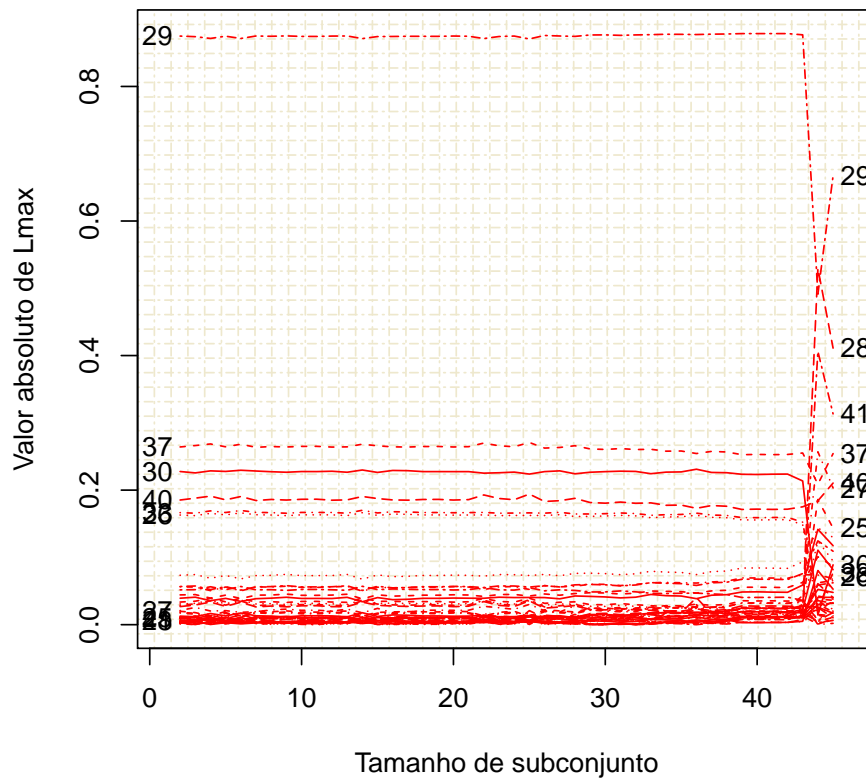


Figura 5.5: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de gancho.

Na Figura 5.5 a observação 29 se destaca do início ao final do processo. Há um grupo de observações que se destaca no processo (37, 30, 40, 26 e 33). Tirando estes

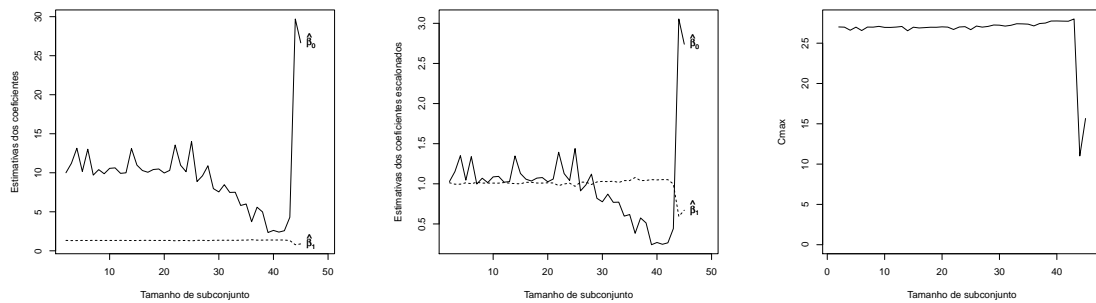


Figura 5.6: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de ganso.

observações obtivemos as estimativas dos parâmetros na Tabela 5.1, quando é inserida as observações 33 e 29 há um pico nos gráficos da Figura 5.6

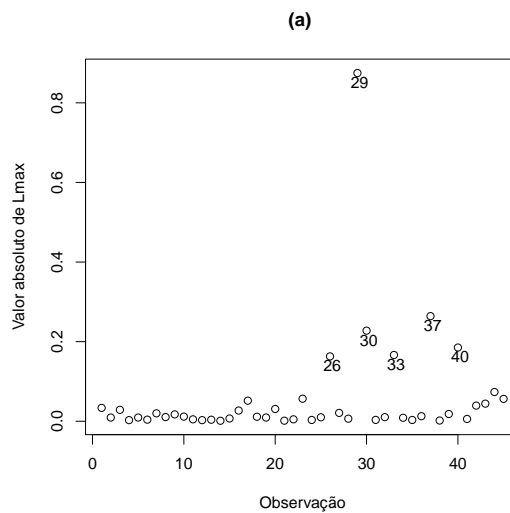


Figura 5.7: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor dos dados de ganso no passo $m = 2$ do processo de influência local com procura “forward”.

Tabela 5.1: Estimativa dos parâmetros dos dados de ganso com a mudança relativa entre parênteses.

	completo	sem (26,30,33,37,40)
$\hat{\beta}_0$	26.65 (0.00)	28.12 (0.06)
p_{valor}	0.00	0.00
$\hat{\beta}_1$	0.88 (0.00)	0.92 (0.04)
p_{valor}	0.00	0.00

() mudança relativa

5.1.4 Perturbação na resposta (σ^2 desconhecido)

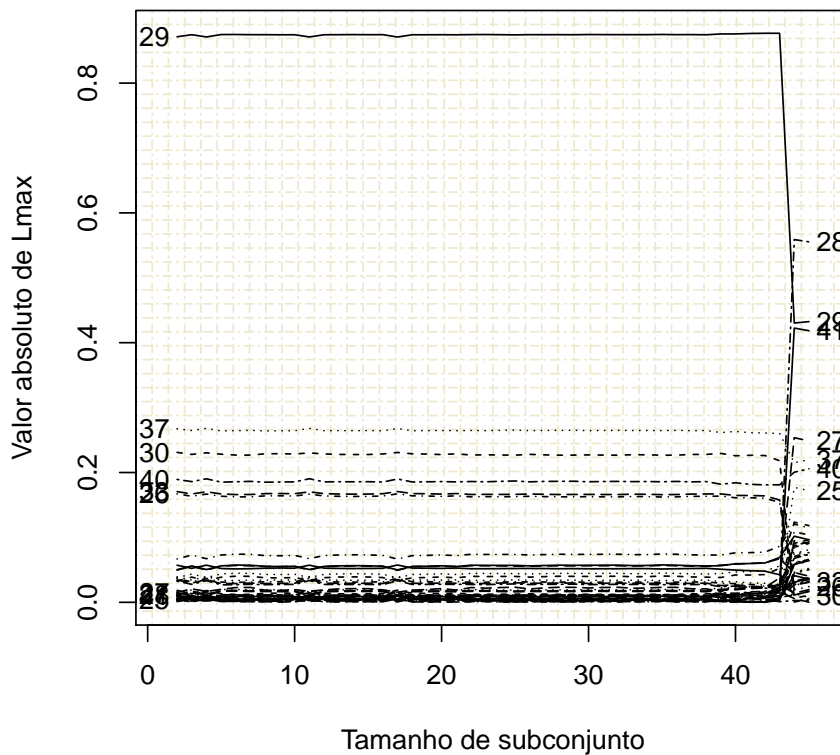


Figura 5.8: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de ganso.

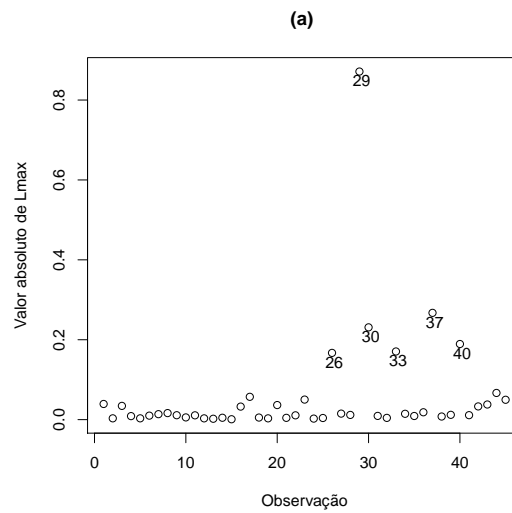


Figura 5.9: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor dos dados de ganso para $m = 2$ da primeira rodada.

Na Figura 5.9 no início da procura “forward”, para $m = 2$, a observação 29 se destaca como um dado atípico e continua destacado até o final do processo. Observamos na Figura 5.8 que o grupo de observações 26, 30, 33, 37 e 40 aparece destacado, assim como havia aparecido considerando a perturbação na covariável com a procura “forward”. Aqui também verificamos um pico nos gráficos da Figura 5.10, quando são inseridas as observações 29 e 30.

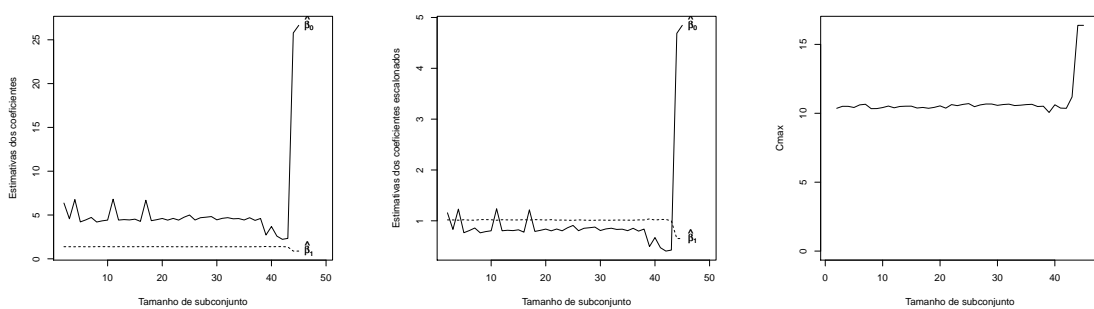


Figura 5.10: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de ganso.

5.1.5 Perturbação na variância (σ^2 desconhecido)

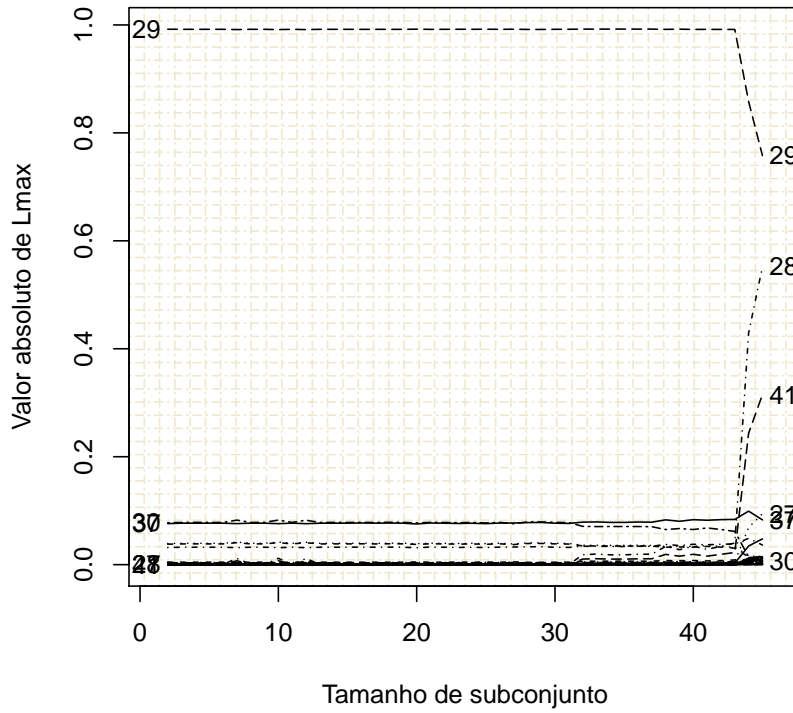


Figura 5.11: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de ganso.

Observando a Figura 5.11 conclui-se que não há dados mascarados neste caso.

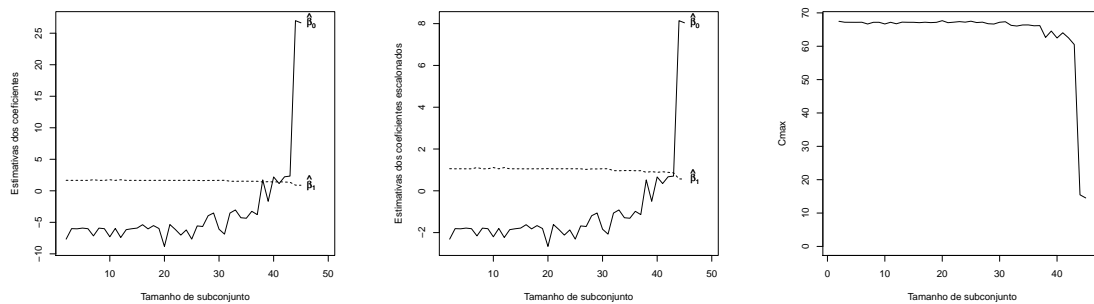


Figura 5.12: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de ganso.

A inclusão das observações 29 e 30 causou picos nos gráficos da Figura 5.12.

5.2 Dados de rato

5.2.1 Ponderação de casos (σ^2 conhecido)

Considerando os dados de rato foi aplicado a metodologia de influência local com procura “forward” descrita no Capítulo 4, para cada tipo de perturbação descrita na seção 2.2.

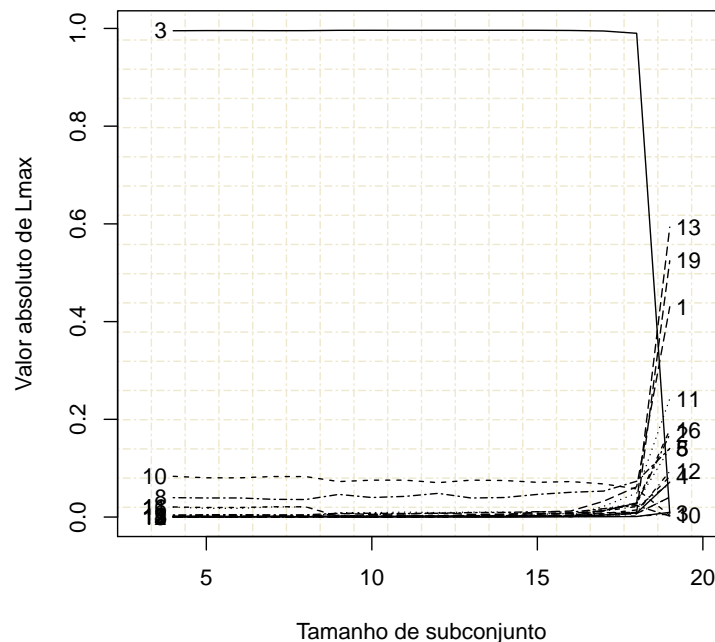


Figura 5.13: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de rato.

Na Figura 5.13, podemos ver claramente que a observação 3 aparece como um dado mascarado.

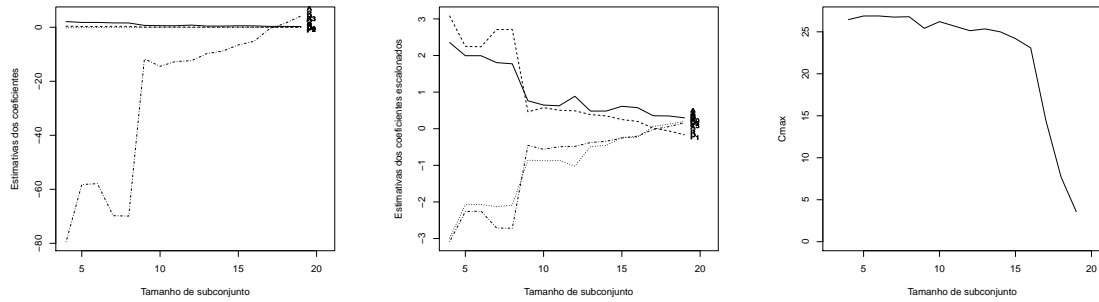


Figura 5.14: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de rato.

5.2.2 Ponderação de casos (σ^2 desconhecido)

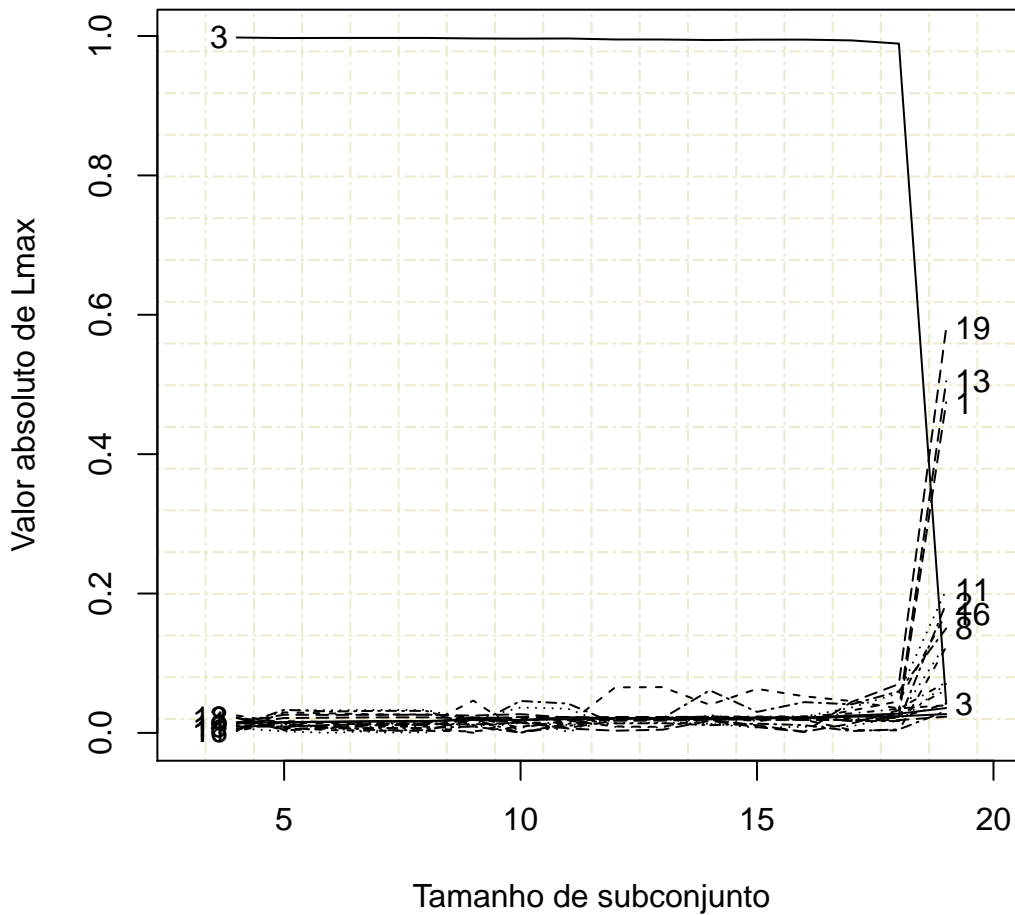


Figura 5.15: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de rato.

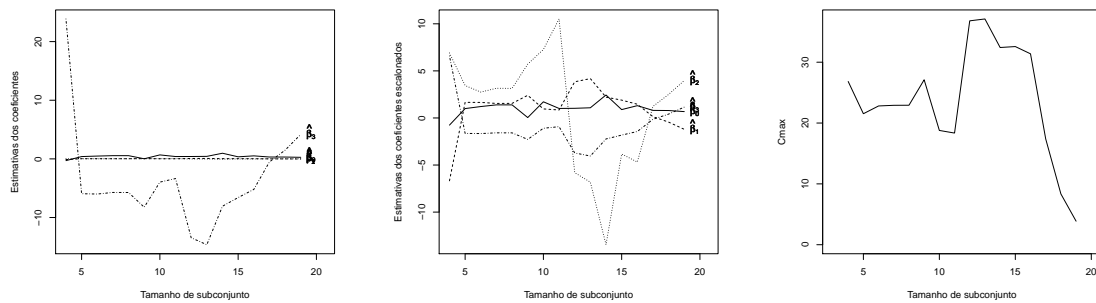


Figura 5.16: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de rato.

Neste caso também, a observação 3 aparece claramente como um dado influente mascarado.

5.2.3 Perturbação na covariável (σ^2 desconhecido)

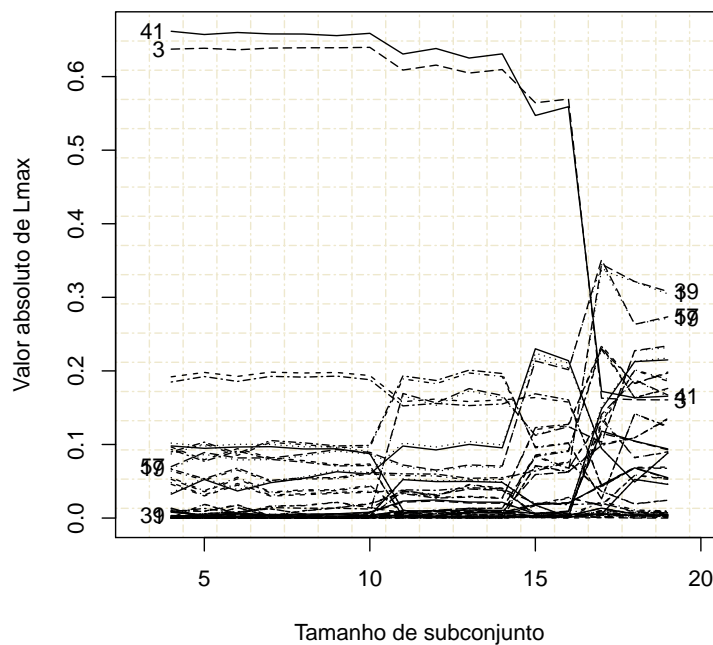


Figura 5.17: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de rato, com as observações. Os valores representados de 1 a 19, 20 a 38 e 39 a 57, no vetor l_{max} , correspondem às covariáveis x_1 , x_2 e x_3 no processo de influência local com procura “forward”.

Na Figura 5.17, observa-se que há uma observação mascarada que é a 3 (representados por 41 e 3).

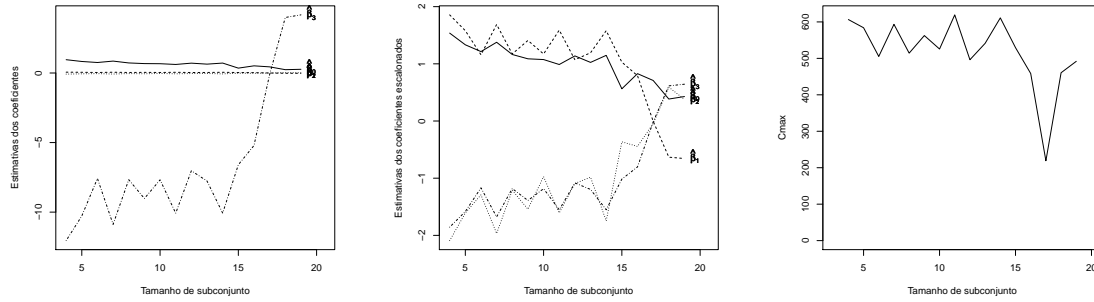


Figura 5.18: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de rato.

5.2.4 Perturbação na resposta (σ^2 desconhecido)

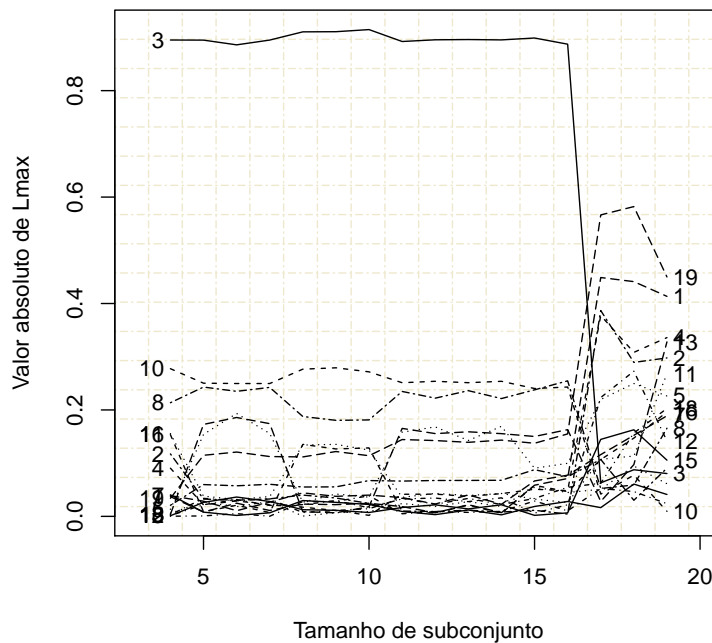


Figura 5.19: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de rato.

Na Figura 5.20 no início do processo da procura “forward” para $m = 2$, aparece a observação 3 como dado atípico e na Figura 5.19 podemos ver que esta observação é mascarada.

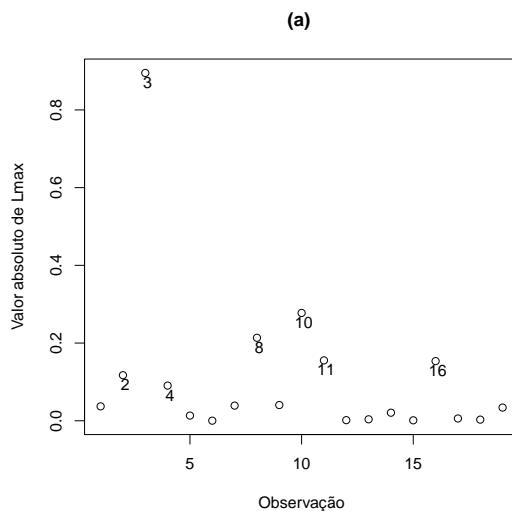


Figura 5.20: Gráfico do valor absoluto do autovetor l_{max} associado ao maior autovalor para $m = 2$ no processo de procura “forward” dos dados de rato.

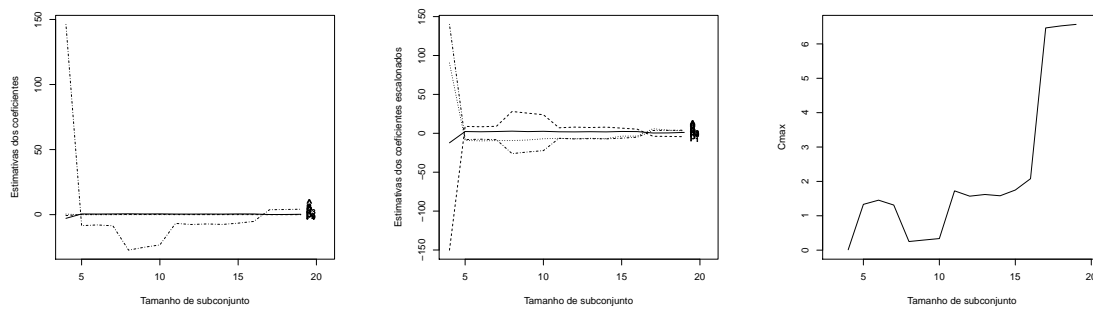


Figura 5.21: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de rato.

5.2.5 Perturbação na variância (σ^2 desconhecido)

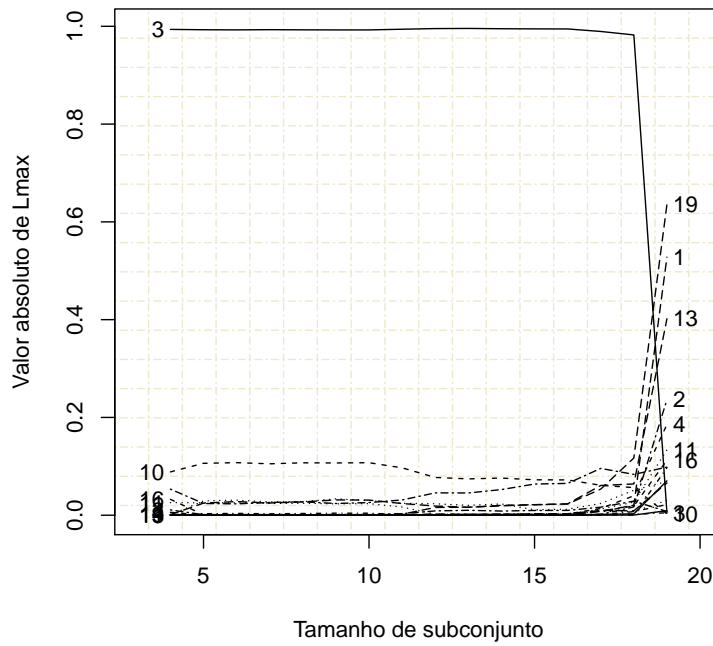


Figura 5.22: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de rato.

Na Figura 5.22, claramente a observação 3 é mascarada.

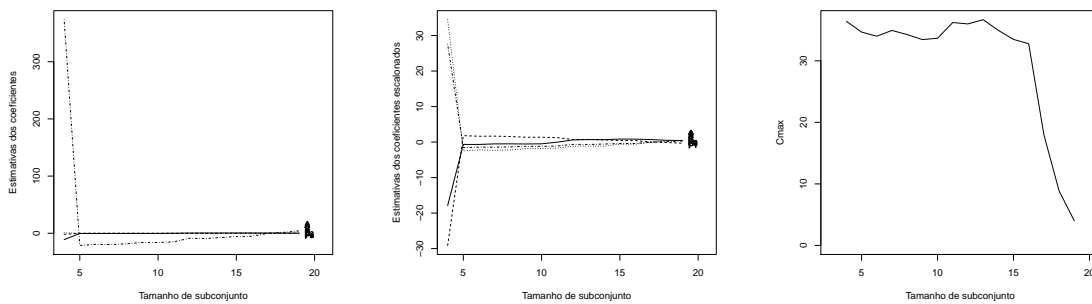


Figura 5.23: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de rato.

5.3 Dados de Stack Loss

Considerando os dados de Stack Loss foi aplicado a metodologia de influência local com procura “forward” descrita na seção 4, para os tipos de perturbações descritas na seção 2.2.

5.3.1 Ponderação de casos (σ^2 conhecido)

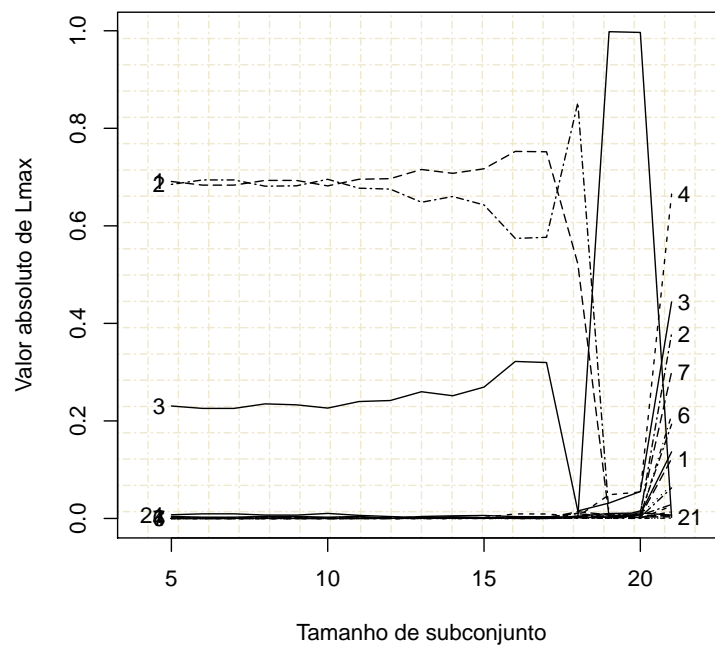


Figura 5.24: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss.

Neste caso, pela Figura 5.24, podemos ver que a observação 1 é mascarada.

5.3.2 Ponderação de casos (σ^2 desconhecido)

Neste caso podemos observar que a observação 21 pode ser um dado mascarado.

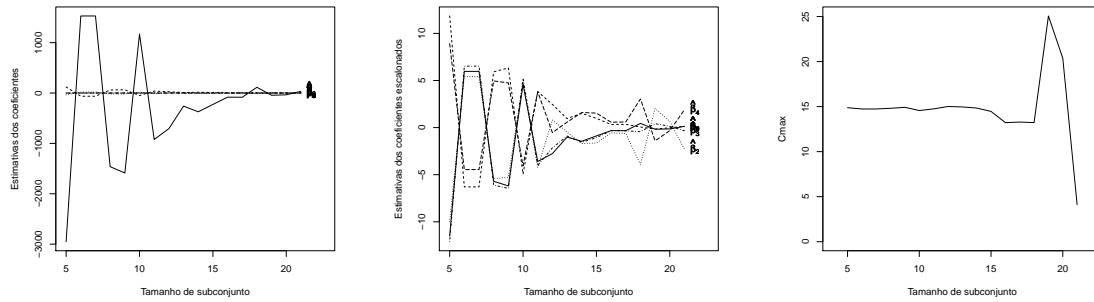


Figura 5.25: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.

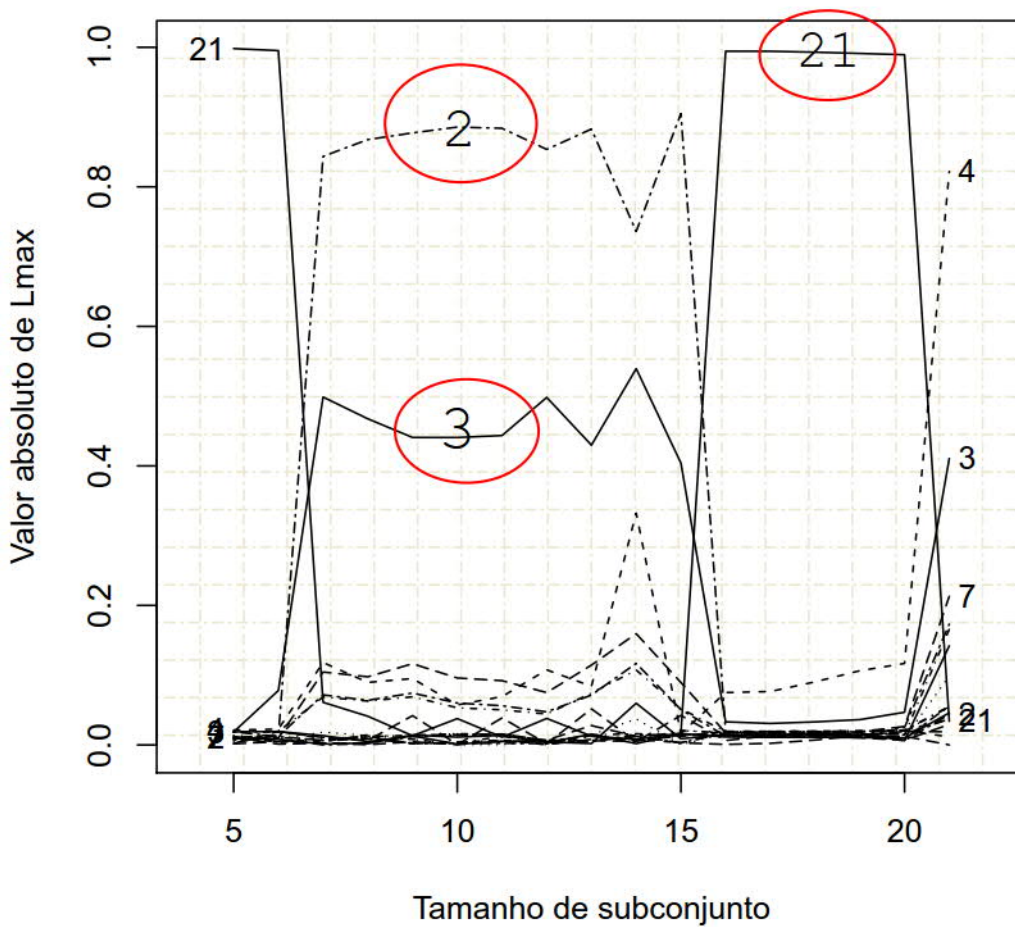


Figura 5.26: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss.

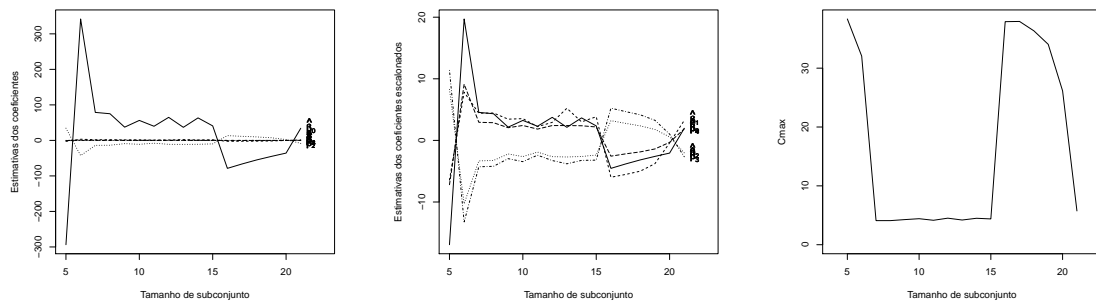


Figura 5.27: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.

5.3.3 Perturbação na covariável (σ^2 desconhecido)

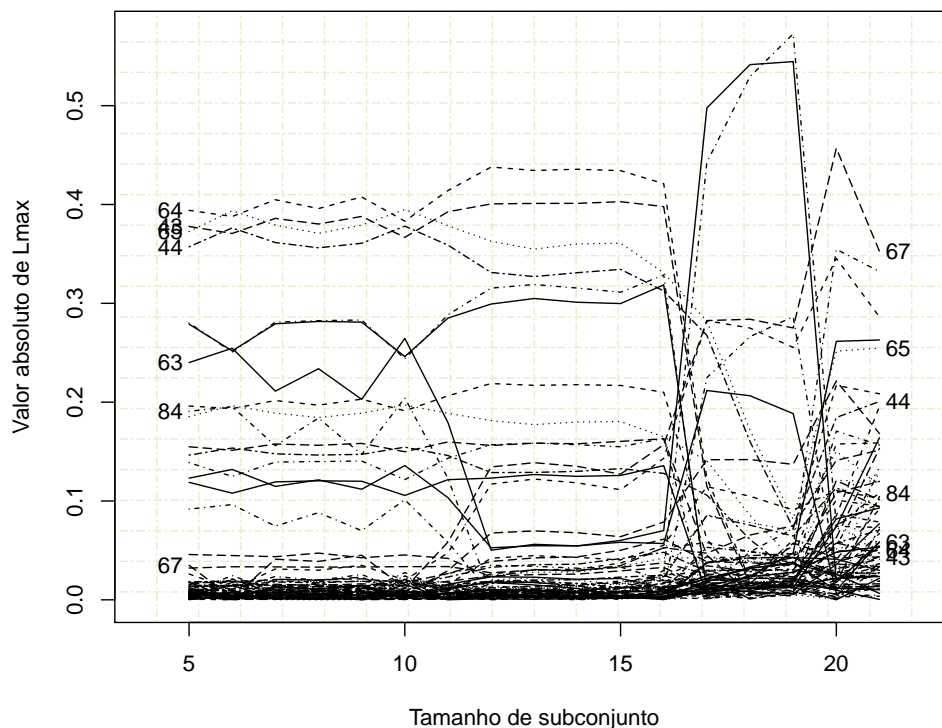


Figura 5.28: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss, com as observações. Os valores representados de 1 a 21, 22 a 42, 43 a 63 e 64 a 84, no vetor l_{max} , correspondem às covariáveis x_1 , x_2 , x_1^2 e x_1x_2 no processo de influência local com procura “forward”.

Observa-se que as observações 1 e 2 são mascaradas representadas pelas observações ((43 e 64) e 44, respectivamente) na Figura 5.28.

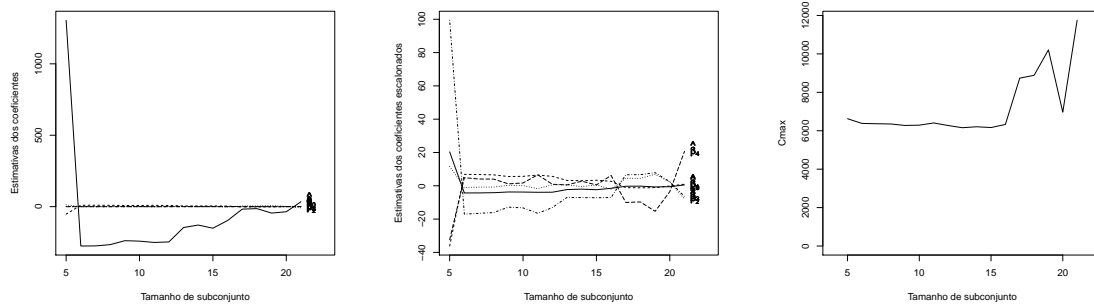


Figura 5.29: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.

5.3.4 Perturbação na resposta (σ^2 desconhecido)

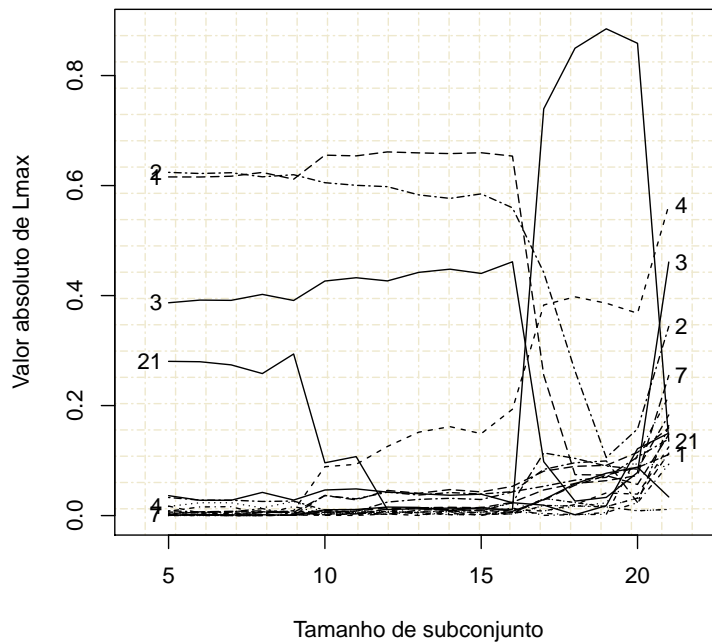


Figura 5.30: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss.

Na Figura 5.30 temos a observação 1 como um dado mascarado.

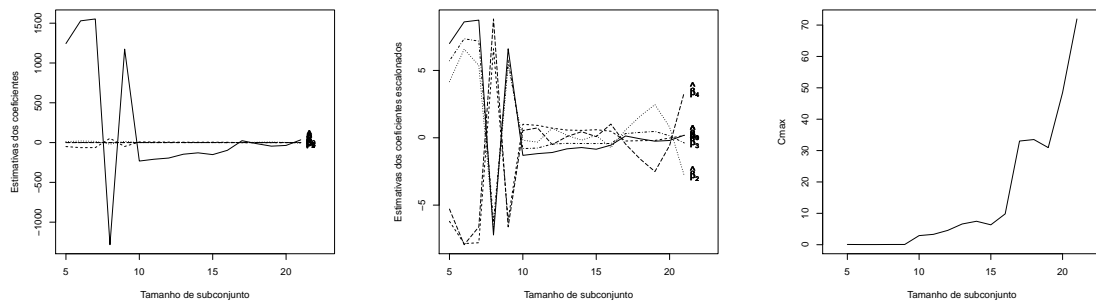


Figura 5.31: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.

5.3.5 Perturbação na variância (σ^2 desconhecido)

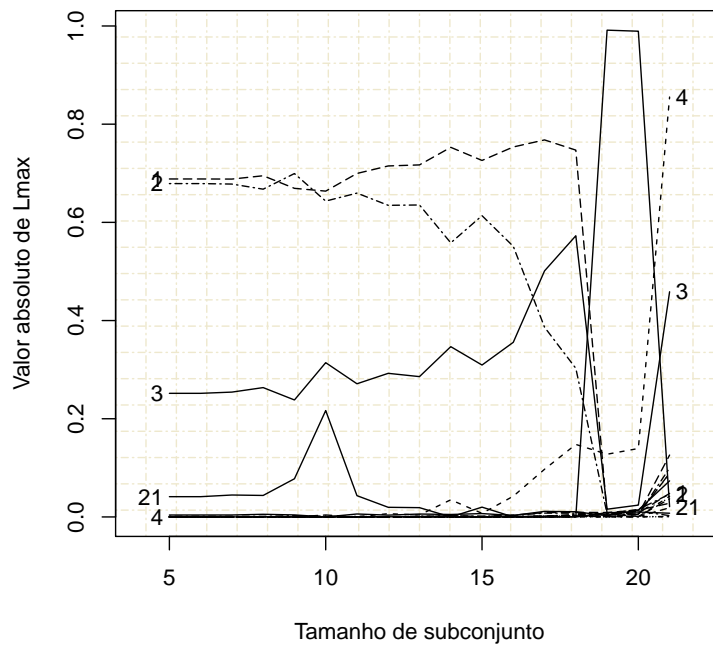


Figura 5.32: Gráfico do valor absoluto do autovetor $l_{max} = L_{max}$ associado ao maior autovalor C_{max} dos dados de Stack Loss.

Na Figura 5.32 observamos que os dados 1 e 2 são mascarados.

Pela Tabela 5.2, retirando-se a observação 1, há uma mudança na estimativa do β_0 e retirando a observação 21 há uma mudança no sinal dos coeficientes dos parâmetros estimados.

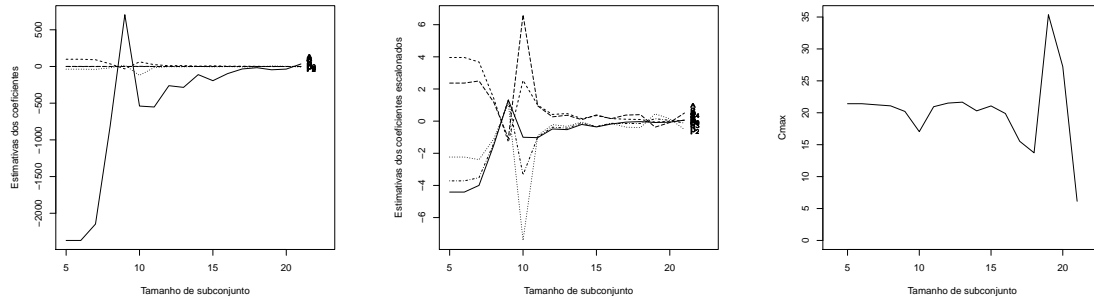


Figura 5.33: Gráficos das estimativas dos coeficientes, coeficientes escalonados e C_{max} dos dados de Stack Loss.

Tabela 5.2: Estimativa dos parâmetros dos dados de Stack Loss com a mudança relativa entre parênteses.

	completo	sem 1	sem 21	sem 1 21
$\hat{\beta}_0$	34.15 (0.00)	25.72 (0.25)	-35.72 (2.05)	-50.70 (2.48)
p_{valor}	0.32	0.49	0.52	0.40
$\hat{\beta}_1$	1.38 (0.00)	1.51 (0.10)	-0.17 (1.13)	-0.08 (1.06)
p_{valor}	0.23	0.21	0.91	0.95
$\hat{\beta}_2$	-8.47 (0.00)	-8.03 (0.05)	2.38 (1.28)	3.58 (1.42)
p_{valor}	0.03	0.05	0.76	0.66
$\hat{\beta}_3$	-0.03 (0.00)	-0.03 (0.00)	0.01 (1.38)	0.02 (1.47)
p_{valor}	0.07	0.07	0.71	0.66
$\hat{\beta}_4$	0.16 (0.00)	0.15 (0.04)	-0.03 (1.17)	-0.05 (1.30)
p_{valor}	0.02	0.03	0.84	0.73

() mudança relativa

Conclusões

6.1 Conclusões

Nesta dissertação discutimos as várias metodologias de diagnóstico de dados atípicos e/ou influentes como a influência global, influência local, curvatura normal conformal e a procura “forward”. Propomos o uso da influência local de Cook (1986) juntamente com a metodologia de procura “forward” de Atkinson e Riani (2000). Estas metodologias foram aplicadas a três conjuntos de dados.

No primeiro conjunto de dados, o conjunto de dados de ganso, tanto a influência global, como a influência local de Cook (1986) e a curvatura normal conformal de Poon e Poon (1999) identificaram as observações 29, 28 e 41 como possíveis pontos influentes (menos na perturbação da covariável (σ^2 conhecido) e resposta (σ^2 desconhecido) na influência local). Estas observações se destacam das demais observações como pode ser visto na Figura 3.2 e são as 3 observações com os maiores valores de x e y . Não foi detectada nenhuma observação mascarada atípica com a utilização do método de procura “forward” e também não foi identificado nenhuma observação mascarada influente utilizando a influência local com procura “forward”.

No segundo conjunto de dados, o conjunto de dados, de ratos, a influência global detectou a observação 3 como sendo atípica, enquanto que na influência local, foram

detectadas as observações 1, 13 e 19 na maioria das perturbações. Somente na perturbação da covariável e variável resposta com σ^2 conhecido destacou-se a observação 3 das demais observações. Considerando a procura “forward” a observação 5 foi identificada como um possível ponto atípico mascarado. Considerando a curvatura normal conformal, foram identificadas as observações 1, 5, 13 e 19. No enfoque de influência local com procura “forward” foi identificado a observação 3 como um possível dado influente mascarado em todos os tipos de perturbações, menos na perturbação na covariável com σ^2 desconhecido, que foi a única perturbação onde a observação 3 se destacou como um possível ponto influente.

As observações 3 e 5 detectadas como dados mascarados nas metodologias de influência local com procura “forward” e procura “forward”, respectivamente, são as observações dos dois ratos com o maior peso corporal e que receberam as maiores doses. O rato 3 teve a maior retenção de dose em relação ao restante do conjunto de dados e o rato 5 teve a menor retenção de dose em relação aos demais ratos, apesar de ter recebido a maior dose.

No terceiro conjunto de dados, o conjunto de dados de Stack Loss, as observações 3 e 4 foram identificadas tanto na influência global, quanto na local e na curvatura normal conformal. Além destas observações, na influência global foram detectadas também as observações 2 e 21 e na influência local na ponderação de casos com σ^2 desconhecido e também na curvatura normal conformal foi detectada a observação 2. Como dados mascarados, foram identificadas as observações 1 e 2 na procura “forward” e na influência local com procura “forward” as observações (1, 21), (2, 21), (1, 2), (1, 21) e (1, 2, 21) dependendo do tipo de perturbação utilizado.

Desta forma, estas diferentes metodologias se complementam, já que em um conjunto de dados pode haver tanto dados atípicos, dados influentes e/ou dados mascarados.

A metodologia proposta de influência local com procura “forward” foi eficaz na detecção de dados influentes mascarados. No conjunto de dados de rato, considerando a ponderação de casos com σ^2 conhecido e desconhecido e a perturbação na variância, foram identificadas as observações 1, 13 e 19. Utilizando estas mesmas perturbações em conjunto com a procura “forward” foi detectado a observação 3 como um dado influente mascarado. O mesmo acontece com as perturbações na covariável e a variável resposta, onde na influência local foram detectadas somente as observações 1 e 19, mas utilizando a metodologia de influência local com procura “forward”, considerando estas perturbações, foram identificados as observações 1, 2 e 21 como pontos influentes mascarados na perturbação da covariável com σ^2 desconhecido. No conjunto de dados

de Stack Loss, também foram identificadas observações influentes mascaradas que não havia sido identificado considerando a influência local de Cook (1986).

6.2 Trabalhos futuros

Uma metodologia bastante utilizada para a análise de diagnóstico em modelos de regressão com erros nas variáveis (Cheng e Van Ness (1999), Fuller (2009), Moran (1971)) é a influência local de Cook (1986) (veja por exemplo, Lee e Zhao (1996), Galea-Rojas et al. (2002), Lee et al. (2006), Lee e Tang (2004), Rasekh (2006), Aoki et al. (2007)). Como uma continuação deste trabalho, vamos considerar os modelos de regressão com erros nas variáveis e utilizar a metodologia de influência local com procura “forward”.

Uma outra extensão a ser considerada é nos modelos de efeitos mistos linear.

Pretendemos também, considerar o enfoque de Poon e Poon (1999) em conjunto com a metodologia procura “forward”.

Referências Bibliográficas

- Aoki, R. (2001). *Modelos de regressão com erros nas variáveis com intercepto nulo*. Tese de Doutorado, Departamento de Estatística, Universidade de São Paulo.
- Aoki, R., J. da Motta Singer, e H. Bolfarine (2007). Local influence for measurement error regression models for the analysis of pretest/posttest data. *Journal of Applied Statistical Science* 15(3), p. 317–330.
- Atkinson, A. e H.-M. Mulira (1993). The stalactite plot for the detection of multivariate outliers. *Statistics and Computing* 3(1), p. 27–35.
- Atkinson, A. e M. Riani (2000). *Robust diagnostic regression analysis*. Springer.
- Atkinson, A. C. (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Clarendon Press Oxford.
- Barnett, V. e T. Lewis (1994). *Outliers in statistical data*, Volume 3. Wiley New York.
- Belsey, D. A., E. Kuh, e R. E. Welsch (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley and Sons.
- Billor, N. e R. Loynes (1993). Local influence: a new approach. *Communications in Statistics-Theory and Methods* 22(6), p. 1595–1611.
- Brownlee, K. A. (1965). *Statistical theory and methodology in science and engineering*, Volume 150. Wiley New York.
- Ceroli, A. e M. Riani (1999). The ordering of spatial data and the detection of multiple outliers. *Journal of Computational and Graphical Statistics* 8(2), p. 239–258.
- Chatterjee, S. e A. S. Hadi (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, p. 379–393.
- Chatterjee, S. e A. S. Hadi (1988). *Sensitivity Analysis in Linear Regression*. Wiley Series in Probability and Statistics.

- Cheng, C.-L. e J. W. Van Ness (1999). *Statistical regression with measurement error*. John Wiley & Sons.
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* 19(1), p. 15–18.
- Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)* 48(2), p. 133–169.
- Cook, R. D. e S. Weisberg (1982). Residuals and influence in regression. *Monographs on Statistics and Applied Probability*.
- Dodge, Y. e A. S. Hadi (1999). Simple graphs and bounds for the elements of the hat matrix. *Journal of Applied Statistics* 26(7), 817–823.
- Donoho, D. L. e P. J. Huber (1983). The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, p. 157–184.
- Fuller, W. A. (2009). *Measurement Error Models*, Volume 305. John Wiley and Sons.
- Fung, W. K. e C. Kwan (1997). A note on local influence based on normal curvature. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59(4), p. 839–843.
- Galea-Rojas, M., H. Bolfarine, e M. de Castro (2002). Local influence in comparative calibration models. *Biometrical journal* 44(1), p. 59–81.
- Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 761–771.
- Hadi, A. S. e J. S. Simonoff (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association* 88(424), p. 1264–1272.
- Hadi, A. S. e P. F. Velleman (1997). Computationally efficient adaptive methods for the identification of outliers and homogeneous groups in large data sets. Em *Joint Statistical Meetings, Anaheim/Orange County*, pp. 10–14.
- Hawkins, D. M. (1980). *Identification of outliers*, Volume 11. Springer.
- Hawkins, D. M., D. Bradu, e G. V. Kass (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26(3), p. 197–208.
- Lee, A. H. e Y. Zhao (1996). Assessing local influence in measurement error models. *Biometrical journal* 38(7), p. 829–841.
- Lee, S.-Y., B. Lu, e X.-Y. Song (2006). Assessing local influence for nonlinear structural equation models with ignorable missing data. *Computational statistics & data analysis* 50(5), p. 1356–1377.

- Lee, S.-Y. e N.-S. Tang (2004). Local influence analysis of nonlinear structural equation models. *Psychometrika* 69(4), p. 573–592.
- Montgomery, D., C. L. Jennings, e M. Kulahci (2011). *Introduction to time series analysis and forecasting*, Volume 526. John Wiley and Sons.
- Montgomery, D., E. Peck, e G. Vining (2001). *Introduction to linear regression analysis*. John Wiley and Sons.
- Moran, P. (1971). Estimating structural and functional relationships. *Journal of Multivariate Analysis* 1(2), p. 232–255.
- Nobre, J. S. (2004). Métodos de diagnóstico para modelos lineares mistos. Dissertação de mestrado, Departamento de Estatística, Universidade de São Paulo, São Paulo, Brasil.
- Pamula, R., J. K. Deka, e S. Nandi (2011). An outlier detection method based on clustering. Em *Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on*, pp. p. 253–256. IEEE.
- Paula, G. (2004). Modelos de regressão: com apoio computacional. São Paulo, Brasil. Universidade de São Paulo.
- Poon, W.-Y. e Y. S. Poon (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(1), p. 51–61.
- Rasekh, A. (2006). Local influence in measurement error models with ridge estimate. *Computational statistics & data analysis* 50(10), p. 2822–2834.
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association* 79(388), p. 871–880.
- Rousseeuw, P. J. e B. C. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85(411), p. 633–639.
- Russo, C. M. (2006). Análise de um modelo de regressão com erros nas variáveis multivariado com intercepto nulo. Dissertação de Mestrado, Universidade de São Paulo.
- Samprit, C. e S. H. Ali (2006). *Regression analysis by example*. John Wiley and Sons.
- Singer, J. M. e D. F. de Andrade (1986). Análise de dados longitudinais. *Simpósio Nacional de Probabilidade e Estatística 7*.
- Soares da Silva Gomes, G. (2005). Análise de influência para a distribuição dirichlet. Dissertação de mestrado, Universidade Federal de Pernambuco.
- Souza, E. C. (2006). Análise de influência local no modelo de regressão logística. Dissertação de mestrado, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, Brasil.

Weisberg, S. (2005). *Applied linear regression*, Volume 528. John Wiley and Sons.

Conjunto de dados

Os conjuntos de dados utilizados nesse trabalho foram os seguintes:

A.1 GANSO

Tabela A.1: Dados ganso.

Número	Foto	obs1	obs2
1	56	50	40
2	38	25	30
3	25	30	40
4	48	35	45
5	38	25	30
6	22	20	20
7	22	12	20
8	42	34	35
9	34	20	30
10	14	10	12
11	30	25	30
12	9	10	10
13	18	15	18
14	25	20	30
15	62	40	50
16	26	30	20
17	88	75	120
18	56	35	60
19	11	9	10
20	66	55	80

Continua na próxima página.

Tabela A.1: Dados de ganso (continuação).

Número	Foto	obs1	obs2
21	42	30	35
22	30	25	30
23	90	40	120
24	119	75	200
25	165	100	200
26	152	150	150
27	205	120	200
28	409	250	300
29	342	500	500
30	200	200	300
31	73	50	40
32	123	75	80
33	150	150	120
34	70	50	60
35	90	60	100
36	110	75	120
37	95	150	150
38	57	40	40
39	43	25	35
40	55	100	110
41	325	200	400
42	114	60	120
43	83	40	40
44	91	35	60
45	56	20	40

A.2 RATO

Os dados de rato são como segue

Tabela A.2: Dados rato

	Peso corporal (x_1)	Peso do fígado (x_2)	Dose relativa (x_3)	y
1	176	6.50	0.88	0.42
2	176	9.50	0.88	0.25
3	190	9.00	1.00	0.56
4	176	8.90	0.88	0.23
5	200	7.20	1.00	0.23
6	167	8.90	0.83	0.32
7	188	8.00	0.94	0.37
8	195	10.00	0.98	0.41
9	176	8.00	0.88	0.33
10	165	7.90	0.84	0.38
11	158	6.90	0.80	0.27
12	148	7.30	0.74	0.36
13	149	5.20	0.75	0.21
14	163	8.40	0.81	0.28
15	170	7.20	0.85	0.34
16	186	6.80	0.94	0.28
17	146	7.30	0.73	0.30
18	181	9.00	0.90	0.37
19	149	6.40	0.75	0.46

A.3 STACK LOSS

Tabela A.3: Dados Stack Loss

	Fluxo de ar	Temperatura de entrada da agua	Concentração de acido	Perda de pilha
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Influência local Gráficos

Neste Apêndice encontra-se os gráficos do autovetor, l_{max} , correspondente ao maior autovalor, C_{max} , para cada tipo de perturbações descrita na Secção 2.2.

Gráficos indicado por (a), (b), (c), (d), (e) e (f) representam, respectivamente,

(a): o diagrama do valor absoluto do autovetor correspondente ao maior autovalor com os resíduos.

(b): o diagrama do valor absoluto do autovetor correspondente ao maior autovalor com os indivíduos.

(c): o diagrama do valor absoluto do autovetor correspondente ao maior autovalor com os dados da variável resposta.

(d), (e) e (f): o diagrama do valor absoluto do autovetor correspondente ao maior autovalor com os dados das covariáveis x_1 , x_2 e x_3 , respectivamente.

B.1 Conjunto de dados de ganso

B.1.1 Ponderação de casos - quando σ^2 é conhecido

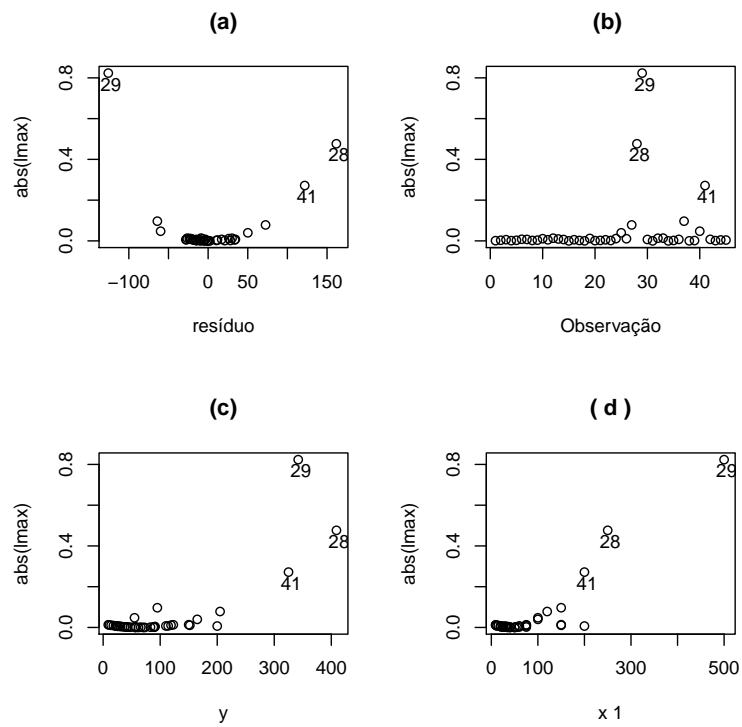


Figura B.1: Gráfico na ponderação de casos (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável no conjunto dos dados de ganso.

B.1.2 Ponderação de casos - quando σ^2 não é conhecido

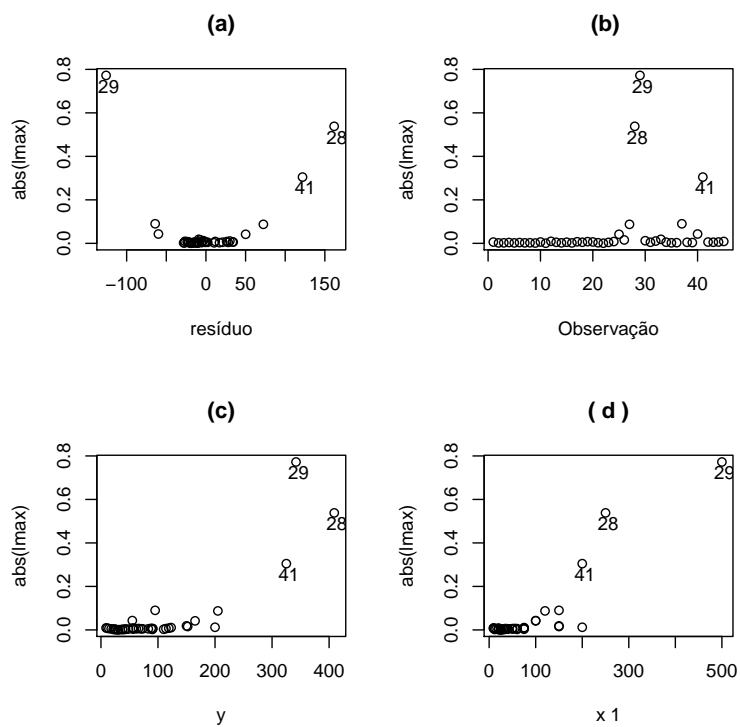


Figura B.2: Gráfico na ponderação de casos (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável dos dados de ganso.

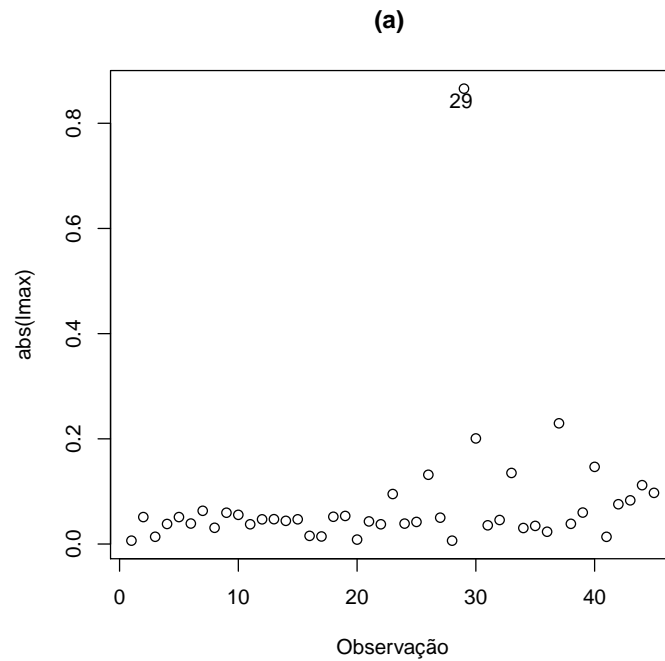
B.1.3 Perturbação na covariável - quando σ^2 é conhecido

Figura B.3: Gráfico na perturbação na covariável (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de ganso.

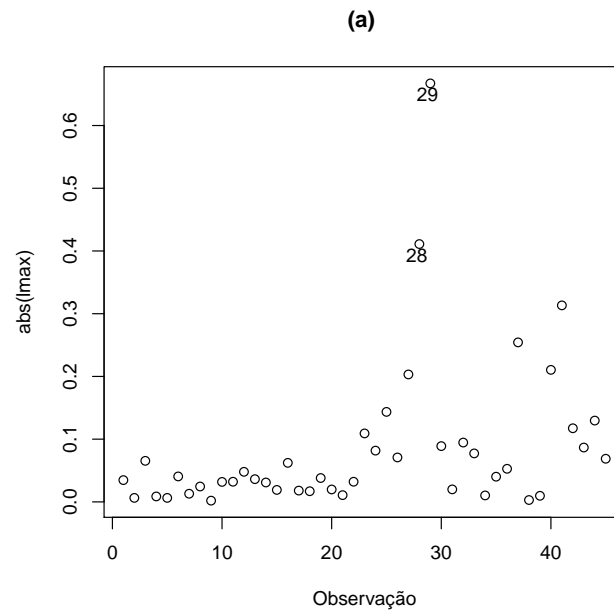
B.1.4 Perturbação na covariável - quando σ^2 não é conhecido

Figura B.4: Gráfico na perturbação na covariável (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de ganso.

B.1.5 Perturbação na variável resposta - quando σ^2 é conhecido

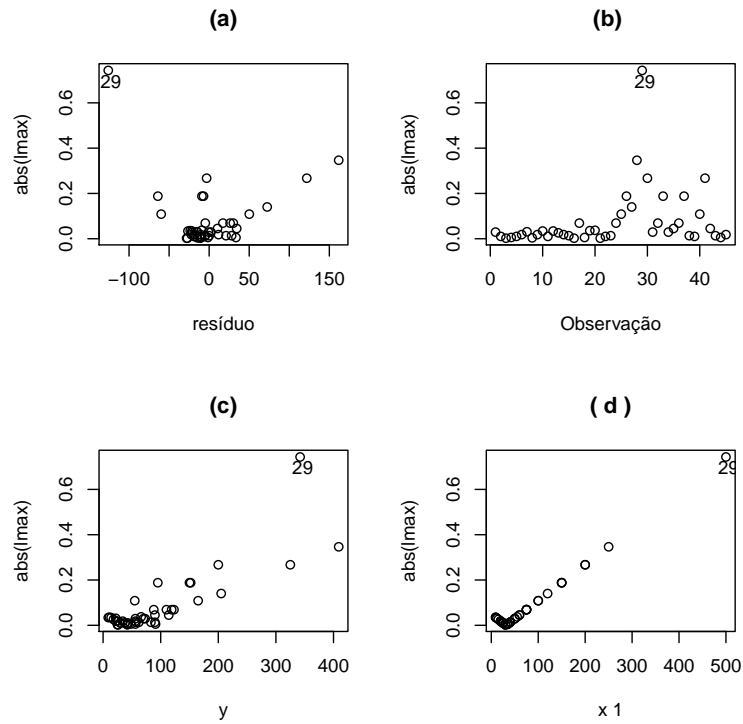


Figura B.5: Gráfico na perturbação na resposta (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) residuo (b) observação (c) resposta y e a covariável dos dados de ganso.

B.1.6 Perturbação na variável resposta - quando σ^2 não é conhecido

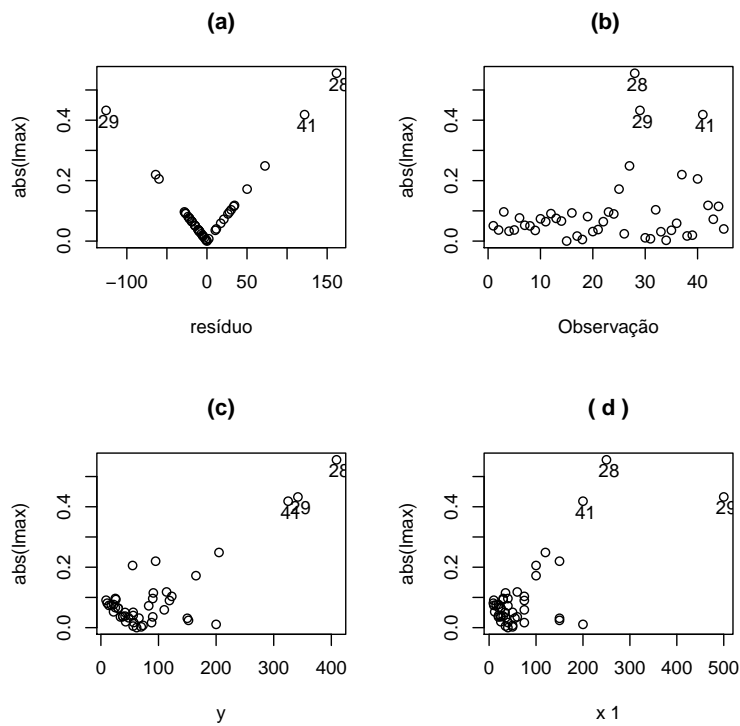


Figura B.6: Gráfico na perturbação na resposta (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável dos dados de ganso.

B.1.7 Perturbação na variância do erro - quando σ^2 é não conhecido

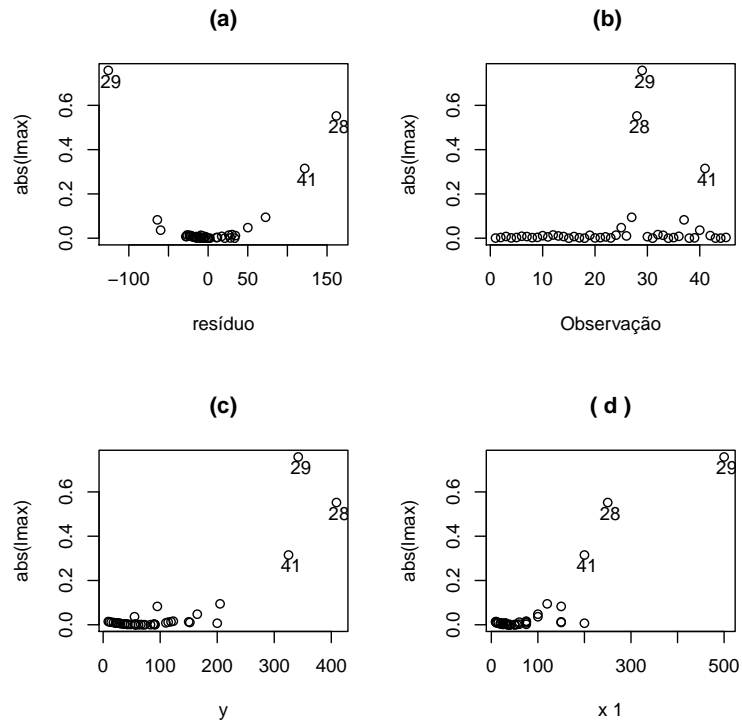


Figura B.7: Gráfico na perturbação na variância (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e a covariável no conjunto dos dados de ganso.

B.2 Conjunto de dados de rato

B.2.1 Ponderação de casos - quando σ^2 é conhecido

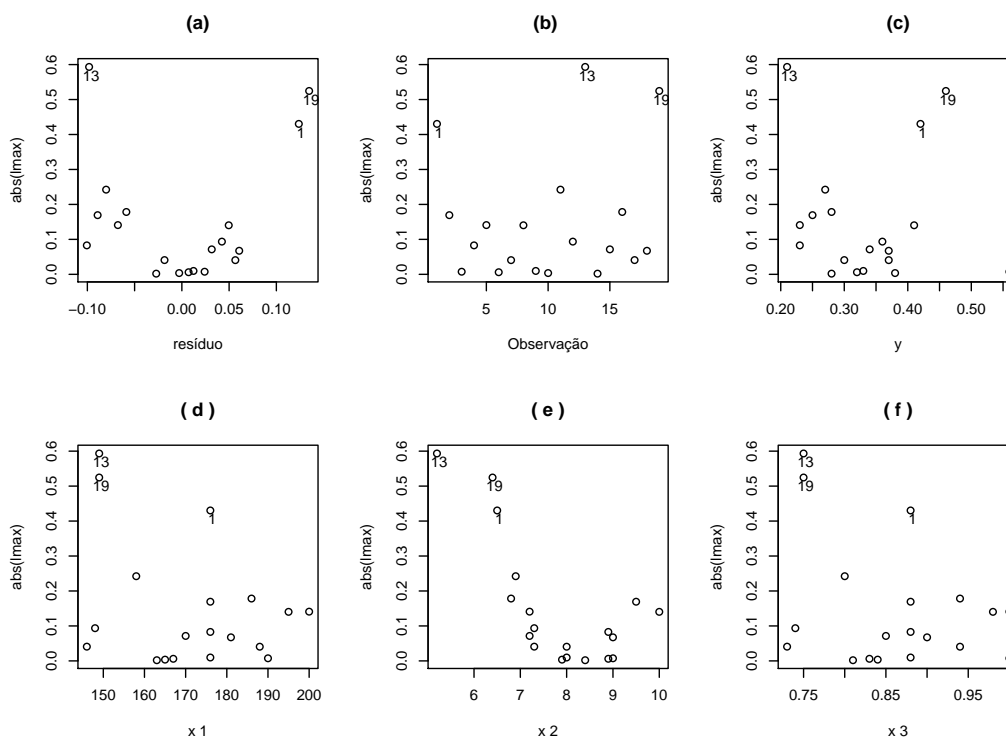


Figura B.8: Gráfico na ponderação de casos (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis dos dados de rato.

B.2.2 Ponderação de casos - quando σ^2 não é conhecido

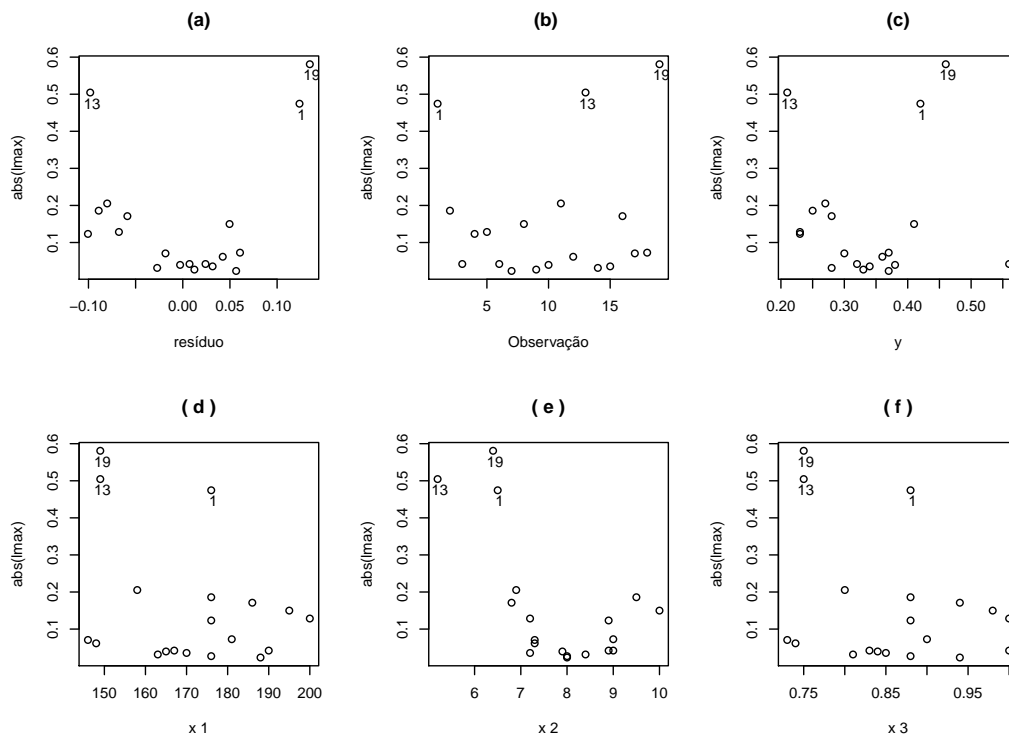


Figura B.9: Gráfico na ponderação de casos (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) residuo (b) observação (c) resposta y e as covariáveis dos dados de rato.

B.2.3 Perturbação na covariável - quando σ^2 é conhecido

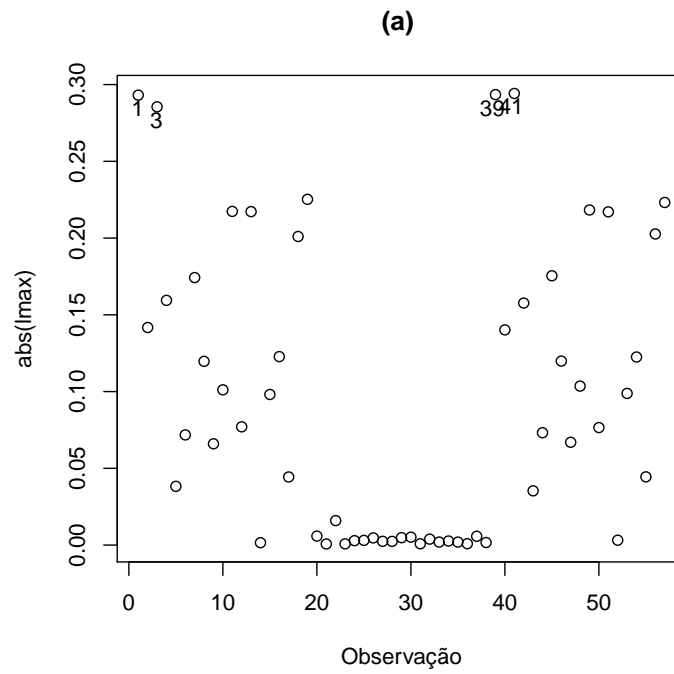


Figura B.10: Gráfico na perturbação na covariável (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de rato. As observações 1 a 19, 20 a 38 e 39 a 57 respectivamente, a covariável x_1 , x_2 e x_3 .

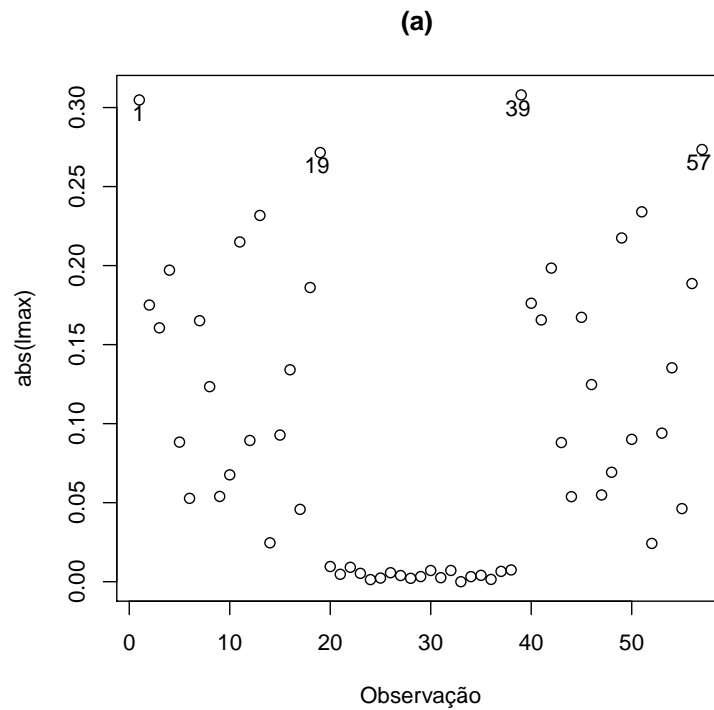
B.2.4 Perturbação na covariável - quando σ^2 não é conhecido

Figura B.11: Gráfico na perturbação na covariável (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de rato. As observações 1 a 19, 20 a 38 e 39 a 57 respectivamente, a covariável x_1 , x_2 e x_3 .

B.2.5 Perturbação na variável resposta - quando σ^2 é conhecido

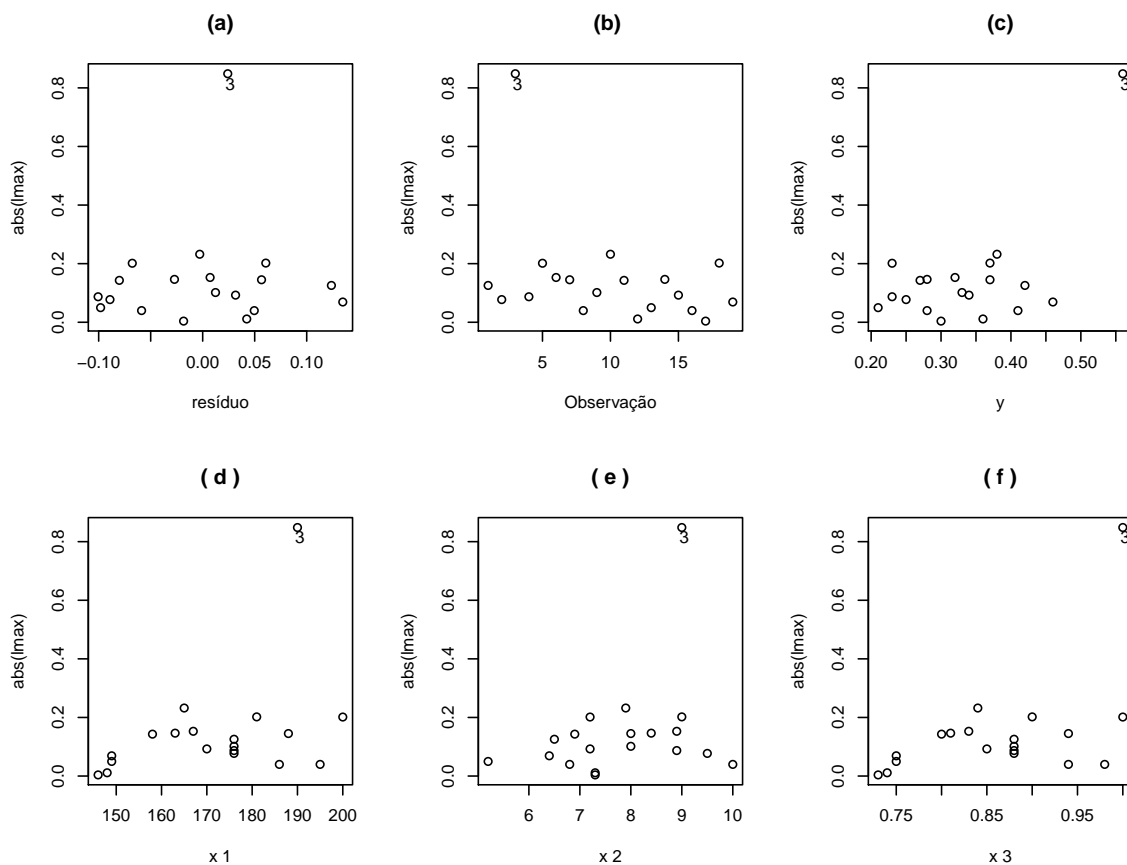


Figura B.12: Gráfico na perturbação na resposta (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) residuo (b) observação (c) resposta y e as covariáveis dos dados de rato.

B.2.6 Perturbação na variável resposta - quando σ^2 não é conhecido

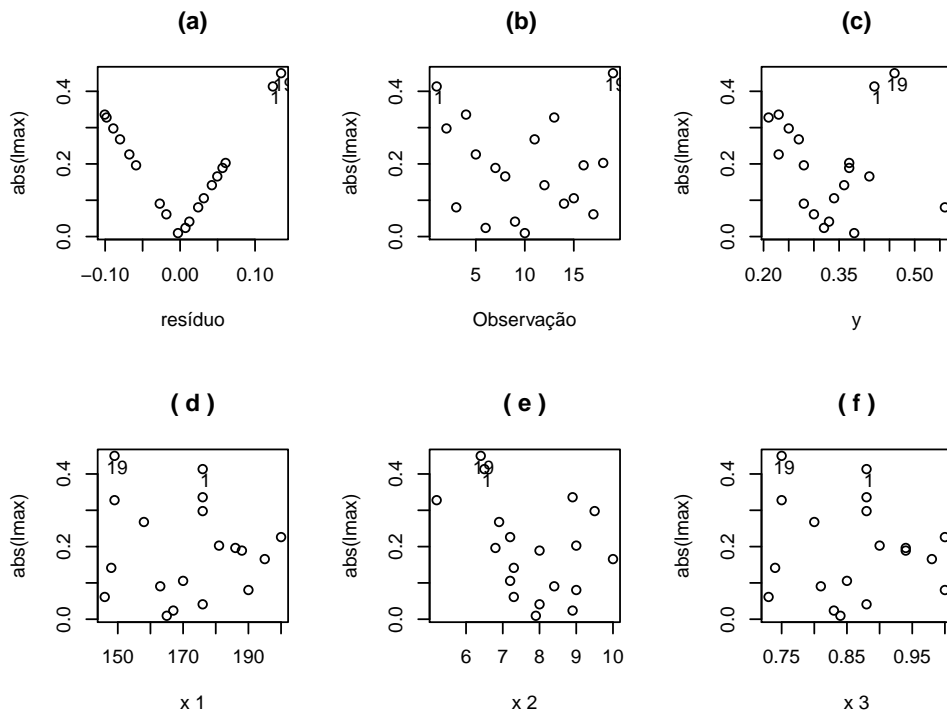


Figura B.13: Gráfico na perturbação na resposta (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis dos dados de rato.

B.2.7 Perturbação na variância do erro - quando σ^2 é não conhecido

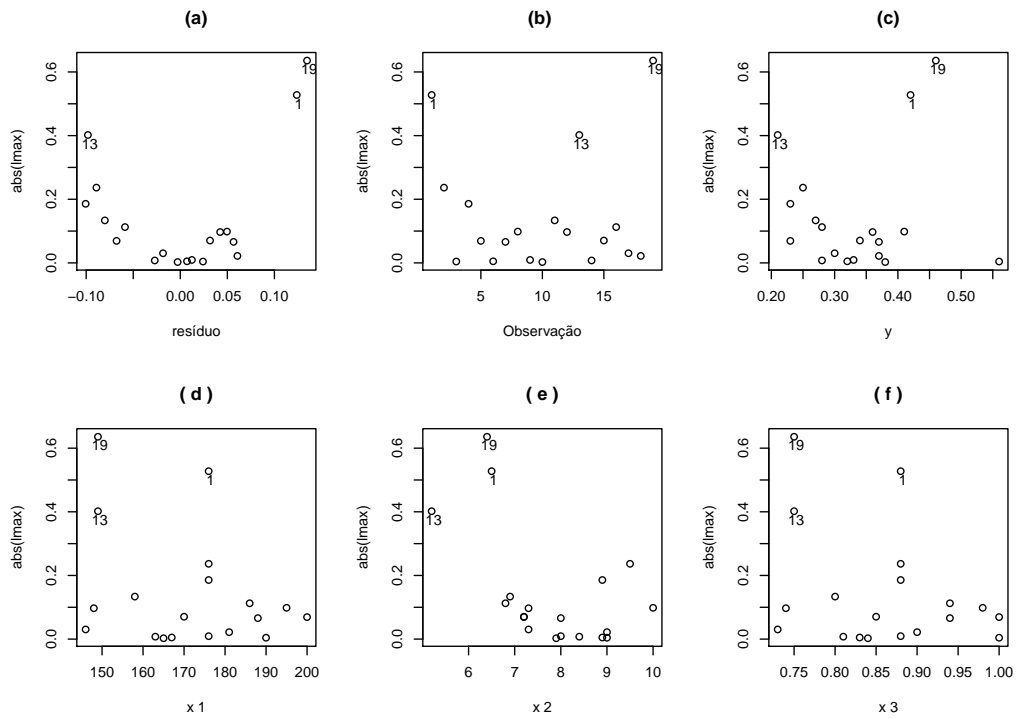


Figura B.14: Gráfico na perturbação na variância (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis do conjunto de dados de rato.

B.3 Conjunto de dados de “Stack Loss”

B.3.1 Ponderação de casos - quando σ^2 é conhecido

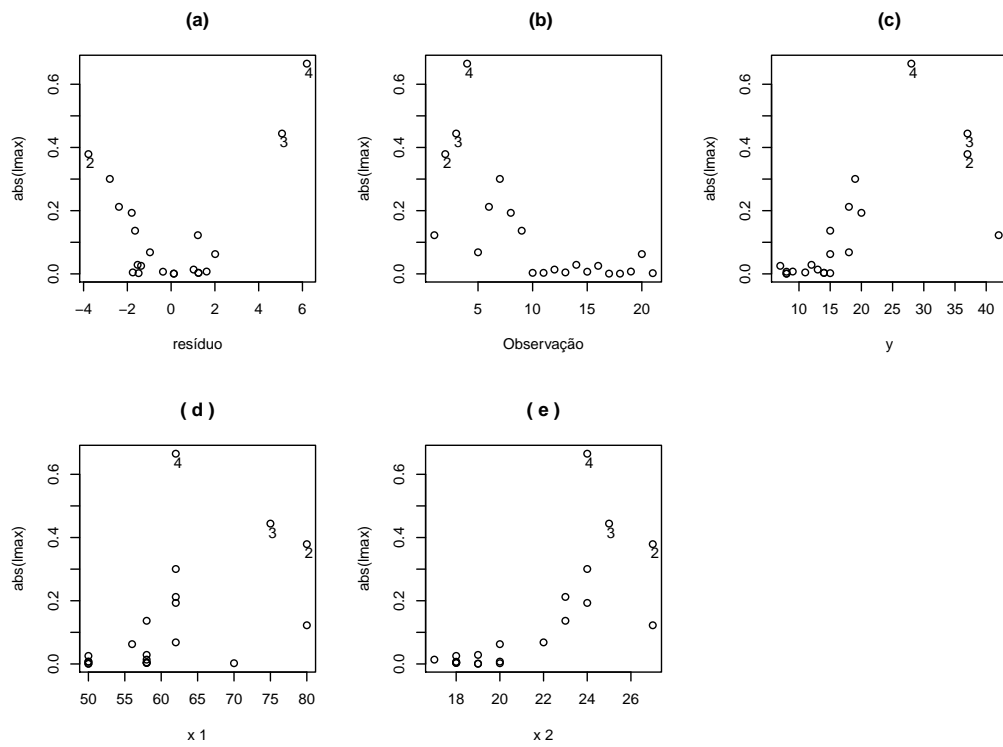


Figura B.15: Gráfico na ponderação de casos (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis no conjunto de dados Stack Loss.

B.3.2 Ponderação de casos - quando σ^2 não é conhecido

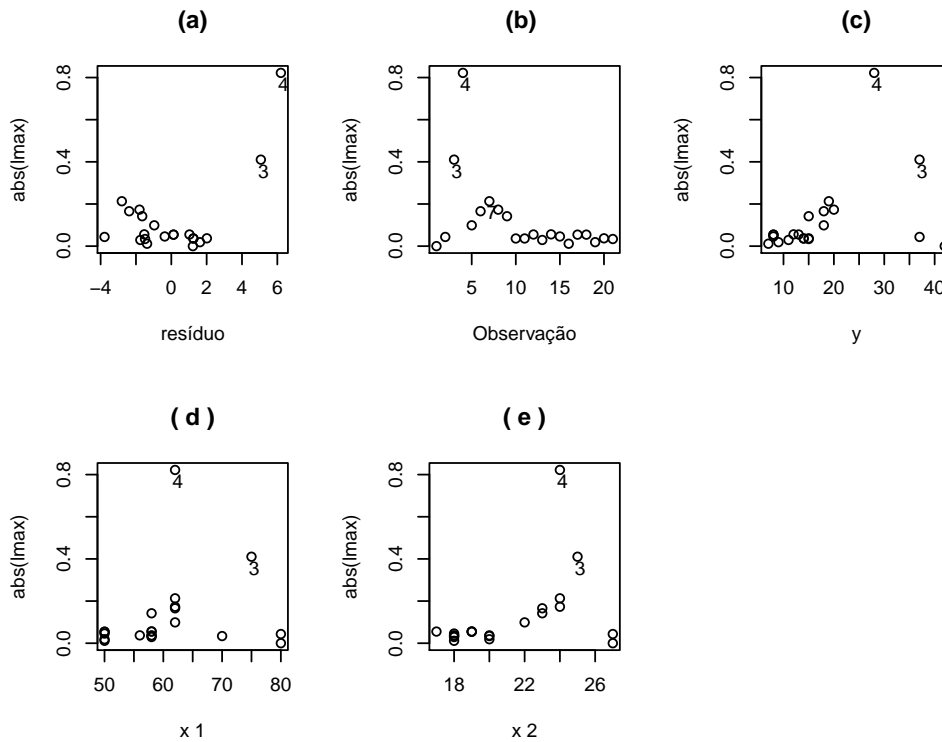


Figura B.16: Gráfico na ponderação de casos (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis no conjunto dos dados de Stack Loss.

B.3.3 Perturbação na covariável - quando σ^2 é conhecido

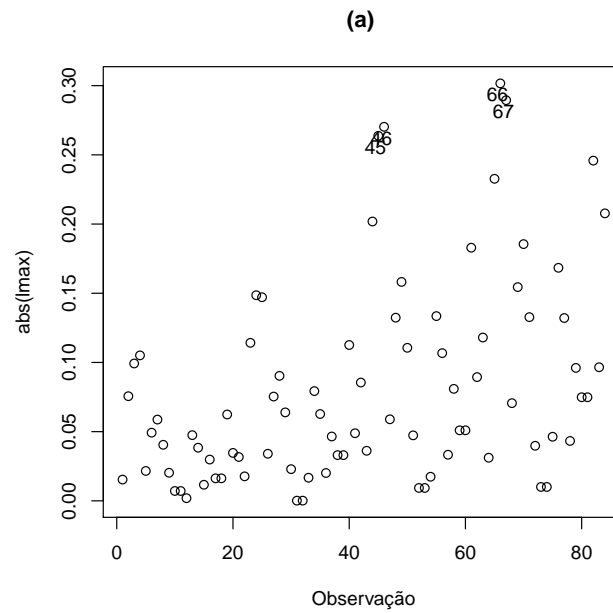


Figura B.17: Gráfico na perturbação na covariável (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de Stack Loss. As observações de 1 a 21, 22 a 42, 43 a 63, 64 a 84 referem-se respectivamente, a covariáveis x_1 , x_2 , x_3 e x_4 .

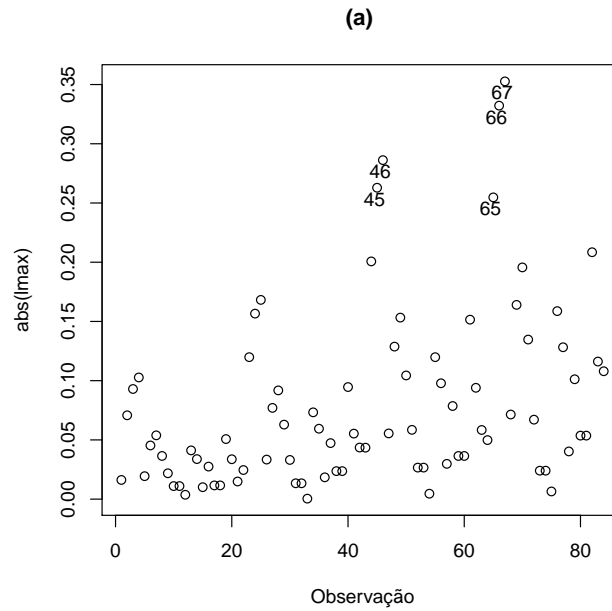
B.3.4 Perturbação na covariável - quando σ^2 não é conhecido

Figura B.18: Gráfico na perturbação na covariável (σ^2 desconhecido) do valor absoluto do autovetor correspondente ao maior autovalor dos dados de Stack Loss. As observações de 1 a 21, 22 a 42, 43 a 63, 64 a 84 referem-se respectivamente, a covariáveis x_1 , x_2 , x_3 e x_4 .

B.3.5 Perturbação na variável resposta - quando σ^2 é conhecido

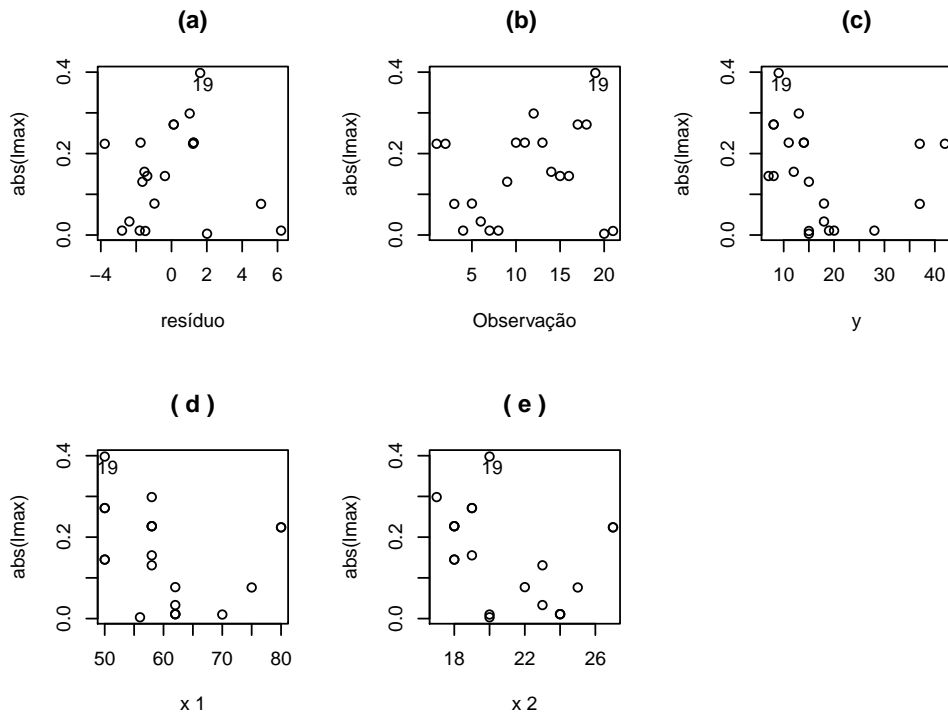


Figura B.19: Gráfico na perturbação na resposta (σ^2 conhecido) do valor absoluto do autovetor correspondente ao maior autovalor com: (a) resíduo (b) observação (c) resposta y e as covariáveis dos dados de Stack Loss.

B.3.6 Perturbação na variável resposta - quando σ^2 não é conhecido

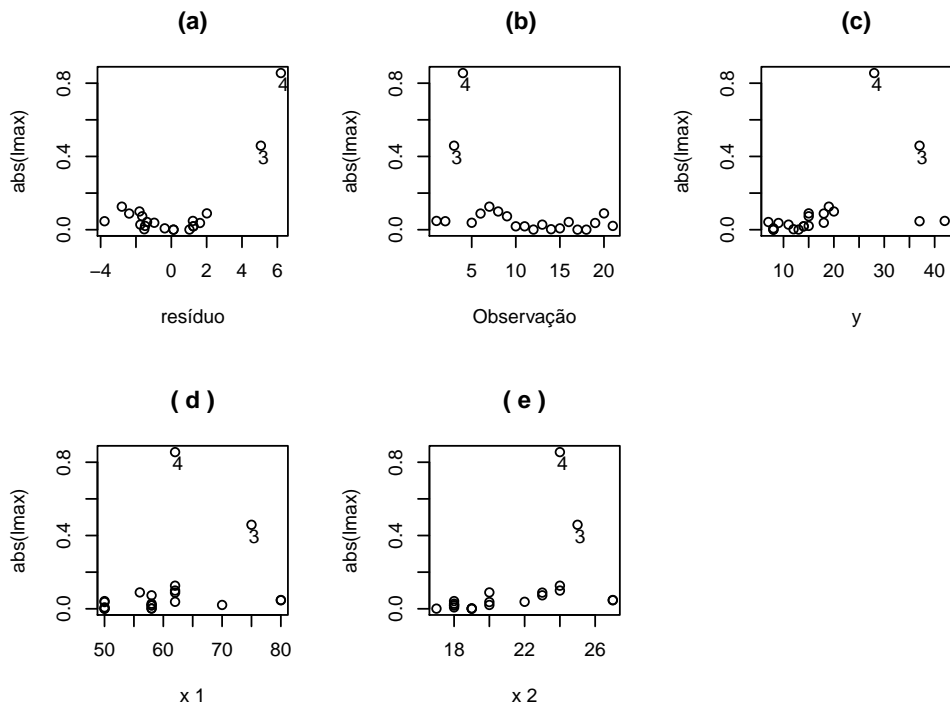


Figura B.20: Gráfico na perturbação na resposta (σ^2 desconhecido) de autovetor máximo dos dados de Stack Loss com: (a) resíduo (b) observação (c) resposta y e as covariáveis.

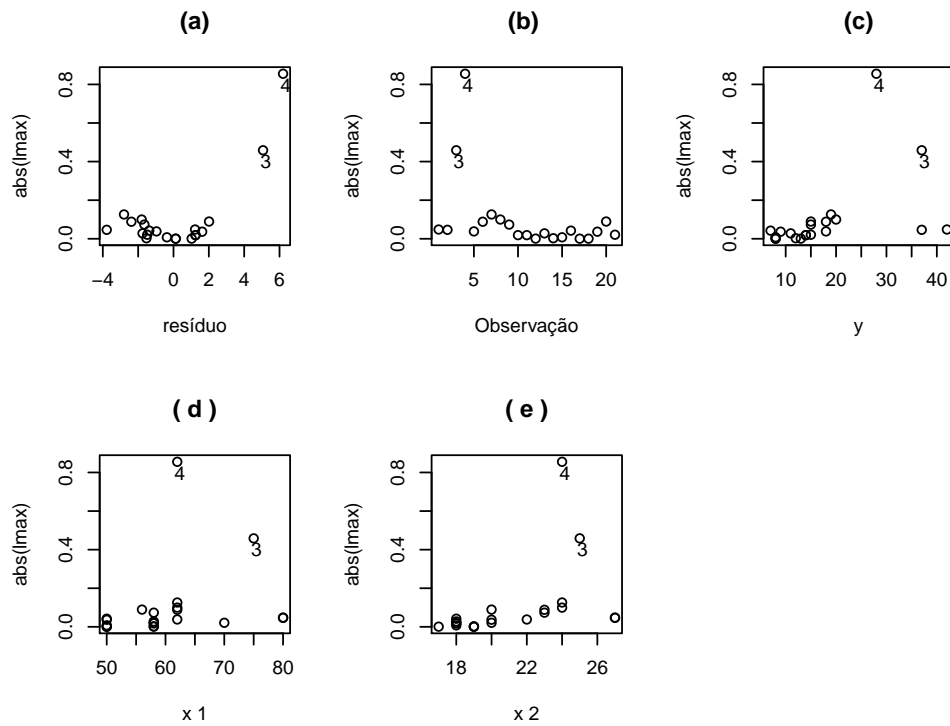
B.3.7 Perturbação na variância do erro - quando σ^2 é não conhecido

Figura B.21: Gráfico na perturbação na variância (σ^2 desconhecido) de autovetor máximo dos dados de Stack Loss com: (a) resíduo (b) observação (c) resposta y e as covariáveis.