Análise de diagnóstico pelo método *forward search*

João Domingos Celeste Junior

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística

Orientadora: Prof. Dra. Lúcia Pereira Barroso

São Paulo, maio de 2010

Análise de diagnóstico pelo método *forward search*

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por João Domingos Celeste Junior
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Profa. Dra. Lúcia Pereira Barroso - IME-USP.
- Profa. Dra. Mônica Carneiro Sandoval - IME-USP.
- Prof. Dr. Rinaldo Artes - INSPER.

Agradecimentos

Agradeço à minha família pelo apoio incondicional em mais esta etapa da minha vida, em especial minha mãe Terezinha. Aos amigos, que nos momentos difíceis me deram força para seguir até o fim. Entre este grande grupo de amigos em especial agradeço a colega Núbia Esteban, Renata Pelissari e Iara Nascimento.

Principalmente agradeço à professora Lúcia que conduziu brilhantemente o trabalho, desde a sugestão do tema como o desenvolvimento; agradeço ainda a forma amiga com que fui tratado durante sua orientação.

Resumo

O processo de análise e modelagem estatística é sensível à presença de observações *outliers* e tal tipo de dado pode trazer erros ou confundimentos neste processo. Com base neste fato, faz-se necessário o uso de técnicas estatísticas que têm a finalidade de detecção destes dados.

Como complemento às técnicas de detecção de *outliers* surge o método *forward search*, proposto por Atkinson e Riani (2000). Neste trabalho é apresentada a abordagem dos autores do método, no contexto de análise descritiva multivariada bem como para modelos de regressão. Também é proposta uma adaptação do método para o contexto de modelos de equações estruturais.

Palavras-chave: *forward search*, equações estruturais.

Abstract

The process of analysis and statistical modeling is sensitive to the presence of outliers observations, this type of data can bring errors or confuse this process. Based on this fact it is necessary to use statistical techniques that have the purpose of detection of this type of data.

In addition to techniques for detection of outlier observations comes the forward search method, proposed by Atkinson and Riani (2000). This dissertation presents the approach of the authors of the method, in the context of descriptive analysis and multivariate regression models. It is also a propose to adapt the method to the context of structural equation models.

Keywords: forward search, structural equation.

Sumário

Agradecimentos iii

Resumo v

Abstract vii

Lista de Figuras xi

Lista de Tabelas xv

1 Introdução 1

2 Método *Forward Search* 3

2.1 QQ plot das Distâncias de Mahalanobis 4

2.2 Boxplot Bivariado 10

2.3 Monitoramento do *Forward Search* 16

2.4 Monitoramento da Matriz de Covariância 19

3 Regressão Linear 23

3.1 Descrição do Método 24

3.2 *Forward Search* na Regressão Linear 26

3.3 Distância de Cook 27
3.4 Estatística t para os Parâmetros do Modelo 28
3.5 Aplicação 30

4 Modelos Lineares Generalizados **45**
4.1 Função de Ligação e Preditores Lineares 46
4.2 Função Desvio 47
4.3 Análise de Resíduo em MLG 47
4.4 Verificação da Função de Ligação 48
4.5 Monitoramento do *Forward Search* para MLG 49
4.6 Aplicação 49

5 Modelos de Equações Estruturais **57**
5.1 Especificação do Modelo Fatorial 57
5.2 Análise Fatorial Exploratória x Análise Fatorial Confirmatória 59
5.3 Modelos de Equações Estruturais 62
5.4 Identificação 64
5.5 Estimação 65
5.6 Medidas de Ajuste 67
5.7 Método *Forward Search* em Modelos de Equações Estruturais 69
5.8 Aplicação 69

6 Conclusões **95**

7 Apêndice **97**

Referências Bibliográficas **101**

Lista de Figuras

2.1 QQ plot das distâncias de Mahalanobis para os recordes de corrida feminina 10

2.2 Gráficos de dispersão para os dados de recordes de corrida feminina 12

2.3 *Boxplots* bivariados, com $\alpha = 0,01$, para os dados de recordes de corrida feminina . . . 13

2.4 *Boxplots* das variáveis da prova de corrida para o *grupo limpo* 15

2.5 Gráfico *forward search* para as distâncias de Mahalanobis dos dados de corrida 17

2.6 Gráfico *forward search* das distâncias de Mahalanobis para os dados de corrida 19

2.7 Gráfico *forward search* para as estimativas da variância generalizada dos dados de corrida 20

2.8 Gráfico *forward search* dos elementos das matrizes de covariância (à esquerda) e correlação (à direita) dos dados de corrida 21

3.1 Gráficos de dispersão para os dados de combustível 31

3.2 *Boxplots* bivariados, com $\alpha = 0,01$, para os dados de combustível 32

3.3 Gráfico *forward search* dos resíduos padronizados dos dados de combustível 36

3.4 Gráfico*forward search* para as distâncias de Cook modificadas dos dados de combustível 37

3.5 Gráfico *forward search* para as medidas de alavanca dos dados de combustível 38

3.6 Gráfico *forward search* para as estimativas dos parâmetros dos dados de combustível . 39

3.7 Gráfico *forward search* para as estatísticas t dos dados de combustível 40

3.8 Gráfico *forward search* para a estatística t sem os estados Nevada, South Dakota e Wyoming 41

3.9 Gráfico de probabilidade normal com todos os estados 42

3.10 Gráfico de probabilidade normal sem os estados South Dakota, Nevada e Wyoming . . 43

4.1 Gráfico *forward search* da *deviance* dos dados de ataques aéreos 52

4.2 Gráfico *forward search* da distância de Cook modificada (à esquerda) e valores de alavanca (à direita) dos dados de ataques aéreos 53

4.3 Gráfico *forward search* das estimativas dos parâmetros dos dados de ataques aéreos . . 54

4.4 Gráfico *forward search* da estatística t dos dados de ataques aéreos 55

4.5 Gráfico *forward search* da estatística do teste para função de ligação dos dados de ataques aéreos 56

5.1 Exemplo de diagrama de caminhos do modelo de análise fatorial exploratória 60

5.2 Exemplo de diagrama de caminhos do modelo de análise fatorial confirmatória 61

5.3 Exemplo de diagrama de caminhos do modelo de equações estruturais 63

5.4 *Boxplots* bivariados para os dados hipotéticos 71

5.5 Gráfico *forward search* para as distâncias de Mahalanobis escalonadas para os dados hipotéticos 72

5.6 Gráfico *forward search* para a variância generalizada dos dados hipotéticos 73

5.7 Gráfico *forward search* dos elementos da matriz de covariância para os dados hipotéticos 74

5.8 Diagrama de caminhos para os dados hipotéticos 75

5.9 Diagrama de caminhos para os dados hipotéticos com as estimativas dos parâmetros . 76

5.10 Gráfico *forward search* da estatística qui-quadrado para os dados hipotéticos 77

5.11 Gráfico *forward search* para o índice de qualidade de ajuste GFI_{MV} para os dados hipotéticos 78

5.12 Gráfico *forward search* para a medida SRMR para os dados hipotéticos 79

5.13 Diagrama de caminhos para os dados de habilidade verbal e espacial 81

5.14 Gráficos de dispersão para os dados de habilidade verbal e espacial 82

5.15 QQ plot das distâncias de Mahalanobis para os dados sobre habilidade verbal e espacial 85

5.16 Gráfico *forward search* das distâncias de Mahalanobis para os dados de habilidade verbal e espacial 86

5.17 Gráfico *forward search* para as variâncias generalizadas para os dados de habilidade verbal e espacial 87

5.18 Gráfico *forward search* dos elementos da matriz de covariância para os dados de habilidade verbal e espacial 88

5.19 Diagrama de caminhos com as estimativas dos parâmetros para os dados de habilidade verbal e espacial 89

5.20 Gráfico *forward search* da estatística qui-quadrado para os dados de habilidade verbal e espacial 91

5.21 Gráfico *forward search* para a medida SRMR para os dados de habilidade verbal e espacial 92

5.22 Gráfico *forward search* para os índices de qualidade de ajuste GFI_{MV} (à esquerda) e $AGFI_{MV}$ (à diretia) para os dados de habilidade verval e espacial 93

Lista de Tabelas

2.1 Dados de recordes de corrida feminina 6

2.2 Dados de recordes de corrida feminina (continuação da Tabela 2.1) 7

2.3 Valores dos componentes necessários para construção do QQ plot 8

2.4 Valores dos componentes necessários para construção do QQ plot (continuação da Tabela 2.3) 9

2.5 Valores observados para os países do *grupo limpo* 14

2.6 Estatísticas descritivas para o *grupo limpo* dos dados de corrida 15

2.7 Inclusão dos países a cada passo do *forward search* 18

3.1 Dados do consumo de combustível 33

3.2 Dados do consumo de combustível (continuação da Tabela 3.1) 34

3.3 Estatísticas descritivas dos dados de combustível 34

3.4 Estimativas dos parâmetros referentes ao *grupo limpo* para os dados de combustível . . 35

3.5 Estimativas dos parâmetros no final do processo para os dados de combustível 35

3.6 Estimativas dos parâmetros referente ao modelo sem a variável estrada 41

3.7 Estimativas dos parâmetros para o ajuste sem os estados South Dakota, Nevada e Wyoming 43

4.1 Funções de ligação canônicas 46

4.2 Dados de ataques aéreos 50
4.3 Estimativas dos parâmetros para os dados de ataques aéreos 52
4.4 Estimativas dos parâmetros do modelo sem as variáveis modelo e experiência 53
4.5 Estimativas dos parâmetros sem as observações 16 e 25 54

5.1 Sumário do modelo de análise fatorial 58
5.2 Dados hipotéticos 70
5.3 Estatísticas descritivas dos dados hipotéticos 71
5.4 Variáveis observáveis do estudo sobre habilidade verbal e espacial 80
5.5 Dados de habilidade verbal e espacial 83
5.6 Dados de habilidade verbal e espacial (continuação da Tabela 5.5) 84
5.7 Estatísticas descritivas dos dados sobre habilidade verbal e espacial 85
5.8 Estatísticas de ajuste do modelo 88

Capítulo 1

Introdução

A modelagem e a análise estatística são cercadas por suposições e hipóteses a respeito do conjunto de dados, tais como homogeneidade das variâncias, estrutura de relacionamento entre as variáveis, distribuição de probabilidade que originou os dados, em suma aspectos que direcionam o tipo de modelo e análise a serem desenvolvidos com o conjunto de dados.

A presença de observações discrepantes ou *outliers* podem trazer confundimento e erros no processo de modelagem e análise estatística, tendo como consequência a não adequação do modelo aos dados, ou um ajuste duvidoso. Dado o empecilho provocado pela presença deste tipo de observação existe hoje uma vasta literatura referente às técnicas de identificação de observações *outliers*, seja no contexto univariado ou multivariado.

Como complemento às técnicas de identificação existentes, surge o método *forward search*, proposto por Atkinson e Riani (2000). O método tem por finalidade descrever explicitamente, através de gráficos, o impacto que cada observação tem sobre o ajuste de modelos e respectivas análises de diagnóstico.

As análises estatísticas presentes neste trabalho foram realizadas no *software* R com o auxílio dos pacotes *Rfwdmv* e *forward*, desenvolvidos pelos autores do método *forward search*; estes pacotes contém os principais comandos nescessários para aplicação do método.

Este trabalho foi desenvolvido em cinco capítulos.

No Capítulo 2 é introduzido o conceito do método *forward search* bem como o algoritmo base do método, é apresentado um exemplo de aplicação para melhor compreensão do processo.

No Capítulo 3 é exibida a abordagem dos autores do método no contexto de regressão linear,

acompanhado de uma aplicação em um cojunto de dados da literatura. De forma similar o capítulo 4 exibe a abordagem dos autores do método para modelos lineares generalizados (MLG).

No Capítulo 5 é desenvolvida uma adaptação do método *forward search* para modelos de equações estruturais. Tal adaptação caracteriza o objetivo do trabalho, uma vez que os autores do método não o abordam neste contexto. A estruturação do método nesta adaptação é acompanhada com aplicações a conjuntos de dados hipotéticos e da literatura, com o intuito de verificar a adequabilidade da adaptação proposta. O algoritmo cosntruído para tal adaptação segue no apêndice, juntamente com os comandos utilizados nas análises estatísticas dos capítulos anteriores.

Capítulo 2

Método *Forward Search*

A detecção e o tratamento de observações *outliers* é um tema abordado com frequência por pesquisadores e estudiosos da área estatística. A presença de tal tipo de observação gera impactos prejudiciais às análises estatísticas propostas, pois o padrão que se procura encontrar no conjunto de dados através da modelagem estatístisca será deturpado por tal observação discrepante.

Há uma vasta literatura na área estatística relativa às técnicas de identificação de tal tipo de dado, em que são abordadas desde técnicas descritivas, por exemplo o *boxplot*, até medidas-diagnóstico em modelos de regressão, como a distância de Cook.

Como proposta de melhoramento das técnicas de identificação existentes, surge o método *forward search*, proposto por Atkinson e Riani (2000), em que através de técnicas gráficas tem-se o objetivo de explicitar a presença e o impacto que observações *outliers* têm sobre as análises estatísticas. Os autores introduziram o método em diversos campos da estatística, como análise de regressão, análise de componentes principais, análise de agrupamentos, estatística espacial, entre outros.

Independentemente da área estatística de aplicação do método, o algoritmo *forward search* segue os seguintes passos:

1. a partir do conjunto de dados bruto é obtido, por meio da análise descritiva, um grupo de observações de dados livres de *outliers*, também denominado de *grupo limpo*;

2. a cada passo do método *forward search* tal grupo de dados limpos é acrescido de uma observação, proveniente do restante do conjunto de dados que o originou. A entrada de tal observação obedece um critério de distância em relação ao *grupo limpo*. A observação candidata a ser

incluída no grupo será aquela que possuir menor medida de distância, ou seja, é a mais próxima ao *grupo limpo* de *outliers*;

3. a entrada da observação leva ao recálculo dos parâmetros da medida de distância adotada.

O processo de inserção das observações no grupo de dados limpos segue até o momento em que o grupo inicial, livre de *outliers*, contenha todos os dados do conjunto bruto.

Os autores do método adotam como medida de distância, a distância de Mahalanobis ao centro dos dados, definida abaixo para a i-ésima observação de uma amostra univariada de tamanho n,

$$D_i^2 = \left(Y_i - \bar{Y}\right)^2 / S^2, \qquad i = 1, 2, ...n \tag{2.1}$$

em que $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ e $S^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n-1}$.

Em uma amostra multivariada em que os dados estão em uma matriz $\mathbf{Y}_{(n \times p)}$, sendo n o número de observações e p o número de variáveis, tem-se que a i-ésima observação tomada para a j-ésima variável seja repesentada pelo elemento y_{ij} da matriz. O vetor que representa a i-ésima observação multivariada é definido como $\mathbf{y}_i = (y_{i1}, ..., y_{ip})^T$, em que y_{ij} representa o valor referente ao indivíduo i na j-ésima variável. Dessa forma, analogamente para a i-ésima observação de uma amostra multivariada, tem-se que a distância de Mahalanobis à média é dada por

$$D_i^2 = \left(\mathbf{y}_i - \bar{\mathbf{y}}\right)^T \mathbf{S}_y^{-1} \left(\mathbf{y}_i - \bar{\mathbf{y}}\right), \qquad i = 1, 2, ...n \tag{2.2}$$

em que $\bar{\mathbf{y}} = \mathbf{Y}^T \mathbf{j}/n$, $\mathbf{j}$ é um vetor de dimensão $(n \times 1)$ de valores iguais a 1, e $\mathbf{S}_y = \frac{\sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T}{n-1}$.

Ao fim do processo tem-se o conjunto de dados ordenado com base nas distâncias das observações inseridas a cada passo do método. O tamanho do grupo inicial é denotado por m_0 e após a inserção da primeira observação, $m = m_0 + 1, ..., n$.

A seção seguinte aborda a distância de Mahalanobis como uma medida estratégica para avaliação da hipótese de normalidade multivariada e detecção de *outliers* multivariados com o uso do QQ plot.

2.1 QQ plot das Distâncias de Mahalanobis

No planejamento das análises estatísticas, muitas vezes, o procedimento para avaliação da normalidade univariada ou multivariada dos dados se faz necessário. Uma ferramenta útil para esse fim

é o gráfico QQ plot, em que os quantis amostrais da distância de Mahalanobis são plotados contra os quantis de uma distribuição χ^2_p. Suspeita-se da normalidade multivariada quando os pontos desviam de um comportamento linear, pontos estes que no princípio do estudo serão avaliados como potenciais *outliers*. É importante ressaltar que o alinhamento dos pontos no QQ plot não garantem a normalidade multivariada.

Como exemplo de aplicação, foram tomadas observações relativas a recordes de corrida feminina. Os dados foram obtidos de Johnson e Wichern (1997) e do manual preparado para as olimpíadas de 1984 em Los Angeles. Os dados são referentes a um interessante período da história do atletismo feminino em que havia, especialmente nos países comunistas, o tratamento de tais atletas com hormônios masculinos. Um aspecto da análise seria observar se os dados evidenciam tal fato - em que se observaria quais países são *outliers*.

Os valores observados são recordes atléticos de mulheres de 55 países, nas seguintes provas de distância: 100 metros em segundos (Y_1), 200 metros em segundos (Y_2), 400 metros em segundos (Y_3), 800 metros em minutos (Y_4), 1500 metros em minutos (Y_5), 3000 metros em minutos (Y_6) e maratona em minutos (Y_7). As Tabelas 2.1 e 2.2 trazem os recordes das provas de distância.

A contrução do QQ plot segue os seguintes passos:

1. Para cada ponto $\mathbf{y}_i$ é calculada a distância $D^2_i = (\mathbf{y}_i - \bar{\mathbf{y}})^T \mathbf{S}_y^{-1} (\mathbf{y}_i - \bar{\mathbf{y}})$, para $i = 1, .., n$;
2. As distâncias obtidas são ordenadas de forma que $D^2_{(1)} \leq D^2_{(2)} \leq ... \leq D^2_{(n)}$;
3. São construídos os níveis de probabilidade através da função empírica $p_{(i)} = \frac{i-0,5}{n}$, em que i representa o número de observações menores ou iguais a $D^2_{(i)}$;
4. É encontrado o quantil $q_{p,p_{(i)}}$ da distribuição qui-quadrado, tal que: $P\left[\chi^2_p \leq q_{p,p_{(i)}}\right] = p_{(i)}$;
5. O gráfico QQ plot mostrará os pares $(D^2_{(i)}, q_{p,p_{(i)}})$, para $i = 1, 2, ..., n$.

Nas Tabelas 2.3 e 2.4 tem-se os valores dos componentes $D^2_{(i)}$, $p_{(i)}$ e $q_{(p,p_{(i)})}$ necessários à construção do QQ plot para os dados dos recordes de corrida feminina.

Como resultado tem-se o gráfico da Figura 2.1, pela qual pode-se concluir pela distância das observações à reta de valores esperados sob normalidade, que há um indício da ausência de normalidade multivariada; e ainda, observa-se que os pontos 53, 54 e 55 relativos respectivamente às

Tabela 2.1: Dados de recordes de corrida feminina

Observação	Países	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
1	Argentina	11,61	22,94	54,5	2,15	4,43	9,79	178,52
2	Austrália	11,20	22,35	51,08	1,98	4,13	9,08	152,37
3	Austria	11,43	23,09	50,62	1,99	4,22	9,34	159,37
4	Bélgica	11,41	23,04	52,00	2,00	4,14	8,88	157,85
5	Bermuda	11,46	23,05	53,3	2,16	4,58	9,81	169,98
6	Brasil	11,31	23,17	52,8	2,10	4,49	9,77	168,75
7	Burma	12,14	24,47	55,00	2,18	4,45	9,51	191,02
8	Canadá	11,00	22,25	50,06	2,00	4,06	8,81	149,45
9	Chile	12,00	24,52	54,90	2,05	4,23	9,37	171,38
10	China	11,95	24,41	54,97	2,08	4,33	9,31	168,48
11	Colombia	11,60	24,00	53,26	2,11	4,35	9,46	165,42
12	Ilhas Cook	12,90	27,10	60,40	2,30	4,84	11,10	233,22
13	Costa Rica	11,96	24,60	58,25	2,21	4,68	10,43	171,80
14	República Tcheca	11,09	21,97	47,99	1,89	4,14	8,92	158,85
15	Dinamarca	11,42	23,52	53,60	2,03	4,18	8,71	151,75
16	Dominica	11,79	24,05	56,05	2,24	4,74	9,89	203,88
17	Finlândia	11,13	22,39	50,14	2,03	4,10	8,92	154,23
18	França	11,15	22,59	51,73	2,00	4,14	8,98	155,27
19	Alemanha Oriental	10,81	21,71	48,16	1,93	3,96	8,75	157,68
20	Alemanha Ocidental	11,01	22,39	49,75	1,95	4,03	8,59	148,53
21	Grã-Betanha	11,00	22,13	50,46	1,98	4,03	8,62	149,72
22	Grécia	11,79	24,08	54,93	2,07	4,35	9,87	182,20
23	Guatemala	11,84	24,54	56,09	2,28	4,86	10,54	215,08
24	Hungria	11,45	23,06	51,50	2,01	4,14	8,98	156,37
25	Índia	11,95	24,28	53,60	2,10	4,32	9,98	188,03
26	Indonésia	11,85	24,24	55,34	2,22	4,61	10,02	201,28
27	Irlanda	11,43	23,51	53,24	2,05	4,11	8,89	149,38
28	Israel	11,45	23,57	54,90	2,10	4,25	9,37	160,48

Tabela 2.2: Dados de recordes de corrida feminina (continuação da Tabela 2.1)

Observação	Países	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
29	Itália	11,29	23,00	52,01	1,96	3,98	8,63	151,82
30	Japão	11,73	24,00	53,73	2,09	4,35	9,20	150,50
31	Quênia	11,73	23,88	52,70	2,00	4,15	9,20	181,05
32	Coréia	11,96	24,49	55,70	2,15	4,42	9,62	164,65
33	Coréia do Norte	12,25	25,78	51,20	1,97	4,25	9,35	179,17
34	Luxemburgo	12,03	24,96	56,10	2,07	4,38	9,64	174,68
35	Malásia	12,23	24,21	55,09	2,19	4,69	10,46	182,17
36	Ilhas Maurício	11,76	25,08	58,10	2,27	4,79	10,9	261,13
37	México	11,89	23,62	53,76	2,04	4,25	9,59	158,53
38	Holanda	11,25	22,81	52,38	1,99	4,06	9,01	152,48
39	Nova Zelândia	11,55	23,13	51,60	2,02	4,18	8,76	145,48
40	Noruega	11,58	23,31	53,12	2,03	4,01	8,53	145,48
41	Papua Nova Guine	12,25	25,07	56,96	2,24	4,84	10,69	233,00
42	Filipinas	11,76	23,54	54,60	2,19	4,60	10,16	200,37
43	Polônia	11,13	22,21	49,29	1,95	3,99	8,97	160,82
44	Portugal	11,81	24,22	54,30	2,09	4,16	8,84	151,20
45	Romênia	11,44	23,46	51,20	1,92	3,96	8,53	165,45
46	Cingapura	12,30	25,00	55,08	2,12	4,52	9,94	182,77
47	Espanha	11,80	23,98	53,59	2,05	4,14	9,02	162,60
48	Suécia	11,16	22,82	51,79	2,02	4,12	8,84	154,48
49	Suiça	11,45	23,31	53,11	2,02	4,07	8,77	153,42
50	Taiwan	11,22	22,62	52,50	2,10	4,38	9,63	177,87
51	Tailândia	11,75	24,46	55,80	2,20	4,72	10,28	168,45
52	Turquia	11,98	24,44	56,45	2,15	4,37	9,38	201,08
53	EUA	10,79	21,83	50,62	1,96	3,95	8,50	142,72
54	USSR	11,06	22,19	49,19	1,89	3,87	8,45	151,22
55	Samoa Ocidental	12,74	25,85	58,73	2,33	5,81	13,04	306,00

Tabela 2.3: Valores dos componentes necessários para construção do QQ plot

Ordem	Países	$D^2_{(i)}$	$p_{(i)}$	$q_{7,p(i)}$
1	Hungria	0,84	0,01	1,20
2	Bélgica	1,41	0,03	1,74
3	Suiça	1,72	0,05	2,09
4	Suécia	1,77	0,06	2,37
5	França	1,84	0,08	2,62
6	Espanha	2,00	0,10	2,83
7	Irlanda	2,01	0,12	3,03
8	Grã-Betanha	2,24	0,14	3,22
9	Corea	2,82	0,15	3,40
10	Austrália	2,85	0,17	3,57
11	Itália	2,90	0,19	3,74
12	China	3,00	0,21	3,90
13	Taiwan	3,12	0,23	4,06
14	Alemanha Ocidental	3,22	0,25	4,22
15	Canadá	3,29	0,26	4,37
16	Holanda	3,46	0,28	4,52
17	Áustria	3,52	0,30	4,67
18	Quênia	3,57	0,32	4,82
19	Chile	3,70	0,34	4,97
20	Cingapura	4,18	0,35	5,12
21	Noruega	4,26	0,37	5,27
22	USRR	4,28	0,39	5,42
23	Portugal	4,30	0,41	5,57
24	Israel	4,33	0,43	5,72
25	Indonésia	4,42	0,45	5,87
26	Colômbia	4,47	0,46	6,03
27	EUA	4,54	0,48	6,19
28	Japão	4,70	0,50	6,35

Tabela 2.4: Valores dos componentes necessários para construção do QQ plot (continuação da Tabela 2.3)

Ordem	Países	$D^2_{(i)}$	$p_{(i)}$	$q_{7,p(i)}$
29	Papua Nova Guine	4,73	0,52	6,51
30	Filipinas	4,83	0,54	6,67
31	Nova Zelândia	4,91	0,55	6,84
32	Brasil	5,03	0,57	7,02
33	Grécia	5,07	0,59	7,19
34	Finlândia	5,13	0,61	7,37
35	Bermudas	5,62	0,63	7,56
36	Polônia	6,49	0,65	7,76
37	Romênia	6,82	0,66	7,96
38	Alemanha Oriental	6,96	0,68	8,17
39	Dinamarca	7,13	0,70	8,38
40	Luxemburgo	7,25	0,72	8,61
41	Argentina	7,68	0,74	8,85
42	Guatemala	8,19	0,75	9,10
43	México	8,66	0,77	9,37
44	República Tcheca	9,60	0,79	9,65
45	Birmânia	10,07	0,81	9,96
46	Índia	10,39	0,83	10,29
47	Turquia	10,52	0,85	10,65
48	Dominica	10,96	0,86	11,05
49	Malásia	11,46	0,88	11,50
50	Tailândia	12,78	0,9	12,02
51	Costa Rica	13,23	0,92	12,62
52	Ilhas Cook	14,00	0,94	13,37
53	Ilhas Maurício	25,83	0,95	14,34
54	Corea do Norte	28,20	0,97	15,77
55	Samoa Ocidental	37,68	0,99	18,73

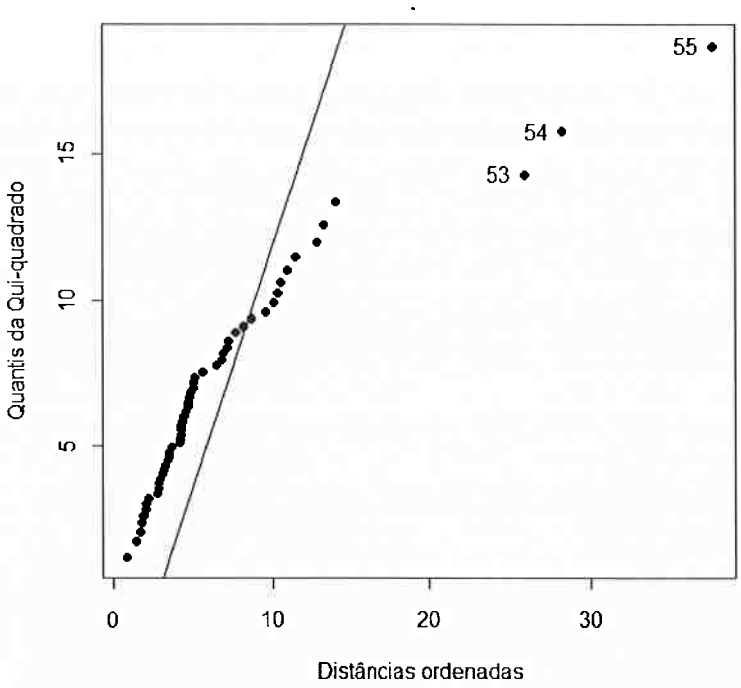

Figura 2.1: QQ plot das distâncias de Mahalanobis para os recordes de corrida feminina

Ilhas Maurício, Coréia do Norte e Samoa Ocidental, se comportam como *outliers* frente às distâncias obtidas.

2.2 Boxplot Bivariado

O *grupo limpo*, ou conjunto livre de *outliers*, de tamanho m_0, é muito importante para o método *forward search*, pois a partir dele resultará a ordenação dos dados e o processo de crescimento deste conjunto. Para não existirem valores que possam distorcer as estimativas do cálculo da distância de Mahalanobis tem-se a necessidade de uma análise descritiva preliminar que garanta que o grupo seja livre de *outliers*; o *boxplot* bivariado tem esta função.

Similar ao objetivo do uso do *boxplot* univariado, em que são produzidas informações a respeito

da localização e espalhamento dos dados, e ainda destacados potenciais *outliers*, tem-se a extensão para o caso bivariado.

O *boxplot* bivariado reflete a extensão bidimensional da distância interquartil de um *boxplot* univariado. Em uma dimensão, tem-se a amplitude da caixa que contém 50% dos valores centrais. Em duas dimensões procura-se por uma região similar, centrada em um estimador robusto de posição, que é definido como sendo a mediana das variáveis, representado por $\tilde{y}_j$ para $j = 1, ..., p$. Os quantis do *boxplot*, representado por linhas horizontais, agora passam a ser representados por elipses. O cálculo das elipses são realizados a partir da matriz de covariância $\mathbf{S}_y^*$, na qual os elementos são dados por

$$s_{jk}^* = \frac{\sum_{i=1}^{n} (y_{ij} - \tilde{y}_j)(y_{ik} - \tilde{y}_k)}{n-1}, \qquad i = 1, 2, ...n \qquad e \qquad j = 1, ..., p, k = 1, ..., p \tag{2.3}$$

em que s_{jk}^* representa a medida de covariância entre as variáveis Y_j e Y_k.

Segundo Zani et al. (1998) o algoritmo para construção do *boxplot* bivariado tem os passos definidos como: especificação da região interna, definição de um centróide robusto e construção de uma região externa. A especificação da região interna é dada pelo cálculo da elipse baseada na matriz $\mathbf{S}_y^*$, como mostra (2.4), em que o contorno para a região interna é dado pelo quantil da distribuição χ^2, em que é esperado 50% das observações pertencerem a esta região.

$$(\mathbf{y}_i - \tilde{\mathbf{y}})^T \mathbf{S}_y^* (\mathbf{y}_i - \tilde{\mathbf{y}}) \leq \chi_p^2(0,5). \tag{2.4}$$

Para construção da região externa os autores sugerem que esta, sob a hipótese de normalidade, possa ser interpretada como um contorno de probabilidade ao nível $(1 - \alpha)$, em que α seja próximo a 0,01. Logo, os pontos encontrados além deste contorno externo são qualificados como *outliers* bivariados.

Aparte o uso da mediana como medida de posição, a teoria é inteiramente baseada na distribuição normal.

A construção de *boxplots* bivariados para cada par de variáveis possibilita definir o conjunto de dados limpos bivariados.

A Figura 2.2 mostra os gráficos de dispersão relativos às variáveis do exemplo. Na Figura 2.3 os mesmos dados são avaliados com o uso dos *boxplots* bivariados.

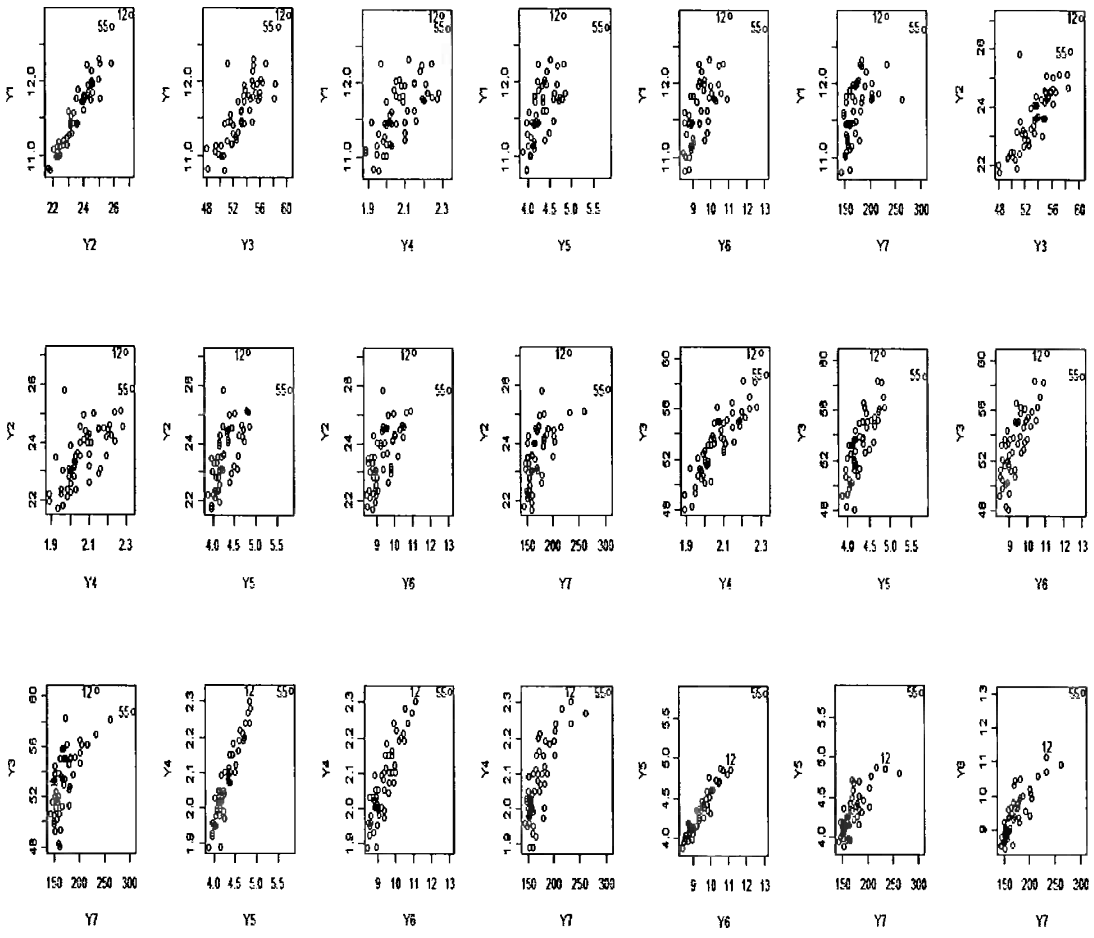

Figura 2.2: Gráficos de dispersão para os dados de recordes de corrida feminina

Com base na análise gráfica destes dados, tem-se que para os valores dos tempos de corrida dentro de cada modalidade, destacam-se as observações 12 e 55 como possíveis *outliers* bivariados; tais observações são referentes às Ilhas Cook e Samoa Ocidental respectivamente, estas observações

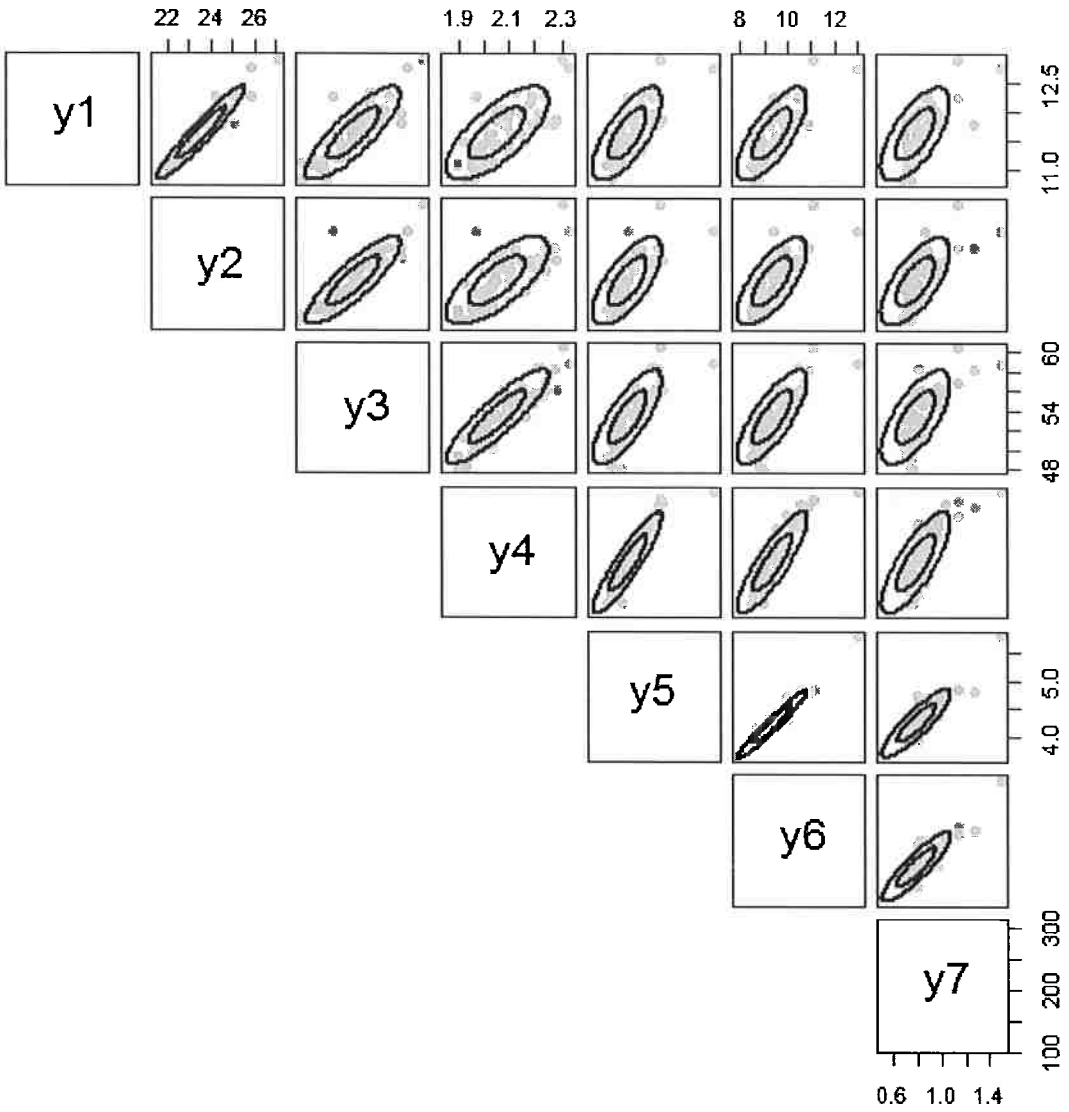

Figura 2.3: *Boxplots* bivariados, com $\alpha = 0,01$, para os dados de recordes de corrida feminina

são referentes aos dados do conjunto bruto exibido nas Tabelas 2.1 e 2.2. A Figura 2.2 traz o posicionamento das observações 12 e 55 nos cruzamentos bivariados e a Figura 2.3 exibe a matriz de *boxplot* bivariados, em que é possível notar a localização das observações 12 e 55 fora do contorno externo em grande parte dos cruzamentos.

O *grupo limpo* é encontrado a partir da intersecção de todos os $p(p-1)/2$ gráficos de dispersão como na Figura 2.2. Este grupo excluirá qualquer observação que seja *outlier* em uma ou duas dimensões, apesar de não excluir as que sejam *outlier* em três ou mais. Segundo o algoritmo adotado pelo autores do método, caso haja conhecimento prévio dos dados, pode-se selecionar as observações que constituirão o *grupo limpo*, uma vez que estas pertençam à região interna dos *boxplots* bivariados. Caso não exista tal conhecimento, as observações pertencentes à região interna das elipses dos *boxplots* bivariados podem ser tomadas de forma aleatória, e o tamanho m_0 é de no mínimo $p+1$ observações.

Para o exemplo da corrida feminina, em que não se tem informações a priori do conjunto de dados, a análise do *boxplot* bivariado leva à seguinte configuração do *grupo limpo* com $p+1$ países, ou seja, oito países: Bélgica, França, Hungria, Irlanda, Itália, Holanda, Espanha e Suíça. A Tabela 2.5 mostra os valores observados para os países integrantes do grupo, seguida da Tabela 2.6 com as estatísticas descritivas observadas.

Tabela 2.5: Valores observados para os países do *grupo limpo*

Países	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
Bélgica	11,41	23,04	52,00	2,00	4,14	8,88	157,85
França	11,15	22,59	51,73	2,00	4,14	8,98	155,27
Hungria	11,45	23,06	51,50	2,01	4,14	8,98	156,37
Irlanda	11,43	23,51	53,24	2,05	4,11	8,89	149,38
Itália	11,29	23,00	52,01	1,96	3,98	8,63	151,82
Holanda	11,25	22,81	52,38	1,99	4,06	9,01	152,48
Espanha	11,80	23,98	53,59	2,05	4,14	9,02	162,60
Suíça	11,45	23,31	53,11	2,02	4,07	8,77	153,42

Com base nas Tabelas 2.5 e 2.6 tem-se que os valores do *grupo limpo* são pouco dispersos em cada variável, exceto para as variáveis Y_3 e Y_7, prova de 400 metros e maratona, respectivamente, que possuem um maior distanciamento entre o primeiro e o terceiro quartil. Com base na Figura 2.4, em que são exibidos os *boxplots* das variáveis para o *grupo limpo*, observa-se que em uma análise univariada o valor para Espanha na variável Y_1, prova dos 100 metros, é *outlier* em relação aos valores

Tabela 2.6: Estatísticas descritivas para o *grupo limpo* dos dados de corrida

Variáveis	Mínimo	Quartil 1	Mediana	Média	Quartil 3	IQ	Máximo
Y_1	11,15	11,28	11,42	11,40	11,45	0,17	11,80
Y_2	22,59	22,95	23,05	23,16	23,36	0,41	23,98
Y_3	51,50	51,93	52,20	52,45	53,14	1,21	53,59
Y_4	1,96	1,99	2,00	2,01	2,02	0,03	2,05
Y_5	3,98	4,06	4,12	4,01	4,14	0,07	4,14
Y_6	8,63	8,85	8,94	8,89	8,98	0,13	9,02
Y_7	149,40	152,30	154,30	154,90	156,70	4,40	162,60

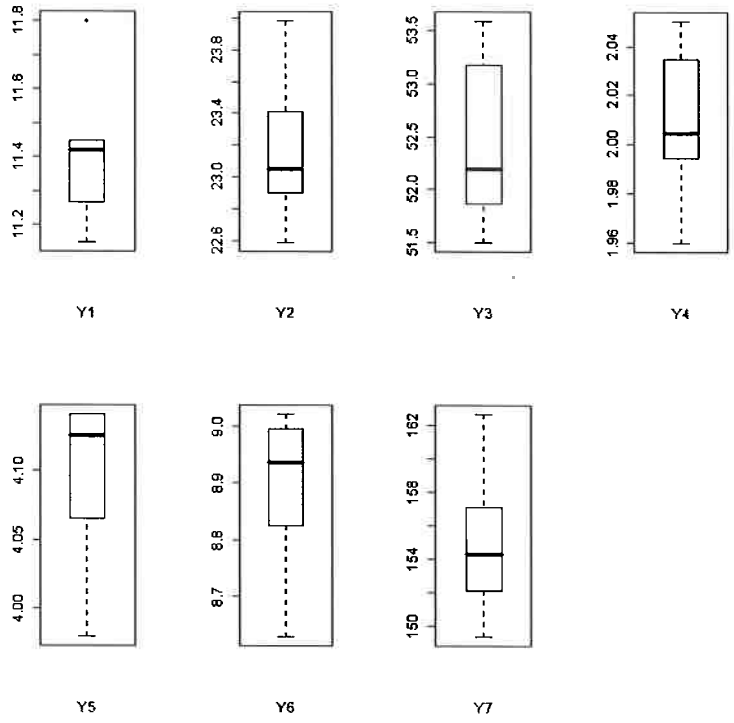

Figura 2.4: *Boxplots* das variáveis da prova de corrida para o *grupo limpo*

dos outros países no *grupo limpo*, mas se comparado ao conjunto de dados total tal comportamento é atribuído às Ilhas Cook. Dessa forma o grupo é livre de *outliers*.

2.3 Monitoramento do *Forward Search*

Durante o processo *forward search*, o vetor de médias e a matriz de covariância sofrem alterações com o crescimento de m, logo os respectivos estimadores da média e matriz de covariância são denotados por $\hat{\mu}_m$ e $\hat{\Sigma}_m$. A cada passo no *forward search* calculam-se todas as distâncias de Mahalanobis D^2_{im}, $i = 1, ..., n$ e $m_0 \leq m \leq n$.

Se o interesse está na detecção de *outliers* é conveniente monitorar a distância mínima de Mahalanobis, definida como os menores valores de distância ao centro do *grupo limpo* assumida pelas observações candidatas a entrar no grupo. Um *outlier* prestes a entrar causará um aumento na distância mínima, o que corresponderá a um pico no gráfico de distâncias mínimas.

A Figura 2.5 mostra que grande parte das distâncias mínimas de Mahalanobis oscilam entre 4 e 6, mas ao fim do processo há um grande salto referente à entrada de Samoa Ocidental, no passo em que $m = n = 55$, precedido dos saltos nos passos 53 e 54, relativos respectivamente às Ilhas Maurício e à Corea do Norte, ou seja, são os mesmos países detectados como potenciais *outliers* no QQ *plot* da Figura 2.1, onde todas as observações foram usadas para estimação dos parâmetros do vetor de médias e matriz de covariância da medida de Mahalanobis. A Tabela 2.7 exibe as distâncias de Mahalanobis no *forward search* para os passos seguintes à criação do grupo inicial, em que é exibido o país a entrar em cada subconjunto de tamanho m, juntamente com a distância de Mahalanobis precedente à sua entrada.

Tem-se que a partir das estimativas de μ_8 e Σ_8 do *grupo limpo*, foi aplicado o cálculo da distância de Mahalanobis ao centro do *grupo limpo* a todo o conjunto de dados. Desta forma, no passo seguinte tem-se o grupo de tamanho $m = 9$ relativo à entrada da Austrália, que apresentou a menor distância de Mahalanobis entre os países que não faziam parte do grupo inicial. A partir desta inserção são calculadas as estimativas de μ_9 e Σ_9 usadas no cálculo das distâncias para todo o conjunto, levando a Portugal como o novo integrante relativo ao grupo com $m = 10$. Tal processo se repetirá até $m = n = 55$, que será relativo à inclusão de Samoa Ocidental.

A Tabela 2.7 mostra a inclusão dos países a cada passo do *forward search*. É interessante e informativo observar o comportamento das distâncias D^2_{im} de cada unidade durante o progresso do método. Para tal, um gráfico que exiba estas distâncias é construído, porém são tomadas as

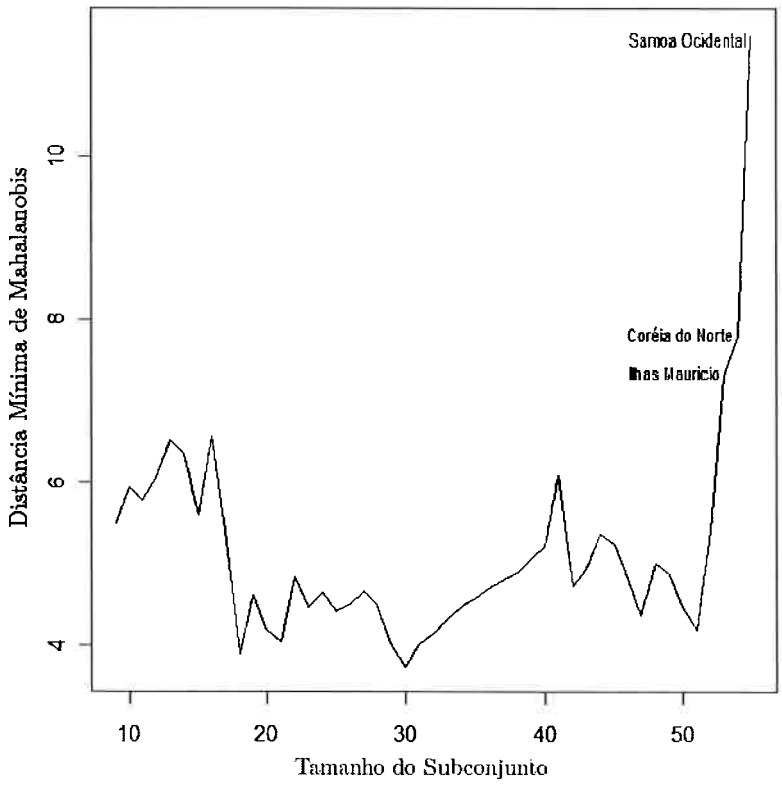

Figura 2.5: Gráfico *forward search* para as distâncias de Mahalanobis dos dados de corrida

distâncias escalonadas de Mahalanobis, em que a medida de distância é ponderada pela razão entre as estimativas de variância durante o processo e a estimativa de variância total do conjunto de dados. O cálculo tem a seguinte forma

$$D_{im}^2 \times \left(\frac{\left| \hat{\boldsymbol{\Sigma}}_m \right|}{\left| \hat{\boldsymbol{\Sigma}}_n \right|} \right)^{\frac{1}{2p}}, \tag{2.5}$$

em que $\hat{\boldsymbol{\Sigma}}_m$ e $\hat{\boldsymbol{\Sigma}}_n$ são respectivamente os estimadores das matrizes de covariância relativas à cada incremento no grupo inicial e a estimativa relativa ao conjunto total de dados. Em consequência, tem-se que nas etapas finais do método as observações receberam maior peso, observações estas

Tabela 2.7: Inclusão dos países a cada passo do *forward search*

m	País	D^2_{im}	m	País	D^2_{im}
9	Austrália	30,31	33	Áustria	18,78
10	Portugal	35,27	34	Quênia	19,74
11	Alemanha Ocidental	48,90	35	Romênia	19,83
12	Suécia	24,26	36	Índia	22,50
13	Nova Zelândia	42,52	37	Tailândia	25,09
14	Israel	40,32	38	Costa Rica	22,54
15	USSR	31,37	39	República Tcheca	26,96
16	Coréia	43,14	40	México	27,28
17	Grã-Betanha	29,27	41	Malásia	38,37
18	EUA	15,18	42	Bermuda	28,02
19	Chile	21,26	43	Taiwan	22,52
20	China	17,52	44	Argentina	31,99
21	Canadá	16,40	45	Filipinas	28,41
22	Grécia	23,56	46	Indonésia	23,01
23	Singapura	19,92	47	Burma	19,13
24	Noruega	21,56	48	Guatemala	25,84
25	Colombia	19,52	49	Papua Nova Guine	24,12
26	Brasil	20,28	50	Turquia	20,09
27	Japão	21,81	51	Dominica	17,96
28	Alemanha Oriental	20,33	52	Ilhas Cook	28,84
29	Finlândia	16,25	53	Ilhas Maurício	53,56
30	Polônia	13,91	54	Coréia do Norte	64,21
31	Dinamarca	16,05	55	Samoa Ocidental	132,70
32	Luxemburgo	17,13			

correspondentes aos *outliers*.

A Figura 2.6, como a Figura 2.5, leva à conclusão de que as observações referentes a Ilhas Maurício, Coréia do Norte e Samoa Ocidental, representados pelas linhas em destaque 36, 33 e 55 respectivamente são observações com comportamento atípico. Quando comparado o histórico de suas distâncias escalonadas de Mahalanobis com as do restante do conjunto de dados, elas estão distantes do padrão em todas as etapas de crescimento do grupo inicial.

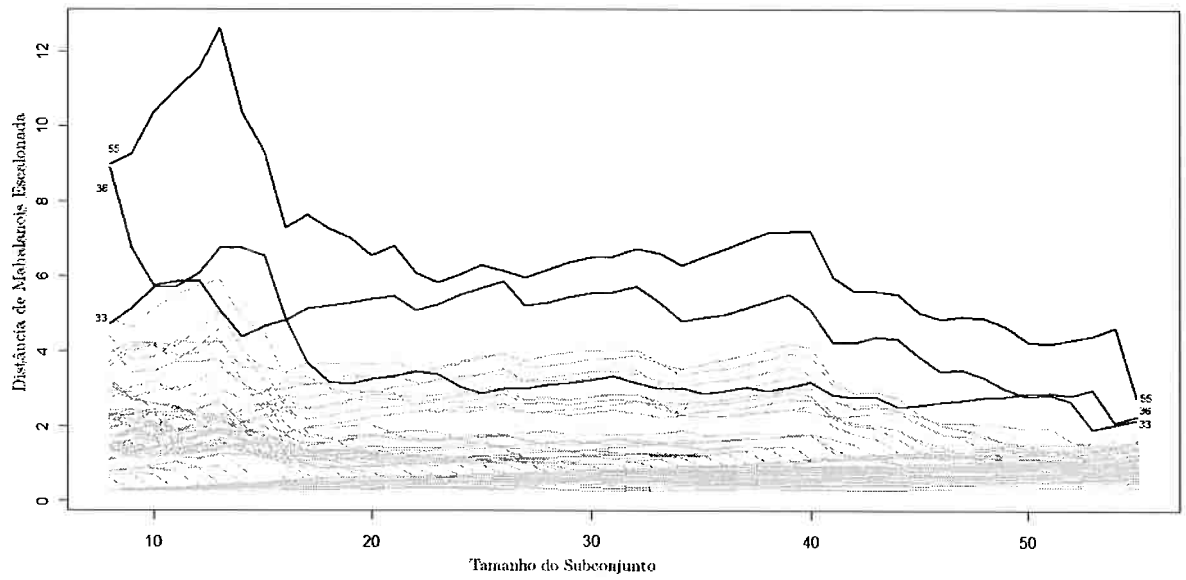

Figura 2.6: Gráfico *forward search* das distâncias de Mahalanobis para os dados de corrida

2.4 Monitoramento da Matriz de Covariância

Enquanto as observações são sequencialmente selecionadas de acordo com as menores distâncias de Mahalanobis, o estimador da matriz de covariância Σ não se matém constante durante o processo de *forward search*. Com fim de observar tal fato, faz-se uso do gráfico da variância generalizada, dada pelo determinante do estimador da matriz de covariância $\hat{\Sigma}_m$ a cada passo do *forward search*.

A Figura 2.7 mostra que a evolução da variância generalizada do grupo inicial até os grupos próximos aos de tamanho 50 foi aproximadamente constante, havendo grandes saltos para os grupos maiores que 50, fato decorrente da entrada dos países Ilhas Maurício, Coréia do Norte e Samoa Ocidental, mencionados na construção da Figura 2.5. Este comportamento leva a concluir que estes países causam uma grande inflação na matriz de covariância quando são incluídos no cálculo das estimativas.

Ainda, com o intuito de estudar o efeito dos subgrupos na estimação da matriz de covariância é

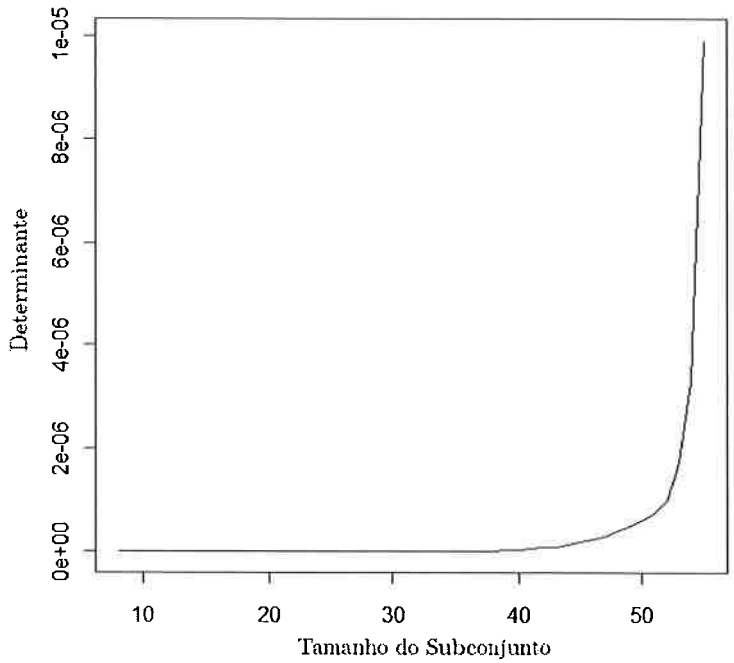

Figura 2.7: Gráfico *forward search* para as estimativas da variância generalizada dos dados de corrida

conveniente analisar os elementos da matriz através de um gráfico que mostre o comportamento das covariâncias e variâncias ao longo do processo, porém deve-se atentar para a escala de variabilidade das variáveis envolvidas; na presença de variáveis com grande variância é preferível tomar a análise com base nas variáveis padronizadas. Em particular, na aplicação dos dados referentes aos recordes de corrida feminina, tem-se que a variável Y_7 referente à prova de maratona, tem a variância discrepante em relação às demais. Dado este fato, é conveniente avaliar o comportamento dos elementos da matriz de covariância resultante das variáveis padronizadas, ou seja, da matriz de correlação.

A Figura 2.8 exibe, à esquerda, o gráfico referente ao comportamento dos elementos da matriz de covariância em que a variância da variável Y_7 mascara o efeito das demais variáveis ao passo que no gráfico à direita tem-se a mesma representação mas com base nas variáveis padronizadas.

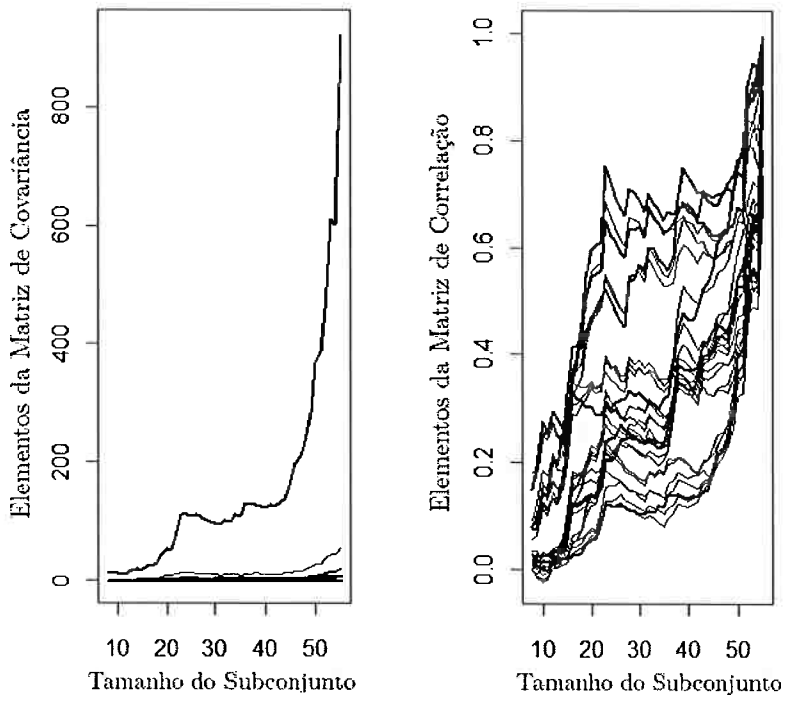

Figura 2.8: Gráfico *forward search* dos elementos das matrizes de covariância (à esquerda) e correlação (à direita) dos dados de corrida

Tem-se como resultado a melhor visualização e representação, do comportamento dos elementos da matriz de covariância que agora passam a representar correlações. Ainda é possível observar neste gráfico padrões de comportamento agregados ao tamanho dos subgrupos, em que é notório para os agrupamentos com tamanho inferior ou igual a 15 estarem agregados a medidas de correlação menores ou iguais a 0,3 ao passo que para os agrupamentos superiores ou iguais a 50 os elementos da matriz de correlação tomam valores maiores ou iguais a 0,5.

Capítulo 3

Regressão Linear

Em estudos observacionais tem-se o interesse de especular sobre as possíveis fontes de variação que produziram os valores observados. Segundo Bussab e Moretim (2006) uma das preocupações estatísticas ao analisar este tipo de dado é a de criar modelos que explicitem a estrutura do fenômeno em observação. Tal preocupação é abordada no ajuste de modelos de regressão, em que se deseja 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 notação matricial a equação de regressão é expressa em (3.1) em que $\mathbf{y}$ representa o vetor de observações $(n \times 1)$, $\mathbf{X}$ é uma matriz $(n \times k)$ de variáveis regressoras mais o intercepto, $\boldsymbol{\beta}$ é um vetor $(k \times 1)$ de coeficientes da regressão e $\boldsymbol{\epsilon}$ é um vetor $(n \times 1)$ de erros aleatórios

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3.1}$$

Em modelos de regressão linear o processo de seleção de variáveis explicativas e estimação dos parâmetros é muito sensível à presença de *outliers*. Tais valores podem vir a camuflar a significância estatística de alguma variável no modelo e principalmente violar as suposições de normalidade 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 *outliers* já são frequentemente utilizadas, como análise de resíduo, identificação de pontos de alavanca e análise

das distâncias de Cook (1977). O método *forward search* 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.

Seguindo o algoritmo apresentado no Capítulo 1, a aplicação do método *forward search* em modelos de regressão linear segue os mesmos passos descritos, ou seja, escolha do grupo de dados livre de *outliers*, *grupo limpo*, que ao longo do processo será acrescido de observações provenenientes do conjunto de dados brutos que o originou.

O diferencial do método, no contexto de modelo de regressão linear, é o fato de que 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 entrada das observações restantes também sofre uma modificação: enquanto que na análise descritiva a medida de distância das observações candidatas a entrar no *grupo limpo* era a distância de Mahalanobis ao centro do *grupo limpo*, agora a medida passa a ser o resíduo obtido a partir do modelo ajustado com o *grupo limpo*. Assim a observação com menor resíduo será incluída no grupo e a entrada da observação leva ao recálculo dos parâmetros do modelo proposto. O processo de entrada das observações no *grupo limpo* segue até o momento em que contenha todas as observações do conjunto de dados brutos que o originou.

Durante o processo, as estimativas dos parâmetros e os resíduos do modelo ajustado sofrem mudanças consideráveis, por isso há o monitoramento das alterações destas quantidades, bem como de várias outras estatísticas relacionadas ao ajuste do modelo. O método tem por objetivo produzir não apenas informação a respeito da detecção do *outlier*, mas também o efeito que cada observação tem no aspecto inferencial do modelo.

3.1 Descrição do Método

A maioria dos métodos para a detecção de *outliers* procura dividir os dados em duas partes, uma referente aos dados limpos e a outra aos *outliers*. 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) a deleção 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 *outliers*, referentes 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 *outliers* são excluídos de uma vez, mas a vasta combinação de casos a serem avaliados leva a uma tarefa onerosa de revisão dos ajustes.

Muitos métodos de detecção de *outliers* múltiplos utilizam técnicas robustas para a classificação das observações em dados limpos e *outliers*. 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, supondo-se erros normalmente distribuídos. Porém, difere significativamente do ajuste por mínimos quadrados quando os erros não são normalmente distribuídos ou quando os dados contêm *outliers*. Seu uso torna-se justificável porque, quanto maior o número de variáveis de um modelo, mais difícil a identificação de *outliers* 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, surge o conceito de ponto de ruptura.

Segundo Montgomery e Vining (2001), o ponto de ruptura em uma amostra finita é a menor fração de dados atípicos que poderia causar perturbação ao estimador. A fração mínima para o ponto de ruptura é de $\frac{1}{n}$, onde apenas com uma observação atípica o estimador do parâmetro já sofre viés dessa observação. O ponto de ruptura do estimador de mínimos quadrados é $\frac{1}{n}$. Um estimador robusto adequado será aquele que possuir o ponto de ruptura igual a 0,5, ou seja, o estimador que resiste a um percentual de até 50% dos dados *outliers*. Um valor de ruptura maior que 0,5 não faria sentido, pois, se mais da metade das observações fosse de valores discrepantes não seria mais possível diferenciar quais seriam os dados *outliers*.

Dentre os estimadores robustos tem-se o estimador LMS - *Least Median of Squares* (em português Mínima Mediana dos Quadrados) que fornece estimativas para os parâmetros que não são afetadas pela presença de observações *outliers*.

A seção seguinte tem como base o uso desse estimador robusto para encontrar o *grupo limpo*.

3.2 *Forward Search* na Regressão Linear

O algoritmo *forward search* 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 *outliers*. Rousseeuw (1984) propõe tomar de forma aleatória subconjuntos de tamanho k. Se $\binom{n}{k}$ for muito grande, Atkinson e Riani (2000) recomendam tomar um grande número de candidatos a *grupo limpo* de tamanho k, ajustar o modelo de regressão robusta aos grupos e tomar como o conjunto limpo de *outliers* 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 $\hat{\beta}^*_{m_0}$ e o estimador de mínimos quadrados ao fim do processo será $\hat{\beta}^*_n = \hat{\beta}$. Na ausência de *outliers* a seguinte relação ocorre

$$E\left(\hat{\beta}^*_{m_0}\right) = E\left(\hat{\beta}\right) = \beta, \tag{3.2}$$

ou seja, ambos os estimadores serão não viesados.

Após a seleção da observação seguinte com menor resíduo quadrático, o grupo passa a ter tamanho $m = m_0 + 1$, no crescendo de uma em uma unidade até n. O estimador *forward search*, $\hat{\beta}_{FS}$ é definido como a coleção de estimativas dos parâmetros produzidas a cada passo do processo

$$\hat{\beta}_{FS} = \left(\hat{\beta}^*_{m_0}, ..., \hat{\beta}^*_n\right). \tag{3.3}$$

O método evita a inclusão de *outliers* 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, como explicado na Seção 3.1, e ao mesmo tempo estimadores de mínimos quadrados. A introdução de observações atípicas é sinalizada por picos nas curvas que monitoram as estimativas dos parâmetros, como o teste t, ou qualquer outra estatística que se esteja monitorando. As curvas presentes nos gráficos do método *forward search* exibem o valor da estatística passo a passo no processo.

O método *forward search* 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 *outliers*. 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 quando m cresce de $m_0 + 1$ a n.

3.3 Distância de Cook

Segundo Atkinson e Riani (2000), a forma mais simples de obter informação sobre o efeito da i-ésima observação na estimativa do parâmetro é monitorar as mudanças individuais das estimativas de $\hat{\beta}*_m$. Os autores também acham informativo obter medidas gerais sobre mudanças no vetor $\hat{\beta}$. Isto é fornecido pela distância de Cook, que é derivada da região de confiança para β.

A região de confiança ao nível $100(1-\alpha)\%$ para o parâmetro β é dada pelos valores dos parâmetros de acordo com a expressão

$$\left(\hat{\beta} - \beta\right)^T X^T X \left(\hat{\beta} - \beta\right) \leq \left(ks^2\right) F_{k,v,\alpha}, \tag{3.4}$$

em que s^2 é uma estimativa de σ^2 e $F_{k,v,\alpha}$ é o (100α)-ésimo quantil da distribuição F com k e v graus de liberdade. Cook (1977) propôs a estatística

$$D_i = \left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T X^T X \left(\hat{\beta}_{(i)} - \hat{\beta}\right) / \left(ks^2\right) \tag{3.5}$$

para detecção de observações influentes, em que o subscrito (i) indica a estimativa do parâmetro sem a i-ésima observação. Grandes valores de D_i indicam observações que são influentes na inferência conjunta sobre todos os parâmetros lineares no modelo.

Uma interpretação para a medida é que ela mensura as mudanças na soma de quadrados na predição quando a observação i não é usada para estimar β.

Segundo os autores essa quantidade-diagnóstico mede o efeito da deleção de uma observação que pode estar mascarada por outras observações *outliers* presentes. O método *forward search* supera esse disfarce de *outliers*, com mudanças abruptas nos parâmetros estimados, indicando a entrada de

observação influente, que é detectada através do monitoramento da versão do *forward search* para distância de Cook D_m^*, chamada Distância de Cook modificada, que é dada por

$$D_m^* = \left(\hat{\beta}_{m-1}^* - \hat{\beta}_m^*\right)^T \left(\mathbf{X}_m^T \mathbf{X}_m\right) \left(\hat{\beta}_{m-1}^* - \hat{\beta}_m^*\right) / \left(k s_m^2\right), \qquad m = k+1, ..., n \tag{3.6}$$

em que $\mathbf{X}_m$ é uma matriz $m \times k$ que contém m linhas da matriz $\mathbf{X}$ e s_m^2 a estimativa de σ^2 referente às unidades que fazem parte do grupo de tamanho m.

Para o cálculo da estatística de Cook modificada, é exigida a medida de alavanca de cada unidade. Segundo Paula (2004) essa medida mensura a influência da i-ésima observação sobre o próprio valor ajustado. A medida de alavanca por si só é útil para detecção de observações influentes. Assim como para as demais estatísticas de monitoramento do ajuste, é construído o gráfico para as medidas de alavanca, em que os valores são dados por

$$h_{i,m} = \boldsymbol{x}_i^T \left(\mathbf{X}_m^T \mathbf{X}_m\right) \boldsymbol{x}_i, \qquad i = 1, ..., n \qquad e \qquad m = k, ..., n. \tag{3.7}$$

No início do processo quando se tem k observações, cada uma delas tem o valor de ponto alavanca igual a um, que decrescerá ao longo do processo.

3.4 Estatística t para os Parâmetros do Modelo

Segundo Atkinson e Riani (2000) o monitoramento da estatística t para cada coeficiente da regressão, no contexto do método *forward search*, é falho ao identificar observações influentes para a significância estatística de cada parâmetro da regressão. Esta falha é devida à ordenação dos dados durante o processo *forward search*, por este fato os autores introduziram o teste de *variável-adicionada*, como correção para análise da significância estatística dos parâmetros. O modelo para o procedimento do teste é exibido em (3.8)

$$E\left(\mathbf{Y}\right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}\gamma, \tag{3.8}$$

em que o vetor $\mathbf{w}$ de dimensão $n \times 1$ é referente à variável a ser testada, γ é um escalar, a matriz $\mathbf{X}$ de dimensão $n \times (k-1)$ é referente ao termo constante da regressão e o restante de variáveis

explicativas presentes no modelo. Ou seja, dada a estrutura de (3.8) a covariável a ser testada entra no modelo como uma variável adicionada. O teste da significância estatística da *variável-adicionada* $\mathbf{w}$ é baseado na verificação do parâmetro γ.

O estimador de mínimos quadrados $\hat{\gamma}$ é encontrado através das equações normais para o modelo particionado, dadas por

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}^T\mathbf{w}\hat{\gamma} = \mathbf{X}^T\mathbf{y} \tag{3.9}$$

e

$$\mathbf{w}^T\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{w}^T\mathbf{w}\hat{\gamma} = \mathbf{w}^T\mathbf{y}. \tag{3.10}$$

A partir de (3.9) tem-se:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{w}\hat{\gamma}. \tag{3.11}$$

Substituindo (3.11) em (3.10), tem-se:

$$\hat{\gamma} = \frac{\mathbf{w}^T\left(\mathbf{I} - \mathbf{H}\right)\mathbf{y}}{\mathbf{w}^T\left(\mathbf{I} - \mathbf{H}\right)\mathbf{w}}, \tag{3.12}$$

em que $\mathbf{I}$ é a matriz identidade de orden n e $\mathbf{H}$ a matriz $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ de dimensão $n \times n$.

Para o cálculo da estatística t é exigido o estimador da variância de $\hat{\gamma}$, que é dado por

$$s_w^2 = \frac{\mathbf{y}^T\mathbf{A}\mathbf{y} - (\mathbf{y}^T\mathbf{A}\mathbf{w})^2/(\mathbf{w}^T\mathbf{A}\mathbf{w})}{n - k - 1}, \tag{3.13}$$

em que $\mathbf{A} = (\mathbf{I} - \mathbf{H})$. A estatítica t para o teste de $\gamma = 0$ é dada por

$$t_w = \frac{\hat{\gamma}}{\sqrt{s_w^2/(\mathbf{w}^T\mathbf{A}\mathbf{w})}} \tag{3.14}$$

cuja distribuição, sob a hipótese nula, é t-Student, com $n - k$ graus de liberdade. A técnica é aplicada a cada variável explicativa presente no modelo. O resultado gráfico são linhas com os

valores da estatística t obtida a cada passo do método *forward search* para cada variável, juntamente com linhas horizontais, baseadas na distribuição t-Student, indicando o nível de significância.

3.5 Aplicação

Como exemplo de aplicação foi utilizado o conjunto de dados de Gray (1989) retirado de Paula (2004), referente a um estudo sobre o consumo de combustível anual por habitante de 48 estados norte-americanos. O objetivo do estudo foi tentar explicar o consumo de combustível, que é dado em número de galões consumidos ao ano, através das seguintes variáveis explicativas: taxa do combustível no estado (%), proporção de motoristas licenciados (%), renda *per-capita* e ajuda federal para as estradas (em dólares).

O modelo proposto no estudo é:

$$Consumo_i = \beta_0 + \beta_1 Taxa_i + \beta_2 Licenca_i + \beta_3 Renda_i + \beta_4 Estrada_i + \epsilon_i. \tag{3.15}$$

As Tabelas 3.1 e 3.2 mostram as observações do estudo e a Tabela 3.3 as estatísticas descritivas correspondentes.

Uma análise gráfica preliminar dos dados, contida nas Figuras 3.1 e 3.2, possibilita uma avaliação da forma e o grau de correlação existente entre a variável combustível e as covariáveis, que poderiam explicar seu comportamento, bem como aponta as observações discrepantes univariadas e bivariadas que poderão vir a ser detectadas como *outliers* no ajuste do modelo.

De acordo com a Figura 3.1, é observada maior evidência de relação linear da variável resposta com a covariável Licença; com as demais covariáveis tal comportamento é menos evidente. Também nesta figura são apontados alguns estados com comportamento discrepante em uma análise gráfica bivariada.

A Figura 3.2 exibe a matriz de *boxplot* bivariados. Os pontos posicionados além do contorno externo são classificados como *outliers* bivariados. Desta forma tem-se que as observações indicadas na Figura 3.1 no mínimo são *outliers* em uma abordagem bivariada, e consequentemente não farão parte do conjunto inicial livre de *outliers*.

Após a análise gráfica preliminar é iniciado o algoritmo *forward search*. O primeiro passo é a escolha do *grupo limpo*. Como apresentado na Seção 3.2, a escolha é realizada após o ajuste de subconjuntos de tamanho k, no caso 5, tomados de forma aleatória. Dado o grande número de

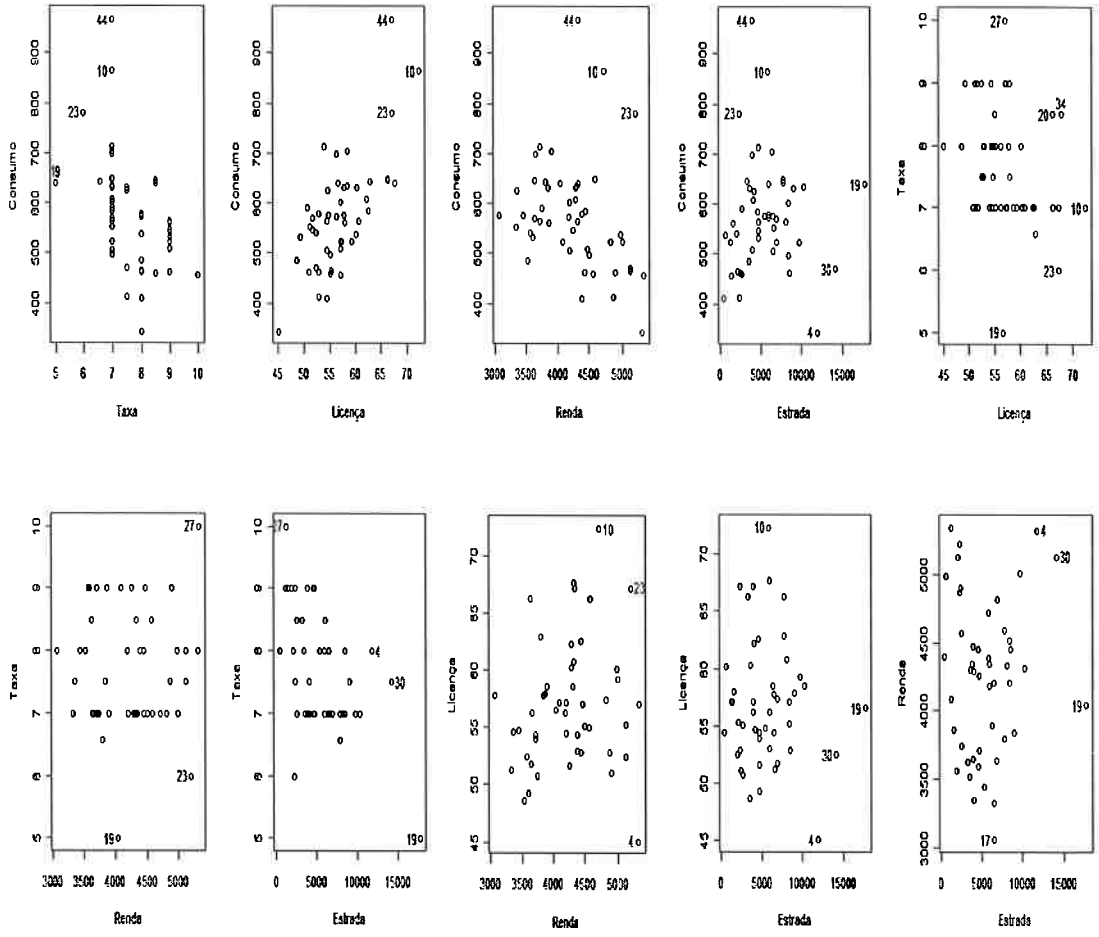

Figura 3.1: Gráficos de dispersão para os dados de combustível

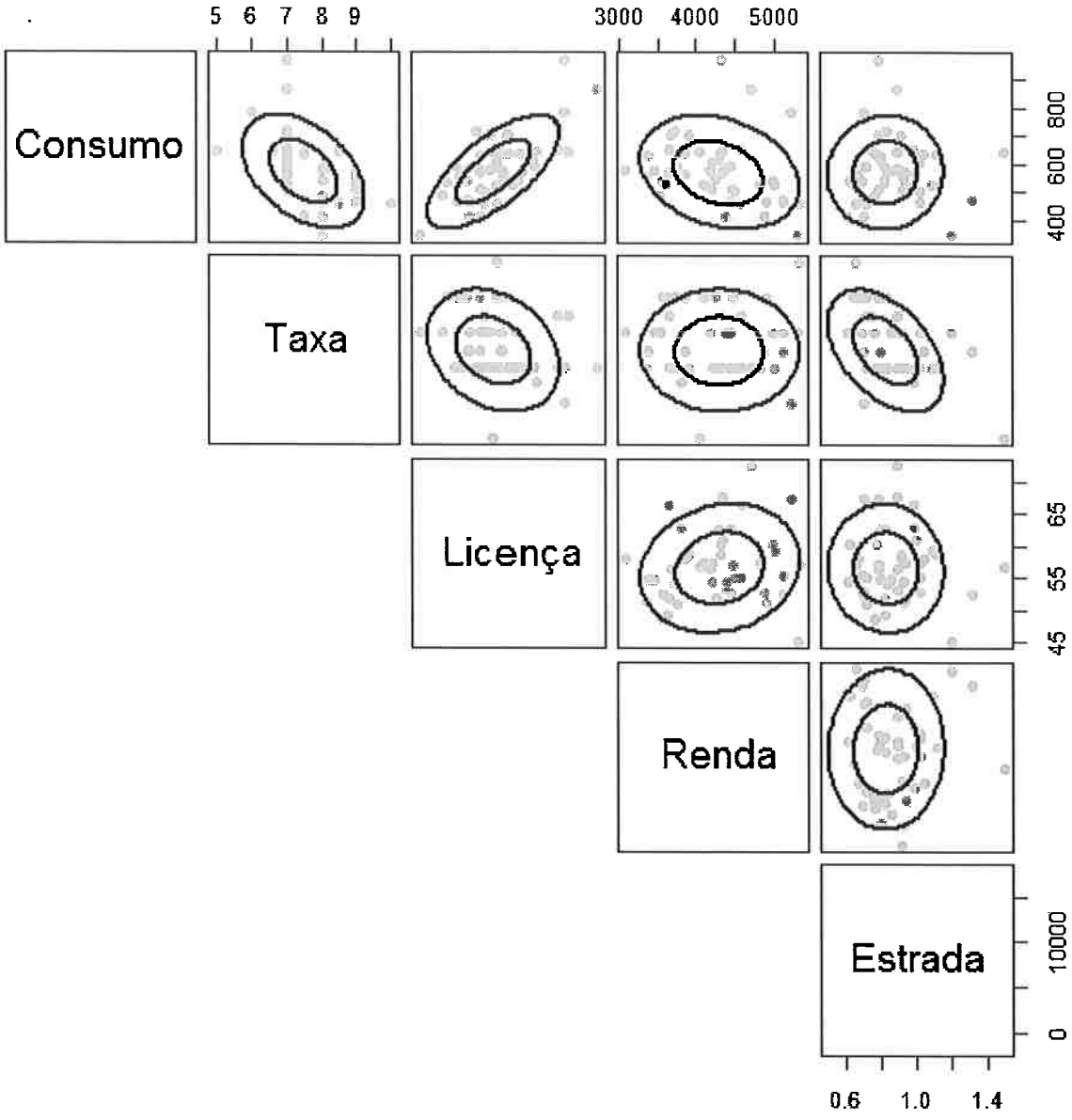

Figura 3.2: *Boxplots* bivariados, com $\alpha = 0,01$, para os dados de combustível

Tabela 3.1: Dados do consumo de combustível

Observação	Estado	Consumo	Taxa	Licença	Renda	Estrada
1	Maine	541	9,00	52,50	3571	1976
2	Vermont	561	9,00	58,00	3865	1586
3	Rhode Island	410	8,00	54,40	4399	431
4	New York	344	8,00	45,10	5319	11868
5	Pennsylvania	464	8,00	52,90	4447	8577
6	Indiana	580	8,00	53,00	4391	5939
7	Michigan	525	7,00	57,40	4817	6930
8	Minnesota	566	7,00	60,80	4332	8159
9	Missouri	603	7,00	57,20	4206	8508
10	South Dakota	865	7,00	72,40	4716	5915
11	Kansas	649	7,00	66,30	4593	7834
12	Maryland	464	9,00	51,10	4897	2449
13	West Virginia	460	8,50	55,10	4574	2619
14	South Carolina	577	8,00	54,80	3448	5399
15	Florida	574	8,00	56,30	4188	5975
16	Tennessee	571	7,00	51,80	3640	6905
17	Mississippi	577	8,00	57,80	3063	6524
18	Los Angeles	487	8,00	48,70	3528	3495
19	Texas	640	5,00	56,60	4045	17782
20	Idaho	648	8,50	66,30	3635	3274
21	Colorado	587	7,00	62,60	4449	4639
22	Arizona	632	7,00	60,30	4300	3635
23	Nevada	782	6,00	67,20	5215	2302

combinações possíveis, $\binom{48}{5} = 1.712.304$, foram tomadas 5000 combinações de amostras de tamanho 5. Dessa forma a amostra que apresentou menor resíduo mediano quadrático foi escolhida como *grupo limpo*.

Os estados que compõem o *grupo limpo* são: Colorado, New York, Los Angeles, Nebraska e California e o modelo referente ao ajuste do *grupo limpo* é dado na Tabela 3.4.

Seguindo o algoritmo do método *forward search*, após encontrado o *grupo limpo*, o processo de inserção das observações restantes do conjunto de dados bruto é o próximo passo. Como já abordado no início do capítulo, o critério de distância será o resíduo obtido do ajuste do modelo,

Tabela 3.2: Dados do consumo de combustível (continuação da Tabela 3.1)

Observação	Estado	Consumo	Taxa	Licença	Renda	Estrada
24	Oregon	610	7,00	62,30	4296	4083
25	New Hampshire	524	9,00	57,20	4092	1250
26	Massachusetts	414	7,50	52,90	4870	2351
27	Connecticut	457	10,00	57,10	5342	1333
28	New Jersey	467	8,00	55,30	5126	2138
29	Ohio	498	7,00	55,20	4512	8507
30	Illinois	471	7,50	52,50	5126	14186
31	Wisconsin	508	7,00	54,50	4207	6580
32	Iowa	635	7,00	58,60	4318	10340
33	North Dakota	714	7,00	54,00	3718	4725
34	Nebraska	640	8,50	67,70	4341	6010
35	Delaware	540	8,00	60,20	4983	602
36	Virginia	547	9,00	51,70	4258	4686
37	North Carolina	566	9,00	54,40	3721	4746
38	Georgia	631	7,50	57,90	3846	9061
39	Kentucky	534	9,00	49,30	3601	4650
40	Alabama	554	7,00	51,30	3333	6594
41	Arkansas	628	7,50	54,70	3357	4121
42	Oklahoma	644	6,58	62,90	3802	7834
43	Montana	704	7,00	58,60	3897	6385
44	Wyoming	968	7,00	67,20	4345	3905
45	New Mexico	699	7,00	56,30	3656	3985
46	Utah	591	7,00	50,80	3745	2611
47	Washington	510	9,00	57,10	4476	3942
48	California	524	7,00	59,30	5002	9794

Tabela 3.3: Estatísticas descritivas dos dados de combustível

Variáveis	Mínimo	Quartil 1	Mediana	Média	Quartil 3	Máximo
Consumo	344,00	509,50	568,50	576,77	632,75	968,00
Taxa	5,00	7,00	7,50	7,66	8,12	10,00
Licença	45,10	52,97	56,45	57,03	59,52	72,40
Renda	3063,00	3739,00	4298,00	4241,83	4578,75	5342
Estrada	431,00	3110,25	4735,50	5565,41	7156,00	17782,00

Tabela 3.4: Estimativas dos parâmetros referentes ao *grupo limpo* para os dados de combustível

Fonte de Variação	Estimativa	Estatística t	R^2
Constante	252,76	186,71	1,00
Taxa	-8,25	-66,09	
Licença	108,8	1003,64	
Renda	-0,07	-308,65	
Estrada	0,0019	41,44	

a observação com menor valor de resíduo será incluída no grupo, o que gerará um novo cálculo do ajuste, lembrando que os parâmetros são estimados pelo método de mínimos quadrados. A Tabela 3.5 mostra as estimativas dos parâmetros do modelo ao fim do processo *forward search.*

Tabela 3.5: Estimativas dos parâmetros no final do processo para os dados de combustível

Fonte de Variação	Estimativas	Estatística t	P-valor	R^2
Constante	377,29	2,033	0,05	0,6787
Taxa	-34,79	-2,682	0,0103	
Licença	13,36	6,950	$1,52e^{-08}$	
Renda	-0,06	-3,867	0,0003	
Estrada	-0,002	-0,716	0,4779	

Assim como se faz usualmente, a análise de diagnóstico tem como ponto inicial o estudo dos resíduos do modelo ajustado. Atkinson e Riani (2000) propõem o estudo do comportamento dos resíduos, mais especificamente o histórico dos resíduos ao longo do processo. Segundo os autores do método, devida à forte dependência existente entre os tamanhos dos grupos formados ao longo do método e os respectivos valores da variância do modelo, é preferível tomar os valores dos resíduos divididos pela estimativa final da variância, dada pelo estimador $\hat{\sigma}^2$,

$$\hat{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-k}. \tag{3.16}$$

A Figura 3.3 exibe o comportamento dos resíduos padronizados; as linhas horizontais são refe-

rentes aos resíduos obtidos para cada observação durante o processo de inclusão das unidades. Os estados South Dakota, Nevada, North Dakota, Wyoming e New Mexico, respectivamente indicados no gráfico pelos pontos 10, 23, 33, 44 e 45, são indicados como valores discrepantes no corpo das medidas de resíduo por ultrapassarem o limite de $2\hat{\sigma}^2$.

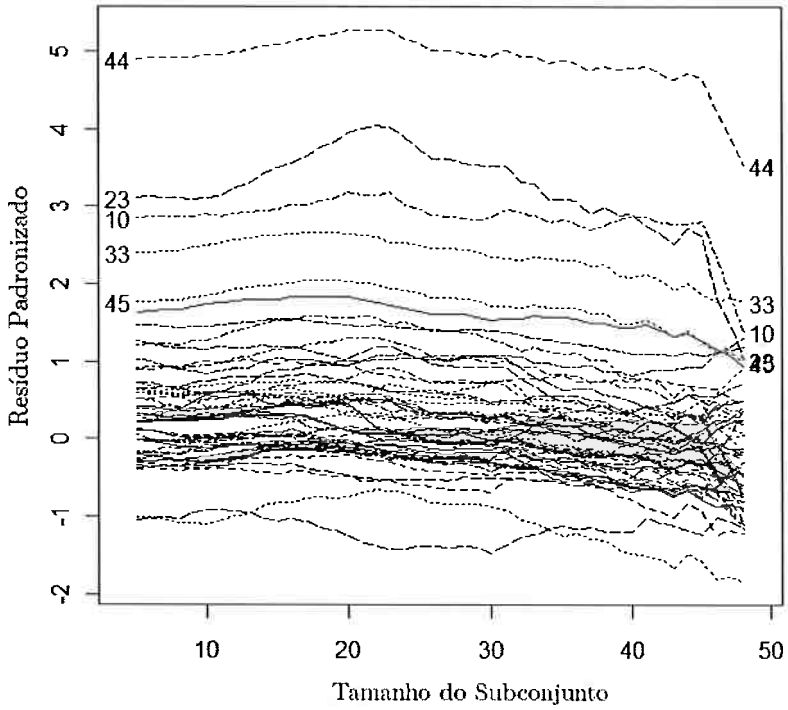

Figura 3.3: Gráfico *forward search* dos resíduos padronizados dos dados de combustível

Com base nas Figuras 3.4 e 3.5, relativas respectivamente aos valores da distância de Cook modificada mínima e às medidas de alavanca, conclui-se que os estados de Nevada, South Dakota e Wyoming são observações influentes para as estimativas dos parâmetros; a entrada destes estados no algoritmo são apontados com altos valores na Figura 3.4 para os passos em que o grupo tem tamanho $m = 46, m = 47$ e $m = 48$.

Para avaliação da medida de alavanca foi adotado o critério de Paula (2004), em que são examinados os valores de alavanca maiores ou iguais a $\frac{2k}{n}$. Tem-se que os estados apontados como influentes para as estimativas dos parâmetros, a menos do estado de Wyoming, são também influentes para seus valores ajustados.

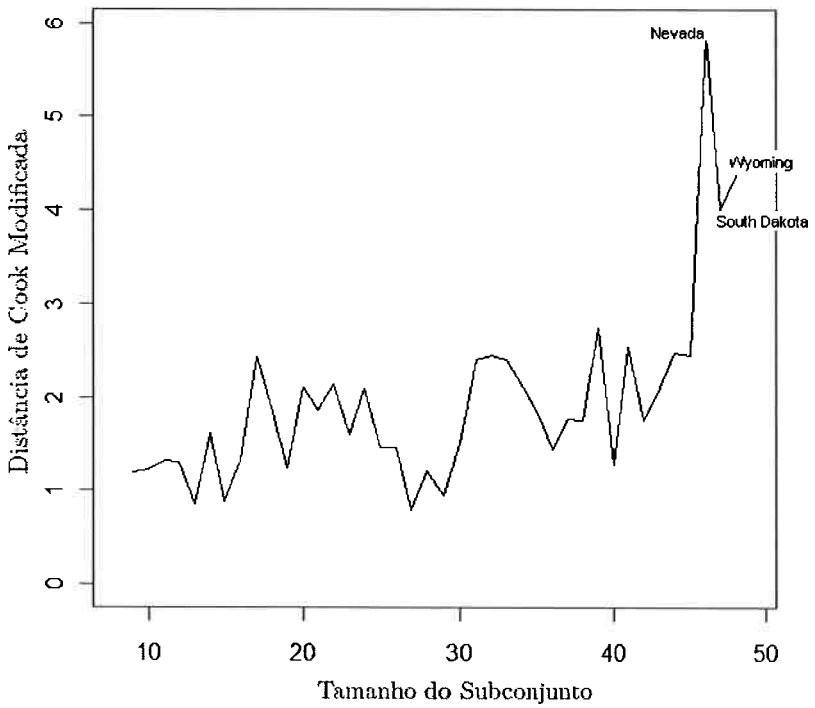

Figura 3.4: Gráfico*forward search* para as distâncias de Cook modificadas dos dados de combustível

Dada a análise de diagnóstico e apontadas as observações *outliers*, é interessante observar a sensibilidade dos parâmetros do modelo frente a tais observações.

A Figura 3.6 exibe o comportamento das estimativas dos parâmetros do modelo durante o *forward search*, em que as linhas representam os valores assumidos de cada parâmetro no decorrer do processo de inserção de observações. É notório que as estimativas dos parâmetros β_2, β_3, β_4 referentes

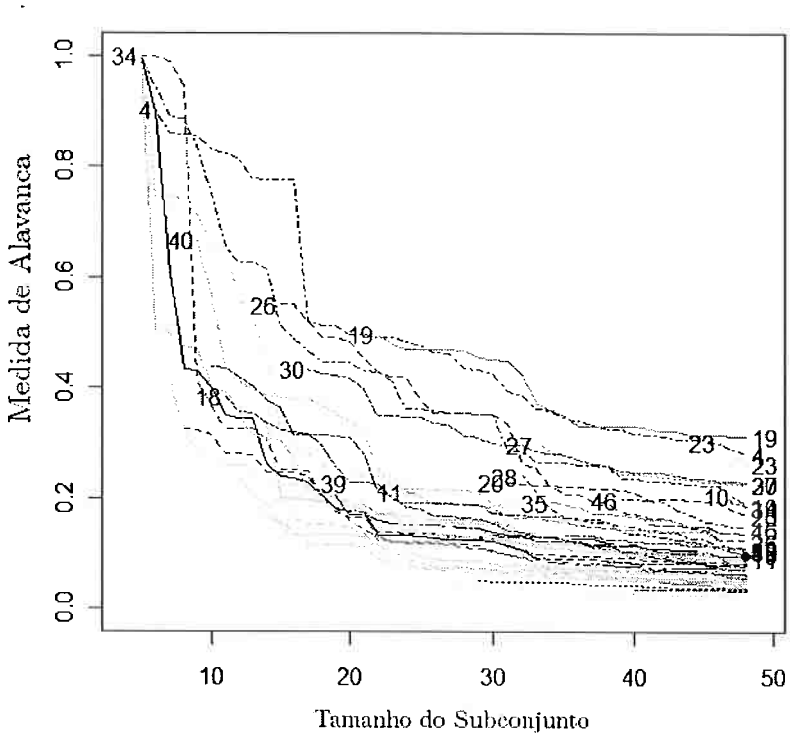

Figura 3.5: Gráfico *forward search* para as medidas de alavanca dos dados de combustível

respectivamente às variáveis Licença, Renda e Estrada, parecem não sofrer grandes variações durante o processo, ao passo que a estimativa de β_0 apresenta comportamento sensível aos dados. A estimativa de β_1, correspondente à variável Taxa, sofre um decréscimo no final do processo. Nos passos $m = 43$ e $m = 46$ são observados picos para a estimativa de β_0; tal comportamento é relativo à entrada das observação dos estados de North Dakota e Nevada.

A Figura 3.7 exibe o comportamento da estatística t, segundo o método da *variável-adicionada*, comentado na Seção 3.4. Inicialmente, quando o conjunto, após alguns passos do processo ainda se mantém homogêneo e de tamanho reduzido, o erro das estimativas dos parâmetros é próximo a zero, e os valores da estatística t são muito grandes; com o aumento do grupo e consequentemente

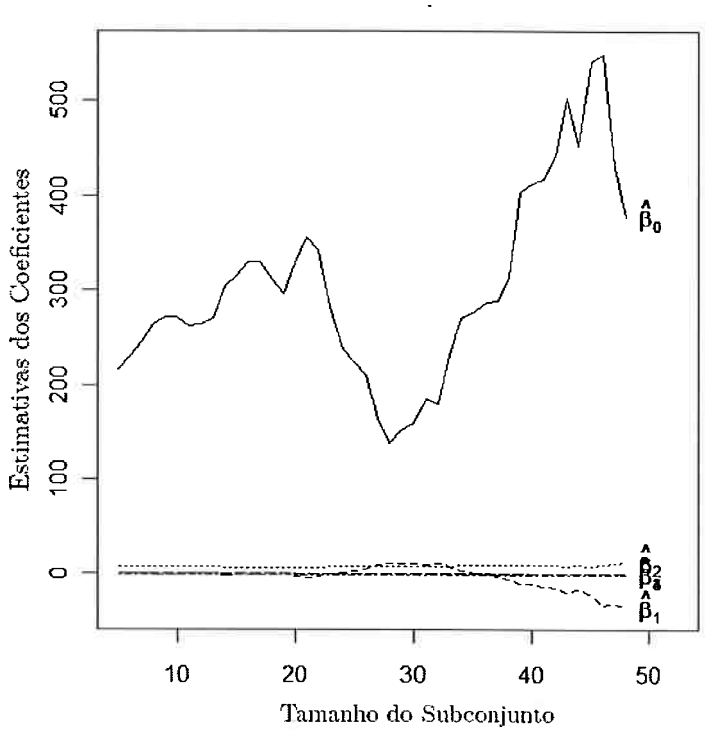

Figura 3.6: Gráfico *forward search* para as estimativas dos parâmetros dos dados de combustível

da sua variabilidade, as estatísticas t decrescem. Este fato é observado na Figura 3.7 e, a menos do parâmetro β_1, as estatísticas t dos demais parâmetros iniciam com grandes valores (absolutos) e decrescem com o aumento de m. O parâmetro β_1 parece sofrer influências das observações relativas aos estados de South Dakota, Nevada e Wyoming, pois a presença destas observações parece ser determinante para a significância da variável Taxa ao fim do processo. Na ausência destas observações a variável perde significância estatística, como mostra a Figura 3.8, em que os valores da estatística t se mantém entre as linhas horizontais, que representam o nível α de significância de 5% na distribuição normal.

O parâmetro β_4 associado à variável Estrada, independentemente da presença dos estados South

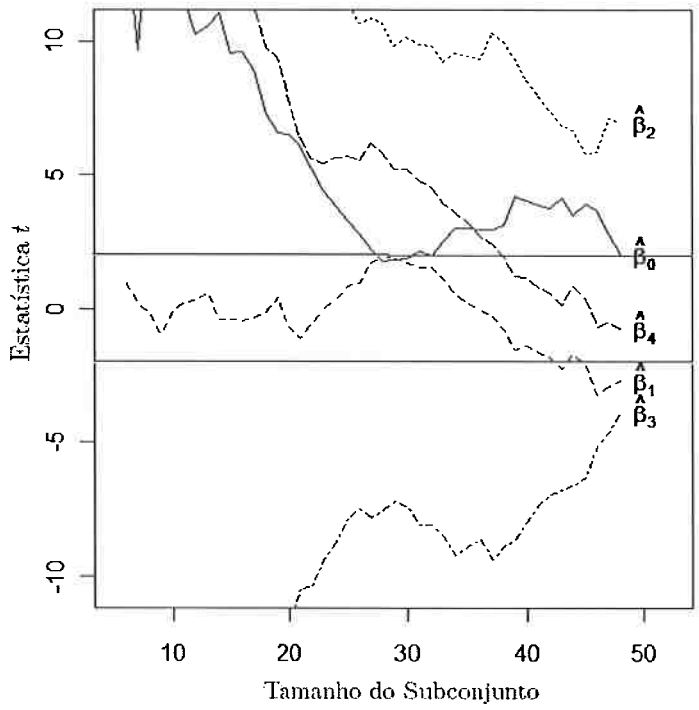

Figura 3.7: Gráfico *forward search* para as estatísticas t dos dados de combustível

Dakota, Nevada e Wyoming, se mantém não significante no final do processo, o que leva a concluir que sua presença no modelo não agrega ganho na explicação da variabilidade da variável resposta. Dado este fato, o modelo foi reajustado sem esta variável explicativa, cujos resultados são apresentados na Tabela 3.6.

Os estados South Dakota, Nevada e Wyoming apontados como *outliers* no modelo anterior continuam com comportamento discrepante e influente para o modelo sem a variável Estrada. A retirada arbitrária dessas observações conduz a uma melhor aderência à hipótese de normalidade dos resíduos, como mostram as Figuras 3.9 e 3.10, em que são gerados envelopes para a distribuição normal, e ainda um aumento de 4% da estatística R^2 e uma redução de 30% de $\hat{\sigma}^2$.

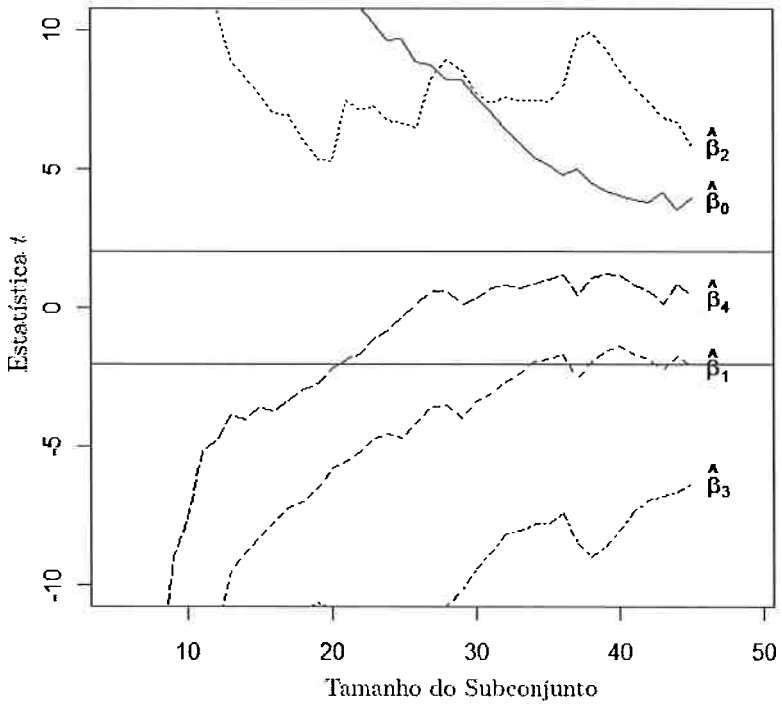

Figura 3.8: Gráfico *forward search* para a estatística t sem os estados Nevada, South Dakota e Wyoming

Tabela 3.6: Estimativas dos parâmetros referente ao modelo sem a variável estrada

Fonte de Variação	Estimativas	Estatística t	P-valor	R^2
Constante	307,32	1,96	0,056	0,67
Taxa	-29,48	-2,78	0,007	
Licença	13,74	7,48	$2,24e^{-09}$	
Renda	-0,07	-3,99	0,0002	

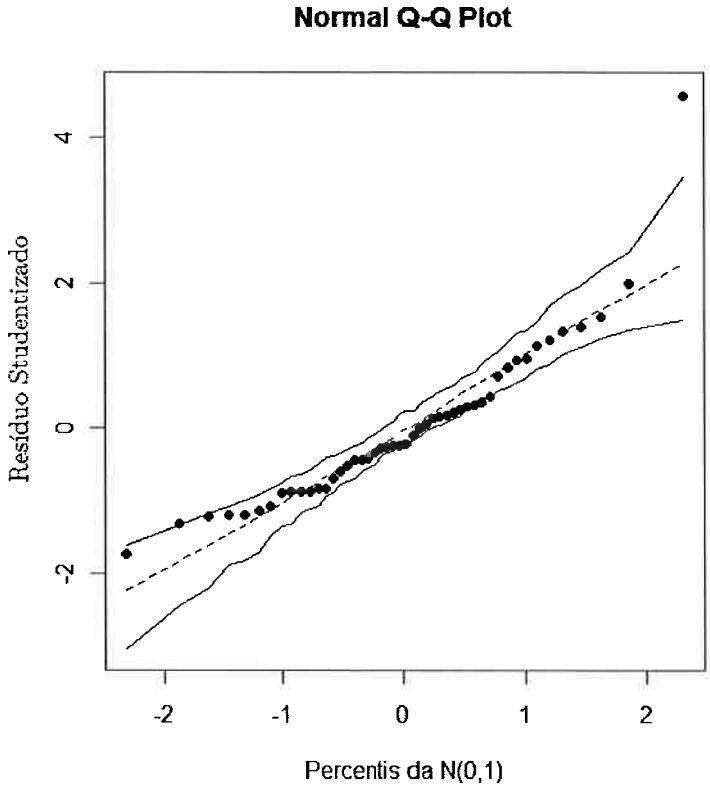

Figura 3.9: Gráfico de probabilidade normal com todos os estados

Com base na Tabela 3.3 tem-se que os valores destes estados são todos acima da média, a menos da variável Taxa. A Tabela 3.7 traz os valores das estimativas do modelo sem os estados South Dakota, Nevada e Wyoming.

Na análise-diagnóstico encontrada em Paula (2004), apenas o estado de Wyoming tem comportamento discrepante nas medidas diagnóstico; a exclusão deste estado leva a um R^2 de 0,675, não muito distante do descrito na Tabela 3.7; já a redução obtida para $\hat{\sigma}^2$ foi de 17,1%, ao passo que a eliminação dos estados de Nevada e South Dakota levou a redução de 30%, que é um indicativo de melhor ajuste.

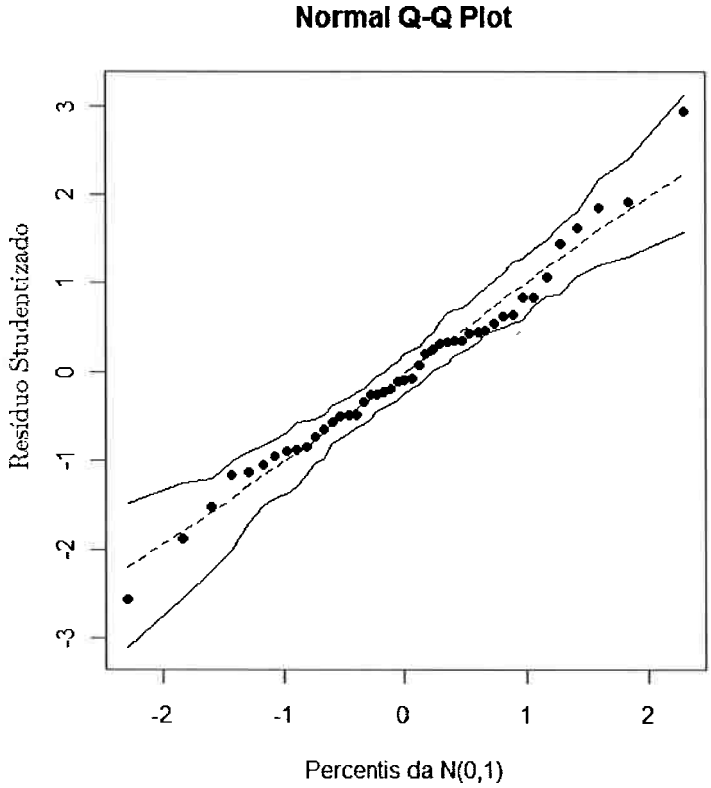

Figura 3.10: Gráfico de probabilidade normal sem os estados South Dakota, Nevada e Wyoming

Tabela 3.7: Estimativas dos parâmetros para o ajuste sem os estados South Dakota, Nevada e Wyoming

Fonte de Variação	Estimativas	Estatística t	P-valor	R^2
Constante	567,41	4,78	$2,25e^{-05}$	0,7
Taxa	-23,19	-3,04	0,004	
Licença	8,92	5,92	$5,58e^{-07}$	
Renda	-0,07	-6,41	$1,13e^{-07}$	

Capítulo 4

Modelos Lineares Generalizados

Segundo Montgomery e Vining (2001), o modelo linear generalizado (MLG) é uma união entre os modelos de regressão linear e não linear, que permite o ajuste a uma variável resposta com distribuição não normal. Nos MLGs, a distribuição da variável resposta deve apenas ser membro da família exponencial.

Distribuições que são membros da família exponencial têm a forma

$$f(y;\theta_i,\phi) = exp\left[\phi\left\{y\theta_i - b(\theta_i) + c(y,\phi)\right\}\right], \tag{4.1}$$

em que ϕ é um parâmetro de dispersão e θ_i é chamado parâmetro de locação. Ainda, tem-se que $E(Y_i) = \mu_i = b'(\theta_i)$, $Var(Y_i) = \phi^{-1}V_i$, $V_i = d\mu_i/d\theta_i$ é a função de variância. Segundo Paula (2004), seu papel é importante, pois caracteriza a distribuição, ou seja, dada a função de variância tem-se uma classe de distribuições correspondentes.

Os modelos lineares generalizados são definidos por (4.1) e pela componente sistemática

$$g(\mu_i) = \eta_i, \tag{4.2}$$

em que $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta}^T = (\beta_0, ..., \beta_{k-1})$, em que $k = p + 1$ e $k < n$, é um vetor de parâmetros desconhecidos a serem estimados e $\mathbf{x}_i = (1, x_{i1}, ...x_{ik-1})^T$ representa o vetor coluna das variáveis explicativas. O conceito de função de ligação será apresentado na próxima seção.

4.1 Função de Ligação e Preditores Lineares

A idéia básica do MLG é desenvolver um modelo linear para uma função do valor esperado da variável resposta, ou seja, estabelecer uma relação funcional entre a média da variável resposta e o preditor linear η. Desta forma, a função **liga** o componente aleatório, relativo à variável resposta, ao componente sistemático, pela relação.

$$\eta_i = g\left[E\left(Y_i\right)\right] = g\left(\mu_i\right) = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{4.3}$$

em que a resposta esperada é

$$E\left(Y_i\right) = g^{-1}\left(\eta_i\right) = g^{-1}\left(\mathbf{x}_i^T \boldsymbol{\beta}\right). \tag{4.4}$$

Há várias escolhas possíveis para a função de ligação. Se a escolha for $\eta_i = \theta_i$ é dito que a ligação é canônica, e, segundo Demétrio (2002), resulta em uma interpretação prática para os parâmetros de regressão. A Tabela 4.1 traz as principais distribuições pertencentes à família exponencial com suas respectivas funções de ligação canônicas.

Tabela 4.1: Funções de ligação canônicas

Distribuição	Ligação Canônica
Normal	$\eta = \mu$
Binomial	$\eta = log\left(\frac{\mu}{1-\mu}\right)$
Poisson	$\eta = log\left(\mu\right)$
Gama	$\eta = \mu^{-1}$
Normal Inversa	$\eta = \mu^{-2}$

Escolhida a função de ligação, a etapa seguinte do processo de modelagem é a verificação do ajuste proposto, cujo roteiro é similar à abordagem dada no contexto de modelos de regressão linear, mas agora tem-se o cuidado adicional relativo à escolha e adequabilidade da função de ligação proposta. As seções seguintes são relativas a esta etapa.

4.2 Função Desvio

Suponha que o logaritmo da função de verossimilhança seja definido por

$$L(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} L(\mu_i; y_i), \tag{4.5}$$

em que $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$. Para o modelo saturado, a função $L(\mu_i; y_i)$ é estimada por

$$L(\mathbf{y}; \mathbf{y}) = \sum_{i=1}^{n} L(y_i; y_i). \tag{4.6}$$

A estimativa de máxima verossimilhança de μ_i fica nesse caso dada por y_i, quando $k < n$, $L(\boldsymbol{\mu}; \mathbf{y})$ é denotado por $L(\hat{\boldsymbol{\mu}}; \mathbf{y})$. A estimativa de máxima verossimilhança de μ_i será dada por $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$, em que $\hat{\eta}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$.

A qualidade do ajuste de um MLG é avaliada através da função desvio ou também chamada *Deviance*, dada por

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2\{L(\mathbf{y}; \mathbf{y}) - L(\widehat{\boldsymbol{\mu}}; \mathbf{y})\}, \tag{4.7}$$

que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com k parâmetros) avaliado na estimativa de máxima verossimilhança de $\hat{\boldsymbol{\beta}}$. Um valor pequeno para a função desvio indica que para um número menor de parâmetros, obtém-se um ajuste tão bom quanto o ajuste com o modelo saturado.

4.3 Análise de Resíduo em MLG

Segundo Montgomery e Vining (2001), a análise residual é importante no ajuste de MLG, pois os resíduos podem fornecer orientação quanto à adequabilidade do modelo, e à função de ligação escolhida.

O resíduo bruto, ou resíduo ordinário, de um MLG é apenas a diferença entre o valor observado e o ajustado

$$e_i = y_i - \hat{\mu}_i. \tag{4.8}$$

Em geral, é recomendado o uso do resíduo baseado na *deviance*. Tal resíduo é definido como a contribuição da i-ésima observação para a função desvio, multiplicada pelo sinal do resíduo bruto, expresso em (4.9) e recebe o nome de *deviance* residual,

$$r_{D_i} = d_i sinal(e_i), \tag{4.9}$$

em que d_i é o valor da função desvio, para a i-ésima observação.

4.4 Verificação da Função de Ligação

Segundo Demétrio (2002), um teste formal para verificação da adequabilidade da função de ligação seria a adição de $u = \hat{\eta}^2$, como uma covariável extra no preditor linear da função de ligação, ou seja,

$$g(\boldsymbol{\mu}) = \mathbf{x}^T\boldsymbol{\beta} + \gamma u. \tag{4.10}$$

A redução drástica da estatística de *deviance* evidencia que a função de ligação é insatisfatória.

No contexto do método *forward search*, os autores adaptaram esta verificação, de tal forma que ficou condicionada a um teste para investigação se o parâmetro $\gamma = 0$.

O teste estatístico é calculado em dois estágios. No primeiro, o modelo é ajustado à função de ligação $g(\boldsymbol{\mu})$, produzindo o preditor linear $\hat{\eta}$. O segundo passo é referente ao ajuste considerando a adição da variável $u = \hat{\eta}^2$, tendo portanto a estimação do parâmetro γ. Desta forma, similar à monitoração dos parâmetros no contexto de modelos lineares, tem-se a monitoração do parâmetro γ. O gráfico referente à monitoração deste parâmetro terá as estimativas de γ passo a passo do processo, juntamente com linhas horizontais, baseadas na distribuição normal, indicando o nível de significância.

4.5 Monitoramento do *Forward Search* para MLG

A obtenção do grupo de tamanho m_0 segue os moldes do contexto de regressão linear, ou seja, é tomado um grande número de combinações de subconjuntos de tamanho k: o grupo inicial será definido pela combinação que resultar em menor mediana para os valores da função desvio. O processo *forward search* é similar aos procedimentos descritos na Seção 3.2, o diferencial está no estudo voltado para o comportamento dos valores da função desvio e a *deviance* residual. O crescimento do conjunto será dado pela seleção da observação com menor *deviance* residual.

No contexto de modelos lineares generalizados também é informativo o monitoramento das medidas de alavanca e distância de Cook. A matriz de estimação dos parâmetros por mínimos quadrados ponderados é dada por

$$\mathbf{H}_m = \mathbf{W}_m^{-\frac{1}{2}} \mathbf{X}_m (\mathbf{X}_m^T \mathbf{W}_m \mathbf{X}_m)^{-1} \mathbf{X}_m^T \mathbf{W}_m^{-\frac{1}{2}}, \tag{4.11}$$

em que $\mathbf{X}_m$ é uma matriz $m \times k$, contendo as variáveis explicativas do grupo de tamanho m e $\mathbf{W}_m = diag\{w_1, ..., w_m\}$ é uma matriz diagonal de pesos definidos por $w_i = (d\mu_i/d\theta_i)^2/V_i$. Desta forma, similarmente ao apresentado na Seção 3.3, tem-se que a distância de Cook modificada para o caso de modelos lineares generalizados é dada por

$$D_m^* = (\hat{\boldsymbol{\beta}}_{m-1}^* - \hat{\boldsymbol{\beta}}_m^*)(\mathbf{X}_m^T \mathbf{W}_m \mathbf{X}_m)(\hat{\boldsymbol{\beta}}_{m-1}^* - \hat{\boldsymbol{\beta}}_m^*)/(k\tilde{\phi}_m), \qquad m = k+1, ...n \tag{4.12}$$

em que $\tilde{\phi}$ é o estimador para o parâmetro de dispersão ϕ.

4.6 Aplicação

Como exemplo de aplicação foi tomado um conjunto de dados de Montgomery e Vining (2001). Os dados são relativos a ataques aéreos na guerra do Vietnã, em que a Marinha dos Estados Unidos operou vários ataques com a utilização de aviões de ataque naval. Os modelos de aviões usados eram dois, um Skyhawk de mecânica simples para ataques durante o dia, e o modelo A-6, com infra-vermelho para ataques noturnos. O conjunto de dados é composto de 30 missões de ataques. A variável resposta é o número de localidades que foram destruídas nessas missões, enquanto as covariáveis são o modelo do avião, codificado como 0 para o modelo A-6 e 1 para o modelo Skyhawk, carga de bombas (em toneladas) e total de meses de experiência da tripulação. A Tabela 4.2 exibe esses dados.

Tabela 4.2: Dados de ataques aéreos

Missão	Localidades	Modelo	Carga	Experiência
1	0	0	4	91,5
2	1	0	4	84,0
3	0	0	4	76,5
4	0	0	5	69,0
5	0	0	5	61,5
6	0	0	5	80,0
7	1	0	6	72,5
8	0	0	6	65,0
9	0	0	6	57,5
10	2	0	7	50,0
11	1	0	7	103,0
12	1	0	7	95,5
13	1	0	8	88,0
14	1	0	8	80,5
15	2	0	8	73,0
16	3	1	7	116,1
17	1	1	7	100,6
18	1	1	7	85,0
19	1	1	10	69,4
20	2	1	10	53,9
21	0	1	10	112,3
22	1	1	12	96,7
23	1	1	12	81,1
24	2	1	12	65,6
25	5	1	8	50,0
26	1	1	8	120,0
27	1	1	8	104,4
28	5	1	14	88,9
29	5	1	14	73,7
30	7	1	14	57,8

Dado que variável resposta é uma contagem, foi utilizado o modelo de regressão Poisson com função de ligação *log*. Desta forma, tem-se o modelo proposto

$$log(\mu_i) = \beta_0 + \beta_1 Modelo_i + \beta_2 Carga_i + \beta_3 Experiência_i. \tag{4.13}$$

Através da função *fwdglm*, que faz parte do pacote *forward*, foi aplicado o método *forward search*. O primeiro passo do processo é a escolha do *grupo limpo*, composto pelas observações 7, 11, 20 e 29 que apresentou menor mediana para a *deviance* residual.

Seguindo o algoritmo do método, os passos seguintes são as inserções das observações restantes do conjunto bruto de dados, tendo-se a cada incremento de observação, o reajuste do modelo e consequentemente o recálculo das estatísticas de análise-diagnóstico.

A Figura 4.1 exibe o histórico do comportamento da *deviance* residual, r_{D_i}. Os resíduos referentes às observações 16 e 25 são distantes do conjunto de medidas de resíduo durante todo o processo, o que levanta a hipótese que estes sejam dados *outliers*.

A Figura 4.2 exibe os gráficos da distância de Cook modificada mínima e o gráfico com os valores de alavanca das observações. O gráfico da distância de Cook modificada mostra ao fim do processo um valor discrepante, relativo à entrada da observação 25 no grupo com $m = 30$. No gráfico dos valores de alavanca, novamente a observação 25 é destacada como ponto de alavanca, segundo critério de Paula (2004) citado na Seção 3.5, juntamente com a observação 30, que entra no processo no passo $m = 8$; tal observação não é influente para as estimativas dos parâmentros, dado o valor não discrepante no gráfico para distância de Cook modificada.

A Tabela 4.3 mostra as estimativas dos parâmetros ao fim do método e a Figura 4.3 traz o comportamento das estimativas dos parâmetros durante o processo *forward search*. Observa-se que as estimativas dos parâmetros β_0 e β_1, relativas à constante do modelo e ao modelo de aeronave, sofrem grande variação ao longo do processo, fato que não se aplica aos parâmetros β_2 e β_3, referentes às variáveis carga e experiência, respectivamente.

Como complemento, a Figura 4.4 mostra o monitoramente da estatística t dos testes sobre os parâmetros β's ao longo do processo *forward search*; nela observam-se as curvas das estatísticas t_0, t_1, t_2 e t_3 relativas aos parâmetros β_0, β_1, β_2 e β_3 respectivamente, ao nível de significância de 5%, representado pelas linhas horizontais; tem-se que os parâmetros β_0 e β_3 não são estatisticamente significantes em nenhuma etapa do processo, o que leva a constatar que nenhuma observação foi

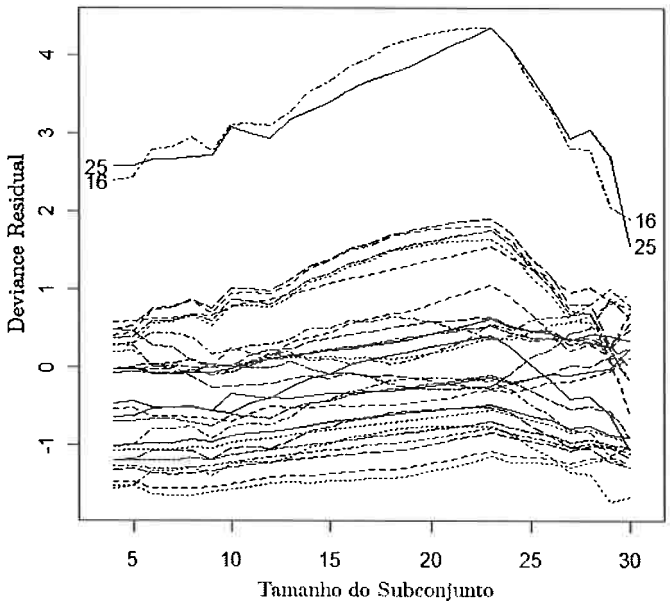

Figura 4.1: Gráfico *forward search* da *deviance* dos dados de ataques aéreos

Tabela 4.3: Estimativas dos parâmetros para os dados de ataques aéreos

Fonte de Variação	Estimativas	Estatística t	P-valor	*Deviance*
Constante	-0,406	-0,463	0,6436	25,953
Modelo	0,568	1,128	0,2595	
Carga	0,165	2,449	0,0143	
Experiência	-0,013	-1,633	0,1025	

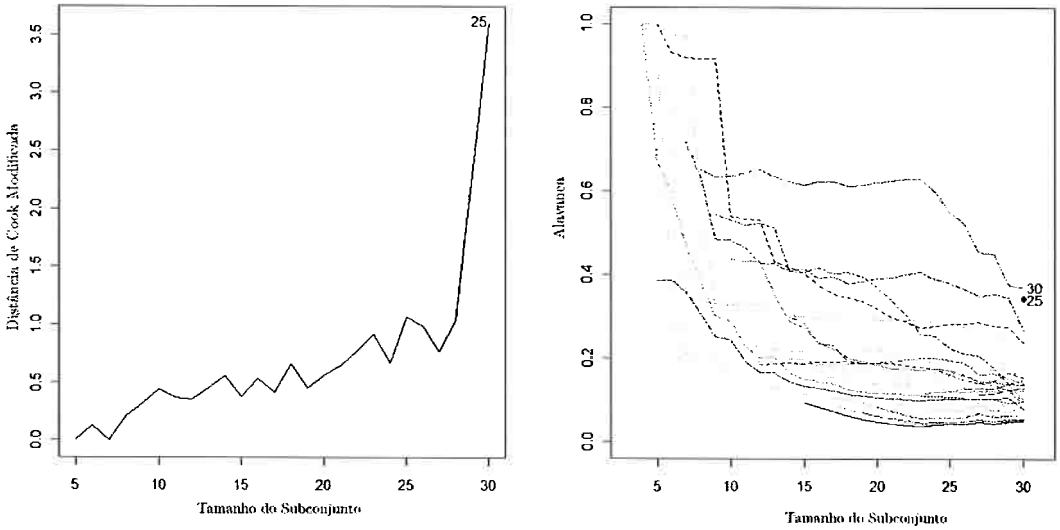

Figura 4.2: Gráfico *forward search* da distância de Cook modificada (à esquerda) e valores de alavanca (à direita) dos dados de ataques aéreos

responsável por este fato e ainda o parâmetro β_1, apesar de apresentar significância estatística entre os passos $m = 16$ e $m = 23$, ao final do processo não exibe significância estatística. Desta forma, foi realizado um ajuste aos dados sem as variáveis modelo e experiência. A Tablela 4.4 mostra as estimativas dos parâmetros obtidas deste novo ajuste.

Tabela 4.4: Estimativas dos parâmetros do modelo sem as variáveis modelo e experiência

Fonte de Variação	Estimativas	Estatística t	P-valor	*Deviance*
Constante	-1,700	-3,356	0,0007	29,206
Carga	0,231	4,942	$7,72e^{-07}$	

Dada a análise de diagnóstico, tem-se que a observação 16 e 25 são apontadas como *outliers*, sendo o modelo reajustado sem estas observações. É constatado, pela Tabela 4.5, que exibe as estimativas deste reajuste, que a estatística da *deviance* tem uma expressiva redução em comparação ao valor encontrado na Tabela 4.4, o que indica um melhor ajuste e maior proximidade ao ajuste produzido

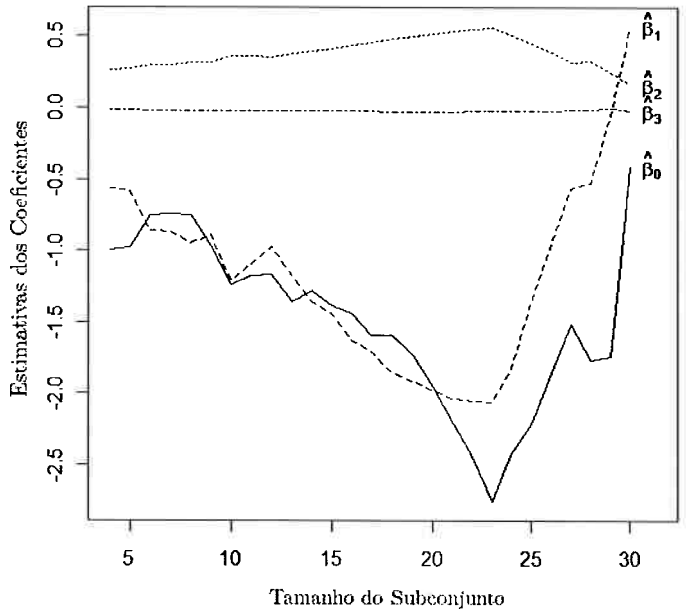

Figura 4.3: Gráfico *forward search* das estimativas dos parâmetros dos dados de ataques aéreos

por um modelo saturado.

Tabela 4.5: Estimativas dos parâmetros sem as observações 16 e 25

Fonte de Variação	Estimativas	Estatística t	P-valor	*Deviance*
Constante	-2,338	0,603	0,0001	17,739
Carga	0,277	5,218	$1,81e^{-07}$	

Tem-se ainda, na Figura 4.5, o gráfico referente ao teste para função de ligação proposta no modelo. Observa-se que as estimativas de γ se mantém dentro do intervalo de confiança de 95%, o que reflete a não significância estatística e consequentemente a conclusão que a ligação escolhida é adequada para o ajuste do modelo.

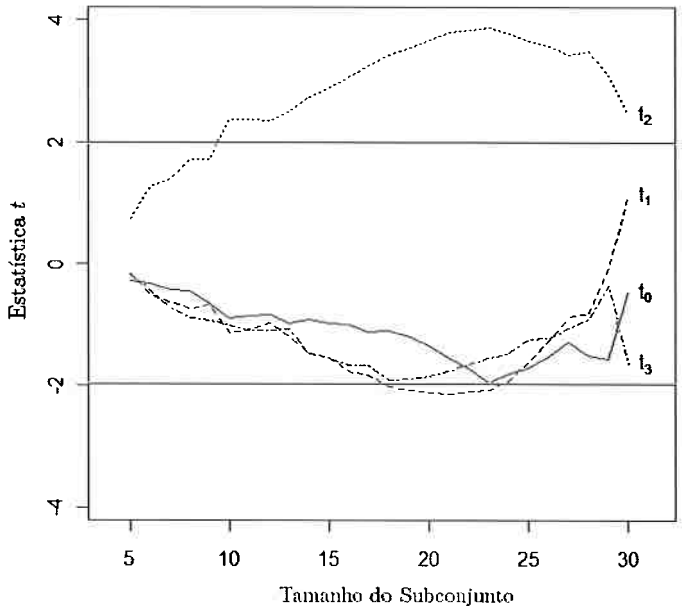

Figura 4.4: Gráfico *forward search* da estatística t dos dados de ataques aéreos

Comparativamente à análise encontrada em Montgomery e Vining (2001), o modelo final tem a mesma formulação, ou seja, apenas a variável Carga é significante no ajuste. O autor não aborda na análise a questão de dados *outliers*, levando ao ajuste encontrado na Tabela 4.4.

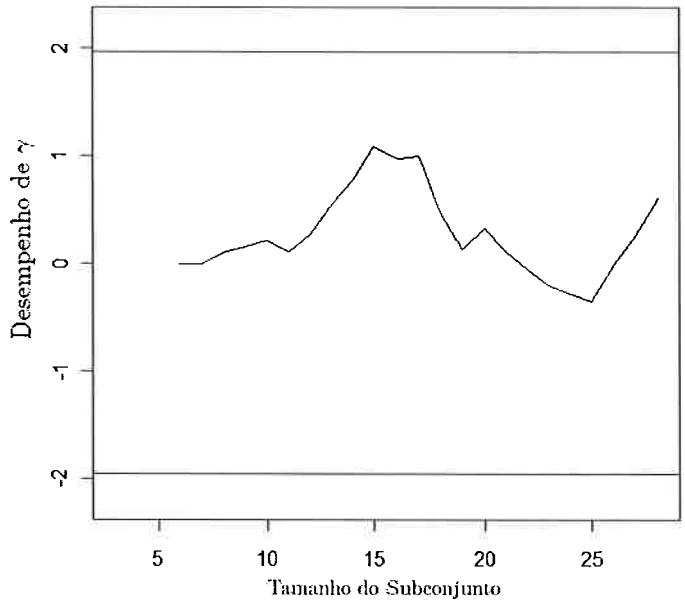

Figura 4.5: Gráfico *forward search* da estatística do teste para função de ligação dos dados de ataques aéreos

Capítulo 5

Modelos de Equações Estruturais

Algumas variáveis de interesse teórico não podem ser observadas diretamente. Essas variáveis são denominadas variáveis latentes ou construtos. Embora esses construtos não possam ser observados diretamente, informações sobre eles podem ser obtidas indiretamente através de seus efeitos sobre variáveis observáveis. Segundo Long (1983) esta é a idéia fundamental que envolve o modelo de análise fatorial. De modo geral, a análise fatorial é um procedimento estatístico que revela um número reduzido de variáveis latentes ou fatores pelo estudo da estrutura de covariância existente entre variáveis observáveis.

5.1 Especificação do Modelo Fatorial

O modelo de análise fatorial procura explicar a estrutura de covariância de um conjunto de variáveis observáveis em termos de um conjunto de fatores. Cada variável observável é considerada uma função linear de um ou mais fatores.

Matematicamente a relação entre as variáveis observáveis e os fatores é expressa por:

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{5.1}$$

em que $\mathbf{x}$ é um vetor $(p \times 1)$ de variáveis observáveis; $\boldsymbol{\xi}$ é um vetor $(q \times 1)$ de fatores comuns; $\boldsymbol{\Lambda}$ é uma matriz $(p \times q)$ de cargas fatoriais que relaciona $\mathbf{x}$ e $\boldsymbol{\xi}$; e $\boldsymbol{\delta}$ é um vetor $(p \times 1)$ de erros ou fatores únicos. Ainda é suposto que o número de variáveis observáveis em $\mathbf{x}$ seja maior que o número de fatores $\boldsymbol{\xi}$, ou seja $p > q$.

Na equação (5.1) é assumido que os valores dos vetores $\mathbf{x}$ e $\boldsymbol{\xi}$, são desvios em relação a suas médias; o que leva o valor esperado de cada vetor ser nulo, ou seja: $E(\mathbf{x}) = \mathbf{0}$; $E(\boldsymbol{\xi}) = \mathbf{0}$; e $E(\boldsymbol{\delta}) = \mathbf{0}$.

A matriz de covariância populacional para as variáveis observáveis contidas em $\mathbf{x}$ é definida como uma matriz $\boldsymbol{\Sigma} = E(\mathbf{x}\mathbf{x}^T)$ de dimensão $(p \times p)$. O elemento (i, j) de $\boldsymbol{\Sigma}$, σ_{ij}, é o valor da covariância populacional entre as variáveis X_i e X_j, e pode ser definido como $\sigma_{ij} = E(X_i X_j)$. Se o vetor $\mathbf{x}$ for padronizado, $E(X_i X_j)$ denotará a correlação entre X_i e X_j e $\boldsymbol{\Sigma}$ será uma matriz de correlação populacional.

A covariância entre os fatores comuns estão contidas em $\boldsymbol{\Phi}$, uma matriz de dimensão $(q \times q)$. O elemento ϕ_{ij} da matriz $\boldsymbol{\Phi}$ denota a covariância entre os fatores ξ_i e ξ_j. Tem-se que $\phi_{ij} = E(\xi_i \xi_j)$ ou na notação matricial $\boldsymbol{\Phi} = E(\boldsymbol{\xi}\boldsymbol{\xi}^T)$. Se for assumido que os fatores são não correlacionados, os elementos fora da diagonal principal de $\boldsymbol{\Phi}$ são zeros. Se os fatores forem padronizados, $\boldsymbol{\Phi}$ representará uma matriz de correlação com valor 1 na diagonal principal e os elementos fora dela representarão correlações entre os fatores.

A matriz de covariância entre os erros, ou fatores únicos, é denotada por $\boldsymbol{\Theta}$, uma matriz de dimensão $(p \times p)$, em que os elementos de $\boldsymbol{\Theta}$, θ_{ij}, representam a covariância entre os erros δ_i e δ_j. De forma similar às matrizes de covariância anteriores, tem-se que o valor $\theta_{ij} = E(\delta_i \delta_j)$ ou, na notação matricial, $\boldsymbol{\Theta} = E(\boldsymbol{\delta}\boldsymbol{\delta}^T)$. Na maioria das abordagens relativas aos modelos de análise fatorial, os elementos fora da diagonal principal de $\boldsymbol{\Theta}$ são assumidos serem nulos, indicando que o erro δ_i da variável X_i é não correlacionado com o erro δ_j relativo à variável X_j (para todo $i \neq j$).

A Tabela 5.1 sumariza estes conceitos.

Tabela 5.1: Sumário do modelo de análise fatorial

Matriz	Dimensão	Média	Covariância	Dimensão	Descrição
$\boldsymbol{\xi}$	$(q \times 1)$	$\mathbf{0}$	$\boldsymbol{\Phi} = E(\boldsymbol{\xi}\boldsymbol{\xi}^T)$	$(q \times q)$	variáveis latentes
$\mathbf{x}$	$(p \times 1)$	$\mathbf{0}$	$\boldsymbol{\Sigma} = E(\mathbf{x}\mathbf{x}^T)$	$(p \times p)$	variáveis observáveis
$\boldsymbol{\Lambda}$	$(p \times q)$	–	–	–	cargas fatoriais
$\boldsymbol{\delta}$	$(p \times 1)$	$\mathbf{0}$	$\boldsymbol{\Theta} = E(\boldsymbol{\delta}\boldsymbol{\delta}^T)$	$(p \times p)$	erros

Por fim, enquanto é permitido que os fatores sejam correlacionados, e em alguns casos, os erros possuam a mesma propriedade, é imposto que os fatores não sejam correlacionados com os erros, ou matematicamente $E(\boldsymbol{\delta}\boldsymbol{\xi}^T) = \mathbf{0}$.

5.2 Análise Fatorial Exploratória x Análise Fatorial Confirmatória

Segundo Barroso e Artes (2003), a AFE (Análise Fatorial Exploratória) não exige formulação de hipóteses a priori a respeito da estrutura de dependência dos dados. Essa estrutura se existir será um dos resultados da AFE.

A análise assume as seguintes restrições:

1. todas as variáveis latentes podem ser correlacionadas (ou, em alguns tipos de AFE todas as variáveis latentes são não-correlacionadas);
2. todas as variáveis observáveis podem ser diretamente afetadas por todas as variáveis latentes;
3. os erros são não-correlacionados entre si;
4. todas as variáveis observáveis são afetadas por um erro;
5. todas as variáveis latentes são não correlacionadas com todos os erros.

Na AFC (Análise Fatorial Confirmatória) as restrições são impostas de acordo com o conhecimento do pesquisador sobre o que se está estudando. Estas restrições determinam:

1. quais pares de variáveis latentes são correlacionadas;
2. quais variáveis observáveis são afetadas por quais variáveis latentes;
3. quais variáveis observáveis são afetadas por um erro;
4. quais pares de erros são correlacionados.

Segundo Barroso e Artes (2003), o que diferencia uma AFE de uma AFC é que na segunda o pesquisador indica a estrutura que ele imagina existir nos dados e, através da aplicação da técnica, terá indícios objetivos para concluir se aquela estrutura é ou não aceitável para explicar o comportamento dos mesmos.

A Figura 5.1 ilustra as relações existentes em um modelo de análise fatorial exploratória. Nela as variáveis observáveis são representadas por quadrados e as variáveis latentes são representadas por círculos. Flechas retas partindo de variáveis latentes para variáveis observáveis indicam o efeito causal

da variável latente sobre a variável observada. Flechas curvas entre duas variáveis latentes indicam que estas são correlacionadas. Esta representação gráfica é denominda diagrama de caminhos.

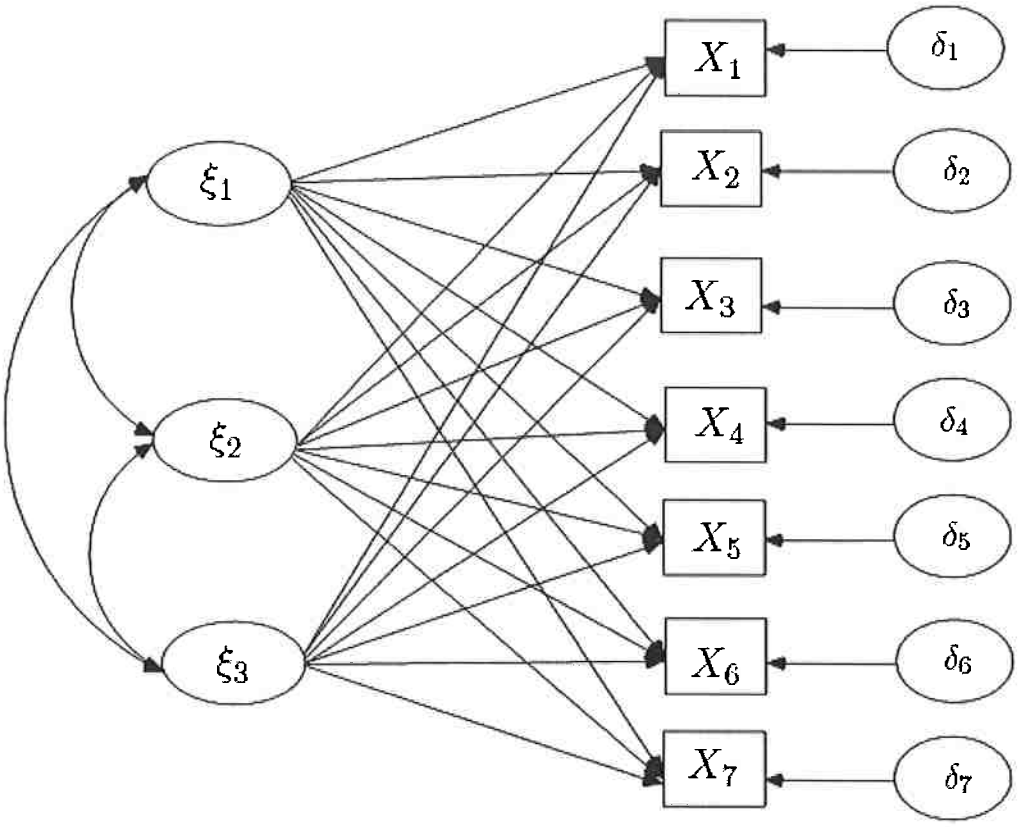

Figura 5.1: Exemplo de diagrama de caminhos do modelo de análise fatorial exploratória

As variáveis latentes ξ_1, ξ_2 e ξ_3 estão ligadas por flechas curvas, o que indica que são correlacionadas. Cada uma dessas variáveis latentes, ou fatores, tem um efeito causal sobre as variáveis observáveis contidas nos quadrados, denominadas de X_1 a X_7; tal efeito é representado pelas flechas retas partindo dos $\xi's$ para os X's. Os fatores $\xi's$ são denominados fatores comuns, pois seus efeitos são partilhados por mais de uma variável observável. As variáveis δ_1 a δ_7, representadas em círculos são denominados fatores únicos, ou erros, e são específicos a cada variável observável.

Na Figura 5.2 tem-se o diagrama de caminhos para um modelo de análise fatorial confirmatória;

os fatores comuns ξ_1 e ξ_3 são não-correlacionados, diferente do observado na Figura 5.1, em que todos os fatores comuns são correlacionados (ou alternativamente poderia se assumir não-correlação). No modelo de AFC as variáveis observáveis são afetadas por apenas alguns fatores (por exemplo é assumido que X_1 não é afetada por ξ_2 e ξ_3). Tambem é assumido neste exemplo que os dois fatores únicos δ_2 e δ_3 são correlacionados, e ainda a variável observável X_6 não é associada a nenhum fator único.

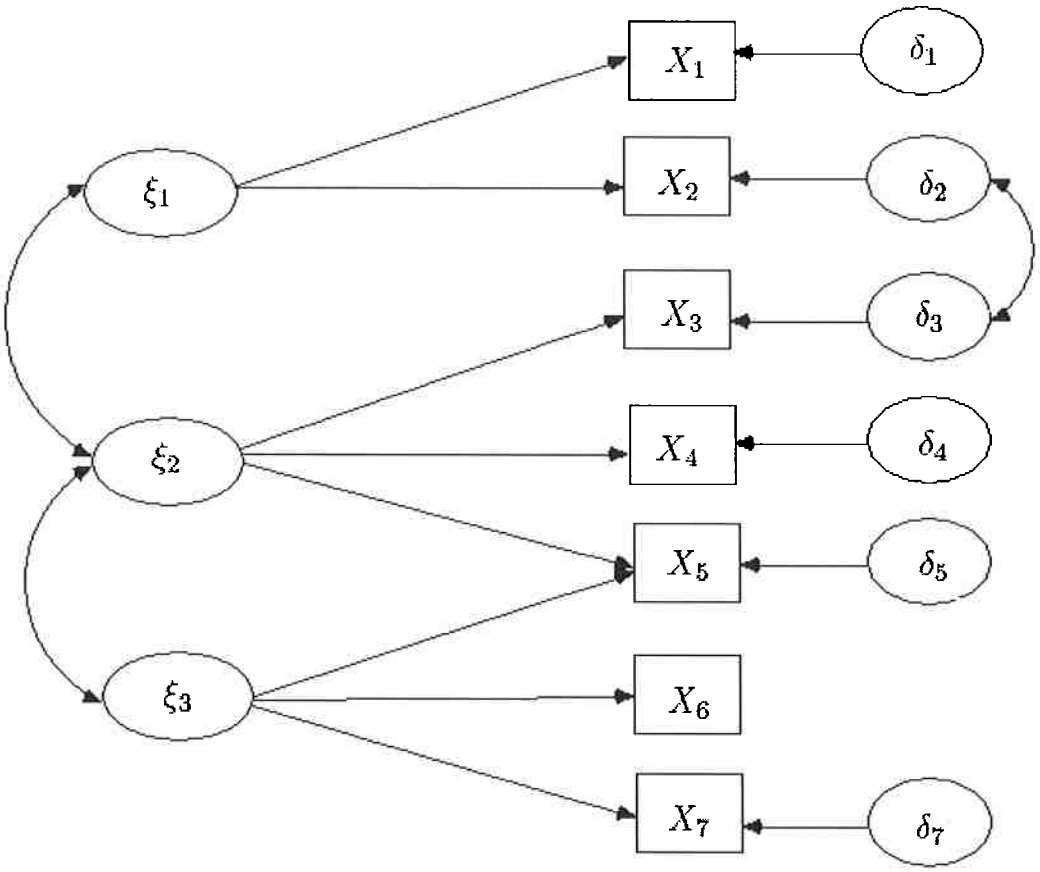

Figura 5.2: Exemplo de diagrama de caminhos do modelo de análise fatorial confirmatória

5.3 Modelos de Equações Estruturais

Segundo Mingoti (2005) a análise fatorial confirmatória está inserida no modelo de equações estruturais, que consiste de um sistema de equações lineares decomposto em dois sub-sistemas: o primeiro chamado de estrutural e que trata da relação entre as variáveis latentes, e o segundo, chamado de modelo de mensuração, que especifica as relações entre as variáveis observadas e as latentes. Matematicamente esses modelos são expressos da seguinte forma:

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \text{ (modelo estrutural)} \\ \mathbf{y} &= \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon} \text{ (modelo de mensuração de } \mathbf{y}) \\ \mathbf{x} &= \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \text{ (modelo de mensuração de } \mathbf{x}) \end{aligned} \tag{5.2}$$

sendo $\boldsymbol{\eta}_{(m\times 1)}$ e $\boldsymbol{\xi}_{(q\times 1)}$ os vetores referentes às variáveis latentes, $\mathbf{y}_{(r\times 1)}$ e $\mathbf{x}_{(p\times 1)}$ relativos ao vetor das variáveis observáveis, $\boldsymbol{\zeta}_{(m\times 1)}$ é o vetor aleatório com os erros latentes, ao passo que $\boldsymbol{\epsilon}_{(r\times 1)}$ e $\boldsymbol{\delta}_{(p\times 1)}$ são os vetores aleatórios representando os erros de mensuração de $\mathbf{y}$ e $\mathbf{x}$, respectivamente. As matrizes $\mathbf{B}_{(m\times m)}$, $\boldsymbol{\Gamma}_{(m\times q)}$, $\boldsymbol{\Lambda}_{y_{(r\times m)}}$ e $\boldsymbol{\Lambda}_{x_{(p\times q)}}$ contêm os parâmetros que descrevem a relação entre as variáveis do modelo. As suposições intrínsecas do modelo são:

- $\boldsymbol{\eta}_{(m\times 1)}$, $\boldsymbol{\xi}_{(q\times 1)}$ e $\boldsymbol{\zeta}_{(m\times 1)}$ são não-correlacionados entre si;
- $\boldsymbol{\epsilon}_{(r\times 1)}$ é não-correlacionado com $\boldsymbol{\eta}_{(m\times 1)}$, $\boldsymbol{\xi}_{(q\times 1)}$ e $\boldsymbol{\delta}_{(p\times 1)}$;
- $\boldsymbol{\delta}_{(p\times 1)}$ é não-correlacionado com $\boldsymbol{\eta}_{(m\times 1)}$, $\boldsymbol{\xi}_{(q\times 1)}$ e $\boldsymbol{\epsilon}_{(r\times 1)}$;
- os vetores de erros aleatórios têm média zero, ou seja, $E(\boldsymbol{\zeta}) = \mathbf{0}$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$ e $E(\boldsymbol{\delta}) = \mathbf{0}$;
- as matrizes de covariância de $\boldsymbol{\zeta}_{(m\times 1)}$, $\boldsymbol{\epsilon}_{(r\times 1)}$ e $\boldsymbol{\delta}_{(p\times 1)}$ são, respectivamente $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}_\epsilon$ e $\boldsymbol{\Theta}_\delta$;
- a matriz de covariância de $\boldsymbol{\xi}_{(q\times 1)}$ é $\boldsymbol{\Phi}$.

Seja $\boldsymbol{\theta}$ o vetor contendo os parâmetros do modelo. Segundo Mingoti (2005) pode ser mostrado que a matriz de covariância populacional $\boldsymbol{\Sigma}$ é uma função de $\boldsymbol{\theta}$, isto é:

$$\Sigma(\theta) = \Sigma_{(r+p)\times(r+p)} = \begin{bmatrix} \text{Var(y)} & Cov(\text{y, x}) \\ Cov(\text{x, y}) & \text{Var(x)} \end{bmatrix} .$$

e que se resume na expressão:

$$\Sigma(\theta) = \begin{bmatrix} \Lambda_y(\mathbf{I}-\mathbf{B})^{-1}(\Gamma\Phi\Gamma^T+\Psi)(\mathbf{I}-\mathbf{B}^T)^{-1}\Lambda_y^T+\Theta_\epsilon & \Lambda_y(\mathbf{I}-\mathbf{B})^{-1}\Gamma\Phi\Lambda_x^T \\ \Lambda_x\Phi\Gamma^T(\mathbf{I}-\mathbf{B}^T)^{-1}\Lambda_y^T & \Lambda_x\Phi\Lambda_x^T+\Theta_\delta \end{bmatrix}.$$

A partir da estimação da matriz $\Sigma(\theta)$ são obtidos os parâmetros que descrevem a relação entre as variáveis do modelo (5.2). Na Seção 5.5 serão abordados os métodos de estimação do modelo.

A Figura 5.3 exibe o diagrama de caminhos para um hipotético modelo de equações estruturais.

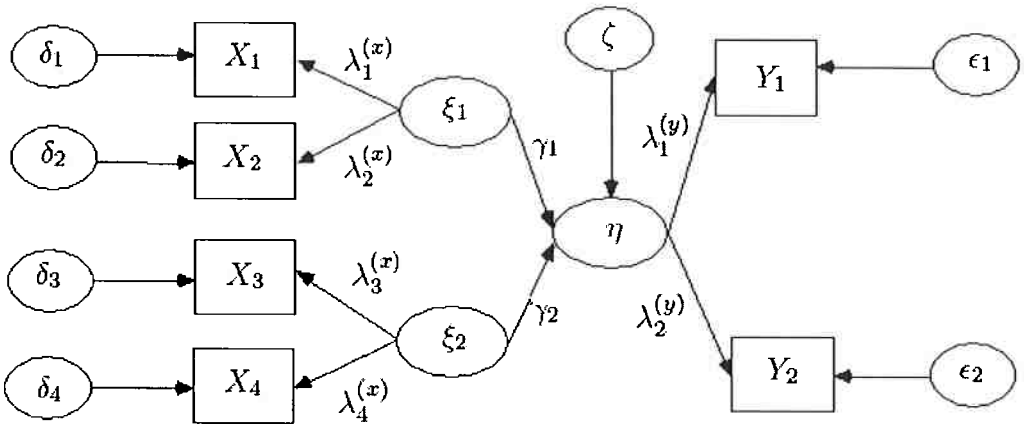

Figura 5.3: Exemplo de diagrama de caminhos do modelo de equações estruturais

A equação estrutural para o diagrama da Figura 5.3 é dada por:

$$\eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \zeta \tag{5.3}$$

e as equações de mensuração para $\mathbf{x}$ e $\mathbf{y}$ são dadas por (5.4).

$$\begin{aligned} X_1 &= \lambda_1^{(x)} \xi_1 + \delta_1 \\ X_2 &= \lambda_2^{(x)} \xi_1 + \delta_2 \\ X_3 &= \lambda_3^{(x)} \xi_2 + \delta_3 \\ X_4 &= \lambda_4^{(x)} \xi_2 + \delta_4 \\ Y_1 &= \lambda_1^{(y)} \eta + \epsilon_1 \\ Y_2 &= \lambda_2^{(y)} \eta + \epsilon_2 \end{aligned} \tag{5.4}$$

5.4 Identificação

Segundo Long (1983) a compreensão do porquê a identicabilidade do modelo fatorial venha ser um problema parte das equações de mensuração indicadas em (5.2), em que o modelo sugere que as variâncias e covariâncias das variáveis observáveis e os parâmetros $\mathbf{\Lambda}_x, \mathbf{\Lambda}_y, \mathbf{\Theta}_\epsilon, \mathbf{\Theta}_\delta$ e $\mathbf{\Phi}$ são relacionados de acordo com as equações de covariâncias expressas em (5.5)

$$\begin{aligned} \mathbf{\Sigma}_x &= \mathbf{\Lambda}_x \mathbf{\Phi} \mathbf{\Lambda}_x^T + \mathbf{\Theta}_\delta, \\ \mathbf{\Sigma}_y &= \mathbf{\Lambda}_y \mathbf{\Phi} \mathbf{\Lambda}_y^T + \mathbf{\Theta}_\epsilon. \end{aligned} \tag{5.5}$$

A menos que sejam impostas restrições sobre os parâmetros $\mathbf{\Lambda}_x, \mathbf{\Lambda}_y, \mathbf{\Theta}_\epsilon, \mathbf{\Theta}_\delta$ e $\mathbf{\Phi}$, o conjunto de soluções para as equações de mensuração de (5.5) é infinito.

Para demonstrar este fato seja $\mathbf{M}_{(q \times q)}$ uma matriz inversível. Tomando o modelo de mensuração de $\mathbf{x}$, segundo Long (1983), se for definido $\ddot{\mathbf{\Lambda}}_x = \mathbf{\Lambda}_x \mathbf{M}^{-1}$; $\ddot{\boldsymbol{\xi}} = \mathbf{M}\xi$ e $\ddot{\mathbf{\Phi}} - \mathbf{M}\mathbf{\Phi}\mathbf{M}^T$, ambos os conjuntos de matrizes, $\mathbf{\Lambda}_x, \mathbf{\Phi}, \mathbf{\Theta}_\delta$ e $\ddot{\mathbf{\Lambda}}_x, \ddot{\mathbf{\Phi}}, \mathbf{\Theta}_\delta$ satisfazem o modelo de mensuração de $\mathbf{x}$ expresso em

(5.5). Realizando as substituições em (5.5), tem-se a demonstração (5.6).

$$\begin{aligned} \ddot{\Lambda}_x \ddot{\Phi} \ddot{\Lambda}_x^T + \Theta_\delta &= \left(\Lambda_x \mathrm{M}^{-1}\right)\left(\mathrm{M}\Phi\mathrm{M}^T\right)\left((\mathrm{M}^T)^{-1}\Lambda_x^T\right) + \Theta_\delta \\ &= \Lambda_x(\mathrm{M}^{-1}\mathrm{M})\Phi(\mathrm{M}^T(\mathrm{M}^T)^{-1})\Lambda_x^T + \Theta_\delta \\ &= \Lambda_x\Phi\Lambda_x^T + \Theta_\delta = \Sigma_x. \end{aligned} \tag{5.6}$$

Desta forma, se $\Sigma_x = \Lambda_x\Phi\Lambda_x^T + \Theta_\delta$, e também valer que $\Sigma_x = \ddot{\Lambda}_x\ddot{\Phi}\ddot{\Lambda}_x^T + \Theta_\delta$, desde que as matrizes indicadas por ¨ não sejam iguais às originais, a menos que M=I, cada uma das infinitas matrizes M pode proporcionar soluções satisfatórias para o modelo fatorial. Por esse fato o modelo é dito não identificável. Tal demonstração é válida para o modelo de mensuração de $\mathbf{y}$.

Os parâmetros desconhecidos (que vêm do modelo estrutural) são ditos identificáveis se eles forem funções apenas de parâmetros identificáveis e se essas funções levarem a soluções únicas.

A hipótese sobre a estrutura de covariância dada pela equação $\Sigma = \Sigma(\theta)$ implica em $\frac{1}{2}(r+p)(r+p+1)$ equações distintas da forma $\sigma_{ij} = \sigma_{ij}(\theta)$, em que σ_{ij} é o ij-ésimo elemento de Σ e $\sigma_{ij}(\theta)$ é o ij-ésimo elemento de $\Sigma(\theta)$. Logo se um elemento de θ pode ser escrito como uma função de um ou mais elementos de σ_{ij}, então ele é identificável. E se todos os elementos de θ satisfizerem essa condição, o modelo é identificável.

5.5 Estimação

Após estabelecida a identificabilidade do modelo, o passo seguinte é referente à estimação dos parâmetros. A partir de uma amostra dos dados observados, o pesquisador pode construir a matriz de covariância S, em que os elementos s_{ii} da diagonal principal são as variâncias amostrais das variáveis observáveis, e os elementos s_{ij} fora da diagonal principal são covariâncias amostrais.

Como definido na Seção 5.4, Σ é relacionada aos parâmetros de variâncias e covariâncias populacionais através das equações (5.5). Da mesma forma uma estimativa de Σ é definida em termos de estimativas populacionais através da equação de covariância $\hat{\Sigma} = \hat{\Sigma}(\hat{\theta})$, em que $\hat{}$ indica as matrizes construídas a partir de estimativas dos parâmetros. Estas estimativas devem satisfazer às restrições impostas ao modelo. De forma geral a estimação do modelo de equações estruturais implica encontrar as estimativas dos parâmetros do vetor θ que reproduza a matriz $\hat{\Sigma}$ mais próxima possível de S.

Considerando todas as matrizes possíveis para o vetor θ, muitas delas serão descartadas por não satisfazerem às restrições impostas ao modelo, desta forma serão consideradas as matrizes $\mathbf{B}^*, \Gamma^*, \Lambda_y^*, \Lambda_x^*, \Psi^*, \Theta_\epsilon^*, \Theta_\delta^*$ que incoporam as restrições impostas. Consequentemente a matriz Σ passa ser

$$\Sigma^* = \Sigma^*(\theta^*). \tag{5.7}$$

Se Σ^* for próximo a $\mathbf{S}$, pode-se concluir que as estimativas de θ^* são razoáveis para os parâmetros populacionais.

A função que mensura a proximidade de Σ^* em relação a $\mathbf{S}$ é chamada função de ajuste e é denotada por $F(\mathbf{S}; \Sigma^*)$. Esta função é definida sob todas as possíveis matrizes geradoras do vetor θ^*. Se um conjunto dessas matrizes poduz Σ_1^*, e um segundo conjunto produz Σ_2^*, se $F(\mathbf{S}; \Sigma_1^*) < F(\mathbf{S}; \Sigma_2^*)$ será considerado que Σ_1^* está mais próxima de $\mathbf{S}$ que Σ_2^*.

Três funções de ajuste são frequentemente utilizadas na análise de modelos de equações estruturais. Estas funções correspondem aos métodos de mínimos quadrados não ponderados (ULS), mínimos quadrados generalizados (GLS) e máxima verossimilhança (ML).

O estimador de ULS para Λ, Φ e Θ é aquele que minimiza a função de ajuste dada em (5.8).

$$F_{ULS}(\mathbf{S}; \Sigma^*) = tr[(\mathbf{S} - \Sigma^*)^2]. \tag{5.8}$$

A função de ajuste por GLS é mais complexa do que por ULS, pois a diferença entre as matrizes $\mathbf{S}$ e Σ é agora ponderada pelos elementos da matriz $\mathbf{S}^{-1}$, como é exibido em (5.9).

$$F_{GLS}(\mathbf{S}; \Sigma^*) = tr[(\mathbf{S} - \Sigma^*)\mathbf{S}^{-1}]^2. \tag{5.9}$$

O estimador ML, sob a hipótese de normalidade multivariada, minimiza a função de ajuste dada por (5.10).

$$F_{ML}(\mathbf{S};\boldsymbol{\Sigma}^*) = tr(\mathbf{S}\boldsymbol{\Sigma}^{*-1}) + [log\,|\boldsymbol{\Sigma}^*| - log\,|\mathbf{S}|] - (r+p). \tag{5.10}$$

Segundo Long (1983), se os vetores $\mathbf{x}$ e $\mathbf{y}$ apresentam distribuição normal multivariada, ambos estimadores, GLS e ML, possuem propriedades assintóticas. Melhado (2004) assegura que $(n-1)F_{GLS}$ e $(n-1)F_{ML}$ avaliadas nas estimativas obtidas têm distribuição assintótica qui-quadrado com $(r+p)(r+p+1)/2 - t$ graus de liberdade, em que t representa o número de parâmetros livres, ou sem restrição.

As três funções de ajuste citadas possuem as seguintes propriedades:

1. o valor mínimo de F é zero;
2. quando $\boldsymbol{\Sigma}^* = \mathbf{S}$ e as suposições sobre a distribuição estão satisfeitas, F é inversamene proporcional a n (ou seja, $F \to 0$ quando $n \to \infty$).

5.6 Medidas de Ajuste

Índices de qualidade do ajuste baseados nas funções de ajuste citadas na Seção 5.5 podem ser usados na indicação do desempenho do modelo frente aos dados. Melhado (2004) traz um estudo detalhado das medidas de ajuste existentes. Neste trabalho serão citadas algumas medidas que o *software* R oferece através do pacote SEM, específico para modelagem de equações estruturais.

A estatística qui-quadrado procedente da função de ajuste pode ser utilizada para avaliar o ajuste geral do modelo. Segundo Melhado (2004) uma possível medida de ajuste para quantificar a diferença entre as matrizes de covariâncias $\boldsymbol{\Sigma}^*$ e $\mathbf{S}$ é a matriz resíduo $\mathbf{V}_{(g\times g)}$, em que $g = r + p$ representa o total de variáveis observáveis. Os elementos dessa matriz são $v_{ij} = s_{ij} - \sigma^*_{ij}$, em que s_{ij} é o ij-ésimo elemento de $\mathbf{S}$ e σ^*_{ij} é o ij-ésimo elemento de $\boldsymbol{\Sigma}^*$.

Uma medida alternativa baseada nos resíduos amostrais v_{ij} é a raiz do quadrado médio residual (SRMR), dada por:

$$SRMR = \left[2\sum_{i=1}^{p}\sum_{j=1}^{i}\frac{(s_{ij}-\sigma^*_{ij})^2}{g(g+1)}\right]^{\frac{1}{2}}. \tag{5.11}$$

Segundo a equação (5.11), tem-se a raiz do quadrado médio dos elementos abaixo da diagonal principal (incluindo a diagonal principal) da matriz de resíduos. Em consequência de um bom ajuste, é esperado que os valores da martiz $\mathbf{V}_{(g\times g)}$ sejam próximos de zero; dessa forma os valores da estatística SRMR póximos de zero indicam um bom ajuste.

Devido aos resíduos amostrais serem afetados por fatores como diferenças de escala entre as variáveis observáveis e tamanho amostral, Jöreskog e Sörbom (1986) propõem uma correção para os resíduos v_{ij}, dada por

$$e_{ij} = \frac{v_{ij}}{\left[\frac{\sigma^*_{ii}\sigma^*_{jj}+\sigma^{*2}_{ij}}{n}\right]^{\frac{1}{2}}}. \tag{5.12}$$

Os valores absolutos mais altos de e_{ij} indicam os elementos s_{ij} que têm pior ajuste pelo modelo.

Jöreskog e Sörbom (1986) propuseram os índices de qualidade de ajuste (GFI) e índice de qualidade de ajuste corrigido (AGFI), para modelos ajustados pelo método de máxima verossimilhança, dados pelas expressões

$$GFI_{MV} = 1 - \frac{tr[(\mathbf{\Sigma}^{*-1}\mathbf{S} - \mathbf{I})]^2}{tr[(\mathbf{\Sigma}^{*-1}\mathbf{S})^2]}, \tag{5.13}$$

$$AGFI_{MV} = 1 - \left[\frac{g(g+1)}{2gl}\right][1 - GFI_{MV}]. \tag{5.14}$$

A expressão relativa ao GFI mensura a quantidade relativa das variâncias e covariâncias em $\mathbf{S}$ que são preditas por $\mathbf{\Sigma}^*$, ao passo que a expressão $AGFI$ corrige o GFI pelos graus de liberdade relativos ao número de variáveis observáveis. Segundo Melhado (2004) os índices alcançam seu valor máximo 1 quando $\mathbf{\Sigma}^* = \mathbf{S}$, em geral são maiores que zero, embora valores negativos sejam possíveis.

5.7 Método *Forward Search* em Modelos de Equações Estruturais

Como mencionado no Capítulo 2, a aplicação do método *forward search*, independentemente da área estatística, segue o mesmo roteiro, ou seja, a criação de um conjunto inicial livre de *outliers* e o aumento deste a cada passo do processo, até a inclusão de todas as observações do conjunto bruto.

Para a abordagem de equações estruturais, o critério de formação do *grupo limpo*, bem como a regra de entrada de observações a cada passo do processo, serão baseados na medida de distância de Mahalanobis ao centro dos dados, dada na equação (2.2). Como consequência o *grupo limpo* será referente às observações mais próximas ao centro multivariado dos dados.

O processo de inserção das observações concernentes ao restante de observações do conjunto bruto obedecerá a ordem das distâncias destas observações em relação ao *grupo limpo*, ou seja, as observações referentes às menores distâncias terão preferência na inclusão nos consecutivos passos. A entrada de observação a cada passo do processo leva ao recálculo dos parâmetros da equação (2.2).

Mavridis e Moustaki (2008) sugerem, em sua abordagem para análise fatorial, que o tamanho do *grupo limpo* seja no mínimo igual a $\frac{1}{2}g(g+1)+1$, quantidade referente ao número de variâncias e covariâncias presentes na matriz $\mathbf{S}$ mais um, em que g representa o número de variáveis observáveis. Pelo fato da análise fatorial estar inserida no modelo de equações estruturais, tal sugestão será adotada. O cuidado sobre o *grupo limpo* recai sobre a hipótese de que um tamanho muito maior ao sugerido aumente a chance de inclusão de observações *outliers*, o que acarretaria erros no método.

Dentro do processo *forward search* sugerido, tem-se que a matriz $\mathbf{S}$ é re-estimada a cada inserção de observações durante o método, o que permitirá o monitoramento de seus elementos a cada passo do processo, bem como as estatísticas de ajuste do modelo proposto, uma vez que tal matriz é base para a construção do modelo. O monitoramento das estatísticas de ajuste será limitado às apresentadas na Seção 5.6, além da estatística qui-quadrado, por serem as medidas mais utilizadas. A informação resultante de tal monitoramento será a avalição do impacto de observações discrepantes sobre o ajuste geral do modelo proposto.

5.8 Aplicação

Como exemplo de aplicação foi sugerida a contrução de um modelo de equações estruturais contendo apenas as equações de mensuração e uma variável latente. Tal sugestão tem a finalidade de proporcionar melhor visualização e compreensão do desempenho do método *forward search* na

aplicação de modelos de equações estruturais.

Inicialmente o conjunto de dados considerados para tal abordagem é hipotético. Os dados e as estatísticas descritivas das variáveis observáveis são exibidos nas Tabelas 5.2 e 5.3. Destaca-se que as observações 21, 22 e 23, da Tabela 5.2 foram imputadas intencionalmente para serem observações *outliers*.

Tabela 5.2: Dados hipotéticos

Observação	X_1	X_2	X_3
1	0,606	0,127	0,494
2	0,557	0,124	0,549
3	0,534	0,114	0,521
4	0,58	0,125	0,546
5	0,536	0,118	0,592
6	0,547	0,114	0,531
7	0,651	0,136	0,527
8	0,448	0,103	0,522
9	0,653	0,140	0,625
10	0,476	0,106	0,429
11	0,528	0,106	0,424
12	0,703	0,134	0,519
13	0,586	0,111	0,505
14	0,573	0,106	0,465
15	0,63	0,115	0,489
16	0,489	0,123	0,562
17	0,685	0,156	0,631
18	0,685	0,156	0,566
19	0,523	0,132	0,505
20	0,664	0,159	0,506
21	0,888	0,075	0,782
22	0,283	0,178	0,631
23	0,507	0,143	0,904

A Figura 5.4 traz os *boxplots* bivariados das variáveis observáveis, assinalando as observações 21, 22 e 23 como *outliers*. Ainda, com base na dispersão dos dados nos respectivos cruzamentos, é possível notar a existência de correlação entre as variáveis.

Tabela 5.3: Estatísticas descritivas dos dados hipotéticos

Variáveis	Mínimo	Quartil 1	Mediana	Média	Quartil 3	IQ	Máximo
X_1	0,283	0,525	0,573	0,578	0,652	0,127	0.888
X_2	0,075	0,112	0,124	0,126	0,138	0,026	0,178
X_3	0,424	0,505	0,527	0,557	0,579	0,074	0,904

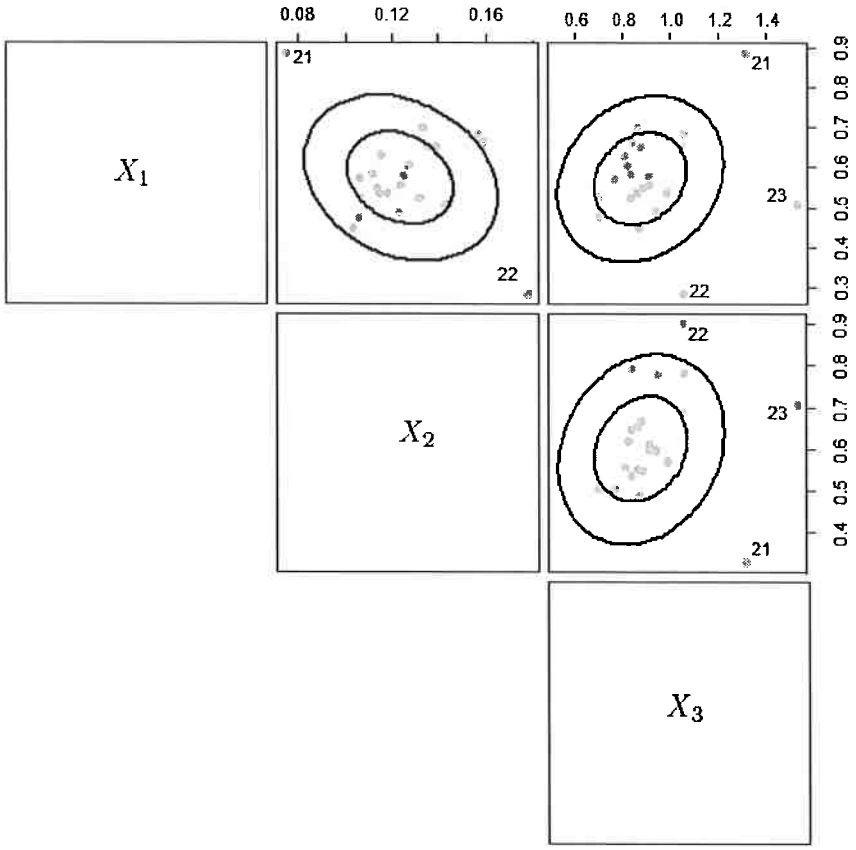

Figura 5.4: *Boxplots* bivariados para os dados hipotéticos

Avaliando as distâncias de Mahalanobis escalonadas, D^2_{im}, constata-se pela Figura 5.5 que as observações apontadas na Figrura 5.4 são de fato distantes da massa de dados, mesmo a observação 21, que no início do processo não tem comportamento discrepante, em poucos passos do processo passa a seguir o padrão das observações 22 e 23, com valores das distâncias de Mahalanobis discrepantes em relação ao observado para o restante dos dados.

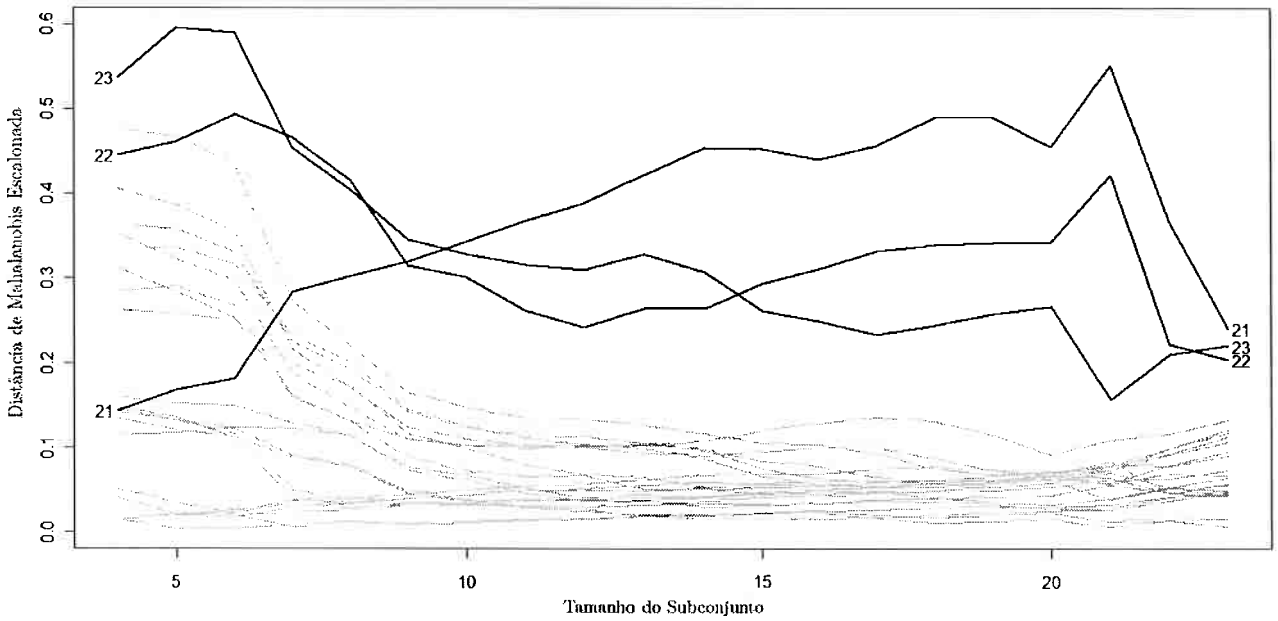

Figura 5.5: Gráfico *forward search* para as distâncias de Mahalanobis escalonadas para os dados hipotéticos

Analisando a estrutura de variância e covariância destes dados, tem-se que a variância generalizada ($|\mathbf{S}|$) exibida na Figura 5.6, se mantém constante até o passo $m = 20$. Logo após, nos passos $m = 21$ $m = 22$ e $m = 23$, é observado o aumento de seu valor; tais passos são respectivamente relativos à entrada das observações 21, 22 e 23.

O acompanhamento em maior detalhe da estrutura de variância e covariância, exibido na Figura 5.7, mostra que os valores das variâncias das variáveis X_1 e X_3, em cada passo do processo, assumem valores maiores e discrepantes em relação ao valores assumidos pela variância de X_2. Já os valores

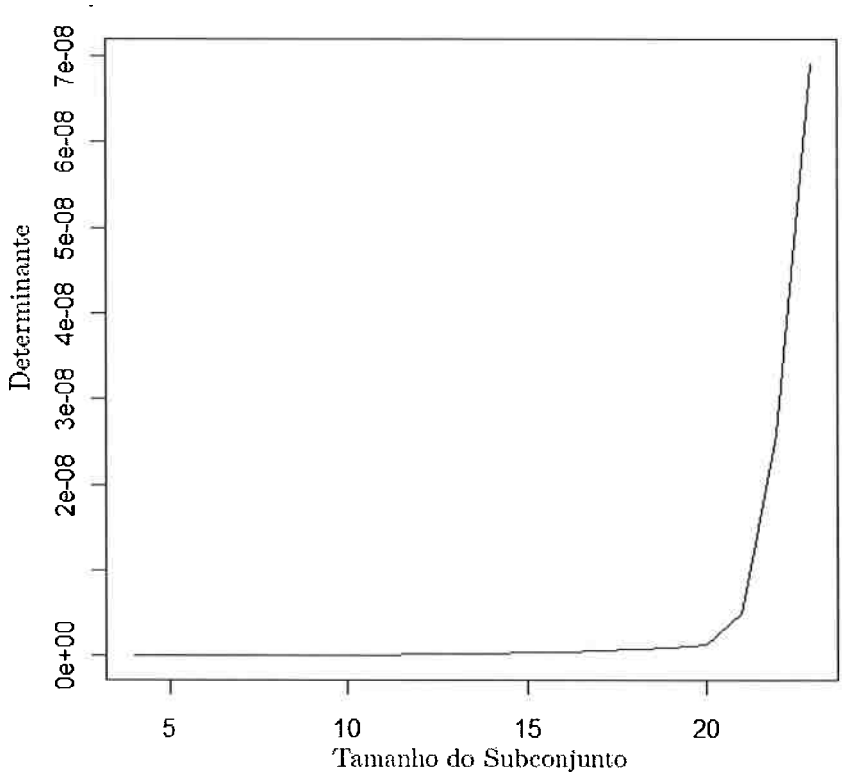

Figura 5.6: Gráfico *forward search* para a variância generalizada dos dados hipotéticos

de covariância das três variáveis em estudo não sofrem grandes oscilações com observações distantes do *grupo limpo*, composto das 7 observações mais próximas ao centro dos dados.

Com base na Figura 5.8, as equações de mensuração são dadas por (5.15)

A Figura 5.9 apresenta o diagrama de caminhos com as estimativas dos parâmetros e os respectivos erros padrões entre parênteses para o ajuste ao conjunto total dos dados.

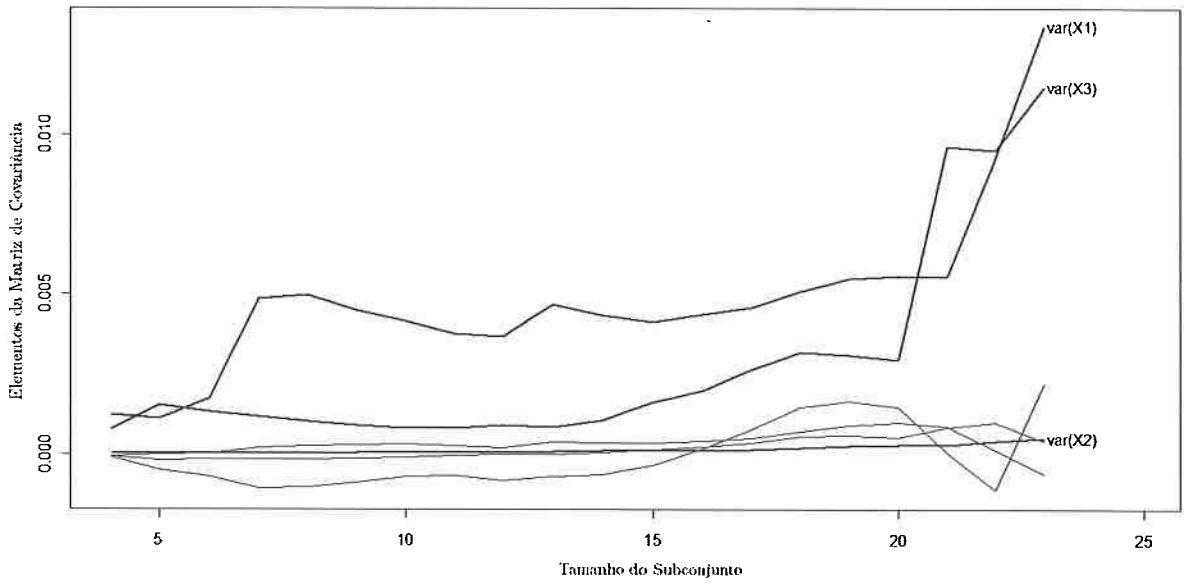

Figura 5.7: Gráfico *forward search* dos elementos da matriz de covariância para os dados hipotéticos

$$\begin{aligned} X_1 &= a_1\eta + \epsilon_1 \\ X_2 &= a_2\eta + \epsilon_2 \\ X_3 &= a_3\eta + \epsilon_3 \end{aligned} \tag{5.15}$$

A Figura 5.10 exibe o comportamento da estatística qui-quadrado, em que observa-se que nos passos iniciais do processo *forward search*, compreendidos entre $m_0 = 7$ a $m = 9$, há um decréscimo da estatística, o que indica que o ajuste do modelo melhora a cada inserção de observação, o valor da estatística entre os passos $m = 9$ a $m = 20$ são aproximadamente zero; tais valores apontam um ajuste ótimo do modelo, em que a hipótese $\mathbf{\Sigma^*} = \mathbf{S}$ se confirma. A entrada das observações 21, 22 e 23, respectivamente nos passos $m = 21$, $m = 22$ e $m = 23$ provocam uma elevação no

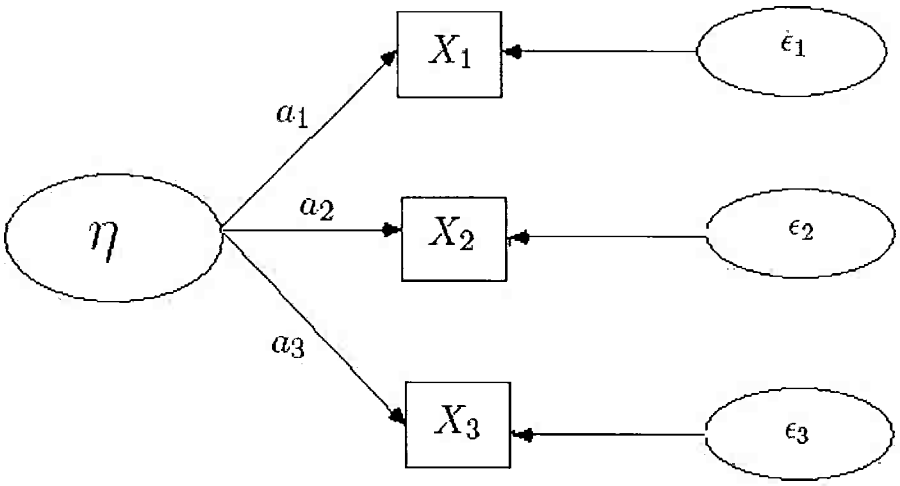

Figura 5.8: Diagrama de caminhos para os dados hipotéticos

valor da estatística qui-quadrado e consequentemente uma defasagem no ajuste do modelo, como é apresentado nas Figuras 5.11 e 5.12 relativas respectivamente ao índice GFI_{MV} e à medida SRMR.

O comportamento observado nas Figuras 5.10, 5.11 e 5.12, em que as medidas de qualidade de ajuste melhoram a cada passo do processo, é explicado pelo algoritmo do método *forward search*. Inicialmente, com a formação do *grupo limpo*, tem-se que os dados contidos neste grupo são homogêneos e não exibem uma razoável estrutura de dependência para o ajuste das equações de mensuração; tal estrutura surge com a entrada de observações nas etapas seguintes do processo, o que dará a forma da estrutura de correlação existente no conjunto de dados. Os dados distantes desta estrutura são dectados através dos impactos negativos na estatística qui-quadrado, no índice de qualidade de ajuste GFI_{MV} e na medida SRMR.

Observado o desempenho do método *forward search* no conjunto de dados hipotéticos, em que foi possível compreender a performance do método, foi realizada uma segunda aplicação, com dados reais e um modelo completo, com equação estrutural e equações de mensuração. O conjunto de dados foi retirado da literatura e é relativo a um estudo de Holzinger e Swineford (1939), acerca dos fatores cognitivos da habilidade espacial e verbal. Tal conjunto pode ser encontrado no pacote específico

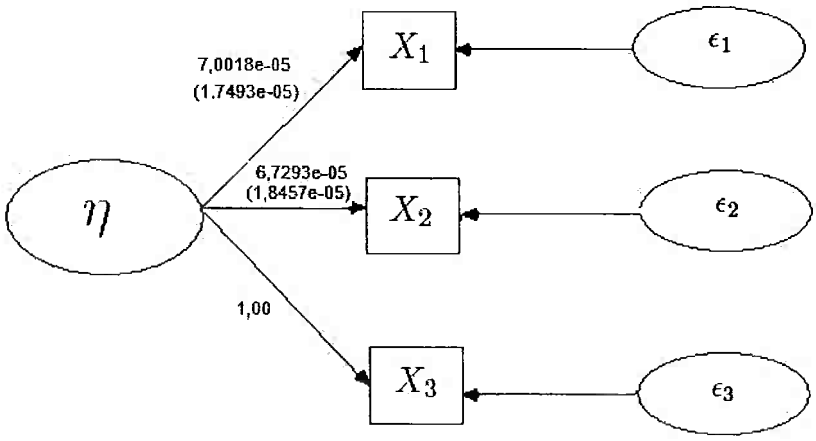

Figura 5.9: Diagrama de caminhos para os dados hipotéticos com as estimativas dos parâmetros

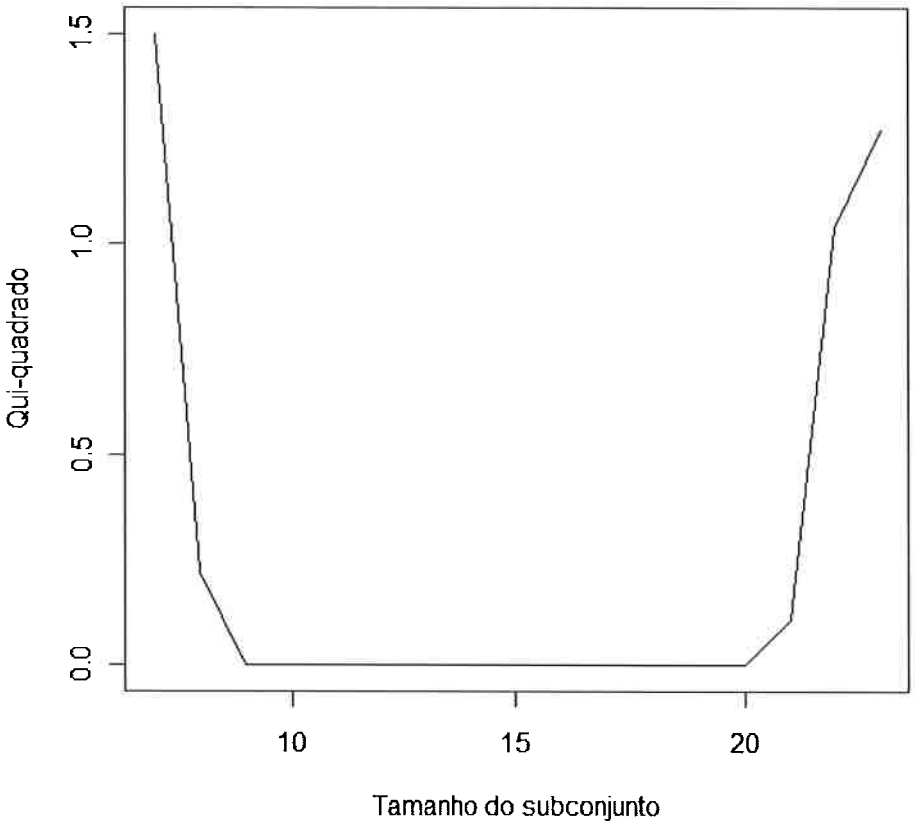

Figura 5.10: Gráfico *forward search* da estatística qui-quadrado para os dados hipotéticos

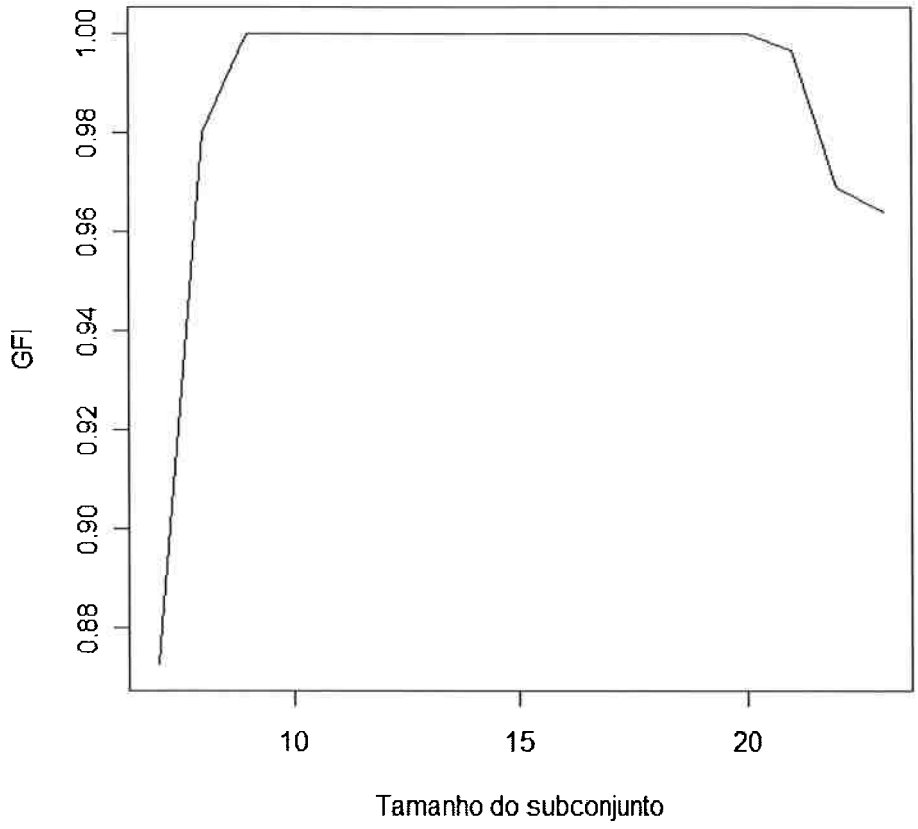

Figura 5.11: Gráfico *forward search* para o índice de qualidade de ajuste GFI_{MV} para os dados hipotéticos

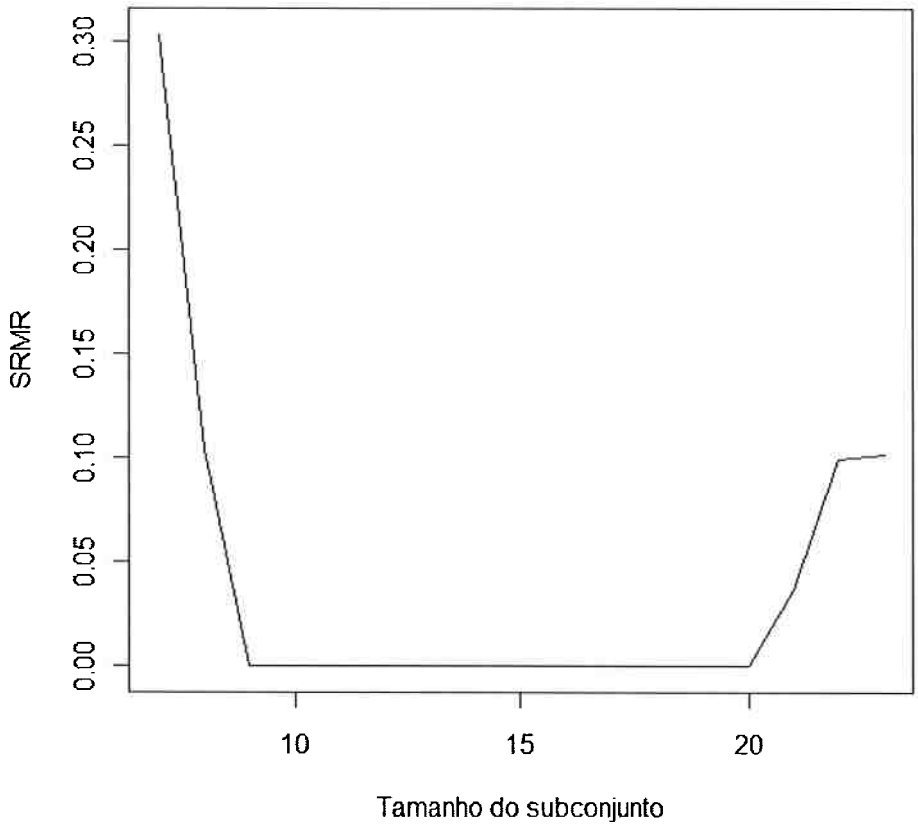

Figura 5.12: Gráfico *forward search* para a medida SRMR para os dados hipotéticos

para modelagem de equações estruturais, AMOS (Arbuckle, 2005). A amostra utilizada nesta análise é composta por 73 estudantes de sétima a oitava séries, do sexo feminino, de uma escola de Chicago (EUA). A Tabela 5.4 exibe as variáveis observáveis presentes no estudo. O conjunto de dados e as estatísticas descritivas das variáveis seguem respectivamente nas Tabelas 5.5, 5.6 e 5.7.

Tabela 5.4: Variáveis observáveis do estudo sobre habilidade verbal e espacial

Variáveis	Descrição
Visperc	escore para percepção visual
Cubo	teste para vizualização espacial
Losango	teste para orientação espacial
Parágrafo	escore para compreensão de parágrafo
Sentença	escore para compreensão de sentenças
Significado	teste de compreensão do significado das palavras

O modelo proposto especifica duas variáveis latentes, denominadas como Espacial e Verbal, cujas relações com as variáves observáveis são dadas de acordo com o diagrama de caminhos da Figura 5.13.

Com base na Figura 5.13 a equação estrutural será dada por (5.16).

$$Verbal = \gamma Espacial + \zeta \tag{5.16}$$

ao passo que as equações de mensuração são dadas por (5.17)

$$\begin{aligned} Visperc &= a_1 Espacial + \epsilon_1 \\ Cubo &= a_2 Espacial + \epsilon_2 \\ Losango &= a_3 Espacial + \epsilon_3 \\ Sentenca &= b_1 Verbal + \delta_1 \\ Significado &= b_2 Verbal + \delta_2 \\ Parágrafo &= b_3 Verbal + \delta_3 \end{aligned}$$

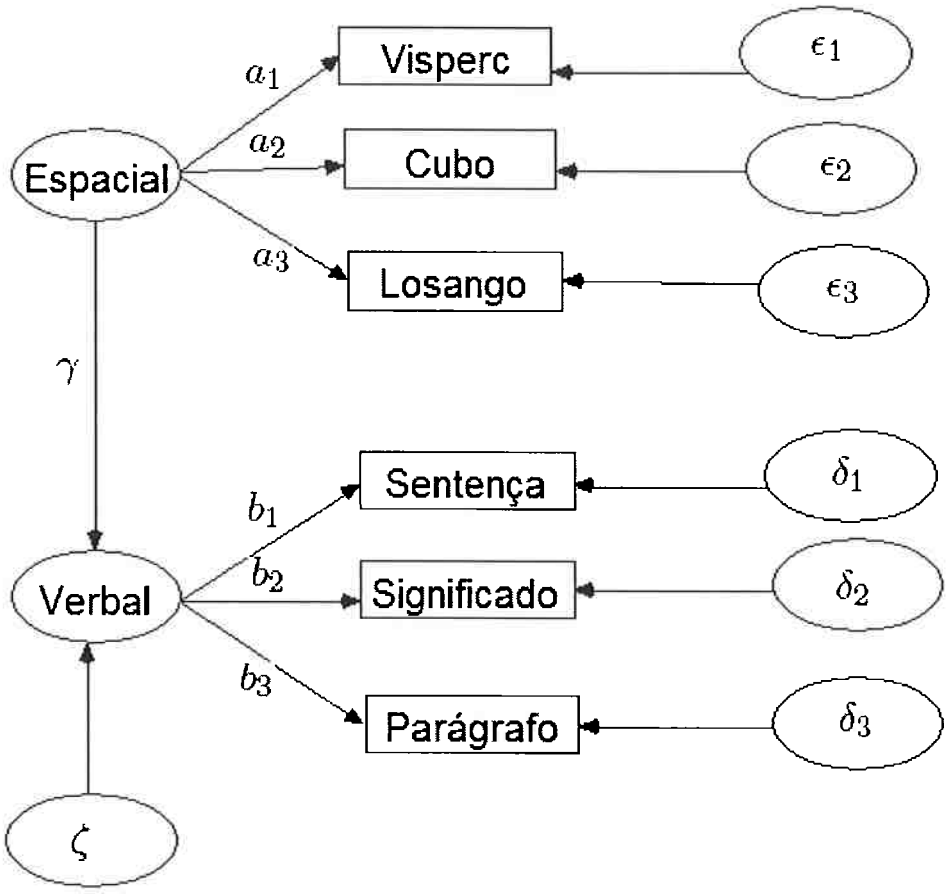

Figura 5.13: Diagrama de caminhos para os dados de habilidade verbal e espacial

(5.17)

Através de uma análise gráfica preliminar, é possível visualizar a estrutura de relação entre as variáves observáveis. A Figura 5.14 apresenta a matriz de gráficos de dispersão das variáveis observáveis, em que é possível notar uma acentuada relação linear dentro dos blocos referentes aos fatores Espacial e Verbal, ao passo que o cruzamento entre as variáveis observáveis destes blocos exibem menor evidência de relação linear.

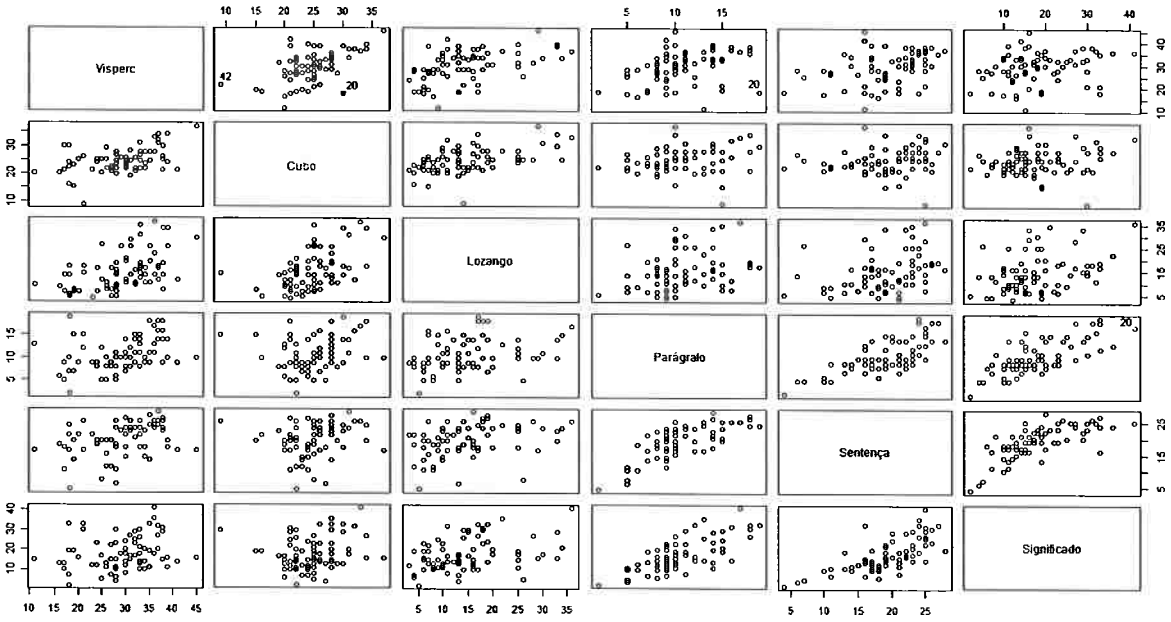

Figura 5.14: Gráficos de dispersão para os dados de habilidade verbal e espacial

Com intuito de avaliar a suposição de normalidade multivariada dos dados, e também destacar observações *outliers*, foi utilizado o gráfico QQ plot, cuja construção é detalhada na Seção 2.1.

A Figura 5.15 exibe o gráfico QQ plot obtido do conjuto de dados bruto em que, a menos dos quantis 72 e 73 em destaque, relativos respectivamente às observações 20 e 42 do conjunto de dados brutos, a suposição de normalidade multivariada ainda pode ser adequada.

Tabela 5.5: Dados de habilidade verbal e espacial

Observações	Visperc	Cubo	Losango	Parágrafo	Sentença	Significado
1	33	22	17	8	17	10
2	30	25	20	10	23	18
3	36	33	36	17	25	41
4	28	25	9	10	18	11
5	30	25	11	11	21	8
6	20	25	6	9	21	16
7	17	21	6	5	10	10
8	33	31	30	11	23	18
9	30	22	20	8	17	20
10	36	28	22	13	24	36
11	30	24	19	14	26	24
12	33	27	16	8	17	13
13	32	22	15	9	20	17
14	27	23	4	9	11	7
15	17	30	13	9	17	13
16	38	25	13	9	23	15
17	34	28	10	9	22	20
18	18	22	5	2	4	2
19	16	20	8	6	18	13
20	18	30	17	19	24	33
21	32	21	9	15	20	25
22	28	20	14	8	18	10
23	39	24	25	14	17	11
24	32	26	10	11	23	23
25	28	24	8	6	10	10
26	31	19	13	8	22	17
27	29	26	25	10	22	18
28	25	25	26	5	7	5
29	37	31	16	14	28	20
30	30	26	19	7	21	14
31	37	34	17	18	24	27
32	28	20	4	8	18	6
33	38	21	18	18	26	31
34	28	27	13	5	6	4
35	11	20	9	13	16	15
36	45	37	29	10	16	16
37	19	15	7	15	19	19

Tabela 5.6: Dados de habilidade verbal e espacial (continuação da Tabela 5.5)

Observações	Visperc	Cubo	Losango	Parágrafo	Sentença	Significado
38	41	21	11	9	16	14
39	34	24	16	9	14	10
40	19	23	6	7	13	19
41	27	21	7	9	19	22
42	21	9	14	15	25	30
43	26	22	8	5	11	11
44	33	22	14	13	25	26
45	38	26	7	16	25	15
46	33	25	34	15	23	21
47	31	25	28	10	25	27
48	36	28	13	11	23	18
49	28	24	6	8	17	15
50	35	25	25	11	24	19
51	23	21	3	9	21	12
52	31	28	25	12	23	15
53	33	28	18	15	23	30
54	21	26	17	12	16	33
55	37	32	11	16	25	32
56	39	34	33	10	19	16
57	18	16	4	10	21	19
58	28	28	18	10	25	23
59	25	20	11	12	19	17
60	32	21	15	14	20	29
61	35	28	19	18	27	33
62	38	30	33	14	22	29
63	26	25	7	10	15	13
64	33	28	12	14	21	14
65	27	22	16	8	19	15
66	26	29	10	8	19	13
67	24	24	16	8	18	12
68	30	23	10	8	14	12
69	24	25	6	9	19	23
70	34	22	13	10	17	14
71	35	28	10	9	13	11
72	18	24	13	7	16	7
73	28	22	15	11	23	30

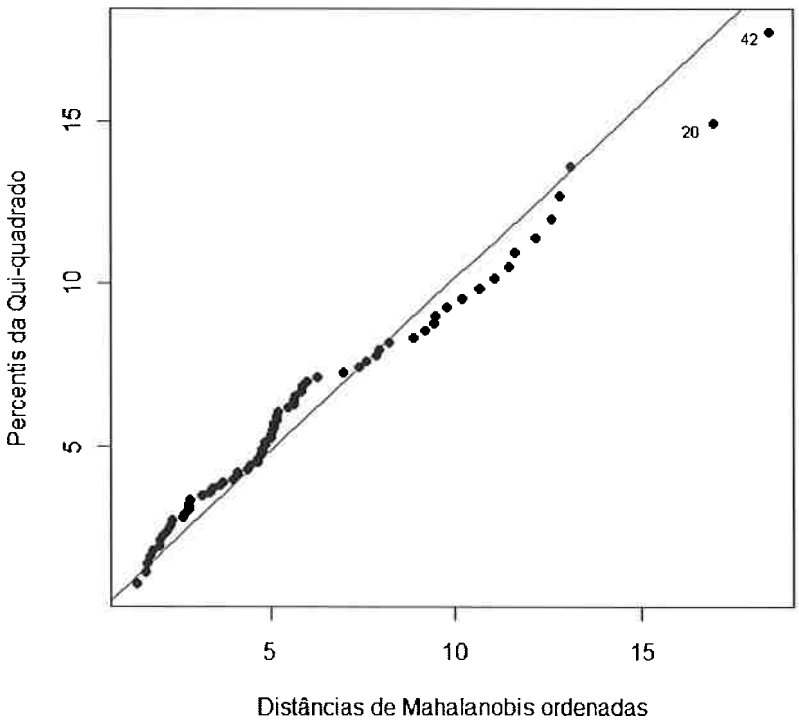

Figura 5.15: QQ plot das distâncias de Mahalanobis para os dados sobre habilidade verbal e espacial

Tabela 5.7: Estatísticas descritivas dos dados sobre habilidade verbal e espacial

Variáveis	Mínimo	Quartil 1	Mediana	Média	Quartil 3	IQ	Máximo
Visperc	11,0	26,0	30,0	29,3	34,0	8,0	45,0
Cubo	9,0	22,0	25,0	24,7	28,0	6,0	37,0
Losango	3,0	9,0	13,0	14,8	18,0	9,0	36,0
Parágrafo	2,0	8,0	10,0	10,6	13,0	5,0	19,0
Sentença	4,0	17,0	20,0	19,3	23,0	6,0	28,0
Significado	2,0	12,0	16,0	18,0	23,0	11,0	41,0

Avaliando o comportamento das distâncias de Mahalanobis escalonadas, D^2_{im}, em que m_0 é composto das 22 observações mais próximas ao centro dos dados, na Figura 5.16, é possível notar que as observações 20 e 42 ao fim do processo *forward search* tomam valores próximos ao padrão observado nos demais dados.

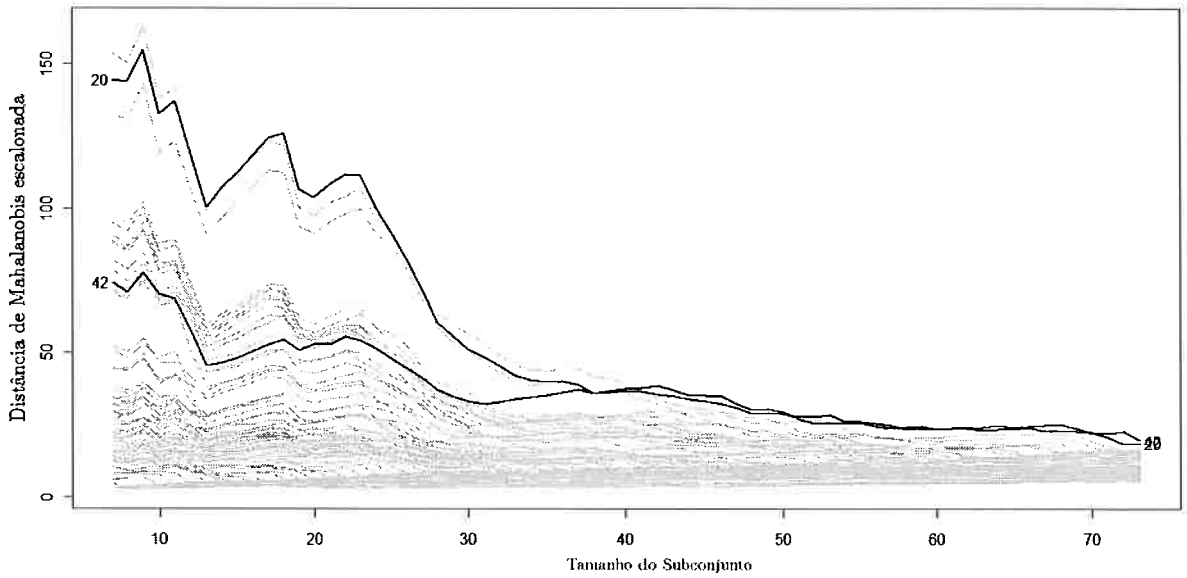

Figura 5.16: Gráfico *forward search* das distâncias de Mahalanobis para os dados de habilidade verbal e espacial

Passando ao estudo da estrutura de variância e covariância, a Figura 5.17 traz a evolução da variância generalizada ($|\mathbf{S}|$) da matriz $\mathbf{S}$ a cada passo do método. Pode-se notar que a variância generalizada se mantém sem grandes mudanças até os grupos de tamanho próximo a 40. Posteriormente é observado maior variação dos valores. Por fim, o gráfico exibe pico com entrada das observação 42, no passo $m = 73$, precedida da observação 20 que entra no passo $m = 72$.

Para maiores detalhes do comportamento da estrutura de variância e covariância dos dados, a Figura 5.18 exibe o histórico das medidas de variância e covariância das variáveis observáveis. Nota-se que as variáveis Significado, Losango e Visperc ao fim do processo, possuem valores de variância em um patamar maior que o do restante das variáveis, o que leva a concluir que estas variáveis são

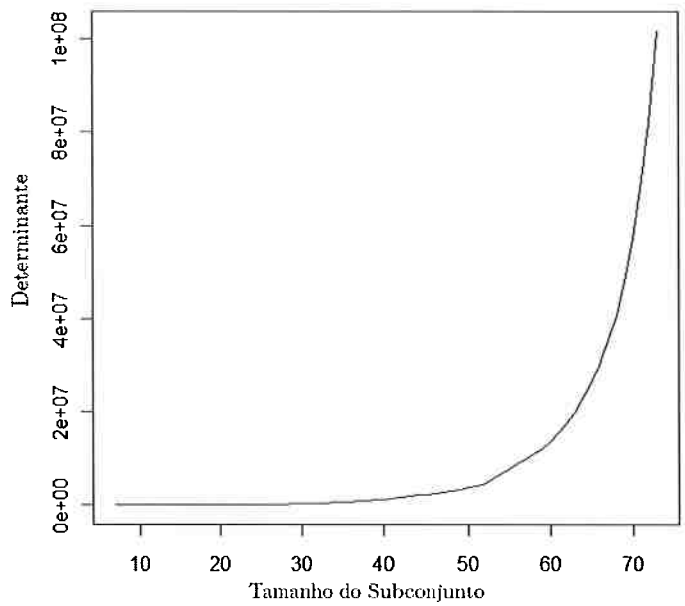

Figura 5.17: Gráfico *forward search* para as variâncias generalizadas para os dados de habilidade verbal e espacial

sensíveis às observações mais distantes do *grupo limpo*, ao passo que as variáveis Sentença, Cubo e Parágrafo se mantém no mesmo patamar de variabilidade ao longo do processo. Quanto às medidas de covariância não há grandes evidências de influência das observações distantes do *grupo limpo*.

Dada a análise descritiva gráfica preliminar, a atenção é voltada ao monitoramento das estatísticas de ajuste do modelo proposto.

A Tabela 5.8 exibe as estatísticas de ajuste do modelo ao conjunto total dos dados e a Figura 5.19 exibe o diagrama de caminhos com as estimativas dos parâmetros e os repectivos erros padrões entre parênteses.

Com base nas estatísticas de ajuste conclui-se que o modelo está adequado aos dados, uma vez que a estatística de χ^2_8 resultante da equação (5.10) tem o nível descritivo igual a 0,447, o que leva

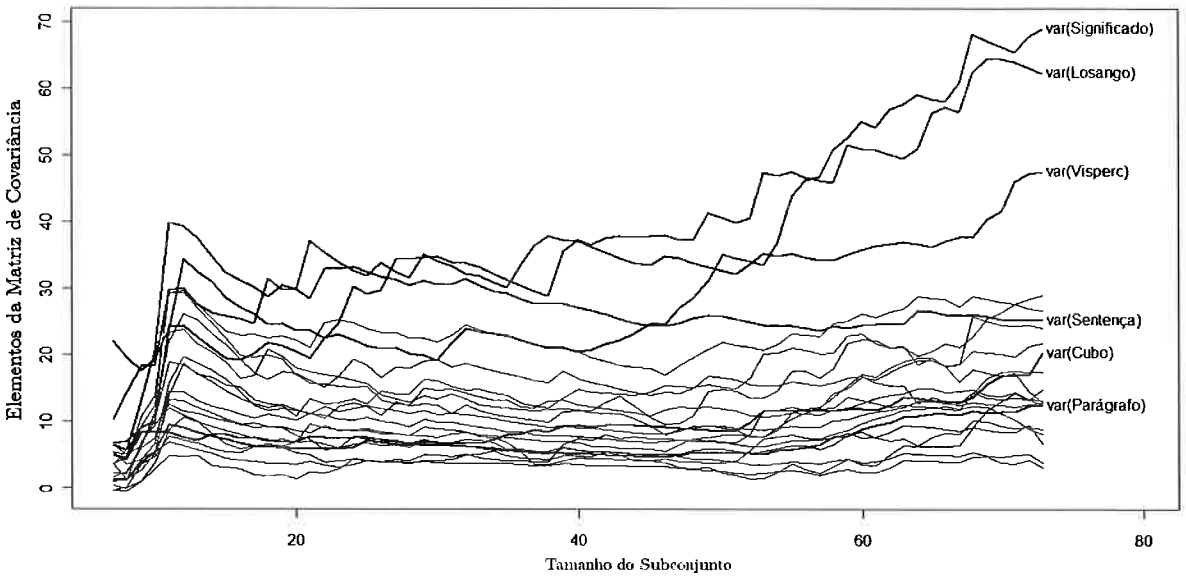

Figura 5.18: Gráfico *forward search* dos elementos da matriz de covariância para os dados de habilidade verbal e espacial

Tabela 5.8: Estatísticas de ajuste do modelo

Estatísticas	Valor
χ^2_8	7,853
GFI_{MV}	0,966
AGFI_{MV}	0,910
SRMR	0,043

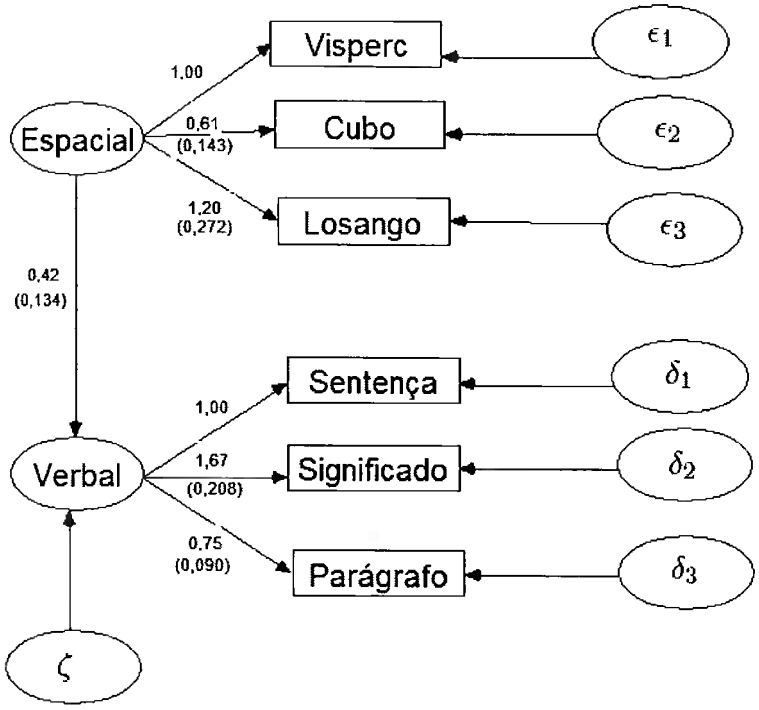

Figura 5.19: Diagrama de caminhos com as estimativas dos parâmetros para os dados de habilidade verbal e espacial

a não rejeitar a hipótese de que $\boldsymbol{\Sigma}^* = \mathbf{S}$. A medida SRMR e os índices de qualidade de ajuste confirmam este fato.

A Figura 5.20 exibe o comportamento do valor da estatística χ^2_8 ao longo do processo; observa-se que a entrada de observações acrescidas ao *grupo limpo* leva à diminuição do valor da estatística, refletindo no melhoramento do ajuste do modelo. Paralelamente, a Figura 5.21 exibe os valores da medida SRMR durante o processo, o que conduz à confirmação do comportamento observado na Figura 5.20. A Figura 5.22 exibe o comportamento dos índices de ajuste GFI_{MV} e $AGFI_{MV}$, que

apoiam o mencionado nas figuras anteriores.

Tal comportamento das medidas de qualidade de ajuste seguem a mesma explicação do ocorrido na aplicação anterior, em que a composição do conjunto inicial não apresenta uma estrutura de dependência razoável para o ajuste do modelo; tal estrutura surge com a entrada de observações nas etapas seguintes do processo, o que dará a forma da estrutura de correlação existente no conjunto de dados, consequentemente levando a melhores valores das medidas de ajuste nestas etapas do processo.

Com relação à Figura 5.20 observa-se que o descréscimo da estatística qui-quadrado é acompanhado de tímidas oscilações. Os grupos de tamanho $m = 61$, $m = 62$, $m = 63$ e $m = 64$, relativos repectivamente à entrada das observações 45, 55, 31 e 28 , tomam valores da estatística qui-quadrado maiores que o padrão observado nesta região do gráfico, o que caracteriza um agrupamento de observações *outliers*. Ainda tem-se que a observação 45, na Figura 5.22, corresponde ao maior déficit nos índices de ajuste GFI_{MV} e $AGFI_{MV}$ dentre as observações deste agrupamento.

É notório que a observação 20, inserida no passo $m = 72$, apresenta uma acentuada queda na estatística qui-quadrado, indicando um bom ajuste ao modelo, e a observação 42, inserida logo após, apresenta similar comportamento, mesmo contribuindo com pequeno decréscimo na qualidade do ajuste.

Embora esses pontos tenham sido os últimos a serem incluídos durante o processo, por apresentarem as maiores distâncias de Mahalanobis, eles encontram-se inseridos no miolo da nuvem de pontos dos diagramas de dispersão da Figura 5.14, quando cruzadas as variáveis Visperc, Cubo e Losango ou Sentença, Significado e Parágrafo, duas a duas, significando que seguem o mesmo padrão de correlação das demais observações. A exceção é o diagrama de dispersão de Visperc x Cubo, em que as observações 20 e 42 estão ligeiramente fora do padrão, mas isso não faz com que o ajuste do modelo seja afetado.

Essas conclusões são reforçadas pelo comentário sobre a Figura 5.18 feito anteriormente, de que não há influência das observações distantes do *grupo limpo* sobre as covariâncias, que são as medidas importantes para o ajuste do modelo de equações estruturais.

Como conclusão, o modelo parece bem ajustado e nenhuma das observações deve ser excluída.

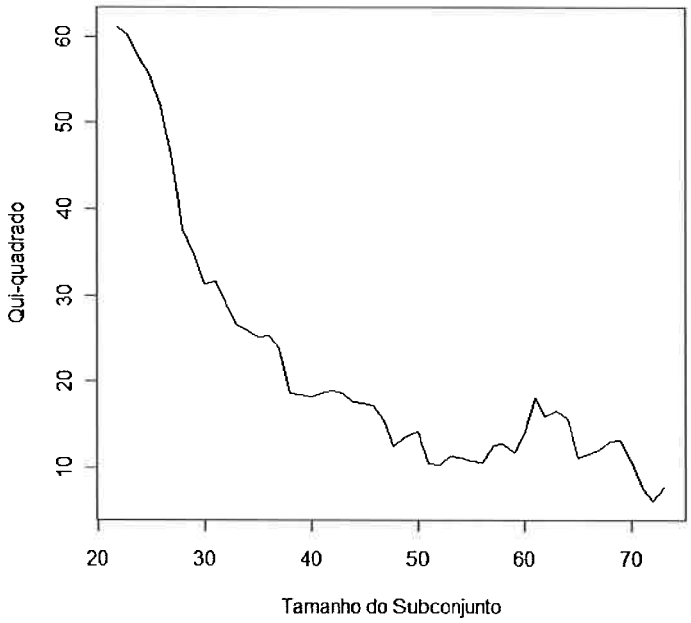

Figura 5.20: Gráfico *forward search* da estatística qui-quadrado para os dados de habilidade verbal e espacial

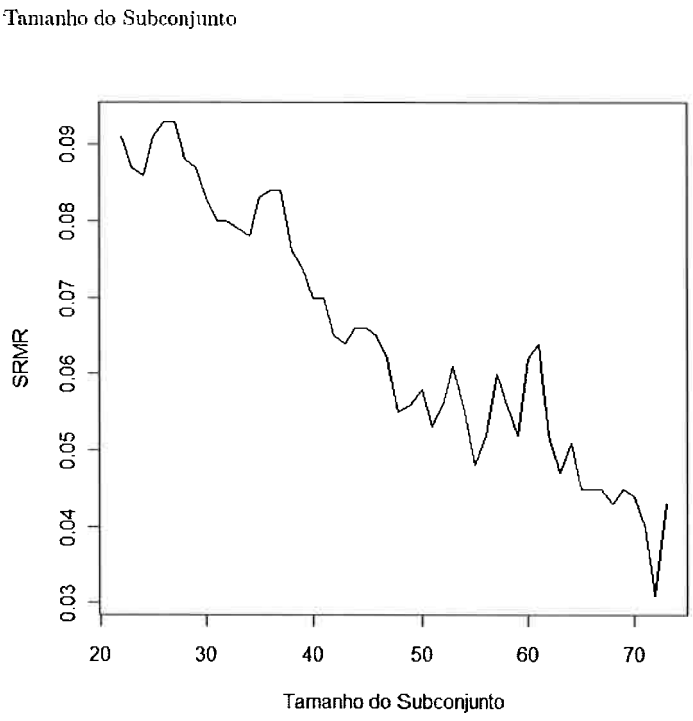

Figura 5.21: Gráfico *forward search* para a medida SRMR para os dados de habilidade verbal e espacial

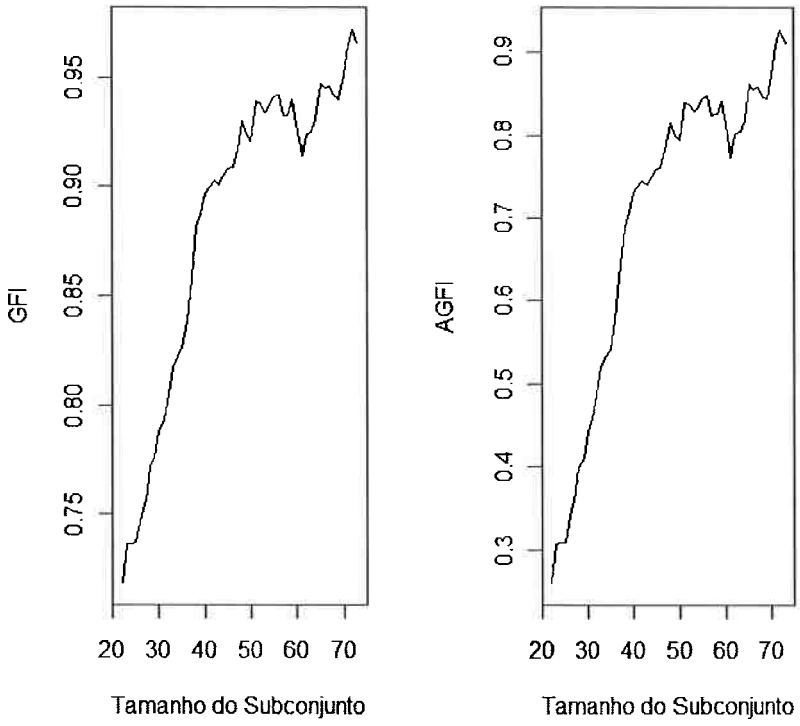

Figura 5.22: Gráfico *forward search* para os índices de qualidade de ajuste GFI_{MV} (à esquerda) e $AGFI_{MV}$ (à diretia) para os dados de habilidade verval e espacial

Capítulo 6

Conclusões

Neste trabalho foi apresentada a técnica de detecção de *outliers* multivariados pelo método *forward search*, que se configura como uma proposta de melhoramento das técnicas de identificação de tal tipo de dado, em que a ênfase está na parte gráfica que tem a finalidade de explicitar a presença do dado discrepante.

A estrutura do algoritmo do método *forward search* foi eficaz na abordagem descritiva do Capítulo 2 e também no contexto de modelos de regressão linear e linear generalizado, apresentados respectivamente nos Capítulos 3 e 4, em que além de identificar as observações *outliers* foi possível avaliar o impacto das observações discrepantes nas estimativas dos parâmetros e no ajuste dos modelos.

Para a abordagem do Capítulo 5, conclui-se que na aplicação aos dados hipotéticos, o método foi coerente com a teoria que envolve o modelo proposto, em que a qualidade do ajuste está ligada à estrutura de correlação existente no conjunto de dados em estudo. A inserção de observações distantes da nuvem de dados que compõem esta estrutura de dependência ocasionou a queda da qualidade do ajuste, o que era esperado.

Na abordagem do método para um modelo estrutural ajustado ao conjunto de dados reais referente a habilidade verbal e espacial, foi observado um desempenho do método *forward search* também coerente com a teoria do modelo estrutural, em que também é exigida uma razoável estrutura de correlação entre as variáveis observáveis. A adaptação do método *forward search* aponta um ajuste insatisfatório no início do processo, fato ocasionado pela homogeneidade dos dados do *grupo limpo* e pela escolha da distância de Mahalanobis para determiná-lo. Isso, entretanto, não afeta o desenvolvimento do método à medida que as observações vão sendo incluídas. Como justificado no caso do ajuste de equações de mensuração da aplicação dos dados hipotéticos, as medidas de ajuste melhoram

nos passos posteriores.

Destaca-se ainda a identificação do agrupamento de valores *outliers* durante o processo, e o fato de as observações 20 e 42, assinaladas como discrepantes, não afetarem o ajuste do modelo, por seguirem o mesmo padrão de correlação das demais observações.

Há que se destacar que o método é descritivo e pode ser utilizado para diagnóstico em outros modelos que não os destacados neste texto. Para isso, basta escolher de forma adequada, as medidas a serem acompanhadas durante o processo de inclusão das unidades experimentais.

Capítulo 7

Apêndice

Comandos do Pacote *Rfwdmv* para Análise Descritiva

fwdmvPrePlot(dados, panel = panel.me,plot.diagonal=T) # Construção dos boxplots bivariados.

fwdmvMinmaxPlot(dados, psfrag.labels =T) # Obtenção dos gráficos para as distâncias mínimas e máxima de Mahalanobis.

fwdmvDistancePlot(dados) # Gráfico das distâncias de Mahalanobis escalonadas.

fwdmvDeterminantPlot(dados) # Gráfico para o processo *forward search* da variância generalizada.

fwdmvCovariancePlot(dados) # Gráfico *forward search* para os elementos da matriz de covariância.

Comandos do Pacote *forward* para Ajuste de Modelos de Regressão

mod1< − - fwdlm(Consumo˜ Taxa+Licença+Renda+Estrada,nsamp=¨ best¨)

plot(mod1, which.plots = 5 , squared = F, scaled = F, th.Res = 2, th.Lev = 0.25, sig.Tst =2.021) # Gráfico para os coeficientes.

plot(mod1, which.plots = 6 , squared = F, scaled = F, th.Res = 2, th.Lev = 0.25, sig.Tst =2.021) # Gráfico para a estatística *t*.

plot(mod1, which.plots = 1 , squared = F, scaled = F, th.Res = 2, th.Lev = 0.25, sig.Tst =2.021) # Gráfico para o resíduo escalonado.

plot(mod1, which.plots = 2 , squared = F, scaled = F, th.Res = 2, th.Lev = 0.16, sig.Tst =2.021) # Gráfico para as medidas de alavanca.

plot(mod1, which.plots = 8 , squared = F, scaled = F, th.Res = 2, th.Lev = 0.16, sig.Tst =2.021) # Gráfico para os valores da distância Cook modificada.

mod2 < − fwdglm(Destruidas Carga+Modelo+Experiencia, family=poisson(log), nsamp="all")

plot(mod2,which.plots = 1,th.Res = 2,squared = F) # Gráfico da *deviance*.

plot(mod2,which.plots = 5) # Gráfico para os coeficientes.

plot(mod2,which.plots = 6,sig.Tst =2.056) # Gráfico para a estatística *t*.

plot(mod2,which.plots = 10) # Gráfico para os valores da distância Cook modificada.

plot(mod2,which.plots = 2,th.Lev = 0.27) # Gráfico para as medidas de alavanca.

plot(mod2,which.plots = 8) # Gráfico para função de ligação.

Algoritmo Construído para Aplicação do Capítulo 5

library(sem) # Carregando o pacote para equações estruturais.

mod.st < − specify.model() # Definindo as esquações de mensuração e estrutural

Espacial − > visperc , NA, 1

Espacial − > cubes, lambda12, NA

Espacial − > lozenges , lambda13, NA

Verbal − > sentences, NA, 1

Verbal − > wordmean,lambda14, NA

Verbal − > paragraph,lambda15, NA

Verbal < − Espacial,beta, NA

visperc < − > visperc,v1,NA

cubes < − > cubes,v2,NA

lozenges < − > lozenges,v3,NA

sentences < − > sentences,v4,NA

wordmean < − > wordmean,v5,NA

paragraph < − > paragraph,v6,NA

Espacial < − > Espacial,vl1, NA

Verbal < − > Verbal,vl2,NA

tab < − matrix(0,nrow=52,ncol=5) # Matriz que receberá os valores das estatísticas de ajuste.

colnames(tab) < − c(¨ Chisq¨ ¨ GFI¨ ¨ AGFI¨ ¨ RMSEA¨ ¨ SRMR¨)

S < − vector('list', 52)

k < − 1

for (i in 22:73) { S[[k]] < − matrix(0,nrow=6,ncol=6) # Matrizes de covariância.

S[[k]] < − read.moments(file = paste(¨ G:/matrizes/m¨ ,i, ¨ .txt¨ , sep = ¨ ¨)# Lendo as matrizes de covariância.

names=c('visperc','cubes','lozenges','paragraph','sentences','wordmean'))

sem.mod < − sem(mod.st, S[[k]], i) # Comando para ajuste do modelo.

summary(sem.mod)

tab[k, 1] < − as.numeric(summary(sem.mod)[1]) # Selecionando o campo referente à estatística qui-quadrado.

tab[k, 2] < − as.numeric(summary(sem.mod)[5]) # Selecionando o campo referente à estatística GFI.

tab[k, 3] < − as.numeric(summary(sem.mod)[6]) # Selecionando o campo referente à estatística

AGFI.

tab[k, 5] < − as.numeric(summary(sem.mod)[12]) # Selecionando o campo referente à estatística SRMR.

k < − k + 1 }

Construção dos gráficos

plot(tab[,1],type= ¨ l¨ , xlab=¨ Tamanho do Subconjunto¨ ylab= ¨ Qui-quadrado¨)

plot(tab[,2],type=¨ l¨ , xlab=¨ Tamanho do Subconjunto¨ ylab=¨ GFI¨)

plot(tab[,3],type=¨ l¨ , xlab=¨ Tamanho do Subconjunto¨ ylab=¨ AGFI¨)

plot(tab[,5],type=¨ l¨ , xlab="Tamanho do Subconjunto", ylab=¨ SRMR¨)

Referências Bibliográficas

[1] Arbuckle, J. L. (2005). *Amos User´s Guide - Version 6.0*, AMOS Development Corporation, USA.

[2] Atkinson, A.C. e Riani, M. (2000). *Robust Diagnostic Regression Analysis.* New York: Springer-Verlag.

[3] Barroso, L. P. e Artes, R. (2003) *Análise Multivariada.* 1. ed. Lavras: Universidade Federal de Lavras.

[4] Bussab, W. O e Morettin, P. A. (2006). *Estatística Básica.* São Paulo: Editora Saraiva.

[5] Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.

[6] Cook, R. D. e Weisberg, S. (1982). *Residual and Influence in Regression* New York: Chapman an Hall.

[7] Demétrio, C. G. B. (2002). *Modelos Lineares Generalizados em Experimentação Agronômica.* Piracicaba: ESALQ/USP.

[8] Gray, J. B. (1989). On the use of regression diagnostics. *The Statistician* **38**, 97-105.

[9] Holzinger, K. J. e Swineford F. A. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary Education Monographs*, No. **48.** Chicago. University of Chicago, Dept. of Education.

[10] Johnson, R. A. e Wichern, D. W (1997). *Applied Multivariate Statistical Analysis*, New York: Prentice-Hall, Inc.

[11] Jöreskog, K. G. e Sörbom, D. (1986). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Square Methods*, Scientific Software, Mooresville.

[12] Long, J. S. (1983). *Confirmatory Factor Analysis.* Beverly Hills: Sage Pub.

[13] Mavridis, D. e Moustaki, I. (2008). Detecting Outliers in Factor Analysis Using the Forward Search Algorithm. *Multivariate Behavioral Research*, **43**, 453-475.

[14] Mingoti, S. A. (2005). *Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada.* Belo Horizonte - Minas Gerais: Editora UFMG.

[15] Melhado, T. T. (2004) *Medidas de Ajuste de Modelos de Equações Estruturais*, Dissertação de mestrado, Instituto de Matemática e Estatística - USP.

[16] Montgomery, D. C. e Vining G. G. (2001). *Introduction to Lienar Regression Analysis, Third Edition* New Yourk: John Wiley.

[17] Paula, G. A. (2004). *Modelos de Regressão com Apoio Computacional.* São Paulo: IME/USP.

[18] Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, **79**, 871-880.

[19] Zani, S., Riani, M. e Coberllini, A. (1998). Robust Bivariate Boxplots and Multiple Outlier Detection. *Computational Statistics and Data Analysis*, **28**, 257-270.