Introdução aos Métodos Estatísticos Robustos

Ana M. Pires
João A. Branco

XV Congresso Anual da Sociedade Portuguesa de Estatística
Lisboa, 19 a 21 de Agosto de 2007

Introdução aos Métodos Estatísticos Robustos

Ana M. Pires e João A. Branco

Edições SPE

FICHA TÉCNICA:

Título: Introdução aos Métodos Estatísticos Robustos

Autores: Ana Pires e João Branco

Editora: Sociedade Portuguesa de Estatística

Concepção gráfica da capa: ISCTE - Instituto Superior de Ciências do Trabalho e da Empresa

Produção gráfica e impressão: Instituto Nacional de Estatística

Tiragem: 550 exemplares

ISBN: 978-972-8890-10-0

Depósito legal: 260674/07

Edições SPE

Manuais

- *Introdução à Probabilidade e à Estatística - com complementos de Excel*, por Maria Eugénia Graça Martins

Minicursos

- *Tópicos de Sondagens*, por Paulo Gomes
- *Controlo Estatístico de Qualidade*, por Ivette Gomes e Isabel Barão
- *Modelos Lineares Generalizados*, por Antónia Turkman e Giovani Silva
- *Inferência sobre Localização e Escala*, por Fátima Brilhante, Dinis Pestana, José Rocha e Sílvio Velosa
- *Modelos Heterocedásticos. Aplicações com o software Eviews*, por Daniel Muller
- *Séries Temporais - Modelações Lineares e Não Lineares*, por Esmeralda Gonçalves e Nazaré Mendes Lopes
- *Uma Introdução à Análise de Clusters*, por João Branco
- *Introdução às Equações Diferenciais Estocásticas e Aplicações*, por Carlos Braumann
- *Outliers em Dados Estatísticos*, por Fernando Rosado
- *Introdução aos Métodos Estatísticos Robustos*, por Ana Pires e João Branco

Actas

- *Afirmar a Estatística. Um Desafio para o Século XXI - Actas do VI Congresso Anual da SPE*. C. Paulino, A. Pacheco, A. Pires e F. da Cunha (Ed.)
- *Um Olhar sobre a Estatística - Actas do VII Congresso Anual da SPE*. P. Oliveira e E. Athayde (Ed.)
- *A Estatística em Movimento - Actas do VIII Congresso Anual da SPE*. M. M. Neves, J. Cadima, M. J. Martins e F. Rosado (Ed.)

- *Novos Rumos em Estatística - Actas do IX Congresso Anual da SPE*. L. Carvalho, F. Brilhante e F. Rosado (Ed.)
- *Literacia e Estatística - Actas do X Congresso Anual da SPE*. P. Brito, A. Figueiredo, F. Sousa, P. Teles e F. Rosado (Ed.)
- *Estatística com Acaso e Necessidade - Actas do XI Congresso Anual da SPE*. P. Rodrigues, E. Rebelo, e F. Rosado (Ed.)
- *Estatística Jubilar - Actas do XII Congresso Anual da SPE*. C. A. Braumann, P. Infante, M. M. Oliveira, R. Alpizar-Jara e F. Rosado (Ed.)
- *Ciência Estatística - Actas do XIII Congresso Anual da SPE*. L. Canto e Castro, E. G. Martins, C. Rocha, M. F. Oliveira, M. M. Leal e F. Rosado (Ed.)
- *Estatística: Ciência Interdisciplinar - Actas do XIV Congresso Anual da SPE*. M. E. Ferrão, C. Nunes e C. A. Braumann (Ed.)

História da Estatística

- *Memorial da Sociedade Portuguesa de Estatística*. F. Rosado (Ed.)

Prefácio

Uma questão concreta, um conjunto de dados e um modelo são os principais ingredientes que fazem funcionar a inferência estatística. Esta estrutura tem-se revelado muito apelativa e extremamente útil pois oferece ao utilizador da estatística a possibilidade de atacar uma multitude de problemas da vida real. Porém, esta abordagem introduz uma dificuldade funcional, a que a inferência estatística tradicional não dá relevo, que pode prejudicar seriamente os resultados de uma análise estatística de dados reais. A dificuldade resulta de um desajustamento entre o que se passa na teoria e o que se passa na realidade e traduz-se no facto de as hipóteses necessárias para o modelo funcionar, de acordo com o previsto, raramente serem encontradas nos dados para os quais o modelo foi afinal concebido. Um exemplo claro é o modelo de regressão linear com todas as hipóteses que se lhe costumam associar para que o método dos mínimos quadrados possa realizar eficientemente o seu trabalho de estimação. Sabe-se que se os dados aos quais se pretende ajustar o modelo de regressão não respeitarem as hipóteses que o modelo exige, as estimativas dos mínimos quadrados poderão dar uma ideia falsa sobre a estrutura dos dados. E quantos conjuntos de dados existem que satisfaçam exactamente tais hipóteses?

Cientes desta realidade incontornável os especialistas inquietos lançaram novos olhares sobre a estatística e fizeram nascer a estatística robusta. Na visão da estatística robusta não há um modelo único, formulado em termos de hipóteses rígidas, que explique a estrutura dos dados, mas há sim um conjunto de modelos que se situam numa vizinhança estreita de um modelo ideal e são esses modelos que estão em condições de explicar a estrutura dos dados. Ao adoptar esta estratégia a estatística robusta fica habilitada a controlar os desvios que os dados reais apresentam em relação ao modelo ideal que se lhe quer impor e que a estatística clássica não é capaz de dominar. É, no fundo, a flexibilidade contra a rigidez da visão clássica.

Este texto constitui o material de suporte ao mini-curso que antecede o XV Congresso Anual da Sociedade Portuguesa de Estatística. Depois de um panorama geral da estatística robusta que é dado no Capítulo 1, segue-se, no Capítulo 2, o conjunto de conceitos básicos da estatística robusta e, no Capítulo 3, aborda-se o problema geral da estimação e apresentam-se os estimadores mais conhecidos. No Capítulo 4 pode apreciar-se a actuação da estatística robusta no estudo da regressão, um dos métodos mais correntemente usados em estatística. O Capítulo 5 mostra o papel das ferramentas da estatística robusta e a sua interligação com a estatística clássica na análise de uma aplicação complexa.

O texto contém abundantes ilustrações gráficas e exemplos com dados simulados e dados reais que permitem compreender melhor os conceitos e mostrar a utilidade dos métodos robustos.

Em face do objectivo deste projecto não foram incluídos vários tópicos de muito interesse para a estatística robusta, em particular a análise multivariada onde se têm verificado avanços de grande relevo.

Este livro é em grande medida o resultado do estudo sobre o tema, iniciado pelos autores em 1989 com a participação no segundo curso ECAS (*European Courses in Advanced Statistics*), intitulado "*Robustness in Statistics – Theory and Applications*", que teve lugar em Schloss Reisensburg, Alemanha, de 2 a 7 de Outubro, e do trabalho de investigação desenvolvido desde essa altura (integrado nas actividades do CEMAT – Centro de Matemática e Aplicações do Instituto Superior Técnico, financiado pela Fundação para a Ciência e a Tecnologia, entidade a quem se agradece o apoio).

Gostaríamos de deixar aqui um agradecimento às nossas colegas Conceição Amado, pela leitura atenta dos Capítulos 2 e 3, e Isabel Rodrigues, co-autora do trabalho em que se baseia o Capítulo 5. É evidente que quaisquer gralhas ou incorrecções são da nossa inteira responsabilidade e desde já agradecemos a todos os que queiram comunicar falhas que encontrarem ou contribuir com comentários e sugestões.

Instituto Superior Técnico,
Lisboa, Junho de 2007

João A. Branco
Ana M. Pires

Índice

1

Um panorama da estatística robusta

1.1 Introdução

Embora o objectivo deste texto seja o estudo da estatística robusta é interessante e útil olhar para as interpretações que o termo "robustez" tem noutros contextos. Como se verá este exercício tem também o mérito de proporcionar uma melhor compreensão do conceito de robustez.

A palavra robustez é usada na linguagem corrente para designar a qualidade daquilo que é robusto (termo derivado da palavra latina *robus*). Por sua vez o significado do termo robusto é (ver Stigler, 1973) entre outros: forte, vigoroso, resistente, saudável, bem constituído. Por exemplo, tanto se fala de um atleta robusto como se fala de um vinho robusto.

Do ponto de vista técnico o termo robusto aplica-se geralmente a um sistema ou processo que é capaz de manter determinado comportamento mesmo quando haja perturbação das condições habituais do seu funcionamento.

Este entendimento genérico do que é um sistema robusto assume interpretações específicas em cada uma das áreas em que o conceito de robustez é relevante. Estudos de robustez são de facto importantes em muitas áreas de trabalho como as ciências naturais, a engenharia, a sociologia, a estatística e outras áreas. É comum considerar-se a robustez de um processo de produção (engenharia), de um *software* (informática), de um sistema social (sociologia), de uma medida económica (economia), de um ecossistema (ecologia), de uma decisão política (política) e de um procedimento estatístico (estatística). Traduzindo a ideia geral de robustez nos vários contextos par-

ticulares leva a que se diga, por exemplo, que: (i) um ecossistema é robusto (Gunderson e Holling, 2001) se tem a capacidade de manter as suas funções e equilíbrio, quando submetido a perturbações, sejam elas ambientais, invasões por espécies diferentes ou outras, (ii) um *software* é robusto (Huhns e Holderfield, 2002) se funciona correctamente dentro das especificações dos seus programas e é ainda capaz de reagir bem em circunstâncias estranhas, fora das especificações definidas no seu delineamento (é capaz, por exemplo, de tolerar um grande volume de erros e *bugs* detectando essas faltas e recuperando dos seus efeitos). Muitos outros exemplos podem encontrar-se em `www.santafe.edu` (Santa Fe Insitute) onde há várias linhas de investigação dedicadas ao tema geral da robustez (em sistemas físicos, em sistemas biológicos, em sistemas sociais, etc.).

Uma outra área em que o conceito de robustez é muito requerido é na área do processo de apoio à decisão. Aí consideram-se vários tipos de robustez e a expressão "análise de robustez" tornou-se parte de muitos estudos nesta área. A análise de robustez pode ter várias interpretações. Segundo Rosenhead (2001) trata-se de um procedimento que consiste em escolher uma acção, de um conjunto de acções possíveis, que seja suficientemente flexível para poder servir muitas das opções que possam surgir no futuro. Por exemplo, sabendo que vou viajar daqui a dois anos, o que devo fazer relativamente a reservar/comprar já o bilhete e que tipo de bilhete, para me resguardar das surpresas que o meu próprio futuro e o do país para onde vou viajar me reservam? Poderei concretizar a viagem e a minha situação económica actual manter-se-á? Poderá acontecer, no país que vou visitar, uma calamidade nacional ou uma mudança política com consequências na alteração do valor do câmbio local?

Como já se percebeu o conceito de robustez não é único, havendo várias interpretações. Aparece muitas vezes ligado às ideias de flexibilidade, insensibilidade, resistência e estabilidade, com as quais por vezes se confunde (ver, por exemplo, Jen, 2003). Em qualquer dos casos o estudo da robustez tem por objectivo principal a construção de sistemas robustos, proporcionando concomitantemente uma melhor compreensão de todo o sistema e um conhecimento das hipóteses cruciais para o seu bom funcionamento, um suproduto da maior importância e que não deve ser menosprezado. A robustez está associada à ideia de tranquilidade do funcionamento ou até de sobrevivência do próprio sistema.

A necessidade de sistemas e processos robustos é muito evidente em diversas actividades. Pensando, por exemplo, na área da visão, em robótica, um bom (robusto) robot de reconhecimento deve ser capaz de absorver e interpretar novos objectos e cenários que não lhe sejam familiares. Trata-se de uma situação em que o número de *outliers* (objectos novos) pode ser muito maior do que o número de observações boas, um caso que poderá deixar muitos estatísticos surpreendidos. No caso da estatística o uso de métodos robustos, isto é, métodos que se comportam bem mesmo quando as hipóteses ideais de funcionamento não são rigorosamente verificadas, é muito apreciado pelos estatísticos uma vez que a aplicação forçada dos métodos tradicionais, quando se verifica a violação daquelas hipóteses, pode conduzir a resultados bastante insatisfatórios.

Este capítulo dá uma visão geral da estatística robusta, destacando os seus aspectos mais importantes, sendo o seu conteúdo parcialmente baseado em três estudos (Pires, 1990; Pires e Branco, 1994; Branco, 2005) cujo material foi em parte alterado e adaptado para ser incluído no presente texto. Após esta introdução começa-se por uma breve referência histórica sobre o aparecimento e desenvolvimento do conceito de robustez em estatística justificando-se simultaneamente a necessidade dos métodos estatísticos robustos. A relação que os métodos robustos têm com outros métodos estatísticos é descrita na secção seguinte. Depois discute-se o papel que a estatística robusta tem na resolução de problemas práticos. Finalmente faz-se um breve comentário relativamente à situação presente e ao futuro da estatística robusta.

1.2 Motivação e nota histórica

Os ingredientes de uma análise estatística incluem, em geral, um conjunto de dados, um modelo e procedimentos estatísticos vários (testes e métodos de estimação). O bom funcionamento destes procedimentos requer que se respeitem certas hipóteses como, por exemplo, a normalidade das observações, a sua independência e identidade em termos de distribuição (i.i.d.), e ainda homogeneidade de variâncias, linearidade e estacionaridade. Se alguma ou várias destas hipóteses forem violadas os resultados de muitos dos procedimentos estatísticos clássicos podem tornar-se tão aberrantes que deixam de merecer

qualquer credibilidade. Procedimentos com este comportamento são designados de não robustos. Por sua vez procedimentos robustos são aqueles cujos resultados não mostram grandes alterações em presença de pequenos desvios das hipóteses assumidas. Tendo em conta que as hipóteses consideradas não passam de simples idealizações e que realmente não se verificam na prática percebe-se imediatamente a importância que têm os procedimentos robustos em qualquer análise estatística.

Antes de prosseguir vale a pena analisar dois exemplos simples que servem para ilustrar os efeitos severos que pequenos desvios da hipótese da normalidade podem acarretar.

Exemplo 1.1. Considere-se uma variável aleatória (v.a.) com distribuição normal, $X \sim \mathcal{N}(\mu, \sigma^2)$, e a partir dela construa-se uma v.a. com distribuição designada normal simetricamente contaminada, $X_{\varepsilon,k}$, cuja função de distribuição (f.d.) é

$$F(x) = (1-\varepsilon)\,\Phi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\,\Phi\left(\frac{x-\mu}{k\sigma}\right), \tag{1.1}$$

onde Φ representa a f.d. da normal reduzida, ε ($0 \leq \varepsilon \leq 1$) a probabilidade de contaminação e $k > 0$, o factor de escala. $X_{\varepsilon,k}$ corresponde assim a um modelo de mistura em que se assume que cada observação provém com probabilidade $1-\varepsilon$ da distribuição de X, $\mathcal{N}(\mu, \sigma^2)$, e com probabilidade ε da distribuição de X_k, $\mathcal{N}(\mu, k^2\sigma^2)$. Comece-se por calcular o valor esperado e a variância de $X_{\varepsilon,k}$, recorrendo à sua função densidade de probabilidade (f.d.p.) dada por

$$f(x) = (1-\varepsilon)\frac{1}{\sigma}\,\varphi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\frac{1}{k\sigma}\,\varphi\left(\frac{x-\mu}{k\sigma}\right),$$

onde φ representa a f.d.p. da normal reduzida. Tem-se então

$$E[X_{\varepsilon,k}] = \int_{-\infty}^{+\infty}\left[(1-\varepsilon)\frac{x}{\sigma}\,\varphi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\frac{x}{k\sigma}\,\varphi\left(\frac{x-\mu}{k\sigma}\right)\right]dx =$$

$$= (1-\varepsilon)E[X] + \varepsilon E[X_k] = \mu,$$

ou seja, o valor esperado do modelo contaminado é igual ao valor esperado do modelo não contaminado. Já em relação à variância não

acontece o mesmo, pois

$$\mathrm{var}(X_{\varepsilon,k}) =$$

$$= \int_{-\infty}^{+\infty} \left[(1-\varepsilon)\frac{x^2}{\sigma}\,\varphi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\frac{x^2}{k\sigma}\,\varphi\left(\frac{x-\mu}{k\sigma}\right)\right] dx - \mu^2 =$$

$$= (1-\varepsilon)E[X^2] + \varepsilon E[X_k^2] - \mu^2 = (1-\varepsilon+\varepsilon k^2)\sigma^2.$$

A nova variável aleatória, embora não sendo normal,[1] desvia-se pouco da normal $X \sim \mathcal{N}(\mu, \sigma^2)$. De facto a distância de Kolmogorov, usada para comparar duas distribuições,

$$d_{\varepsilon,k} = \max_x \left| F(x) - \Phi\left(\frac{x-\mu}{\sigma}\right)\right|, \tag{1.2}$$

é muito pequena (pode mostrar-se que é inferior a $\varepsilon/2$ quaisquer que sejam os valores dos outros parâmetros) tomando, por exemplo, o valor 0.024, para $k = 3$ e $\varepsilon = 0.1$.

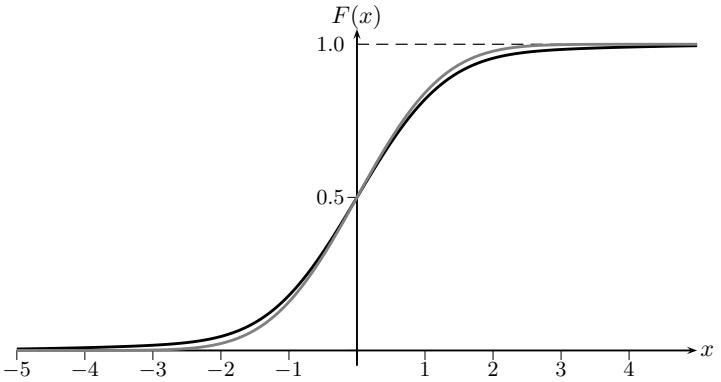


Figura 1.1 *$F(x)$ dada por (1.1), com $\varepsilon = 0.1$, $k = 3$, $\mu = 0$ e $\sigma = 1$, e $\Phi(x)$ (traço menos marcado).*

O gráfico da Figura 1.1 mostra as duas funções de distribuição em causa, $F(x)$, com $\varepsilon = 0.1$, $k = 3$, $\mu = 0$ e $\sigma = 1$, e $\Phi(x)$, podendo

[1] Pode-se mostrar que se $0 < \varepsilon < 1$ e $k \neq 1$, ou seja, com excepção dos casos em que ainda se obtém uma distribuição normal, então a distribuição normal simetricamente contaminada tem caudas mais pesadas do que a distribuição normal.

observar-se a pequena distância entre elas. É de notar que a distância dada em (1.2) não depende de μ e σ. Isso pode perceber-se em termos gráficos pois uma alteração do valor de μ produz apenas uma translação horizontal no gráfico (mudança de origem das abcissas) enquanto que uma alteração no valor de σ corresponde a uma mudança de escala das abcissas em todas as funções envolvidas, não se alterando em qualquer dos casos a diferença máxima em causa.

No entanto os efeitos dessa pequena diferença entre as funções de distribuição são bem visíveis, se observarmos o que se passa com a variância. Usando ainda $\varepsilon = 0.1$ e quatro valores distintos de k, obtêm-se os seguintes valores para a distância de Kolmogorov e para a variância da normal contaminada:

k	$d_{\varepsilon,k}$	$(1-\varepsilon+\varepsilon k^2)\sigma^2$
3	0.024	1.8 σ^2
6	0.035	4.5 σ^2
10	0.040	10.9 σ^2
16	0.043	26.5 σ^2

O que se pode concluir é que a variância é muito sensível a desvios, mesmo que pequenos, efectuados na distribuição, neste caso concreto nas caudas da distribuição. Ou seja, uma pequena proporção da população pode ter efeitos dominantes em certos aspectos da população, como, por exemplo, a variância, o que vai, por sua vez, reflectir-se em certos procedimentos estatísticos como é o caso dos habituais intervalos de confiança. Uma amostra da distribuição contaminada daria intervalos de confiança para o valor médio da população com amplitudes maiores do que as amplitudes produzidas a partir da amostra da distribuição não contaminada, mesmo considerando, como é o caso, que a variância é conhecida (aproximadamente 30% maior se $k = 3$, 2 vezes maior se $k = 6$, 3 vezes maior se $k = 10$ e 5 vezes maior se $k = 16$).

Exemplo 1.2. Quanto a estimadores sabe-se que o desvio padrão amostral é, sob vários aspectos, o estimador óptimo do desvio padrão populacional para populações normais, mas o que acontece quando a população se desvia da normalidade? Para ilustrar a falta de robustez

do desvio padrão amostral,

$$S_n = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2}{n-1}},$$

Tukey (1960) comparou este com o desvio médio (também conhecido por desvio absoluto médio),

$$D_n = \frac{\sum_{i=1}^{n} |X_i - \overline{X}|}{n},$$

para uma amostra aleatória $(X_1, \ldots, X_n)$ de uma população $X_{\varepsilon,3}$ com distribuição normal simetricamente contaminada (1.1) em que o factor de escala é igual a 3 e a probabilidade de contaminação pode variar. Há aqui um pequeno detalhe para o qual convém desde já chamar a atenção, enquanto que o valor esperado de S_n, $E[S_n]$, é aproximadamente igual ao desvio padrão de $X_{\varepsilon,3}$ (como se viu no exemplo anterior este desvio padrão depende de ε e é em geral diferente do parâmetro σ do modelo), o valor esperado de D_n, $E[D_n]$, estará próximo do desvio médio populacional que é distinto do desvio padrão correspondente. Isto significa que os dois estimadores não estão a estimar directamente a mesma característica da população e que a sua comparação não pode ser feita usando o habitual critério da eficiência relativa dada pelo quociente entre os erros quadráticos médios (que se reduz ao quociente entre as variâncias para estimadores centrados). O critério que Tukey (1960) utilizou para comparar estes dois estimadores de dispersão com valores esperados distintos foi uma eficiência relativa assintótica dada pelo quociente entre os quadrados dos coeficientes de variação, ou seja

$$ARE(\varepsilon) = \lim_{n\to\infty} \frac{\mathrm{var}(S_n)/E^2(S_n)}{\mathrm{var}(D_n)/E^2(D_n)} \tag{1.3}$$

o que ao fim e ao cabo corresponde a comparar as variâncias de $S_n/E[S_n]$ e $D_n/E[D_n]$. Na Tabela 1.1 apresentam-se os resultados para alguns valores de ε. A primeira linha da tabela refere-se à situação de não contaminação para a qual se verifica, como esperado, uma superioridade (de cerca de 12%) do desvio padrão. As restantes linhas da metade esquerda da tabela correspondem à situação de contaminação nas caudas enquanto as da metade direita correspondem a contaminação ao "centro", e se aqui não há valores surpreendentes (o desvio padrão continua a ser cerca de 12 % mais eficiente) já o

mesmo não se pode dizer em relação à contaminação nas caudas. De facto, basta uma média de duas observações contaminantes em 1000 ($\varepsilon = 0.002$) para fazer desaparecer a vantagem de S_n, a qual é literalmente pulverizada para $\varepsilon = 0.05$.

Tabela 1.1 *Eficiência assintótica relativa do desvio médio relativamente ao desvio padrão para a distribuição normal simetricamente contaminada com factor de escala 3.*

ε	$ARE(\varepsilon)$	ε	$ARE(\varepsilon)$
0	0.876	1	0.876
0.001	0.948	0.999	0.876
0.002	1.016	0.998	0.876
0.005	1.198	0.995	0.876
0.01	1.439	0.99	0.876
0.02	1.751	0.98	0.876
0.05	2.035	0.95	0.876
0.1	1.903	0.9	0.878
0.15	1.689	0.85	0.882
0.25	1.371	0.75	0.900
0.5	1.017		

O cenário é particularmente inquietante quando se sabe (Huber, 1981) que muitas situações reais são especialmente bem modeladas pelo modelo (1.1) com $0.01 \leq \varepsilon \leq 0.1$ e ainda que, com as dimensões amostrais correntes, se torna virtualmente impossível distinguir a situação de contaminação da situação pura. A conclusão óbvia é que, contrariamente à prática estabelecida, se deveria provavelmente preferir o desvio médio ao desvio padrão. No entanto, apesar da vantagem aqui demonstrada em relação ao desvio padrão, verifica-se no Exemplo 2.9 que o estimador D_n também não é robusto pelo que será necessário estudar outros estimadores para a característica de dispersão (ou parâmetro de escala). No Exemplo 2.9 também são explicados os cálculos que permitem obter os resultados apresentados na Tabela 1.1.

A questão que agora se coloca é a de saber se desvios em relação à normal são frequentes e o recurso a procedimentos robustos é inevitável ou se a normalidade deve prevalecer. Como já se deixou entender

a distribuição normal não se encontra na prática, trata-se de uma distribuição ideal. No entanto o "mito" de que os erros de medição seguem uma distribuição normal encontra-se ainda largamente difundido. A este propósito é interessante a seguinte afirmação:

> *Everyone believed in the normal distribution, the mathematicians because they thought it was an experimental fact, the experimenters because they thought it was a mathematical theorem.* (Lippmann, segundo Poincaré, citado em Stigler, 1975)

Isto deve-se em parte ao Teorema do Limite Central o qual, de facto, apenas fala de um limite, não dando indicação da distância a que se encontra esse limite. Qualquer verificação estatística poderá quando muito provar que os dados têm uma distribuição pertencente a uma ***vizinhança*** da distribuição normal. Vem ainda a propósito referir o modo como a distribuição normal, ou Gaussiana, foi introduzida pelo próprio Gauss. De acordo com Huber (1972), que se baseia numa citação de Gauss, este introduz a distribuição normal como a que mais se ajusta ao uso da média aritmética e não como a distribuição que melhor se ajusta às observações!

Desvios em relação à normalidade ocorrem com frequência na prática e consoante a sua natureza podem separar-se em dois grupos:

(i) Ocorrência de erros: segundo Hampel *et al.* (1986) os conjuntos de dados em que não se consegue encontrar nenhum erro são muito raros. Diversos estudos citados por aqueles autores mostraram que mesmo em dados de alta qualidade a percentagem de erros pode ir até 1% enquanto que para a maior parte dos dados de rotina varia entre 1% e 10%. Este facto, infelizmente, parece ser a regra e não a excepção.

(ii) A distribuição que melhor se ajusta aos dados de alta qualidade tem de facto caudas mais pesadas que a normal (veja-se exemplo em Huber, 1981).

E se tudo o que se disse é verdade para dados univariados com mais propriedade ainda se aplica a dados multivariados (podendo mesmo afirmar-se que a multinormalidade nunca se encontra na prática).

A necessidade de procedimentos robustos foi sentida desde longa data por estatísticos famosos que demonstraram a não robustez de

várias estatísticas e usaram métodos mais robustos (segundo Stigler, 1986, o método baseado na minimização do desvio absoluto médio dos resíduos era usado em regressão antes do método dos mínimos quadrados). Entre outros destacam-se Newcomb (1886), Student (1927), Pearson (1929, 1931), Box (1953), a quem é atribuída a introdução do termo "robustez", e o já referido Tukey (1960) que é bastante radical ao afirmar:

> *Nearly imperceptible non-normalities may make conventional relative efficiencies of estimates of location and scale entirely useless... If contamination is a real possibility (and when is it not?), neither the mean nor variance is likely to be a wisely chosen basis for making estimates from a large sample.*

As vantagens oferecidas pela robustez foram despertando o interesse de outros estatísticos e o conhecimento foi-se acumulando dando lugar a uma teoria própria da estatística robusta. Essa visão moderna da robustez surgiu só na década de sessenta do século XX, pelas mãos de Huber e de Hampel.

Huber (1964) introduz o conceito de vizinhança de um modelo paramétrico e sugere que o comportamento dos procedimentos estatísticos deve ser analisado não só usando o próprio modelo paramétrico mas também os modelos pertencentes a uma vizinhança desse modelo. Partindo do princípio que os modelos estatísticos são apenas aproximações da realidade, conclui que um bom método de estimação é aquele que se comporta bem na vizinhança do modelo paramétrico assumido. Esta ideia é depois explorada de forma a permitir a construção de estimadores robustos com boas propriedades. Hampel (1968) apresenta uma via diferente baseada num instrumento novo, a função de influência, que será amplamente discutido no Capítulo 2.

Pode dizer-se que a estatística robusta é um ramo da estatística que se ocupa do estudo dos efeitos dos desvios em relação a certas hipóteses ideais normalmente assumidas no estudo de modelos estatísticos, encontrando-se parcialmente formalizada em "teorias da robustez", e tendo como principal objectivo a construção de métodos estatísticos robustos. Diz-se que um método estatístico é robusto quando é pouco sensível a pequenos afastamentos em relação àquelas hipóteses ideais.

Como até agora todas as teorias da robustez têm considerado os desvios em relação a modelos paramétricos, pode dizer-se, como Ham-

pel, que:

> *Robust statistics, as a collection of related theories, is the statistics of approximate parametric models. It is thus an extension of classical parametric statistics, taking into account that parametric models are only approximations to reality.* (Hampel *et al.*, 1986)

O aparecimento dos trabalhos de Huber e Hampel dá início a um novo período na história do estudo da robustez. Este período é caracterizado pela divulgação generalizada de conhecimentos sobre robustez e por uma intensa produção científica. Os artigos de Stahel (1991) e de Portnoy e He (2000), com uma diferença de cerca de 10 anos, são elucidativos quanto ao número de publicações produzidas. Surgem mais investigadores na área e começa a haver uma maior comunicação entre eles. Alguns grupos destacam-se pelo interesse continuado nos estudos de robustez como, por exemplo, o grupo suíço do ETH (Swiss Federal Institute of Technology), a escola argentina e certamente o grupo norte-americano e ainda o grupo belga, talvez o grupo mais numeroso e activo na Europa.

Mais recentemente também a comunidade de estatísticos interessados no estudo da robustez sentiu força e dimensão para dar início às conferências anuais ICORS (International Conference on Robust Statistics) iniciadas em 2001 com o objectivo de constituir um fórum para apresentação de novos desenvolvimentos e aplicações da estatística robusta e interacção com outros campos da ciência em geral.

A par de um volumoso número de artigos[2] a produção de livros tem sido notável: para além dos já referidos Huber (1981) e Hampel *et al.* (1986), podem citar-se, por ordem cronológica, Andrews *et al.* (1972); Rey (1978); Launer e Wilkinson (1979); Bustos e James (1980); Bierens (1981); Box *et al.* (1983); Hoaglin *et al.* (1983); Kadane (1984); Rasch e Tiku (1984); Franke *et al.* (1985); Tiku e Balakrishnan (1986); Rousseeuw e Leroy (1987); Kariya e Sinha (1989); Lawrence e Arthur (1990); Staudte e Sheather (1990); Stahel e Weisberg (1991); Marazzi (1993); Morgenthaler *et al.* (1993); Rieder (1994); Huber (1996); Jurečková e Sen (1996); Rieder (1996); Maddala e Rao (1997); Muller (1997); Hettmansperger e McKean (1998); Atkinson e Riani (2000); Insua e Ruggeri (2000); Shevlyakov

[2] 1617 artigos entre 1987 e 2001, de acordo com uma pesquisa no Current Index of Statistics (Ronchetti, 2006).

e Vilchevski (2002); Dutter *et al.* (2003); Hubert *et al.* (2004); Wilcox (2004); Lucas *et al.* (2005); Maronna *et al.* (2006). Além disso têm também sido editados vários números especiais de revistas relacionados com a estatística robusta. Por exemplo, só no ano de 2007, pode dar-se conta de três (Riani *et al.*, 2007; Pires e Souto de Miranda, 2007; Wilcox *et al.*, 2007).

Dos 35 livros citados destacam-se como obras de referência, Huber (1981), Hampel *et al.* (1986), Rousseeuw e Leroy (1987), Staudte e Sheather (1990) e Maronna *et al.* (2006).

1.3 Relações com outros métodos

Os métodos robustos são muitas vezes referidos a par com outros métodos podendo dar a ideia de que são equivalentes. É por isso conveniente clarificar a posição dos métodos robustos especialmente em relação aos métodos não paramétricos e aos métodos de rejeição de *outliers*.

Observa-se também por vezes na literatura o uso indiferenciado dos termos robustez e resistência para designar aparentemente a mesma propriedade. Embora conceptualmente haja distinção — enquanto para falar de robustez é necessário ter um modelo subjacente e um conjunto de critérios bem definidos para a sua avaliação (ver Capítulo 2), para falar de resistência basta apenas um conjunto de observações e uma ideia vaga de estabilidade — não parece haver grande perigo em usar os dois termos como sinónimos.

Quanto à estatística não paramétrica pode-se dizer que ela não tem uma relação directa com a estatística robusta, pois enquanto que na estatística robusta são considerados modelos paramétricos, embora aproximados, na estatística não paramétrica, por seu lado, colocam-se hipóteses que apesar de fracas são exactas, tais como distribuição contínua, distribuição simétrica ou independência das observações. A confusão referida parece ter resultado essencialmente de dois factores: em primeiro lugar o aparecimento da estatística não paramétrica foi em parte uma tentativa de resolver o problema posto pela não robustez de alguns métodos clássicos, e, em segundo lugar, alguns métodos não paramétricos acabam por ter, como se verá no seguimento, propriedades de robustez, pelo que são por vezes aplica-

dos a problemas paramétricos e classificados como robustos.

A relação com os métodos de identificação de *outliers* já não é tão simples. É pretensão da estatística robusta o englobar a questão dos *outliers*, dado que o aparecimento destes pode ser visto como um desvio do modelo ideal, quer através da ocorrência de um erro ou da observação acidental de elementos de uma população distinta, quer como resultado de o modelo exacto ter caudas mais pesadas que o habitual. As teorias existentes na área da detecção de *outliers* preocupam-se com a identificação dos *outliers* deixando ao critério do estatístico a posterior rejeição dessas observações discordantes. Os métodos robustos por seu lado constituem não só um poderoso meio de identificação/diagnóstico como procedem à chamada acomodação dos *outliers* atribuindo-lhes geralmente um peso muito baixo ou mesmo nulo no conjunto da análise.

A aplicação de um bom método robusto é em geral preferível à rejeição de *outliers* seguida da aplicação de um método clássico. Esta afirmação pode ser justificada por, entre outras, as seguintes razões (note-se que algumas das explicações avançadas só ficarão mais claras após a leitura dos capítulos seguintes, em especial dos exemplos):

(i) A maior parte dos testes de detecção de *outliers* são muito dependentes do modelo de geração de *outliers* admitido e não são robustos no sentido em que a ocorrência de um *outlier* pode mascarar totalmente a ocorrência de outros (efeito de mascaramento — *masking effect*) ou tende a arrastar consigo outras observações (*swamping effect*).

(ii) A aplicação de um método robusto procede a uma acomodação "suave" dos *outliers* ao contrário da rejeição que é um processo "abrupto".

(iii) A atribuição de pesos diferentes às diversas observações nada tem de estranho se se pensar que isso é precisamente o que fazem por exemplo os métodos de estimação do tipo L_1 ou o método da máxima verosimilhança quando aplicado a modelos de caudas pesadas, os quais, como se sabe, são susceptíveis de produzir observações aparentemente discordantes.

(iv) O procedimento constituído pela rejeição de *outliers* seguida da aplicação de um método clássico pode ser considerado, no seu

conjunto, um procedimento robusto mas a sua eficiência pode ser bastante inferior à de um bom método robusto.

(v) Em modelos mais estruturados e em especial no caso multivariado os métodos robustos têm provado ser muito mais eficazes como meio de detecção do que qualquer outro método (Rousseeuw e Van Zomeren, 1990).

Apesar de não haver a intenção de instigar a polémica tem-se consciência de que as afirmações aqui feitas em relação aos métodos de rejeição de *outliers* não são aceites pacificamente havendo autores que advogam posições contrárias. Julga-se no entanto que as razões apresentadas justificam plenamente a opção pelos métodos robustos, mesmo que seja apenas como meio de diagnóstico. Uma referência interessante sobre os métodos de rejeição de *outliers* onde também é abordada a relação com os métodos robustos é Beckman e Cook (1983). Outras referências relevantes na mesma área são Barnett e Lewis (1994), Hawkins (1980) e, em português e muito recente, Rosado (2006).

1.4 A estatística robusta na prática

Perante o remédio que a estatística robusta fornece para tratar o deficiente funcionamento de muitos procedimentos estatísticos clássicos, em consequência do irrealismo das hipóteses em que assentam, cabe perguntar se as capacidades da estatística robusta são de facto bem aproveitadas pelos utilizadores na prática. A resposta talvez seja "ainda não". Huber e Hampel introduziram os fundamentos de uma teoria longa e complexa que muitos investigadores continuam a construir, possivelmente atraídos pelas muitas questões em aberto e pelo aliciante trabalho matemático requerido na procura de soluções. Um interesse e esforço semelhantes não parece terem sido devotados à preparação dos resultados dessas teorias para utilização imediata dos praticantes da estatística. Esta preocupação existe desde longa data como se percebe, por exemplo, lendo o artigo "*Do robust estimators work with real data?*" de Stigler (1977), o artigo de Hogg (1979) que questiona "*Why is it that robust methods are not used more frequently today?*", o artigo de Yohai *et al.* (1991) cujo resumo começa com a frase elucidativa "*Even if robust regression estimators have been around for nearly 20 years, they have not found widespread application.*" ou ainda a comunicação de Stromberg (2002) que afirma

"*Hundreds, and perhaps thousands, of papers have been published in the area of robust statistics, yet robust methods are still not used routinely by most applied statisticians.*"

Os exemplos desta preocupação são muitos e as razões para o menor sucesso da utilização da estatística robusta na prática são várias, entre as quais se destacam:

(i) Os métodos robustos disponíveis são em grande número e não há geralmente uma directiva para que o utilizador possa fazer a selecção que lhe convém. No caso da regressão linear, por exemplo, foam criados aproximadamente vinte novos estimadores robustos desde a introdução dos estimadores-M (Stromberg, 2002). E qual deles é o melhor para usar na prática?

(ii) O *software* necessário à operacionalidade dos métodos robustos não está suficientemente divulgado, com excepção feita aos programas disponíveis no S-Plus, R e SAS.[3] No entanto, o *output* dos métodos estatísticos robustos nem sempre é semelhante ao *output* dos métodos clássicos (muitas vezes o valor-p não é fornecido) o que dificulta o trabalho do analista corrente. Recentemente esta situação tem preocupado a comunidade estatística que iniciou a realização de encontros científicos com vista a debater este tema: International Workshop on Robust Statistics and R (primeira edição em 2005, Treviso, Itália, segunda edição em 2007, Banff, Canadá). O objectivo destes encontros é o de promover a organização de uma biblioteca de programas de métodos estatísticos robustos que seja de uso amigável, que possa servir o ensino e a aprendizagem de métodos robustos e que possa ser usada com facilidade por toda a comunidade: professores, alunos ou simples utilizadores.

(iii) A divulgação dos métodos robustos não tem sido eficiente, isto é, feita em termos acessíveis e atractivos para o utilizador. Basta observar que a maior parte dos livros publicados são de índole teórica e os livros sobre aplicações são menos e não são simples (uma excepção é certamente o livro de Rousseeuw e Leroy, 1987) e que a inclusão de conceitos básicos de estatística robusta e de explicações sobre os seus objectivos em livros de es-

[3]Ver S-Plus (2000), S-Plus (2002), R Development Core Team (2006), Maechler (2006), SAS Institute Inc. (2006).

> tatística de nível elementar ou intermédio é praticamente inexistente (a nível nacional registam-se duas excepções, a tradução Hoaglin *et al.*, 1992, e Murteira, 1993).

A falta de robustez de certos procedimentos clássicos e a razão dessa falta de robustez deve ser denunciada cedo para que não continue a perpetuar-se o mito da normalidade e de outras hipóteses irrealistas em que estes procedimentos assentam. A compreensão dos perigos da falta de robustez é, por isso, muito útil para que se possa conduzir com consciência uma análise estatística e talvez mais importante do que o uso inseguro de métodos robustos sofisticados.

Ao comum utilizador da estatística convêm respostas claras relativamente às dúvidas que lhe possa suscitar o funcionamento dos métodos que utiliza no seu dia a dia. Por exemplo, o teste t para comparação de duas médias ou o teste F em que se baseia a análise de variância são robustos?

Pensando no teste t as hipóteses envolvidas são a independência (quer entre as observações, quer entre as duas amostras), a normalidade e a igualdade das variâncias. Quais as consequências se alguma ou várias destas hipóteses são violadas e o que fazer nesse caso? No caso da falha da independência entre as duas amostras o teste não é válido embora possa ser substituído pelo teste t para amostras emparelhadas. No caso da normalidade não se verificar sabe-se que pequenos desvios da normal não perturbam significativamente a estatística t, principalmente se as amostras têm dimensões grandes e aproximadamente iguais, e a hipótese nula é verdadeira. No entanto a potência do teste é muito sensível até a pequenos desvios da normalidade. Pode dizer-se que o teste é robusto em relação a desvios da normalidade (desvios esses que podem assumir a forma de *outliers*), sob H_0, mas pode apresentar-se com fraca potência para certas distribuições não normais, isto é, não é robusto sob H_1. É sabido, no entanto, que o teste é o mais potente no caso das hipóteses ideais se verificarem. A transformação dos dados é por vezes a salvação desta situação indesejável. Outra possibilidade consiste no uso de testes não paramétricos (ou ainda no uso dos métodos de regressão robusta, ver Secção 4.6).

Quanto à hipótese da igualdade das variâncias sabe-se que o teste t é razoavelmente robusto se as dimensões das amostras são iguais. Mas se as dimensões das amostras forem diferentes e a amostra de

menor dimensão tiver a maior variância a falha da igualdade das variâncias pode conduzir a grandes alterações nos resultados. Neste caso aconselha-se o teste t de Welch-Satterhwaite que tem propriedades de robustez semelhantes às do teste t na situação das variâncias serem iguais.

Perante esta descrição, como reagir à afirmação corrente, encontrada principalmente na literatura das aplicações da estatística: os métodos de análise de variância são robustos. Qual o significado desta afirmação? E o que fazer quando se lê?

> *The power (of the t-test) is very sensitive even to small deviations from normality. For not too small samples, there are other tests, such as the Wilcoxon-(Mann-Whitney U)-test, with a much better behavior.* (Hampel, 2000)

> *On the other hand the t-test is so robust against non-normality that there is really no need to use the Wilcoxon test.* (Rasch e Guiard, 2004)

1.5 O presente e o futuro da estatística robusta

Os fundamentos da estatística robusta ficaram estabelecidos na década de sessenta do século XX. Passados cerca de 40 anos é natural querer saber qual foi a evolução da estatística robusta, qual é o seu papel no contexto da ciência estatística actual, qual é a sua situação presente e quais são as suas perspectivas futuras.

O que se segue é uma breve opinião sobre questões que requerem uma reflexão profunda.

Os princípios da estatística robusta foram ao longo deste período amplamente divulgados entre cientistas e embora a produção de literatura científica nesta área tenha sido muito grande, a estatística robusta não conseguiu penetrar ainda ao nível do ensino da estatística, nem conseguiu tornar-se numa nova maneira de fazer estatística, como alguns entusiastas da estatística robusta sonharam.

A acção da estatística robusta tem-se feito sentir sobretudo na análise correctiva de *outliers* continuando a ser relegados para um plano secundário o controlo dos desvios associados à violação de

várias hipóteses fundamentais inerentes à formulação de muitos modelos estatísticos. O desenvolvimento de métodos robustos de estimação paramétrica tem sido intenso, com destaque para a análise multivariada, em detrimento de outros temas essenciais da inferência estatística, como os testes de hipóteses e a estimação de erros padrão que continuam a requerer mais atenção.

Enquanto que os avanços teóricos em estatística robusta têm sido notórios a sua repercussão na aplicação prática é muito menos visível.

O primeiro mérito destes desenvolvimentos teóricos é o de ter influenciado de forma incontestável a estatística moderna. A propriedade "robustez" é hoje tida em conta por muitos consumidores da estatística aplicada e é frequente verificar que em muitos trabalhos há por parte dos seus autores um reconhecimento do interesse da estatística robusta que se traduz pelo menos em discutir a propriedade de robustez dos procedimentos estatísticos adoptados.

Já a utilização directa de métodos robustos em problemas aplicados é menos frequente. Contudo está a notar-se que a estatística robusta é muito apelativa para o trabalho em áreas em que as aplicações envolvem a produção de dados com um elevado grau de ruído. Entre outras áreas destacam-se:

(i) **Visão computacional:** Black e Rangarajan (1996); Ong e Spann (1999); Meer *et al.* (2000) (edição especial da revista *Computer Vision and Image Understanding* sobre o tema *Robust Statistical Techniques in Image Understanding*); Black *et al.* (2000); De La Torre e Black (2001).

(ii) **Quimiometria:** Hubert (2006); Engelen *et al.* (2007);

(iii) **Detecção remota:** Bosdogianni *et al.* (1997); Chau e Parker (2004);

(iv) **Controlo de qualidade:** de Mast e Roes (2004);

(v) **Processamento de imagens:** Black *et al.* (1998);

(vi) **Bioinformática:** Fomenko *et al.* (2006);

(vii) **Alocação de activos:** Zhou (2001).

Outras actividades em que a estatística robusta tem um papel de relevo são: aquelas em que a análise de dados é feita de forma au-

tomática e rotineira com a ajuda de *software* especializado e ainda a actividade de exploração de grandes bases de dados (*data mining*).

São estas áreas e estas actividades, todas elas de certa forma ligadas à acção das novas tecnologias que podem ser o novo motor de desenvolvimento da estatística robusta, pelo proveito que retiram da sua utilização e pelos desafios que lhe colocam.

Vejamos com mais atenção o caso *data mining*. A estatística robusta é especialmente hábil na análise de grandes volumes de dados para os quais a estatística clássica não está preparada. Enquanto que a estatística clássica tem por objectivo ajustar modelos a todo o conjunto de dados, pequeno ou grande, a estatística robusta procura os padrões revelados apenas pela maioria dos dados eliminando ou suavizando a acção daqueles que exercem uma influência extrema. Mas será que os métodos robustos existentes, que foram concebidos para conjuntos de dimensão pequena ou mediana, estão preparados para, em termos teóricos, práticos e computacionais, ser aplicados directamente a grandes volumes de dados? Será viável aplicar os actuais algoritmos, que são tão exigentes computacionalmente, a grandes volumes de dados? Será que a definição de *outlier* não merece ser retocada quando a par de muitos dados a situação envolve também muitas variáveis? Se se considerar que um objecto é *outlier* se o for em alguma das suas variáveis então pode acontecer que mesmo num grande conjunto de objectos haja apenas um pequeno subconjunto de objectos inteiramente não contaminados. Esta situação é contrária à ideia básica da estatística robusta que considera que a maioria dos dados é não contaminada.

Não parece haver dúvidas sobre a necessidade da estatística robusta para *data mining* e sobre os desafios que a área *data mining* lança à estatística robusta. Resta saber se os especialistas em estatística robusta estão interessados em conhecer bem os problemas levantados por *data mining*. Pois se estiverem isso significa que a estatística robusta conhecerá novos desenvolvimentos no futuro e aumentará a sua credibilidade e os seus méritos como metodologia de pensar e fazer estatística.

2
Conceitos básicos

2.1 Introdução

Neste capítulo apresentam-se os principais conceitos utilizados em análise de robustez. Começa-se pelo conceito de sensibilidade por ser o mais simples e intuitivo. A avaliação da sensibilidade, em particular a avaliação da sensibilidade de uma estimativa, ou estimador, efectuada através da chamada curva de sensibilidade, está muito ligada ao conceito de resistência e não necessita de um grande formalismo teórico. Em seguida apresenta-se a função de influência. Trata-se de uma extensão de certa forma natural da curva de sensibilidade mas que requer um enquadramento mais cuidado, nomeadamente a consideração de um modelo paramétrico subjacente e a noção de funcional estatístico. Outro conceito importante discutido neste capítulo é o de ponto de rotura, o qual de certa forma complementa a informação veiculada pela função de influência. É também apresentada, se bem que de uma forma abreviada, a definição de robustez qualitativa.

Para exemplificar estes conceitos recorre-se a estimadores clássicos como a média, a mediana, o desvio padrão ou o desvio médio, sendo simultaneamente apresentados estimadores simples baseados nestes mas com melhores propriedades em termos de robustez (por exemplo médias e desvios padrões aparados). O capítulo termina com uma breve síntese relativa às propriedades desejáveis para um estimador sob o ponto de vista da robustez.

2.2 Curva de sensibilidade

A avaliação da sensibilidade de uma estimativa (T_n) é muito simples. Considere-se uma dada amostra univariada fixa $(x_1, \ldots, x_{n-1})$ e a perturbação dessa amostra por acréscimo de uma nova observação (x). A ideia consiste em calcular a diferença entre os valores das estimativas obtidas na amostra perturbada, $T_n(x_1, \ldots, x_{n-1}, x)$, e na amostra original, $T_{n-1}(x_1, \ldots, x_{n-1})$. Veja-se o que acontece no caso da média aritmética, $T_n = \overline{x}_n$:

$$\begin{aligned} T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1}) = \\ = \frac{x + \sum_{i=1}^{n-1} x_i}{n} - \frac{\sum_{i=1}^{n-1} x_i}{n-1} = \\ = \frac{(n-1)x - \sum_{i=1}^{n-1} x_i}{n(n-1)} = \frac{x - \overline{x}_{n-1}}{n}. \end{aligned}$$

O resultado mistura dois efeitos, o da dimensão da amostra e o da perturbação introduzida na amostra. O mesmo sucede para a grande maioria das estimativas habituais. Para remover o efeito da dimensão da amostra multiplica-se por n a diferença entre as estimativas, obtendo-se então a curva de sensibilidade.

Definição 2.1. *Dada uma amostra de dimensão $n-1$, $(x_1, \ldots, x_{n-1})$, e um estimador $T_n(X_1, \ldots, X_n)$ a curva de sensibilidade do estimador T_n para essa amostra é a seguinte função de x*

$$SC_n(x; T_n) = n\left[T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})\right]. \tag{2.1}$$

A curva de sensibilidade da média aritmética é assim dada por

$$SC_n(x; \overline{X}) = x - \overline{x}.$$

Esta função é ilimitada, mostrando que uma única observação pode ter um efeito devastador sobre este tipo de estimativa da tendência central de um conjunto de dados. Outro tipo de estimadores de

tendência central não sofrem do mesmo problema. É o caso da mediana, ou da média aparada ("trimmed mean") consideradas no Exemplo 2.1. A média aparada a $100 \times \alpha\%$, $0 < \alpha < 1/2$, representada por $\overline{x}_\alpha$, é calculada do modo seguinte: ordenam-se as observações por ordem crescente e em seguida calcula-se a média aritmética das observações centrais, omitindo as $[\alpha n]$ inferiores e as $[\alpha n]$ superiores (onde $[t]$ significa o "maior inteiro contido em t"). Ou seja,

$$\overline{x}_\alpha = \frac{1}{n - 2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} x_{(i)}, \tag{2.2}$$

onde $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ representam as observações ordenadas. Com algum abuso de notação representa-se a mediana por $\overline{x}_{0.5}$ (só é exacto se n for ímpar).

O cálculo da curva de sensibilidade da média aparada é fastidioso pelo que não se apresenta aqui, optando-se por apresentar resultados numéricos no exemplo seguinte. É no entanto fácil concluir que, ao contrário do que sucede para a média aritmética e desde que $[\alpha n] \geq 1$, o efeito de uma observação contaminante (x) nunca pode ser ilimitado, sendo uma função contínua, linear por troços e que é constante para $x > x_{(n-1)}$ e para $x < x_{(1)}$.

No caso das estimativas tradicionais de dispersão, a variância ou o desvio padrão, o resultado é ainda mais dramático do que para a média aritmética. De facto, cálculos elementares permitem mostrar que sendo a variância dada por

$$S_n^2(x_1, \ldots, x_n) = \frac{\sum_{i=1}^n x_i^2 - n\,\overline{x}^2}{n-1}, \tag{2.3}$$

se tem

$$S_n^2(x_1, \ldots, x_{n-1}, x) = s^2 + \frac{(x - \overline{x})^2 - \left(2 - \dfrac{n-2}{n-1}\right) s^2}{n},$$

com $s^2 = S_n^2(x_1, \ldots, x_{n-1})$. É assim imediato que

$$SC_n(x; S_n^2) = (x - \overline{x})^2 - \left(2 - \frac{n-2}{n-1}\right) s^2 \tag{2.4}$$

e que

$$SC_n(x; S_n) = \frac{(x - \overline{x})^2 - \left(2 - \frac{n-2}{n-1}\right) s^2}{\sqrt{s^2 + \frac{(x - \overline{x})^2 - \left(2 - \frac{n-2}{n-1}\right) s^2}{n}} + s}. \qquad (2.5)$$

A curva de sensibilidade da variância é do tipo quadrático, sendo o efeito de uma nova observação x da ordem do quadrado da distância entre x e a média da amostra original $\overline{x}$. No caso do desvio padrão o crescimento da curva de sensibilidade é um pouco atenuado, em relação à curva de sensibilidade da variância, pela existência de um termo adicional no denominador, mas é ainda assim bastante mais rápido que o da curva de sensibilidade da média. No exemplo seguinte ilustram-se numericamente estes aspectos.

Exemplo 2.1. Considere-se a seguinte amostra (Abbey, 1988; Analytical Methods Committee, 1989a,b; Ripley, 2004) de 24 observações relativas à concentração de cobre num certo alimento (em $\mu g\, g^{-1}$):

2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90	3.03	3.03	3.10	3.37
3.40	3.40	3.40	3.50	3.60	3.70	3.70	3.70	3.70	3.77	5.28	28.95

Evidentemente que com uma amostra univariada desta dimensão é fácil constatar, mesmo sem o recurso a qualquer método de diagnóstico, que a última observação se distingue bastante das restantes, pelo que qualquer analista sensato verificaria a sua origem e consideraria a hipótese de a remover antes de prosseguir a análise dos dados, o que é de certa forma equivalente a utilizar, em vez das medidas tradicionais de localização e dispersão (média e desvio padrão, respectivamente), medidas alternativas com menor sensibilidade. As dificuldades surgem, como se disse anteriormente, quando se processam de forma automática, ou semi-automática, centenas ou milhares de conjuntos de dados como este. De qualquer modo, este exemplo pretende apenas ilustrar o cálculo de curvas de sensibilidade num contexto o mais simples possível.

Na Tabela 2.1 apresentam-se diversas medidas de tendência central e de dispersão, calculadas com a amostra completa e apenas com as 23 primeiras observações. De entre as medidas de tendência central seleccionaram-se a média, as médias aparadas a 5% e a 10% e a

mediana. De entre as medidas de dispersão escolheram-se o desvio padrão, o desvio médio e o desvio absoluto mediano. O desvio padrão representa-se por s e é a raiz quadrada da expressão (2.3), ou seja é dado pela média quadrática dos desvios em relação à média. O desvio médio (já referido no Exemplo 1.2, página 6) é usualmente calculado como a média dos desvios absolutos em relação à média, $\sum_{i=1}^{n} |x_i - \overline{x}|/n$, mas aqui opta-se por usar uma expressão que é múltipla desta,

$$d = D(x_1, \ldots, x_n) = \sqrt{\frac{\pi}{2}} \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}.$$

A multiplicação pela constante $\sqrt{\pi/2}$ garante que para $X_i \sim \mathcal{N}(\mu, \sigma^2)$ e n grande $E[D(X_1, ..., X_n)] \simeq \sigma$. Por sua vez o desvio absoluto mediano é geralmente representado pela sigla MAD (de *Median Absolute Deviation*) e é dado pela mediana dos desvios absolutos em relação à mediana, multiplicando depois o resultado por uma constante[1]

$$\text{MAD}(x_1, \ldots, x_n) = 1.4826 \times \text{med}\,|x_i - \text{med}(x_1, \ldots, x_n)| \qquad (2.6)$$

(tal como para d, a multiplicação pela constante $1.4826 \simeq 1/\Phi^{-1}(3/4)$ garante que para $X_i \sim \mathcal{N}(\mu, \sigma^2)$ e n grande $E[\text{MAD}(X_1, ..., X_n)] \simeq \sigma$).

Tabela 2.1 *Diversas medidas de localização e de dispersão para os dados do Exemplo 2.1.*

Amostra	$\overline{x}$	$\overline{x}_{0.05}$	$\overline{x}_{0.1}$	$\overline{x}_{0.5}$	s	d	MAD
Completa	4.280	3.254	3.205	3.385	5.297	2.681	0.526
Sem a última obs.	3.208	3.157	3.175	3.370	0.687	0.657	0.504

Analisando os resultados apresentados na Tabela 2.1 pode verificar-se que a observação número 24 tem um grande efeito sobre a média, o desvio padrão e o desvio médio mas um efeito bastante diminuto nas restantes medidas.

Uma visão mais global do efeito de uma única observação é dada pelas curvas de sensibilidade, as quais podem ser facilmente obtidas

[1] Ver função `mad` do R ou do S-PLUS.

numericamente.[2] Neste exemplo, como já se percebeu que a última observação é atípica, o que faz sentido é calcular as curvas de sensibilidade usando a amostra constituída pelas primeiras 23 observações. Desta forma estuda-se o efeito do presumível *outlier*, enquanto que se se calculasse a curva de sensibilidade com todas as observações se estava a avaliar o efeito de uma outra observação contaminante para além da já detectada. Note-se ainda que, assim, o valor de cada uma das curvas de sensibilidade no ponto 28.95 coincide com a diferença entre os valores da primeira e da segunda linha da Tabela 2.1, na coluna respectiva, multiplicada por 24.

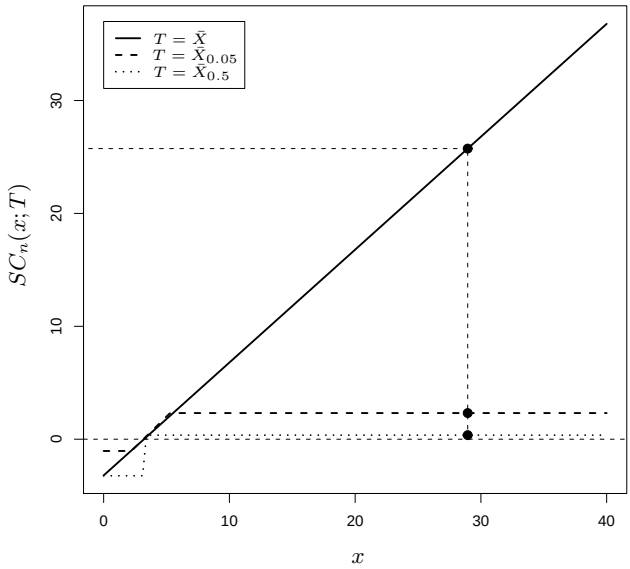


Figura 2.1 *Curvas de sensibilidade de diversos estimadores de localização baseadas nas primeiras 23 observações da amostra do Exemplo 2.1.*

Nas Figuras 2.1 e 2.2 estão representadas as curvas de sensibilidade das medidas de localização e das medidas de dispersão, respectivamente, no intervalo $[0, 40]$. A grande diferença de comportamento

[2]Como se disse, estas curvas poderiam também ter sido calculadas analiticamente, à semelhança do que foi feito para a média, mas o processo é mais trabalhoso e de certa forma desnecessário, dado que as curvas de sensibilidade podem com vantagens ser substituídas pela função de influência que já não depende de uma amostra particular.

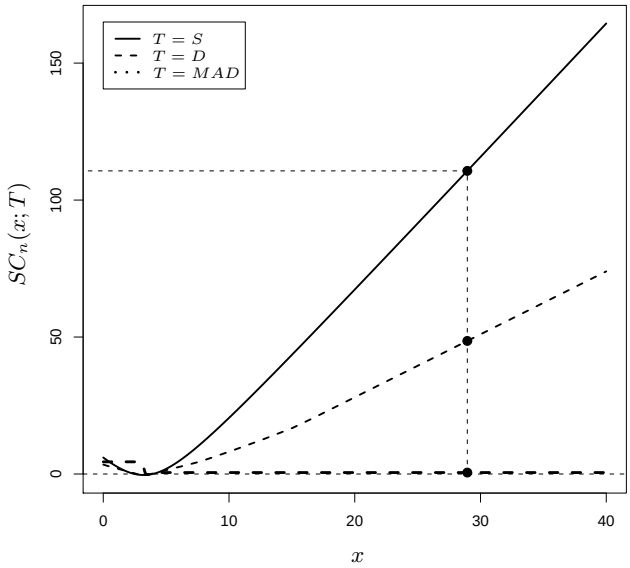


Figura 2.2 *Curvas de sensibilidade de diversos estimadores de dispersão baseadas nas primeiras 23 observações da amostra do Exemplo 2.1.*

entre a média, o desvio padrão e o desvio médio, por um lado, e as restantes medidas, por outro, é agora mais clara. Enquanto no caso das primeiras estamos perante funções ilimitadas, no caso das segundas acontece o contrário. É também claro que o efeito de uma observação aberrante sobre o desvio padrão pode ser bastante maior do que sobre a média ou sobre o desvio médio, o que evidentemente se deve ao efeito da componente quadrática na expressão de cálculo do desvio padrão.

Em jeito de conclusão, pode afirmar-se que a utilização das medidas com curva de sensibilidade limitada é um método seguro e objectivo para controlar o efeito nefasto de possíveis observações contaminantes. Como se disse no início do exemplo este método pode ser na prática quase equivalente ao método de eliminação de *outliers*. No entanto, o primeiro método oferece vantagens, uma vez que o tratamento dos *outliers* é feito de forma objectiva e automática e, na maior parte dos casos, "suave" (não variando abruptamente entre "inclusão" e "eliminação"). Pelo contrário, o método manual de "eliminação" de *outliers* encerra subjectividade, não é automático e processa-se sempre de forma descontínua ou abrupta.

2.3 Função de influência

A função de influência proposta por Hampel (1968, 1974) representou um marco fundamental no desenvolvimento da teoria da robustez, como foi já referido, e é um dos mais importantes instrumentos disponíveis para o estudo das propriedades de robustez de um estimador. A ideia inicial foi a de criar uma medida da alteração provocada numa estimativa quando se junta a uma amostra de grande dimensão uma nova observação num ponto x. Isto é a mesma ideia presente na curva de sensibilidade, mas fazendo a dimensão da amostra tender para infinito, ou seja, passando a trabalhar num contexto assintótico. Esta informação é por si só importante, mas como se verá no desenvolvimento o estudo da função de influência possibilita ir bastante mais longe no conhecimento do estimador do que o estudo da curva de sensibilidade e permite inclusivamente criar novos estimadores com propriedades de robustez predefinidas.

No entanto, a passagem ao contexto assintótico exige algum cuidado e o uso de ferramentas matemáticas mais elaboradas. Em primeiro lugar considera-se a existência de um **modelo paramétrico subjacente ou central**, o qual não se espera que seja rigorosamente verificado pelas observações reais, antes se espera que seja verificado apenas aproximadamente. Este conceito vago pode ser tornado rigoroso com recurso a teoria avançada (veja-se, por exemplo, Huber, 1981).

Para fixar notação, introduz-se agora o modelo paramétrico. Seja $(X_1, \ldots, X_n)$ uma amostra aleatória de dimensão n, constituída por variáveis aleatórias independentes e identicamente distribuídas a uma v.a. X podendo tomar valores num espaço métrico geral Ω (normalmente Ω será coincidente com $\mathbb{R}^m$, ou com um seu subconjunto, ou seja, admite-se desde já a situação multivariada em que X, e consequentemente cada um dos X_i, $i = 1, \ldots, n$, pode representar um vector aleatório de dimensão m). Um modelo paramétrico $\{F_\theta, \theta \in \Theta\}$ consiste numa família de distribuições F_θ sobre Ω, onde θ é um parâmetro desconhecido (pertencente a um espaço paramétrico Θ) que se pretende estimar a partir de uma concretização de $(X_1, \ldots, X_n)$. No caso mais geral θ será um vector. Supõe-se que Θ é um subconjunto aberto e convexo de $\mathbb{R}^p$, onde p é o número de parâmetros, e ainda

que F_θ possui uma "densidade" de probabilidade[3] que se representa por $f_\theta(x)$ ou por $f(x, \theta)$.

Na Figura 2.3 ilustra-se, usando uma representação esquemática, a ideia de modelo paramétrico central e de modelo paramétrico aproximado, formado pelo conjunto das distribuições contidas numa "vizinhança" ε do modelo central. $\mathcal{F}$ representa o conjunto de todas as distribuições sobre $\mathbb{R}$ (note-se que é um espaço de dimensão infinita) enquanto a linha representa as distribuições que pertencem ao modelo central escolhido (neste caso a distribuição normal com parâmetros μ e σ^2, e corresponde a um subconjunto de $\mathcal{F}$ com dimensão 2). A zona sombreada representa as distribuições que pertencentes à "vizinhança"-ε do modelo central, ou seja, que se situam a uma "distância" inferior a ε de algum elemento do modelo central. O ponto F^* identifica um elemento particular do modelo central, com parâmetros μ^* e $(\sigma^*)^2$ enquanto que o ponto G identifica uma distribuição não pertencente ao modelo central mas contida na "vizinhança"-ε deste.

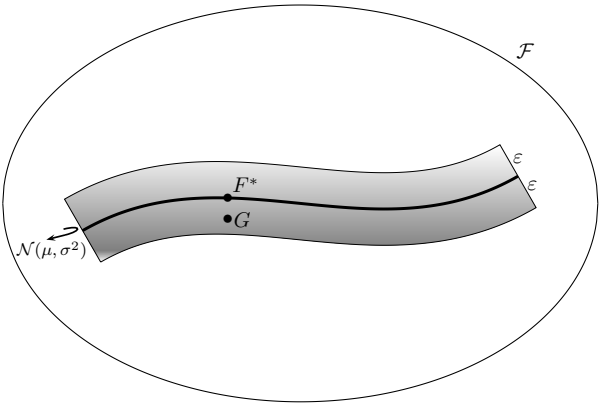


Figura 2.3 *Representação esquemática de um modelo paramétrico central, $\mathcal{N}(\mu, \sigma^2)$, e do modelo paramétrico aproximado formado pelo conjunto das distribuições contidas numa "vizinhança"-ε do modelo central.*

[3]Entenda-se num sentido generalizado, isto é, uma função de densidade no sentido usual se X for uma v.a. contínua, uma função de probabilidade se X for discreta ou ainda o objecto apropriado com as duas componentes se X for mista.

Para melhor se compreender a noção de distância e de vizinhança observe-se a Figura 2.4. A linha representa a f.d. da distribuição $\mathcal{N}(0,1)$, $\Phi(x)$. Se se considerar a distância de Kolmogorov,

$$d(F,\Phi) = \sup_x |F(x) - \Phi(x)|\,,$$

então qualquer v.a. cuja função de distribuição se situe na zona sombreada pertence à vizinhança-ε da distribuição $\mathcal{N}(0,1)$, definida como o conjunto das distribuições para as quais $d(F,\Phi) < \varepsilon$.

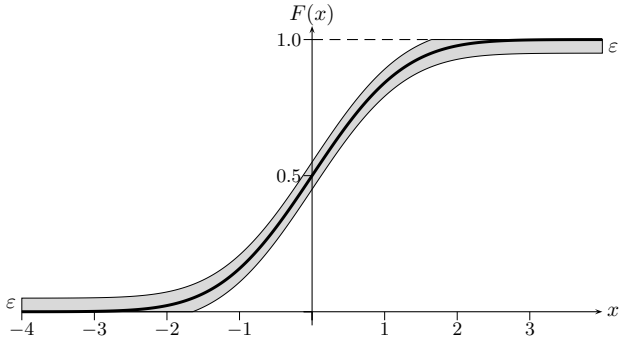


Figura 2.4 .

2.3.1 Conceito de funcional

Para elevar o conceito de sensibilidade da dimensão finita, onde foi definida a curva de sensibilidade, à dimensão infinita, onde vai ser definida a função de influência, é indispensável usar a noção de funcional estatístico.

Recorde-se que matematicamente um funcional é uma aplicação que a cada elemento de um espaço de funções faz corresponder um elemento de um espaço métrico. Por sua vez, um funcional estatístico, é uma aplicação que a cada elemento de um espaço de funções de distribuição (ou simplesmente de distribuições) faz corresponder um valor real, ou um vector de valores reais. Ou seja,

$$T:\ F \in \mathcal{D}(T) \subset \mathcal{F} \longrightarrow T(F) \in \mathbb{R}^p,$$

onde $\mathcal{D}(T)$ representa o domínio de T e $\mathcal{F}$ o espaço das distribuições sobre $\mathbb{R}$, ou, duma forma mais geral, sobre $\mathbb{R}^m$.

Exemplo 2.2. Considere-se uma variável aleatória X com função de distribuição F.

- Um exemplo imediato de um funcional é o valor esperado de X, $\mu(F) = E_F(X)$. $\mathcal{D}(\mu)$ é o conjunto de todas as distribuições com valor esperado finito. O funcional pode escrever-se explicitamente como
$$\mu(F) = \int_{-\infty}^{+\infty} x\, dF(x), \tag{2.7}$$
o que abarca, se o integral for entendido como um integral de Riemann-Stieltjes, todo o tipo de variáveis aleatórias, contínuas, discretas ou mistas.[4]
- Outro exemplo é dado pela variância de X, pois tem-se
$$\sigma^2(F) = \mathrm{var}(X) = \int_{-\infty}^{+\infty} (x - \mu(F))^2\, dF(x) = \\ = \int_{-\infty}^{+\infty} x^2 dF(x) - \mu^2(F). \tag{2.8}$$
O desvio padrão pode simplesmente definir-se como
$$\sigma(F) = \sqrt{\sigma^2(F)}.$$
$\mathcal{D}(\sigma^2) = \mathcal{D}(\sigma)$ é o conjunto de todas as distribuições com variância finita.
- Para a mediana pode considerar-se o funcional definido por
$$m(F) = F^{-1}\left(\frac{1}{2}\right). \tag{2.9}$$
Com $F^{-1}(a) = \inf\{x :\ F(x) \geq a\}$, $0 < a < 1$, obtém-se a mediana inferior, enquanto que com $F^{-1}(a) = \inf\{x :\ F(x) > a\}$,

[4] O integral de Riemann-Stieltjes da função real de variável real f com respeito à função (também real de variável real) g no intervalo $[a, b]$ representa-se por $\int_a^b f(x)\, dg(x)$.

Enquanto o integral de Riemann é definido como o limite, quando a distância entre os pontos sucessivos da partição $\mathcal{P} = \{a = x_0 < x_1 < x_2 < \cdots < x_n = b\}$ tende para zero, das somas $\sum_{\mathcal{P}} f(c_i)(x_{i+1} - x_i)$, com $x_i \leq c_i \leq x_{i+1}$, o integral de Riemann-Stieltjes é definido como um limite do mesmo tipo para as somas $\sum_{\mathcal{P}} f(c_i)\,(g(x_{i+1}) - g(x_i))$. Se g for contínua, crescente e tiver derivada em quase toda a parte, tem-se $\int_a^b f(x)\, dg(x) = \int_a^b f(x) g'(x)\, dx$, o que permite recuperar a definição usual de valor esperado de uma v.a. contínua. Por outro lado, se g corresponder à função de distribuição de uma v.a. discreta obtém-se $\int_{-\infty}^{+\infty} f(x)\, dg(x) = \sum_x f(x) P(X = x)$.

$0 < a < 1$, se obtém a mediana superior. Em muitos casos, por exemplo quando F é invertível, mas não só, ambas coincidem. Para recuperar a definição usual basta considerar

$$F^{-1}(a) = \frac{\inf\{x:\ F(x) \geq a\} + \inf\{x:\ F(x) > a\}}{2}.$$

Muitos outros exemplos poderiam ser aqui apresentados. Em particular, não é difícil imaginar as generalizações em que F representa a distribuição de um vector aleatório de dimensão m e o resultado, $T(F)$, pertence a $\mathbb{R}^p$. Mas não convém fazer mais desvios em relação ao objectivo que se pretende alcançar e que é, recorde-se, a representação de um estimador à custa de um funcional e a definição da função de influência.

Considere-se novamente a amostra de dimensão n, $(X_1, \ldots, X_n)$, e um estimador de θ, $T_n = T_n(X_1, \ldots, X_n)$. Para muitos estimadores verifica-se que T_n depende de $(x_1, \ldots, x_n)$ apenas através da função de distribuição empírica da amostra,

$$G_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_{x_i},$$

onde Δ_x representa a distribuição discreta degenerada em x ($\Delta_x(u) = 1$, se $u \geq x$ e $\Delta_x(u) = 0$, se $u < x$).[5] Ou seja, a estimativa tem o mesmo valor para todas as amostras que tenham a mesma função de distribuição empírica. Isto acontece, por exemplo, se a estimativa não depender da ordem porque são recolhidas as observações e também se não for alterada por duplicações/multiplicações da amostra.

Definição 2.2. *Dada a amostra de dimensão n, $(X_1, \ldots, X_n)$, e um estimador de θ, $T_n = T_n(X_1, \ldots, X_n)$, se*

$$T_n(x_1, \ldots, x_n) = T(G_n) \quad \forall_{n,G_n}, \tag{2.10}$$

para algum funcional T definido pelo menos no espaço das distribuições empíricas, diz-se que o estimador T_n é equivalente ao funcional T.

[5] Também se pode dizer que G_n é a distribuição da variável aleatória discreta que toma os valores $x_1, \ldots, x_n$, todos com probabilidade $1/n$.

Com esta abordagem pretende substituir-se uma dada sequência de estimadores, $\{T_n, n \geq 1\}$, por um funcional independente de n. Para algumas sequências de estimadores este funcional existe (ver Exemplo 2.3 a seguir), para outras essa substituição envolve apenas uma correcção de ordem $1/n$, isto é, existe um funcional, $T(\cdot)$, tal que $T_n(x_1, \ldots, x_n) \neq T(G_n)$ mas $T_n(x_1, \ldots, x_n) - T(G_n) = O(1/n)$ (Exemplo 2.4). Neste caso diz-se que o estimador é assintoticamente equivalente ao funcional. Felizmente a maior parte dos estimadores habituais, incluindo os desenvolvidos no âmbito da teoria da robustez, pertence a um destes dois grupos, ou seja, ou é equivalente a um funcional ou é assintoticamente equivalente a um funcional.

Em principio existe uma extensão natural de T a um subconjunto convexo de $\mathcal{F}$ que constitui o domínio de T. Se G representar a distribuição exacta de X então $T(G)$ define o vector dos parâmetros de facto estimados enquanto que $T(G_n)$ define o vector das estimativas. É evidente que para que esta afirmação faça sentido deve verificar-se

$$\lim_{n\to\infty} T_n = T(G) \quad \text{(em probabilidade)},$$

o que significa que o estimador T_n é um estimador consistente da "quantidade" $T(G)$. Assume-se, como se disse no início desta secção, que a distribuição G pertence a uma "vizinhança" do modelo paramétrico adoptado mas pode não coincidir com nenhum elemento do modelo. Deve porém exigir-se que sob o modelo o estimador T_n estime correctamente o parâmetro, pelo menos assintoticamente. Isto pode traduzir-se em termos formais dizendo que o funcional em questão, T, deve ser consistente segundo Fisher.

Definição 2.3. *Um funcional T, equivalente ou assintoticamente equivalente a um estimador T_n do parâmetro θ de um modelo paramétrico F_θ, é consistente segundo Fisher, nesse modelo, sse*

$$T(F_\theta) = \theta, \quad \forall_{\theta\in\Theta}. \tag{2.11}$$

Sob condições de regularidade fracas a consistência segundo Fisher é equivalente à consistência usual. Trata-se essencialmente de uma diferença de notação que é conveniente no contexto da estatística robusta, quando se faz a distinção entre modelo paramétrico central e modelo paramétrico aproximado.

Exemplo 2.3. Considere-se para $\Omega \subset \mathbb{R}$ o estimador

$$T_n = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Para o funcional definido por $\mu(F) = \int x\, dF(x)$ verifica-se (2.10), pois

$$\mu(G_n) = \int x\, dG_n(x) = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Considere-se novamente a situação representada esquematicamente na Figura 2.3, isto é, suponha-se que o modelo paramétrico central é $\{F_\mu,\, \mu \in \mathbb{R}\}$, em que F_μ corresponde à distribuição $\mathcal{N}(\mu, \sigma^2)$ (admitindo por simplicidade que σ é conhecido). O funcional $\mu(\cdot)$ é consistente segundo Fisher pois

$$\mu(F_\mu) = \mu, \quad \forall_{\mu \in \mathbb{R}}.$$

Portanto se as observações forem i.i.d. de acordo com uma distribuição do modelo paramétrico o parâmetro estimado é μ. Por outro lado se as observações forem i.i.d. de acordo com uma distribuição G pertencente ao domínio de $\mu(\cdot)$ o "parâmetro" que de facto está a ser estimado é $\mu(G)$, pois $\lim_{n\to\infty} T_n = \mu(G)$ (em probabilidade). O estimador T_n seria robusto (em certo sentido, a definir mais claramente numa das secções seguintes) se o facto de G estar próximo de algum F^* implicasse que $\mu(G)$ estaria próximo de $\mu(F^*) = \mu^*$. No entanto para este estimador isso não acontece, isto é, pode ter-se G arbitrariamente próximo de F^* mas $\mu(G)$ arbitrariamente afastado de μ^*. Para confirmar esta afirmação pode fazer-se um exercício semelhante ao que foi apresentado no Exemplo 1.1 (página 4) relativamente à variância mas considerando em vez da contaminação simétrica (1.1), uma contaminação assimétrica, por exemplo,

$$F(x) = (1-\varepsilon)\, \Phi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\, \Phi\left(\frac{x-k\mu}{\sigma}\right).$$

Pode também concluir-se o mesmo a partir da análise da Figura 2.4. De facto, na vizinhança-ε da distribuição $\mathcal{N}(0,1)$ representada no gráfico podem encontrar-se distribuições com valor esperado tão elevado quanto se queira, inclusive ∞, basta, por exemplo, considerar funções de distribuição com a cauda direita encostada à parte inferior da zona sombreada.[6]

[6] Se se fizer o mesmo exercício com a mediana, outro estimador do parâmetro

Exemplo 2.4. Para o estimador usual da variância

$$S_n^2 = \frac{\sum_{i=1}^{n}(X_i - T_n)^2}{n-1}$$

não é possível definir um funcional tal que se verifique (2.10). No entanto tem-se, para $X_i \sim G$,

$$\lim_{n\to\infty} S_n^2 = \sigma^2(G) \quad \text{(em probabilidade)},$$

com $\sigma^2(\cdot)$ definido em (2.8), desde que $G \in \mathcal{D}(\sigma^2)$. Ou seja, o estimador S_n^2 é assintoticamente equivalente ao funcional $\sigma^2(\cdot)$. Para amostras de dimensão finita a diferença entre a estimativa e o valor do funcional é apenas da ordem de $1/n$ pois

$$S_n^2(x_1, \ldots, x_n) = \frac{n}{n-1}\sigma^2(G_n).$$

Ao longo do capítulo surgirão outros exemplos que ilustrarão melhor estes conceitos.

2.3.2 Definição e propriedades

Encontram-se agora reunidas as condições para compreender a definição de função de influência. Por uma questão de simplicidade de notação considera-se inicialmente o caso univariado e uniparamétrico ($m = 1$ e $p = 1$).

Definição 2.4. *Seja F uma distribuição pertencente ao domínio de T. Chama-se função de influência (IF) do funcional T em F à função*

$$IF(x; T, F) = \lim_{\varepsilon \to 0^+} \frac{T\left((1-\varepsilon)F + \varepsilon\Delta_x\right) - T(F)}{\varepsilon} \qquad (2.12)$$

definida pontualmente nos pontos $x \in \Omega \subset \mathbb{R}$ para os quais o limite existe.[7]

μ, tem-se uma conclusão diferente, qualquer distribuição considerada na mesma faixa tem mediana que não pode variar arbitrariamente entre $-\infty$ e $+\infty$, antes fica sempre compreendida num intervalo $]-\delta, +\delta[$, com δ dependente de ε.

[7] Recorde-se que Δ_x representa a distribuição discreta degenerada em x ($\Delta_x(u) = 1$, se $u \geq x$ e $\Delta_x(u) = 0$, se $u < x$).

Esta definição suscita imediatamente um conjunto importante de observações:

- Note-se que $(1-\varepsilon)F + \varepsilon\Delta_x$ representa a distribuição de uma variável aleatória que é a mistura, nas proporções de $1-\varepsilon$ para ε, das variáveis aleatórias com distribuição F e Δ_x, respectivamente. Dito de outra maneira, o que se está a considerar é um modelo de contaminação em x, em que com probabilidade $1-\varepsilon$ se tem uma observação proveniente do verdadeiro modelo, F, e com probabilidade ε se observa x. Repare-se que se F corresponder a uma variável aleatória contínua, $(1-\varepsilon)F+\varepsilon\Delta_x$ corresponde a uma variável aleatória mista.

- Embora rigorosamente ε deva ser maior que zero para que $(1-\varepsilon)F + \varepsilon\Delta_x$ seja uma distribuição legítima para qualquer F, em certos casos ainda se obtém uma distribuição legítima com $\varepsilon < 0$ (embora não interpretável como uma mistura), pelo que na definição da função de influência se poderia considerar $\lim_{\varepsilon\to 0}$ em vez de $\lim_{\varepsilon\to 0^+}$. As definições seriam coincidentes excepto quando $\lim_{\varepsilon\to 0^+} \neq \lim_{\varepsilon\to 0^-}$. Sem falta de rigor pode-se usar sempre $\lim_{\varepsilon\to 0}$, caso este não exista verifica-se então se existe o limite lateral direito.

- Uma definição equivalente da função de influência, por vezes mais conveniente em termos de cálculo, é

$$IF(x;T,F) = \left.\frac{\partial}{\partial\varepsilon}T(F_{\varepsilon,x})\right|_{\varepsilon=0}, \tag{2.13}$$

  com $F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$. Note-se que com x, F e T fixos $T(F_{\varepsilon,x})$ é uma função real da variável real ε.

- Substituindo em (2.12) F por F_{n-1}, (se n grande $F \simeq F_{n-1}$) ε por $1/n$ e omitindo o limite pode verificar-se que de facto $IF(x;T,F)$ é aproximadamente igual a n vezes a alteração provocada na estimativa $T_n(x_1,\ldots,x_{n-1})$ pelo aparecimento de uma nova observação no ponto x, isto é,

$$IF(x;T,F) \simeq \frac{T\left(\left(1-\frac{1}{n}\right)F_{n-1} + \frac{1}{n}\Delta_x\right) - T(F_{n-1})}{\frac{1}{n}} =$$

$$= n\left[T_n(x_1,\ldots,x_{n-1},x) - T_{n-1}(x_1,\ldots,x_{n-1})\right].$$

Ou seja, se existir a função de influência e se o estimador for pelo menos assintoticamente equivalente a um funcional regular, verifica-se que $SC_n(x;T_n) \simeq IF(x;T,F)$ e que (em probabilidade) $\lim_{n\to\infty} SC_n(x;T_n) = IF(x;T,F)$. A curva de sensibilidade permite ter uma ideia do comportamento da função de influência em casos em que esta é muito difícil de calcular, basta para isso que se calcule $SC_n(x;T_n)$ para uma amostra grande gerada a partir da distribuição F. É, no entanto, necessário ter algum cuidado com a regularidade do funcional, na Secção 2.3.4 apresentam-se dois exemplos em que apesar de existir $IF(x;T,F)$ não existe $\lim_{n\to\infty} SC_n(x;T_n)$, devido ao facto de o funcional não ser regular.

- Outra versão da função de influência para dimensão finita pode obter-se a partir do conjunto das pseudo-observações *jackknife* (Quenouille, 1949, 1956). Recorde-se que a i-ésima pseudo--observação *jackknife*, para um estimador T_n e uma amostra $(x_1,\ldots,x_n)$, é dada por

$$T^*_{ni} = nT_n(x_1,\ldots,x_n) - (n-1)T_{n-1}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n).$$

Então, se existir a função de influência e se o estimador for pelo menos assintoticamente equivalente a um funcional regular, tem-se que

$$IF(x_i;T,F) \simeq \frac{T\left(\left(1-\frac{-1}{n-1}\right)F_n + \left(\frac{-1}{n-1}\right)\Delta_{x_i}\right) - T(F_n)}{\frac{-1}{n-1}} =$$

$$= T^*_{ni} - T_n(x_1,\ldots,x_n). \tag{2.14}$$

Esta é uma situação em que faz sentido considerar $\varepsilon < 0$, embora ε não seja interpretável como uma probabilidade de contaminação, pois

$$\left(1-\frac{-1}{n-1}\right)F_n + \left(\frac{-1}{n-1}\right)\Delta_{x_i}$$

ainda é uma distribuição de probabilidade.

- Tem interesse para melhor compreensão do conceito de função de influência e das suas propriedades, a consideração da seguinte

questão: e se em vez de uma mistura (contaminação) com uma massa pontual se considerasse a mistura com uma outra distribuição? Se essa distribuição for discreta com suporte finito, $\{x_1, \ldots, x_n\}$, e probabilidades associadas $\{\alpha_1, \ldots, \alpha_n\}$, com $\sum_{i=1}^{n} \alpha_i = 1$ e $0 \leq \alpha_i \leq 1, \forall_i$, pode representar-se por $G_n = \sum_{i=1}^{n} \alpha_i \Delta_{x_i}$. Pode então mostrar-se que[8]

$$\lim_{\varepsilon \to 0} \frac{T\left((1-\varepsilon)F + \varepsilon G_n\right) - T(F)}{\varepsilon} = \sum_{i=1}^{n} \alpha_i IF(x_i; T, F) = \\ = \int IF(x; T, F)\, dG_n(x).$$

Para uma distribuição contaminante geral, G, verifica-se[9]

$$\lim_{\varepsilon \to 0} \frac{T\left((1-\varepsilon)F + \varepsilon G\right) - T(F)}{\varepsilon} = \int IF(x; T, F)\, dG(x). \tag{2.15}$$

A expressão anterior exprime a diferencialidade do funcional T, em F, na direcção de G. Neste contexto faz sentido interpretar $IF(x; T, F)$ como a derivada (parcial) do funcional T, em F, na direcção de Δ_x.

As propriedades mais importantes da função de influência são as seguintes:

(P1) O valor esperado em F de IF é nulo, isto é,

$$E_F\left[IF(X; T, F)\right] = \int IF(x; T, F)\, dF(x) = 0.$$

Esta propriedade decorre imediatamente de (2.15) com $G = F$.

(P2) Para G próxima de F tem-se, com base no desenvolvimento em fórmula de Taylor com os termos de primeira ordem,

$$T(G) \simeq T(F) + \int IF(x; T, F)\, dG(x). \tag{2.16}$$

[8] Sob condições de regularidade muito fracas e que geralmente se verificam na prática. O leitor curioso pode consultar Pires (1995), em relação a este aspecto específico, ou Fernholz (1983) sobre o tema mais geral da diferenciabilidade de funcionais estatísticos.

[9] Idem.

Esta expressão obtém-se facilmente considerando a função auxiliar

$$\varphi(\varepsilon) = T\left((1-\varepsilon)F + \varepsilon G\right),$$

que é uma função real da variável real ε. Se esta função for diferenciável até à segunda ordem tem-se[10]

$$\varphi(1) = \varphi(0) + \varphi'(0) + \text{Resto},$$

mas $\varphi(1) = T(G)$, $\varphi(0) = T(F)$ e $\varphi'(0) = \int IF(x;T,F)\, dG(x)$, por (2.15). Se em (2.16) se considerar em vez de G a distribuição empírica, F_n, a qual está tanto mais próxima de F quanto maior for n, obtém-se

$$\begin{aligned} T_n = T(F_n) \simeq T(F) + \int IF(x;T,F)\, dF_n(x) = \\ = T(F) + \sum_{i=1}^{n} \frac{IF(x_i;T,F)}{n}. \end{aligned} \tag{2.17}$$

(P3) Se $X_i \overset{\text{iid}}{\sim} F$ então $\sqrt{n}(T_n - T(F))$ tem, pelo Teorema do Limite Central,[11] distribuição assintoticamente normal com valor médio nulo e variância dada por

$$\begin{aligned} V(T,F) &= \operatorname{var}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(X_i;T,F)\right) = \\ &= E\left[IF(X;T,F)^2\right] = \\ &= \int IF(x;T,F)^2 dF(x). \end{aligned} \tag{2.18}$$

Ou seja,

$$\sqrt{n}(T_n - T(F)) \overset{a}{\sim} \mathcal{N}\left(0, V(T,F)\right).$$

A $V(T,F) = \lim_{n\to\infty} n \operatorname{var}(T_n)$ chama-se **variância assintótica** do estimador T_n em F.

[10] $\exists_{0<c<1}$: Resto $= \varphi''(c)/2$.

[11] Se os termos de ordem superior da expansão de Taylor forem assintoticamente desprezáveis.

(P4) Seja $\beta(\theta)$ uma transformação diferenciável do parâmetro θ, $\beta : \Theta \subset \mathbb{R} \longrightarrow \beta(\Theta) \subset \mathbb{R}$, com derivada $\beta'(\theta)$ e $\beta(T_n)$ um estimador de $\beta(\theta)$, equivalente ao funcional $\beta(T(F))$, então

$$IF(x; \beta(T), F) = \beta'(T(F)) IF(x; T, F) \tag{2.19}$$

e

$$V(\beta(T), F) = [\beta'(T(F))]^2 V(T, F). \tag{2.20}$$

A fórmula (2.19) decorre imediatamente de (2.13) por aplicação da regra de derivação da função composta.

(P5) A partir da propriedade **(P2)** é ainda possível demonstrar (ver, por exemplo, Hampel *et al.*, 1986, p. 86, ou Huber, 1981, p. 69) o seguinte importante resultado: Seja $J(\theta)$ a informação de Fisher em relação a determinado modelo paramétrico F_θ, dada por

$$J(\theta) = \int \left(\frac{\partial \log f(x, \theta)}{\partial \theta} \right)^2 dF_\theta(x) = - \int \frac{\partial^2 \log f(x, \theta)}{\partial \theta^2} dF_\theta(x).$$

Se $0 < J(\theta) < \infty$ e T for consistente à Fisher, tem-se

$$\int IF(x; T, F_\theta)^2 dF_\theta(x) \geq \frac{1}{J(\theta)},$$

verificando-se a igualdade se e só se

$$IF(x; T, F_\theta) = J(\theta)^{-1} \frac{\partial \log f(x, \theta)}{\partial \theta}. \tag{2.21}$$

Em relação à propriedade **(P2)**, pode pensar-se à primeira vista que o facto de se considerarem apenas os termos de primeira ordem, torna as expressões (2.16) e (2.17) aproximações de muito fraca qualidade, mas a verdade é que isso não acontece. Na realidade essas expressões podem até ser exactas (para funcionais lineares num certo sentido), como se verá nos Exemplos 2.5 e 2.6.

A propriedade **(P3)** é bastante importante, pois mostra que a função de influência permite, para além da avaliação da influência relativa das observações individuais nas estimativas, a dedução explícita e expedita da distribuição assintótica do estimador. O resultado obtido depende, porém, do comportamento dos termos de ordem

superior, ou se se preferir, do resto em (2.17). A análise desses termos é normalmente bastante complicada. Assim o que se costuma fazer é obter a expressão da variância assintótica, dada por (2.18), e se for necessário demonstrar a sua validade duma forma rigorosa então fazê-lo por outros processos. Esta observação não retira importância ao resultado pois, e citando Hampel *et al.* (1986), "o importante é que aquela fórmula fornece a resposta correcta em todos os casos práticos que se conhecem".

A este propósito recorde-se que a fórmula proposta por Tukey (1958), no contexto do método *jackknife*, para estimar a variância de T_n é V_n/n, com

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(T_{ni}^* - T_n^*)^2 \tag{2.22}$$

e

$$T_n^* = \frac{1}{n}\sum_{i=1}^{n} T_{ni}^*.$$

A fórmula (2.22) pode ser deduzida a partir da expressão (2.18) utilizando (2.14) para estimar os valores da função de influência:

$$V(T,F) = \mathrm{var}[IF(X;T,F)] \simeq \mathrm{var}\{IF(x_i;T,F); i = 1,\ldots,n\} \simeq V_n.$$

A propriedade **(P5)** é também de grande importância, pois afirma que a variância assintótica de um estimador de θ, num determinado modelo paramétrico, atinge o seu mínimo, o limite inferior de Cramér-Rao, quando a função de influência do estimador coincide com determinado múltiplo da derivada do logaritmo da função de verosimilhança (mais adiante se verá que esta é precisamente a função de influência do estimador de máxima verosimilhança). Usando (2.18) é simples calcular a eficiência assintótica absoluta de um estimador T_n (equivalente a um funcional consistente à Fisher, T, em relação a determinado modelo paramétrico F_θ):

$$e(T,F_\theta) = \left[V(T,F_\theta)J(\theta)\right]^{-1}. \tag{2.23}$$

Pode também utilizar-se a função de influência para calcular a eficiência assintótica relativa do estimador S_n em relação ao estimador T_n para estimação de um mesmo parâmetro (equivalentes, respectivamente, aos funcionais S e T ambos consistentes à Fisher),

$$ARE_{S,T}(F) = \frac{V(T,F)}{V(S,F)}.$$

2.3.3 Medidas de robustez baseadas na função de influência

A popularidade da função de influência não pode ser atribuída apenas às propriedades acabadas de discutir, por mais úteis que elas sejam. Os motivos principais dessa popularidade devem-se ao facto de a função de influência permitir o estudo local das propriedades de robustez, contribuindo para o aprofundamento do conhecimento sobre os estimadores, e possibilitar a construção de novos estimadores com características pré-definidas, quer em termos de robustez, quer em termos de eficiência. Nesse sentido são relevantes alguns valores numéricos, para além de $V(T,F)$ e $e(T,F_\theta)$, que resumem outras características importantes da função de influência e cujas definições se apresentam em seguida.

Definição 2.5. *Chama-se "sensibilidade a grandes erros" (gross-error sensitivity) de T em F a*

$$\gamma^* = \gamma^*(T,F) = \sup_x |IF(x;T,F)|. \tag{2.24}$$

Este valor mede aproximadamente a pior influência que uma determinada contaminação, ε, próxima de zero, pode ter no valor de uma estimativa, o que pode ser visto como uma medida aproximada do enviesamento máximo (estandardizado) do estimador T_n. De facto, se $\mathcal{P}_\varepsilon(F)$ representar a "vizinhança de contaminação" proposta por Huber,[12]

$$\mathcal{P}_\varepsilon(F) = \{G : G = (1-\varepsilon)F + \varepsilon H,\, H \in \mathcal{F}\}, \tag{2.25}$$

então

$$b_{T,F}(\varepsilon) = \sup_{G \in \mathcal{P}_\varepsilon(F)} |T(G) - T(F)|, \tag{2.26}$$

representa o enviesamento assintótico máximo de T na vizinhança $\mathcal{P}_\varepsilon(F)$, isto é, o erro máximo que se pode cometer quando se pretende estimar $T(F)$ mas as observações não seguem exactamente o modelo

[12] Também conhecida por "gross error model". Note-se que não é uma vizinhança no sentido topológico do termo. Em (2.25), $(1-\varepsilon)F + \varepsilon H$ representa como habitualmente uma mistura das distribuições F e H nas proporções de $(1-\varepsilon)$ para ε e $\mathcal{F}$ representa o espaço de todas as distribuições sobre Ω.

F, podendo estar "contaminadas" com probabilidade ε por uma outra qualquer distribuição H. Por (2.15) e para $\varepsilon \approx 0$ pode concluir-se[13] que

$$\sup_{G\in\mathcal{P}_\varepsilon(F)} |T(G)-T(F)| \simeq \varepsilon \sup_{H\in\mathcal{F}} \left| \int IF(x;T,F)\, dH(x) \right| =$$

$$= \varepsilon \sup_x |IF(x;T,F)| = \varepsilon\gamma^*(T,F),$$

ou seja

$$b_{T,F}(\varepsilon) \simeq \varepsilon\gamma^*(T,F). \tag{2.27}$$

Esta expressão mostra que

- Se $\gamma^*(T,F) = +\infty$, o enviesamento assintótico de T na vizinhança de contaminação $\mathcal{P}_\varepsilon(F)$ pode ser ilimitado, por mais pequeno que seja o valor de ε.

- Se $\gamma^*(T,F) < +\infty$, então esse mesmo enviesamento é limitado e no pior dos casos é dado por $\varepsilon\gamma^*(T,F)$.

Logo pode dizer-se que um estimador com função de influência limitada é robusto, mais concretamente B-robusto.[14]

Definição 2.6. *T é B-robusto em F se $\gamma^*(T,F) < \infty$.*

Se para um dado modelo paramétrico central o estimador de máxima verosimilhança for equivalente a um funcional B-robusto então tem-se o melhor dos dois mundos, melhor eficiência e robustez. No Exemplo 2.7 é apresentado um modelo paramétrico para o qual se verifica esta situação. Se, pelo contrário, se admite um modelo paramétrico central para o qual o estimador de máxima verosimilhança é equivalente a um funcional com função de influência ilimitada, tem-se uma situação perigosa na medida em que a mais pequena percentagem de contaminação do modelo pode provocar um grande enviesamento. Por

[13] A passagem da primeira para a segunda linha é justificada pelo facto de se estar a trabalhar com $\sup_{H\in\mathcal{F}}$, ou seja, deve pensar-se na distribuição H para a qual $\left|\int IF(x;T,F)\, dH(x)\right| = |E_H[IF(X;T,F)]|$ toma o maior valor possível.

[14] B vem de "bias".

outro lado, nestas condições ao usar um estimador alternativo com função de influência limitada ela estará necessariamente afastada da função de influência do estimador de máxima verosimilhança, o que provoca um acréscimo na variância assintótica e um decréscimo na eficiência assintótica. É assim necessário, para resolver este dilema eficiência-robustez, procurar um estimador que represente um compromisso entre os dois extremos: por um lado um estimador com $\gamma^*(T,F)$ pequena mas que pode ter grande variância assintótica e por outro lado um estimador com pequena variância assintótica mas $\gamma^*(T,F)$ muito grande. Num dos extremos está o estimador de máxima verosimilhança com $V(T,F)$ mínima mas $\gamma^*(T,F)$ infinita e no outro o estimador chamado "mais B-robusto" com $\gamma^*(T,F)$ mínima mas maior variância assintótica. Os estimadores compreendidos entre estes dois extremos e que não podem ser melhorados simultaneamente em relação a $V(T,F)$ e a $\gamma^*(T,F)$ designam-se por "B-robustos óptimos".

Outras medidas calculadas a partir da função de influência são a sensibilidade local e o ponto de rejeição.

Definição 2.7. *A sensibilidade local de T em F é*

$$\lambda^*(T,F) = \sup_{x \neq y} \frac{|IF(y;T,F) - IF(x;T,F)|}{|y-x|}. \tag{2.28}$$

A sensibilidade local está relacionada com pequenas flutuações das observações, pois λ^* mede aproximadamente o pior efeito possível (estandardizado) provocado na estimativa por variações locais nas observações. Estas variações ocorrem, por exemplo, quando se efectuam arredondamentos ou agrupamentos dos dados. É fácil verificar que se F corresponder a uma distribuição contínua e $IF(x;T,F)$ tiver algum ponto de descontinuidade então $\lambda^*(T,F) = \infty$. Note-se, no entanto, que um estimador ter $\lambda^*(T,F) = \infty$ não é considerado tão grave como ter $\gamma^*(T,F) = \infty$.

Definição 2.8. *Seja F uma distribuiçao simétrica em relação a um ponto c. O ponto de rejeição de T em F é*

$$\rho^*(T,F) = \inf\{r > 0 : IF(x;T,F) = 0 \text{ para } x \text{ tal que } |x-c| > r\}. \tag{2.29}$$

O valor de ρ^* representa a menor distância ao centro de simetria a partir da qual a função de influência é identicamente nula, ou seja, a partir da qual as observações são totalmente "rejeitadas", não tendo nenhum efeito sobre a estimativa. Note-se ainda que a definição pode ser modificada de modo a incluir distribuições não simétricas. Pode obter-se um estimador com $\rho^*(T,F)$ finito se, por exemplo, se aplicar previamente uma regra de rejeição de *outliers*. Outros exemplos serão apresentados na Secção 3.2. Também neste caso, tal como em relação a $\lambda^*(T,F)$, se considera que $\rho^* = \infty$ é menos grave que $\gamma^* = \infty$.

Resumindo, pode dizer-se que, sob o ponto de vista da robustez, é desejável que a função de influência de um estimador T_n (equivalente ao funcional T) do parâmetro θ de determinado modelo paramétrico $\{F_\theta,\, \theta \in \Theta\}$ seja tal que

(i) $\gamma^*(T,F_\theta) < \infty$, ou seja, tenha função de influência limitada.
(ii) $V(T,F_\theta) \simeq J^{-1}(\theta)$, para que a eficiência assintótica absoluta dada por (2.23) seja próxima de 100%.

e que, se possível, se tenha

(iii) $\lambda^*(T,F_\theta) < \infty$.
(iv) $\rho^*(T,F_\theta) < \infty$.

Convém salientar que, como se verá na Secção 3.2, é possível e até bastante simples construir um estimador cuja função de influência tenha uma forma pré-definida (a menos de um factor multiplicativo).

2.3.4 Exemplos

Apresentam-se de seguida uma série de exemplos ilustrativos dos conceitos acabados de apresentar. Nos primeiros exemplos consideram-se

todos os estimadores utilizados no Exemplo 2.1 a propósito da curva de sensibilidade.

Exemplo 2.5. Na sequência do Exemplo 2.3 (página 34) apresenta-se o cálculo da função de influência do estimador

$$T_n = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

equivalente ao funcional $\mu(F) = \int x\, dF(x)$, cujo domínio é o conjunto de todas as distribuições com média finita. O valor do funcional na distribuição contaminada $(1-\varepsilon)F + \varepsilon\Delta_x$ é dado por

$$\begin{aligned}\mu\left((1-\varepsilon)F + \varepsilon\Delta_x\right) &= (1-\varepsilon)\int u\, dF(u) + \varepsilon \int u\, d\Delta_x(u) = \\ &= (1-\varepsilon)\mu(F) + \varepsilon x = \\ &= \mu(F) + \varepsilon(x - \mu(F)),\end{aligned}$$

donde

$$IF(x;\mu,F) = \lim_{\varepsilon\to 0} \frac{\mu(F) + \varepsilon(x-\mu(F)) - \mu(F)}{\varepsilon} = x - \mu(F).$$

Esta função mostra que, desde que o suporte de F, Ω, seja ilimitado, a influência de uma única observação no valor da estimativa pode ser tão grande quanto se queira. Nestas circunstâncias tem-se

$$\gamma^*(\mu,F) = \infty, \quad \lambda^*(\mu,F) = 1 \quad \text{e} \quad \rho^*(\mu,F) = \infty.$$

Além disso, como não podia deixar de ser,

$$V(\mu,F) = E_F\left[IF(X;\mu,F)^2\right] = \text{var}(X) = \sigma^2(F) \quad \text{(se existir)}.$$

Também é óbvio que (em probabilidade)

$$\lim_{n\to\infty} SC_n(x,\overline{X}) = \lim_{n\to\infty}\left(x - \overline{X}\right) = x - \mu(F) = IF(x;\mu,F).$$

Note-se que a função de influência obtida é neste caso independente do modelo paramétrico considerado, pelo que as conclusões são válidas para todos os modelos (paramétricos ou não paramétricos, discretos, contínuos, ou mistos) com suporte ilimitado para os quais se utilize o estimador em causa. É simples também verificar que

$$\mu\left((1-\varepsilon)F + \varepsilon G\right) = (1-\varepsilon)\,\mu(F) + \varepsilon\,\mu(G),$$

ou seja, o funcional é linear no espaço das distribuições, e a expressão (2.15) é verificada para qualquer F e $G \in \mathcal{D}(\mu)$ e qualquer $0 \leq \varepsilon \leq 1$ e não apenas no limite, o que tem como consequência que os desenvolvimentos (2.16) e (2.17) sejam exactos e não meras aproximações de primeira ordem.

Exemplo 2.6. Veja-se agora o caso dos estimadores variância e desvio padrão amostrais. Como se viu no Exemplo 2.4 (página 35) a variância amostral representada pelo estimador S_n^2 é assintoticamente equivalente ao funcional $\sigma^2(F)$ definido em (2.8). Então conclui-se, com $F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$, que

$$\begin{aligned}\sigma^2\left(F_{\varepsilon,x}\right) &= \int u^2 dF_{\varepsilon,x}(u) - \mu^2(F_{\varepsilon,x}) = \\ &= (1-\varepsilon)(\sigma^2(F) + \mu^2(F)) + \varepsilon x^2 - (\mu(F) + \varepsilon(x - \mu(F)))^2 = \\ &= (1-\varepsilon)\sigma^2(F) + \varepsilon(x-\mu(F))^2 - \varepsilon^2(x-\mu(F))^2.\end{aligned}$$

Logo

$$IF(x;\sigma^2,F) = \frac{\partial}{\partial\varepsilon}\sigma^2(F_{\varepsilon,x})\bigg|_{\varepsilon=0} = (x-\mu(F))^2 - \sigma^2(F). \qquad (2.30)$$

É imediato verificar que, para modelos com suporte ilimitado,

$$\gamma^*(\sigma^2,F) = \infty, \quad \lambda^*(\sigma^2,F) = \infty \quad \text{e} \quad \rho^*(\sigma^2,F) = \infty.$$

Além disso, é bastante fácil calcular a variância assintótica,

$$\begin{aligned}V(\sigma^2,F) = E_F\left[\left[(X-\mu(F))^2 - \sigma^2(F)\right]^2\right] &= \\ = E_F\left[(X-\mu(F))^4\right] - \sigma^4(F),&\end{aligned}$$

expressão que é válida se existir o quarto momento central da distribuição F, $\mu_4(F) = E_F\left[(X-\mu(F))^4\right]$.

É também simples verificar que $\lim_{n\to\infty} SC_n(x;S_n^2) = IF(x;\sigma^2,F)$ (em probabilidade), com $SC_n(x;S_n^2)$ dada por (2.4).

Por sua vez o desvio padrão amostral S_n será assintoticamente equivalente ao funcional $\sigma(F)$. Neste caso o mais simples é utilizar a propriedade **(P4)**, uma vez que se tem $\sigma = \beta(\sigma^2)$ e $S_n = \beta(S_n^2)$,

com $\beta(\theta) = \sqrt{\theta}$, função que é diferenciável em $\Theta = \mathbb{R}^+$. Como $\beta'(\theta) = 1/(2\sqrt{\theta})$, tem-se

$$IF(x;\sigma,F) = \frac{1}{2\sigma(F)}\left[(x-\mu(F))^2 - \sigma^2(F)\right]$$

e

$$V(\sigma,F) = \frac{V(\sigma^2,F)}{4\,\sigma^2(F)} = \frac{\mu_4(F) - \sigma^4(F)}{4\,\sigma^2(F)}. \tag{2.31}$$

É importante realçar que, diferindo apenas por uma constante, ambas as funções de influência possuem as mesmas características, nomeadamente, nos modelos com suporte ilimitado,

$$\gamma^*(\sigma,F) = \infty, \quad \lambda^*(\sigma,F) = \infty \quad \text{e} \quad \rho^*(\sigma,F) = \infty.$$

Também mais uma vez é simples confirmar que (em probabilidade) $\lim_{n\to\infty} SC_n(x;S_n) = IF(x;\sigma,F)$, com $SC_n(x;S_n)$ dada por (2.5).

Com este exemplo ilustra-se o seguinte importante aspecto: as propriedades essenciais de um estimador, em termos de robustez, não podem ser alteradas por meio de uma transformação diferenciável, por mais limitativa que esta seja. A este propósito tem também interesse considerar o momento amostral de ordem r na origem

$$M_r^* = \frac{\sum_{i=1}^n X_i^r}{n},$$

o qual é equivalente ao funcional $\mu_r^*(F) = \int x^r dF(x)$ e pode ser utilizado para estimar o correspondente momento populacional $\mu_r^*(F) \equiv \mu_r^*$. O domínio de $\mu_r^*(F)$, $\mathcal{D}(\mu_r^*)$, é o conjunto das distribuições para as quais μ_r^* existe. É fácil verificar que $\mathcal{D}(\mu_r^*)$ contém todas as distribuições empíricas e que é convexo. É também imediato que

$$IF(x;\mu_r^*,F) = x^r - \mu_r^*(F),$$

com variância assintótica

$$V(\mu_r^*,F) = \mu_{2r}^*(F) - \left[\mu_r^*(F)\right]^2,$$

desde que exista $\mu_{2r}^*(F)$. Conclui-se então que $\mu_r^*(\cdot)$ não é B-robusto para as distribuições com suporte ilimitado e que o mesmo acontece com qualquer função diferenciável destes momentos, ou seja, com qualquer estimador regular obtido pelo método dos momentos.

Tal como $\mu(F) \equiv \mu_1^*(F)$, $\mu_r^*(F)$ é um funcional linear para o qual são exactas as expansões de primeira ordem (2.16) e (2.17), mas o mesmo não acontece com funções não lineares destes funcionais.

Exemplo 2.7. Neste exemplo procede-se com algum detalhe à determinação da função de influência da mediana no caso em que o modelo F corresponde a uma distribuição contínua, invertível no intervalo $]0,1[$, com densidade $f = F'$. O funcional equivalente ao estimador mediana amostral foi já apresentado no Exemplo 2.2 (ver página 31), não havendo no caso em apreço qualquer ambiguidade quanto à definição de F^{-1}. Comece-se por calcular explicitamente $F_{\varepsilon,x}$,

$$F_{\varepsilon,x}(u) = \begin{cases} (1-\varepsilon)F(u), & u < x \\ \\ (1-\varepsilon)F(u) + \varepsilon, & u \geq x \end{cases}$$

Esta função encontra-se representada (genericamente) na Figura 2.5. Para determinação do valor da mediana na distribuição contaminada, $m(F_{\varepsilon,x})$, é necessário considerar separadamente os casos $x > m(F)$ e $x < m(F)$:

Se $x > m(F)$ (ver Figura 2.5):

$$(1-\varepsilon)F(m(F_{\varepsilon,x})) = \frac{1}{2} \quad \Leftrightarrow \quad m(F_{\varepsilon,x}) = F^{-1}\left(\frac{1/2}{1-\varepsilon}\right).$$

Se $x < m(F)$:

$$(1-\varepsilon)F(m(F_{\varepsilon,x})) + \varepsilon = \frac{1}{2} \quad \Leftrightarrow \quad m(F_{\varepsilon,x}) = F^{-1}\left(\frac{1/2-\varepsilon}{1-\varepsilon}\right).$$

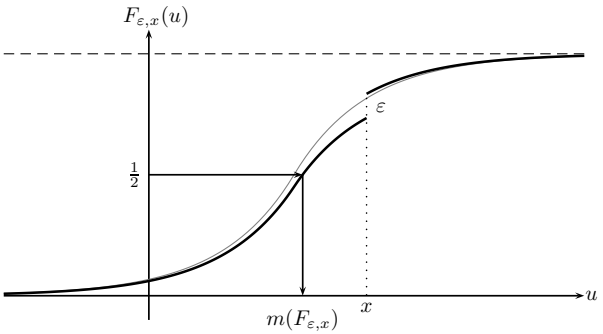


Figura 2.5 *Representação genérica de $F_{\varepsilon,x}$ (a traço cheio) sobre F (a traço fino) e determinação de $m(F_{\varepsilon,x})$ quando $x > m(F)$.*

Derivando em ordem a ε obtém-se

$$\frac{\partial}{\partial\varepsilon}m(F_{\varepsilon,x}) = \begin{cases} \dfrac{-1/2-2\varepsilon}{(1-\varepsilon)^2}\left[f\left(F^{-1}\left(\dfrac{1/2-\varepsilon}{1-\varepsilon}\right)\right)\right]^{-1}, & x < m(F) \\ \dfrac{1/2}{(1-\varepsilon)^2}\left[f\left(F^{-1}\left(\dfrac{1/2}{1-\varepsilon}\right)\right)\right]^{-1}, & x > m(F) \end{cases}.$$

Finalmente, fazendo $\varepsilon = 0$, obtém-se[15]

$$IF(x;m,F) = \frac{\partial}{\partial\varepsilon}m(F_{\varepsilon,x})\Big|_{\varepsilon=0} = \begin{cases} -\dfrac{1}{2f(m(F))}, & x < m(F) \\ 0, & x = m(F) \\ \dfrac{1}{2f(m(F))}, & x > m(F) \end{cases}.$$

Quanto às sensibilidades tem-se

$$\gamma^*(m,F) = \frac{1}{2f(m(F))}, \quad \lambda^*(m,F) = \infty \quad \text{e} \quad \rho^*(m,F) = \infty,$$

nos dois primeiros casos para qualquer modelo nas condições enunciadas e no caso de ρ^* se além disso F tiver suporte ilimitado. A variância assintótica calcula-se muito facilmente,

$$V(m,F) = E_F\left[IF(X;m,F)^2\right] = \frac{1}{4f^2(m(F))}.$$

Considerem-se agora dois casos particulares para ilustrar o cálculo da eficiência relativa da mediana amostral em relação à média amostral:

- $F_N \sim \mathcal{N}(\mu,\sigma^2)$, com $m(F_N) = \mu$, $f_N(m(F_N)) = (2\pi\sigma^2)^{-1/2}$, $\mu(F_N) = \mu$ e $\sigma^2(F_N) = \sigma^2$,

$$ARE_{m,\mu}(F_N) = \frac{V(\mu,F_N)}{V(m,F_N)} = \frac{\sigma^2}{(2\pi\sigma^2)/4} = \frac{2}{\pi} \simeq 0.637.$$

- $F_L \sim Laplace(\mu,b)$, com

$$f_L(x;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right),$$

[15] Para $x = m(F)$ é fácil verificar que $m(F_{\varepsilon,x}) \equiv m(F)$, $\forall_\varepsilon$.

$$m(F_L) = \mu,\ f_L(m(F_L)) = (2b)^{-1},\ \mu(F_L) = \mu \text{ e } \sigma^2(F_L) = 2b^2,$$

$$ARE_{m,\mu}(F_L) = \frac{V(\mu, F_L)}{V(m, F_L)} = \frac{2b^2}{1/(4(2b)^{-2})} = 2.$$

Enquanto o primeiro facto parece ser muito conhecido, "a eficiência da mediana é cerca de 64% da eficiência da média amostral", o facto de essa afirmação só ser válida para observações com distribuição exactamente normal parece ser ignorado. O segundo facto, "a eficiência da mediana é duas vezes a eficiência da média amostral para observações com distribuição de *Laplace*", já é aparentemente muito menos vezes referido, embora se analisado com mais cuidado, não seja nada estranho. O que acontece é que para a distribuição de *Laplace* a mediana é o estimador de máxima verosimilhança do parâmetro μ, portanto não é só mais eficiente que a média, é o estimador mais eficiente, sendo simples confirmar que se verifica (2.21).

Note-se ainda que estas comparações só fazem sentido porque os dois estimadores estão a estimar o mesmo parâmetro, uma vez que tratando-se de distribuições simétricas a média e a mediana teóricas coincidem. O mesmo tipo de comparação não seria válido tratando-se, por exemplo, de distribuições assimétricas.

Para terminar o exemplo, e com o objectivo de ilustrar algumas dificuldades da metodologia, não se pode deixar de estudar um caso discreto. Para isso considere-se para F o modelo discreto mais simples, $Bernoulli(p)$.

Se $p > 1/2$, $m(F) = 1$ e se $p < 1/2$, $m(F) = 0$. Para $p = 1/2$ a mediana não está bem definida pois a mediana inferior é zero, a mediana superior é um e a mediana dada pela definição do Exemplo 2.2 é $1/2$ (ver página 31).

A função de influência deve ser calculada apenas para os valores de x pertencentes a Ω, ou seja, $x = 0$ e $x = 1$.

Se $x = 0$ vem $F_{\varepsilon,x} \sim Bernoulli\,((1-\varepsilon)p))$ enquanto que se $x = 1$ vem $F_{\varepsilon,x} \sim Bernoulli\,((1-\varepsilon)p + \varepsilon))$. É fácil verificar que para ε suficientemente pequeno e desde que $p \neq 1/2$ nenhuma das contaminações produz alteração da mediana, ou seja, se $p > 1/2$, $m(F_{\varepsilon,x}) = 1$ e $p < 1/2$, $m(F_{\varepsilon,x}) = 0$, em ambos os casos quer para $x = 0$ quer para $x = 1$. Conclui-se então que

$$\text{se} \quad p \neq 1/2, \quad IF(x; m, F) = 0, \quad \forall_{x=0,1}.$$

Para $p = 1/2$ a situação é ainda mais estranha, pois para $x = 0$ vem $m(F_{\varepsilon,x}) = 0$ e para $x = 1$, $m(F_{\varepsilon,x}) = 1$, isto para qualquer $\varepsilon > 0$, o que significa, aplicando a definição, e considerando $m(F) = 1/2$, que

$$IF(x; m, F) = \begin{cases} -\infty, & x = 0 \\ \\ +\infty, & x = 1 \end{cases}$$

(se se pudesse considerar $\varepsilon < 0$, concluir-se-ia o oposto). Esta situação muito simples pode sem dificuldade ser generalizada a todas as distribuições discretas, permitindo afirmar que:

- Se a mediana está bem definida, isto é, se a mediana inferior é igual à mediana superior, então a função de influência da mediana é identicamente nula;

- Caso contrário, a função de influência vale $-\infty$ para valores de x inferiores ou iguais à mediana inferior e $+\infty$ nos restantes casos.

Como interpretar estes factos?

O facto de uma função de influência ser identicamente nula, implica que a aproximação de primeira ordem (2.16) não é válida e que para estudar o estimador por este tipo de métodos será necessário considerar termos de ordem superior naquele desenvolvimento. Significa também que $\text{var}(T_n)$ não é da ordem de $1/n$, mas converge para zero mais rapidamente. Obviamente que nesse caso $V(T, F)$ tal como foi definida será nula.

O segundo aspecto é mais grave e está relacionado com a falta de regularidade (não diferenciabilidade) do estimador.

Em qualquer um dos casos os problemas agora detectados não são mais do que a outra face de um problema conhecido em relação ao estimador mediana amostral e que consiste na impossibilidade de estimar a sua variância por *jackknife* através da expressão (2.22) (ver, por exemplo, Efron, 1982).

É fácil também perceber que as dificuldades encontradas com distribuições discretas se devem traduzir em dificuldades ao nível da curva de sensibilidade, uma vez que todas as amostras são discretas. De facto, este é um caso em que se verifica que não existe

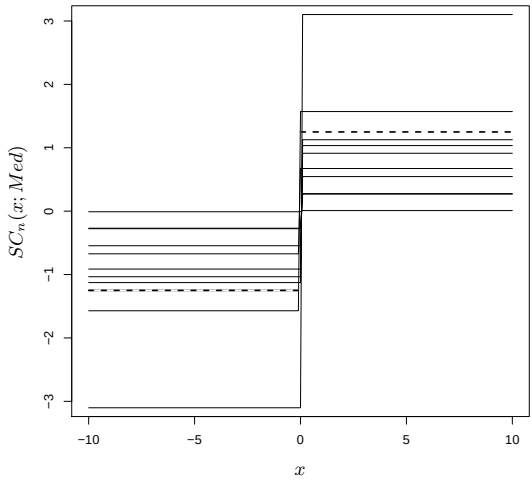


Figura 2.6 *Curvas de sensibilidade da mediana para 10 amostras de dimensão* $n = 100000$ *geradas a partir da distribuição* $\mathcal{N}(0,1)$ *e* $IF(x; m, \Phi)$ *(a tracejado).*

$\lim_{n\to\infty} SC_n(x; m)$, mesmo nos casos em que existe $IF(x; m, F)$ e se tem $X \sim F$. Uma pequena experiência de simulação permite observar este facto. Geraram-se aleatoriamente[16] 10 amostras com 10000 observações a partir da distribuição $\mathcal{N}(0,1)$ e calculou-se, para cada amostra, a curva de sensibilidade da mediana para $x \in [-10, 10]$. Em seguida repetiu-se o procedimento com 10 amostras de dimensão 100000. No gráfico da Figura 2.6 representam-se as curvas obtidas neste último caso ($n = 100000$) juntamente com a função de influência sob $F = \Phi$. O gráfico para $n = 10000$ (não apresentado) tem exactamente o mesmo aspecto. Embora este resultado experimental não faça demonstração da não convergência permite, no entanto, perceber que não se pode confiar na curva de sensibilidade para "estimar" a função de influência da mediana, mesmo usando uma amostra de dimensão extremamente elevada.

Exemplo 2.8. Como se viu na Secção 2.2 um estimador de localização de certa forma "intermédio" entre a média e a mediana (e que pode

[16] Ou, mais rigorosamente, pseudo-aleatoriamente, recorrendo à função `rnorm` do R.

até incluí-las como casos extremos) é a média aparada a $100 \times \alpha\%$, $0 < \alpha < 1/2$, dada por

$$T_{\alpha,n} = \frac{1}{n - 2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} X_{(i)}, \tag{2.32}$$

onde $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ representam as estatísticas de ordem da amostra $(X_1, \ldots, X_n)$. Este estimador é equivalente ao funcional

$$T_\alpha(F) = \frac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x \, dF(x) = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(x) \, dx,$$

a menos de uma correcção que desaparece quando n aumenta e que se deve ao facto de em (2.32) o número de observações que compõem a soma variar discretamente.[17] Através de cálculos algo semelhantes aos efectuados nos exemplos anteriores mas mais longos,[18] mostra-se que a função de influência de T_α é dada por

$$IF(x; T_\alpha, F) = \begin{cases} \frac{1}{1-2\alpha} \left(F^{-1}(\alpha) - W(F) \right), & x < F^{-1}(\alpha) \\ \frac{1}{1-2\alpha} \left(x - W(F) \right), & F^{-1}(\alpha) \leq x \leq F^{-1}(1-\alpha) \\ \frac{1}{1-2\alpha} \left(F^{-1}(1-\alpha) - W(F) \right), & x > F^{-1}(1-\alpha) \end{cases}$$

com $W(F) = (1-2\alpha) T_\alpha(F) + \alpha F^{-1}(\alpha) + \alpha F^{-1}(1-\alpha)$.

Na Figura 2.7 apresentam-se os gráficos das funções de influência dos estimadores média $(T_0(F))$, $T_{0.05}(F)$, $T_{0.10}(F)$ e mediana $(T_{0.50}(F))$ para a distribuição normal padrão $(F = \Phi)$.

Os valores da sensibilidade e da variância assintótica para a distribuição normal padrão, obtidos pelas fórmulas (2.24) e (2.18), respectivamente, são apresentados na Tabela 2.2. Analisando estes valores verifica-se, como esperado, a existência de uma relação inversa entre a sensibilidade e a variância assintótica: quanto menor é a sensibilidade maior é a variância assintótica, uma vez que maior é o afastamento em relação à função de influência do estimador de máxima de verosimilhança sob o modelo que está a ser considerado. Com $\alpha = 0$ tem-se

[17] É possível generalizar a definição de $T_{\alpha,n}$ de modo a permitir que aquele número varie continuamente, basta que as observações $x_{([\alpha n]+1)}$ e $x_{(n-[\alpha n])}$ entrem em (2.32) com peso $1 - p$ onde $p = \alpha n - [\alpha n]$, nesse caso já o estimador é exactamente equivalente ao funcional.

[18] Ver, por exemplo, Huber (1981, p. 57).

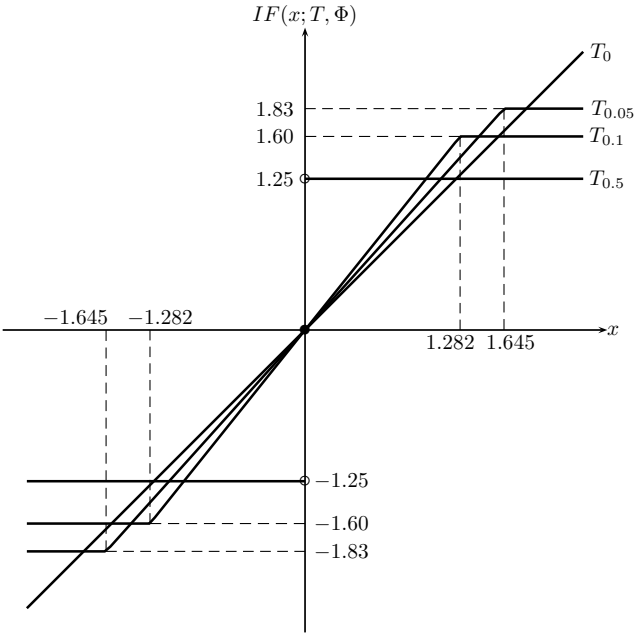


Figura 2.7 *Funções de influência da média* (T_0)*, de* $T_{0.05}$*,* $T_{0.10}$ *e da mediana* $(T_{0.50})$ *para a distribuição normal padrão.*

Tabela 2.2 *Valores da sensibilidade e da variância assintótica de* $T_{\alpha,n}$ *bem como da eficiência relativa de* $T_{\alpha,n}$ *em relação a* $\overline{X}$ *na distribuição* Φ*, para* $\alpha = 0,\ 0.05,\ 0.1,\ 0.5$*.*

T	$\gamma^*(T,\Phi)$	$V(T,\Phi)$	$ARE_{T,T_0}(\Phi) \times 100\%$
T_0	∞	1.000	100
$T_{0.05}$	1.83	1.026	97.5
$T_{0.10}$	1.60	1.060	94.3
$T_{0.50}$	1.25	1.571	63.7

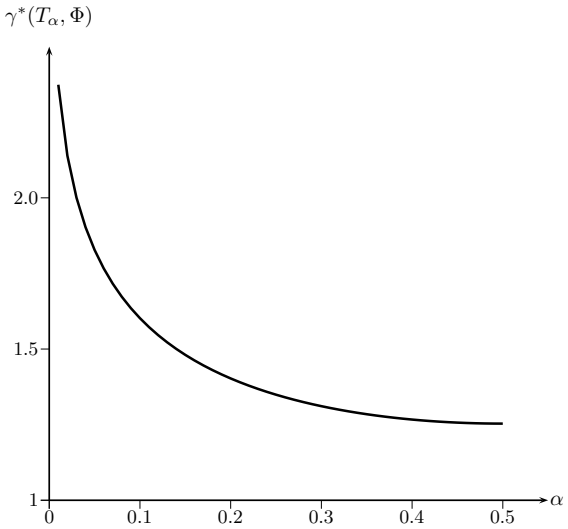


Figura 2.8 $\gamma^*(T_\alpha, \Phi)$ *em função de* α *para* $0.01 \leq \alpha \leq 0.5$.

um estimador não robusto com sensibilidade infinita, enquanto que com $\alpha = 0.05$ e $\alpha = 0.1$ se obtêm estimadores com baixa sensibilidade à custa de um decréscimo de eficiência que pode ser considerado baixo (2.5% no primeiro caso e 5.7% no segundo). Em relação à mediana que apresenta a menor sensibilidade (pode mesmo mostrar-se que é o estimador mais B-robusto do parâmetro de localização no modelo normal, ou seja aquele que apresenta sensibilidade mínima) já se constata uma perda de eficiência que pode ser considerada como um preço exagerado a pagar pela segurança extra oferecida.

Para completar a informação da Tabela 2.2 apresentam-se nas Figuras 2.8 e 2.9 gráficos representando, respectivamente, $\gamma^*(T_\alpha, \Phi)$ e $V(T_\alpha, \Phi)$, para $0.01 \leq \alpha \leq 0.5$.

Como se referiu anteriormente a propósito da definição de estimador B-robusto, este tipo de relação inversa entre sensibilidade e variância assintótica depende do modelo central considerado. Sendo geralmente verdadeira quando este é o modelo normal, pode não o ser noutros casos. Veja-se o exemplo anterior: a mediana para a distribuição de *Laplace* tem menor sensibilidade e menor variância assintótica que a média. Para este modelo pode pensar-se que não

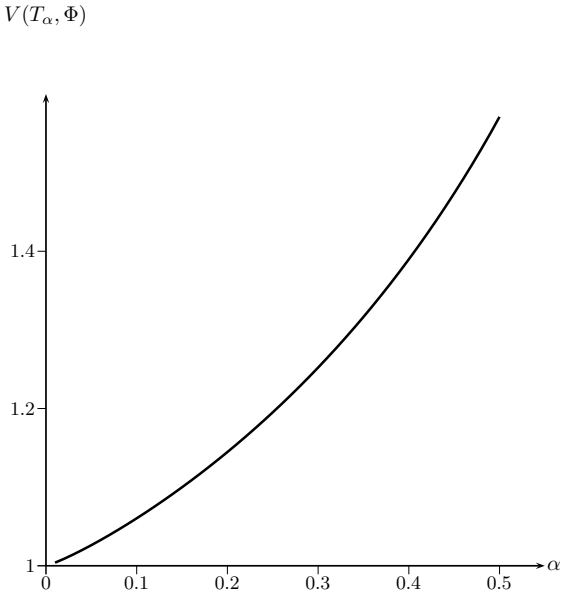


Figura 2.9 $V(T_\alpha, \Phi)$ *em função de* α *para* $0.01 \leq \alpha \leq 0.5$.

fará provavelmente sentido procurar estimadores alternativos ao estimador de máxima verosimilhança, no entanto a questão não é tão simples pois o que se pretende é precisamente não estar excessivamente dependente de um modelo mas usar um estimador que seja seguro numa vizinhança dele. Neste contexto faz sentido a utilização de estimadores sub-óptimos para o modelo mas que se comportem bem numa sua vizinhança, e as médias aparadas podem ser uma opção.

Exemplo 2.9. Considera-se agora o estimador desvio médio, já introduzido como medida de dispersão no Exemplo 2.1,

$$D_n = c\,\frac{\sum_{i=1}^{n} |X_i - \overline{X}|}{n}.$$

Este estimador é equivalente ao funcional

$$D(F) = c \int_{-\infty}^{+\infty} |x - \mu(F)|\; dF(x).$$

O papel da constante c é tornar o funcional consistente à Fisher para um dado modelo com um parâmetro de dispersão que se pretende estimar usando D_n. No caso do modelo normal para estimar σ tem-se $c = \sqrt{\pi/2}$, enquanto que no caso do modelo de *Laplace* se tem $c = 1$ para estimar b (como curiosidade refira-se que o estimador de máxima verosimilhança do parâmetro b do modelo de *Laplace* é muito semelhante ao desvio médio, a diferença é que os desvios são calculados em relação à mediana e não em relação à média).

Para calcular a função de influência comece-se por calcular o valor do funcional na distribuição contaminada $F_{\varepsilon,x}$ (para simplificar a notação usa-se $\mu \equiv \mu(F)$):

$$\begin{aligned} D(F_{\varepsilon,x}) &= c \int_{-\infty}^{+\infty} |u - \mu(F_{\varepsilon,x})| \, dF_{\varepsilon,x}(u) = \\ &= c \int_{-\infty}^{+\infty} |u - (\mu + \varepsilon(x-\mu))| \, dF_{\varepsilon,x}(u) = \\ &= c \left(\int_{-\infty}^{+\infty} (1-\varepsilon) \, |u - (\mu + \varepsilon(x-\mu))| \, dF(u) + \varepsilon(1-\varepsilon)|x-\mu| \right). \end{aligned}$$

Derivando esta expressão em ordem a ε obtém-se

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} \frac{D(F_{\varepsilon,x})}{c} = & \int - |u - \mu - \varepsilon(x-\mu)| \, dF(u) + \\ & + (1-\varepsilon) \int \frac{\partial}{\partial \varepsilon} |u - \mu - \varepsilon(x-\mu)| \, dF(u) + \\ & + (1-2\varepsilon)|x-\mu|, \end{aligned}$$

e, fazendo $\varepsilon = 0$, tem-se

$$\left. \frac{\partial}{\partial \varepsilon} \frac{D(F_{\varepsilon,x})}{c} \right|_{\varepsilon=0} = \int - |u-\mu| \, dF(u) + (\mu - x) \int \frac{|u-\mu|}{u-\mu} dF(u) + |x-\mu|,$$

pelo que

$$IF(x; D, F) = c \, |x-\mu| - D(F) + c \, (x-\mu)(2F(\mu) - 1).$$

Note-se que o último termo desaparece se F for uma distribuição simétrica, pois nesse caso $\mu(F)$ coincide com a mediana e $F(\mu) = 1/2$. Facilmente se verifica que, para um modelo com suporte ilimitado,

$$\gamma^*(D, F) = \infty, \quad \lambda^*(D, F) = c \, (1 + |2F(\mu) - 1|) \quad \text{e} \quad \rho^*(D, F) = \infty,$$

pelo que nesse caso o estimador D_n não é B-robusto (o que vem justificar a afirmação feita no Capítulo 1 no final do Exemplo 1.2, ver página 8). Por outro lado, em comparação com o desvio padrão, constata-se que, apesar de ambos terem função de influência ilimitada, o crescimento de $IF(x;\sigma,F)$ é muito mais rápido que o de $IF(x;D,F)$. É este aspecto que justifica o melhor comportamento de D_n em comparação com S_n nas distribuições contaminadas nas caudas consideradas naquele exemplo.

A variância assintótica calcula-se agora com alguma facilidade. Para F simétrica vem

$$V(D,F) = E_F\left[IF(X;D,F)^2\right] = \int \left(c\,|x-\mu(F)| - D(F)\right)^2 dF(x) =$$

$$= \int \left(c^2\,|x-\mu(F)|^2 - 2c\,D(F)\,|x-\mu(F)| + D^2(F)\right) dF(x) =$$

$$= c^2\sigma^2(F) - D^2(F),$$

expressão que é válida se existir $\sigma^2(F)$. Para F assimétrica a única diferença é que aparecem termos adicionais envolvendo $(2F(\mu)-1)$.

Esta expressão juntamente com a expressão de $V(\sigma,F)$, (2.31), obtida no Exemplo 2.6, permitem calcular os valores de $ARE(\varepsilon)$, dados por (1.3), e apresentados no Exemplo 1.2. De facto, com

$$F(x) = (1-\varepsilon)\Phi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\Phi\left(\frac{x-\mu}{3\,\sigma}\right) = (1-\varepsilon)F_1 + \varepsilon F_2,$$

e $c=\sqrt{\pi/2}$, cálculos relativamente simples[19] permitem concluir que, quando $n\to\infty$,

$$E(S_n) \to \sigma(F) = \sqrt{(1-\varepsilon)\sigma^2(F_1) + \varepsilon\sigma^2(F_2)} = \sqrt{1+8\varepsilon}\,\sigma,$$

$$E(D_n) \to D(F) = (1-\varepsilon)D(F_1) + \varepsilon D(F_2) = (1+2\varepsilon)\,\sigma,$$

$$n\,\mathrm{var}(S_n) \to V(\sigma,F) = \frac{3+240\varepsilon-(1+8\varepsilon)^2}{4(1+8\varepsilon)}\sigma^2,$$

$$n\,\mathrm{var}(D_n) \to V(D,F) = [c^2(1+8\varepsilon)-(1+2\varepsilon)^2]\sigma^2,$$

[19]Notar que F_1 e F_2 têm o mesmo valor esperado, pelo que a variância da mistura é a mistura das variâncias e o mesmo acontece para qualquer momento central, incluindo $D(F)$ e $\mu_4(F)$. Usou-se ainda o facto de que para a distribuição $\mathcal{N}(\mu,\sigma^2)$, $\mu_4(F)=3\sigma^4$.

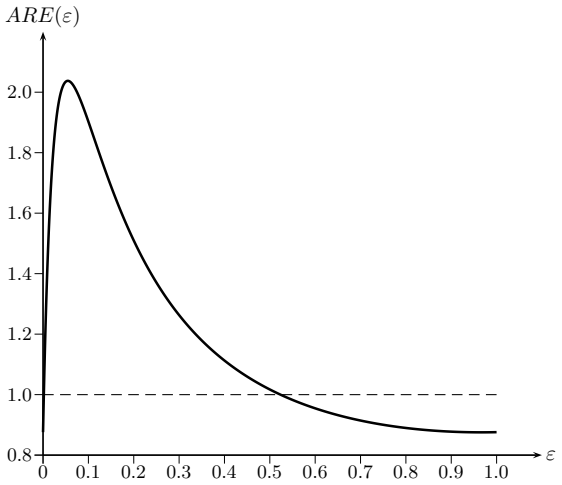


Figura 2.10 *$ARE(\varepsilon)$ em função de ε para $0 \leq \varepsilon \leq 1$.*

pelo que

$$ARE(\varepsilon) = \frac{[3 + 240\varepsilon - (1 + 8\varepsilon)^2](1 + 2\varepsilon)^2}{4[\frac{\pi}{2}(1 + 8\varepsilon) - (1 + 2\varepsilon)^2](1 + 8\varepsilon)^2},$$

de onde se obtêm os valores apresentados na Tabela 1.1. Note-se que o facto de no presente exemplo se utilizar a definição do desvio médio que inclui a multiplicação pela constante c, enquanto que no Exemplo 1.2 se omitiu essa constante, é irrelevante pois no cálculo de ARE c cancela, isto é, $\mathrm{var}(D_n)/E^2(D_n) = \mathrm{var}(cD_n)/E^2(cD_n)$.

Na Figura 2.10 apresenta-se como complemento da Tabela 1.1 o gráfico de $ARE(\varepsilon)$ em função de ε.

Exemplo 2.10. Em relação ao desvio absoluto mediano, definido em (2.6), tem-se que o funcional equivalente é

$$\mathrm{MAD}(F) = c\, m(G) = c\, G^{-1}\left(\frac{1}{2}\right),$$

com $c = 1/\Phi^{-1}(3/4)$ e onde F representa a distribuição de X e G a distribuição de $|X - m(F)|$. O funcional pode ainda ser definido

como a solução da equação

$$F\left(m(F)+\frac{\text{MAD}(F)}{c}\right)-F\left(m(F)-\frac{\text{MAD}(F)}{c}\right)=\frac{1}{2},$$

uma vez que $G(u)=F\left(m(F)+u\right)-F\left(m(F)-u\right)$.

Antecipam-se para este funcional e estimador o mesmo tipo de dificuldades encontradas em relação à mediana no Exemplo 2.7. De qualquer modo ilustra-se o cálculo da função de influência do MAD para um modelo F correspondente a uma distribuição contínua com suporte em $\mathbb{R}$, invertível no intervalo $]0,1[$ e com densidade $f=F'$.

Após alguns cálculos conclui-se que se X tiver a distribuição contaminada $F_{\varepsilon,x}$, a distribuição de $|X-m(F_{\varepsilon,x})|$ é dada por

$$G_{\varepsilon,x}(u)=\begin{cases}(1-\varepsilon)\left[F(m+u)-F(m-u)\right], & u<|x-m|\\ \\ (1-\varepsilon)\left[F(m+u)-F(m-u)\right]+\varepsilon, & u\geq|x-m|\end{cases},$$

onde, para simplificar a notação, $m\equiv m(F)$. Derivando $G^{-1}_{\varepsilon,x}(1/2)$, em ordem a ε e fazendo $\varepsilon=0$, tal como no Exemplo 2.7, obtém-se então

$$IF(x;\text{MAD},F)=\begin{cases}\dfrac{-c/2}{f\left(m+\frac{\text{MAD}}{c}\right)+f\left(m-\frac{\text{MAD}}{c}\right)}, & |x-m|<\frac{\text{MAD}}{c}\\ \\ \dfrac{c/2}{f\left(m+\frac{\text{MAD}}{c}\right)+f\left(m-\frac{\text{MAD}}{c}\right)}, & |x-m|>\frac{\text{MAD}}{c}\end{cases}.$$

Concretizando para a distribuição $N(0,1)$, com $F=\Phi$ e $f=\varphi$,

$$IF(x;\text{MAD},\Phi)=\begin{cases}-\dfrac{c}{4\varphi\left(\Phi^{-1}(\frac{1}{4})\right)}, & |x|<\Phi^{-1}(\frac{3}{4})\\ \\ \dfrac{c}{4\phi\left(\Phi^{-1}(\frac{1}{4})\right)}, & |x|>\Phi^{-1}(\frac{3}{4})\end{cases}.$$

Tem-se então,

$$\gamma^*(\text{MAD},\Phi)\simeq 1.166,\quad \lambda^*(\text{MAD},\Phi)=\infty\quad\text{e}\quad\rho^*(\text{MAD},\Phi)=\infty,$$

pelo que o estimador é B-robusto, aliás é o estimador mais B-robusto para o parâmetro σ do modelo normal (ver, por exemplo, Hampel *et al.*, 1986, p. 142).

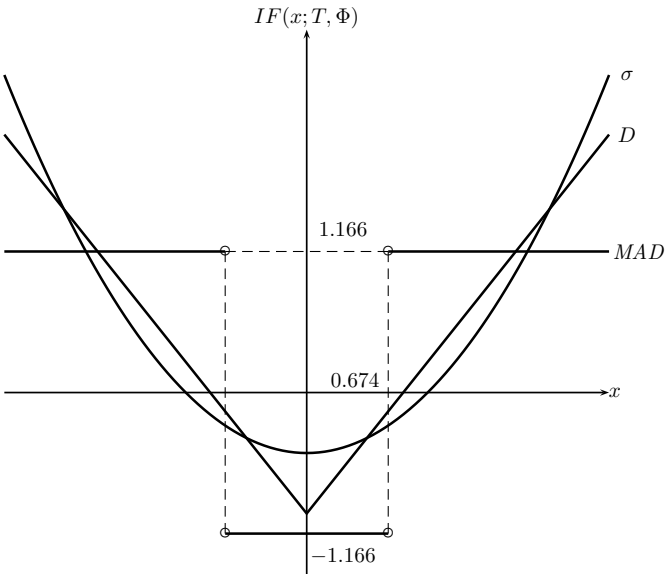


Figura 2.11 *Funções de influência do desvio absoluto mediano (MAD), do desvio padrão (σ) e do desvio médio (D) para a distribuição normal padrão.*

Na Figura 2.11 apresentam-se para $F = \Phi$ as funções de influência dos vários estimadores do parâmetro σ considerados até ao momento.

Para cálculo da eficiência tem-se $V(\text{MAD}, \Phi) = [\gamma^*(\text{MAD}, \Phi)]^2 \simeq 1.36$. Como termo de comparação, e utilizando os cálculos efectuados no exemplo anterior, tem-se $V(\sigma, \Phi) = 1/2$ e $V(D, \Phi) = \pi/2 - 1 \simeq 0.571$. Logo o MAD é para a distribuição normal muito pouco eficiente (37% em relação ao desvio padrão e 42% em relação ao desvio médio, mas claro que estes estimadores têm a desvantagem de não serem B-robustos). A seu favor o MAD tem apenas a muito baixa sensibilidade. Há ainda a considerar como desvantagem adicional, a instabilidade associada à não diferenciabilidade que é resultante do uso do funcional mediana, e cujos sintomas são, como se viu, a não existência de função de influência para distribuições discretas e a não convergência da curva de sensibilidade. Uma experiência de simulação semelhante à descrita no final do Exemplo 2.7 (ver página 53) conduziu aos resultados que se podem observar na Figura 2.12. A

situação parece ser ainda pior que em relação à mediana, pois deixa de se observar a simetria e surge mais uma descontinuidade. Globalmente o facto de o MAD ter pior comportamento que a mediana seria provavelmente de antecipar, uma vez que resulta da aplicação dupla, sucessiva, do funcional mediana.

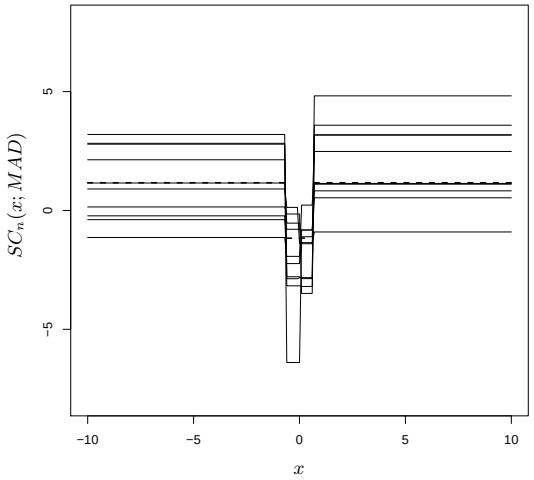


Figura 2.12 *Curvas de sensibilidade do desvio absoluto mediano para 10 amostras de dimensão* $n = 100000$ *geradas a partir da distribuição* $\mathcal{N}(0,1)$ *e* $IF(x; \text{MAD}, \Phi)$ *(a tracejado).*

Até este momento não foi considerado nenhum estimador razoável de dispersão. Enquanto o desvio padrão e o desvio médio não são robustos, o desvio absoluto mediano é robusto mas é pouco eficiente, muito instável e totalmente desaconselhado para dados discretos. Um outro estimador de escala por vezes referido na literatura é a distância ou amplitude inter-quartis (IQR, de *inter-quartile range*), definida como a diferença entre o terceiro e o primeiro quartis amostrais. No entanto, este estimador não constitui alternativa ao MAD, uma vez que sendo baseado em dois quantis amostrais sofre essencialmente dos mesmos problemas.[20] Nos exemplos seguintes consideram-se outros

[20] Aliás, pode provar-se que para distribuições simétricas são assintoticamente equivalentes, desde que o IQR seja multiplicado por uma constante equivalente à que aparece no MAD. Por exemplo para estimar o desvio padrão de dados normais o IQR deve ser multiplicado por $1/(\Phi^{-1}(3/4) - \Phi^{-1}(1/4)) \simeq 0.741$.

estimadores de escala com melhores propriedades que os até agora considerados.

Exemplo 2.11. Em relação à estimação de localização obteve-se um estimador B-robusto, "aparando" a média. Um procedimento do mesmo tipo pode ser aplicado à variância. Assim, define-se a variância aparada a $100 \times \alpha\%$, $0 < \alpha < 1/2$, como

$$S^2_{\alpha,n} = \frac{\gamma_n(\alpha)}{n - 2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} \left(X_{(i)} - T_n\right)^2, \tag{2.33}$$

onde $T_n \equiv T_n(X_1, \ldots X_n)$ representa um estimador de localização conveniente que, naturalmente, deverá ser robusto. Faz todo o sentido considerar a média aparada com o mesmo parâmetro α, mas poderia também utilizar-se a mediana. O estimador $S^2_{\alpha,n}$ é assintoticamente equivalente ao funcional

$$S^2_\alpha(F) = \frac{\gamma(\alpha)}{1 - 2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [x - T(F)]^2 \, dF(x). \tag{2.34}$$

Nas definições anteriores $\gamma_n(\alpha)$ e $\gamma(\alpha)$ são constantes tais que quando $X_i \sim F$, $E[S^2_{\alpha,n}] = \sigma^2(F)$ e $S^2_\alpha(F) = \sigma^2(F)$, ou seja, estas constantes fazem, respectivamente, com que o estimador seja centrado para estimar a variância de X e com que o funcional seja consistente segundo Fisher para o parâmetro $\sigma^2 = \sigma^2(F)$. Para a distribuição normal obtém-se

$$\gamma(\alpha) = \frac{1 - 2\alpha}{1 - 2\alpha - 2\xi\varphi(\xi)}, \tag{2.35}$$

com $\xi = \Phi^{-1}(1-\alpha)$ e $\varphi = \Phi'$. Se se pretender utilizar este estimador da variância em amostras de dimensão pequena convém calcular o valor das constantes $\gamma_n(\alpha)$, sendo para isso necessário usar os valores esperados dos quadrados das estatísticas de ordem. Para a distribuição normal padrão estes valores podem consultar-se em Teichroew (1956). Em Pires e Branco (1991) apresentam-se mais detalhes sobre o cálculo das constantes e alguns valores concretos precisamente para o modelo central normal (reproduzidos na Tabela 2.3). Para n grande pode utilizar-se $\gamma(\alpha)$ uma vez que $\lim_{n\to\infty} \gamma_n(\alpha) = \gamma(\alpha)$ (para $n \geq 25$ e $\alpha \leq 0.15$ pode considerar-se que o erro cometido é pequeno).

Tabela 2.3 *Valores das constantes $\gamma_n(\alpha)$ e $\gamma(\alpha)$ para correcção de $S^2_{\alpha,n}$ e S^2_α sob o modelo central normal.*

	α			
n	0.05	0.10	0.15	0.20
10	1.41	2.14	2.69	4.11
12	1.44	2.07	2.82	3.88
14	1.47	2.06	2.93	4.03
16	1.50	2.08	2.90	4.12
18	1.53	2.12	2.93	4.12
20	1.56	2.18	3.03	4.29
∞	1.61	2.28	3.23	4.66

É importante salientar que, ao contrário do que sucedia com a média aparada, no caso da variância aparada não basta dividir a soma aparada por $n - 2[\alpha n]$ (ou o integral por $1 - 2\alpha$) para estimar essencialmente o mesmo valor, sendo a constante $\gamma(\alpha)$ absolutamente necessária para esse efeito. A diferença entre os dois casos é que ao aparar a média, se "cortam" as $[\alpha n]$ parcelas inferiores e as $[\alpha n]$ parcelas superiores, quando se apara a variância, "cortam-se" aproximadamente as $2[\alpha n]$ maiores parcelas (em termos de $(x_i - T_n)^2$) pelo que não multiplicando pela constante se incorre sempre em subestimação da variabilidade dos dados.

Uma outra diferença em relação à média aparada é que no caso da variância aparada não faz sentido fazer $\alpha \to 1/2$, pois no limite fica apenas a observação central, cuja variância (nula) não fornece nenhuma informação sobre dispersão. Esta observação tem a ver com a anterior, pois ao fazer $\alpha \to 1/2$ observa-se $\gamma(\alpha) \to +\infty$.

A função de influência de $S^2_\alpha(F)$ (veja-se, por exemplo, Huber, 1981, p. 112) é dada por

$$IF(x; S^2_\alpha, F) = \begin{cases} \frac{\gamma(\alpha)}{1-2\alpha}\left(F^{-1}(\alpha)^2 - C(F)\right), & x_T < F^{-1}(\alpha) \\ \frac{\gamma(\alpha)}{1-2\alpha}\left(x_T^2 - C(F)\right), & F^{-1}(\alpha) \le x_T \le F^{-1}(1-\alpha) \\ \frac{\gamma(\alpha)}{1-2\alpha}\left(F^{-1}(1-\alpha)^2 - C(F)\right), & x_T > F^{-1}(1-\alpha) \end{cases}$$

com $x_T = x - T(F)$ e

$$C(F) = (1-2\alpha)S_\alpha^2(F)/\gamma(\alpha) + \alpha\left[F^{-1}(\alpha)\right]^2 + \alpha\left[F^{-1}(1-\alpha)\right]^2.$$

Para comparação com os estimadores anteriores tem mais interesse considerar o estimador $S_{\alpha,n} = \sqrt{S_{\alpha,n}^2}$ e o funcional $S_\alpha(F) = \sqrt{S_\alpha^2(F)}$, os quais vêm expressos nas unidades da variável e não no seu quadrado. Tal como se viu no Exemplo 2.6 em relação à variância e ao desvio padrão (ver páginas 47/48), utilizando a propriedade **(P4)**, tem-se

$$IF(x; S_\alpha, F) = \frac{1}{2\,S_\alpha(F)} IF(x; S_\alpha^2, F).$$

Na Figura 2.13 podem observar-se para $F = \Phi$ os gráficos das funções de influência do desvio padrão ($S_0(F)$) e de $S_{0.05}(F)$, $S_{0.10}(F)$ e $S_{0.15}(F)$. Verifica-se que para $\alpha > 0$ a sensibilidade é limitada, no entanto percebe-se que à medida que α cresce diminui o patamar superior mas simultaneamente diminui o ponto de mínimo (em $x = 0$), ou seja, a influência das observações centrais vai-se tornando cada vez maior. No limite $\alpha = 0.5$ ter-se-ia uma função de influência identicamente nula com um pico a valer $-\infty$ em $x = 0$, ou seja, um estimador não robusto e ainda por cima com variância infinita. Note-se que o valor $\alpha = 0.15$ corresponde aproximadamente a um ponto de equilíbrio em que $\max IF(x, S_\alpha, \Phi) = |\min IF(x, S_\alpha, \Phi)|$, pelo que se pode indicar $\alpha = 0.15$ como o valor a partir do qual a aplicação do estimador começa a não ser muito segura devido ao facto de as observações centrais começarem a ter demasiada influência.

Na Tabela 2.4 apresentam-se alguns valores da sensibilidade e da variância assintótica bem como da eficiência assintótica relativa de $S_{\alpha,n}$ em relação a S_n na distribuição Φ. Pode verificar-se que a sensibilidade começa por diminuir mas a partir de $\alpha = 0.15$ volta a crescer. A variância assintótica por seu lado cresce sempre e bastante mais rapidamente que no caso das médias aparadas. Para $\alpha = 0.15$ tem-se já uma eficiência de apenas 50% (relativamente ao desvio padrão), o que apesar de ser muito baixo é ainda assim bastante melhor que a do MAD.

Para completar a informação da Tabela 2.4 apresentam-se nas Figuras 2.14 e 2.15 gráficos representando, respectivamente, $\gamma^*(S_\alpha, \Phi)$ e $V(S_\alpha, \Phi)$, para $0.01 \leq \alpha \leq 0.4$. Não se fez a representação para $0.4 < \alpha < 0.5$ devido ao forte crescimento de ambas as funções, como

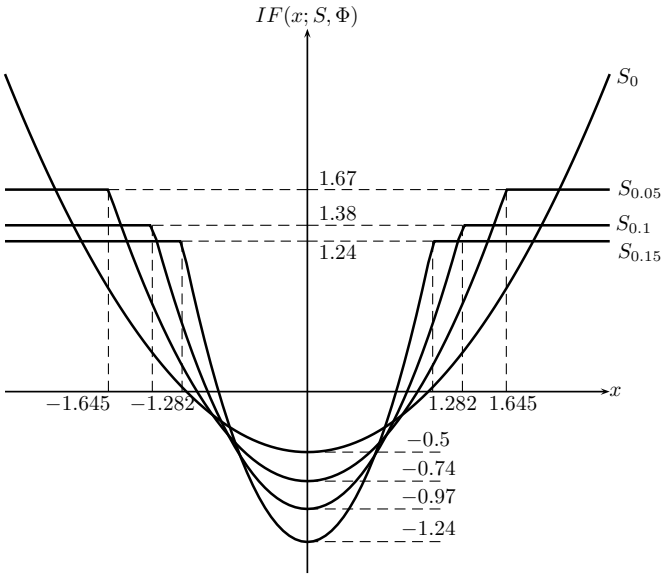


Figura 2.13 *Funções de influência do desvio padrão* (S_0) *e dos desvios padrões aparados* $S_{0.05}$*,* $S_{0.10}$ *e* $S_{0.15}$ *para a distribuição normal padrão.*

Tabela 2.4 *Valores da sensibilidade e da variância assintótica de* S_α *bem como da eficiência assintótica relativa de* $S_{\alpha,n}$ *em relação a* S_n *na distribuição* Φ*, para* $\alpha = 0,\ 0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25$.

S	$\gamma^*(S, \Phi)$	$V(S, \Phi)$	$ARE_{S,S_0}(\Phi) \times 100\%$
S_0	∞	0.50	100
$S_{0.05}$	1.67	0.64	78.0
$S_{0.10}$	1.38	0.80	62.8
$S_{0.15}$	1.24	0.99	50.3
$S_{0.20}$	1.60	1.26	39.8
$S_{0.25}$	2.09	1.63	30.7

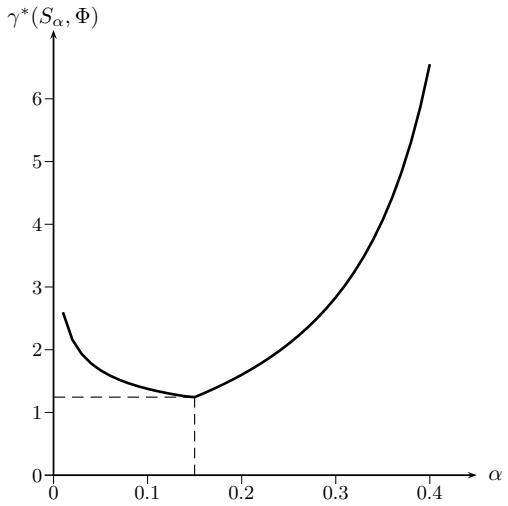


Figura 2.14 $\gamma^*(S_\alpha, \Phi)$ *em função de* α *para* $0.01 \leq \alpha \leq 0.4$.

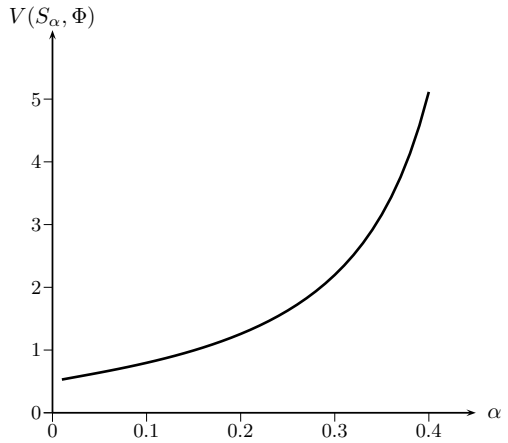


Figura 2.15 $V(S_\alpha, \Phi)$ *em função de* α *para* $0.01 \leq \alpha \leq 0.4$.

se referiu acima ambas tendem para $+\infty$ quando $\alpha \to 1/2$. O gráfico de $\gamma^*(S_\alpha, \Phi)$ tem além do mais uma assíntota vertical em $\alpha = 0$, pois $\gamma^*(S_0, \Phi) = +\infty$.

Exemplo 2.12. Rousseeuw e Croux (1993) propuseram um estimador de escala B-robusto e com muito melhores propriedades que o MAD e que o desvio padrão aparado. Esse estimador denotado por Q_n é definido explicitamente como o primeiro quartil da amostra formada pelos valores absolutos das diferenças entre todos os pares de observações, ou seja da amostra

$$(D_1, \ldots, D_{\binom{n}{2}}) = \{|X_i - X_j|,\ 1 \leq i < j \leq n\}\,.$$

Representando por $D_{(1)} \leq \cdots \leq D_{\left(\binom{n}{2}\right)}$ as correspondentes estatísticas de ordem, tem-se

$$Q_n(X_1, \ldots, X_n) = c\, D_{\left(\binom{[n/2]+1}{2}\right)}. \tag{2.36}$$

A função `Qn` do *package* `robustbase` do R permite calcular as estimativas correspondentes usando um algoritmo de ordenação eficiente. O funcional (assintoticamente) equivalente é dado por

$$Q(F) = c\, G^{-1}\left(\frac{1}{4}\right) = c\, H^{-1}\left(\frac{5}{8}\right),$$

onde G representa a distribuição de $|X - Y|$ e H a distribuição de $X - Y$, com X e Y variáveis aleatórias independentes com distribuição F (a igualdade resulta do facto de $X - Y$ ter distribuição simétrica para qualquer F). Mais uma vez o papel da constante multiplicativa c é fazer com que o funcional seja consistente à Fisher para um dado modelo. Se se considerar o modelo normal vem

$$c = \frac{1}{\sqrt{2}\Phi^{-1}(5/8)} \simeq 2.2191,$$

o que conduz a $Q(\Phi) = 1$. Tem interesse notar que Q_n não depende de nenhum estimador de localização, pelo que se diz que é *location--free* (dos outros estimadores de escala mencionados até agora apenas o IQR possui esta propriedade).

Rousseeuw e Croux (1993) mostraram que a função de influência de Q é dada, para uma distribuição contínua F, com densidade f,

por

$$IF(x;Q,F) = c\,\frac{\frac{1}{4} - F\left(x+\frac{1}{c}\right) + F\left(x-\frac{1}{c}\right)}{\displaystyle\int_{-\infty}^{+\infty} f\left(y+\frac{1}{c}\right) f(y)dy},$$

e que para a distribuição normal se tem $\gamma^*(Q,\Phi) \simeq 2.07$ e $V(Q,\Phi) = 0.608$, pelo que a eficiência assintótica relativamente ao estimador de máxima verosimilhança de σ é de 82.3% (pois $V(\sigma,\Phi)$=0.5). Conclui-se assim que, de todos os estimadores robustos de σ apresentados até agora (excluindo por razões óbvias o desvio padrão aparado com $\alpha \approx 0$), este é o que tem melhor eficiência. Na Figura 2.16 apresenta-se o gráfico de $IF(x;Q,\Phi)$, o qual em comparação com os gráficos apresentados nas Figuras 2.11 e 2.13 revela aparentemente boas propriedades deste estimador em termos de regularidade e permite perceber porque é que tem boa eficiência assintótica (porque a função de influência de Q é a que de mais perto acompanha a de σ na zona central do gráfico).

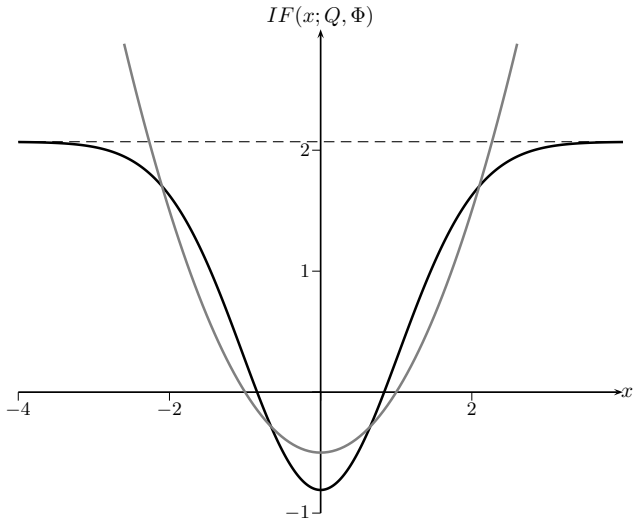


Figura 2.16 *Função de influência do estimador Q_n para a distribuição normal padrão (a curva menos carregada corresponde à função de influência do desvio padrão para o mesmo modelo).*

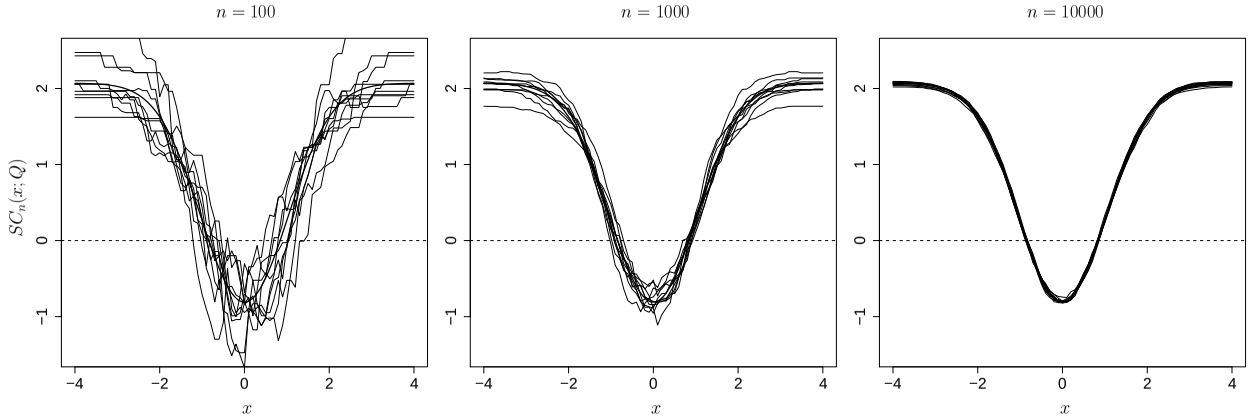


Figura 2.17 *Curvas de sensibilidade do estimador Q_n para $n = 100$, $n = 1000$ e $n = 10000$. Em cada caso usam-se 10 amostras geradas a partir da distribuição $\mathcal{N}(0, 1)$. Em cada um dos gráficos representa-se também $IF(x; Q_n, \Phi)$ (a traço mais forte).*

Apesar de tudo suspeita-se que os problemas de regularidade não ficaram completamente resolvidos, uma vez que continua a ser usado na definição do estimador um quantil amostral (se bem que da amostra das diferenças que tem uma dimensão muito maior que a amostra original) e não existe a função de influência para variáveis discretas. Para compreender melhor esta questão procedeu-se a um exercício de simulação sobre a curva de sensibilidade semelhante ao apresentado para o MAD no Exemplo 2.10. Assim, geraram-se aleatoriamente 10 amostras com 100 observações a partir da distribuição $\mathcal{N}(0,1)$ e calculou-se, para cada amostra, a curva de sensibilidade de Q_n para $x \in [-4,4]$. Em seguida repetiu-se o procedimento com 10 amostras de dimensão 1000 e 10 amostras de dimensão 10000. Nos gráficos da Figura 2.17 representam-se as curvas obtidas, juntamente com a função de influência sob $F = \Phi$. Verifica-se que embora cada curva seja irregular, essa irregularidade vai-se "desvanecendo" com o aumento de n e parece observar-se, embora lentamente, a convergência em probabilidade de $SC_n(x; Q_n)$ para $IF(x; Q, \Phi)$. Este comportamento é portanto diverso do observado no caso do MAD, e um indicador de que o estimador Q_n é mais "regular" que o MAD.

Exemplo 2.13. Para terminar esta série de exemplos analisam-se de novo os dados do Exemplo 2.1, apresentando mais algumas estimativas de localização e escala para os dados aí considerados e fazendo uma análise mais global à luz dos conceitos entretanto introduzidos.

Como se referiu no início da Secção 2.3 na estatística robusta também se admite a existência de um modelo paramétrico central gerador dos dados, embora se espere que esse modelo seja verificado apenas aproximadamente. Ou seja, toleram-se pequenos desvios do modelo, como por exemplo uma pequena percentagem de observações provenientes de outro modelo e que se podem manifestar nas observações como *outliers*. Isto significa que mesmo numa situação tão simples como a que está a ser analisada (24 observações univariadas) se deve especificar esse modelo e ter a ideia clara do que se pretende estimar. Se se ignorar o último valor da amostra e tendo em conta que é apenas uma aproximação, não é despropositado tomar como modelo central o modelo normal, $X \sim \mathcal{N}(\mu, \sigma^2)$. O que se pretende então é **estimar de forma robusta** os parâmetros μ e σ deste modelo, pelo que devem apenas ser usados estimadores consistentes destes parâmetros.

O facto de se considerar um modelo simétrico facilita a questão

em termos do parâmetro μ. Se tivesse havido necessidade de optar por um modelo assimétrico já seria preciso ter mais cuidado pois, por exemplo, a média e a mediana são consistentes para a média e a mediana populacionais, que seriam nesse caso diferentes. Obviamente que só se deve admitir a situação simétrica se ela fizer de facto sentido, o que aqui parece acontecer.[21]

Na Tabela 2.5 repetem-se as estimativas do parâmetro μ já apresentadas como medidas de localização ou tendência central na Tabela 2.1 e apresentam-se ainda as estimativas dadas pelas médias aparadas a 15%, 20% e 25%. Na Tabela 2.6 apresentam-se as estimativas do parâmetro σ obtidas anteriormente (s, d, MAD) e ainda o IQR, os desvios aparados com $\alpha = 0.05$, 0.1 e 0.15, bem como a estimativa q correspondente ao estimador Q_n. Como se usam as correcções para a distribuição normal todos os valores correspondem a estimativas do parâmetro σ do modelo central e são portanto comparáveis.

Tabela 2.5 *Diversas estimativas do parâmetro de localização para os dados do Exemplo 2.1.*

Amostra	$\overline{x}$	$\overline{x}_{0.05}$	$\overline{x}_{0.1}$	$\overline{x}_{0.15}$	$\overline{x}_{0.2}$	$\overline{x}_{0.25}$	$\overline{x}_{0.5}$
$(x_1, \ldots, x_{24})$	4.28	3.25	3.21	3.22	3.24	3.27	3.39
$(x_1, \ldots, x_{23})$	3.21	3.16	3.18	3.19	3.21	3.23	3.37

Tabela 2.6 *Diversas estimativas do parâmetro de dispersão para os dados do Exemplo 2.1.*

Amostra	s	d	MAD	IQR	q	$s_{0.05}$	$s_{0.1}$	$s_{0.15}$
$(x_1, \ldots, x_{24})$	5.30	2.68	0.53	0.69	0.63	0.78	0.66	0.70
$(x_1, \ldots, x_{23})$	0.69	0.66	0.50	0.67	0.63	0.61	0.66	0.70

Os resultados apresentados nas Tabelas 2.5 e 2.6 parecem coerentes com o que foi visto nos exemplos anteriores sobre regularidade,

[21] Se uma distribuição simétrica não parecer de todo razoável é sempre de considerar a possibilidade de transformar os dados de forma a que uma distribuição simétrica seja adequada para os dados transformados. Uma transformação que costuma ser usada com sucesso para este objectivo é a transformação logarítmica, ou mais raramente uma outra transformação da família de transformações de Box-Cox.

eficiência e robustez dos diversos estimadores. Em termos do parâmetro de localização todos as estimativas se situam à volta de 3.2 a 3.3, com excepção da média da amostra completa (sintoma de não robustez) e da mediana (pode ser sintoma quer da falta de eficiência quer da irregularidade). Quanto às estimativas do parâmetro de dispersão observa-se algo semelhante, todos os valores se situam à volta de 0.7, com excepção dos não robustos desvio padrão e desvio médio para a amostra completa e do MAD (também neste caso o mais ineficiente e irregular). Para terminar apresenta-se na Figura 2.18 o gráfico quantil-quantil (para a distribuição normal) da amostra completa juntamente com duas rectas correspondentes a dois modelos possíveis: a tracejado a recta correspondente à distribuição normal com parâmetros iguais às estimativas de máxima verosimilhança ($\hat{\mu} = 4.28$ e $\hat{\sigma} = 5.3$) e a cheio a recta correspondente às estimativas robustas consensuais ($\hat{\mu} = 3.2$ e $\hat{\sigma} = 0.7$). Verifica-se que a segunda não só se ajusta muito melhor à maioria das observações como revela o *outlier* de forma muito mais clara.

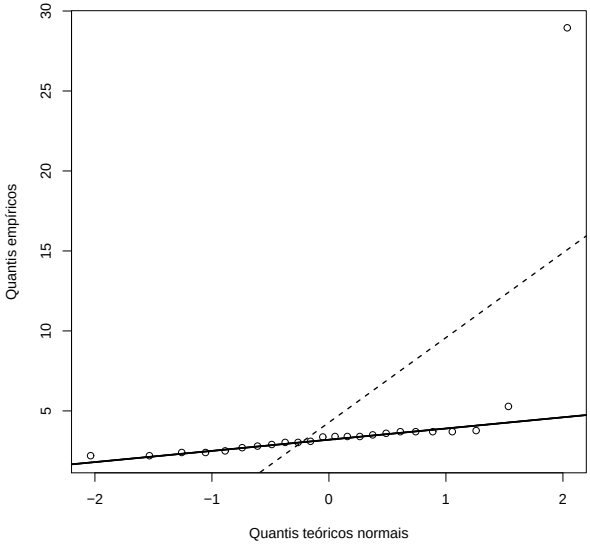


Figura 2.18 *Gráfico quantil-quantil (para a distribuição normal) da amostra completa do Exemplo 2.1. A tracejado: recta correspondente à distribuição normal com* $\hat{\mu} = 4.28$ *e* $\hat{\sigma} = 5.3$. *A cheio: recta correspondente à distribuição normal com* $\hat{\mu} = 3.2$ *e* $\hat{\sigma} = 0.7$.

2.3.5 Problemas multivariados e multiparamétricos

Nesta secção apresenta-se a generalização da função de influência e dos conceitos dela derivados para estimadores de parâmetros vectoriais com base em observações que podem ser ou não multivariadas.

Seja $\mathbf{\Omega} \subset \mathbb{R}^m$ o espaço amostra, $\mathbf{\Theta} \subset \mathbb{R}^p$ o espaço paramétrico e $\mathbf{T}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ um estimador (vectorial) do parâmetro $\boldsymbol{\theta}$, equivalente a um funcional $\mathbf{T}$.[22]

Definição 2.9. *A função de influência, $IF : \mathbf{\Omega} \subset \mathbb{R}^m \longrightarrow \mathbb{R}^p$, do funcional $\mathbf{T}$ em F é a função*

$$IF(\mathbf{x}; \mathbf{T}, F) = \lim_{\varepsilon \to 0^+} \frac{\mathbf{T}((1-\varepsilon)F + \varepsilon \Delta_{\mathbf{x}}) - \mathbf{T}(F)}{\varepsilon}, \qquad (2.37)$$

definida pontualmente nos pontos $\mathbf{x} \in \Omega \subset \mathbb{R}^m$ para os quais o limite existe e onde $\Delta_{\mathbf{x}}$ representa a distribuição (multivariada) de uma variável aleatória degenerada no ponto $\mathbf{x}$.

As observações feitas na Secção 2.3 imediatamente a seguir à definição da função de influência para o caso univariado (Definição 2.4, página 35) mantêm-se válidas após as necessárias adaptações de notação. Quanto às propriedades **(P1)** a **(P5)** tem-se (sob condições de regularidade que incluem a diferenciabilidade do funcional):

(P1) O valor esperado em F de IF é nulo, isto é,

$$E_F\left[IF(\mathbf{X}; \mathbf{T}, F)\right] = \int IF(\mathbf{x}; \mathbf{T}, F) dF(\mathbf{x}) = \mathbf{0},$$

onde $\mathbf{0}$ representa um vector com p zeros.

(P2) Para G próxima de F tem-se (fórmula de Taylor com um termo),

$$\mathbf{T}(G) \simeq \mathbf{T}(F) + \int IF(\mathbf{x}; \mathbf{T}, F) dG(\mathbf{x}),$$

[22] Adaptando a Definição 2.2 com $\mathbf{T} : \ F \in \mathcal{D}(\mathbf{T}) \subset \mathcal{F} \longrightarrow \mathbf{T}(F) \in \mathbb{R}^p$, onde $\mathcal{D}(\mathbf{T})$ representa o domínio de $\mathbf{T}$ e $\mathcal{F}$ o espaço das distribuições sobre $\mathbb{R}^m$.

em particular

$$\mathbf{T}_n = \mathbf{T}(F_n) \simeq \mathbf{T}(F) + \int IF(\mathbf{x}; \mathbf{T}, F) dF_n(\mathbf{x}) = \\ = \mathbf{T}(F) + \sum_{i=1}^{n} \frac{IF(\mathbf{x}_i; \mathbf{T}, F)}{n},$$

onde F_n representa a distribuição empírica multivariada associada à amostra $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

(P3) Se $\mathbf{X}_i \overset{\text{iid}}{\sim} F$ então $\sqrt{n}(\mathbf{T}_n - \mathbf{T}(F))$ tem distribuição assintoticamente normal (multivariada) com vector valor médio nulo e matriz de covariâncias $(p \times p)$ dada por

$$V(\mathbf{T}, F) = \lim_{n \to \infty} n \operatorname{var}(\mathbf{T}_n) = \\ = E_F \left[IF(\mathbf{X}; \mathbf{T}, F) IF(\mathbf{X}; \mathbf{T}, F)^T \right] = \\ = \int IF(\mathbf{x}; \mathbf{T}, F) IF(\mathbf{x}; \mathbf{T}, F)^T dF(\mathbf{x}) \qquad (2.38)$$

(que se designará daqui por diante indiferenciadamente por variância assintótica ou por matriz de covariâncias assintótica). Ou seja,

$$\sqrt{n}(\mathbf{T}_n - \mathbf{T}(F)) \overset{a}{\sim} \mathcal{N}_p \left(\mathbf{0}_p, V(\mathbf{T}, F) \right).$$

(P4) Seja $\beta(\boldsymbol{\theta})$ uma transformação regular do parâmetro, $\beta : \boldsymbol{\Theta} \subset \mathbb{R}^p \longrightarrow \beta(\boldsymbol{\Theta}) \subset \mathbb{R}^k$, com matriz Jacobiana $(k \times p)$,

$$\mathbf{B}(\boldsymbol{\theta}) = \frac{\partial \beta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

e $\beta(\mathbf{T}_n)$ um estimador de $\beta(\boldsymbol{\theta})$, então

$$IF(\mathbf{x}; \beta(\mathbf{T}), F) = \mathbf{B}(\mathbf{T}(F)) IF(\mathbf{x}; \mathbf{T}, F) \qquad (2.39)$$

e

$$V(\beta(\mathbf{T}), F) = \mathbf{B}(\mathbf{T}(F)) V(\mathbf{T}, F) \mathbf{B}(\mathbf{T}(F))^T. \qquad (2.40)$$

(P5) Desigualdade assintótica de Cramér-Rao para estimadores consistentes segundo Fisher (em relação ao parâmetro $\boldsymbol{\theta}$ de um dado modelo paramétrico $F_{\boldsymbol{\theta}}$ com "densdade" $f(\mathbf{x}, \boldsymbol{\theta})$). Seja

$$s(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\mathbf{x})$$

a função "score" de Fisher (gradiante da log-verosimilhança) e

$$\mathbf{J}(\boldsymbol{\theta}) = \int s(\mathbf{x}, \boldsymbol{\theta}) s(\mathbf{x}, \boldsymbol{\theta})^T dF_{\boldsymbol{\theta}}(\mathbf{x})$$

a matriz de informação de Fisher. Sob condições de regularidade verifica-se

$$V(\mathbf{T}, F_{\boldsymbol{\theta}}) \geq \mathbf{J}(\boldsymbol{\theta})^{-1},$$

onde "$\geq$" representa a relação de ordem usual (chamada forte) entre matrizes de covariâncias, isto é, dadas duas matrizes, $\mathbf{U}$ e $\mathbf{V}$, $\mathbf{U} \geq \mathbf{V}$ sse $\mathbf{U} - \mathbf{V}$ é semidefinida positiva.

Como se viu no caso univariado e uniparamétrico a sensibilidade, dada pelo supremo do valor absoluto da função de influência, é uma medida importante da robustez de um estimador. Para o caso multivariado e multiparamétrico tem-se a definição seguinte.

Definição 2.10. *A sensibilidade (não estandardizada) a grandes erros do estimador* $\mathbf{T}$ *em* F *é*

$$\gamma_u^*(\mathbf{T}, F) = \sup_{\mathbf{x}} \|IF(\mathbf{x}; \mathbf{T}, F)\|,$$

com $\|\cdot\|$ *representando a norma euclidiana.*

Em muitos modelos a parametrização é de certa forma arbitrária e é desejável que uma medida de robustez como a sensibilidade seja invariante em relação a pelo menos algumas transformações dos parâmetros. A sensibilidade definida anteriormente não satisfaz essa propriedade de invariância, daí que tenham surgido duas novas definições de sensibilidade.

Definição 2.11. *Se* $V(\mathbf{T}, F)$ *existir e for não singular chama-se sensibilidade auto-estandardizada (self-standardized sensitivity) a*

$$\gamma_s^*(\mathbf{T}, F) = \sup_{\mathbf{x}} \left(IF(\mathbf{x}; \mathbf{T}, F)^T V(\mathbf{T}, F)^{-1} IF(\mathbf{x}; \mathbf{T}, F) \right)^{1/2}.$$

Definição 2.12. *Se* $\mathbf{J}(\boldsymbol{\theta})$ *existir (para todo o* $\boldsymbol{\theta}$*) chama-se sensibilidade estandardizada pela informação (information standardized sensitivity) a*

$$\gamma_i^*(\mathbf{T}, F) = \sup_{\mathbf{x}} \left(IF(\mathbf{x}; \mathbf{T}, F)^T \mathbf{J}(\mathbf{T}(F)) IF(\mathbf{x}; \mathbf{T}, F) \right)^{1/2}.$$

Não é difícil verificar, usando (2.39) e (2.40), que γ_s^* e γ_i^* são invariantes para transformações bijectivas e não singulares de $\boldsymbol{\theta}$. Para efeitos da definição de B-robustez é indiferente qual das definições se usa.

Definição 2.13. $\mathbf{T}$ *é B-robusto em* F *se* $\gamma_u^*(\mathbf{T}, F) < \infty$.

Apresentam-se de seguida alguns exemplos relativos a estimadores multivariados.[23]

Exemplo 2.14. Em primeiro lugar considera-se o estimador clássico do valor esperado de um vector aleatório. Seja F a distribuição conjunta de um vector aleatório de dimensão m, $\mathbf{x} = (x_1, \ldots, x_m)^T$, com valor esperado

$$\boldsymbol{\mu} = E(\mathbf{x}) = (E(x_1), \ldots, E(x_m))^T = (\mu_1, \ldots, \mu_m)^T.$$

Dada uma amostra $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, com $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^T$, o estimador clássico de $\boldsymbol{\mu}$ é

$$\mathbf{T}_n = \overline{\mathbf{x}} = \frac{\sum_{i=1}^n \mathbf{x}_i}{n} = \begin{bmatrix} \dfrac{\sum_{i=1}^n x_{i1}}{n} \\ \vdots \\ \dfrac{\sum_{i=1}^n x_{im}}{n} \end{bmatrix} = \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_n \end{bmatrix}.$$

[23] A partir daqui torna-se difícil manter a convenção habitual de usar letras maiúsculas para variáveis aleatórias e estimadores e minúsculas para concretizações e estimativas, especialmente quando se trata de matrizes e vectores (que são, como é usual em análise multivariada, representados a **negrito**, as primeiras por maiúsculas e os segundos por minúsculas). Espera-se que a distinção fique clara a partir do contexto.

É imediato que este estimador é equivalente ao funcional (com domínio constituído pelas distribuições m-variadas com valor esperado finito)

$$\boldsymbol{\mu}(F) = \int \mathbf{x}\, dF(\mathbf{x}) = \begin{bmatrix} \int x_1 dF(\mathbf{x}) \\ \vdots \\ \int x_m dF(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \int x_1 dF_1(x_1) \\ \vdots \\ \int x_m dF_m(x_m) \end{bmatrix} = \begin{bmatrix} \mu(F_1) \\ \vdots \\ \mu(F_m) \end{bmatrix},$$

onde F_i representa a distribuição marginal de x_i. É simples verificar que

$$\boldsymbol{\mu}((1-\varepsilon)F + \varepsilon\Delta_{\mathbf{x}}) = (1-\varepsilon)\boldsymbol{\mu}(F) + \varepsilon\mathbf{x},$$

pelo que

$$IF(\mathbf{x}; \boldsymbol{\mu}, F) = \mathbf{x} - \boldsymbol{\mu}(F) = \begin{bmatrix} IF_1(\mathbf{x}; \boldsymbol{\mu}, F) \\ \vdots \\ IF_m(\mathbf{x}; \boldsymbol{\mu}, F) \end{bmatrix} = \begin{bmatrix} x_1 - \mu_1 \\ \vdots \\ x_m - \mu_m \end{bmatrix},$$

onde IF_i representa a i-ésima componente da função vectorial IF. Conclui-se imediatamente que, para um modelo com suporte ilimitado em pelo menos uma componente, a sensibilidade não estandardizada é infinita e que o estimador $\overline{\mathbf{x}}$ não é B-robusto. Quanto à variância assintótica tem-se, por (2.38),

$$V(\boldsymbol{\mu}, F) = E\left[(\mathbf{x} - \boldsymbol{\mu}(F))(\mathbf{x} - \boldsymbol{\mu}(F))^T\right],$$

pelo que, se existir a matriz de covariâncias,

$$\mathrm{var}(\mathbf{x}) = \boldsymbol{\Sigma} = E\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\right] = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & \ddots & \vdots \\ \sigma_{m1} & \cdots & \sigma_{mm} \end{bmatrix},$$

com $\sigma_{ii} = \mathrm{var}(x_i)$ e $\sigma_{ij} = \sigma_{ji} = \mathrm{cov}(x_i, x_j)$, então $V(\boldsymbol{\mu}, F) = \boldsymbol{\Sigma}$. Relativamente à normalidade assintótica ela está garantida, como no caso univariado, pela linearidade do estimador, a qual implica que a fórmula de Taylor com um termo é exacta. Este exemplo serviu basicamente para ilustrar os cálculos multivariados, não trazendo grandes novidades, uma vez que cada uma das componentes do estimador coincide com o estimador média amostral univariada já analisado exaustivamente.

Exemplo 2.15. Neste exemplo obtém-se a função de influência para o estimador clássico da matriz de covariâncias. Seja F a distribuição de um vector aleatório de dimensão m, $\mathbf{x} = (x_1, \ldots, x_m)^T$, com

valor esperado $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. Dada uma amostra $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, com $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^T$, o estimador clássico de $\boldsymbol{\Sigma}$ é

$$\mathbf{S}_n = \frac{\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T}{n-1} = \begin{bmatrix} S_{11} & \cdots & S_{1m} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mm} \end{bmatrix},$$

com $S_{jj} = (\sum_{i=1}^n x_{ij}^2 - n\overline{x}_j^2)/(n-1) = S_n^2(x_{1j}, \ldots, x_{nj})$ e

$$S_{jl} = \frac{\sum_{i=1}^n x_{ij}x_{il} - n\overline{x}_j\overline{x}_l}{n-1}.$$

Este estimador é assintoticamente[24] equivalente ao funcional (com domínio constituído pelas distribuições m-variadas com variância finita)

$$\boldsymbol{\Sigma}(F) = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T dF(\mathbf{x}) = \begin{bmatrix} \sigma_{11}(F) & \cdots & \sigma_{1m}(F) \\ \vdots & \ddots & \vdots \\ \sigma_{m1}(F) & \cdots & \sigma_{mm}(F) \end{bmatrix},$$

com

$$\sigma_{ij}(F) = \int x_i x_j dF(\mathbf{x}) - \mu(F_i)\mu(F_j), \tag{2.41}$$

tendo-se como é evidente $\sigma_{ii}(F) = \sigma^2(F_i)$.

Sendo $\boldsymbol{\Sigma}(F)$ um funcional matricial em vez de vectorial não se pode aplicar directamente a expressão (2.37) para o cálculo da sua função de influência. O que se pode fazer nestes casos é calcular a função de influência componente a componente e proceder depois à sua reorganização e à do funcional num vector conveniente (usando por exemplo o operador vec que transforma uma matriz num vector por justaposição das colunas da matriz). Com algum abuso de notação também se pode organizar numa matriz a função de influência de funcionais matriciais, só sendo necessário ter cuidados especiais na aplicação da fórmula (2.38) para cálculo da variância assintótica. Procedendo então ao cálculo da função de influência da componente ij de $\boldsymbol{\Sigma}(F)$, tem-se, com $F_{\varepsilon,\mathbf{x}} = (1-\varepsilon)F + \varepsilon\Delta_{\mathbf{x}}$, e após alguns cálculos,

$$\sigma_{ij}\left(F_{\varepsilon,\mathbf{x}}\right) = (1-\varepsilon)\sigma_{ij}(F) + \varepsilon(1-\varepsilon)(x_i - \mu_i(F))(x_j - \mu_j(F)),$$

[24] Seria exactamente equivalente se se tivesse usado na definição o denominador n em vez de $n-1$.

pelo que

$$IF(\mathbf{x}, \sigma_{ij}, F) = \frac{\partial}{\partial \varepsilon} \sigma_{ij}\left(F_{\varepsilon,\mathbf{x}}\right)\Big|_{\varepsilon=0} = (x_i - \mu_i(F))(x_j - \mu_j(F)) - \sigma_{ij}(F).$$

Note-se que com $i = j$ se recupera a função de influência de $\sigma^2(F)$. Se, abusando da notação, se organizar a função de influência de forma matricial, vem

$$IF(\mathbf{x}, \boldsymbol{\Sigma}, F) = (\mathbf{x} - \boldsymbol{\mu}(F))(\mathbf{x} - \boldsymbol{\mu}(F))^T - \boldsymbol{\Sigma}(F).$$

Novamente se conclui que para modelos com suporte ilimitado em alguma das componentes os estimadores em causa não são B-robustos.

Exemplo 2.16. Considere-se agora a determinação da função de influência do estimadores clássicos do coeficiente de correlação (coeficiente de correlação amostral) e da matriz de correlações (matriz de correlações amostral). Seja novamente F a distribuição de um vector aleatório de dimensão m, $\mathbf{x} = (x_1, \ldots, x_m)^T$, com valor esperado $\boldsymbol{\mu}$, matriz de covariâncias $\boldsymbol{\Sigma}$ e matriz de correlações

$$\mathbf{R} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1m} \\ \rho_{21} & 1 & \cdots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1} & \rho_{m2} & \cdots & 1 \end{bmatrix},$$

com $\mathbf{D} = \operatorname{diag}(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{mm}})$ e

$$\rho_{ij} = \rho_{ji} = \operatorname{cor}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Dada uma amostra $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, com $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^T$, o estimador clássico de ρ_{ij} é

$$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} = \frac{\sum_{l=1}^n x_{li}x_{lj} - n\overline{x}_i\overline{x}_j}{\sqrt{\left(\sum_{l=1}^n x_{li}^2 - n\overline{x}_i^2\right)\left(\sum_{l=1}^n x_{lj}^2 - n\overline{x}_j^2\right)}}$$

e o de $\mathbf{R}$ é

$$\mathbf{R}_n = \mathbf{D}_n^{-1}\mathbf{S}_n\mathbf{D}_n^{-1} = \begin{bmatrix} 1 & R_{12} & \cdots & R_{1m} \\ R_{21} & 1 & \cdots & R_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ R_{m1} & R_{m2} & \cdots & 1 \end{bmatrix},$$

com $\mathbf{D}_n = \text{diag}(\sqrt{S_{11}}, \ldots, \sqrt{S_{mm}})$. O estimador R_{ij} é equivalente ao funcional (com domínio constituído pelas distribuições m-variadas com variância finita)

$$\rho_{ij}(F) = \frac{\sigma_{ij}(F)}{\sqrt{\sigma_{ii}(F)\sigma_{jj}(F)}},$$

onde $\sigma_{ij}(F)$ é o funcional definido em (2.41). Como a função de influência deste funcional já foi determinada no exemplo anterior este é um caso em que é bastante útil a propriedade **(P4)**, em particular a expressão (2.39). Obtém-se então da aplicação dessa expressão,[25]

$$\begin{aligned}
&IF(\mathbf{x}; \rho_{ij}, F) = \\
&= \frac{\partial \rho_{ij}}{\partial \sigma_{ii}} IF(\mathbf{x}; \sigma_{ii}, F) + \frac{\partial \rho_{ij}}{\partial \sigma_{jj}} IF(\mathbf{x}; \sigma_{jj}, F) + \frac{\partial \rho_{ij}}{\partial \sigma_{ij}} IF(\mathbf{x}; \sigma_{ij}, F) = \\
&= \left[(x_i - \mu_i)^2 - \sigma_{ii}\right] \left(-\frac{1}{2} \frac{\sigma_{ij}}{\sigma_{ii}\sqrt{\sigma_{ii}\sigma_{jj}}}\right) + \\
&\quad + \left[(x_j - \mu_j)^2 - \sigma_{jj}\right] \left(-\frac{1}{2} \frac{\sigma_{ij}}{\sigma_{jj}\sqrt{\sigma_{ii}\sigma_{jj}}}\right) + \\
&\quad + \left[(x_i - \mu_i)(x_j - \mu_j) - \sigma_{ij}\right] \frac{1}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \\
&= \frac{(x_i - \mu_i)}{\sqrt{\sigma_{ii}}} \frac{(x_j - \mu_j)}{\sqrt{\sigma_{jj}}} - \frac{\rho_{ij}}{2} \left(\frac{(x_i - \mu_i)^2}{\sigma_{ii}} + \frac{(x_j - \mu_j)^2}{\sigma_{jj}} \right).
\end{aligned}$$

Esta expressão pode ser escrita de uma forma mais compacta como

$$IF(\mathbf{x}; \rho_{ij}, F) = z_i z_j - \frac{\rho_{ij}}{2}(z_i^2 + z_j^2),$$

onde z_h, $h = i, j$, representa o valor de x_h estandardizado da forma usual, isto é,

$$z_h = \frac{x_h - \mu_h}{\sqrt{\sigma_{hh}}} = \frac{x_h - \mu(F_h)}{\sigma(F_h)}.$$

Note-se que quando $i = j$, $R_{ij} \equiv 1$ e a função de influência é identicamente nula como não podia deixar de ser.

Mais uma vez se verifica que, para modelos com suporte ilimitado, a sensibilidade a grandes erros é infinita. Este exemplo vem mostrar

[25] Por uma questão de espaço omitiu-se o argumento F em todos os funcionais.

que, ao contrário do que se poderia ingenuamente pensar, mesmo um estimador que só toma valores num intervalo limitado, como é o caso da correlação, pode ter sensibilidade infinita. Isto acontece porque a influência é uma derivada ou taxa de variação face a contaminações pontuais e não um valor absoluto dessa variação.

Para ilustrar o comportamento da função de influência acabada de calcular apresentam-se na Figura 2.19 o gráfico de curvas de nível e o gráfico tridimensional de $IF(\mathbf{x}; \rho_{ij}, F)$, para uma distribuição bivariada F com $\mu_1 = \mu_2 = 0$, $\rho = 0.5$, $\sigma_{11} = 0.9$ e $\sigma_{22} = 1.2$, podendo observar-se que mesmo na gama limitada de valores de x_1 e x_2 considerada (o intervalo $[-3, 3]$) há pontos com influência muito elevada, no caso negativa. Os parâmetros escolhidos para a distribuição F são de certa forma arbitrários. Uma alteração em μ_1 e μ_2 provoca apenas uma translação sem alteração na forma. A alteração nos outros parâmetros provoca a rotação das rectas em que $IF(\mathbf{x}; \rho_{ij}, F) = 0$ (pode mostrar-se que estas coincidem com as rectas de regressão $E(x_1|x_2)$ e $E(x_2|x_1)$) e o crescimento da função ao longo das bissectrizes dos quadrantes mas não as suas características essenciais.

Mostrou-se nos exemplos acabados de apresentar que os estimadores tradicionais da localização multivariada, da matriz de covariâncias e do coeficiente de correlação não são B-robustos, pelo que convém agora dar alguma indicação relativa a estimadores B-robustos destes parâmetros, tão importantes na análise de dados multivariados.

Em relação à estimação do coeficiente de correlação, alguns estimadores não-paramétricos tradicionais, como por exemplo o coeficiente de correlação de Spearman ou o coeficiente de correlação de Kendall, representam um bom compromisso entre eficiência e robustez sobre o modelo central normal bivariado (Croux e Dehon, 2005).

Em relação à localização e dispersão multivariadas a situação é muito mais complexa. Poderia ingenuamente pensar-se que bastaria considerar estimadores univariados robustos para cada elemento do vector de localização (por exemplo a mediana) ou para cada elemento da matriz de covariâncias (por exemplo substituindo as variâncias amostrais pelo quadrado de algum dos estimadores de dispersão já apresentados e estimando as covariâncias com base na expressão $\sigma_{ij}(F) = \rho_{ij}(F)\sqrt{\sigma_{ii}(F)\sigma_{jj}(F)}$). No entanto verifica-se que os estimadores multivariados obtidos desta forma não têm propriedades desejáveis, nomeadamente não apresentam um bom comportamento

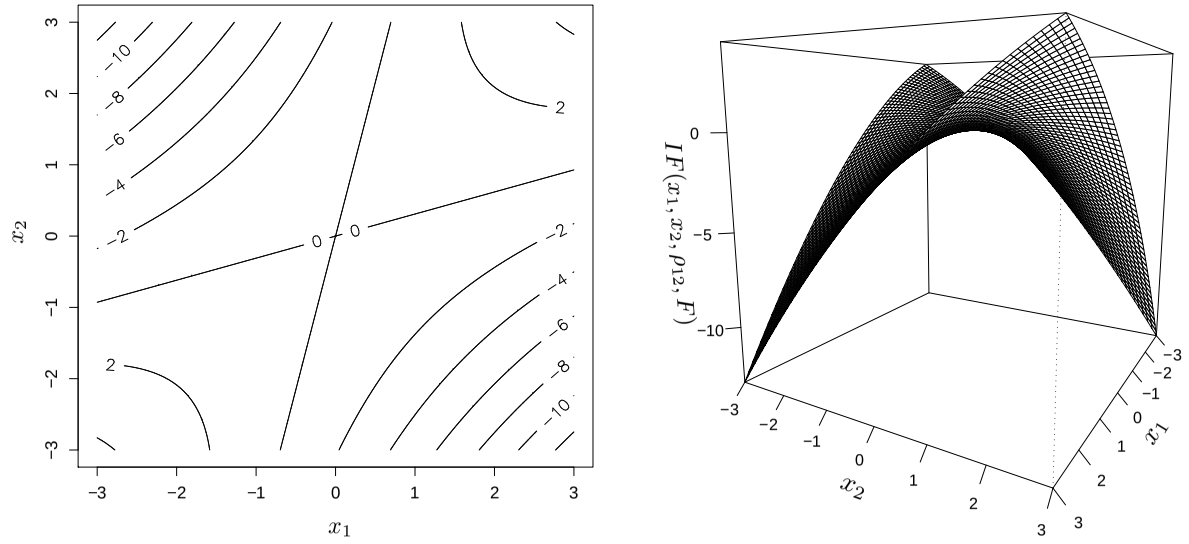


Figura 2.19 *Gráfico de curvas de nível e gráfico tridimensional de* $IF(\mathbf{x}; \rho_{ij}, F)$*, para uma distribuição bivariada* F *com* $\mu_1 = \mu_2 = 0$*,* $\rho = 0.5$*,* $\sigma_{11} = 0.9$ *e* $\sigma_{22} = 1.2$*.*

relativamente a transformações lineares das variáveis, pelo que não devem ser considerados. Infelizmente e por razões de espaço não é possível abordar em mais detalhe a questão da robustez em análise multivariada. Para o leitor interessado recomenda-se a leitura de Hubert *et al.* (2007) ou a consulta do Capítulo 6 de Maronna *et al.* (2006).

2.3.6 Generalizações

As definições da função de influência apresentadas nas secções anteriores podem ser aplicadas apenas a estimadores equivalentes a funcionais, ou seja a estimadores que dependem de uma única amostra apenas através da função de distribuição empírica. Isto exclui várias classes de estimadores:

- os que dependem de várias amostras, de que são exemplo as variâncias combinadas usadas na comparação de duas amostras, ou mais geralmente em análise de variância;

- os que dependem da ordem porque são recolhidas as observações e são usados em situações de dados dependentes, como por exemplo os estimadores da auto-covariância e auto-correlação em séries temporais.

São também excluídas da análise as estatísticas de teste, como por exemplo a conhecida estatística $T = \sqrt{n}\,\overline{X}/S_n$ usada para testar H_0: $\mu = 0$ contra H_1: $\mu \neq 0$ quando $X \sim \mathcal{N}(\mu, \sigma^2)$. Esta estatística não é equivalente a um funcional uma vez que duas amostras com a mesma função de distribuição empírica mas diferente dimensão conduzem a valores diferentes da estatística.

Generalizações da definição de função de influência destinadas a abarcar estas situações têm sido consideradas na literatura especializada. Por serem demasiado técnicas e para não alongar demasiado esta secção não se apresentam aqui os detalhes indicando-se apenas as referências bibliográficas mais relevantes.

Funções de influência para testes foram consideradas por Lambert (1981), Rousseeuw e Ronchetti (1981) e Michael e Schucany (1985). Um tratamento detalhado dessas funções pode também encontrar-se em Hampel *et al.* (1986, Cap. 3).

Künsch (1984) e Martin e Yohai (1986) propuseram generalizações para a situação de não independência, nomeadamente no contexto das séries temporais.

Pires e Branco (2002) estudaram o caso de funcionais dependentes de mais de uma distribuição, situação aplicável aos estimadores que dependem de mais do que uma amostra. Um exemplo deste tipo de estimadores foi já referido acima. Outros exemplos importantes são dados por diversos estimadores usados em análise multivariada, tais como, a distância de Mahalanobis entre os centróides de dois grupos, ou o vector dos coeficientes discriminantes em análise discriminante linear.

A função de influência que se acabou de estudar não constitui o único instrumento para avaliar a robustez. A robustez qualitativa e em especial o ponto de rotura são outros instrumentos usados para essa avaliação.

2.4 Robustez qualitativa

A noção de robustez qualitativa foi também introduzida por Hampel (1968, 1971). Esta noção tem a ver com continuidade e complementa de certa maneira a noção de diferenciabilidade contida na função de influência. A ideia é simples e pode de uma forma simplista traduzir-se como: um estimador será robusto se dadas duas distribuições próximas, F e G, as estimativas baseadas em amostras obtidas dessas populações também estiverem "próximas". Esta ideia é formalizada nas definições seguintes.

Definição 2.14. *Uma sequência de estimadores* $\{T_n; n \geq 1\}$ *é qualitativamente robusta em* F *se*

$$\forall_{\varepsilon>0}\exists_{\delta>0,n_0} : \forall_{G\in\mathcal{F}}\forall_{n>n_0} d(F,G) < \delta \Rightarrow d\left(\mathcal{L}_F(T_n), \mathcal{L}_G(T_n)\right) < \varepsilon, \tag{2.42}$$

onde $d(F,G)$ *representa uma distância adequada no espaço das distribuições*[26] *e* $\mathcal{L}_F(T_n)$ *significa "a distribuição de* T_n *quando as observações têm distribuição* F*".*

[26] Uma distância adequada é a distância de Prohorov cuja definição se optou

A robustez qualitativa corresponde em termos matemáticos à equicontinuidade da distribuição de T_n. Uma noção estreitamente relacionada com a de robustez qualitativa é a de continuidade do estimador que equivale em termos práticos a dizer que se duas distribuições empíricas estão "próximas" então, para que o estimador seja contínuo, as estimativas baseadas nas amostras correspondentes também devem estar próximas.

Definição 2.15. *Uma sequência de estimadores* $\{T_n; n \geq 1\}$ *é contínua em* F *se*

$$\forall_{\varepsilon>0}\exists_{\delta>0,n_0} : \forall_{n,m>n_0}\forall_{F_n,F_m} \\ (d(F,F_n) < \delta \wedge d(F,F_m) < \delta) \Rightarrow |T_n(F_n) - T_m(F_m)| < \varepsilon. \tag{2.43}$$

A robustez qualitativa e a continuidade do estimador são conceitos quase equivalentes. De facto são equivalentes se o estimador for consistente numa vizinhança do modelo. Este resultado constitui o teorema de Hampel e permite justificar a afirmação feita no Capítulo 1 de que resistência de uma estimativa e robustez de um funcional são noções praticamente equivalentes.

Por outro lado B-robustez e robustez qualitativa já não são equivalentes, a não ser em certas classes particulares de estimadores. Como se disse B-robustez tem a ver com diferenciabilidade enquanto robustez qualitativa tem a ver com continuidade. É possível dar exemplos de estimadores que são qualitativamente robustos mas não são B-robustos e de outros que são B-robustos mas não são qualitativamente robustos.

Note-se ainda que nenhum dos conceitos apresentados está relacionado com a continuidade (usual) de $T_n(x_1, \ldots, x_n)$ como função das observações. Por exemplo a média aritmética é uma função contínua das observações mas como estimador não é contínuo nem qualitativamente robusto para nenhuma distribuição F com suporte ilimitado.

---

por não apresentar por ser demasiado técnica. Para detalhes ver, por exemplo, Huber (1981).

Apesar de teoricamente úteis, e por isso não se quis deixar de os apresentar aqui, estes conceitos são de tratamento delicado sob o ponto de vista matemático pelo que a sua aplicação se tem restringido muito aos estimadores mais simples.

2.5 Ponto de rotura

Os conceitos de robustez derivados da função de influência são estritamente locais, ou seja descrevem o comportamento do estimador na presença de uma pequena percentagem de contaminação. Esta informação deve ser complementada com uma medida global que indique até que distância do modelo o estimador fornece alguma informação relevante. Por outras palavras essa medida deve indicar qual é a maior percentagem de contaminação que não produz valores arbitrários da estimativa. É este valor que se designa por ponto de rotura (*breakdown point*). Além disso o ponto de rotura dá uma indicação da distância até à qual é possível utilizar a aproximação linear dada pela função de influência através da equação (2.16). Rigorosamente há várias definições possíveis que normalmente dão informações coincidentes. Apresentam-se aqui apenas duas delas, uma para amostras finitas e outra em termos assintóticos. A definição assintótica foi introduzida também por Hampel na sua tese de doutoramento (Hampel, 1968) e posteriormente apresentada em Hampel (1971, 1974). Segundo o próprio Hampel, esta definição é baseada numa de Hodges (1967), "*tolerance for location estimators*". A definição para amostras finitas a seguir apresentada é também da autoria de Hampel (ver Hampel *et al.*, 1986, p. 97) e resulta da adaptação duma proposta de Donoho e Huber (1983).

Definição 2.16. *Numa dada amostra* $(x_1, \ldots, x_n)$ *substituam-se* m *observações (quaisquer)* $x_{i_1}, \ldots, x_{i_m}$ *por valores arbitrários* $y_1, \ldots, y_m$ *e designe-se a nova amostra por* $(z_1, \ldots, z_n)$*. O* ***ponto de rotura em dimensão finita*** *de* T_n *é dado por*

$$\varepsilon_n^*(T_n) = \varepsilon_n^*(T_n; x_1, \ldots, x_n) = \\ = \frac{1}{n} \max \left\{ m : \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots y_m} |T_n(z_1, \ldots, z_n)| < \infty \right\}. \quad (2.44)$$

Observações:

- A definição original de Donoho e Huber (1983) considera em vez da "maior percentagem de contaminação que não produz valores arbitrários da estimativa", a "menor percentagem que produz valores arbitrários da estimativa". Na prática as duas definições são equivalentes, sendo a diferença entre ambas de $1/n$.

- Note-se que habitualmente $\varepsilon_n^*(T_n)$ não depende de $(x_1, \ldots, x_n)$.

- A Definição 2.16 pode sem dificuldades ser adaptada para amostras e/ou estimadores multidimensionais.

- Esta definição foi no entanto proposta a pensar em estimadores de parâmetros com espaço paramétrico ilimitado, em particular $\Theta = \mathbb{R}$, como é o caso dos parâmetros de localização em modelos com suporte ilimitado. A definição pode também ser adaptada para estimadores de parâmetros de outro tipo. Pense-se por exemplo num estimador de um parâmetro de dispersão ($\Theta = \mathbb{R}^+$) ou num estimador do coeficiente de correlação ($\Theta =]-1, 1[$). Para traduzir a ideia que se pretende deve para estes casos escrever-se em (2.44), $T_n(z_1, \ldots, z_n) \in \Theta$ em vez de $|T_n(z_1, \ldots, z_n)| < \infty$.

- Embora não explicitamente mencionado deve também considerar-se que os valores arbitrários $y_1, \ldots, y_m$ são arbitrários no espaço amostra em causa (Ω). Isto significa que se, por exemplo, se estiver a trabalhar com proporções todos os valores mencionados devem pertencer ao intervalo $[0, 1]$.

Para a definição do ponto de rotura (assintótico) considere-se a vizinhança de contaminação (2.25) e o enviesamento assintótico máximo já definido em (2.26), dado por

$$b_{T,F}(\varepsilon) = \sup_{G \in \mathcal{P}_\varepsilon(F)} |T(G) - T(F)|. \tag{2.45}$$

Como $\mathcal{P}_1(F) = \mathcal{F}$ conclui-se que $b_{T,F}(1)$ é o pior valor possível de $b_{T,F}(\varepsilon)$. Este valor é igual a $+\infty$ no caso dos estimadores de localização em que $\Omega = \mathbb{R}$. Note-se que em (2.25) $\mathcal{F}$ representa o espaço de todas as distribuições sobre Ω, portanto se o modelo central F tiver suporte Ω, também H e por conseguinte G têm como suporte Ω.

Definição 2.17. *O ponto de rotura assintótico do estimador T_n, equivalente ao funcional T, é dado por*

$$\varepsilon^*(T, F) = \sup \{\varepsilon : b_{T,F}(\varepsilon) < b_{T,F}(1)\}.$$

Em muitos casos $\varepsilon^*(T, F)$ é independente de F, desde que F esteja numa certa classe. Por exemplo no caso dos estimadores de localização isso acontece para F com suporte ilimitado, ou seja para $\Omega = \mathbb{R}$. Também, na maior parte dos casos, se verifica que

$$\lim_{n\to\infty} \varepsilon_n^*(T_n) = \varepsilon^*(T, F).$$

Exemplo 2.17. Em relação aos estimadores de localização considerados na Secção 2.3.4 é muito fácil verificar que, se Ω for ilimitado,

- para a média aritmética, $\varepsilon_n^* = \varepsilon^* = 0$,
- para a média aparada a $100 \times \alpha\%$, $\varepsilon_n^* = [\alpha n]/n$ e $\varepsilon^* = \alpha$,
- para a mediana $\varepsilon_n^* = ([(n+1)/2] - 1)/n$ e $\varepsilon^* = 1/2$.

Se o modelo central for o modelo normal observa-se aqui mais uma vez o conflito entre eficiência e robustez, verificando-se que o ponto de rotura varia inversamente com a eficiência. Se se pretender usar, por exemplo, uma média aparada com um valor baixo de α, para maior eficiência, é conveniente ter em conta que a estimativa pode ser completamente arruinada por uma percentagem de contaminação superior a α.

Se Ω for limitado a situação muda completamente de figura. Suponha-se a título ilustrativo que F corresponde à distribuição uniforme em $[0, 1]$ e que se pretendia estimar $\mu(F) = 1/2$ usando a média aritmética. Para cálculo do ponto de rotura assintótico é necessário

determinar $b_{T,F}(\varepsilon)$. É fácil verificar que as contaminações, H, mais desfavoráveis para cálculo de $b_{T,F}(\varepsilon)$ são as distribuições degeneradas em 0 ou em 1 que conduzem ambas a $b_{T,F}(\varepsilon) = 0.5\varepsilon$. Por outro lado, $b_{T,F}(1) = 0.5$ donde se conclui que $\varepsilon^*(T, F) = 1$. De igual modo se concluiria que $\varepsilon_n^*(T_n) = 1$. Idêntico raciocínio permitiria concluir que o mesmo acontece para os restantes estimadores, médias aparadas e mediana.

Exemplo 2.18. Em relação aos estimadores de escala que foram também considerados na Secção 2.3.4 tem-se que, novamente se Ω for ilimitado,

- para a variância, desvio padrão e desvio médio, $\varepsilon_n^* = \varepsilon^* = 0$,
- para a variância e desvio padrão aparados a $100 \times \alpha\%$, $\varepsilon_n^* = [\alpha n]/n$ e $\varepsilon^* = \alpha$,
- para o desvio absoluto mediano, $\varepsilon_n^* = ([(n+1)/2] - 1)/n$ e $\varepsilon^* = 1/2$,
- para o IQR, $\varepsilon_n^* = ([(n+3)/4] - 1)/n$ e $\varepsilon^* = 1/4$,
- para o estimador Q_n, Rousseeuw e Croux (1993) mostraram que $\varepsilon_n^* = ([(n+1)/2] - 1)/n$ e $\varepsilon^* = 1/2$.

São válidos os comentários feitos no exemplo anterior a propósito dos estimadores de localização.

Neste caso é ainda importante referir que, em relação ao ponto de rotura, em dimensão finita, dos três últimos estimadores, MAD, IQR e Q_n, os valores apresentados só são válidos se se considerarem amostras genéricas provenientes de dados contínuos, isto é, para as quais a probabilidade de haver observações repetidas é nula. Croux e Rousseeuw (1992) apresentam um exemplo de uma amostra concreta em que, com $n = 11$, o ponto de rotura do MAD é zero devido ao facto de existirem observações repetidas.

Para outro tipo de estimadores a determinação do ponto de rotura (quer assintótico, quer em dimensão finita) pode não ser tão simples. Nesses casos é possível e usual avaliar o ponto de rotura em dimensão

finita pelo método de Monte Carlo, em estudos de simulação adequadamente delineados.

Como é fácil de concluir a partir das definições, e também dos exemplos apresentados, em geral $\varepsilon^* \in [0, 1]$. No entanto, em termos práticos, $\varepsilon^* = 1/2$ é o maior valor que tem significado uma vez que para contaminações superiores a 50% não há nenhum estimador que consiga distinguir entre a parte "boa" e a parte "má" da amostra.

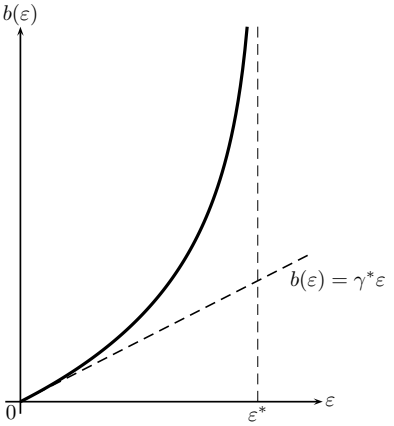


Figura 2.20 *Representação gráfica (qualitativa) de* $b(\varepsilon)$*, definido em (2.45), em função de* ε *para um estimador e um modelo hipotéticos tais que* $\gamma^* < \infty$ *e* $\varepsilon^* > 0$.

A relação entre todos os conceitos relacionados com robustez até agora apresentados pode ser explicada geometricamente através do gráfico do enviesamento assintótico máximo,[27] $b_{T,F}(\varepsilon)$ definido em (2.45), em função de ε. Na Figura 2.20 mostra-se um gráfico tipo para um estimador e um modelo hipotéticos tais que $\gamma^* < \infty$ e $\varepsilon^* > 0$.

O valor de ε onde $b(\varepsilon)$ regista uma assíntota vertical corresponde ao ponto de rotura ε^* (para que o gráfico faça sentido é necessário supor que $\varepsilon^* > 0$). Por sua vez, o facto de $b(\varepsilon)$ ser uma função contínua em $\varepsilon = 0$ significa que o estimador é qualitativamente robusto. Finalmente, e relacionando-se com a B-robustez, o declive da recta tangente à curva no ponto $\varepsilon = 0$ é γ^* (por 2.27).

[27] Costuma ser designado na literatura por "maxbias curve".

Exemplo 2.19. Nos casos em que a contaminação com uma massa pontual é a contaminação mais desfavorável e isso acontece frequentemente – em particular acontece sempre para os estimadores de localização – $b_{T,F}(\varepsilon)$ coincide com o supremo em x do módulo do numerador da expressão que define a função de influência, antes da passagem ao limite. Tomando como exemplo a mediana tem-se então (usando os cálculos apresentados no Exemplo 2.7 e para uma distribuição F nas condições aí consideradas) que

$$\begin{aligned} b_{m,F}(\varepsilon) &= \max\left\{b^+_{m,F}, \left|b^-_{m,F}\right|\right\} = \\ &= \max\left\{F^{-1}\left(\tfrac{1/2}{1-\varepsilon}\right) - F^{-1}\left(\tfrac{1}{2}\right), \left|F^{-1}\left(\tfrac{1/2-\varepsilon}{1-\varepsilon}\right) - F^{-1}\left(\tfrac{1}{2}\right)\right|\right\}. \end{aligned}$$

Se F for simétrica vem $b^+_{m,F} = \left|b^-_{m,F}\right|$, em particular para $F = \Phi$, como $m(\Phi) = \Phi^{-1}(1/2) = 0$, obtém-se

$$b_{m,\Phi}(\varepsilon) = \Phi^{-1}\left(\frac{1/2}{1-\varepsilon}\right),$$

função esta que se representa na Figura 2.21, juntamente com a recta $1.25\,\varepsilon$ (recorde-se que $\gamma^*(m,\Phi) \simeq 1.25$).

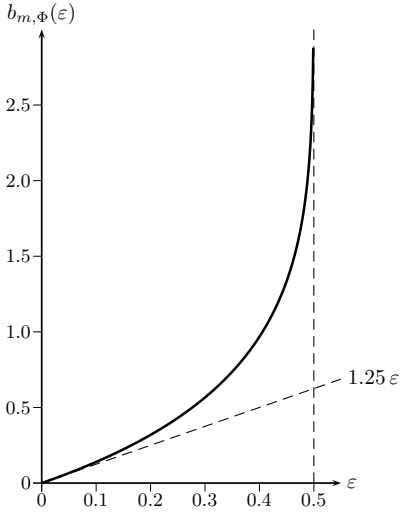


Figura 2.21 *Gráfico de* $b_{m,\Phi}(\varepsilon)$.

As Figuras 2.20 e 2.21 mostram, por um lado, como as diversas noções de robustez estão relacionadas com propriedades elementares de funções de uma variável – continuidade, diferenciabilidade, distância à singularidade – e por outro lado a interpretação do ponto de rotura como o valor a partir do qual deixa de fazer qualquer sentido aproximar linearmente o estimador utilizando a função de influência.

Enquanto o ponto de rotura é o ponto a partir do qual o estimador fornece informação totalmente desadequada, a função $b_{T,F}(\varepsilon)$ fornece uma melhor perspectiva, permitindo também dar uma ideia do ponto a partir do qual um estimador deixa de dar informação credível apesar de não totalmente arbitrária. Segundo Hampel *et al.* (1986) para contaminações superiores a cerca de 1/4 do ponto de rotura as estimativas podem não ser arbitrárias mas deixam de ser credíveis. A Figura 2.21 justifica de certa forma esta afirmação.

São este tipo de considerações que justificam a procura de estimadores com alto ponto de rotura,[28] em moda nalgumas escolas, mas por vezes mal compreendida. De facto, o usar um estimador com um ponto de rotura de 50% não significa que se esteja à espera de dados com contaminação próxima de 50%, significa sim que se espera uma protecção razoável para precentagens de contaminação até, digamos, cerca de 12.5%.

Embora a existência de ponto de rotura positivo esteja bastante relacionada com robustez qualitativa verifica-se que não há uma equivalência exacta entre os dois conceitos. Também não existe equivalência entre B-robustez ($\gamma^* < \infty$ ou função de influência limitada) e ponto de rotura positivo. De facto existem estimadores com função de influência ilimitada mas ponto de rotura positivo. Alguns exemplos concretos são mencionados no Capítulo 4. O que se passa nestes casos é que $b(\varepsilon)$ é contínua em $\varepsilon = 0$ mas tem derivada infinita nesse ponto. Na Figura 2.22 apresenta-se um gráfico tipo para um estimador e um modelo hipotéticos tais que $\gamma^* = \infty$ e $\varepsilon^* > 0$, ou seja, correspondente a um estimador que não é B-robusto (tem função de influência ilimitada) mas é robusto segundo o critério do ponto de rotura. É evidente que um estimador desde tipo pode ser preferível a um estimador não robusto mas será em princípio preferível usar, caso exista, um estimador com o mesmo ponto de rotura e $\gamma^* < \infty$.

Para terminar esta secção é importante referir que o conceito de

[28] Por alto ponto de rotura entende-se geralmente um valor próximo de 50%.

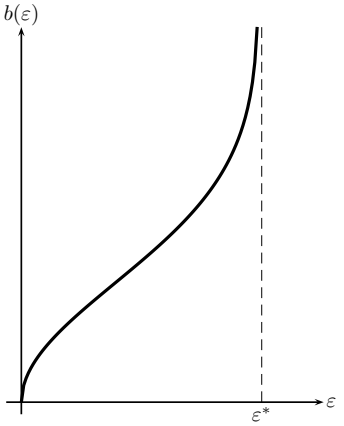


Figura 2.22 *Representação gráfica (qualitativa) de* $b(\varepsilon)$*, definido em (2.45), em função de* ε *para um estimador e um modelo hipotéticos tais que* $\gamma^* = \infty$ *e* $\varepsilon^* > 0$*.*

ponto de rotura, apesar de conceptualmente simples, é delicado, quer sob o ponto de vista puramente matemático, quer sob os pontos de vista probabilístico e estatístico. Para a maior parte dos métodos não é de todo simples a sua determinação. Noutros casos há a necessidade de adaptar a definição à situação concreta. Estes aspectos fazem com que este tópico registe bastante actividade em termos de investigação (veja-se, por exemplo, Zuo, 2004, ou Davies e Gather, 2007).

2.6 Síntese das propriedades mais relevantes

Para concluir este capítulo apresenta-se a lista das principais propriedades que um bom método de estimação deve possuir pela ordem em que, segundo as tendências actuais, devem ser consideradas:

Ponto de rotura de um estimador é a percentagem de contaminação (dos dados ou do modelo) a partir da qual se obtêm estimativas completamente arbitrárias. Deve ser **o mais elevado possível** tendo em consideração que a partir de cerca de 1/4 do seu valor as estimativas, apesar de não serem arbitrárias, deixam de ser fiáveis.

Eficiência (assintótica absoluta) em relação ao modelo central que se considera é o quociente entre a variância assintótica mínima sob esse modelo, geralmente igual ao inverso da informação de Fisher, e a variância assintótica do estimador em causa. Deve ser **o mais elevada possível**, sujeita à condição anterior (ponto de rotura elevado).

Sensibilidade a grandes erros mede aproximadamente o pior efeito que uma pequena contaminação pode ter nas estimativas, medido em termos de taxa de variação do enviesamento. É dada pelo supremo do valor absoluto da função de influência e daí a designação de estimadores de influência limitada para os estimadores que têm esta propriedade. Deve ser **o mais baixa possível**, sujeita às condições anteriores.

Enviesamento assintótico máximo indica a máxima alteração produzida nas estimativas em função da percentagem de contaminação. Deve ser compatibilizado com as propriedades anteriores tendo em conta que é desejável que seja **o mais reduzido possível**.

Sensibilidade local mede aproximadamente o pior efeito estandardizado provocado por alterações locais dos dados e está relacionada com a instabilidade das estimativas. Deve ser **o mais reduzida possível**, sujeita às condições anteriores.

Ponto de rejeição indica a distância (ao centro dos dados) a partir da qual as observações são completamente rejeitadas, não contribuindo de todo para as estimativas. Deve ser **finito**, sujeito às condições anteriores.

A harmonização e conciliação de todas estas propriedades não é, como claramente se adivinha, tarefa fácil e tem constituído o cerne do trabalho de investigação em estatística robusta. Esse trabalho tem conduzido à criação de novas classes de estimadores onde se procura então o estimador de compromisso que satisfaz tantos e tão contraditórios requisitos. Dessa busca se dá conta nos dois capítulos seguintes, o Capítulo 3 dedicado essencialmente à estimação univariada e o Capítulo 4 relativo ao modelo de regressão.

3
Estimação

3.1 Introdução

Este capítulo é dedicado principalmente à estimação pontual robusta num contexto univariado e em grande parte à introdução e estudo detalhado de uma classe de estimadores desenvolvida no âmbito da estatística robusta, os estimadores-M.

Os estimadores-M foram propostos por Huber (1964) como uma generalização dos estimadores de máxima verosimilhança. O artigo de Huber (1964) constituiu um marco fundamental no desenvolvimento da estatística robusta e os estimadores-M são e serão sempre um assunto incontornável em qualquer curso de robustez. Apesar de inicialmente propostos para os problemas univariados de localização e dispersão ou escala, eles acabam por servir de base a procedimentos muito mais complexos.

Como exemplificação importante das propriedades destes estimadores é considerada precisamente a sua aplicação nos modelos univariados de localização e escala, sendo importante referir que de entre os estimadores apresentados até aqui para a estimação de localização e de dispersão, não se encontrou nenhum que reunisse em si todas as propriedades relevantes enunciadas no final do Capítulo 2. A classe dos estimadores-M, pela sua riqueza, vem aumentar as possibilidades de escolha.

No final do capítulo apresenta-se uma descrição sucinta de outras classes de estimadores e são feitas considerações relativas a intervalos de confiança e a testes de hipóteses.

3.2 Os estimadores-M

3.2.1 Definição geral

Como se disse os estimadores-M não são mais do que uma generalização dos estimadores de máxima verosimilhança, e daí a sua designação.

Recorde-se então, e para fixar notação, a definição de estimador de máxima verosimilhança (EMV). Para tanto onsidere-se um determinado modelo paramétrico, com densidade $f(x,\theta)$. Por enquanto supõe-se que tanto Ω como Θ são unidimensionais. O estimador de máxima verosimilhança de θ é o estimador $T_n = T_n(X_1,\ldots,X_n)$ que torna máximo $\prod_{i=1}^n f(X_i,T_n)$ ou, duma forma equivalente, T_n é tal que

$$T_n \text{ maximiza} \sum_{i=1}^n \log f(X_i,T_n) \;\Leftrightarrow\; T_n \text{ minimiza} \sum_{i=1}^n \left(-\log f(X_i,T_n)\right). \tag{3.1}$$

A determinação de T_n é feita normalmente, nos casos regulares de estimação, recorrendo à condição de primeira ordem

$$\sum_{i=1}^n \psi(X_i,T_n) = 0, \tag{3.2}$$

com

$$\psi(x,\theta) = -\frac{\partial \log f(x,\theta)}{\partial\theta}, \tag{3.3}$$

e verificando se a solução obtida é um máximo (mínimo) global de (3.1).

Definição 3.1. *Um estimador-M é um estimador T_n tal que*

$$T_n \textit{ minimiza } \sum_{i=1}^n \rho(X_i,T_n), \tag{3.4}$$

onde ρ é uma função arbitrária (chamada função objectivo) definida em $\Omega\times\Theta$. Se existir $\psi(x,\theta) = (\partial/\partial\theta)\rho(x,\theta)$, então T_n satisfaz a equação implícita (3.2).

Assim todo o EMV é um estimador-M, mas a inversa não é verdadeira, existem estimadores-M que não são estimadores de máxima verosimilhança para nenhum modelo, como se verá mais adiante.

Embora as condições (3.2) e (3.4) só sejam equivalentes nos casos regulares de estimação, é a condição (3.2) que é geralmente utilizada para definir estimador-M. Note-se também que o estimador não é alterado se a função ψ for multiplicada por uma constante $r \neq 0$. A menos dessa constante pode identificar-se o estimador-M com a função ψ que o define.

O funcional T equivalente ao estimador-M, T_n, é definido também implicitamente por

$$\int \psi(x, T(F))dF(x) = 0, \tag{3.5}$$

uma vez que $T_n(x_1, \ldots, x_n) = T(G_n)$, $\forall_{n,G_n}$. O domínio de T é o conjunto de todas as distribuições para as quais aquele integral existe. Recorde-se que o estimador é consistente segundo Fisher se $T(F_\theta) = \theta$ para todo o θ, ou seja, se se verificar

$$\int \psi(x, \theta)dF_\theta(x) = 0, \quad \forall_\theta.$$

Para obter a função de influência procede-se como habitualmente, substituindo em (3.5) F por $F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$ e derivando em ordem a ε. Ou seja,

$$\frac{\partial}{\partial\varepsilon}\left[\int \psi\left(u, T(F_{\varepsilon,x})\right) dF_{\varepsilon,x}(u)\right] = 0$$

que é equivalente a

$$\frac{\partial}{\partial\varepsilon}\left[(1-\varepsilon)\int \psi\left(u, T(F_{\varepsilon,x})\right) dF(u) + \varepsilon\psi\left(x, T(F_{\varepsilon,x})\right)\right] = 0$$

por sua vez, admitindo que é possível trocar as ordens de derivação

e integração, equivalente a

$$-\int \psi\left(u, T(F_{\varepsilon,x})\right) dF(u)+ \\ +(1-\varepsilon)\int \frac{\partial T(F_{\varepsilon,x})}{\partial \varepsilon}\left.\frac{\partial \psi\left(u,\theta\right))}{\partial \theta}\right|_{\theta=T(F_{\varepsilon,x})} dF(u)+ \\ +\psi\left(x, T(F_{\varepsilon,x})\right)+\varepsilon\left.\frac{\partial T(F_{\varepsilon,x})}{\partial \varepsilon}\,\frac{\partial \psi\left(u,\theta\right)}{\partial \theta}\right|_{\theta=T(F_{\varepsilon,x})}=0.$$

Fazendo então $\varepsilon = 0$ e notando que nesse caso o integral da primeira linha da expressão anterior é nulo (por definição) e que

$$\left.\frac{\partial T(F_{\varepsilon,x})}{\partial \varepsilon}\right|_{\varepsilon=0} = IF(x;T,F),$$

também por definição, obtém-se

$$IF(x;T,F)\int \left.\frac{\partial \psi\left(u,\theta\right)}{\partial \theta}\right|_{\theta=T(F)} dF(u)+\psi\left(x,T(F)\right)=0,$$

e, finalmente, admitindo que o denominador é diferente de zero,

$$IF(x;T,F)=\frac{\psi(x,T(F))}{-\int\left[(\partial/\partial\theta)\psi(u,\theta)\right]_{\theta=T(F)}dF(u)}. \tag{3.6}$$

Se ψ for a função correspondente ao EMV obtém-se a função de influência dada pela expressão (2.21) da Secção 2.3.2.

Está agora também justificada a afirmação feita anteriormente de que é simples criar um estimador cuja função de influência tenha uma forma pré-definida, uma vez que, por (3.6), IF é proporcional a ψ.

Por outro lado, dado um estimador suficientemente regular, T, existe um estimador-M que lhe é assintoticamente equivalente sob o modelo considerado. Basta que se tome $\psi(x,\cdot) = IF(x;\cdot,F)$. Os dois estimadores terão as mesmas propriedades locais (função de influência, sensibilidade, variância assintótica) mas poderão diferir em relação a propriedades globais (tais como robustez qualitativa e ponto de rotura).

As ideias fundamentais relacionadas com estimadores-M podem ser mais facilmente compreendidas no contexto de dois modelos simples

que são o modelo de localização e o modelo de escala (ou dispersão). O interesse destes modelos e dos estimadores correspondentes não se resume ao caso unidimensional pois eles próprios, ou generalizações suas, desempenham um papel importante nos problemas de estimação em análise de regressão tratados no Capítulo 4.

3.2.2 Modelo de localização

Antes de apresentar a definição matemática de modelo de localização convém discutir um pouco melhor o conceito, o qual está de certa forma ligado ao conceito de medida de localização usado de forma informal e intuitiva até agora. Pensando sobre o conceito de medida de localização percebe-se que a propriedade básica que está em jogo se relaciona com translações dos dados, do seguinte modo: somando uma certa quantidade a todas as observações de um conjunto de dados univariados, a medida de localização deve sofrer o mesmo efeito. Ou seja, sendo t uma medida de localização genérica ela deve verificar

$$t(x_1 + a, \ldots, x_n + a) = t(x_1, \ldots, x_n) + a,$$

para qualquer $a \in \mathbb{R}$ e quaisquer que sejam as observações $(x_1, \ldots, x_n)$. A esta propriedade chama-se equivariância em relação a translações ou ainda equivariância em relação à localização.[1] A mesma ideia pode transportar-se para o campo teórico das distribuições de probabilidade, dizendo que: dado um certo modelo com um parâmetro θ, F_θ, o parâmetro θ (pode haver outros parâmetros no modelo) é um parâmetro de localização e o modelo é um modelo de localização, se dada uma variável X com distribuição desse modelo e parâmetro $\theta = b$, a variável que se obtém fazendo uma translação igual a a, $X + a$, ainda tem, para qualquer a, uma distribuição da mesma família mas com parâmetro $\theta = b + a$.

Por exemplo, conclui-se facilmente que o modelo normal com o parâmetro μ é um modelo de localização, pois se $X \sim \mathcal{N}(\mu, \sigma^2)$, então $X + a \sim \mathcal{N}(\mu + a, \sigma^2)$. Podia-se pensar, erradamente, que um modelo de localização é um modelo em que há um parâmetro que coincide com o valor esperado. Considere-se então, por exemplo o modelo de Poisson, para o qual o valor esperado coincide com o parâmetro λ.

[1] É simples verificar que todas as medidas de localização consideradas anteriormente (média aritmética, mediana e médias aparadas) verificam esta propriedade.

Este modelo seria um modelo de localização se dada $X \sim Poisson(\lambda)$ se verificasse que $X + a \sim Poisson(\lambda + a)$, para qualquer a, o que obviamente não acontece. Ainda outro contra-exemplo é dado pela distribuição exponencial, $X \sim Exp(\beta)$, com densidade

$$f(x) = \begin{cases} \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right), & x > 0 \\ 0, & x \leq 0 \end{cases} . \tag{3.7}$$

O valor esperado de X é β mas $X + a$ não tem distribuição $Exp(\beta + a)$, logo β não é um parâmetro de localização. Poder-se-ia então concluir que os modelos de localização disponíveis são em número muito reduzido. Nada é mais falso, pois dada uma variável aleatória com uma qualquer distribuição pode criar-se a partir dela um modelo de localização, usando como parâmetro o próprio valor da translação, a. Pense-se por exemplo na variável aleatória $X \sim Exp(\beta)$ e na transformação $X_a = X + a$. Da igualdade

$$F_{X_a}(x) = P(X + a \leq x) = P(X \leq x - a) = F_X(x - a)$$

obtém-se o modelo pretendido, o qual tem densidade

$$f_a(x) = f(x - a) = \begin{cases} \frac{1}{\beta} \exp\left(-\frac{x - a}{\beta}\right), & x > a \\ 0, & x \leq a \end{cases} . \tag{3.8}$$

e costuma ser referido como distribuição exponencial deslocada. A definição seguinte formaliza o que se acabou de exemplificar.

Definição 3.2. *Um modelo de localização, com parâmetro de localização θ, consiste numa família de distribuições $\{F_\theta, \theta \in \mathbb{R}\}$ tal que*

$$F_\theta(x) = F(x - \theta),$$

onde $F = F_0$ representa uma função de distribuição univariada genérica que define o tipo ou família do modelo.

É fácil verificar que, tal como no exemplo antes da definição, se $X \sim F$, ou seja com $\theta = 0$, então $X + \theta \sim F_\theta$, ou ainda se $Y \sim F_\theta$ então $Y - \theta \sim F$ e, para qualquer $a \in \mathbb{R}$, $Y + a \sim F_{\theta+a}$.

Admite-se que F possui uma densidade $f = F'$, que pode ser uma densidade num sentido generalizado (por exemplo uma função de probabilidade). $f(x)$ representa então a densidade quando $\theta = 0$, enquanto que $f_\theta(x) = f(x - \theta)$ representa a densidade para um valor genérico do parâmetro θ. Muitas vezes também se admite que f é simétrica em relação a zero, nesses casos f_θ será simétrica em relação a θ.

Dada uma amostra aleatória $(X_1, \ldots, X_n)$ o estimador de máxima verosimilhança, $T_n = T_n(X_1, \ldots, X_n)$, do parâmetro θ do modelo de localização minimiza

$$\sum_{i=1}^{n} \rho(X_i - T_n), \tag{3.9}$$

com $\rho(u) = -\log f(u)$, ou, de forma equivalente, nos casos regulares de estimação, é solução de

$$\sum_{i=1}^{n} \psi(X_i - T_n) = 0, \tag{3.10}$$

com $\psi(u) = -f'(u)/f(u)$.

Para definir estimador-M no contexto do modelo de localização usa-se a expressão (3.10) em que ψ é uma função genérica de apenas uma variável.[2] Para que o funcional equivalente, T, definido implicitamente por

$$\int \psi(x - T(F))\, dF(x) = 0,$$

seja consistente segundo Fisher deve verificar-se

$$\int \psi(x - \theta) dF(x - \theta) = \int \psi(u) dF(u) = 0.$$

Nos casos em que F corresponde a uma distribuição simétrica (em relação à origem), esta condição verifica-se automaticamente se ψ for uma função ímpar. Note-se que esta condição é também verificada pela função correspondente ao estimador de máxima verosimilhança nesses casos, pois se f é par f' e f'/f são ímpares.

Uma propriedade desejável para os estimadores do parâmetro de localização de um modelo de localização é a equivariância em relação

[2] Ou seja, passa-se da definição geral para a definição particular adequada a este modelo fazendo, por analogia com o estimador da máxima verosimilhança, $\psi(x, \theta) = \psi(x - \theta)$.

à localização apresentada na definição seguinte.

Definição 3.3. *Um estimador T_n do parâmetro de localização de um modelo de localização é equivariante em relação à localização (location equivariant) se dada a amostra aleatória $(X_1, \ldots, X_n)$,*

$$T_n(X_1 + k, \ldots, X_n + k) = T_n(X_1, \ldots, X_n) + k \quad \forall_{k \in \mathbb{R}}.$$

É imediato verificar que os estimadores-M definidos por (3.10) possuem esta propriedade (e esta pode ser outra justificação para utilizar $\psi(x, \theta) = \psi(x - \theta)$).

Não é muito difícil concluir que para estimadores (funcionais) equivariantes se verifica, sob as distribuições do modelo,

$$IF(x; T, F_\theta) = IF(x - \theta; T, F) = IF(u; T, F),$$

com $u = x - \theta$. Isto significa que todas as propriedades do estimador que recorrem à função de influência podem ser avaliadas na distribuição central F.

Partindo da expressão da função de influência do estimador-M geral, (3.6), têm-se então os seguintes resultados para um estimador-M, equivalente a um funcional T, do parâmetro de localização num modelo de localização (na distribuição central F):

- $$IF(x; T, F) = \frac{\psi(x)}{\int \psi'(u) dF(u)} = \frac{\psi(x)}{M} \tag{3.11}$$
  e todas as medidas (quer as relacionadas com a eficiência quer as relacionadas com a robustez) obtidas a partir da função de influência são determinadas pela forma da função ψ.

- A variância assintótica é dada por
  $$V(T, F) = \frac{\int \psi^2(u) dF(u)}{M^2}. \tag{3.12}$$
- O estimador é B-robusto sse ψ for limitada sendo a sensibilidade a grandes erros dada por
  $$\gamma^*(T, F) = \frac{\sup_x |\psi(x)|}{M}.$$

- A sensibilidade local é finita sse ψ for contínua e ψ' limitada e nesse caso pode ser calculada por

$$\lambda^*(T,F) = \frac{\sup_x |\psi'(x)|}{M}.$$

- O ponto de rejeição é finito sse ψ for identicamente nula a partir de determinado ponto, tendo-se então

$$\rho^* = \inf\{r > 0 : \psi(x) = 0, \forall_{|x|>r}\}.$$

No caso de ψ ser uma função não decrescente é ainda possível garantir que (Hampel *et al.*, 1986, p. 104):

- se ψ é limitada então o estimador é qualitativamente robusto e tem ponto de rotura $\varepsilon^*(T,F) = 1/2$;

- se ψ é ilimitada então o estimador não é qualitativamente robusto e o seu ponto de rotura é $\varepsilon^*(T,F) = 0$.

Nos dois exemplos seguintes ilustram-se estes conceitos com os dois estimadores de localização mais comuns, a média aritmética e a mediana que, como já se viu, são estimadores de máxima verosimilhança de modelos de localização particulares e por conseguinte estimadores-M.

Exemplo 3.1. No modelo de localização normal assume-se que X_i tem distribuição $\mathcal{N}(\theta, \sigma^2)$, com σ^2 conhecido. Sem perda de generalidade pode estudar-se o caso particular $\sigma^2 = 1$, ou seja, $X_i \sim \mathcal{N}(\theta, 1)$. Tem-se então $F_\theta(x) = F(x-\theta)$ com $F(u) = \Phi(u)$ e

$$f(u) = \Phi'(u) = \varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

Para este modelo o estimador de máxima verosimilhança minimiza (3.9) com

$$\rho(u) = -\log f(u) = \frac{u^2}{2} + \sqrt{2\pi},$$

ou, de forma equivalente, com

$$\rho(u) = \frac{u^2}{2},$$

o que é ainda equivalente a resolver (3.10) com $\psi(u) = u$. Logo

$$\sum_{i=1}^{n} \psi(X_i - \hat{\theta}_{MV}) = 0 \;\Leftrightarrow\; \sum_{i=1}^{n} (X_i - \hat{\theta}_{MV}) = 0 \;\Leftrightarrow\; \hat{\theta}_{MV} = \overline{X}.$$

O raciocínio anterior, baseado em equivalências, permite responder à questão em sentido contrário: qual é o modelo de localização (regular) para o qual o estimador de máxima verosimilhança é a média aritmética? E a resposta é: unicamente o modelo normal.

É também interessante notar que a minimização de (3.9) com $\rho(u) \propto u^2$ corresponde à utilização do critério dos mínimos quadrados para a estimação do parâmetro θ do modelo de localização e conduz sempre ao estimador $\hat{\theta}_{MQ} = \overline{X}$. Ora o que se acabou de demonstrar é que o único modelo de localização para o qual este estimador tem propriedades de eficiência assintótica máxima é precisamente o modelo normal pois é o único onde coincide com o estimador de máxima verosimilhança.

Vem a propósito recordar a afirmação feita a seguir à Definição 3.1 a respeito da existência de estimadores-M que não são estimadores de máxima verosimilhança para nenhum modelo. Para o modelo de localização tem-se que dada uma função ψ, essa função corresponde a um modelo com densidade f tal que

$$-(\log f(u))' = \psi(u) \quad \Leftrightarrow \quad \log f(u) = -P\left(\psi(u)\right),$$

ou seja,

$$f(u) \propto \exp\left[-P\left(\psi(u)\right)\right] \quad \text{ou} \quad f(u) \propto \exp[-\rho(u)]. \tag{3.13}$$

Se, por exemplo, ψ tiver ponto de rejeição finito, isto é se existir um ponto r tal que $\psi(u)$ é identicamente nula para $|u| > r$, conclui-se que $\rho(u)$ toma um valor constante para $|u| > r$ e que f não pode ser uma densidade própria. Isto significa que uma condição necessária, embora não suficiente para que o estimador-M seja equivalente a um estimador de máxima verosimilhança é que $\rho(u)$ seja uma função estritamente crescente para $u > 0$ e estritamente decrescente para $u < 0$.

Exemplo 3.2. Considere-se agora o modelo de localização de *Laplace*, $X_i \sim Laplace(\theta, 1)$, com

$$f(u) = \frac{1}{2} \exp\left(-|u|\right).$$

A menos de uma constante tem-se $\rho(u) = |u|$ e

$$\psi(u) = \text{sinal}(u) = \begin{cases} 1, & u > 0 \\ 0, & u = 0 \\ -1, & u < 0 \end{cases},$$

onde o valor $\psi(0) = 0$ é imposto apenas para que ψ seja uma função ímpar, uma vez que $\rho'(0)$ não existe. Então o EMV de θ obtém-se de

$$\sum_{i=1} \psi(X_i - \hat{\theta}_{MV}) = 0 \Leftrightarrow \sum_{i=1}^{n} \text{sinal}(X_i - \hat{\theta}_{MV}) = 0.$$

Como a segunda expressão só se anula se houver tantas observações superiores a $\hat{\theta}_{MV}$ como inferiores, obtém-se então, se n for ímpar, $\hat{\theta}_{MV} = \text{med}(X_i)$. Se n for par todos os pontos do intervalo cujos extremos são as duas estatísticas de ordem centrais são solução, podendo escolher-se como solução o ponto médio desse intervalo, ou seja novamente $\hat{\theta}_{MV} = \text{med}(X_i)$.

À semelhança do exemplo anterior, é interessante notar que a minimização de (3.9) com $\rho(u) \propto |u|$ corresponde à utilização do critério dos mínimos desvios absolutos (também conhecido por L_1 ou LAD, de *Least Absolute Deviations*) para a estimação do parâmetro θ do modelo de localização e conduz sempre ao estimador $\hat{\theta}_{L_1} = \text{med}(X_i)$. Ora o que se acabou de constatar é que este critério conduz a um estimador robusto, ao contrário do critério dos mínimos quadrados, no entanto este estimador sofre como já se viu de outros defeitos, nomeadamente falta de eficiência sob outros modelos e instabilidade local. O que se acabou de explicar permite além disso concluir que o modelo de *Laplace* é precisamente o único modelo de localização para o qual este estimador tem propriedades de eficiência assintótica máxima.

Uma vez admitido um determinado modelo de localização para as observações, o qual como já várias vezes se referiu, se supõe verificado apenas aproximadamente, é fácil ter uma ideia clara da forma da função ψ que conduz a um estimador-M com boas propriedades. Para isso deve atentar-se nas propriedades acabadas de descrever e ter em conta o que foi dito no final do capítulo anterior sobre propriedades desejáveis para um estimador.

Admitindo então o modelo de localização normal como modelo central ou modelo aproximado, conclui-se que a função ψ deve ser:

- limitada, garantindo assim ponto de rotura máximo (igual a 1/2);

- linear próximo da origem de modo a ser próxima da função ψ do EMV e conseguir assim uma boa eficiência assintótica (ou seja uma variância assintótica próxima da da média aritmética) exactamente sob o modelo;

- tal que $\sup_x |\psi(x)|$ não seja muito elevado para garantir uma baixa sensibilidade a grandes erros e simultaneamente uma curva de enviesamento assintótico máximo bem comportada;

- contínua e eventualmente nula para valores afastados da origem, o que garante, respectivamente, baixa sensibilidade local e ponto de rejeição finito.

Apresentam-se e discutem-se em seguida diversas funções ψ que têm sido propostas na literatura e que têm por base o modelo de localização normal, ou seja, possuem as características acabadas de referir. Os nomes dados às funções são também geralmente associados aos estimadores-M de localização a que elas dão origem.

(i) Função ψ de Huber (estimador-M de Huber)

$$\psi_b(x) = [x]_{-b}^{b} = \begin{cases} -b, & x < -b \\ x, & |x| \leq b \\ b, & x > b \end{cases} .$$

(ii) Função ψ biponderada[3] de Tukey (estimador biponderado de Tukey ou estimador-M de Tukey)

$$\psi(x) = \begin{cases} x(r^2 - x^2)^2, & |x| \leq r \\ 0, & |x| > r \end{cases} .$$

[3]Tradução de *biweight*.

(iii) Função ψ de Hampel (estimador-M de Hampel)

$$\psi(x) = \begin{cases} -\psi(-x), & x < 0 \\ x, & 0 \leq x < a \\ a, & a \leq x < b \\ \dfrac{c-x}{c-b}a, & b \leq x < c \\ 0, & x \geq c \end{cases} .$$

(iv) Função ψ de Andrews (estimador-M de Andrews)

$$\psi(x) = \begin{cases} \text{sen}(x/a), & |x| \leq a\pi \\ 0, & |x| > a\pi \end{cases} .$$

Os gráficos respectivos apresentam-se na Figura 3.1. Note-se que em todos os casos fazendo variar convenientemente as constantes se pode obter como limite (pontual) a função ψ correspondente à média aritmética. Além disso, o estimador de Huber aproxima-se da mediana quando $b \to 0$, acontecendo o mesmo com o estimador de Hampel quando $a \to 0$ e $c \to +\infty$. As funções ρ correspondentes a cada uma daquelas funções ψ são as indicadas a seguir e os respectivos gráficos encontram-se na Figura 3.2.

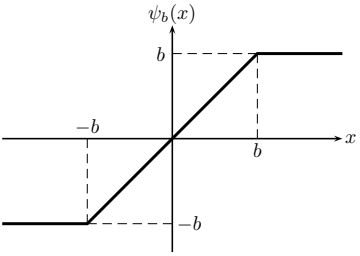
(i) Função ρ de Huber

$$\rho_b(x) = \begin{cases} \dfrac{x^2}{2}, & |x| < b \\ b|x| - \dfrac{b^2}{2}, & |x| > b \end{cases} . \tag{3.14}$$

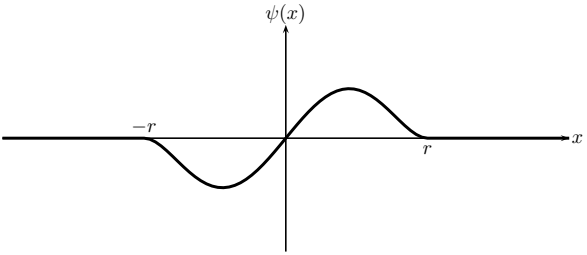
(ii) Função ρ biponderada de Tukey

$$\rho(x) = \begin{cases} \dfrac{x^6}{6} - \dfrac{r^2x^4}{2} + \dfrac{r^4x^2}{2}, & |x| \leq r \\ \dfrac{r^6}{6}, & |x| > r \end{cases} .$$

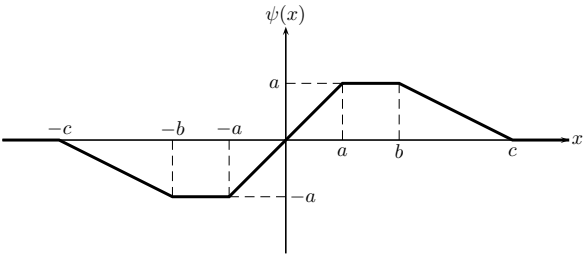
Huber:



Tukey:



Hampel:



Andrews:

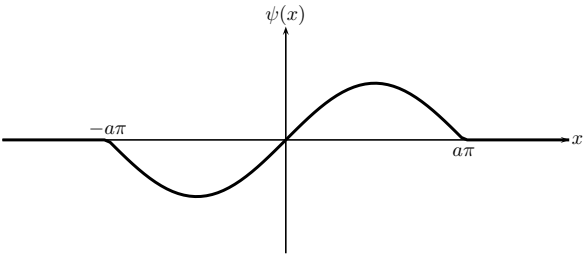
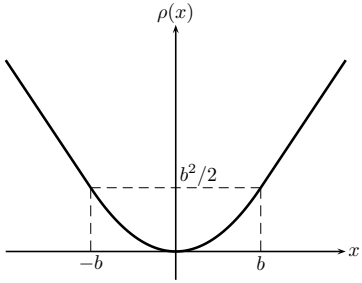
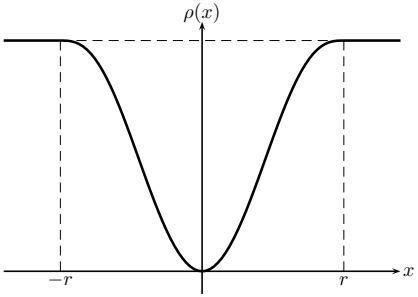


Figura 3.1 *Gráficos de várias funções ψ associadas a estimadores-M do parâmetro de localização (modelo de localização normal).*

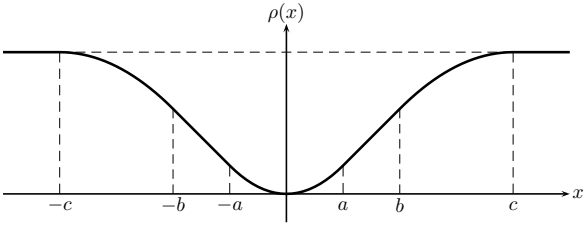
Huber:



Tukey:



Hampel:



Andrews:

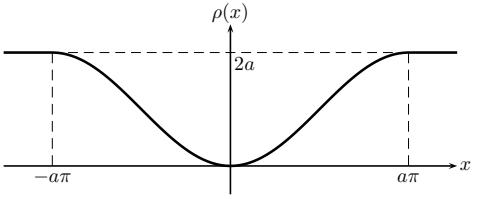


Figura 3.2 *Gráficos de várias funções ρ associadas a estimadores-M do parâmetro de localização (modelo de localização normal).*

(iii) Função ρ de Hampel

$$\rho(x) = \begin{cases} \rho(-x), & x < 0 \\ \dfrac{x^2}{2}, & 0 \leq x < a \\ ax - \dfrac{a^2}{2}, & a \leq x < b \\ ab - \dfrac{a^2}{2} + \dfrac{ac(x-b)}{c-b} - \dfrac{x^2 - b^2}{2(c-b)}, & b \leq x < c \\ ab + ac - \dfrac{a^2}{2} - \dfrac{c}{2} - \dfrac{b}{2}, & x \geq c \end{cases} .$$

(iv) Função ρ de Andrews

$$\rho(x) = \begin{cases} a - a\cos(x/a), & |x| \leq a\pi \\ 2a, & |x| > a\pi \end{cases} .$$

De seguida analisam-se em mais detalhe estas funções. É de notar que, à excepção da função ψ de Huber, todas têm ponto de rejeição finito. Isto significa, de acordo com a observação feita a seguir à expressão (3.13), que o estimador de Huber é o único que corresponde ao estimador de máxima verosimilhança para um certo modelo. Utilizando aquela expressão é fácil concluir, com $\rho(x) = \rho_b(x)$ dada por (3.14), que a densidade desse modelo corresponde a

$$f(x) = K\exp(-\rho(x)) = \begin{cases} K\exp\left(-\frac{x^2}{2}\right), & |x| \leq b \\ K\exp\left(-b|x| + \frac{b^2}{2}\right), & |x| > b \end{cases} . \quad (3.15)$$

A constante K é tal que $\int_{-\infty}^{+\infty} f(x)dx = 1$ e pode ser escrita como

$$K = \frac{1-\varepsilon}{\sqrt{2\pi}},$$

com ε dependente de b através de

$$\frac{\varepsilon}{1-\varepsilon} = \frac{2\varphi(b)}{b} - 2\Phi(-b). \quad (3.16)$$

A escrita da constante nesta forma possibilita interpretar a densidade f como a mistura, nas proporções de $(1-\varepsilon)$ para ε, de uma distribuição normal com uma distribuição contaminante de tipo exponencial nas caudas (isto é, para $|x| > b$).

Historicamente a função de Huber foi a primeira a aparecer, tendo sido proposta logo em Huber (1964) juntamente com a definição de estimador-M. As considerações que conduziram a essa proposta não foram baseadas na função de influência mas na chamada abordagem minimax de Huber. Nessa abordagem Huber considerou a vizinhança de contaminação da distribuição normal, $\mathcal{P}_\varepsilon(\Phi)$, com ε fixo, e propôs-se determinar o estimador, T, tal que

$$\sup_{F \in \mathcal{P}_\varepsilon(\Phi)} V(T, F) \text{ é mínima.}$$

Ou seja, procura-se a distribuição mais desfavorável em $\mathcal{P}_\varepsilon(\Phi)$ e a solução será dada pelo estimador de máxima verosimilhança para essa distribuição. O que Huber mostrou foi que essa distribuição mais desfavorável (também chamada "menos informativa" porque minimiza a informação de Fisher em $\mathcal{P}_\varepsilon(\Phi)$) corresponde precisamente à distribuição que tem a densidade f dada por (3.15).

As restantes funções ψ podem ser vistas como modificações de ψ_b de modo a conduzirem a estimadores com ponto de rejeição finito. Na função ψ de Hampel isso é feito através de uma transição linear (entre os pontos b e c). As outras duas propostas surgiram com o intuito de "suavizar" as funções ψ, no sentido de as tornar diferenciáveis, esperando-se dessa forma melhorar as propriedades de regularidade dos estimadores.

É importante salientar que o estimador de Huber é assintoticamente equivalente, em termos locais, a uma média aparada a $100 \times \alpha\%$, dado que as suas funções de influência coincidem. Comparando a função ψ_b com a expressão da função de influência da média aparada a $100 \times \alpha\%$ apresentada no Exemplo 2.8 pode concluir-se que isso acontece precisamente quando α e b verificam $b = \Phi^{-1}(1-\alpha)$. No entanto, é preciso realçar que, apesar desta equivalência assintótica local (isto é, numa vizinhança muito próxima do modelo central) os dois tipos de estimadores diferem em termos de ponto de rotura, que é uma propriedade global: enquanto o ponto de rotura do estimador de Huber é 50% o da média aparada é $100 \times \alpha\%$. Esta é uma razão extremamente importante para se preferir o estimador de Huber sobre

a média aparada que lhe é assintoticamente equivalente.

Na prática, para utilizar estes estimadores, é ainda necessário, para além da escolha do tipo de função ψ, fixar o(s) valor(es) da(s) constante(s) que entra(m) na definição (essas constantes são chamadas vulgarmente constantes de afinação). A grande diversidade, quer de funções, quer de constantes, é uma das razões que pode ser apontada como responsável pela fraca utilização destes métodos.

A escolha da(s) constante(s) a utilizar em cada função ψ depende por um lado da eficiência que se pretende obter sob o modelo e por outro lado da protecção a garantir. Uma vez que não é possível alcançar, simultaneamente, protecção e eficiência máxima é necessário procurar uma solução de compromisso.

Para o caso do estimador de Huber, dada a relação existente entre as constantes b e ε, (3.16), e se houver um conhecimento *a priori* acerca da percentagem de contaminação deve utilizar-se o valor de b correspondente. Para $\varepsilon = 0.01$ e $\varepsilon = 0.1$ obtêm-se, respectivamente, $b = 1.945$ e $b = 1.140$. Um valor vulgarmente utilizado é $b = 1.645$, correspondente a $\varepsilon = 0.025$ e ao percentil 95% de Φ (deste modo, tal como na média aparada a 10%, só 10% das observações duma amostra "normal" é que têm influência inferior à que teriam na estimativa de máxima verosimilhança).

Para o estimador de Hampel a constante a pode ser escolhida pelo mesmo critério. Para b é usual adoptar um percentil de Φ próximo de 100% e c deve ser tal que $c - b \geq 2a$.

No caso do estimador de Andrews têm sido propostos, com base em resultados experimentais, valores de $a \in [1.0; 1.6]$ e para o estimador de Tukey valores de $r \in [3.5; 6]$.

Um método menos ambíguo consiste em determinar a ou as constantes de afinação que garantem uma eficiência elevada (maior ou igual a 90%) sob o modelo central. Este processo pode ser designado por calibração das constantes de afinação ou do estimador. Como a variância da média sob o modelo central é 1, a eficiência assintótica do estimador-M em causa vai ser simplesmente dada pelo inverso de $V(T, \Phi)$ que por sua vez pode ser calculada por (3.12).[4] Em alternativa, no caso do estimador de Huber cuja variância assintótica coin-

[4] A função `chb` do S-Plus faz estes cálculos para os estimadores de Huber e Tukey.

Tabela 3.1 *Valores das constantes de afinação que garantem para os diversos estimadores eficiências assintóticas de 90%, 95% e 97.5% sob o modelo central normal e respectivas sensibilidades (γ^*).*

	Huber			Tukey	
Eficiência	b	α	γ^*	r	γ^*
90%	0.982	0.163	1.457	3.883	1.664
95%	1.345	0.089	1.637	4.685	1.770
97.5%	1.655	0.049	1.835	5.596	1.944

	Hampel				Andrews	
Eficiência	a	b	c	γ^*	a	γ^*
90%	1.10	$2a$	$4a$	1.539	1.11	1.664
95%	1.38	$2a$	$4a$	1.664	1.34	1.770
97.5%	1.67	$2a$	$4a$	1.846	1.60	1.945

cide com a da média aparada equivalente devido à coincidência das funções de influência, pode usar-se o gráfico da Figura 2.9 (pág. 57): fixada V, lê-se α e faz-se $b = \Phi^{-1}(1-\alpha)$. Na Tabela 3.1 apresentam-se valores das constantes que garantem valores elevados de eficiência assintótica para os diversos estimadores, bem como os valores da sensibilidade a grandes erros (γ^*) correspondentes. Recorde-se que todos estes estimadores têm ponto de rotura igual a 1/2.

Uma outra questão de natureza operacional, que pode na prática dificultar a utilização destes estimadores, tem a ver com a mudança de escala das observações. Apesar de para o estudo teórico se poder assumir que a escala é conhecida, como foi feito nos Exemplos 3.1 e 3.2, em termos práticos é desejável que os estimadores de localização sejam equivariantes em relação à escala, de acordo com a definição seguinte.

Definição 3.4. *Um estimador T_n do parâmetro de localização de um modelo de localização é equivariante em relação à escala (scale equivariant) se dada a amostra aleatória $(X_1, \ldots, X_n)$,*

$$T_n(k\,X_1, \ldots, k\,X_n) = k\,T_n(X_1, \ldots, X_n) \quad \forall_{k\neq 0}.$$

É imediato verificar que os estimadores-M dos Exemplos 3.1 e 3.2, respectivamente média aritmética e mediana, cuja solução é explícita, possuem esta propriedade. No entanto, para os estimadores-M com uma função ψ genérica ela não se verifica automaticamente. Por exemplo, no caso do estimador de Huber, fazendo k suficientemente grande obtém-se como solução a mediana e não k vezes a estimativa original. No caso dos outros estimadores, também para k suficientemente grande a estimativa passa a ser indeterminada.

Para tornar estes estimadores equivariantes em relação à escala considera-se simultaneamente um estimador auxiliar de escala, S_n, que deve ele próprio ser equivariante em relação a mudanças de escala, isto é tal que

$$S_n(k\,X_1, \ldots, k\,X_n) = |k|\,S_n(X_1, \ldots, X_n) \quad \forall_{k\neq 0},$$

e define-se T_n como solução de

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - T_n}{S_n}\right) = 0. \tag{3.17}$$

É simples verificar que se T_n é solução da equação anterior então $k\,T_n$, com $k > 0$, é solução da equação que se obtém substituindo X_i por $k\,X_i$ (se se permitir $k < 0$ é além disso necessário que ψ seja ímpar). O estimador auxiliar de escala, S_n, deve ser o mais robusto possível, não sendo muito importante a sua eficiência, uma vez que o parâmetro de interesse é o de localização. O estimador normalmente recomendado para este efeito é o MAD.

Uma vez que os estimadores-M são definidos por equações implícitas é necessário, na maior parte dos casos, utilizar métodos iterativos para cálculo das estimativas. Dispondo de uma solução inicial, $T_n^{(0)}$ (a mediana é o estimador mais indicado para este efeito por ser o que tem menor sensibilidade), o método de Newton conduz a

$$T_n^{(k+1)} = T_n^{(k)} + S_n \frac{\sum_{i=1}^{n} \psi(u_i^{(k)})}{\sum_{i=1}^{n} \psi'(u_i^{(k)})}, \text{ com } u_i^{(k)} = \frac{x_i - T_n^{(k)}}{S_n}, \tag{3.18}$$

obtendo-se uma sucessão de estimativas que em princípio converge para a estimativa desejada. Pode verificar-se que se ψ for não decrescente ρ é convexa e a equação (3.17) tem um único zero que

corresponde ao único mínimo da função objectivo,

$$\sum_{i=1}^{n} \rho\left(\frac{X_i - T_n}{S_n}\right). \tag{3.19}$$

Nestas condições o único efeito de uma solução inicial desadequada é aumentar o número de iterações necessárias para obter uma estimativa razoável. Se, por outro lado, ψ não for monótona em sentido lato[5] então ρ não é convexa, pelo que (3.19) pode ter vários mínimos e máximos locais correspondentes a vários zeros de (3.17). Neste contexto, para que o algoritmo (3.18) tenha boas propriedades, isto é, convirja para a solução que é o mínimo global da função objectivo, é fundamental que ψ' seja regular e que a solução inicial seja adequada.

Outro processo para obter as estimativas-M pretendidas consiste em calcular em cada iteração uma média ponderada das observações, com pesos variando de iteração para iteração, através da fórmula

$$T_n^{(k+1)} = \frac{\sum_{i=1}^{n} x_i w(u_i^{(k)})}{\sum_{i=1}^{n} w(u_i^{(k)})}, \tag{3.20}$$

com $u_i^{(k)}$ como em (3.18). Se

$$w(u) = \begin{cases} \psi(u)/u, & u \neq 0 \\ 1, & u = 0 \end{cases},$$

e a sucessão das estimativas for convergente, então o limite é solução de (3.17). De facto, fazendo $\psi(u) = u\, w(u)$, (3.17) pode escrever-se como

$$\sum_{i=1}^{n} w\left(\frac{X_i - T_n}{S_n}\right) \times \left(\frac{X_i - T_n}{S_n}\right) = 0 \;\Leftrightarrow\; T_n = \frac{\sum_{i=1}^{n} X_i w\left(\frac{X_i - T_n}{S_n}\right)}{\sum_{i=1}^{n} w\left(\frac{X_i - T_n}{S_n}\right)}$$

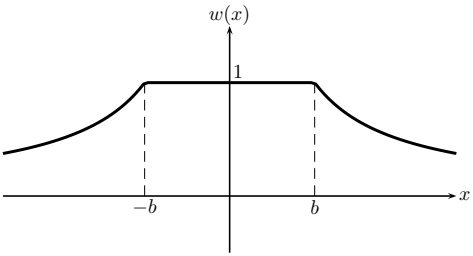
e conclui-se que a equação (3.20) corresponde ao método do ponto fixo para obtenção da solução de (3.17).[6]

Desta forma vê-se que se pode fazer uma caracterização alternativa dos estimadores-M indicando directamente a função $w(u)$. Na Figura 3.3 apresentam-se as funções w correspondentes aos estimadores-M até agora considerados.

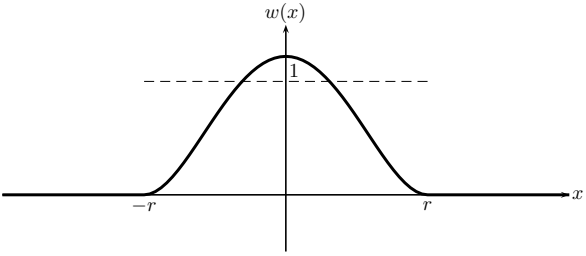
[5] Em inglês costuma dizer-se se ψ for *redescending*.

[6] O que se disse acima acerca da importância da solução inicial para ψ não monótona aplica-se também obviamente a este método.

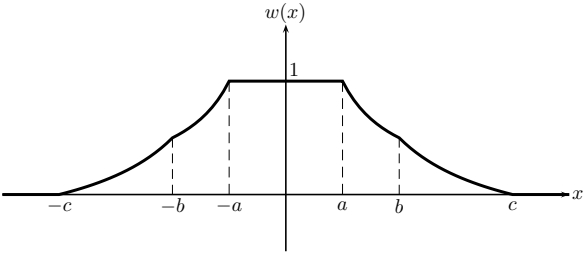
Huber:



Tukey:



Hampel:



Andrews:

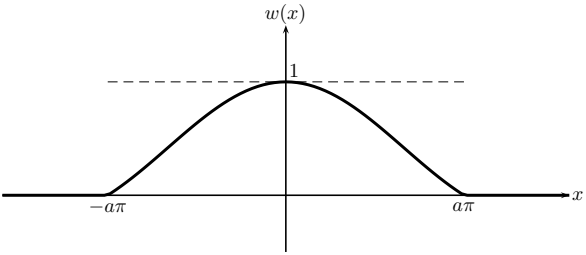


Figura 3.3 *Gráficos de várias funções w associadas a estimadores-M do parâmetro de localização (modelo de localização normal).*

Os estimadores obtidos por intermédio do algoritmo baseado em (3.20) podem também ser vistos como uma generalização dos estimadores dos mínimos quadrados, muito utilizados em regressão, e conhecidos como *Iteratively Reweighted Least Squares* (IRWLS). Recorde-se que a estimativa obtida por aplicação do método dos mínimos quadrados é a média aritmética, sendo a média ponderada com pesos fixos resultante de um critério de mínimos quadrados pesados (com pesos fixos), e a média com pesos a variar iterativamente, a resultante de um critério de "mínimos quadrados iterativamente pesados".

Alguns autores (por exemplo, Rousseeuw e Croux, 1994) têm sugerido que se obtêm boas estimativas apenas com um número fixo de passos daqueles algoritmos. Esse número de passos pode ser tão reduzido como um ou dois. Os estimadores resultantes costumam ser designados por "estimadores-M a um passo"/"estimadores-M a dois passos" (*one-step M-estimators/two step M-estimators*).

É importante também salientar que o ponto de rotura do estimador obtido pelo processo iterativo pode não coincidir com o ponto de rotura teórico de T_n verificando-se que

$$\varepsilon^*(T_n^{(k)}) = \min\left\{\varepsilon^*(S_n); \varepsilon^*(T_n^{(0)}); \varepsilon^*(T_n)\right\},$$

daí que se tenha recomendado utilizar como estimador inicial $T_n^{(0)} = \text{med}(X_i)$ e como estimador auxiliar de escala $S_n = \text{MAD}$, estimadores estes que possuem ponto de rotura igual a $1/2$.

Para cálculo das estimativas não há necessidade de efectuar trabalho de programação pois pode-se recorrer a *software* onde esses cálculos estão implementados (por exemplo, o comando `location.m` do S-PLUS ou as funções `huberM` do *package* `robustbase` e a função `huber` do *package* `MASS`, em R).

Exemplo 3.3. Na Tabela 3.2 apresentam-se várias estimativas-M de localização para os dados do Exemplo 2.1 (utilizando o *software* mencionado acima). Verifica-se que todos os valores obtidos são da ordem de grandeza da estimativa "consensual" ($\hat{\mu} = 3.2$) encontrada no Exemplo 2.13 e praticamente não se nota efeito da constante de afinação. A pequena diferença observada entre os resuultados obtidos pelo estimador de Huber e pelo estimador de Tukey podem ser justificados pelo facto deste último ter ponto de rejeição finito, o que faz com que não se faça sentir o efeito da última observação.

Tabela 3.2 *Estimativas-M de localização para os dados do Exemplo 2.1.*

b	1.30	1.45	2.00
Huber(b)	3.22	3.21	3.21
r	4	5	6
Tukey(r)	3.15	3.15	3.16

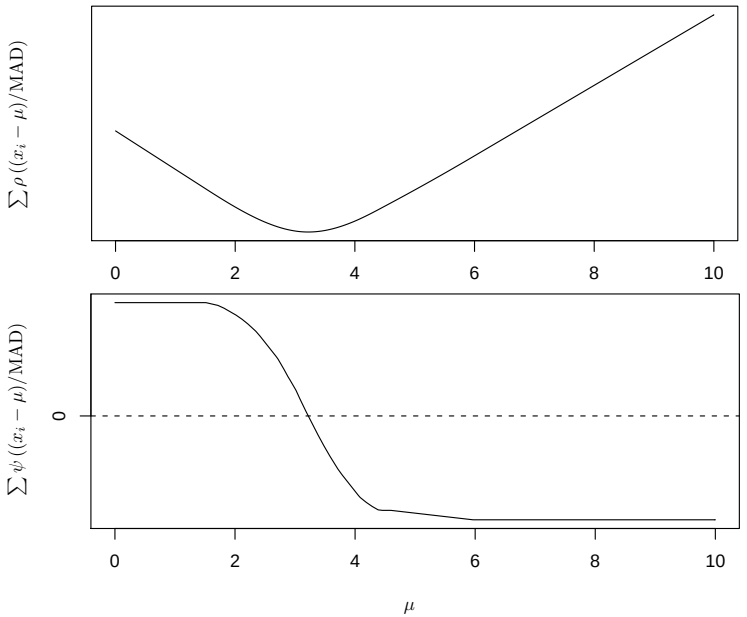


Figura 3.4 *Função objectivo e respectiva derivada correspondentes ao estimador de Huber* ($b = 1.3$) *para os dados do Exemplo 3.3.*

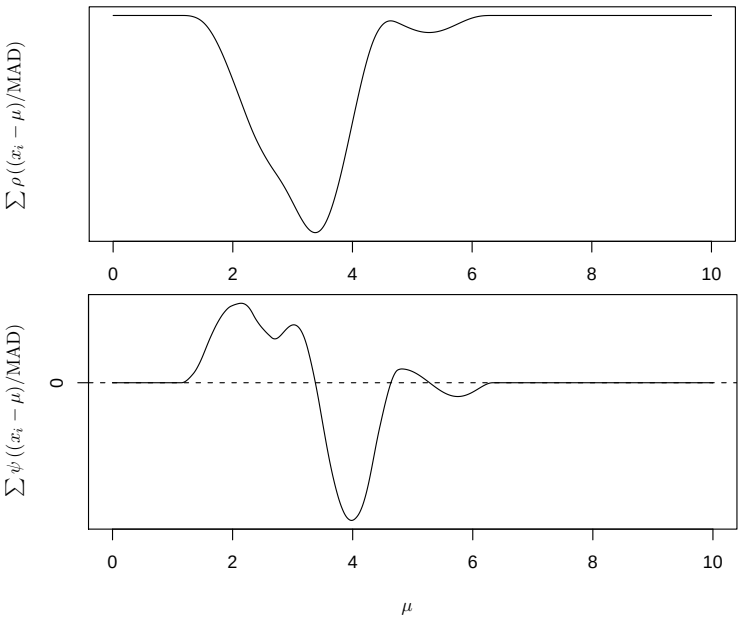


Figura 3.5 *Função objectivo e respectiva derivada correspondentes ao estimador de Tukey* ($r = 2$) *para os dados do Exemplo 3.3.*

Para ilustrar a importância das soluções iniciais e a convexidade (ou não) das funções objectivo calcularam-se para estes dados as funções

$$\sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\text{MAD}}\right) \quad \text{e} \quad \sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\text{MAD}}\right),$$

com $\mu \in [0; 10]$ e para os estimadores de Huber (com $b = 1.3$) e Tukey (com $r = 2$). Os gráficos respectivos apresentam-se nas Figuras 3.4 e 3.5. Pode verificar-se que no caso do estimador de Huber a função objectivo é convexa, pelo que tem um único ponto de mínimo, correspondente à única solução ($\hat{\mu} \simeq 3.22$) da equação

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\text{MAD}}\right) = 0.$$

Já no caso do estimador de Tukey a função objectivo não é convexa, possuindo um mínimo absoluto que corresponde a uma "boa" solução e um mínimo local que corresponde a uma "má" solução. A derivada

da função objectivo tem em consequência vários zeros. Se se usar uma estimativa inicial desadequada o método iterativo pode convergir para o mínimo local e não para o mínimo absoluto (é o que acontece se com a função `location.m` do S-PLUS se indicar como estimativa inicial o valor 5, a estimativa final vem igual a 5.28).

É importante notar que que se ψ for monótona em sentido lato, o que acontece com o estimador de Huber, pode garantir-se que ρ é convexa. Por outro lado, se ψ não for monótona, como é o caso do estimador de Tukey, ρ não é convexa mas pode ter, dependendo da constante de afinação, um único mínimo. De qualquer modo mantém-se a importância da estimativa inicial pois a derivada terá sempre uma região onde é constante e igual a zero e é importante evitar essa região.

Exemplo 3.4. Uma outra forma de obter um estimador-M robusto consiste em considerar o estimador de máxima verosimilhança para um modelo de localização com caudas pesadas. O modelo de localização baseado na distribuição t-Student com ν graus de liberdade tem essa característica, sendo as caudas tanto mais pesadas quanto menor for ν. Para esse modelo tem-se

$$f_\theta(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{(x-\theta)^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

com $\nu > 0$ não necessariamente inteiro. Tem-se então, como anteriormente,

$$f(x) = K(\nu) \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

pelo que se pode considerar

$$\rho(x) = -\log f(x) + \log K(\nu) = \frac{\nu+1}{2} \log\left(1 + \frac{x^2}{\nu}\right)$$

e

$$\psi(x) = \rho'(x) = \frac{(\nu+1)x}{\nu + x^2}. \tag{3.21}$$

Nas Figuras 3.6 e 3.7 apresentam-se os gráficos das funções ρ e ψ, respectivamente, para vários valores de ν. No limite ($\nu = \infty$) tem-se como se sabe o modelo normal.

Este exemplo é interessante por mostrar que sob um modelo propenso à ocorrência de *outliers*, o estimador de máxima verosimilhança

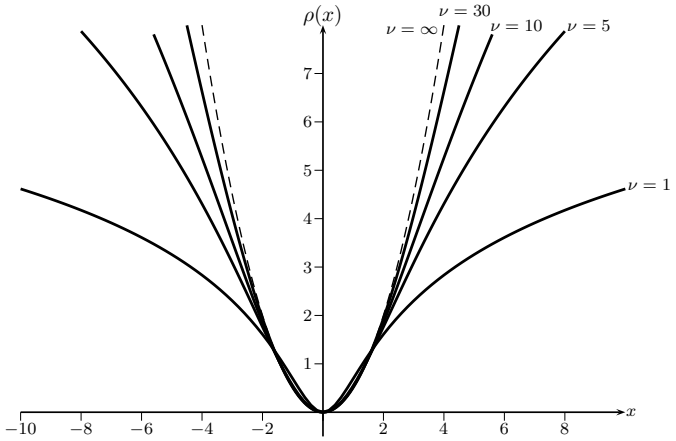


Figura 3.6 *Funções ρ para os estimadores de máxima verosimilhança do parâmetro de localização sob o modelo t-Student com ν graus de liberdade.*

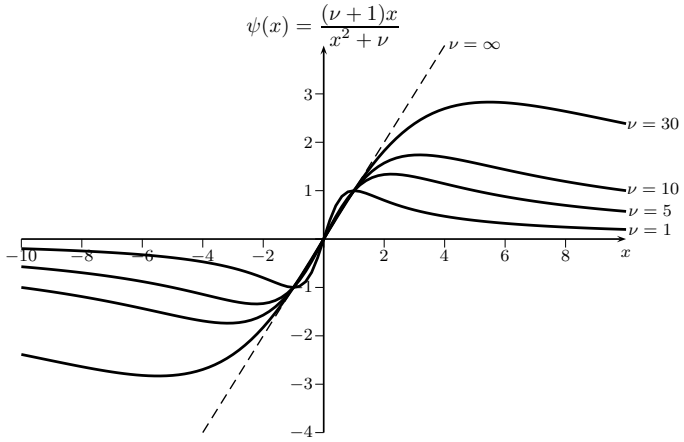


Figura 3.7 *Funções ψ para os estimadores de máxima verosimilhança do parâmetro de localização sob o modelo t-Student com ν graus de liberdade.*

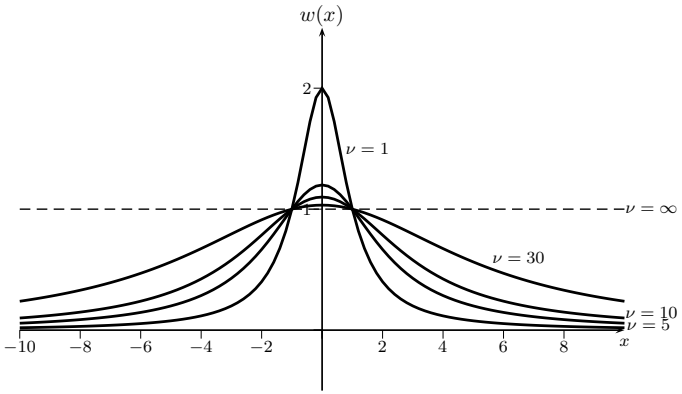


Figura 3.8 *Funções w para os estimadores de máxima verosimilhança do parâmetro de localização sob o modelo t-Student com ν graus de liberdade.*

encarrega-se automaticamente da atribuição de um "peso" pequeno (próximo de zero) às observações mais afastadas. Este efeito pode ver-se nos gráficos apresentados na Figura 3.8.

Desta forma obtém-se uma outra de classe de funções ψ que não se destina a ser usada apenas sob este modelo exacto mas, do mesmo modo que as funções apresentadas anteriormente, pode ser aplicada para protecção contra *outliers* também sob o modelo central normal, funcionando o parâmetro ν como uma constante de afinação. Para ajudar na escolha desta constante apresentam-se na Tabela 3.3 os valores de ν que garantem para o estimador-M em causa eficiências de 90%, 95% e 97.5% sob o modelo central normal, bem como os valores da sensibilidade a grandes erros (γ^*) correspondentes.

Depois de apresentado este conjunto de estimadores-M surge naturalmente a questão: qual é o que deve ser escolhido numa situação concreta? Na maior parte dos casos a forma da função ψ (ou ρ, ou w) acaba por não ser muito relevante, como se viu no Exemplo 3.3. Além disso, também como se viu, todos os estimadores podem ser calibrados para uma dada eficiência sob o modelo central e todos têm ponto de rotura de 50%. As diferenças em termos de robustez acabam por se notar apenas na sensibilidade a grandes erros (γ^*) e na curva de enviesamento assintótico máximo. Para atingir maior

Tabela 3.3 *Valores do parâmetro ν (constante de afinação) que garantem para o estimador com função ψ dada por (3.21) eficiências de 90%, 95% e 97.5% sob o modelo central normal e respectivas sensibilidades (γ^*).*

Eficiência	ν	γ^*
90%	2.98	1.519
95%	5.69	1.710
97.5%	9.82	1.985

eficiência paga-se um preço em relação a essas duas características, sendo por vezes mais adequado escolher um estimador com menor eficiência mas com menos sensibilidade.

A função de Huber parece ter sido popular durante muito tempo (talvez por razões históricas), mas actualmente regista-se uma maior preferência por funções ψ com ponto de rejeição finito ou pelo menos decrescentes a partir de certo ponto para oferecerem uma protecção extra contra *outliers* extremos. A diferenciabilidade é também uma vantagem acrescida. Uma função que verifica estas duas condições e tem ganho popularidade nos últimos anos é a função ψ de Tukey. A única desvantagem desta escolha consiste numa maior probabilidade de ocorrência de mínimos locais na função objectivo, o que implica que é necessário ter um cuidado acrescido com a escolha das soluções iniciais.

3.2.3 Modelo de escala

Do mesmo modo que o conceito de modelo de localização está ligado à ideia intuitiva de medida de localização, o conceito de modelo de escala (também chamado de dispersão) está ligado à ideia intuitiva de medida de dispersão (ou escala). E do mesmo modo que o conceito de medida de localização se relaciona com translações dos dados, o de medida de dispersão relaciona-se com mudanças de escala dos dados, uma vez que é natural exigir a uma medida de dispersão que reflicta as mudanças de escala dos dados. Mais concretamente, ao mudar a escala de um conjunto de dados, multiplicando todas as observações por uma constante positiva, espera-se que a medida de dispersão sofra

o mesmo efeito. Ou seja, sendo s uma medida de dispersão genérica ela deve verificar

$$s(b\,x_1, \ldots, b\,x_n) = b\,s(x_1, \ldots, x_n),$$

para qualquer $b \in \mathbb{R}^+$ e quaisquer observações $(x_1, \ldots, x_n)$. A esta propriedade chama-se equivariância em relação a mudanças de escala ou simplesmente equivariância em relação à escala.[7] Passando ao campo teórico esta ideia pode ser traduzida por: dado um certo modelo com um parâmetro θ, F_θ, o parâmetro θ (pode haver outros parâmetros no modelo) é um parâmetro de escala e o modelo é um modelo de escala, se dada uma variável X com distribuição desse modelo e parâmetro $\theta = a$, a variável que se obtém multiplicando-a por b, $b\,X$, ainda tem, para qualquer $b > 0$, uma distribuição da mesma família mas com parâmetro $\theta = a\,b$.

Conclui-se facilmente que o modelo normal considerando o parâmetro desvio padrão, σ, é um modelo de escala, pois se X tem distribuição normal com desvio padrão σ, então $b\,X$, com $b > 0$, tem distribuição normal com desvio padrão $b\,\sigma$. Um outro exemplo é dado pela distribuição exponencial, $X \sim Exp(\beta)$, com densidade dada em (3.7), pois $X_b = b\,X$, com $b > 0$, tem distribuição $X \sim Exp(b\,\beta)$, o que se pode ver facilmente fazendo

$$F_{X_b}(x) = P(b\,X \leq x) = P\left(X \leq \frac{x}{b}\right) = F_X\left(\frac{x}{b}\right).$$

Então, tal como para o modelo de localização, dada uma variável aleatória com uma qualquer distribuição pode criar-se a partir dela um modelo de escala, usando como parâmetro o próprio valor da mudança de escala, b. Basta pensar na transformação $X_b = b\,X$ e da igualdade

$$F_{X_b}(x) = F_X\left(\frac{x}{b}\right)$$

obtém-se o modelo pretendido, o qual, se corresponder a uma variável aleatória contínua, tem densidade

$$f_{X_b}(x) = \frac{1}{b}\,f_X\left(\frac{x}{b}\right).$$

A definição seguinte formaliza o que se acabou de dizer.

[7] Também neste caso é simples verificar que todas as medidas de dispersão consideradas anteriormente (desvio padrão, desvio médio, desvio absoluto mediano, desvios padrões aparados e amplitude inter-quartis) verificam esta propriedade.

Definição 3.5. *Um modelo de escala, com parâmetro de escala θ, consiste numa família de distribuições $\{F_\theta,\, \theta \in \mathbb{R}^+\}$ tal que*

$$F_\theta(x) = F\left(\frac{x}{\theta}\right),$$

onde $F = F_1$ representa uma função de distribuição univariada genérica que define o tipo ou família do modelo.

É fácil verificar que se $X \sim F$, ou seja com $\theta = 1$, então $\theta X \sim F_\theta$, ou ainda que se $Y \sim F_\theta$ então $Y/\theta \sim F$, e em geral, com $b > 0$, $bY \sim F_{b\theta}$.

Admite-se que F possui uma densidade $f = F'$, pelo que $f(x)$ representa a densidade quando $\theta = 1$, enquanto que

$$f_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$$

representa a densidade para um valor genérico do parâmetro θ.

Dada uma amostra aleatória $(X_1, \ldots, X_n)$ o estimador de máxima verosimilhança, $S_n = S_n(X_1, \ldots, X_n)$, do parâmetro θ do modelo de escala maximiza

$$\sum_{i=1}^{n} \left[\log f\left(\frac{X_i}{S_n}\right) - \log S_n\right]$$

o que é equivalente a minimizar

$$n \log S_n + \sum_{i=1}^{n} \rho\left(\frac{X_i}{S_n}\right), \tag{3.22}$$

com $\rho(u) = -\log f(u)$. Ou ainda de forma equivalente, nos casos regulares de estimação, S_n é solução de

$$\sum_{i=1}^{n} \psi\left(\frac{X_i}{S_n}\right) = 0, \tag{3.23}$$

com

$$\psi(u) = -u\frac{f'(u)}{f(u)} - 1. \tag{3.24}$$

Pois

$$\frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{1}{\theta}\left(-\frac{x}{\theta}\frac{f'\left(\frac{x}{\theta}\right)}{f\left(\frac{x}{\theta}\right)} - 1\right),$$

multiplicando por θ, o que obviamente não altera a solução, e fazendo $u = \frac{x}{\theta}$ obtém-se a função dada em (3.24). A equação (3.23) pode também ser escrita como

$$\sum_{i=1}^{n} \chi\left(\frac{X_i}{S_n}\right) = n, \tag{3.25}$$

com

$$\chi(u) = -u\frac{f'(u)}{f(u)} \quad \text{ou seja,} \quad \psi(u) = \chi(u) - 1. \tag{3.26}$$

Para definir um estimador-M no contexto do modelo de escala considera-se a expressão (3.23) em que ψ é uma função genérica de apenas uma variável. Ou seja, passa-se da definição geral para a definição particular adequada a este modelo começando por considerar, por analogia com o estimador da máxima verosimilhança, a forma $\psi(x, \theta) = \psi(x/\theta)$. Há no entanto que ter, em comparação com o modelo de localização tratado na secção anterior, um cuidado adicional. Para que o funcional equivalente, S, definido implicitamente por

$$\int \psi\left(\frac{x}{S(F)}\right) dF(x) = 0,$$

seja consistente segundo Fisher deve verificar-se

$$\int \psi\left(\frac{x}{\theta}\right) dF\left(\frac{x}{\theta}\right) = 0 \Leftrightarrow$$

$$\Leftrightarrow \int \psi(u) dF(u) = 0 \Leftrightarrow$$

$$\Leftrightarrow \int (\chi(u) - 1) dF(u) = 0,$$

e se se mudar apenas a forma funcional de χ, passando por exemplo de uma função ilimitada para uma função limitada, não se obtém em geral consistência, pois não se pode garantir que $\int \chi(u) dF(u) = 1$, nem mesmo nos casos em que F corresponde a uma distribuição simétrica em relação à origem. Suponha-se então que se define primeiro a forma da função χ e se conclui que (sob o modelo central)

$$\int \chi(u) dF(u) = \beta. \tag{3.27}$$

Para garantir a consistência basta então considerar na definição do estimador $\psi(u) = \chi(u) - \beta$. Tem-se então que, dada uma amostra

aleatória $(X_1, \ldots, X_n)$ um estimador-M, $S_n = S_n(X_1, \ldots, X_n)$, do parâmetro θ do modelo de escala é solução de

$$\sum_{i=1}^{n} \chi\left(\frac{X_i}{S_n}\right) = n\,\beta, \tag{3.28}$$

onde χ é uma função genérica de uma variável e β é a constante dada por (3.27) e que garante a consistência à Fisher do funcional equivalente sob o modelo central F. A equação (3.28) é ainda equivalente a (3.23) desde que se tenha o cuidado de considerar $\psi(u) = \chi(u) - \beta$. Pode ainda mostrar-se que esta definição é equivalente, nos casos regulares de estimação, à definição de S_n como o estimador que minimiza

$$n\,\beta \log S_n + \sum_{i=1}^{n} \rho\left(\frac{X_i}{S_n}\right), \tag{3.29}$$

com $\rho(u) = P\left(\frac{\chi(u)}{u}\right)$.

Uma propriedade desejável para os estimadores do parâmetro de escala de um modelo de escala é a equivariância em relação à escala já referida na secção anterior mas apresentada formalmente na definição seguinte.

Definição 3.6. *Um estimador S_n do parâmetro de escala de um modelo de escala é equivariante em relação à escala (scale equivariant) se dada a amostra aleatória $(X_1, \ldots, X_n)$,*

$$S_n(k\,X_1, \ldots, k\,X_n) = k\,S_n(X_1, \ldots, X_n) \quad \forall_{k \in \mathbb{R}^+}.$$

É simples verificar que se S_n é solução de (3.28) então $k\,S_n$, com $k > 0$, é solução da equação que se obtém substituindo X_i por $k\,X_i$.

Não é muito difícil concluir que para estimadores (funcionais) equivariantes se verifica, sob as distribuições do modelo,

$$IF(x; S, F_\theta) = \theta\,IF\left(\frac{x}{\theta}; S, F\right) = \theta\,IF(u; S, F),$$

com $u = x/\theta$. Isto significa que todas as propriedades do estimador que recorrem à função de influência podem ser avaliadas na distribuição central F.

Partindo da expressão da função de influência do estimador-M geral, (3.6), têm-se então os seguintes resultados para um estimador-M, equivalente a um funcional S, do parâmetro de escala num modelo de escala (na distribuição central F):

- $$IF(x; S, F) = \frac{\psi(x)}{\int u\psi'(u)dF(u)} = \frac{\psi(x)}{M}$$
  e todas as medidas (quer as relacionadas com a eficiência quer as relacionadas com a robustez) obtidas a partir da função de influência são determinadas pela forma da função ψ.

- A variância assintótica é dada por
$$V(S, F) = \frac{\int \psi^2(u)dF(u)}{M^2}. \tag{3.30}$$

- O estimador é B-robusto sse ψ for limitada sendo a sensibilidade a grandes erros dada por
$$\gamma^*(S, F) = \frac{\sup_x |\psi(x)|}{M}. \tag{3.31}$$

Também aqui, no caso de $\psi(x)$ ser uma função não decrescente para $x > 0$, é possível garantir que (Hampel *et al.*, 1986, p. 107):

- se ψ é limitada então o estimador é qualitativamente robusto e tem ponto de rotura dado por
$$\varepsilon^*(S, F) = \min\left\{\frac{-\psi(0)}{\psi(\infty) - \psi(0)}; \frac{1}{2}\right\} = \min\left\{\frac{\beta}{\chi(\infty)}; \frac{1}{2}\right\}. \tag{3.32}$$
- se ψ é ilimitada então o estimador não é qualitativamente robusto e o seu ponto de rotura é $\varepsilon^*(S, F) = 0$.

Nos exemplos seguintes aplicam-se os conceitos acabados de apresentar a modelos de escala particulares. Começa-se pelo modelo de escala normal, de longe o mais importante.

Exemplo 3.5. No modelo de escala normal considera-se que X_i tem distribuição $\mathcal{N}(\mu, \theta^2)$, com μ conhecido. Sem perda de generalidade

pode estudar-se o caso particular $\mu = 0$, ou seja, $X_i \sim \mathcal{N}(0, \theta^2)$ (se $\mu \neq 0$ mas conhecido trabalha-se com $X_i - \mu$). A importância deste modelo deriva do facto de ser geralmente o modelo assumido para os erros nos modelos de regressão (ver Capítulo 4). Tem-se então que $F_\theta(x) = F(x/\theta)$ com $F(u) = \Phi(u)$ e

$$f_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right) \quad \text{com} \quad f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

Para este modelo o estimador de máxima verosimilhança minimiza (3.22) com

$$\rho(u) = -\log f(u) = \frac{u^2}{2} + \sqrt{2\pi},$$

ou, de forma equivalente, com

$$\rho(u) = \frac{u^2}{2},$$

o que é ainda equivalente, por (3.24), a resolver (3.23) com

$$\psi(u) = u^2 - 1,$$

ou ainda (3.25) com $\chi(u) = u^2$. Logo

$$\sum_{i=1}^{n} \psi\left(\frac{X_i}{\hat{\theta}_{MV}}\right) = 0 \;\Leftrightarrow\; \sum_{i=1}^{n} \left(\frac{X_i}{\hat{\theta}_{MV}}\right)^2 = n \;\Leftrightarrow\; \hat{\theta}_{MV} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}},$$

ou seja, como não podia deixar de ser, conclui-se que o estimador de máxima verosimilhança de θ no modelo de escala normal é o desvio padrão amostral (em relação à média suposta conhecida e igual a zero). Tal como no Exemplo 3.1, o raciocínio anterior, baseado em equivalências, permite responder à questão em sentido contrário: qual é o modelo de escala (regular) para o qual o estimador de máxima verosimilhança é o desvio padrão? E a resposta é: unicamente o modelo normal.

Exemplo 3.6. Considere-se agora o modelo de escala de *Laplace*, com $X_i \sim Laplace(0, \theta)$ e

$$f(u) = \frac{1}{2} \exp\left(-|u|\right).$$

É fácil confirmar que $\rho(u) = |u|$ (a menos de uma constante) e que

$$\psi(u) = |u| - 1 \qquad (\text{ou } \chi(u) = |u|),$$

pelo que

$$\sum_{i=1}^{n} \psi\left(\frac{X_i}{\hat{\theta}_{MV}}\right) = 0 \;\Leftrightarrow\; \sum_{i=1}^{n} \left|\frac{X_i}{\hat{\theta}_{MV}}\right| = n \;\Leftrightarrow\; \hat{\theta}_{MV} = \frac{\sum_{i=1}^{n} |X_i|}{n},$$

e conclui-se, tal como esperado, que o estimador de máxima verosimilhança de θ no modelo de escala de *Laplace* é o desvio médio amostral (em relação à localização suposta conhecida e igual a zero).

Também se conclui, como aliás já se sabia, que tanto este estimador-M de escala como o do exemplo anterior, têm função de influência ilimitada pelo que não são nem B-robustos, nem qualitativamente robustos e têm ponto de rotura nulo.

Para ilustrar a questão da consistência verifique-se que:

- Para f do modelo *Laplace*,

$$\int \psi(u)dF(u) = \int_{-\infty}^{+\infty} (|u| - 1)f(u)du = E(|Z|) - 1 = 0$$

  (onde $Z \sim Laplace(0, 1)$), o que significa que o estimador desvio médio é consistente para o parâmetro θ neste modelo, como não podia deixar de ser.

- Para f do modelo normal, tem-se igualmente

$$\int \psi(u)dF(u) = \int_{-\infty}^{+\infty} (u^2 - 1)f(u)du = E(Z^2) - 1 = 0$$

  (onde $Z \sim \mathcal{N}(0, 1)$).
- Mas se se utilizar o estimador-M definido por $\psi(u) = |u| - 1$ no modelo normal, tem-se com f do modelo normal

$$\int \psi(u)dF(u) = \int_{-\infty}^{+\infty} (|u| - 1)f(u)du = E(|Z|) - 1 = \sqrt{\frac{2}{\pi}} - 1$$

  (onde novamente $Z \sim \mathcal{N}(0, 1)$), o que mostra que o desvio médio não é consistente para o parâmetro de escala do modelo normal. No entanto, se se substituir a função ψ acima por $\psi(u) = |u| - \sqrt{2/\pi}$, o integral $\int \psi(u)dF(u)$ já se anula, obtendo-se então o estimador, consistente sob o modelo normal,

$$\hat{\theta} = \sqrt{\frac{\pi}{2}} \sum_{i=1}^{n} \frac{|X_i|}{n}.$$

Novamente terá interesse a obtenção de estimadores robustos para o parâmetro de escala quando o modelo paramétrico central é o normal (o qual, como se disse, se supõe verificado apenas aproximadamente).

Tendo em atenção as propriedades desejáveis um estimador-M do parâmetro de escala sob o modelo central normal deve ter função $\chi(u)$:

- limitada para garantir ponto de rotura finito;
- positiva, simétrica, com $\chi(0) = 0$ e próxima de u^2 na vizinhança da origem, isto de forma a ser nessa região semelhante à função χ do EMV e conduzir assim a um estimador com boa eficiência assintótica exactamente sob o modelo (ou seja com variância assintótica próxima da do desvio padrão);
- tal que $\sup_x |\chi(x) - \beta|$ não seja muito elevado para garantir uma baixa sensibilidade a grandes erros e simultaneamente uma curva de enviesamento assintótico máximo bem comportada;
- contínua, e eventualmente nula para valores afastados da origem, garantindo, respectivamente, baixa sensibilidade local e ponto de rejeição finito.

Note-se que como habitualmente a multiplicação de ψ (ou de χ e β) por uma constante não altera o estimador. As funções apresentadas na secção anterior para o modelo de localização fornecem algumas ideias para esta nova situação. Em particular pode-se pensar em $\chi = \psi_L^2$ ou em $\chi = \rho_L$ se esta for do tipo limitado.[8]

O primeiro estimador-M com boas propriedades de robustez, no contexto do modelo de escala normal, foi igualmente proposto em Huber (1964). Trata-se do estimador definido por

$$\chi(x) = [\psi_b(x)]^2 = \begin{cases} u^2, & |u| \leq b \\ \\ b^2, & |u| > b \end{cases} \tag{3.33}$$

ou por $\psi(x) = \chi(x) - \beta$. A constante β deve ser determinada precisamente de modo a que o estimador seja consistente sob aquele

[8] O índice L refere-se a localização.

modelo, isto é, faz-se $\beta = \int \chi(u) d\Phi(u)$ de modo a que se tenha $\int \psi(u) d\Phi(u) = 0$. Esta condição conduz à expressão

$$\beta(b) = F(b^2; \chi^2_3) + 2\, b^2 \Phi(-b), \tag{3.34}$$

onde $F(\cdot; \chi^2_\nu)$ representa a função de distribuição de uma variável aleatória com distribuição do qui-quadrado com ν graus de liberdade.

Como ψ é limitada e não decrescente (para $x > 0$) o estimador correspondente é B-robusto, qualitativamente robusto e tem, por (3.32), ponto de rotura dado por

$$\varepsilon^* = \min\left\{ \frac{\beta(b)}{b^2}, \frac{1}{2} \right\}.$$

Este estimador é um caso particular do estimador determinado por Huber ao formular o problema sob a perspectiva minimax. Do mesmo modo que foi descrito na secção anterior para o modelo de localização, é possível determinar, dada uma vizinhança-ε do modelo normal, a distribuição mais desfavorável (ou seja, a que minimiza a informação de Fisher) para o parâmetro de escala e o correspondente estimador de máxima verosimilhança, o qual é definido por

$$\chi(x) = \begin{cases} 0, & |x| < b_0 \\ x^2 - b_0^2, & b_0 \le x \le b \\ b^2 - b_0^2, & |x| > b \end{cases} . \tag{3.35}$$

A expressão (3.33) obtém-se desta última fazendo $b_0 = 0$ (o que acontece na formulação minimax para $\varepsilon \le 0.205$).

Um outro estimador popular é baseado na função ρ correspondente ao estimador biponderado de Tukey para a localização. Para o parâmetro de escala essa função costuma ser escrita, sem alteração das características do estimador, na forma normalizada tal que $0 \le \chi \le 1$,

$$\chi(x) = \begin{cases} \left(\frac{x}{r}\right)^6 - 3\left(\frac{x}{r}\right)^4 + 3\left(\frac{x}{r}\right)^2, & |x| \le r \\ 1, & |x| > r \end{cases} . \tag{3.36}$$

Finalmente considerando o modelo t-Student obtém-se a seguinte função

$$\chi(x) = \frac{(\nu+1)x^2}{x^2 + \nu}.$$

Na Figura 3.9 apresentam-se os gráficos das quatro funções χ acabadas de descrever. Note-se que em todos os casos fazendo variar convenientemente as constantes de afinação obtém-se como limite uma função que coincide com a função $\chi(u) = u^2$ que define o desvio padrão (no caso da função de Tukey é preciso considerar a versão não normalizada mas, como já se referiu, ambas são equivalentes). Além disso o estimador de Huber generalizado aproxima-se do MAD, que também pode ser visto como um estimador-M de escala, quando se multiplica a função pela constante $b^2 - b_0^2$ e se faz $b_0 \to b = \Phi^{-1}(3/4)$. De facto nesse caso obtém-se a função χ dada por

$$\chi(x) = \begin{cases} 0, & |x| < b \\ \\ 1, & |x| \geq b \end{cases},$$

$\beta = 1/2$ e, a menos de uma constante, a função $\psi(x) = \text{sinal}(|x| - b)$. O estimador-M de escala correspondente é portanto solução de

$$\sum_{i=1}^{n} \text{sinal}\left(\left|\frac{X_i}{\hat{\theta}}\right| - b\right) = 0 \;\Leftrightarrow\; \sum_{i=1}^{n} \text{sinal}\left(|X_i| - b\,\hat{\theta}\right) = 0,$$

onde a equivalência é justificada por se ter necessariamente $\hat{\theta} > 0$. A solução da segunda equação é dada, usando o argumento usado no Exercício 3.2 para concluir que a mediana é o EMV no modelo de localização de *Laplace*, por

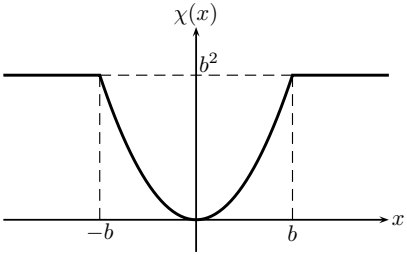
$$\hat{\theta} = \text{med}(|X_i|)/b = 1.4826\,\text{med}(|X_i|),$$

ou ainda, pelo facto de se ter assumido a localização conhecida e igual a zero, $\hat{\theta} = \text{MAD}(X_i)$.

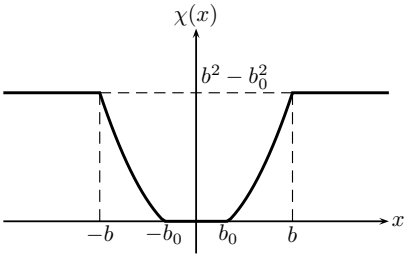
Nenhuma das funções apresentadas tem ponto de rejeição finito pelo que todas correspondem ao estimador de máxima verosimilhança para algum modelo. Embora tenham surgido propostas de estimadores-M de escala com ponto de rejeição finito (ver Hampel *et al.*, 1986, p. 168) elas nunca tiveram grande popularidade, provavelmente porque na situação de escala os problemas de convergência dos algoritmos se agravam.

É interessante constatar que tal como o estimador de Huber de localização é assintoticamente localmente equivalente à média aparada a $100 \times \alpha\%$, quando se escolhe $b = \Phi^{-1}(1 - \alpha)$, também o estimador de Huber de escala é equivalente a um desvio padrão aparado a

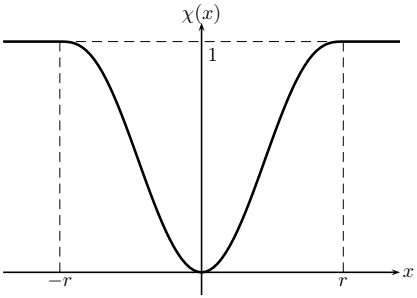
Huber:



Huber geral:



Tukey:



EMV t-Student:

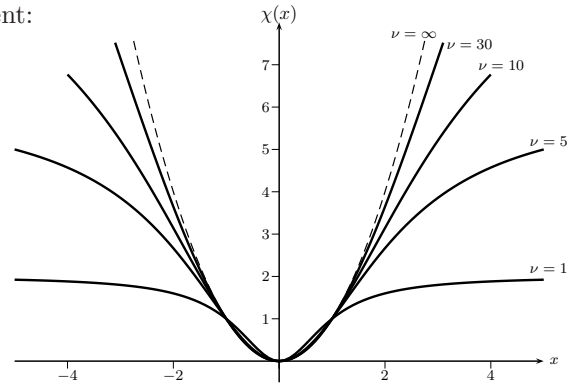


Figura 3.9 *Gráficos de várias funções χ associadas a estimadores-M do parâmetro de escala (modelo de escala normal).*

$100 \times \alpha\%$, com a mesma relação entre α e b (basta verificar que as funções de influência respectivas coincidem). No entanto, mais uma vez se verifica que em termos de ponto de rotura o estimador-M tem muito melhor comportamento, sendo a diferença muito mais marcada que no caso dos estimadores de localização.

Apresentam-se na Tabela 3.4 alguns resultados numéricos que ajudam a escolher as constantes de afinação envolvidas nas diversas funções χ. Para além dos valores das constantes que garantem ou o máximo ponto de rotura ou valores elevados de eficiência assintótica para os diversos estimadores sob o modelo central normal, indica-se ainda em cada caso o valor da constante β, do ponto de rotura (ε^*) e da sensibilidade a grandes erros (γ^*). A constante de consistência é dada no caso do estimador de Huber pela expressão (3.34). Para os estimadores de Tukey e MV t-Student há também fórmulas explícitas dadas, respectivamente, por

$$\beta(r) = 2\Phi(-r) + \frac{15}{r^6} F(r^2; \chi^2_7) - \frac{9}{2r^4} F(r^2; \chi^2_5) + \frac{3}{r^2} F(r^2; \chi^2_3)$$

e

$$\beta(\nu) = (\nu + 1)\left(1 - e^{\nu/2}\sqrt{2\nu\pi}\,\Phi(-\sqrt{\nu})\right).$$

O ponto de rotura foi determinado em todos os casos recorrendo à expressão (3.32), enquanto que a eficiência assintótica foi calculada a partir da variância assintótica dada por (3.30) e a sensibilidade por (3.31).

Aqui se revela mais uma vez que na estimação do parâmetro de escala é mais difícil conciliar eficiência e robustez do que na estimação do parâmetro de localização. Atente-se por exemplo nos resultados relativos ao desvio padrão aparado que coincidem com os do estimador de Huber relativamente à eficiência e sensibilidade mas que são muito piores no que diz respeito ao ponto de rotura (que é igual a α e quase nulo para as eficiências mais elevadas).[9] Estes estimadores são por esta razão totalmente desaconselhados em face da existência de um estimador assintoticamente equivalente (sob o modelo central) e com muito melhor ponto de rotura que é o estimador de Huber. Mas mesmo para este estimador e ainda mais para os estimadores de Tukey e MV t-student é difícil conciliar alta eficiência com alto ponto de rotura. De entre os estimadores-M de escala considerados

[9] Ver também a Tabela 2.4 na página 67.

Tabela 3.4 *Valores das constantes de afinação que garantem para os diversos estimadores ou ponto de rotura de 50%, ou eficiências assintóticas de 85%, 90%, 95% e 97.5% sob o modelo central normal.*

Huber					
Eficiência	b	α	β	ε^*	γ^*
50.5%	1.04	0.1492	0.541167	0.500	1.238
85%	1.86	0.0314	0.891519	0.258	1.905
90%	2.07	0.0197	0.930859	0.219	2.184
95%	2.38	0.0087	0.968936	0.171	2.696
97.5%	2.65	0.0040	0.985341	0.140	3.250

	Tukey			EMV t-Student			
Eficiência	r	$\beta = \varepsilon^*$	γ^*	ν	β	ε^*	γ^*
53.9%	1.548	0.49991	1.284	–	–	–	–
52.2%	–	–	–	0.375	0.68724	0.500	1.599
85%	2.50	0.30916	1.803	2.86	0.75549	0.196	3.289
90%	3.21	0.22004	2.402	4.46	0.79668	0.146	4.201
95%	3.92	0.16125	3.192	8.41	0.85630	0.091	6.326
97.5%	4.70	0.11876	4.274	14.35	0.89990	0.059	9.404

até aqui o que reúne, apesar de tudo melhores propriedades é o de Huber. No final desta secção este assunto será de novo discutido, sendo apresentadas outras alternativas

Em muitas situações práticas, mesmo que o parâmetro de interesse seja apenas o de escala, o parâmetro de localização não é conhecido. Nesses casos é necessário determinar uma estimativa auxiliar de localização $T_n(x_1, \ldots, x_n)$ e em seguida determina-se a estimativa de escala para as observações centradas $x_i^* = x_i - T_n$. Por esta via obtém-se (se o estimador de localização T_n for equivariante em relação à localização e à escala, ver Definições 3.3 e 3.4) um estimador do parâmetro de escala que é invariante em relação à localização, de acordo com a definição seguinte, e equivariante em relação à escala de acordo com a Definição 3.6.

Definição 3.7. *Um estimador S_n do parâmetro de escala de um modelo de escala é invariante em relação à localização (location invariant) se dada a amostra aleatória $(X_1, \ldots, X_n)$,*

$$S_n(X_1 + k, \ldots, X_n + k) = S_n(X_1, \ldots, X_n) \quad \forall_{k \in \mathbb{R}}.$$

O estimador auxiliar T_n deve ser o mais robusto possível de modo a não afectar as propriedades de robustez do estimador de escala, pelo que usualmente se recomenda $T_n = \text{med}(X_i)$.

Finalmente, para cálculo das estimativas nos casos em que a solução não é explícita, pode recorrer-se ao método do ponto fixo, tal como foi feito na secção anterior em relação aos estimadores de localização. Recorde-se que se pretende resolver em ordem a S_n a equação

$$\sum_{i=1}^{n} \chi\left(\frac{x_i - T_n}{S_n}\right) = n\,\beta. \tag{3.37}$$

Escrevendo $\chi(u) = u^2 w^2(u)$ (note-se que $\chi(u) \geq 0$), ou de forma equivalente, definindo

$$w(u) = \begin{cases} \dfrac{\sqrt{\chi(u)}}{|u|}, & u \neq 0 \\ 1, & u = 0 \end{cases},$$

tem-se que (3.37) se pode escrever como

$$\sum_{i=1}^{n} w^2\left(\frac{x_i - T_n}{S_n}\right) \times \left(\frac{x_i - T_n}{S_n}\right)^2 = n\,\beta \;\Leftrightarrow$$

$$\Leftrightarrow\; S_n^2 = \frac{1}{n\,\beta} \sum_{i=1}^{n} w^2\left(\frac{x_i - T_n}{S_n}\right)(x_i - T_n)^2. \tag{3.38}$$

O que permite, com base no método do ponto fixo, propor o seguinte procedimento iterativo, para obtenção da estimativa que é solução de (3.37) a partir de uma estimativa inicial $S_n^{(0)}$:

$$\left[S_n^{(k+1)}\right]^2 = \frac{1}{n\,\beta} \sum_{i=1}^{n} w^2(u_i^{(k)})(x_i - T_n)^2 \tag{3.39}$$

com

$$u_i^{(k)} = \frac{x_i - T_n}{S_n^{(k)}}.$$

Este método é equivalente ao cálculo de uma variância ponderada com pesos a variar de iteração para iteração, um processo semelhante ao proposto para obter a solução no caso do modelo de localização, fórmula (3.20).[10] Conclui-se também que o estimador-M de escala pode em alternativa ser definido pela função $w(x)$. É interessante notar que as funções w dos estimadores estudados (Huber, Tukey e MV t-Student) coincidem com as dos estimadores-M de localização com o mesmo nome (Figuras 3.3 e 3.8).

Alguns autores usam uma expressão equivalente à expressão (3.39) que é mais rápida em termos de cálculo mas que não tem a interpretação intuitiva daquela:

$$\left[S_n^{(k+1)}\right]^2 = \frac{1}{n\beta} \sum_{i=1}^{n} \chi(u_i^{(k)}) \left[S_n^{(k)}\right]^2.$$

Tal como para os estimadores de localização, tem-se que

$$\varepsilon^*(S_n^{(k)}) \leq \min\left\{\varepsilon^*(T_n); \varepsilon^*(S_n^{(0)})\right\},$$

pelo que $T_n = \text{med}(X_i)$ e $S_n^{(0)} = \text{MAD}(X_i)$ constituem novamente as escolhas mais indicadas. No entanto, contrariamente àquele caso verifica-se (Rousseeuw e Croux, 1994) que se pode ter para k finito,

$$\varepsilon^*(S_n^{(k)}) > \varepsilon^*(S_n).$$

Exemplo 3.7. Para os habituais dados do Exemplo 2.1 usou-se a função `hubers` do package `MASS` do R para determinar estimativas-M de Huber do parâmetro de escala com alguns dos valores de b da Tabela 3.4 e o valor $b = 1.5$ que é o valor que a função usa por defeito (tal como acabou de ser recomendado escolheu-se a mediana como estimativa auxiliar de localização e o MAD como estimativa inicial de escala). Os resultados obtidos foram os seguintes:

[10] Há uma pequena diferença, enquanto que no caso de localização é possível manter a consistência colocando no denominador a soma dos pesos, se se fizesse isso na expressão anterior incorrer-se-ia em subestimação. Aplica-se aqui a explicação dada no Exemplo 2.11 (página 65) a propósito do uso da constante $\gamma(\alpha)$ nas variâncias aparadas.

b	1.50	1.86	2.07	2.38	2.65
Huber(b)	0.697	0.721	0.735	0.794	0.817

Pode observar-se que para os valores mais baixos da constante de afinação ($b = 1.5$, $b = 1.86$, $b = 2.07$) se obtêm valores das estimativas próximos de 0.7, o que está de acordo com as estimativas anteriormente obtidas (ver Exemplo 2.13). No entanto para os valores mais elevados da constante de afinação observa-se um aumento das estimativas para 0.8. Levanta-se naturalmente a questão, será que afinal a estimativa mais correcta da escala para estes dados é 0.8, ou pode haver outra justificação para a aparente divergência? Para perceber melhor o que está a acontecer considerou-se o valor mais elevado $b = 2.65$ e calcularam-se as estimativas do mesmo tipo mas para os dados com $0, 1, 2, \ldots, 12$ observações iguais ao valor extremo (28.95). No primeiro caso tem-se a amostra constituída pelas 23 primeiras observações, no segundo a amostra original (valor já calculado) e nos restantes a amostra com as últimas duas, três, $\ldots$, doze observações iguais a 28.95. Os resultados obtidos apresentam-se na Tabela 3.5. Nessa tabela apresentam-se ainda as estimativas-M correspondentes a apenas 1 e 2 passos do processo iterativo, $s_n^{(1)}$ e $s_n^{(2)}$, respectivamente. O valor de ε indicado corresponde à percentagem de contaminação, ou seja (Número de *outliers*)/24 (em %).

Uma representação gráfica simples permite uma leitura clara dos resultados apresentados na tabela: subtrai-se a todas as estimativas o valor de s_n com $\varepsilon = 0$ para ter uma espécie de enviesamento devido à contaminação, e em seguida representa-se esse valor em função de ε. Obtêm-se assim os gráficos da Figura 3.10. A ideia é ter uma curva de enviesamento empírica correspondente à curva de enviesamento assintótico máximo (ver Figura 2.20). O que se pode concluir, quer da análise da tabela, quer da figura é que:

- As linhas quase verticais indicam a ocorrência de rotura do estimador, situando-se imediatamente a seguir ao ponto de rotura em dimensão finita. Nota-se aqui o efeito já várias vezes referido em relação ao ponto de rotura, ainda relativamente longe desse ponto as estimativas começam a apresentar um enviesamento elevado, apesar de permanecerem limitadas.

- Os resultados para as estimativas completamente iteradas estão de acordo com o comportamento teórico assintótico do estima-

Tabela 3.5 *Estimativas de dispersão para várias contaminações da amostra original (Exemplo 3.7), usando o estimador de Huber com* $b = 2.65$.

Nº. de *outliers*	ε	s_n	$s_n^{(1)}$	$s_n^{(2)}$
0	0	0.685	0.635	0.670
1	4.2	0.817	0.695	0.777
2	8.3	0.885	0.695	0.777
3	12.5	1.690	0.747	0.899
4	16.7	10.529	0.876	1.104
5	20.8	11.768	1.112	1.463
6	25.0	12.889	1.370	1.909
...	...	...	...	...
11	45.8	15.105	3.221	5.847
12	50.0	13.219	13.219	13.219

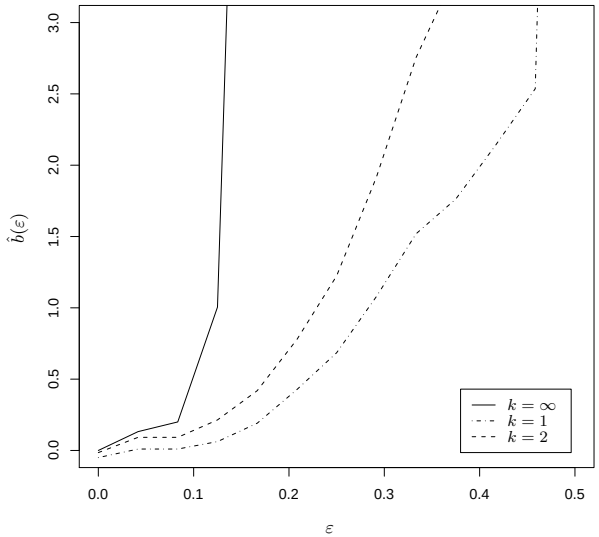


Figura 3.10 *Curvas de enviesamento empírico em função da contaminação* (ε), *usando o estimador de Huber com* $b = 2.65$ *e número de passos* k ($k = \infty$ *significa até convergência).*

dor descrito na Tabela 3.4, onde se lê que para o estimador de Huber com $b = 2.65$ o ponto de rotura é $\varepsilon^* = 14\%$ e a sensibildade a grandes erros é $\gamma^* = 3.25$. Da curva de enviesamento empírica estima-se um valor da sensibilidade ("derivada" de $\hat{b}(\varepsilon)$ em $\varepsilon = 0$) de $24\times(0.817-0.685) = 3.17$, muito próximo daquele.

- Globalmente as estimativas a um passo têm o melhor comportamento. O ponto de rotura aproxima-se de 50% e as sensibilidades estimadas como no ponto anterior são dadas por 1.44 e 2.57, respectivamente para 1 e 2 passos.
- A grande eficiência pode ser enganadora porque além de menor ponto de rotura acarreta elevada sensibilidade e faz com que o enviesamento devido a uma pequena percentagem de contaminação possa não ser desprezável.
- É preferível usar um estimador-M de escala a um ou dois passos do que a versão completamente iterada. Pode haver uma ligeira perda de eficiência (segundo Rousseeuw e Croux, 1994, essa perda de eficiência é de facto ligeira) mas os ganhos em termos de redução da sensibilidade e do aumento do ponto de rotura são consideráveis.

De tudo o que se viu até aqui, em especial dos valores apresentados na Tabela 3.4 e dos resultados do último exemplo, conclui-se em primeiro lugar que os estimadores-M do parâmetro de escala são mais problemáticos que os estimadores-M de localização, em particular parece impossível conciliar elevada eficiência com elevado ponto de rotura e baixa sensibilidade. Será mesmo impossível? Teoricamente, a resposta é negativa. Croux (1994) construiu, a partir do estimador de Huber geral com função χ dada por (3.35), uma função χ não trivial, que conduz a um estimador-M de escala com ponto de rotura de 50%, independentemente das constantes de afinação, as quais podem ser calibradas para uma eficiência arbitrariamente alta ($\rightarrow 1$) sob o modelo normal. Só que o próprio autor não recomenda esses estimadores pois acabam por não ter melhores propriedades em termos da sensibilidade a grandes erros e da sensibilidade local.

O que se recomenda em termos práticos é a utilização dos estimadores a um passo que, como se viu no exemplo, têm um comportamento muito melhor em termos de robustez que as versões completamente iteradas. Há ainda outra possibilidade que consiste em escolher sem-

pre a constante que conduz ao maior ponto de rotura, de acordo com a Tabela 3.4. Esta estratégia é especialmente adequada quando a estimação da escala não é o interesse principal da análise e é, por exemplo, a seguida para estimar o erro padrão dos resíduos em alguns métodos de regressão robusta (tais como a regressão-S e a regressão-MM, descritos no Capítulo 4). Em termos de escolha da função concreta para estes efeitos parece actualmente ser bastante popular a função χ de Tukey (o que, recorde-se, já acontecia em relação ao modelo de localização).

3.2.4 Situações multivariadas e multiparamétricas

Os estimadores-M do parâmetro vectorial $\boldsymbol{\theta}$ contendo p parâmetros (isto é, pretende-se estimar simultaneamente mais do que um parâmetro, podendo as observações ser relativas a um modelo uni ou multivariado) são definidos do mesmo modo que no caso uniparamétrico tratado até agora.

Definição 3.8. *Um estimador-M é um estimador,* $\mathbf{T}_n$*, que minimiza*

$$\sum_{i=1}^{n} \rho(\mathbf{X}_i, \mathbf{T}_n), \tag{3.40}$$

onde ρ *é uma função arbitrária,* $\rho : \boldsymbol{\Omega} \times \boldsymbol{\Theta} \longrightarrow \mathbb{R}$*. Chama-se também estimador-M a um estimador,* $\mathbf{T}_n$*, que seja solução da equação vectorial*

$$\sum_{i=1}^{n} \psi(\mathbf{X}_i, \mathbf{T}_n) = \mathbf{0} \tag{3.41}$$

com $\psi : \boldsymbol{\Omega} \times \boldsymbol{\Theta} \longrightarrow \mathbb{R}^p$*.*

O estimador de máxima verosimilhança corresponde ao caso particular $\rho(\mathbf{x}, \boldsymbol{\theta}) = -\log f(\mathbf{x}, \boldsymbol{\theta})$ e

$$\psi(\mathbf{x}, \boldsymbol{\theta}) = -s(\mathbf{x}, \boldsymbol{\theta}) = -\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\mathbf{x}).$$

Note-se que se um estimador-M for definido por (3.40) através duma

função ρ diferenciável, então a definição por (3.41), com

$$\psi(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial \rho(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

resulta equivalente nos casos regulares de estimação. No entanto a definição por (3.41) é mais geral pois as componentes de ψ não são necessariamente as derivadas parciais duma função ρ. Tal como no caso unidimensional usar-se-á daqui em diante a definição através da função ψ. Esta função pode ser multiplicada à esquerda por uma matriz não singular que não dependa de $\mathbf{x}$ sem que o estimador seja alterado.

A função de influência do funcional equivalente ao estimador-M, definido por

$$\int \psi(\mathbf{x}, \mathbf{T}(F))\, dF(\mathbf{x}) = \mathbf{0},$$

pode ser deduzida da mesma forma que (3.6), obtendo-se

$$IF(\mathbf{x}; \mathbf{T}, F) = \mathbf{M}(\psi, F)^{-1} \psi(\mathbf{x}, \mathbf{T}(F)), \tag{3.42}$$

com a matriz $\mathbf{M}$ $(p \times p)$ definida por

$$\mathbf{M}(\psi, F) = -\int \left[\frac{\partial}{\partial \boldsymbol{\theta}} \psi(\mathbf{y}, \boldsymbol{\theta}) \right]_{\boldsymbol{\theta} = \mathbf{T}(F)} dF(\mathbf{y}). \tag{3.43}$$

Como aplicação da situação multiparamétrica trata-se a seguir resumidamente o modelo de localização e escala, generalização natural dos modelos de localização e de escala tratados nas secções anteriores.

3.2.5 Modelo de localização e escala

Parece natural juntar numa mesma definição os modelos de localização e de escala. Assim, recordando a forma como foram introduzidos esses modelos, pode-se pensar numa variável aleatória arbitrária X, e considerar uma transformação linear envolvendo simultaneamente uma translação e uma mudança de escala, $X_{\mu,\sigma} = \sigma X + \mu$. Constrói-se assim um modelo em que a função de distribuição é dada por

$$F_{X_{\mu,\sigma}}(x) = P(\sigma X + \mu \leq x) = P\left(X \leq \frac{x - \mu}{\sigma}\right) = F_X\left(\frac{x - \mu}{\sigma}\right).$$

Se a variável aleatória X for contínua o modelo tem densidade dada por

$$f_{X_{\mu,\sigma}}(x) = \frac{1}{\sigma} f_X\left(\frac{x-\mu}{\sigma}\right).$$

O exemplo que surge imediatamente é o modelo normal, que já é um modelo desta forma com os parâmetros habituais. Outro exemplo é dado pelo modelo de $Laplace(\mu, b)$ (definido no Exemplo 2.7, página 50), em que μ é o parâmetro de localização e b é o parâmetro de escala. A distribuição exponencial deslocada, com densidade dada por (3.8) corresponde também a um modelo de localização e escala, em que a é o parâmetro de localização e β é o parâmetro de escala. A definição seguinte formaliza o que se acabou de exemplificar.

Definição 3.9. *Um modelo de localização e escala, com parâmetro de localização μ e parâmetro de escala σ, consiste numa família de distribuições $\{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, com $\boldsymbol{\theta} = (\mu, \sigma)^T$ e $\boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^+$, tal que*

$$F_{\boldsymbol{\theta}}(x) = F\left(\frac{x-\mu}{\sigma}\right),$$

onde $F = F_{0,1}$ representa uma função de distribuição univariada genérica que define o tipo ou família do modelo.

Note-se que, apesar da notação, μ e σ não coincidem necessariamente com o valor esperado e desvio padrão, que até podem não existir. Basta, por exemplo, pensar no modelo de localização e escala baseado na distribuição de Cauchy, que é um modelo com densidade dada por

$$f_{\mu,\sigma}(x) = \frac{1}{\pi\,\sigma}\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}.$$

Por comparação com o estimador de máxima verosimilhança, um estimador-M de $\boldsymbol{\theta}$ com as propriedades de invariância e equivariância usualmente requeridas[11] deverá ser definido por uma função vectorial ψ, com duas componentes, da forma

$$\psi(x, \mu, \sigma) = \psi(z) = (\psi_1(z); \psi_2(z))^T, \quad \text{com} \quad z = \frac{x-\mu}{\sigma}.$$

[11] Equivarìância do estimador de μ em relação à localização e à escala, de acordo com as Definições 3.3 e 3.4; equivariância do estimador de σ em relação à escala e invariância em relação à localização, de acordo com as Definições 3.6 e 3.7.

Se F for uma distribuição simétrica é natural escolher ψ_1 ímpar e ψ_2 par. Como casos particulares considerem-se os seguintes:

(i) Os estimadores "clássicos" de μ e σ sob o modelo normal, respectivamente, média e desvio padrão amostrais, são estimadores-M (simultâneos) definidos por:

$$\psi(z) = \begin{pmatrix} z \\ z^2 - 1 \end{pmatrix}.$$

(ii) Outro par de estimadores é formado pela mediana e desvio absoluto mediano (amostrais) definidos por

$$\psi(z) = \begin{pmatrix} \text{sinal}(z) \\ \text{sinal}(|z| - \Phi^{-1}(\frac{3}{4})). \end{pmatrix}$$

Estes estimadores são também consistentes para os parâmetros do modelo normal e são os estimadores mais B-robustos dos parâmetros daquele modelo, como se viu atrás.

(iii) Como (3.41) conduz, nos casos anteriores, a soluções explícitas, cada um dos estimadores anteriores pode ser obtido considerando separadamente os modelos de localização e de escala. O mesmo já não acontece com o chamado estimador de Huber de localização e escala definido por

$$\psi(z) = \begin{pmatrix} \psi_b(z) \\ \psi_c(z)^2 - \beta(c) \end{pmatrix}.$$

Este estimador corresponde a uma combinação dos resultados minimax obtidos independentemente para os modelos de localização e de escala. Ao contrário dos estimadores (i) e (ii) as estimativas têm de ser obtidas por processos iterativos. Usualmente é considerada uma versão simplificada do estimador em que $c = b$, conhecida como Huber-proposta 2. Para esse estimador o ponto de rotura global (tanto para a componente de localização como para a de escala) é dado por (Huber, 1981, p. 143),

$$\varepsilon^* = \frac{\beta(b)}{\beta(b) + b^2}, \tag{3.44}$$

onde $\beta(b)$ é calculado por (3.34). Este valor é inferior ao ponto de rotura quer do estimador-M de localização quer do estimador-M de escala quando considerados isoladamente (respectivamente, 50% e $\beta(b)/b^2$).

O procedimento iterativo para cálculo das estimativas pode obter-se por combinação dos métodos utilizados para os estimadores separados, expressões (3.20) e (3.39), mas em que agora tanto T_n como S_n são actualizados em cada iteração. Dadas estimativas iniciais, $T_n^{(0)}$ e $S_n^{(0)}$ (pelas razões já anteriormente apontadas os pontos de rotura de $T_n^{(0)}$ e $S_n^{(0)}$ devem ser máximos), para as sucessivas iterações faz-se

$$T_n^{(k+1)} = \frac{\sum_{i=1}^n x_i w(u_i^{(k)})}{\sum_{i=1}^n w(u_i^{(k)})}$$

e

$$\left[S_n^{(k+1)}\right]^2 = \frac{\sum_{i=1}^n w^2(u_i^{(k)})(x_i - T_n^{(k)})^2}{n\beta},$$

com

$$u_i^{(k)} = \frac{x_i - T_n^{(k)}}{S_n^{(k)}}$$

e

$$w(u) = \begin{cases} 1, & |u| \leq b \\ \dfrac{b}{|u|}, & |u| > b \end{cases}.$$

Como este procedimento é pior em termos de ponto de rotura que a estimação separada de localização e escala (sendo equivalente nos restantes aspectos, eficiência e sensibilidade), não é actualmente recomendado.

3.3 Breve referência a outras classes de estimadores

Apesar de os estimadores-M serem os mais conhecidos e importantes entre os métodos de estimação robusta, existem muitas outras classes de estimadores com propriedades de robustez. Nesta secção faz-se uma descrição breve das principais.

Estimadores-L

Chamam-se estimadores-L aos estimadores que são combinações lineares de estatísticas de ordem. Exemplos de estimadores deste tipo são

a média aritmética, as médias aparadas e a mediana (estimadores de localização) ou a amplitude inter-quartis (estimador de escala). Alargando um pouco a definição para uma função, não necessariamente linear, das estatísticas de ordem abarca-se também os desvios padrões aparados.

Como se viu ao longo das secções anteriores, os estimadores mencionados no parágrafo anterior ou são equivalentes a um estimador-M (média e mediana) ou existe um estimador-M que lhe é assintoticamente equivalente mas mais robusto no sentido de ter um ponto de rotura muito mais elevado (o estimador de Huber de localização para as médias aparadas, o estimador de Huber de escala para os desvios padrões aparados, o desvio absoluto mediano para a amplitude inter-quartis).

É uma conclusão geral que nesta classe de estimadores não existem alternativas credíveis aos estimadores-M. Acresce ainda que a generalização a problemas com mais do que uma variável (regressão ou análise multivariada) não é directa, como se percebe pelas dificuldades em estabelecer relações de ordem com mais do que uma variável (há no entanto algumas tentativas nesse sentido como, por exemplo, Fraiman *et al.*, 1999).

Estimadores-R

Os estimadores-R surgiram no contexto não paramétrico de estimação da localização e devem o seu nome ao facto de estarem relacionados com testes baseados em postos (*rank tests*). O primeiro estimador deste tipo foi proposto por Hodges e Lehmann (1963) e é actualmente conhecido por estimador de Hodges-Lehmann. Esse estimador foi obtido considerando o teste de Wilcoxon e é definido como a mediana das médias de todos os pares de observações. Este estimador tem função de influência limitada, ponto de rotura de cerca de 29% e pode ser visto como uma espécie de regularização da mediana.

Não existe uma ligação directa entre os testes *rank* e a estimação de escala, no entanto o estimador Q_n de escala (ver Exemplo 2.12) foi inspirado no estimador de Hodges-Lehmann.

Para este tipo de estimadores registam-se o mesmo tipo de dificuldades de generalização que em relação aos estimadores-L.

Estimadores-A

Os estimadores-A (Lax, 1985) são estimadores de escala baseados na constatação de que a variância assintótica de um estimador de localização é igual ao quadrado de um parâmetro de escala. Por exemplo, para a média amostral a variância assintótica é o quadrado do desvio padrão. Então, estimando a variância assintótica de um estimador de localização e tomando a raiz quadrada obtém-se uma estimativa de escala. É evidente que com a média amostral isto não conduz a nenhum resultado novo. Mas o que Lax (1985) propôs foi um estimador baseado na expressão (3.12) que dá a variância assintótica de um estimador-M de localização. A ideia é então aplicar aquela expressão a uma amostra finita e tomar a raiz quadrada. Para esse efeito as observações são em primeiro lugar estandardizadas,

$$u_i = \frac{x_i - \text{med}(x_j)}{\text{MAD}(x_j)},$$

e em seguida faz-se

$$S_n = \text{MAD}(x_j) \left[\frac{1}{n}\sum_{i=1}^{n} \psi^2(u_i)\right]^{1/2} \Bigg/ \left|\frac{1}{n}\sum_{i=1}^{n} \psi'(u_i)\right|. \qquad (3.45)$$

Estes estimadores fornecem uma forma natural de estimar a escala quando se estima a localização através de um estimador-M. Em consequência, deve escolher-se a mesma função ψ (incluindo a constante ou constantes de afinação). Como se referiu no final da Secção 3.2.2 uma escolha popular e com boas propriedades é a função biponderada de Tukey com uma das constantes da Tabela 3.1.

Estimadores-D

A ideia dos estimadores-D (D de "distância mínima") remonta a Wolfowitz (1957). A definição destes estimadores recorre a uma distância entre a função de distribuição empírica e a distribuição teórica, $\pi(G_n, F_\theta)$, e diz que a estimativa de θ é o valor que minimiza aquela distância. π pode ser por exemplo a distância de Kolmogorov ou uma outra qualquer distância obtida a partir de um teste de ajustamento. O princípio da distância mínima enquadra-se nos métodos

não-paramétricos e generaliza-se bem a muitos modelos. Estes estimadores têm continuado a ser estudados mais no âmbito não paramétrico do que no contexto da teoria da robustez, talvez devido à observação (Hampel *et al.*, 1986) de que muitos dos estimadores resultantes da aplicação deste critério acabam por poder ser considerados casos especiais de estimadores-M.

Estimadores-P

Os estimadores-P são generalizações dos estimadores de Pitman (Pitman, 1937). Estes estimadores ao contrário dos estimadores-M têm uma forma explícita, não necessitando de procedimentos iterativos, contudo essa forma explícita envolve integração numérica o que também não os torna muito populares. Tal como o anterior este método é bastante geral e tem continuado a ser objecto de investigação fora do âmbito da teoria da robustez (cita-se a título de exemplo Chaturvedi e Shalabh, 2004).

Estimadores-S

Genericamente chama-se estimador-S a um estimador que minimiza uma estimativa de escala conveniente. No caso de se querer estimar um parâmetro de localização, θ, a partir de uma amostra univariada $x_1, \ldots, x_n$, a definição é

$$\hat{\theta} \text{ minimiza } S(x_1 - \theta, \ldots, x_n - \theta).$$

Note-se que se S representar o desvio padrão usual se obtém $\hat{\theta} = \overline{x}$, enquanto que se S representar o desvio médio se obtém a mediana. É muito comum escolher para S o estimador-M de escala que usa a função biponderada de Tukey e tem ponto de rotura 1/2 (ver Tabela 3.4).

Os estimadores-S foram originalmente propostos por Rousseeuw e Yohai (1984) no contexto do modelo de regressão e como tal serão descritos mais detalhadamente no Capítulo 4. Estes estimadores são também facilmente generalizáveis ao caso multivariado (Rousseeuw e Leroy, 1987; Davies, 1987).

Estimadores-W

Os estimadores-W consistem em versões ponderadas dos estimadores usuais, por exemplo da média ou da variância. A expressão é atribuída a Tukey (1977). Viu-se já, por exemplo em (3.20), que se os pesos forem sendo actualizados sucessivamente se obtém um estimador-M. Se se realizar a ponderação apenas uma vez então obtém-se um estimador-M a um passo, também designado por estimador-w.

Estimadores-τ

Os estimadores-τ foram propostos por Yohai e Zamar (1988) no contexto do modelo de escala para resolver o conflito existente entre ponto de rotura e eficiência para os estimadores-M de escala (e que foi detalhadamente analisado no final da Secção 3.2.3). A ideia acaba por ser simples e consiste em tomar primeiro um estimador-M com o máximo ponto de rotura e em seguida, usando esse como ponto de partida, calcular um estimador-M com elevada eficiência mas apenas a um passo. O estimador esultante herda o ponto de rotura do primeiro e a eficiência do segundo. Esta ideia foi também generalizada aos estimadores multivariados (Lopuhaä, 1991).

3.4 Para além da estimação pontual

Até este ponto, e o texto já vai longo, só se falou de estimação pontual. Este facto reflecte de certo modo a história da estatística robusta. Começou por se dar toda a atenção à estimação pontual e só mais recentemente foram abordadas seriamente as questões relativas à estimação da variabilidade dos estimadores (erros padrão), à construção de intervalos de confiança e de testes de hipóteses. Todas estas questões se resolvem facilmente uma vez conhecida a distribuição dos estimadores. Percebe-se a dificuldade do problema, se muitos estimadores não têm forma explícita o que dizer da sua distribuição? Descrevem-se em seguida três estratégias que têm sido seguidas para resolver este problema, e que consistem em: (i) uso de distribuições assintóticas, (ii) uso de outras aproximações e (iii) uso de métodos de reamostragem (*jackknife* e *bootstrap*).

Uma vez resolvidas estas dificuldades, ou seja, estimado o erro padrão de uma estimativa robusta, construído o intervalo de confiança respectivo ou efectuado um teste, é ainda preciso responder às questões: será o resultado obtido ele próprio robusto? Em que sentido deve ser entendida a robustez destes procedimentos?

3.4.1 Distribuições assintóticas

Como se viu no Capítulo 2, se um estimador T_n de um parâmetro $\theta \equiv T(F)$ tiver função de influência (e verificar condições de regularidade), então pela propriedade (**P3**) pode escrever-se que

$$\frac{T_n - \theta}{\sqrt{V(T,F)/n}} \overset{a}{\sim} \mathcal{N}(0,1), \tag{3.46}$$

onde $V(T,F)$ representa a variância assintótica dada pela fórmula (2.18). Para que este resultado seja de alguma utilidade prática, ou seja para que possa ser usado como variável fulcral para a construção de intervalos de confiança ou transformado em estatística de teste, é necessário que se verifiquem duas condições: (i) seja possível estimar $V(T,F)$ de modo a que a distribuição fique completamente especificada a menos do parâmetro desconhecido θ,[12] (ii) a aproximação seja razoável a partir de valores relativamente baixos de n.

Quanto à estimação de $V(T,F) = \int IF(x;T,F)^2 dF(x)$, ela é possível e podem seguir-se várias vias:

- A primeira consiste em recorrer à aproximação da função de influência dada pela curva de sensibilidade, Definição 2.1, e fazer

$$\widehat{V(T,F)} = V_n = \frac{1}{n}\sum_{i=1}^{n} \left[SC(x_i,T)\right]^2, \tag{3.47}$$

  o que pode ser considerado uma espécie de estimativa não paramétrica pois o modelo central F não precisa de ser especificado. Não é conveniente usar esta estimativa quando o estimador é irregular, o que está geralmente associado à existência de descontinuidades na função de influência, pois sabe-se que nesses

[12] Pode então invocar-se o Teorema de Slutsky para usar o resultado (3.46) com $V(T,F)$ substituída por $\widehat{V(T,F)}$.

casos a curva de sensibilidade não converge para a função de influência. É o que acontece por exemplo com a mediana e o MAD (ver Exemplos 2.7 e 2.10, respectivamente). Como ilustração aplique-se este método com o estimador média amostral, como $SC(x; \overline{X}) = x - \overline{x}$, obtém-se

$$V_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2,$$

o que conduz, como se sabe, a uma aproximação razoável da distribuição assintótica.

Apesar de muito simples e de funcionar bem em certos casos, não é muito recomendável o uso indiscriminado da expressão (3.47) por razões que ficarão claras um pouco mais adiante quando se falar do método *jackknife*.

- A segunda consiste em especificar o modelo central F e avaliar numericamente o integral do quadrado da função de influência através de

$$\widehat{V(T,F)} = \frac{1}{n} \sum_{i=1}^{n} \left[IF(x_i; T, F) \right]^2 .$$

Neste caso pode acontecer que surjam parâmetros/funcionais desconhecidos que é preciso estimar. Veja-se novamente o que acontece com a média amostral (Exemplo 2.5),

$$\widehat{V(T,F)} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\mu(F)})^2,$$

o que coincide com o resultado obtido usando a curva de sensibilidade.

- A terceira consiste em calcular analiticamente o integral do quadrado da função de influência sob o modelo central F. Se tal for possível obtém-se então uma expressão, geralmente envolvendo parâmetros desconhecidos que devem em seguida ser convenientemente estimados. Exemplificando novamente com a média aritmética conclui-se que este método conduz a $\widehat{V(T,F)} = \widehat{\sigma^2(F)}$, que é o resultado geralmente utilizado.

Em relação às duas últimas modalidades é importante salientar que, para que o resultado seja robusto, os estimadores adicionais que é preciso utilizar devem eles próprios ser robustos.

Como ilustração menos trivial que a média aritmética considere-se um estimador-M de localização com função ψ genérica e recordem-se as expressões obtidas na Secção 3.2.2 para a função de influência e variância assintótica, respectivamente (3.11) e (3.12). A primeira via é de aplicação directa mas só deve ser usada se a função ψ não tiver descontinuidades. A segunda conduz a uma expressão do tipo da expressão (3.45) do estimador-A de escala em que em vez do MAD pode figurar outro qualquer estimador robusto de escala A_n (auxiliar) e em vez da mediana deve figurar o estimador de localização em causa, T_n. Finalmente pela terceira via obtém-se

$$\widehat{V(T,F)} = A_n \left[\frac{1}{n}\sum_{i=1}^{n}\psi^2(u_i)\right]^{1/2} \Big/ M,$$

com

$$u_i = \frac{x_i - T_n}{A_n},$$

Exemplo 3.8. Para perceber se a utilização destes três métodos de estimação da variância assintótica dão resultados muito diferentes eles foram usados para estimar o erro padrão ($ep(T_n) = \sqrt{V(T,F)/n}$) associado à estimativa de localização obtida para os dados do Exemplo 2.1 com o estimador de Huber com constante de afinação $b = 1.35$, para o qual $\hat{\mu} = 3.22$. Usando a fórmula baseada na curva de sensibilidade obtém-se $ep(\hat{\mu}) = 0.12$. A segunda e a terceira fórmula conduzem ambas a $ep(\hat{\mu}) = 0.14$. Há uma ligeira diferença entre os valores que se julga não ser muito relevante, no entanto como se disse o método baseado na curva de sensibilidade não é muito fiável.

Os intervalos de confiança aproximados a um nível de confiança de $100 \times (1-\alpha)\%$ podem agora obter-se fazendo

$$IC_{100\times(1-\alpha)\%}(\mu) \simeq \left(\hat{\mu} - ep(\hat{\mu}) \times z_{1-\alpha/2}; \hat{\mu} + ep(\hat{\mu}) \times z_{1-\alpha/2}\right),$$

com $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$, o que dá para 95% de confiança, respectivamente, $(2.98, 3.46)$ e $(2.94, 3.49)$. O intervalo de confiança a 95%, baseado na média amostral e com quantis da distribuição normal, é igual a $(2.16, 6.40)$, com todas as observações, e a $(2.93, 3.48)$, sem a observação 24. Se em vez dos quantis da distribuição normal se

usarem os da distribuição t_{23} obtêm-se os intervalos $(2.04, 6.52)$, com todas as observações, e $(2.92, 3.50)$, sem a observação 24, com comprimentos cerca de 5% maiores. O que se conclui é que os intervalos obtidos usando o estimador robusto parecem muito mais razoáveis em face da existência do *outlier*.

Quanto à condição (ii), é difícil dar uma resposta única. Mas segundo Hampel (2000) a aproximação pode ser razoável a partir de n tão reduzido como 10! Convém acrescentar que a qualidade da aproximação depende essencialmente do tipo de estimador/parâmetro em causa, sendo a situação mais propícia para atingir uma boa aproximação a da estimação univariada de localização. Para que a aproximação seja razoável num problema de estimação univariada de escala já é necessário ter amostras com dimensões bem superiores, da ordem de pelo menos 30 observações. Esse número pode ser ainda maior noutros casos (por exemplo em relação ao coeficiente de correlação sabe-se que é necessário um número de observações da ordem de várias centenas).

3.4.2 Correcções e outras aproximações

Quando o número de observações é reduzido, não há resultados exactos e não é aconselhável usar resultados assintóticos procura-se encontrar métodos também aproximados mas que entrem em conta com a dimensão da amostra.

Uma correcção muito simples que pode melhorar os resultados sem contudo os alterar substancialmente, consiste em usar para os testes e intervalos de confiança, por analogia com o resultado clássico, a aproximação baseada em

$$\frac{T_n - \theta}{\sqrt{\widehat{V(T,F)}/n}} \overset{a}{\sim} t_{n-1}. \tag{3.48}$$

Outra possibilidade de correcção é ao nível das estimativas do erro padrão, podendo ser aí introduzidas correcções dependentes da dimensão da amostra e que melhoram a qualidade da aproximação para dimensões baixas. Uma correcção elementar deste tipo consiste em substituir em todas as fórmulas apresentadas na Secção 3.4.1 o denominador n por $n - 1$.

Outras correcções mais sofisticadas podem ser conseguidas usando métodos importados de outras áreas como é o caso da econometria (ver por exemplo Croux *et al.*, 2004).

Por fim é ainda possível recorrer a técnicas genéricas para obter os chamados resultados assintóticos para pequenas amostras.[13] Por exemplo Field e Hampel (1982) obtêm, com base nos métodos de ponto de sela, distribuições aproximadas dos estimadores-M de localização que são muito boas a partir de $n = 3$. É no entanto de referir que este tipo de resultados é de difícil utilização e não tem tido muito sucesso na prática.

3.4.3 *Jackknife* e *bootstrap*

Numa definição algo simplista pode dizer-se que as técnicas de reamostragem consistem no cálculo repetido de estimativas, um número elevado de vezes. Trata-se no fundo de métodos, em geral não paramétricos, de estimação da distribuição amostral do estimador correspondente ou de algumas das suas características. O *jackknife* já referido no capítulo anterior a propósito da função de influência e o *bootstrap* (Efron, 1979) são talvez os procedimentos de reamostragem mais famosos.[14]

Em face da descrição poder-se-ia concluir que os métodos de reamostragem podem duma forma fácil e eficaz resolver o problema da determinação da distribuição de um estimador robusto e, em consequência, as questões levantadas no início desta secção (estimação do erro padrão, intervalos de confiança e testes de hipóteses). Infelizmente, e talvez espantosamente, tal não é verdade. De facto certos métodos de reamostragem, em particular o *jackknife* e o *bootstrap*, não funcionam bem e não devem ser usados quando há *outliers* nos dados. Ora essa é precisamente a situação em que se devem usar estimadores robustos! Existe portanto um conflito entre as duas metodologias que tem como consequência a perda de propriedades do *jackknife* e do *bootstrap* ordinários quando aplicados a estimadores robustos e a dados com *outliers*. Aparentemente Efron teve consciência disto quando

[13] *Small sample asymptotics*, baseados em métodos do tipo aproximações empíricas de ponto de sela ou em expansões de Edgeworth.

[14] Outros procedimentos de reamostragem são as permutações, utilizadas no exemplo de aplicação do Capítulo 5, e a validação cruzada.

propôs técnicas de diagnóstico para o *bootstrap* (Efron, 1992) mas essas técnicas só funcionam se houver um único *outlier*. Posteriormente surgiram avisos mais claros sobre os perigos da utilização dos métodos de reamostragem para estimar a variabilidade de estimadores robustos e para realizar outras inferências e foram propostas algumas medidas correctivas (Stronberg, 1997; Singh, 1998).[15] É de facto possível fazer com que o *bootstrap* e o *jackknife* funcionem adequadamente com estimadores robustos e na presença de *outliers* mas à custa de se perder a simplicidade que constituía um dos atractivos principais destes métodos (Salibian-Barrera, 2000, 2003; Amado, 2003; Amado e Pires, 2004).

3.4.4 Robustez de intervalos de confiança e testes de hipóteses

Após esta breve descrição de métodos que podem ser usados para construir intervalos de confiança ou efectuar testes de hipóteses após a estimação pontual com estimadores robustos, importa ainda responder às questões colocadas no início da Secção 3.4:

- Em que sentido deve ser entendida a robustez destes procedimentos?
  e
- Será o resultado obtido ele próprio robusto?

Dada a dualidade existente entre testes de hipóteses e intervalos de confiança a explicação concentra-se nestes últimos.

A robustez dos intervalos de confiança deve ser averiguada, em face da não verificação das hipóteses subjacentes a um dado modelo paramétrico, sob dois pontos de vista:

Robustez de validade: manutenção do nível de confiança (em termos de testes é equivalente à manutenção do nível de significância);

Robustez de eficiência: manutenção do comprimento esperado do intervalo em valores aceitáveis e se possível próximo do mínimo

[15] Dado que a curva de sensibilidade é uma espécie de *jackknife*, este facto justifica a afirmação feita na Secção 3.4.1 de que as estimativas do erro padrão baseadas na curva de sensibilidade não são fiáveis.

(em termos de testes é equivalente à manutenção da potência e se possível próximo da potência máxima). [16]

Admitindo que os intervalos de confiança aleatórios são da forma geral

$$IC_{100\times(1-\alpha)\%}(\theta) = \left(\hat{\theta} - ep(\hat{\theta}) \times z_{1-\alpha/2}; \hat{\theta} + ep(\hat{\theta}) \times z_{1-\alpha/2}\right),$$

vê-se que a manutenção do nível de confiança é assegurada se $\hat{\theta}$ "estiver próximo" de θ (ou seja se for centrado ou assintoticamente centrado), se $\sqrt{n}\, ep(\hat{\theta})$ for maior do que $V(T,F)$ e se os quantis $z_{1-\alpha/2}$ utilizados forem superiores aos valores correctos. Por outro lado o comprimento do intervalo mantém-se em níveis aceitáveis se $\sqrt{n}\, ep(\hat{\theta}) = \widehat{V(T,F)}$ corresponder a um estimador robusto.

Percebe-se que, embora se possam a partir desta análise tirar linhas de orientação gerais (por exemplo, devem ser usados estimadores robustos de θ e de $V(T,F)$), o estudo teórico da robustez de intervalos de confiança é muito complexo. Assim não admira que em geral este assunto seja analisado recorrendo a estudos de simulação. Um dos primeiros estudos desse tipo foi realizado por Gross (1976) que avaliou a robustez, quer em termos de validade quer em termos de eficiência, dos intervalos relativos ao parâmetro de localização em populações univariadas simétricas com caudas pesadas. Para o efeito, utilizou vários estimadores robustos e concluiu que, para as situações estudadas, os melhores desempenhos foram obtidos para um estimador-M com função ψ de Tukey. Para além disso, também concluiu que os estimadores baseados em *jackknife* não tiveram um desempenho satisfatório (mesmo o relacionado com o estimador de Tukey).

Hampel (2000) chama a atenção para uma outra questão importante e muitas vezes ignorada que tem a ver com o efeito do enviesamento do estimador ($\hat{\theta}$) em grandes amostras. Enquanto que o erro padrão tende para zero quando a dimensão da amostra aumenta, o mesmo pode não acontecer com o enviesamento (por exemplo devido a contaminações assimétricas ou a erros sistemáticos), o que implica que a probabilidade de cobertura tenda para zero, por mais robusto que seja o método utilizado. Citando Hampel (2000) "*this*

[16] É esta dualidade de critérios que é responsável pela confusão existente na literatura relativamente, por exemplo, à robustez dos testes-t, e já referida na Secção 1.4.

implies that it does not make sense to collect too many bad data for the same information". Branco e Pires (2007) vão mesmo um pouco mais longe analisando os perigos relacionados com a elevada dimensão das amostras (não necessariamente más) no contexto geral dos testes de hipóteses.

Após a visão geral sobre os fundamentos da estatística robusta, dada neste capítulo e no anterior, considera-se no capítulo seguinte a análise sob o ponto de vista da robustez de um dos modelos mais importantes da estatística ao nível das aplicações: o modelo de regressão.

4
Regressão

4.1 Introdução

Quando se fala do modelo de regressão admite-se quase automaticamente que a estimação dos parâmetros do modelo é feita com base no princípio dos mínimos quadrados. Tal é no fundo consequência de uma longa tradição do uso deste princípio, nascido nos finais do século XVIII (1795), segundo Gauss, ou nos princípios do século XIX (1805), segundo Legendre, e que viria a dominar o cenário da actividade estatística. No longo percurso dos mínimos quadrados os seus méritos foram sempre mais distinguidos do que os seus defeitos, mas por volta dos anos 70, quando o interesse pela estatística robusta se manifestou de forma inequívoca, os mínimos quadrados não foram poupados à denúncia da sua acção, por vezes desastrosa, como método de estimação em regressão, o que veio favorecer definitivamente o desenvolvimento da regressão robusta. De facto a regressão robusta propõe-se enfrentar e remover os males de que a estimação pelos mínimos quadrados enferma, principalmente quando em presença de *outliers* e erros não normais. Sendo a regressão um dos primeiros modelos que a estatística pôs ao serviço dos seus utilizadores e certamente um dos mais requeridos na análise de problemas práticos, a proposta robusta, com vista à afinação e melhoria da eficácia do modelo, constitui um acontecimento de grande relevo na longa história da regressão. A sua divulgação está feita nos principais livros de robustez já mencionados no Capítulo 1 e mais recentemente a regressão robusta passou a figurar em livros sobre regressão, como Draper e Smith (1998) e Ryan (1997), e também em livros de índole mais geral, como Neter *et al.* (1996). Um tratamento bastante completo da regressão robusta encontra-se em Rousseeuw e Leroy (1987). Esta

disponibilização, juntamente com a crescente oferta de *software* para regressão robusta, sobretudo em R e S-Plus, está a contribuir para que a regressão robusta esteja a ser usada por um número de utilizadores cada vez maior.

Neste capítulo discutem-se os perigos a que pode levar o uso cego ou automático do método dos mínimos quadrados, dando relevo às vantagens que a prática da regressão robusta tem sobre este método tradicional. Em seguida apresentam-se alguns dos vários métodos robustos para regressão e um sumário das propriedades mais relevantes desses métodos. Para ilustrar melhor os conceitos é dada ênfase à regressão simples, mas a regressão múltipla é também contemplada. O capítulo termina com a análise de dois conjuntos de dados reais e uma apreciação global do papel da regressão robusta.

4.2 Méritos e defeitos do método dos mínimos quadrados

O modelo clássico de regressão linear múltipla assume a existência da relação linear

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.1}$$

entre um conjunto de observações de uma variável resposta (Y) e de $p-1$ variáveis explicativas $(X_1, \ldots, X_{p-1})$, onde n é o número de indivíduos ou objectos observados, ε_i é o erro associado à resposta y_i e $\beta_0, \beta_1, \ldots, \beta_{p-1}$ são os parâmetros da relação.

Por ser conveniente para o estudo global do modelo, a equação (4.1) pode escrever-se na forma matricial,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.2}$$

onde

$$\mathbf{y} = (y_1, \ldots, y_n)^T$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$$

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$$

e

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix}$$

é a chamada matriz de especificação das variáveis explicativas.

O caso em que $p-1=1$, $y_i = \beta_0+\beta_1 x_i+\varepsilon_i$, designa-se por regressão linear simples e tem a vantagem de permitir antever e compreender de forma fácil e muito clara o funcionamento e grande parte das propriedades do modelo geral representado pela equação (4.1).

São conhecidos os objectivos do estudo de regressão, identificação e descrição da relação, calibração, previsão e predição, mas quaisquer que eles sejam o primeiro passo do estudo é estimar o vector $\boldsymbol{\beta}$ e a própria equação de regressão, isto é,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{p-1} x_{i,p-1},$$

com base nos dados disponíveis. No seguimento são obtidos os resíduos, $e_i = y_i - \hat{y}_i$, cujo conhecimento é indispensável para analisar as propriedades que se queiram atribuir aos erros ε_i do modelo.

Como já se mencionou na secção anterior, o método dos mínimos quadrados dominou desde sempre o processo de estimação em regressão. Na verdade existem muitos outros métodos que podem ser usados para estimar os parâmetros deste modelo e o primeiro de que há notícia, conhecido actualmente como método de regressão L_1 ou método dos mínimos desvios absolutos e que consiste em minimizar a soma dos valores absolutos dos resíduos (em vez da soma dos quadrados dos resíduos, como nos mínimos quadrados), é anterior, em várias dezenas de anos, ao método dos mínimos quadrados. O sucesso dos mínimos quadrados, confirmado pela popularidade da sua incessante utilização, deve-se em parte à facilidade do tratamento matemático, à simplicidade computacional das expressões a que a sua aplicação conduz (era um dos poucos métodos de estimação que fornecia expressões explícitas para os estimadores e por isso permitia calcular facilmente as estimativas antes do aparecimento dos computadores) e ainda a certas propriedades óptimas, embora estas se revelem apenas em condições que frequentemente não se encontram na prática.

De facto, a aplicação dos mínimos quadrados exige que $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ e $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, isto é, que os erros sejam não correlacionados e te-

nham variância constante (homocedásticos). Nestas condições o teorema de Gauss-Markov afirma que: **(i)** as estimativas dos mínimos quadrados, $\hat{\boldsymbol{\beta}}$, minimizam a soma dos quadrados dos resíduos, qualquer que seja a distribuição dos erros, **(ii)** os estimadores dos mínimos quadrados, $\hat{\boldsymbol{\beta}}$, são combinações lineares das respostas $Y_1, \ldots, Y_n$, são estimadores centrados para $\boldsymbol{\beta}$ e apresentam a variância mínima entre todos os estimadores de $\boldsymbol{\beta}$ que sejam centrados e obtidos como combinações lineares de $Y_1, \ldots, Y_n$. Esta é uma propriedade muito desejável mas que não se encontra na prática sempre que as hipóteses exigidas pelo método dos mínimos quadrados não são satisfeitas. Quando isso acontece não só a optimalidade não é atingida como os resultados da aplicação indevida dos mínimos quadrados podem ser perigosamente enganadores.

Outra hipótese que, embora não sendo exigida para a estimação dos parâmetros da regressão pelo método dos mínimos quadrados, é indispensável para uma análise mais completa do modelo consiste em assumir que os erros são normais, isto é, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Esta condição torna possível o trabalho de inferência sobre os parâmetros, o que é essencial para o estudo compreensivo do modelo. Com a normalidade os erros ficam independentes[1] e uma vez que há uma distribuição associada aos erros torna-se possível estimar os parâmetros com base no princípio da máxima verosimilhança, sendo então fácil verificar que o método dos mínimos quadrados conduz precisamente à mesma solução que o método da máxima verosimilhança. Esta coincidência é referida como mais uma boa razão para justificar o uso dos mínimos quadrados, uma vez que assumir que os erros têm distribuição normal é uma decisão razoável em muitas situações práticas.

É também importante referir que as propriedades óptimas dos estimadores dos mínimos quadrados dizem respeito apenas ao grupo de estimadores que são lineares nas observações da variável resposta. Por isso é pertinente perguntar: e o que se passa se os estimadores não forem lineares? Ou, dizendo de outra maneira mais conveniente: haverá estimadores não lineares que têm um desempenho melhor que os estimadores lineares no contexto do modelo de regressão? Em princípio nada impede que existam esses estimadores e como se verá ao introduzir a regressão robusta na Secção 4.3 eles existem mesmo,

[1] A distribuição normal é a única distribuição multivariada para a qual a não correlação entre as componentes é equivalente à independência das mesmas.

tudo depende do verdadeiro modelo.[2]

Apresentados que foram os méritos do método dos mínimos quadrados, interessa agora saber como é que o método se comporta em condições que não são as ideais, isto é, quando as hipóteses em que o método assenta não são respeitadas, o que de facto acontece correntemente na prática.

Para constatar como a violação das hipóteses pode influenciar os resultados de uma análise de regressão apresentam-se a seguir algumas ilustrações.

4.2.1 Heterocedasticidade e não normalidade

Uma situação em que a hipótese da variância constante falha é aquela em que a variável resposta tem uma distribuição cuja variância depende da média, como se verifica em várias aplicações nas áreas da biologia, medicina, engenharia e muitas outras. A tensão arterial diastólica, por exemplo, apresenta uma variabilidade que é crescente com a idade e o mesmo acontece quando relacionamos a altura (ou o peso) de crianças com a idade. Outros casos há em que a variabilidade é decrescente com a variável explicativa em estudo. Em qualquer dessas situações verifica-se que num gráfico de resíduos (resíduos contra a variável explicativa) estes se apresentam dispersos em forma de altifalante (ou funil), forma que é geralmente invocada para identificar estas situações de heterocedasticidade.

O exemplo que se segue ilustra os efeitos que a falta de homogeneidade da variância e da normalidade dos erros podem ter nas estimativas dos parâmetros da regressão.

Exemplo 4.1. Escolheu-se o modelo de regressão simples com parâmetros $\beta_0 = 0$ e $\beta_1 = 1$,

$$y_i = x_i + \varepsilon_i$$

e a partir dele foram geradas 10 observações para cada um dos possíveis valores de $x = \{1, 2, 3, 4, 5\}$, ou seja o vector das observações

[2]Mais um resultado singular da distribuição normal: é a única distribuição dos erros para a qual os estimadores óptimos são lineares, e por conseguinte os dos mínimos quadrados.

da variável explicativa é

$$\mathbf{x}^T = (1, \ldots, 1, 2, \ldots, 2, 3, \ldots, 3, 4, \ldots, 4, 5, \ldots, 5)$$

e a matriz $\mathbf{X}$ é dada por

$$\mathbf{X}^T = \begin{bmatrix} 1 & \ldots & 1 & 1 & \ldots & 1 & 1 & \ldots & 1 & 1 & \ldots & 1 & 1 & \ldots & 1 \\ 1 & \ldots & 1 & 2 & \ldots & 2 & 3 & \ldots & 3 & 4 & \ldots & 4 & 5 & \ldots & 5 \end{bmatrix}. \tag{4.3}$$

Para definir completamente o modelo foram associados aos erros, ε_i, quatro tipos de distribuições, contemplando as quatro possibilidades que interessam ao estudo:

(i) duas normais, uma com variância constante, $\text{var}(\varepsilon_i) = \sigma^2 = 0.5^2$, e outra cuja variância é proporcional ao quadrado da abcissa, ou seja, $\text{var}(\varepsilon_i) = \sigma_i^2 = 0.5^2\, x_i^2$;
(ii) duas distribuições t_5, isto é, não normais e com maior potencial no que respeita à geração de *outliers*, mas com estrutura de variância igual à das duas normais anteriores, respectivamente.

Na Tabela 4.1 estão descritas as várias situações (o factor multiplicativo $\sqrt{3/5}$ nos casos da distribuição t tem por objectivo fazer com que $\text{var}(\varepsilon_i)$ seja igual nos casos normal e t, uma vez que a variância de uma variável aleatória com distribuição t_ν é $\nu/(\nu-2)$).

Tabela 4.1 *Quatro estruturas para os erros de* $y_i = x_i + \varepsilon_i$.

	Erros	Situação
NH	$\varepsilon_i \sim \mathcal{N}(0, 0.5^2)$	Normal homocedástica
NnH	$\varepsilon_i \sim \mathcal{N}(0, 0.5^2 x_i^2)$	Normal heterocedástica
T5H	$\varepsilon_i \sim t_5 \times 0.5 \times \sqrt{\frac{3}{5}}$	Não normal homocedástica
T5nH	$\varepsilon_i \sim t_5 \times 0.5 \times \sqrt{\frac{3}{5}}\, x_i$	Não normal heterocedástica

Com base em cada um dos modelos assim definidos foram realizadas 1000 simulações com 50 observações cada uma, geradas da forma já descrita. A Figura 4.1 mostra os gráficos de dispersão de 50 observações correspondentes a uma única simulação, nos quatro casos

considerados. Nota-se claramente a dispersão afunilada nas situações de heterogeneidade e também o aparecimento de algumas observações mais afastadas quando se passa das distribuições normais para as distribuições t.

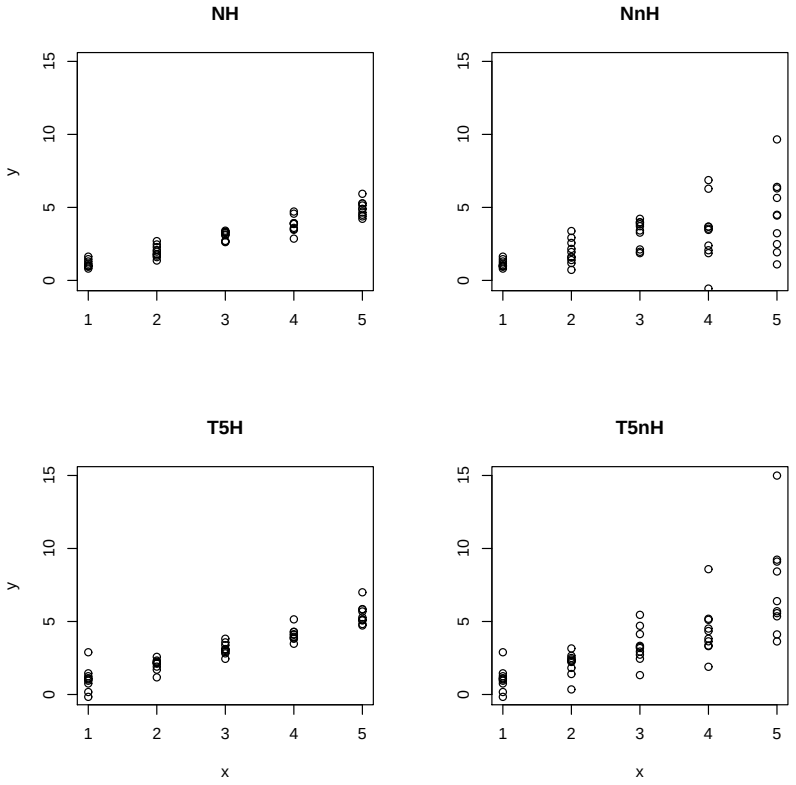


Figura 4.1 *Gráficos de dispersão das 50 observações de uma simulação para cada uma das distribuições dos erros descritas na Tabela 4.1.*

A Tabela 4.2 apresenta os resultados das 1000 simulações. Relativamente às estimativas pontuais dos parâmetros são indicadas as médias das 1000 estimativas obtidas (deve estar próximo do verdadeiro valor se os estimadores forem centrados) e entre parêntesis o respectivo desvio padrão (que corresponde a uma estimativa do erro padrão associado à estimativa pontual). Para os intervalos de confi-

Tabela 4.2 *Resultados da simulação relativa ao modelo* $y_1 = x_i + \varepsilon_i$ *(método dos mínimos quadrados).*

	NH	NnH	T5H	T5nH
$\hat{\beta}_0$	0.004 (0.165)	0.001 (0.393)	−0.002 (0.165)	−0.008 (0.396)
$IC(\beta_0)$	0.659 (0.068)	2.206 (0.278)	0.660 (0.110)	2.179 (0.425)
$\hat{\alpha}$	0.944	0.995	0.953	0.994
$\hat{\beta}_1$	0.998 (0.050)	0.993 (0.176)	1.000 (0.050)	1.007 (0.179)
$IC(\beta_1)$	0.199 (0.021)	0.665 (0.084)	0.199 (0.033)	0.657 (0.128)
$\hat{\alpha}$	0.946	0.937	0.956	0.938

ança indicam-se a média dos comprimentos dos intervalos, o desvio padrão dos mesmos (entre parêntesis) e, na segunda linha, a estimativa da probabilidade de cobertura ($\hat{\alpha}$), calculada como a proporção de vezes, nas 1000 simulações, em que o intervalo obtido contém o valor do parâmetro. $\hat{\alpha}$ corresponde a uma estimativa do verdadeiro nível de confiança dos intervalos (α), o qual pode ser diferente do nível especificado (95%) se o método não for adequado.

Note-se que no caso NnH a falta de homogeneidade pode ser controlada usando o método dos mínimos quadrados ponderados, que é o método correcto para esta situação (também aqui se pode mostrar que coincide com o método da máxima verosimilhança se for assumida a normalidade dos erros). A não homogeneidade pode ser incorporada no modelo (4.2), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, alterando a estrutura dos erros de $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ para $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$. No caso em consideração, erros com variâncias distintas, $\mathbf{V}$ é uma matriz diagonal onde cada elemento da diagonal corresponde a $\text{var}(\varepsilon_i)/\sigma^2$ ($\mathbf{V}$ não diagonal corresponde a uma estrutura com erros correlacionados, nesse caso o método associado costuma ser designado por mínimos quadrados generalizados, podendo os mínimos quadrados ponderados ser considerados como um caso particular dos mínimos quadrados generalizados). Na Tabela 4.3 apresentam-se em paralelo os resultados que permitem fazer inferências sobre os parâmetros do modelo de regressão nas duas situações, mínimos quadrados e mínimos quadrados

Tabela 4.3 *Expressões que permitem fazer inferências sobre os parâmetros do modelo de regressão quando se usa o método dos mínimos quadrados e o método dos mínimos quadrados ponderados/generalizados.*

	Mínimos quadrados	Mínimos quadrados ponderados/generalizados
Modelo	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Estrutura dos erros	$E(\boldsymbol{\varepsilon}) = \mathbf{0}$ $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$	$E(\boldsymbol{\varepsilon}) = \mathbf{0}$ $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$
Função a minimizar	$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$	$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
Estimativas	$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
Prop. dos estimadores	$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ $\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$	$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ $\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2$
Outras estimativas	$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2$ $\hat{\sigma}^2 = \dfrac{SSE}{n-p}$ $SSE = = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$	$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\hat{\sigma}^2$ $\hat{\sigma}^2 = \dfrac{SSE}{n-p}$ $SSE = = \mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
Se $\boldsymbol{\varepsilon} \sim \mathcal{N}_n$	$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}$ $\dfrac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{\text{var}}(\beta_i)}} \sim t_{n-p}$	$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}$ $\dfrac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{\text{var}}(\beta_i)}} \sim t_{n-p}$
	$\left(\text{com } \widehat{\text{var}}(\beta_i) \equiv \left[\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\right]_{ii}\right)$	

Tabela 4.4 *Resultados da simulação relativa ao modelo* $y_1 = x_i + \varepsilon_i$ *(método dos mínimos quadrados ponderados).*

	NnH	T5nH
$\hat{\beta}_0$	0.008 (0.240)	−0.003 (0.253)
$IC(\beta_0)$	0.969 (0.101)	0.970 (0.162)
$\hat{\alpha}$	0.946	0.949
$\hat{\beta}_1$	0.995 (0.132)	1.000 (0.134)
$IC(\beta_1)$	0.524 (0.054)	0.525 (0.087)
$\hat{\alpha}$	0.943	0.952

ponderados/generalizados.[3] Note-se que os resultados da segunda coluna (mínimos quadrados ponderados/generalizados) coincidem com os resultados dos mínimos quadrados se aplicados a um modelo transformado que resulta do modelo original (4.2) pela multiplicação à esquerda por $\mathbf{V}^{-1/2}$. No presente estudo tem-se $\sigma^2 = 0.5^2$ e $\mathbf{V}$ é uma matriz diagonal de dimensão 50×50, com $v_{ii} = x_i^2$, $i = 1, \ldots, 50$. Assim $\mathbf{V}^{-1/2}$ também é uma matriz diagonal, em que os elementos da diagonal são os pesos $w_i = 1/\sqrt{v_{ii}} = 1/x_i$. O método dos mínimos quadrados ponderados foi também aplicado ao caso de não homogeneidade, T5nH, embora aqui os resultados relativos à precisão das estimativas não sejam exactos, devido à não normalidade.

Os resultados da simulação usando o método dos mínimos quadrados ponderados são apresentados na Tabela 4.4 em formato semelhante ao descrito para a Tabela 4.2 (só para as situações NnH e T5nH, como é evidente).

Da análise das Tabelas 4.2 e 4.4 podem obter-se várias conclusões:

(i) As médias das estimativas de β_0 e β_1 são muito próximas dos verdadeiros valores em todos os casos experimentados. Este resultado não é de estranhar, uma vez que a teoria garante que os estimadores dos mínimos quadrados (e dos mínimos quadrados ponderados) são centrados seja qual for a distribuição dos

[3] Quando $\mathbf{V}$ envolve parâmetros, as expressões indicadas só são válidas se estes forem conhecidos.

erros, a única condição necessária é que o valor esperado dos erros seja nulo, o que acontece nos quatro cenários simulados.

(ii) Apesar das médias das estimativas dos parâmetros sugerirem um bom desempenho em condições que não são as ideais, verifica-se que as médias das amplitudes dos intervalos de confiança (as quais estão relacionadas com a variância, ou com a eficiência, dos estimadores) nas duas situações não homogéneas mostram aumentos substanciais em relação às correspondentes situações homogéneas, mais concretamente, aumentos superiores a 300%. O mesmo se passa, claro, com as estimativas do erro padrão, cujo acréscimo chega a ultrapassar os 400%.

(iii) As estimativas da probabilidade de cobertura estão geralmente próximas do valor nominal $\alpha = 95\%$. No entanto, em certos casos, como por exemplo em NnH para β_0, tem-se que $\hat{\alpha}$ é bastante superior a 95%, o que é compatível com o facto dos respectivos intervalos terem grande amplitude.

(iv) A acção dos mínimos quadrados ponderados sobre a heterogeneidade melhora a precisão das estimativas e reduz os comprimentos dos intervalos de confiança, os quais ficam, no entanto, sempre superiores aos correspondentes no caso homogéneo. Este resultado é de certa foram esperado, uma vez que se estão a introduzir no modelo erros com maior variabilidade para todas as observações com excepção das 10 primeiras.

(v) O caso com resultados mais gravosos é o caso T5nH onde a heterocedasticidade se junta à não normalidade, mas ainda assim estes resultados não são muito diferentes dos que se obtêm no caso NnH, o que leva a pensar que os desvios da situação ideal NH são causados essencialmente, neste exemplo, pela inclusão da heterogeneidade, sendo a falta de normalidade introduzida pela distribuição t pouco significativa. Este resultado acaba por não ser de todo estranho pois por construção os erros normais e não normais têm a mesma variância, e quer a variabilidade das estimativas, quer o comprimento médio dos intervalos de confiança apenas reflectem aquele parâmetro. Assim, o que é natural é que a não normalidade se faça sentir em efeitos de ordem superior, tais como a variabilidade do comprimento dos

intervalos de confiança. De facto comparando os valores dos desvios padrões dos comprimentos dos intervalos de confiança apresentados nas Tabelas 4.2 e 4.4, entre os casos normais e t correspondentes (isto é, NH com T5H e NnH com T5nH), verifica-se que os associados aos erros com distribuição t são cerca de 50% a 60% superiores aos associados aos erros com distribuição normal.

(vi) É também de referir que os resultados experimentais observados estão em grande concordância com os valores teóricos calculados a partir das fórmulas da Tabela 4.3, o que atesta a validade do estudo de simulação. No caso homogéneo (independentemente de a distribuição de ε_i ser normal ou t) tem-se, com $\mathbf{X}^T$ dada em (4.3) e $\sigma^2 = 0.5^2$,

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 = \begin{bmatrix} 0.0275 & -0.0075 \\ -0.0075 & 0.0025 \end{bmatrix},$$

o que significa que os verdadeiros erros padrão de $\hat{\beta}_0$ e $\hat{\beta}_1$ são, respectivamente, $\sqrt{0.0275} \simeq 0.166$ e 0.05, enquanto que, consultando a Tabela 4.2 se verifica que os valores estimados com base na simulação foram, respectivamente, 0.165 e 0.050. Já no caso não homogéneo, e utilizando os estimadores dos mínimos quadrados ponderados, conclui-se que

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2 = \begin{bmatrix} 0.0593981 & -0.0271251 \\ -0.0271251 & 0.01738714 \end{bmatrix},$$

pelo que os valores teóricos são 0.244 e 0.132, e os valores experimentais são (da Tabela 4.4) dados por 0.240 e 0.132, no caso NnH, e por 0.253 e 0.134, no caso T5nH. Finalmente quando se aplica o método dos mínimos quadrados aos dados não homogéneos, conclui-se que a variância de $\hat{\boldsymbol{\beta}}$ não é dada por nenhuma daquelas expressões, mas sim, uma vez que se usa o estimador $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ e se tem $\text{var}(\mathbf{y}) = \text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, por

$$\begin{aligned} \text{var}(\hat{\boldsymbol{\beta}}) &= \text{var}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) = \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T\sigma^2 = \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2. \quad (4.4) \end{aligned}$$

Fazendo os cálculos obtém-se

$$\text{var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.154 & -0.063 \\ -0.063 & 0.031 \end{bmatrix},$$

pelo que os valores teóricos são 0.392 e 0.176, e os valores experimentais são (da Tabela 4.2) dados por 0.393 e 0.176, no caso NnH, e por 0.396 e 0.179, no caso T5nH. Deve ainda notar-se que a fórmula usada para estimar os erros padrão que são posteriormente usados em cada simulação para calcular os intervalos de confiança quando se está nesta situação (aplicação dos mínimos quadrados ordinários ao caso não homogéneo), é a fórmula do lado esquerdo da Tabela 4.3 a qual é deduzida assumindo que os erros são homogéneos e não está de facto a estimar a verdadeira variância de $\hat{\boldsymbol{\beta}}$, dada por (4.4). Esta constitui mais uma fonte de erro para os intervalos de confiança.

4.2.2 Presença de *outliers*

Outliers são observações extremas que se encontram de tal forma afastadas da maioria dos dados que surgem dúvidas sobre se elas poderão ou não ter sido geradas pelo modelo proposto para explicar essa maioria dos dados. Os *outliers* tanto podem corresponder a erros de medição, a erros de cálculo ou outros erros de origem diversa, como a observações genuínas contendo informação relevante que surge em consequência de interacções muitas vezes difíceis de localizar.

Note-se que a presença de *outlier*s genuínos pode querer dizer que o modelo proposto não é adequado e que deve ser reformulado. Por exemplo, no contexto do modelo normal a presença de *outliers* pode significar que o verdadeiro modelo é um modelo de caudas mais pesadas, o que equivale a dizer que a hipótese de normalidade não é respeitada.

A presença de *outliers* num conjunto de dados é sempre um motivo de reflexão para o cientista, preocupado em entender o verdadeiro sinal que lhe é dado pelo aparecimento de um *outlier*. O cientista desejará saber se:

(i) afinal o *outlier* não é *outlier* mas sim um ponto do próprio modelo podendo portanto ser acomodado por ele;
(ii) o *outlier* é claramente um erro;
(iii) o *outlier* é o resultado de considerações ou hipóteses inadequadas sobre o modelo;
(iv) o modelo deverá ser reformulado e como fazê-lo.

O interesse pelos *outliers* no contexto da estimação pelos mínimos quadrados resulta do facto da regressão poder ser atraída de forma eventualmente exagerada para o ou os *outliers*, em face do critério de minimização dos quadrados dos resíduos, conduzindo a um modelo enganador se o *outlier* for um verdadeiro erro e não houver evidência clara para o reconhecer. É claro que o reconhecimento de um *outlier* não informativo levaria eventualmente à sua eliminação do conjunto inicial dos dados.

Para perceber melhor os problemas que a presença de observações extremas pode causar à estimação dos mínimos quadrados é conveniente distinguir vários tipos de *outliers* e exemplificar as consequências da sua acção, o que será feito aqui usufruindo das vantagens de simplicidade e clareza que a regressão linear simples proporciona:

(i) *Outlier* de regressão – trata-se de um ponto que se afasta bastante da estrutura linear evidenciada pela massa principal dos dados e que influencia de forma inconveniente a estimação, conduzindo a modelos ajustados impróprios. Este tipo de *outlier* é identificado pela análise simultânea da variável resposta e da variável explicativa, mas há outros tipos de *outliers* que são identificados considerando as variáveis separadamente.

(ii) *Outlier* em x (ponto de *leverage* ou alavanca) é um ponto que é um *outlier* em relação à coordenada x, isto é, cuja coordenada x está muito afastada das suas congéneres. Este é um ponto que pode ter uma grande influência na estimação (mau ponto de *leverage*), ou não ter uma grande influência na estimação (bom ponto de *leverage*), sendo por isso um potencial *outlier* de regressão.

(iii) *Outlier* em y – um ponto que é *outlier* em relação à coordenada y. Na prática pode ser ou não um *outlier* de regressão.

(iv) *Outlier* em (x, y) – um ponto que é *outlier* nas duas coordenadas. Este pode também ser ou não um *outlier* de regressão.

Exemplo 4.2. Para ilustrar as várias situações acabadas de descrever considerou-se um conjunto de 12 pontos (artificiais) ao qual foi ajus-

tada a recta

$$\hat{y}_{so} = \hat{\beta}_0 + \hat{\beta}_1\, x.$$

Ao conjunto de 10 pontos juntou-se um *outlier* W_i, $i = 1, 2, 3, 4$, e ajustou-se a recta

$$\hat{y}_{co} = \hat{\beta}_0^* + \hat{\beta}_1^*\, x.$$

Os *outliers* considerados foram os seguintes:

W_i	x	y	Descrição
W_1	4	8	*Outlier* de regressão
W_2	5.5	10	*Outlier* em y e de regressão
W_3	10	11	*Outlier* em x e em y mas não de regressão
W_4	10	4	*Outlier* em x e de regressão

Note-se que o ponto W_3 corresponderá a um bom ponto de *leverage* enquanto o ponto W_4 corresponderá a um mau ponto de *leverage*.

A Figura 4.2 mostra o diagrama de dispersão dos 12 pontos originais, a recta de regressão estimada só com esses pontos ($\hat{y}_{so}$) e a posição relativa dos *outliers*, $W_1, \ldots, W_4$. Na Tabela 4.5 apresentam-se as estimativas dos parâmetros do modelo (incluindo σ) e os valores do coeficiente de determinação (R^2) para cada situação. O efeito dos *outliers* pode também ser apreciado nos gráficos da Figura 4.3 que mostram a posição relativa das rectas $\hat{y}_{so}$ e $\hat{y}_{co}$ para cada um dos cenários.

Tabela 4.5 *Estimativas dos parâmetros do modelo de regressão e valores do coeficiente de determinação para os 12 pontos originais (sem outliers) e com cada um dos outliers,* $W_1, \ldots, W_4$.

Caso	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	R^2
sem *outliers*	0.821	1.041	0.374	0.936
com W_1	2.005	0.859	0.911	0.622
com W_2	0.936	1.067	1.062	0.634
com W_3	0.946	1.015	0.359	0.966
com W_4	4.824	0.240	1.498	0.083

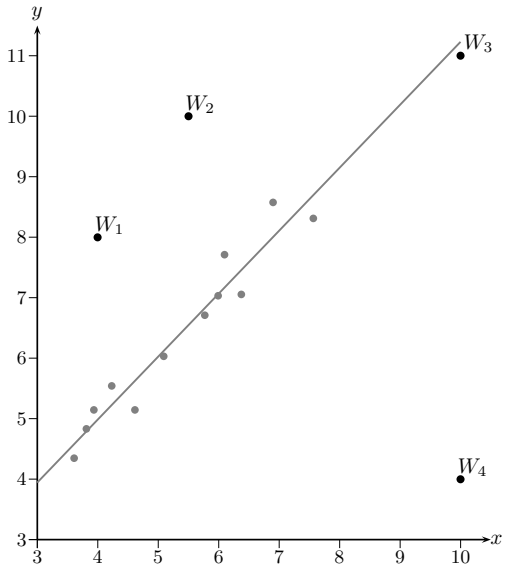


Figura 4.2 *Diagrama de dispersão dos 12 pontos originais, recta de regressão estimada com esses pontos ($\hat{y}_{so}$) e posição relativa dos outliers, $W_1, \ldots, W_4$.*

Da observação da Tabela 4.5 e dos gráficos da Figura 4.3 podem avançar-se os seguintes comentários.

- Como esperado o pior efeito sobre as estimativas pontuais de β_0 e β_1 é produzido pelo mau ponto de *leverage* W_4. Pelo contrário, o bom ponto de *leverage* W_3 provoca apenas uma pequena alteração naquelas estimativas (a alteração seria nula se o ponto estivesse exactamente sobre a recta $\hat{y}_{so}$).
- Em relação à estimativa de σ e ao valor do coeficiente de determinação verifica-se que o pior efeito é também o do ponto W_4 (que em relação ao valor de R^2 pode mesmo qualificar-se como desastroso). Por seu lado o bom ponto de *leverage* acaba por ter um efeito positivo nestes indicadores.
- Aparentemente os pontos W_1 e W_2 têm uma pequena influência nas rectas ajustadas (ver Figura 4.3), no entanto analisando com mais cuidado verifica-se que há efeitos elevados quer em $\hat{\sigma}$

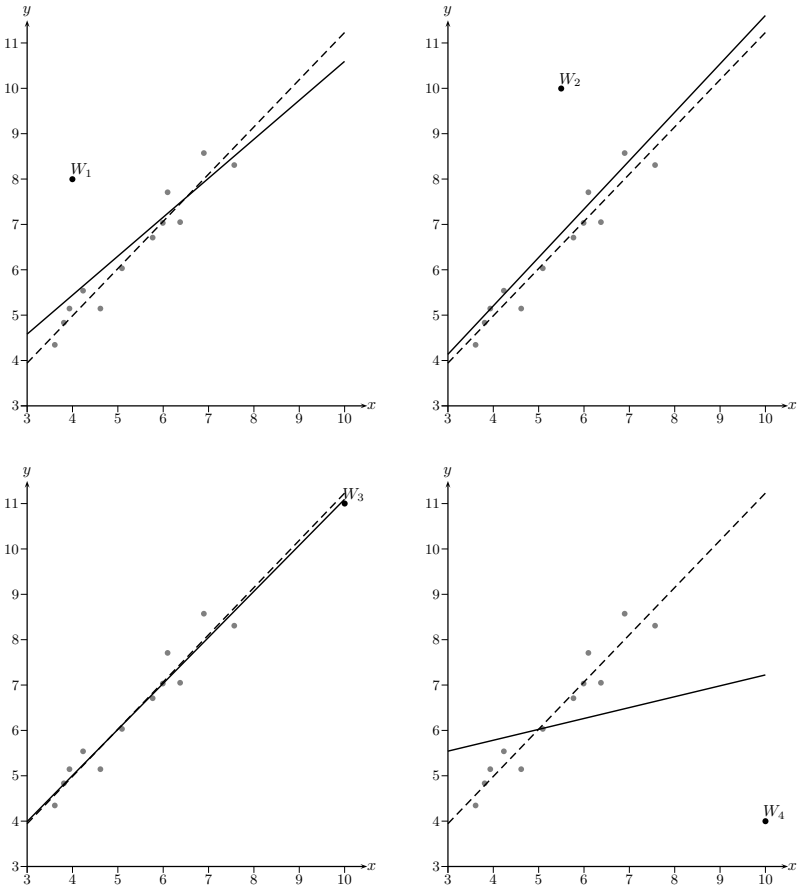


Figura 4.3 *Posição relativa das rectas $\hat{y}_{so}$ (a tracejado) e $\hat{y}_{co}$ (a cheio) para cada um dos outliers, $W_1, \ldots, W_4$.*

quer em R^2 que teriam consequências nefastas na análise global do modelo (qualidade do ajustamento e testes de hipóteses ou intervalos de confiança para β_0 e β_1).

- É também importante analisar a influência ao nível dos resíduos. Por exemplo os resíduos de $\hat{y}_{co}$ com os pontos W_1 e W_4 poderiam sugerir a correlação sucessiva dos erros.

Em Neter *et al.* (1996) apresentam-se vários procedimentos, uns informais[4] e outros mais ou menos formais,[5] para usar de foram sistemática, com vista a efectuar o diagnóstico das deficiências já referidas, e indicam-se os respectivos remédios.[6] À aplicação destes remédios deve seguir-se novo processo de diagnóstico para verificar se as deficiências detectadas foram ou não eliminadas.

Como se refere na última nota de rodapé um dos remédios preconizados para atacar as deficiências do método dos mínimos quadrados quando as condições ideais não se verificam é o método da regressão robusta cujo estudo vai ser feito a seguir.

4.3 Métodos robustos de regressão

A proposta de regressão robusta surge como alternativa a um procedimento que não é invulgar na prática e que consiste em retirar, do conjunto de dados a que se vai ajustar um modelo de regressão, os *outliers* que, aparentemente não satisfazendo o modelo, também não são claramente verdadeiros erros. A razão deste procedimento resulta da indesejável sensibilidade que o método dos mínimos quadrados, tradicionalmente usado no modelo de regressão, apresenta em relação a *outliers*. De facto, uma única observação excepcionalmente distante da massa central dos dados pode distorcer por completo o resultado do ajustamento, dando uma indicação errada da associação entre as variáveis em estudo.

Traduzindo a referida ideia de distorção em termos mais formais e

[4] Análise gráfica dos resíduos e de combinações variadas, em pares, de variáveis explicativas.

[5] Diversas análises de resíduos, estandardizados e semi-studentizados, para identificação de *outliers*; matriz chapéu e medidas de identificação de casos influentes – *DFITS*, *DBETAS* e distância de Cook; testes relativos a variância constante, teste de Levene modificado; testes relativos às correlações entre os erros, teste de Durbin-Watson. A utilização destes procedimentos nem sempre tem o sucesso esperado, especialmente em presença de vários *outliers*, devido ao efeito de mascaramento (*masking*).

[6] Transformações com vista a tornar os erros com distribuição próxima da normal, aproximadamente não correlacionados, com variâncias aproximadamente iguais e a tornar a relação de regressão linear. Usar mínimos quadrados pesados para tratar o problema da desigualdade das variâncias dos erros e a regressão robusta para tratar do problema criado pela existência de *outliers* influentes no conjunto de dados em estudo.

quantitativos pode dizer-se que os estimadores dos mínimos quadrados têm um ponto de rotura igual a zero e uma função de influência ilimitada.

O objectivo da regressão robusta consiste em reduzir a influência dos *outliers* permitindo que eles façam parte dos dados e permaneçam no processo de estimação, em vez da alternativa radical de, pura e simplesmente os eliminar.[7]

Em termos muito genéricos e algo simplistas pode dizer-se que a regressão robusta actua de modo a atribuir "pesos" às observações, consoante a influência que exercem no processo de estimação. A observações a que correspondem resíduos de grande magnitude será atribuído um menor "peso". O método dos mínimos quadrados não tem essa preocupação e, por isso, ao atribuir o mesmo "peso" (igual a 1) a todas as observações, deixa os *outliers* livres de exercerem a sua total influência. Ao limitar a influência dos *outliers* a regressão robusta propõe-se realizar um ajustamento que reflicta principalmente a contribuição da maioria dos dados.

Exemplo 4.3. Para espreitar ainda de maneira informal a acção da regressão robusta consideram-se os dados da Tabela 4.6 retirados da colecção de exercícios usada nas aulas práticas da disciplina de Probabilidades e Estatística do Instituto Superior Técnico, sendo que y_1 foi alterado de 4.3 para 6.0. Os dados representam o consumo médio de energia (Y), em KW, por agregado familiar durante 10 dias de um mês de Inverno numa cidade, e a temperatura média diária (X), em ${}^0C$.

A Figura 4.4 mostra o gráfico dos dados e a recta de regressão dos mínimos quadrados: **(a)** com todos os pontos e **(b)** sem a primeira observação da tabela. O gráfico sugere que a observação (15, 6.0) é um *outlier* de regressão, embora não seja *outlier* nem em x nem em y. A recta foi claramente atraída pelo *outlier* e as estimativas dos coeficientes e respectivos intervalos de confiança foram alterados de forma significativa, quando comparados com as estimativas obtidas sem aquele ponto, como mostra a Tabela 4.7.

Em seguida acrescentaram-se ao gráfico da Figura 4.4 rectas de regressão robustas (obtidas pelos métodos aqui representados pelas

[7] Até porque em problemas com muitas variáveis explicativas a sua detecção é muito difícil e portanto a sua eliminação é impossível.

Tabela 4.6 *Dados relativos a consumo médio de energia (y_i) e temperatura média diária (x_i).*

i	1	2	3	4	5	6	7	8	9	10
x_i	15	14	12	14	12	11	11	10	12	13
y_i	6.0	4.4	5.3	4.6	5.5	5.9	5.7	6.2	5.2	5.0

Tabela 4.7 *Resultados do ajustamento do modelo de regressão linear simples (pelo método dos mínimos quadrados) aos dados da Tabela 4.6, com todas as observações e retirando a primeira.*

Dados	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2	$IC_{95\%}(\beta_0)$	$IC_{95\%}(\beta_1)$
Todos	7.88	−0.20	0.53	$(4.63, 11.13)$	$(-0.46, 0.06)$
Sem a 1ª. obs.	10.45	−0.42	0.97	$(9.58, 11.31)$	$(-0.50, -0.35)$

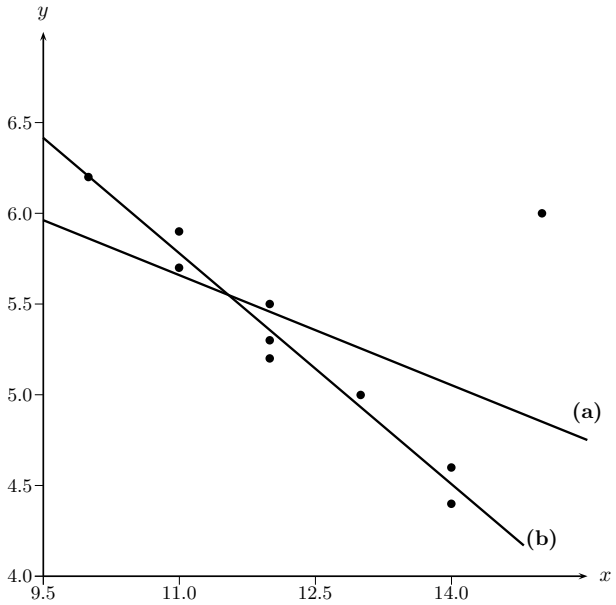


Figura 4.4 *Consumo de energia versus temperatura: gráfico de dispersão com regressões dos minímos quadrados,* ***(a)*** *com todos os pontos e* ***(b)*** *sem o primeiro ponto.*

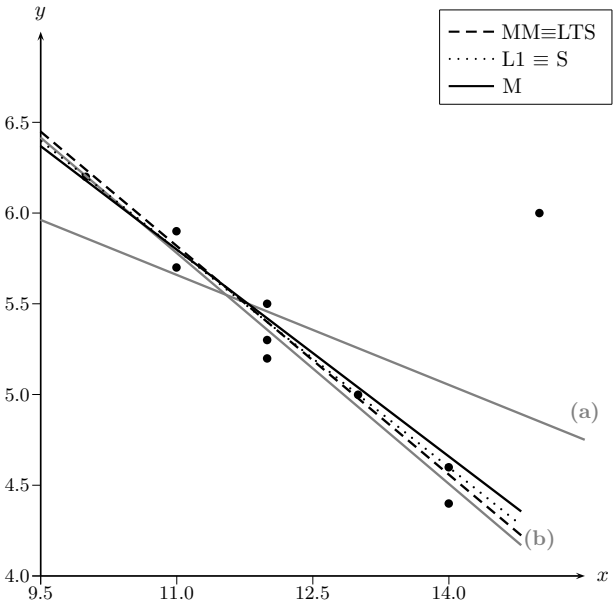


Figura 4.5 *Consumo de energia versus temperatura: gráfico de dispersão com regressões robustas.*

siglas L1, LTS, M, MM e S e que serão descritos a seguir). Como se observa na Figura 4.5 as rectas de regressão robustas estão próximas da recta de regressão dos mínimos quadrados obtida com base nos 9 pontos. Isto significa que os métodos de regressão robusta utilizados dão um "peso" muito pequeno ao *outlier*, o que é consistente com o resultado dos mínimos quadrados com 9 pontos, que por sua vez é equivalente a dar "peso" zero ao ponto que foi retirado.

O interesse pela regressão robusta tem levado ao desenvolvimento de muitos métodos robustos de estimação dos parâmetros do modelo de regressão.[8] A seguir descrevem-se vários dos métodos mais conhecidos.

[8]Sem perigo de confusão usam-se indistintamente os termos "métodos" e "estimadores".

4.3.1 Mínimos desvios absolutos

O método dos mínimos desvios absolutos (ou método LAD, de *Least Absolute Deviations*), também conhecido por mínimos erros absolutos, mínimo valor absoluto ou método de regressão L_1, é semelhante ao método dos mínimos quadrados no sentido em que procura encontrar a regressão que se ajusta o melhor possível a um conjunto de dados. Porém, em vez de operar sobre a soma dos quadrados dos resíduos, este método opera antes sobre a soma dos resíduos absolutos, isto é, pretende obter $\hat{\boldsymbol{\beta}}$ que minimiza

$$\sum_{i=1}^{n} \left| y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{p-1} x_{i,p-1}) \right| = \sum_{i=1}^{n} |e_i|,$$

ou seja, o método dá a mesma importância a todos os resíduos quer sejam grandes quer sejam pequenos, ao passo que os mínimos quadrados ampliam a importância dos resíduos de grande valor absoluto e reduzem a importância dos resíduos de pequeno valor absoluto. Note-se ainda que ao usar o critério dos mínimos quadrados se pode garantir que a soma (ou a média) dos resíduos é zero, enquanto que quando se usa o critério dos mínimos desvios absolutos o que se pode garantir é que a mediana dos resíduos é nula.

Como já se referiu, o método é anterior ao método dos mínimos quadrados e, contrariamente a este, largamente procurado e utilizado, tem-se verificado que o seu uso ao longo dos tempos tem sido modesto, possivelmente devido às complicações computacionais que no passado eram geralmente difíceis de ultrapassar. Contudo o método tem vantagens sobre o método dos mínimos quadrados por ser robusto na presença de certo tipo de *outliers*, concretamente os *outliers* em y. Este comportamento pode ser explicado pelo facto do valor absoluto dos resíduos dar menos ênfase às observações extremas do que o critério dos mínimos quadrados que considera o quadrado dos resíduos. Contudo o método é extremamente sensível a *outliers* em $\mathbf{x}$, podendo a presença destes, no caso da sua magnitude ser grande, deturpar completamente o ajustamento. Por esta razão o seu ponto de rotura é zero.[9] Outro aspecto que não abona em favor deste método é a sua

[9]No entanto quando, por imposição de delineamento experimental, não há possibilidade de haver *outliers* em $\mathbf{x}$, o ponto de rotura dos estimadores LAD da regressão é positivo. Ellis e Morgenthaler (1992) mostram que nesse contexto o ponto de rotura pode ser igual a 1/4 independentemente do número de variáveis explicativas.

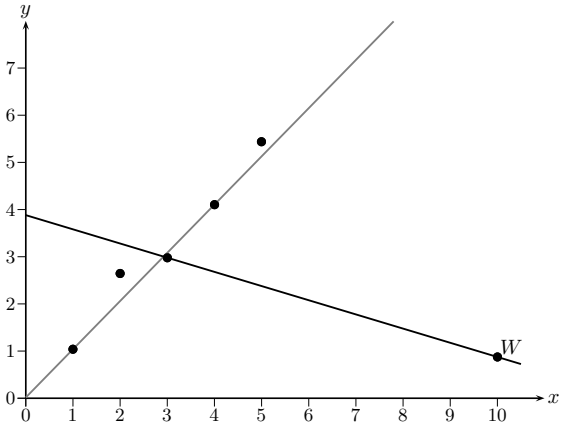


Figura 4.6 *Diagrama de dispersão e recta de regressão obtida pelo método LAD para um conjunto de cinco pontos artificiais. Acrescentou-se ao conjunto original o ponto W e obteve-se a recta de regressão dos mínimos desvios absolutos também representada e que passa por W.*

instabilidade, contrariamente ao método dos mínimos quadrados que é estável.[10] A situação é semelhante ao que se passa com os mínimos quadrados aplicados a dados colineares, mas com o método LAD os dados podem estar mesmo longe da colinearidade. Acresce ainda que, contrariamente ao que acontece com os mínimos quadrados, que produzem uma solução que é única e pode ser obtida explicitamente por via analítica, este método pode conduzir a múltiplas soluções e quaisquer que elas sejam só podem ser aproximadas por via numérica, iterativa e habitualmente assente em programação linear.

Há ainda uma particularidade curiosa do método LAD: a recta dos mínimos desvios absolutos ajustada ao conjunto de dados

$$\{(x_i, y_i),\, i = 1, \ldots, n\}\,,$$

contém pelo menos dois dos seus pontos, a menos que haja soluções múltiplas, caso em que essas soluções ficam limitadas por duas rectas que contêm pelo menos dois pontos dos dados. Esta propriedade

[10] Um método de estimação diz-se instável se pequenas alterações nos dados provocam variações súbitas (descontínuas) nas estimativas.

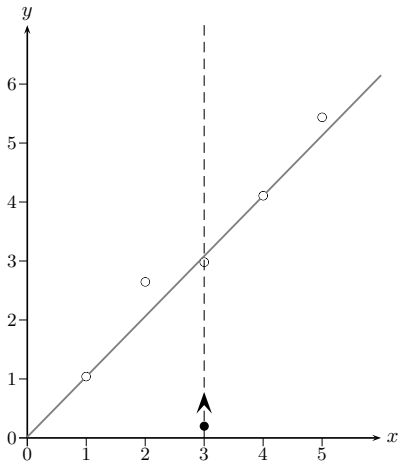


Figura 4.7 *Diagrama de dispersão e recta de regressão obtida pelo método LAD para um conjunto de cinco pontos. Fez-se variar a ordenada do ponto de abcissa* $x = 3$ *ao longo da linha a tracejado e observaram-se as variações para os dois coeficientes da recta dos mínimos desvios absolutos (ver Figura 4.8).*

única ajuda também a compreender a instabilidade e a robustez do método relativamente a *outliers* em y. Se a recta passa sempre por dois pontos, então logo que o conjunto de dados é alterado, a recta poderá saltar bruscamente em vez de reagir movendo-se continuamente, mas certamente que evitará os pontos *outliers* pois a inclusão destes não contribui para diminuir a soma dos resíduos absolutos, a não ser que eles estejam muito afastados em $\mathbf{x}$ (pontos de *leverage*), caso em que constituem uma atracção que pode ser fatal. Nas Figuras 4.6, 4.7 e 4.8 ilustram-se estas afirmações recorrendo a dados artificiais com *outliers* em $\mathbf{x}$ e y, respectivamente. Com os dados reais do Exemplo 4.5, apresentado mais adiante, observa-se também a atracção da recta LAD pelos pontos de *leverage* (ver Figura 4.11), o que mostra que este problema não é um problema só de dados artificiais de muito pequena dimensão.

Finalmente é preciso destacar que os estimadores obtidos por esta via perdem muita eficiência em presença de erros normais. É evidente que se os erros não forem normais pode suceder o contrário. Bassett

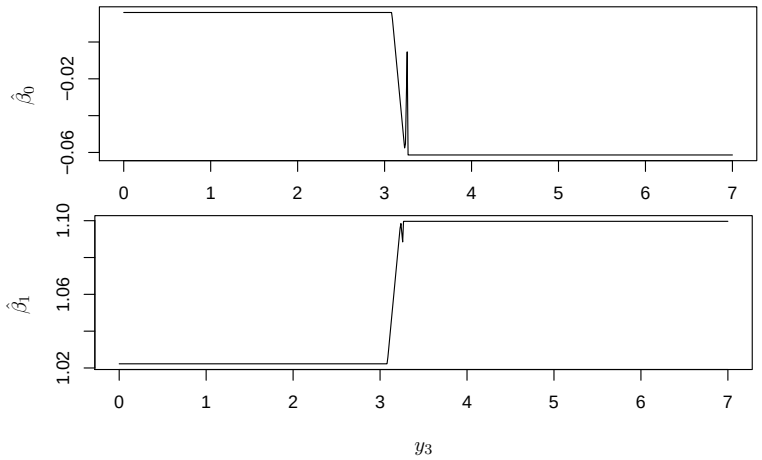


Figura 4.8 *Variação observada nos coeficientes da recta dos mínimos desvios absolutos quando se faz variar a ordenada do ponto de abcissa* $x = 3$*. Estes resultados ilustram simultaneamente a instabilidade e a robustez, pois registam-se variações súbitas (instabilidade) mas a variação global é pequena e limitada (robustez).*

e Koenker (1978) mostram que este estimador é mais eficiente que o dos mínimos quadrados para todas as distribuições dos erros para as quais a mediana é um estimador do parâmetro de localização mais eficiente que a média. Isto remete directamente para a observação final feita no Exemplo 3.2 (página 106), permitindo concluir que se os erros seguirem uma distribuição de *Laplace* então o estimador dos mínimos desvios absolutos é o estimador mais eficiente dos parâmetros da regressão devido a coincidir com o estimador de máxima verosimilhança. Por estas razões a recta dos mínimos desvios absolutos é considerada uma generalização da mediana univariada. Não é assim de estranhar que sofra dos mesmos problemas de instabilidade local.

Com tantas dificuldades associadas ao critério LAD, não admira a postura de Gauss ao defender que os mínimos quadrados são preferíveis quando os erros são normais.

O método LAD é geralmente invocado, muitas vezes talvez de forma leviana, quando se pretende sugerir um método sem os perigos dos mínimos quadrados, mas a história da sua actuação na prática é curta. O método não é muito requerido na prática, mas tem servido para

inspirar a procura de outros métodos que não se deixem perturbar por *outliers* em $\mathbf{x}$, isto é, que tenham um maior ponto de rotura, e melhorem a sua eficiência perante dados normais.

4.3.2 Estimadores-M

Como já se deixou antever no Capítulo 3, os estimadores-M constituem uma ferramenta de grande potencial para os mais variados problemas da estatística robusta. Foi já vista a sua contribuição fundamental na estimação da localização e escala de modelos estatísticos. Pretende-se agora apreciar o papel destes estimadores no contexto da regressão.

Os estimadores-M em regressão, introduzidos em Huber (1973), são obtidos com base num procedimento que pode ser visto como uma extensão natural do critério dos mínimos quadrados. Em vez de estimar $\boldsymbol{\beta}$, no modelo (4.2), minimizando a soma dos quadrados dos resíduos, isto é,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right)^2 ,$$

porque não pensar em minimizar antes outra função dos resíduos, ou seja,

$$\hat{\boldsymbol{\beta}} \ \text{minimiza} \ \sum_{i=1}^{n} \rho(e_i) \, ?$$

Note-se que, uma vez que e_i depende de y_i, $\mathbf{x}_i$ e $\hat{\boldsymbol{\beta}}$, esta expressão é um caso particular da expressão (3.40) que figura na definição geral (Definição 3.8) de um estimador-M, com $\mathbf{X}_i \equiv (y_i, \mathbf{x}_i)$.

Sendo assim, $\hat{\boldsymbol{\beta}}$, o estimador dos mínimos quadrados é um estimador-M, onde $\rho(u) = u^2$. E o mesmo acontece quando $\hat{\boldsymbol{\beta}}$ é o estimador LAD, pois

$$\hat{\boldsymbol{\beta}} \ \text{minimiza} \ \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} \rho(e_i) \, ,$$

com $\rho(u) = |u|$.

Como já se explicou na Secção 3.2, minimizar $\sum_{i=1}^{n} \rho(e_i)$ é equivalente a resolver o sistema de equações

$$\sum_{i=1}^{n} \frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \rho(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) = \mathbf{0},$$

ainda equivalente a

$$\sum_{i=1}^{n} \psi(e_i)\,\mathbf{x}_i = \sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})\,\mathbf{x}_i = \mathbf{0}, \tag{4.5}$$

onde $\psi = \rho'$. Contrariamente ao que acontece com os estimadores dos mínimos quadrados e com os estimadores LAD, os estimadores-M com função ψ geral não são equivariantes em relação à escala, de forma que para conseguir esta propriedade desejável é preciso proceder a uma transformação substituindo e_i por $e_i/\hat{\sigma}$, onde $\hat{\sigma}$ é uma estimativa robusta (normalmente é usado o MAD) do factor de escala, ou melhor do parâmetro de escala (σ) da distribuição dos erros, ε_i (ver a Secção 3.2.3). Nestas condições (4.5) toma a forma

$$\sum_{i=1}^{n} \psi\left(\frac{e_i}{\hat{\sigma}}\right)\,\mathbf{x}_i = \sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\,\mathbf{x}_i = \mathbf{0}. \tag{4.6}$$

Para conseguir estimadores-M com níveis de robustez e eficiência desejáveis é preciso escolher judiciosamente a função ψ, de forma idêntica àquela que foi explicada para os estimadores-M de localização. Por exemplo, escolhendo para ψ a função de Huber (ver página 108) o estimador-M de $\boldsymbol{\beta}$ é mais eficiente do que o correspondente estimador LAD para erros com distribuição normal e é robusto em relação a *outliers* em y. Contudo o estimador não é robusto em $\mathbf{x}$, a não ser que o delineamento possa evitar o aparecimento de *outliers* em $\mathbf{x}$. No caso das variáveis explicativas serem aleatórias, *outliers* em $\mathbf{x}$ afectam directamente as equações (4.5) (ou as equações (4.6)). As estimativas tornam-se pouco credíveis e um só desses *outliers* pode conduzir a estimadores sem sentido, o que é equivalente a dizer que o ponto de rotura relativamente às variáveis explicativas é zero.

E é pena que estimadores tão ricos e abrangentes sofram deste defeito. Embora o seu interesse para a obtenção de estimadores de regressão, no contexto que acaba de ser apresentado, seja actualmente considerado apenas de valor histórico, a verdade é que a estrutura conceptual dos estimadores-M proporciona a construção de novos e melhores estimadores, como se verá a seguir.

4.3.3 Estimadores-M generalizados

Para contornar esta dificuldade dos estimadores-M foi pensada uma estratégia que consiste em associar um peso w_i a $\mathbf{x}_i$ com o objectivo

de limitar a influência de $\mathbf{x}_i$, sendo que w_i deve ser tanto menor quanto mais extremo, mais influente, for $\mathbf{x}_i$.

Assim surgiram os estimadores-M generalizados (estimadores-GM), por vezes designados por estimadores de influência limitada. Duas famílias deste tipo de estimadores bem conhecidas (ver Hampel *et al.*, 1986, Cap. 6) são:

- Estimadores de Mallows que são solução das equações

$$\sum_{i=1}^{n} w_i(\mathbf{x}_i)\psi\left(\frac{e_i}{\hat{\sigma}}\right)\mathbf{x}_i = \mathbf{0},$$

  equação do tipo dos mínimos quadrados pesados.
- Estimadores de Schweppe, que são solução das equações

$$\sum_{i=1}^{n} w_i(\mathbf{x}_i)\psi\left(\frac{e_i}{w_i(\mathbf{x}_i)\hat{\sigma}}\right)\mathbf{x}_i = \mathbf{0},$$

  perseguindo a ideia de dar pesos não só a $\mathbf{x}_i$ como também aos resíduos de forma a que pesos menores são dados a resíduos e_i correspondentes a $\mathbf{x}_i$ com pesos pequenos.

Os estimadores-GM têm algumas propriedades com muito interesse (Maronna et al., 2006), mas apresentam várias dificuldades, sendo talvez a mais impressionante a que se refere ao ponto de rotura, que se mostra ser decrescente para zero com o valor de p.

Esta desvantagem dos estimadores de influência limitada não arrefece a pesquisa para encontrar novos e melhores estimadores, em particular, e naturalmente, estimadores que se desejam com elevado ponto de rotura (independentemente do número de variáveis explicativas). Os estimadores da mínima mediana dos quadrados (ou LMS de *Least Median of Squares*), dos mínimos quadrados aparados (ou LTS de *Least Trimmed Squares*) e estimadores-S são exemplos de estimadores caracterizados por possuírem ponto de rotura elevado e vão ser apresentados a seguir.

4.3.4 Mínima mediana dos quadrados (LMS)

Os vários estimadores apresentados anteriormente foram obtidos pensando na função objectivo dos resíduos dos mínimos quadrados que

é preciso minimizar, e alterando a parte que envolve os resíduos, $\rho(e_i) = e_i^2$, mas mantendo o somatório. O estimador LMS foi criado, Rousseeuw (1984), ao ser perseguida uma ideia diferente: em vez de alterar $\rho(e_i) = e_i^2$ porque não modificar também o somatório substituindo-o por outro operador que seja robusto, como a mediana, por exemplo? O resultado é a definição de estimador LMS, ou seja o estimador que minimiza a mediana dos quadrados dos resíduos:

$$\hat{\boldsymbol{\beta}}_{LMS} \text{ minimiza } \text{med}_i\,(e_i)^2\,.$$

A mesma definição pode ser conseguida seguindo o raciocínio: se os mínimos quadrados minimizam a soma dos quadrados dos resíduos então minimizam a média dos quadrados dos resíduos e, como a média é muito sensível à presença de *outliers*, o melhor é substituí-la pela mediana que resiste bem à sua presença.

Quando a regressão envolve apenas uma variável explicativa a recta LMS é a recta que é paralela aos lados de uma faixa que contém a maioria dos dados e passa pelo centro da faixa. Esta visão intuitiva é útil e imediatamente extensível ao caso de duas e mais variáveis.

O estimador LMS é robusto em relação a *outliers* em y e em $\mathbf{x}$ e tem o ponto de rotura máximo (50%). Contudo prova-se (Rousseeuw e Leroy, 1987) que a sua eficiência assintótica é zero, o que desaconselha o uso inadvertido do método LMS para estimar a regressão.

4.3.5 Mínimos quadrados aparados (LTS)

Em face das limitações, em termos de eficiência, do estimador LMS é natural continuar a procura para encontrar um estimador que não tenha essas limitações. Rousseeuw (1984) apresenta também o estimador LTS construído com base na minimização da soma aparada dos quadrados dos resíduos, isto é,

$$\hat{\boldsymbol{\beta}}_{LTS} \text{ minimiza } \sum_{i=1}^{h} e_{(i)}^2,$$

onde $e_{(1)}^2 \leq e_{(2)}^2 \leq \cdots \leq e_{(n)}^2$, são os quadrados dos resíduos ordenados e o somatório não inclui os $n - h$ maiores resíduos absolutos que foram aparados.

A constante h que representa o número de observações (correspondentes aos menores valores dos quadrados dos resíduos) a incluir no

somatório determina o ponto de rotura, pois as restantes observações podem ser alteradas sem perturbar o estimador LTS.

Rousseeuw e Leroy (1987) mostram que o valor máximo do ponto de rotura em dimensão finita é $([(n-p)]/2]+1)/n$, o qual é atingido quando $h = [n/2] + [(p-1)/2]$, e ainda que para h genérico maior do que este valor o ponto de rotura é aproximadamente $(n-h)/n$. É claro que quando $h = n$ o ponto de rotura é zero pois o estimador LTS coincide com o estimador dos mínimos quadrados.

Para obter a solução exacta do estimador LTS basta aplicar o método dos mínimos quadrados a todos os subconjuntos com h observações, em número de $\binom{n}{h}$, e escolher o subconjunto com a menor soma de quadrados dos resíduos, ao qual corresponde a solução final. Porém, este procedimento não é geralmente seguido porque é muito dispendioso em termos computacionais e em vez dele usam-se procedimentos que conduzem a soluções aproximadas. Entre outros, Hawkins (1994) fornece um algoritmo para obter soluções aproximadas. Mais recentemente Rousseeuw e Van Driessen (1999) apresentaram um algoritmo de computação rápida que está implementado em R, `ltsReg(robustbase)`.

O estimador LTS, embora mais eficiente do que o estimador LMS, tem ainda eficiência relativamente baixa. No entanto é conhecido o seu papel, como estimador inicial, no processo de construção de um estimador-M, altamente eficiente, com elevado ponto de rotura e função de influência limitada (Coakley e Hettmansperger, 1993).

4.3.6 Estimadores-S

O nome desta família de estimadores deriva do facto do processo de construção destes estimadores integrar um estimador de escala (*scale*).

Pensando que o modo de operar dos mínimos quadrados, minimizar a soma dos quadrados dos resíduos equivale a minimizar a sua variância, e portanto o desvio padrão (estimador de escala) dos resíduos, porque não aproveitar esta linha de raciocínio e minimizar outro qualquer estimador de escala? É precisamente esta a estratégia usada para definir os estimadores-S, introduzidos por Rousseeuw e Yohai (1984).

Assim o estimador-S é $\hat{\boldsymbol{\beta}}_S$ tal que

$$\hat{\boldsymbol{\beta}}_S \text{ minimiza } S\left(e_1(\hat{\boldsymbol{\beta}}), \ldots, e_n(\hat{\boldsymbol{\beta}})\right), \tag{4.7}$$

onde $S\left(e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})\right)$ é o estimador-M de escala, solução de

$$\sum_{i=1}^{n} \chi\left(\frac{e_i}{\hat{S}}\right) = nK,$$

e K é precisamente a constante de consistência associada à função χ que figura na equação (3.28), sendo aí denotada por β. Neste processo é preciso escolher a função χ. Na Secção 3.2.3 foram apresentados vários exemplos de funções χ adequadas quando o modelo central de escala em causa é o modelo normal (neste contexto é o modelo central associado à distribuição dos erros). Em consequência há que escolher para além da forma da função χ, as constantes de afinação envolvidas para depois ser possível determinar K. Para escolher as constantes de afinação é preciso jogar, como habitualmente, com o equilíbrio que se pretende entre o ponto de rotura e a eficiência. A Tabela 3.4 apresenta um conjunto de resultados que ajudam a fazer esta escolha. No contexto dos estimadores-S, Rousseeuw e Leroy (1987) recomendam que se use a função χ dada por (3.36), cuja derivada é a função ψ biponderada de Tukey, fixando o ponto de rotura em 50%, o que pela Tabela 3.4, determina que a constante de afinação é $r = 1.548$ e $K = 1/2$.

O tempo consumido e as complicações para o cálculo destes estimadores, situação que também se verifica para os estimadores LTS, levam Rousseeuw e Leroy (1987) a eleger o estimador LMS, de entre os três estimadores com elevado ponto de rotura aqui estudados, como aquele que menos esforço computacional requer. Contudo Ruppert (1992) é de opinião, baseada em várias análises com dados simulados e dados reais, de que os estimadores-S têm um desempenho mais satisfatório.

É interessante notar que não só os estimadores dos mínimos quadrados como vários outros estimadores que foram descritos até agora podem ser vistos como casos particulares de (4.7), considerando S um estimador genérico de escala. De facto, o método LAD minimiza o desvio médio dos resíduos e os estimadores-M minimizam uma espécie de desvio padrão ponderado dos resíduos. Para o estimador LMS já não é tão evidente mas sucede o mesmo, pois $\text{med}(e_i^2)$ é proporcional

ao quadrado de MAD(e_i), assumindo como é usual que os erros têm localização conhecida e igual a zero. Também a soma aparada dos resíduos, que se minimiza para obter o estimador LTS pode ser vista como uma modificação da variância aparada.

O interesse principal que a observação anterior tem é o de permitir indicar, para cada um dos métodos referidos, uma estimativa directa e adequada do parâmetro σ, que será precisamente o valor mínimo da função objectivo em (4.7).

4.3.7 Estimadores-MM

Duas das propriedades que um bom estimador robusto deve possuir são, como se viu no final do Capítulo 3, ter um elevado ponto de rotura e uma grande eficiência em relação ao modelo central ou modelo de referência que no caso da regressão é invariavelmente o modelo normal. Dos estimadores até agora apresentados nenhum deles satisfaz as duas propriedades:

(i) O estimador dos mínimos quadrados tem eficiência máxima mas o ponto de rotura é igual a zero.
(ii) O estimador LAD é pouco eficiente e tem ponto de rotura igual a zero (em $\mathbf{x}$).
(iii) Os estimadores LMS, LTS e S têm baixa eficiência mas elevado ponto de rotura.
(iv) Os estimadores-M podem ter grande eficiência, dependendo de ψ (ou de ρ) mas o seu ponto de rotura é zero (em $\mathbf{x}$).
(v) Os estimadores-GM têm baixa eficiência (depende da distribuição de $\mathbf{x}$) e o seu ponto de rotura é decrescente para zero com o número de variáveis.

Na linha de preocupações em harmonizar um elevado ponto de rotura com uma alta eficiência surge uma tentativa, Rousseeuw (1984) com a proposta do método LTS, que não resultou, e depois Yohai (1987) que propôs os estimadores-MM que satisfazem aquelas duas propriedades.

Em termos práticos os estimadores-MM não são mais do que estimadores-M calculados a partir de estimativas iniciais convenientes. Como se descreveu em 4.3.2 um estimador-M de regressão é aquele

$\hat{\boldsymbol{\beta}}$ tal que

$$\hat{\boldsymbol{\beta}} \text{ minimiza } \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\hat{\sigma}}\right),$$

onde $\hat{\sigma}$ é um estimador de escala, sendo que nas propostas iniciais se recomendava o uso do MAD para esse efeito. Exceptuando os casos de valores baixos de p a obtenção do mínimo absoluto é tarefa muito difícil. No entanto, o que Yohai (1987) provou foi que se for encontrado um "bom" mínimo local então é possível conseguir simultaneamente um elevado ponto de rotura e uma alta eficiência por escolha adequada da função ρ que deve ser limitada.

Um "bom" mínimo local pode ser encontrado aplicando o método IRWLS (mínimos quadrados iterativamente pesados, ver o que foi dito a respeito deste método no Capítulo 3, página 119) a uma estimativa inicial "competente" $\hat{\boldsymbol{\beta}}_0$. Usualmente recorre-se aos estimadores--S para determinar a estimativa inicial adequada, o que equivale a usar para $\hat{\sigma}$ um estimador-M de escala e a garantir à partida que o ponto de rotura é elevado. Em consequência pode dizer-se que um estimador-MM corresponde ao mínimo local da função objectivo de um estimador-M que está mais próximo do estimador-S inicial.

Programas para o cálculo de estimadores-MM estão implementados em R e S-Plus o que faz destes estimadores uma preferência dos utilizadores. Ao longo deste texto foram usados, sempre que possível, estimadores-MM para resolver vários problemas de estimação.

Para terminar esta secção dedicada aos métodos robustos de regressão, apresenta-se um último método que não segue a linha de desenvolvimento dos anteriores. Trata-se de um método recente que assenta na ideia de profundidade já conhecida desde os anos setenta do século XX noutros contextos que não o de regressão.

4.3.8 Regressão mais profunda

Este método está ligado à noção de profundidade de um ponto relativamente a um conjunto de dados, introduzida por John Tukey (Tukey, 1975).

É conveniente começar por introduzir a profundidade de um ponto usando um espaço bidimensional. A profundidade de um ponto P do plano (também chamada profundidade de localização, profundidade

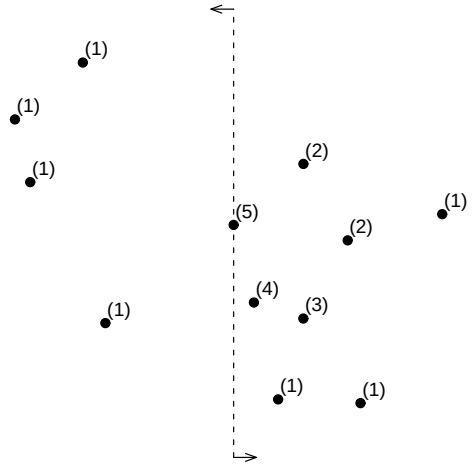


Figura 4.9 *Exemplo de um conjunto de dados bivariado com os valores da profundidade de Tukey indicados entre parêntesis. Todos os pontos situados sobre a fronteira do invólucro convexo que contém os dados têm profundidade 1. Há um único ponto do conjunto de dados que tem profundidade 5 (a tracejado representa-se uma das rectas que determina a profundidade desse ponto). De notar que o sistema de coordenadas dos pontos é totalmente irrelevante.*

semi-espaço ou profundidade de Tukey) relativamente a um conjunto de dados bivariados, S_n, representado em $\mathbb{R}^2$ por n pontos, é o número mínimo de pontos de S_n, situado num dos dois lados de qualquer recta passando por P, incluindo os próprios pontos que estão sobre a recta. Um ponto que tenha profundidade máxima é chamado mediana de Tukey, o que generaliza a noção de mediana conhecida no caso univariado ao caso bivariado.[11] A profundidade de um ponto é assim uma medida quantitativa da centralidade do ponto relativamente ao conjunto de dados ou à distribuição de probabilidade que lhe está subjacente. Na Figura 4.9 apresenta-se um exemplo.

Desta forma, e como o conceito de profundidade é imediatamente extensível a espaços de dimensão maior do que dois, pode concluir-se

[11]Para um conjunto de dados univariado $S_n = \{x_1, \ldots, x_n\}$ a profundidade de um qualquer ponto x é dada por $\min\{\#\{x_i \leq x\}, \#\{x_i \geq x\}\}$ e a mediana corresponde ao ponto, ou pontos, com profundidade máxima.

que a profundidade induz uma ordem no conjunto de dados multivariados, generalizando o conceito de ordem, natural em dados univariados, a dados multivariados, um velho sonho de todo o analista que trabalha com análise multivariada.

A generalização da mediana e da noção de ordem a conjuntos de dados multivariados proporciona uma maior abrangência da acção da análise multivariada, uma vez que torna possível o tratamento não paramétrico dos dados e a resolução de problemas como testes de hipóteses e estimação, nomeadamente na área da regressão robusta.

Levando a ideia de profundidade para o campo univariado verifica-se, como já se referiu, que o ponto de profundidade máxima de um conjunto de dados univariados é a mediana, que como se sabe é uma medida de localização altamente robusta. São conhecidas outras funções de profundidade, sendo muitas delas robustas, o que é muito útil para o estudo de problemas reais envolvendo dados em espaços multidimensionais onde proliferam *outliers*. É o caso da regressão mais profunda cuja definição assenta na noção de profundidade de regressão relativamente a um conjunto de dados (Rousseeuw e Hubert, 1999). Esta representa o envolvimento que uma regressão tem com os dados que a rodeiam.

Definição 4.1. *Profundidade de regressão, de uma regressão identificada por um conjunto de coeficientes de regressão $\boldsymbol{\beta}$, relativamente a um conjunto de dados S_n, é o número mínimo de observações de S_n que é preciso remover para que a regressão definida por $\boldsymbol{\beta}$ deixe de ser considerada um ajustamento, isto é, o ajustamento deixe de ter sentido.*

No caso bivariado (regressão linear simples), a definição é equivalente a dizer que a profundidade de regressão de uma recta em relação ao conjunto de pontos S_n é o menor número de pontos que é preciso remover para que a recta possa rodar (em torno de algum ponto) para a posição vertical sem passar por qualquer ponto de S_n (ver um exemplo na Figura 4.10). Uma recta vertical é uma regressão inútil uma vez que não estabelece qualquer relação de dependência entre as variáveis nem pode servir para predizer valores da variável dependente.

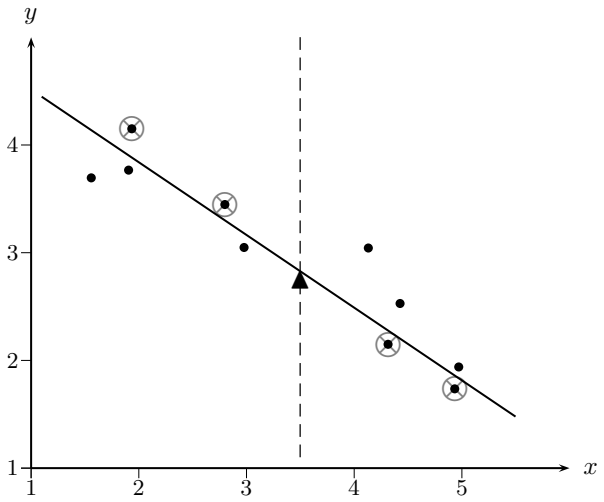


Figura 4.10 *Conjunto de pontos bivariado e recta com profundidade de regressão igual a 4, com indicação dos pontos que têm de ser removidos para tornar a recta vertical (com o fulcro da rotação representado).*

De todas as regressões que se possam produzir aquela que tiver a profundidade máxima é a regressão mais profunda e o método usado para a conseguir é o método da regressão mais profunda. A regressão mais profunda representa afinal o grau de linearidade intrínseco aos dados e é uma generalização do conceito de mediana univariada.

A regressão mais profunda é robusta (Van Aelst e Rousseeuw, 2000) em relação a *outliers*, em particular a pontos de *leverage* e o seu ponto de rotura é no mínimo 1/3. Aqueles autores também apresentam as funções de influência dos funcionais correspondentes, os quais são consistentes para a mediana condicional, $\text{med}(y|\mathbf{x})$, e mostram que elas são limitadas e têm baixa sensibilidade.

Uma estratégia para encontrar a regressão mais profunda no caso bivariado, ou seja na regressão linear simples, consiste em considerar as rectas que se podem definir com todos os pares de pontos de S_n, calcular as respectivas profundidades, e escolher a recta de profundidade máxima. No caso de k variáveis explicativas, com $k > 1$, pode-se pensar nos hiperplanos que se podem definir com todos os

subconjuntos de $k+1$ pontos, mas este método tem uma grande complexidade computacional. Van Aelst *et al.* (2002) apresentam um algoritmo para computação eficiente da regressão mais profunda em mais do que duas dimensões e derivam testes e regiões de confiança simultâneas para os verdadeiros valores dos parâmetros do modelo.

4.4 Comparação de estimadores

Para além dos estimadores que aqui se descrevem (LAD, M, GM, LMS, LTS, S e regressão mais profunda), considerados os mais populares na literatura, existem outros potenciais candidatos a estimadores robustos que não foram considerados.

A questão que agora se coloca é a de comparar os estimadores descritos para seleccionar os melhores. O processo não é fácil pois existem vários critérios de comparação e os próprios estimadores podem ter comportamentos diferentes, quando a análise é feita considerando a variável resposta ou as variáveis explicativas, ou ainda quando a experimentação é ou não delineada.

A Tabela 4.8 contém características relevantes dos vários estimadores que são indispensáveis para a sua comparação global. Na tabela pode ver-se que os estimadores-MM são de facto os que mais interessam na prática pois satisfazem os dois critérios com maior peso na avaliação dos estimadores robustos, elevado ponto de rotura e alta eficiência.

Apresentados os estimadores é indispensável saber como prosseguir a análise do modelo com vista à realização das habituais inferências sobre os parâmetros. Isso vai ser explicado na secção seguinte.

4.5 Análise dos resultados de uma regressão robusta

4.5.1 Estimação de σ e coeficiente de determinação

Quando se ajusta um modelo de regressão a um conjunto de dados é habitual calcular, para além das estimativas $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$ da regressão

Tabela 4.8 *Propriedades mais relevantes dos métodos de regressão estudados.*

Método	Robustez	Ponto de rotura	Função de influência	Estabilidade local	Solução	Eficiência (modelo normal)
Mínimos quadrados	Não	0	Ilimitada	Sim	Única	Máxima
Mínimos desvios absolutos (LAD)	Em y: Sim Em $\mathbf{x}$: Não	$\leq 1/4$ 0	Limitada em y Ilimitada em $\mathbf{x}$	Não	Pode ter múltiplas	Baixa
Estimadores-M	Em y: Sim Em $\mathbf{x}$: Não	$\leq 1/2$ 0	Limitada em y Ilimitada em $\mathbf{x}$	Sim[(1)]	Pode ter múltiplas	Depende de ψ
Estimadores-M generalizados	Sim[(2)]	$1/p$	Limitada	Sim[(1)]	Pode ter múltiplas	Depende de ψ e da dist. de $\mathbf{x}$
Mínima mediana dos quadrados (LMS)	Sim	1/2	Não tem	Não	Pode ter múltiplas	Nula
Mínimos quadrados aparados (LTS)	Sim	$\leq 1/2$ [(3)]	Limitada	Não	Pode ter múltiplas	Muito baixa
Estimadores-S	Sim	$\leq 1/2$ [(4)]	Limitada em y Ilimitada em $\mathbf{x}$	Sim[(4)]	Pode ter múltiplas	Baixa
Estimadores-MM	Sim	$\leq 1/2$ [(4)]	Limitada em y Ilimitada em $\mathbf{x}$	Sim	Pode ter múltiplas	Elevada
Regressão mais profunda	Sim	1/3	Limitada	Não	Pode ter múltiplas	Baixa[(5)]

[(1)] ψ contínua. [(2)] Depende de p. [(3)] Depende de h. [(4)] Depende de S. [(5)] Inferior à de LAD.

linear propriamente dita, a estimativa do parâmetro σ, $\hat{\sigma}$, que se costuma designar por erro padrão dos resíduos, e o coeficiente de determinação, R^2. É importante que o utilizador de um método de regressão robusta também tenha acesso a estas duas quantidades.

Quanto à estimação de σ, ela foi já brevemente abordada no final da Secção 4.3.6 a propósito dos estimadores-S. Além do método aí indicado é sempre possível, após o ajustamento do modelo de regressão, calcular os resíduos e estimar o respectivo parâmetro de escala através de um qualquer estimador robusto dos estudados no Capítulo 2 ou no Capítulo 3.

Quanto ao coeficiente de determinação, considere-se a variante ajustada (mais adequada para problemas de regressão múltipla e equivalente na regressão simples) dada por

$$R^2_{aj} = 1 - \frac{s_n^2(e_i)}{s_n^2(y_i)}.$$

Para obter uma versão robusta deste coeficiente basta considerar em vez da variância amostral um estimador robusto de escala elevado ao quadrado. Em Croux e Dehon (2003) analisam-se detalhadamente as propriedades de coeficientes de determinação robustos obtidos por este processo. Esses autores recomendam que para substituir $s_n^2(y_i)$ se use o método de regressão usado para estimar o modelo completo mas aplicado a um modelo só com ordenada na origem, $y_i = \alpha_0 + \varepsilon_i$, uma vez que para este modelo o erro padrão dos resíduos fornece uma estimativa do parâmetro de escala de Y (já agora refere-se que $\hat{\alpha}_0$ é uma estimativa do parâmetro de localização de Y).

4.5.2 Intervalos de confiança e testes de hipóteses

Como é sabido o método dos mínimos quadrados, na situação de normalidade dos erros do modelo de regressão, é equivalente ao método da máxima verosimilhança. Esta equivalência torna possível a realização de inferências sobre os parâmetros do modelo. De facto, a Tabela 4.3 tem a chave do problema pois a expressão

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{\mathrm{var}}(\beta_i)}} \sim t_{n-p}$$

permite obter intervalos de confiança para β_i e testar hipóteses sobre β_i, sendo a hipótese $H_0\colon\ \beta_i = 0$ contra $H_1\colon\ \beta_i \neq 0$ a mais comum e

mais útil e cujo teste se baseia agora na estatística

$$\frac{\hat{\beta}_i}{\sqrt{\widehat{\text{var}}(\beta_i)}} \sim t_{n-p}.$$

A questão que agora se levanta é a de saber como realizar essas inferências no caso de estimação dos parâmetros do modelo de regressão usando métodos robustos. Isto é, como obter intervalos de confiança e realizar testes de hipóteses relativos aos verdadeiros valores dos parâmetros? Em segundo lugar interessa saber se as estimativas do erro padrão são robustas e se os intervalos de confiança obtidos e os resultados dos testes realizados também são robustos.

Estas questões foram já abordadas num contexto geral na Secção 3.4. Particularizando o que aí foi dito para os estimadores dos parâmetros da regressão, pode afirmar-se que se os estimadores robustos, $\hat{\boldsymbol{\beta}}_R$ (onde R se refere a um dado método robusto) do parâmetro $\boldsymbol{\beta}$ tiverem distribuição assintótica normal (o que acontece com todos os considerados com excepção do LMS), então pode usar-se a expressão

$$\frac{\hat{\beta}_{R,i} - \beta_i}{\sqrt{\widehat{\text{var}}(\beta_{R,i})}} \overset{a}{\sim} \mathcal{N}(0,1)$$

ou a variante corrigida

$$\frac{\hat{\beta}_{R,i} - \beta_i}{\sqrt{\widehat{\text{var}}(\beta_{R,i})}} \overset{a}{\sim} t_{n-p}.$$

Resta saber como estimar o erro padrão. Isso pode ser feito usando um dos procedimentos gerais baseados na função de influência e que também foram apontados na Secção 3.4. Uma descrição mais detalhada tornar-se-ia fastidiosa, refere-se apenas que em relação aos estimadores-MM, que são os utilizados daqui em diante, Yohai *et al.* (1991) apresentam todos os resultados necessários (que além disso se encontram implementados na função `lmRobMM` do S-Plus). Uma outra abordagem foi seguida por Croux *et al.*, (2004), que apresentam expressões para estimação dos erros padrão que produzem estimativas consistentes mesmo quando os erros são heterocedásticos ou autocorrelacionados. Estas são as estimativas implementadas na função `lmrob(robustbase)` do R.

Quanto à validade das aproximações para amostras de baixa dimensão o que se pode afirmar é que neste caso, e à semelhança da

estimação da localização, as aproximações são razoáveis a partir de dimensões da amostra relativamente baixas. Prova-se em qualquer dos casos que os intervalos de confiança e os testes resultantes são robustos.

Uma outra abordagem consiste na utilização do método *bootstrap*, mas como se explicou na Secção 3.4.3 isso não pode ser feito directamente. Salibian-Barrera e Zamar (2002) apresentam as necessárias correcções.

Exemplo 4.4. Neste exemplo repete-se o estudo de simulação descrito no Exemplo 4.1 (página 165) mas usando na estimação dos parâmetros e na construção dos intervalos de confiança o método MM em vez do método dos mínimos quadrados.[12] Os resultados encontram-se nas Tabelas 4.9 e 4.10. O método designado por método de regressão MM ponderado corresponde à aplicação do método de regressão MM ao modelo transformado pela multiplicação à esquerda por $\mathbf{V}^{-1/2}$ tal como se explicou no Exemplo 4.1, página 170.

Tabela 4.9 *Resultados da simulação relativa ao modelo* $y_1 = x_i + \varepsilon_i$ *(método de regressão MM).*

	NH	NnH	T5H	T5nH
$\hat{\beta}_0$	0.004 (0.168)	0.013 (0.371)	−0.020 (0.151)	−0.005 (0.317)
$IC(\beta_0)$	0.672 (0.118)	1.532 (0.420)	0.598 (0.125)	1.293 (0.331)
$\hat{\alpha}$	0.943	0.946	0.939	0.942
$\hat{\beta}_1$	0.998 (0.052)	0.996 (0.189)	1.000 (0.045)	1.006 (0.154)
$IC(\beta_1)$	0.203 (0.033)	0.764 (0.264)	0.181 (0.035)	0.622 (0.204)
$\hat{\alpha}$	0.944	0.937	0.946	0.928

As Tabelas 4.9 e 4.10 devem ser comparadas com as Tabelas 4.2 e 4.4, respectivamente, originando os seguintes comentários:

- Em termos da média das estimativas de β_0 e β_1 obtêm-se, tal como com o método dos mínimos quadrados, valores muito próximos dos verdadeiros, indicando que os estimadores-MM

[12] Foi utilizada a função `lmrob` do R com todas as opções por defeito.

Tabela 4.10 *Resultados da simulação relativa ao modelo* $y_1 = x_i + \varepsilon_i$ *(método de regressão MM ponderado).*

	NnH	T5nH
$\hat{\beta}_0$	0.009 (0.249)	−0.003 (0.225)
$IC(\beta_0)$	0.978 (0.199)	0.877 (0.207)
$\hat{\alpha}$	0.941	0.932
$\hat{\beta}_1$	0.995 (0.137)	0.999 (0.119)
$IC(\beta_1)$	0.538 (0.073)	0.481 (0.078)
$\hat{\alpha}$	0.946	0.950

destes parâmetros são centrados em qualquer das situações (quer seja usando o método ponderado ou não).

- Em relação aos erros padrão de β_0 e β_1 e aos intervalos de confiança respectivos observa-se como esperado uma perda de eficiência sob as distribuições normais (caso NH na Tabela 4.9 e caso NnH na Tabela 4.10) mas que é muito baixa (2% a 4%) e uma melhoria de eficiência sob as distribuições t (caso T5H na Tabela 4.9 e caso T5nH na Tabela 4.10) mais substancial (8% a 11%).

- Conclui-se também que o método de regressão MM ordinário sofre dos mesmos problemas que o método dos mínimos quadrados ordinários quando aplicado a dados não homogéneos, ou seja, intervalos de confiança com comprimento muito superior ao necessário (pouco eficientes) embora mantendo o nível de confiança nominal.

- Finalmente observa-se que o método de regressão MM ponderado é um método robusto indicado quando os erros não são homogéneos, conduzindo a intervalos de confiança muito melhores e perdendo pouca eficiência quando os dados são normais.

4.5.3 Exemplos

Nesta secção exemplifica-se o uso da regressão robusta (MM) e faz-se a comparação com os resultados dos mínimos quadrados recorrendo a dois conjuntos de dados reais, o primeiro relativo a uma regressão linear simples e o segundo a uma regressão linear múltipla.

Exemplo 4.5. A medição rigorosa da obliquidade da eclíptica[13] é uma tarefa que tem interessado os astrónomos desde sempre, não só para compreenderem certos factos de interesse mais imediato, como as estações do ano, mas também para poderem investigar a evolução de fenómenos astronómicos complexos.

Os dados que se encontram na Tabela 4.11 foram compilados em 1740 pelo astrónomo Jacques Cassini e constituem uma sequência de 15 observações da obliquidade da eclíptica efectuadas ao longo de praticamente 2000 anos.

Tabela 4.11 *Dados relativos à obliquidade da eclíptica ao longo de cerca de 2000 anos.*

Ano (x)	Obliquidade (y)	Ano (x)	Obliquidade (y)
-140	23.853	1500	23.488
-140	23.856	1570	23.499
390	23.500	1570	23.525
880	23.583	1600	23.517
1070	23.567	1656	23.484
1300	23.533	1672	23.482
1460	23.500	1738	23.472
1500	23.473		

Os dados reflectem o rigor da medição, as dificuldades da medição (por três vezes, 140 A.C., 1500 e 1570 são indicados, para o mesmo ano, valores diferentes, presumivelmente obtidos por astrónomos diferentes) e mostram uma tendência decrescente da obliquidade durante o período de quase 2000 anos, começando em cerca de 23.8^o e

[13] A eclíptica é definida como a circunferência imaginária correspondente à trajectória aparente do Sol na esfera celeste. A obliquidade é o ângulo que o plano da eclíptica forma com o plano do equador da Terra.

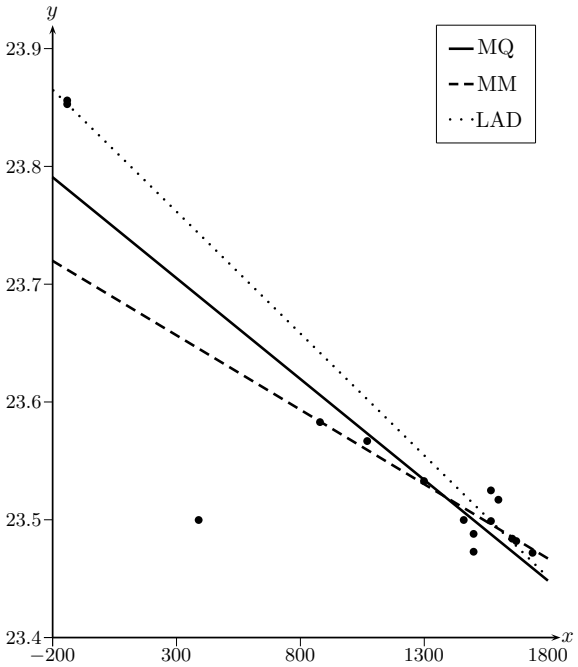


Figura 4.11 *Representação dos dados relativos à obliquidade da eclíptica com recta dos mínimos quadrados (MQ), recta MM e recta LAD.*

chegando a 23.5^o aproximadamente. Esta diferença tão magra, mas tão importante e significativa do ponto de vista astronómico, reflecte a delicadeza dos dados e indicia eventuais dificuldades na análise. Vejamos como é que os métodos estatísticos podem ajudar a interpretar estes dados.

O gráfico da Figura 4.11 confirma que as duas primeiras observações, e eventualmente a terceira observação, estão muito afastadas do corpo central dos dados, sugerindo que poderão ser *outliers*. Vê-se ainda que a recta dos mínimos quadrados é claramente atraída pelas duas primeiras observações e que a recta robusta obtida com base no método MM refreia, e bem, essa atracção.[14]

[14]Como curiosidade inclui-se também a recta dos mínimos desvios absolutos, a qual exibe um comportamento ainda pior que a dos mínimos quadrados, sendo

Tabela 4.12 *Obliquidade da eclíptica: parâmetros do modelo de regressão e respectivos intervalos de confiança a 95%.*

	Mínimos quadrados	Regressão MM
$\hat{\beta}_0$	23.76	23.69
$IC_{95\%}(\beta_0)$	(23.681,23.832)	(23.669,23.711)
	$(l = 0.151)$	$(l = 0.042)$
$\hat{\beta}_1$	-1.713×10^{-4}	-1.264×10^{-4}
$IC_{95\%}(\beta_1)$	$(-2.27, -1.15) \times 10^{-4}$	$(-1.44, -1.08) \times 10^{-4}$
	$(l = 1.12 \times 10^{-4})$	$(l = 0.36 \times 10^{-4})$
$\hat{\sigma}$	6.30×10^{-2}	2.48×10^{-2}

Na Tabela 4.12 apresentam-se os valores das estimativas dos parâmetros do modelo de regressão

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, 15$$

obtidas pelo método dos mínimos quadrados e pelo método MM, bem como os respectivos intervalos de confiança a 95%, calculados assumindo que $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.[15]

Apesar de $\hat{\beta}_1$ estar extremamente próximo de zero,[16] a hipótese $H_0 \colon \beta_1 = 0$ (contra a alternativa bilateral) é rejeitada com base num valor-p praticamente nulo (ver *outputs* dos dois métodos apresentados nas Figuras 4.12 e 4.13, respectivamente). Note-se que o facto de os intervalos de confiança correspondentes terem o mesmo sinal nos dois

totalmente dominada pelas duas primeiras observações.

[15] Como foi já referido, em relação ao método dos mínimos quadrados os resultados só são válidos se esta hipótese de trabalho for exactamente verificada, enquanto que para a validade dos resultados obtidos pelo método MM apenas é necessário que ela se verifique aproximadamente.

[16] Há aqui uma questão de escala em x que é importante não esquecer, por exemplo a estimativa $\hat{\beta}_1 = -1.264 \times 10^{-4}$ tem a ver com o decréscimo anual da obliquidade, o que corresponde a um decréscimo de 1.264×10^{-2} graus por século, ou a um decréscimo de 0.1264 graus por milénio.

```
lm(formula = obli ~ date)

Residuals:
      Min        1Q    Median        3Q       Max
-0.189943 -0.009234  0.010914  0.023641  0.075272

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.376e+01  3.484e-02 681.974  < 2e-16 ***
date        -1.713e-04  2.622e-05  -6.533 1.90e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06297 on 13 degrees of freedom
Multiple R-Squared: 0.7665,     Adjusted R-squared: 0.7486
F-statistic: 42.69 on 1 and 13 DF,  p-value: 1.902e-05
```

Figura 4.12 *Obliquidade da eclíptica: resultados do ajustamento pelo método do mínimos quadrados obtidos no R (output parcial da função* `lm`*).*

```
lmrob(formula = obli ~ date)

Weighted Residuals:
       Min         1Q     Median         3Q        Max
-0.1454352 -0.0066462 -0.0005108  0.0159905  0.1435853

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.369e+01  9.637e-03 2458.78  < 2e-16 ***
date        -1.264e-04  8.247e-06  -15.32 1.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.02478
```

Figura 4.13 *Obliquidade da eclíptica: resultados do ajustamento pelo método MM obtidos no software R (output parcial da função* `lmrob` *do package* `robustbase`*).*

extremos é concordante com esta conclusão. Este teste sustenta assim a hipótese de uma relação linear entre as variáveis que deram origem a estes dados.

Os dois métodos usados na estimação dão resultados de acordo com os seus objectivos (o valor absoluto de $\hat{\beta}_1$ é mais pequeno no caso MM, reflectindo a menor importância que este método dá aos possíveis *outliers*), notando-se que a maior discrepância ocorre ao nível dos intervalos de confiança (o comprimento dos intervalos com base nos estimadores MM é inferior a 1/3 do comprimento dos correspondentes intervalos dos mínimos quadrados), em consequência de uma maior precisão associada à estimação MM dos parâmetros do modelo.

É evidente que a análise do modelo não está completa sem a correspondente análise dos resíduos, especialmente importante quando se usa o método dos mínimos quadrados. Na Figura 4.14 mostra-se o gráfico dos resíduos studentizados externamente (os mais adequados para detectar *outliers*, para mais detalhes ver Neter *et al.*, 1996). A conclusão é que a terceira observação é um *outlier* importante e deve ser retirada pois está certamente a prejudicar os resultados. Se isso for feito, obtêm-se para estimativas dos parâmetros,

$$\hat{\beta}_0 = 23.81, \quad \hat{\beta}_1 = -2.026 \times 10^{-4} \quad \text{e} \quad \hat{\sigma} = 2.57 \times 10^{-2}$$

e um valor de $R^2 = 0.961$. O gráfico de resíduos studentizados externamente para este novo modelo é apresentado na Figura 4.15 indicando que não há nenhum resíduo especialmente elevado. A única indicação de que algo pode não estar bem tem a ver com a existência de três grupos de resíduos mas este facto pode passar despercebido ao utilizador mais incauto, especialmente tendo em conta o reduzido número de observações, propício à ocorrência de fenómenos deste tipo sem grande significado. Conclui-se assim ser altamente plausível que um utilizador, mesmo cuidadoso, aceite este modelo, que acaba por ser ainda pior, em termos de ajustamento global que o modelo com todas as observações (graficamente esta nova recta situa-se muito próximo da recta dos mínimos desvios absolutos, ver Figura 4.11).

Por último resta observar o que acontece com os resíduos calculados após o ajustamento da recta de regressão MM, representados na Figura 4.16. As observações 1, 2 e 3 surgem claramente como *outliers*, mas os restantes resíduos são bem comportados. Obviamente que se se tivesse em algum momento decidido usar o método dos mínimos quadrados sem as três primeiras observações se obteriam resul-

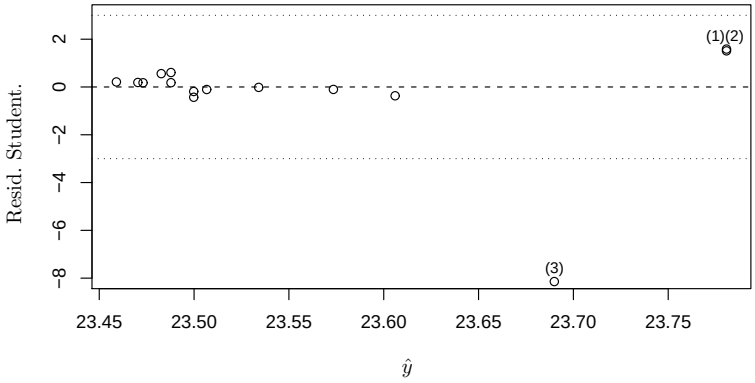


Figura 4.14 *Obliquidade da eclíptica: resíduos studentizados versus valores ajustados pelo método dos mínimos quadrados com todas as observações.*

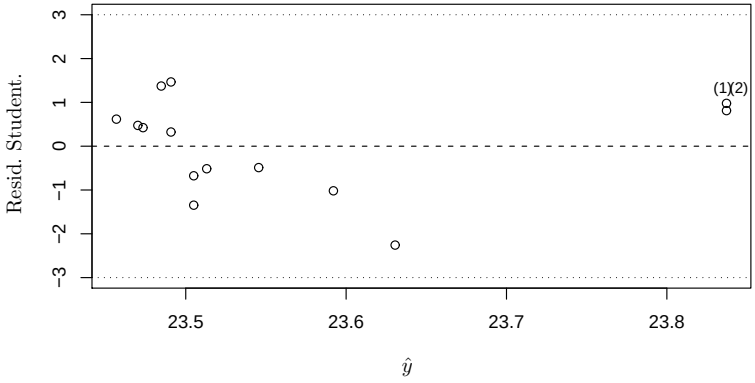


Figura 4.15 *Obliquidade da eclíptica: resíduos studentizados versus valores ajustados pelo método dos mínimos quadrados após retirar a terceira observação.*

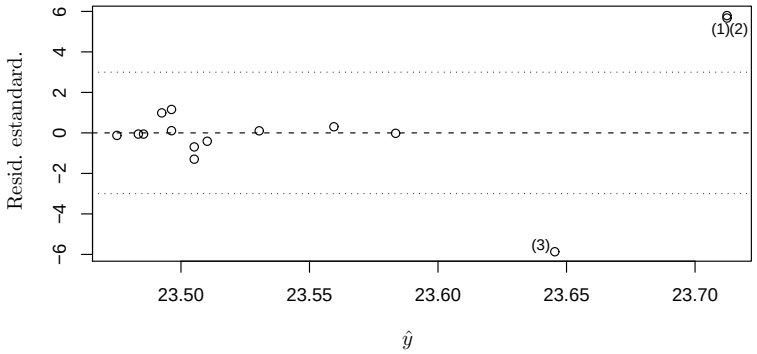


Figura 4.16 *Obliquidade da eclíptica: resíduos estandardizados versus valores ajustados pelo método MM com todas as observações.*

tados semelhantes aos obtidos com os estimadores MM ($\hat{\beta}_0 = 23.69$, $\hat{\beta}_1 = -1.258 \times 10^{-4}$). A dificuldade reside em chegar de forma objectiva a esse resultado e em avaliar correctamente a variabilidade adicional introduzida pelo processo de "limpeza". Ao utilizar um método robusto o analista liberta-se de decisões difíceis relativas à rejeição de observações e simultaneamente pode detectar os *outliers* muito mais facilmente.

Exemplo 4.6. Os dados que vão ser analisados neste exemplo estão descritos em Myers (1990, p. 224) e resultaram de um processo experimental conduzido com vista a compreender o efeito de três factores quantitativos na capacidade de resposta de um sistema de limpeza de carvão (a experiência consiste em usar um polímero para limpar carvão). A variável resposta (y) e as variáveis experimentais (x_1, x_2, x_3), que influenciam a resposta são:

- y: quantidade de matéria sólida em suspensão na solução resultante depois da operação de limpeza (é uma medida da eficiência da operação), mg/l.
- x_1: percentagem de sólidos na solução inicial.
- x_2: pH do tanque onde se encontra a solução.
- x_3: taxa de fluxo do polímero.

Os dados encontram-se na Tabela 4.13.

Tabela 4.13 *Dados produzidos por 12 operações de limpeza de carvão.*

Experiência	x_1	x_2	x_3	y
1	1.5	6.0	1315	243
2	1.5	6.0	1315	261
3	1.5	9.0	1890	244
4	1.5	9.0	1890	285
5	2.0	7.5	1575	202
6	2.0	7.5	1575	180
7	2.0	7.5	1575	183
8	2.0	7.5	1575	207
9	2.5	9.0	1315	216
10	2.5	9.0	1315	160
11	2.5	6.0	1890	104
12	2.5	6.0	1890	110

Tabela 4.14 *Matriz de correlações entre as variáveis na experiência de limpeza de carvão.*

	x_1	x_2	x_3	y
x_1	1	0	0	-0.841
x_2	0	1	0	0.355
x_3	0	0	1	-0.255
y	-0.841	0.355	-0.255	1

Uma vez que o número de variáveis e o número de observações não são grandes, o exame directo dos próprios dados, do gráfico da Figura 4.17 e da matriz de correlações entre as variáveis, na Tabela 4.14, ajuda a compreender a estrutura da experimentação: variáveis de controlo não correlacionadas e correlação máxima negativa (-0.841) entre y e x_3. Esta correlação sugere uma relação linear entre estas variáveis, talvez menos expressiva devida à maior magnitude da resposta relativa à experiência número 9, como se depreende da análise visual do gráfico (y versus x_1).

De facto, segundo o relato apresentado em Myers (1990), ainda antes de ser feita a análise estatística dos dados os engenheiros respon-

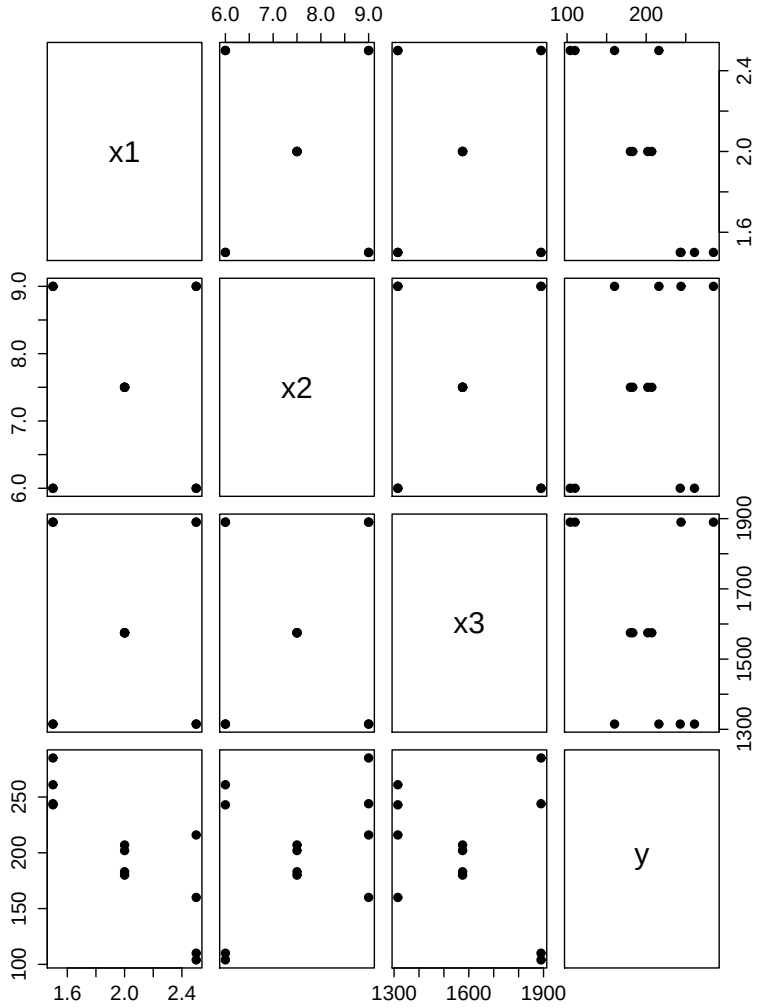


Figura 4.17 *Gráficos de dispersão entre todos os pares de variáveis na experiência de limpeza de carvão.*

Tabela 4.15 *Análise do modelo de regressão pelo método dos mínimos quadrados com todas as observações e sem a observação 9 (e.p. significa erro padrão).*

β	$\hat{\beta}$	$e.p.(\hat{\beta})$	valor de t	valor-p
Com todas as observações				
β_0	397.087	62.757	6.327	0.0002
β_1	−110.750	14.763	−7.502	6.91×10^{-5}
β_2	15.583	4.921	3.167	0.0133
β_3	−0.058	0.026	−2.274	0.0526
	$\hat{\sigma} = 20.88$		$R^2_a = 0.862$	
Sem a observação 9				
β_0	418.927	46.110	9.085	4.01×10^{-5}
β_1	−125.386	11.851	−10.580	1.47×10^{-5}
β_2	10.705	3.951	2.710	0.0302
β_3	−0.034	0.020	−1.649	0.1432
	$\hat{\sigma} = 15.13$		$R^2_a = 0.933$	

sáveis pelo projecto experimental manifestaram preocupações quanto à validade daqueles resultados pois as condições experimentais não teriam sido mantidas constantes como era exigido. Em face desta suspeita pensou-se que o estudo ganharia com a eliminação da observação 9, mas inicialmente foi ajustado o modelo de regressão

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \text{com } \operatorname{var}(\varepsilon_i) = \sigma^2, \qquad (4.8)$$

a todas as observações, usando o método dos mínimos quadrados. A análise dos resíduos resultantes deste ajustamento revela que o valor do resíduo correspondente à observação 9 é 32.2, precisamente o resíduo mais extremo e o seu valor studentizado externamente (Neter *et al.*, 1996) é $2.8695 > t_{7,0.975} = 2.365$ o que mostra que há evidência para considerar que a observação 9 não pertence ao corpo central dos dados e é portanto um *outlier*. A Tabela 4.15 mostra os resultados do ajustamento do modelo de regressão pelo método dos mínimos quadrados, com e sem a observação 9.

As conclusões são semelhantes mas nota-se uma melhoria do ajustamento quando a observação 9 é retirada: o R^2 ajustado passa de 0.862 para 0.933, o erro padrão dos resíduos passa de 20.88 para 15.13

e em consequência o erro padrão é menor para cada um dos parâmetros analisados. Isto mostra que a variável x_1 passa a dar, depois da eliminação da observação 9, uma contribuição mais acentuada para a explicação da variável resposta. Simultaneamente observa-se também uma redução da contribuição das outras variáveis (embora o valor-p associado a β_2 se mantenha inferior a 5% há mais dúvidas sobre a importância desta variável na explicação da variável resposta).

Os dados originais (12 observações) foram também analisados ajustando o modelo de regressão linear (4.8) com base no método MM. Os resultados globais não são muito diferentes dos resultados dos mínimos quadrados quando estes actuam sobre as 12 observações, excepto que a variável x_2 é considerada não significativa, com valor-$p = 0.053$, na estimação robusta. Olhando então para os pesos, w_i, atribuídos às observações pela estimação MM (Tabela 4.16), verifica-se que à observação 9 é dado um peso apenas ligeiramente menor que os pesos dados às outras observações, o que significa que a observação 9 não teve um tratamento muito diferente das restantes e portanto é lógico que o resultado da estimação esteja mais próximo do resultado obtido com os mínimos quadrados com todas as observações do que do resultado correspondente à eliminação da observação 9.

Tabela 4.16 *Pesos atribuídos às observações pelo método MM.*

i	1	2	3	4	5	6	7	8	9	10	11	12
w_i	.993	.963	.938	.866	.999	.909	.934	.989	**.736**	.901	1.000	0.992

O que se passa é que a função `lmrob` usa por defeito para a componente-M do método MM, a função ψ de Tukey com constante de afinação, tuning.psi = 4.685, que garante uma eficiência assintótica dos estimadores igual a 95% sob a distribuição normal dos erros (ver Tabela 3.1). Modificando a constante de afinação (para o valor 3.0, a que corresponde uma eficiência de cerca de 77%) observa-se uma alteração nos pesos, em especial no peso atribuído à observação 9 (ver Tabela 4.17).

O que acontece é que agora o peso atribuído à observação 9 é quase zero, o que corresponde praticamente a ter desprezado aquela observação, como aconteceu no caso da segunda tentativa de estimação dos mínimos quadrados. Os resultados da estimação robusta tornaram-se mais consistentes com os resultados dos mínimos quadrados sem

Tabela 4.17 *Pesos atribuídos às observações pelo método MM com constante de afinação 3.0.*

i	1	2	3	4	5	6	7	8	9	10	11	12
w_i	.966	.943	.822	.728	.975	.872	.916	.924	.003	.998	.997	.992

Tabela 4.18 *Análise do modelo de regressão pelo método com constante de afinação 4.685 e 3.0 (e.p. significa erro padrão).*

β	$\hat{\beta}$	$e.p.(\hat{\beta})$	valor de t	valor-p
Com constante de afinação 4.685				
β_0	401.038	46.561	8.613	2.56×10^{-5}
β_1	-111.886	19.217	-5.822	3.95×10^{-4}
β_2	14.838	6.528	2.273	0.0526
β_3	-0.056	0.033	-1.720	0.1237
	$\hat{\sigma} = 19.85$		$R_a^2 = 0.900$	
Com constante de afinação 3.0				
β_0	420.502	53.405	7.874	4.90×10^{-4}
β_1	-124.345	15.209	-8.175	3.73×10^{-4}
β_2	10.561	5.269	2.004	0.080
β_3	-0.035	0.027	-1.287	0.234
	$\hat{\sigma} = 19.85$		$R_a^2 = 0.900$	

a observação 9. A Tabela 4.18 apresenta os resultados da estimação robusta com as duas constantes de afinação.

Nas Figuras 4.18, 4.19 e 4.20 apresentam-se os gráficos de resíduos para, respectivamente, ajustamento dos mínimos quadrados com todas as observações, ajustamento dos mínimos quadrados sem a observação 9 e ajustamento MM com a constante de afinação 3.0. Estes gráficos confirmam que se obtém um ajustamento muito melhor que o inicial, ou eliminando a observação 9 ou usando o método MM mas reduzindo o valor por defeito da constante de afinação. De notar a quase simetria dos resíduos observada nos dois últimos gráficos, que é esperada quando os erros têm distribuição simétrica e se usa o delineamento experimental escolhido (altamente incompleto, das 27 combinações possíveis dos níveis dos 3 factores apenas se fizeram observações em 5).

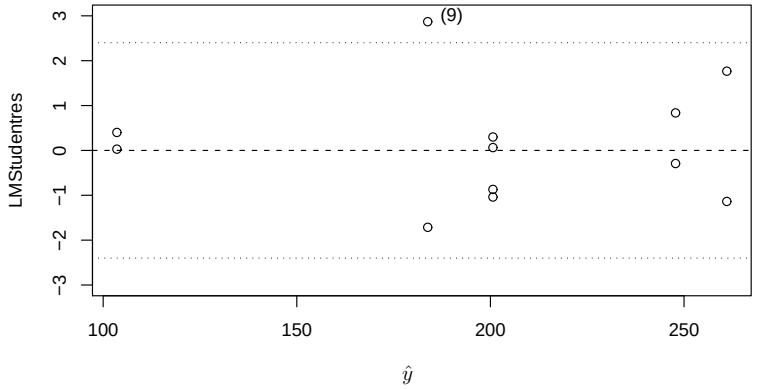


Figura 4.18 *Lavagem de carvão: resíduos studentizados versus valores ajustados pelo método dos mínimos quadrados com todas as observações.*

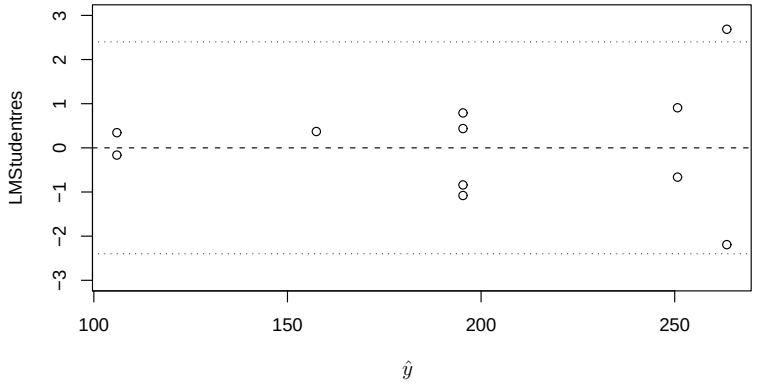


Figura 4.19 *Lavagem de carvão: resíduos studentizados versus valores ajustados pelo método dos mínimos quadrados sem a observação 9.*

Nesta discussão ilustram-se atitudes que o analista deve ter presente sempre que é confrontado na prática com o problema que o exemplo revela:

(i) A suspeita de *outlier* sobre uma observação específica é correctamente analisada (se houver um único *outlier*) usando o

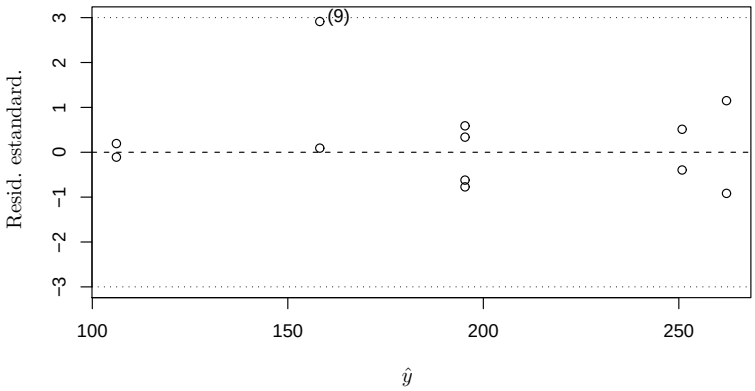


Figura 4.20 *Lavagem de carvão: resíduos estandardizados versus valores ajustados pelo método MM com todas as observações (com constante de afinação 3.0).*

valor residual studentizado externamente associado a essa observação, o qual deve ser comparado com um quantil adequado da distribuição t. A aplicação cega deste procedimento a todas as observações é incorrecta e requer a correcção de Bonferroni para ajustar o nível de significância.

(ii) A prática da eliminação de uma observação suspeita que se confirmou ser *outlier* é em si um procedimento robusto que deve ser usado quando o método de estimação não é robusto, como nos mínimos quadrados.

(iii) O uso da regressão robusta requer um conhecimento razoável do seu funcionamento e da natureza dos vários estimadores que proporciona, se assim não for a eficácia do método pode não ser de facto bem aproveitada.

4.6 Apreciação geral

Como ficou esclarecido na Secção 4.2 o uso automático do método dos mínimos quadrados em regressão pode prejudicar a estimação tanto dos parâmetros como das variâncias dos respectivos estimadores e comprometer assim toda a inferência sobre o modelo baseada quer

em intervalos de confiança, quer em testes de hipóteses.

A regressão robusta tem como objectivo obter estimadores:

(i) que tenham um comportamento muito próximo do comportamento dos estimadores dos mínimos quadrados quando este é o método adequado para ser utilizado (erros normais e i.i.d.);

(ii) que permitam realizar inferências sobre os verdadeiros parâmetros do modelo;

(iii) que tenham um comportamento melhor do que os dos mínimos quadrados quando as hipóteses associadas à optimalidade deste método não são satisfeitas;

(iv) cuja construção seja fácil de compreender e cujas estimativas sejam, computacionalmente, fáceis de obter.

Aparentemente são muitas as metas a cumprir e por isso não admira que a metodologia da regressão robusta seja complexa e como tal não seja acessível ao utilizador comum.

No passado recente a principal explicação para a fraca utilização de regressão robusta pelos analistas era a falta de *software* apropriado. Mas hoje isso já não é razão, uma vez que o software já existe em packages de estatística populares, como SAS, STATA, S-Plus e R.

Apesar disso o uso rotineiro deste software está longe de ser o procedimento correcto e eficaz de usar a regressão robusta. Os próprios mínimos quadrados não podem ser definitivamente descartados pois, mesmo fora das condições ideais, podem ainda assim ser superiores aos métodos robustos. É o caso, relatado por Cook *et al.* (1992), em que o modelo incorpora componentes não lineares, embora o modelo linear seja o modelo a ser ajustado, sendo o desempenho dos métodos LMS e LTS inferior ao dos mínimos quadrados.

A regressão robusta é um processo complexo mas serve bem àquele utilizador que tenha compreendido o conceito e a maneira de operar do procedimento e seja capaz de:

(i) escolher entre os vários métodos disponíveis,
(ii) decidir qual a função $\rho(u)$ que lhe convém e qual o estimador robusto a usar para o factor de escala a incorporar em $\rho(e_i/\hat{\sigma})$,

(iii) determinar a constante de afinação mais adequada.

O desenvolvimento da regressão robusta tem-se centrado principalmente no problema do controlo dos *outliers* cuja presença pode ser provocada pela falta de normalidade dos dados. Mas há outras dificuldades: heterogeneidade e dependência dos erros. Qual o comportamento da regressão robusta em relação a cada uma estas dificuldades, ou às duas em conjunto?

Uma resposta parcial a estas questões é dada com o estudo de simulação apresentado no Exemplo 4.4 (em relação à heterogeneidade) e com o exemplo estudado no capítulo que se segue (em relação à dependência). De facto, a regressão robusta tal como é apresentada, só trata o problema dos *outliers*, para a obrigar a tratar dos outros problemas é preciso usar as ideias que também servem para tratar do problema da heterogeneidade e não independência em relação aos mínimos quadrados (respectivamente mínimos quadrados ponderados e mínimos quadrados generalizados). A resposta consiste em estimar a matriz $\mathbf{V}$,[17] transformar o modelo multiplicando-o por $\mathbf{V}^{-1/2}$ e a seguir aplicar o método robusto ao modelo transformado.

Para terminar este capítulo importa referir que os métodos de regressão robusta acabam por permitir dar resposta a vários procedimentos estatísticos básicos. Como se deixou já antever na Secção 4.5.1 o modelo de localização e o modelo de localização e escala são casos particulares do modelo de regressão o que permite utilizar os métodos aqui apresentados para estimar e realizar inferências relativamente a esses parâmetros. Nalguns casos cai-se nos métodos estudados anteriormente mas noutros obtêm-se novos métodos (é possível por exemplo falar da estimativa LTS de localização univariada). A vantagem principal desta abordagem relaciona-se com a obtenção de intervalos de confiança e a realização de testes de hipóteses e principalmente com a possibilidade de utilizar o *software* já desenvolvido (por exemplo todas as funções para regressão robusta do R). O que se acabou de dizer aplica-se igualmente ao teste-t de diferença de médias ou aos modelos de análise de variância (neste caso é preciso ter alguns cuidados adicionais pois há métodos de regressão robusta que não funcionam quando todas as variáveis explicativas são do tipo qualitativo, ver observações adicionais a este respeito na Secção 5.3.3.).

[17] Como se verá já a seguir este passo é que pode dificultar a resolução do problema.

5
Uma aplicação

5.1 Introdução

Neste capítulo apresenta-se a análise exaustiva de um conjunto de dados reais com alguma complexidade em que os métodos robustos foram aplicados com sucesso. O que se pensava inicialmente ser um problema de regressão múltipla trivial acabou por constituir um desafio e conduziu ao desenvolvimento de vários métodos, quer clássicos, quer robustos, para estimação e diagnóstico no modelo de regressão múltipla em que alguns dos erros podem não ser independentes. O que se vai mostrar é no fundo um possível modo de actuação quando acontecem simultaneamente duas violações importantes das hipóteses de trabalho do modelo de regressão estudado no Capítulo 4: presença de *outliers* (ou distribuição dos erros com caudas mais pesadas que a normal) e não independência.[1]

Os dados reais em análise foram obtidos num estudo observacional realizado com o objectivo de identificar factores que afectam o resultado de um método cirúrgico para correcção da escoliose (curvatura lateral anormal da coluna vertebral). Os dados contêm 392 observações mas algumas dessas observações são referentes ao mesmo paciente. A princípio pareceu adequado usar um modelo de regressão linear múltipla para responder aos objectivos do estudo. No entanto, como não era conveniente eliminar as observações duplas sobre o mesmo paciente (por razões que se vão perceber mais adiante aquando da descrição dos dados), a hipótese de trabalho dos erros não correlacionados, associada ao modelo de regressão usual (4.1) é claramente

[1] Os resultados aqui apresentados baseiam-se em grande parte no artigo Pires e Rodrigues (2007).

violada pois espera-se que haja alguma espécie de relação entre os erros relativos a duas observações efectuadas sobre o mesmo doente. Para confirmar estas suspeitas começou-se por ajustar o modelo usual tendo o diagnóstico dos resíduos correspondentes revelado de facto a existência de problemas de associação entre os resíduos.

Foi então necessário pensar num modelo mais adequado. E o que se decidiu foi manter a estrutura linear mas permitir a existência de correlações não nulas entre os erros associados ao mesmo doente. A seguir foi preciso usar uma estratégia adequada para a estimação dos parâmetros de um tal modelo. Foram analisadas dois procedimentos para estimação desses parâmetros (ou seja, os parâmetros do modelo linear em si e os parâmetros de correlação): (i) uso do método da máxima verosimilhança assumindo normalidade dos erros e (ii) uma variante robusta inspirada na máxima verosimilhança mas recorrendo à regressão robusta. O segundo procedimento tem como objectivo oferecer uma protecção extra contra *outliers* e acabou por conduzir, como se verá, aos resultados mais satisfatórios.

A questão da estimação dos parâmetros do modelo de regressão quando os erros não podem ser considerados independentes, não é novidade na literatura mas a maior parte do trabalho tem sido direccionado para erros auto-correlacionados, ou seja, erros em que se verifica uma correlação fixa entre cada erro e o anterior. Esta estrutura de erros ocorre frequentemente quando as observações são feitas ao longo do tempo e é muito comum em dados de tipo económico. O problema do conjunto de dados que se pretendia analisar, é que para a maior parte das observações pode-se assumir que os erros são independentes mas, para um número ainda razoável de observações, aos pares, e que não interessa aos objectivos do estudo deixar de parte, essa independência não parece natural. A situação mais próxima desta a que se conseguiu encontrar referência na literatura foi estudada por Hsu e Mei (1998). Estes autores consideraram conjuntos de dados com observações correlacionadas aos pares e compararam, num estudo de simulação, a performance de três métodos de estimação dos parâmetros da regressão: mínimos quadrados usuais, mínimos quadrados usuais usando as médias das observações dos pares e mínimos quadrados generalizados, com estimação prévia do parâmetro de correlação. Nesse caso os autores dizem que também pode ser usado um procedimento iterativo de estimação mas não indicam qual.

Para a análise aqui descrita foi também necessário considerar um

procedimento de diagnóstico para detecção de afastamentos da situação ideal de não correlação entre os erros, baseado no teste clássico de Durbin-Watson, e que é robusto em relação à presença de *outliers*, sendo portanto adequado para usar em conjunto com os métodos de estimação robusta.

O resto do capítulo está organizado do modo seguinte: na Secção 5.2 faz-se o enquadramento do problema médico em questão e uma breve descrição dos dados; na Secção 5.3 descrevem-se os métodos estatísticos utilizados, começando por um resumo dos métodos clássicos baseados nos mínimos quadrados, e dos diagnósticos geralmente associados, apresentando-se em seguida os métodos robustos especialmente desenvolvidos para analisar os dados em questão; na Secção 5.4 são apresentados os resultados da aplicação desses métodos e finalmente na Secção 5.5 discutem-se os resultados e apresentam-se algumas conclusões.

5.2 Enquadramento e descrição dos dados

A escoliose consiste, com já se referiu, numa curvatura lateral anormal da coluna vertebral. A gravidade desta doença pode ser quantificada através da medição, numa radiografia à coluna, do chamado ângulo de Cobb (ver a Figura 5.1, retirada de Richardson, 2001).

Os casos mais graves (definidos como aqueles que têm ângulo de Cobb $\geq 30^{\circ}$), e que ocorrem muitas vezes na infância, são geralmente corrigidos cirurgicamente. Existem vários métodos cirúrgicos para correcção da escoliose, um deles é conhecido como o "método português" (Resina, 1963, Resina e Ferreira-Alves, 1977, 1985). Os dados que se vão analisar foram obtidos no contexto de um estudo observacional relativo a 301 doentes de escoliose que foram operados por esse método. As variáveis registadas no estudo foram as seguintes:

A_{pre}: ângulo de Cobb pré-operatório ($^{\circ}$).
A_{pos}: ângulo de Cobb pós-operatório ($^{\circ}$).
A_{f-u}: ângulo de Cobb após 3 anos de *follow-up* ($^{\circ}$).
v_1: idade (anos).
v_2: tipo de curva (quanto à localização).
v_3: sexo (0 - masculino; 1 - feminino).
v_4: cirurgião (1, 2, 3).

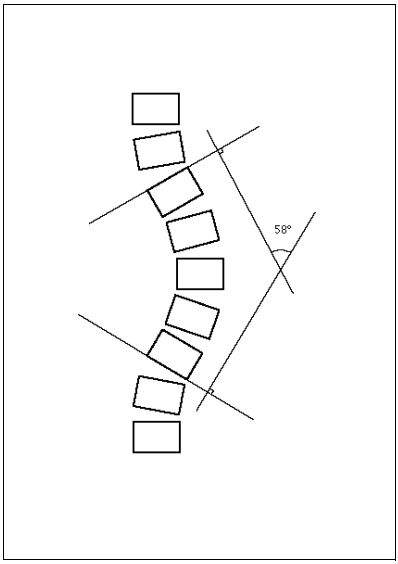


Figura 5.1 *Ilustração da medição do ângulo de Cobb para quantificação da escoliose.*

A variável "tipo de curva" (v_2) tem a seguinte codificação: 1 - curva simples torácica; 2 - curva simples lombar ou toraco-lombar; 3 - ramo superior de curva dupla toraco-lombar; 4 - ramo inferior de curva dupla toraco-lombar; 5 - ramo superior de curva dupla torácica; 6 - ramo inferior de curva dupla torácica. Note-se que os códigos 1 e 2 correspondem a colunas vertebrais com uma única malformação, enquanto que os códigos 3, 4, 5 e 6 correspondem a colunas vertebrais com duas malformações, as quais foram operadas simultaneamente. Na Figura 5.2 (também retirada de Richardson, 2001) mostram-se exemplos de alguns dos tipos de curva.

Um dos objectivos do estudo prende-se com a avaliação dos resultados após a cirurgia para cada tipo de curva. Isto quer dizer que para alguns doentes existem duas linhas na tabela de dados, com valores comuns para as variáveis v_1, v_3 e v_4, mas valores diferentes para as variáveis A_{pre}, A_{pos}, A_{f-u} e v_2. O número total de curvas no conjunto de dados é de 392, enquanto o número de doentes é de 301, o que significa que há 91 casos de doentes com curvas duplas. Para discussão

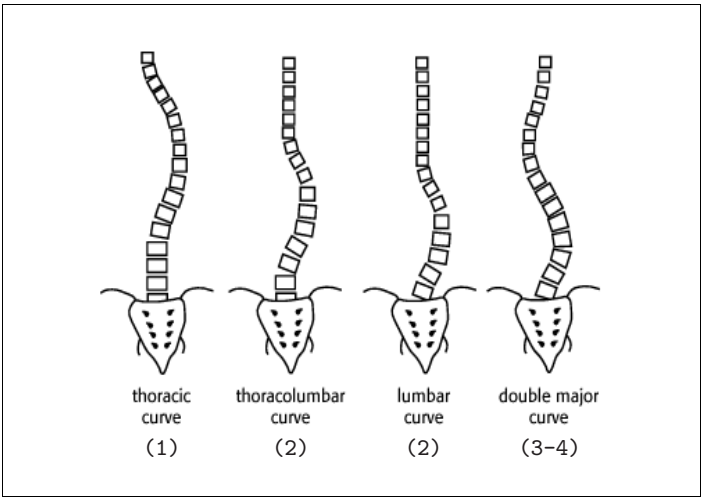


Figura 5.2 *Exemplos de tipos de curvas quanto à localização.*

posterior regista-se aqui a informação de que os dados se encontram ordenados por cirurgião e para cada cirurgião por ordem crescente da data de realização da operação. Na Tabela 5.1 apresenta-se um excerto dos dados (na primeira coluna é indicado um número de código atribuído ao doente por cada cirurgião) percebendo-se que os doentes com os números 3, 9 e 33 (do cirurgião 3) têm curvas duplas e portanto duas linhas na tabela. Pelo valor de v_3 pode concluir-se que essas curvas são do tipo toraco-lombar nos três casos.

O objectivo global do estudo é avaliar a influência de cada uma das possíveis variáveis explicativas $(A_{pre}, v_1, \ldots, v_4)$ no resultado da operação (A_{pos} ou A_{f-u}). Nestas condições parece razoável considerar um modelo de regressão linear múltipla relacionando cada uma das duas variáveis resposta com as variáveis explicativas. Não se pode a partir deste momento ignorar que para as curvas relativas ao mesmo doente os erros do modelo podem ser correlacionados. O modelo de regressão pode vir a ser modificado mas em qualquer caso deverá haver muito cuidado relativamente ao diagnóstico sobre a independência dos resíduos.

Após uma análise exploratória inicial dos dados foi decidido, devido à presença de alguma heterocedasticidade e assimetria, aplicar

Tabela 5.1 *Excerto do conjunto de dados.*

Doente	v_4	v_1	v_2	v_3	A_{pre}	A_{pos}	A_{f-u}
...	...	...	...	...	...	...	...
47	2	27	2	0	90	48	50
48	2	17	2	1	40	25	24
49	2	16	1	1	55	27	30
50	2	15	2	1	50	28	30
51	2	14	2	0	95	54	60
52	2	13	1	1	68	35	38
53	2	16	1	1	55	35	36
54	2	23	1	1	90	45	45
55	2	21	2	0	80	40	40
3	3	11	4	1	50	20	20
3	3	11	5	1	50	30	25
4	3	14	1	1	40	30	28
6	3	14	1	1	80	35	33
9	3	20	4	0	90	70	65
9	3	20	5	0	95	65	60
10	3	18	1	1	50	20	28
13	3	18	1	1	45	18	25
26	3	11	1	0	49	14	18
31	3	14	1	0	74	40	44
33	3	15	4	1	52	28	28
33	3	15	5	1	43	10	17
...	...	...	...	...	...	...	...

a transformação logarítmica a todas as variáveis quantitativas:

$$y_p = \log A_{pos}, \quad y_f = \log A_{f-u}, \quad x_1 = \log A_{pre}, \quad x_2 = \log v_1.$$

O modelo de regressão que se vai considerar pode ser escrito, usando para as variáveis que são factores (variáveis qualitativas) a formulação em termos de efeitos dos diversos níveis, como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3(v_{i2}) + \beta_4(v_{i3}) + \beta_5(v_{i4}) + \varepsilon_i \,, i = 1, \ldots, n, \tag{5.1}$$

onde y pode representar quer y_p quer y_f e os parâmetros do modelo são β_0, β_1, β_2, $\beta_3(1)$, ... , $\beta_3(6)$, $\beta_4(0)$, $\beta_4(1)$, $\beta_5(1)$, $\beta_5(2)$ e $\beta_5(3)$.

Para que este modelo faça sentido e os parâmetros associados aos factores possam ser interpretados como efeitos dos níveis[2] é necessário impor as restrições

$$\sum_{j=1}^{6} \beta_3(j) = 0, \quad \sum_{j=0}^{1} \beta_4(j) = 0, \quad \sum_{j=1}^{3} \beta_5(j) = 0\,. \tag{5.2}$$

Note-se que as variáveis que são factores não aparecem explicitamente nesta formulação do modelo mas aparecem através do parâmetro associado ao nível correspondente. Por exemplo, para a variável v_2, consoante o nível observado assim surge um dos parâmetros $\beta_3(1)$, $\ldots$, $\beta_3(6)$, sendo que por (5.2) se tem de escrever algum desses parâmetros como função dos restantes, por exemplo,

$$\beta_3(6) = -\beta_3(1) - \ldots - \beta_3(5).$$

Esta formulação é equivalente à criação de $k-1$ variáveis indicadoras para cada factor, sendo k o número de níveis do factor. Assim, para codificar a variável v_2 (tipo de curva) são criadas cinco novas variáveis:

$$x_3 = \begin{cases} 1, & \text{se } v_2 = 1 \\ 0, & \text{se } v_2 = 2,3,4,5 \\ -1, & \text{se } v_2 = 6 \end{cases} \qquad x_4 = \begin{cases} 1, & \text{se } v_2 = 2 \\ 0, & \text{se } v_2 = 1,3,4,5 \\ -1, & \text{se } v_2 = 6 \end{cases}$$

$$x_5 = \begin{cases} 1, & \text{se } v_2 = 3 \\ 0, & \text{se } v_2 = 1,2,4,5 \\ -1, & \text{se } v_2 = 6 \end{cases} \qquad x_6 = \begin{cases} 1, & \text{se } v_2 = 4 \\ 0, & \text{se } v_2 = 1,2,3,5 \\ -1, & \text{se } v_2 = 6 \end{cases}$$

$$x_7 = \begin{cases} 1, & \text{se } v_2 = 5 \\ 0, & \text{se } v_2 = 1,2,3,4 \\ -1, & \text{se } v_2 = 6 \end{cases},$$

estando cada uma delas associada, respectivamente, aos parâmetros $\beta_3(1)$, $\ldots$, $\beta_3(5)$. Para a variável v_3 (sexo) é apenas criada uma variável indicadora

$$x_8 = \begin{cases} 1, & \text{se } v_3 = 0 \\ -1, & \text{se } v_3 = 1 \end{cases},$$

associada ao parâmetro $\beta_4(0)$, enquanto que para codificar a variável v_4 (cirurgião) são criadas as variáveis

$$x_9 = \begin{cases} 1, & \text{se } v_4 = 1 \\ 0, & \text{se } v_4 = 2 \\ -1, & \text{se } v_4 = 3 \end{cases} \quad \text{e} \quad x_{10} = \begin{cases} 0, & \text{se } v_4 = 1 \\ 1, & \text{se } v_4 = 2 \\ -1, & \text{se } v_4 = 3 \end{cases}$$

[2]Ou seja, como diferenças em relação à média global.

associadas, respectivamente, aos parâmetros $\beta_5(1)$ e $\beta_5(2)$. Depois desta recodificação o modelo (5.1) pode ser escrito na forma matricial usual dos modelos de regressão, equação (4.2),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{5.3}$$

onde $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X}$ é uma matriz $n \times p$ que tem em conta (5.1) e (5.2), ou seja, é construída de acordo com a descrição acima. No caso de se considerarem todas as variáveis e todas as observações as dimensões de $\mathbf{X}$ são 392×11, sendo a linha i o vector $\mathbf{x}_i{=}(1, x_{i1}, \ldots, x_{i,10})$. Por exemplo, para a primeira linha da Tabela 5.1, a linha correspondente da matriz $\mathbf{X}$ é

$$1 \quad \log(90) \quad \log(27) \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1\,.$$

Em (5.3) tem-se ainda que o vector dos parâmetros é

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3(1), \ldots, \beta_3(5), \beta_4(0), \beta_5(1), \beta_5(2))^T$$

e o vector dos erros é $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Assume-se daqui por diante que a matriz $\mathbf{X}$ tem característica completa (isto é, característica igual ao número de parâmetros). A estrutura assumida para os erros será apresentada na secção seguinte.

Para concluir esta secção apresentam-se nas Tabelas 5.2, 5.3 e 5.4, descrições sumárias de vários aspectos do conjunto de dados. Analisando a Tabela 5.2 é possível verificar que a transformação logarítmica das variáveis quantitativas melhorou a simetria tanto para as variáveis explicativas como para as variáveis resposta.

5.3 Métodos

Nesta secção descrevem-se os métodos estatísticos utilizados, começando por um resumo dos métodos clássicos baseados nos mínimos quadrados e dos diagnósticos que lhe estão geralmente associados. Apresentam-se em seguida os métodos robustos especialmente desenvolvidos para analisar os dados em apreço.

O modelo (5.3) foi objecto de estudo e discussão detalhada no Capítulo 4, tendo-se aí apresentado os principais resultados relativos ao método dos mínimos quadrados. No entanto, para facilitar a leitura

Tabela 5.2 *Estatísticas sumárias univariadas para as variáveis quantitativas.*

Variável	Mín.	q_1	Mediana	Média	q_3	Máx.
A_{pos}	9.00	25.00	32.50	35.20	42.25	92.00
A_{f-u}	10.00	28.00	36.00	37.94	45.00	90.00
A_{pre}	30.00	52.00	65.00	67.94	80.00	147.00
v_1	6.00	13.00	15.00	15.38	17.00	33.00
y_p	2.197	3.219	3.481	3.483	3.744	4.522
y_f	2.303	3.332	3.584	3.569	3.807	4.500
x_1	3.401	3.951	4.174	4.176	4.382	4.990
x_2	1.792	2.565	2.708	2.711	2.833	3.497

Tabela 5.3 *Correlações entre as variáveis quantitativas.*

Originais

	A_{f-u}	A_{pre}	v_1
A_{pos}	0.93	0.83	0.30
A_{f-u}		0.81	0.29
A_{pre}			0.15

Transformadas

	y_f	x_1	x_2
y_p	0.92	0.80	0.29
y_f		0.79	0.28
x_1			0.14

Tabela 5.4 *Tabela de contingência para as variáveis qualitativas.*

	$v_3 = 0$				$v_3 = 1$				Marginal			
$v_2 \backslash v_4$	1	2	3		1	2	3		1	2	3	
1	32	7	23		49	11	47		81	18	70	169
2	7	8	3		11	7	5		18	15	8	41
3	7	6	4		20	15	24		27	21	28	76
4	7	6	4		20	15	24		27	21	28	76
5	1	0	4		5	1	4		6	1	8	15
6	1	0	4		5	1	4		6	1	8	15
				124				268	165	77	150	392

e estabelecer notação para uso posterior apresenta-se aqui um breve resumo dos principais resultados relativos a este método e que se encontram concentrados na Tabela 4.3. Como se sabe, se em relação ao modelo (5.3) se assumir que

$$E(\varepsilon_i) = 0, \;\; \text{var}(\varepsilon_i) = \sigma^2 \;\; \text{e} \;\; \text{cov}(\varepsilon_i, \varepsilon_j) = 0, \;\; \text{para todo o } i \neq j,$$

o que é equivalente a escrever $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ e $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, os estimadores dos mínimos quadrados (MQ) de $\boldsymbol{\beta}$ são dados por

$$\hat{\boldsymbol{\beta}}_{MQ} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{5.4}$$

e a matriz de covariâncias estimada de $\hat{\boldsymbol{\beta}}_{MQ}$ é dada por

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_{MQ}) = \hat{\sigma}^2_{MQ}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}, \tag{5.5}$$

com

$$\hat{\sigma}^2_{MQ} = \frac{SSE_{MQ}}{n-p} \quad \text{e} \quad SSE_{MQ} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T_{MQS}\mathbf{X}^T\mathbf{y}\,. \tag{5.6}$$

Se, além disso, se assumir a normalidade multivariada de $\boldsymbol{\varepsilon}$, o método da máxima verosimilhança conduz aos mesmos resultados. Nesse caso $\hat{\boldsymbol{\beta}}_{MQ}$ também tem distribuição normal o que permite prosseguir a inferência sobre os parâmetros do modelo, usando essencialmente a variável com distribuição t_{n-p} indicada na Tabela 4.3. A aplicação do método dos mínimos quadrados para estimação dos parâmetros do modelo (5.1), com os dados da escoliose, conduziu aos resultados apresentados na Secção 5.4 (Tabelas 5.6 e 5.7, para y_p e y_f, respectivamente, colunas com título MQ). Porém, antes de analisar estes resultados convém introduzir os métodos de diagnóstico relativos à independência dos erros.

5.3.1 Diagnóstico

Quando existe alguma dúvida sobre a hipótese de trabalho de não correlação entre os erros, é fundamental aplicar algum meio de diagnóstico especialmente concebido para detectar a falha dessa hipótese. Esses diagnósticos são baseados nos resíduos,

$$\mathbf{e}_{MQ} = (e_1, \ldots, e_n)^T = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{MQ}\,, \tag{5.7}$$

de um modelo que foi ajustado assumindo tal hipótese. Um procedimento muito conhecido e utilizado é o teste de Durbin-Watson (Durbin e Watson, 1950), cuja estatística de teste é

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}. \tag{5.8}$$

É muito simples verificar que para n grande se tem

$$DW \simeq 2(1 - r_1), \tag{5.9}$$

onde r_1 representa o coeficiente de correlação empírica entre

$$(e_1, \ldots, e_{n-1})^T \quad \text{e} \quad (e_2, \ldots, e_n)^T.$$

Para que esta aproximação seja válida é necessário que as médias dessas amostras sejam próximas de zero. Este teste foi especialmente concebido para detectar auto-correlação em erros com distribuição normal, ou seja, que verificam

$$\varepsilon_i = \rho\,\varepsilon_{i-1} + \xi_i, \quad \text{com } |\rho| < 1 \ \text{ e } \xi_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\xi^2)$$

(as variâncias de ε_i e ξ_i estão relacionadas por $\sigma_\xi^2 = \sigma^2(1 - \rho^2)$). Dada a forma da estatística do teste ele pode também ser aplicado a outras estruturas de correlação que se possam manifestar, como no contexto presente, através de correlação entre resíduos sucessivos.

A hipótese nula é sempre $H_0 :\ \rho = 0$, enquanto a hipótese alternativa pode ser unilateral, $H_1 :\ \rho > 0$ ou $H_1 :\ \rho < 0$, ou bilateral, $H_1 :\ \rho \neq 0$. Na maior parte das situações (geralmente observações ao longo do tempo) escolhe-se a alternativa unilateral do primeiro tipo. No caso presente julga-se que é mais sensato considerar a alternativa bilateral.

A expressão (5.9) mostra que DW tende a tomar valores próximo de 2 se $\rho = 0$ e menores (maiores) do que 2 se $\rho > 0$ ($\rho < 0$). A hipótese nula deverá então ser rejeitada, contra a alternativa bilateral de DW se afastar demasiado de 2. Durbin e Watson (1951) e Durbin e Watson (1971) apresentam tabelas com valores críticos para o teste. Em vez de usar essas tabelas é preferível estimar o valor-p associado a um valor observado da estatística, $DW = dw$, usando permutações dos resíduos (Schmoyer, 1994, e Canjels, 2002). O procedimento para obter estes valores-p é o seguinte:

- Para a $j-$ésima permutação, o vector original dos resíduos, $\mathbf{e}_{MQ}$ (que origina o valor observado da estatística, dw) é permutado, conduzindo a $\mathbf{e}_j^*$.
- Um novo vector de respostas é então construído fazendo

$$\mathbf{y}_j^* = \mathbf{X}\hat{\boldsymbol{\beta}}_{MQ} + \mathbf{e}_j^*;$$

  o modelo é de novo ajustado, e em seguida calculam-se os resíduos, $\mathbf{e}_{MQ,j}^*$, e o valor da estatística, dw_j^*, correspondentes.
- Os dois passos anteriores são repetidos um número elevado de vezes, $j = 1, \ldots, B$.

Para o teste bilateral o valor-p pode então ser estimado por intermédio de

$$p_{DW,B} = \begin{cases} \dfrac{\#\left\{j : dw_j^* < dw\right\}}{B} \times 2, & \text{se } dw \leq 2 \\ \dfrac{\#\left\{j : dw_j^* > dw\right\}}{B} \times 2, & \text{se } dw > 2 \end{cases} . \qquad (5.10)$$

Também é possível efectuar apenas as permutações dos resíduos sem reajustar o modelo mas segundo Schmoyer (1994) o teste resultante apenas é válido assintoticamente e a aproximação resultante pode não ser boa na prática. O mesmo autor recomenda um procedimento de simulação com base na distribuição empírica dos resíduos que é muito semelhante ao procedimento acabado de descrever. Uma vantagem adicional, ou talvez até a mais importante, do uso do método de permutações é que deixa de ser necessária a hipótese de normalidade dos erros, pois a distribuição da estatística do teste sob a hipótese nula passa a ser determinada sob a distribuição empírica dos resíduos que é em geral uma boa aproximação não paramétrica para a distribuição dos erros, desde que o número de observações não seja muito reduzido.

Apesar disso, o valor da estatística (5.8) é extremamente sensível a *outliers*, pois como já se viu o coeficiente de correlação amostral na qual se baseia não é robusto. A não robustez da estatística de Durbin-Watson relativamente a desvios das hipóteses usuais foi investigada por vários autores (por exemplo, Evans, 1992, e Ali e Sharma, 1993). Esses autores concluíram que o teste pode ter um desempenho muito mau quando os erros seguem distribuições assimétricas ou com caudas pesadas.

A partir de (5.9) é muito fácil construir uma versão robusta do teste de Durbin-Watson, aplicável quando n é elevado. Basta substituir r_1 (correlação entre $(e_1, \ldots, e_{n-1})^T$ e $(e_2, \ldots, e_n)^T$) por qualquer estimador robusto do coeficiente de correlação usando as mesmas amostras. No que se vai seguir adoptou-se o estimador proposto por Rousseeuw (1985) e designado pela sigla RMCD (de *Reweighted Minimum Covariance Determinant*) calibrado para um ponto de rotura de 25%, o qual segundo Croux e Haesbroeck (1999) para além de ter alto ponto de rotura é razoavelmente eficiente para dados multinormais. Daqui em diante o teste de Durbin-Watson robusto refere-se ao teste baseado na estatística

$$DW_R = 2(1 - r_{MCD,1}), \tag{5.11}$$

onde $r_{MCD,1}$ significa o estimador RMCD25 do coeficiente de correlação calculado com as amostras $(e_1, \ldots, e_{n-1})^T$ e $(e_2, \ldots, e_n)^T$. Assume-se, como anteriormente, que as estimativas de localização baseadas em cada uma dessas amostras são aproximadamente zero.

Os argumentos que justificam a versão baseada em permutações para o teste de Durbin-Watson usual mantêm-se válidos para o teste robusto, pelo que também se utilizou o procedimento já descrito, e a expressão (5.10) para estimar o valor-p do teste de Durbin-Watson robusto (representado por $p_{DWR,B}$).

Kalina (2002), propôs uma outra variante robusta do teste de Durbin-Watson, contudo a abordagem usada é diferente da que foi aqui considerada, uma vez que o que aquele autor propõe é a utilização da expressão (5.8) após um alisamento dos resíduos com vista a reduzir o efeito dos potenciais *outliers*. A versão aqui proposta tem o mesmo objectivo recorrendo a métodos já estabelecidos para o coeficiente de correlação.

Aplicou-se então o teste (nas duas versões, usual e robusta) aos resíduos do modelo (5.3) para os dados da escoliose com todas as variáveis explicativas e assumindo, como se disse, $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Concluiu-se que para ambas as variáveis resposta, y_p e y_f, a hipótese nula é rejeitada para todos os níveis de significância usuais, pelas duas versões do teste. Os resultados podem ler-se novamente nas Tabelas 5.6 e 5.7, para y_p e y_f, respectivamente, nas colunas MQ, linhas $p_{DW,B}$ e $p_{DWR,B}$. Para estimar os valores-p apresentados usou-se um número de permutações $B = 10000$.

Suspeitou-se que este resultado fosse devido à existência das curvas duplas no conjunto de dados. Uma estratégia possível para ultrapassar o problema, uma vez que os resultados da estimação não podem ser aceites nestas circunstâncias, é considerar outra estrutura para $\text{var}(\boldsymbol{\varepsilon})$ que permita a existência de correlações não nulas entre erros relativos ao mesmo doente. É disto que se trata na secção seguinte.

5.3.2 Estimação com erros correlacionados

O modelo de regressão que se considera agora tem a mesma equação de regressão, (5.3), mas admite-se que $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$ onde $\mathbf{V}$ é uma matriz quadrada de ordem n, simétrica e definida positiva, que por enquanto se supõe conhecida. Este modelo também foi já mencionado no Capítulo 4 a propósito do método dos mínimos quadrados generalizados. Modelos deste tipo surgem frequentemente na literatura para, tal como no Exemplo 4.1, acomodar a situação de heterocedasticidade, caso em que a formulação adequada conduz a uma matriz $\mathbf{V}$ de tipo diagonal. Outra situação bastante referida na literatura, principalmente na área da econometria, é a situação de erros autocorrelacionados, ou seja, em que os erros, ε_i, seguem o modelo AR(1), $\varepsilon_i = \rho\varepsilon_{i-1} + \xi_i$ com $|\rho| < 1$ e ξ_i não correlacionadas com variância constante dada por $\sigma^2_\xi = \sigma^2(1-\rho^2)$. Nesse caso $\mathbf{V}$ é da forma seguinte

$$\mathbf{V} = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix}. \tag{5.12}$$

Como na maior parte dos casos os parâmetros que surgem em $\mathbf{V}$ não são de facto conhecidos e têm de ser estimados, é aconselhável manter $\mathbf{V}$ o mais simples possível (o número total de parâmetros, incluindo os da regressão deve manter-se bastante inferior ao número de observações). O que se deve fazer, usando o princípio da parcimónia, é tentar captar os aspectos mais relevantes da estrutura de correlação com o menor número possível de parâmetros. Fazendo uso deste princípio propôs-se, para resolver o problema detectado no modelo de regressão para os dados da escoliose, assumir que os erros continuam a ter variância constante mas que os erros relativos ao mesmo doente podem ter correlação não nula. Devido à existência de dois

tipos de curvas duplas, que correspondem a cirurgias de grau de dificuldade diferente (as curvas duplas torácicas são mais difíceis de corrigir simultaneamente), propõe-se a consideração de dois parâmetros de correlação: ρ_1 associado à escoliose dupla toraco-lombar e ρ_2, associado à escoliose dupla torácica. Ou seja, assume-se que as covarìâncias entre pares de erros são as seguintes:

- $\text{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2$, se $i = j$;
- $\text{cov}(\varepsilon_i, \varepsilon_j) = \rho_1\sigma^2$, se $i \neq j$, i e j referem-se ao mesmo doente e $v_{i3} = 3$ ou $v_{i3} = 4$;
- $\text{cov}(\varepsilon_i, \varepsilon_j) = \rho_2\sigma^2$, se $i \neq j$, i e j referem-se ao mesmo doente e $v_{i3} = 5$ ou $v_{i3} = 6$;
- $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, para todos os restantes casos.

Esta estrutura é representada por uma matriz $\mathbf{V}$ com a seguinte forma genérica (caracterizada pela existência de alguns blocos 2×2 na diagonal principal)

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdot \\ 0 & 1 & \rho_1 & 0 & 0 & 0 & \cdot \\ 0 & \rho_1 & 1 & 0 & 0 & 0 & \cdot \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdot \\ 0 & 0 & 0 & 0 & 1 & \rho_2 & \cdot \\ 0 & 0 & 0 & 0 & \rho_2 & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

Para o modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ com $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, $\mathbf{V}$ conhecida, os parâmetros podem, como se sabe, ser estimados pelo método dos mínimos quadrados generalizados (MQG). Repete-se aqui, tal como foi feito para os mínimos quadrados, a informação relevante da Tabela 4.3. Os estimadores são obtidos resolvendo o problema de minimização

$$\hat{\boldsymbol{\beta}}_{MQG} \ \text{ minimiza } \ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

o que conduz a

$$\hat{\boldsymbol{\beta}}_{MQG} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{5.13}$$

e à matriz de covariâncias estimada de $\hat{\boldsymbol{\beta}}_{MQG}$,

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_{MQG}) = \hat{\sigma}^2_{MQG}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}, \tag{5.14}$$

com

$$\hat{\sigma}^2_{MQG} = \frac{SSE_{MQG}}{n-p} \quad \text{e} \quad SSE_{MQG} = \mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} - \hat{\boldsymbol{\beta}}^T_{MQG}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}\,. \tag{5.15}$$

Os resíduos para este modelo são dados por $\mathbf{e}_{MQG} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{MQG}$. Como foi também referido no Capítulo 4 pode chegar-se aos mesmos resultados aplicando o método dos mínimos quadrados ao modelo transformado

$$\mathbf{V}^{-1/2}\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\boldsymbol{\varepsilon}\,, \tag{5.16}$$

ou

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

com

$$\mathbf{y}^* = \mathbf{V}^{-1/2}\mathbf{y}, \quad \mathbf{X}^* = \mathbf{V}^{-1/2}\mathbf{X} \quad \text{e} \quad \boldsymbol{\varepsilon}^* = \mathbf{V}^{-1/2}\boldsymbol{\varepsilon},$$

onde $\mathbf{V}^{-1/2}$ é uma matriz tal que

$$(\mathbf{V}^{-1/2})^T\mathbf{V}^{-1/2} = \mathbf{V}^{-1}$$

que pode ser obtida , por exemplo, pela decomposição de Cholesky de $\mathbf{V}^{-1}$. Nestas condições

$$\text{var}(\mathbf{V}^{-1/2}\boldsymbol{\varepsilon}) = \text{var}(\boldsymbol{\varepsilon}^*) = \sigma^2\mathbf{I}$$

e a estimativa dos mínimos quadrados $\hat{\boldsymbol{\beta}}_{MQ}$ obtida com os dados $(\mathbf{y}^*, \mathbf{X}^*)$, representada por $\hat{\boldsymbol{\beta}}^*_{MQ}$, coincide com a estimativa dos mínimos quadrados generalizados, $\hat{\boldsymbol{\beta}}_{MQG}$, obtida a partir dos dados originais $(\mathbf{y}, \mathbf{X})$. Os resíduos do modelo original e os do modelo transformado estão relacionados através de

$$\mathbf{e}_{MQG} = \mathbf{V}^{1/2}\mathbf{e}^*_{MQ} \quad \text{com} \quad \mathbf{e}^*_{MQ} = \mathbf{y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}^*_{MQ}.$$

É de realçar que os diagnósticos para verificar a não existência de auto-correlação ou para validar a independência dos erros devem ser realizados sobre os resíduos do modelo transformado, pois é em relação aos erros desse modelo que se assumem tais hipóteses. Desta maneira pode validar-se, ou não, a estrutura de covariâncias adoptada, consubstanciada na matriz $\mathbf{V}$.

O uso da transformação (5.16) é também muito conveniente em termos computacionais pois possibilita o uso do *software* usual para

análise de regressão pelos mínimos quadrados. Só é preciso ter em atenção que a matriz transformada $\mathbf{X}^*$ não terá na maior parte das vezes, e ao contrário de $\mathbf{X}$, uma coluna constante pelo que quando se ajusta o modelo transformado se deve indicar que se trata de uma regressão a passar pela origem.

Em geral, mesmo que a estrutura de $\mathbf{V}$ seja conhecida, o valor exacto dos parâmetros envolvidos nessa estrutura (no caso em estudo ρ_1 e ρ_2) dificilmente será conhecido, pelo que haverá necessidade de os estimar a partir dos dados. É importante recordar que quando se assume $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{V})$ e $\mathbf{V}$ é conhecida, conclui-se que $\hat{\boldsymbol{\beta}}_{MQG} \sim N_p(\boldsymbol{\beta}, \text{var}(\hat{\boldsymbol{\beta}}_{MQG}))$, o que permite realizar inferências sobre $\boldsymbol{\beta}$. No caso em que os parâmetros de $\mathbf{V}$, ρ_1 e ρ_2, são estimados essas inferências deixam de ser exactas passando a ser válidas apenas aproximadamente (a distribuição referida é assintótica).

É preciso pensar agora num processo para estimar ρ_1 e ρ_2. Embora se pudesse imaginar algum procedimento *ad-hoc* para fazer essa estimação[3] a estratégia mais adequada consiste em estimar conjuntamente e pelo método da máxima verosimilhança todos os parâmetros do modelo de regressão, incluindo os referentes a var($\boldsymbol{\varepsilon}$), σ^2, ρ_1 e ρ_2.

Assumindo então que $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{V})$, as estimativas de máxima verosimilhança (MV) de todos os parâmetros envolvidos maximiza a função de verosimilhança dada por

$$L(\boldsymbol{\beta}, \sigma^2, \rho_1, \rho_2 | \mathbf{X}, \mathbf{y}) = (2\pi\sigma^2)^{-n/2} |\mathbf{V}|^{-1/2} \times \\ \times \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right], \quad (5.17)$$

com

$$|\mathbf{V}|^{-1/2} = \left(1 - \rho_1^2\right)^{-n_1/2} \left(1 - \rho_2^2\right)^{-n_2/2}, \quad (5.18)$$

onde n_i é o número de casos com observações duplas e correlação ρ_i (nos dados da escoliose, $n_1 = 76$ e $n_2 = 15$, representando, respectivamente, o número de doentes com escoliose dupla do primeiro tipo e o número de doentes com escoliose dupla do segundo tipo, ver Tabela 5.4). A função de log-verosimilhança correspondente a (5.17)

[3]Por exemplo, em relação ao modelo com erros auto-correlacionados usa-se por vezes o procedimento de Cochrane-Orcutt (Cochrane e Orcutt, 1949). No entanto sabe-se que esse procedimento tende a subestimar o verdadeiro valor da auto-correlação (Neter *et al.*, 1996, p. 512) ou pode ficar "preso" num mínimo local da função objectivo (Dufour et al., 1980).

é dada por

$$\log L(\boldsymbol{\beta}, \sigma^2, \rho_1, \rho_2 | \mathbf{X}, \mathbf{y}) = C - \frac{n}{2} \log \sigma^2 - \\ - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \\ - \frac{n_1}{2} \log(1 - \rho_1^2) - \frac{n_2}{2} \log(1 - \rho_2^2). \quad (5.19)$$

Não parece ser possível obter soluções explícitas para os estimadores

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\rho}_1, \hat{\rho}_2) \ \text{ maximiza } \ \log L(\boldsymbol{\beta}, \sigma^2, \rho_1, \rho_2 | \mathbf{X}, \mathbf{y}) \, .$$

É, no entanto, simples calcular uma função log-verosimilhança perfilada (de *profile log-likelihood*) para uma grelha de pontos

$$(\rho_1, \rho_2) \in (-1, 1) \times (-1, 1),$$

uma vez que, dado (ρ_1, ρ_2), se verifica que

$$\hat{\boldsymbol{\beta}}(\rho_1, \rho_2) = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} = \hat{\boldsymbol{\beta}}_{MQG}, \quad (5.20)$$

com $\mathbf{V} = \mathbf{V}(\rho_1, \rho_2)$ e

$$\hat{\sigma}^2(\rho_1, \rho_2) = \hat{\sigma}^2_{MQG} \frac{n - p}{n} \, . \quad (5.21)$$

A função log-verosimilhança perfilada pode portanto escrever-se como

$$\log \hat{L}(\rho_1, \rho_2) = \log L(\hat{\boldsymbol{\beta}}(\rho_1, \rho_2), \hat{\sigma}^2(\rho_1, \rho_2), \rho_1, \rho_2 | \mathbf{X}, \mathbf{y}) = \\ = C - \frac{n}{2} - \frac{n}{2} \log \hat{\sigma}^2(\rho_1, \rho_2) - \frac{n_1}{2} \log(1 - \rho_1^2) - \frac{n_2}{2} \log(1 - \rho_2^2) \, , \quad (5.22)$$

e as estimativas são dadas por

$$(\hat{\rho}_1, \hat{\rho}_2) \ \text{ maximiza } \ \log \hat{L}(\rho_1, \rho_2), \quad (5.23)$$

e $\hat{\boldsymbol{\beta}}_{MV} = \hat{\boldsymbol{\beta}}(\hat{\rho}_1, \hat{\rho}_2)$ e $\hat{\sigma}^2_{MV} = \hat{\sigma}^2(\hat{\rho}_1, \hat{\rho}_2)$ obtidas a partir das expressões (5.20) e (5.21).

Para os dados da escoliose obtiveram-se as seguintes estimativas pontuais de (ρ_1, ρ_2): $(0.38, 0.68)$, para o modelo com variável resposta

Figura 5.3 *Gráficos de curvas de nível de* $\log \hat{L}(\rho_1, \rho_2)$ *para* y_p *e* y_f *com todas as variáveis explicativas. Os pontos* × *indicam as estimativas pontuais e as linhas a tracejado delimitam regiões de confiança aproximadas (a 95%) para os verdadeiros valores de* (ρ_1, ρ_2).

y_p e com todas as variáveis explicativas; $(0.46, -0.33)$, para o modelo com variável resposta y_f e com todas as variáveis explicativas.

O procedimento de busca foi o seguinte: efectuou-se uma busca preliminar grosseira na grelha $(\rho_1, \rho_2) \in \{-0.9, -0.8, \ldots, 0.8, 0.9\}^2$ para identificar a localização aproximada de $(\hat{\rho}_1, \hat{\rho}_2)$; conduziu-se então uma busca mais fina nessa região, com intervalos de 0.01 nas duas direcções. A Figura 5.3 mostra os gráficos de curvas de nível das funções log-verosimilhança perfiladas obtidas e as localizações das estimativas pontuais obtidas. Observando a figura nota-se que para y_p existe um máximo global e um máximo local da função objectivo, enquanto que para y_f apenas se observa um máximo global. As linhas a tracejado delimitam regiões de confiança aproximadas (a 95%) para os verdadeiros valores de (ρ_1, ρ_2) obtidas a partir de

$$\log \hat{L}(\hat{\rho}_1, \hat{\rho}_2) - \log \hat{L}(\rho_1, \rho_2) \leq \frac{\chi^2_{2,0.95}}{2}\,.$$

Estas regiões mostram que existe uma maior incerteza associada à estimativa de ρ_2 do que à estimativa de ρ_1. Este resultado não é surpreendente dado que $n_1 = 76$ é muito maior do que $n_2 = 15$ Pode também concluir-se que, ao nível de significância de 5% e em ambos os casos, ρ_1 é significativamente diferente de zero mas ρ_2 não. Apesar disso decidiu-se manter os dois parâmetros de correlação nos modelos.

Se após este trabalho se aplicar o teste de Durbin- Watson aos resíduos do modelo transformado, para os dados da escoliose com todas as variáveis explicativas, e assumindo que $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 \hat{\mathbf{V}}$ com as estimativas de MV de ρ_1 e ρ_2, verifica-se que para ambas as variáveis resposta, y_p e y_f, os resultados melhoraram em termos do valor observado da estatística mas ainda conduzem à rejeição da hipótese nula de não correlação entre os erros. Estes resultados estão apresentados nas Tabelas 5.6 e 5.7, para y_p e y_f, respectivamente, nas colunas MV, linhas $p_{DW,B}$, $p_{DWR,B}$ ($B = 10000$ permutações).

5.3.3 Estimação robusta com erros correlacionados

Os resultados descritos no final da última secção constituem de alguma forma uma desilusão pois parece que não se conseguiu com o modelo generalizado resolver o problema, os resíduos do modelo transformado continuam a indicar a não independência nos erros. Decidiu-se então completar a análise de diagnóstico do modelo, tendo

em conta também os aspectos de normalidade dos erros e influência das observações individuais nas estimativas.

A Figura 5.4 apresenta os gráficos quantil-quantil (para a distribuição normal) dos resíduos do melhor modelo obtido até ao momento, ou seja com o último método de estimação descrito (recorda-se mais uma vez que se têm de analisar os resíduos do modelo transformado). Os gráficos apontam para a existência de um certo desvio da normalidade, tanto em termos de enviesamento como de caudas mais pesadas, bem como para a possível ocorrência de alguns *outliers*. Em consequência decidiu-se usar métodos de estimação robusta para todos os parâmetros do modelo, tratando separadamente as hipóteses de erros correlacionados e não correlacionados.

Considera-se em primeiro lugar a hipótese de erros não correlacionados com o objectivo de comparar os resultados com os resultados obtidos usando o método dos mínimos quadrados ordinários. De entre os métodos de regressão robusta associados ao modelo (5.3) e descritos no Capítulo 4 há alguns que não podem ser aplicados quando há variáveis explicativas do tipo qualitativo devido à possibilidade de surgirem matrizes singulares durante a execução dos algoritmos. Este é um problema bastante ignorado na literatura sobre regressão robusta, no entanto há algumas referências a ele dedicadas, como por exemplo Hubert e Rousseeuw (1997) ou Maronna e Yohai (2000) que propuseram mesmo métodos robustos para modelos de regressão em que há variáveis explicativas contínuas e categorizadas. Segundo os últimos autores os estimadores-S da regressão também podem ser usados nessa situação, embora não sejam recomendados devido à sua baixa eficiência. Como os estimadores-MM da regressão, recomendados no Capítulo 4 pelo conjunto das suas propriedades, são obtidos sobre os estimadores-M e os estimadores-S e ambos suportam variáveis explicativas de vários tipos conclui-se que são o método adequado também neste caso. Para obter os resultados apresentados daqui em diante e relacionados com este método de estimação usou-se a função `lmRobMM` do S-Plus, com todas as opções por defeito (para mais detalhes ver S-Plus, 2000, Cap. 9). Deixa-se só aqui a nota de que, ao contrário da implementação em R a do S-Plus calcula os erros padrão com base nas distribuições assintóticas dos estimadores-M (Yohai *et al.*, 1991).

Os resultados relativos à estimação dos parâmetros do modelo (5.3) para os dados da escoliose admitindo a não correlação dos erros

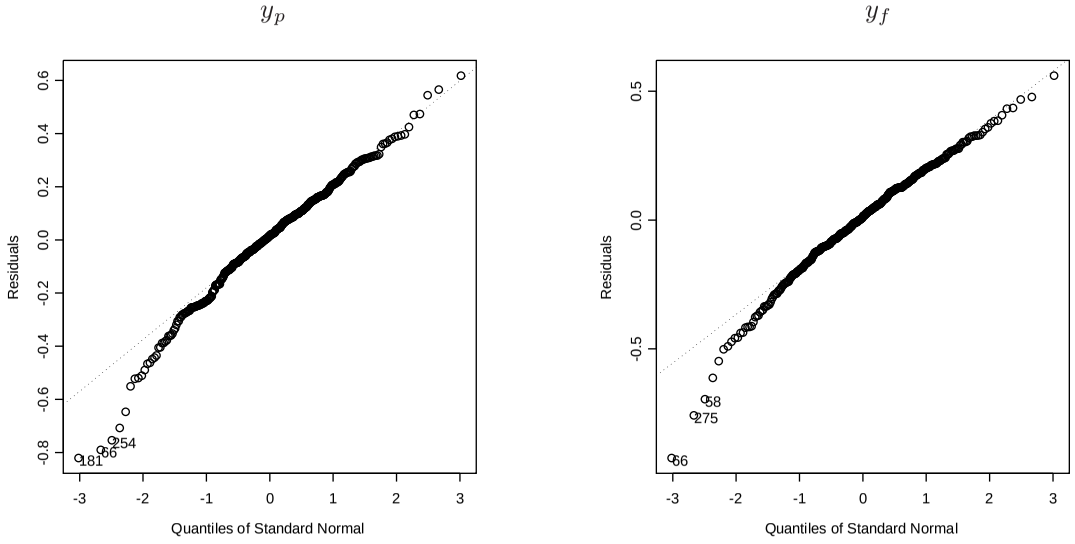


Figura 5.4 *Gráficos quantil-quantil (para a distribuição normal) dos resíduos do modelo transformado para y_p e y_f, após estimação por máxima verosimilhança de (ρ_1, ρ_2).*

encontram-se também nas Tabelas 5.6 e 5.7, para y_p e y_f, respectivamente, nas colunas referenciadas MM. Em relação aos testes de permutação usou-se $B = 5000$, devido a problemas de tempo de computação. Analisando os resultados dos testes de Durbin-Watson conclui-se que se rejeita a hipótese de correlação nula, quer com a versão clássica, quer com a versão robusta, o que significa que os resultados anteriores não se devem a nenhum efeito criado artificialmente devido à existência de *outliers*.

Para confirmar se a estrutura que está a ser detectada pelo teste de Durbin-Watson tem de facto a ver com a existência de curvas duplas procedeu-se ao seguinte exercício: ajustou-se o modelo com erros não correlacionados ao subconjunto dos dados constituído só pelas curvas simples (210 casos). Se o valor-p significativo do teste de Durbin-Watson for devido unicamente à existência de curvas duplas, então para este subconjunto de dados os valores devem deixar de ser significativos. Os resultados são apresentados na Tabela 5.5 para as duas variáveis resposta y_p e y_f com todas as variáveis explicativas, usando os métodos de estimação MQ e MM (o número de permutações usado foi $B = 10000$ para os MQ e $B = 5000$ para os MM).

Tabela 5.5 *Resultados dos testes de diagnóstico para os resíduos do modelo linear com erros não correlacionados ajustado por mínimos quadrados e regressão MM a um subconjunto dos dados da escoliose que contém apenas as curvas simples ($v_2 = 1$ ou $v_2 = 2$).*

	y_p (MQ)	y_f (MQ)	y_p (MM)	y_f (MM)
dw	1.55	1.58	1.46	1.56
$p_{DW,B}$	0.000	0.002	0.000	0.005
dw_R	1.41	1.55	1.60	1.80
$p_{DWR,B}$	0.000	0.009	0.018	0.171

O que se pode concluir da análise dos resultados apresentados na tabela é o seguinte:

(i) Quando os parâmetros do modelo linear são estimados por MQ observa-se que tanto o teste de Durbin-Watson clássico como o robusto produzem valores-p altamente significativos. Pode-se imaginar que afinal existe alguma auto-correlação ou tendência

temporal nos dados (recorde-se o modo como as observações estão ordenadas, referido na Secção 5.2).

(ii) No entanto, quando os parâmetros são estimados pelo método robusto MM, apenas o teste clássico rejeita claramente a não correlação.

(iii) A única explicação possível para estes resultados é: o método dos mínimos quadrados foi perturbado por alguns *outliers* que produziram artificialmente uma tendência moderada nos resíduos, a qual é detectada tanto pelo teste clássico como pelo teste robusto. Os resíduos resultantes do ajustamento após a estimação MM, pelo contrário, devem ter *outliers* mas não apresentam auto-correlação nem tendência, o que explica que o teste de Durbin-Watson clássico forneça resultados significativos e o teste robusto não.

(iv) De salientar que foi necessário usar simultaneamente métodos robustos de estimação e de diagnóstico para chegar a estas conclusões satisfatórias.

O que se discute em seguida é o processo de adaptação do procedimento MM para obter estimadores robustos dos parâmetros do modelo de regressão quando há erros correlacionados.

Quando os parâmetros de correlação em $\mathbf{V}$ são conhecidos o que é adequado é usar o método MM sobre o modelo transformado (5.16). As estimativas obtidas desta forma representam-se daqui em diante por $\hat{\boldsymbol{\beta}}_{MMG}$ e $\hat{\sigma}^2_{MMG}$.

Para estimar conjuntamente os parâmetros de correlação e os do modelo linear o que se propõe é uma adaptação do método da máxima verosimilhança descrito na secção anterior, e que pode ser visto como uma espécie de método "plug-in" ("substituição") frequentemente usado em estatística robusta. Este método consiste em substituir numa dada expressão ou função objectivo os estimadores não robustos por versões robustas adequadas, e foi o procedimento usado atrás para obter o teste de Durbin-Watson robusto. Na situação em causa este método consiste em fazer o mesmo tipo de procura na grelha (ρ_1, ρ_2) usando a função log-verosimilhança perfilada (5.24) mas com $\hat{\boldsymbol{\beta}}(\rho_1, \rho_2)$ de (5.20) e $\hat{\sigma}^2(\rho_1, \rho_2)$ de (5.21) estimados por

MMG em vez de MQG:

$$\begin{aligned}\hat{\mathcal{L}}(\rho_1,\rho_2) = \mathcal{L}(\hat{\boldsymbol{\beta}}_{MMG}(\rho_1,\rho_2), \hat{\sigma}^2_{MMG}(\rho_1,\rho_2), \rho_1, \rho_2|\mathbf{X},\mathbf{y}) = \\ = C - \frac{n}{2} - \frac{n}{2}\log\hat{\sigma}^2_{MMG}(\rho_1,\rho_2) - \\ - \frac{n_1}{2}\log(1-\rho_1^2) - \frac{n_2}{2}\log(1-\rho_2^2)\,, \quad (5.24)\end{aligned}$$

A solução pode ser vista como um ponto que minimiza

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, s, \rho_1, \rho_2|\mathbf{X},\mathbf{y}) = C - n\log s(\boldsymbol{\beta}) - \\ - \eta\left[\frac{1}{2s^2(\boldsymbol{\beta})}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] - \\ - \frac{n_1}{2}\log(1-\rho_1^2) - \frac{n_2}{2}\log(1-\rho_2^2), \quad (5.25)\end{aligned}$$

onde s representa um estimador robusto da escala dos resíduos e η é uma função objectivo limitada relacionada com a regressão MM. Pela sua construção este procedimento herda a eficiência e as propriedades de robustez, nomeadamente o ponto de rotura, dos estimadores robustos da regressão usados nos vários passos. Este método será daqui em diante identificado pela sigla MMV. Aplicando o procedimento de busca acabado de descrever obtiveram-se as estimativas pontuais,

$$(\hat{\rho}_1, \hat{\rho}_2) = (0.68, -0.07) \quad \text{e} \quad (\hat{\rho}_1, \hat{\rho}_2) = (0.71, -0.78),$$

para y_p e y_f, respectivamente. A Figura 5.5 contém os gráficos de curvas de nível da função $\hat{\mathcal{L}}(\rho_1,\rho_2)$. Nota-se a existência de vários máximos locais nos dois gráficos e que, como é usual com este tipo de modificações, a superfície não tem um aspecto tão regular como a baseada na verosimilhança.

As restantes estimativas e os resultados dos testes são apresentados nas Tabela 5.6 e 5.7 na coluna MMV (novamente com $B = 5000$ permutações). Em relação aos diagnósticos os resultados são muito melhores que quaisquer dos anteriores e completamente satisfatórios (ou quase, considerando que $p_{DWR,B}$ para y_f é 0.044, ligeiramente inferior a 0.05).

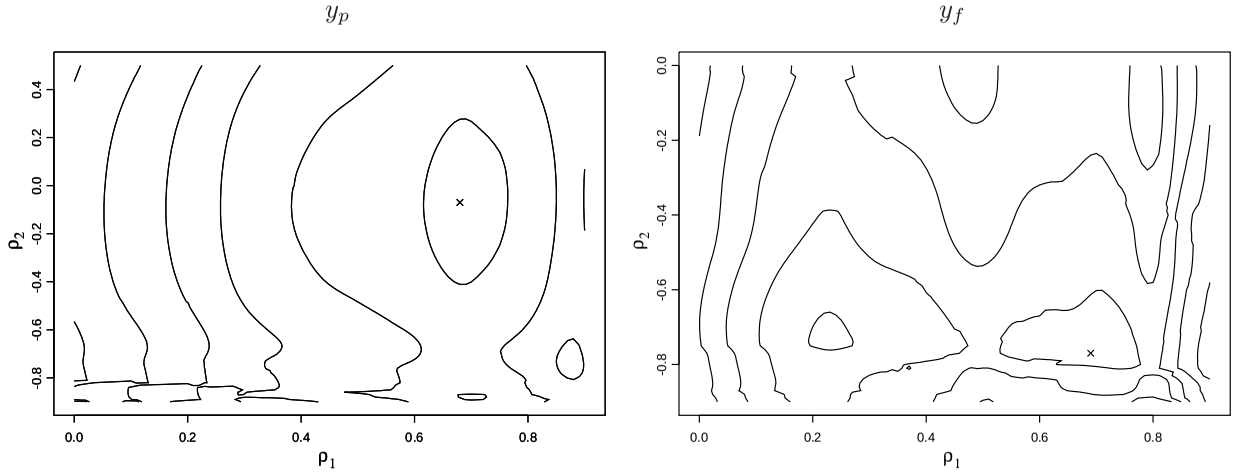


Figura 5.5 *Gráficos de curvas de nível de* $\hat{\mathcal{L}}(\rho_1, \rho_2)$ *para* y_p *e* y_f *com todas as variáveis explicativas. As estimativas pontuais estão marcadas com* $\times$.

5.4 Resultados

Nesta secção apresentam-se os resultados completos da aplicação dos métodos descritos na secção anterior aos dados da escoliose.

As Tabelas 5.6 e 5.7 contêm o conjunto completo dos resultados para as variáveis resposta y_p e y_f com todas as variáveis explicativas. Alguns comentários em relação aos resultados dos testes de diagnóstico foram já sendo feitos durante a apresentação dos métodos. O comentário geral mais importante é que é possível observar uma das principais consequências indesejáveis (e conhecidas) de desprezar a estrutura de correlação: há variáveis que não têm efeito significativo mas que parecem tê-lo. Isto acontece com a variável v_4 (cirurgião) para y_f quando a estimação é feita por MQ e com as variáveis v_3 (sexo) e v_4 para y_p quando se faz a estimação pelo método MM.

Por todas as razões já apresentadas conclui-se que os resultados mais satisfatórios são os obtidos através do método MMV. No entanto para concluir a análise é necessário considerar a eliminação das variáveis que foram consideradas como não tendo efeito significativo na explicação da variável resposta (v_3 e v_4). Isso foi feito mas não se apresentam os resultados por serem muito semelhantes aos obtidos com todas as variáveis, incluindo as estimativas das correlações (ρ_1, ρ_2). As estimativas finais dos parâmetros para interpretação médica seriam essas mas para os objectivos deste capítulo são irrelevantes.

Para concluir refira-se uma interpretação curiosa dos resultados que tem a ver com os parâmetros de correlação. As estimativas finais (das tabelas não apresentadas) são $(\hat{\rho_1}, \hat{\rho_2}) = (0.69, 0.10)$ e $(\hat{\rho_1}, \hat{\rho_2}) = (0.69, -0.77)$, para y_p e y_f, respectivamente. Estes valores indicam que para o ângulo pós-operatório existe, tal como esperado, uma correlação positiva moderada a elevada entre os resultados das correcções cirúrgicas para o mesmo doente, se este sofrer de escoliose dupla toraco-lombar, mas uma correlação positiva insignificante se sofrer de escoliose dupla torácica. Para o ângulo após *follow-up* há uma diferença importante, enquanto para as escolioses duplas do tipo toraco-lombar (bem separadas ao longo da coluna) se observa o mesmo valor de correlação que em relação ao pós-operatório, para as escolioses duplas torácicas, pelo contrário, obteve-se uma estimativa negativa da correlação (mas que em termos absolutos é a mais

Tabela 5.6 *Resultados para y_p com todas as variáveis explicativas (erros padrão entre parêntesis). Código de significância (valor-p):* $0 < *** < 0.001 < ** < 0.01 < * < 0.05$.

	MQ	MV	MM	MMV
$\hat{\rho}_1$	0.00	0.38	0.00	0.68
$\hat{\rho}_2$	0.00	0.68	0.00	−0.07
$\hat{\beta}_0$	-1.975^{***} (0.211)	-1.808^{***} (0.219)	-1.148^{***} (0.216)	-1.679^{***} (0.251)
$\hat{\beta}_1$	1.096^{***} (0.041)	1.054^{***} (0.041)	1.039^{***} (0.042)	1.069^{***} (0.045)
$\hat{\beta}_2$	0.339^{***} (0.056)	0.342^{***} (0.060)	0.134^{*} (0.060)	0.275^{***} (0.070)
$\hat{\beta}_3(1)$	-0.066^{**} (0.022)	-0.065^{**} (0.025)	-0.065^{**} (0.022)	-0.066^{**} (0.025)
$\hat{\beta}_3(2)$	−0.051 (0.033)	−0.048 (0.036)	−0.017 (0.033)	−0.048 (0.038)
$\hat{\beta}_3(3)$	−0.032 (0.027)	−0.029 (0.028)	−0.031 (0.027)	−0.022 (0.028)
$\hat{\beta}_3(4)$	−0.035 (0.027)	−0.040 (0.028)	0.021 (0.028)	−0.020 (0.028)
$\hat{\beta}_3(5)$	0.185^{***} (0.051)	0.176^{***} (0.044)	0.095^{*} (0.046)	0.166^{**} (0.056)
$\hat{\beta}_3(6)$	−0.001 (0.050)	0.006 (0.044)	−0.003 (0.046)	−0.010 (0.055)
$\hat{\beta}_4(0)$	0.012 (0.012)	0.012 (0.013)	0.033^{**} (0.013)	0.013 (0.015)
$\hat{\beta}_4(1)$	−0.012 (0.012)	−0.012 (0.013)	-0.033^{**} (0.013)	−0.013 (0.015)
$\hat{\beta}_5(1)$	0.004 (0.016)	0.010 (0.017)	−0.006 (0.016)	0.026 (0.019)
$\hat{\beta}_5(2)$	−0.035 (0.020)	−0.036 (0.021)	-0.058^{**} (0.020)	−0.046 (0.025)
$\hat{\beta}_5(3)$	0.031 (0.016)	0.026 (0.018)	0.064^{***} (0.017)	0.020 (0.020)
$\hat{\sigma}$	0.222	0.222	0.207	0.207
dw	1.481	1.721	1.462	1.954
$p_{DW,B}$	0.000	0.004	0.000	0.701
dw_R	1.461	1.649	1.542	1.783
$p_{DWR,B}$	0.000	0.002	0.000	0.061

Tabela 5.7 *Resultados para y_f com todas as variáveis explicativas (erros padrão entre parênteses). Código de significância (valor-p):* $0 < *** < 0.001 < ** < 0.01 < * < 0.05$.

	MQ	MV	MM	MMV
$\hat{\rho}_1$	0.00	0.46	0.00	0.71
$\hat{\rho}_2$	0.00	−0.33	0.00	−0.78
$\hat{\beta}_0$	-1.337^{***} (0.202)	-1.205^{***} (0.205)	-1.494^{***} (0.214)	-1.301^{***} (0.229)
$\hat{\beta}_1$	1.012^{***} (0.039)	0.982^{***} (0.039)	1.084^{***} (0.044)	1.032^{***} (0.043)
$\hat{\beta}_2$	0.263^{***} (0.053)	0.261^{***} (0.055)	0.216^{***} (0.056)	0.226^{***} (0.059)
$\hat{\beta}_3(1)$	-0.057^{**} (0.021)	-0.056^{**} (0.020)	-0.063^{**} (0.022)	-0.065^{**} (0.021)
$\hat{\beta}_3(2)$	-0.090^{**} (0.032)	-0.090^{**} (0.031)	-0.119^{***} (0.034)	-0.110^{**} (0.035)
$\hat{\beta}_3(3)$	0.002 (0.026)	0.005 (0.023)	0.005 (0.027)	0.006 (0.024)
$\hat{\beta}_3(4)$	0.005 (0.026)	0.003 (0.023)	0.038 (0.028)	0.032 (0.024)
$\hat{\beta}_3(5)$	0.164^{***} (0.048)	0.158^{**} (0.051)	0.191^{***} (0.054)	0.183^{**} (0.064)
$\hat{\beta}_3(6)$	−0.024 (0.048)	−0.020 (0.050)	−0.052 (0.050)	−0.046 (0.061)
$\hat{\beta}_4(0)$	0.003 (0.012)	0.008 (0.012)	0.010 (0.012)	0.016 (0.013)
$\hat{\beta}_4(1)$	−0.003 (0.012)	−0.008 (0.012)	−0.010 (0.012)	−0.016 (0.013)
$\hat{\beta}_5(1)$	-0.040^{**} (0.015)	−0.031 (0.016)	−0.018 (0.016)	−0.014 (0.017)
$\hat{\beta}_5(2)$	0.019 (0.019)	0.017 (0.020)	−0.007 (0.020)	−0.003 (0.022)
$\hat{\beta}_5(3)$	0.021 (0.015)	0.014 (0.016)	0.025 (0.016)	0.017 (0.018)
$\hat{\sigma}$	0.212	0.210	0.194	0.207
dw	1.454	1.728	1.467	1.956
$p_{DW,B}$	0.000	0.005	0.000	0.682
dw_R	1.374	1.693	1.440	1.767
$p_{DWR,B}$	0.000	0.011	0.000	0.044

elevada). Estes resultados reflectem, muito provavelmente, a grande dificuldade em manter, simultaneamente, ou seja, para as duas curvas, a correcção conseguida após a cirurgia, se as curvas estiverem muito próximas.

5.5 Discussão e conclusões

Neste capítulo apresentou-se uma aplicação dos métodos robustos na análise de um conjunto de dados com várias peculiaridades. Julga-se que o exemplo é útil porque ilustra várias metodologias e toca em vários aspectos que foram focados ao longo do texto.

É de referir que as ideias apresentadas neste capítulo podem encontrar, após adaptações ligeiras, aplicação em várias outras situações de regressão não *standard*, tais como erros correlacionados aos pares (é um caso particular do apresentado), outros casos de medições repetidas, erros auto-correlacionados ou erros auto-regressivos mais gerais. Basta para cada uma dessas situações seleccionar a matriz $\mathbf{V}$ apropriada. O único aspecto que é preciso salvaguardar tem a ver com o número de parâmetros desconhecidos da matriz, se ele for excessivo os métodos são computacionalmente mais difíceis e podem também tornar-se instáveis.

Particularmente interessante é o facto de os dados reais aqui analisados violarem simultaneamente duas das suposições do modelo de regressão linear múltipla usual (que era o modelo indicado para os analisar): a normalidade e a independência. O que se mostrou foi, por um lado, a dificuldade sentida pelos métodos clássicos (mínimos quadrados e teste de Durbin-Watson) devida essencialmente à presença de *outliers* e, por outro lado, que a aplicação de um método robusto "de pacote" não permite resolver simultaneamente todas as falhas das hipóteses. Neste caso foi preciso adaptar o método de regressão MM (que funciona bem na presença de *outliers*) mas não resolve só por si a questão quando falha a independência. Esta conclusão vem, como também foi referido na Secção 4.6, na linha das conclusões do Exemplo 4.4 onde se mostrou que aquele método *per si* também não responde satisfatoriamente na presença simultânea de heterocedasticidade e erros com caudas pesadas.

Referências Bibliográficas

Abbey, S. (1988). Robust measures and the estimator limit. *Geostandards Newsletter*, **12**, 241–248.

Ali, M. e Sharma, S. (1993). Robustness to nonnormality of the Durbin-Watson test for autocorrelation. *Journal of Econometrics*, **57**, 117–136.

Amado, C. (2003). *Bootstrap Robusto com base na Função de Influência*. Tese de Doutoramento, Instituto Superior Técnico, UTL, Lisboa.

Amado,C. e Pires, A.M. (2004). Robust bootstrap with non-random weights based on the influence function. *Communications in Statistics – Simulation and Computation*, **33**, 377–396.

Analytical Methods Committee (1989a). Robust statistics – how not to reject outliers. Part 1. Basic concepts. *The Analyst*, **114**, 1693–1697.

Analytical Methods Committee (1989b). Robust statistics – how not to reject outliers. Part 2. Inter-laboratory trials. *The Analyst*, **114**, 1699–1702.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. e Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton, NJ: Princeton University Press.

Atkinson, A. e Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer.

Barnett, V. e Lewis, T. (1994). *Outliers in Statistical Data*, 3ª. Edição. Chichester: Wiley.

Bassett, G. e Koenker, R. (1978). Asymptotic theory for least absolute error regression. *Journal of the American Statistical Association*, *73*, 618–622.

Beckman, R.J. e Cook, R.D. (1983). Outlier...s (com discussão).

Technometrics, **25**, 119–163.

Bierens, H.J. (1981). *Robust Methods and Asymptotic Theory in Nonlinear Econometrics*. New York: Springer.

Black, M.J., Fleet, D.J. e Yacoob, Y. (2000). Robustly estimating changes in image appearance. *Computer Vision and Image Understanding*, **78**, 8–31.

Black, M.J. e Rangarajan, A. (1996). On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, **19**, 57–91.

Black, M.J., Sapiro, G., Marimont, D.H. e Heeger, D. (1998). Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, **7**, 421–432.

Bosdogianni, P., Petrou, M. e Kittler, J. (1997). Mixed pixel classification with robust statistics. *IEEE Transactions on Geoscience and Remote Sensing*, **35**, 551–559.

Box, G.E.P. (1953). Non-normality and tests on variances. *Biometrika*, **40**, 318–335.

Box, G.E.P., Leonard, T. e Wu, C.F. (editores) (1983). *Scientific Inference, Data Analysis and Robustness*. New York: Academic Press.

Branco, J.A. (2005). Estatística robusta: contribuição portuguesa. Em Rosado, F. (editor), *Memorial da Sociedade Portuguesa de Estatística*. Lisboa: Edições SPE, 73–90.

Branco, J.A. e Pires, A.M. (2007). Poucos dados não é derrota e muitos dados não é vitória. Aceite para publicação em *Actas do XIV Congresso Anual da SPE*.

Bustos, O. e James, K.L. (1980). *Procedimentos Robustos*. 4º Simpósio Nacional de Probabilidade e Estatística. Rio de Janeiro.

Canjels, E. (2002). A permutation version of the Durbin-Watson test for serial correlation. Working paper, New School University.

Chaturvedi, A. e Shalabh (2004). Risk and Pitman closeness properties of feasible generalized double k-class estimators in linear regression models with non-spherical disturbances under balanced loss function. *Journal of Multivariate Analysis*, **90**, 229–256.

Chau, T. e Parker, K. (2004). On the robustness of stride frequency estimation. *IEEE Transactions on Biomedical Engineering*, **51**,

294-303.

Coakley, C.W. e Hettmansperger, T.P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*, **88**, 872–880.

Cochrane, D. e Orcutt, G.H. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, **44**, 32–61.

Cook, R.D., Hawkins, D.N. e Weisberg, S. (1992). Comparison of model misspecification diagnostics using residuals from least mean of squares and least median of squares fits. *Journal of the American Statistical Association*, **87**, 419–424.

Croux, C. (1994). Efficient high-breakdown M-estimators of scale. *Statistics and Probability Letters*, **19**, 371–379.

Croux, C. e Dehon, C. (2003). Estimators of the multiple correlation coefficient: local robustness and confidence intervals. *Statistical Papers*, **44**, 315–334.

Croux, C. e Dehon, C. (2005). Robustness versus efficiency for nonparametric correlation measures. Trabalho não publicado.

Croux, C., Dhaene, G. e Hoorelbeke, D. (2004). Robust standard errors for robust estimators. Preprint, Dept. of Applied Economics, Faculty of Economics and Applied Economics, Katholieke Universiteit Leuven.

Croux, C. e Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, **71**, 161–190.

Croux, C. e Rousseeuw, P.J. (1992). A class of high-breakdown estimators based on subranges. *Communications in Statistics – Theory and Methods*, **21**, 1935–1951.

Davies, P.L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics*, **15**, 1269–1292.

Davies, P.L. e Gather, U. (2007). The breakdown point – examples and counterexamples. *REVSTAT - Statistical Journal*, **5**, 1–17.

De La Torre, F. e Black, M. (2001). Robust principal component analysis for computer vision. Em *Proceedings of the International Conference on Computer Vision*. ICCV-2001, 362–369.

de Mast, J. e Roes, K.C.B. (2004). Robust individuals control chart for exploratory analysis. *Quality Engineering*, **16**, 407–421.

Donoho, D.L. e Huber, P.J. (1983). The notion of breakdown point.

Em Bickel, P., Doksum, K. e Hodges, J. (editores), *A Festschrift for Erich Lehmann*. Belmont, California: Wadsworth, 157–184.

Draper, N.R. e Smith, H. (1998). *Applied Regression Analysis*, 3ª. Edição. New York: Wiley.

Dufour, J.M., Gaudry, M.J.I. e Liem, T.C. (1980). The Cochrane-Orcutt procedure numerical examples of multiple admissible minima. *Economics Letters*, **6**, 43–48.

Durbin, J. e Watson, G.S. (1950). Testing for serial correlation in least squares regression I. *Biometrika*, **37**, 409–428.

Durbin, J. e Watson, G.S. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, **38**, 159–178.

Durbin, J. e Watson, G.S. (1971). Testing for serial correlation in least squares regression III. *Biometrika*, **58**, 1–19.

Dutter, R., Filzmoser, P., Gather, U. e Rousseeuw, P.J. (editores) (2003). *Developments in Robust Statistics*. Heidelberg: Physica-Verlag.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: SIAM.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society B*, **54**, 83–127.

Ellis, S.P. e Morgenthaler, S. (1992). Leverage and breakdown in L_1 regression. *Journal of the American Statistical Association*, **87**, 143–148.

Engelen, S., Frosch Møller, S. e Hubert, M. (2007). Automatically identifying scatter in fluorescence data using robust techniques. *Chemometrics and Intelligent Laboratory Systems*, **86**, 35–51.

Evans, E. (1992). Robustness of size of tests of autocorrelation and heteroscedasticity to nonnormality. *Journal of Econometrics*, **51**, 7–24.

Fernholz, L.T. (1983). *von Mises Calculus for Statistical Functionals*. Lecture Notes in Statistics **19**. New York: Springer.

Field, C.A. e Hampel, F.R. (1982). Small sample asymptotic distributions of M-estimators of location. *Biometrika*, **69**, 29–46.

Fomenko, I., Durst, M. e Balaban, D. (2006). Robust regression for high throughput drug screening. *Computer Methods and Programs in Biomedicine*, **82**, 31–37.

Fraiman, R., Meloche, J., García-Escudero, L.A., Gordaliza, A., He, X., Maronna, R., Yohai, V.J., Sheather, S.J., McKean, J.W., Small, C.G. e Wood, A. (1999). Multivariate L-estimation. *Test*, **8**, 255–317.

Franke, J., Härdle, W. e Martin, D. (editores) (1985). *Robust and Nonlinear Time Series Analysis*. New York: Springer.

Gross, A.M. (1976). Confidence interval robustness with long-tailed symmetric distributions. *Journal of the American Statistical Association*, **71**, 409–416.

Gunderson, L.H. e Holling, C.S. (editores) (2001). *Panarchy: Understanding Transformations in Human and Natural Systems*. Washington, D.C.: Island Press.

Hampel, F.R. (1968). *Contributions to the Theory of Robust Estimation*. PhD Thesis, University of California, Berkeley.

Hampel, F.R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, **42**, 1887–1896.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383–393.

Hampel, F.R. (2000). Robust Inference. Research Report No. 93, Seminar für Statistik, ETH, Zürich.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. e Stahel, W.A. (1986). *Robust Statistics: The Approach based on Influence Functions*. New York: Wiley.

Hawkins, D.M. (1980). *Identification of Outliers*. London: Chapman and Hall.

Hawkins, D.M. (1994). The feasible solution algorithm for least trimmed squares regression. *Computational Statistics and Data Analysis*, **17**, 185–196.

Hettmansperger, T.P. e McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*. London: Arnold.

Hoaglin, D.C., Mosteller, F. e Tukey, J.W. (editores) (1983). *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Hoaglin, D.C., Mosteller, F. e Tukey, J.W. (editores) (1992). *Análise Exploratória de Dados. Técnicas Robustas*. Lisboa: Edições Salamandra.

Hodges, J.L. Jr. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. Em *Fifth*

Berkeley Symposium on Mathematical Statistics and Probability, Volume 1. Berkeley: University of California Press, 163–186.

Hodges, J.L. Jr. e Lehmann, E.L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics*, **34**, 598–611.

Hogg, R.V. (1979). Statistical robustness: one view of its use in applications today. *The American Statistician*, **33**, 108–115.

Hsu, Y. e Mei, H. (1998). Comparisons among three estimation methods in linear models when observations are pairwise correlated. *Journal of Statistical Computation and Simulation*, **60**, 223–236.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

Huber, P.J. (1972). Robust statistics: a review. *Annals of Mathematical Statistics*, **43**, 1041–1067.

Huber, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, **1**, 729–821.

Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.

Huber, P.J. (1996). *Robust Statistical Procedures*, 2ª. Edição. Philadelphia: SIAM.

Hubert, M. (2006). Robust calibration. Em Gemperline, P. (editor), *Practical Guide to Chemometrics*. Boca Raton, FL: CRC Press, 167–215.

Hubert, M., Pison, G., Struyf, A. e Van Aelst, S. (editores) (2004). *Theory and Application of Recent Robust Methods*. Basel: Birkhäuser.

Hubert, M. e Rousseeuw, P.J. (1997). Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference*, **57**, 153–163.

Hubert, M., Rousseeuw, P.J. e Van Aelst, S. (2007). High-breakdown robust multivariate methods. Aceite para publicação em *Statistical Science*.

Huhns, M.N. e Holderfield, V.T. (2002). Robust software. *IEEE Internet Computing*, **6**, 80–82.

Insua, D.R. e Ruggeri, F. (editores) (2000). *Robust Bayesian Analysis*. New York: Springer.

Jen, E. (2003). Stable or robust? What's the difference?. *Complexity*, **8**, 12–18.

Jurečková, J. e Sen, P.K. (1996). *Robust Statistical Procedures: Asymptotics and Intercorrelations*. New York: Wiley.

Kadane, J.B. (editor) (1984). *Robustness of Bayesian Analysis*. Amsterdam: North-Holland.

Kalina, J. (2002). Autocorrelated residuals of least trimmed squares regression. Em Safránková, J. (editor), *Proceedings of the 11th Annual Conference of Doctoral Students – WDS 2002, Prague, 11th June–14th June, 2002*. Prague: MATFYZPRESS, 198–203.

Kariya, T. e Sinha, B.K. (1989). *Robustness of Statistical Tests*. New York: Academic Press.

Künsch, H. (1984). Infinitesimal robustness for autoregressive processes. *Annals of Statistics*, **12**, 843–863.

Lambert, D. (1981). Influence functions for testing. *Journal of the American Statistical Association*, **76**, 649–657.

Launer, E. e Wilkinson G. (Eds.) (1979). *Robustness in Statistics*. New York: Academic Press.

Lawrence, K.D. e Arthur, J.L. (1990). *Robust Regression: Analysis and Applications*. New York: Marcel Dekker.

Lax, D.A. (1985). Robust estimators of scale: finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, **80**, 736–741.

Lopuhaä, H.P. (1991). Multivariate τ-estimators for location and scatter. *The Canadian Journal of Statistics*, **19**, 307–321.

Lucas, A., Franses, P.H. e Van Dijk, D. (2005). *Outlier Robust Analysis of Economic Time Series*. Oxford: Oxford University Press.

Maddala, G.S. e Rao, C.R. (editores) (1997). *Robust Inference*. Handbook of Statistics **15**. Amesterdam: Elsevier.

Maechler, M. (2006). *Basic Robust Statistics: The robustbase Package*. Reference Manual.

Marazzi, A. (1993). *Algorithms, Routines and S Functions for Robust Statistics*. Belmont, CA: Wadsworth.

Maronna, R.A., Martin, R.D. e Yohai, V.J. (2006). *Robust Statistics, Theory and Methods*. Chichester: Wiley.

Maronna, R.A. e Yohai, V.J. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, **89**, 197–214.

Martin, R.D. e Yohai, V.J. (1986). Influence functionals for time series. *Annals of Statistics*, **14**, 781–818.

Meer, P., Stewart, C.V. e Tyler, D.E. (editores convidados) (2000). Special Issue on Robust Statistical Techniques in Image Understanding. *Computer Vision and Image Understanding*, **78**.

Michael, J.R. e Schucany, W.R. (1985). The influence curve and goodness-of-fit. *Journal of the American Statistical Association*, **80**, 678–682.

Morgenthaler, S., Ronchetti, E. e Stahel, W.A. (editores) (1993). *New Directions in Statistical Data Analysis and Robustness*. Basel: Birkhäuser.

Muller, C.H. (1997). *Robust Planning and Analysis of Experiments*. New York: Springer.

Murteira, B.J. (1993). *Análise Exploratória de Dados. Estatística Descritiva*. Lisboa: McGraw-Hill.

Myers, R.H. (1990). *Classical and Modern Regression with Applications*, 2ª. Edição. Boston: PSW-Kent.

Neter, J., Kutner, M.H., Nachtsheim, C.J. e Wasserman, W. (1996). *Applied Linear Statistical Models*, 4ª. Edição. Chicago: Irwin.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, **8**, 343–366.

Ong, E.P. e Spann, M. (1999). Robust optical flow computation based on least-median-of-squares regression. *International Journal of Computer Vision*, **31**, 51–82.

Pearson, E.S. (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika*, **21**, 259–286.

Pearson, E.S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, **23**, 114–133.

Pires, A.M. (1990). *Estimação Robusta e sua Aplicação a Componentes Principais*. Tese de Mestrado, Instituto Superior Técnico, UTL, Lisboa.

Pires, A.M. (1995). *Análise Discriminante: Novos Métodos Robustos de Estimação*. Tese de Doutoramento, Instituto Superior Técnico, UTL, Lisboa.

Pires, A.M. e Branco, J.A. (1991). Estimadores robustos da variância. Em Braumman, C. (editor), *Actas das XV Jornadas*

Luso-Espanholas de Matemática, Volume IV. Évora: Universidade de Évora, 175–180.

Pires, A.M. e Branco, J.A. (1994). Estatística robusta: passado, presente e futuro. Em Pestana, D. *et al.* (editores), *A Estatística e o Futuro e o Futuro da Estatística*. Lisboa: Edições Salamandra, 531–549.

Pires, A.M. e Branco, J.A. (2002). Partial influence functions. *Journal of Multivariate Analysis*, **83**, 451–468.

Pires, A.M. e Rodrigues, I.M. (2007). Multiple linear regression with some correlated errors: classical and robust methods. *Statistics in Medicine*, **26**, 2901–2918.

Pires, A.M. e Souto de Miranda, M.M. (editores convidados) (2007). Special Issue on Robust Statistics. *REVSTAT – Statistical Journal*, **5**.

Pitman, E.J.G. (1937). The closest estimator of statistical parameters. *Proceedings of the Cambridge Philosophical Society*, **33**, 212–222.

Portnoy, S. e He, X. (2000). A robust journey in the New Millennium. *Journal of the American Statistical Association*, **95**, 1331–1335.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society B*, **11**, 18–84.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353–360.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. URL http://www.R-project.org.

Rasch, D. e Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, **46**, 175–208.

Rasch, D. e Tiku, M.L. (editores) (1984). *Robustness of Statistical Methods and Nonparametric Statistics*. Dordrecht: Reidel.

Resina, J. (1963). Redressement et stabilisation des scolioses par un tuteur métallique. Em *Association Européene contre la poliomyelite, 9ème Symposium*. Paris: Masson et Cie, 421–429.

Resina, J. e Ferreira-Alves, A. (1977). A technique of correction and internal fixation for scoliosis. *Journal of Bone and Joint Surgery [British]*, **59-B**, 159–165.

Resina, J. e Ferreira-Alves, A. (1985). Portuguese method of correction of scoliosis and kyphosis. Em Bradford, D.S. e Hensinger,

R.M. (editores) *The Pediatric Spine*. New York: Thieme Inc., 518–528.

Rey, W.J. (1978). *Robust Statistical Methods*. New York: Springer.

Riani, M., Ceriolli, A. e Chiandotto, B. (editores convidados) (2007). Special Issue on Robust Multivariate Analysis and Classification. *Statistical Methods and Applications*, **15**.

Richardson, M.L. (2001). *Approaches To Differential Diagnosis In Musculoskeletal Imaging*. E-book, University of Washington. URL http://www.rad.washington.edu/mskbook/scoliosis.html

Rieder, H. (1994). *Robust Asymptotic Statistics*. New York: Springer.

Rieder, H. (editor) (1996). *Robust Statistics, Data Analysis and Computer Intensive Methods*. New York: Springer.

Ripley, B.D. (2004). *Robust Statistics*. Notas de um curso disponíveis em http://www.stats.ox.ac.uk/pub/StatMeth/Robust.pdf.

Ronchetti, E. (2006). The historical development of robust statistics. ICOTS7, 7th International Conference on Teaching Statistics, Salvador, Brasil.

Rosado, F. (2006). *Outliers em Dados Estatísticos*. Lisboa: Edições SPE.

Rosenhead, J. (2001). Robustness analysis: keeping your options open. Em Rosenhead, J. e Mingers, J. (editores), *Rational Analysis for a Problematic World Revisited: problem structuring methods for complexity, uncertainty and conflit*. Chichester: Wiley, 181–207.

Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. Em Grossmann, W., Pflug, G., Vincze, I. e Wertz, W. (editores), *Mathematical Statistics and Applications*, Volume B. Dordrecht: Reidel, 283–297.

Rousseeuw, P.J. e Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273–1283.

Rousseeuw, P.J. e Croux, C. (1994). The bias of k-step M-estimators. *Statistics and Probability Letters*, **20**, 411–420.

Rousseeuw, P.J. e Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, **94**, 388–402.

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

Rousseeuw, P.J. e Ronchetti (1981). Influence curves for general statistics. *Journal of Computational and Applied Mathematics*, **7**, 161–166.

Rousseeuw, P.J. e Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

Rousseeuw, P.J. e Van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **85**, 633–651.

Rousseeuw, P.J. e Yohai, V.J. (1984). Robust regression by means of S-estimators. Em Franke, J., Härdle, W. e Martin, R.D. (editores), *Robust and Nonlinear Time Series Analysis*. New York: Springer, 256–272.

Ruppert, D. (1992). Computing S-estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*, **1**, 253–270.

Ryan, T.P. (1997). *Modern Regression Methods*. New York: Wiley.

Salibian-Barrera, M. (2000). *Contributions to the Theory of Robust Inference*. PhD Thesis, University of British Columbia, Vancouver, Canada.

Salibian-Barrera, M. (2003). Fast and stable bootstrap methods for robust estimates. Em Wegman, E. e Braverman, A. (editores), *Computing Science and Statistics*, **34**. Fairfax Station, VA: Interface Foundation of North America, 346–359.

Salibian-Barrera, M. e Zamar, R.H. (2002). Bootstrapping robust estimates of regression. *Annals of Statistics*, **30**, 556–582.

SAS Institute Inc. (2004). *What's New in SAS® 9.0, 9.1, 9.1.2 and 9.1.3*. Cary, NC: SAS Institute Inc.

Schmoyer, R.L. (1994). Permutation tests for correlation in regression errors. *Journal of the American Statistical Association*, **89**, 1507–1516.

Shevlyakov, G.L. e Vilchevski, N.O. (2002). *Robustness in Data Analysis: Criteria and Methods*. Leiden: Brill Academic Publishers.

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *Annals of Statistics*, **26**, 1719–1732.

S-Plus (2000). *S-Plus Guide to Statistics*, Volume 1. Seattle: Insightful Corporation.

S-Plus (2002). *S-Plus 6 Robust Library User's Guide*. Seattle: Insightful Corporation.

Stahel, W.A. (1991). Research directions in robust statistics. Em Stahel, W.A. e Weisberg, S. (editores), *Directions in Robust Statistics and Diagnostics*, Volume 2. New York: Springer, 243–278.

Stahel, W.A. e Weisberg, S. (editores) (1991). *Directions in Robust Statistics and Diagnostics*, Volumes 1 e 2. New York: Springer.

Staudte, R.G. e Sheather, S.J. (1990). *Robust Estimation and Testing*. New York: Wiley.

Stigler, S.M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. *Journal of the American Statistical Association*, **68**, 872–879.

Stigler, S.M. (1975). Contributions to the discussion of the meeting on robust statistics. Em *Proceedings of the 40th session of the ISI*, Book 1. 383–384.

Stigler, S.M. (1977). Do robust estimators work with real data? (com discussão). *Annals of Statistics*, **5**, 1055–1077.

Stigler, S.M. (1986). *The History of Statistics*. Cambridge, MA: Belknap Press.

Stromberg, A.J. (1997). Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference*, **57**, 321–334.

Stromberg, A.J. (2002). Computational issues in robust statistics. Comunicação apresentada em International Conference on Robust Statistics 2002, Vancouver, British Columbia, Canada, Maio 2002.

Student (W. S. Gosset) (1927). Errors of routine analysis. *Biometrika*, **19**, 151–164.

Teichroew, D. (1956). Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution. *Annals of Mathematical Statistics*, **27**, 410–426.

Tiku, M.L. e Balakrishnan, N. (1986). *Robust Inference*. New York: Marcel Dekker.

Tukey, J.W. (1958). Bias and confidence in not-quite large samples (abstract). *Annals of Mathematical Statistics*, **29**, 614.

Tukey, J.W. (1960). A survey of sampling from contaminated distributions. Em Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., e Mann, H.B. (editores), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford: Stanford University Press, 448–485.

Tukey, J.W. (1975). Mathematics and the picturing of data. Em *Proceedings of the International Congress of Mathematicians*, **2**, Vancouver, 523–531.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Van Aelst, S. e Rousseeuw, P.J. (2000). Robustness of deepest regression. *Journal of Multivariate Analysis*, *73*, 82–106.

Van Aelst, S. Rousseeuw, P.J., Hubert, M. e Struyf, A. (2002). The deepest regression method. *Journal of Multivariate Analysis*, **81**, 138–166.

Wilcox, R.R. (2004). *Introduction to Robust Estimation and Testing*, 2ª. Edição. San Diego, CA: Academic Press.

Wilcox, R.R., Sheather, S., Brunner, E. e Schimek, M.G. (editores convidados) (2007). Special Issue on Nonparametric and Robust Methods. *Computational Statistics and Data Analysis*, **51**.

Wolfowitz, J. (1957). The minimum distance method. *Annals of Mathematical Statistics*, **28**, 75–88.

Yohai, V.J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics*, **15**, 642–656.

Yohai, V.J., Stahel, W.A. e Zamar, R.H. (1991). A procedure for robust estimation and inference in linear regression. Em Stahel, W.A. e Weisberg, S.W. (editores), *Directions in Robust Statistics and Diagnostics*, Volume 2. New York: Springer, 365–374.

Yohai, V.J. e Zamar, R.H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, **83**, 406–413.

Zhou, X. (2001). *Application of Robust Statistics to Asset Allocation Models*. Master of Science Thesis in Operations Research, MIT, Cambridge, MA.

Zuo, Y. (2004). Projection based affine equivariant multivariate location estimators with the best possible finite sample breakdown point. *Statistica Sinica*, **14**, 1199–1208.

Agradecimentos

A Sociedade Portuguesa de Estatística agradece às seguintes entidades o valioso apoio dado à realização do XV Congresso Anual:

Banco de Portugal

Caixa Geral de Depósitos

Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)

CERGER

Cruz Vermelha Portuguesa

Departamento de Métodos Quantitativos do ISCTE

Fundação Calouste Gulbenkian

Fundação para a Ciência e Tecnologia

Grupo de Investigação em Estatística e Análise de Dados do ISCTE (GIESTA)

INDEG / ISCTE *Business School*

Instituto Nacional de Estatística

Livraria Escolar Editora

PSE - Produtos e Serviços de Estatística, Lda.

Timberlake Consultores

WideScope