

[Home Page](#)

[Title Page](#)

[Contents](#)



[Page 1 of 104](#)

Estatística Aplicada em Ciências e Engenharia

A. Rita Gaio

Departamento de Matemática - FCUP

argai@fc.up.pt

November 9, 2016

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Home Page

Title Page

Contents

◀◀ ▶▶

◀ ▶

Page 2 of 104

Go Back

Full Screen

Close

Quit

Contents

[Home Page](#)
[Title Page](#)
[Contents](#)

[Page 3 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

1 Regressão Linear	5
1.1 Modelo de regressão linear	6
1.2 Interpretação dos parâmetros de regressão	10
1.3 Algumas regras de derivação vectorial/matricial	12
1.4 Estimação de parâmetros	13
1.5 Testes de hipóteses	19
1.5.1 Teste sobre um coeficiente β_j (teste de Wald)	19
1.5.2 Teste sobre um grupo de $p - r$ coeficientes	19
1.6 Intervalos de confiança	25
1.6.1 IC para os coeficientes de regressão	25
1.6.2 IC para a predição	25
1.6.3 Exemplo	27
1.7 Coeficiente de determinação	28
1.7.1 Definição	28
1.7.2 Significância de R^2	31
1.8 Multicolinearidade	32
1.9 MÉTODOS DE SELEÇÃO DE VARIÁVEIS	34
1.9.1 Forward	34
1.9.2 Backward	34
1.9.3 Stepwise	34
1.10 Comparação de modelos	37
1.10.1 Modelos encaixados	37

1.10.2 Modelos não encaixados	38
1.11 Gráfico seminormal	40
1.12 Diagnósticos	41
1.12.1 Outliers	41
1.12.2 Leverages	41
1.12.3 Resíduos Studentizados	44
1.12.4 Observações influentes	48
1.13 Análises gráficas	52
1.14 Instruções práticas	57
1.15 Variáveis explicativas categóricas	61
1.16 Confundimento	63
1.17 Interacções	64
1.18 Transformações para a linearidade	66
1.19 Transformações não lineares	67
2 Exemplos de Regressão Linear em R	69
2.1 Exemplo: dimensões corporais	70
2.2 Exemplo: funções respiratórias e tabaco	94
2.3 Exemplo: resultados eleitorais na Georgia (EUA) nas eleições presidenciais de 2000	103
2.4 Exercícios	104

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 4 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Chapter 1

Régressão Linear

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

Page 5 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.1. Modelo de regressão linear

Problema: modelação da relação entre uma variável contínua Y e um conjunto de variáveis (de qualquer natureza) X_1, \dots, X_p .

Objectivos:

- avaliação do efeito das variáveis X_1, \dots, X_p sobre Y
- predição de observações futuras.

Uma solução possível para o problema considera um modelo probabilístico em que a distribuição de probabilidade de Y depende do valor que as variáveis $X = (X_1, \dots, X_p)$ tomam:

$$P(Y \leq y | X = x) = F(y; \theta(x), \varphi).$$

Aqui:

- $F(\cdot; \theta, \varphi)$ é uma distribuição de probabilidade condicionada de $Y|X$ com parâmetros
- $(\theta, \varphi) = (\theta_1, \dots, \theta_q, \varphi_1, \dots, \varphi_r) \in \mathbb{R}^{q+r}$
- $\theta(x) = (\theta_1(x), \dots, \theta_q(x))$ são parâmetros de F que dependem dos valores de X
- $\varphi = (\varphi_1, \dots, \varphi_r)$ são parâmetros de F que não dependem de X .

O modelo de regressão linear assume que $Y|X$ segue uma distribuição normal

$$Y|X \sim N(\mu(X), \sigma^2(X))$$

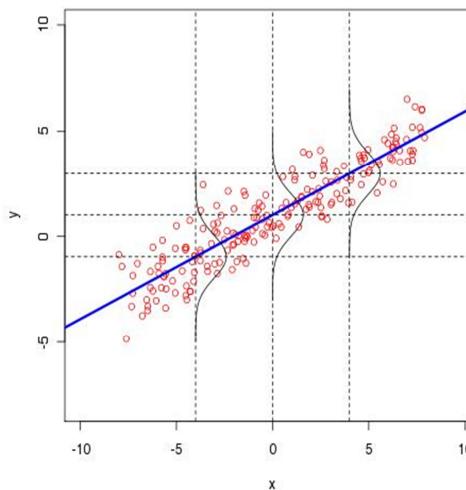
com média que depende linearmente das variáveis X

$$\begin{aligned}\mu(X) &= E(Y|X) \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\end{aligned}$$

e variância que poderá depender ou não de X . Portanto, para qualquer $x \in X$,

$$\begin{aligned}Y(x) &= \mu(x) + u \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + u\end{aligned}$$

onde $u \sim N(0, \sigma^2(x))$.



[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

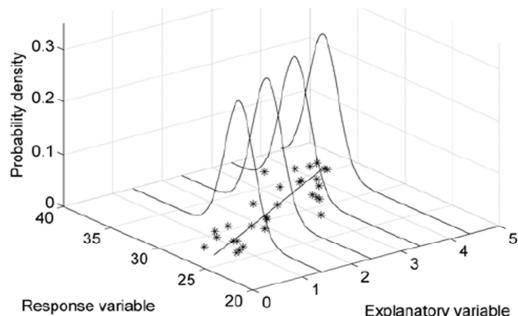
[Page 6 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



Neste contexto temos as seguintes designações:

- **Y : resposta ou variável dependente**
- X_1, \dots, X_p : **variáveis explicativas, independentes ou preditores**. Variáveis explicativas contínuas são frequentemente designadas por **covariáveis**.
- u : **erros do modelo (\hat{u} são os resíduos ...)**
- $\beta = (\beta_0, \beta_1, \dots, \beta_p)$: **coeficientes ou parâmetros da regressão** (a estimar); β_0 diz-se o **coeficiente independente ou termo constante (intercept)**
- $p = 1$: **regressão simples**
 $p > 1$: **regressão múltipla**
- $Y \in \mathbb{R}^k$ ($k > 1$): regressão multivariada múltipla.

Para n observações

$$(y_i; x_{i1}, \dots, x_{ip}) \quad i = 1, \dots, n$$

a forma matricial do modelo é

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

ou, de forma mais abreviada,

$$y = X\beta + u.$$

Cada **resposta y_i** corresponde a uma realização de uma v.a. Y_i , com

$$Y_i | x_i \sim N(\mu(x_i), \sigma^2(x_i)).$$

Analogamente, cada resíduo u_i corresponde também a uma realização de uma v.a. Se

as v.a. Y_i forem independentes entre si ^a

então as v.a. u_i são também independentes entre si e

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \sim N(0, \Sigma) \quad \text{com } \Sigma = \begin{pmatrix} \sigma^2(x_1) & & \\ & \ddots & \\ & & \sigma^2(x_n) \end{pmatrix}.$$

^asatisfaz se o mesmo indivíduo não fornece mais de uma observação para o modelo, por exemplo

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

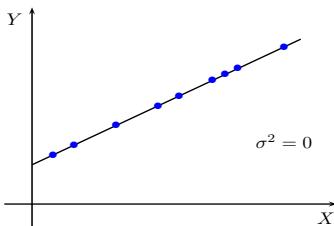
[Page 7 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



Para simplificação da exposição teórica,
assumimos daqui por diante que, para $i = 1, \dots, n$,

(H₁) as v.a. Y_i são independentes

(H₂) a variância de $Y_i | x_i$ não depende de x_i .

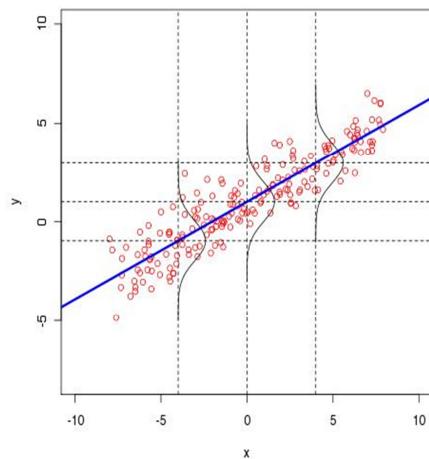
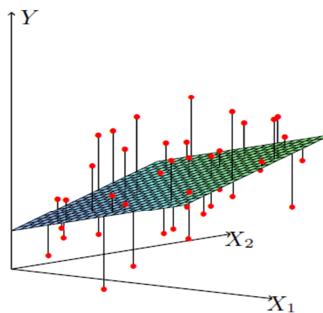
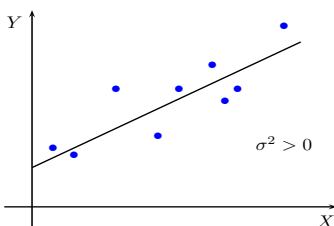
Tem-se portanto

$$\mathbf{Y} | \mathbf{x} \sim N_n(\mu(x), \sigma^2 \text{Id}).$$

ou, equivalentemente,

$$\mathbf{u} \sim N_n(0, \sigma^2 \text{Id}).$$

A matriz X é usualmente designada por matriz do modelo ou matriz do desenho^a, e o vector $X\beta$ por preditor linear.



[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 8 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

^amodel matrix ou design matrix

O modelo de regressão linear mais simples possui um único coeficiente de regressão e nenhuma variável explicativa. Para a observação i , tem-se

$$Y_i = \beta_0 + u_i, \quad u_i \sim N(0, \sigma^2).$$

Em particular,

$$E(Y_i) = \beta_0 + E(u_i) = \beta_0$$

ou seja, o modelo estima a mesma constante para as várias médias das respostas.

Este modelo é usualmente designado por **modelo nulo** e graficamente corresponde a uma recta horizontal.

Um outro modelo de referência é o **modelo saturado**, que consta de um parâmetro para cada observação. As variáveis explicativas deste modelo são as variáveis indicatrizes de cada uma das unidades experimentais. Para cada observação, resulta que a resposta prevista é igual à resposta observada.

Método generalizado dos mínimos quadrados:

Considere-se um modelo de regressão linear do tipo

$$Y = X\beta + u \quad \text{onde} \quad u \sim N(0, \sigma^2 V) \quad (1.1)$$

e $\sigma^2 V$ é a matriz de variância-covariância dos erros.

Estamos portanto a assumir eventualmente

erros correlacionados e heterocedásticos.

A ideia do método generalizado dos mínimos quadrados consiste em encontrar uma transformação que reduza o modelo (1.1) a um modelo em que os erros sejam independentes e homocedásticos.

Ora, a matriz V é simétrica e definida positiva. Então V é invertível e a sua inversa é também definida positiva. Pelo teorema da **decomposição de Cholesky**, existe uma matriz P (não necessariamente única), não singular, tal que

$$V^{-1} = P^t P.$$

Sejam agora

$$\tilde{Y} = PY, \quad \tilde{X} = PX, \quad \tilde{u} = Pu.$$

Obtém-se

$$\tilde{Y} = \tilde{X}\beta + \tilde{u}, \quad \text{com } \tilde{u} \sim N(0, \sigma^2 I),$$

isto é, os erros são agora independentes e homocedásticos.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 9 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.2. Interpretação dos parâmetros de regressão

Considere-se novamente o modelo de regressão linear

$$\begin{aligned}Y &= X\beta + u \\&= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + u\end{aligned}$$

sob a condição $u \sim N(0, \sigma^2 \text{Id})$ de os erros serem independentes e normalmente identicamente distribuídos.

O coeficiente β_0 representa a resposta esperada na situação em que todas as variáveis explicativas são nulas. Na impossibilidade de isto acontecer, esta constante deixa de ser particularmente interessante; aí pode-se, por exemplo, centrar cada uma das variáveis explicativas. A constante passará a representar a resposta esperada na situação em que todas as variáveis explicativas tomam o valor da média amostral correspondente.

Para $j = 1, \dots, p$, o **coeficiente** β_j representa o incremento na média de Y quando a variável explicativa X_j é aumentada de uma unidade e as restantes variáveis explicativas são mantidas constantes.

Consequentemente, β_j é uma medida da intensidade da relação entre Y e X_j quando o modelo está controlado para as restantes variáveis, isto é, após remoção do efeito dessas variáveis.

Sem perda de generalidade, mostramos a interpretação para $j = 1$.

Sejam x_i e x_k os vectores de observações dados por

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}), \quad x_k = (x_{i1}+1, x_{i2}, \dots, x_{ip}).$$

O indivíduo k tem valores observados para as variáveis explicativas iguais aos do indivíduo i , excepto para X_1 .

Para o valor esperado da resposta, tem-se agora

$$\begin{aligned}E(Y|x_k) &= \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + \cdots + \beta_p x_{kp} \\&= \beta_0 + \beta_1(x_i + 1) + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \\&= (\beta_0 + \beta_1 x_i + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) + \beta_1 \\&= E(Y|x_i) + \beta_1\end{aligned}$$

e portanto

$$\beta_1 = E(Y|x_k) - E(Y|x_i).$$

É importante notar que esta a interpretação apresentada para β_j só é possível graças à **linearidade do modelo seguido**. Em particular, a interpretação de um coeficiente β_j num modelo de regressão não linear não coincide necessariamente com a dada acima.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

Page 10 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Para além das estimativas usuais para os coeficientes de regressão β_j , $j = 0, 1, \dots, p$, alguns softwares de estatística, incluindo o **SPSS** e o **R**, apresentam também estimativas para os **coeficientes estandardizados**.

Os coeficientes estandardizados são úteis para comparar efeitos de variáveis explicativas. Correspondem aos coeficientes de regressão obtidos para as variáveis explicativas quando quer estas quer a resposta foram estandardizadas (têm média 0 e variância 1) antes do processo de regressão.

Para $j = 1, \dots, p$, o **coeficiente estandardizado** de β_j representa portanto o incremento na média de Y , em termos de desvios-padrão, resultante do aumento de um desvio-padrão na variável explicativa X_j e mantendo constantes as restantes variáveis explicativas.

Os coeficientes estandardizados de um modelo de regressão múltipla são portanto directamente comparáveis entre si, sendo que a variável explicativa que apresenta um maior coeficiente é a que possui maior influência sobre a resposta.

Note-se que o processo de estandardização não faz sentido para o coeficiente constante.

A questão da opção entre coeficientes estandardizados ou não estandardizados é deixada a cargo do investigador. Enquanto que os coeficientes estandardizados de modelos diferentes e/ou de estudos diferentes podem ser directamente comparáveis, também é verdade que, à partida, não há razões para crer que a alteração de um desvio-padrão numa variável explicativa seja equivalente à alteração de um desvio-padrão noutra variável.

No **R**, a instrução que permite obter os coeficientes estandardizados está contida no pacote *QuantPsyc*:

```
> library(QuantPsyc)
```

```
> lm.beta(model)
```

onde *model* é um objecto obtido de uma instrução *lm*.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

Page 11 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.3. Algumas regras de derivação vectorial/matricial

Sejam $x \in \mathbb{R}^n$ e $y \in \mathbb{R}^n$ dois vectores e $A \in \mathbb{R}^{n \times n}$ uma matriz (que não depende de x). Temos as seguintes regras de cálculo, para a derivação de algumas das operações mais comuns nos cálculos matriciais que se seguirão:

1. $\frac{\partial}{\partial z}(Ax) = A\frac{\partial x}{\partial z}$, para qualquer vector $z \in \mathbb{R}^n$
2. $\frac{\partial}{\partial x}(y^t Ax) = y^t A$
3. $\frac{\partial}{\partial x}(x^t Ax) = x^t(A + A^t)$.

Em particular, se A é simétrica tem-se

$$\frac{\partial}{\partial x}(x^t Ax) = 2x^t A.$$

$$4. \frac{\partial}{\partial z}(y^t x) = x^t \frac{\partial y}{\partial z} + y^t \frac{\partial x}{\partial z}.$$

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

Page 12 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.4. Estimação de parâmetros

O problema que agora se coloca é o da estimação dos parâmetros β e σ^2 . No caso do modelo linear, o que se vai mostrar já de seguida é que estimar o parâmetro β pelo método da máxima verosimilhança, mantendo σ^2 fixo, é equivalente a estimar β pelo método dos mínimos quadrados.

Muito sucintamente, o

método da máxima verosimilhança

consiste da determinação de estimativas para os parâmetros que maximizam a probabilidade de ocorrência dos dados, a chamada **verosimilhança**, no modelo probabilístico assumido para esses dados.

No caso da regressão linear, pela independência das observações, tem-se

$$L(\beta, \sigma^2 | (y_i, x_i)_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y_i - \hat{y}_i)^2}{\sigma^2}\right).$$

Dado que o logaritmo é uma função crescente, maximizar L é equivalente a maximizar $\log(L)$, dado por

$$\begin{aligned} & \log(L(\beta, \sigma^2 | (y_i, x_i)_i)) \\ &= \ell(\beta, \sigma^2 | (y_i, x_i)_i) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{\sigma^2} \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - X_i.\beta)^2. \end{aligned}$$

Daqui resulta que maximizar a função ℓ em relação a β é equivalente a minimizar a soma dos quadrados dos erros

$$RSS(\beta) = \sum_{i=1}^n (y_i - X_i.\beta)^2 = (y - X\beta)^t(y - X\beta),$$

uma vez que a primeira parcela de $\ell(\beta, \sigma^2 | (y_i, x_i)_i)$ é independente de β .

Aplicando regras de derivação vectorial,

$$\begin{aligned} \frac{\partial}{\partial \beta} RSS(\beta) &= 2(y - X\beta)^t \frac{\partial}{\partial \beta}(y - X\beta) \\ &= -2(y^t - \beta^t X^t)X \\ &= -2(y^t X - \beta^t X^t)X. \end{aligned}$$

Igualando esta expressão a zero e tomando a transposta obtém-se as chamadas **equações normais** para o estimador $\hat{\beta}$ (são $p+1$ equações):

$$X^t X \hat{\beta} - X^t y = 0$$

e portanto o estimador dos mínimos quadrados para β ou estimador de máxima verosimilhança é:

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

que só existe se as colunas X_1, \dots, X_p da matriz X não forem linearmente dependentes.

O facto de $\hat{\beta}$ ser um mínimo global resulta de a função $RSS(\beta)$ ser ilimitada.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 13 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

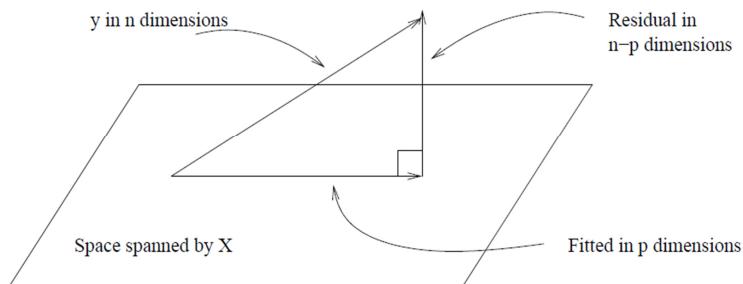
Para o estimador $\hat{\beta}$ tem-se

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ &= X(X^t X)^{-1} X^t y \\ &= Hy\end{aligned}$$

sendo $H \in \mathbb{R}^{n \times n}$ usualmente designada por **matriz-chapéu**^a. A matriz H corresponde à **projecção ortogonal de $y \in \mathbb{R}^n$ sobre o sub-espaco de \mathbb{R}^n de dimensão $p + 1$ gerado pelas colunas de X** .

Este resultado fornece uma interpretação geométrica do modelo de regressão linear. Em particular, observe-se que os resíduos são ortogonais aos valores previstos.

^ahat matrix


[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 14 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

O estimador $\hat{\beta} \in \mathbb{R}^{p+1}$ de β tem as seguintes propriedades:

$$(1) \quad E(\hat{\beta}) = \beta \quad (\text{i.e. é não enviesado})$$

$$(2) \quad \text{Var}(\hat{\beta})^a = \sigma^2(X^t X)^{-1}$$

$$(3) \quad \hat{\beta} \sim N(\beta, \text{Var}(\hat{\beta}))$$

(4) Teorema (Gauss-Markov):

$\hat{\beta}$ é o mais preciso de entre os estimadores lineares não enviesados de β , no sentido de ter menor variância (é um estimador eficiente).

As propriedades (1) e (2) acima resultam simplesmente de se ter

$$\begin{aligned} E(\hat{\beta}) &= (X^t X)^{-1} X^t E(y) \\ &= (X^t X)^{-1} X^t (X\beta) \\ &= (X^t X)^{-1} (X^t X)\beta \\ &= \beta \end{aligned}$$

e

$$\begin{aligned} \text{Var}(\hat{\beta}) &= (X^t X)^{-1} X^t \text{Var}(y)((X^t X)^{-1} X^t)^t \\ &= (X^t X)^{-1} X^t \sigma^2 I X (X^t X)^{-t} \\ &= \sigma^2 (X^t X)^{-1} (X^t X) ((X^t X)^t)^{-1} \\ &= \sigma^2 (X^t X)^{-1}. \end{aligned}$$

^amatriz de variância-covariância

Observações:

- mais à frente precisaremos de calcular o desvio-padrão de β_i , $i = 1, \dots, p$. Da propriedade (2), tem-se

$$se(\hat{\beta}_i) = \sigma \sqrt{(X^t X)_{ii}^{-1}}.$$

Aqui usamos a notação mais comum no contexto da regressão linear que consiste em usar $se(\hat{\beta}_j)$ ^a para um desvio-padrão e não um erro padrão da média.

- o Teorema de Gauss-Markov só é válido na condição de os erros serem independentes e terem a mesma variância.
- quando os erros são correlacionados ou têm variâncias diferentes, há estimadores melhores (no sentido de terem menor variância) do que $\hat{\beta}$ definido atrás. Deve-se usar estimação pelo método generalizado dos mínimos quadrados.
- quando a distribuição é assimétrica, deve-se usar estimação robusta, que normalmente não é linear em y
- quando há a necessidade de incluir vários preditores correlacionados, então devem-se usar estimadores enviesados, como por exemplo o fornecido pela regressão ridge.

^ase: standard error

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 15 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

O estimador para σ^2 obtém-se da expressão do logaritmo da verosimilhança em $\hat{\beta}$. Mais precisamente,

$$\frac{\partial}{\partial \sigma^2} \ell(\sigma^2 | \hat{\beta}, (y_i, x_i)_i) = \frac{1}{2\sigma^2} \left(-n + \frac{RSS(\hat{\beta})}{\sigma^2} \right)$$

que é igual a 0 para

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n}.$$

Este é o **estimador de máxima verosimilhança de σ^2** .

O estimador $\hat{\sigma}^2 \in \mathbb{R}$, de σ^2 , é enviesado. Um estimador não enviesado para σ^2 é

$$\bar{\sigma}^2 = \frac{RSS(\hat{\beta})}{n - (p + 1)}$$

e tem-se

$$(n - (p + 1)) \frac{\bar{\sigma}^2}{\sigma^2} \sim \chi^2(n - (p + 1)).$$

Mostramos agora que

$\bar{\sigma}^2$ é um estimador não enviesado de σ^2 .

Recorde-se que

$$X\hat{\beta} = Hy, \quad \text{com } H = X(X^t X)^{-1} X^t$$

e H é designada por matriz chapéu. Têm-se agora os seguintes factos (exercício):

- $H^2 = H$ e $(1 - H)^2 = (1 - H)$
Diz-se que H e $1 - H$ são idempotentes.
- $\hat{u} = y - X\hat{\beta} = (1 - H)y$
- $\hat{u} = (1 - H)y = \dots = (1 - H)u$
- $\hat{u}^t \hat{u} = u^t (1 - H)u$
- $RSS(\hat{\beta}) = \hat{u}^t \hat{u}$
- $$\begin{aligned} E(\hat{u}^t \hat{u}) &= E(u^t (1 - H)u) \\ &= \text{tr}E(u^t (1 - H)u), \text{ porque } E(u^t (1 - H)u) \in \mathbb{R} \\ &= E(\text{tr}(u^t (1 - H)u)), \text{ porque tr comuta com } E \\ &= E(\text{tr}((1 - H)uu^t)), \text{ porque tr}(AB) = \text{tr}(BA) \\ &= \text{tr}(E((1 - H)uu^t)), \text{ porque tr comuta com } E \\ &= \text{tr}((1 - H)E(uu^t)), \text{ porque } 1 - H \text{ não é v.a.} \\ &= \text{tr}((1 - H)\sigma^2 I) \\ &= \sigma^2 \text{tr}(1 - H) \\ &= \sigma^2(n - (p + 1)). \end{aligned}$$

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

[Page 16 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

O traço de uma matrix comuta com o valor esperado porque o traço é um operador linear.

Para calcular o traço de $1 - H$ usou-se novamente $\text{tr}(AB) = \text{tr}(BA)$:

$$\begin{aligned}\text{tr}(1 - H) &= \text{tr}(1_{n \times n}) - \text{tr}(X(X^t X)^{-1} X^t) \\ &= \text{tr}(1_{n \times n}) - \text{tr}(1_{(p+1) \times (p+1)}) \\ &= n - (p + 1).\end{aligned}$$

Finalmente, resulta de todas as igualdades anteriores que

$$E(\bar{\sigma}^2) = \frac{E(\hat{u}^t \hat{u})}{n - (p + 1)} = \sigma^2$$

e que portanto $\bar{\sigma}^2$ é um estimador não enviesado de σ^2 .

Pode-se provar ainda que

cada $\hat{\beta}_j$ é independente de $\bar{\sigma}^2$.

Resulta agora da propriedade (3) de $\hat{\beta}$ que

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t(n - (p + 1)), \quad j = 0, 1, \dots, p$$

onde $\text{se}(\hat{\beta}_j) = \bar{\sigma} \sqrt{(X^t X)_{jj}^{-1}}$ estima o desvio padrão de $\hat{\beta}_j$.

Na distribuição da estatística T_j usou-se o seguinte facto: se U e V são variáveis aleatórias independentes com $U \sim N(0, 1)$ e $V \sim \chi^2(n - 1)$ então

$$T = U / \sqrt{V/(n - 1)} \sim t(n - 1).$$

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 17 of 104

[Go Back](#)

[Full Screen](#)

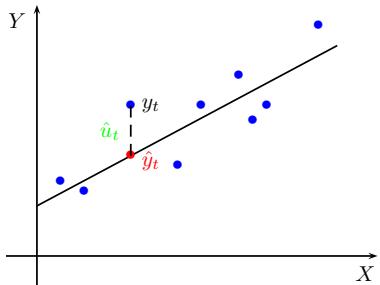
[Close](#)

[Quit](#)

Representando por $\hat{\beta}$ o estimador de máxima verosimilhança de β e por \hat{y}_i o **valor ajustado** (ou **previsto**) para a observação i , os **resíduos brutos** correspondem a

$$\hat{u}_i = y_i - \hat{y}_i$$

para $i = 1, \dots, n$.



Proposição: Num modelo de regressão linear com termo constante, tem-se:

1. $E(\hat{u}) = 0$
2. $X^t \hat{u} = 0$
3. $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$; em particular, a soma dos valores observados é igual à soma dos valores previstos

Observe-se que, para efeitos de estimação de parâmetros pelo método dos mínimos quadrados, teve que se definir a soma dos quadrados dos resíduos, já que a soma dos resíduos é sempre zero.

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 18 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

1.5. Testes de hipóteses

1.5.1. Teste sobre um coeficiente β_j (teste de Wald)

Suponhamos que se pretende testar a hipótese nula

$$H_0 : \beta_j = 0,^a \text{ para algum } j = 0, 1, \dots, p$$

parâmetro

que indica que o **coeficiente** independente $\beta_0 = 0$ é irrelevante para o modelo ($j = 0$) ou que a variável explicativa X_j não deve constar do modelo de regressão ($j \neq 0$).

Pelos resultados da secção anterior, pode-se usar a estatística de teste

$$T_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t(n - (p + 1)), \quad j = 0, 1, \dots, p$$

e proceder de forma usual.

Suponhamos agora que pelo menos duas variáveis explicativas, digamos X_j e X_k , não são significativas para o modelo ($p > 0.05$ para cada uma delas no teste anterior). Isso não implica necessariamente que ambas as variáveis devam ser excluídas do modelo. A eliminação de apenas uma variável pode alterar as estimativas dos restantes coeficientes pelo que pode acontecer que retirando X_j do modelo, por exemplo, faça com que X_k passe a ser significativa.

^amais geralmente, podemos testar $H_0 : \beta_j = \beta_j^0$ para um qualquer valor β_j^0 . A estatística de teste correspondente será

$$T_j = \frac{\hat{\beta}_j - \beta_j^0}{se(\hat{\beta}_j)} \sim t(n - (p + 1)).$$

Por este motivo, cada uma das variáveis explicativas não significativas deve ser retirada do modelo vez à vez. Caso se considere que a presença de uma dada variável explicativa é fundamental para a explicação/previsão da resposta, essa variável deve permanecer no preditor linear, mesmo que seja não significativa no teste anterior.

1.5.2. Teste sobre um grupo de $p - r$ coeficientes

Suponhamos agora que escrevemos β na forma

$$\beta = (\beta', \beta'') \quad \text{com} \quad \begin{cases} \beta' = & (\beta_0, \dots, \beta_r) \\ \beta'' = & (\beta_{r+1}, \dots, \beta_p) \end{cases}$$

onde $p - r > 1$, e que queremos testar a hipótese nula

$$H_0 : \beta'' = \beta''^0. \quad ^a$$

Sejam

$$\hat{\beta}^0 = (\hat{\beta}'^0, \hat{\beta}''^0) \quad \text{e} \quad \hat{\sigma}^0$$

os estimadores de máxima verosimilhança de β e σ^2 , respectivamente, na situação em que se assume H_0 . Pode-se mostrar que

$$F^0 = \frac{(RSS(\hat{\beta}^0) - RSS(\hat{\beta})) / ((p + 1) - (r + 1))}{RSS(\hat{\beta}) / (n - (p + 1))}$$

segue a distribuição $F(p - r, n - (p + 1))$.

^aNesta situação, H_0 incide sobre mais de um coeficiente. Em particular, a hipótese $\beta'' = 0$ indica que as variáveis explicativas X_{r+1}, \dots, X_p não são significativas para o modelo.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 19 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

O teste descrito averigua se um determinado grupo de variáveis explicativas é ou não significativo para o modelo de regressão. Equivalentemente, compara a qualidade do ajustamento do **modelo completo** (com todas as variáveis explicativas) com a do **modelo reduzido** (que substitui β do modelo completo pelo valor especificado na hipótese nula). A rejeição de H_0 significa que o modelo completo tem melhor ajustamento.

- No caso do grupo de variáveis explicativas testadas consistir apenas de uma variável, o teste coincide com o apresentado na subsecção anterior.
- No caso do grupo de variáveis explicativas testadas consistir de todas as variáveis

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

o modelo que se está a testar é aquele que consiste apenas do coeficiente independente β_0 . Equivalentemente, não existe nenhuma relação linear entre Y e as variáveis explicativas X .

A hipótese alternativa é

$$H_1: \text{existe } j \in \{1, \dots, p\} \text{ tal que } \beta_j \neq 0.$$

Nessa situação, tem-se então $\hat{\beta}^0 = (\hat{\beta}_0, 0, \dots, 0)$ e

$$F^0 = \frac{(RSS(\hat{\beta}^0) - RSS(\hat{\beta}))/p}{RSS(\hat{\beta})/(n - (p + 1))} \sim F(p, n - (p + 1)).$$

Observe-se que:

- (a) a falha de rejeição de H_0 nesta situação não indica necessariamente o fim da análise de regressão que se está a efectuar.

A análise pode prosseguir com transformações não lineares de variáveis ou identificação de outliers. Pode ainda acontecer que os dados sejam insuficientes para demonstrar a existência de um efeito real entre Y e X .

De qualquer forma, não é adequado testar cada um dos β_j individualmente.

- (b) a rejeição de H_0 não significa necessariamente que o modelo testado seja bom. Pode acontecer que se deva na verdade reter todas as variáveis explicativas mas também pode acontecer que apenas seja de reter algumas delas. A conclusão a tirar estará associada ao resultado de cada um dos testes de hipóteses efectuados sobre os coeficientes individuais (parciais) de regressão.

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀](#) [▶](#)
[◀](#) [▶](#)
[Page 20 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

Como resultado de uma análise de regressão, a maioria dos softwares de estatística apresenta uma tabela de análise da variância, ANOVA, onde se indicam alguns valores necessários ao desenvolvimento do teste de hipóteses descrito anteriormente, com hipótese nula

$$H_0 : \beta_1 = \cdots = \beta_p = 0.$$

A tabela da ANOVA tem geralmente o seguinte formato:

Modelo	Soma de quadrados	graus de liberdade	Quadrados médios	\hat{F}	valor- <i>p</i>
Regressão	$RegSS = \sum(\hat{y}_i - \bar{y}_Y)^2$	p	$RegSS/p$	$\frac{RegSS/p}{RSS/(n-(p+1))}$	$P(F > \hat{F})$
Residual	$RSS = \sum(y_i - \hat{y}_i)^2$	$n - (p + 1)$	$RSS/(n - (p + 1))$		
Total	$TSS = \sum(y_i - \bar{y}_Y)^2$	$n - 1$			

As várias somas de quadrados resultam da seguinte decomposição da variação total de Y (exercício):

$$\text{variação total de } Y = \text{variação devida à regressão} + \text{variação residual}$$

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ TSS &= RegSS + RSS. \end{aligned}$$

Um modelo com um bom ajustamento aos dados (portanto rejeitando a hipótese nula anterior) é um modelo em que a variação total é *essencialmente* devida à regressão, apresentando uma baixa variação residual.
Notar também que, sob H_0 referido acima, $RSS(\beta^0) = TSS$ e portanto

$$RSS(\hat{\beta}^0) - RSS(\hat{\beta}) = TSS - RSS = RegSS,$$

daí a expressão do numerador da estatística F apresentada na tabela.

Notação:

- RegSS: **R**egression **S**um of **S**quares,
- RSS: **R**esidual **S**um of **S**quares
- TSS: **T**otal **S**um of **S**quares

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

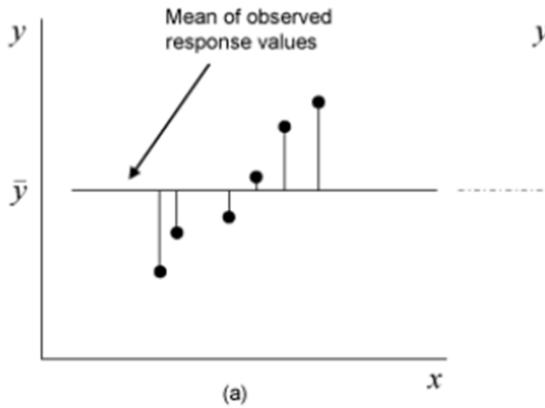
Page 21 of 104

[Go Back](#)

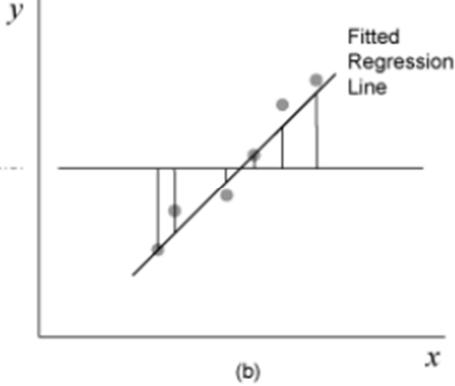
[Full Screen](#)

[Close](#)

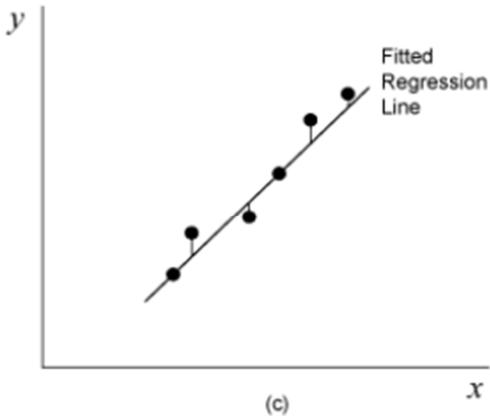
[Quit](#)



(a)



(b)



(c)

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

Page 22 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Em resumo:

- TSS: funciona como a soma dos quadrados dos resíduos para o modelo linear que consiste apenas de uma constante. Neste sentido, TSS representa a variabilidade da resposta que fica por explicar quando se ajusta um modelo que conste apenas de uma constante (a média amostral da resposta, \bar{y}).
- RSS: mede a variabilidade da resposta que fica por explicar quando o modelo em causa é ajustado
- RegSS: mede a variabilidade da resposta que é explicada pelo modelo linear em causa.

No R, a tabela da anova é sequencial, sendo que a segunda coluna (soma de quadrados) representa:

- para a primeira variável, X_1 , a soma dos quadrados RegSS da regressão $Y = X_1 + u$.
- para X_j , com $j > 1$, a redução na variabilidade que fica por explicar ao passar do modelo que usa X_1, \dots, X_{j-1} para o modelo que usa X_1, \dots, X_{j-1}, X_j . Isto corresponde à redução na soma dos quadrados dos resíduos, RSS, à medida que vamos adicionando mais uma variável.
- para os resíduos, a soma dos quadrados dos resíduos, RSS.

Variable	Df	Sum Sq
X_1	1	RegSS(X_1)
X_2	1	RegSS($X_2 X_1$)
:	:	:
X_p	1	RegSS($X_p X_1, \dots, X_{p-1}$)
Residuals	$n - p - 1$	RSS

Variable	Mean Sq	F value	Pr(> F)
X_1	xxx	xxx	xxx
X_2	xxx	xxx	xxx
:			
X_p	xxx	xxx	xxx
Residuals	xxx	xxx	xxx

A estatística de teste associada a cada uma das variáveis X_j , $j > 1$, é

$$\frac{\text{RegSS}(\mathcal{X}_j|\mathcal{X}_1, \dots, \mathcal{X}_{j-1})}{\text{RSS}/(n-p-1)} \sim F(1, n-p-1)$$

sob a hipótese nula $\beta_j = 0$.

O teste de Wald é coincidente com este teste desde que a variável em causa (cujo coeficiente está a ser testado pelo teste de Wald) seja a última variável da regressão.

Uma outra instrução para obtenção da soma RSS é
`>deviance(model.name)`

uma vez que, nos modelos lineares gaussianos, a desviância coincide com a soma RSS (veremos mais tarde).

Uma instrução em **R** que permite efectuar o teste de Wald sobre um coeficiente e o teste (de Wald, generalizado) sobre um grupo ds coeficientes é

```
> library(lmtest)
> waldtest(mod1,mod2)
```

onde *mod1* e *mod2* são os modelos em causa; um que resulta da assumpção da hipótese nula, e outro que resulta da rejeição da hipótese nula.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

[Page 24 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.6. Intervalos de confiança

1.6.1. IC para os ~~coeficientes~~ de regressão

Na secção sobre estimação de ~~coeficientes~~ viu-se que

$$T_j = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t(n - (p + 1)), \quad j = 0, 1, \dots, p$$

onde $se(\hat{\beta}_j) = \bar{\sigma} \sqrt{(X^t X)_{jj}^{-1}}$.

A partir daqui, e da forma usual, pode-se deduzir que os extremos de um **intervalo com coeficiente de confiança $1 - \alpha$ para β_j** são

$$\hat{\beta}_j \pm se(\hat{\beta}_j) t_{1-\frac{\alpha}{2}}(n - (p + 1)).$$

Cálculos adicionais mostram que **uma região ^a com coeficiente de confiança $1 - \alpha$ para o vetor de ^{parâmetros} coeficientes $\beta = (\beta_0, \dots, \beta_p)$** (todos os coeficientes em simultâneo!) é dada por:

$$\{\beta \in \mathbb{R}^{p+1} \mid (\hat{\beta} - \beta)^t X^t X (\hat{\beta} - \beta) \leq \bar{\sigma}^2 \chi^2_{1-\alpha}(p+1)\}.$$

Instruções para a determinação destas regiões de confiança para o vetor β não estão disponíveis no **SPSS**. Existem, por exemplo, no **R**.

^ajá não se pode falar em intervalo

1.6.2. IC para a predição

Começamos por observar que um modelo que tenha sido ajustado a partir de uns certos dados não deve ser utilizado para fazer previsões correspondentes a valores das variáveis explicativas que sejam muito diferentes dos usados inicialmente.

- **Predição de valores futuros**

Esta é a situação mais comum. Pretende-se obter um intervalo de predição ^a para a resposta dados novos valores para as variáveis explicativas.

Seja $x_0 = (x_{0,1}, \dots, x_{0,p})^t$ o vector de novas observações. O correspondente valor previsto é

$$\begin{aligned}\hat{y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 x_{0,1} + \dots + \hat{\beta}_p x_{0,p} \\ &= x_0^t \hat{\beta}.\end{aligned}$$

Considerando a variância do estimador $\hat{\beta}$ tem-se

$$\text{Var}(\hat{y}_0) = x_0^t (X^t X)^{-1} x_0 \sigma^2$$

e portanto

$$\text{Var}(\hat{y}_0 + u) = (1 + x_0^t (X^t X)^{-1} x_0) \sigma^2.$$

^aUsa-se a notação "intervalo de confiança" para um parâmetro; para os valores de uma variável, usa-se a notação "intervalo de predição".

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 25 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Um intervalo a $(1 - \alpha)\%$ de predição para y_0 é

$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-p-1) \bar{\sigma} \sqrt{1 + x_0^t (X^t X)^{-1} x_0}.$$

Se a previsão for feita para um conjunto de valores de variáveis explicativas em número inferior ao considerado inicialmente (x_0 é um vetor de comprimento inferior a p), a matriz X a considerar deve ser a correspondente a esse número de variáveis explicativas.

- **Intervalo de confiança para a resposta média**

Por vezes, podemos estar interessados em estimar o valor médio da resposta dados valores específicos para as variáveis explicativas. Tecnicamente, a variância associada a esta média é inferior à variância associada à estimativa de uma resposta futura pois a variância dos resíduos já não terá de ser considerada.

De forma análoga à anterior, um **intervalo de confiança a $(1 - \alpha)\%$ para o valor médio de Y dado x_0** é

$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-p-1) \bar{\sigma} \sqrt{x_0^t (X^t X)^{-1} x_0}.$$



Instruções em R:

- intervalo de confiança para os coeficientes:

```
> confint(model, parm, level=0.95,...)
```

onde model é o objecto que representa o modelo estimado, parm é o conjunto dos parâmetros para os quais se pretendem os intervalos de confiança (por defeito, os intervalos de confiança são calculados para todos os coeficientes do modelo), e level é o nível de significância que se pretende.

- intervalos de predição:

```
> pr.c <- predict(model, interval="confidence", ...)
```



O objecto pr.c criado é uma matriz que, para cada indivíduo (linha), indica o valor esperado da resposta e os limites inferior e superior do intervalo de confiança para essa resposta média.

- intervalos de predição individual:

```
> pr.p <- predict(model, interval="prediction", ...)
```

[Go Back](#)

O objecto pr.p criado é uma matriz que, para cada indivíduo (linha), indica o valor esperado da resposta e os limites inferior e superior do intervalo de predição para a resposta com essa variáveis explicativas.

[Full Screen](#)

[Close](#)

[Quit](#)

[Home Page](#)

[Title Page](#)

[Contents](#)

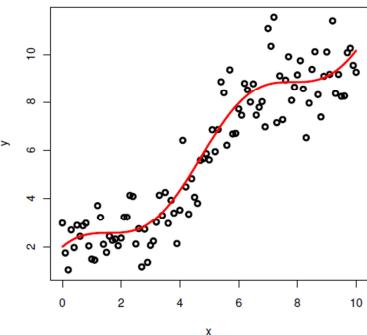
Page 26 of 104

1.6.3. Exemplo

Simulamos dados do modelo

$$y_i = 1 + x_i + \cos(x_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2 Id)$$

com $x_i = i/10$, $i = 1, \dots, 100$.



Na figura acima, a curva a vermelho representa a função a partir da qual os dados foram gerados.

Na figura ao lado:

- os valores ajustados aparecem a rosa
 - os intervalos de confiança para a média da resposta estão indicados a azul
 - os intervalos de predição para valores individuais estão indicados a verde.

As instruções em **R** para a determinação do intervalo de confiança para a resposta média e do intervalo de predição para observações individuais são apresentadas a seguir, respectivamente:



[Home Page](#)

Title Page

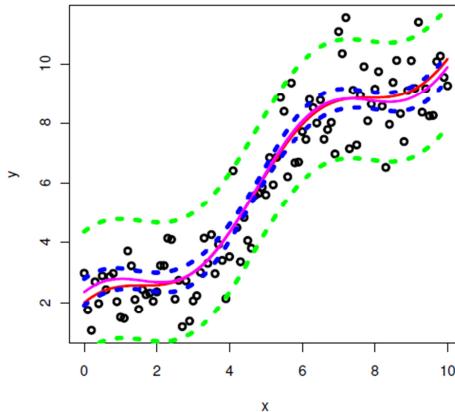
Contents



Page 27 of 104

[Go Back](#)

Full Screen



A. Rita Gaio, DM-FCUP

- (1) R^2 não tem unidades
- (2) $0 \leq R^2 \leq 1$
- (3) o ajustamento do modelo é tanto melhor quanto mais próximo R^2 estiver de 1
- (4) R^2 é o quadrado do coeficiente de correlação linear de Pearson amostral entre os valores observados y e os valores ajustados \hat{y} , i.e.,

$$\begin{aligned} R^2 &= r_{Y,\hat{Y}}^2 = \frac{(SS_{Y,\hat{Y}})^2}{SS_{YY} SS_{\hat{Y}\hat{Y}}} \\ &= \frac{\left[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) \right]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}. \end{aligned}$$

- (5) no caso da regressão linear simples,

$$R^2 = r_{Y,X}^2.$$

Em particular, $R^2 = 0$ corresponde à situação em que a curva de regressão é uma recta paralela ao eixo dos xx (nenhuma informação sobre Y se obtém a partir de X).

- (6) no caso da regressão linear múltipla em que as variáveis explicativas são não correlacionadas duas a duas,

$$R^2 = r_{Y,X_1}^2 + \cdots + r_{Y,X_p}^2.$$

- (7) a inexistência de uma relação linear entre Y e $X = (X_1, \dots, X_p)$ conduz a valores de R^2 próximos de 0.

[Home Page](#)[Title Page](#)[Contents](#)[◀](#) [▶](#)[◀](#) [▶](#)[Page 28 of 104](#)[Go Back](#)[Full Screen](#)[Close](#)[Quit](#)

1.7. Coeficiente de determinação

1.7.1. Definição

Considere-se o modelo

$$y = X\beta + u, \quad u \sim N(0, \sigma^2 \text{Id})$$

e sejam

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{e} \quad TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

a soma dos quadrados dos resíduos e a soma dos quadrados total, respectivamente.

O coeficiente de determinação ^a (múltiplo) R^2 é definido por

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\hat{\sigma}_Y^2}{\hat{\sigma}_Y^2}$$

onde

$$\hat{\sigma}_Y^2 = \frac{RSS}{n} \quad \text{e} \quad \hat{\sigma}_Y^2 = \frac{TSS}{n}$$

são os estimadores de máxima verosimilhança da variância residual e da variância de Y , respectivamente.

O coeficiente R^2 representa a percentagem da variância de Y explicada pelo modelo de regressão. Não deve ser usado como uma medida da qualidade do ajustamento do modelo.

^aou o quadrado do coeficiente de correlação múltipla, uma vez que $R^2 = r_{Y,\hat{Y}}^2$ e $r_{Y,\hat{Y}}$ é designado o coeficiente de correlação múltipla

O coeficiente de determinação R^2 apresenta o seguinte **inconveniente**: se acrescentarmos ao modelo mais uma variável explicativa, X_{p+1} , o novo R^2 é superior ou igual ao anterior, uma vez que a soma dos quadrados dos resíduos não pode aumentar e a soma total dos quadrados dos desvios de Y relativamente à sua média se mantém constante.

O **coeficiente de determinação ajustado aos graus de liberdade**, \bar{R}^2 , é calculado tendo em conta o número de graus de liberdade envolvidos no cálculo da variância amostral de Y e de \hat{Y} :

$$\bar{R}^2 = 1 - \frac{RSS/(n - (p + 1))}{TSS/(n - 1)} = 1 - \frac{\bar{\sigma}^2}{\bar{\sigma}_Y^2}$$

onde

$$\bar{\sigma}^2 = \frac{RSS}{n - (p + 1)} \quad \text{e} \quad \bar{\sigma}_Y^2 = \frac{TSS}{n - 1}$$

são estimadores não enviesados da variância residual e da variância de Y , respectivamente.

O inconveniente apontado para R^2 já não se verifica para \bar{R}^2 . Com efeito, ao adicionar-se a variável explicativa X_{p+1} , \bar{R}^2 cresce apenas se a inclusão de X_{p+1} melhora o modelo mais do que seria esperado devido ao acaso; *i.e.*, se essa inclusão induz uma diminuição da soma dos quadrados dos resíduos suficiente para compensar o decréscimo de uma unidade no denominador de $\bar{\sigma}^2$.

Propriedades:

- (1) \bar{R}^2 tem o inconveniente de poder ser negativo
- (2) $\bar{R}^2 \leq R^2 \leq 1$
- (3) Quanto mais próximo de 1 estiver \bar{R}^2 mais ajustado é o modelo aos dados.

(4) $\bar{R}^2 = \frac{\bar{\sigma}_Y^2 - \bar{\sigma}^2}{\bar{\sigma}_Y^2}$; Em particular:

- \bar{R}^2 aumenta quando $\bar{\sigma}^2$ diminui.
- \bar{R}^2 pode ser interpretado como a percentagem da variância estimada de Y que é explicada pelo modelo linear.

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{n\bar{\sigma}^2/(n - (p + 1))}{n\bar{\sigma}_Y^2/(n - 1)} \\ &= 1 - \frac{n - 1}{n - (p + 1)}(1 - R^2). \end{aligned}$$

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀](#) [▶](#)
[◀](#) [▶](#)
[Page 29 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

Observações:

- a indicação do valor de \bar{R}^2 em estudos de regressão é tão mais indicada quanto maior for o número de variáveis explicativas usadas no modelo
- \bar{R}^2 pode ser usado como uma medida da qualidade do ajustamento do modelo aos dados
- \bar{R}^2 pode ser usado como uma medida de comparação entre modelos envolvendo um número diferente de variáveis explicativas (usando R^2 , o modelo com um maior número de variáveis explicativas ficaria beneficiado à partida)
- \bar{R}^2 pode ser usado como uma medida da qualidade de predição do modelo em relação a observações futuras.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

[Page 30 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.7.2. Significância de R^2

Testar a significância de R^2

$$H_0 : R^2 = 0 \quad (1.2)$$

é equivalente a testar a significância de R

$$H_0 : R = 0$$

e equivalente a testar a significância dos coeficientes de regressão (β_1, \dots, β_p)

$$H_0 : \beta_1 = \dots = \beta_p = 0. \quad (1.3)$$

Qualquer uma destas hipóteses nulas corresponde à situação de **inexistência de uma relação linear entre a resposta e as variáveis explicativas**.

Supondo que os resíduos são independentes e normalmente identicamente distribuídos, sabemos já que a estatística de teste correspondente a (1.3), e portanto também a (1.2), é

$$F^0 = \frac{RegSS/p}{RSS/(n - (p + 1))} \sim F(p, n - (p + 1))$$

ou ainda, usando a igualdade $R^2 = RegSS/TSS$,

$$F^0 = \frac{R^2/p}{(1 - R^2)/(n - (p + 1))} \sim F(p, n - (p + 1)).$$

Nota:

Tal como definido aqui, R^2 não faz sentido em **modelos de regressão que não incluem o parâmetro independente β_0** , para os quais é usual observarem-se valores de R^2 muito elevados.

Modelos desse tipo só devem aliás ser considerados em situações muito especiais; em particular, há que ter a certeza de que valores nulos das variáveis explicativas conduzem a valores nulos da resposta.

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 31 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

1.8. Multicolinearidade

O fenómeno de **multicolinearidade**, corresponde à existência de associações lineares entre as variáveis explicativas.

Multicolinearidade elevada indica redundância na informação fornecida pelas variáveis explicativas. Em termos de análise estatística, isso reflecte-se numa matriz X com colunas quase linearmente dependentes, e portanto na causa de alguns problemas:

- problemas numéricos no cálculo da matriz $(X^t X)^{-1}$ com consequências negativas sobre a qualidade do ajustamento do modelo e da estimação de coeficientes
- variâncias grandes para algumas das estimativas $\hat{\beta}_j$ dos coeficientes de regressão e portanto muita instabilidade na inferência.

Deve portanto suspeitar-se de multicolinearidade elevada quando

- alguns dos coeficientes de regressão apresentam sinais contrários ou valores demasiado grandes ou demasiado significativos em relação ao que seria de esperar
- ocorrem mudanças substanciais nalguns dos coeficientes de regressão quando se excluem ou introduzem novas variáveis explicativas

- ocorrem mudanças substanciais nalguns dos coeficientes de regressão quando se excluem ou alteram observações
- os coeficientes das variáveis que se esperam ser importantes para o modelo apresentam desvios padrão grandes
- a estatística F da regressão é significativa sem que nenhum dos parâmetros de regressão o seja
- alguns dos parâmetros de regressão sejam significativos sem que a estatística F da regressão o seja
- a matriz de correlação entre as variáveis explicativas apresenta elevados coeficientes de correlação. Aqui, será conveniente excluir uma ou várias dessas variáveis muito correlacionadas, deixando evidentemente aquelas que façam mais sentido no contexto do problema.

É usual considerar-se multicolinearidade elevada a situação em que a matriz de correlação apresenta valores superiores a 0.75

Contudo, pode existir uma relação linear substancial entre várias variáveis explicativas do modelo sem que isso seja reflectido nos coeficientes de correlação, que são obtidos para o conjunto das variáveis, duas a duas.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 32 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

A questão da multicolinearidade pode ser também averiguada através da regressão de cada uma das variáveis explicativas sobre as restantes.

Seja R_i^2 o coeficiente de determinação associado à regressão de X_i sobre as restantes variáveis explicativas. O **fator de inflação da variância para X_i** , **VIF(X_i) - Variance Inflation Factor**, é

$$\text{VIF}(X_i) = \frac{1}{1 - R_i^2}.$$

Observe-se que cada variável explicativa tem um valor de VIF associado.

Relações lineares fortes entre X_i e X_j ($j \neq i$) reflectem-se num valor de R_i^2 próximo de 1 e portanto num valor alto de VIF.

Regra empírica:

A observação de pelo menos um valor de VIF superior a 10 sugere a existência de multicolinearidade elevada.

Pode-se mostrar que (exercício) o VIF é uma medida do aumento da variância do coeficiente de regressão estimado para a variável em causa quando as restantes variáveis explicativas são correlacionadas. A amplitude dos intervalos de confiança do coeficiente aumenta de \sqrt{VIF} , por comparação com a situação em que as variáveis explicativas não são correlacionadas.

Uma outra medida de colinearidade consiste do **índice de condição**, **CI - Condition Index**. É definido à custa da matriz de correlação Cor através da fórmula

$$CI = \sqrt{\frac{\text{maior valor próprio de } C}{\text{menor valor próprio de } C}}.$$

Por definição, tem-se sempre $CI \geq 1$.

Regra empírica:

Um índice de condição superior a 30 com proporções de variâncias grandes em pelo menos duas das variáveis é indicador de problemas de multicolinearidade elevada.

De qualquer forma é difícil confiar neste valor de corte de 30 porque o CI depende da escala em que as variáveis estão a ser consideradas...

Instruções em R:

Para as correlações entre as várias variáveis as instruções

```
> cor(x, method = c("pearson", "kendall", "spearman"))
> pairs(x)
```

efectuam, respectivamente, o cálculo das correlações segundo o método especificado e os gráficos de dispersão das variáveis na matriz ou data frame x, duas a duas.

Os factores de inflação da variância podem ser obtidos da instrução

```
> library(HH)
> vif(x)
```

onde x é o objecto que representa o modelo lm em causa.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 33 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.9. Métodos de selecção de variáveis

Mesmo apresentando um bom ajustamento aos dados, um modelo de regressão com **muitas variáveis explicativas** traz quase sempre associados **problemas de interpretação**. O que se gostaria de ter era um modelo mais simples, com menos variáveis, que evidenciasse os efeitos mais fortes.

A este propósito, existem vários métodos de selecção de variáveis, nem todos conduzindo aos mesmos resultados. No que se segue, apresentamos alguns desses procedimentos.

1.9.1. *Forward*

Este método, também designado por *forward selection procedure*, parte do modelo que consta apenas do coeficiente independente - **modelo nulo** - e vai adicionando sequencialmente a variável explicativa que mais melhora o ajustamento do modelo aos dados.

No R, através dos comandos que se seguem, a melhoria na qualidade do ajustamento é avaliada através dos critérios de informação AIC ou BIC. Suponhamos que num determinado passo o modelo contém k variáveis explicativas e que, no passo seguinte, tem $k + 1$ variáveis explicativas. Concretamente, a variável X_{k+1} a adicionar é aquela que conduz ao menor valor do critério de informação inicialmente escolhido. O algoritmo pára a partir do momento em que essa condição não é satisfeita.

1.9.2. *Backward*

Método inverso ao anterior, também designado por *backward elimination procedure*. Parte do modelo que consta de todas as variáveis explicativas - **modelo completo** - e depois retira sequencialmente variáveis explicativas, uma a uma, de acordo com o ajustamento do novo modelo encontrado. No R, com as instruções que se apresentam a seguir, o critério de remoção continua a usar um critério de informação: em cada passo, o algoritmo vai eliminando a variável explicativa que conduz a um maior valor do critério, parando quando a eliminação de qualquer variável explicativa não diminui o critério de informação anterior.

1.9.3. *Stepwise*

É o procedimento mais sofisticado e mais utilizado. Envolve **inclusão e exclusão de variáveis** e pode partir do modelo nulo ou do modelo completo. Em cada passo, avalia-se a adição de uma nova variável ao modelo. Se essa variável contribuir para um modelo melhor (segundo um critério definido a priori), a variável é retida e todas as variáveis explicativas do novo modelo são testadas para avaliar esse novo modelo. Aquelas que já não contribuírem de forma significativa são excluídas. Idealmente, o método identifica o menor conjunto de variáveis explicativas a considerar na regressão.

[Home Page](#)

[Title Page](#)

[Contents](#)

« »

« »

Page 34 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Não existe unanimidade entre a comunidade científica quanto ao método de selecção de variáveis a usar nem tão pouco a opinião sobre a sua utilidade é convergente.

- Vários investigadores aconselham a que o procedimento stepwise seja usado numa fase exploratória da investigação mas não necessariamente na definição e interpretação teórica para o modelo final. Nessa última fase, a opção de incluir ou excluir variáveis deve ser feita pelo investigador, com base em pressupostos teóricos do estudo, e não partindo apenas de um algoritmo (de entre vários possíveis). Kleimbaum (1994) refere mesmo que as variáveis explicativas a usar devem ser especificadas a priori.
 - Um problema adicional com este procedimentos de selecção tem a ver com o nível de significância. Como $\alpha = 0.05$ em cada passo da iteração e o método consiste de vários passos (não independentes), o nível de significância final pode sair muito inflacionado.
 - Os métodos stepwise são os mais sensíveis ao fenómeno de multicolinearidade elevada.
 - Em dados não colineares, é esperado que os métodos forward, backward e stepwise forneçam a mesma selecção de variáveis.
- Há autores que preferem o método Backward ao Forward. Isto porque:
 - o modelo completo é ajustado no início, ficando depois disponível para comparações
 - lida com o problema da multicolinearidade de uma forma melhor
 - Os processos de selecção de modelos são baseados em critérios arbitrários. Podem-se escolher, por exemplo, critérios baseados no resultado de testes de Wald para adicionar ou retirar variáveis; ou em estatísticas F de comparação de modelos; ... Não existe portanto garantia de que usando um processo de selecção para a escolha do modelo resulte num bom modelo (em termos de boas previsões), ou num modelo sensato (em termos do que se conhece teoricamente sobre o processo em causa).
 - Os processos de selecção de modelos lidam mal com a presença de outliers e/ou pontos influentes, na medida em que o modelo final pode não ser mesmo o mais deseável.
 - Para qualquer um dos métodos, o modelo final terá de ser devidamente inspeccionado e diagnosticado.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

[Page 35 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

A instrução em **R** para a escolha do método de selecção de variáveis é (library(MASS))

```
step(object, scope, direction = c("both",
  "backward", "forward"), k, trace, ...)
```

Aqui:

- a escolha do modelo final é feita com base no critério de informação AIC, que é uma medida da qualidade do ajustamento do modelo que penaliza modelos com muitos parâmetros. **Quanto menor for o AIC, mais preferível será o modelo.**
- object: objecto representando o modelo inicial do processo de selecção (pode ser, por exemplo, o modelo completo com algumas potências das variáveis explicativas e/ou algumas interacções)
- scope: define o tipo de modelos a serem examinados no processo de selecção; por exemplo:
 - scope=list(upper=mod5): indica que o maior modelo que o utilizador está disposto a aceitar é mod5
 - scope=list(lower=mod0): indica que o menor modelo que o utilizador está disposto a aceitar é mod0
- direction: escolher de entre "forward", "backward" ou "both" (step-wise); se scope não estiver definido, a direcção escolhida por defeito é "backward"; se scope estiver definido, a direcção escolhida por defeito é "both".

- k: múltiplo do número de graus de liberdade usados para penalização; $k = 2$ corresponde ao AIC; $k = \log(n)$ corresponde ao BIC.

- trace: indica se os resultados dos modelos intermédios são ou não para mostrar (respectivamente, trace igual a TRUE ou FALSE).

Há um pacote no **R**, *leaps*, que contém algoritmos que devolvem o melhor modelo para cada um dos tamanhos possíveis (para as várias combinações das variáveis explicativas). Está integrado nos processos de selecção que são designados por *best subset regression*. O critério de selecção do melhor modelo é baseado no valor do coeficiente de determinação ajustado.

Para modelos com um número relativamente pequeno de variáveis explicativas, este processo é razoável. À medida que o número de variáveis explicativas vai aumentando, o número de subconjuntos de variáveis explicativas cresce exponencialmente e torna-se bastante difícil avaliar todos os modelos que utilizam esses subconjuntos.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 36 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.10. Comparação de modelos

A teoria aqui apresentada aplica-se a **modelos genéricos de regressão, lineares ou não lineares**.

1.10.1. Modelos encaixados

Dois modelos genéricos M_1 e M_2 dizem-se **encaixados** ou **aninhados** se os parâmetros de M_1 constituirem um subconjunto dos parâmetros de M_2 . Designando por $p_1 + 1$ o número de parâmetros de M_1 e por $p_2 + 1$ o número de parâmetros de M_2 , tem-se $p_1 < p_2$. Para uma situação de comparação de modelos, é aconselhável usar o teste F de análise da variância ou o teste da razão de verosimilhanças entre M_1 e M_2 .

Repare-se que, para cada modelo, estamos a designar o número de variáveis explicativas por p , daí que o número total de parâmetros do modelo seja $p + 1$, uma vez que temos também de considerar o termo constante.

Sejam M_1 e M_2 dois modelos encaixados. Os parâmetros do modelo M_2 podem escrever-se na forma

$$\eta = (\eta', \eta''), \quad \text{com } \eta' \in \mathbb{R}^{p_1}, \eta'' \in \mathbb{R}^{p_2-p_1}$$

em que η' são os parâmetros do modelo M_1 , que na maior parte das vezes se obtêm dos parâmetros do modelo M_2 fazendo $\eta'' = 0$.

(Esta é a situação mais usada. Mais geralmente, o critério pode ser formulado para a situação em que os parâmetros de M_1 se obtêm de η fazendo $\eta'' = \eta_0''$ para algum valor conhecido η_0'' .)

- Seja RSS_j a soma dos quadrados dos resíduos do modelo M_j , $j = 1, 2$. Então, assintoticamente,

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/(p_2 - p_1)}{\text{RSS}_2/(n - (p_2 + 1))} \sim F(p_2 - p_1, n - (p_2 + 1))$$

onde n representa o número total de observações.

Este resultado permite efectuar um teste de hipóteses de comparação de modelos, **ANOVA**, com hipótese nula dada por

$$H_0: \text{a qualidade do ajustamento de } M_1 \text{ é igual à de } M_2. \quad (\eta'' = 0)$$

A **não rejeição de H_0** significa que não se pode rejeitar a hipótese de o modelo reduzido explicar a variância da resposta de forma tão adequada quanto o modelo completo.

A **rejeição de H_0** significa que o modelo reduzido não é preferível ao modelo completo, na medida em que o último explica uma maior percentagem de variância da resposta.

Uma nota importante: o teste só pode ser efectuado se ambos os modelos tiverem sido estimados com os mesmos dados.

Aplicação mais comum: o modelo M_1 é obtido de M_2 por exclusão de algumas variáveis explicativas.

- Representando por LL_j o logaritmo da verosimilhança L_j do modelo M_j , $j = 1, 2$, tem-se, assimptoticamente,

$$2 \log \left(\frac{L_2}{L_1} \right) = 2(LL_2 - LL_1) \sim \chi^2(p_2 - p_1).$$

No caso de um modelo não linear, é preferível usar o critério da razão de verosimilhanças, eventualmente acompanhado de um dos critérios de informação descritos na secção a seguir.

Instruções em R:

A instrução para o teste F de análise da variância para a comparação de modelos encaixados é

```
> anova(modelo1, modelo2)
```

onde modelo1 e modelo2 são os modelos a comparar. A ordem pela qual os modelos são apresentados é irrelevante.

A instrução para o teste da razão de verosimilhanças para a comparação de modelos encaixados obriga ao carregamento da biblioteca lmtest e é

```
> library(lmtest)
> lrtest(modelo1, modelo2)
```

onde modelo1 e modelo2 são os modelos a comparar. A ordem pela qual os modelos são apresentados é irrelevante.

1.10.2. Modelos não encaixados

Se os modelos M_1 e M_2 não forem encaixados, é aconselhável utilizar-se um critério que meça a **quantidade de informação** que o modelo recolhe dos dados, por oposição ao ruído que o modelo não consegue explicar. Uma medida da quantidade de informação é o logaritmo da verosimilhança do modelo. Como a log-verosimilhança cresce com o número de parâmetros, um tal critério tem de ser parcimonioso quanto ao número de parâmetros para evitar o ajustamento de modelos saturados. Desse modo, um tal critério incorpora um termo que penaliza um número grande de parâmetros e é também função do tamanho da amostra.

Dois critérios de informação usuais são:

- **critério de informação de Akaike**

$$AIC = -2 LL + 2p$$

- **critério de informação Bayesiana**

$$BIC = -2 LL + p \log(n)$$

onde LL é o logaritmo da verosimilhança do modelo, p é o número total de parâmetros, e n é o número total de observações.

Para qualquer destes dois critérios, quanto menor for o seu valor mais preferível é o modelo.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 38 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Observe-se que o critério BIC penaliza mais o número de parâmetros do que o critério AIC porque o número de parâmetros é multiplicado por $\log(n)$ que é maior do que 1 quando $n > 3$. O facto da penalização também crescer com o número de observações serve para penalizar os modelos ajustados a muitas observações e que sejam "pouco verosímeis", i.e., que tenham uma log-verosimilhança pequena. A ideia é que, moralmente, com muitas observações há a "obrigação" de se encontrar um modelo que extraia muita informação. Entre os dois critérios de informação, é portanto preferível usar-se o critério BIC.

Nota: a comparação entre modelos **lineares** pode ser feita usando o coeficiente de determinação ajustado. O senão dessa abordagem consiste no facto de não existir uma estatística de teste que avalie a significância estatística dessa comparação.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

Page 39 of 104

Instruções em R:

```
extractAIC(object, ..., k)
```

onde

- object é o modelo em causa
- k é a penalização a ser usada por parâmetro;
 $k = 2$ dá o AIC; $k = \log(n)$ dá o BIC, onde n é o tamanho amostral.

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.11. Gráfico seminormal

O gráfico **seminormal**^a é uma variação do gráfico normal de probabilidades ^b que é usado para dados positivos.

Os gráficos semi-normais são usados para **detectar observações extremas (outliers)**, em vez de avaliar a adesão à normalidade que é o que fazem os gráficos de normalidade. Mais precisamente, um gráfico semi-normal representa as estatísticas de ordem da variável em causa contra $\Phi^{-1} \left(\frac{n+i}{2n+1} \right)$, isto é, os quantis $\left(\frac{n+i}{2n+1} \right)_{i=1,\dots,n}$ da distribuição normal reduzida $N(0,1)$ - usou-se a notação Φ para representar a função de distribuição de probabilidade de $N(0,1)$. Repare-se que, para qualquer $i = 1, \dots, n$,

$$\begin{aligned} \frac{n+i}{2n+1} &\geq \frac{n+1}{2n+1} \\ &\geq \frac{n+1}{2n+2} \\ &= \frac{1}{2} \end{aligned}$$

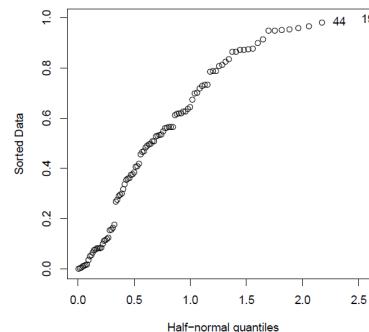
e

$$\begin{aligned} \frac{n+i}{2n+1} &\leq \frac{2n}{2n+1} \\ &\leq 1 \end{aligned}$$

Daqui se verifica que o gráfico seminormal só devolve quantis maiores do que o quantil 0.5 logo apenas números positivos.

Em R:

```
> library(faraway)
> halfnorm(...vector...)
```



Por defeito, o gráfico identifica as duas observações correspondentes aos maiores quantis. Essa opção pode ser alterada.

^ahalf-normal plot

^bprobability normal plot, PP-plot

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 40 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.12. Diagnósticos

1.12.1. Outliers

Um outlier é um indivíduo com observações que não se ajustam ao modelo encontrado, apresentando portanto resíduos grandes quando comparados com os resíduos das restantes observações.

Podem corresponder a erros cometidos aquando da recolha ou inclusão dos dados, ou a observações genuínas que legitimamente reflectem a relação existente entre a resposta e as variáveis explicativas.

De forma a garantir que as conclusões do modelo não dependem fortemente da presença de observações extremas, esse pontos devem ser inicialmente identificados e depois detalhadamente examinados. Mesmo serem legítimos, é uma boa prática considerar a sua exclusão do modelo, a reestimação dos parâmetros do modelo de regressão em causa, e a comparação com as estimativas iniciais.

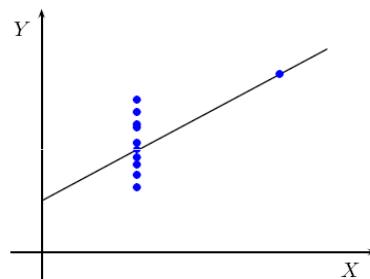
Uma regra empírica sugere que as observações com resíduos estandardizados superiores a 3.3 (correspondente a um nível de significância de 0.001 para a distribuição normal) sejam analisadas pormenorizadamente, e eventualmente excluídas do estudo. Nessa análise, é fundamental ter em consideração quer a magnitude dos resíduos quer o padrão de dispersão gráfica.

Sempre que houver lugar a exclusão de indivíduos, isso deve ser claramente referido no respectivo relatório de estudo.

1.12.2. Leverages

No caso da regressão múltipla, há observações que influenciam demasiado a estimativa dos parâmetros do modelo mas que não são outliers nem são facilmente identificáveis através de análises gráficas dos dados.

Considere-se por exemplo a figura abaixo.



O ponto mais à direita não é um outlier (tem resíduo nulo) mas a sua posição afecta de forma decisiva a recta de regressão a considerar. Observe-se também que esse ponto é um outlier da variável explicativa X , na medida em que está muito distanciado de todos os restantes valores de X . O ponto em causa corresponde a um ponto com leverage alta.

De forma breve, a **leverage** de uma observação reflecte a possibilidade de essa observação:

- ser um outlier para as variáveis explicativas
- influenciar o seu próprio valor ajustado.

Contudo pontos com leverages altas não correspondem necessariamente a pontos com muita influência sobre a estimação dos parâmetros do modelo! É preciso investigar esses pontos.

No que se segue precisamos a teoria de forma matemática.

Considere-se o modelo de regressão linear múltipla

$$Y = X\beta + u, \quad u \sim N(0, \sigma^2 \text{Id}).$$

Vimos já que o estimador de máxima verosimilhança de β é

$$\hat{\beta} = (X^t X)^{-1} X^t y.$$

Em particular, o vector de **valores ajustados** é

$$\hat{y} = X \hat{\beta} = H y \tag{1.4}$$

onde

$$H = X(X^t X)^{-1} X^t \in \mathbb{R}^{n \times n}$$

é a **matriz-chapéu**, correspondente à projecção ortogonal de y no espaço gerado pelas colunas de X .

A equação (1.4) diz-nos que

$$\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

e portanto que h_{ii} captura a contribuição da observação y_i sobre o seu próprio valor ajustado \hat{y}_i . As entradas da diagonal de H

$$h_i := h_{ii}$$

são por isso designadas por **leverages**. Mais ainda, tem-se

$$\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$$

portanto pontos com leverage alta conduzem a resíduos com variância baixa; os seus valores ajustados tenderão então a estar próximos das observações originais.

Sendo H uma matriz de projecção ortogonal sobre um subespaço vectorial de dimensão $p+1$, tem-se

$$\sum_{i=1}^n h_i = p+1$$

e portanto

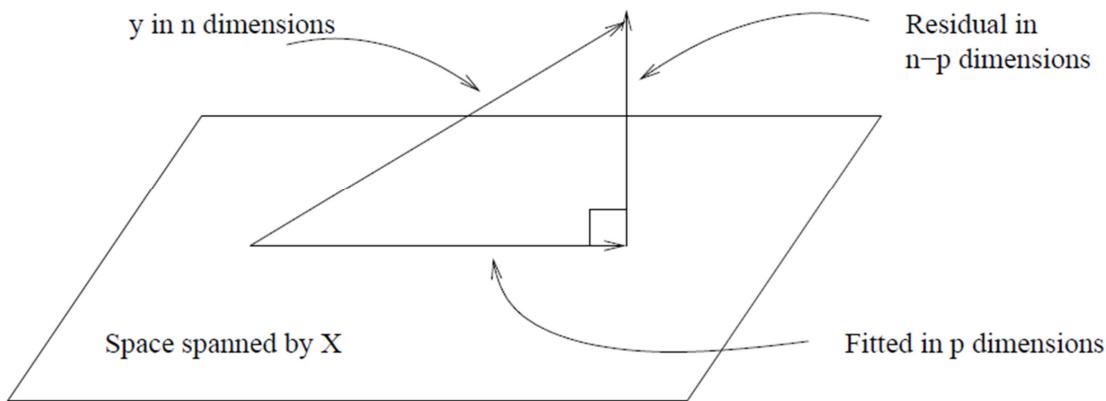
$$\frac{1}{n} \leq h_i \leq 1, \quad \text{média de } \{h_i\}_{i=1,\dots,n} = \frac{p+1}{n}.$$

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀](#) [▶](#)
[◀](#) [▶](#)
[Page 42 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

Regra empírica:

É comum dizer que uma observação tem **leverage alta** se a sua leverage é superior a $\frac{2(p+1)}{n}$. Observações com leverages altas devem ser identificadas e a sua influência sobre a estimativa dos coeficientes do modelo deve ser averiguada. Mais precisamente, o modelo deve ser ajustado com e sem esses pontos e os coeficientes obtidos para cada uma das situações devem ser comparados.

A partida, de entre essas observações, aquelas que não apresentarem resíduos grandes não serão preocupantes.


[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)

Page 43 of 104

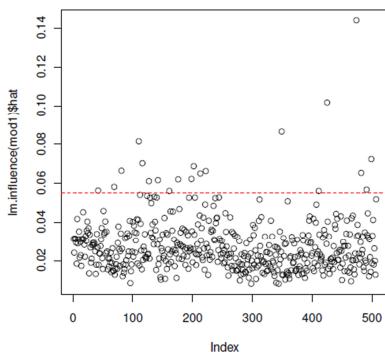
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

Uma forma gráfica de o conseguir no **R** consiste da análise do gráfico das leverages:

```
> plot(lm.influence(modelo)$hat)
```

ou

```
> plot(hatvalues(mod1))
```



Sabendo que a figura corresponde a uma regressão com duas variáveis explicativas numa amostra de 25 indivíduos, devem ser reconsideradas as três observações (indivíduos 2, 17 e 23) que apresentam leverages superiores a $2 \times 3/25 = 0.24$.

A tradução para português de *leverage point* é **observação alavancada**.

1.12.3. Resíduos Studentizados

O vector dos resíduos estimados é

$$\hat{u} = y - \hat{y} = (\text{Id} - H)y$$

que tem média 0 e matriz de covariância $\sigma^2(\text{Id} - H)$. Em particular, os resíduos estimados \hat{u} não são independentes e não têm variância constante, contrariamente aos erros u .

Os resíduos estimados estandardizados ^a são

$$\frac{\hat{u}_i}{\sigma\sqrt{1-h_i}}, \quad i = 1, \dots, n$$

que não têm grande aplicação prática pelo facto de desconhecermos σ .

Tudo isto conduz à formação dos **resíduos (internamente) studentizados**, ou ainda **resíduos estandardizados**,

$$u'_i = \frac{\hat{u}_i}{\sigma\sqrt{1-h_i}}, \quad i = 1, \dots, n$$

os quais, nas hipóteses de normalidade e independência dos u_i com que temos estado a trabalhar, têm média 0, igual variância (igual a 1) e correlações baixas entre eles.

Nos casos em que as leverages não são excessivamente elevadas, não existe grande diferença entre os gráficos dos resíduos brutos e dos resíduos studentizados, excluindo a escala em que se apresentam. Por vezes há uma preferência pelos resíduos studentizados em análises gráficas por estes terem igual variância.

^aou resíduos estimados padronizados ou resíduos de Pearson padronizados

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 44 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Observações:

- o estimador $\bar{\sigma}$ de σ apresentado acima é o definido na secção sobre estimação de parâmetros,

$$\begin{aligned}\bar{\sigma}^2 &= \frac{RSS(\hat{\beta})}{n - p - 1} \\ &= \frac{1}{n - p - 1} \sum_{j=1}^n \hat{u}_j^2.\end{aligned}$$

- o processo de "studentização" (interna) só corrige a variância não constante dos resíduos brutos na situação em que os erros u têm igual variância.
- Por vezes, observações individuais têm uma grande influência sobre a variância estimada e isto pode mascarar outliers.

Definem-se então os **resíduos externamente studentizados**:

$$\hat{u}_i'' = \frac{\hat{u}_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_i}}$$

com

$$\hat{\sigma}_{(i)}^2 = \frac{1}{n - p - 2} \sum_{j=1(j \neq i)}^n \hat{u}_j^2$$

i.e., o resíduo correspondente à observação i usa uma estimativa para o desvio padrão de u que exclui a i -ésima observação. Tem-se que

$$\hat{u}_i'' \sim t(n - p - 2).$$

No R:

Conforme definido atrás, os resíduos estandardizados não se conseguem calcular na situação geral em que se desconhece a variância σ^2 dos erros.

A linguagem e software de estatística **R** adopta então outras definições para resíduos estandardizados e studentizados. A saber:

- resíduos estandardizados (R)**

$$\frac{\hat{u}_i}{\bar{\sigma} \sqrt{1 - h_i}}$$

- resíduos studentizados (R)**

$$\frac{\hat{u}_i}{\bar{\sigma}_{(i)} \sqrt{1 - h_i}}$$

onde $\bar{\sigma}_{(i)}$ representa o desvio-padrão amostral de u quando a i -ésima observação é excluída.

Depois de ajustado um modelo linear através da instrução `lm`, digamos `model`, as instruções para os vários tipos de resíduos são as seguintes:

- residuals(model)**
`resid(model)`
`model$resid`

representam os resíduos brutos

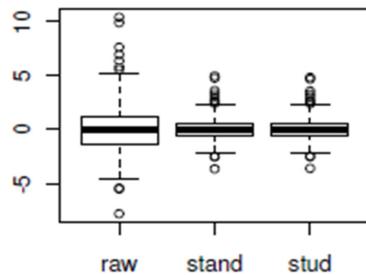
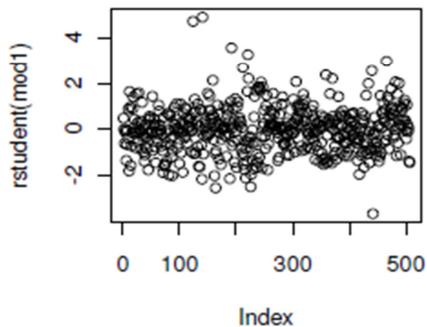
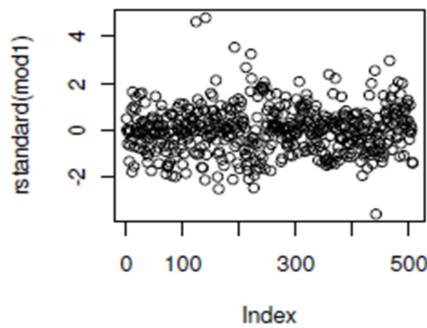
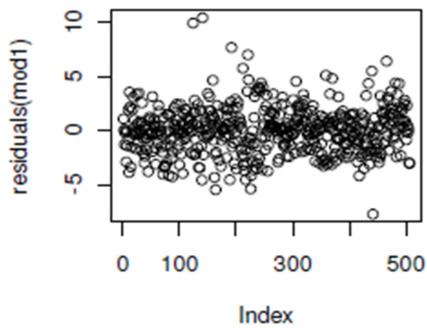
- rstandard(model)**

representa os resíduos estandardizados

- rstudent(model)**

representa os resíduos studentizados.

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀](#) [▶](#)
[◀](#) [▶](#)
[Page 45 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

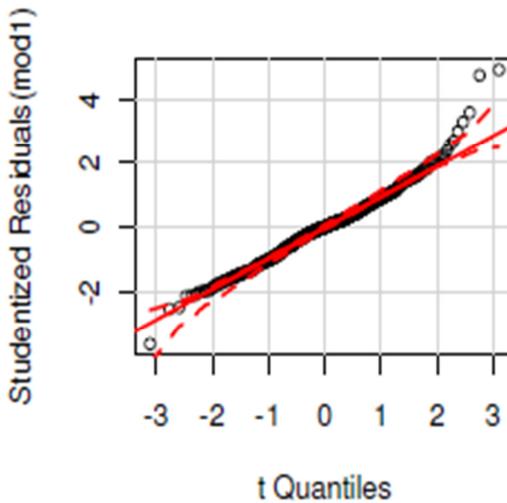
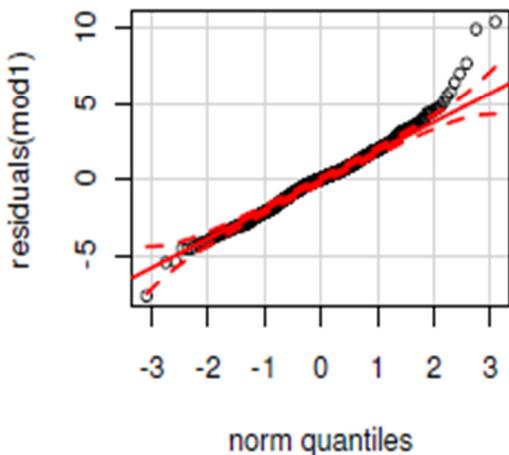
Page 46 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

Page 47 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



1.12.4. Observações influentes

Uma observação diz-se **influente** se tem uma grande influência sobre os parâmetros de regressão estimados, i.e., se o valor de pelo menos um desses parâmetros determinados a partir do conjunto de dados incluindo e excluindo a observação influente são muito diferentes.

Uma observação influente pode ou não ser um outlier e pode ou não ter uma leverage grande mas tenderá a ter pelo menos uma destas propriedades.

A estatística usada comumente para avaliar a influência da observação i é a **distância de Cook**

$$\begin{aligned} C_i^2 &= \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{(p+1)\bar{\sigma}^2} {}^a \\ &= \frac{1}{p+1} \cdot \hat{u}_i^2 \cdot \frac{h_i}{1-h_i} \end{aligned}$$

A última igualdade mostra que esta estatística contém simultaneamente informação sobre os resíduos (studentizados) e sobre as leverages.

Regra empírica:

os pontos com **distância de Cook superior a 1** podem ser pontos influentes pelo que o seu impacto no ajustamento do modelo deve ser averiguado.

Um forma prática de detectar a existência de observações influentes consiste da análise do gráfico das distâncias de Cook contra a indexação dos indivíduos.

^a $\hat{y}_{j(i)}$ representa o valor ajustado para y_j quando o modelo é estimado excluindo a observação i

- quando a magnitude dos valores é aproximadamente a mesma, nada precisa de ser feito
- quando existem pontos que se destacam dos restantes, os seus efeitos sobre o ajustamento do modelo devem ser averiguados (comparando os coeficientes estimados quando os pontos são considerados e quando são excluídos).

Uma medida de influência análoga a C_i^2 é

$$\text{DFITS}_i {}^a = \frac{\hat{u}_i}{\bar{\sigma}_{(i)}} \sqrt{\frac{h_i}{1-h_i}}.$$

Da relação aproximada

$$C_i^2 \approx \frac{1}{p+1} \text{DFITS}_i^2$$

resulta que quando

$$|\text{DFITS}_i| > 2\sqrt{\frac{p+1}{n-p-1}}.$$

a observação i pode ser um ponto influente pelo que o seu efeito deverá ser avaliado.

Dada a relação entre as duas medidas de influência, C_i^2 e DFITS, a detecção de pontos influentes pode ser feita apenas com base numa delas.

^aNo SPSS, aparece designada por **DfFIT**.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

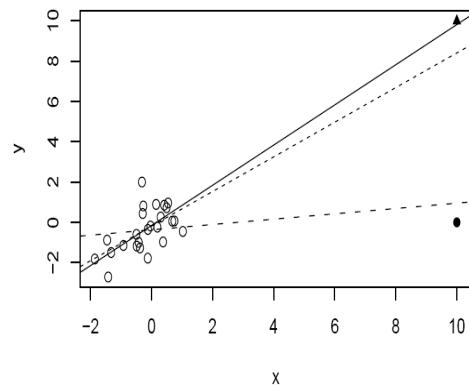
Page 48 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



Na figura, o conjunto de pontos na nuvem à esquerda é sempre considerado. Adicionalmente:

- a recta a **cheio** representa a equação de regressão **incluindo o ponto ▲ e excluindo o ponto ●**
- a recta a **ponteado** representa a equação de regressão **excluindo ambos os pontos ▲ e ●**
- a recta a **tracejado** representa a equação de regressão **incluindo o ponto ● e excluindo o ponto ▲**.

Tem-se que:

- os dois pontos adicionais ▲ e ● têm leverages altas porque estão longe dos restantes dados em relação à variável explicativa X
- o ponto ▲ não é um outlier para a recta a cheio

- o ponto ● não é um outlier para a recta a tracejado
- o ponto ● é um outlier para a recta a ponteado pois apresenta resíduos elevados para esse modelo.
- o ponto ▲ não é um ponto influente porque a diferença entre as rectas a cheio e a ponteado é pequena
- o ponto ● é um ponto influente porque a diferença entre as rectas a tracejado e ponteado é grande.

Conclusão:

- os outliers para o modelo de regressão devem ser sempre identificados e analisados
- os pontos com leverages altas devem ser identificados (não imediatamente excluídos!);
 - se apresentarem medidas de influência (C^2 ou DFITS) baixas, não causam problemas
 - se apresentarem medidas de influência (C^2 ou DFITS) altas, os seus efeitos sobre a qualidade do ajustamento do modelo devem ser averiguados.

A este respeito, o gráfico das leverages contra uma medida de influência (C^2 ou DFITS) poderá ser muito informativo. Há que identificar os pontos com valores altos em ambos os eixos coordenados.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

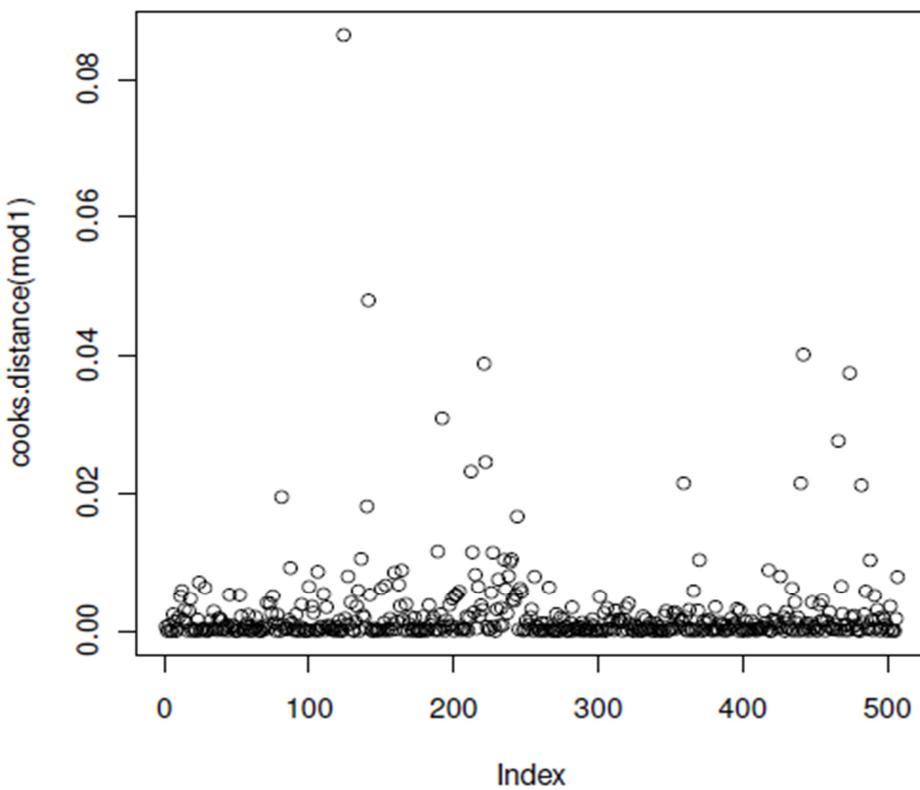
[Page 49 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)

[!\[\]\(98a0f62050c8ae5b6b5f206bfc69317a_img.jpg\)](#) [!\[\]\(fdf21e9ac78a82d793efc2dca9b630e5_img.jpg\)](#)

[!\[\]\(b59c51a1865446c8f7a5093cc693b46d_img.jpg\)](#) [!\[\]\(deb0d593c48c6dc8516dc6b875353bf3_img.jpg\)](#)

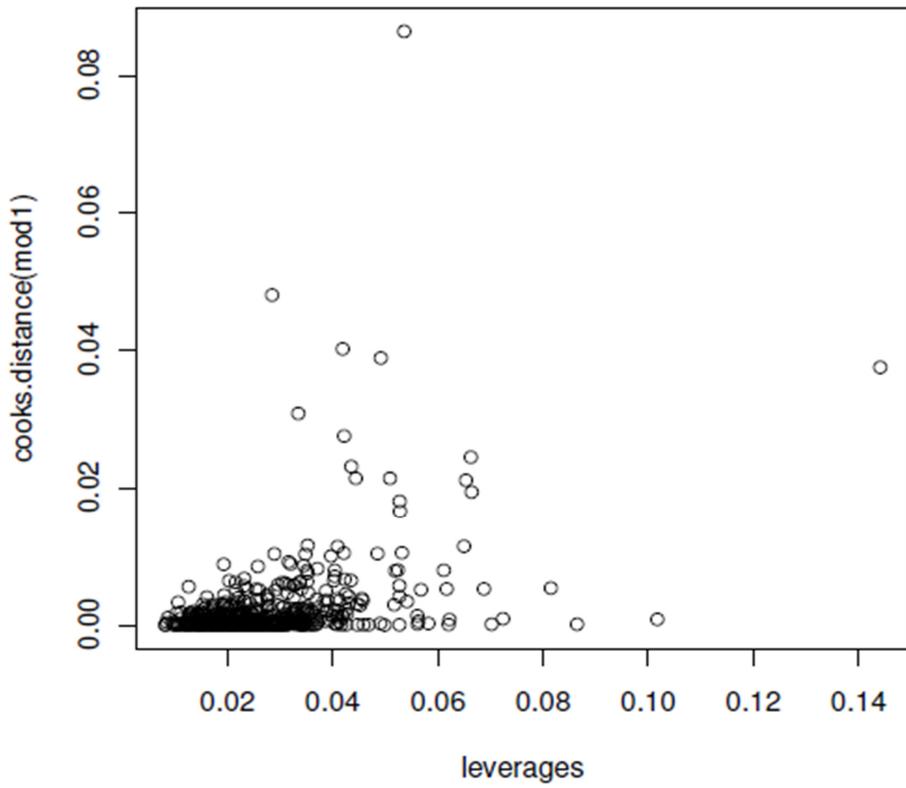
Page 50 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)

[!\[\]\(e83b2296dc4c9009963f494078a8a780_img.jpg\)](#) [!\[\]\(b927e50c4b8bd01e93641b827f471bb3_img.jpg\)](#)

[!\[\]\(1656df77d289379892e822a5b403488d_img.jpg\)](#) [!\[\]\(24944fa47cb6cfae18ca670c03a7dd6e_img.jpg\)](#)

Page 51 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.13. Análises gráficas

Nos cálculos anteriores assumimos que o modelo de regressão era

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + u, \quad u \sim N(0, \sigma^2 \text{Id})$$

com resíduos independentes e identicamente normalmente distribuídos, com média 0 e variância σ^2 .

Contudo, uma ou mais dessas hipóteses podem não se verificar. Mais precisamente:

- (a) os resíduos podem não ser normalmente distribuídos
- (b) os resíduos podem ser heterocedásticos, apresentando variâncias diferentes (para os vários valores das variáveis explicativas)
- (c) os resíduos podem ser correlacionados e portanto não independentes
- (d) a relação entre a resposta e as variáveis explicativas pode não ser linear
- (e) os resíduos podem possuir relações adicionais com as variáveis explicativas
- (f) os resíduos podem ser *grandes* demais, sugerindo um mau ajustamento do modelo aos dados
- (g) os problemas anteriores podem estar relacionados com a presença de outliers e/ou observações influentes.
- (h) multicolinearidade ...

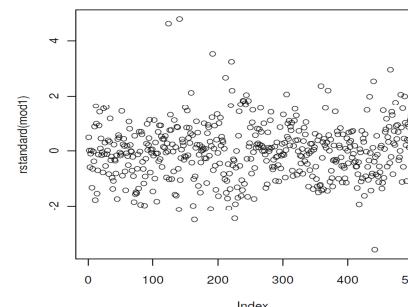
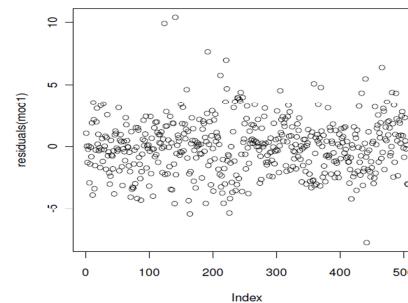
A consideração e interpretação de gráficos adequados pode ajudar à detecção das situações problemáticas referidas anteriormente.

Nesta secção sugerimos a análise de vários gráficos, apontando para as possíveis informações a retirar de cada um deles.

- gráfico dos resíduos estandardizados contra a ordem de recolha das observações

Devemos analisar com cuidado as observações que apresentem resíduos *muito grandes* (> 3.3).

Abaixo apresentamos os gráficos dos resíduos brutos e dos estandardizados; não há grandes diferenças, à parte a escala.



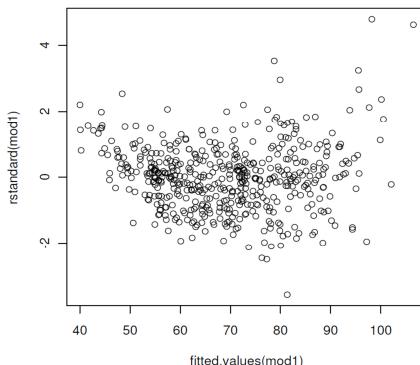
[Home Page](#)
[Title Page](#)
[Contents](#)

[Page 53 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

Resíduos **grandes** sugerem que o **ajustamento do modelo** aos dados não é bom. Essa situação pode ser indicadora de que as variáveis correctas não foram (ainda) incluídas no modelo ou de que seja necessário transformar (de forma não linear) algumas das variáveis já incluídas.

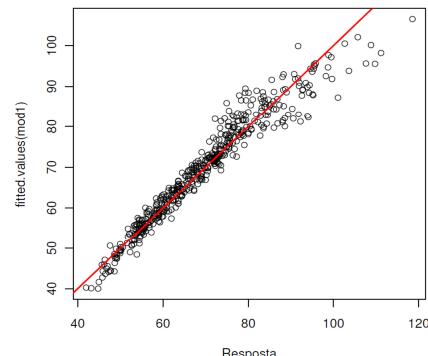
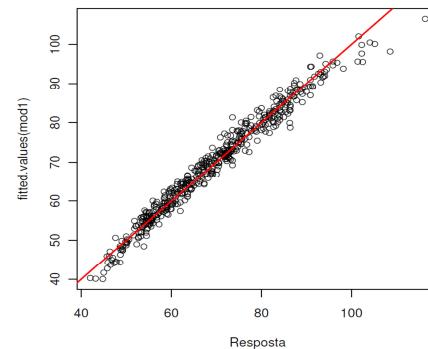
- gráfico dos resíduos estandardizados contra os valores ajustados

É talvez o mais importante de todos os gráficos com carácter de diagnóstico; deve-se observar variância constante na direcção dos resíduos, as observações devem formar uma nuvem rectangular, sem nenhuma tendência específica, e devem ser simétricas em torno do 0. A não verificação deste padrão sugere **heterocedasticidade** e/ou **não linearidade**, pelo que algumas transformações no modelo devem ser consideradas (por ex., logaritmo e raiz quadrada).

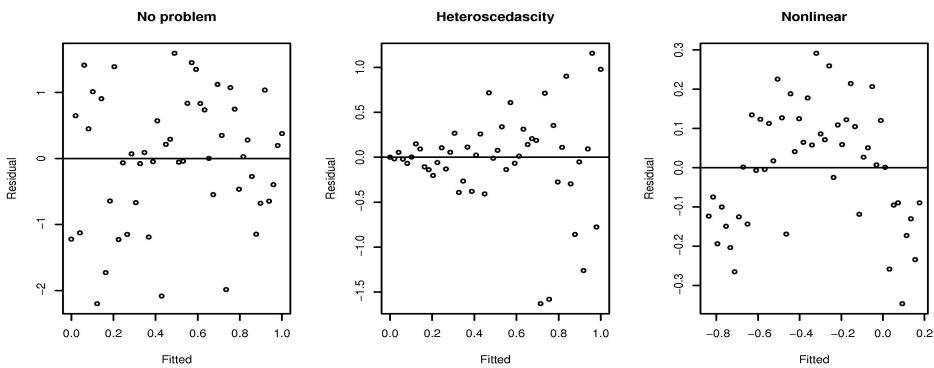


A figura não satisfaz os requisitos mencionados porque a dispersão dos pontos é em forma de U.

- gráfico dos valores ajustados contra a resposta
- Deverá ser possível identificar uma relação linear entre as duas variáveis. Para que a hipótese de **homocedasticidade** seja satisfeita, as observações devem distribuir-se pela (imaginada) recta de regressão de igual forma ao longo dos valores da resposta.



O gráfico de baixo ilustra uma situação heteroscedástica: a dispersão das observações relativamente à recta é diferente para valores baixos e altos da resposta.



(de Faraway, *Practical Regression and ANOVA using R*)

A figura apresenta três situações possíveis para o gráfico dos resíduos estandardizados contra os valores ajustados, com indicação do problema (ou não) associado.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

Page 54 of 104

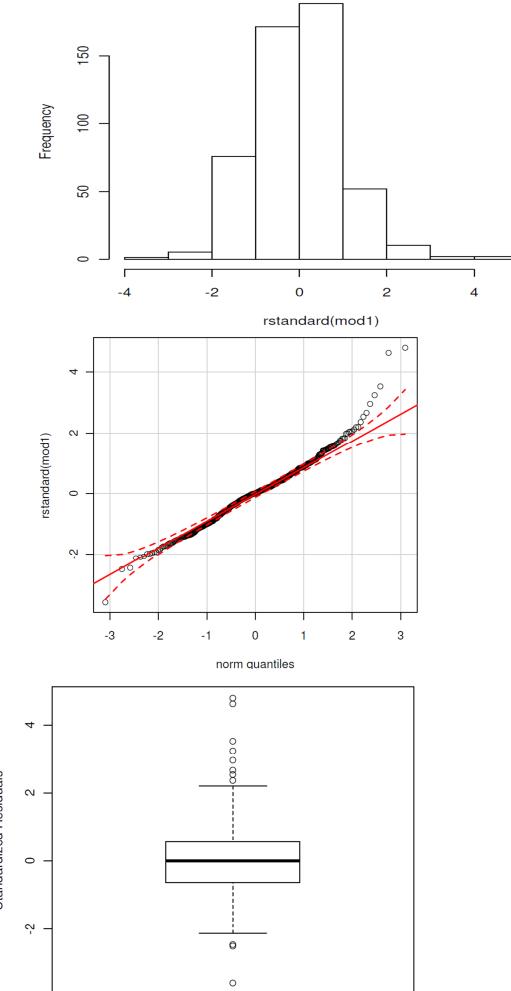
[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

- gráficos de normalidade (histograma, pp-plot) para os resíduos estandardizados



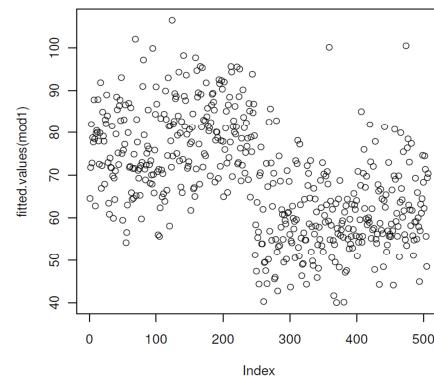
A situação de uma não normalidade *suave* pode não ser grave em amostras de tamanho *grande* uma vez que a distribuição dos estimadores $\hat{\beta}$ tenderá para a distribuição normal, essencialmente pelo teorema do limite central.

Sobre a **normalidade**, temos também que:

- (i) não é necessária para a estimação dos parâmetros de regressão pois pode-se sempre seguir o método dos mínimos quadrados; mas de qualquer forma também é sabido que as estimativas obtidas podem não ser as óptimas.
- (ii) é necessária para a realização de testes de hipóteses e/ou intervalos de confiança.

Em situações de falha clara de normalidade, podem-se considerar transformações sobre a resposta ou proceder a alterações no modelo.

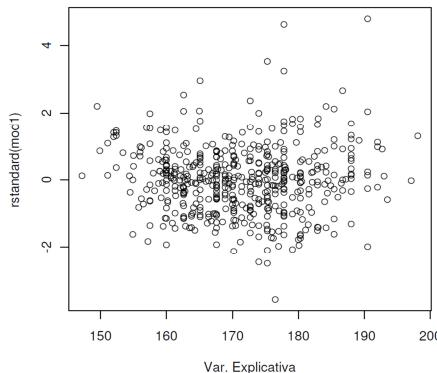
- **gráfico dos valores ajustados contra a ordem de recolha das observações**



A dispersão dos pontos não deve apresentar nenhuma tendência especial caso as observações tenham sido recolhidas de forma **independente**. Essa não parece ser a situação no gráfico anterior.

- gráfico dos resíduos estandardizados contra cada uma das variáveis explicativas ^a

Nas condições do modelo de regressão, espera-se que os resíduos não possuam qualquer relação com as variáveis explicativas e que portanto o gráfico considerado não evidencie nenhum padrão específico.



No caso da figura sugerir, por exemplo, existência de curvatura na dispersão, podemos suspeitar da necessidade de inclusão do termo quadrático X_i^2 na equação de regressão, simultaneamente à inclusão de X_i .

Nota: No R, alguns dos gráficos anteriores correspondentes a um modelo *lm()* podem ser obtidos de

```
>par(mfrow=c(2,2))
>plot(mod, which=1:4)
```

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 56 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

1.14. Instruções práticas

Nesta secção apresentamos algumas instruções de análise de regressão de índole prática.

- antes de formular qualquer modelo convém
 - enunciar o problema de forma precisa
 - descrever os dados recolhendo medidas sumárias descritivas numéricas e gráficas de cada uma das variáveis separadamente; **as distribuições observadas não devem ser muito assimétricas (usar transformações não-lineares caso exista assimetria)**
 - avaliar a qualidade dos dados (por ex., os valores são todos aceitáveis?)
 - avaliar gráfica e numericamente as associações entre a resposta e cada uma das variáveis explicativas contínuas: gráfico de dispersão, coeficiente de correlação, ... As associações são lineares?
 - avaliar gráfica e numericamente as associações entre a resposta e cada uma das variáveis explicativas categóricas: box-plots lado-a-lado, médias da resposta por categorias do factor, ... No caso de factores ordinais, verifica-se um aumento/diminuição consistente dessas médias?
 - avaliar as associações entre pares de variáveis explicativas (matriz de correlação). Existe multicolinearidade elevada?
- no modelo de regressão, não se deve incluir:
 - mais do que uma variável que diga respeito à mesma característica, para evitar problemas de multicolinearidade elevada; por exemplo, não se deve incluir peso, altura e IMC uma vez que IMC se deduz das duas variáveis anteriores
 - mais variáveis do que o número de observações disponíveis.
- uma regra empírica afirma que, por cada variável explicativa de um modelo de regressão linear múltipla, devem existir cerca de 20 observações disponíveis.
- para controlar efeitos de confundimento, as variáveis explicativas de principal interesse (usualmente um tratamento ou exposição a factores de risco) devem ser ajustadas para outras variáveis previamente identificadas. Esta identificação não é normalmente uma tarefa fácil...
- se não tiver nenhuma consideração teórica sobre os dados, ajuste o modelo de regressão linear completo (caso contrário, ajuste uma equação de regressão aos dados)
- elimine as variáveis que não são estatisticamente significativas para o modelo completo. Obtém um modelo mais simples e com um menor erro de previsão. Para o modelo reduzido:
 - realize análises gráficas que testem as hipóteses estatísticas subjacentes ao modelo (linearidade, independência das observações, normalidade e homocedasticidade dos resíduos)

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 57 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

- identifique pontos com leverages altas, outliers e pontos influentes, e estude os modelos obtidos com e sem esses pontos.
 - averigue se ainda pode retirar ou adicionar mais alguma variável explicativa sem alterar a integridade do modelo. Repita a análise residual descrita no passo anterior e itere o processo. Certifique-se que o modelo final satisfaz todos os requisitos da análise de regressão linear.
 - há situações em que é razoável manter a presença de uma (ou várias) variáveis explicativas na equação de regressão mesmo sem o respetivo coeficiente ser estatisticamente significativo. Isto acontece quando o investigador considera essa variável de grande interesse teórico para a descrição do fenômeno, e os coeficientes ajudam a uma melhor explicação dos valores da resposta em função das variáveis explicativas.
 - é aconselhável experimentar vários modelos de regressão, para obter descrições alternativas dos dados, e averiguar se existem efeitos semelhantes em todos eles.
 - na escolha do modelo final, tenha em consideração o princípio de parcimónia que refere a simplicidade da descrição do fenômeno
 - o número de variáveis explicativas a considerar pode ser determinado por um algoritmo de selecção de variáveis mas é importante conhecer as limitações desses algoritmos e usá-los unicamente como uma ferramenta de análise exploratória.
- Pode acontecer que:
- existam variáveis explicativas que possuem associações baixas com a resposta quando consideradas isoladamente e associações fortes quando consideradas em conjunto.
 - de entre duas variáveis significativamente associadas com a resposta, o algoritmo escolhe aquela que apresenta melhores propriedades estatísticas mas não necessariamente as melhores propriedades clínicas.
- na interpretação dos resultados sobre os coeficientes de regressão, intervalos de confiança com amplitudes grandes são sinal de um tamanho amostral insuficiente ou de multicolinearidade elevada.
 - mesmo que a amostra considerada seja grande, a associação estimada entre um factor de risco e a resposta pode ser imprecisa pelo facto de o factor de risco apresentar uma prevalência baixa.
- Por exemplo, se apenas 10 de 800 indivíduos forem consumidores de drogas injectáveis, o modelo não conseguirá estimar de forma precisa a relação entre o consumo de drogas injetáveis e a resposta.
- tal como num teste-*t* usual para a média, a significância dos coeficientes de regressão pode ser uma consequência de um tamanho amostral grande. Se necessário, devem usar-se também intervalos de confiança.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 58 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

- o modelo assume que as variáveis explicativas são medidas sem erro. Esta hipótese é na prática quase nunca satisfeita mas impossível de quantificar. Quando são *grandes*, os erros de medição das variáveis explicativas afetam a variância residual, o coeficiente de determinação e as estimativas dos coeficientes de regressão, fazendo descer a qualidade de predição do modelo. É preciso ter cuidado com a interpretação dos coeficientes de regressão nesta situação.
- na interpretação final dos resultados, convém não esquecer que um modelo de regressão reflecte apenas uma aproximação da verdadeira relação existente entre as variáveis em causa.
- a disponibilidade de uma amostra grande não é garantia absoluta de reproduzibilidade dos resultados de previsão relativamente a novos dados.

Na impossibilidade de recolha de novos dados, a análise poderá incluir um dos seguinte métodos de validação: *split-group*, *jackknife* ou *bootstrap*.

- *split-group* - o conjunto de dados é dividido em duas partes; o modelo é estimado a partir de uma das partes e validado com a outra.
- *jackknife* - itera o seguinte procedimento: um indivíduo é escolhido aleatoriamente, o modelo é estimado para toda a amostra excepto esse indivíduo e é feita uma avaliação dos resultados

– *bootstrap* - são escolhidas aleatoriamente várias subamostras da amostra inicial com repetição e o modelo é estimado para cada uma dessas subamostras

- Os modelos de regressão linear podem incluir polinómios de ordem superior a 1 nas variáveis explicativas. Por exemplo, se X_1 e X_2 são variáveis explicativas, podemos tentar incluir X_1^2 ou X_2^2 no modelo de forma a tentar capturar efeitos não lineares das variáveis explicativas sobre a resposta. Existe agora um senão: os termos quadrados são, geralmente, muito correlacionados com os termos originais pelo que, em alguns modelos, isto poderá conduzir a problemas numéricos de estimação! Uma solução simples que reduz a correlação consiste em considerar as variáveis X_1 e $(X_1 - \bar{\mu}_1)^2$, em lugar de X_1 e X_1^2 (usámos a notação $\bar{\mu}_1 = \bar{\mu}_{X_1} = \overline{X_1}$). Um processo mais elaborado que elimina totalmente a correlação consiste da utilização de polinómios ortogonais...

- Considere-se um modelo de regressão simples com duas variáveis predictivas, X_1 e X_2 . Pode ser instructivo comparar os resultados da regressão múltipla (considerando X_1 e X_2) com os da regressão simples, para X_1 e X_2 individualmente.

Os coeficientes obtidos das regressões simples são usualmente designados por **efeitos brutos**, ^a pois representam o efeito provocado unicamente por aquela variável predictiva sobre a resposta.

^agross effects

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 59 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 60 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Os coeficientes obtidos das regressões múltiplas são mais interessantes porque representam alterações provocadas na resposta pela variável predictiva associada para valores fixos das restantes variáveis, portanto, depois da equação ter sido ajustada para outras variáveis predictivas de interesse. Esses coeficientes são usualmente designados por **efeitos ajustados**.

A interpretação dos efeitos ajustados para outras variáveis é mais precisa e completa do que a interpretação dos efeitos brutos mas é preciso ter cuidado e não esquecer que poderão existir variáveis de confundimento que ainda não foram consideradas...

Os efeitos brutos e ajustados podem ser apresentados numa tabela, na forma

Variável Preditiva	Efeito Bruto	Efeito Ajustado
X_1
X_2

- Quanto maior for a amostra, maior é a probabilidade de identificar um efeito de uma variável explicativa que seja estatisticamente significativo sobre a resposta. Isto porque o teste de hipóteses usado (o teste de Wald) usa o tamnho amostral em denominador: quanto maior é a amostra, maior é a probabilidade de rejeitarmos $\beta = 0$ e identificarmos diferenças.

Em última instância, até se poderia dizer que $\beta = 0$ não foi rejeitado numa situação específica porque o tamanho amostral não era suficientemente grande!

De acordo com estas considerações, os testes de hipóteses feitos aos coeficientes de regressão resumem-se a testes sobre o tamanho amostral e há autores que preferem então usar intervalos de confiança para os coeficientes, em lugar dos testes de hipóteses.

- Ainda na sequência do item anterior, quanto maior for a amostra, menor será o valor-*p* obtido para os vários coeficientes portanto há que ter cuidado e não confundir valores-*p* pequenos com efeitos associados a grandes capacidades preditivas.

1.15. Variáveis explicativas categóricas

Até agora tratámos apenas da situação em que as variáveis explicativas são contínuas contudo a situação de variáveis explicativas categóricas pode também ser considerada.

Variáveis explicativas categóricas são representadas por um conjunto de variáveis auxiliares, designadas por **variáveis dummy** ou **indicatrizes**. Mais precisamente, uma variável explicativa X com $k + 1$ categorias pode ser representada por k variáveis dicotómicas dummies Z_1, \dots, Z_k da seguinte forma.

- Atribuam-se os valores numéricos $0, 1, \dots, k$ às categorias de X e designe-se a categoria correspondente ao valor 0 por **classe de referência**
- As observações pertencentes à categoria 1 de X vão ser identificadas pela variável dummy Z_1 , isto é, Z_1 é a variável dicotómica definida por

$$Z_1 = \begin{cases} 1, & \text{se } X = 1 \\ 0, & \text{se } X \neq 1 \end{cases}$$

- As observações pertencentes à categoria 2 de X vão ser identificadas pela variável dummy Z_2 , isto é, Z_2 é a variável dicotómica definida por

$$Z_2 = \begin{cases} 1, & \text{se } X = 2 \\ 0, & \text{se } X \neq 2 \end{cases}$$

- o processo anterior é iterado até à k -ésima categoria de X ; Z_k é a variável dicotómica definida por

$$Z_k = \begin{cases} 1, & \text{se } X = k \\ 0, & \text{se } X \neq k \end{cases}$$

Para uma melhor elucidação do processo, apresentamos de seguida dois exemplos.

- (a) suponhamos que X representa o sexo dos indivíduos e que se escolheu a codificação

$$0 : \text{mulher}, \quad 1 : \text{homem}.$$

Como X é um factor com duas categorias, X será representado por uma dummy, Z . De acordo com o descrito anteriormente, tem-se

$$Z = \begin{cases} 1, & \text{se } X = 1 \\ 0, & \text{se } X = 0 \end{cases}$$

- (b) suponhamos que W representa o grupo etário de um indivíduo de acordo com a classificação

$$0 : < 40 \text{ anos}, \quad 1 : 40 \leq \text{anos} \leq 65, \quad 2 : > 65 \text{ anos}.$$

O factor W será representado por duas dummies, nomeadamente I_1 e I_2 definidas por

$$I_1 = \begin{cases} 1, & \text{para indivíduos do grupo etário 1} \\ 0, & \text{caso contrário} \end{cases}$$

$$I_2 = \begin{cases} 1, & \text{para indivíduos do grupo etário 2} \\ 0, & \text{caso contrário} \end{cases}$$

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

[Page 61 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Para as duas variáveis sexo e grupo etário consideradas atrás,

- (i) um homem com 44 anos é representado pelo vector de dummies $(z, i_1, i_2) = (1, 1, 0)$
- (ii) um homem com 19 anos é representado pelo vector de dummies $(z, i_1, i_2) = (1, 0, 0)$
- (iii) uma mulher com 66 anos é representada pelo vector de dummies $(z, i_1, i_2) = (0, 0, 1)$.

Em termos de regressão linear, supondo que estávamos interessados em **usar as variáveis sexo, grupo etário e índice de massa corporal (BMI) para prever a quantidade de glicose no sangue em jejum**, a equação

$$\text{glicose} = \beta_0 + \beta_1 \text{sexo} + \beta_2 \text{gp etário} + \beta_3 \text{BMI}$$

com coeficientes $(\beta_0, \beta_1, \beta_2, \beta_3)$, seria equivalente a

$$\text{glicose} = \beta_0 + \alpha_1 Z + \gamma_1 I_1 + \gamma_2 I_2 + \beta_3 \text{BMI}$$

com coeficientes $(\beta_0, \alpha_1, \gamma_1, \gamma_2, \beta_3)$. Depois de ajustado o modelo e de obtidas estimativas $(\hat{\beta}_0, \hat{\alpha}_1, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\beta}_3)$, os indivíduos nas situações (i)-(iii) referidas acima teriam glicose prevista de

- (i) $\widehat{\text{glicose}} = \hat{\beta}_0 + \hat{\alpha}_1 + \hat{\gamma}_1 + \hat{\beta}_3 \text{BMI}$
- (ii) $\widehat{\text{glicose}} = \hat{\beta}_0 + \hat{\alpha}_1 + \hat{\beta}_3 \text{BMI}$
- (iii) $\widehat{\text{glicose}} = \hat{\beta}_0 + \hat{\gamma}_2 + \hat{\beta}_3 \text{BMI}$.

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 62 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

1.16. Confundimento

Uma **variável confundidora** é uma variável correlacionada com uma variável predictiva e com a resposta.

Atentemos nos seguintes modelos:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1.5)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 V_i + \varepsilon_i \quad (1.6)$$

Nestes modelos, X_i é o valor da variável predictiva X no indivíduo i .

Seja $\hat{\beta}_1$ uma estimativa do coeficiente β_1 .

A condição de erros normais e identicamente distribuídos implica que e_i seja independente de X_i . Ora isto não acontece quando existe uma variável de confundimento, V , que seja correlacionada simultaneamente com ε e X .

A existir confundimento, $\hat{\beta}_1$ pode ser estatisticamente significativo no modelo 1 mas não ser estatisticamente significativo no modelo 2. Neste caso, $\hat{\beta}_1$ no modelo 1 não pode ser interpretado como um indicador de um efeito de X sobre Y .

Existe alguma controvérsia quanto aos processos de identificação de variáveis confundidoras e à forma como estas devem ser abordadas quantitativamente e qualitativamente. Contudo, o **critério CE**, *Change in Estimate*, tem sido usado quase como regra para avaliar efeitos individuais de potenciais variáveis confundidoras.

A regra empírica utilizada tem sido a seguinte: uma covariável é considerada confundidora se o coeficiente de regressão parcial do termo principal mudar mais de 10% quando a covariável é adicionada (ou retirada) do modelo.

Nos modelos anteriores, existe confundimento se

$$\frac{\hat{\beta}_1(\text{Mod 1}) - \hat{\beta}_1(\text{Mod 2})}{\hat{\beta}_1(\text{Mod 1})} > 10\%.$$

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

Page 63 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.17. Interacções

Considere-se um modelo de regressão de uma variável resposta Y contra duas variáveis explicativas contínuas (covariáveis) X_1 e X_2 ,

$$Y \sim X_1 + X_2 \quad \text{ou} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

onde $\varepsilon \sim N(0, \sigma^2)$ são os resíduos do modelo.

Os coeficientes β_1 e β_2 são os **efeitos principais associados às variáveis X_1 e X_2** , respectivamente. Representam a influência da variável explicativa em causa sobre a resposta, quando o modelo está controlado para a restante variável explicativa.

Pode-se também considerar a inclusão no modelo anterior de termos não lineares que resultem de produtos cruzados, do tipo $X_1 X_2$, para capturar a interacção entre X_1 e X_2 . O modelo passará a ser

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

que continua a ser um modelo linear! (a linearidade é nos parâmetros.)

Para cada unidade experimental i , a variável $X_1 X_2$ que representa o termo da interacção toma o valor

$$x_{1i} x_{2i},$$

onde x_{1i} , x_{2i} representam os valores das variáveis X_1 e X_2 na unidade i , respectivamente. Em particular, resulta que a interacção é simétrica, *i.e.*, que

$$X_1 X_2 = X_2 X_1$$

A equação de regressão que permite avaliar a significância estatística da interacção $X_1 : X_2$ é

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 : X_2 + \varepsilon'$$

com $\varepsilon' \sim N(0, \sigma'^2)$.

Existe interacção se o coeficiente β_3 é estatisticamente significativo (diferente de zero).

Uma característica importante dos modelos que consideram **interacções** é que o efeito de qualquer uma das variáveis, X_1 ou X_2 , passa agora a depender dos valores da outra variável.

De facto, fixando X_1 podemos escrever o modelo com a interacção na forma

$$Y = (\beta_0 + \beta_1 X_1) + (\beta_2 + \beta_3 X_1) X_2 + \varepsilon$$

e agora vemos que o efeito de X_2 sobre a resposta é $\beta_2 + \beta_3 X_1$, que depende de X_1 !! Este efeito começa por ser nulo quando X_1 vale zero, e depois aumenta de β_3 unidades sempre que X_1 aumenta de uma unidade.

Sempre que se quer testar uma interacção, o modelo a considerar deve ser o anterior. De facto, considerando apenas

$$Y \sim \beta_0 + \beta_3 X_1 : X_2$$

a significância estatística de β_3 não é facilmente interpretável e confunde-se com os efeitos principais de X_1 e X_2 : mantendo X_1 constante, um aumento de uma unidade em X_2 provoca um aumento médio de Y de $\beta_3 X_1$.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 64 of 104](#)

[Go Back](#)

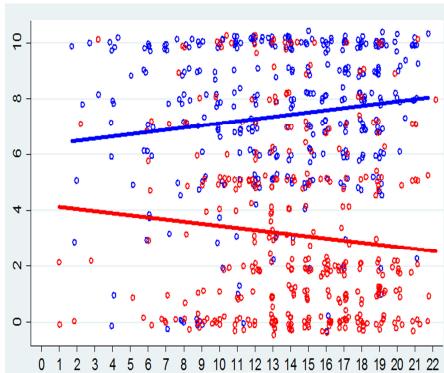
[Full Screen](#)

[Close](#)

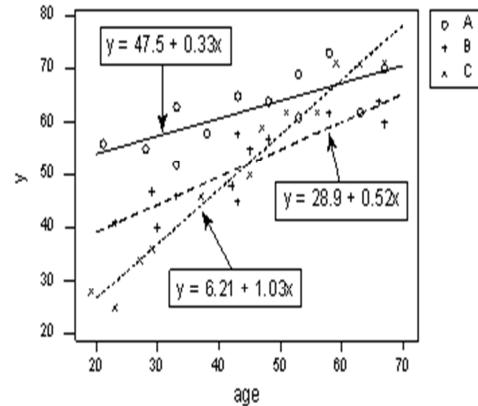
[Quit](#)

Havendo mais variáveis, podem-se considerar interacções de ordem superior a 2, tendo o cuidado de incluir no modelo todas as variáveis e interacções de ordem inferior. Contudo, é raro encontrar interacções de ordem superior a três por serem difíceis de interpretar.

No gráfico abaixo, existe uma interacção entre o factor, representado pelas cores vermelha e azul, e a variável que está no eixo dos xx's. Para a classe azul é estimada uma tendência crescente com os valores de X enquanto que para a classe vermelha é estimada uma tendência descrecente.



Suponha que a figura abaixo reflecte a resposta Y de acordo com a idade dos indivíduos, para diferentes categorias de índice de massa corporal (IMC). Pode-se ver, por exemplo, que a categoria IMC=A é a que apresenta um menos crescimento de Y ao longo dos anos, enquanto que a categoria IMC=C é a que apresenta o maior crescimento de Y ao longo do tempo. Diz-se que existe interacção entre a idade e o IMC dos indivíduos.



1.18. Transformações para a linearidade

Há muitas situações em que a variável resposta se pode escrever na forma

$$Y(x) = f(x; \beta) + u(x)$$

ou na forma

$$Y(x) = f(x; \beta) u(x)$$

onde f é uma função não necessariamente linear, i.e., não necessariamente da forma

$$f(x) = (1, x). \beta = \beta_0 + x_1 \beta_1 + \cdots + x_p \beta_p,$$

e $u(x) \sim F(\varphi(x))$ é uma variável aleatória que segue uma distribuição F dependendo de um parâmetro φ que por sua vez pode depender das variáveis explicativas x .

Numa tal situação, dizemos que estamos perante um **modelo de regressão não linear**. A teoria subjacente a tais modelos é mais complexa do que a teoria do modelo linear geral que estamos a tratar neste capítulo ou à do chamado modelo linear generalizado (regressão logística, regressão de Poisson, regressão de Cox, ...) que trataremos no capítulo seguinte. Porém, há situações muito comuns em que uma transformação da variável resposta Y ou das variáveis explicativas x reconduz-nos ao modelo linear geral. É de algumas dessas transformações que vamos tratar nesta secção.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

Page 66 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

1.19. Transformações não lineares

Há situações em que a relação entre a resposta e (pelo menos) uma das variáveis explicativas não é linear, mas que pode ser tornada linear se se considerarem transformações adequadas de uma ou ambas as variáveis em causa.

Noutras situações, a normalidade (ou pelo menos simetria) dos dados, ou mesmo a homocedasticidade, é apenas conseguida com transformações de variáveis.

De entre as transformações mais comuns, destacam-se as seguintes:

- **logaritmo:** aplica-se quando os aumentos de uma variável estão relacionados com a proporção com que a outra variável aumenta. Por exemplo, se quando o valor de uma variável explicativa aumenta de 1 unidade se tem que o valor da resposta passa para o dobro, então uma [transformação logarítmica da resposta](#) pode tornar a relação linear.

Usualmente, as respostas que representam [tempos de reacção](#) precisam de ser logaritmizadas para as relações com as outras variáveis serem do tipo linear (Newell & Rosenbloom, 1981).

Nas situações em que aumentos proporcionais na variável explicativa provocam aumentos proporcionais na resposta (por exemplo, a resposta duplica sempre que a variável explicativa duplica), então a relação pode ser linearizada aplicando [transformações logarítmicas a ambas as variáveis](#).

Nos casos anteriores, a escolha da base da função logarítmica é indiferente, sendo que as mais usuais são a base 10 (\log_{10}) e a base e ($\ln = \log_e$).

- **inverso:** se necessário, é aplicado às variáveis que representam taxas (número médio de ocorrências por unidade de tempo), especialmente quando a outra variável em causa está relacionada os períodos de tempo decorridos entre as ocorrências.
- **raiz quadrada:** é indicada para as variáveis que representam o número de ocorrências de um evento num determinado período de tempo. Na situação em que a frequência das ocorrências é baixa, Freeman and Tukey (1950) sugerem o uso de uma transformação do tipo $\sqrt{X} + \sqrt{X+1}$, que conduz mais facilmente a um contexto de homocedasticidade.
- ...

Mais geralmente temos a **transformação de Box-Cox**: o método de Box-Cox (1964) permite [escolher a melhor transformação da resposta](#), por potência, com base directamente nos dados. A ideia geral consiste em restringirmo-nos a transformações indexadas por parâmetros desconhecidos λ (usualmente, um único parâmetro), e depois estimar λ e os outros parâmetros do modelo por métodos standard de inferência estatística.

Para variáveis dependentes que tomem apenas valores positivos, considera-se a família paramétrica de transformações de y a $y^{(\lambda)}$ dada por

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & (\lambda \neq 0) \\ \log(y) & (\lambda = 0). \end{cases}$$

Os autores assumem para um valor de λ desconhecido as observações transformadas $y_i^{(\lambda)}$, $i = 1, \dots, n$, satisfazem a totalidade das hipóteses do modelo normal, i.e., são independentemente e normalmente distribuídas com variância constante, e a média condicional da resposta relaciona-se linearmente com as variáveis explicativas. A estimativa para λ é obtida por maximização da função de verosimilhança dos dados originais, escrita à custa dos dados transformados.

No **R**, a instrução a usar é

```
> boxcox()
```

A transformação devolvida pelo método de Box-Cox não deve ser aceite de forma cega mas apenas como uma indicação. É aconselhável que o uso de transformações seja considerado essencialmente quando exista suporte científico que o justifique. Contudo, análises exploratórias gráficas podem também conduzir a transformações de variáveis, para garantir a normalidade dos resíduos e a homocedasticidade, por exemplo.

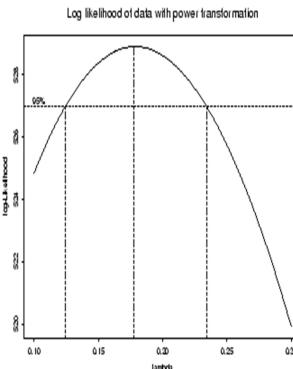
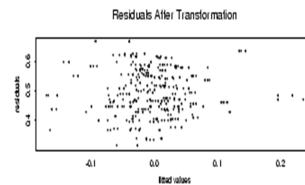
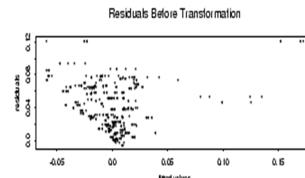


Figure 3: Residuals before and after transformation

Figure 4: Likelihood for power transformation

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀](#) [▶](#)
[◀](#) [▶](#)
[Page 68 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

Chapter 2

Exemplos de Regressão Linear em R

[Home Page](#)

[Title Page](#)

[Contents](#)

[« «](#) [» »](#)

[◀](#) [▶](#)

Page 69 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

2.1. Exemplo: dimensões corporais

O ficheiro **bodyLM.sav** contém dados recolhidos em 507 indivíduos adultos jovens fisicamente activos (praticando várias horas de exercício físico por semana) de vários estados americanos. A amostra consta de 247 homens e de 260 mulheres e, de entre as variáveis consideradas, constaram 12 medidas corporais de circunferência, a idade, o peso, a altura e o sexo. Um dos objectivos do estudo consistiu da modelação do peso através das várias medidas de circunferência recolhidas e da altura.

O ficheiro contém as seguintes variáveis:

- **ShoulderG**: shoulder girth over deltoid muscles (cm)
- **ChestG**: chest girth, nipple line in males and just above breast tissue in females, mid-expiration (cm)
- **WaistG**: waist girth, narrowest part of torso below the rib cage, average of contracted and relaxed position (cm)
- **NavelG**: navel (or "abdominal") girth at umbilicus and iliac crest, iliac crest as a landmark (cm)
- **HipG**: hip girth at level of bitrochanteric diameter (cm)
- **ThighG**: thigh girth below gluteal fold, average of right and left girths (cm)
- **BicepG**: bicep girth, flexed, average of right and left girths (cm)
- **ForearmG**: forearm girth, extended, palm up, average of right and left girths (cm)
- **KneeG**: knee girth over patella, slightly flexed position, average of right and left girths (cm)
- **CalfMaxG**: calf maximum girth, average of right and left girths (cm)
- **AnkleMinG**: ankle minimum girth, average of right and left girths (cm)
- **WristMinG**: wrist minimum girth, average of right and left girths (cm)
- **age**: idade (anos)
- **weight**: peso (Kg)
- **height**: altura (cm)
- **sex**: sexo (0- mulher; 1- homem)

Considerando o objectivo do estudo e não tendo nenhuma indicação extra sobre as variáveis em causa, podemos começar por considerar o modelo completo, com todas as variáveis explicativas.

$$\text{weight} \sim \text{ShoulderG} + \text{ChestG} + \text{WaistG} + \text{NavelG} \\ + \text{HipG} + \text{ThighG} + \text{BicepG} + \text{ForearmG} \\ + \text{KneeG} + \text{CalfMaxG} + \text{AnkleMinG} \\ + \text{WristMinG} + \text{height}$$

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

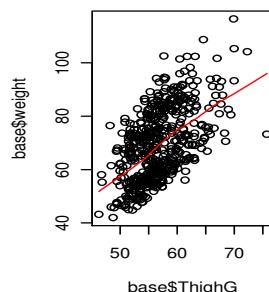
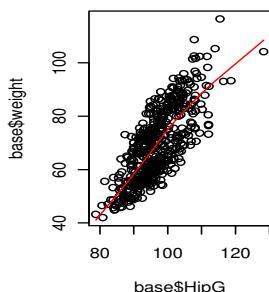
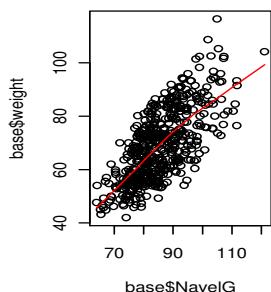
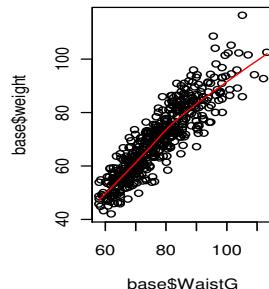
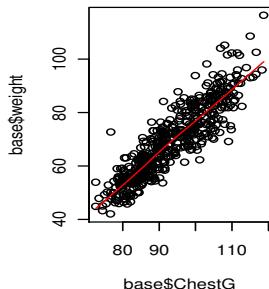
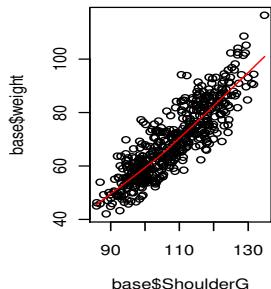
[Page 70 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)



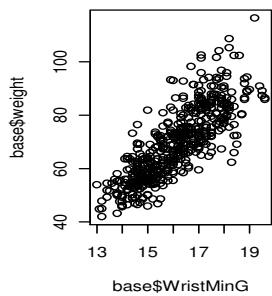
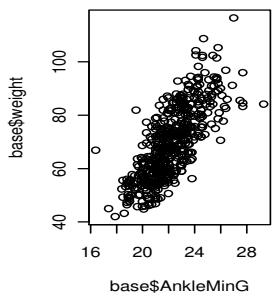
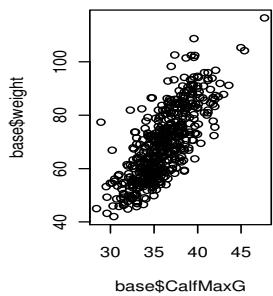
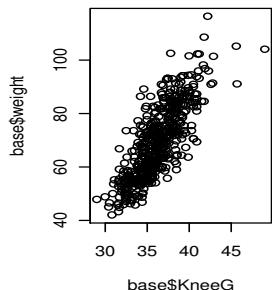
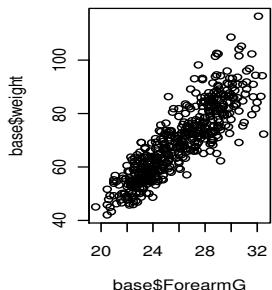
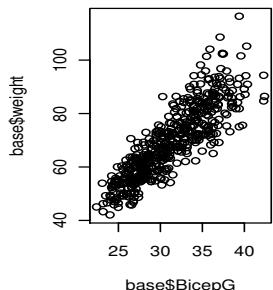
Page 71 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)



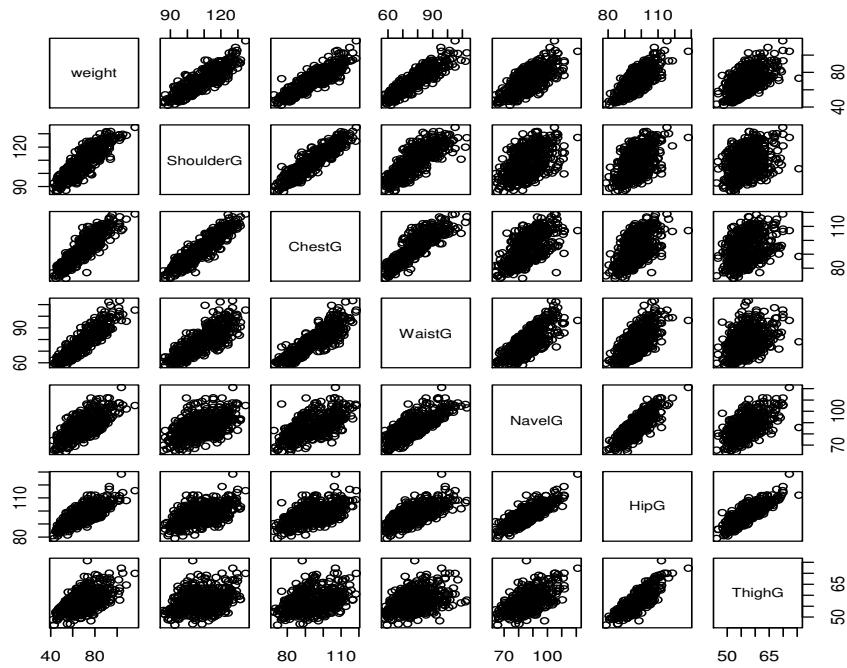
[Page 72 of 104](#)

[Go Back](#)

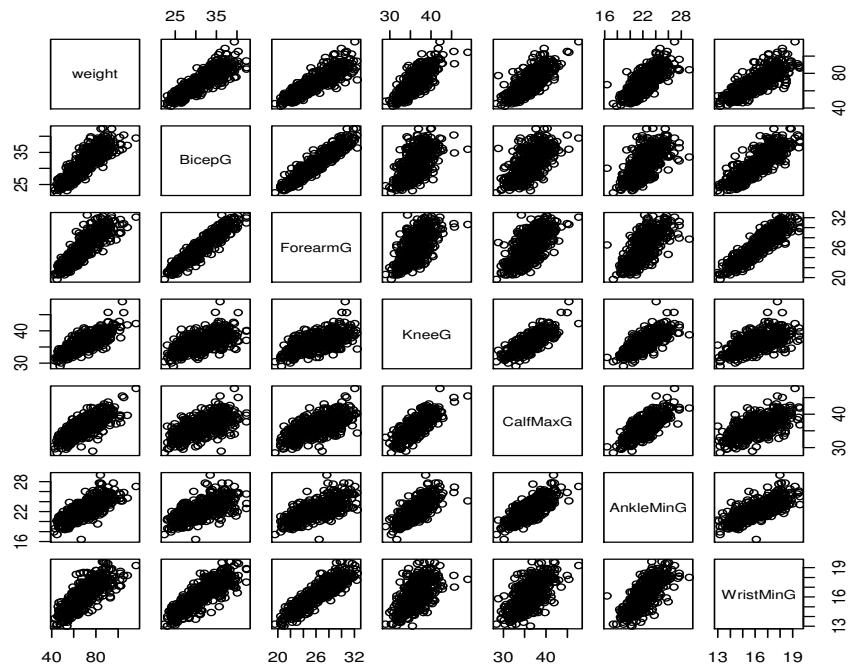
[Full Screen](#)

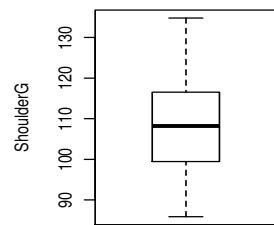
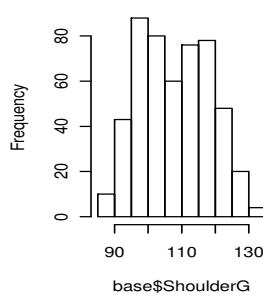
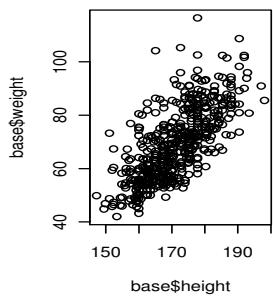
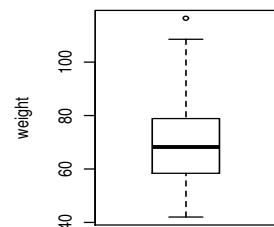
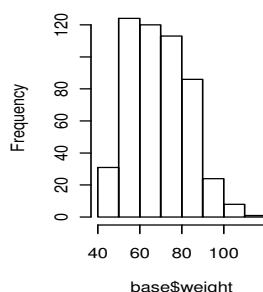
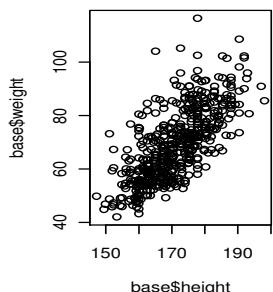
[Close](#)

[Quit](#)


[Home Page](#)
[Title Page](#)
[Contents](#)

[Page 73 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)


[Home Page](#)
[Title Page](#)
[Contents](#)
[!\[\]\(44c673e7df06f4f5e0fe437046ae9fad_img.jpg\)](#)
[!\[\]\(5a086b6f6d1da930544089509dc3b02a_img.jpg\)](#)
[!\[\]\(a282a674e7588c8b9a0290c8b253a834_img.jpg\)](#)
[!\[\]\(5b5b748f2000b5a4c200c59ac0c71c78_img.jpg\)](#)
[Page 74 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)



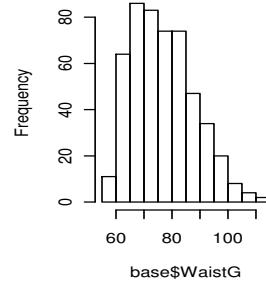
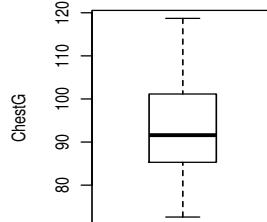
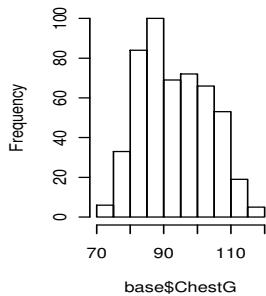
[Page 75 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)



[Page 76 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

```
> base <- read.spss("BodyLM.sav", to.data.frame=TRUE, use.value.labels=FALSE)
Warning message:
In read.spss("BodyLM.sav", to.data.frame = TRUE, use.value.labels = FALSE) :
  BodyLM.sav: Unrecognized record type 7, subtype 18 encountered in system file
>
> names(base)
[1] "ShoulderG"   "ChestG"      "WaistG"      "NavelG"      "HipG"       "ThighG"
[7] "BicepG"       "ForearmG"    "KneeG"       "CalfMaxG"    "AnkleMinG"  "WristMinG"
[13] "age"          "weight"      "height"     "sex"
> dim(base)
[1] 507  16
```

[Home Page](#)

```
> par(mfrow=c(2,3))
plot(base$ShoulderG, base$weight)
lines(lowess(base$ShoulderG, base$weight), col="red")
```

[Title Page](#)

```
plot(base$ChestG, base$weight)
lines(lowess(base$ChestG, base$weight), col="red")
```

[Contents](#)

```
plot(base$WaistG, base$weight)
lines(lowess(base$WaistG, base$weight), col="red")
```



```
plot(base$NavelG, base$weight)
lines(lowess(base$NavelG, base$weight), col="red")
```



```
plot(base$HipG, base$weight)
lines(lowess(base$HipG, base$weight), col="red")
```

[Page 77 of 104](#)

```
plot(base$ThighG, base$weight)
lines(lowess(base$ThighG, base$weight), col="red")
```

[Go Back](#)

[Full Screen](#)

```
> cor(base[,c(14, 1:12)])
      weight ShoulderG   ChestG   WaistG   NavelG     HipG    ThighG
weight    1.0000000 0.8788342 0.8989595 0.9039908 0.7118165 0.7629691 0.5585626
ShoulderG 0.8788342 1.0000000 0.9271923 0.8234546 0.5154661 0.5336717 0.3234272
ChestG    0.8989595 0.9271923 1.0000000 0.8837994 0.6229823 0.5834991 0.3630508
WaistG    0.9039908 0.8234546 0.8837994 1.0000000 0.7547704 0.6923506 0.4210849
NavelG    0.7118165 0.5154661 0.6229823 0.7547704 1.0000000 0.8258924 0.6026428
HipG      0.7629691 0.5336717 0.5834991 0.6923506 0.8258924 1.0000000 0.8289411
ThighG    0.5585626 0.3234272 0.3630508 0.4210849 0.6026428 0.8289411 1.0000000
BicepG    0.8666722 0.8951884 0.9081845 0.8047044 0.5578071 0.5598848 0.4114580
ForearmG  0.8695531 0.8949838 0.8875909 0.7807924 0.4862181 0.5143585 0.3452848
KneeG     0.7955518 0.6247826 0.6140547 0.6582072 0.6120932 0.7349017 0.6384400
CalfMaxG  0.7692826 0.6270538 0.6088643 0.6313445 0.5247789 0.6745805 0.6288901
AnkleMinG 0.7619985 0.6797568 0.6691396 0.6558891 0.5194785 0.5770429 0.4217687
WristMinG 0.8164884 0.8407085 0.8246754 0.7289813 0.4354197 0.4588567 0.2416102
```

```
      BicepG ForearmG   KneeG   CalfMaxG AnkleMinG WristMinG
weight    0.8666722 0.8695531 0.7955518 0.7692826 0.7619985 0.8164884
ShoulderG 0.8951884 0.8949838 0.6247826 0.6270538 0.6797568 0.8407085
ChestG    0.9081845 0.8875909 0.6140547 0.6088643 0.6691396 0.8246754
WaistG    0.8047044 0.7807924 0.6582072 0.6313445 0.6558891 0.7289813
NavelG    0.5578071 0.4862181 0.6120932 0.5247789 0.5194785 0.4354197
HipG      0.5598848 0.5143585 0.7349017 0.6745805 0.5770429 0.4588567
ThighG    0.4114580 0.3452848 0.6384400 0.6288901 0.4217687 0.2416102
BicepG    1.0000000 0.9423755 0.6207299 0.6374041 0.6693240 0.8479443
ForearmG  0.9423755 1.0000000 0.6575450 0.6701918 0.7125539 0.9047086
KneeG     0.6207299 0.6575450 1.0000000 0.7958277 0.7377154 0.6409596
CalfMaxG  0.6374041 0.6701918 0.7958277 1.0000000 0.7622219 0.6476269
AnkleMinG 0.6693240 0.7125539 0.7377154 0.7622219 1.0000000 0.7536365
WristMinG 0.8479443 0.9047086 0.6409596 0.6476269 0.7536365 1.0000000
```

```
> pairs(base[,c(14, 1:6)])
> pairs(base[,c(14, 7:12)])
```

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)

Page 78 of 104

[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

```
> mod1 <- lm(weight ~ ShoulderG+ChestG+WaistG+NavelG+HipG+ThighG+BicepG+
+                 ForearmG+KneeG+CalfMaxG+AnkleMinG+WristMinG+height, data=base)
> summary(mod1)
```

Call:

```
lm(formula = weight ~ ShoulderG + ChestG + WaistG + NavelG +
    HipG + ThighG + BicepG + ForearmG + KneeG + CalfMaxG + AnkleMinG +
    WristMinG + height, data = base)
```

[Home Page](#)

Residuals:

Min	1Q	Median	3Q	Max
-7.7332	-1.3677	0.0069	1.2171	10.3976

[Title Page](#)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.2020e+02	2.4890e+00	-48.306	< 2e-16 ***
ShoulderG	7.8130e-02	2.9790e-02	2.622	0.009001 **
ChestG	1.9790e-01	3.5690e-02	5.544	4.83e-08 ***
WaistG	3.4040e-01	2.4380e-02	13.960	< 2e-16 ***
NavelG	1.1720e-03	2.2910e-02	0.051	0.959225
HipG	2.4040e-01	4.3340e-02	5.547	4.76e-08 ***
ThighG	3.1410e-01	5.1480e-02	6.103	2.11e-09 ***
BicepG	5.4680e-02	8.5260e-02	0.641	0.521631
ForearmG	5.3210e-01	1.3710e-01	3.882	0.000118 ***
KneeG	3.0130e-01	7.7400e-02	3.892	0.000113 ***
CalfMaxG	4.0390e-01	7.0050e-02	5.765	1.44e-08 ***
AnkleMinG	-9.6350e-03	9.9920e-02	-0.096	0.923221
WristMinG	-1.1800e-01	1.9590e-01	-0.602	0.547135
height	3.2820e-01	1.5600e-02	21.033	< 2e-16 ***

Signif. codes: 0 `***' 0.001 `*' 0.01 `.' 0.05 `.' 0.1 ` ' 1

Residual standard error: 2.204 on 493 degrees of freedom
 Multiple R-squared: 0.9734, Adjusted R-squared: 0.9727
 F-statistic: 1390 on 13 and 493 DF, p-value: < 2.2e-16

[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 79 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

```
> anova(mod1)
Analysis of Variance Table
```

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ShoulderG	1	69607	69607	14332.992	< 2.2e-16 ***
ChestG	1	4544	4544	935.680	< 2.2e-16 ***
WaistG	1	4822	4822	992.990	< 2.2e-16 ***
NavelG	1	1131	1131	232.923	< 2.2e-16 ***
HipG	1	3089	3089	636.027	< 2.2e-16 ***
ThighG	1	226	226	46.615	2.549e-11 ***
BicepG	1	201	201	41.442	2.884e-10 ***
ForearmG	1	997	997	205.275	< 2.2e-16 ***
KneeG	1	755	755	155.542	< 2.2e-16 ***
CalfMaxG	1	151	151	30.991	4.263e-08 ***
AnkleMinG	1	24	24	4.851	0.028092 *
WristMinG	1	34	34	6.929	0.008747 **
height	1	2148	2148	442.391	< 2.2e-16 ***
Residuals	493	2394	5		

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

[Page 80 of 104](#)

coeficientes estandardizados

ShoulderG	ChestG	WaistG	NavelG	HipG	ThighG
0.060735509	0.148661084	0.280903890	0.000827652	0.120334352	0.104980810
BicepG	ForearmG	KneeG	CalfMaxG	AnkleMinG	WristMinG
0.017400311	0.112861443	0.059086773	0.086176095	-0.001344473	-0.012212489
height					
0.231311627					

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

```
> regressionSS <- sum(anova(mod1)[1:13,2] )  
> regressionSS  
[1] 87729.14  
  
> residualSS <- anova(mod1)[14,2]  
> residualSS  
[1] 2394.205  
  
> totalSS <- residualSS + regressionSS  
> totalSS  
[1] 90123.34  
  
# resíduos estandardizados  
> hist(rstandard(mod1))  
> boxplot(rstandard(mod1))  
> rqnorm(rstandard(mod1))  
> qqline(rstandard(mod1))  
> hist(rstandard(mod1), breaks=20)  
  
> which(rstandard(mod1)>3.3)  
# 4 observações; poder-se-ia explorar a relevância e o significado físico  
# dessas observações no contexto dos dados  
  
> library(car)  
> qqPlot(residuals(mod1))  
  
# resíduos studentizados; só funciona para modelos lm  
> qqPlot(mod1)  
  
> plot(fitted.values(mod1), rstandard(mod1))  
> abline(0,0,lty="dashed")  
> plot(fitted.values(mod1), base$weight)  
> abline(0,1, lwd=2, col="red")  
> plot(base$ShoulderG, rstandard(mod1))  
> plot(base$height, rstandard(mod1))
```

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

[Page 81 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

A tabela da ANOVA mostra que o modelo é estatisticamente significativo ($F=1389.588$, $p=0.000$) sendo que o efeito de quatro variáveis explicativas (NavelG, BicepG, AnkleMinG, WristMinG) sobre o peso não é significativo. Os coeficientes estandardizados indicam que as variáveis WaistG e height são as que mais influenciam o peso dos indivíduos.

O coeficiente de determinação é significativo e bastante elevado $R^2 = 0.973$ ($p=0.000$), $\bar{R}^2 = 0.973$.

Antes de retirarmos mais conclusões sobre o modelo verificamos se os seus pressupostos são satisfeitos (gráficos na página seguinte):

- o histograma, o boxplot e o qq-plot dos resíduos estandardizados sugerem ligeiros desvios da normalidade, que parecem contudo toleráveis dado o elevado tamanho amostral (estamos a usar o teorema do limite central); identificam-se ainda algumas observações com resíduos superiores a 3.3 unidades.
- o gráfico dos resíduos estandardizados contra os valores ajustados indica pequenos problemas com a hipótese de homocedasticidade e evidencia 3 pontos com resíduos grandes; os gráficos dos resíduos estandardizados contra cada uma das variáveis regressoras e dos valores ajustados contra a resposta indicam conclusões análogas. A hipótese de homocedasticidade não parece estar a ser grandemente violada.
- o gráfico das leverages contra as distâncias de Cook (nas páginas seguintes) não revela a existência de pontos influentes.

O ponto de corte tradicional para os leverages é de $2(p + 1)/n = 2(13 + 1)/507 = 0.055$ e para as distâncias de Cook é 1.0. Há alguns pontos com leverages superior a 0.055 mas não parecem ser problemáticos porque as suas distâncias de Cook são inferiores a 1.

Acontece porém que

- uma análise à multicolinearidade do modelo revela altos coeficientes de correlação entre várias das variáveis explicativas
(gráfico de dispersão das variáveis explicativas duas a duas e tabela dos coeficientes de correlação)
- alguns dos coeficientes de regressão não são significativos, o que sugere que pelo menos algumas das variáveis explicativas que lhes estão associadas possam ser excluídas do modelo
- a interpretação do modelo seria mais simples caso este contivesse menos variáveis.

Uma forma de reduzir o número de variáveis explicativas poderia ser por eliminação sucessiva daquelas que apresentam um maior valor- p , portanto as menos significativas. Em cada eliminação ter-se-ia de interpretar as várias estimativas obtidas, incluindo estimativas relacionadas com o ajustamento do modelo, e averiguar a satisfação das hipóteses das equações de regressão através de análises gráficas adequadas.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

[Page 82 of 104](#)

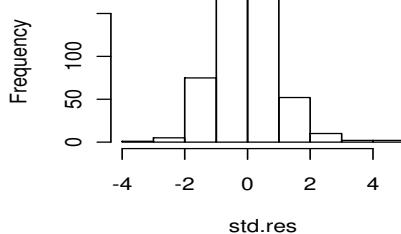
[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Histogram of std.res



[Home Page](#)

[Title Page](#)

[Contents](#)



[Page 83 of 104](#)

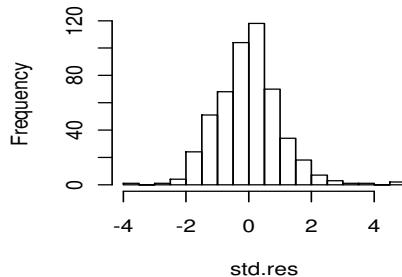
[Go Back](#)

[Full Screen](#)

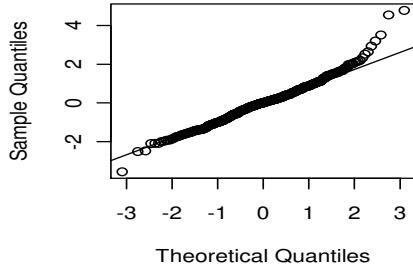
[Close](#)

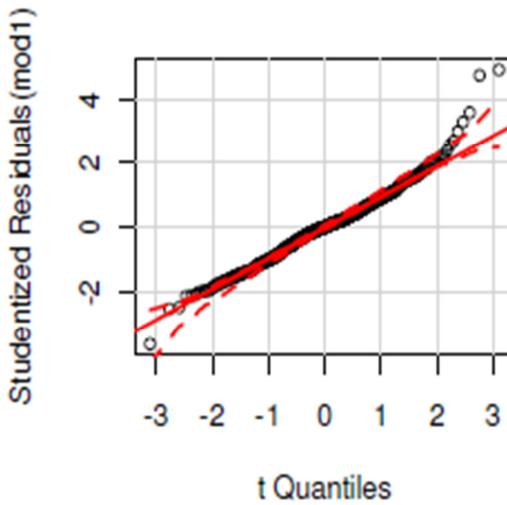
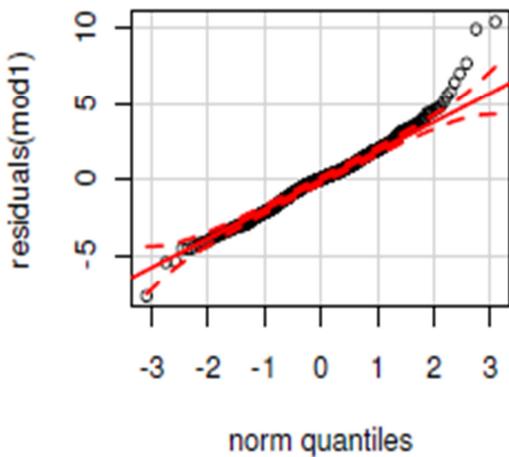
[Quit](#)

Histogram of std.res



Normal Q-Q Plot





[Home Page](#)

[Title Page](#)

[Contents](#)

[!\[\]\(8892ec72c0bc57672fb7190d39b54289_img.jpg\)](#) [!\[\]\(bb22d49e68e521365301991896f26f1f_img.jpg\)](#)

[!\[\]\(f405b1e8bd04ee8aed9e864a3e5b89df_img.jpg\)](#) [!\[\]\(2786933c353b2fdce008450bc3738216_img.jpg\)](#)

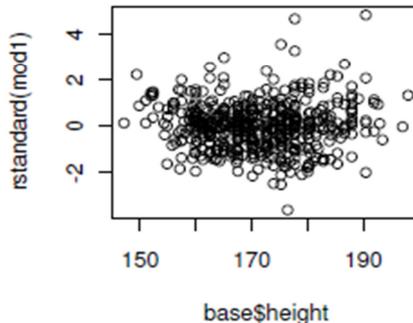
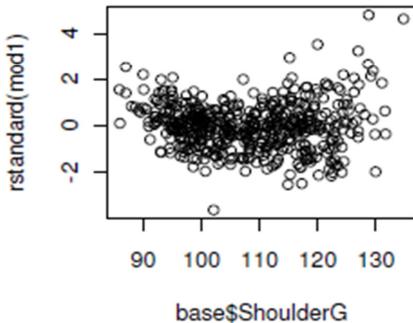
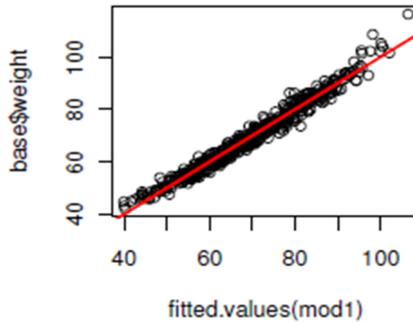
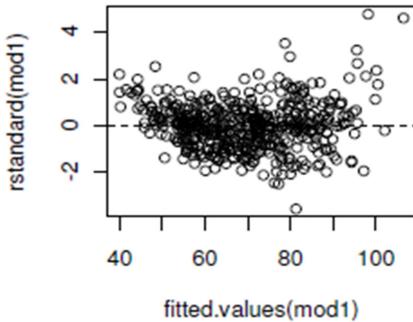
Page 84 of 104

[Go Back](#)

[Full Screen](#)

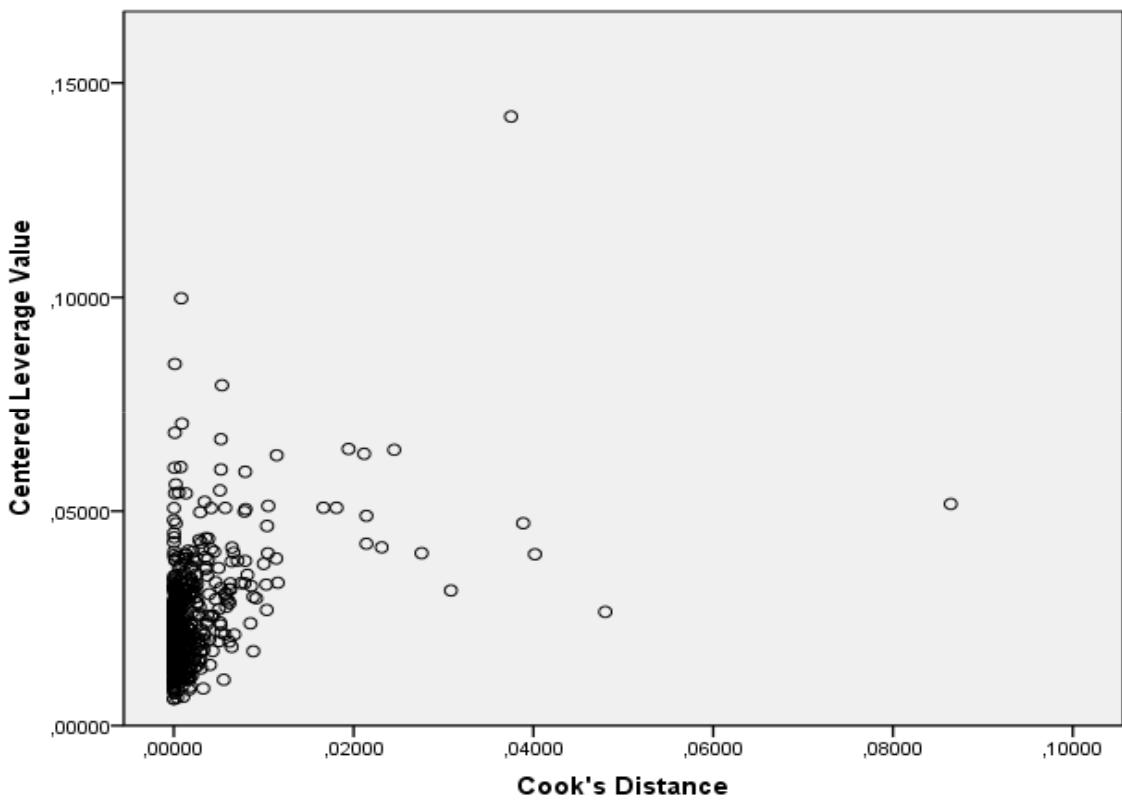
[Close](#)

[Quit](#)



Há 4 observações com resíduos superiores a 3.3 desvios-padrão. Poder-se-ia explorar a relevância e significado físico dessas observações no conjunto de dados.

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 85 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)



[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

Page 86 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Aproveitamos este exemplo para aplicar alguns métodos de selecção de variáveis. Seguiremos o método stepwise, em ambas as direcções (instruções e resultados mais pormenorizados na página seguinte).

O modelo escolhido é

$$\text{weight} \sim \text{ShoulderG} + \text{ChestG} + \text{WaistG} \\ + \text{HipG} + \text{ThighG} + \text{ForearmG} \\ + \text{KneeG} + \text{CalfMaxG} + \text{height}$$

com 9 variáveis explicativas.

Apesar do problema de multicolinearidade persistir, há diagnósticos que se podem fazer que levam a crer que isso não esteja a afectar o ajustamento do modelo:

- os erros padrão dos coeficientes não são exageradamente grandes
- não há incongruências entre os resultados das estatísticas F e t
- retiraram-se algumas variáveis do modelo e os coeficientes das outras variáveis não sofreram alterações substanciais.

Análises gráficas standard (gráficos na página seguinte) não revelam violação das hipóteses de normalidade, homocedasticidade, linearidade e independência.

Receamos apenas que os coeficientes estejam a ser demasiado significativos devido à multicolinearidade elevada pelo que, de seguida, tentaremos reduzir essa multicolinearidade e ajustaremos um terceiro modelo.

- De entre ShoulderG, ChestG e WaistG, todas muito correlacionados, escolhemos ficar apenas com a variável WaistG por apresentar o maior coeficiente estandardizado e o valor mais significativo da estatística t e portanto nos estar a dar a indicação de que será essa variável que mais contribui para explicar o peso.

- De entre HipG e ThighG escolhemos ThighG por apresentar o maior valor para a estatística t
- De entre KneeG e CalfMaxG escolhemos CalfMaxG por apresentar o maior valor para a estatística t
- De entre WaistG e ForearmG não eliminamos nenhuma porque ambas as variáveis parecem estar a ser muito significativas para o modelo e o coeficiente de correlação não é exageradamente elevado.

Os restantes coeficientes de correlação já parecem ser razoáveis pelo que prosseguimos então com o modelo

$$\text{weight} \sim \text{WaistG} + \text{ThighG} + \text{ForearmG} \\ + \text{CalfMaxG} + \text{height}$$

agora com 5 variáveis explicativas.

Todos os coeficientes permanecem muito significativos o que parece indicar que a multicolinearidade do modelo anterior não está a provocar problemas.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 87 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Em relação a este último modelo (exercício):

- $R^2 = 0.966$ ($p = 0.000$), $R^2 = 0.966$
- $F = 2836.678$ ($p = 0.000$)
- não existe multicolinearidade elevada entre as variáveis (coeficientes de correlação razoáveis e $VIF < 10$ para todas as variáveis explicativas; o índice de condição é elevado mas as proporções de variância não são elevadas em mais de duas variáveis portanto não é de valorizar)
- todos os pressupostos do modelo parecem ser satisfeitos
- o gráfico das leverages contra as distâncias de Cook revela a existência de um ponto influente. O efeito da remoção desse ponto sobre as estimativas dos coeficientes deverá ser analisada.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

[Page 88 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

```
> mod2 <- step(mod1, direction="both")

Start: AIC=815.01
weight ~ ShoulderG + ChestG + WaistG + NavelG + HipG + ThighG +
      BicepG + ForearmG + KneeG + CalfMaxG + AnkleMinG + WristMinG +
      height
```

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 89 of 104](#)

	Df	Sum of Sq	RSS	AIC
- NavelG	1	0.01	2394.2	813.02
- AnkleMinG	1	0.05	2394.3	813.02
- WristMinG	1	1.76	2396.0	813.39
- BicepG	1	2.00	2396.2	813.44
<none>			2394.2	815.01
- ShoulderG	1	33.40	2427.6	820.04
- ForearmG	1	73.17	2467.4	828.28
- KneeG	1	73.56	2467.8	828.36
- ChestG	1	149.26	2543.5	843.68
- HipG	1	149.40	2543.6	843.70
- CalfMaxG	1	161.43	2555.6	846.10
- ThighG	1	180.86	2575.1	849.94
- WaistG	1	946.48	3340.7	981.91
- height	1	2148.43	4542.6	1137.72

Step: AIC=813.02

```
weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + BicepG +
      ForearmG + KneeG + CalfMaxG + AnkleMinG + WristMinG + height
```

	Df	Sum of Sq	RSS	AIC
- AnkleMinG	1	0.04	2394.3	811.03
- WristMinG	1	1.80	2396.0	811.40
- BicepG	1	2.09	2396.3	811.46
<none>			2394.2	813.02
+ NavelG	1	0.01	2394.2	815.01
- ShoulderG	1	35.06	2429.3	818.39
- KneeG	1	74.28	2468.5	826.51
- ForearmG	1	74.34	2468.6	826.52
- ChestG	1	153.89	2548.1	842.60

[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

```
- CalfMaxG   1    163.46 2557.7  844.50
- ThighG     1    183.09 2577.3  848.38
- HipG       1    200.26 2594.5  851.74
- WaistG     1   1080.33 3474.5  999.83
- height     1   2153.30 4547.5 1136.27
```

Step: AIC=811.03

weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + BicepG +
 ForearmG + KneeG + CalfMaxG + WristMinG + height

.

.

Step: AIC=807.87

weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + ForearmG +
 KneeG + CalfMaxG + height

	Df	Sum of Sq	RSS	AIC
<none>		2398.2	807.87	
+ WristMinG	1	1.88	2396.4	809.47
+ BicepG	1	1.87	2396.4	809.47
+ AnkleMinG	1	0.36	2397.9	809.79
+ NavelG	1	0.13	2398.1	809.84
- ShoulderG	1	38.00	2436.2	813.84
- KneeG	1	72.48	2470.7	820.96
- ChestG	1	169.16	2567.4	840.42
- ForearmG	1	177.77	2576.0	842.12
- CalfMaxG	1	180.18	2578.4	842.60
- HipG	1	196.99	2595.2	845.89
- ThighG	1	240.80	2639.0	854.38
- WaistG	1	1103.10	3501.3	997.72
- height	1	2219.38	4617.6	1138.03

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

Page 90 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

```
> summary(mod2)
```

Call:

```
lm(formula = weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG +  
ForearmG + KneeG + CalfMaxG + height, data = base)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.7663	-1.3694	0.0286	1.2146	10.3239

[Home Page](#)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-120.83845	2.37022	-50.982	< 2e-16 ***
ShoulderG	0.08007	0.02853	2.806	0.005211 **
ChestG	0.20152	0.03404	5.921	5.98e-09 ***
WaistG	0.34283	0.02267	15.120	< 2e-16 ***
HipG	0.23724	0.03713	6.389	3.84e-10 ***
ThighG	0.33178	0.04697	7.064	5.48e-12 ***
ForearmG	0.54867	0.09040	6.070	2.55e-09 ***
KneeG	0.28703	0.07406	3.876	0.000121 ***
CalfMaxG	0.38924	0.06370	6.111	2.01e-09 ***
height	0.32519	0.01516	21.446	< 2e-16 ***

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

[Title Page](#)

Residual standard error: 2.197 on 497 degrees of freedom
Multiple R-squared: 0.9734, Adjusted R-squared: 0.9729
F-statistic: 2020 on 9 and 497 DF, p-value: < 2.2e-16

[Contents](#)

```
> anova(mod2, mod1) # F test for nested models  
Analysis of Variance Table
```

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

Model 1: weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + ForearmG +
KneeG + CalfMaxG + height

Model 2: weight ~ ShoulderG + ChestG + WaistG + NavelG + HipG + ThighG +
BicepG + ForearmG + KneeG + CalfMaxG + AnkleMinG + WristMinG +
height

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

```
Res.Df      RSS Df Sum of Sq      F Pr(>F)
1      497 2398.2
2      493 2394.2  4     4.0352 0.2077 0.9341
```

```
> lrtest(mod2, mod1)
Likelihood ratio test
```

```
Model 1: weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + ForearmG +
KneeG + CalfMaxG + height
```

```
Model 2: weight ~ ShoulderG + ChestG + WaistG + NavelG + HipG + ThighG +
BicepG + ForearmG + KneeG + CalfMaxG + AnkleMinG + WristMinG +
height
```

```
#Df LogLik Df Chisq Pr(>Chisq)
1 11 -1113.3
2 15 -1112.9  4 0.8538    0.9311
```

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

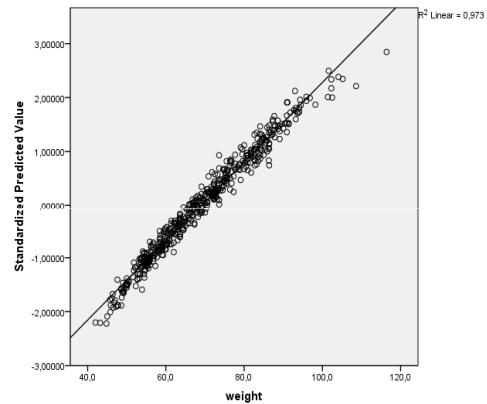
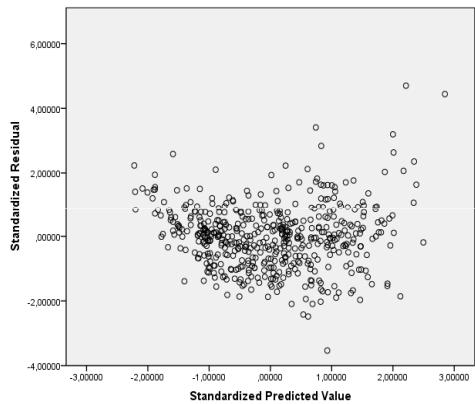
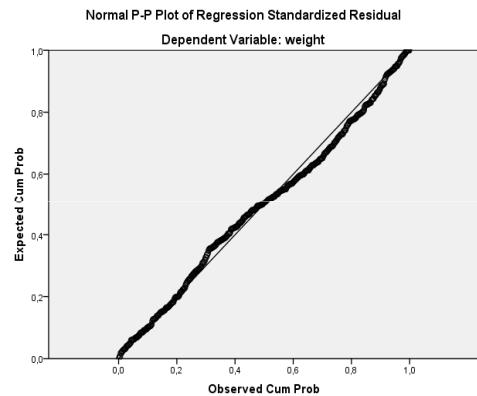
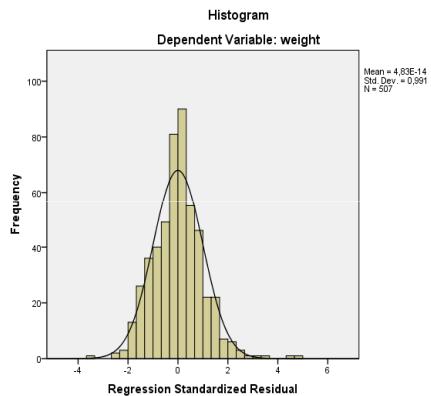
Page 92 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)


[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 93 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

2.2. Exemplo: funções respiratórias e tabaco

O ficheiro FEV.sav ^a contém dados recolhidos entre 1975 e 1980 respeitantes a 654 crianças e adolescentes da área de Boston Oriental. O objectivo do estudo consistiu da avaliação das funções pulmonares na presença ou ausência de exposição a fumo de cigarros (quer por exposição activa do próprio por fumar, quer por exposição passiva por contacto com um progenitor fumador).

O ficheiro contém as seguintes variáveis:

- age: idade (anos)
- fev: volume expiratório forçado (litros)
- ht: altura (polegadas)
- sex: sexo (0-rapariga; 1-rapaz)
- smoke: fuma cigarros de forma regular? resposta auto-declarada: 0- não; 1-sim. ^b

Tendo em conta o objectivo da análise, começamos por descrever numericamente e graficamente o volume expiratório forçado observado no grupo dos fumadores e dos não fumadores.

O comando

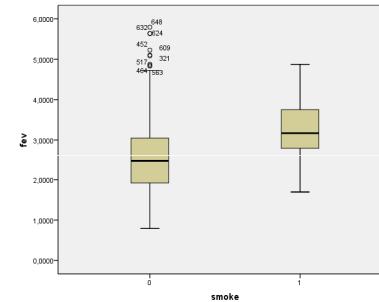
Analyze → Descript Stats → Explore

com a separação de fev por sexo evidencia uma distribuição assimétrica para os não fumadores, rejeitando a hipótese de normalidade, e uma distribuição relativamente simétrica para os fumadores.

^a Rosner, B. (1999), Fundamentals of Biostatistics, 5th Ed., Pacific Grove, CA: Duxbury

^b esta variável diz apenas respeito à exposição activa, apesar de outros dados terem sido também levantados no estudo.

Var.	Smoke	n	min	max
fev	0	589	0.7910	5.7930
	1	65	1.6940	4.8720
Var	Smoke	$q_{0.25}$	mediana	$q_{0.75}$
	fev	0	1.9195	2.4650
		1	2.7770	3.1690
			3.7680	



Os resultados sugerem que, de uma forma geral, os fumadores apresentem valores superiores de fev (i.e. melhores funções pulmonares) do que os não fumadores!

Mas:

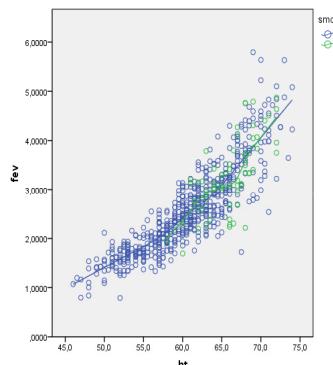
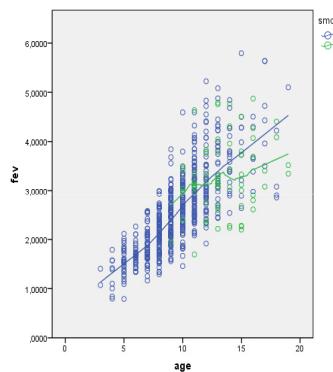
- a classificação em fumador vs não-fumador é auto-declarada...
- é sabido que os valores de fev dependem da estrutura corporal e a análise anterior não teve isso em consideração...

Tentamos então obter gráficos mais elucidativos sobre a questão, ajustados para variáveis que são julgadas de interesse.

O gráfico abaixo obteve-se das instruções usuais de

Graphs → Chart Builder → Scatter/Dot

com separação das observações, por cor, de acordo com a classe de smoke. As curvas representadas correspondem a um modelo de regressão não linear local. Obtém-se clicando sobre a figura, no output, escolhendo depois o ícone que representa duas retas de regressão diferentes ajustadas às observações e considerando a opção loess.



Repare-se que agora há conclusões diferentes a retirar, sendo que algumas já parecem ir ao encontro daquilo que se esperava.

- a função pulmonar cresce com a idade de uma forma aproximadamente linear; para os jovens a partir dos 14 anos, aproximadamente, parece existir uma distinção entre os valores de fev entre fumadores e não-fumadores, sendo que agora a relação de ordem parece ser a correcta.

Esta questão ultrapassa a análise aqui apresentada mas **os resultados estão a sugerir a existência de uma interacção age*smoke.**

Nota: Diz-se que existe **interacção** $X_1 * X_2$ entre duas variáveis explicativas X_1 e X_2 quando o efeito de uma delas sobre a resposta depende do valor da outra variável.

Havendo interacção, os efeitos principais de cada uma das variáveis não devem ser retirados do modelo de regressão, i.e., devemos considerar

$$Y \sim X_1 + X_2 + X_1 * X_2 + \dots$$

- a função pulmonar cresce com a altura de uma forma que parece quadrática mas não há indicação de que essa relação esteja a ser influenciada pelo facto de o indivíduo ser ou não fumador.

Outra questão que se pode levantar: **terá interesse considerar crianças com menos de, por exemplo, 6 anos?!** Se por um lado essa crianças são num certo sentido irrelevantes para qualquer avaliação do efeito de fumar, também é verdade que os seus dados podem contribuir para uma melhor explicação do fev em função das outras medidas recolhidas...

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

[Page 95 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Apesar de já se ter visto que a regressão

$$\text{fev} = \beta_0 + \beta_1 \text{smoke} + u, \quad u \sim N(0, \sigma^2 \text{Id})$$

não vai conduzir a resultados concordantes com o expectável, a análise dessa equação vai ser aqui apresentada de forma muito rápida, por motivos pedagógicos.

A variável *smoke* é um factor com duas categorias (portanto coincide com a dummy que lhe está associada), sendo a classe dos não fumadores a classe de referência.

Os resultados para o ajustamento do modelo aos dados são os seguintes

Modelo	coef	s.e. coef	t	valor-p
constante	2.566	0.035	74.037	0.000
smoke	0.711	0.110	6.464	0.000

$$F = 41.789, p = 0.000; R^2 = 0.060, \bar{R}^2 = 0.059$$

(Por o modelo não ter interesse, não avançamos sequer para as análises gráficas e numéricas de verificação dos pressupostos do modelo.)

Supondo que as condições do modelo eram satisfeitas, a tabela permitiria concluir:

- a existência de diferenças significativas para o fev de acordo com a classe de fumador (*observar que a estatística de teste da tabela coincide com a estatística de teste de um teste de comparação de médias em amostras independentes assumindo variâncias iguais - exercício*)

- o valor médio de fev para um indivíduo fumador é de $2.566 + 0.711$ enquanto que o correspondente valor médio para um indivíduo não fumador é de 2.566. Em média, um fumador tem um volume expiratório superior em 0.711 litros a um não fumador. Um intervalo a 95% de confiança para esta diferença é aproximadamente $0.7 \pm 2(0.1)$, isto é, (0.5, 0.9).

- caso a variável *smoker* tivesse mais de duas categorias, o que esta análise de regressão permitiria analisar era a **comparação entre os vários volumes expiratórios médios para as diferentes classes de smoker** (ANOVA - análise da variação).

Os comandos em **SPSS** correspondentes à tabela apresentada acima são os seguintes:

```
Analyze → Regression → Linear
Dependent: fev
Independent(s): smoke
```

deixando todas as opções por defeito que o software tem incluídas.

De acordo com os últimos gráficos considerados, analisamos agora a regressão do volume expiratório contra *smoke*, ajustada para a idade e a altura.

$$\text{fev} \sim \text{smoke} + \text{age} + \text{height} + \text{height}^2. ^a$$

^a Um dos gráficos anteriores sugere uma dependência quadrática do volume expiratório em relação à altura. Nestas situações, é usual considerarem-se ambas as variáveis "altura" e "altura"².

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀](#) [▶](#)
[◀](#) [▶](#)
[Page 96 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

*Output1 [Document1] - PASW Statistics Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,890 ^a	,791	,790	,3972925	,791	615,303	4	649	,000

a. Predictors: (Constant), ht.sq, smoke, age, ht
 b. Dependent Variable: fev

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	388,481	4	97,120	615,303	,000 ^a
Residual	102,439	649	,158		
Total	490,920	653			

a. Predictors: (Constant), ht.sq, smoke, age, ht
 b. Dependent Variable: fev

Coefficients^a

Model	Unstandardized Coefficients			t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	7,868	1,469		5,356	,000		
smoke	-,150	,057	-,052	-,2,635	,009	,827	1,210
age	,067	,009	,227	7,313	,000	,334	2,992
ht	-,307	,049	-2,019	-6,310	,000	,003	318,399
ht.sq	,003	,000	2,726	8,589	,000	,003	313,366

a. Dependent Variable: fev

PASW Statistics Processor is ready

As variáveis explicativas consideradas são todas significativas para o modelo. Pela observação da magnitude dos coeficientes estandardizados observamos que a altura parece ser a variável que mais influencia o volume expiratório. Em relação ao factor smoke, observamos que, para indivíduos da mesma idade e com a mesma altura, o volume expiratório dos fumadores é, em média, 0.15 litros mais baixo do que o dos não fumadores. Um intervalo a 95% de confiança para esta diferença entre os volumes médios, condicionada pela idade e pela altura, é aproximadamente

$$(0.150 - 2(0.057), 0.150 + 2(0.057)) = (-0.264, -0.036).$$

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀](#) [▶](#)
[◀](#) [▶](#)
[Page 97 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 98 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Acresce que o modelo na generalidade é significativo ($F=615.303$, $p=0.000$), com um coeficiente de determinação $R^2 = 0.791$ que é significativamente diferente de zero e um coeficiente de determinação ajustado de $\bar{R}^2 = 0.790$.

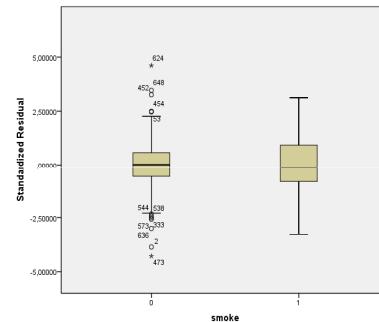
Nas instruções efectuadas em SPSS pedimos que fossem guardados os objectos: resíduos estandardizados, valores ajustados estandardizados, leverages e distâncias de Cook. Seleccionámos ainda a identificação das observações com resíduos estandardizados superiores a 3.3, fazendo

Statistics → Residuals → Casewise Diagnostics

e incluindo o valor de referência de 3.3 desvios padrão. O output respeitante aos resíduos é apresentado na página seguinte.

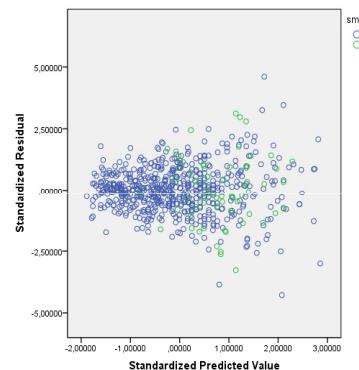
As conclusões anteriores só são válidas uma vez satisfeitos os pressupostos do modelo. No que se segue, analisamos graficamente a satisfação desse pressuposto para cada uma das classes de smoke.

Poderíamos começar por considerar o gráfico de dispersão dos resíduos estandardizados contra a ordem das observações, quer para detecção de linearidade e homocedasticidade, quer para a eventual detecção de outliers. Contudo, esse gráfico exige a criação de um variável com a indexação dos indivíduos, que não existe no ficheiro, e portanto não vai ser aqui considerado. De qualquer forma podemos observar o que se passa com o boxplot dos resíduos estandardizados para cada uma das categorias de smoke



Existem 4 indivíduos com resíduos superiores a 3.3 sendo que dois deles têm resíduos grandes negativos e outros têm resíduos grandes positivos. A remoção destas observações do modelo depende de uma análise crítica dos valores de todas as variáveis recolhidas para essas observações.

O gráfico a considerar de seguida é o gráfico dos resíduos estandardizados (ZRESID) contra os valores ajustados estandardizados (ZPRED), por categoria de smoke.



*Output1 [Document1] - PASW Statistics Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help

Descriptives
Tests of Normality
Standardized Residuals
Histograms
Normal Q-Q Plots
Detrended Fit
Boxplot
Log
CGraph
Title
Notes
Active Dataset
Graph
Regression
Title
Notes
Active Dataset
Variables Entered/Removed
Model Summary
ANOVA
Coefficients
Casewise Diagnostics
Residuals Statistics

Casewise Diagnostics^a

Case Number	Std. Residual	fev	Predicted Value	Residual
2	-3,856	1,7240	3,26609	-1,5320098
473	-4,279	2,5380	4,238029	-1,7000294
624	4,810	5,7930	3,981472	1,8315283
648	3,461	5,6380	4,262780	1,3752196

a. Dependent Variable: fev

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1,181591	4,838557	2,636780	,7713086	654
Std. Predicted Value	-1,887	2,855	,000	1,000	654
Standard Error of Predicted Value	,021	,083	,032	,013	654
Adjusted Predicted Value	1,186635	4,877048	2,638666	,7712649	654
Residual	-1,7000294	1,8315283	,00000000	,3960738	654
Std. Residual	-4,279	4,610	,000	,997	654
Stud. Residual	-4,309	4,633	,000	1,002	654
Deleted Residual	-1,7236316	1,8500280	-,0000866	,4000394	654
Stud. Deleted Residual	-4,368	4,708	,000	1,005	654
Mahal. Distance	,750	27,739	3,994	4,652	654
Cook's Distance	,000	,060	,002	,006	654
Centered Leverage Value	,001	,042	,006	,007	654

a. Dependent Variable: fev

PASW Statistics Processor is ready

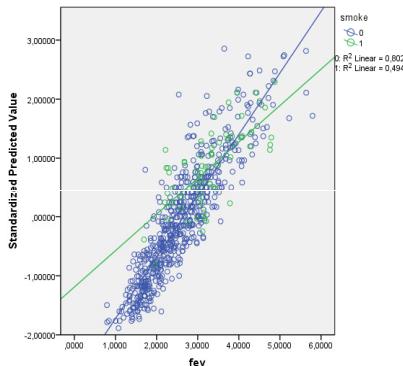
[Home Page](#)[Title Page](#)[Contents](#)[◀◀](#) [▶▶](#)[◀](#) [▶](#)

Page 99 of 104

[Go Back](#)[Full Screen](#)[Close](#)[Quit](#)

Parece existir uma tendência para um aumento da variância dos resíduos com um aumento dos valores ajustados, mais nos não fumadores do que nos fumadores, o que poderá violar a hipótese de homocedasticidade dos resíduos. Eventualmente poder-se-ia transformar a resposta e ver se isso resolvia esta questão.

O gráfico dos valores ajustados contra a resposta apresentam uma relação linear para cada uma das categorias de smoke e mais uma vez sugerem uma pequena violação do pressuposto de homocedasticidade.



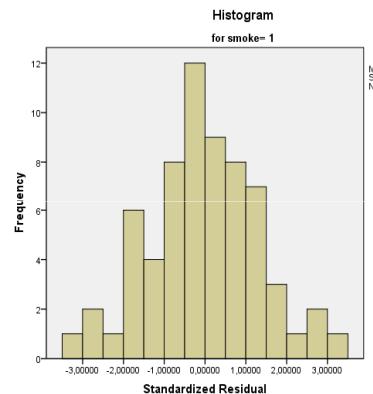
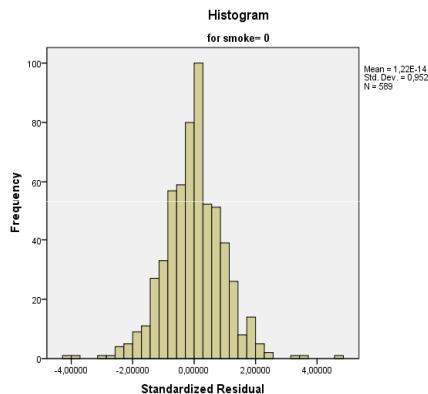
Os gráficos usuais de averiguação de normalidade são obtidos do comando

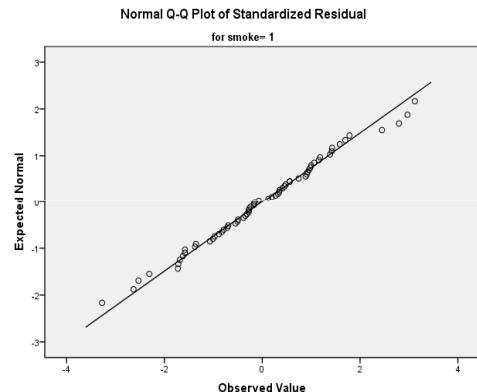
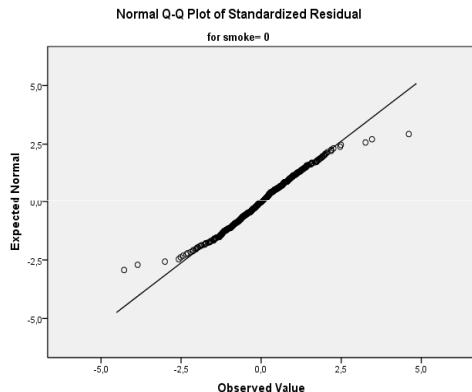
Analyze → Descriptive Stats → Explore

escolhendo os resíduos estandardizados com variável dependente e smoke como factor (página seguinte)

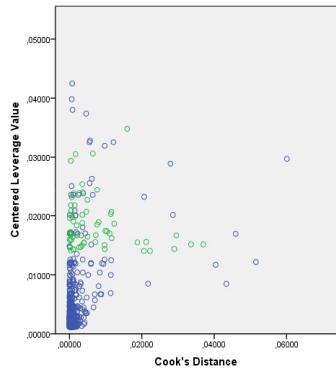
Observa-se a existência de vários outliers para os resíduos no grupo dos não fumadores o que inevitavelmente acaba por comprometer a normalidade dos resíduos neste grupo mas notamos simultaneamente que o tamanho amostral é grande e a distribuição não é marcadamente assimétrica pelo que, na verdade, não parecem existir problemas com este aspecto. Quanto aos resíduos no grupo dos fumadores, não são identificadas violações de normalidade.

[Home Page](#)
[Title Page](#)
[Contents](#)
[◀◀](#) [▶▶](#)
[◀](#) [▶](#)
[Page 100 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)


[Home Page](#)
[Title Page](#)
[Contents](#)

[Page 101 of 104](#)
[Go Back](#)
[Full Screen](#)
[Close](#)
[Quit](#)


Finalmente fazemos a detecção de pontos influentes considerando o gráfico das leverages contra as distâncias de Cook (dispondo de uma variável índice para os indivíduos, consideraríamos também os gráficos de dispersão de cada uma destas medidas contra o índice).



Os "pontos de corte" para as leverages e distâncias de Cook são $2(p + 1)/n$ e 1, respectivamente, onde p representa o número de variáveis explicativas do modelo e n o tamanho amostral. Ora, para cada nível do factor smoke, existem 3 variáveis no modelo portanto teremos de analisar com cuidado pontos com leverages superiores a $2(3 + 1)/589 = 0.014$ nos não-fumadores e $2(3 + 1)/65 = 0.123$ nos fumadores. De entre esses pontos, fortes candidatos a pontos influentes são aqueles que apresentarem também distância de Cook superior a 1.

Não há portanto observações especiais a considerar.

Outro modelo a considerar poderia ser

$$\text{fev} \sim \text{smoke} + \text{age} + \text{height} + \text{height}^2 + \text{sex}$$

fazendo uso de todas as variáveis explicativas disponibilizadas.

Os resultados para o ajustamento do modelo aos dados são os seguintes

Modelo	coef	s.e. coef	<i>t</i>	valor- <i>p</i>
constante	6.895	1.499	4.60	0.000
age	0.069	0.009	7.63	0.000
sex	0.095	0.033	2.88	0.004
ht	-0.274	0.050	-5.52	0.000
ht.sq	0.003	0.000	7.65	0.000
smoke	-0.133	0.057	-2.33	0.020

$$F = 499.416, p = 0.000; R^2 = 0.794 \quad (p = 0.000), \\ \bar{R}^2 = 0.792$$

O modelo é muito semelhante ao anterior em termos de interpretação. Quando em presença também do sexo, o efeito do tabaco sobre o volume expiratório é essencialmente o mesmo que no caso anterior, quer em termos de magnitude quer em termos de variância. O mesmo se passa em relação à percentagem da variância total explicada pela regressão.

As análises gráficas desta equação de regressão são deixadas como exercício, sendo que os 4 outliers identificados na situação anterior mantêm-se também neste modelo.

Entre os dois modelos o anterior parece ser preferível, pela simplicidade da apresentação.

Outros modelos com interacções de variáveis poderiam também ser considerados.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

[Page 102 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

2.3. Exemplo: resultados eleitorais na Georgia (EUA) nas eleições presidenciais de 2000

- (a) Leia a biblioteca *faraway* no R. Se não a tiver instalada, terá de o fazer previamente. Esta biblioteca contém vários ficheiros de dados, a maior parte usados no livro *Extending the Linear Model with R*, de J.J. Faraway.
- (b) Considere as instruções

```
> data(gavote)
> help(gavote)
```

A primeira lê um ficheiro designado por *gavote*; a segunda fornece elementos descritivos das variáveis que constam no ficheiro.

A coluna *votes* representa a totalidade de votos úteis, e a coluna *ballots* representa a totalidade de boletins de voto preenchidos (dos quais nem todos resultaram num voto útil). Mais precisamente, um eleitor dirige-se à sua secção de voto, onde se identifica a sua legitimidade para a votação. Estando recenseado, é emitido um boletim de voto. Contudo, votos em branco e votos nulos não são considerados. Por vezes, o equipamento que lê e guarda os resultados da votação também tem falhas. A diferença

$\text{ballots} - \text{votes}$

é designada por *undercount* (digamos, sub-contagem).

O objectivo deste exercício consiste da determinação dos factores que afectam a sub-contagem.

- (c) Efectue uma análise estatística descritiva dos dados, usando gráficos e medidas estatísticas numéricas adequados. Observe que o comando

```
> plot(density(gavote$votes), main="Votes")
> rug(gavote$votes)
```

produz um gráfico que pode ser visto como uma suavização do histograma, e portanto como um bom complemento deste último. O "tapete" que aparece no fundo do gráfico indica a localização das várias observações recolhidas.

Consegue identificar alguns problemas com os dados que possam vir a dificultar a análise posterior? Defina também a variável resposta.

- (d) Encontre um modelo para a sub-contagem que lhe pareça adequado ao problema, e estude a sua validade e qualidade do ajustamento. Interprete também os coeficientes de regressão obtidos no modelo anterior.

[Home Page](#)

[Title Page](#)

[Contents](#)

◀ ▶

◀ ▶

Page 103 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

2.4. Exercícios

1. Determine explicitamente o valor dos estimadores usuais para β e σ nas seguintes situações ($i = 1, \dots, n$):

- (a) $y_i = \beta_0 + u_i$, $u_i \sim N(0, \sigma^2)$
- (b) $y_i = \beta_0 + \beta_1 x_i + u_i$, $u_i \sim N(0, \sigma^2)$

Sugestão: reescreva a equação com a variável explicativa centrada.

2. Mostre que, num modelo de regressão linear com termo constante, se tem:

- (a) $E(\hat{u}) = 0$
- (b) $X^t \hat{u} = 0$
- (c) A soma dos valores previstos é igual à soma dos valores observados
- (d) O ponto (\bar{x}, \bar{y}) satisfaz a equação de regressão.
- (e) $H(1 - H) = 0$, onde H representa a matriz chapéu.
- (f) Os valores previstos e os resíduos são ortogonais.

3. Mostre que o estimador de máxima verosimilhança $\hat{\beta}$ de β é um minimizante da soma dos quadrados de resíduos, mostrando que, para um qualquer outro vector $b \in \mathbb{R}^p$ se tem

$$(y - Xb)^t(y - Xb) \geq (y - X\hat{\beta})^t(y - X\hat{\beta}).$$

4. Mostre que, para qualquer matriz real $A \in \mathbb{R}^{n \times n}$, o produto $A^t A$ é uma matriz semi-definida positiva.

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀](#) [▶](#)

[◀](#) [▶](#)

[Page 104 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)