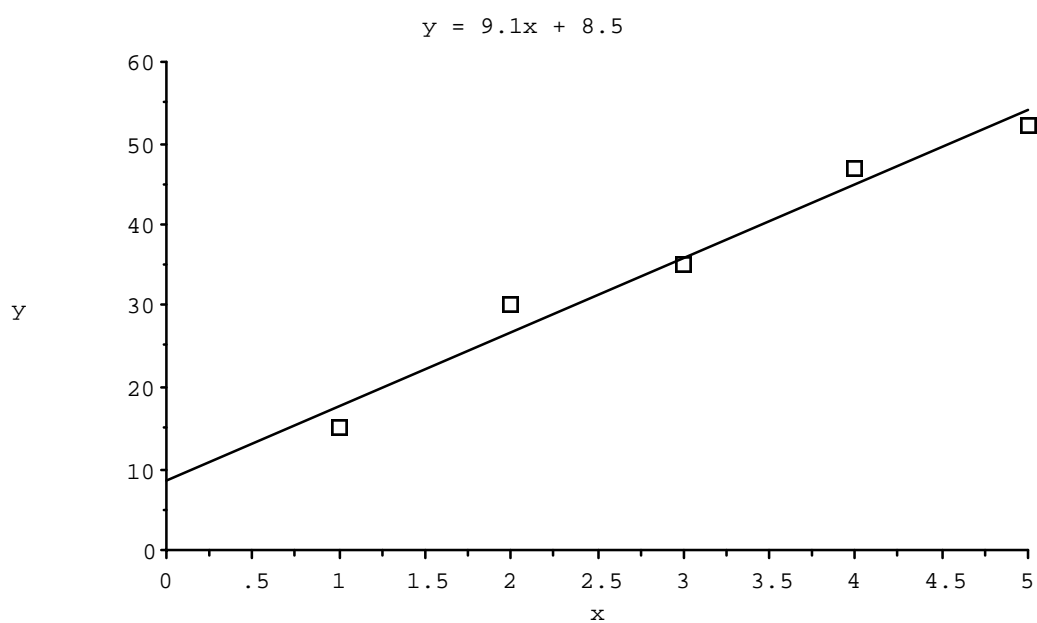


# Manual Operacional para a Regressão Linear

*Manuel António Matos*



FEUP 1995

# Índice

<b>1. Introdução</b>	<b>3</b>
<b>2. Preliminares</b>	<b>3</b>
2.1. Convenções.....	3
2.2. Modelo da regressão linear.....	3
2.3. Pressupostos .....	5
2.4. Médias e variâncias .....	5
<b>3. Modelização</b>	<b>6</b>
3.1. Variáveis não-numéricas.....	6
3.2. Interacções.....	7
3.3. Suavizar .....	7
3.4. Variáveis ortogonais .....	7
<b>4. Transformações dos dados</b>	<b>8</b>
4.1. Centragem.....	8
4.2. Estandarização.....	8
4.3. Norma unitária.....	8
<b>5. Estimação de parâmetros</b>	<b>9</b>
5.1. Passagem obrigatória na origem.....	9
5.2. Interpretação dos parâmetros estimados .....	9
<b>6. Avaliação da qualidade do ajuste</b>	<b>10</b>
6.1. Erro quadrático .....	10
6.2. Variância do erro .....	10
6.3. Coeficiente de Determinação.....	10
6.4. ANOVA.....	11
<b>7. Testes e intervalos de confiança</b>	<b>11</b>
7.1. Distribuições de probabilidade dos parâmetros .....	12
7.2. Correlações e Matriz Covariância .....	12
7.3. Testes de hipóteses .....	13
7.4. Intervalos de confiança dos parâmetros.....	14
7.5. Intervalo da resposta.....	14
7.6. Intervalo de predição .....	15
<b>8. Análise de resíduos</b>	<b>15</b>
8.1. Eliminação de observações.....	16
8.2. Verificação de pressupostos .....	16
8.3. Expressão do modelo.....	18
<b>9. Selecção de variáveis</b>	<b>18</b>
9.1. Medida $F$ e estatística $C_k$ .....	18
9.2. Pesquisa t-dirigida .....	19
9.3. Selecção para a frente ( <i>forward selection</i> ).....	19
9.4. Eliminação para trás ( <i>backward elimination</i> ).....	20
9.5. Procedimento passo a passo .....	21
<b>10. Multicolinearidade</b>	<b>21</b>
10.1. Detecção .....	22
10.2. Regressão de componentes principais .....	22
<b>Bibliografia</b>	<b>26</b>

## 1. Introdução

Este texto destina-se à utilização no ensino e investigação, como manual operacional para a regressão linear. Nesse sentido, prescinde-se de qualquer tipo de desenvolvimento teórico, e centra-se a atenção nos procedimentos a efectuar para realizar um exercício completo de regressão, ou seja, incluindo alguns cuidados com a selecção de variáveis e a validação de resultados (testes de hipóteses e intervalos de confiança), nem sempre tidos em conta na prática corrente.

Os leitores interessados em maior profundidade teórica ou em aspectos adicionais podem recorrer à lista bibliográfica indicada no fim do texto, ou a qualquer livro sobre esta matéria. O primeiro livro da lista serviu de base a muito do presente texto, cuja organização e redacção são, no entanto, bastante diferentes, dados os seus objectivos.

Procurou-se ser o mais conciso possível, partindo do princípio que o utilizador já teve algum contacto com esta técnica. No entanto, o texto também pode ser usado por quem não tenha tido qualquer contacto prévio com a regressão.

## 2. Preliminares

### 2.1. Convenções

As variáveis são designadas por letras maiúsculas em itálico ( $Y$ ,  $X_k$ ), o mesmo se passando com as suas médias ( $\bar{Y}$ ,  $\bar{X}_k$ ). Os vectores são representados em letra minúscula carregada ( $\mathbf{y}$ ,  $\mathbf{x}_k$ ), e os seus elementos em itálico e letra minúscula ( $y_i$ ,  $x_{ik}$ ). Usam-se letras maiúsculas carregadas para as matrizes ( $\mathbf{X}$ ,  $\mathbf{M}$ ,  $\mathbf{W}$ ), sendo os seus elementos representados como os dos vectores. O acento circunflexo é usado para indicar que se trata de valores estimados ( $\hat{\sigma}$ ,  $\hat{\mathbf{y}}$ ). A transposição de vectores e matrizes é indicada por uma plica ( $\mathbf{y}'$ ,  $\mathbf{M}'$ ). Constantes e outros valores matemáticos são apresentados em letra minúscula e itálico. Outras convenções são referidas à medida que aparecem no texto.

As referências a distribuições estatísticas são feitas do modo usual, indicando-se entre parêntesis ou em índice os parâmetros necessários. Por exemplo,  $N(\mu, \sigma)$  designa uma distribuição normal com média  $\mu$  e variância  $\sigma^2$ , enquanto que  $t_{1-\gamma/2}(n-p-1)$  indica o valor crítico da distribuição t de Student com  $n-p-1$  graus de liberdade e nível de significância  $\gamma$  (teste bilateral).

### 2.2. Modelo da regressão linear

A regressão nasce da tentativa de relacionar um conjunto de observações de certas variáveis,

designadas genericamente por  $X_k$  ( $k=1..p$ ), com as leituras de uma certa grandeza  $Y$ . No caso da regressão linear, está subjacente uma relação do tipo:

$$Y = a + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

onde  $a, b_1, b_2, \dots, b_p$  seriam os parâmetros da relação linear procurada. O objectivo pode ser explicativo (demonstrar uma relação matemática que pode indicar, *mas não prova*, uma relação de causa-efeito) ou preditivo (obter uma relação que nos permita, perante futuras observações das variáveis  $X_k$ , *prever* o correspondente valor de  $Y$ , sem necessidade de o medir). Dadas as características deste texto, não se aprofundará esta questão, mas a distinção básica entre as duas situações é fundamental. Independentemente dos objectivos, as variáveis  $X_k$  são muitas vezes designadas por variáveis explicativas, uma vez que tentam *explicar* as razões da variação de  $Y$ .

Supondo que se dispõe de  $n$  conjuntos de medidas com as correspondentes observações, a utilização do modelo incluirá sempre uma parcela de erro. Utilizando o índice  $i$  ( $i=1..n$ ) para indicar cada conjunto, ter-se-á então:

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \dots + b_p x_{ip} + e_i \quad i=1..n$$

Independentemente das motivações, a versão da regressão linear que aqui se apresenta consiste em estimar os valores dos parâmetros  $a, b_1, b_2, \dots, b_p$ , através da minimização da soma dos quadrados dos desvios. Daí o nome de *método dos mínimos quadrados* que às vezes se utiliza, nomeadamente para a *regressão simples* ( $p=1$ ). O termo *multi-regressão* é usado para explicitar o caso  $p>1$ .

Neste ponto, é conveniente definir:

$\mathbf{y} = [y_1 \ y_2 \ \dots \ y_n]'$	vector das leituras
$\mathbf{x}_k = [x_{1k} \ x_{2k} \ \dots \ x_{nk}]'$	vector das observações de cada variável $X_k$
$\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_p]$	matriz de observações (elementos $x_{ik}$ , $i=1..n$ , $k=1..p$ )
$\mathbf{b} = [a \ \mathbf{b}_0]' = [a \ b_1 \ b_2 \ \dots \ b_p]'$	vector dos parâmetros
$\mathbf{e} = [e_1 \ e_2 \ \dots \ e_n]'$	vector dos erros
$\mathbf{1} = [1 \ \dots \ 1]'$	vector unitário de dimensão $n$
$\mathbf{X}_a = [\mathbf{1} \ \mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_p]$	matriz aumentada de observações ( $i=0..n$ , $k=0..p$ )

Com estas definições, é possível escrever a expressão anterior de forma compacta:

$$\mathbf{y} = a \cdot \mathbf{1} + \mathbf{X} \cdot \mathbf{b}_0 + \mathbf{e}$$

ou

$$\mathbf{y} = \mathbf{X}_a \cdot \mathbf{b} + \mathbf{e}$$

Uma vez obtida a estimativa  $\hat{\mathbf{b}}$  dos parâmetros  $\mathbf{b}$ , a expressão operacional da regressão permite obter estimativas  $\hat{y}$  das leituras correspondentes às observações  $x_1 x_2 \dots x_p$ :

$$\hat{y} = \hat{a} + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \dots + \hat{b}_p x_p$$

ou

$$\hat{y} = \hat{a} + \mathbf{x}' \cdot \hat{\mathbf{b}}_0$$

O cálculo simultâneo dos valores estimados correspondentes às observações usadas na parametrização da regressão pode ser feito com base na expressão matricial correspondente:

$$\hat{\mathbf{y}} = \mathbf{X}_a \cdot \hat{\mathbf{b}}$$

Nas restantes secções deste texto, dão-se indicações sobre a selecção de variáveis, obtenção de estimativas dos parâmetros e interpretação e validação de resultados.

### 2.3. Pressupostos

Para além de pressupostos gerais acerca da correcta especificação do modelo e da medição sem erros das variáveis observadas, um pressuposto importante para todo o desenvolvimento é de que os erros do modelo  $\mathbf{e}$  têm média nula, não estão correlacionados e têm variância constante  $\sigma$ . Se estes pressupostos não forem verificados, muitas das expressões utilizadas neste texto podem deixar de fazer sentido, pois foram deduzidas a partir dessa hipótese.

Uma condição adicional para os erros do modelo é de que estejam normalmente distribuídos. Não sendo essencial para a derivação das expressões de cálculo das estimativas dos parâmetros, este pressuposto é indispensável para toda a matéria respeitante a testes de hipóteses e derivação de intervalos de confiança e, em geral, para toda a validação estatística dos resultados.

Para contemplar esse importante aspecto, este texto inclui alguns procedimentos de verificação, *a posteriori*, dos pressupostos respeitantes aos erros do modelo (análise de resíduos).

### 2.4. Médias e variâncias

Sobretudo com o intuito de fixar notações, recordam-se, a seguir, algumas definições que

são utilizadas no resto do texto.

Define-se a média de uma variável através de:

$$\bar{X}_k = \frac{\sum_{i=1}^n x_{ik}}{n} \text{ (observações)} \quad \text{ou} \quad \bar{Y} = \frac{\sum_{i=1}^n y_i}{n} \text{ (leituras)}$$

Note-se que, em rigor, as expressões anteriores referem-se à média *amostral*, que é uma estimativa não tendenciosa da média das variáveis. Com esse facto em conta, utilizar-se-ão estas designações no texto, por não haver possibilidade de confusão, uma vez que as médias populacionais não são acessíveis.

Por outro lado, definindo

$$d_k^2 = \sum_{i=1}^n (x_{ik} - \bar{X}_k)^2$$

soma dos quadrados dos desvios em relação à média de  $X_k$ , a estimativa não tendenciosa da variância de  $X_k$  é dada por:

$$s_k^2 = \frac{d_k^2}{n-1} = \frac{\sum_{i=1}^n (x_{ik} - \bar{X}_k)^2}{n-1}$$

De forma análoga se calcularia a estimativa da variância de  $Y$ .

### 3. Modelização

#### 3.1. Variáveis não-numéricas

A inclusão de categorias no modelo da regressão deve fazer-se recorrendo a variáveis binárias. No caso mais simples, em que há duas categorias (A e B), cria-se um variável  $X_1$ , com dois valores possíveis, correspondendo cada um a uma das categorias. Os dois valores costumam ser 0 e 1, mas pode ser usado qualquer par de números. Se houver  $c$  categorias, deverão criar-se as variáveis binárias necessárias para definir todas as categorias. O número de variáveis a criar é o inteiro imediatamente superior (ou igual) a  $\log_2 c$ . Por exemplo, para 3 categorias A, B e C, poderão criar-se 2 variáveis  $X_1$  e  $X_2$ , definidas como na tabela 1:

**Tabela 1: Variáveis binárias para 3 categorias**

	A	B	C
$X_1$	1	0	0
$X_2$	0	1	0

Um erro frequente consiste em usar variáveis com mais de dois valores, o que institui uma ordem *a priori* e uma relação fixa entre classes. No caso do exemplo, seria portanto errado usar apenas uma variável que tomasse os valores (0, 1, 2) para as três classes.

### 3.2. Interacções

Os efeitos conjuntos de variáveis podem ter de ser incluídos no modelo linear, se elas não forem independentes. No caso de variáveis numéricas, o gráfico de  $Y$  vs  $X_1, X_2$  deve ser linear, se o termo for de incluir. No caso de  $X_1$  representar uma categoria, sendo  $X_2$  uma variável numérica, os gráficos de  $Y$  vs  $X_2$  para diversos valores de  $X_1$  devem ter distintas inclinações e ordenadas na origem, se a interacção entre as duas variáveis for importante.

### 3.3. Suavizar

Para facilitar a visualização as tendências dos dados, nomeadamente em gráficos, podem usar-se mecanismos de suavização de irregularidades em dados ordenados, como médias móveis ou medianas de três pontos. A regularização por médias móveis consiste em substituir cada ponto  $(x_i, y_i)$  por  $(x_i, z_i)$ , onde  $z_i$  é a média dos valores de  $Y$  nos 3 ou 5 pontos centrados em  $(x_i, y_i)$ , por exemplo  $z_i = (y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2})/5$ , sendo ignorados, neste caso, necessariamente os dois primeiros e dois últimos pontos da lista, previamente ordenada pelos valores de  $X$ . Na utilização da mediana de 3 pontos, substitui-se cada valor de  $y_i$  pela mediana de  $(y_{i-1}, y_i, y_{i+1})$ , repetindo-se o processo até estabilizar. Neste caso, os pontos extremos da lista inicial, ordenada pelos valores de  $X$ , mantêm-se fixos ao longo do processo.

### 3.4. Variáveis ortogonais

Há vantagem em que o maior número possível de variáveis sejam ortogonais, pois permite simplificações e separabilidade no cálculo. Recordar-se que duas variáveis  $X_u$  e  $X_v$  são ortogonais se  $\mathbf{x}'_u \cdot \mathbf{x}_v = 0$ . Note-se, por outro lado, que as variáveis que representam categorias não são obrigatoriamente ortogonais. No caso do exemplo da tabela 1,  $X_1$  e  $X_2$  são ortogonais, mas o mesmo não se passaria se fossem definidas como na tabela 2:

**Tabela 2: Variáveis binárias não-ortogonais**

	A	B	C
$X_1$	0	1	1
$X_2$	0	0	1

## 4. Transformações dos dados

Em alternativa ao uso das variáveis originais ("raw"), podem ser usadas variáveis centradas ("centered"), estandardizadas ("standardized") ou com norma unitária ("unit length"), obtidas através das transformações indicadas a seguir. Todos estes procedimentos visam compatibilizar, de algum modo, variáveis que podem ter escalas e dispersões muito diferentes. Em particular, a comparação da influência relativa das diversas variáveis, com base nos parâmetros estimados, só faz sentido se as variáveis forem normalizadas.

Como se verá noutra local deste texto, os resultados obtidos depois de qualquer das transformações que se descrevem a seguir são sempre iguais aos da versão com os dados originais. Também os parâmetros têm relações simples entre si, permitindo passar facilmente de uma formulação a outra.

### 4.1. Centragem

Uma transformação simples consiste em centrar cada variável em relação à sua média. A variável transformada  $M_k$  obtém-se de  $X_k$  através de:

$$m_{ik} = x_{ik} - \bar{X}_k$$

Semelhantemente ao que se fez para  $\mathbf{X}$ , também aqui se define  $\mathbf{M}=[\mathbf{m}_1 \mathbf{m}_2 \dots \mathbf{m}_p]$ .

### 4.2. Estandarização

A estandarização corresponde a uma transformação para média nula e desvio padrão unitário de cada variável original  $X_k$ . A nova variável  $Z_k$  é obtida através de:

$$z_{ik} = \frac{x_{ik} - \bar{X}_k}{s_k}$$

Neste caso, define-se  $\mathbf{Z}=[\mathbf{z}_1 \mathbf{z}_2 \dots \mathbf{z}_p]$ .

### 4.3. Norma unitária

Esta transformação substitui os valores de cada variável  $X_k$  por uma nova variável  $W_k$ , obtida pela seguinte regra:

$$w_{ik} = \frac{x_{ik} - \bar{X}_k}{d_k}$$

Definindo aqui também  $\mathbf{W}=[\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_p]$ , verifica-se que a matriz  $\mathbf{W}'\mathbf{W}$  apresenta diagonal unitária (daí o nome da transformação). Os restantes elementos  $(\mathbf{W}'\mathbf{W})_{uv}$  correspondem à correlação entre  $X_u$  e  $X_v$ . Note-se ainda que  $\mathbf{Z}'\mathbf{Z}=(n-1) \mathbf{W}'\mathbf{W}$ .

## 5. Estimação de parâmetros

A estimativa não tendenciosa de  $\mathbf{b}$  pelo método dos mínimos quadrados é dada por:

$$\hat{\mathbf{b}} = (\mathbf{X}'_a \mathbf{X}_a)^{-1} \mathbf{X}'_a \mathbf{y}$$

No caso de variáveis centradas, estandardizadas ou de norma unitária, o processo de obtenção da estimativa dos parâmetros  $\mathbf{b}_0$  utiliza uma expressão análoga à anterior, substituindo-se  $\mathbf{X}_a$  respectivamente por  $\mathbf{M}$ ,  $\mathbf{Z}$  ou  $\mathbf{W}$ . A estimativa de  $a$  é, em todos esses casos, igual à média de  $Y$ . Os valores de  $\hat{b}_k$  obtidos se as variáveis forem centradas são iguais aos do caso geral. Para variáveis estandardizadas e de norma unitária, cada  $\hat{b}_k$  vem multiplicado respectivamente por  $s_k$  e  $d_k$  em relação ao caso geral. A menos de erros de arredondamento, os valores estimados com qualquer dos modelos são rigorosamente correspondentes.

### 5.1. Passagem obrigatória na origem

No caso de se pretender que o estimador passe pela origem,  $\hat{a}=0$ , e

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

Os resultados obtidos com esta imposição são sempre piores do que os do modelo geral.

### 5.2. Interpretação dos parâmetros estimados

Os coeficientes do modelo linear representam a variação na resposta prevista que resulta de uma variação de uma unidade no valor ajustado das respectivas variáveis. Se todas as variáveis forem ortogonais, o aumento de uma unidade em  $x_k$  (supondo as outras iguais) teria como resultado um aumento de  $\hat{b}_k$  em  $\hat{y}$ . No entanto, no caso geral das variáveis não serem ortogonais, não faz sentido variar só uma variável, pois aquelas que estão correlacionadas com ela também terão que variar. Em consequência, a variação de  $\hat{y}$  é afectada pelos coeficientes de correlação aplicáveis.

## 6. Avaliação da qualidade do ajuste

### 6.1. Erro quadrático

O valor minimizado do quadrado dos erros pode ser calculado através de:

$$\sum_i r_i^2 = \sum_i (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}' \cdot \mathbf{y} - \mathbf{y}' \cdot \mathbf{X}_a \cdot \hat{\mathbf{b}}$$

A simplificação no último membro da igualdade deve-se a ser  $\hat{\mathbf{y}}'(\mathbf{y} - \hat{\mathbf{y}}) = \hat{\mathbf{y}}' \cdot \mathbf{e} = 0$ .

### 6.2. Variância do erro

Supondo que os erros são independentes e têm a mesma variância, a estimativa da variância do erro é dada por:

$$\hat{\sigma}^2 = \frac{\sum_i r_i^2}{n - p - 1}$$

Em princípio, todas as futuras observações de  $Y$  estarão no intervalo  $\pm 3\sigma$  centrado no valor predito  $\hat{y}$ . Mais adiante se verá uma melhor definição de intervalos de confiança para  $\hat{y}$ .

### 6.3. Coeficiente de Determinação

A soma dos quadrados das observações pode ser decomposta em:

$$\sum_i y_i^2 = n \cdot \bar{Y}^2 + \sum_i (\hat{y}_i - \bar{Y})^2 + \sum_i r_i^2$$

ou

$$TSS = SSM + SSR + SSE$$

onde se usam as iniciais, em língua inglesa, de "Total Sum of Squares" (soma quadrática total), "Sum of Squares due to the Mean" (soma quadrática devida à média), "Sum of Squares due to the Regression" (soma quadrática devida à regressão) e "Sum of Squares due to the Error" (soma quadrática devida ao erro). À soma  $SSR+SSE$  chama-se "Adjusted Total Sum of Squares" (soma quadrática total ajustada), com a sigla inglesa  $TSS(adj)$ .

O coeficiente de determinação, usado como medida de qualidade do ajuste, é dado por:

$$R^2 = \frac{SSR}{SSR + SSE} = \frac{\sum_i (\hat{y}_i - \bar{Y})^2}{\sum_i (\hat{y}_i - \bar{Y})^2 + \sum_i r_i^2}$$

ou seja, o coeficiente mede a proporção da variação de  $Y$  em relação à média que é explicada pela regressão. Um resultado a reter é que  $R^2 = \rho_{Y\hat{Y}}^2$  (quadrado do coeficiente de correlação entre  $Y$  e  $\hat{Y}$ ). Em princípio, a qualidade do ajuste será tanto maior quanto mais  $R^2$  se aproximar da unidade.

#### 6.4. ANOVA

As tabelas de análise de variância ("Analysis Of Variance") são comuns em diversos tipos de estudos estatísticos, sendo frequentemente incorporadas nos programas dedicados à regressão e nas folhas de cálculo que incluem este tipo de estudos. A organização dos valores tem normalmente o aspecto indicado na tabela 3, onde são usadas algumas iniciais referidas no ponto anterior. As médias dos quadrados são obtidas dividindo as somas de quadrados pelos graus de liberdade correspondentes, como em  $MSE = SSE/(n-p-1)$ . Repare-se que  $MSE = \hat{\sigma}^2$ .

**Tabela 3: Quadro típico de ANOVA**

	Graus de liberdade	Soma dos quadrados	Média dos quadrados	F	R <sup>2</sup>
Média	1	SSM	MSM	MSM/MSE	SSR/(SSR+SSE)
Regressão	p	SSR	MSR	MSR/MSE	
Erro	n-p-1	SSE	MSE		
Total	n	TSS			

Os dois valores de  $F$  apresentados na tabela permitem realizar testes de nulidade dos parâmetros. O valor na linha da média é em geral muito elevado, não conduzindo a qualquer resultado com interesse; o valor na linha da regressão é usado no teste de  $\mathbf{b}_0 = \mathbf{0}$ , descrito noutra secção do presente texto. Alguns programas e folhas de cálculo incluem na tabela o valor da probabilidade do teste  $F$ , permitindo uma avaliação imediata da rejeição ou não da hipótese  $\mathbf{b}_0 = \mathbf{0}$ . A rejeição dá-se quando o valor da probabilidade é pequeno, correspondendo a valores elevados de  $F$ .

### 7. Testes e intervalos de confiança

Os exercícios de validade (testes e intervalos de confiança) que se apresentam a seguir permitem ter uma ideia indirecta da qualidade da regressão. Para além de uma validação geral do modelo obtido, os testes podem servir para confirmar hipóteses de valores particulares para os parâmetros, estabelecidas por via teórica ou em anteriores experiências.

As versões habituais baseiam-se na distribuição normal e, em alguns casos, do  $\chi^2$ , aqui substituídas respectivamente pelas distribuições t de Student e F, dado que a variância  $\sigma$  é sempre estimada.

Esclareça-se, também, que todas as expressões que se seguem apenas são válidas se se verificar o pressuposto de normalidade dos erros, para além das outras condições de aplicabilidade.

### 7.1. Distribuições de probabilidade dos parâmetros

Os parâmetros do modelo linear apresentam distribuições normais, com as seguintes características:

$$\begin{aligned}\hat{a} &\sim N(a, c_{00} \cdot \sigma^2) \\ \hat{b}_k &\sim N(b_k, c_{kk} \cdot \sigma^2)\end{aligned}$$

### 7.2. Correlações e Matriz Covariância

Definindo a matriz  $\mathbf{C} = (\mathbf{X}'_a \cdot \mathbf{X}_a)^{-1}$ , na qual o índice 0 corresponde ao parâmetro  $a$ , respeitando os restantes índices aos parâmetros  $b_1 \dots b_p$

$$\mathbf{C} = (\mathbf{X}'_a \cdot \mathbf{X}_a)^{-1} = \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0p} \\ c_{10} & c_{11} & \cdots & c_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p0} & c_{p1} & \cdots & c_{pp} \end{bmatrix}$$

podem calcular-se com facilidade as correlações entre parâmetros, a partir das regras:

$$\text{corr}(\hat{a}, \hat{b}_k) = \frac{c_{0k}}{\sqrt{c_{00} \cdot c_{kk}}} \quad \text{e} \quad \text{corr}(\hat{b}_j, \hat{b}_k) = \frac{c_{jk}}{\sqrt{c_{jj} \cdot c_{kk}}}$$

Relacionada com  $\mathbf{C}$  está a matriz covariância dos parâmetros,  $\Sigma_b = \sigma^2 \cdot \mathbf{C}$ . A diagonal principal de  $\Sigma_b$  é constituída pelas variâncias dos parâmetros, a partir das quais se calculam os desvios padrão usados nos testes de hipóteses e no cálculo de intervalos de confiança:  $\sigma\sqrt{c_{00}}$  para  $a$ , e  $\sigma\sqrt{c_{kk}}$  para cada um dos  $b_k$ . Os elementos fora da diagonal correspondem às covariâncias entre parâmetros. Como habitualmente se desconhece  $\sigma$ , este valor é substituído por  $\hat{\sigma}$ , o que permite obter uma estimativa não tendenciosa de  $\Sigma_b$ .

### 7.3. Testes de hipóteses

Os testes indicados a seguir seguem, em geral, o princípio habitual de propor uma hipótese nula, uma hipótese alternativa e uma regra de rejeição, para um certo nível de significância (tipicamente 5%, embora possam ser usados outros valores). Isto significa que a probabilidade do teste rejeitar uma hipótese nula que fosse verdadeira (erro tipo I) é inferior a 0.05, mas não nos diz nada sobre o erro complementar (erro tipo II) ou seja, não rejeitar a hipótese nula, sendo esta falsa. Os resultados positivos dos testes devem ser, portanto, utilizados com prudência, a menos que se possua uma estimativa da probabilidade do erro do tipo II.

#### 7.3.1. O valor do parâmetro $b_k$ é igual a $b_x$ ?

Este teste permite excluir ou não a hipótese do verdadeiro valor de  $b_k$  ser um certo valor  $b_x$ , por exemplo um valor teórico que se pretende confirmar, ou então o valor nulo, correspondente a não incluir a variável  $X_k$  no modelo. Claro que o teste não serve para verificar se o parâmetro tem exactamente o valor estimado, pois  $t=0$  e a hipótese nula nunca seria rejeitada.

$H_0: b_k = b_x$	$H_a: b_k \neq b_x$	$t = \frac{\hat{b}_k - b_x}{\hat{\sigma} \sqrt{c_{kk}}}$
Rejeição de $H_0$ se $ t  >  t_{1-\gamma/2}(n-p-1) $		

O teste anterior também pode ser aplicado ao parâmetro  $a$ , com as alterações evidentes (mesmos graus de liberdade).

#### 7.3.2. Os coeficientes $b_0$ são todos simultaneamente nulos?

Permite uma verificação genérica da adequação do modelo, neste caso pela rejeição da hipótese nula. Quanto maior é o valor calculado de  $F$ , mais fácil é aquela rejeição, por ser mais pequeno o valor de  $\gamma$  para a qual o valor da tabela é menor ou igual ao valor calculado de  $F$ .

$H_0: b_0 = 0$	$H_a: b_0 \neq 0$	$F = \frac{MSR}{MSE}$
Rejeição de $H_0$ se $\gamma < \gamma_0$ , sendo $F_{1-\gamma}(p, n-p-1) \leq F$		

O valor limite  $\gamma_0$  a utilizar depende das circunstâncias (0.05 ou 0.10), mas pode ir até 0.25, numa opção cautelosa (no sentido de manter o modelo) que torna mais difícil não rejeitar a

hipótese nula.

#### 7.4. Intervalos de confiança dos parâmetros

Os intervalos de confiança indicados a seguir são válidos apenas para parâmetros considerados individualmente. Se se pretendesse considerar simultaneamente vários parâmetros, teriam que ser usadas distribuições de probabilidade conjuntas. Como é óbvio, os intervalos serão tanto mais apertados quanto menor for o nível de confiança  $100.(1-\gamma)\%$ . Repare-se que os intervalos de confiança podem funcionar como teste de hipóteses: se o intervalo contém a hipótese nula, esta não é rejeitada. Neste caso, o valor de  $\gamma$  funciona como nível de significância.

##### 7.4.1. Intervalo de $a$

$$\hat{a} - \Delta_a \leq a \leq \hat{a} + \Delta_a \quad \text{onde} \quad \Delta_a = \hat{\sigma} \sqrt{c_{00}} \cdot t_{1-\gamma/2}(n-p-1)$$

##### 7.4.2. Intervalos dos $\mathbf{b}_0$

$$\hat{b}_k - \Delta_{b_k} \leq b_k \leq \hat{b}_k + \Delta_{b_k} \quad \text{onde} \quad \Delta_{b_k} = \hat{\sigma} \cdot \sqrt{c_{kk}} \cdot t_{1-\gamma/2}(n-p-1)$$

O uso combinado destes intervalos dá uma ideia otimista do conjunto dos  $\mathbf{b}_0$ . Uma alternativa ao uso de distribuições conjuntas de probabilidade será usar a expressão:

$$(\hat{\mathbf{b}}_0 - \mathbf{b}_0)' \mathbf{M}' \mathbf{M} (\hat{\mathbf{b}}_0 - \mathbf{b}_0) \leq p \hat{\sigma}^2 \cdot F_{1-\gamma}(p, n-p-1)$$

que define a região de  $100.(1-\gamma)\%$  confiança (em geral um elipsoide) para o conjunto dos  $\mathbf{b}_0$ . A expressão é mais facilmente utilizada para verificar se um conjunto particular de valores está ou não incluído na região de confiança.

#### 7.5. Intervalo da resposta

Uma vez estabelecidos os parâmetros do modelo, é possível, como se disse inicialmente, estimar o valor de  $Y$  correspondente a uma dada observação das variáveis  $X_k$ . Se designarmos por  $\mathbf{u} = [1 \ u_1 \ u_2 \ \dots \ u_p]'$  o vector alargado das observações das  $p$  variáveis, a estimativa de  $y$  será dada por:

$$\hat{y} = \mathbf{u}' \cdot \hat{\mathbf{b}}$$

Se os erros tiverem distribuição normal, também  $\hat{Y} \sim N(E[Y], \text{Var}[\hat{Y}])$ . Uma vez que a variância é estimada, o intervalo de  $100.(1-\gamma)\%$  de confiança para  $E[Y] = E[\hat{Y}]$  será dado por:

$$\hat{Y} - \Delta_{E[Y]} \leq E[Y] \leq \hat{Y} + \Delta_{E[Y]} \quad \text{onde} \quad \Delta_{E[Y]} = t_{1-\gamma/2}(n-p-1) \sqrt{\mathbf{u}' \cdot \Sigma_b \cdot \mathbf{u}}$$

## 7.6. Intervalo de predição

O intervalo de  $100 \cdot (1-\gamma)\%$  de confiança para futuras leituras de  $Y$  é dado, em função das observações  $\mathbf{u}$ , por:

$$\hat{Y} - \Delta_Y \leq Y \leq \hat{Y} + \Delta_Y \quad \text{onde} \quad \Delta_Y = t_{1-\gamma/2}(n-p-1) \sqrt{\sigma^2 + \mathbf{u}' \cdot \Sigma_b \cdot \mathbf{u}}$$

A extrapolação da região onde foram obtidos os valores de  $\mathbf{X}$  e  $\mathbf{y}$  que levaram à estimação dos parâmetros conduz, normalmente, a um aumento substancial da largura do intervalo, pelo que tal exercício deve ser realizado com prudência.

## 8. Análise de resíduos

De acordo com os pressupostos da regressão, os resíduos devem distribuir-se aleatoriamente em torno de 0, tanto no modelo global como em relação a cada variável. Caso tal não se verifique, será normalmente necessário alterar o modelo, incluindo ou retirando variáveis, ou realizando alguma transformação que adeque melhor o modelo aos dados (por exemplo  $X_k^2$  em vez de  $X_k$ ).

Para além dos resíduos correspondentes directamente aos erros do modelo,  $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ , é usual calcular também os resíduos estandardizados,  $\mathbf{s} = (1/\hat{\sigma}) \cdot \mathbf{r}$  (que, apesar do nome, não têm uma distribuição normal standard, porque  $\hat{\sigma}$  não é a variância individual de cada resíduo). Para o cálculo de outros tipos de resíduos convém introduzir a matriz simétrica  $\mathbf{H} = \mathbf{X}_a \cdot \mathbf{C} \cdot \mathbf{X}'_a$ , na qual  $0 \leq h_{ii} \leq 1$  e  $-1 \leq h_{ik} \leq 1$  ( $i \neq k$ ). Repare-se que  $\hat{\mathbf{y}} = \mathbf{H} \cdot \mathbf{y}$ .

Podem agora calcular-se os resíduos "Student"  $t_i$  e resíduos de eliminação  $r_{(-i)}$  (resíduos que se obteriam estimando  $y_i$  sem incluir a observação correspondente no cálculo dos parâmetros. Ter-se-á, então:

$$t_i = \frac{r_i}{\hat{\sigma}_{r_i}} = \frac{r_i}{\hat{\sigma} \sqrt{1 - h_{ii}}} \quad \text{e} \quad r_{(-i)} = \frac{r_i}{1 - h_{ii}}$$

Repare-se, no denominador da primeira expressão, que a variância individual de cada resíduo é igual a  $(1-h_{ii}) \cdot \sigma^2$ . Os resíduos "Student", apesar do nome, distribuem-se mais perto da distribuição normal do que da  $t$  de Student ( $n-p-1$  graus de liberdade). É possível, ainda, definir resíduos "Student" de eliminação  $t_{(-i)}$ , que seguem exactamente uma distribuição  $t$  com  $n-p-2$  graus de liberdade, através de:

$$t_{(-i)} = r_i \cdot \sqrt{\frac{n-p-2}{(1-h_{ii})SSE - r_i^2}}$$

### 8.1. Eliminação de observações

Valores elevados de um determinado resíduo (sob qualquer das formas) aconselham uma inspecção cuidadosa da observação correspondente, com vista à sua eventual eliminação. mais formalmente, se o valor de um ou mais  $t_{(-i)}$  corresponder a uma probabilidade pequena na tabela da distribuição com  $n-p-2$  graus de liberdade, os pontos em causa poderão estar muito fora da regressão, podendo justificar-se a sua eliminação, sobretudo se houver razões físicas que ponham em causa as observações ou leituras correspondentes.

Certos traçados gráficos também podem ser utilizados na referida detecção. Por exemplo:

- Histogramas de resíduos "Student". Possível eliminação dos pontos que estejam para lá de três desvios padrão, na distribuição (aproximadamente normal) destes resíduos;
- Resíduos em função das respostas ou em função de variáveis. Permitem uma detecção visual qualitativa de situações a investigar;
- Resíduos em função de resíduos de eliminação. Os pontos "normais" deverão estar sobre uma linha recta de inclinação  $1$ , que passa pela origem, ou seja, a eliminação da observação respectiva não faz variar sensivelmente os resíduos.

É possível definir, também, testes estatísticos aproximados para detecção de isolados. No entanto, as decisões de eliminação devem ser sempre tomadas com muita prudência, pois correspondem a uma diminuição do volume inicial de dados. Eliminações apressadas são facilmente sujeitas a crítica.

### 8.2. Verificação de pressupostos

Apresentam-se, a seguir, alguns testes que permitem verificar se os pressupostos em relação aos erros do modelo são verificados pelos resíduos. Trata-se de verificações *a posteriori* que poderão levar à revisão do modelo.

#### 8.2.1. Aleatoriedade

Uma forma corrente de verificar a aleatoriedade dos resíduos é o teste às sequências de sinais dos resíduos, através do "runs test" (teste de corridas), importante sobretudo quando as observações dependem do tempo. Considerando apenas os sinais (+ ou -) dos resíduos, pela ordem em que foram recolhidos, haverá  $n_1$  sinais (+),  $n_2$  sinais (-) e  $r$  corridas (sequências máximas de sinais iguais seguidos). Na sequência (+ - - + + + - - - + + -), por

exemplo, será  $n_1=7$ ,  $n_2=6$  e  $r=6$ . Usando em seguida tabelas para o "runs test", determinam-se valores críticos que ajudam a determinar, com nível de significância 5%, se a sequência é ou não aleatória. Em função de  $n_1$  e  $n_2$ , as tabelas dão dois valores (inferior e superior) que terão que enquadrar o valor de  $r$ . Caso contrário, suspeita-se de não-aleatoriedade. No caso do exemplo, os dois valores são 3 e 12, concluindo-se pela aleatoriedade, uma vez que  $3 \leq r \leq 12$ .

As tabelas referidas para este teste só abrangem, geralmente, até um máximo de 20 para  $n_1$  ou  $n_2$ . Para valores superiores, usa-se a distribuição normal da forma habitual nos testes, com

$$Z = \frac{r - \frac{2n_1n_2}{n_1 + n_2} - \frac{1}{2}}{\sqrt{\frac{2n_1n_2 \cdot (2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2 \cdot (n_1 + n_2 - 1)}}$$

### 8.2.2. Correlação sucessiva

A verificação de independência é usualmente feita através do teste de Durbin-Watson à correlação entre resíduos sucessivos. O teste é útil sobretudo em dados dependentes do tempo. A partir de  $\varepsilon_i = \rho \cdot \varepsilon_{i-1} + \delta_i$ , onde os  $\delta_i \sim N(0, \sigma^2)$ , a estatística a usar é:

$$d = \frac{\sum_{i=2}^n (r_i - r_{i-1})^2}{\sum_{i=1}^n r_i^2}$$

O teste  $H_0: \rho=0$ ,  $H_a: \rho>0$  baseia-se em tabelas próprias, que fornecem dois índices  $d_L$  e  $d_U$ , rejeitando-se  $H_0$  se  $d < d_L$ , e não se rejeitando se  $d > d_U$  (dentro do intervalo não se podem retirar conclusões). Para testar  $H_0$  contra  $H_a: \rho < 0$ , usa-se a estatística  $d' = 4 - d$ , com as mesmas tabelas.

### 8.2.3. Heteroscedaticidade

A detecção de desigualdades de variância dos erros pode ser realizada a partir de um gráfico dos resíduos  $r_i$  em função dos  $\hat{y}_i$ . Se o aspecto não for uma mancha de largura uniforme, por exemplo alargando com o aumento de  $\hat{y}_i$ , poderá ser necessário transformar  $Y$  ( $\ln Y$ ,  $1/Y$ , etc) ou alterar o modelo. Um gráfico semelhante, mas dos quadrados dos resíduos, pode confirmar suspeitas e ajudar a detectar isolados.

#### 8.2.4. Normalidade

A verificação visual da normalidade é feita ordenando os resíduos de forma crescente, e desenhando-os em papel de distribuição normal. Se a presunção de normalidade se verificar, os resíduos deverão estar aproximadamente em linha recta.

### 8.3. Expressão do modelo

São úteis alguns gráficos de resíduos em relação a variáveis, para verificação visual da correcção da expressão do modelo. Os gráficos potencialmente mais interessantes são:

- Resíduos em função das variáveis. Permitem verificar se é necessário transformar as variáveis ( $\ln X$ ,  $\sqrt{X}$ , etc.);
- Resíduos em função de produtos de variáveis. No caso de ser detectado um padrão, deve ser incluído no modelo um novo termo com o produto em causa ( $X_u \cdot X_v$ , por exemplo);
- Resíduos parciais. Gráfico dos resíduos obtidos sem incluir  $X_k$ , em função de  $X_k$ . Permitem detectar não-linearidades que levem à transformação de  $X_k$ . Se o ajuste for bom, o gráfico tem o aspecto de uma recta com inclinação igual ao parâmetro da variável na regressão.

## 9. Selecção de variáveis

As técnicas apresentadas a seguir podem ser úteis para confirmar ou afastar hipóteses acerca da inclusão de variáveis explicativas, produzidas a partir do conhecimento do sistema e do seu comportamento. Essa selecção primária de variáveis é, portanto, fundamental para o eventual sucesso do exercício de regressão. Note-se, também, que todas as técnicas se dirigem a variáveis numa determinada forma, ou seja, a rejeição de  $X_j$  não significa que  $1/X_j$ , por exemplo, não devesse ser incluída no modelo.

### 9.1. Medida $F$ e estatística $C_k$

Decisão sobre a inclusão ou não de um conjunto de  $r$  variáveis, cujas observações estão agrupadas numa matriz  $\mathbf{X}_b$ , correspondendo aos parâmetros  $\mathbf{b}_2$ . Podem calcular-se os valores de  $SSR$  e  $SSE$  do modelo contendo estas variáveis, comparando-os com os valores  $SSR_1$  e  $SSE_1$  que se obteriam com o modelo reduzido (sem as  $r$  variáveis em causa). O teste a realizar usa a estatística  $F$ , com  $\gamma_0$  típico da ordem de 0.05:

$\mathbf{H}_0: \mathbf{b}_2=0$	$\mathbf{H}_a: \mathbf{b}_2 \neq 0$	$F = \frac{SSR - SSR_1}{r.MSE}$
Rejeição de $H_0$ se $\gamma < \gamma_0$ , sendo $F_{1-\gamma}(r, n-p-1) \leq F$		

Ou seja, valores elevados de  $F$  conduzem à rejeição da hipótese  $\mathbf{b}_2=\mathbf{0}$ , e as variáveis em causa são mantidas no modelo.

Uma alternativa ao uso de  $F$  é o cálculo de

$$c_k = \frac{SSE_l}{MSE} - (n - 2k)$$

onde  $k$  é o número de variáveis do modelo reduzido (sem as  $r$  variáveis em causa). Se o valor de  $c_k$  for muito superior a  $k$ , deve suspeitar-se que algumas variáveis importantes serão rejeitadas, caso se opte pelo modelo reduzido. Devem procurar-se, portanto, subconjuntos de variáveis que conduzam a valores de  $c_k$  próximos de  $k$ .

Menos formalmente, podem comparar-se os valores de  $MSE$  e  $R^2$  para os modelos "completo" e reduzido. Se não diferirem muito, será mais económico usar o modelo reduzido.

## 9.2. Pesquisa t-dirigida

Este método parte do modelo completo, calculando-se as estatísticas  $t$  correspondentes à eliminação de cada uma das variáveis  $X_k$ , através de

$$t_k = \frac{\hat{\beta}_k}{\hat{\sigma} \cdot \sqrt{c_{kk}}}$$

Um critério habitual é conservar todas as variáveis para as quais  $|t| > 3$ . Depois desta selecção inicial, que permite limitar muito o número de regressões alternativas a experimentar, é investigada a inclusão de cada uma das restantes variáveis, com recurso, por exemplo, às técnicas da secção anterior. Para além do esquema aqui indicado, podem ser utilizados diferentes procedimentos com base no mesmo princípio.

## 9.3. Selecção para a frente (*forward selection*)

Neste método, as variáveis candidatas  $X_1 \dots X_p$  vão sendo introduzidas progressivamente no modelo, com base na comparação das somas dos quadrados dos resíduos ( $SSE$ ) que resultam da sua introdução. Não garantindo a descoberta do melhor subconjunto de variáveis, o método é fácil de usar e permite obter, em geral, resultados bastante bons, com muito menos esforço do que ensaiar todas as possíveis regressões. O procedimento geral é o seguinte:

0. Escolher  $X_u$  que conduz ao menor valor de SSE da regressão  $y=a+b_uX_u$   
*Repetir*
1. Escolher  $X_v$  que conduz ao menor valor de SSE da regressão  $y=a+b_uX_u+b_vX_v$   
*até SSE não diminuir mais ou até estarem incluídos todos os termos.*

A decisão sobre paragem pode ser baseada num teste com uma estatística próxima de  $F$ , calculada para cada uma das variáveis candidatas no estágio  $(s+1)$ , quando há  $s$  variáveis anteriormente seleccionadas. A expressão para uma variável candidata  $X_k$  será:

$$F_k = \frac{SSE_s - SSE_{s+1}^k}{MSE_{s+1}^k} = \frac{SSE_s - SSE_{s+1}^k}{\frac{SSE_{s+1}^k}{n-s-2}}$$

A variável a entrar no modelo será a que tiver maior valor de  $F_k$ , mas só será adicionada se se verificar a condição

$$\max_k F_k = F_{max} \geq F_{1-\gamma}(1, n-s-2)$$

sendo usual fixar-se um valor pouco exigente para  $\gamma$  (p.ex. 0.25). Se a condição não for satisfeita, o procedimento pára. No limite, será necessário calcular  $p!$  regressões, com um número de parâmetros crescendo de 2 até  $p+1$  ao longo dos estágios.

#### 9.4. Eliminação para trás (*backward elimination*)

Filosofia complementar da anterior. O processo inicia-se com a regressão completa (todas as  $p$  variáveis) e em cada estágio é eliminada a variável cuja saída do modelo conduz à regressão reduzida com menor SSE. A eliminação pode ser feita usando a estatística:

$$F_k = \frac{SSE_{s+1}^k - SSE_s}{MSE}$$

onde  $MSE$  é sempre o da regressão completa, e  $SSE_{s+1}^k$  não obriga a executar a regressão reduzida (sem  $X_k$ ), toda a vez que

$$SSE_{s+1}^k - SSE_s = t_k^2 \cdot MSE_s$$

onde  $t_k$  é a estatística usada para o teste de  $b_k=b_x$  (ver testes), neste caso com  $b_x=0$ . Em consequência, só é necessário calcular uma regressão em cada estágio, o que torna este processo bastante económico.

Depois de seleccionada a variável com o menor  $F_k$ , a variável é eliminada se for verificada a condição:

$$\min_k F_k = F_{min} \leq F_{1-\gamma}(1, n - p - 1)$$

Caso contrário, o processo termina, e mais nenhuma variável é eliminada. Tal como no caso anterior, este método também não garante a melhor selecção, mas comporta-se bastante bem, sendo, em princípio, preferível quando o número de variáveis não é excepcionalmente grande.

## 9.5. Procedimento passo a passo

A combinação dos dois procedimentos anteriores conduz a uma estratégia algo mais complexa, mas que conduz a melhores resultados. O princípio operacional é semelhante ao da selecção para a frente, mas em cada estágio realiza-se um passo de eliminação para trás, que pode conduzir à manutenção de todas as variáveis ou à eliminação de uma delas, de acordo com o teste exposto na secção anterior. A regra de paragem é igual à do método de selecção para a frente.

## 10. Multicolinearidade

Se existir dependência linear entre pelo menos dois vectores  $\mathbf{x}_u$  e  $\mathbf{x}_v$ , o processo de regressão não é possível tecnicamente, dado que  $\mathbf{X}'_a \cdot \mathbf{X}_a$  é singular. Evidentemente que, em tal caso, que corresponde a redundância na informação, a eliminação de variáveis resolve o problema. Sucede, no entanto, que podem surgir situações de dependência linear *aproximada*, ou seja, existe pelo menos um  $\mathbf{c} \neq \mathbf{0}$  para o qual  $\mathbf{X} \cdot \mathbf{c} \approx \mathbf{0}$ . Esta situação designa-se por multicolinearidade e tem efeitos nocivos nos modelos, nomeadamente pela perturbação da ligação entre os fenómenos estudados e os valores matemáticos dos parâmetros.

Podem surgir, por exemplo, modelos alternativos de qualidade de ajuste semelhante, mas com valores completamente díspares (até no sinal) nos parâmetros das mesmas variáveis. Os valores das estatísticas usadas nos testes tendem a baixar, podendo levar à eliminação de variáveis importantes não-colineares (como remédio, sugere-se subir o nível de confiança para 0.25). A extrapolação pode, na situação geral de multicolinearidade, ser desastrosa.

Uma vez detectada a multicolinearidade, podem seguir-se duas estratégias: eliminação de variáveis redundantes, ou regressões tendenciosas que eliminam os efeitos da redundância, sem eliminar variáveis. Dada a especialização deste último tópico, apenas se descreverá, no presente texto, um desses modelos, baseado na análise de componentes principais.

## 10.1. Detecção

As multicolinearidades estão relacionadas com a correlação entre variáveis, podendo ser detectadas na matriz de correlação dada pelo produto  $\mathbf{W}'\mathbf{W}$ . No entanto, o fenómeno pode não ser evidente na matriz, sobretudo quando são envolvidas mais do que duas variáveis. A análise dos valores próprios de  $\mathbf{W}'\mathbf{W}$  próximos de zero permite uma detecção mais eficaz, à custa dos vectores próprios correspondentes. Na verdade, se for  $\mathbf{v}_k$  um desses vectores próprios, correspondendo ao valor próprio  $\lambda_k \approx 0$ , pode mostrar-se que:

$$\mathbf{W}\mathbf{v}_k \approx \mathbf{0}$$

Os elementos de  $\mathbf{v}_k$  correspondem aproximadamente, portanto, aos coeficientes de uma combinação linear "quase nula", ressaltando aqueles que tiverem maior valor absoluto, por indicarem as variáveis multicolineares.

A detecção também pode recorrer à matriz  $\mathbf{Q}=(\mathbf{W}'\mathbf{W})^{-1}$ , nomeadamente aos elementos da diagonal principal ( $q_{kk}$ ), designados por VIF ("variance inflation factors" - factores de aumento da variância). Valores elevados de  $q_{kk}$  sugerem que a variável  $X_k$  está envolvida em multicolinearidades, podendo detectar-se as relações com outras variáveis  $j$  a partir de valores elevados de  $q_{kj}$ .

## 10.2. Regressão de componentes principais

A eliminação de variáveis multicolineares é um exercício sempre arriscado, dado que muitas vezes não são claros os limites aceitáveis. Por outro lado, a situação de multicolinearidade não deve ignorar-se, pois os indicadores de qualidade (como  $R^2$ ) e os testes são afectados, podendo levar a tomar decisões erradas sobre variáveis importantes para o modelo, etc.

Uma alternativa às duas opções anteriores é a utilização de regressões tendenciosas, ou seja, aceita-se que  $E[\hat{\mathbf{b}}_0] \neq \mathbf{b}_0$ , em troca de uma grande redução da variância dos parâmetros (ou seja, dos VIF). Os resultados "visíveis", ao nível da qualidade do ajuste, são semelhantes aos que se obteriam com a regressão normal, mas a supressão das multicolinearidades faz com que os parâmetros reflectam mais correctamente a importância relativa das diversas variáveis explicativas.

Das várias hipóteses existentes, apresenta-se aqui a regressão de componentes principais. Outras variantes podem ser vistas na bibliografia indicada. Chama-se a atenção para que, ao contrário do restante texto, se utiliza nesta secção a redução para norma unitária.

### 10.2.1. Princípio

O princípio da regressão de componentes principais consiste em eliminar os vectores próprios de  $\mathbf{W}'\mathbf{W}$  que correspondam a valores próprios próximos de zero. Partindo da igualdade:

$$\mathbf{W}'\mathbf{W} = \sum_{k=1}^p \frac{1}{\lambda_k} \mathbf{v}_k \cdot \mathbf{v}_k'$$

e admitindo que os valores próprios a ignorar são os  $s$  primeiros, define-se uma nova matriz:

$$(\mathbf{W}'\mathbf{W})^+ = \sum_{k=s+1}^p \frac{1}{\lambda_k} \mathbf{v}_k \cdot \mathbf{v}_k'$$

As estimativas dos parâmetros da regressão serão então (em termos das variáveis reduzidas para norma unitária):

$$\hat{a}^* = \bar{Y}$$

$$\hat{\mathbf{b}}_0^* = (\mathbf{W}'\mathbf{W})^+ \cdot \mathbf{W}'\mathbf{y}$$

Tal como referido anteriormente, os parâmetros usuais  $a$  e  $\mathbf{b}_0$  podem ser obtidos a partir destes, através de:

$$\hat{b}_k = \frac{\hat{b}_k^*}{d_k} \quad \text{e} \quad \hat{a} = \bar{Y} - \sum_{k=1}^p \hat{b}_k \cdot \bar{X}_k$$

### 10.2.2. Componentes principais

As decisões de eliminação baseadas simplesmente na proximidade de zero dos valores próprios podem não ser fáceis de tomar. Para facilitar essa tarefa, é possível exprimir a regressão directamente a partir das componentes principais, e aplicar testes estatísticos semelhantes aos do caso geral.

Partindo de  $\mathbf{U}=\mathbf{W}\mathbf{V}=\mathbf{W}\cdot[\mathbf{v}_1 \mathbf{v}_2 \dots \mathbf{v}_p]$ , e admitindo, como anteriormente, que os  $s$  primeiros valores próprios de  $\mathbf{W}'\mathbf{W}$  são próximos de zero, fica definida a matriz de componentes principais de  $\mathbf{W}$ , dada por  $\mathbf{U}_P=[\mathbf{u}_{s+1} \mathbf{u}_{s+2} \dots \mathbf{u}_p]$ , que reúne as  $p-s$  últimas colunas de  $\mathbf{U}$ , e uma matriz  $\mathbf{U}_E=[\mathbf{u}_1 \mathbf{u}_2 \dots \mathbf{u}_s]$  de componentes eliminados de  $\mathbf{W}$ . Ou seja,  $\mathbf{U}=[\mathbf{U}_E \mathbf{U}_P]$ .

É possível, então, escrever a expressão da regressão em termos dos componentes de  $\mathbf{W}$ , com parâmetros  $\mathbf{c}_P$  e  $\mathbf{c}_E$  correspondentes às componentes principais e às componentes a eliminar:

$$\mathbf{y} = \mathbf{a}^* \cdot \mathbf{1} + \mathbf{U}_E \cdot \mathbf{c}_E + \mathbf{U}_P \cdot \mathbf{c}_P + \mathbf{e}$$

onde novamente  $\hat{a}^* = \bar{Y}$ . Dada a ortogonalidade de  $\mathbf{U}_E$  e  $\mathbf{U}_P$ , a expressão dos estimadores de  $\mathbf{c}_E$  e  $\mathbf{c}_P$  pode ser obtida separadamente:

$$\hat{\mathbf{c}}_E = (\mathbf{U}'_E \cdot \mathbf{U}_E)^{-1} \cdot \mathbf{U}'_E \cdot \mathbf{y}$$

$$\hat{\mathbf{c}}_P = (\mathbf{U}'_P \cdot \mathbf{U}_P)^{-1} \cdot \mathbf{U}'_P \cdot \mathbf{y}$$

sendo de notar a relação  $\hat{\mathbf{b}}_0^* = [\mathbf{v}_{s+1} \dots \mathbf{v}_p] \hat{\mathbf{c}}_P$ .

Em face de uma hipótese de eliminação de componentes, pode construir-se um quadro de ANOVA e tirar conclusões sobre a eliminação, através dos testes com a estatística  $F$ .

**Tabela 4: Quadro de ANOVA (componentes principais)**

	Graus de liberdade	Soma dos quadrados	Média dos quadrados	F	$R^2$
Média	1	SSM	MSM	MSM/MSE	SSR <sub>P</sub> /TSS(adj)
Regressão					
Comp P	p-s	SSR <sub>P</sub>	MSR <sub>P</sub>	MSR <sub>P</sub> /MSE	
Comp E	s	SSR <sub>E</sub>	MSR <sub>E</sub>	MSR <sub>E</sub> /MSE	
Erro	n-p-1	SSE	MSE		
Total	n	TSS			

O valor de  $R^2$  não fica, neste caso, obrigatoriamente no intervalo  $[0, 1]$ , nem será exactamente igual à correlação entre  $\mathbf{y}$  e  $\hat{\mathbf{y}}$ . Quanto aos testes, a hipótese de eliminação das componentes E não será de rejeitar se o valor de  $F$  for suficientemente pequeno (tal como no caso geral). Uma vez que a regressão (completa) nas componentes de  $\mathbf{W}$  não difere de uma regressão normal, também podem ser usados os testes descritos em secções anteriores para a inclusão de variáveis, aplicados aqui à inclusão de componentes.

A análise da versão final da regressão de componentes principais pode ser feita com o quadro de ANOVA anterior, ou alternativamente considerando as parcelas dos componentes eliminados incluídas no erro ( $SSE_p = SSE + SSE_E$ ). O quadro correspondente será:

**Tabela 5: Quadro final de ANOVA (componentes principais)**

	Graus de liberdade	Soma dos quadrados	Média dos quadrados	F	$R^2$
Média	1	SSM	MSM	MSM/MSE <sub>P</sub>	SSR <sub>P</sub> /TSS(adj)
Comp P	s	SSR <sub>P</sub>	MSR <sub>P</sub>	MSR <sub>P</sub> /MSE <sub>P</sub>	
Erro	n-p+s-1	SSE <sub>P</sub>	MSE <sub>P</sub>		
Total	n	TSS			

Os valores deste quadro devem ser usados com alguma prudência, tendo em conta que as estatísticas da penúltima coluna não são exactamente  $F$  excepto se  $\mathbf{c}_E = \mathbf{0}$ . Por outro lado, a estimativa de  $\sigma$  na tabela 5 é normalmente melhor do que a da tabela 4, sobretudo quando há poucos graus de liberdade em  $SSE$ .

## Bibliografia

Gunst, R.F., Mason, R.L. (1980), *Regression Analysis and Its Application: A Data-Oriented Approach*, Marcel Dekker, New York.

Marques de Sá, J.P. (1993), *Análise de Dados*, apontamentos para a disciplina de Análise de Dados, FEUP, Porto.

Gmurman, V.E. (1983), *Teoria das Probabilidades e Estatística Matemática*, Ed. Mir, Moscovo.

Taylor, J.R. (1982), *An Introduction to Error Analysis*, University Science Books, Mill Valley